ALIBABA CLOUD

Alibaba Cloud

Apsara Stack Enterprise

User Guide - Cloud Essentials and Security

Product Version: V3.15.0

Document Version: 20220526

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	**Danger:** Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	**Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	**Notice:** If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	**Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.Apsara Uni-manager Management Console

1.1. User Guide

1.1.1. What is the Apsara Uni-manager Management Console?

The Apsara Uni-manager Management Console is a service capability platform based on the Alibaba Cloud Apsara Stack platform and designed for government and enterprise customers. This platform improves IT management and troubleshooting and is dedicated to providing a leading service capability platform of the cloud computing industry. It provides large-scale and cost-efficient end-to-end cloud computing and big data services for customers in industries such as government, education, healthcare, finance, and enterprise.

Overview

The Apsara Uni-manager Management Console simplifies the management and deployment of physical and virtual resources by building an Apsara Stack platform that supports various business types of government and enterprise customers. The console helps you build your business systems in a simple and quick manner, fully improve resource utilization, and reduce O&M costs. This allows you to shift your focus from O&M to business. The console brings the Internet economy model to government and enterprise customers, and builds a new ecosystem chain based on cloud computing.

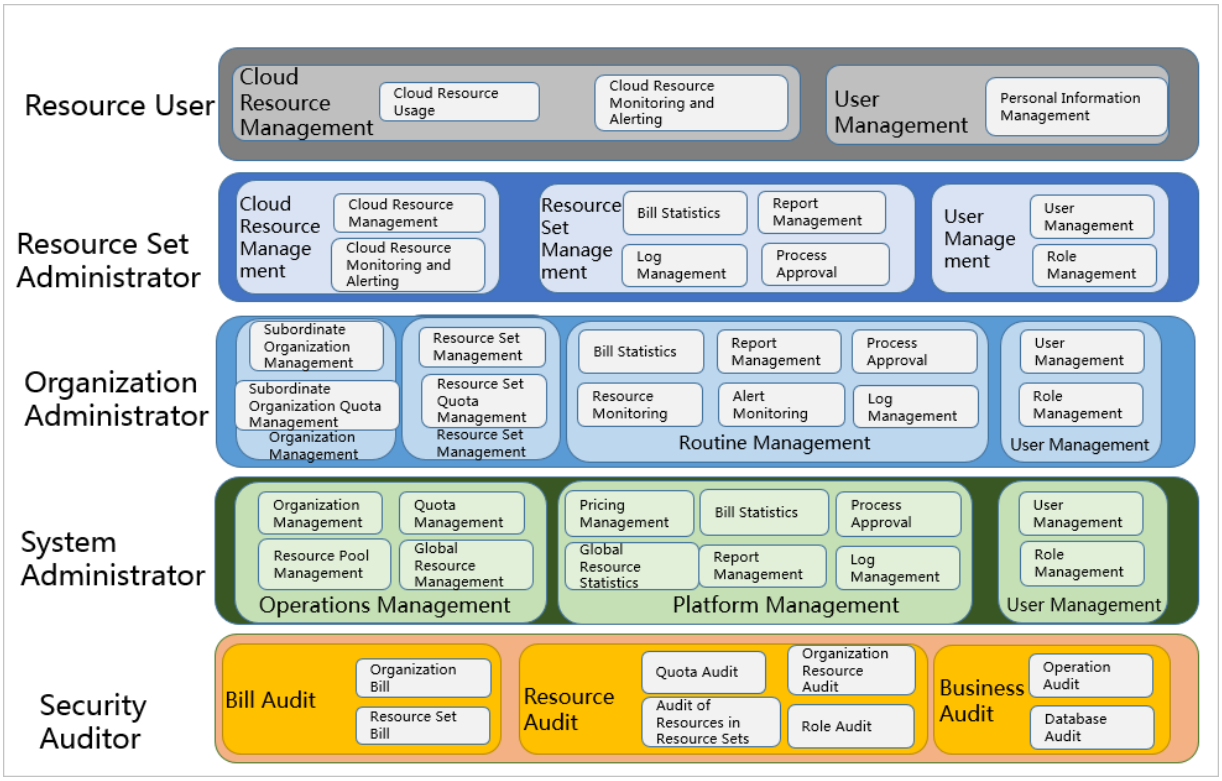
Workflow

Operations in the Apsara Uni-manager Management Console are divided into the following parts:

- System initialization: This part is designed to complete basic system configurations, such as creating organizations, resource sets, and users, creating basic resources such as VPCs, and creating contacts and contact groups in Cloud Monitor.
- Cloud resource creation: This part is designed to create resources.
- Cloud resource management: This part is designed to complete resource management operations, such as starting, using, and releasing resources, and changing resource configurations and resource quotas.

1.1.2. User roles and permissions

This topic describes roles and their permissions.



Roles and permissions

Role	Role permission
Resource user	This role has permissions to view and modify resources in a resource set and create alert rules.
Resource set administrator	This role has permissions to create, modify, and delete resources in a resource set and manage the users of the resource set.
Organization administrator	This role has permissions to manage an organization and its subordinate organizations, create, modify, and delete the resources of organizations, create and view alert rules for resources, and export reports.
Operations administrator	This role has read and write permissions on all resources.
Security auditor	This role performs security audits on the Apsara Uni-manager Management Console and has read-only permissions on the operation logs of the Apsara Uni-manager Management Console.
Platform administrator	This role has permissions to initialize the system and create operations administrators.
Resource auditor	This role has read-only permissions on all resources in the Apsara Uni-manager Management Console.
Organization security administrator	This role manages the security of an organization, including the security of hosts, applications, and networks. This role has read-only permissions on the operation logs of the Apsara Uni-manager Management Console and read and write permissions on ApsaraDB RDS, Elastic Compute Service (ECS), and Apsara Stack Security.

Role	Role permission
Security system configuration administrator	This role configures system security features such as the upgrade center and global configurations. This role has read and write permissions on the upgrade, protection, and configuration features of Apsara Stack Security.
Global organization security administrator	This role manages the security of global tenants and the platform by using Cloud Security Operation Center (SOC). This role has read and write permissions on RDS, ECS, and Apsara Stack Security.
Platform security administrator	This role manages the security of the Apsara Uni-manager Management Console by using SOC.
Global organization security auditor	This role checks the security conditions of all organizations by using SOC. This role has read-only permissions on the operation logs of the Apsara Uni-manager Management Console and all features of Apsara Stack Security.
Platform security auditor	This role checks the security conditions of the Apsara Uni-manager Management Console by using SOC. This role has read-only permissions on the operation logs of the Apsara Uni-manager Management Console, Server Guard, Cloud Firewall, Sensitive Data Discovery and Protection, SOC, system configurations, and Web Application Firewall (WAF) configurations. This role also has read and write permissions on Anti-DDoS, Threat Detection, and Update Center of Apsara Stack Security.
Platform security configuration administrator	This role configures security services in the Apsara Uni-manager Management Console, such as Server Guard and WAF, and has read and write permissions on these services.
Organization resource auditor	This role has read-only permissions on all resources in an organization to which it belongs.

1.1.3. Log on to the Apsara Uni-manager Management Console

This topic describes how to log on to the Apsara Uni-manager Management Console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

> **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
>
> - Uppercase or lowercase letters
> - Digits
> - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

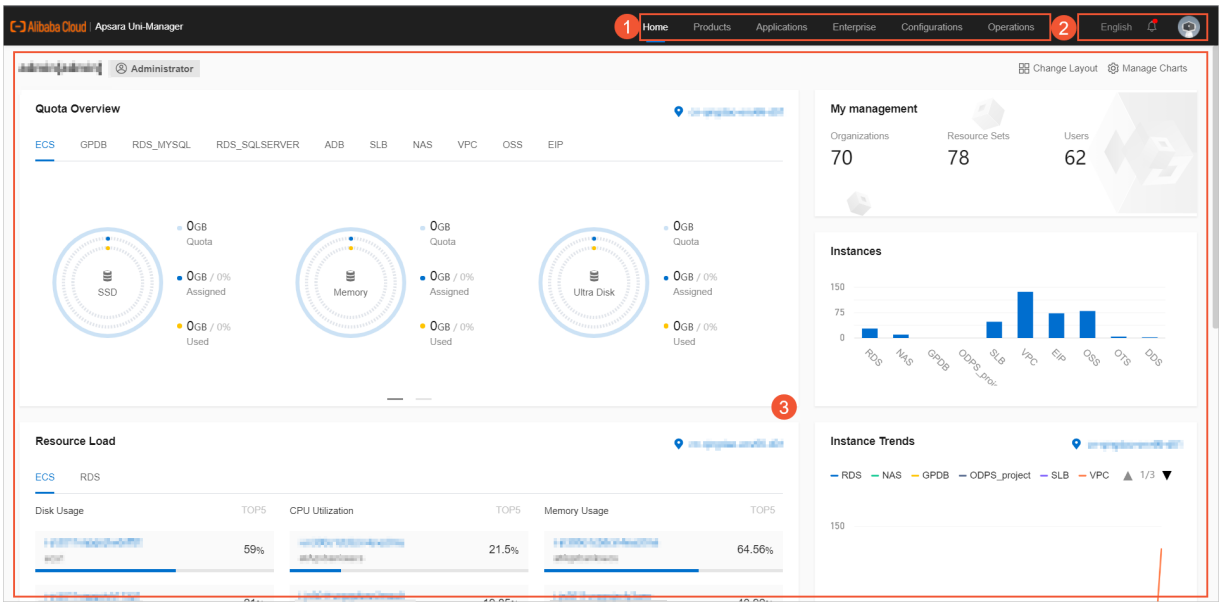
     Enter a six-digit MFA authentication code and click **Authenticate**.

> **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

1.1.4. Web page introduction

The web page of the Apsara Uni-manager Management Console consists of the search box, top navigation bar, information section of the current logon user, and operation section.

Apsara Uni-manager Management Console page



Functional sections of the web page

Section		Description
1	Top navigation bar	This section includes the following modules: • **Home**: uses charts to show the usage and monitoring data of existing system resources in each region. • **Products**: manages all types of basic cloud services and resources. • **Applications**: manages cloud applications of enterprises. • **Enterprise**: manages organizations, resource sets, roles, users, logon policies, user groups, ownership, and resource pools. • **Configurations**: manages resource pools, password policies, specifications, menus, and Resource Access Management (RAM) roles. • **Operations**: manages the routine operations of cloud resources, including usage statistics and quotas. • **Security**: provides operation and system logs.
2	Information section of the current logon user	• English: allows you to switch between English, simplified Chinese, and traditional Chinese. • : provides message notifications. • Personal information: When you click the icon of the current logon user, the **Personal information**, **View version information**, and **Log out** menu items are displayed. If you click **Personal information**, you can perform the following operations on the User Information page: ◦ View basic information. ◦ Modify personal information. ◦ Change the logon password. ◦ View the AccessKey pair of your Apsara Stack tenant account. ◦ Switch the current role. ◦ Enable or disable alert notification. If you click **View version information**, you can view the authorization status and build number of the Apsara Uni-manager Management Console.
3	Operation section	**Operation section**: shows the information and operations.

1.1.5. Initial configuration

1.1.5.1. Configuration description

Before you use the Apsara Uni-manager Management Console, you must complete a series of basic configuration operations as an administrator, such as creating organizations, resource sets, users, and roles and initializing resources. This is the initial system configuration.

The Apsara Uni-manager Management Console manages the organizations, resource sets, users, and roles of cloud data centers in a centralized and service-oriented manner to grant different resource access permissions to different users.

- Organization

After the Apsara Uni-manager Management Console is deployed, a root organization is automatically generated. You can create other organizations under the root organization.

Organizations are displayed in a hierarchical structure. You can create subordinate organizations under each organization level.

- Resource Set

  A resource set is a container used to store resources. Each resource must belong to a resource set.

- User

  A user is a resource manager and user.

- Role

  A role is a set of access permissions. You can assign different roles to different users to implement system access control to meet a variety of different requirements.

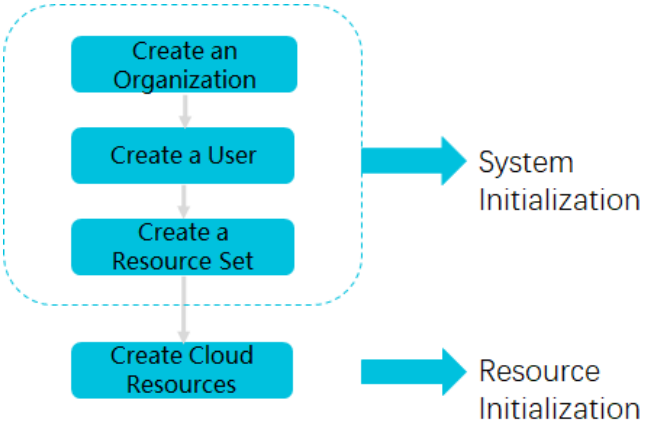
The following table describes the relationships among organizations, resource sets, users, roles, and cloud resources.

Relationship between two items	Relationship type	Description
Organization and resource set	One-to-many	An organization can have multiple resource sets, but each resource set can belong to only a single organization.
Organization and user	One-to-many	An organization can have multiple users, but each user can belong to only a single organization.
Resource set and user	Many-to-many	A user can have multiple resource sets, and a resource set can be assigned to multiple users under the same level-1 organization.
User and role	Many-to-many	A user can have multiple roles, and a role can be assigned to multiple users.
Resource set and resource	One-to-many	A resource set can have multiple resources, but each cloud resource can belong to only a single resource set.

1.1.5.2. Configuration process

This topic describes the initial configuration process.

Before you use the Apsara Uni-manager Management Console, you must complete the initial system configurations as an administrator based on the process shown in the following figure.



1. Create an organization

Create an organization to store resource sets and their resources.

2. Create a user

   Create a user and assign the user different roles to meet different requirements for system access control.

3. Create a resource set

   Create a resource set before you apply for resources.

4. Create cloud resources

   Create instances in each service console based on project requirements. For more information about how to create cloud service instances, see the user guide of each cloud service.

1.1.6. Monitoring

1.1.6.1. View the workbench

The Apsara Uni-manager Management Console uses charts to keep you up to date on the current usage and monitoring information of resources.

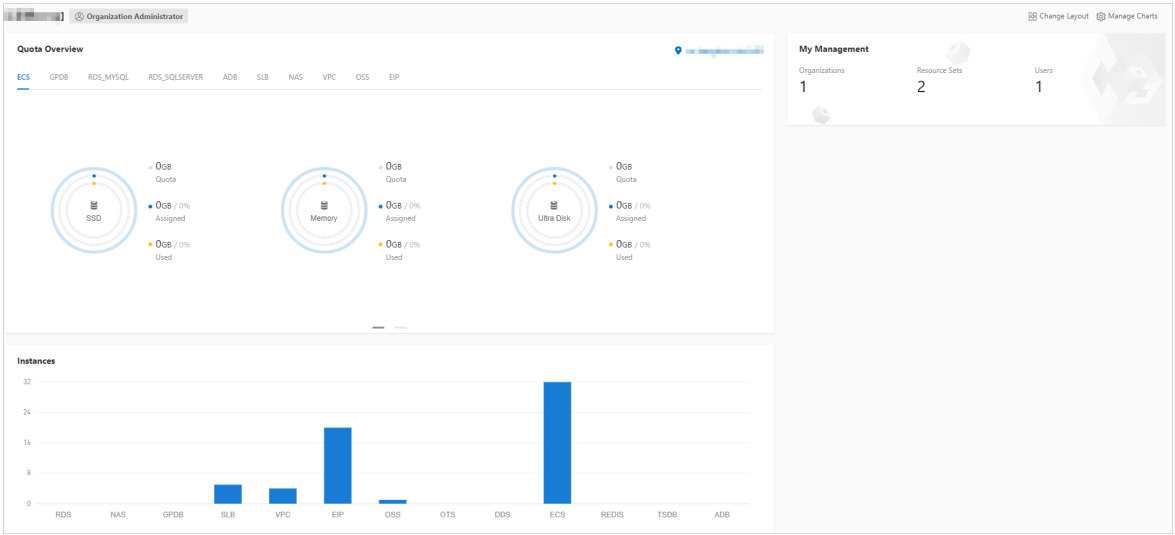
Context

The resource types displayed may vary with region types. See your dashboard for available resource types.

Procedure

1. Log on to the Apsara Uni-manager Management Console.

   By default, the **workbench** page appears when you log on to the Apsara Uni-manager Management Console. To return to the workbench page from other pages, click **Home** in the top navigation bar.



2. On the **workbench** page, you can view the instance summary information for all regions of the Apsara Stack environment.

   You can click **Manage Charts** in the upper-right corner of the page to select all or individual modules to view relevant information. You can also click **Change Layout** in the upper-right corner of the page and drag a specific module to the desired location.

   - **Quota Overview**

     Shows the usage and quotas of Elastic Compute Service (ECS), ApsaraDB RDS, Object Storage Service (OSS), and Server Load Balancer (SLB) resources.

   - **Instances**

Shows the numbers of ECS instances, ApsaraDB RDS instances, OSS buckets, and SLB instances in each region.

- **Instance Trends**

  Shows the numbers of recent ECS instances, ApsaraDB RDS instances, OSS buckets, and SLB instances.

- **Resource Load**

  Shows the top five ECS and ApsaraDB RDS instances in terms of disk usage, CPU utilization, and memory usage.

- **Alert Rules**

  Shows the number of alerts and details of the alerts.

- **My Management**

  Shows the numbers of organizations, resource sets, and users.

- **Multi-cloud Regions**

  Shows the information of all primary and secondary nodes in Apsara Stack. The network connection status and related alerts are displayed for each secondary node.

- **Multi-cloud Resources**

  Shows the cloud services and the number of instances in each secondary node.

1.1.6.2. Configure the workbench

If your homepage is the default homepage, you can customize the dashboards for the homepage.

Background information

If Default Homepage is selected on the Menu Permissions tab of your role, no dashboards are displayed in the workbench on the homepage. You can manually configure a URL to be displayed on the homepage.

> **Note**
>
> This URL supports only domain names of the oneconsole type. You can view and access a URL only by using the account and the role that are used to specify the URL. For more information, see Create a menu.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click Home.
3. Click **Configure Iframe**.
4. Enter a URL in the field.
5. Click **OK**.

1.1.6.3. CloudMonitor

1.1.6.3.1. Cloud Monitor overview

Cloud Monitor provides real-time monitoring, alerting, and notification services for resources to protect your services and businesses.

Cloud Monitor can monitor metrics for a variety of services such as ECS, ApsaraDB RDS, SLB, OSS, KVStore for Redis, VPN Gateway, AnalyticDB for PostgreSQL, ApsaraDB for MongoDB, EIP, and API Gateway.

You can use the metrics of cloud services to configure alert rules and notification policies. This way, you can stay up to date on the running status and performance of your service instances and scale resources in a timely manner when resources are insufficient.

1.1.6.3.2. Metrics

This topic describes the metrics available for each service.

CloudMonitor checks the availability of services based on their metrics. You can configure alert rules and notification policies for these metrics to stay up to date on the running status and performance of monitored service instances.

CloudMonitor can monitor resources of other services, including Elastic Compute Service (ECS), ApsaraDB RDS, Server Load Balancer (SLB), Object Storage Service (OSS), KVStore for Redis, VPN Gateway, AnalyticDB for PostgreSQL, ApsaraDB for MongoDB, Elastic IP Address (EIP), and API Gateway. The following tables list the metrics for each service.

Operating system metrics for ECS

Metric	Description	Unit
Host.cpu.total	The total CPU utilization of an ECS instance.	%
Host.mem.usedutilization	The memory usage of an ECS instance.	%
Host.load1	The system loads over the last 1 minute. This metric is unavailable for Windows operating systems.	N/A
Host.load5	The system loads over the last 5 minutes. This metric is unavailable for Windows operating systems.	N/A
Host.load15	The system loads over the last 15 minutes. This metric is unavailable for Windows operating systems.	N/A
Host.disk.utilization	The disk usage of an ECS instance.	%
Host.disk.readbytes	The number of bytes read from the disk per second.	byte/s
Host.disk.writebytes	The number of bytes written to the disk per second.	byte/s
Host.disk.readIops	The number of read requests received by the disk per second.	count/s
Host.disk.writeIops	The number of write requests received by the disk per second.	count/s
Host.fs.inode	The inode usage.	%

Basic metrics for ECS

Metric	Description	Unit
CPU utilization	The CPU utilization of an ECS instance.	%

Metric	Description	Unit
Inbound bandwidth to the Internet	The average rate of inbound traffic to the Internet.	bit/s
Inbound bandwidth to the internal network	The average rate of inbound traffic to the internal network.	bit/s
Outbound bandwidth from the Internet	The average rate of outbound traffic from the Internet.	bit/s
Outbound bandwidth from the internal network	The average rate of outbound bandwidth from the internal network.	bit/s
System disk BPS	The number of bytes read from and written to the system disk per second.	byte/s
System disk IOPS	The number of reads from and writes to the system disk per second.	count/s
Advance CPU credits	The changes in advance CPU credits. Advance CPU credits can be used only when the unlimited mode is enabled.	N/A
CPU credit consumption	The changes in CPU credit consumption. Consumption trends are consistent with CPU utilization.	N/A
Overdrawn CPU credits	The changes in overdrawn CPU credits. Overdrawn CPU credits can be used only when the unlimited mode is enabled.	N/A
CPU credit balance	The changes in CPU credit balance. The CPU credit balance is used to maintain CPU credit usage.	N/A

> **Note**
>
> For ECS instances, you must install a monitoring plug-in to collect metric data at the operating system level.
>
> Installation method: On the **Cloud Monitor** page, select the instance that you want to monitor from the ECS instance list and click **Batch Install** in the lower part of the page.
>
> Metric data is displayed in the monitoring chart within 5 to 10 minutes after the monitoring plug-in is installed.

Metrics for ApsaraDB RDS for PostgreSQL

Metric	Description	Apsara Stack service	Formula
CPU utilization	The CPU utilization of an ApsaraDB RDS for PostgreSQL instance. Unit: %.	ApsaraDB RDS for PostgreSQL	Used CPU cores of an ApsaraDB RDS for PostgreSQL instance/Total CPU cores of the ApsaraDB RDS for PostgreSQL instance

Metric	Description	Apsara Stack service	Formula
Memory usage	The memory usage of an ApsaraDB RDS for PostgreSQL instance. Unit: %.	ApsaraDB RDS for PostgreSQL	Used memory of an ApsaraDB RDS for PostgreSQL instance/Total memory of the ApsaraDB RDS for PostgreSQL instance
Disk usage	The disk usage of an ApsaraDB RDS for PostgreSQL instance. Unit: %.	ApsaraDB RDS for PostgreSQL	None
IOPS usage	The number of I/O requests for an ApsaraDB RDS for PostgreSQL instance per second. Unit: %.	ApsaraDB RDS for PostgreSQL	Number of I/O requests for an ApsaraDB RDS for PostgreSQL instance/Statistical period
Connection usage	The number of connections between an application and an ApsaraDB RDS for PostgreSQL instance per second. Unit: %.	ApsaraDB RDS for PostgreSQL	Number of connections between an application and an ApsaraDB RDS for PostgreSQL instance/Statistical period

Metrics for ApsaraDB RDS for MySQL

Metric	Description	Apsara Stack service	Formula
CPU utilization	The CPU utilization of an ApsaraDB RDS for MySQL instance. Unit: %.	ApsaraDB RDS for MySQL	Used CPU cores of an ApsaraDB RDS for MySQL instance/Total CPU cores of the ApsaraDB RDS for MySQL instance
Memory usage	The memory usage of an ApsaraDB RDS for MySQL instance. Unit: %.	ApsaraDB RDS for MySQL	Used memory of an ApsaraDB RDS for MySQL instance/Total memory of the ApsaraDB RDS for MySQL instance
Disk usage	The disk usage of an ApsaraDB RDS for MySQL instance. Unit: %.	ApsaraDB RDS for MySQL	None
IOPS usage	The number of I/O requests for an ApsaraDB RDS for MySQL instance per second. Unit: %.	ApsaraDB RDS for MySQL	Number of I/O requests for an ApsaraDB RDS for MySQL instance/Statistical period
Connection usage	The number of connections between an application and an ApsaraDB RDS for MySQL instance per second. Unit: %.	ApsaraDB RDS for MySQL	Number of connections between an application and an ApsaraDB RDS for MySQL instance/Statistical period
Inbound bandwidth to ApsaraDB RDS for MySQL	The inbound traffic to an ApsaraDB RDS for MySQL instance per second.	ApsaraDB RDS for MySQL	None
Outbound bandwidth from ApsaraDB RDS for MySQL	The outbound traffic from an ApsaraDB RDS for MySQL instance per second.	ApsaraDB RDS for MySQL	None

Metrics for ApsaraDB RDS for SQL Server

Metric	Description	Apsara Stack service	Formula
CPU utilization	The CPU utilization of an ApsaraDB RDS for SQL Server instance. Unit: %.	ApsaraDB RDS for SQL Server	Used CPU cores of an ApsaraDB RDS for SQL Server instance/Total CPU cores of the ApsaraDB RDS for SQL Server instance
Memory usage	The memory usage of an ApsaraDB RDS for SQL Server instance. Unit: %.	ApsaraDB RDS for SQL Server	Used memory of an ApsaraDB RDS for SQL Server instance/Total memory of the ApsaraDB RDS for SQL Server instance
Disk usage	The disk usage of an ApsaraDB RDS for SQL Server instance. Unit: %.	ApsaraDB RDS for SQL Server	None
IOPS usage	The number of I/O requests for an ApsaraDB RDS for SQL Server instance per second. Unit: %.	ApsaraDB RDS for SQL Server	Number of I/O requests for an ApsaraDB RDS for SQL Server instance/Statistical period
Connection usage	The number of connections between an application and an ApsaraDB RDS for SQL Server instance per second. Unit: %.	ApsaraDB RDS for SQL Server	Number of connections between an application and an ApsaraDB RDS for SQL Server instance/Statistical period
Inbound bandwidth to ApsaraDB RDS for SQL Server	The inbound traffic to an ApsaraDB RDS for SQL Server instance per second.	ApsaraDB RDS for SQL Server	None
Outbound bandwidth from ApsaraDB RDS for SQL Server	The outbound traffic from an ApsaraDB RDS for SQL Server instance per second.	ApsaraDB RDS for SQL Server	None

Metrics for PolarDB

Metric	Description	Apsara Stack service	Formula
CPU utilization	The CPU utilization of a PolarDB instance. Unit: %.	PolarDB	Used CPU cores of a PolarDB instance/Total CPU cores of the PolarDB instance
Memory usage	The memory usage of a PolarDB instance. Unit: %.	PolarDB	Used memory of a PolarDB instance/Total memory of the PolarDB instance
Disk usage	The disk usage of a PolarDB instance. Unit: %.	PolarDB	None
IOPS usage	The number of I/O requests for a PolarDB instance per second. Unit: %.	PolarDB	Number of I/O requests for a PolarDB instance/Statistical period
Connection usage	The number of connections between an application and a PolarDB instance per second. Unit: %.	PolarDB	Number of connections between an application and a PolarDB instance/Statistical period

Metrics for SLB

Metric	Description	Unit
Inbound bandwidth on a port	The average rate of inbound traffic on a port.	bit/s
Outbound bandwidth on a port	The average rate of outbound traffic on a port.	bit/s
Number of new connections on a port	The average number of new TCP connections established between clients and SLB instances in a statistical period.	N/A
Number of inbound packets received on a port	The number of packets received by an SLB instance per second.	count/s
Number of outbound packets sent on a port	The number of packets sent by an SLB instance per second.	count/s
Number of active connections on a port	The number of TCP connections in the ESTABLISHED state. If persistent connections are used, a connection can transfer multiple file requests at one time.	N/A
Number of inactive connections on a port	The number of TCP connections that are not in the ESTABLISHED state. You can run the **netstat -an** command to view the connections for both Windows and Linux instances.	N/A
Number of concurrent connections on a port	The number of established TCP connections.	count/s
Number of dropped connections on a port	The number of connections dropped per second.	count/s
Number of dropped inbound packets on a port	The number of inbound packets dropped per second.	count/s
Number of dropped outbound packets on a port	The number of outbound packets dropped per second.	count/s
Dropped inbound bandwidth on a port	The amount of inbound traffic dropped per second.	bit/s
Dropped outbound bandwidth on a port	The amount of outbound traffic dropped per second.	bit/s

Metrics for monitoring service overview of OSS

Metric	Description	Unit

Metric	Description	Unit
Availability	The metric that describes the system availability of OSS. You can obtain the metric value based on the following formula: Metric value = `1 - Server error requests with the returned HTTP status code 5xx/All requests`.	%
Valid request percentage	The percentage of valid requests out of all requests.	%
Total number of requests	The total number of requests that are received and processed by the OSS server.	N/A
Number of valid requests	The total number of requests with HTTP status codes 2xx and 3xx returned.	N/A
Outbound traffic from the Internet	The amount of outbound traffic from the Internet.	byte
Inbound traffic to the Internet	The amount of inbound traffic to the Internet.	byte
Outbound traffic from the internal network	The amount of outbound traffic from the internal network.	byte
Inbound traffic to the internal network	The amount of inbound traffic to the internal network.	byte
CDN outbound traffic	The amount of outbound traffic sent over CDN after CDN is activated. Such outbound traffic over CDN is back-to-origin traffic.	byte
CDN inbound traffic	The amount of inbound traffic received over CDN after CDN is activated.	byte
Outbound traffic of cross-region replication	The amount of outbound traffic generated during data replication after cross-region replication is enabled.	byte
Inbound traffic of cross-region replication	The amount of inbound traffic generated during data replication after cross-region replication is enabled.	byte
Storage size	The amount of total storage occupied by the buckets of a specified user before the statistics collection deadline.	byte

Metric	Description	Unit
Number of PUT requests	The total number of PUT requests made by the user between 00:00:00 on the first day of the current month and the statistics collection deadline.	N/A
Number of GET requests	The total number of GET requests made by the user between 00:00:00 on the first day of the current month and the statistics collection deadline.	N/A

Metrics for request status details of OSS

Metric	Description	Unit
Number of requests with server-side errors	The total number of system-level error requests with the returned HTTP status code 5xx.	N/A
Percentage of requests with server-side errors	The percentage of requests with server-side errors out of all requests.	%
Number of requests with network errors	The total number of requests with the returned HTTP status code 499.	N/A
Percentage of requests with network errors	The percentage of requests with network errors out of all requests.	%
Number of requests with client-side authorization errors	The total number of requests with the returned HTTP status code 403.	N/A
Percentage of requests with client-side authorization errors	The percentage of requests with authorization errors out of all requests.	%
Number of requests with client-side errors indicating resources not found	The total number of requests with the returned HTTP status code 404.	N/A
Percentage of requests with client-side errors indicating resources not found	The percentage of requests with errors indicating resources not found out of all requests.	%
Number of requests with client-side timeout errors	The total number of requests with the returned HTTP status code 408 or OSS error code RequestTimeout.	N/A
Percentage of requests with client-side timeout errors	The percentage of requests with client-side timeout errors out of all requests.	%
Number of requests with other client-side errors	The total number of requests other than the foregoing client-side error requests with the returned HTTP status code 4xx.	N/A
Percentage of requests with other client-side errors	The percentage of requests with other client-side errors out of all requests.	%

Metric	Description	Unit
Number of successful requests	The total number of requests with the returned HTTP status code 2xx.	N/A
Percentage of successful requests	The percentage of successful requests out of all requests.	%
Number of redirected requests	The total number of requests with the returned HTTP status code 3xx.	N/A
Percentage of redirected requests	The percentage of redirected requests out of all requests.	%

Metrics for maximum latency of OSS

Metric	Description	Unit
Maximum end-to-end latency of GetObject requests	The maximum end-to-end latency of successful GetObject requests.	ms
Maximum server latency of GetObject requests	The maximum server latency of successful GetObject requests.	ms
Maximum end-to-end latency of HeadObject requests	The maximum end-to-end latency of successful HeadObject requests.	ms
Maximum server latency of HeadObject requests	The maximum server latency of successful HeadObject requests.	ms
Maximum end-to-end latency of PutObject requests	The maximum end-to-end latency of successful PutObject requests.	ms
Maximum server latency of PutObject requests	The maximum server latency of successful PutObject requests.	ms
Maximum end-to-end latency of PostObject requests	The maximum end-to-end latency of successful PostObject requests.	ms
Maximum server latency of PostObject requests	The maximum server latency of successful PostObject requests.	ms
Maximum end-to-end latency of AppendObject requests	The maximum end-to-end latency of successful AppendObject requests.	ms
Maximum server latency of AppendObject requests	The maximum server latency of successful AppendObject requests.	ms
Maximum end-to-end latency of UploadPart requests	The maximum end-to-end latency of successful UploadPart requests.	ms
Maximum server latency of UploadPart requests	The maximum server latency of successful UploadPart requests.	ms
Maximum end-to-end latency of UploadPartCopy requests	The maximum end-to-end latency of successful UploadPartCopy requests.	ms
Maximum server latency of UploadPartCopy requests	The maximum server latency of successful UploadPartCopy requests.	ms

Metrics for successful request category of OSS

Metric	Description	Unit
Number of successful GetObject requests	The number of successful GetObject requests.	N/A
Number of successful HeadObject requests	The number of successful HeadObject requests.	N/A
Number of successful PostObject requests	The number of successful PostObject requests.	N/A
Number of successful AppendObject requests	The number of successful AppendObject requests.	N/A
Number of successful UploadPart requests	The number of successful UploadPart requests.	N/A
Number of successful UploadPartCopy requests	The number of successful UploadPartCopy requests.	N/A
Number of successful DeleteObject requests	The number of successful DeleteObject requests.	N/A
Number of successful DeleteObjects requests	The number of successful DeleteObjects requests.	N/A

Metrics for KVStore for Redis

Metric	Description	Apsara Stack service	Unit
CPU utilization	The CPU utilization of a KVStore for Redis instance.	KVStore for Redis	%
Memory usage	The percentage of memory that is in use.	KVStore for Redis	%
Used memory	The amount of memory that is in use.	KVStore for Redis	byte
Number of used connections	The total number of client connections that are in use.	KVStore for Redis	N/A
Percentage of used connections	The percentage of connections that are in use.	KVStore for Redis	%
Write bandwidth	The write traffic per second.	KVStore for Redis	byte/s
Read bandwidth	The read traffic per second.	KVStore for Redis	byte/s
Number of failed operations per second	The number of failed operations on a KVStore for Redis instance per second.	KVStore for Redis	count/s
Write bandwidth usage	The percentage of total bandwidth used by write operations.	KVStore for Redis	%
Read bandwidth usage	The percentage of total bandwidth used by read operations.	KVStore for Redis	%
Used QPS	The number of queries per second (QPS).	KVStore for Redis	count/s

Metric	Description	Apsara Stack service	Unit
QPS usage	The QPS usage.	KVStore for Redis	%
Average response time	The average response time.	KVStore for Redis	ms
Maximum response time	The maximum response time.	KVStore for Redis	ms
Number of failed commands	The number of failed commands.	KVStore for Redis	N/A
Hit rate	The current hit rate.	KVStore for Redis	%
Inbound traffic	The inbound traffic to a KVStore for Redis instance.	KVStore for Redis	byte
Inbound bandwidth usage	The inbound bandwidth usage of a KVStore for Redis instance.	KVStore for Redis	%
Outbound traffic	The outbound traffic from a KVStore for Redis instance.	KVStore for Redis	byte
Outbound bandwidth usage	The outbound bandwidth usage of a KVStore for Redis instance.	KVStore for Redis	%

Metrics for VPN Gateway

Metric	Dimension	Monitoring period	Unit
Number of inbound packets in a connection per second	User and instance	1 minute	pps
Number of outbound packets in a connection per second	User and instance	1 minute	pps
Inbound bandwidth of a connection	User and instance	1 minute	bit/s
Outbound bandwidth of a connection	User and instance	1 minute	bit/s
Number of connections	User and instance	1 minute	N/A

Metrics for AnalyticDB for PostgreSQL

Metric	Description	Unit
Connection usage	The number of connections between an application and an AnalyticDB for PostgreSQL instance per second.	%
CPU utilization	The CPU utilization of an AnalyticDB for PostgreSQL instance.	%
Disk usage	The disk usage of an AnalyticDB for PostgreSQL instance.	%

Metric	Description	Unit
IOPS usage	The number of I/O requests for an AnalyticDB for PostgreSQL instance per second.	%
Memory usage	The memory usage of an AnalyticDB for PostgreSQL instance.	%

Metrics for ApsaraDB for MongoDB

Tab	Metric	Description	Unit
Basic metric	CPU utilization	The CPU utilization of an ApsaraDB for MongoDB instance.	%
	Memory usage	The memory usage of an ApsaraDB for MongoDB instance.	%
	Disk usage	The disk usage of an ApsaraDB for MongoDB instance.	%
	IOPS usage	The percentage of the IOPS used by an ApsaraDB for MongoDB instance out of the maximum available IOPS.	%
	Connection usage	The number of connections between an application and an ApsaraDB for MongoDB instance per second.	%
	QPS	The number of queries per second.	N/A
	Number of used connections	The number of current connections to an ApsaraDB for MongoDB instance.	N/A
Disk capacity	Disk space occupied by an instance	The total used space.	byte
	Disk space occupied by data	The disk space occupied by data.	byte
	Disk space occupied by logs	The disk space occupied by logs.	byte
Network request	Inbound traffic to the internal network	The inbound traffic.	byte
	Outbound traffic from the internal network	The outbound traffic.	byte

Tab	Metric	Description	Unit
	Number of requests	The number of processed requests.	N/A
Number of operations	Number of Insert operations	None	N/A
	Number of Query operations	None	N/A
	Number of Update operations	None	N/A
	Number of Delete operations	None	N/A
	Number of Getmore operations	None	N/A
	Number of Command operations	None	N/A

Metrics for EIP

Metric	Description	Dimension	Monitoring period	Unit
Inbound bandwidth	The traffic that passes through EIP to ECS per second.	Instance	1 minute	bit/s
Outbound bandwidth	The traffic that passes through EIP from ECS per second.	Instance	1 minute	bit/s
Number of inbound packets per second	The number of packets that pass through EIP to ECS per second.	Instance	1 minute	pps
Number of outbound packets per second	The number of packets that pass through EIP from ECS per second.	Instance	1 minute	pps
Packet loss rate due to throttling	The packet loss rate when the actually used bandwidth exceeds the configured upper limit.	Instance	1 minute	pps

Metrics for API Gateway

Metric	Description	Dimension	Unit	Monitoring period

Metric	Description	Dimension	Unit	Monitoring period
Error distribution	The number of 2xx, 4xx, and 5xx status codes returned for an API in the monitoring period.	User and API	N/A	1 minute
Inbound traffic	The total traffic of requests received by an API in the monitoring period.	User and API	byte	1 minute
Outbound traffic	The total traffic of responses sent by an API in the monitoring period.	User and API	byte	1 minute
Response time	The latency between the time when API Gateway calls the backend service of an API and the time when the result is received from the backend service in the monitoring period.	User and API	s	1 minute
Number of total requests	The total number of requests received by an API in the monitoring period.	User and API	N/A	1 minute

1.1.6.3.3. View monitoring charts

You can view monitoring charts to obtain up-to-date information about each instance.

Procedure

1. Log on to the Apsara Uni-manager console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > CloudMonitor**.
3. In the left-side navigation pane of the CloudMonitor page, click **Cloud Service Monitoring**.
4. Click a cloud service.
5. Click **Monitoring Charts** in the **Actions** column corresponding to an instance.

   On the Monitoring Charts page that appears, you can select a date and time to view the monitoring data of each metric.

1.1.6.4. Alerts

1.1.6.4.1. View alert overview

On the **Overview** page in Cloud Monitor, you can view the alert status statistics and alert logs.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.

2. In the top navigation bar, choose **Products > Monitoring and O&M > CloudMonitor**.
3. In the left-side navigation pane of the Cloud Monitor page, click **Overview**.
4. On the **Overview** page, view the alert status statistics and alert logs that are generated in the last 24 hours.

1.1.6.4.2. Enable or disable alert notification

You can choose whether to enable alert notification by email or DingTalk.

Prerequisites

Valid contact information is specified when you create a user. If the contact information has changed, you must modify your personal information. For more information, see Modify personal information.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and click **Personal information**.
3. In the **Messages** section, select **Email** or **DingTalk** to enable alert notification.

   To disable alert notification, clear the corresponding check box.

1.1.6.4.3. View alert logs

You can view alert information to stay up to date on the running status of Elastic Compute Service (ECS), ApsaraDB RDS, Server Load Balancer (SLB), KVStore for Redis, VPN Gateway, AnalyticDB for PostgreSQL, ApsaraDB for MongoDB, Elastic IP Address (EIP), API Gateway, and Object Storage Service (OSS).

Context

Alert information contains information for all items that do not comply with your configured alert rules.

> **Note**
> - The system can retain up to one million alert items generated within the last three months.
> - This topic describes how to view alert information for ECS. You can view the alert information for other cloud resources in a similar manner.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > Cloud Monitor**.
3. In the left-side navigation pane of the Cloud Monitor page, choose **Alerts > Alert History**.
4. On the **Alert Rule History List** page, filter alert information by rule ID, rule name, service, and date.

   The following table describes the fields in the query result. Alert information fields

Parameter	Description
Product	The service for which the alert was triggered.
Fault Instance	The instance for which the alert was triggered.
Occurred At	The time when the alert was triggered.
Rule Name	The name of the alert rule.

Parameter	Description
Status	The status of the alert rule.
Notification Contact	The recipient of the alert notification.

1.1.6.4.4. Alert rules

1.1.6.4.4.1. View alert rules

After you create alert rules, you can view your alert rules on the Alert Rules page.

Context

The system provides alert rules for ECS, ApsaraDB RDS, SLB, OSS, KVStore for Redis, VPN Gateway, AnalyticDB for PostgreSQL, ApsaraDB for MongoDB, EIP, and API Gateway.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > CloudMonitor**.
3. In the left-side navigation pane of the CloudMonitor page, click **Cloud Service Monitoring**.
4. Click a cloud service.
5. Click **Alert Rules** in the **Actions** column corresponding to an instance.

   On the **Alert Rules** page, view the detailed information of alert rules.

1.1.6.4.4.2. Create an alert rule

You can create an alert rule to monitor an instance.

Prerequisites

For Elastic Compute Service (ECS) instances, you must install a monitoring plug-in to collect metric data at the operating system level.

You can use the following method to install a monitoring plug-in:

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > Cloud Monitor**.
3. In the left-side navigation pane, choose **Cloud Service Monitoring > ECS**.
4. In the ECS instance list, select the instances that you want to monitor and click **Batch Install**.

> **Note**
>
> The monitoring chart displays monitoring data 5 to 10 minutes after the monitoring plug-in is installed.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > Cloud Monitor**.
3. In the left-side navigation pane of the Cloud Monitor page, click **Cloud Service Monitoring**.
4. Click a cloud service.
5. Click **Alert Rules** in the **Actions** column corresponding to an instance.

> **Note** You can also use the search feature to query specific instances for which you want to create alert rules.

6. On the **Alert Rules** page, click **Create Alert Rule**.

   Parameters for creating an alert rule

Parameter	Description
Product	The monitored cloud service.
Resource Range	The range of resources that is associated with the alert rule.
Rule Description	The description of the alert rule.
Add Rule Description	Click **Add Rule Description** to go to the rule configuration panel. For more information, see Parameters for adding rule description.
Effective Time	Only a single alert is sent during each mute duration, even if the metric value exceeds the alert rule threshold several times in a row.
Effective Period	An alert is sent only when the threshold is crossed during the effective period.
HTTP CallBack	The callback URL when the alert conditions are met.
Alert Contact Group	The group to which alerts are sent.

   Parameters for adding rule description

Parameter	Description
Rule Name	The name of the alert rule. The name must be 1 to 64 characters in length and can contain letters and digits.
Metric Name	Different products have different monitoring metrics. For more information, see Metrics.
Comparison	The comparison between thresholds and observed values. The comparison operators include >, >=, <, and <=. When the comparison rule is satisfied, an alert rule is triggered.
Threshold and Alert Level	Different metrics have different reference thresholds.

7. Click **OK**.

1.1.6.4.4.3. Disable an alert rule

You can disable one or more alert rules.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > CloudMonitor**.
3. In the left-side navigation pane of the Cloud Monitor page, click **Cloud Service Monitoring**.
4. Click a cloud service.
5. Click **Alert Rules** in the **Actions** column corresponding to an instance.
6. On the **Alert Rules** page, choose **More > Disable** in the **Actions** column corresponding to the alert rule to be

disabled.

7. In the Disable Alert Rule message, click **Confirm**.

1.1.6.4.4.4. Enable an alert rule

You can re-enable alert rules at any time after they have been disabled.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > Cloud Monitor**.
3. In the left-side navigation pane of the CloudMonitor page, click **Cloud Service Monitoring**.
4. Click a cloud service.
5. Click **Alert Rules** in the **Actions** column corresponding to an instance.
6. On the **Alert Rules** page, choose **More > Enable** in the **Actions** column corresponding to the alert rule to be enabled.
7. In the message that appears, click **Confirm**.

1.1.6.4.4.5. Delete an alert rule

You can delete alert rules that are no longer needed.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, choose **Products > Monitoring and O&M > CloudMonitor**.
3. In the left-side navigation pane of the Cloud Monitor page, click **Cloud Service Monitoring**.
4. Click a cloud service.
5. Click **Alert Rules** in the **Actions** column corresponding to an instance.
6. On the **Alert Rules** page, click **Delete** in the **Actions** column corresponding to the alert rule to be deleted.
7. In the Delete Alert message, click **Confirm**.

1.1.7. VMware Cloud on Alibaba Cloud

1.1.7.1. VMware Cloud on Alibaba Cloud

1.1.7.1.1. Log on to the VMware Cloud on Alibaba Cloud console

This topic describes how to log on to the VMware Cloud on Alibaba Cloud console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.

2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Elastic Computing > VMware Cloud on Alibaba Cloud**.

1.1.7.1.2. Bind a VMware Cloud on Alibaba Cloud region

Before you use VMware Cloud on Alibaba Cloud, you must bind a VMware Cloud on Alibaba Cloud region to an organization.

Prerequisites

A VMware Cloud on Alibaba Cloud region is managed. For more information, see Add a VMware node.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, click **Resource Pools**.
4. In the organization navigation tree, click an organization. In the **Regions** section, select the region that you want to bind.
5. Click **Update Association**.

1.1.7.1.3. Instructions

1.1.7.1.3.1. Limits

Before you use VMware Cloud on Alibaba Cloud virtual machine (VM) templates, you must familiarize yourself with the instance limits.

General limits

- You must select appropriate operating systems for VMware Cloud on Alibaba Cloud VM templates.

  The following operating systems are confirmed to be supported and available in the Apsara Uni-manager Management Console:

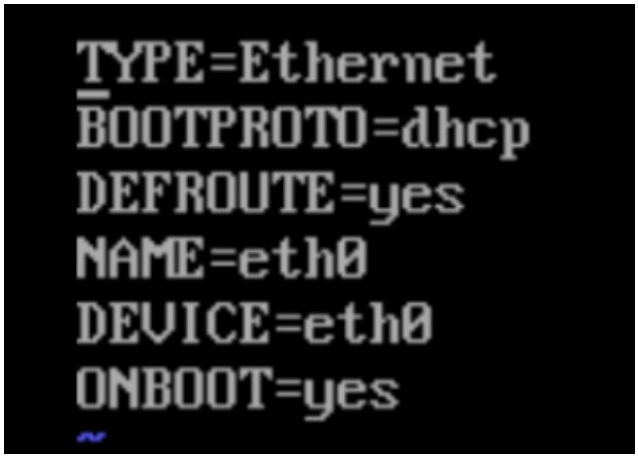
  - CentOS8.2.2004
  - CentOS7.2003
  - CentOS6.10
  - Ubuntu-20.04.1
  - Ubuntu-18.04.5
  - Ubuntu-16.04.7
  - WindowsServer2016
  - WindowsServer2019

- The Apsara Uni-manager Management Console supports VMware vSphere 6.x. Other versions of VMware vSphere such as 5.x or 7.x can be supported in theory. However, whether or not specific versions are supported depends on the compatibility of the VMware Cloud on Alibaba Cloud API and must be evaluated by the R&D team of the Apsara Uni-manager Management Console.
- You must install VMware Tools.

  For more information, see the VMware documentation. Select Full Installation in the installation process.
- You must modify network interface controller (NIC) configurations in the operating system of the VM.

  When you create a VM in the Apsara Uni-manager Management Console, you can specify the IP address of the operating system. This feature is supported by valid NIC configurations.

  Operating systems of VM templates must be in Dynamic Host Configuration Protocol (DHCP) mode. Information such as the MAC address and universally unique identifier (UUID) in the NIC configurations must be removed. The following information can be retained.



> **Note**
>
> Some additional configurations are required for the following operating systems:
>
> - CentOS 6: You must clear the content in the NIC configuration file named 70-persistent-net.rules. The file is stored in the /etc/udev/rules.d/ directory.
> - CentOS 7: The system generates the name for a NIC, such as ifcfg-ens160. You must modify the name to ifcfg-eth0 to make the name take effect.
> - Ubuntu 18.04, 20.04, and later: You must run the `sudo rm /etc/netplan/*.yaml` command to remove the NIC configurations.

1.1.7.1.3.2. Suggestions

Before you use VMware Cloud on Alibaba Cloud virtual machine (VM) templates, take note of the following operation suggestions:

- Select the latest version of VM hardware.
- Select thin provision for VM disks.

  Disk replication is required when you create VMs based on templates. Files of disks of the thin provision type are small in size. This helps accelerate the creation of VMs.

> **Note**
>
> Large sizes of disk files in VM templates or slow storage write speeds may cause VM creation to time out and fail. The maximum timeout period supported by the Apsara Uni-manager Management Console is 10 minutes.

1.1.7.1.4. Instances

1.1.7.1.4.1. Create a VMware Cloud on Alibaba Cloud instance

A VMware Cloud on Alibaba Cloud instance is a virtual machine (VM) that contains the basic computing components of a server, such as CPU, memory, operating system, network, and disks.

Prerequisites

- The region where VMware Cloud on Alibaba Cloud is deployed is managed. For more information, see Add a VMware node.
- The region where VMware Cloud on Alibaba Cloud is deployed is bound to an organization. For more information, see Bind a VMware Cloud on Alibaba Cloud region.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Click **Create Instance** in the upper-right corner.
5. Configure parameters listed in the following table to create an instance.

Section	Parameter	Required	Description
Basic Settings	**Organization**	Yes	The organization in which to create the instance.
	Resource Set	Yes	The resource set in which to create the instance.
Region	**Region**	Yes	The region in which to create the instance.
	Zone	Yes	The zone in which to create the instance.
	VPC	Yes	The VPC in which to create the instance.
	vSwitch	Yes	Select the vSwitch to which the instance belongs. The vSwitch corresponds to the port group of a VMware ESXi host or a distributed switch, and maps to the VLAN of a physical switch.
	Private IP Address	Yes	The private IPv4 address of the instance. The private IPv4 address must be within the CIDR block of the vSwitch.
	Private Subnet Mask	Yes	The private subnet mask. Example: 255.255.255.0. The specified subnet mask must be within the CIDR block of the selected vSwitch.
	Private IP Address of Gateway	Yes	The private IP address of the gateway. Example: 192.168.100.1. The IP address must be within the CIDR block of the selected vSwitch.
	Private IP Address of DNS Server	No	The private IP address of the DNS server. Example: 114.114.114.114. The IP address must be within the CIDR block of the selected vSwitch.

Section	Parameter	Required	Description
Instance	**Instance Family**	No	The instance family of the instance. Valid values: ○ Memory Optimized ○ Compute Optimized ○ General Purpose
	Instance Type	Yes	The instance type of the instance. You can specify the vCPUs and memory.
Image	**Image Type**	No	The type of the image. Default value: **Public Image**.
	Public image	Yes	The public image of the instance.
	System Disk (GB)	No	The system disk to which the operating system is installed. You can configure different storage types for the disk. Valid values: ○ **Shared Storage: All**: The system selects an available shared storage. We recommend that you select this type. ○ **Shared Storage: storageA**: The storage named storageA of the VMware Cloud on Alibaba Cloud instance is used. Administrators must make sure the storage is appropriate. If the storage capacity is insufficient, the instance fails to be created.

Section	Parameter	Required	Description
Storage	**Data Disk (GB)**	No	You can also add data disks after the instance is created. You can configure different storage types for the disk. Valid values: ◦ **Shared Storage: All**: The system selects an available shared storage. We recommend that you select this type. ◦ **Shared Storage: storageA**: The storage named storageA of the VMware Cloud on Alibaba Cloud instance is used. Administrators must make sure the storage is appropriate. If the storage capacity is insufficient, the instance fails to be created. You must also specify the provision type when you create the instance. Valid values: ◦ Thin Provision: Storage space increases with the use of the disk. ◦ Thick Provision Lazy Zeroed: Storage space is equal to the size of the disk and does not increase. The disk is formatted when data is written. ◦ Thick Provision Eager Zeroed: Storage space is equal to the size of the disk and does not increase. The storage of the disk is immediately formatted when the disk is created.
Password	**Password Setting**	No	Select **Set after Purchase**.
Instance Name	**Instance Name**	Yes	The name of the instance. The name must be 2 to 128 characters in length and can contain letters, periods (.), underscores (_), hyphens (-), and colons (:). It must start with a letter and cannot start with http:// or https://.

6. Click **Submit**.

1.1.7.1.4.2. View instance information

You can view the list of created instances as well as details of individual instances, such as their basic configurations, disks, and elastic network interfaces (ENIs).

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.

   You can view the list of VMware Cloud on Alibaba Cloud instances that are deployed in the current region.
4. Use one of the following methods to go to the details page of an instance:
   - In the **Instance ID/Name** column, click the instance ID.
   - Click **Manage** in the **Actions** column corresponding to the instance.
   - Choose **More** > **Show Details** in the **Actions** column corresponding to the instance.

1.1.7.1.4.3. Modify an instance

You can modify the name and description of a created VMware Cloud on Alibaba Cloud instance.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance that you want to modify and choose **More** > **Modify** in the **Actions** column.
5. Modify the name and description of the instance.
6. Click **OK**.

1.1.7.1.4.4. Remotely connect to an instance

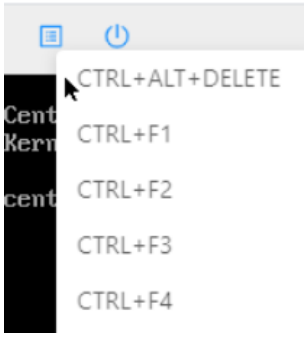
You can remotely connect to and manage added VMware Cloud on Alibaba Cloud instances.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance that you want to manage and click **Remote Connection** in the **Actions** column.
5. Enter the username and password.
   - For a Linux instance, enter the username *root* and the logon password.

     > **Note**
     >
     > When you log on to the Linux instance, the password is not displayed as you enter it. Press the Enter key after you enter the password.

- For a Windows instance, to use a key combination such as Ctrl+Alt+Delete, click the List icon in the upper-right corner of the page and select the corresponding composite key from the drop-down list.



Enter the username and password, and click the Log On icon.

1.1.7.1.4.5. Stop an instance

You can stop VMware Cloud on Alibaba Cloud instances that are not in use. The stop operation interrupts services that are running on the instances. Exercise caution when you perform this operation.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Use one of the following methods to stop instances:
   - To stop a single instance, find the instance and choose **More** > **Instance Status** > **Stop** in the **Actions** column.
   - To stop one or more instances at a time, select the instances and click **Stop** in the lower part of the Instances page.
5. Click **OK**.

Execution results

When the instance is being stopped, its state in the **Status** column changes from **Running** to **Stopping**. After the instance is stopped, its state changes to **Stopped**.

1.1.7.1.4.6. Start an instance

You can start a stopped instance.

Prerequisites

The instance is in the **Stopped** state.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.

3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Use one of the following methods to start instances:
   - To start a single instance, find the instance and choose **More** > **Instance Status** > **Start** in the **Actions** column.
   - To start one or more instances at a time, select the instances and click **Start** in the lower part of the Instances page.
5. Click **OK**.

Execution results

When the instance is being started, its state in the **Status** column changes from **Stopped** to **Starting**. After the instance is started, its state changes to **Running**.

1.1.7.1.4.7. Restart an instance

After you change the logon password of an instance or install system updates, you must restart the instance. The restart operation stops the instances for a short period of time and interrupts services that are running on the instance. Exercise caution when you perform this operation.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Use one of the following methods to restart the instance:
   - To restart a single instance, find the instance and choose **More** > **Instance Status** > **Restart** in the **Actions** column.
   - To restart one or more instances at a time, select the instances and click **Restart** in the lower part of the Instances page.
5. In the Restart Instance dialog box, select a restart mode.
   - **Restart**: restarts the instance normally.
   - **Force Restart**: forces the instance to restart. This may result in the loss of unsaved data.
6. Click **OK**.

1.1.7.1.4.8. Delete an instance

You can delete instances that are no longer needed to release their resources. Deleted instances cannot be recovered. We recommend that you back up data before you delete an instance. If data disks are released with the instances, the disk data cannot be recovered.

Prerequisites

The instance is in the **Stopped** state.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.

2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Select the instance and click **Delete** in the lower part of the Instances page.
5. Click **OK**.

1.1.7.1.4.9. Change the instance type of an instance

You can change the instance types of instances to suit your business needs. This eliminates the need to create VMware Cloud on Alibaba Cloud instances.

Prerequisites

The instance is in the **Stopped** state.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, select an organization, a resource set, and a region.
4. Find the instance whose instance type you want to change and click **Upgrade/Downgrade** in the **Actions** column.
5. On the page that appears, select a new instance type and click **Submit**.

   The page that appears shows the instance types available for selection.
6. Start the instance to make the new instance type take effect.

   For more information, see Start an instance.

1.1.7.1.5. Images

1.1.7.1.5.1. Create a custom image

You can perform the following steps to create a custom image.

Background information

You can create a custom image from an instance to replicate the data of all disks on the instance.

> **Note**
>
> To avoid data security risks, we recommend that you delete sensitive data from an instance before you use the instance to create a custom image.

Create a custom image from an instance

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance from which you want to create a custom image and choose **More** > **Create Custom Image** in the **Actions** column.
5. Set the name, sharing scope, and description for the custom image, and click **OK**.

The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), and colons (:). It cannot start with a special character or digit.

You can set the sharing scope to the permission scope of the image.

The description must be 2 to 256 characters in length and cannot start with http:// or https://.

1.1.7.1.5.2. View images

You can view the list of created images.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, choose **Images** > **Images**.
3. In the top navigation bar, move the pointer over **Region** and select the region where the image is created.
4. Select a filter option, enter the corresponding information in the search box, and then click **Search**.

   You can select multiple filter options to narrow down search results.

Parameter	Description
Image Name	The image name used to search for the image.
Image ID	The image ID used to search for the image.

1.1.7.1.6. Snapshots

1.1.7.1.6.1. Create a snapshot

You can manually create a snapshot for a disk to back up disk data.

Prerequisites

- The instance to which the disk is attached is in the **Running** or **Stopped** state.
- The disk is in the **Running** state.

Background information

A snapshot of a disk can be used to roll back data of the disk.

When you create a snapshot, take note of the following items:

- For each disk, the first snapshot taken is a full snapshot and subsequent snapshots are incremental snapshots. It takes longer to create the first full snapshot than it does subsequent incremental snapshots. The amount of taken time depends on the amount of data that has been changed since the previous snapshot. The more data that has been changed, the longer it takes to create an incremental snapshot.
- Avoid creating snapshots during peak hours.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.

3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Click the instance from which you want to create a snapshot. On the page that appears, click the **Snapshots** tab.
5. Click **Create and Bind Snapshot**.
6. Set the name, type, and description for the snapshot, and click **Submit**.

Parameter	Description
Snapshot Name	The name of the snapshot.
Snapshot Type	The snapshot type. Valid values: ○ Disk Snapshot: stores information in disks. ○ Memory Snapshot: stores information in memory.
Snapshot Description	The description of the snapshot.

1.1.7.1.6.2. Delete a snapshot

You can delete a snapshot that is no longer needed. After the snapshot is deleted, it cannot be recovered.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance whose snapshot is to be deleted and click the **Snapshots** tab.
5. Use one of the following methods to delete the snapshot:
   - To delete a single snapshot, find the snapshot and click **Delete** in the **Actions** column.
   - To delete one or more snapshots at a time, select the snapshots and click **Delete** in the lower part of the Snapshots tab.
6. Click **OK**.

1.1.7.1.6.3. View snapshots

You can view the list of created snapshots.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance in which you want to view snapshots and click the **Snapshots** tab.
5. Select a filter option, enter the corresponding information in the search box, and then click **Search**.

You can select multiple filter options to narrow down search results.

Parameter	Description
Snapshot Name	The snapshot name used to search for the snapshot.
Snapshot ID	The snapshot ID used to search for the snapshot.

1.1.7.1.7. Disks

1.1.7.1.7.1. Create a disk

To increase the storage space of VMware Cloud on Alibaba Cloud instances, you can create standalone data disks and then attach them to the instances. This topic describes how to create an empty data disk. You cannot create standalone system disks.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance for which you want to create a disk and click the **Disks** tab.
5. Click **Create and Attach Disk**.
6. Configure parameters listed in the following table to create a disk.

Section	Parameter	Required	Description
Region	**Zone**	Yes	The zone in which to create the disk.
	Specifications	Yes	The disk category and the disk size.

Section	Parameter	Required	Description
Basic Settings	**Provision Type**	Yes	The provision type. Valid values: ○ Thin Provision: Storage space increases with the use of the disk. ○ Thick Provision Lazy Zeroed: Storage space is equal to the size of the disk and does not increase. The disk is formatted when data is written. ○ Thick Provision Eager Zeroed: Storage space is equal to the size of the disk and does not increase. The storage of the disk is immediately formatted when the disk is created.

7. Click **Submit**.

Execution results

The created disk is displayed in the disk list and in the **Running** state.

1.1.7.1.7.2. View disks

You can view the list of created disks and the details of individual disks.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Click the instance for which you want to view disks. On the page that appears, click the **Disks** tab.
5. Select a filter option from the drop-down list, enter the relevant information in the search box, and then click **Search**.

   You can select multiple filter options to narrow down search results.

Parameter	Description
Disk Name	The disk name used to search for the disk.

Parameter	Description
Disk ID	The disk ID used to search for the disk.
Disk Properties	The disk type used to search for disks of that type. Valid values: ○ All ○ System Disk ○ Data Disk

1.1.7.1.7.3. Detach a data disk

You can detach data disks. System disks cannot be detached.

Procedure

> **Warning**
> Resources are released after disks are detached. Make sure that the data of a disk is backed up before you detach it.

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Click the instance from which you want to detach a data disk. On the page that appears, click the **Disks** tab.
5. Find the data disk that you want to detach and choose **More** > **Detach** in the **Actions** column.
6. Click **OK**.

1.1.7.1.8. ENIs

1.1.7.1.8.1. Create an ENI

You can create and bind elastic network interfaces (ENIs) to VMware Cloud on Alibaba Cloud instances.

Prerequisites

A virtual private cloud (VPC) and a vSwitch are created. For more information, see Create a VPC and Create a vSwitch in *Apsara Stack VPC User Guide*.

Background information

ENIs are classified into primary and secondary ENIs.

A primary ENI is created by default when an instance is created in a VPC. This primary ENI shares the lifecycle of the instance with which it is created and cannot be unbound from the instance.

ENIs that are separately created are secondary ENIs. This topic describes how to create a secondary ENI.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Click the instance for which you want to create an ENI. On the page that appears, click the **ENIs** tab.
5. Click **Create and Bind ENI**.
6. Configure parameters listed in the following table to create an ENI.

Section	Parameter	Required	Description
Region	**Organization**	Yes	The organization in which to create the ENI.
	Resource Set	Yes	The resource set in which to create the ENI.
	Region	Yes	The region in which to create the ENI.
	Zone	Yes	The zone in which to create the ENI.
Basic Settings	**VPC**	Yes	The VPC in which to create the ENI. The secondary ENI can be bound only to an instance in the same VPC. **Note** After an ENI is created, its VPC cannot be changed.
	vSwitch	Yes	The vSwitch in which to create the ENI. The secondary ENI can be bound only to an instance within the same VPC. Select a vSwitch that is deployed within the same zone as the instance to which the ENI is bound. The vSwitch of the ENI can be different from that of the instance. **Note** After an ENI is created, you cannot change its vSwitch.

7. Click **Submit**.

Execution results

The created ENI is displayed on the ENIs tab and is in the **Bound** state.

1.1.7.1.8.2. View ENIs

You can view the list of created elastic network interfaces (ENIs).

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Click the instance for which you want to view ENIs. On the page that appears, click the **ENIs** tab.
5. Select a filter option, enter the corresponding information in the search box, and then click **Search**.

   You can select multiple filter options to narrow down search results.

Parameter	Description
ENI Name	The ENI name used to search for the ENI.
ENI ID	The ENI ID used to search for the ENI.
VSwitch ID	The vSwitch ID used to search for the ENIs that are associated with the vSwitch.

1.1.7.1.8.3. Delete an ENI

You can delete secondary elastic network interfaces (ENIs) that are no longer needed.

Background information

Only secondary ENIs can be deleted. Primary ENIs share the same lifecycle as instances and cannot be deleted.

Procedure

1. Log on to the VMware Cloud on Alibaba Cloud console.
2. In the left-side navigation pane, click **Instances**.
3. In the top navigation bar, move the pointer over **Region** and select the region where VMware Cloud on Alibaba Cloud is deployed.
4. Find the instance whose secondary ENI is to be deleted and click the **ENIs** tab.
5. Find the secondary ENI and click **Delete** in the **Actions** column.
6. Click **OK**.

1.1.8. Enterprise

1.1.8.1. Organizations

1.1.8.1.1. Create an organization

You can create organizations to store resource sets and their resources. This topic describes how to create an organization.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Organizations**.
4. In the organization navigation tree, click the name of the parent organization under which you want to create an organization. Click **Create Organization** in the upper-right corner of the page.
5. In the Add Organization dialog box, enter an organization name and click **OK**.

1.1.8.1.2. View organization information

You can view information of an organization on the Organizations page.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Organizations**.
4. (Optional)Enter the name of the organization that you want to view in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the organization that you want to view.

   You can click the **Resource Sets**, **Users**, or **User Groups** tab to query the resource sets, users, and user groups under the organization.

   - Click the **Resource Sets** tab to view the resource set names, resource set IDs, creation time, creators, and user groups. Click the name of a resource set to view its details.
   - Click the **Users** tab to view the usernames, user status, and roles. Click a username to view the user details.
   - Click the **User Groups** tab to view the user group names, corresponding organizations, roles, users, and creation time.

Click icons in the upper-right corner of the **Resource Sets**, **Users**, and **User Groups** tabs to perform the following operations on the lists:

Icon	Description
	Shows the list in full screen mode. You can click the icon in the upper-right corner to exit the full screen mode.
	Selects the size of cells and fonts in the list. Valid values: • Default • Compact
	Refreshes the list.
	Selects or clears check boxes to determine the items to be displayed in the list. You can also click **Reset** to restore the list to its original settings.

1.1.8.1.3. Change the name of an organization

Users that have operation permissions on an organization can change the name of the organization.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Organizations**.
4. (Optional)Enter the organization name in the search box for fuzzy search above the organization tree.
5. Click the organization name that you want to change.
6. In the upper-right corner of the page, click **Editing**.
7. In the Edit Organization dialog box, change the organization name.
8. Click **OK**.

1.1.8.1.4. Change organization ownership

Users that have operation permissions on an organizations can change the ownership of the organization.

Prerequisites

- Make sure that each sub-organization under the organization that you want to change the ownership has a unique name.
- The ownership of an organization cannot be changed across level-1 organizations.

Context

Users can change the ownership of an organization across parent organizations. This way, the ownership of subordinate organizations, users, and resources are also changed in a cascading manner.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, choose **Resources > Changes** .
4. (Optional)Enter the name of the organization that you want to change in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the organization whose ownership you want to change.
6. Click **Change Organization** on the right side of the page.
7. In the **Change Organization To** dialog box, select a new organization and click **OK** to change the ownership of the original organization along with that of its resources sets and users.

1.1.8.1.5. Obtain AccessKey pairs of an organization

An AccessKey pair that consists of an AccessKey ID and an AccessKey secret is used to implement symmetric encryption. This helps you verify the identity of a requester. The AccessKey ID is used to identify a user. The AccessKey secret is used to encrypt the signature string. This topic describes how to obtain AccessKey pairs of an organization.

Prerequisites

Only operations administrators and level-1 organization administrators can obtain AccessKey pairs of an organization.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.

3. In the left-side navigation pane, choose **Resources > Organizations**.
4. In the organization navigation tree, click the name of the level-1 organization for which you want to obtain AccessKey pairs.
5. Click **Management AccessKey** on the right side of the page.
6. In the Management AccessKey dialog box, view the AccessKey pairs of the organization.

1.1.8.1.6. Create an AccessKey pair for an organization

You can create AccessKey pairs for organizations and delete the original ones to implement the rotation of your AccessKey pairs.

Prerequisites

- Level-1 organization administrators can create AccessKey pairs for organizations to which they belong.
- Operations administrators can create AccessKey pairs for all level-1 organizations.
- A maximum of two AccessKey pairs can be created for each organization.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Organizations**.
4. (Optional) Enter the name of the organization for which you want to create an AccessKey pair in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the level-1 organization for which you want to create an AccessKey pair.
6. Click **Management AccessKey** on the right side of the page.
7. In the **Management AccessKey** dialog box, click **Create AccessKey**.

1.1.8.1.7. Delete an AccessKey pair from an organization

You can delete AccessKey pairs that are no longer needed.

Prerequisites

- Level-1 organization administrators can delete AccessKeys from organizations to which they belong.
- Operations administrators can delete AccessKey pairs from all level-1 organizations.
- At least one AccessKey pair is retained for an organization.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Organizations**.
4. (Optional) Enter the name of the organization from which you want to delete an AccessKey pair in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the level-1 organization from which you want to delete an AccessKey pair.
6. Click **Management AccessKey** on the right side of the page.
7. In the **Management AccessKey** dialog box, find the AccessKey pair that you want to delete and click **Delete** in the **Operation** column.
8. In the message that appears, click **OK**.

1.1.8.1.8. Disable an AccessKey pair for an organization

You can disable AccessKey pairs that are no longer needed. Newly created AccessKey pairs are in the Enable state by default.

Prerequisites

- Level-1 organization administrators can disable AccessKey pairs for organizations to which they belong.
- Operations administrators can disable AccessKey pairs for all level-1 organizations.
- At least one AccessKey pair is enabled for each level-1 organization.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources** > **Organizations**.
4. (Optional) Enter the name of the organization for which you want to disable an AccessKey pair in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the level-1 organization for which you want to disable an AccessKey pair.
6. Click **Management AccessKey** on the right side of the page.
7. Find the AccessKey pair that you want to disable and click **Disable** in the **Operation** column.
8. In the message that appears, click **OK**.

1.1.8.1.9. Enable an AccessKey pair for an organization

Disabled AccessKey pairs must be enabled before they can be used. Newly created AccessKey pairs are in the Enable state by default.

Prerequisites

- Level-1 organization administrators can enable AccessKey pairs for organizations to which they belong.
- Operations administrators can enable AccessKey pairs for all level-1 organizations.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources** > **Organizations**.
4. (Optional) Enter the name of the organization for which you want to enable an AccessKey pair in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the level-1 organization for which you want to enable an AccessKey pair.
6. Click **Management AccessKey** on the right side of the page.
7. Find the AccessKey pair that you want to enable and click **Enable** in the **Operation** column.

1.1.8.1.10. Delete an organization

Administrators can delete organizations that are no longer needed.

Prerequisites

Before you delete an organization, make sure that the organization does not contain users, resource sets, or

subordinate organizations. Otherwise, the organization cannot be deleted.

Warning After the organization is deleted, the data of the organization cannot be recovered. Proceed with caution.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Organizations**.
4. (Optional)Enter the name of the organization that you want to delete in the search box above the organization navigation tree. Fuzzy search is supported.
5. Click the name of the organization that you want to delete.
6. Click **Delete** on the right side of the page.
7. In the Confirm message, click **OK**.

1.1.8.2. Resource sets

1.1.8.2.1. Create a resource set

You must create a resource set before you apply for resources.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Resource Sets**.
4. In the upper-left corner, click **Create Resource Set**.
5. In the **Create Resource Set** dialog box, set **Name** and **Organization**.
6. Click **OK**.

1.1.8.2.2. View the details of a resource set

When you want to use a cloud resource in your organization, you can view the details of the resource set that contains the resource, including all resource instances and users of the resource set.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Resource Sets**.
4. (Optional)Select an **organization** from the drop-down list, or enter a **resource set** name in the search box, and then click **Search**.
5. Click the name of the resource set.
6. On the **Resource Set Details** page, click the **Resources** and **Users** tabs to view all resource instances and users of the resource set.
7. On the **Resources** tab, click the number next to a resource in a service card. You are redirected to the instance list page.

   On the page that appears, view the details of the instance list. The list is automatically filtered and displayed based on the organization and resource set.

1.1.8.2.3. Change the name of a resource set

You can change the name of a resource set to keep it up-to-date as an administrator.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Resource Sets**.
4. (Optional)Select an **organization** from the drop-down list, or enter a **resource set** name in the search box, and then click **Search**.
5. Find the resource set that you want to change and click **Edit Name** in the **Actions** column.
6. In the **Modify Resource Set** dialog box, enter a new name.
7. Click **OK**.

1.1.8.2.4. Add a member to or delete a member from a resource set

You can add a member to or delete a member from a resource set to manage member access to resources within the resource set. A member can be a user or a user group.

Prerequisites

- A resource set is created. For more information, see Create a resource set.
- A user is created. For more information, see Create a user.
- A user group is created. For more information, see Create a user group.

Context

Members of a resource set have the permissions to use resources in the resource set. If you delete resources from a resource set, members of the resource set are not affected. Similarly, If you delete members from a resource set, resources in the resource set are not affected.

You can delete a member that is no longer in use from a resource set. After the member is deleted, it is no longer able to access the resource set.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Resource Sets**.
4. (Optional)Select an **organization** from the drop-down list, or enter a **resource set** name in the search box, and then click **Search**.
5. Find the resource set that you want to manage and click **Member authorization** in the **Actions** column.
6. (Optional)In the **Member authorization** dialog box, select the role type of the user or the user group that you want to add.
7. Add or remove a member.
   - Add a user or a user group: In the lower-left section, select the user or the user group that you want to add.
   - Delete a user or a user group: In the lower-right section, click **Remove** in the **Operation** column.
8. Click **OK**.

1.1.8.2.5. Delete a resource set

You can delete resource sets that are no longer needed as an administrator.

Prerequisites

The resource set to be deleted does not contain resources, users, or user groups.

> Notice
> - A resource set cannot be deleted if it contains resources, users, or user groups.
> - The default resource set cannot be deleted.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Resource Sets**.
4. Find the resource set that you want to delete and click **Delete** in the **Actions** column.
5. In the **Confirm** message, click **OK**.

1.1.8.3. Roles

1.1.8.3.1. Create a custom role

You can create custom roles in the Apsara Uni-manager Management Console to more efficiently grant permissions to users so that different personnel can work with different features.

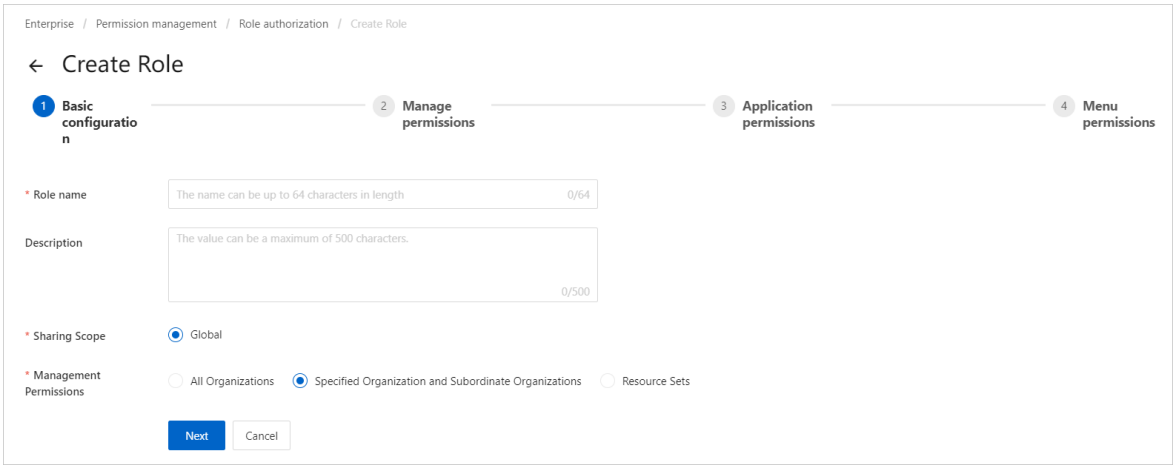
Context

A role is a set of access permissions. Each role has a range of permissions. A user can have multiple roles, which means that the user is granted all of the permissions defined for each role. A role can be used to grant the same set of permissions to a group of users.

The total number of custom and default roles cannot exceed 20.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Click **Create a role** in the upper-left corner.
5. In the **Basic configuration** step, set Role name, Description, Sharing Scope, and Management Permissions. Click **Next**.



The following table describes the role parameters.

Role parameters

Parameter	Description
Role name	The name of the role. The name can be up to 64 characters in length and can contain only letters and digits.
Description	Optional. The description of the role. The description can be up to 100 characters in length and can contain letters, digits, commas (,), semicolons (;), and underscores (_).
Sharing Scope	**Global**: This role is visible and valid for all organizations.
Management Permissions	○ **All Organizations** These permissions apply to all organizations involved. ○ **Specified Organization and Subordinate Organizations** These permissions apply to the organization to which the user belongs and its subordinate organizations. ○ **Resource Sets** These permissions apply to the resource sets that are assigned to the user.

6. In the **Manage permissions** step, select the operation permissions of this role on the Apsara Uni-manager Operations Console, and click **Next**.

   You can click permission categories on the left side to filter permissions.

7. In the **Application Permissions** step, select the operation permissions of this role on cloud services, and click **Next**.

   You can click service names on the left side to filter application permissions.

8. In the **Menu Permissions** step, select the operation permissions of this role on menus and homepage dashboard templates corresponding to the role, and click **Next**.

   You can click the parent menu on the left side to filter permissions on cascading menus.

Result

After the role is created, you can click **Authorize now** in the Submission result message to authorize users or user groups. You can also authorize users or user groups on the details page of the role.

1.1.8.3.2. View the details of a role

If you are uncertain about the specific permissions of a role, you can go to the **Role Authorization** page to view the role permissions.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Click the name of the role that you want to view. On the **Role Authorization** page, view the information of the role.

   You can click the **Management Permissions**, **Application Permissions**, **Menu Permissions**, and **Authorized Personnel** tabs to view the role permissions.

1.1.8.3.3. Modify custom role information

You can modify the name and permissions of a custom role as an administrator.

Context

You cannot modify information of preset roles.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. In the role list, click **Modify** in the **Actions** column.
5. On the **Role Authorization** page, modify the name, description, management permission scope, related permissions, and authorized personnel of the role.
   - Modify role information: In the **Basic information** section, click the icon to the right of the item that you want to modify. After the item is modified, click **OK**.
   - Modify permissions: Click the **Management Permissions**, **Application Permissions**, or **Menu Permissions** tab, select or clear related permissions from the corresponding tab, and then click **Update**.
   - Authorize users: Click the **Authorized Personnel** tab. In the **Authorized User** section, click **Authorized**. In the dialog box that appears, select the users that you want to authorize and click **Confirm authorization**. To cancel the authorization from a user, click **Remove** in the **Operation** column.
   - Authorize user groups: Click the **Authorized Personnel** tab. In the **Authorized User Group** section, click **Authorized**. In the dialog box that appears, select the user groups that you want to authorize and click **Confirm authorization**. To cancel the authorization from a user group, click **Remove** in the **Operation** column.

1.1.8.3.4. Copy a role

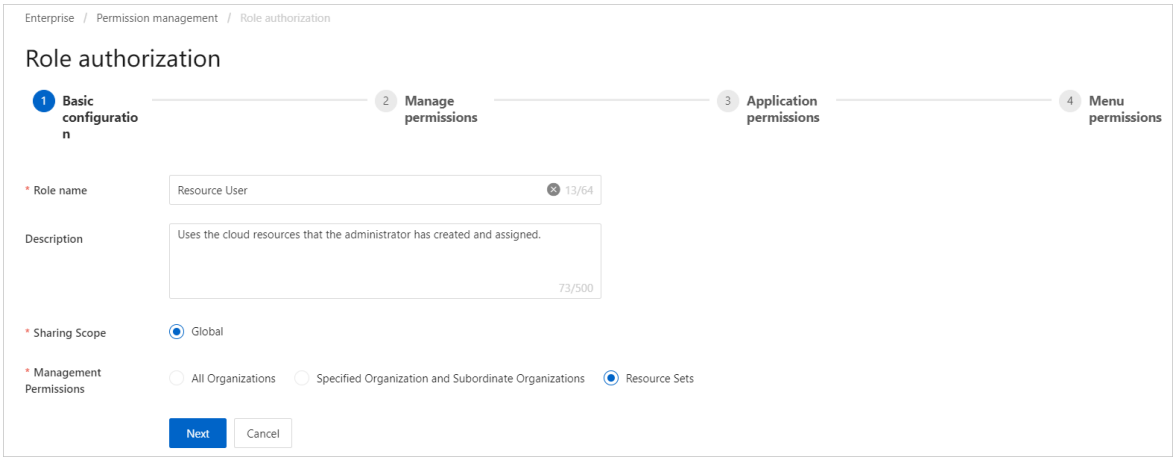
You can copy a preset role or a custom role to create a role that has the same permissions.

Context

Operations on the **Role authorization** page are the same as those for creating a custom role. You can add, modify, and remove the role permissions in the copied role. By default, if you do not modify the role permissions, the sharing scope, management permissions, application permissions, menu permissions, and authorized users are all the same as those of the source role.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. In the role list, choose **More** > **Copy** in the **Actions** column corresponding to a role.
5. On the **Role authorization** page, set the new role name, sharing scope, and management permissions.



> **Note** The role name must be unique.

6. Complete the basic configurations of the new role, and click **Next**.
7. In the **Manage permissions** step, select the operation permissions that the new role has on the Apsara Uni-manager Management Console, and click **Next**.
8. In the **Application permissions** step, select the operation permissions that the new role has on cloud services, and click **Next**.
9. In the **Menu permissions** step, select the operation permissions that the new role has on menus, and click **Next**.

Result

After the role is created, you can click **Authorize now** in the Submission result message to authorize users or user groups. You can also authorize users or user groups on the details page of the role.

1.1.8.3.5. Disable a role

When you disable a role, all of its permissions are also disabled.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. In the role list, find the role that you want to disable and click **Disable** in the **Actions** column.

1.1.8.3.6. Enable a role

When you enable a disabled role, the permissions of the role are restored.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.

3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. In the role list, find the role that you want to enable and click **Enable** in the **Actions**.

1.1.8.3.7. Delete a custom role

You can delete custom roles that are no longer needed.

Prerequisites

- Default or preset roles cannot be deleted.
- To delete a role, you must unbind all user groups from the role.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. In the role list, find the role that you want to delete and choose **More** > **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

1.1.8.4. Users

1.1.8.4.1. System users

1.1.8.4.1.1. Create a user

You can create a user and assign the user different roles as an administrator to meet different requirements for system access control.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab. Click **Create a user**.
5. In the dialog box that appears, set the following parameters.

Parameter	Required	Description
User name	Yes	The Apsara Stack tenant account name of the user. The name must be 1 to 64 characters in length and can contain letters, digits, hyphens (-), underscores (_), and periods (.).
Display name	Yes	The display name of the user. The name must be 1 to 128 characters in lengthand can contain letters, digits, hyphens (-), underscores (_), periods (.), and at signs (@).
Role	Yes	The role to be assigned to the user. **Note** You can enter role names in the field. Fuzzy match is supported.

Parameter	Required	Description
Tissue	Yes	The organization to which the user belongs.
Logon policy	Yes	The logon policy that restricts the logon time and IP addresses of the user. The default policy is automatically bound to new users. **Note** The default policy does not restrict the logon time and IP addresses of the user. To restrict the allowed logon time and IP addresses of the user, you can modify the logon policy of the user or create a logon policy for the user. For more information, see Create a logon policy.
Phone	Yes	The mobile number of the user. This mobile number is used to notify users of resource application and usage. Make sure that this mobile number is valid. **Note** If the mobile number of the user changes, update it on the system in a timely manner.
Landline	No	The landline number of the user. The landline number must be 4 to 20 characters in length and can contain only digits and hyphens (-).
Email	Yes	The email address of the user. Emails about the resource application and usage are sent to this email address. Make sure that this email address is valid. **Note** If the email address changes, update it on the system in a timely manner.
DingTalk Key	No	The key of the chatbot for the DingTalk group to which the user belongs. For more information about how to configure the key, see DingTalk development documentation.
Notify User by Email	No	If this option is selected, the Apsara Uni-manager Management Console informs the user configured as the alert contact by email whenever an alert is generated. **Note** You must configure an email server to receive emails. For more information, contact on-site O&M engineers.
Notify User by DingTalk	No	If this option is selected, the Apsara Uni-manager Management Console informs the user configured as the alert contact by DingTalk whenever an alert is generated.

6. Click **OK**.

1.1.8.4.1.2. Query a user

You can view user information such as the name, organization, mobile number, email address, role, logon time, and initial password.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Enter a **username** in the search box or select an **organization**, and then click **Search**.

   You can also click **Advanced** on the right side and select a role or enter a display name to search for the user that you want to manage.
6. Click the username to view the basic information and the resource sets to which it belongs.

1.1.8.4.1.3. Modify user information

You can modify user information such as the display name, mobile number, and email address to keep it up to date.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Find the user that you want to modify and click **Edit** in the **Actions** column.
6. In the **Edit User** dialog box, modify the relevant information and click **OK**.

1.1.8.4.1.4. Change user roles

You can add, change, and delete roles for a user.

Change user roles by using user management

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Find the user that you want to modify and choose **More** > **Authorize** in the **Actions** column.
6. In the **Role** field, add, delete, or change user roles.

   You can enter role names in this field. Fuzzy match is supported.
7. Click **OK**.

Change user roles by changing ownership

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Changes**.
4. Click the  icon to the left of the organization that you want to manage and click **Users**.
5. In the right-side **Users** section, specify a **logon policy**, a **role**, or a **username**. Then, click **Search** to find the user that you want to change.
6. Find the user and click **Change** in the **Actions** column.
7. In the **Organization to Change** dialog box, select the original or a new organization and select the role to be added or removed from the **Assigned Roles** drop-down list.

> **Note** If you change only roles without changing the organization, select the original organization.

8. Click **OK**.

1.1.8.4.1.5. Modify the information of a user group

On the Users page, you can view the user group information and modify the ownership of users in user groups.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab, select the user that you want to modify, and then click **More** in the **Actions** column.
   - Select **Add to User Group**. In the dialog box that appears, select a user group and click **OK** to add the user to the user group.
   - Select **Remove from User Group**. In the dialog box that appears, select a user group and click **OK** to remove the user from the user group.

1.1.8.4.1.6. Modify a user logon policy

An administrator can modify a user logon policy to control the permitted logon time and IP addresses of the user.

Prerequisites

A new logon policy is created. For more information, see Create a logon policy.

Modify the logon policy of a single user

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Find the user that you want to modify and choose **More > Logon Policy** in the **Actions** column.
6. In the **Assign Logon Policy** dialog box, select a logon policy and click **OK**.

Modify multiple user logon policies at a time

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Select users that you want to modify.
6. In the lower-left corner of the page, click **Change Logon Policy**.
7. In the **Batch Assign Logon Policy** dialog box, select a logon policy and click **OK**.

1.1.8.4.1.7. View the initial password of a user

After a user is created, the system generates an initial password for the user.

Context

Organization administrators can view the initial passwords of all users in the organizations they manage.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Find the user that you want to manage and choose **More** > **View Initial Password** in the **Actions** column.

> **Note** After the initial password is changed, the initial password is not displayed.

1.1.8.4.1.8. Reset the password of a user

If users forget their logon passwords, the organization administrator can reset the logon passwords for them.

Prerequisites

Only organization administrators can reset the password of a user.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an organization administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the user for which you want to reset the password.
5. In the **User Information** panel, click **Reset password**.

   After the password is reset, a message is displayed, which indicates that the password has been reset.

1.1.8.4.1.9. Disable or enable a user account

You can disable a user account to prevent the user account from logging on to the Apsara Uni-manager Management Console. User accounts that are disabled must be re-enabled before they can be used to log on to the Apsara Uni-manager Management Console again.

Context

By default, user accounts are enabled when they are created.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Perform the following operations on the current tab:
   - Select a user account whose **Status** is **Enabled** and click **Disable** in the **Actions** column. In the **Are you sure that you want to disable the specified user** message, click **Disable** to disable the user account.
   - Select a user whose **Status** is **Disabled** and click **Enable** in the **Actions** column. In the **Are you sure you want to enable the specified user** message, click **Enable** to enable the user account.

1.1.8.4.1.10. Delete a user

You can delete a specific user as an administrator.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **System Users** tab.
5. Find the user that you want to delete and choose **More** > **Delete** in the **Actions** column.
6. In the **Confirm** message, click **OK**.

1.1.8.4.2. Historical users

1.1.8.4.2.1. Query historical users

You can query deleted users.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **Historical Users** tab.
5. Enter the username that you want to query in the **Username** search box and click **Search**.

> **Note** Fuzzy search is supported.

1.1.8.4.2.2. Restore historical users

An administrator can restore a deleted user account from the **Historical Users** tab.

Context

The basic information such as the logon password of a restored user is the same as it was before the user was deleted, except for its organization and role.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > Users**.
4. Click the **Historical Users** tab.
5. Find the user that you want to restore and click **Restore** in the **Actions** column.
6. In the **Restore User** dialog box, select an organization and a role.
7. Click **OK**.

1.1.8.5. Access control management

1.1.8.5.1. Create an access policy

To improve the security of the Apsara Uni-manager Management Console, you can create an access policy as an administrator to control logon access based on the logon time and IP addresses of users.

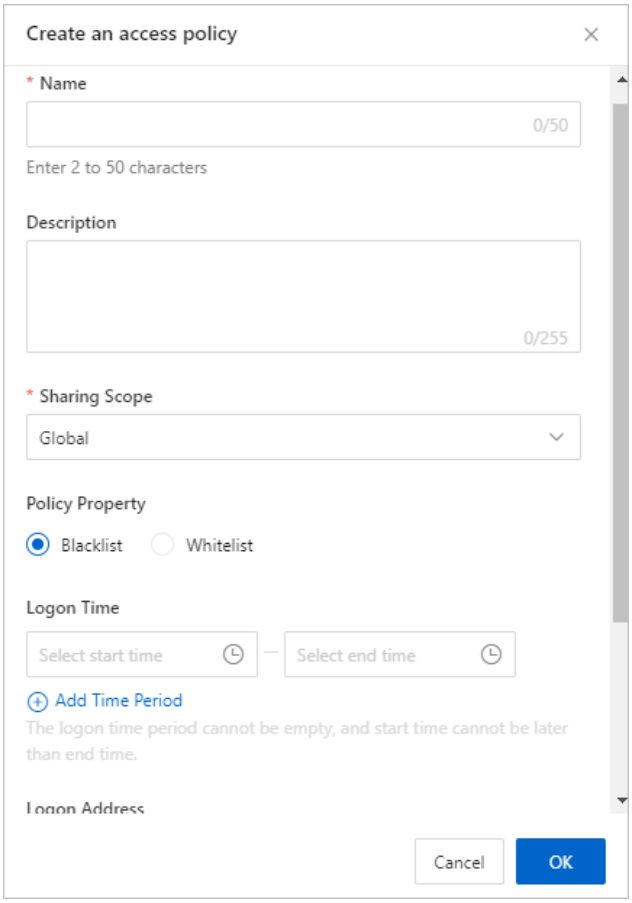
Context

Access policies are used to control the time periods and IP addresses valid for user logon. After a user is bound to an access policy, user logons are limited based on the logon time and IP addresses specified in the policy.

The default policy generated by the Apsara Uni-manager Management Console does not have limits on the time and IP addresses valid for user logon. This policy cannot be deleted.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Permission Policies**.
4. Click **Create an access policy** in the upper-right corner.
5. In the **Create an access policy** dialog box, set Name, Sharing Scope, Policy Property, Logon Time, and Logon Address.



Parameters for creating an access policy

Parameter	Description
Name	The name of the access policy. The name must be 2 to 50 characters in length and can contain only letters and digits. The name must be unique in the system.
Description	The description of the access policy.

Parameter	Description
Sharing Scope	The scope in which the role is visible. Valid values: ○ **Global**: The role is globally visible. The default value is Global. ○ **Current Organization**: The role is visible only in the current organization and is invisible in subordinate organizations. ○ **Subordinate Organization**: The role is visible in the current organization and all its subordinate organizations.
Policy Property	The authentication method of the access policy. Valid values: ○ **Whitelist**: Logon is allowed if the parameter settings are met. ○ **Blacklist**: Logon is denied if the parameter settings are met.
Logon Time	The permitted logon time period. If this policy is configured, users can log on to the Apsara Uni-manager Management Console only within the configured period. Specify the time period in minutes in a 24-hour clock. Example: `16:32`. **Note** When the Policy Property parameter is set to Whitelist, you can select No Time Limit.
Logon Address	The permitted CIDR block. ○ If the **Policy Property** parameter is set to **Whitelist**, IP addresses within this CIDR block are allowed to log on to the Apsara Uni-manager Management Console. ○ If the **Policy Property** parameter is set to **Blacklist**, IP addresses within this CIDR block are not allowed to log on to the Apsara Uni-manager Management Console. **Note** When the Policy Property parameter is set to Whitelist, you can select No CIDR Block Limit.

6. Click **OK**.

1.1.8.5.2. Query access policies

You can query the detailed information of an access policy in the Apsara Uni-manager Management Console.

Context

When the Apsara Uni-manager Management Console provides services, it automatically generates a default policy without limits on the logon time and IP addresses.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Permission Policies**.
4. Enter the name of the policy that you want to view and click **Search**.

   View the access policy, including the permitted logon time and IP addresses.

You can click the following icons in the upper-right corner to manage the access policy list.

Icon	Description
	Shows the list in full screen mode. You can click the icon in the upper-right corner to exit the full screen mode.
	Selects the size of cells and fonts in the list. Valid values: • Default • Compact
	Refreshes the access policy list.
	Selects or clears check boxes to determine the items to be displayed in the list. You can also click **Reset** to restore the list to its original settings.

1.1.8.5.3. Modify a logon policy

You can modify the policy name, policy properties, permitted logon time period, and IP addresses of a logon policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Permission Policies**.
4. Find the policy that you want to modify and click **Modify** in the **Actions** column.
5. In the **Modify an access policy** dialog box, modify the information of the policy.
6. Click **OK**.

1.1.8.5.4. Disable an access policy

You can disable access policies that are no longer needed.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Permission Policies**.
4. Find the access policy that you want to disable and click **Disable** in the **Actions** column.
5. In the message that appears, click **OK**.

1.1.8.5.5. Enable a logon policy

You can re-enable disabled logon policies.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Permission Policies**.

4. Find the policy that you want to enable and click **Enable** in the **Actions** column.

1.1.8.5.6. Delete an access policy

You can delete access policies that are no longer needed.

Prerequisites

The access policy to be deleted is not bound to users. If an access policy is bound to a user, the access policy cannot be deleted.

Context

> **Note** The default policy cannot be deleted.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Permission Policies**.
4. Find the policy that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

1.1.8.6. User groups

1.1.8.6.1. Create a user group

You can create a user group in a selected organization and batch authorize users within the group.

Prerequisites

An organization is created. For more information, see Create an organization.

Context

Relationships between user groups and users:

- A user group can contain zero or more users.
- A user does not need to belong to a user group.
- You can add a user to multiple user groups.

Relationships between user groups and organizations:

- A user group belongs to only a single organization.
- You can create multiple user groups within an organization.

Relationships between user groups and roles:

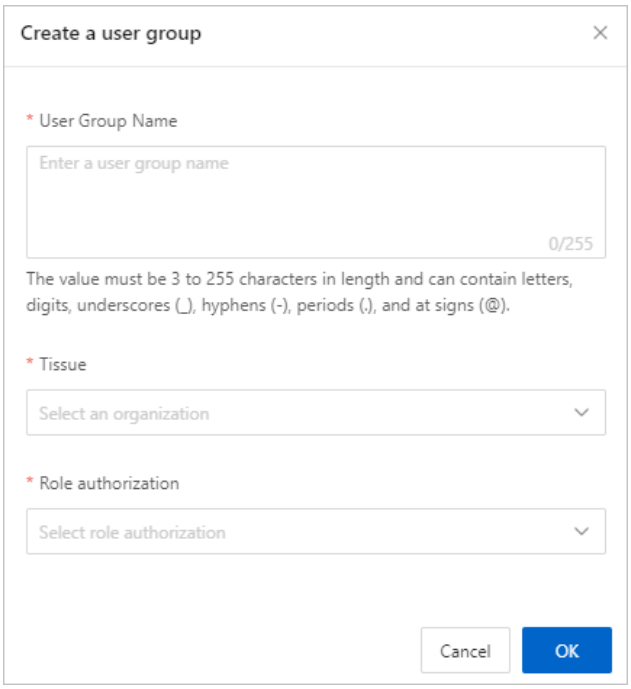
- A role can be assigned to multiple user groups.
- When a role is assigned to a user group, the permissions that the role has are automatically granted to users within the user group.

Relationships between user groups and resource sets:

- You can add zero or more user groups to a resource set.
- A user group can be added to multiple resource sets.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > User Groups**.
4. Click **Create a user group** in the upper-left corner.
5. In the dialog box that appears, set **User Group Name**, **Tissue**, and **Role authorization**.



Parameter	Description
User Group Name	The name of the user group. The name must be 3 to 255 characters in length and can contain letters, digits, underscores (_), hyphens (-), periods (.), and at signs (@).
Tissue	The organization to which the user group belongs.
Role authorization	The roles that are assigned to the user group.

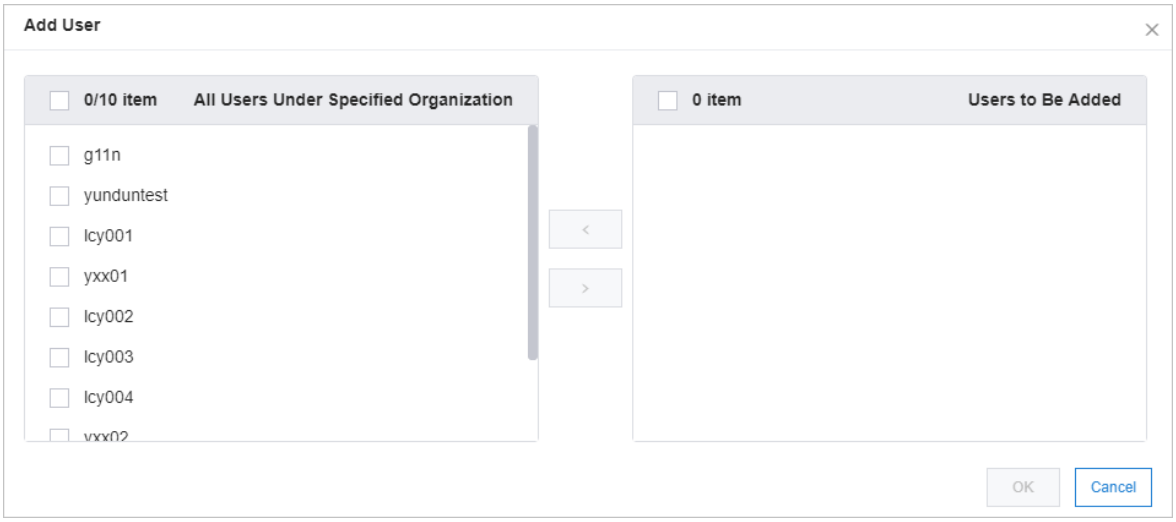
6. Click **OK**.

1.1.8.6.2. Add users to a user group

You can add users to a user group.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > User Groups**.
4. Find the user group to which you want to add users and click **Add User** in the **Actions** column.
5. Select users to be added in the left-side section and click the right arrow to move them to the right-side section.



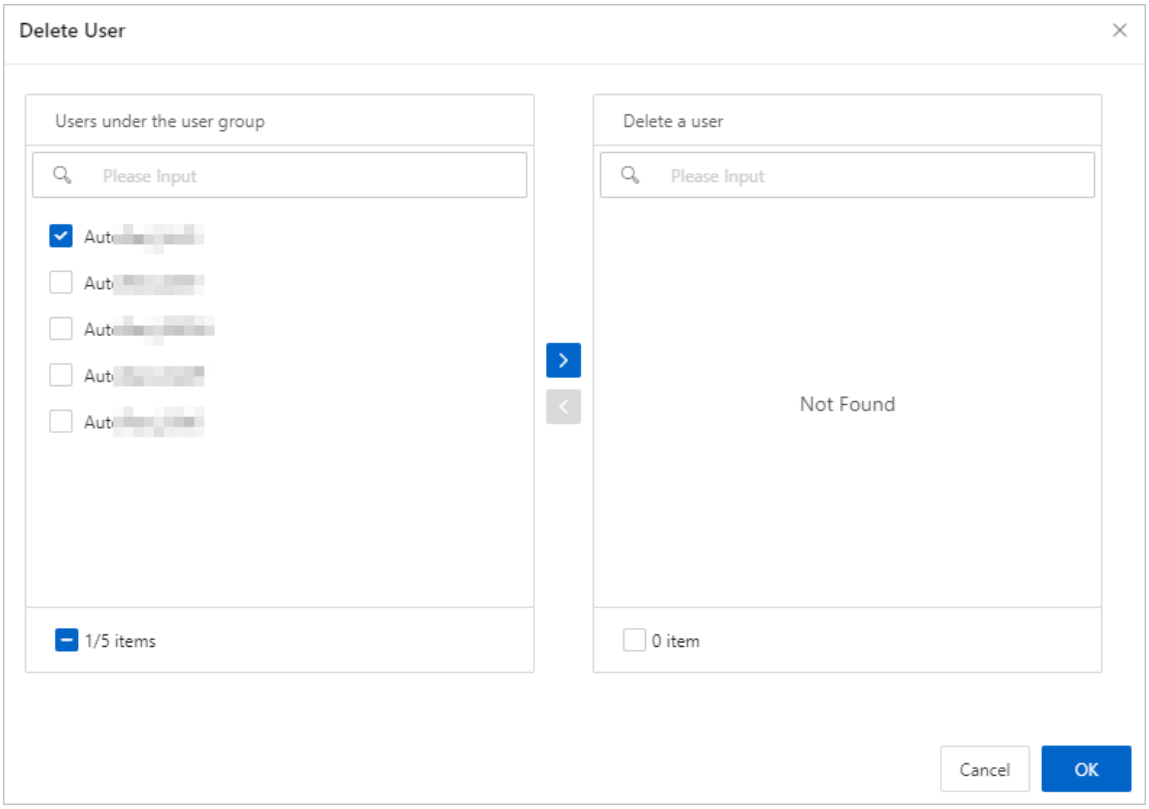
6. Click **OK**.

1.1.8.6.3. Remove a user

You can remove users from a user group.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > User Groups**.
4. Find the user group from which you want to remove a user and choose **More** > **Remove User** in the **Actions** column.
5. Select the user that you want to remove in the **Users under the user group** section and click the right arrow to move the user to the **Remove User** section.



6. Click **OK**.

1.1.8.6.4. Add or remove a role

You can add a role to a user group to grant the role permissions to all users within the group. You can also remove a role from a user group to revoke the role permissions from all users within the group.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > User Groups**.
4. Find the user group that you want to manage and choose **More** > **Authorized** in the **Actions** column.
5. In the **Authorized** dialog box, set **Role authorization**.
6. In the drop-down list, you can select or clear a role to grant permissions to or revoke permissions from the user group.
7. Click **OK**.

1.1.8.6.5. Modify the name of a user group

You can modify the names of user groups.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > User Groups**.
4. Find the user group that you want to modify and click **Editing** in the **Actions** column.

5. In the dialog box that appears, modify **User Group Name**.
6. Click **OK**.

1.1.8.6.6. Delete a user group

You can delete user groups that are no longer needed.

Prerequisites

The user group to be deleted is unbound from all roles. If a role is bound to a user group, the user group cannot be deleted.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users > User Groups**.
4. Find the user group that you want to delete and choose **More** > **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

1.1.8.7. Region management

1.1.8.7.1. Update associations

The Apsara Uni-manager Management Console allows you to manage resources across regions. You can update the associations between organizations and regions.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Regions**.
4. In the left-side organization navigation tree, click the name of the organization that you want to update.
5. In the right-side **Regions** section, select the names of regions to be associated.
6. Click **Update Association**.

1.1.8.8. Change the ownership of an instance

If you want to transfer resource instances to another resource set within the same organization, you can change the resource set ownership of the instances.

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Resources > Changes**.
4. Click the + icon to the left of the organization that you want to manage and click a resource set.
5. In the right-side **Product Type** section, click the name of the service that contains the instance you want to manage.
6. Set **Resource Type**, enter the instance ID or name in the search box, and then click **Search** to search for the instance.
7. Click **Change Ownership** in the **Actions** column corresponding to the instance to change the ownership of the instance to another resource set.

8. In the **Change Resource Set** dialog box, select the desired resource set and click **OK**.

1.1.8.9. Cloud instances

1.1.8.9.1. Manage Apsara Stack cloud instances

1.1.8.9.1.1. Export data of the current cloud

You can export the data of secondary Apsara Stack nodes to a configuration file. This can be used by the primary node to manage nodes in a centralized manner.

Procedure:

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **Apsara Stack Management** tab.
5. Click **Collect Data of Current Cloud** to collect the deployment information of the current cloud.
6. Click **Export** to export the information in the JSON format.

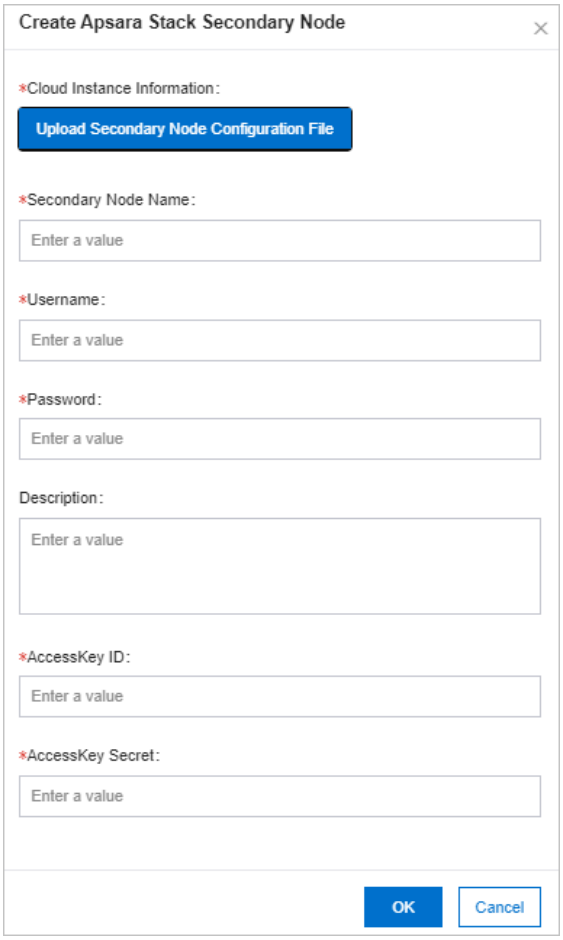
1.1.8.9.1.2. Add a secondary Apsara Stack node

You can add the configuration information of secondary Apsara Stack nodes to the multi-cloud configuration of the primary Apsara Stack node for centralized management.

Procedure:

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **Apsara Stack Management** tab.
5. Click **Import**.

6. In the **Create Apsara Stack Secondary Node** dialog box, enter the configuration information of a secondary node and click **OK**.



Parameter	Description
Cloud Instance Information	The configuration file of the secondary node. For more information, see Export data of the current cloud.
Secondary Node Name	The name of the secondary node.
Username	The username of the operations administrator that manages the secondary node.
Password	The password of the operations administrator that manages the secondary node.
Description	The description of the secondary node.

Parameter	Description
AccessKey ID	The AccessKey ID of the operations administrator that manages the secondary node. For more information, see View the AccessKey pair of your Apsara Stack tenant account.
AccessKey Secret	The AccessKey secret of the operations administrator that manages the secondary node. For more information, see View the AccessKey pair of your Apsara Stack tenant account.

> **Notice**
> You must create an operations administrator account in the secondary node. This account is for dedicated use by the primary node and cannot be the default operations administrator account.

1.1.8.9.1.3. View managed cloud instances

You can use the multi-cloud management feature to view the details of all managed cloud instances.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **Apsara Stack Management** tab.

   You can view the name, description, cloud type, cloud role, and address of all managed cloud instances.
5. Enter a cloud instance name in the search box and click **Search** to search for the cloud instance.
6. Click **View Details** in the **Actions** column corresponding to the cloud instance.

   In the Manage Cloud Instance message, you can view the version, Apsara Stack API (ASAPI) address, and region of the cloud.

1.1.8.9.1.4. Modify a cloud instance

If you want to change the information of a cloud instance for more efficient management, you can modify it in the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **Apsara Stack Management** tab.
5. Enter the name of the cloud instance that you want to modify in the search box and click **Search** to search for the cloud instance.
6. Click **Edit** in the **Actions** column corresponding to the cloud instance.

7. In the **Edit Cloud Instance** dialog box, set **Cloud Name**, **Username**, **Password**, **Description**, **AccessKey ID**, **AccessKey Secret**, **Longitude**, and **Latitude**, and click **OK**.

1.1.8.9.1.5. Manage cloud instances

You can manage Apsara Stack cloud instances to check whether they can be connected.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **Apsara Stack Management** tab.
5. Enter a cloud instance name in the search box and click **Search** to search for the cloud instance.
6. Click **Manage** in the **Actions** column corresponding to the cloud instance.
7. In the **Manage Cloud Instance** dialog box, click **Test Connectivity**.

1.1.8.9.2. Manage VMware nodes

1.1.8.9.2.1. Add a VMware node

You can add the configuration information of VMware nodes to the Apsara Stack VMware management configuration for centralized management.

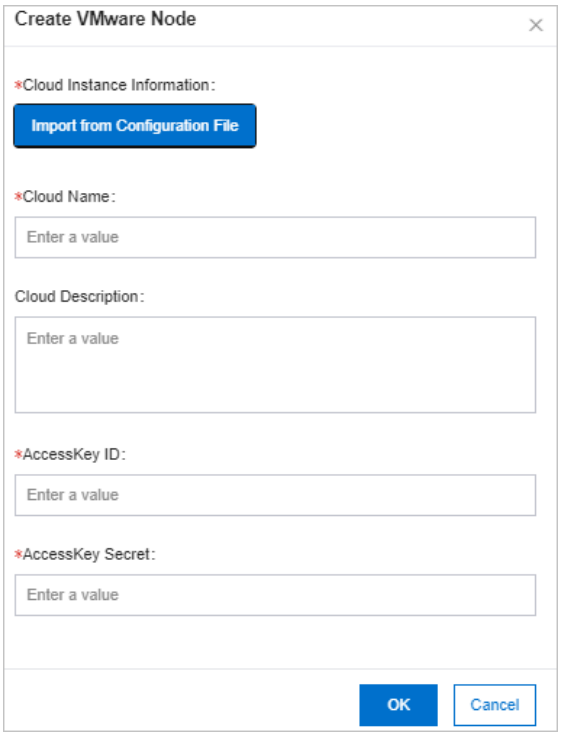
Prerequisites

- The configuration file of a VMware node is obtained from the deployment personnel.
- The VMware node is configured.

Procedure:

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **VMware Management** tab.
5. Click **Create VMware Node**.

6. In the **Create VMware Node** dialog box, enter the configuration information of a VMware node and click **OK**.



Parameter	Description
Cloud Instance Information	The configuration file of the VMware node.
Cloud Name	The name of the VMware node.
Cloud Description	The description of the VMware node.
AccessKey ID	The AccessKey ID in the configuration file of the VMware node.
AccessKey Secret	The AccessKey secret in the configuration file of the VMware node.

1.1.8.9.2.2. Modify a VMware node

If you want to change the information of a VMware node for more efficient management, you can modify it in the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **VMware Management** tab.

5. Enter the name of the VMware node that you want to modify in the search box and click **Search** to search for the VMware node.
6. Click **Edit** in the **Actions** column corresponding to the VMware node.
7. In the **Edit Cloud Instance** dialog box, set **Cloud Name**, **Cloud Description**, **AccessKey ID**, and **AccessKey Secret**, and click **OK**.

1.1.8.9.2.3. Test VMware node connectivity

You can manage VMware nodes to check whether they can be connected.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the Enterprise page, click **Cloud Instances**.
4. Click the **VMware Management** tab.
5. Enter a VMware node name in the search box and click **Search** to search for the VMware node.
6. Click **Manage** in the **Actions** column corresponding to the VMware node.
7. In the **Manage Cloud Instance** dialog box, click **Test Connectivity**.

1.1.8.10. Data permissions

1.1.8.10.1. Overview

Data permission management allows you to specify which users can access instances of a specific service, grant data access permissions to the users, and view and modify the data permissions in all the RAM policies attached to specified users.

Apsara Stack controls users and permissions by managing their visibility and operability in the Apsara Uni-manager Management Console. Many Apsara Stack cloud services are directly used by calling their API operations or SDKs instead of in the console. In this case, data access permissions must be controlled by RAM permission verification provided by the cloud services.

RAM policies are configured for such cloud service instances for access control. Automatic judgment is used when personnel are added to or removed from resource sets. However, this judgement method can affect performance and has a high error rate in complex scenarios. To solve this problem, the authorization of cloud services that require data access permissions is separately managed. Organization administrators can configure the data permissions granted to related personnel on the data authorization page.

1.1.8.10.2. Set the data permissions of resource instances

Organization administrators can set the data permissions of resource instances to allow or deny access to and operations on cloud services in the Apsara Uni-manager Management Console.

Prerequisites

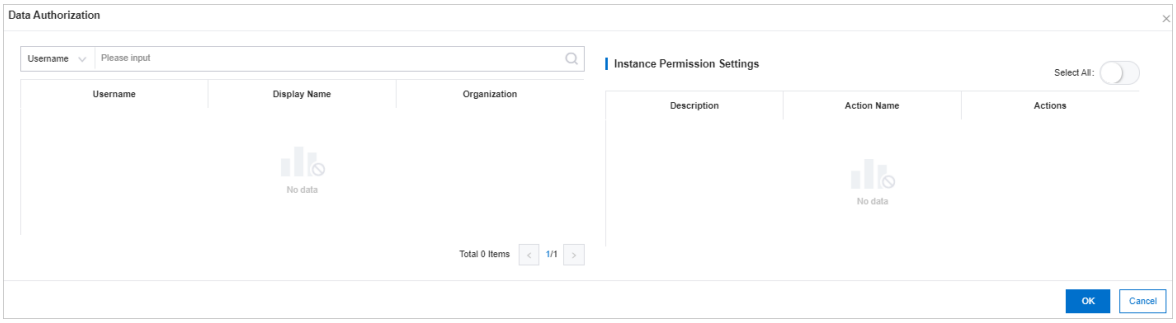
The cloud services that support data authorization include Message Queue (MQ), Object Storage Service (OSS), Log Service, DataHub, Container Service for Kubernetes, and Key Management Service (KMS).

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.

3. In the left-side navigation pane, choose **Permissions** > **Data Permissions**.
4. Click the resource set that you want to manage and click a service type on the right side of the page.
5. Find the instance that you want to manage and click **Authorize** in the **Actions** column.
6. In the **Data Authorization** dialog box, select the user that you want to manage on the left side.
7. Turn on or off the data permission switches in the Actions column on the right side.

   You can also select or clear Enable All to batch manage permissions.



8. Click **OK**.

1.1.8.10.3. Edit user permissions

You can use JSON statements to edit user permissions.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions** > **Data Permissions**.
4. In the organization navigation tree, click the icon to the left of the organization that contains the user you want to manage.
5. Click **Users**.
6. Enter the username in the search box and click **Search**.
7. Click **Edit Permissions** in the **Actions** column corresponding to the user.
8. In the Edit Permissions dialog box, select a data permission on the left side and click OK.

   If no permissions are available, specify a policy in the text editor. For more information about the syntax and structure of a policy, see Permission policy structure and syntax.

1.1.8.10.4. View the permissions of a user

You can view the existing permissions of a user.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions** > **Data Permissions**.
4. In the organization navigation tree, click the icon to the left of the organization that contains the user you want to manage.

5. Click **Users**.
6. Enter the username in the search box and click **Search**.
7. Click **View Permissions** in the **Actions** column corresponding to the user.

1.1.9. Configurations

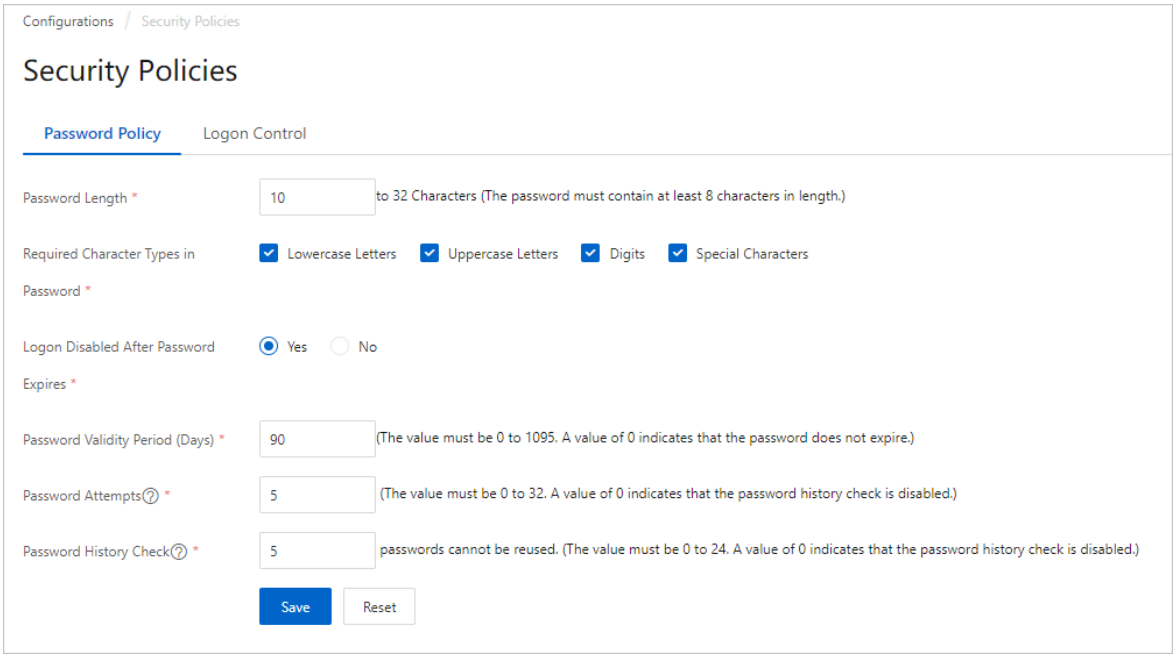
1.1.9.1. Security policies

1.1.9.1.1. Configure password policies

You can configure password policies for user logons.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Security Policies**.
4. Click the **Password Policy** tab.
5. On the **Password Policy** tab, set the password policy parameters.



Parameter	Description
Password Length	The minimum length of the password. Minimum value: 8.
Required Character Types in Password	The character types that must be contained in the password. You can select multiple options.
Logon Disabled After Password Expires	Specifies whether to allow logon after the password expires.
Password Validity (Days)	The validity period of the password. Unit: days. Maximum value: 1095. A value of 0 indicates that the password never expires.

Parameter	Description
Password Attempts	The maximum number of logon attempts within 1 hour. Maximum value: 32. After the maximum number is reached, the account is locked and cannot be used to log on for a specific period of time.
Password History Check	The number of previous passwords that cannot be reused. Maximum value: 24. A value of 0 indicates the password history check is disabled.

6. Click **Save**.

   To restore to the default password policy, click **Reset**.

1.1.9.1.2. Configure logon control

You can limit the access from multiple clients of users.

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Security Policies**.
4. Click the **Logon Control** tab.
5. Perform the following operations:
   - Select **Single Session**. A single session means that a user is allowed to log on only by using a single client at the same time.
   - Select **Multi-session**. A multi-session means that a user is allowed to log on by using multiple clients at the same time.

1.1.9.2. Menus

1.1.9.2.1. Create a menu

You can create a menu and add its URL to the Apsara Uni-manager Management Console for quick access.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Menu Settings**.
4. On the **Main Menu** page, click **Create** in the upper-right corner.
5. In the **Create** dialog box, configure the menu parameters.

   Menu parameters

Parameter	Description
Title	The display name of the menu.
URL	The URL of the menu.

Parameter	Description
Console Type	Different console types correspond to different domain names. Valid values: ◦ **oneconsole**: You need only to enter the path in the URL field. The domain name is automatically matched. The URL starts with `one.console`. ◦ **asconsole**: You need only to enter the path in the URL field. The domain name is automatically matched. The URL starts with `asc.`. ◦ **asconsole 2.0**: You need only to enter the path in the URL field. The domain name is automatically matched. The URL starts with `one.console`. ◦ **other**: You must enter the domain name in the URL field.
Icon	The icon displayed in the left-side navigation pane. The icon cannot be changed.
Identifier	The unique identifier of the menu in the system. This identifier can be used to indicate whether the menu is selected in the navigation bar. The identifier cannot be changed.
Sort	The display order among the same-level menus. The larger the value, the lower the display order. Leave the Order field empty.
Parent Level	The displayed tree structure.
Open With	Specifies whether to open the menu in the current window or in a new window.
Description	The description of the menu.

6. Click **OK**.

1.1.9.2.2. Modify a menu

You can modify an existing menu, including the menu name, URL, icon, and menu order.

Prerequisites

Default menus cannot be modified.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Menu Settings**.
4. Find the menu that you want to modify and click **Edit** in the **Actions** column.
5. In the **Edit** dialog box, modify the information of the menu.

   For information about the menu parameters, see Create a menu.
6. Click **OK**.

1.1.9.2.3. Delete a menu

You can delete menus that are no longer needed.

Prerequisites

Default menus cannot be deleted.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.

2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane of the **Configurations** page, click **Menu Settings**.
4. Find the menu that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

1.1.9.2.4. Show or hide menus

You can perform the following operations to show or hide menus:

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane of the **Configurations** page, click **Menu Settings**.
4. Select or clear the check box in the **Displayed** column corresponding to a menu.

1.1.9.3. Specifications

1.1.9.3.1. Specification parameters

This topic describes the specification parameters of each resource type.

NAT Gateway

Parameter	Description
Specifications	The specifications that can be configured for NAT Gateway.
Specifications Description	The description of the specifications that can be configured for NAT Gateway.

AnalyticDB for PostgreSQL

Parameter	Description
Specifications	The specifications that can be configured for AnalyticDB for PostgreSQL.
Specifications Name	The name of the instance type that can be configured for AnalyticDB for PostgreSQL.
CPU	The total number of CPU cores that can be configured for AnalyticDB for PostgreSQL.
Memory	The memory size that can be configured for AnalyticDB for PostgreSQL.
Storage Space	The total storage size that can be configured for AnalyticDB for PostgreSQL.
Version	The version number of AnalyticDB for PostgreSQL.
Node	The number of nodes that can be configured for AnalyticDB for PostgreSQL.

SLB

Parameter	Description
Specifications	The instance type that can be configured for Server Load Balancer (SLB).
Specifications Name	The name of the instance type that can be configured for SLB.
Maximum Connections	The maximum number of connections that can be configured for SLB.
New Connections	The number of new connections that can be configured for SLB.
QPS	The queries per second (QPS) that can be configured for SLB.
Description	The description of the specifications that can be configured for SLB.

ApsaraDB RDS

Parameter	Description
Engine Type	The engine type that can be configured for ApsaraDB RDS.
Minimum Storage (GB)	The minimum amount of storage space that can be configured for ApsaraDB RDS.
Maximum Storage (GB)	The maximum amount of storage space that can be configured for ApsaraDB RDS.
Specifications Name	The name of the instance type that can be configured for ApsaraDB RDS.
Version	The version number of ApsaraDB RDS.
CPUs	The number of CPU cores that can be configured for ApsaraDB RDS.
Maximum Connections	The maximum number of connections that can be configured for ApsaraDB RDS.
Memory (GB)	The memory size that can be configured for ApsaraDB RDS.
Share Type	The share type that can be configured for ApsaraDB RDS.

PolarDB-X

Parameter	Description
Instance Type	The instance type that can be configured for PolarDB-X.
Instance Type Name	The name of the instance type that can be configured for PolarDB-X.

Parameter	Description
Specifications	The specifications that can be configured for PolarDB-X.
Specifications Name	The name of the specifications that can be configured for PolarDB-X.

ECS

Parameter	Description
Instance Family	The instance family that is divided into different instance types based on the scenarios for which they are suitable.
Specifications Level	The level of the instance type that can be configured for Elastic Compute Service (ECS).
vCPUs	The maximum number of vCPUs that can be configured for ECS.
Memory (GB)	The memory size that can be configured for ECS.
Instance Specifications	The instance type that can be configured for ECS.
GPU Type	The GPU type that can be configured for ECS.
GPUs	The number of GPUs that can be configured for ECS.
Supported ENIs	The number of elastic network interface (ENIs) that can be configured for ECS.
Number Of Private IP Addresses	The number of private IP addresses that can be configured for ECS.

KVStore for Redis

Parameter	Description
Specifications Name	The name of the specifications that can be configured for KVStore for Redis.
Instance Specifications	The instance type that can be configured for KVStore for Redis.
Maximum Connections	The maximum number of connections that can be configured for KVStore for Redis.
Maximum Bandwidth	The maximum bandwidth that can be configured for KVStore for Redis.
CPUs	The number of CPU cores that can be configured for KVStore for Redis.
Version	The version number of KVStore for Redis.
Architecture	The architecture of KVStore for Redis.
Node Type	The node type of KVStore for Redis.

Parameter	Description
Service Plan	The service plan that can be configured for KVStore for Redis.

ApsaraDB for MongoDB

Parameter	Description
Specifications	The instance type that can be configured for ApsaraDB for MongoDB.
Instance Specifications	The name of the specifications that can be configured for ApsaraDB for MongoDB.
Engine Type	The engine type that can be configured for ApsaraDB for MongoDB.
Version	The version number of ApsaraDB for MongoDB.
Serial Number	The serial number of ApsaraDB for MongoDB.
Sequence Description	The description of the serial number of ApsaraDB for MongoDB.
Maximum Connections	The maximum number of connections that can be configured for ApsaraDB for MongoDB.
IOPS	The IOPS of ApsaraDB for MongoDB.
Storage Space	The amount of storage space that can be configured for ApsaraDB for MongoDB.
Minimum Storage	The minimum amount of storage space that can be configured for ApsaraDB for MongoDB.
Maximum Storage	The maximum amount of storage space that can be configured for ApsaraDB for MongoDB.

1.1.9.3.2. Create specifications

You can customize specifications for each resource type.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Specifications**.
4. Select the resource type for which you want to create specifications and click the **Resource Specifications** tab.
5. On the **Resource Specifications** tab, click **Create Specifications** in the upper-left corner.
6. In the dialog box that appears, configure the specifications parameters.

   For more information, see Specification parameters.
7. Click **OK**.

1.1.9.3.3. View specifications

You can view the specifications of each resource type.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane of the **Configurations** page, click **Specifications**.
4. Click the resource type for which you want to view specifications.
5. On the **Resource Specifications** tab, set a **region**, **column**, and **value**. The corresponding information is displayed in the specifications list.
6. Click the **Existing Specifications** tab and view the existing specifications and their quantity.

1.1.9.3.4. Disable specifications

By default, newly created specifications are in the Enabled state.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane of the **Configurations** page, click **Specifications**.
4. Select the resource type for which you want to disable specifications.
5. Click **Disable** in the **Actions** column corresponding to the specifications that you want to disable.
6. In the message that appears, click **OK**.

1.1.9.3.5. Export specifications

You can export specifications that you want to view and share.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Specifications**.
4. Select the resource type for which you want to export specifications.
5. Click the **Resource Specifications** tab and click **Export** in the upper-right corner to save the specifications to your computer as a CSV file.

1.1.9.3.6. View specifications of each resource type in previous versions

You can view specifications of each resource type in previous versions.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane of the **Configurations** page, click **Specifications**.

4. On the **Specifications** page, click the resource type for which you want to view specifications.
5. Click the **Specifications History** tab. View the detailed information in the specifications list.

1.1.9.4. Message center

1.1.9.4.1. View internal messages

You can view all internal messages, including read and unread messages.

Context

When changes are made to or alerts are generated for an instance in a resource, all users that have read and operation permissions on this resource receive corresponding messages.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, move the pointer over the icon and click **More**.
3. In the left-side navigation pane of the **Message Center** page, click the target message scope.
   - Choose **Internal Messages > All Messages** to view all messages, including unread and read messages.
   - Choose **Internal Messages > Unread Messages** to view unread messages.
   - Choose **Internal Messages > Read Messages** to view read messages.

1.1.9.4.2. Mark messages as read

You can mark unread messages as read messages to facilitate message management.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, move the pointer over the icon and click **More**.
3. In the left-side navigation pane of the **Message Center** page, choose **Internal Messages > Unread Messages**.

   In the upper part of the **Unread Messages** page, click different message types to filter messages.
4. On the Unread Messages page, find the message that you want to mark as read and click **Mark as Read** in the **Actions** column.

   You can also select the check boxes to the left of messages and click **Batch Read** in the upper-left corner of the page.
5. In the **Mark as Read** message, click **OK**.

1.1.9.4.3. Delete a message

You can delete messages that are no longer needed.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, move the pointer over the icon and click **More**.
3. In the left-side navigation pane of the **Message Center** page, choose **Internal Messages > All Messages**.

4. Find the message that you want to delete on the All Messages tab or other tabs and click **Delete**.

   You can also select the check boxes to the left of messages and click **Batch Delete** in the upper-left corner of the page.

1.1.9.5. Resource pool management

You can modify the maximum usage of each resource.

Prerequisites

- If the physical inventory is unlimited, the logical inventory cannot be less than the used inventory.
- If the physical inventory is limited, the logical inventory cannot be less than the used inventory and cannot be greater than the physical inventory.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Resource Pool Configuration**.
4. On the right side of the **Resource Pools** page, click the service that you want to modify.

   You can also click the name of a service on the left side of the page.
5. Find the service that you want to modify, and click the icon in the Logical Inventory column, and then modify resource quantities of the service.
6. Click the icon to complete modification.

1.1.9.6. Custom configurations

1.1.9.6.1. Configure brands

You can customize the icon, platform name, and logon page of the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a platform administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Custom Configurations**.
4. In the **Brand Settings** section, click the language tab of the page that you want to modify.
5. Configure the following parameters and click **OK**.

Parameter	Description

Parameter	Description
Browser ICO Image	You can click **Upload Image**, select an image file, and then click **Open**. We recommend that you select an image whose resolution is 160 × 36 pixels in the PNG or JPG format. After the image is uploaded, you can click the icon to preview it. Before you upload a new image, you must click the icon to delete the existing image.
Platform Logo	You can click **Upload Image**, select an image file, and then click **Open**. We recommend that you select an image whose resolution is 160 × 36 pixels in the PNG or JPG format. After the image is uploaded, you can click the icon to preview it. Before you upload a new image, you must click the icon to delete the existing image.
Platform Name	You can customize a platform name. The name can be up to 8 characters in length.
Logon Page	You can click + Add Screen to increase the number of scrolling screens. You can also click the icon to the left of a screen to decrease the number of scrolling screens. One to three scrolling screens can be allowed on the logon page. After you click the tab of a screen, you can perform the following operations: ◦ **Background Image**: Click Upload Image, select a background image file, and then click Open. We recommend that you select an image whose resolution is 1,880 × 1,600 pixels in the PNG or JPG format. After the image is uploaded, you can click the icon to preview it. Before you download a new image, you must click the icon to delete the existing image. ◦ **Image Copywriting**: Customize the image content. The image content can be up to 40 characters in length.
Copyright Notice	You can customize the information of the copyright notice.

1.1.10. Operations

1.1.10.1. Quotas

1.1.10.1.1. Quota parameters

This topic describes the quota parameters of each service.

An organization administrator can set resource quotas and create resources within the allowed quotas for its organization. When the quotas for the organization are used up, the system does not allow the organization administrator to create more resources for the organization. To create more resources, you must first increase the quotas for the organizations.

If no quotas are set, you can create an unlimited amount of resources.

ECS

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for Elastic Compute Service (ECS) and the number of used cores.
Memory quota (G)	The total memory size that you can configure for ECS.
GPU quota (number of cards)	The total number of GPU cores that you can configure for ECS.
SSD quota (G)	The total SSD capacity that you can configure for ECS.
Efficient cloud disk quota (G)	The total number of disks that you can configure for an ECS instance.

VPC

Parameter	Description
VPC quota (units)	The maximum number of virtual private clouds (VPCs) that you can configure.

OSS

Parameter	Description
OSS quota (G)	The maximum capacity that you can allocate for Object Storage Service (OSS).

ApsaraDB RDS for MySQL

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for ApsaraDB RDS for MySQL and the number of used cores.
Memory quota (G)	The total memory size that you can configure for ApsaraDB RDS for MySQL.
Disk quota (G)	The total storage size that you can configure for ApsaraDB RDS for MySQL.

PolarDB

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for PolarDB and the number of used cores.
Memory quota (G)	The total memory size that you can configure for PolarDB.
Disk quota (G)	The total storage size that you can configure for PolarDB.

ApsaraDB RDS for SQL Server

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for ApsaraDB RDS for SQL Server and the number of used cores.
Memory quota (G)	The total memory size that you can configure for ApsaraDB RDS for SQL Server.
Disk quota (G)	The total storage size that you can configure for ApsaraDB RDS for SQL Server.

ApsaraDB RDS for PostgreSQL

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for ApsaraDB RDS for PostgreSQL and the number of used cores.
Memory quota (G)	The total memory size that you can configure for ApsaraDB RDS for PostgreSQL.
Disk quota (G)	The total storage size that you can configure for ApsaraDB RDS for PostgreSQL.

SLB

Parameter	Description
VIP quota (units)	The maximum number of internal IP addresses that you can configure for Server Load Balancer (SLB).
Public network VIP quota (units)	The maximum number of public IP addresses that you can configure for SLB.

EIP

Parameter	Description
EIP quota (units)	The maximum number of elastic IP addresses (EIPs) that you can configure.

MaxCompute

Parameter	Description

Parameter	Description
CU quota (units)	The total number of capacity units (CUs) that you can configure for MaxCompute.
Disk quota (G)	The total storage size that you can configure for MaxCompute.

KVStore for Redis

Parameter	Description
Memory quota (G)	The total memory size that you can configure for KVStore for Redis.

PolarDB-X

Parameter	Description
CPU quota (core)	The total number of CPUs that you can configure for PolarDB-X.

AnalyticDB for PostgreSQL

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for AnalyticDB for PostgreSQL and the number of used cores.
Memory quota (G)	The total memory size that you can configure for AnalyticDB for PostgreSQL.
Disk quota (G)	The total storage size that you can configure for AnalyticDB for PostgreSQL.

ApsaraDB for MongoDB

Parameter	Description
CPU quota (core)	The total number of CPU cores that you can configure for ApsaraDB for MongoDB and the number of used cores.
Memory quota (G)	The total memory size that you can configure for ApsaraDB for MongoDB.
Disk quota (G)	The total storage size that you can configure for ApsaraDB for MongoDB.

Dedicated hosts

Parameter	Description
DDH quota (units)	The number of dedicated hosts.

1.1.10.1.2. Set quotas for a cloud service

You can set quotas for each organization to allocate and manage resources for each organization in an appropriate manner.

Prerequisites

You must set quotas for a parent organization before you can set quotas for its subordinate organizations.

Context

If the parent organization (not including level-1 organizations) has quotas, the available quotas for a subordinate organization are equal to the quotas for the parent organization minus the quotas for other subordinate organizations.

This topic describes how to set quotas for Elastic Compute Service (ECS). You can set quotas for other cloud resources in a similar manner.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an organization administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, click **Quotas**.
4. In the **Organizational Architecture** section, click the name of the organization that you want to manage.
5. Click **ECS**.
6. In the upper-right corner of the quota section, click the  icon.
7. Set the total quotas and click **Save**.

   For information about quota parameters, see Quota parameters.

1.1.10.1.3. Modify quotas

Administrators can adjust quotas for cloud resources based on their organizational requirements.

Context

This topic describes how to modify quotas for Elastic Compute Service (ECS). You can modify quotas for other cloud resources in a similar manner.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an organization administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, click **Quotas**.
4. In the **Organizational Architecture** section, click the name of the organization that you want to manage.
5. Click **ECS**.
6. In the upper-right corner of the quota section, click the  icon.
7. After you modify the quotas, click **Save**.

   For more information about specification parameters, see Quota parameters.

1.1.10.1.4. Reset quotas

Administrators can reset quotas to remove quota limits on cloud resources.

Prerequisites

Before you delete a quota for an organization, make sure that no subordinate organizations have quotas.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an organization administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, click **Quotas**.
4. In the **Organizational Architecture** section, click the name of the organization that you want to manage.
5. Click the cloud service for which to reset quotas.
6. In the upper-right corner of the quota section, click the icon.
7. In the message that appears, click **OK**.

1.1.10.2. Usage statistics

1.1.10.2.1. View the usage statistics of cloud resources

The Apsara Uni-manager Management Console shows the number of resource instances that run in the Apsara Stack environment by time, organization, resource set, and region. You can also export statistical reports from the Apsara Uni-manager Management Console.

Context

The following cloud resources can be measured: Elastic Compute Service (ECS), Virtual Private Cloud (VPC), Server Load Balancer (SLB), Object Storage Service (OSS), ApsaraDB RDS for MySQL, Elastic IP Address (EIP), Tablestore, PolarDB-X, KVStore for Redis, AnalyticDB for PostgreSQL, ApsaraDB for MongoDB, ApsaraDB RDS for PostgreSQL, ApsaraDB RDS for SQL Server, Log Service, ECS disks, ECS snapshots, scaling group rules, API gateways, Key Management Service (KMS), dedicated hosts, and ApsaraDB for OceanBase.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, click **Usage Statistics**.
4. Click **Elastic Compute Service**.
5. In the **query condition** section, specify a **time range** and click **Search**.

   You can also click **Advanced** in the upper-right corner, set **Tissue**, **Resource set**, **Geographical**, and **Instance ID**, and then click **Search**.

   > **Note** In the console, you can view or export up to 1,000 statistical records to an XLS file. If you want more statistical data, you can call the MeteringQuery operation.

6. (Optional)Click the icon to save the displayed information to your computer as an XLS file.

   The exported file is named *<Resource type name>.xls*. Find the downloaded file from the download path of the browser.

1.1.10.3. Statistical analysis

1.1.10.3.1. View reports of current data

You can use reports to view the most recent resources, quotas, and CloudMonitor data of each service.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Statistical Analysis** > **Reports**.
4. Click the tab that you want to view.

   You can click the **Resource Report**, **Quota report**, or **Cloudmonitor report** tab.
5. (Optional) Set **Organization and resource set** and **Geographical** and click **Search**.
6. Click the service that you want to view.

   View the most recent data of the service.

1.1.10.3.2. Export reports of current data

You can batch export the resources, quotas, or monitoring data that you want to view based on cloud service types.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Statistical Analysis** > **Reports**.
4. Click the tab that you want to view.

   You can click the **Resource Report**, **Quota report**, or **Cloudmonitor report** tab.
5. (Optional) Set **Organization and resource set** and **Geographical** for which you want to view data and click **Search**.

   Click **Reset** to clear all filter properties.
6. Use one of the following methods to export data:
   - Export data by service.
     a. Click the  icon on the right side.
     b. In the **Select the product to export** dialog box, select the service that you want to view and click **OK**.

        You can also select **Select All** in the lower-left corner and click **OK**.
   - Export data by instance.
     a. Click the name of the service that you want to view.
     b. Select the instance that you want to view and click **Export Selected reports** in the lower-left corner of the page.

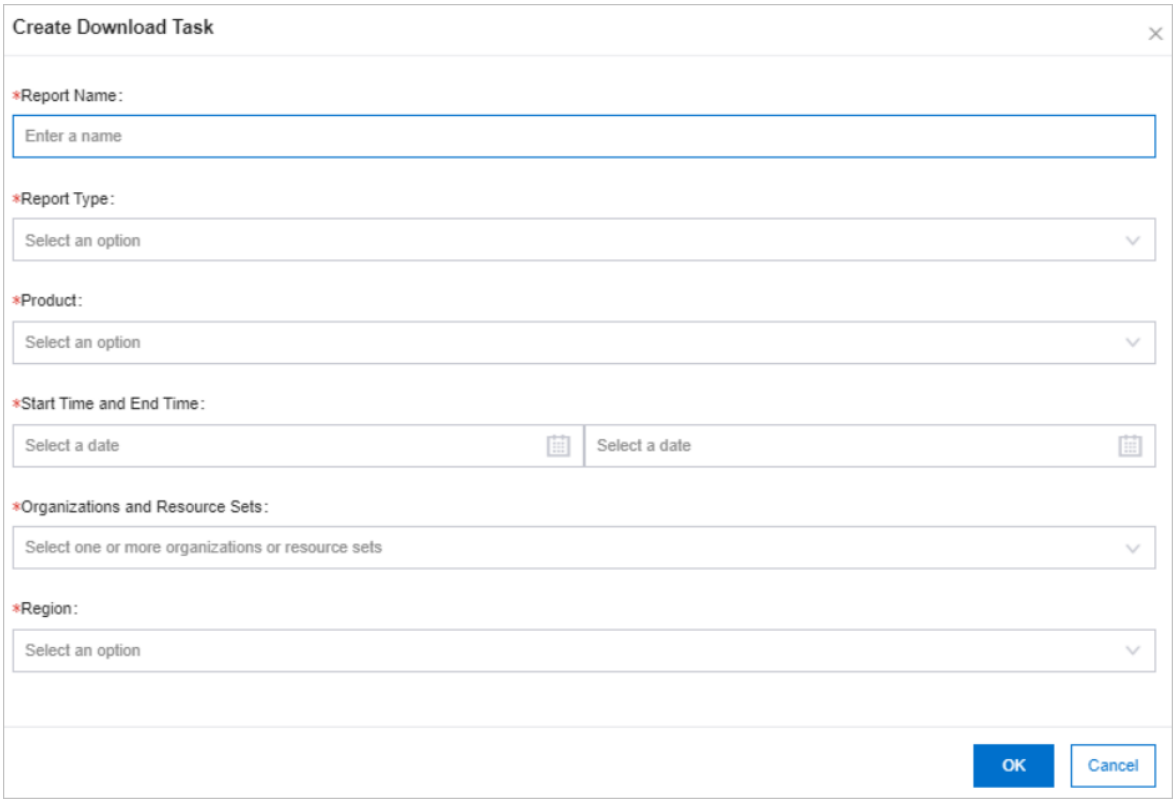
1.1.10.3.3. Download reports of historical data

You can download data reports of cloud services within the specified period of time, resource set, and region by creating download tasks.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane of the Operations page, choose **Statistical Analysis** > **Download Center**.

4. Click the icon in the upper-right corner.
5. Enter the information of the download task.



Parameter	Description
Report name	The name of the report.
Report type	The type of the report. Valid values: ○ Resource Report ○ Cloudmonitor report
Product	The cloud service for which you want to download reports. You can select multiple cloud services.
Start time and end time	The start time and end time of the data.
Organization Resource set	The organization to which the data belongs. You can select multiple organizations.
Area	The region of the data. You can select multiple regions.

6. Click **OK**.

7. After the **Creation successful** message appears, the Download Center page appears. Enter the information of the created report in the search box and click **Search** to search for the created download task.
8. After **In progress** changes to **Completed** in the **State** column, click **Download reports** in the Operation column.

1.1.10.4. Billing management

1.1.10.4.1. Billing overview

The billing system collects resource usage statistics for each service, organization, or resource set per month. You are charged for the consumed resources based on the billing rules.

Billing configurations

Billing configurations involve four terms to meet complex billing requirements. These configurations cover a variety of aspects such as basic prices, time policies, organization policies, and discount policies.

The following terms from bottom up are used for different billing logics:

- Billable item: specifies prices and billing methods.
- Billing rule: specifies the billing range of services.
- Billing policy: specifies organization policies and discount capabilities of services.
- Billing plan: specifies billing time policies and discount capabilities.

Billing overview

On the Billing overview page, you can view the numbers of billing plans, policies, and rules that are available and in use. You can also view the effective time of billing plans and policies.

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Overview**.

   In the upper part of the Billing overview page, you can view the numbers of billing plans, policies, and rules that are available and in use in the **Billing plan**, **Billing Policy**, and **Billing rules** sections.
4. In the lower part of the page, the effective time of billing plans and policies is displayed in a Gantt chart. You can select an option from the drop-down list in the upper-right corner to change the time span of the chart.
5. You can slide the Gantt chart left and right to view whether billing plans and policies are effective within different periods of time. Click **LocateDay** in the upper-right corner of the page to reset the time to the current day.

1.1.10.4.2. Billable items

1.1.10.4.2.1. Create a billable item

Billable items define the billing methods of services. This topic describes how to create a billable item.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Rules**.
4. Click the billing rule to which you want to add a billable item.

5. On the **Billing rule details** page, click **Create a billing item**.
6. In the **Create a billing item** panel, set the following parameters.

Parameter	Required	Description
Billing item name	Yes	The name of the custom billable item. The name can be up to 64 characters in length and must be unique.
Billing item description	No	The description of the custom billable item. The description can be up to 256 characters in length.
Unit price information settings	Yes	The pricing logic of a specific service feature within a unit quantity. Valid values: ○ **Fixed pricing**: You are charged for each instance on an hourly basis at a fixed rate that does not change with variables within a unit quantity. For example, the storage feature does not have properties such as specifications. You are charged at a fixed rate for each GB of stored data. This pricing logic is fixed pricing. ○ **Pricing**: Service prices change with properties such as specifications. Prices vary based on types or property values. For example, Elastic Compute Service (ECS) provides a variety of instance types at different prices. Metering fields and values can be found in the metering data. The system obtains the values of specific metering fields from the metering data to determine the unit price of an instance. Additionally, the prices of some services are not entirely determined by specifications. For example, the prices of resources with identical specifications may vary based on their region. In metering pricing mode, prices are determined by a combination of **Primary pricing** and **From pricing**, which is similar to MySQL composite indexes. After you set **Primary pricing** and **From pricing**, you can click **Download pricing file**. Then, you can edit the billable items and prices in the file and click **Upload pricing file** to upload the file again. **Note** A maximum of two fields can be combined to determine prices.

Parameter	Required	Description
Quantity information settings	Yes	The quantity of resources used in a service within 1 hour. Valid values: ○ **Fixed quantity**: The number of instances is fixed in each metering data record. For example, you are charged for an ECS instance based on its instance type. Each metering data record represents an instance. Then, the quantity is 1. ○ **Metering quantity**: The number of instances is defined in each metering data record. A field in the metering data must be used to specify the quantity. During billing, the system obtains the value of this field in the metering data and calculates fees. For example, for Object Storage Service (OSS), the value of the Storage field in the metering data indicates the amount of used storage space. If you are charged for the amount of used storage space, the Storage field must be specified to define the quantity property. Conversion factor: The unit used to measure metering data may not be the same as that used to determine the price. For example, the storage price is calculated based on the amount of GBs consumed in OSS. In the metering data, the unit of the Storage value may be byte. Therefore, a conversion factor must be declared in the definition of each quantity. You can divide a quantity in the metering data by a conversion factor to convert the original unit to the unit in a price.
Billing unit	Yes	The unit of the custom billable item. Example: USD/hour.

7. Click **Determine**.

1.1.10.4.2.2. Clone a billable item

You can clone an existing billable item to create the same or a similar billable item.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Rules**.
4. Click the name of the billing rule that you want to manage.
5. In the **Billing item** section, find the billable item that you want to clone and click **Clone** in the **Operation** column.
6. In the **Create a billing item** panel, modify parameters of the billable item.

   For more information about billable item parameters, see Create a billable item.
7. Click **Determine**.

1.1.10.4.2.3. Modify a billable item

This topic describes how to modify a billable item.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Rules**.
4. Click the name of the billing rule that you want to manage.
5. In the **Billing item** section, find the billable item that you want to modify and click **Editing** in the **Operation** column.
6. In the **Modify billing items** panel, modify parameters of the billable item.
7. Click **Determine**.

> **Warning**
>
> If you modify billing items that are in use, billing data may change. This change cannot be undone. Proceed with caution.

1.1.10.4.2.4. Delete a billable item

You can delete billable items that are incorrect or no longer needed.

Prerequisites

At least one billable item is retained in a billing rule.

> **Warning**
>
> If you delete billable items, billing data may change. This change cannot be undone. Proceed with caution.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Rules**.
4. Click the name of the billing rule that you want to manage.
5. In the **Billing item** section, find the billable item that you want to delete and click **Delete** in the **Operation** column.

1.1.10.4.3. Billing rules

1.1.10.4.3.1. Create a billing rule

A billing rule consists of all billable items of a service. Billing rules define which billing dimensions are used for services. This topic describes how to create a billing rule for a cloud service.

Background information

Service fees are accumulated from multiple billable items. A billing rule must contain at least one billable item. The number of billable items of a cloud service is unlimited. However, we recommend that you do not configure large numbers of billable items.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.

2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings Management** > **Billing Rules**.
4. Click **Create billing rules**.
5. In the **Create billing rules** panel, set the following parameters.

Parameter	Required	Description
Rule name	Yes	The name of the custom billing rule. The name can be up to 64 characters in length and must be unique.
Rule description	No	The description of the custom billing rule. The description can be up to 256 characters in length.
Associated cloud products	Yes	The associated cloud service that you want to manage.
Select billing item	Yes	After you select an associated cloud service, you can select billable items. A maximum of three billable items can be selected. If no appropriate billable item is available, you can click Create billing items to create billable items based on actual requirements. For more information, see Create a billable item.

6. Click **Determine**.

1.1.10.4.3.2. View billing rules

This topic describes how to view existing billing rules and their details.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings Management** > **Billing Rules**.
4. (Optional) Use the following methods to filter billing rules:
   - Select a cloud service associated with billing rules from the service drop-down list and click **Search** to view all billing rules of the cloud service.
   - Select **All statuses**, **Not used**, **In use**, or **Used** from the status drop-down list and click **Search** to view all billing rules in the corresponding state.
   - Enter a billing rule name in the search box and click **Search** to search for the billing rule. Fuzzy search is supported.

> **Note**
>
> These filter methods can be used individually or in combination to more easily view and classify existing billing rules.

5. On the Billing rules page, view **Rule NAME/Status**, **Associated cloud products**, **Rule description**, **Number of associated policies**, **Included billing items**, and **Fee**.
6. Click the name of the billing rule that you want to view.

   On the billing rule details page, you can view **Billing Policy**, **Billing field**, and **Billing item** of the billing rule.
7. Click the + icon to the left of the billable item name.

   You can view **status**, **Unit Price type**, **Metering field**, **Pricing**, **Bill presentation**, **Quantity type**, **Quantity value**, and **Conversion factor** of the billable item.

1.1.10.4.3.3. Clone a billing rule

You can clone an existing billing rule to create the same or a similar billing policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings Management** > **Billing Rules**.
4. Find the billing rule that you want to clone and click **Clone** in the **Operation** column.
5. In the **Create billing rules** panel, modify billing rule parameters based on your requirements.

   For more information about billing rule parameters, see Create a billing rule.
6. Click **Determine**.

1.1.10.4.3.4. Modify a billing rule

This topic describes how to modify a billing rule.

Prerequisites

After a billing rule is created, you cannot delete billable items that are in effect from the billing rule. You can create billable items in the billing rule, but these newly created billable items do not take effect. To modify billable items that are in effect in a billing rule, you must clone the billing rule. For more information, see Clone a billing rule.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings Management** > **Billing Rules**.
4. Click the name of the billing rule that you want to modify.
5. Modify the basic information of the billing rule.
   i. In the **Basic information** section, click **Editing**.

      You can modify **Rule name** and **Rule description**.

   ii. Click **Save**.
6. Modify the billable items of the billing rule.
   i. In the **Billing item** section, find the billable item that you want to modify and click **Editing** in the **Operation** column.

   ii. In the **Modify billing items** panel, modify related parameters and click **Determine**.

      For more information about billable item parameters, see Create a billable item.

> **Warning**
>
> If you modify billing rules that are in use, billing data may change. This change cannot be undone. Proceed with caution.

1.1.10.4.3.5. Delete a billing rule

You can delete billing rules that are incorrect or no longer needed.

Prerequisites

A billing rule is preconfigured for each service. This rule is used as the default billing rule and cannot be deleted. Default billable items in the default billing rule cannot be deleted. You can create other billable items.

> **Warning**
>
> If you delete a billing rule, billing data may change. This change cannot be undone. Proceed with caution.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings Management** > **Billing Rules**.
4. Find the billing rule that you want to delete and click **Delete** in the **Operation** column.
5. In the **Do you want to delete this rule** message, click **Delete**.

1.1.10.4.4. Billing policies

1.1.10.4.4.1. View billing policies

This topic describes how to view existing billing policies.

Background information

A billing policy consists of billing rules of multiple cloud services. A billing policy defines billing rules for a group of services, including organizations for which the billing policy is in effect and discounts for all services.
A billing policy includes a billing rule for each service, a list of organizations for which the billing policy is in effect, and a discount for each service. The discount for each service can be different.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Policies**.
4. (Optional) Use one of the following methods to filter billing policies:
   - Select **All statuses**, **Not used**, **In use**, or **Used** from the status drop-down list and click **Search** to view all billing policies in the corresponding state.
   - Enter a billing policy name in the search box and click **Search** to search for the billing policy. Fuzzy search is supported.

> **Note**
>
> These filtering methods can be used individually or in combination to more easily view and classify existing billing policies.

5. On the Billing Policy page, view **Policy name/Status**, **Include rules**, **Association plan**, and **Associated organizations**.
   - Move the pointer over the number of rules in the **Include rules** column to view the list of rule names. You can click a rule name to view its details.
   - Move the pointer over the number of plans in the **Association plan** column to view the list of plan names. You can click a plan name to view its details.
6. Click the name of the billing policy that you want to view.

   On the **Billing Policy details** page, you can view **Basic Information** and **Billing rules** of the billing policy.
7. In the **Billing rules** section, billing rules included in the billing policy are displayed as a table.

   Filter and search for billing rules by **rule status**, **product name**, and **rule name**.
8. Click icons in the upper-right corner of the **Billing rules** section to customize how the **Billing rules** section is displayed.

Icon	Description
	Shows the list in full screen mode. You can click the icon in the upper-right corner to exit the full screen mode.
	Selects the size of cells and fonts in the list. Valid values: ○ Default ○ Compact
	Selects or clears check boxes to determine the items to be displayed in the list. You can also click **Reset** to restore the list to its original settings.

1.1.10.4.4.2. Create a billing policy

This topic describes how to create a billing policy.

Background information

The system requires that each billing policy includes all services for which billing is enabled. If a service in a billing policy is provided free of charge, the discount rate can be set to 0. This discount property takes effect only in the billing policy and does not belong to billing rules.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Policies**.
4. Click **Create a billing policy**.

5. In the **Create a billing policy** panel, set the following parameters.

Parameter	Description
Policy name	The name of the custom billing policy. The name can be up to 64 characters in length and must be unique.
Note	The description of the custom billing policy. The description can be up to 256 characters in length.
Policy effective range	The level-1 organization for which the custom billing policy takes effect. You can select **All organizations take effect** to assign the billing policy to all level-1 organizations. If a metering data record belongs to an organization or a subordinate organization in the organization list to which a billing policy is assigned, the billing policy is the only one that is applicable to the metering data record.
Bind billing rules	You can perform the following operations to select a billing rule: i. Click a service name. ii. Select the billing rule that you want to use on the right. iii. In the **Set discount rate** column, enter a discount rate. If this service is provided free of charge, set the discount rate to 0. The discount rate cannot exceed 10. The default discount rate is set to 1 for all services, which indicates no discount. After fees of metering data of a service are calculated based on billing rules, the fees are multiplied by the discount rate of the service in the billing policy to calculate the price, discount, and original price. The original price is the sum of the price and the discount.

6. Click **Determine**.

1.1.10.4.4.3. Clone a billing policy

You can clone an existing billing policy to create the same or a similar billing policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Policies**.
4. Find the billing policy that you want to clone and click **Clone** in the **Operation** column.
5. In the **Create a billing policy** panel, modify the configurations of the billing policy based on your requirements.

   For information about billing policy parameters, see Create a billing policy.
6. Click **Determine**.

1.1.10.4.4.4. Modify a billing policy

This topic describes how to modify a billing policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Policies**.
4. Click the name of the billing policy that you want to modify.
5. Modify the basic information of the billing policy.
   i. In the **Basic information** section, click **Editing**.
   Policy name, **Associated organizations**, and **Note** can be modified.
   ii. Click **Save**.
6. Modify the billing rules and discount rates for each service.
   i. In the **Billing rules** section, click **Editing**.
   ii. Separately modify the billing rules and discount rates.
   - Select the desired billing rule in the **Rule name** column corresponding to the service that you want to manage.
   - Enter the expected discount rate in the **Discount rate** column corresponding to the service that you want to manage.

   iii. Click **Save**.

> **Warning**
> If you modify billing rules that are in use, billing data may change. This change cannot be undone. Proceed with caution.

1.1.10.4.4.5. Delete a billing policy

You can delete billing policies that are incorrect or no longer needed.

Prerequisites

A billing policy is preconfigured in the system. This policy is used as the default billing policy and cannot be deleted.

> **Warning**
> If you delete a billing policy, billing data may change. This change cannot be undone. Proceed with caution.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Policies**.
4. Find the billing policy that you want to delete and click **Delete** in the **Operation** column.
5. In the **Do you want to delete the policy** message, click **Delete**.

1.1.10.4.5. Billing plans

1.1.10.4.5.1. View billing plans

This topic describes how to view existing billing plans.

Background information

A billing plan consists of multiple billing policies that are configured with an effective time and priorities.
A billing plan specifies a time range for billing policies. This enables the billing system to use different billing policies at different periods of time.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Plans**.
4. (Optional) Use one of the following methods to filter billing plans:
   - Select **All statuses**, **Not used**, **In use**, or **Used** from the status drop-down list and click **Search** to view all billing plans in the corresponding state.
   - Set **Start Date** and **End Date** from the date picker and click **Search** to view all billing plans that are in effect within the specified period of time.

   > **Note**
   > These filtering methods can be used individually or in combination to more easily view and classify existing billing plans.
5. On the Billing plan page, view **Plan NAME/Status**, **Billing Policy**, **Billing rules**, **Priority**, and **Effective time** of a billing plan.
   - Move the pointer over the number of policies in the **Billing Policy** column to view the list of policy names. You can click a policy name to view its details.
   - Move the pointer over the number of rules in the **Billing rules** column to view the list of rule names. You can click a rule name to view its details.
   - The greater the number in the **Priority** column, the higher the priority of the plan.
6. Click the name of the billing plan that you want to view.

   On the **Billing plan details** page, you can view **Basic information** and **Billing Policy list** of the billing plan.
7. In the **Billing Policy list** section, view the billing policies contained in the billing plan.

   You can filter and search for billing policies by **status** and **policy name**.
8. Click icons in the upper-right corner of the **Billing Policy list** section to customize how the **Billing Policy list** section is displayed.

Icon	Description
	Shows the list in full screen mode. You can click the icon in the upper-right corner to exit the full screen mode.

Icon	Description
	Selects the size of cells and fonts in the list. Valid values: ○ Default ○ Compact
	Refreshes the list of billing policies.
	Selects or clears check boxes to determine the items to be displayed in the list. You can also click **Reset** to restore the list to its original settings.

1.1.10.4.5.2. Create a billing plan

This topic describes how to create a billing plan.

Prerequisites

- To simplify billing configurations and avoid unexpected charges due to excessive billing plans, a maximum of nine billing pans can be created (not including the default billing plan).
- To avoid errors caused by empty policies in a billing plan, the default billing policy is included in each billing plan. By default, the default billing policy has the lowest priority. This ensures that at least one billing policy is available after the billing plan is assigned to metering data.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Plans**.
4. Click **Create a billing plan**.
5. In the **Create a billing plan** panel, set the following parameters.

Parameter	Required	Description
Plan Name	Yes	The name of the custom billing plan. The name can be up to 64 characters in length and must be unique.
Effective time	Yes	The period of time for which the billing plan is in effect. You can select **Permanently effective**. This way, the billing plan is in effect all the time.
Plan priority	Yes	The priority of the billing plan in the billing system. The priority must be an integer greater than or equal to 0. The greater the number, the higher the priority.

Parameter	Required	Description
Plan Remarks	No	The description of the custom billing plan. The description can be up to 256 characters in length.
Bind a billing policy	Yes	You can add a billing policy and set the effective time and discount rate for each policy. Perform the following operations: i. Click **Add a billing policy**. ii. (Optional) Select a state from the **All statuses** drop-down list and click **Search**. iii. Select the billing policies that you want to add. iv. Click **Determine**. v. Set **Policy Effective Time** and **Discount rate** for each billing policy. The discount rate must be greater than 0 and less than or equal to 10.

6. Click **Determine**.

> **Note**
> Each billing policy in a billing plan has a priority. When metering data is matched to a billing plan, billing policies in the billing plan are selected based on their priorities. When a billing policy is selected, bills are generated only based on this policy.

1.1.10.4.5.3. Clone a billing plan

You can clone an existing billing plan to create the same or a similar billing plan.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Plans**.
4. Find the billing plan that you want to clone and click **Clone** in the **Operation** column.
5. In the **Create a billing plan** panel, modify billing plan parameters based on your requirements.

   For more information about billing plan parameters, see Create a billing plan.
6. Click **Determine**.

1.1.10.4.5.4. Modify a billing plan

This topic describes how to modify a billing plan.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Plans**.
4. Click the name of the billing plan that you want to modify.

5. Modify the basic information of the billing plan.
   i. In the **Basic information** section, click **Editing**.

      You can change **Plan Name**, **Effective range**, **Priority**, and **Remarks**.

   ii. Click **Save**.
6. Change **Effective time**, **Priority**, and **Discount rate** of the billing policy or delete the billing policy
   i. In the **Billing Policy list** section, click **Editing**.
   ii. Separately change the effective time, priority, and discount rate of a billing policy or delete the billing policy.
      - Select a new effective time from the date picker in the **Effective time** column.

        > **Note**
        > The effective time of the default billing policy cannot be changed.

      - Enter a desired discount rate in the **Discount rate** column corresponding to the billing policy.
      - Drag and drop the icon corresponding to the billing policy to prioritize it. The closer to the top of the list, the higher the priority.
      - Click **Delete** in the **Operation** column. In the **Do you want to delete the policy?** message, click Delete.

        > **Note**
        > The default billing policy cannot be deleted.

   iii. Click **Save**.
7. Add a billing policy.
   i. Click **Add a billing policy**.
   ii. (Optional) Select a state from the **All statuses** drop-down list and click **Search**.
   iii. Select the billing policies that you want to add.
   iv. Click **Determine**.
   v. Set **Effective time** and **Discount rate** for each billing policy.

> **Warning**
> If you modify billing plans that are in use, billing data may change. This change cannot be undone. Proceed with caution.

1.1.10.4.5.5. Delete a billing plan

You can delete billing plans that are incorrect or no longer needed.

Prerequisites

A billing plan is preconfigured in the system. This policy is used as the default billing policy and cannot be deleted.

> **Warning**
> If you delete a billing plan, billing data may change. This change cannot be undone. Proceed with caution.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billing Management** > **Billing Plans**.
4. Find the billing plan that you want to delete and click **Delete** in the **Operation** column.
5. In the **Do you want to delete the plan** message, click **Delete**.

1.1.10.4.6. Bills

1.1.10.4.6.1. View cloud service bills

You can view bills by cloud service type.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings** > **Cloud Service Bills**.
4. View the bill overview.
   i. Click the **Bill overview** tab.
   ii. Select the year and the month for which you want to view bills from the **Account period** date picker. You can view different data in the following sections:
   - **Cloud products bill summary**: View the number of cloud services that has billing enabled, the number of cloud services that incur charges, the total fees generated within the selected month, and the total fees generated within the selected year.
   - **Consumption trends of cloud products.Last six months**: View the consumption trend of each cloud service within the last six months. Click **Top 5** or **Top 10** in the upper-right corner to show the top five or top ten cloud services in terms of consumption within each month. Click the  or  icon to show the consumption trends of cloud services by using a column chart or a line chart.
   - **Top 10Consumption distribution of cloud products**: View the generated fees and their proportions of the top ten cloud services within the selected month.
   - **Consumption details of all cloud products**: View the fees of all cloud services within the selected month. Select Sort by fee from high to low or Sort by fee from low to high in the upper-right corner to sort cloud services by fees.
5. View bill statistics.
   i. Click the **Bill statistics** tab.
   ii. Select the year and the month for which you want to view bills from the date picker.
   iii. (Optional) Select the cloud service and the region that you want to view.

iv. Click **Search**.

You can view the billing information that meets the search conditions. The billing information includes the billing period, cloud service, region, fees, and billing cycle.

v. (Optional) Click the icon to modify fees as an operations administrator.

vi. (Optional) Click **View details** in the Operation column to view bill details.

vii. Save the queried data to your computer.

- Click **Export all bills** to save all queried data to your computer as an XLS file.
- Select the bills that you want to export and click **Export Selected bills** to save the selected data to your computer as an XLS file.

6. View bill details.

i. Click the **Bill details** tab.

ii. Select the year and the month for which you want to view bills from the **Account period** date picker.

iii. (Optional) Set Cloud products, Area, and Instance ID.

iv. Click **Search**.

You can view the billing information that meets the search conditions. The billing information includes **Account period**, **Cloud products**, **Tissue**, **Resource set**, **Instance ID**, **Area**, **Billing cycle**, **Free amount**, and **Amount payable**.

v. Save the queried data to your computer.

- Click **Export all bills** to save all queried data to your computer as an XLS file.
- Select the bills that you want to export and click **Export Selected bills** to save the selected data to your computer.

1.1.10.4.6.2. View organization and resource set bills

You can view bills by organization and resource set.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Operations**.
3. In the left-side navigation pane, choose **Billings** > **Organizations and Resource Set Bills**.
4. View the bill overview.

i. Click the **Bill overview** tab.

ii. Select the year and the month for which you want to view bills from the **Account period** date picker. You can view different data in the following sections:

- **root(and all subordinate organizations) Bill overview**: View the number of organizations, the number of organizations that incur charges, the total fees generated within the selected month, and the total fees generated within the selected year.
- **Organizational consumption trend.Last 6 months**: View the consumption trend of each organization within the last six months. Click **Top 5** or **Top 10** in the upper-right corner to show the top five or top ten organizations in terms of consumption within each month. Click the  or  icon to show the consumption trends of organizations by using a column chart or a line chart.
- **TOP 10 Resource set consumption distribution**: View the generated fees and their proportions of the top ten organizations within the selected month.
- **Organization consumption details**: View the fees of all organizations within the selected month. Select Sort by fee from high to low or Sort by fee from low to high in the upper-right corner to sort organizations by fees.

5. View bill statistics.
   i. Click the **Bill statistics** tab.
   ii. Select the year and the month for which you want to view bills from the date picker.
   iii. (Optional) Select the organization and the resource set that you want to view.
   iv. Click **Search**.

   You can view the billing information that meets the search conditions. The billing information includes **Account period**, **Organization name**, **Resource set name**, **Amount**, and **Billing cycle**.

   v. (Optional) Click the icon to modify fees as an operations administrator.
   vi. (Optional) Click **View details** in the Operation column to view bill details.
   vii. Save the queried data to your computer.
   - Click **Export all bills** to save all queried data to your computer as an XLS file.
   - Select the bills that you want to export and click **Export Selected bills** to save the selected data to your computer as an XLS file.
6. View bill details.
   i. Click the **Bill details** tab.
   ii. Select the year and the month for which you want to view bills from the **Account period** date picker.
   iii. (Optional) Set **Organization and resource set**, **Cloud products**, **Area**, and Instance ID.
   iv. Click **Search**.

   You can view the billing information that meets the search conditions. The billing information includes **Account period**, **Cloud products**, **Tissue**, **Resource set**, **Instance ID**, **Area**, **Billing cycle**, **Free amount**, and **Amount payable**.

   v. Save the queried data to your computer.
   - Click **Export all bills** to save all queried data to your computer as an XLS file.
   - Select the bills that you want to export and click **Export Selected bills** to save the selected data to your computer as an XLS file.

1.1.11. Security

1.1.11.1. View operation logs

You can view operation logs to obtain up-to-date information for a variety of resources and functional modules in the Apsara Uni-manager Management Console. You can also export operation logs to your computer.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a security administrator.
2. In the top navigation bar, click **Security**.
3. Select an **organization and a resource set**, a **resource type**, and a **product** and then click **Search**.

   You can also click **Advanced** in the upper-right corner to obtain more filter conditions, such as **User name**, **Resource ID**, **Source IP address**, **Start Date**, **End Date**, and **Keyword search**.

   The following table describes the fields in the query result.

   Fields in the query result

Log field	Description
Tissue	The organization of the object on which operations are performed.
Resource set	The resource set of the object on which operations are performed.
Resource type	The resource type of the object on which operations are performed.
Resource id	The ID of the object on which operations are performed.
State	The current status of the operation.
User	The name of the operator.
Event type	The operation performed on an Apsara Stack service. Operations include create, modify, delete, query, update, bind, unbind, enable or disable service instances, apply for or release service instances, or change ownership of service instances.
Source IP address	The IP address of the operator.
Details	A brief introduction of the operation.
Time	The start and end time of the operation.

   Note You can click the icon and select the fields that you want to show in the query result from the drop-down list.

4. (Optional)Click the icon to save the logs displayed on the current page to your computer as a CSV file.

1.1.12. RAM

1.1.12.1. RAM introduction

Resource Access Management (RAM) is a resource access control service provided by Apsara Stack.

You can use RAM to manage users and control which resources are accessible to employees, systems, and applications.

RAM provides the following features:

- RAM role

  To authorize a cloud service in a level-1 organization to use other resources in the organization, you must create a RAM role. This role specifies the operations that the cloud service can perform on resources.

  Only system administrators and level-1 organization administrators can create RAM roles.

- User group

  You can create multiple users within an organization and grant them different operation permissions on cloud resources.

  You can create RAM user groups to classify and authorize RAM users within your Apsara Stack tenant account. This simplifies the management of RAM users and their permissions.

  You can create RAM permission policies to grant different operation permissions to different user groups.

1.1.12.2. Permission policy structure and syntax

This topic describes the structure and syntax used to create or update permission policies in Resource Access Management (RAM).

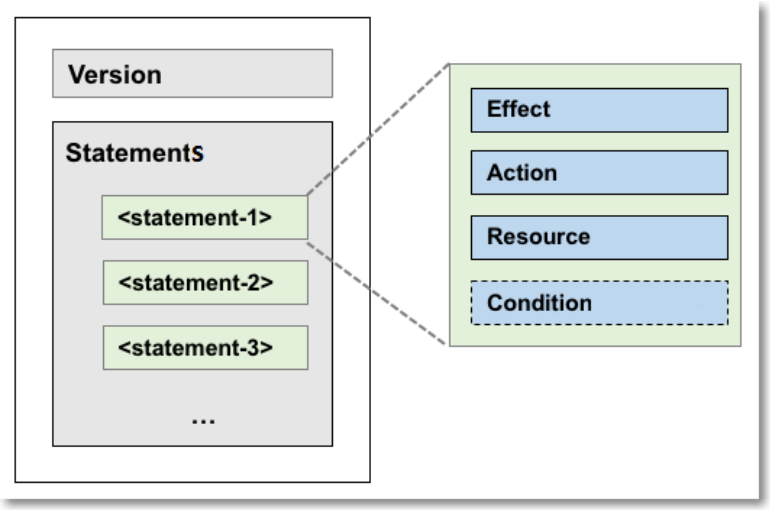
Policy characters and usage rules

- Characters in a policy
  - The following characters are JSON tokens and are included in policies: `{ } [ ] " , :`.
  - The following characters are special characters in the syntax and are not included in policies: `= < > ( ) |`.
- Use of characters
  - If an element can have more than one value, you can perform the following operations:
    - Separate multiple values by using commas (,) as delimiters between each value and use an ellipsis (...) to describe the remaining values. Example: `[ <action_string>, <action_string>, ...]`.
    - Include only one value. Examples: `"Action": [<action_string>]` and `"Action": <action_string>`.
  - A question mark (?) following an element indicates that the element is optional. Example: `<condition_block?>`.
  - A vertical bar (`|`) between elements indicates multiple options. Example: `("Allow" | "Deny")`.
  - Elements that must be text strings are enclosed in double quotation marks (""). Example: `<version_block> = "Version" : ("1")`.

Policy structure

The policy structure includes the following components:

- The version number.
- A list of statements. Each statement contains the following elements: Effect, Action, Resource, and Condition. The Condition element is optional.



Policy syntax

```
policy  = {
     <version_block>,
     <statement_block>
}
<version_block> = "Version" : ("1")
<statement_block> = "Statement" : [ <statement>, <statement>, ... ]
<statement> = {
    <effect_block>,
    <action_block>,
    <resource_block>,
    <condition_block? >
}
<effect_block> = "Effect" : ("Allow" | "Deny")
<action_block> = ("Action" | "NotAction") :
    ("*" | [<action_string>, <action_string>, ...])
<resource_block> = ("Resource" | "NotResource") :
    ("*" | [<resource_string>, <resource_string>, ...])
<condition_block> = "Condition" : <condition_map>
<condition_map> = {
  <condition_type_string> : {
      <condition_key_string> : <condition_value_list>,
      <condition_key_string> : <condition_value_list>,
      ...
  },
  <condition_type_string> : {
      <condition_key_string> : <condition_value_list>,
      <condition_key_string> : <condition_value_list>,
      ...
  }, ...
}
<condition_value_list> = [<condition_value>, <condition_value>, ...]
<condition_value> = ("String" | "Number" | "Boolean")
```

Description:

- The current policy version is 1.
- The policy can have multiple statements.

- The effect of each statement can be either `Allow` or `Deny`.

  > **Note** In a statement, both the Action and Resource elements can have multiple values.

- Each statement can have its own conditions.

  > **Note** A condition block can contain multiple conditions with different operators and logical combinations of these conditions.

- You can attach multiple policies to a RAM user. If policies that apply to a request include an `Allow` statement and a `Deny` statement, the Deny statement overrides the Allow statement.
- Element value:
  - If an element value is a number or Boolean value, it must be enclosed in double quotation marks ("") in the same way as strings.
  - If an element value is a string, characters such as the asterisk (`*`) and question mark (`?`) can be used for fuzzy matching.
    - The asterisk (`*`) indicates any number (including zero) of allowed characters. For example, `ecs:Describe*` indicates all ECS API operations that start with `Describe`.
    - `The question mark (?)` indicates an allowed character.

Policy format check

Policies are stored in RAM as JSON documents. When you create or update a policy, RAM first checks whether the JSON format is valid.

- For more information about JSON syntax standards, see RFC 7159.
- We recommend that you use tools such as JSON validators and editors to check whether the policies meet JSON syntax standards.

1.1.12.3. RAM roles

1.1.12.3.1. View basic information of a RAM role

You can view basic information of a RAM role, including its user groups and existing permission policies.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Click the name of the RAM role that you want to view.
5. On the **Role Authorization** page, click the **User Groups** and **Permissions** tabs to view related information.

1.1.12.3.2. Create a RAM role

To authorize a cloud service in a level-1 organization to use other resources in the organization, you must create a Resource Access Management (RAM) role. This role contains the operations that the cloud service can perform on resources.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.

3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Click **Create RAM Role** in the upper-left corner.
5. On the **Create a RAM role** page, set **Role name**, **Description**, and **Sharing Scope**.

   Valid values of the **Sharing Scope** parameter:

   - **Global**

     The role is visible and valid to all organizations involved. The default value is Global.
   - **Current Organization**

     The role is visible and valid to the organization to which the user belongs.
   - **Subordinate Organization**

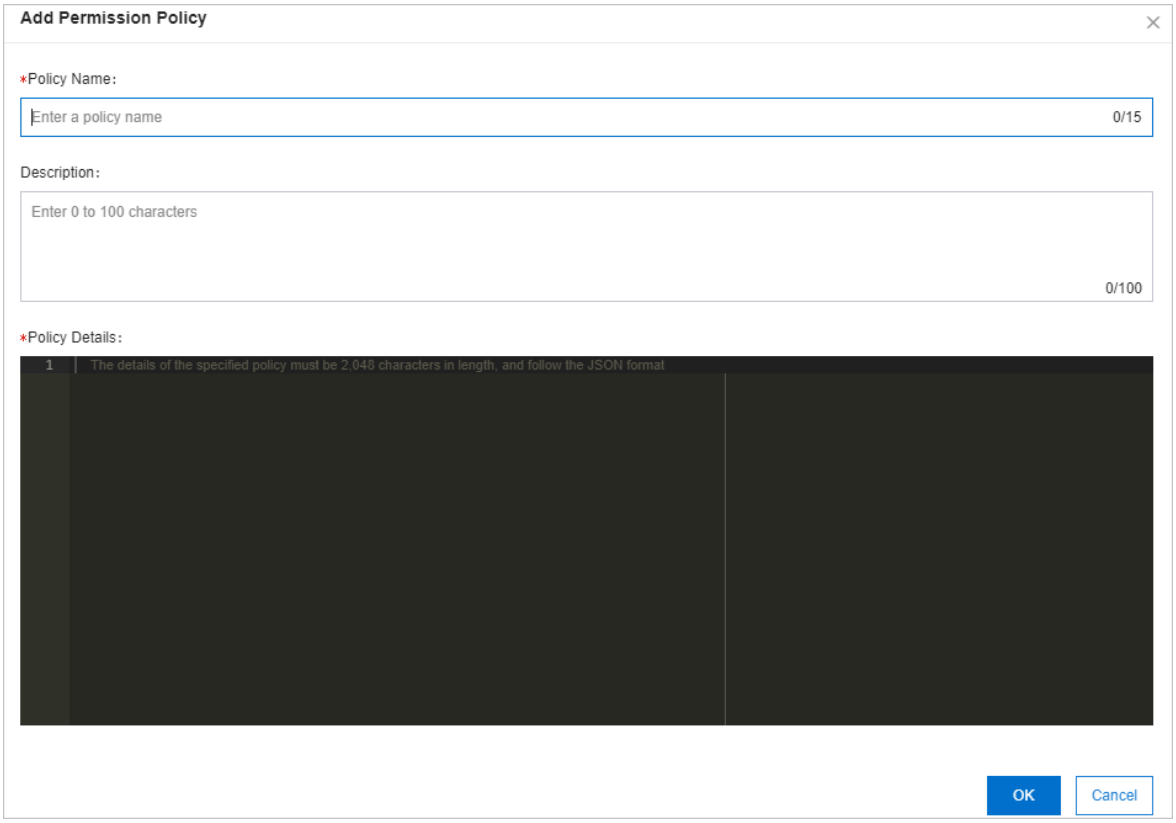
     The role is visible and valid to the organization to which the user belongs and its subordinate organizations.
6. Click **Create and configure a RAM role**.

1.1.12.3.3. Create a policy

To use a cloud service to access other cloud resources, you must create a policy and attach it to a user group.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the Resource Access Management (RAM) role that you want to modify and click **Modify** in the **Actions** column.
5. Click the **Permissions** tab.
6. Click **Add Permission Policy**.
7. In the Add Permission Policy dialog box, enter information of the policy.



For more information about how to enter the policy content, see Permission policy structure and syntax.

8. Click **OK**.

1.1.12.3.4. Modify the content of a RAM policy

You can modify the content of a Resource Access Management (RAM) policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the RAM role that you want to modify and click **Modify** in the **Actions** column.
5. Click the **Permissions** tab.
6. Click the name of a policy in the **Permission Policy Name** column.
7. In the **Modify Permission Policy** dialog box, modify the relevant information and click **OK**.

   For more information about how to modify the policy content, see Permission policy structure and syntax.

1.1.12.3.5. Modify the name of a RAM policy

You can modify the name of a Resource Access Management (RAM) policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.

4. Find the RAM role that you want to modify and click **Modify** in the **Actions** column.
5. Click the **Permissions** tab.
6. Click the name of a policy in the **Permission Policy Name** column.
7. In the **Modify Permission Policy** dialog box, modify the policy name.

1.1.12.3.6. Add a RAM role to a user group

You can bind Resource Access Management (RAM) roles to user groups.

Prerequisites

A user group is created. For more information, see Create a user group.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the RAM role that you want to manage and click **Modify** in the **Actions** column.
5. Click the **User Groups** tab.
6. Click **Add User Group**. In the Add User Group dialog box, select a user group.
7. Click **OK**.

1.1.12.3.7. Grant permissions to a RAM role

When you grant permissions to a RAM role, all users in the user groups that are assigned this role share the granted permissions.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the RAM role to which you want to grant permissions and click **Modify** in the **Actions** column.
5. Click the **Permissions** tab.
6. Click **Select Existing Permission Policy**.
7. In the dialog box that appears, select a RAM policy and click **OK**.

   If no RAM policy is available, you must add a policy. For more information, see Add a permission policy.

1.1.12.3.8. Remove permissions from a RAM role

You can remove permissions that are no longer needed from Resource Access Management (RAM) roles.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the RAM role that you want to manage and click **Modify** in the **Actions** column.
5. Click the **Permissions** tab.
6. Find the policy that you want to remove and click **Remove** in the **Actions** column.

1.1.12.3.9. Change a RAM role name

Administrators can change the names of Resource Access Management (RAM) roles.

Context

The names of preset roles cannot be changed.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the RAM role whose name you want to change and click **Modify** in the **Actions** column.
5. Enter a new role name.

1.1.12.3.10. Delete a RAM role

You can delete a Resource Access Management (RAM) role that is no longer need.

Prerequisites

No policies are attached to the RAM role.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Permissions > Role Permissions**.
4. Find the RAM role that you want to delete and click **Modify** in the **Actions** column.
5. In the message that appears, click **OK**.

1.1.12.4. RAM authorization policies

1.1.12.4.1. Create a service-linked role

You can create authorization policies and grant them to organizations.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Service-linked Roles**.
4. In the upper-left corner of the page, click **Create Service-linked Role**.
5. On the **Create Service-linked Role** page, set **Organization Name** and **Service Name**.
6. Click **OK**.

1.1.12.4.2. View the details of a service-linked role

You can view the details of a Resource Access Management (RAM) role, including its role name, creation time, description, and Alibaba Cloud Resource Name (ARN).

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Service-linked Roles**.
4. On the **Service-linked Roles** page, set **Role Name**, **Service Name**, or **Organization Name**, and click **Search**.

   To perform another search, click **Reset**.
5. Find the service-linked role that you want to view and click **View Details** in the **Actions** column.

1.1.12.4.3. View RAM policies

You can view the details of a Resource Access Management (RAM) policy, including its policy name, policy type, default version, description, association time, and policy content.

Prerequisites

A RAM policy is created. For more information, see Create a RAM role.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an operations administrator.
2. In the top navigation bar, click **Configurations**.
3. In the left-side navigation pane, click **Service-linked Roles**.
4. On the **Service-linked Roles** page, set **Role Name** or **Service Name** and click **Search** in the upper-right corner.

   To perform another search, click **Reset**.
5. Find the service-linked role that you want to view and click **View Details** in the **Actions** column.
6. Click the **Role Policy** tab to view the information of the role policy. Click **View Details** in the Actions column to view the policy details.

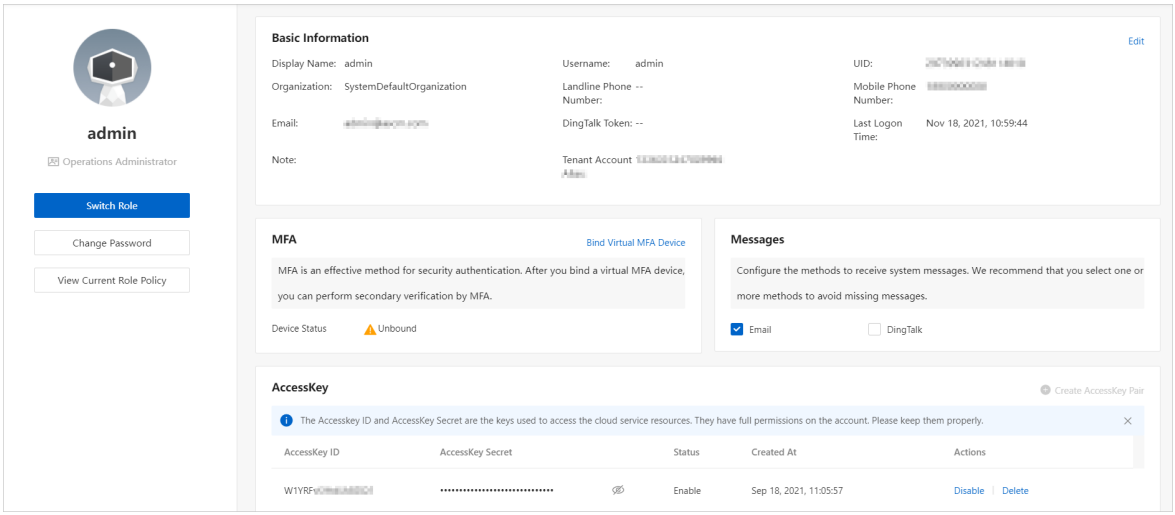
1.1.13. Personal information management

1.1.13.1. Modify personal information

You can modify your personal information to keep it up to date.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and click **User Information**.



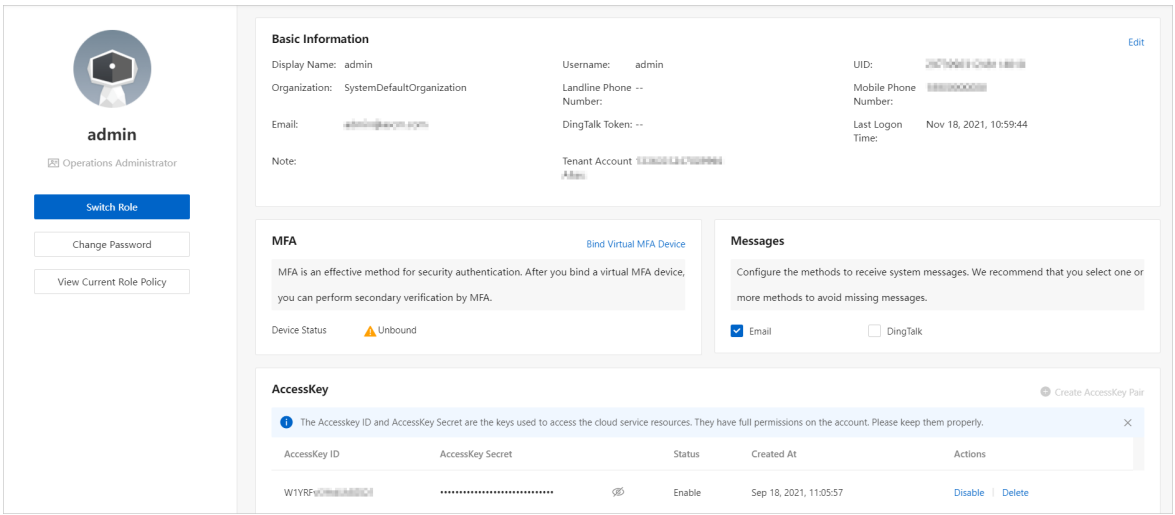
3. In the upper-right corner of the **Basic Information** section, click **Edit**.
4. In the **Modify User Information** dialog box, modify the personal information.
5. Click **OK**.

1.1.13.2. Change the logon password

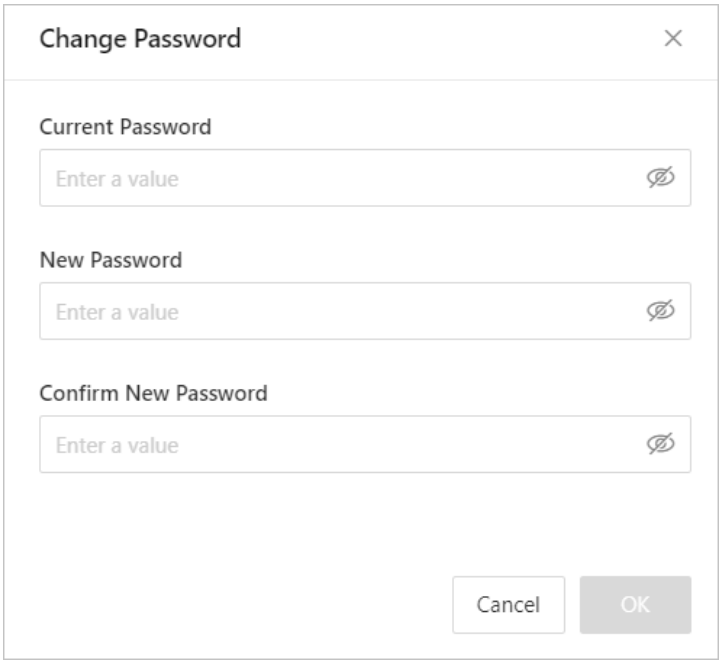
To improve security, you must change your logon password on a regular basis. This topic describes how to change your logon password.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and click **User Information**.



3. Click **Change Password**.
4. In the **Change Password** dialog box, set **Current Password**, **New Password**, and **Confirm Password**.



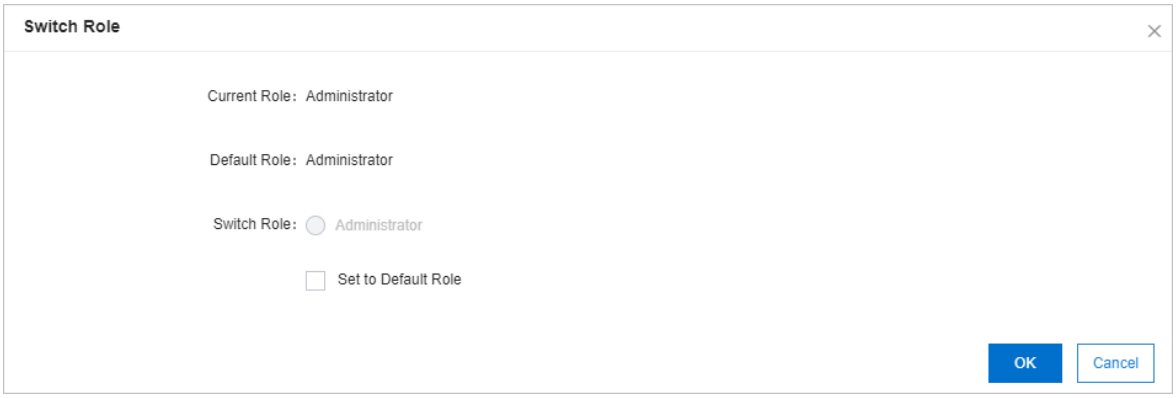
5. Click **OK**.

1.1.13.3. Switch the current role

You can switch the scope of your current role.

Procedure

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the upper-right corner of the homepage, move the pointer over the user profile picture and choose **User Information** from the shortcut menu.
3. Click **Switch Role**.
4. In the **Switch Role** dialog box that appears, select the role that you want to switch to.



You can also switch back to the default role.

1.1.13.4. View the current role policy

You can view the information of your current role policy.

Procedure

1. Log on to the Apsara Uni-manager Management Console.

2. In the upper-right corner of the homepage, move the pointer over the profile picture and click **Personal information**.
3. Click **View Current Role Policy** on the left side of the page.

   In the **View Policies of Current Role** dialog box, view your RAM role.
4. Select the organization and the policy that you want to view.

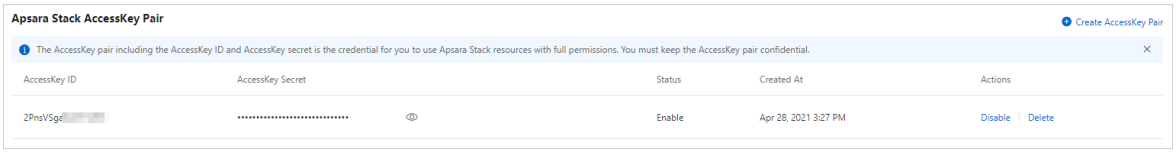
   In the **View Policies of Current Role** message, view the policy content.

1.1.13.5. View the AccessKey pair of your Apsara Stack tenant account

To ensure the security of cloud resources, the system must verify the identity of visitors and ensure that they have the relevant permissions. You must obtain the AccessKey ID and AccessKey secret of your Apsara Stack tenant to access cloud resources.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.
3. In the **Apsara Stack AccessKey Pair** section, view your AccessKey pair.



> **Note** The AccessKey pair consists of an AccessKey ID and an AccessKey secret. These credentials provide you with full permissions on Apsara Stack resources. You must keep the AccessKey pair confidential.

1.1.13.6. Create an AccessKey pair

You can delete your old AccessKey pairs and create new ones to implement the rotation of your AccessKey pairs.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.
3. In the upper-right corner of the **Apsara Stack AccessKey Pair** section, click **Create AccessKey Pair**.

> **Note**
>
> Each user can have up to two AccessKey pairs.

1.1.13.7. Delete an AccessKey pair

You can delete AccessKey pairs that are no longer needed.

Prerequisites

- Each user can delete only its own AccessKey pairs. The administrator cannot delete the AccessKey pairs of users.
- Each user retains at least one AccessKey pair.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.
3. In the **Apsara Stack AccessKey Pair** section, find the AccessKey pair that you want to delete and click **Delete** in the **Actions** column.
4. In the message that appears, click **OK**.

1.1.13.8. Disable an AccessKey pair

You can disable AccessKey pairs that are no longer needed. Newly created AccessKey pairs are in the Enable state by default.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.
3. In the **Apsara Stack AccessKey Pair** section, find the AccessKey pair that you want to disable and click **Disable** in the **Actions** column.

   Each user can disable only its own AccessKey pairs. The administrator cannot disable the AccessKey pairs of users.
4. In the message that appears, click **OK**.

> **Note**
>
> Make sure that at least one AccessKey pair is in the Enable state.

1.1.13.9. Enable an AccessKey pair

Disabled AccessKey pairs must be re-enabled before you can continue to use them. Newly created AccessKey pairs are in the Enable state by default.

Procedure

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.
3. In the **Apsara Stack AccessKey Pair** section, find the AccessKey pair that you want to enable and click **Enable** in the **Actions** column.
4. In the message that appears, click **OK**.

> **Note**
>
> Make sure that at least one AccessKey pair is in the Enable state.

1.1.13.10. MFA

1.1.13.10.1. Overview

Multi-factor authentication (MFA) is an identity authentication method in computer systems. It requires users to provide two or more verification factors to gain access to a resource. This topic introduces the principles and use scenarios of MFA.

Introduction to MFA

When MFA is enabled, you must enter your username and password (first security factor) and then a variable verification code (second security factor) from an MFA device when you log on to the Apsara Uni-manager Management Console. Two-factor authentication enhances security for your account.

MFA devices use the Time-based One-time Password (TOTP) algorithm to generate time-dependent 6-digit dynamic verification codes. The Apsara Uni-manager Management Console supports software-based virtual MFA devices. You can install software such as the Alibaba Cloud app that supports MFA on your mobile device such as your mobile phone to act as a virtual MFA device.

Limits

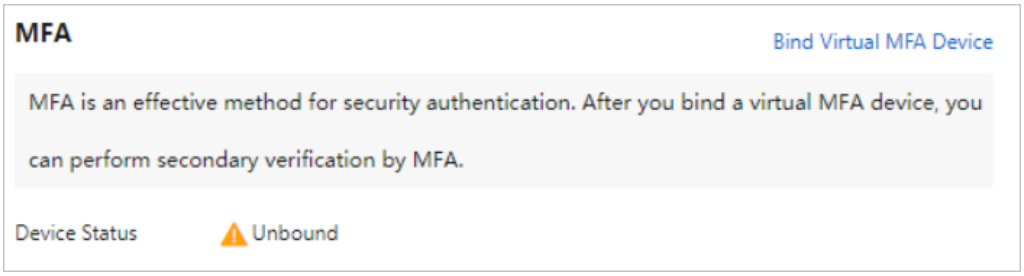
The TOTP algorithm requires that the time of the system clock of the Apsara Uni-manager Management Console remain consistent with the standard time on the Internet. Otherwise, discrepancies in time can lead to inconsistent MFA verification codes and leave you unable to log on to the Apsara Uni-manager Management Console.

1.1.13.10.2. Bind a virtual MFA device to enable MFA

Multi-factor authentication (MFA) is automatically enabled after you bind an MFA device. This topic describes how to bind a virtual MFA device.

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.
3. In the upper-right corner of the **MFA** section, click **Bind Virtual MFA Device**.
4. On the **Bind Virtual MFA Device** page, follow the instructions to bind an MFA device.

   The following figure shows the MFA section after the MFA device is bound.



Results

After a virtual MFA device is bound, you must enter a 6-digit MFA verification code in addition to your username and password before you can log on to the Apsara Uni-manager Management Console.

1.1.13.10.3. Unbind a virtual MFA device to disable MFA

To disable multi-factor authentication (MFA), you must disable the MFA device that is bound. This topic describes how to unbind a virtual MFA device to disable MFA.

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and choose **User Information** from the shortcut menu.

3. In the upper-right corner of the **MFA** section, click **Disable Virtual MFA Device**.
4. In the dialog box that appears, click **Disable Virtual MFA Device**.

   After you disable MFA, you need only to enter your username and password the next time you log on to the Apsara Uni-manager Management Console.

1.1.13.10.4. Forcibly enable MFA

Administrators including the platform administrator, operations administrator, and organization administrator can check whether their users have multi-factor authentication (MFA) enabled. If MFA is disabled for the users, the administrators can forcibly enable MFA for the users.

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users** > **Users**.
4. Click the **System Users** tab.
5. Find the user for whom you want to forcibly enable MFA and choose **More** > **View MFA Status** in the **Actions** column.

   You can search for the user by username, organization, or role.
6. In the MFA Status Prompt message, click **OK**.

> **Note**
>
> - After MFA is forcibly enabled for a user, the user must go to the Bind Virtual MFA Device page to bind a virtual MFA device before the user can log on to the Apsara Uni-manager Management Console.
> - MFA can be forcibly enabled for users, but cannot be forcibly disabled. Before MFA is enabled, the MFA status of the user is **Not Enabled**.
> - After MFA is enabled, the MFA status of the user is **Enabled but Not Bound**.

1.1.13.10.5. Reset MFA

If you lose your multi-factor authentication (MFA) key, you must have the administrator reset the key. The MFA key is automatically reset while the password is reset.

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane, choose **Users** > **Users**.
4. Click the **System Users** tab.
5. Find the user for whom you want to reset MFA and click the username.

   You can search for the user by username, organization, or role.
6. In the user details panel, click **Reset password**.

2.Elastic Compute Service (ECS)

2.1. User Guide

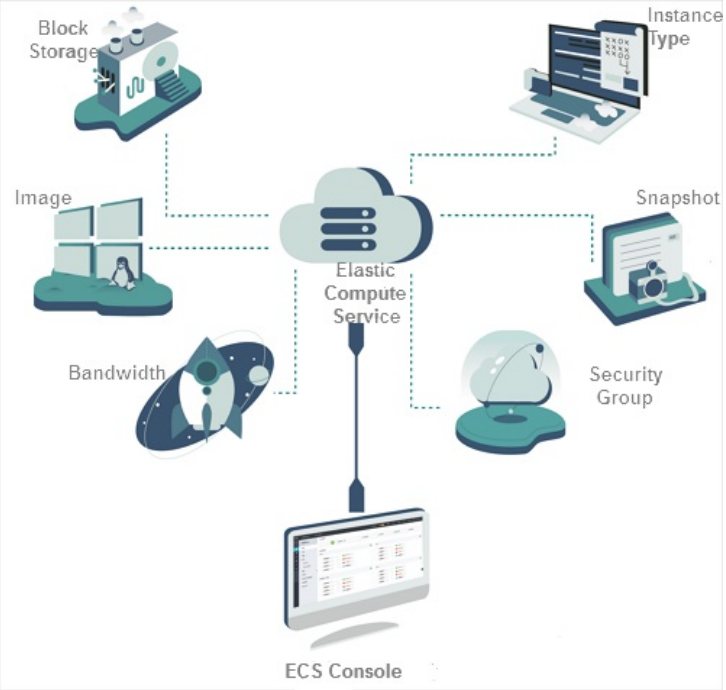
2.1.1. What is ECS?

2.1.1.1. Overview

Elastic Compute Service (ECS) is a type of computing service that features elastic processingcapabilities. Compared with physical servers, ECS can be more efficientlymanaged and is more user-friendly. You can create instances, resize disks, and add or release any number of ECS instances at any time based on your business needs.

An ECS instance is a virtual computing environment that contains the most basic components of computers such as the CPU, memory, and storage. Users perform operations on ECS instances. Instances are core components of ECS, and operations can be performed on instances by using the ECS console. Other resources such as block storage, images, and snapshots can only be used after they are integrated into ECS instances. For more information, see ECS components.

ECS components



2.1.1.2. Instance lifecycle

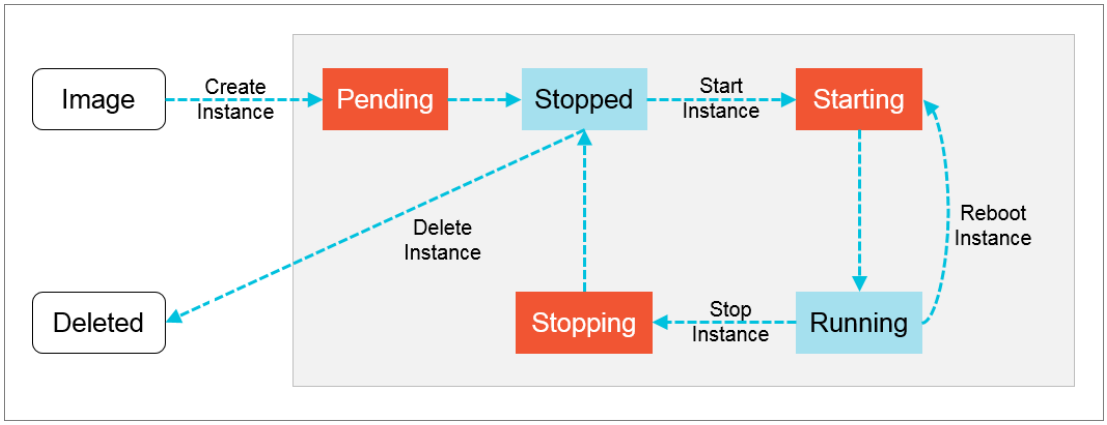
The lifecycle of an ECS instance begins when the instance is created and ends when the instance is released. This topic describes the instance states in the ECS console as well as state attributes and their corresponding instance states in API responses.

The following table describes the instance states in the ECS console and their corresponding instance states in API responses.

State	State attribute	Description	State in an API response

State	State attribute	Description	State in an API response
Instance being created	Intermediate	The instance is being created and waiting to start. If an instance remains in this state for an extended period of time, an exception has occurred.	Pending
Starting	Intermediate	When you start or restart an instance by using the ECS console or by calling an API operation, the instance enters this state before it enters the Running state. If an instance remains in the Starting state for an extended period of time, an exception has occurred.	Starting
Running	Stable	While an instance is in the Running state, the instance can function normally and can accommodate your business needs.	Running
Stopping	Intermediate	When you stop an instance by using the ECS console or by calling an API operation, the instance enters this state before it enters the Stopped state. If an instance remains in the Stopping state for an extended period of time, an exception has occurred.	Stopping
Stopped	Stable	An instance enters this state when it is stopped. Instances in the Stopped state cannot provide external services.	Stopped
Reinitializing	Intermediate	When you re-initialize the system disk or a data disk of an instance by using the ECS console or calling an API operation, the instance enters this state before it enters the Running state. If an instance remains in the Reinitializing state for an extended period of time, an exception has occurred.	Stopped
Changing system disk	Intermediate	When you replace the system disk of an instance by using the ECS console or by calling an API operation, the instance enters this state before it enters the Running state. If an instance remains in the Changing system disk state for an extended period of time, an exception has occurred.	Stopped

The following figure shows the transitions between instance states in API responses.



2.1.2. Instructions

2.1.2.1. Restrictions

Learn about restrictions before performing operations on ECS instances.

- Do not upgrade the kernel or operating system version of an ECS instance.
- Do not start SELinux for Linux systems except CentOS and RedHat.
- Do not detach PVDriver.
- Do not arbitrarily modify the MAC address of the network interface.

2.1.2.2. Suggestions

Consider the following suggestions to make more efficient use of ECS:

- ECS instances with 4 GiB or higher memory must use a 64-bit operating system. 32-bit operating systems have a maximum of 4 GiB of memory addressing.
- A 32-bit Windows operating system supports a maximum of 4 CPU cores.
- To ensure service continuity and avoid failover-induced service unavailability, we recommend that you configure service applications to boot automatically at system startup.

2.1.2.3. Limits

Before using ECS instances, you must be familiar with the limits of instance families.

General limits

- Windows operating systems support a maximum of 64 vCPUs in instance specifications.
- ECS instances do not support the installation of virtualization software and secondary virtualization.
- Sound card applications are not supported. Only GPU instances support virtual sound cards. External hardware devices, such as hardware dongles, USB flash drives, external hard disks, and bank U keys, cannot be directly connected to ECS instances.
- ECS does not support multicast protocols. If multicasting services are required, we recommend that you use unicast instead.

Instance family ga1

To create a ga1 instance, you must use one of the following images pre-installed with drivers:

- Ubuntu 16.04 with an AMD GPU driver pre-installed
- Windows Server 2016 English version with an AMD GPU driver pre-installed
- Windows Server 2008 R2 English version with an AMD GPU driver pre-installed

Note:

- A ga1 instance uses an optimized driver provided by Alibaba Cloud and AMD. The driver is installed in images provided by Alibaba Cloud and is currently unavailable for download.
- If the GPU driver malfunctions due to improper removal of related components, you need to replace the system disk to restore GPU related functions.

  > **Note** This operation causes data loss.

- If the driver malfunctions because an improper image is selected, you need to replace the system disk to reselect an image with an AMD GPU driver pre-installed.
- For Windows Server 2008 or earlier, you cannot connect to the VNC after the GPU driver takes effect. The VNC is irresponsive with a black screen or stuck at the splash screen. You can use other methods such as Remote

Desktop Protocol (RDP) to access the system.

- RDP does not support DirectX, OpenGL, or other related applications. You need to install the VNC and a client, or use other supported protocols such as PCOIP and XenDesktop HDX 3D.

Instance families gn4, gn5i, and gn5

- Bandwidth: If you use an image of Windows Server 2008 R2 for a gn4 instance, you cannot use the Connect to VNC function in the ECS console to connect to the instance after the installed GPU driver takes effect. You need to set the bandwidth to a non-zero value or attach an Elastic IP address to the created instance.
- Image: If an NVIDIA GPU driver is not required, you can select any image, and then Install the CUDA and GPU drivers for a Linux instance or Install the CUDA and GPU drivers for a Windows instance.

2.1.2.4. Notice for Windows users

Before using Windows-based ECS instances, you must consider the following points:

- Data loss may occur if a local disk is used as the data disk of an instance. We recommend that you use a cloud disk to create your instance if you are not sure about the reliability of the data architecture.
- Do not close the built-in shutdownmon.exe process. Otherwise, the server may require a longer time to restart.
- Do not rename, delete, or disable Administrator accounts or it may affect the use of the server.
- We do not recommend that you use virtual memory.
- When you modify your computer name, you must synchronize the following key values in the registry. Otherwise, the computer name cannot be modified, causing failure when installing certain third-party programs. The following key values must be modified in the registry:

```
HKEY_LOCAL_MACHINE\SYSTEM\ControlSet001\Control\ComputerName\ActiveComputerName
```

```
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\ComputerName\ComputerName
```

2.1.2.5. Notice for Linux users

Before using Linux-based ECS instances, you must consider the following points:

- Do not modify content of the default /etc/issue files under a Linux instance. Otherwise, the custom image created from the instance cannot be recognized, and instances created based on the image cannot start as expected.
- Do not arbitrarily modify the permissions of each directory in the partition where the root directory is located, especially permissions of /etc, /sbin, /bin, /boot, /dev, /usr, and /lib directories. Improper modification of permissions can cause errors.
- Do not rename, delete, or disable Linux root accounts.
- Do not compile or perform any other operations on the Linux kernel.
- We do not recommend the use of Swap for partitioning.
- Do not enable the NetWorkManager service. This service conflicts with the internal network service of the system, causing network errors.

2.1.2.6. Notice on defense against DDoS attacks

You need to purchase Anti-DDoS Pro to defend against DDoS attacks. For more information, see *Apsara Stack Security Product Introduction*.

2.1.3. Quick start

2.1.3.1. Overview

This topic describes how to create and connect to an Elastic Compute Service (ECS) instance.

Perform the following operations:

1. Create a security group

   A security group acts as a virtual firewall to control the inbound and outbound traffic of ECS instances. Each ECS instance must belong to at least one security group. When you create an instance, you must select a security group to control network access of the instance.

2. Create an instance

   An ECS instance is a virtual machine that contains the basic computing components of a server, such as CPU, memory, operating system, network settings, and disks. After a security group is created, you can select an instance type to create instances based on your business needs.

3. Connect to an instance

   Select a method to connect to an instance based on the network configurations and operating system of the instance and the operating system of your computer. After you connect to the instance, you can perform operations on it such as installing applications.

2.1.3.2. Log on to the ECS console

This topic describes how to log on to the ECS console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. In the top navigation bar, choose **Products > Elastic Computing > Elastic Compute Service**.

2.1.3.3. Create a security group

Security groups are an important means for network security isolation. They implement network access control for one or more ECS instances.

Prerequisites

A virtual compute cloud (VPC) is created. For more information, see *VPC User Guide*.

Context

Security groups determine whether the instances in the same account that are deployed within the same VPC and region can communicate with each other. By default, if the instances belong to the same security group, they can communicate with each other over the internal network. If the instances belong to different security groups, you can authorize mutual access between the security groups to allow the instances to communicate with each other over the internal network.

Procedure

1.
2.
3.
4. Click **Create Security Group**.
5. Configure the parameters listed in the following table to create a security group.

Type	Parameter	Required	Description
Region	Organization	Yes	Select an organization in which to create the security group. Make sure that the security group and the VPC belong to the same organization.
	Resource Set	Yes	Select a resource set in which to create the security group. Make sure that the security group and the VPC belong to the same resource set.
	Region	Yes	Select a region in which to create the security group. Make sure that the security group and VPC belong to the same region.
	Zone	Yes	Select a zone in which to create the security group.
Basic Settings	VPC	Yes	Select a VPC in which to create the security group.
	Security Group Name	No	Enter a name for the security group. The name must be 2 to 128 characters in length and start with a letter. It can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,). It cannot start with http:// or https://.
	Description	No	Enter a description for the security group. To simplify future management operations, we recommend that you provide an informational description. The description must be 2 to 256 characters in length and start with a letter. It can contain letters, digits, periods (.), underscores (_), hyphens (-), and commas (,). It cannot start with http:// or https://.

6. Click **Submit**.

2.1.3.4. Create an instance

An Elastic Compute Service (ECS) instance is a virtual machine that contains the basic computing components of a server, such as CPU, memory, operating system, network settings, and disks.

Prerequisites

- A virtual private cloud (VPC) and a vSwitch are created. For more information, see the "Create a VPC" and "Create a vSwitch" topics in *Apsara Stack VPC User Guide*.

- If you want to assign an IPv6 address to the instance that you want to create, make sure that the VPC and vSwitch are associated with IPv6 CIDR blocks. For more information, see the "Enable an IPv6 CIDR block for a VPC" and "Enable an IPv6 CIDR block for a vSwitch" topics in *Apsara Stack VPC User Guide*.
- One or more security groups are available. If no security groups are available, create one. For more information, see Create a security group.

Context

Some limits apply when you create GPU-accelerated instances. For more information, see Limits.

Procedure

1.
2.
3.
4. Click **Create Instance**.
5. Configure the parameters described in the following tables.
   i. Configure the basic settings of the instance.

Parameter	Required	Description
Organization	Yes	Select an organization.
Resource Set	Yes	Select a resource set.

   ii. Select a region and zone for the instance.

Parameter	Required	Description
Region	Yes	Select a region.
Zone	Yes	Select a zone. Zones are the physical zones with separate power supplies and networks within the same region. The internal networks of zones are connected, and faults in one zone are isolated from the other zones. To increase the availability of your applications, we recommend that you create instances in different zones.

iii. Configure network settings for the instance.

Parameter	Required	Description
Network Type	Yes	Select a network type. Only **VPC** is supported.
VPC	Yes	Select a VPC.
vSwitch	Yes	Select a vSwitch.
Private IP Address	No	Specify a private IPv4 address for the instance. The private IPv4 address must be within the CIDR block of the selected vSwitch. If you do not specify a private IP address, the system allocates one to the instance.
IPv6	No	Specify whether to assign an IPv6 address to the instance.

iv. Configure instance settings such as security group, instance family, and instance type and specify the number of instances that you want to create.

Parameter	Required	Description
Security Group	Yes	Select a security group.
Deployment Set	No	Select a deployment set. You can use deployment sets to disperse or aggregate the instances involved in your business.
User Data	No	In the User Data field, enter the user data to be automatically run on instance startup. ■ Windows supports batch and PowerShell scripts. Before you perform Base64 encoding of user data, make sure that the data to be encoded includes `[bat]` or `[powershell]` as the first line. ■ Linux supports shell scripts.
Quantity	Yes	Specify the number of instances that you want to create. The number must be an integer within the range of 1 to 100.
Instance Family	Yes	Select an instance family. After you select an instance family, you must select an instance type.
Instance Type	Yes	Select an instance type. Information such as CPU, memory, and instance family level are displayed in the Instance Type list. Select an instance type based on your business needs. Instance types that have specific CPU and memory combinations do not support Windows Server images. For more information, see the "Limits" topic in *ECS Product Introduction*.

v. Configure the image to be used by the instance.

Parameter	Required	Description
Image Type	Yes	Select an image type. Valid values: **Public Image**, **Custom Image**, and **Shared Custom Image**.
Public Image	Subject to the image type	Select a public image. Public images provided by Alibaba Cloud are licensed, secure, and stable. Public images include Windows Server images and major Linux images. This parameter is required when you set Image Type to **Public Image**. When you use an image that supports Dynamic Host Configuration Protocol version 6 (DHCPv6) to create an instance, an IPv6 address is automatically assigned to the instance. The instance can use this IPv6 address to communicate over the internal network. When you use an image that does not support DHCPv6 to create an instance, you must manually assign an IPv6 address to the instance. The following images support DHCPv6: ■ Linux images: ■ CentOS 7.6 IPV6 64Bit ■ CentOS 6.10 64Bit ■ SUSE Linux Enterprise Server 12 SP4 64Bit ■ Windows Server images **Note** To use an IPv6 address to communicate over the Internet, you must enable public bandwidth for the IPv6 address. For more information, see the "Enable Internet bandwidth for an IPv6 address" topic in *Apsara Stack VPC User Guide*.
Custom Image	Subject to the image type	Select a custom image. Custom images are created from instances or snapshots or are imported from your device. This parameter is required when you set Image Type to **Custom Image**.
Shared Custom Image	Subject to the image type	Select a custom image that is shared by another Apsara Stack tenant. This parameter is required when you set Image Type to **Shared Custom Image**.

vi. Configure the storage settings for the instance.

Parameter	Required	Description
System Disk	Yes	Select a disk category from the drop-down list and specify the system disk capacity. Valid values for the disk category: **Ultra Disk** and **Standard SSD**. The system disk capacity must range from 20 GiB to 500 GiB.
Data Disk	No	You can click Data Disk to create and attach data disks. For each data disk, select a disk category from the drop-down list and specify the disk capacity. Valid values for the disk category: **Ultra Disk** and **Standard SSD**. A maximum of 16 data disks can be attached to an instance. The maximum capacity of each data disk is 32 TiB. You can optionally select **Release with Instance** and **Encryption** for each data disk. To encrypt a data disk, configure the following parameters: ■ **Encryption Method**: Select **AES256**. ■ **Encryption Key**: You can select a key created in **Key Management Service (KMS)**. You can also add data disks after the instance is created. For more information, see Create a disk.

vii. Configure the logon password settings for the instance.

Parameter	Required	Description
Password Setting	Yes	Specify when to set a password. Valid values: **Set Now** and **Set After Purchase**. If you select Set After Purchase, you can use the password reset feature to set a password after the instance is created. For more information, see Change the logon password of an instance. **Note** The password is used to log on to the instance, not to a VNC management terminal.
Logon Password	No	Set the password to be used to log on to the instance. The password must be 8 to 30 characters in length and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. The supported special characters include `( ) ` ~ ! @# $%^&*-_+=|{}[]:;'<>,.?/` **Note** This parameter is required when you set **Password Setting** to **Set Now**.
Confirm Password	No	Enter the password again. **Note** This parameter is required when you set **Password Setting** to **Set Now**.

viii. (Optional) Enter a name for the instance.

The name must be 2 to 128 characters in length and can contain `hyphens (-), underscores (_), and colons (:)`. It must start with a letter and cannot start with `http://` or `https://`.

If you do not specify a name, the system assigns one at random.

6. Click **Submit**.

Result

The new instance appears in the instance list. When the instance is being created, it is in the **Preparing** state. When the instance is created, it enters the **Running** state.

2.1.3.5. Connect to an instance

2.1.3.5.1. Instance connecting overview

After an instance is created, you can connect to the instance to perform operations such as installing applications.

You can use one of the following methods to connect to an instance:

- Use remote connection tools to connect to instances that have public IP addresses. For more information about the procedure, see the following topics:
  - Connect to a Linux-based instance by using SSH commands in Linux or Mac OS X
  - Connect to a Linux-based instance by using remote connection tools in Windows

- Connect to a Windows-based instance by using RDP

- Use the VNC feature in the ECS console. For more information, see Connect to an instance by using a VNC management terminal.

The username of a Windows instance is Administrator, and that of a Linux instance is root.

2.1.3.5.2. Connect to a Linux instance by using SSH commands in Linux or Mac OS X

This topic describes how to use SSH commands to connect to a Linux instance.

Prerequisites

- The instance and the security group are created.
- The instance is in the **Running** state.
- A logon password is set for the instance.
- An Elastic IP address (EIP) is bound with the instance.
- An inbound security group rule is added to the security group to allow the SSH port.

Rule direction	Authorization policy	Protocol type	Port range	Priority	Authorization type	Authorization object
Inbound	Accept	TCP	22/22	1	IPv4 CIDR block	0.0.0.0/0

Procedure

1. Enter the following command and press the Enter key.

```
ssh root@instance IP
```

2. Enter the instance password of the root user and press the Enter key.

2.1.3.5.3. Connect to a Linux-based instance by using remote connection tools in Windows

This topic describes how to connect to an instance by using the PuTTY tool.

Prerequisites

Remote connection tools are designed with similar logics. In this example, PuTTY is used to connect to an instance. Download PuTTY at the following URL: .

Procedure

1. Download and install PuTTY for Windows.
2. Start the PuTTY client and complete the following settings:
   - Host Name (or IP Address): Enter the EIP of the instance to be connected.
   - Port: Select the default port 22.
   - Connection Type: Select SSH.
   - Saved Session: Enter the name of the session. Click **Save**. After the settings are saved, PuTTY remembers the name and IP address of the instance. This eliminates the need to enter them every time you connect to the instance.

3. Click **Open** to connect to the instance.

   When you connect to the instance for the first time, PuTTY displays security alerts. Click **Yes** to proceed.
4. Enter the username root and press Enter.
5. Enter the password for the instance and press Enter.

   If a message similar to the following one appears, a connection to the instance is established.

```
Welcome to aliyun Elastic Compute Server!
```

2.1.3.5.4. Connect to a Windows instance by using RDC

This topic describes how to connect to a Windows instance by using Remote Desktop Connection (RDC).

Prerequisites

- A security group and a Windows instance are created.
- The instance is in the **Running** state.
- A logon password is set for the instance.
- An Elastic IP address is associated with the instance.
- An inbound security group rule is added to the security group to allow traffic on the RDP port.

Rule direction	Action	Protocol	Port range	Priority	Authorization type	Authorization object
Inbound	Allow	tcp	3389/3389	1	IPv4 addresses	0.0.0.0/0

Procedure

1. Use one of the following methods to enable RDC:
   - Click **Start**, enter *mstsc* in the search box, and click **mstsc** in the search result.
   - Press the Windows logo key+R. In the **Run** dialog box that appears, enter *mstsc* and click **OK**.
2. In the **Remote Desktop Connection** dialog box, enter the Elastic IP address of the instance and click **Show Options**.
3. Enter the username.

   The default username is administrator.
4. (Optional)If you do not want to enter the password upon subsequent logons, select **Allow me to save credentials**.
5. Click **Connect**.
6. In the **Windows Security** dialog box that appears, enter the password corresponding to the username you entered and click **OK**.

Result

If the Windows desktop appears, a connection to the Windows instance is established.

If an error message is returned indicating that an authentication error has occurred and the function requested is not supported, install CredSSP updates and try again. Follow these steps to install the updates:

1. Connect to an ECS instance by using the VNC.
2. Choose **Start > Control Panel**.
3. Click **System and Security**.
4. Click **Check for updates** in the **Windows Updates** section.

5. If updates are available, click **Install updates**.
6. Restart the instance.

2.1.3.5.5. Connect to an instance by using a VNC management terminal

If other remote connection tools such as PuTTy, Xshell, and SecureCRT are not installed or do not work properly, you can access your instances by using a VNC management terminal in the ECS console.

Prerequisites

- The instance to which you want to connect is in the **Running** state.
- The root certificate is imported to your web browser. For more information, see Install the certificate for VNC in Windows.
- The VNC password is reset if it is your first time to connect to the instance after the instance is created. For more information, see Change the VNC password.

Context

The VNC password is used to log on to a VNC management terminal in the ECS console, whereas the instance password is used to log on to the instance.

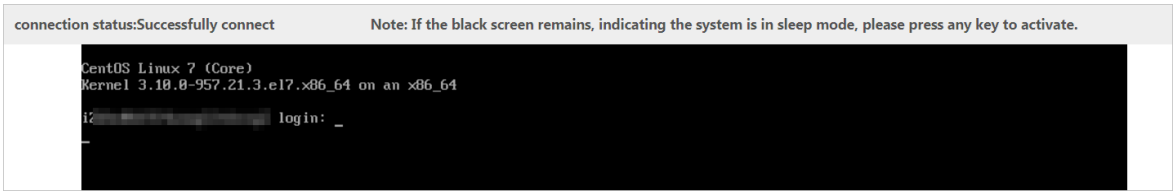
You can use a VNC management terminal to connect to an instance to solve specific issues. The following table lists some of the issues.

Issue	Solution
The instance starts slowly due to self-check on startup.	Check the self-check progress.
The firewall of the operating system is enabled by mistake.	Disable the firewall.
Abnormal processes appear and consume large amounts of CPU or bandwidth resources.	Troubleshoot and terminate the abnormal processes.

Procedure

1.
2.
3.
4. Find the instance to which you want to connect and click **Remote Connection** in the **Actions** column.
5. Enter the VNC password and click **OK**.

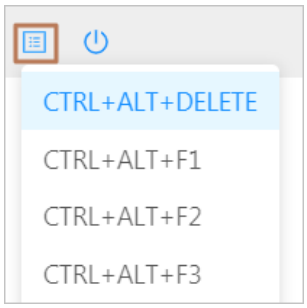
   After you are logged on to the VNC management terminal, a logon page similar to the following one appears.



6. Enter your username and password.
   - Fo a Linux instance, enter the username *root* and the logon password.

> **Note** Passwords in Linux are not displayed as you type. Press the Enter key after you enter the password.

- For a Windows instance, to use a key combination such as Ctrl+Alt+Delete, click the List icon in the upper-right corner of the VNC page and select the corresponding key combination from the drop-down list.



Enter the username and password as prompted, and click the Log On icon such as  .

2.1.4. Instances

2.1.4.1. Create an instance

An Elastic Compute Service (ECS) instance is a virtual machine that contains the basic computing components of a server, such as CPU, memory, operating system, network settings, and disks.

Prerequisites

- A virtual private cloud (VPC) and a vSwitch are created. For more information, see the "Create a VPC" and "Create a vSwitch" topics in *Apsara Stack VPC User Guide*.
- If you want to assign an IPv6 address to the instance that you want to create, make sure that the VPC and vSwitch are associated with IPv6 CIDR blocks. For more information, see the "Enable an IPv6 CIDR block for a VPC" and "Enable an IPv6 CIDR block for a vSwitch" topics in *Apsara Stack VPC User Guide*.
- One or more security groups are available. If no security groups are available, create one. For more information, see Create a security group.

Context

Some limits apply when you create GPU-accelerated instances. For more information, see Limits.

Procedure

1.
2.
3.
4. Click **Create Instance**.
5. Configure the parameters described in the following tables.
   i. Configure the basic settings of the instance.

Parameter	Required	Description
Organization	Yes	Select an organization.
Resource Set	Yes	Select a resource set.

ii. Select a region and zone for the instance.

Parameter	Required	Description
Region	Yes	Select a region.
Zone	Yes	Select a zone. Zones are the physical zones with separate power supplies and networks within the same region. The internal networks of zones are connected, and faults in one zone are isolated from the other zones. To increase the availability of your applications, we recommend that you create instances in different zones.

iii. Configure network settings for the instance.

Parameter	Required	Description
Network Type	Yes	Select a network type. Only **VPC** is supported.
VPC	Yes	Select a VPC.
vSwitch	Yes	Select a vSwitch.
Private IP Address	No	Specify a private IPv4 address for the instance. The private IPv4 address must be within the CIDR block of the selected vSwitch. If you do not specify a private IP address, the system allocates one to the instance.
IPv6	No	Specify whether to assign an IPv6 address to the instance.

iv. Configure instance settings such as security group, instance family, and instance type and specify the number of instances that you want to create.

Parameter	Required	Description
Security Group	Yes	Select a security group.
Deployment Set	No	Select a deployment set. You can use deployment sets to disperse or aggregate the instances involved in your business.
User Data	No	In the User Data field, enter the user data to be automatically run on instance startup. ■ Windows supports batch and PowerShell scripts. Before you perform Base64 encoding of user data, make sure that the data to be encoded includes `[bat]` or `[powershell]` as the first line. ■ Linux supports shell scripts.
Quantity	Yes	Specify the number of instances that you want to create. The number must be an integer within the range of 1 to 100.
Instance Family	Yes	Select an instance family. After you select an instance family, you must select an instance type.
Instance Type	Yes	Select an instance type. Information such as CPU, memory, and instance family level are displayed in the Instance Type list. Select an instance type based on your business needs. Instance types that have specific CPU and memory combinations do not support Windows Server images. For more information, see the "Limits" topic in *ECS Product Introduction*.

v. Configure the image to be used by the instance.

Parameter	Required	Description
Image Type	Yes	Select an image type. Valid values: **Public Image**, **Custom Image**, and **Shared Custom Image**.

Parameter	Required	Description
Public Image	Subject to the image type	Select a public image. Public images provided by Alibaba Cloud are licensed, secure, and stable. Public images include Windows Server images and major Linux images. This parameter is required when you set Image Type to **Public Image**. When you use an image that supports Dynamic Host Configuration Protocol version 6 (DHCPv6) to create an instance, an IPv6 address is automatically assigned to the instance. The instance can use this IPv6 address to communicate over the internal network. When you use an image that does not support DHCPv6 to create an instance, you must manually assign an IPv6 address to the instance. The following images support DHCPv6: ■ Linux images: ■ CentOS 7.6 IPV6 64Bit ■ CentOS 6.10 64Bit ■ SUSE Linux Enterprise Server 12 SP4 64Bit ■ Windows Server images **Note** To use an IPv6 address to communicate over the Internet, you must enable public bandwidth for the IPv6 address. For more information, see the "Enable Internet bandwidth for an IPv6 address" topic in *Apsara Stack VPC User Guide*.
Custom Image	Subject to the image type	Select a custom image. Custom images are created from instances or snapshots or are imported from your device. This parameter is required when you set Image Type to **Custom Image**.
Shared Custom Image	Subject to the image type	Select a custom image that is shared by another Apsara Stack tenant. This parameter is required when you set Image Type to **Shared Custom Image**.

vi. Configure the storage settings for the instance.

Parameter	Required	Description
System Disk	Yes	Select a disk category from the drop-down list and specify the system disk capacity. Valid values for the disk category: **Ultra Disk** and **Standard SSD**. The system disk capacity must range from 20 GiB to 500 GiB.
Data Disk	No	You can click Data Disk to create and attach data disks. For each data disk, select a disk category from the drop-down list and specify the disk capacity. Valid values for the disk category: **Ultra Disk** and **Standard SSD**. A maximum of 16 data disks can be attached to an instance. The maximum capacity of each data disk is 32 TiB. You can optionally select **Release with Instance** and **Encryption** for each data disk. To encrypt a data disk, configure the following parameters: ■ **Encryption Method**: Select **AES256**. ■ **Encryption Key**: You can select a key created in **Key Management Service (KMS)**. You can also add data disks after the instance is created. For more information, see Create a disk.

vii. Configure the logon password settings for the instance.

Parameter	Required	Description
Password Setting	Yes	Specify when to set a password. Valid values: **Set Now** and **Set After Purchase**. If you select Set After Purchase, you can use the password reset feature to set a password after the instance is created. For more information, see Change the logon password of an instance. **Note** The password is used to log on to the instance, not to a VNC management terminal.
Logon Password	No	Set the password to be used to log on to the instance. The password must be 8 to 30 characters in length and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. The supported special characters include `( ) ` ~ ! @# $%^&*-_+=|{}[]:;'<>,.?/` **Note** This parameter is required when you set **Password Setting** to **Set Now**.
Confirm Password	No	Enter the password again. **Note** This parameter is required when you set **Password Setting** to **Set Now**.

viii. (Optional) Enter a name for the instance.

The name must be 2 to 128 characters in length and can contain `hyphens (-), underscores (_), and colons (:)`. It must start with a letter and cannot start with `http://` or `https://`.

If you do not specify a name, the system assigns one at random.

6. Click **Submit**.

Result

The new instance appears in the instance list. When the instance is being created, it is in the **Preparing** state. When the instance is created, it enters the **Running** state.

2.1.4.2. Connect to an instance

2.1.4.2.1. Instance connecting overview

After an instance is created, you can connect to the instance to perform operations such as installing applications.

You can use one of the following methods to connect to an instance:

- Use remote connection tools to connect to instances that have public IP addresses. For more information about the procedure, see the following topics:
  - Connect to a Linux-based instance by using SSH commands in Linux or Mac OS X
  - Connect to a Linux-based instance by using remote connection tools in Windows

- Connect to a Windows-based instance by using RDP

- Use the VNC feature in the ECS console. For more information, see Connect to an instance by using a VNC management terminal.

The username of a Windows instance is Administrator, and that of a Linux instance is root.

2.1.4.2.2. Connect to a Linux instance by using SSH commands in Linux or Mac OS X

This topic describes how to use SSH commands to connect to a Linux instance.

Prerequisites

- The instance and the security group are created.
- The instance is in the **Running** state.
- A logon password is set for the instance.
- An Elastic IP address (EIP) is bound with the instance.
- An inbound security group rule is added to the security group to allow the SSH port.

Rule direction	Authorization policy	Protocol type	Port range	Priority	Authorization type	Authorization object
Inbound	Accept	TCP	22/22	1	IPv4 CIDR block	0.0.0.0/0

Procedure

1. Enter the following command and press the Enter key.

```
ssh root@instance IP
```

2. Enter the instance password of the root user and press the Enter key.

2.1.4.2.3. Connect to a Linux-based instance by using remote connection tools in Windows

This topic describes how to connect to an instance by using the PuTTY tool.

Prerequisites

Remote connection tools are designed with similar logics. In this example, PuTTY is used to connect to an instance. Download PuTTY at the following URL: .

Procedure

1. Download and install PuTTY for Windows.
2. Start the PuTTY client and complete the following settings:
   - Host Name (or IP Address): Enter the EIP of the instance to be connected.
   - Port: Select the default port 22.
   - Connection Type: Select SSH.
   - Saved Session: Enter the name of the session. Click **Save**. After the settings are saved, PuTTY remembers the name and IP address of the instance. This eliminates the need to enter them every time you connect to the instance.

3. Click **Open** to connect to the instance.

   When you connect to the instance for the first time, PuTTY displays security alerts. Click **Yes** to proceed.
4. Enter the username root and press Enter.
5. Enter the password for the instance and press Enter.

   If a message similar to the following one appears, a connection to the instance is established.

```
Welcome to aliyun Elastic Compute Server!
```

2.1.4.2.4. Connect to a Windows instance by using RDC

This topic describes how to connect to a Windows instance by using Remote Desktop Connection (RDC).

Prerequisites

- A security group and a Windows instance are created.
- The instance is in the **Running** state.
- A logon password is set for the instance.
- An Elastic IP address is associated with the instance.
- An inbound security group rule is added to the security group to allow traffic on the RDP port.

Rule direction	Action	Protocol	Port range	Priority	Authorization type	Authorization object
Inbound	Allow	tcp	3389/3389	1	IPv4 addresses	0.0.0.0/0

Procedure

1. Use one of the following methods to enable RDC:
   - Click **Start**, enter *mstsc* in the search box, and click **mstsc** in the search result.
   - Press the Windows logo key+R. In the **Run** dialog box that appears, enter *mstsc* and click **OK**.
2. In the **Remote Desktop Connection** dialog box, enter the Elastic IP address of the instance and click **Show Options**.
3. Enter the username.

   The default username is administrator.
4. (Optional)If you do not want to enter the password upon subsequent logons, select **Allow me to save credentials**.
5. Click **Connect**.
6. In the **Windows Security** dialog box that appears, enter the password corresponding to the username you entered and click **OK**.

Result

If the Windows desktop appears, a connection to the Windows instance is established.

If an error message is returned indicating that an authentication error has occurred and the function requested is not supported, install CredSSP updates and try again. Follow these steps to install the updates:

1. Connect to an ECS instance by using the VNC.
2. Choose **Start > Control Panel**.
3. Click **System and Security**.
4. Click **Check for updates** in the **Windows Updates** section.

5. If updates are available, click **Install updates**.
6. Restart the instance.

2.1.4.2.5. Install the certificate for VNC in Windows

Before you log on to the VNC management terminal, you must export the relative certificate from the site such as the Apsara Uni-manager Management Console and install the certificate in your local browser.

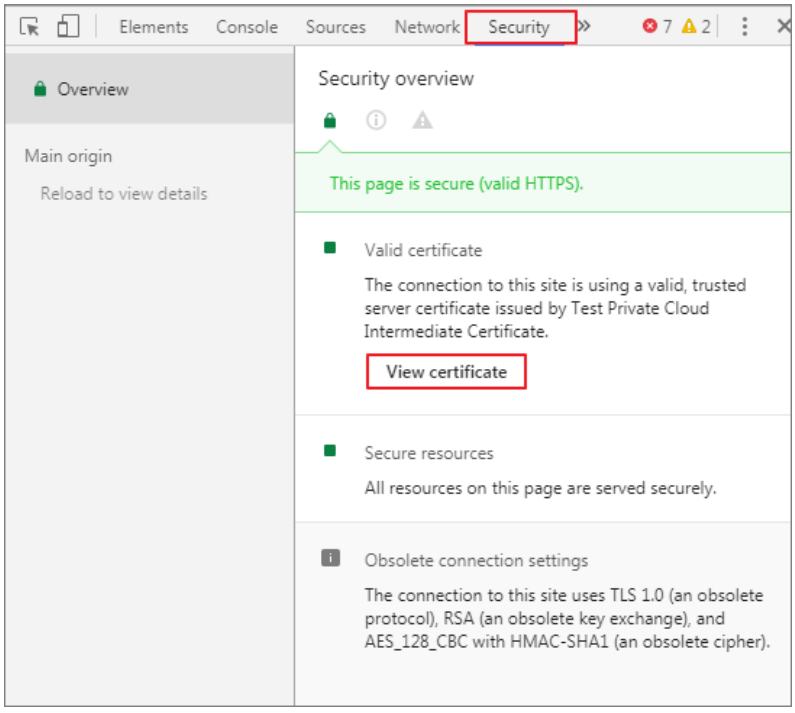
Context

The VNC feature is provided by the VNC proxy service. The VNC proxy service uses a certificate different from that of Apsara Infrastructure Management Framework. The certificate of the VNC proxy service must be manually imported.

Procedure

1. Export the certificate from the Apsara Uni-manager Management Console.
   i. Log on to the Apsara Uni-manager Management Console. Press the **F12** key or **Fn+F12** to view and select the certificate.

      For example, in the Chrome browser, press the **F12** key to open Chrome DevTools.



   ii. In the **Certificate** dialog box, click the **Certificate Path** tab, select the root certificate, and then click **View Certificate**.
   iii. In the **Certificate** dialog box, click the **Details** tab and then click **Copy to File**.
   iv. In the **Certificate Export Wizard** dialog box, click **Next**.
   v. Select **DER encoded binary X.509 (.CER)** as the format and then click **Next**.
   vi. Click **Browse**, choose where to store the certificate, enter a file name, and then click **Save**.
   vii. Click **Next**.
   viii. Click **Finish**.
   ix. Click **OK**.
2. Install the certificate in your local browser.

i. Double-click the certificate.
ii. In the **Certificate** dialog box, click **Install Certificate**.
iii. In the **Certificate Import Wizard** dialog box, click **Next**.
iv. Select **Place all certificates in the following store** and click **Browse**.
v. In the **Select Certificate Store** dialog box, select **Trusted Root Certificate Authority** and then click **OK**.
vi. In the **Certificate Import Wizard** dialog box, click **Next**.
vii. Click **Finish**.
viii. If a security warning message is displayed, click **Yes**.

3. Restart your browser and log on to the Apsara Uni-manager Management Console.

   After the certificate is installed, the security warning message is no longer displayed on the left of the URL when you log on to the Apsara Uni-manager Management Console.

2.1.4.2.6. Connect to an instance by using a VNC management terminal

If other remote connection tools such as PuTTy, Xshell, and SecureCRT are not installed or do not work properly, you can access your instances by using a VNC management terminal in the ECS console.

Prerequisites

- The instance to which you want to connect is in the **Running** state.
- The root certificate is imported to your web browser. For more information, see Install the certificate for VNC in Windows.
- The VNC password is reset if it is your first time to connect to the instance after the instance is created. For more information, see Change the VNC password.

Context

The VNC password is used to log on to a VNC management terminal in the ECS console, whereas the instance password is used to log on to the instance.

You can use a VNC management terminal to connect to an instance to solve specific issues. The following table lists some of the issues.

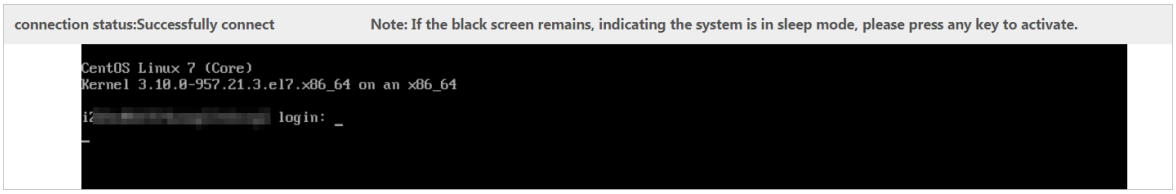
Issue	Solution
The instance starts slowly due to self-check on startup.	Check the self-check progress.
The firewall of the operating system is enabled by mistake.	Disable the firewall.
Abnormal processes appear and consume large amounts of CPU or bandwidth resources.	Troubleshoot and terminate the abnormal processes.

Procedure

1.
2.

3.
4. Find the instance to which you want to connect and click **Remote Connection** in the **Actions** column.
5. Enter the VNC password and click **OK**.

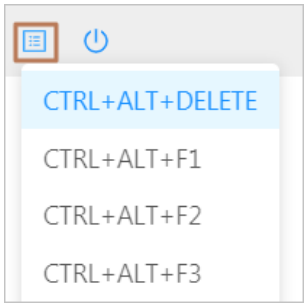
   After you are logged on to the VNC management terminal, a logon page similar to the following one appears.



6. Enter your username and password.
   - Fo a Linux instance, enter the username *root* and the logon password.

     Note Passwords in Linux are not displayed as you type. Press the Enter key after you enter the password.

   - For a Windows instance, to use a key combination such as Ctrl+Alt+Delete, click the List icon in the upper-right corner of the VNC page and select the corresponding key combination from the drop-down list.



   Enter the username and password as prompted, and click the Log On icon such as  .

2.1.4.3. View instances

You can view the list of instances and the details of individual instances. The details of an instance include basic configurations, disks, snapshots, security groups, and elastic network interfaces (ENIs).

Procedure

1.
2.
3.
4. Select a filter option from the drop-down list, enter the relevant information in the search bar, and click **Search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
Instance Name	Enter an instance name to search for the instance.
Instance ID	Enter an instance ID to search for the instance.
IP Address	Enter the IP address of an instance to search for the instance.

Filter option	Description
VPC ID	Enter a VPC ID to search for instances that belong to the VPC.
Image ID	Enter an image ID to search for instances that use the image.
Status	Select an instance status to search for instances in that status. Valid values: ○ **Running** ○ **Stopped** ○ **Starting** ○ **Stopped**
Security Group ID	Enter a security group ID to search for instances that belong to the security group.
Operating System	Enter the name of operating system to search for instances that use the operating system.

5. Use one of the following methods to go to the Instance Details page of an instance:
   - In the **Instance ID/Name** column, click the instance ID.
   - Click **Manage** in the **Actions** column corresponding to the instance.
   - Choose **More > Show Details** in the **Actions** column corresponding to the instance.

2.1.4.4. Modify an instance

You can modify the names, descriptions, and user data of created instances.

Procedure

1.
2.
3.
4. Find the instance that you want to modify and choose **More > Modify** in the **Actions** column.
5. Modify the name, description, and user data of the instance.

   The name must be 2 to 128 characters in length. The description must be 2 to 256 characters in length. The user data must be 2 to 999 characters in length.
6. Click **OK**.

2.1.4.5. Stop an instance

You can stop instances that are not in use. The stop operation interrupts the services that are running on the instances. Exercise caution when you perform this operation.

Prerequisites

The instance that you want to stop is in the **Running** state.

Procedure

1.
2.
3.

4. Use one of the following methods to stop instances:
   - To stop a single instance, find the instance and choose **More > Instance Status > Stop** in the **Actions** column.
   - To stop one or more instances at a time, select the instances and click **Stop** in the lower-left corner of the Instances page.
5. Click **OK**.

Result

When the instance is being stopped, its status in the **Status** column changes from **Running** to **Stopping**. After the instance is stopped, its status changes to **Stopped**.

2.1.4.6. Start an instance

You can start stopped instances.

Prerequisites

The instance that you want to start is in the **Stopped** state.

Procedure

1.
2.
3.
4. Use one of the following methods to start instances:
   - To start a single instance, find the instance and choose **More > Instance Status > Start** in the **Actions** column.
   - To start one or more instances at a time, select the instances and click **Start** in the lower-left corner of the Instances page.
5. Click **OK**.

Result

When the instance is being started, its status in the **Status** column changes from **Stopped** to **Starting**. After the instance is started, its status changes to **Running**.

2.1.4.7. Restart an instance

You must restart instances after you change their logon passwords or install system updates for the instances. The restart operation stops the instances for a period of time. This causes the services that are running on the instances to be interrupted. Exercise caution when you perform this operation.

Prerequisites

The instance that you want to restart is in the **Running** state.

Procedure

1.
2.
3.
4. Use one of the following methods to restart instances:
   - To restart a single instance, find the instance and choose **More > Instance Status > Restart** in the **Actions** column.

- To restart one or more instances at a time, select the instances and click **Restart** in the lower-left corner of the Instances page.

5. In the Restart Instance dialog box, select a restart mode.
   - **Restart**: restarts the instances normally.
   - **Force Restart**: forcibly restarts the instances. This may result in loss of unsaved data.
6. Click **OK**.

2.1.4.8. Delete an instance

You can delete instances that are no longer needed to release their resources. Deleted instances cannot be recovered. We recommend that you back up data before you delete instances. If data disks are released along with instances, the data on the disks cannot be recovered.

Prerequisites

The instance to be deleted is in the **Stopped** state.

Procedure

1.
2.
3.
4. Select the instance that you want to delete and click **Delete** in the lower-left corner of the Instances page.
5. Click **OK**.

2.1.4.9. View the monitoring information of an instance

You can view monitoring charts in the CloudMonitor console to learn about the running conditions of Elastic Compute Service (ECS) instances. This topic describes how to go to the CloudMonitor console to view the monitoring information of an ECS instance.

Context

CloudMonitor provides real-time monitoring, alerting, and notification services for resources to protect your services and business. For more information, see **CloudMonitor overview** in *Apsara Uni-manager Management Console User Guide*.

Procedure

1.
2.
3.
4. Find the ECS instance whose monitoring information you want to view and click **Monitor** in the Monitoring column.
5. On the **Monitoring Charts** page, view the monitoring information of the ECS instance.

2.1.4.10. Change the instance type of an instance

You can change the instance types of instances to suit your business needs. This eliminates the need to create new instances.

Hot configuration change

You can perform a hot configuration change on an instance that is not in the Stopped state to change its instance type.

> **Note** Only specific instance types support hot configuration changes. The instance types that support hot configuration changes are displayed in the Elastic Compute Service (ECS) console.

1.
2.
3.
4. Find the ECS instance whose instance type you want to change and click **Hot Configuration Change** in the **Actions** column.
5. On the Change Configurations of ECS Instance without Service Interruption page, select a new instance type and click **Submit**.

   The Change Configurations of ECS Instance without Service Interruption page shows the instance types available for selection.

Upgrade and downgrade

You can upgrade or downgrade the instance type of an instance that is in the **Stopped** state.

1.
2.
3.
4. Find the instance whose instance type you want to change and click **Upgrade/Downgrade** in the **Actions** column.
5. On the Change Configurations of ECS Instance page, select a new instance type and click **Submit**.

   The Change Configurations of ECS Instance page shows the instance types available for selection.
6. Restart the instance by using the ECS console or by calling an API operation for the new instance type to take effect.

   For more information, see Start an instance or the "StartInstance" topic in *ECS Developer Guide*.

2.1.4.11. Change the logon password of an instance

If you did not set a logon password when you created an instance or have forgotten the password, you can reset the password.

Procedure

1.
2.
3.
4. Find the instance whose logon password you want to change and use one of the following methods to go to the Instance Details page.
   - In the **Instance ID/Name** column, click the ID of the instance.
   - Click **Manage** in the **Actions** column.
   - In the **Actions** column, choose **More > Show Details**.
5. Click **Change Password**.
6. Enter and confirm the new password, and then click **OK**.

   The password must be 8 to 30 characters in length and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. The supported special characters include `( ) ' ~ ! @ # $ % ^ & * - _ + = | { } [ ] : ; ' < > , . ? /`
7. Restart the instance by using the ECS console or by calling an API operation for the new password to take

effect.

For more information, see Restart an instance or the "RebootInstance" topic in *ECS Developer Guide*.

2.1.4.12. Change the VNC password

If you use Virtual Network Console (VNC) to connect to an Elastic Compute Service (ECS) instance for the first time or if you forget the VNC password, you can reset the VNC password.

Procedure

1.
2.
3.
4. Find the instance whose VNC password you want to change and use one of the following methods to go to the Instance Details page:
   - On the Instances page, click the instance ID in the **Instance ID/Name** column.
   - On the Instances page, click **Manage** in the **Actions** column.
   - On the Instances page, choose **More > Show Details** in the **Actions** column.
5. Click **Change VNC Password**.
6. Enter and confirm the new password, and click **OK**.

   The password must be 8 to 14 characters in length and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. The supported special characters include `! / | \@~`#$%^*()_-+={[}]:;,.?` .

   > **Note** If your instance is a non-I/O optimized instance, you must restart the instance in the ECS console for the new password to take effect.

2.1.4.13. Add an instance to a security group

You can add created instanced to security groups and use security group rules to control network access for the instances.

Context

Security groups act as virtual firewalls to provide security isolation and implement network access control for instances.

Security groups determine whether the instances in the same account that are deployed within the same VPC and region can communicate with each other. By default, if the instances belong to the same security group, they can communicate with each other over the internal network. If the instances belong to different security groups, you can authorize mutual access between the security groups to allow the instances to communicate with each other over the internal network.

Procedure

1.
2.
3.
4. Find the security group to which you want to add an instance and click **Manage Instances** in the **Actions** column.
5. Click **Add Instance**.
6. In the Add Instance dialog box, select an instance and click **OK**.

An instance can belong to up to five security groups. After an instance is added to a security group, the rules of the security group automatically apply to the instance.

2.1.4.14. Customize instance data

ECS allows you to run the instance customization script upon startup and import data into instances.

Context

The instance data customization feature is applicable to both Windows and Linux instances. It allows you to:

- Run the instance customization script upon startup.
- Import data into instances.

Usage instructions

- **Limits**

  The instance data customization feature can only be used when an instance meets all the following requirements:

  - Network type: VPC
  - Image: a system image or a custom image that is inherited from the system image
  - Operating system: one type included in Supported operating systems Supported operating systems

Windows	Linux
■ Windows Server 2016 64-bit ■ Windows Server 2012 64-bit ■ Windows Server 2008 64-bit	■ CentOS ■ Ubuntu ■ SUSE Linux Enterprise ■ OpenSUSE ■ Debian ■ Aliyun Linux

  - When you configure instance data customization scripts, you must enter custom data based on the type of operating system and script.

    Note Only English characters are allowed.

    - If your data is Base64 encoded, select **Enter Base64 Encoded Information**.

      Note The size of the customization script cannot exceed 16 KB before the data is Base64 encoded.

    - For Linux instances, the script format must meet the requirements described in Types of Linux instance customization scripts.
    - For Windows instances, the script can only start with `[bat]` or `[powershell]`.

- After starting an instance, run a command to view the following information:
  - Execution result of the instance customization script
  - Data imported to instances

- **Console**: You can modify the custom instance data in the console. Whether the modified instance customization script needs to be re-executed depends on the script type. For example, if the `bootcmd` script in Cloud Config is modified for Linux instances, the script is automatically executed each time instances are restarted.

- **OpenAPI**: You can also use OpenAPI to customize instance data. For more information, see **CreateInstance** and **ModifyInstanceAttribute** in *ECS Developer Guide*.

Linux instance data customization scripts

Linux instance data customization scripts provided by Alibaba Cloud are designed based on the cloud-init architecture. They are used to automatically configure parameters of Linux instances. Customization script types are compatible with the cloud-init.

Description of Linux instance data customization scripts

- Linux instance customization scripts are executed after instances are started and before `/etc/init` is executed.
- Linux instance customization scripts can only be executed with root permissions by default.

Types of Linux instance customization scripts

- **User-Data Script**
  - Description: A script, such as shell script, is used to customize data.
  - Format: The first line must include `#!`, such as `#! /bin/sh`.
  - Limit: The script size (including the first line) cannot exceed 16 KB before the data is Base64 encoded.
  - Frequency: The script is executed only when instances are started for the first time.
  - Example:

```
#! /bin/sh
echo "Hello World. The time is now $(date -R)!" | tee /root/output10.txt
```

- **Cloud Config Data**
  - Description: Predefined data is used to configure services, such as specifying yum sources or importing SSH keys.
  - Format: The first line must be `#cloud-config`.
  - Limit: The script size (including the first line) cannot exceed 16 KB before the data is Base64 encoded.
  - Frequency: The script execution frequency varies with the specific service.
  - Example:

```
#cloud-config
apt:
primary:
- arches: [default]
uri: http://us.archive.ubuntu.com/ubuntu/
```

- **Include**
  - Description: The configuration content can be saved in a text file and imported into cloud-init as a URL.
  - Format: The first line must be `#include`.
  - Limit: The script size (including the first line) cannot exceed 16 KB before the data is Base64 encoded.
  - Frequency: The script execution frequency depends on the script type in the text file.
  - Example:

```
#include
http://ecs-image-test.oss-cn-hangzhou.aliyuncs.com/userdata/cloudconfig
```

- **GZIP format**
  - Description: Cloud-ini limits the size of customization scripts to 16 KB. You can compress and import the script file into the customization script if the file size exceeds 16 KB.

- Format: The .gz file is imported into the customization script as a URL in `#include`.
- Frequency: The script execution frequency depends on the script content contained in the GZIP file.
- Example:

```
#include
http://ecs-image-test.oss-cn-hangzhou.aliyuncs.com/userdata/config.gz
```

View the custom data of a Linux instance

Run the following command in the instance:

```
curl http://100.100.100.200/latest/user-data
```

Windows instance customization scripts

Windows instance customization scripts independently developed by Alibaba Cloud can be used to initialize Windows instances.

There are two types of Windows instance customization scripts:

- Batch processing program: starts with `[bat]` and serves as the first line. The script size must be smaller than 16 KB before the data is Base64 encoded.
- PowerShell script: starts with `[powershell]` and serves as the first line. The script size must be smaller than 16 KB before the data is Base64 encoded.

View the custom data of a Windows instance

Run the following PowerShell command in the instance:

```
Invoke-RestMethod http://100.100.100.200/latest/user-data/
```

2.1.4.15. Change the private IP address of an instance

Each instance is assigned a private NIC and associated with a private IP address. You can change the private IP address of an instance. The new private IP address must be within the CIDR block of the vSwitch to which the instance is connected and must not be in use by another instance.

Prerequisites

The instance whose private IP address you want to change is in the **Stopped** state.

Procedure

1.
2.
3.
4. Find the instance whose private IP address you want to change and choose **More > Change Private IP Address** in the **Actions** column.
5. Enter a new private IP address.

   The new private IP address must be within the CIDR block of the vSwitch to which the instance is connected. This IP address must not be in use by another instance or be reserved for a specific purpose.

   For example, if the CIDR block of the vSwitch is 192.168.1.0/24, you can use an IP address within the range from 192.168.1.1 to 192.168.1.254. The first address 192.168.1.0 is reserved to identify the subnet itself, and the last address 192.168.1.255 is reserved as the broadcast address. Neither address can be used.
6. Click **OK**.

2.1.4.16. Enable IPv6

Compared with IPv4 addresses, IPv6 addresses are more sufficient and allow more types of devices to access the Internet. If your network environment supports IPv6, you can assign IPv6 addresses for exiting Elastic Compute Service (ECS) instances.

Prerequisites

- The virtual private cloud (VPC) of the ECS instance to which you want to assign an IPv6 address is associated with an IPv6 CIDR block. For more information, see the "Enable an IPv6 CIDR block for a VPC" topic in *Apsara Stack VPC User Guide*.
- The vSwitch connected to the ECS instance to which you want to assign an IPv6 address is associated with an IPv6 CIDR block. For more information, see the "Enable an IPv6 CIDR block for a vSwitch" topic in *Apsara Stack VPC User Guide*.
- The instance family of the ECS instance supports IPv6.

Procedure

1. Log on to the ECS console.
2.
3.
4. Find the instance to which you want to assign an IPv6 address and choose **More > Manage Secondary Private IPv6 Addresses** in the **Actions** column.
5. In the **Manage Secondary Private IPv6 Addresses** dialog box, click **Open Now** next to the **IPv6 Addresses** field.

   > **Note** Specific instance families do not support IPv6.

6. On the vSwitch page, check whether IPv6 is enabled for the vSwitch.

   If IPv6 is not enabled for the vSwitch, click **Enable IPv6 CIDR Block** in the **IPv6 CIDR Block** column corresponding to the vSwitch.
7. In the **Manage Secondary Private IPv6 Addresses** dialogue box, click **OK**.

Result

After the configuration is complete, you can click the instance ID and check whether an **IPv6** address is assigned to the instance on the **Instance Details** page.

2.1.4.17. Install the CUDA and GPU drivers for a Linux instance

You must install a GPU driver on GPU instances to use the GPU. If the image you use does not contain a pre-installed GPU driver, you must manually install the CUDA and GPU drivers for the instance.

Prerequisites

If your instance cannot connect to the Internet, the installation file cannot be downloaded. You can install an FTP client on the instance to transfer the installation file to the instance.

Context

When installing NVIDIA drivers, you must install the kernel package that contains the kernel header file before you install the CUDA and GPU drivers on the instance.

Procedure

1. Install the kernel package.

i. Run the **uname -r** command to view the current kernel version.

A similar output is displayed:

- CentOS: `3.10.0-862.14.4.el7.x86_64`
- Ubuntu: `4.4.0-117-generic`

ii. Copy the kernel package of the corresponding version to the instance and install the package.

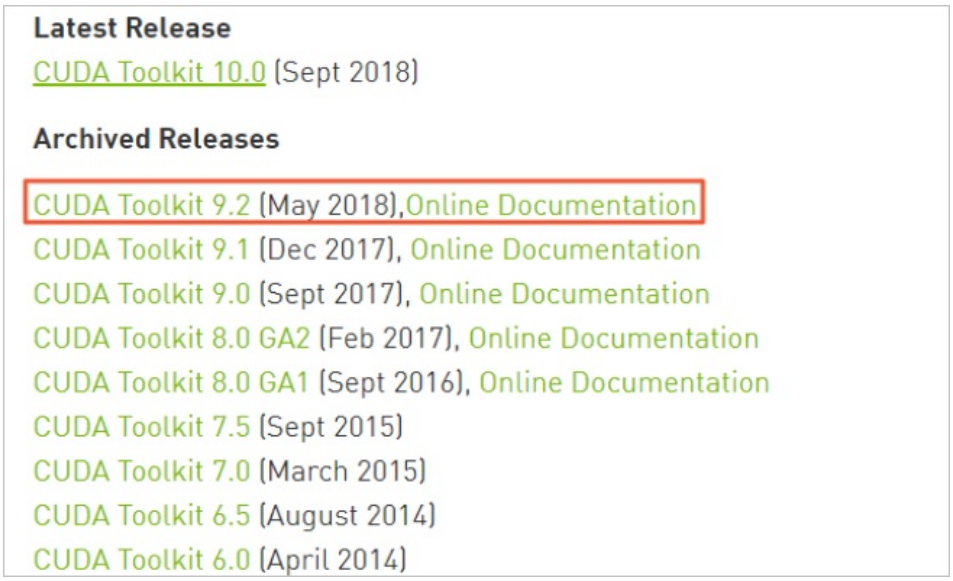
- CentOS: Copy the RPM package of the `kernel-devel` component and run the **rpm -ivh 3.10.0-862.14.4.el7.x86_64.rpm** command to install the package. *3.10.0-862.14.4.el7.x86_64.rpm* is used as an example. Replace it with the actual package name.
- Ubuntu: Copy the DEB package of the `linux-headers` component and run the **dpkg -i 4.4.0-117-generic.deb** command to install the package. *4.4.0-117-generic.deb* is used as an example. Replace it with the actual package name.

2. Download the CUDA Toolkit.

i. Access the official CUDA download page. Choose the version based on the GPU application requirements for CUDA.

This example uses CUDA Toolkit 9.2.

Download the CUDA Toolkit



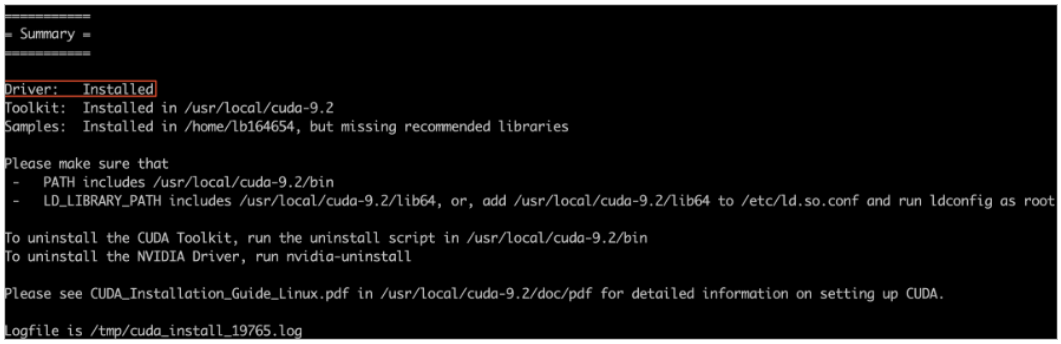
ii. Choose a platform based on your operating system. Select **Installer Type** to **runfile (local)** and click **Download**.

NVIDIA drivers are already included in the CUDA Toolkit.

Download the drivers

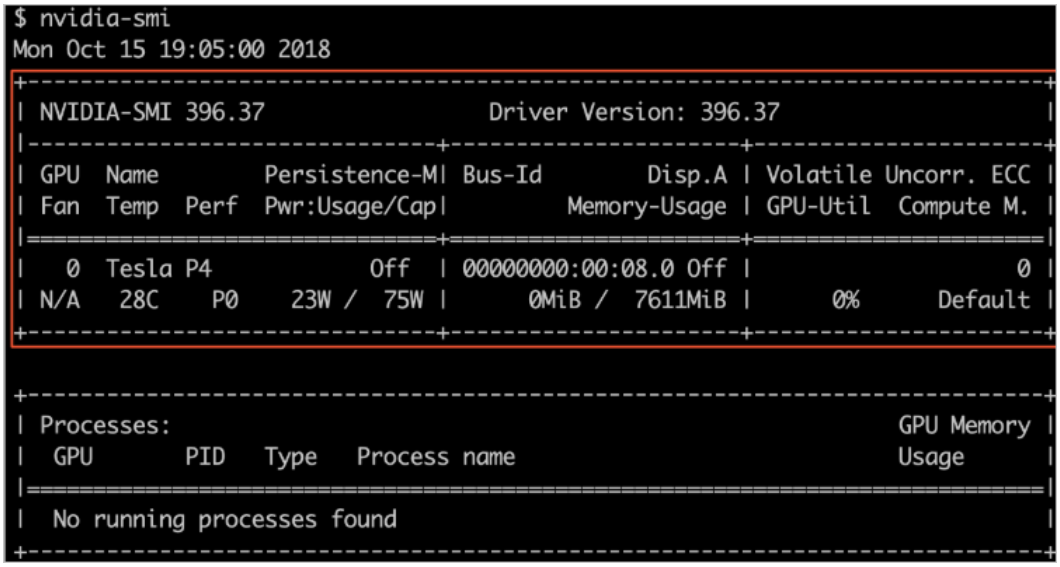
3. Copy the downloaded *cuda_9.2.148_396.37_linux.run* file to the instance. *cuda_9.2.148_396.37_linux.run* is used as an example. Replace it with the actual file name.
4. Run the **sudo sh ./cuda_9.2.148_396.37_linux.run --silent --verbose --driver --toolkit --samples** command to install the CUDA driver. *cuda_9.2.148_396.37_linux.run* is used as an example. Replace it with the actual file name.
   The installation takes about 10 to 20 minutes. When `Driver: Installed` is displayed, the installation is successful.

   View the CUDA installation result



5. Run the **nvidia-smi** command to view the GPU driver status.
   If the output displays the details of the GPU driver, the driver is running properly.

   View the GPU driver status



What's next

If you want to run the OpenGL program, you must first purchase the licenses and install the GRID drivers. For information about the installation procedure, see the official NVIDIA documentation.

2.1.4.18. Install the CUDA and GPU drivers for a Windows instance

You must install a GPU driver on GPU instances to use the GPU. If the image you use does not contain a pre-installed GPU driver, you must manually install the CUDA and GPU drivers for the instance.

Prerequisites

- If your instance cannot connect to the Internet, the installation file cannot be downloaded. You can install an FTP client on the instance to transfer the installation file to the instance.
- To compile CUDA programs, first install a Windows compiling environment, such as Visual Studio 2015. If you do not need to compile CUDA programs, ignore it.

Procedure

1. Download the CUDA Toolkit.
   i. Access the official CUDA download page. Choose the version based on the GPU application requirements for CUDA.

      This example uses CUDA Toolkit 9.2.

   ii. Choose a platform based on your operating system. Set **Installer Type** to **exe (local)** and click **Download**.

      NVIDIA drivers are already included in the CUDA Toolkit.
2. Copy the downloaded *cuda_9.2.148_windows.exe* file to the instance. *cuda_9.2.148_windows.exe* is used as an example. Replace it with the actual file name.
3. Double-click *cuda_9.2.148_windows.exe* and follow the installation wizard to install the CUDA driver. *cuda_9.2.148_windows.exe* is used as an example. Replace it with the actual file name.
   The installation takes about 10 to 20 minutes. When `Installed: - Nsight Monitor and HUD Launcher` is displayed, the driver is installed.
4. Press *Win + R* and enter **devmgmt.msc**.
   The NVIDIA device is displayed in **Display Adapter**.
5. Press *Win + R*, enter **cmd**, and run the **"C:\Program Files\NVIDIA Corporation\NVSMI\nvidia-smi"**

command.
If the output displays the details of the GPU driver, the driver is running properly.

What's next

If you want to run the OpenGL and DirectX programs, you must first purchase the licenses and install the GRID drivers. For information about the installation procedure, see the official NVIDIA documentation.

2.1.5. Disks

2.1.5.1. Create a disk

You can create standalone data disks and then attach them to ECS instances to increase the storage space of the instances. This topic describes how to create an empty data disk. You cannot create standalone system disks.

Context

We recommend that you determine the number and sizes of data disks before you create them. The following limits apply to data disks:

- A maximum of 16 data disks can be attached to an instance. Disks and Shared Block Storage devices share this quota.
- Each Shared Block Storage device can be attached to up to four ECS instances at the same time.
- Each ultra disk, standard SSD, Ultra Shared Block Storage device, or SSD Shared Block Storage device can have a maximum capacity of 32 TiB.
- Disks cannot be combined in ECS. They are independent of each other. You cannot combine multiple disks into one by formatting them.

We recommend that you do not use Logical Volume Manager (LVM) to create logical volumes across multiple disks, because a snapshot can back up data only of a single disk. If you create a logical volume across several disks, data discrepancies may occur when you restore these disks.

Procedure

1.
2.
3.
4. Click **Create Disk**.
5. Configure the parameters listed in the following table to create a disk.

Type	Parameter	Required	Description
Region	Organization	Yes	Select an organization in which to create the disk.
	Resource Set	Yes	Select a resource set in which to create the disk.
	Region	Yes	Select a region in which to create the disk.
	Zone	Yes	Select a zone in which to create the disk.
	Name	Yes	Enter a name for the disk. The name must be 2 to 128 characters in length. It must start with a letter and cannot start with http:// or https://. It can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,).

Type	Parameter	Required	Description
Basic Settings	Specifications	Yes	Select a disk category and specify the disk size. Valid values for the disk category: ○ **Standard SSD**: standard SSD ○ **Ultra Disk**: ultra disk ○ **Shared SSD**: SSD Shared Block Storage device ○ **Shared Ultra Disk**: Ultra Shared Block Storage device The disk size must range from 20 GiB to 32,768 GiB.
	Encrypted	No	Specify whether to encrypt the disk.
	Encryption Method	No	Select an encryption algorithm. This parameter is required when you set **Encrypted** to **Yes**. Valid values: ○ **AES256** ○ **SM4**
	Encryption Key	No	Select a key to be used to encrypt the disk. This parameter is required when you set **Encrypted** to **Yes**. **Note** If no key is available, create a key in KMS.
	Use Snapshot	No	Specify whether to create the disk from a snapshot. If you select **Yes**, you must specify a snapshot. The size of the created disk depends on the size of the specified snapshot. ○ If the disk size that you specify is greater than the snapshot size, the disk is created with the size you specify. ○ If the disk size that you specify is smaller than the snapshot size, the disk is created with the snapshot size.

6. Click **Submit**.

Result

The created disk is displayed in the disk list and in the **Pending** state.

What's next

After the disk is created, you must attach the disk to an instance and partition and format the disk. For more information, see the following topics:

- Attach a disk
- Format a data disk for a Linux instance
- Format a data disk of a Windows instance

2.1.5.2. Attach a disk

You can attach separately created disks as data disks to instances. To attach a disk to an instance, make sure that the disk and the instance are located within the same region and zone.

Prerequisites

The disk that you want to attach is in the **Available** state.

Context

- You do not need to attach data disks that are created together with instances.
- A disk can be attached only to a single instance within the same region and zone.
- Each disk can be attached only to a single instance at the same time.
- Each Shared Block Storage device can be attached to up to four instances at the same time.

Attach a disk on the instance details page

To attach multiple disks to an instance, you can go to the details page of the instance.

1.
2.
3.
4. Find the instance to which you want to attach a disk and click the instance ID.
5. Click the **Disks** tab.
6. Click **Attach**.
7. In the Attach dialog box, select a disk from the **Disk** drop-down list.
8. Click **OK**.

Attach a disk on the Disks page

To attach multiple disks to different instances, you can go to the Disks page.

1.
2.
3.
4. Find the disk that you want to attach and choose **More > Attach** in the **Actions** column.
5. Specify the destination instance and configure the release mode.
   - If you select **Release Disk with Instance**, the disk will be released when its associated instance is deleted.
   - If you do not select **Release Disk with Instance**, the disk will be retained and enter the **Available** state when its associated instance is deleted.
6. Click **OK**.

2.1.5.3. Partition and format disks

2.1.5.3.1. Format a data disk for a Linux instance

Data disks created separately are not partitioned or formatted. This topic describes how to partition and format a data disk of a Linux instance.

Prerequisites

The disk has been attached to the instance.

Procedure

1. Connect to the instance.

2. Run the **fdisk -l** command to view all data disks attached to the ECS instance.

   If */dev/vdb* is not displayed in the command output, the ECS instance does not have a data disk. Check whether the data disk is attached to the instance.

```
[root@iZ********eZ ~]# fdisk -l
Disk /dev/vda: 42.9 GB, 42949672960 bytes
255 heads, 63 sectors/track, 5221 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00078f9c
Device Boot      Start         End      Blocks   Id  System
/dev/vda1   *           1        5222    41940992   83  Linux
Disk /dev/vdb: 21.5 GB, 21474836480 bytes
16 heads, 63 sectors/track, 41610 cylinders
Units = cylinders of 1008 * 512 = 516096 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

3. Create partitions for the data disk.
   i. Run the **fdisk /dev/sdb** command.
   ii. Enter *n* to create a new partition.
   iii. Enter *p* to set the partition as the primary partition.
   iv. Enter a partition number and press the Enter key. In this example, *1* is entered to create Partition 1.
   v. Enter the number of the first available sector. This example uses the default value. This is done by pressing the Enter key. You can also enter a value from 1 to 41610 and press the Enter key.
   vi. Enter the number of the last sector. This example uses the default value. This is done by pressing the Enter key. You can also enter a value from 1 to 11748 and press the Enter key.
   vii. (Optional)Optional. To create multiple partitions, repeat steps b through f until all four primary partitions are created.
   viii. Run the **wq** command to start partitioning.

```
[root@iZ********eZ ~]# fdisk /dev/vdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF disklabel
Building a new DOS disklabel with disk identifier 0x01ac58fe.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
Warning: invalid flag 0x0000 of partition table 4 will be corrected by w(rite)
WARNING: DOS-compatible mode is deprecated. It's strongly recommended to
         switch off the mode (command 'c') and change display units to
         sectors (command 'u').
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-41610, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-41610, default 41610):
Using default value 41610
Command (m for help): wq
The partition table has been altered!
```

4. Run the **fdisk -l** command to view the partitions.

If */dev/vdb1* is displayed in the command output, new partition *vdb1* is created.

```
[root@iZ********eZ ~]# fdisk -l
Disk /dev/vda: 42.9 GB, 42949672960 bytes
255 heads, 63 sectors/track, 5221 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00078f9c
Device Boot      Start         End      Blocks   Id  System
/dev/vda1   *           1        5222    41940992   83  Linux
Disk /dev/vdb: 21.5 GB, 21474836480 bytes
16 heads, 63 sectors/track, 41610 cylinders
Units = cylinders of 1008 * 512 = 516096 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x01ac58fe
Device Boot      Start         End      Blocks   Id  System
/dev/vdb1               1       41610    20971408+  83  Linux
```

5. Format the new partition. In this example, the new partition is formatted as ext3 after you run the **mkfs.ext3 /dev/vdb1** command.

   The time required for formatting depends on the disk size. You can also format the new partition to another file system. For example, you can run the **mkfs.ext4 /dev/vdb1** command to format the partition as ext4.

   Compared with ext2, ext3 only adds the log function. Compared with ext3, ext4 improves on some important data structures. ext4 provides better performance and reliability, and more functions.

```
[root@iZ********leZ ~]# mkfs.ext3 /dev/vdb1
mke2fs 1.41.12 (17-May-2010)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
1310720 inodes, 5242852 blocks
262142 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
160 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000
Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done
This filesystem will be automatically checked every 25 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

6. Run the `echo '/dev/vdb1 /mnt ext3 defaults 0 0' >> /etc/fstab` command to write the information of the new partition to the /etc/fstab file. You can run the `cat /etc/fstab` command to view the new partition information.

   Ubuntu 12.04 does not support barriers. To write the information of the new partition into the /etc/fstab file, you must run the `echo '/dev/vdb1 /mnt ext3 barrier=0 0 0' >> /etc/fstab` command.

   In this example, the partition information is added to the ext3 file system. You can also modify the ext3 parameter to add the partition information to another type of file system.

To attach the data disk to a specific folder, for example, to store web pages, modify the /mnt part of the preceding command.

```
[root@iZ********eZ ~]# echo '/dev/vdb1 /mnt ext3 defaults 0 0' >> /etc/fstab
[root@iZbp19cdhgdj0aw5r2izleZ ~]# cat /etc/fstab
#
# /etc/fstab
# Created by anaconda on Thu Aug 14 21:16:42 2014
#
# Accessible filesystems, by reference, are maintained under '/dev/disk'
# See man pages fstab(5), findfs(8), mount(8) and/or blkid(8) for more info
#
UUID=94e4e384-0ace-437f-bc96-057dd64f**** / ext4 defaults,barrier=0 1 1
tmpfs                   /dev/shm                tmpfs   defaults        0 0
devpts                  /dev/pts                devpts  gid=5,mode=620  0 0
sysfs                   /sys                    sysfs   defaults        0 0
proc                    /proc                   proc    defaults        0 0
/dev/vdb1 /mnt ext3 defaults 0 0
```

7. Mount the new partitions. Run the **mount -a** command to mount all the partitions listed in */etc/fstab* and run the **df -h** command to view the result.

   If the following information is displayed, the new partitions are mounted and available for use.

```
[root@iZ********eZ ~]# mount -a
[root@iZ********eZ ~]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/vda1        40G  5.6G   32G  15% /
tmpfs           499M     0  499M   0% /dev/shm
/dev/vdb1        20G  173M   19G   1% /mnt
```

2.1.5.3.2. Format a data disk of a Windows instance

Data disks created separately are not partitioned or formatted. This topic describes how to partition and format a data disk of a Windows instance. This example uses Windows Server 2008.

Prerequisites

The disk has been attached to an instance.

Procedure

1. In the lower-left corner of the screen, click the **Server Manager** icon.
2. In the left-side navigation pane of the Server Manager window, choose **Storage > Disk Management**.
3. Right-click an empty partition and select **New Simple Volume** from the shortcut menu.

   If the disk status is **Offline**, change it to **Online**.
4. Click **Next**.
5. Set the size of the simple volume, which is the partition size. Then click **Next**.

   The default value is the maximum value of the disk space. You can specify the partition size as needed.
6. Specify the drive letter and then click **Next**.
7. Specify the formatting options and then click **Next**.

   We recommend that you format the partition with the default settings provided by the wizard.
8. When the wizard prompts that the partition has been completed, click **Finish** to close the wizard.

2.1.5.4. View disks

You can view the list of disks and the details of individual disks.

Procedure

1.
2.
3.
4. Select a filter option from the drop-down list, enter the relevant information in the search box, and click **search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
Disk Name	Enter a disk name to search for the disk.
Disk ID	Enter a disk ID to search for the disk.
Instance ID	Enter an instance ID to search for disks that are attached to the instance.
Disk Status	Select a disk status to search for disks in that status. Valid values: ○ **Running** ○ **Pending** ○ **Attaching** ○ **Detaching** ○ **Creating** ○ **Deleting** ○ **Deleted** Note Deleted disks are no longer displayed in the disk list. ○ **Initializing** ○ **All Statuses**
Disk Properties	Select a disk type to search for disks of that type. Valid values: ○ **All** ○ **System Disk** ○ **Data Disk**
Policy ID	Enter the ID of an automatic snapshot policy to search for disks that use the policy.
Encryption Key ID	Enter the ID of an encryption key to search for disks that are encrypted by using the key.

5. In the **Disk ID/Name** column, click a disk ID to go to the Disk Details page of the disk.

   The properties and mount information of the disk are displayed on the Disk Details page.

2.1.5.5. Restore a disk

If you have created snapshots for a disk, you can use one of the snapshots to restore the disk to the state when the snapshot was created. The disk restore operation is irreversible. After the disk is restored, the data stored on the disk before the restore operation is performed cannot be recovered. Exercise caution when you perform this operation.

Prerequisites

- Snapshots are created for the disk that you want to restore.
- The disk is not released.
- The instance to which the disk is attached is in the **Stopped** state.

Procedure

1.
2.
3.
4. Select a filtering option from the drop-down list, enter the relevant information in the search bar, and click **Search**.

   You can select multiple filtering options to narrow down the search results.

Filtering option	Description
Snapshot Name	Enter a snapshot name to search for the snapshot.
Snapshot ID	Enter a snapshot ID to search for the snapshot.
Instance ID	Enter an instance ID to search for the snapshots related to the instance.
Disk ID	Enter a disk ID to search for the snapshots related to the disk.
Snapshot Type	Select a snapshot type to search for the snapshots of that type. Options include: ○ **All** ○ **User Snapshots**: manual snapshots. ○ **Automatic snapshots**: automatic snapshots.
Creation Time	Enter a creation time to search for the snapshots that were created at that time.

5. Find the snapshot that you want to use to restore the specified disk and click **Restore Disk** in the **Actions** column.
6. Click **OK**.

2.1.5.6. Modify the attributes of a disk

You can modify the attributes of created disks, such as the settings of the Release Disk with Instance and Release Automatic Snapshots with Disk options.

Procedure

1.
2.
3.

4. Find the disk whose attributes you want to modify and choose **More > Modify Disk Properties** in the **Actions** column.
5. Modify the Release Mode settings.
   - **Release Disk with Instance**: If this option is selected, the disk will be released when its associated instance is deleted. If this option is not selected, the disk will be retained and enter the **Pending** state when its associated instance is deleted.
   - **Release Automatic Snapshots with Disk**: If this option is selected, the automatic snapshots of the disk will be released when the disk is deleted. If this option is not selected, the automatic snapshots will be retained when the disk is deleted.
6. Click **OK**.

2.1.5.7. Modify the description of a disk

You can modify the names and descriptions of created disks.

Procedure

1.
2.
3.
4. Find the disk that you want to modify and choose **More > Modify Disk Description** in the **Actions** column.
5. Modify the name and description of the disk.

   The name must be 2 to 128 characters in length and start with a letter. It can contain letters, digits, periods (.), underscores (_), colons (:), and hyphens (-).

   The description must be 2 to 256 characters in length and cannot start with http:// or https://.
6. Click **OK**.

2.1.5.8. Expand a disk

You can expand system disks or data disks online. After a disk is expanded, you do not need to restart the instance to which the disk is attached for the new disk capacity to take effect.

Prerequisites

- To avoid data loss, we recommend that you create a snapshot to back up disk data before you expand a disk. For more information, see Create a snapshot.
- No snapshot is being created for the disk to be expanded.
- The following requirements are met:
  - If the disk is a system disk, the associated instance is in the **Running** state.
  - If the disk is a data disk, one of the following requirements is met:
    - The disk is in the **Pending** state.
    - If the disk is attached to an instance, the instance is in the **Running** state.
  - If the disk is a Shared Block Storage device, it is in the **Pending** state.

Context

The following limits apply when you expand a disk.

Limit	Description

Limit	Description
Disk category	• Ultra disks and standard SSDs can be expanded. • SSD Shared Block Storage and Ultra Shared Block Storage devices can be expanded.
Operating system	The system disks of Windows Server 2003 instances cannot be expanded.
Partitioning mode	Data disks that use the MBR partitioning scheme cannot be expanded to larger than 2 TiB in size. To expand an MBR data disk to larger than 2 TiB in size, we recommend that you create and attach a new data disk larger than 2 TiB in size, use the GPT partitioning scheme to partition this new data disk, and then copy data from the original MBR data disk to the new GPT data disk.
File system	For Windows instances, only disks that use NTFS file systems can be expanded.
Maximum capacity	• Ultra disk and standard SSD: 32,768 GiB • SSD Shared Block Storage device and Ultra Shared Block Storage device: 32,768 GiB
Related operations	• When you expand disks, only the capacity of the disks is expanded. The sizes of partitions and file systems do not change. You must manually re-allocate the storage space on a disk after the disk is expanded. • You cannot shrink an expanded disk by any means, such as by rolling it back.

Procedure

1.
2.
3.
4. Find the disk that you want to expand and choose **More > Expand Disk** in the **Actions** column.
5. In the Expand Disk dialog box, specify a new capacity for the disk.

   The new capacity must be greater than the current capacity.
6. Click **OK**.

Result

When you expand disks, only the capacity of the disks is expanded. The sizes of partitions and file systems do not change. You must manually re-allocate the storage space on a disk after the disk is expanded.

2.1.5.9. Encrypt a disk

2.1.5.9.1. Encrypt a system disk

In the scenarios that require data security and regulatory compliance, you can encrypt disks to secure your data stored on the disks. To encrypt system disks, you can encrypt custom images and then use the encrypted custom images to create instances. The system disks of the created instances are automatically encrypted.

Context

You can encrypt system disks only by encrypting custom images.

Step 1: Create a custom image from an instance

1.

2.
3.
4. Find the snapshot from which you want to create a custom image and click **Create Custom Image** in the **Actions** column.
5. Configure the parameters listed in the following table to create a custom image.

Parameter	Description
Custom Image Name	Enter a name for the custom image. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), and colons (:). It must start with a letter.
Sharing Scope	Select the scope in which to share the custom image. ○ **Current Organization and Subordinate Organizations** ○ **Current Resource Set** ○ **Current Organization**
Description	Enter a description for the custom image. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

6. Click **OK**.

Step 2: Encrypt the custom image

1.
2.
3.
4. Click the **Custom Images** tab.
5. Find the custom image that you want to encrypt and click **Encrypt Image** in the **Actions** column.
6. In the **Encrypt Image** dialog box, configure the parameters listed in the following table.

Parameter	Description
Image ID	The system automatically obtains the ID of the image to be encrypted. You do not need to configure this parameter.
Custom Image Name	Enter a name for the new encrypted custom image. The name must be 2 to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-). It must start with a letter.
Description	Enter a description for the new encrypted custom image. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

7. Click **OK**.

Step 3: Use the encrypted custom image to create an instance

1.
2.
3.
4. Click **Create Instance**.
5. Configure the parameters for the instance.

   For more information about how to configure these parameters, see Create an instance

   In the **Image** section, set Image Type to **Custom Image**. Then, select the image that you encrypted from the **Custom Image** drop-down list.
6. Click **Submit**.

Result

After the instance is created, you can click its ID to go to the **Instance Details** page. Then, you can click the **Disks** tab and check the value in the **Disk Encryption** column corresponding to the system disk. If the value is **Yes**, the system disk is encrypted.

2.1.5.9.2. Encrypt a data disk

In the scenarios that require data security and regulatory compliance, you can encrypt disks to secure your data stored on the disks. After a data disk is created, you cannot change its encryption state. If you want to encrypt a data disk, enable encryption for the disk when you create it.

Context

We recommend that you determine the number and sizes of data disks before you create them. The following limits apply to data disks:

- A maximum of 16 data disks can be attached to an instance. Disks and Shared Block Storage devices share this quota.
- Each Shared Block Storage device can be attached to up to four ECS instances at the same time.
- Each ultra disk, standard SSD, Ultra Shared Block Storage device, or SSD Shared Block Storage device can have a maximum capacity of 32 TiB.
- Disks cannot be combined in ECS. They are independent of each other. You cannot combine multiple disks into one by formatting them.

We recommend that you do not use Logical Volume Manager (LVM) to create logical volumes across multiple disks, because a snapshot can back up data only of a single disk. If you create a logical volume across several disks, data discrepancies may occur when you restore these disks.

Procedure

1.
2.
3.
4. Click **Create Disk**.
5. On the **Create Disk** page, configure the parameters for the disk.

   When you encrypt the disk, take note of the following parameters:

   - **Encrypted**: Select **Yes**.
   - **Encryption Method**: Select an encryption algorithm.
     - **AES256**: indicates the AES256 encryption algorithm.
   - **Encryption Key**: Select an encryption key.

   For information about how to configure the other parameters to create a disk, see Create a disk.

6. Click **Submit**.

2.1.5.10. Re-initialize a disk

You can re-initialize disks to restore them to their initial states.

Prerequisites

- The disk to be re-initialized is in the **Running** state.
- The instance to which the disk is attached is in the **Stopped** state.
- After a disk is re-initialized, the data stored on the disk is lost and cannot be recovered. Exercise caution when you perform this operation. We recommend that you back up disk data or create snapshots before you re-initialize a disk. For more information, see Create a snapshot.

Context

The result of disk re-initialization depends on the disk type and how the disk was created.

- System disk:
  - The disk is restored to the initial state of the image from which the disk was created.
  - If the corresponding image has been deleted, the disk cannot be re-initialized.
- Data disk:
  - If the disk is empty when created, the disk is restored to an empty disk.
  - If the disk was created from a snapshot, the disk is restored to the state of the snapshot.
  - If the snapshot from which a disk was created has been deleted, the disk cannot be re-initialized.

Procedure

1.
2.
3.
4. Find the data disk that you want to re-initialize and click **Re-initialize** in the **Actions** column.
5. In the Re-initialize Disk dialog box, perform operations based on the disk type.
   - For a system disk, enter and confirm a new instance logon password. Select or clear Instance After Re-initializing Disk, and click **OK**.

     The password must be 8 to 30 characters in length, and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. The supported special characters include `( ) ' ~ ! @#$%^&*-_+=|{}[]:;'<>,.? /`
   - For a data disk, click **OK**.

Result

When the disk is being re-initialized, the disk is in the **Initializing** state. After the disk is re-initialized, it enters the **Running** state.

2.1.5.11. Detach a data disk

You can detach data disks but cannot detach system disks.

Prerequisites

- Before you can detach a data disk from a Windows instance, you must bring the data disk offline by using Disk Management.

> **Note** To avoid data loss and ensure data integrity, we recommend that you stop read and write operations on the disk before you detach it.

- Before you can detach a data disk from a Linux instance, you must connect to the instance and unmount all partitions on the disk.

  > **Note** If you have added an entry in the */etc/fstab* file for the disk partitions to be automatically mounted on instance startup, you must delete the entry from the */etc/fstab* file before you detach the disk. Otherwise, you are unable to connect to the instance after the instance is restarted.

- The data disk that you want to detach is in the **Running** state.

Procedure

1.
2.
3.
4. Find the data disk that you want to detach and choose **More** > **Detach** in the **Actions** column.
5. Click **OK**.

2.1.5.12. Release a data disk

You can release data disks that are no longer needed. Released disks cannot be recovered. Exercise caution when you release data disks.

Prerequisites

The data disk to be released is in the **Pending** state. If the data disk is attached to an instance, you must detach the disk from the instance before you can release the disk.

Procedure

1.
2.
3.
4. Find the disk that you want to release and choose **More** > **Release** from the **Actions** column.
5. Click **OK**.

2.1.6. Images

2.1.6.1. Create a custom image

You can create a custom image and use it to create identical instances or replace the system disks of existing instances. This way, you can quickly obtain multiple instances with the same operating system and data environment.

Create a custom image from a snapshot

You can create a custom image from a system disk snapshot to load the operating system and data environment of the snapshot to the image. Before you perform this operation, make sure that a system disk snapshot is used. You cannot create custom images from data disk snapshots.

1.
2.

3.
4. Find the snapshot from which you want to create a custom image and click **Create Custom Image** in the **Actions** column.
5. Configure the parameters listed in the following table to create a custom image.

Parameter	Description
Custom Image Name	Enter a name for the custom image. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), and colons (:). It must start with a letter.
Sharing Scope	Select the scope in which to share the custom image. ○ **Current Organization and Subordinate Organizations** ○ **Current Resource Set** ○ **Current Organization**
Description	Enter a description for the custom image. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

6. Click **OK**.

Create a custom image from an instance

You can create a custom image from an instance to replicate the data of all disks of the instance, including the system disk and data disks.

> **Note** To avoid data security risks, we recommend that you delete sensitive data from an instance before you use the instance to create a custom image.

When you create a custom image from an instance, a snapshot is automatically generated for each disk on the instance, and all the snapshots constitute a complete custom image.

1.
2.
3.
4. Find the instance from which you want to create a custom image and choose **More > Create Custom Image** in the **Actions** column.
5. Configure the parameters listed in the following table to create a custom image.

Parameter	Description
Custom Image Name	Enter a name for the custom image. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), and colons (:). It must start with a letter.

Parameter	Description
Sharing Scope	Select the scope in which to share the custom image. ○ **Current Organization and Subordinate Organizations** ○ **Current Resource Set** ○ **Current Organization**
Description	Enter a description for the custom image. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

6. Click **OK**.

2.1.6.2. View images

You can view a list of images.

Procedure

1.
2.
3.
4. Select a tab based on the type of images that you want to view.

   You can select the **Custom Images** or **Public Images** tab.
5. Select a filter option from the drop-down list, enter the relevant information in the search bar, and click **Search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
Image Name	Enter an image name to search for the image.
Image ID	Enter an image ID to search for the image.
Snapshot ID	Enter a snapshot ID to search for images associated with the snapshot. This option is not available for public images.

2.1.6.3. View instances related to an image

You can view the instances that use a specified image.

Procedure

1.
2.
3.
4. Select a tab based on the type of the image that you want to view.

   You can select the **Custom Images** or **Public Images** tab.
5. Find the image and click **Related Instances** in the **Actions** column.

Result

The Instances page appears and shows the instances that use the image. You can perform operations on these instances, such as updating the image.

2.1.6.4. Modify the description of a custom image

You can modify the descriptions of created custom images.

Procedure

1.
2.
3.
4. Find the custom image that you want to modify and click **Modify Description** in the **Actions** column.
5. In the Modify Description dialog box, modify the image description in the Basic Settings field.

   The description must be 2 to 256 characters in length and cannot start with http:// or https://.
6. Click **OK**.

2.1.6.5. Share a custom image

You can share a created custom image to the organizations that you manage. Then, the organizations can use the shared image to create multiple instances that have identical environments.

Context

Only custom images can be shared. Shared images do not count against the image quotas of the organizations to which the images are shared.

The organizations to which images are shared can use the shared images to create instances or replace the system disks of existing instances.

You can delete shared images. After a shared image is deleted, the image is no longer visible to the organizations to which the image was shared and the system disks of the instances created from this image can no longer be re-initialized.

Procedure

1.
2.
3.
4. Find the image that you want to share and click **Share Image** in the **Actions** column.
5. In the Share Image dialog box, select a department from the **Department** drop-down list.

   > **Note** If no department list is available due to lack of permissions, you can enter the name of a level-1 organization in the Department field to share the image.
6. Click **OK**.

2.1.6.6. Encrypt a custom image

This topic describes how to encrypt a custom image to generate a new identical encrypted custom image.

Prerequisites

The custom image that you want to encrypt is in the Available (Available) state.

Context

To meet the requirements for data security compliance, you can use encrypted custom images to create instances. The system disks of the created instances are automatically encrypted.

Procedure

1.
2.
3.
4. Click the **Custom Images** tab.
5. Find the custom image that you want to encrypt and click **Encrypt Image** in the **Actions** column.
6. In the **Encrypt Image** dialog box, configure the parameters listed in the following table.

Parameter	Description
Image ID	The system automatically obtains the ID of the image to be encrypted. You do not need to configure this parameter.
Custom Image Name	Enter a name for the new encrypted custom image. The name must be 2 to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-). It must start with a letter.
Description	Enter a description for the new encrypted custom image. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

7. Click **OK**.

Result

After you encrypt the custom image, a new identical encrypted custom image is generated and displayed on the Custom Images tab.

2.1.6.7. Import custom images

2.1.6.7.1. Limits on importing images

This topic describes the limits on importing images. You must understand the limits to ensure that the imported images are available and make the import operation more efficient.

When you import images, take note of the limits described in the following sections:

- Linux images
- Windows images

Linux images

When you import Linux images, take note of the following limits:

- Multiple network interfaces are not supported.
- IPv6 addresses are not supported.
- Each password must be 8 to 30 characters in length. It must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.
- Firewalls must be disabled. By default, port 22 is enabled.

- Each Linux system disk must range from 40 GiB to 500 GiB in size.
- DHCP must be enabled in the images.
- SELinux must be disabled.
- The Kernel-based Virtual Machine (KVM) driver must be installed.
- We recommend that you install cloud-init to configure hostnames, NTP sources, and YUM sources.

Limits

Item	Standard operating system image	Non-standard operating system image
Definition	The supported standard 32-bit and 64-bit operating systems include: • CentOS • Ubuntu • SUSE • openSUSE • RedHat • Debian • CoreOS • Aliyun Linux **Note** Support for standard operating systems may be subject to version changes. You can log on to the ECS console to view the latest supported operating systems.	Non-standard operating systems include: • Operating systems that are not supported by Alibaba Cloud • Standard operating systems that do not meet the requirements on critical system configuration files, basic system environments, or applications To use non-standard operating system images, select Others Linux when you import images. If the images that you import are non-standard operating system images, Alibaba Cloud does not process the instances created from these images. After you create an instance from a non-standard operating system image, you must connect to the instance by using the VNC feature in the ECS console, and then configure the IP address, route, and password of the instance.
Critical system configuration file	• Do not modify `/etc/issue*`. Otherwise, the version of the operating system cannot be identified, which leads to a failure to create the system. • Do not modify `/boot/grub/menu.lst`. Otherwise, the system fails to start. • Do not modify `/etc/fstab`. Otherwise, partitions cannot be loaded, which causes the system to fail to start. • Do not change `/etc/shadow` to read-only. Otherwise, the password file cannot be modified, which leads to a failure to create the system. • Do not modify `/etc/selinux/config` to enable SELinux. Otherwise, the system fails to start.	

Item	Standard operating system image	Non-standard operating system image
Basic system environment	• Do not adjust the system disk partitions. Only disks with a single root partition are supported. • Make sure that the system disk has sufficient free space. • Do not modify critical system files such as `/sbin`, `/bin`, and `/lib*`. • Before you import an image, confirm the integrity of the file system. • Only ext3 and ext4 file systems are supported.	Requirements for standard operating system images are not met.
Application	Do not install `qemu-ga` on a custom image. Otherwise, some of the services that Alibaba Cloud uses may be unavailable.	
Image file format	Only images in the RAW, VHD, or QCOW2 format can be imported. To import images in other formats, use a tool to convert the images to a supported format. We recommend that you import images in the VHD format, which has a smaller transmission footprint.	
Image file size	We recommend that you configure the system disk size based on the virtual disk size rather than the image size. The configured system disk size must be at least 40 GiB.	

Windows images

When you import Windows images, take note of the following limits:

- Each password must be 8 to 30 characters in length. It must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.
- Imported Windows images do not provide the Windows Activation Service.
- Firewalls must be disabled. Otherwise, remote logon is not supported. Port 3389 must be enabled.
- Each Windows system disk must range from 40 GiB to 500 GiB in size.

Limits

Item	Description

Item	Description

Operating system version	Alibaba Cloud allows you to import the following 32-bit and 64-bit versions of Windows operating system images:
	• Microsoft Windows Server 2016
	• Microsoft Windows Server 2012, including:
	  ◦ Microsoft Windows Server 2012 R2 (Standard Edition)
	  ◦ Microsoft Windows Server 2012 (Standard Edition and Datacenter Edition)
	• Microsoft Windows Server 2008, including:
	  ◦ Microsoft Windows Server 2008 R2 (Standard Edition, Datacenter Edition, and Enterprise Edition)
	  ◦ Microsoft Windows Server 2008 (Standard Edition, Datacenter Edition, and Enterprise Edition)
	• Microsoft Windows Server 2003, including:
	  ◦ Microsoft Windows Server 2003 R2 (Standard Edition, Datacenter Edition, and Enterprise Edition)
	  ◦ Microsoft Windows Server 2003 (Standard Edition, Datacenter Edition, and Enterprise Edition) or later, including Service Pack 1 (SP1)
	• Microsoft Windows 7, including:
	  ◦ Microsoft Windows 7 (Professional Edition)
	  ◦ Microsoft Windows 7 (Enterprise Edition)
	Note Support for standard operating systems may be subject to version changes. You can log on to the ECS console to view the latest supported operating systems.

Basic system environment	• Multi-partition system disks are supported. • Make sure that the system disk has sufficient free space. • Do not modify critical system files. • Before you import an image, confirm the integrity of the file system. • Disks can be partitioned in the MBR format and formatted to NTFS file systems.
Application	Do not install qemu-ga on an imported image. Otherwise, some of the services that Alibaba Cloud uses may be unavailable.
Supported image file format	• RAW • VHD • QCOW2 We recommend that you configure the system disk size based on the virtual disk size rather than the image size. The configured system disk size must range from 40 GiB to 500 GiB. **Note** We recommend that you import images in the VHD format, which has a smaller transmission footprint.

2.1.6.7.2. Convert the image file format

You can only import image files in the RAW, VHD, and qcow2 formats to ECS. If you want to import images in other formats, you must convert the image into a supported format. This topic describes how to convert the image format in Windows and Linux.

Context

You can use the qemu-img tool to convert an image from VMDK, VDI, VHDX, qcow1, or QED to RAW, VHD, or qcow2, or implement conversion between RAW, VHD, and qcow2.

> **Note** We recommend that you use the qcow2 format if your application environment supports this format.

Windows

1. Download qemu.

   Visit QEMU Binaries for Windows (64 bit) to download the qemu tool. Select a qemu version based on your operating system.

2. Install qemu.

   The installation path in this example is *C:\Program Files\qemu*.

3. Configure the environment variables for qemu.

   i. Choose **Start > Computer**, right-click Computer, and choose **Properties** from the shortcut menu.

   ii. In the left-side navigation pane, click **Advanced System Settings**.

   iii. In the **System Properties** dialog box that appears, click the **Advanced** tab and then click **Environment Variables**.

   iv. In the **Environment Variables** dialog box that appears, find the **Path** variable from the **System variables** section.

   - If the **Path** variable exists, click **Edit**.
   - If the **Path** variable does not exist, click **New**.

   v. Add a system variable value.

   - In the Edit System Variable dialog box that appears, add *C:\Program Files\qemu* to the **Variable value** field, separate different variable values with semicolons (;), and then click **OK**.
   - In the New System Variable dialog box that appears, enter *Path* in the **Variable name** field, enter *C:\Program Files\qemu* in the **Variable value** field, and then click **OK**.

4. Open Command Prompt in Windows and run the `qemu-img --help` command. If a successful response is displayed, the tool is installed.

5. In the Command Prompt window, run the `cd [Directory of the source image file]` command to switch to a new file directory,

   for example, `cd D:\ConvertImage`.

6. In the Command Prompt window, run the `qemu-img convert -f raw -O qcow2 centos.raw centos.qcow2` command to convert the image file format.

   The parameters are described as follows:

   - The `-f` parameter is followed by the source image format.
   - The `-O` parameter (case-sensitive) is followed by the destination image format, source file name, and destination file name.

   After the conversion is complete, the destination file appears in the directory of the source image file.

Linux

1. Install the qemu-img tool.
   - For Ubuntu, run the `apt install qemu-img` command.
   - For CentOS, run the `yum install qemu-img` command.
2. Run the `qemu-img convert -f raw -O qcow2 centos.raw centos.qcow2` command to convert the image file format.

   The parameters are described as follows:
   - The `-f` parameter is followed by the source image format.
   - The `-O` parameter (case-sensitive) is followed by the destination image format, source file name, and destination file name.

2.1.6.7.3. Import an image

After you upload a local image to an Object Storage Service (OSS) bucket, you can import the image as a custom image to Elastic Compute Service (ECS).

Prerequisites

- An image is made. It meets the limits and requirements for image import and is in the RAW, VHD, or QCOW2 format. For more information, see Limits on importing custom images and Convert the image file format.
- You are authorized to import images. For more information, see the "RAM" chapter in *Apsara Uni-manager Management Console User Guide*.
- A local image is uploaded to a bucket by using the OSS console or by calling an OSS API operation. For more information, see the "Upload objects" topic in *OSS User Guide* or the "PutObject" topic in *OSS Developer Guide*.

> **Note** Make sure that the bucket resides in the region to which you want to import the image as a custom image.

Procedure

1.
2.
3.
4. Click **Import Image**.
5. Configure the parameters described in the following table.

Parameter	Required	Description
Region	Yes	The region to which to import the image as a custom image.
Organization	Yes	The organization in which to use the custom image.
Resource Set	Yes	The resource set to which to assign the custom image.
OSS Bucket Name	Yes	The name of the OSS bucket where the image to be imported is stored.
OSS Object Name	Yes	The URL of the object as which the image to be imported is stored in the OSS bucket. For information about how to obtain the URL of an OSS object, see the "Obtain object URLs" topic in *OSS User Guide*.

Parameter	Required	Description
Image Name	Yes	The name of the custom image. The name must be 2 to 128 characters in length. It must start with a letter and can contain letters, periods (.), underscores (_), and hyphens (-).
Sharing Scope	Yes	The scope in which to share the custom image. ○ **Current Organization and Subordinate Organizations** ○ **Current Resource Set** ○ **Current Organization**
Operating system	Yes	Valid values: **Linux** and **Windows**.
System Disk (GiB)	Yes	The size of the system disk on an instance. Unit: GiB.
System Architecture	Yes	Valid values: **x86_64** and **i386**.
Platform	Yes	Linux: ○ **CentOS** ○ **Ubuntu** ○ **SUSE** ○ **OpenSUSE** ○ **Debian** ○ **CoreOS** ○ **Aliyun** ○ **Others Linux** ○ **Customized Linux** Windows: ○ **Windows Server 2003** ○ **Windows Server 2008** ○ **Windows Server 2012**
Description	No	The description of the custom image.
Add Data Disk Image	No	Imports another image that contains data from data disks. If you select Add Data Disk Image, you must specify parameters including OSS Object Address, Image Format, Device Name, and Disk Capacity.

6. Click **OK**.

Result

You can go to the Images page to view the creation progress of the custom image. For more information, see View images. When the custom image is created, 100% is displayed in the Progress column.

2.1.6.8. Export a custom image

You can export custom images to OSS buckets and then download the images to your local device.

Prerequisites

- OSS is activated and an OSS bucket is created. For more information, see the "Create buckets" topic in *OSS User Guide*.
- You are authorized to export images. For more information, see the "RAM" chapter in *Apsara Uni-manager Management Console User Guide*.

Context

You can export custom images to the RAW, VHD, or QCOW2 format. After a custom image is exported to an OSS bucket, you can download the image to your local device. For more information, see the "Obtain object URLs" topic in *OSS User Guide*.

Procedure

1.
2.
3.
4. Find the custom image that you want to export and click **Export Image** in the **Actions** column.
5. Configure the parameters listed in the following table.

Parameter	Required	Description
OssBucket	Yes	The name of the OSS bucket where to store the exported image.
Image Type	No	The format in which to export the image. Valid values: **RAW**, **VHD**, and **QCOW2**.
OSS Prefix	No	The prefix of the OSS object to which to export the image. must be 1 to 30 characters in length and can contain digits and letters.

6. Click **OK**.

2.1.6.9. Delete a custom image

You can delete custom images that are no longer needed. Custom images can be deleted but public images cannot.

Procedure

1.
2.
3.
4. Use one of the following methods to delete custom images:
   - To delete a single custom image, find the image and click **Delete Image** in the **Actions** column.
   - To delete one or more custom images at a time, select the images and click **Delete** in the lower-left corner of the image list.
5. Click **OK**.

2.1.7. Snapshots

2.1.7.1. Create a snapshot

You can manually create snapshots for disks to back up disk data.

Prerequisites

- The associated instance of the disk for which you want to create a snapshot is in the **Running** or **Stopped** state.
- The disk is in the **Running** state.

Context

Up to 64 snapshots can be retained for each disk.

Snapshots can be used in the following scenarios:

- Roll back data on disks

  For more information, see Restore a disk.

- Create a custom image

  For more information, see Create a custom image from a snapshot. Data disk snapshots cannot be used to create custom images.

- Create a data disk from a data disk snapshot

  To create a data disk from a snapshot, set Use Snapshot to Yes and then specify a snapshot on the Create Disk page. For more information, see Create a disk. The created disk size is determined by the size of the specified snapshot and cannot be changed. When you re-initialize a data disk created from a snapshot, the disk is restored to the state of the snapshot.

When you create a snapshot, take note of the following items:

- For each disk, the first snapshot is a full snapshot and subsequent snapshots are incremental snapshots. It takes longer to create the first snapshot. The amount of time it takes to create an incremental snapshot depends on the volume of data that has been changed since the last snapshot. The more data that has changed, the longer it takes.

  > **Note** If you want to use a snapshot immediately after it is created, you can enable the instant access feature.

- We recommend that you create snapshots during off-peak hours.

Procedure

1.
2.
3.
4. On the Disks page, find the disk for which you want to create a snapshot and click **Create Snapshot** in the **Actions** column.
5. Enter a snapshot name and description and then click **OK**.

Parameter	Description
Snapshot Name	The name of the snapshot. **Note** The name cannot start with auto because snapshots whose names start with auto are recognized as automatic snapshots.

Parameter	Description
Instant Access	Specifies whether to enable the instant access feature. After you enable the instance access feature for a snapshot, you can immediately use the snapshot to perform operations such as rolling back the disk.
Description	The description of the snapshot.

You can go to the Snapshots page to check the creation progress of the snapshot. For more information, see View snapshots. When the snapshot is created, 100% is displayed in the Progress column.

2.1.7.2. View snapshots

You can view the list of snapshots.

Procedure

1.
2.
3.
4. Select a filter option from the drop-down list, enter the relevant information in the search bar, and click **Search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
Snapshot Name	Enter a snapshot name to search for the snapshot.
Snapshot ID	Enter a snapshot ID to search for the snapshot.
Instance ID	Enter an instance ID to search for snapshots related to the instance.
Disk ID	Enter a disk ID to search for snapshots related to the disk.
Snapshot Type	Select a snapshot type to search for snapshots of that type. Valid values: ◦ **All** ◦ **User Snapshots**: manual snapshots ◦ **Automatic Snapshots**: automatic snapshots
Creation Time	Enter a time to search for snapshots that were created at that time.

2.1.7.3. Delete a snapshot

You can delete snapshots that are no longer needed. Deleted snapshots cannot be recovered. You cannot delete system disk snapshots that have been used to create custom images.

Procedure

1.
2.
3.
4. Use one of the following methods to delete snapshots:

- To delete a single snapshot, find the snapshot and click **Delete** in the **Actions** column.
- To delete one or more snapshots at a time, select the snapshots and click **Delete** in the lower-left corner of the Snapshots page.

5. Click **OK**.

2.1.8. Automatic snapshot policies

2.1.8.1. Create an automatic snapshot policy

Automatic snapshot policies can be applied to system disks and data disks to create periodical snapshots of the disks. You can use automatic snapshot policies to improve data security and tolerance against operation faults.

Context

Automatic snapshot policies can effectively eliminate the following risks associated with manual snapshots:

- When applications such as personal websites or databases deployed on an ECS instance encounter attacks or system vulnerabilities, you may be unable to manually create snapshots. In this case, you can use the latest automatic snapshots to roll back the affected disks to restore your data and reduce loss.
- You can also specify an automatic snapshot policy to create snapshots before regular system maintenance tasks are performed. This eliminates the need to manually create snapshots and ensures that snapshots are always created before maintenance.

You can retain up to 64 snapshots for each disk. If the maximum number of snapshots for a disk is reached while a new snapshot is being created, the system deletes the oldest automatic snapshot.

Procedure

1.
2.
3.
4. Click **Create Policy**.
5. Configure the parameters listed in the following table to create an automatic snapshot policy.

Parameter	Required	Description
Region	Yes	The ID of the region in which to apply the automatic snapshot policy.
Organization	Yes	The organization in which to apply the automatic snapshot policy.
Resource Set	Yes	The resource set in which to apply the automatic snapshot policy.
Policy Name	Yes	The name of the automatic snapshot policy. The name must be 2 to 128 characters in length and cannot start with a special character or digit. It can contain periods (.), underscores (_), hyphens (-), and colons (:).

Parameter	Required	Description
Creation Time	Yes	The time of the day at which to create an automatic snapshot. Valid values: 00:00 to 23:00 (the start of each hour). You can select multiple values. **Note** The default time zone for the snapshot policy is UTC+8. You can change the time zone to suit your business requirements. If the time scheduled for creating an automatic snapshot is due while a previous automatic snapshot is being created, the new snapshot creation task is skipped. This may occur when a disk contains a large volume of data. For example, assume that an automatic snapshot policy is applied to a disk that contains a large volume of data, and the policy specifies to create snapshots at 00:00, 01:00, and 02:00. If the system starts to create a snapshot at 00:00 and takes 70 minutes to complete the snapshot creation task, the system skips the automatic snapshot task scheduled for 01:00 and creates the next automatic snapshot at 02:00.
Frequency	Yes	The day of the week when to create automatic snapshots. The valid values range from Monday to Sunday. You can select multiple values.
Retention Policy	No	The retention period of the automatic snapshots. The default value is Keep for 30 Days. You can select one of the following options: ◦ **Keep for**: You can select this option and then specify the number of days during which to retain the automatic snapshots. Valid values: 1 to 65536. ◦ **Always Keep the Snapshots Until the Number of Snapshots Reaches the Upper Limit**: You can select this option to retain the automatic snapshots until the maximum number of snapshots is reached.

6. Click **OK**.

What's next

After the automatic snapshot policy is created, you must apply it to a disk for snapshots to be automatically created. For more information, see Configure an automatic snapshot policy for multiple disks.

2.1.8.2. View automatic snapshot policies

You can view the list of automatic snapshot policies.

Procedure

1.
2.
3.
4. View the list of automatic snapshot policies.

2.1.8.3. Modify an automatic snapshot policy

You can modify the attributes of automatic snapshot policies. The attributes of each automatic snapshot policy that can be modified include the name, creation time, frequency, and retention policy.

Procedure

1.
2.
3.
4. Find the automatic snapshot policy that you want to modify and click **Modify Policy** in the **Actions** column.
5. Modify the attributes of the policy.

   Changes made to the retention policy do not affect existing snapshots but take effect only on subsequent snapshots.
6. Click **OK**.

2.1.8.4. Configure an automatic snapshot policy

After you apply an automatic snapshot policy to a disk, snapshots will be created automatically for the disk based on the policy settings. You can cancel an applied automatic snapshot policy at any time.

Context

We recommend that you configure the automatic snapshot policy to create automatic snapshots during off-peak hours. You can also manually create a snapshot for the disk that already has an automatic snapshot policy applied. When an automatic snapshot is being created, you must wait until the snapshot is complete before you can create a manual snapshot. The automatic snapshot is named in the following format: auto_yyyyMMdd_1, such as auto_20140418_1.

Procedure

1.
2.
3.
4. Find the disk and click **Configure Automatic Snapshot Policy** in the **Actions** column.
5. Select a procedure based on the operation you want to perform on the policy.
   - To apply an automatic snapshot policy, turn on **Automatic Snapshot Policy**, select a policy, and then click **OK**.
   - To cancel an automatic snapshot policy, turn off **Automatic Snapshot Policy** and click **OK**.

2.1.8.5. Configure an automatic snapshot policy for multiple disks

After you apply automatic snapshot policies to disks, snapshots are created automatically for the disks based on the policies. You can disable applied automatic snapshot policies at any time.

Context

We recommend that you configure automatic snapshot policies to create automatic snapshots during off-peak hours. You can manually create snapshots for disks that already have automatic snapshot policies applied. When an automatic snapshot is being created for a disk, you must wait for the snapshot to be complete before you can create a manual snapshot for the disk. Each automatic snapshot is named in the following format:

auto_yyyyMMdd_1. Example: auto_20140418_1.

Procedure

1.
2.
3.
4. Find the automatic snapshot policy that you want to configure and click **Apply Policy** in the **Actions** column.
5. Apply the policy to disks or disable the policy for disks.
   - To apply the automatic snapshot policy, select the **Disks Without Policy Applied** tab, select one or more disks, and then click **Apply Policy** in the lower part of the tab.
   - To disable the automatic snapshot policy, select the **Disks With Policy Applied** tab, select one or more disks, and then click **Disable Policy** in the lower part of the tab.

2.1.8.6. Delete an automatic snapshot policy

You can delete automatic snapshot policies that are no longer needed. After you delete an automatic snapshot policy, the policy is automatically canceled for the disks that have it applied.

Procedure

1.
2.
3.
4. Find the automatic snapshot policy that you want to delete and click **Delete Policy** in the **Actions** column.
5. Click **OK**.

2.1.9. Security groups

2.1.9.1. Create a security group

Security groups are an important means for network security isolation. They implement network access control for one or more ECS instances.

Prerequisites

A virtual compute cloud (VPC) is created. For more information, see *VPC User Guide*.

Context

Security groups determine whether the instances in the same account that are deployed within the same VPC and region can communicate with each other. By default, if the instances belong to the same security group, they can communicate with each other over the internal network. If the instances belong to different security groups, you can authorize mutual access between the security groups to allow the instances to communicate with each other over the internal network.

Procedure

1.
2.
3.
4. Click **Create Security Group**.
5. Configure the parameters listed in the following table to create a security group.

Type	Parameter	Required	Description
Region	Organization	Yes	Select an organization in which to create the security group. Make sure that the security group and the VPC belong to the same organization.
	Resource Set	Yes	Select a resource set in which to create the security group. Make sure that the security group and the VPC belong to the same resource set.
	Region	Yes	Select a region in which to create the security group. Make sure that the security group and VPC belong to the same region.
	Zone	Yes	Select a zone in which to create the security group.
Basic Settings	VPC	Yes	Select a VPC in which to create the security group.
	Security Group Name	No	Enter a name for the security group. The name must be 2 to 128 characters in length and start with a letter. It can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,). It cannot start with http:// or https://.
	Description	No	Enter a description for the security group. To simplify future management operations, we recommend that you provide an informational description. The description must be 2 to 256 characters in length and start with a letter. It can contain letters, digits, periods (.), underscores (_), hyphens (-), and commas (,). It cannot start with http:// or https://.

6. Click **Submit**.

2.1.9.2. View security groups

You can view the list of security groups.

Procedure

1.
2.
3.
4. Select a filter option from the drop-down list, enter the relevant information in the search bar, and click **Search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
Security Group ID	Enter a security group ID to search for the security group.
Security Group Name	Enter a security group name to search for the security group.
VPC ID	Enter a VPC ID to search for security groups that belong to the VPC.

2.1.9.3. Modify a security group

You can modify the names and descriptions of created security groups.

Procedure

1.
2.
3.
4. Find the security group that you want to modify and click **Modify** in the **Actions** column.
5. Modify the name and description of the security group.
6. Click **OK**.

2.1.9.4. Add a security group rule

You can use security group rules to control access to and from the ECS instances in a security group over the Internet and the internal network.

Procedure

1.
2.
3.
4. Find the security group to which you want to add a rule and click **Rules** in the **Actions** column.
5. Click **Create Rule**.
6. Configure the parameters listed in the following table to create a security group rule.

Parameter	Required	Description
ENI Type	Yes	Valid value: **Internal Network ENI**. In VPCs, you cannot find public NICs in ECS instances and can add only internal security group rules. However, the added security group rules apply to both the Internet and the internal network.
Direction	Yes	◦ **Outbound**: access from the ECS instances in the current security group to other ECS instances on the internal network or to resources on the Internet. ◦ **Inbound**: access from other ECS instances on the internal network or from resources on the Internet to the ECS instances in the current security group.
Action	Yes	◦ **Allow**: allows access requests on specified ports. ◦ **Deny**: discards requests received on specified ports without returning messages. If two security group rules are different only in the Action parameter, the **Deny** rule takes effect whereas the **Allow** rule is ignored.

Parameter	Required	Description
Protocol	Yes	◦ **All**: This value can be used in total trust scenarios. ◦ **TCP**: This value can be used to allow or deny traffic on one or several successive ports. ◦ **UDP**: This can be used to allow or deny traffic on one or several successive ports. ◦ **ICMP**: This value can be used when the `ping` command is used to test the status of the network connection between instances. ◦ **ICMPv6**: This value can be used when the `ping6` command is used to test the status of the network connection between instances. ◦ **GRE**: This value can be used for VPN.
Port Range	Yes	The port range depends on the protocol type. ◦ When you set Protocol to **ALL**, the value of **-1/-1** is displayed, which indicates all ports. You cannot specify a port range in this case. ◦ When you set Protocol to **TCP**, you can specify a port range in the <start port number>/<end port number> format. Valid port numbers: 1 to 65535. You can set the start port number and end port number to the same value to specify a single port. For example, use 22/22 to specify port 22. ◦ When you set Protocol to **UDP**, you can specify a port range in the <start port number>/<end port number> format. Valid port numbers: 1 to 65535. You can set the start port number and end port number to the same value to specify a single port. For example, use 3389/3389 to specify port 3389. ◦ When you set Protocol to **ICMP**, the value of **-1/-1** is displayed, which indicates all ports. You cannot specify a port range in this case. ◦ When you set Protocol to **ICMPv6**, the value of **-1/-1** is displayed, which indicates all ports. You cannot specify a port range in this case. ◦ When you set Protocol to **GRE**, the value of **-1/-1** is displayed, which indicates all ports. You cannot specify a port range in this case.
Priority	Yes	The priority of the rule. Valid values: 1 to 100. The default value is 1, which indicates the highest priority.
Authorization Type	Yes	◦ **IPv4 Addresses**: IPv4 addresses or CIDR blocks. ◦ **IPv6 Addresses**: IPv6 addresses or CIDR blocks. ◦ **Security Groups**: security groups. This authorization type takes effect only on the internal network. You can select another security group in the current account as the authorization object for the instances in the current security group. This way, you can control the access to or from the ECS instances in that security group over the internal network.

Parameter	Required	Description
Authorization object	Yes	Authorization objects depend on the authorization type. When you set Authorization Type to **IPv4 Addresses**: ◦ Enter single IPv4 addresses or CIDR blocks. Example: *192.0.2.1* or *192.0.2.0/24*. ◦ You can enter up to 10 authorization objects at a time. Separate multiple objects with commas (,). ◦ If you enter *0.0.0.0/0*, all IPv4 addresses are allowed or denied based on the Action parameter. Exercise caution when you specify 0.0.0.0/0. When you set Authorization Type to **IPv6 Addresses**: ◦ Enter single IPv6 addresses or CIDR blocks. Example: *2001:db8:1:1:1:1:1:1* or *2001:db8::/32*. ◦ You can enter up to 10 authorization objects at a time. Separate multiple objects with commas (,). ◦ If you enter *::/0*, all IPv6 addresses are allowed or denied based on the Action parameter. Exercise caution when you specify ::/0. When you set Authorization Type to **Security Groups**, select a security group ID. If the current security group is of the VPC type, the selected security group must be in the same VPC as the current security group.
Description	No	The description of the security group rule. To simplify future management operations, we recommend that you provide an informational description. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

7. Click **OK**.

2.1.9.5. Clone a security group rule

You can clone a security group rule to quickly create a similar rule.

Procedure

1.
2.
3.
4. Find the target security group and click **Rules** in the **Actions** column.
5. On the Rules page that appears, click the **Inbound** or **Outbound** tab.
6. Find the target security group rule and click **Clone** in the **Actions** column.
7. In the **Clone Security Group Rule** dialog box, modify the attributes of the security group rule.

   For more information about the attributes of security group rules, see Add a security group rule.
8. Click **OK**.

2.1.9.6. Modify a security group rule

You can modify improper rules in a security group to ensure the security of ECS instances in the security group.

Procedure

1.

2.
3.
4. Find the target security group and click **Rules** in the **Actions** column.
5. On the Rules page that appears, click the **Inbound** or **Outbound** tab.
6. Find the target security group rule and click **Modify** in the **Actions** column.
7. In the **Modify Security Group Rule** dialog box, modify the attributes of the security group rule.

   For more information about the attributes of security group rules, see Add a security group rule.
8. Click **OK**.

2.1.9.7. Export security group rules

You can export security group rules of a security group to a local device for backup.

Procedure

1.
2.
3.
4. Find the target security group and click **Rules** in the **Actions** column.
5. On the Rules page that appears, click the **Inbound** or **Outbound** tab.
6. Click **Export** in the upper-right corner to download and save the rules to a local device.

2.1.9.8. Import security group rules

You can import a local backup file of security group rules into a security group to create or restore security group rules.

Context

You can first download a template file (Excel file), configure security group rules in the template file based on the template requirements, and then import the file.

Procedure

1.
2.
3.
4. On the Security Groups page, find the security group into which you want to import rules and click **Rules** in the **Actions** column.
5. Click **Import** in the upper-right corner.
6. In the **Import Rule** dialog box, click **Upload File**.
7. Select a local backup file of security group rules and click **Open**.

2.1.9.9. Add an instance to a security group

You can add an existing instance to a security group in the same region. After the instance is added, the rules of the security group automatically apply to the instance.

Procedure

1.

2.
3.
4. Find the security group to which you want to add an instance and click **Manage Instances** in the **Actions** column.
5. Click **Add Instance**.
6. Select an instance and click **OK**.

2.1.9.10. Remove instances from a security group

You can remove instances from security groups, but each of the instances must always belong to at least one security group.

Prerequisites

The instances to be removed belong to two or more security groups.

Context

After an instance is removed from a security group, the instance is isolated from the other instances in the security group. We recommend that you perform all tests in advance to ensure that services can continue to run properly after you remove the instance from the security group.

Procedure

1.
2.
3.
4. Find the security group from which you want to remove instances and click **Manage Instances** in the **Actions** column.
5. On the Instances page, select one or more instances and click **Remove** in the lower-left corner.
6. Click **OK**.

2.1.9.11. Delete a security group

You can delete security groups that are no longer needed.

Prerequisites

No instances exist in the security group that you want to delete.

Procedure

1.
2.
3.
4. Use one of the following methods to delete security groups:
   - To delete a single security group, find the security group and click **Delete** in the **Actions** column.
   - To delete one or more security groups at a time, select the security groups and click **Delete** in the lower-left corner of the Security Groups page.
5. Click **OK**.

2.1.10. Elastic Network Interfaces

2.1.10.1. Create an ENI

You can bind elastic network interfaces (ENIs) to instances to create high-availability clusters and implement fine-grained network management. You can also unbind an ENI from an instance and then bind the ENI to another instance to implement a low-cost failover solution.

Prerequisites

- A virtual private cloud (VPC) and a VSwitch are created. For more information, see Create a VPC and Create a VSwitch in *Apsara Stack VPC User Guide* .
- A security group is available in the VPC. If no security group is available in the VPC, create a security group. For more information, see Create a security group.

Context

ENIs are classified into primary and secondary ENIs.

A primary ENI is created by default when an instance is created in a VPC. This primary ENI has the same lifecycle as the instance and cannot be unbound from the instance.

ENIs created separately are secondary ENIs. You can bind secondary ENIs to or unbind them from instances. This topic describes how to create a secondary ENI.

Procedure

1.
2.
3.
4. Click **Create ENI**.
5. Configure parameters listed in the following table to create an ENI.

Section	Parameter	Required	Description
Region	Organization	Yes	The organization in which to create the ENI.
	Resource Set	Yes	The resource set in which to create the ENI.
	Region	Yes	The region in which to create the ENI.
	Zone	Yes	The zone in which to create the ENI.

Section	Parameter	Required	Description
Basic Settings	VPC	Yes	The VPC in which to create the ENI. The secondary ENI can be bound only to an instance in the same VPC. **Note** After the ENI is created, you cannot change its VPC.
	VSwitch	Yes	The VSwitch to be associated with the ENI. The secondary ENI can be bound only to an instance in the same VPC. Select a VSwitch that is in the same zone as the instance to which the ENI will be bound. The VSwitch of the ENI can be different from that of the instance. **Note** After an ENI is created, you cannot change its VSwitch.
	Security Group	Yes	The security group in which to create the ENI within the specified VPC. The rules of the security group automatically apply to the ENI.
	ENI Name	Yes	The name of the ENI. The name must be 2 to 128 characters in length. It must start with a letter and cannot start with http:// or https://. It can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,).

Section	Parameter	Required	Description
	Description	No	The description of the ENI. We recommend that you provide an informational description to simplify future management operations. The description must be 2 to 256 characters in length. It must start with a letter and cannot start with http:// or https://. It can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,).
	Primary Private IP	No	The primary private IPv4 address of the ENI. The IPv4 address must be within the CIDR block of the specified VSwitch. If you do not specify a primary private IP address, the system automatically assigns a private IP address to the ENI.

6. Click **Submit**.

Result

The created ENI is displayed on the ENIs page and is in the **Available** state.

2.1.10.2. View ENIs

You can view the list of ENIs.

Procedure

1.
2.
3.
4. Select a filter option from the drop-down list, enter the relevant information in the search bar, and click **Search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
ENI Name	Enter an ENI name to search for the ENI.
ENI ID	Enter an ENI ID to search for the ENI.
VSwitch ID	Enter a vSwitch ID to search for ENIs that are associated with the vSwitch.

Filter option	Description
Security Group ID	Enter a security group ID to search for ENIs that belong to the security group.
Instance ID	Enter an instance ID to search for ENIs that are bound to the instance.

2.1.10.3. Modify a secondary ENI

You can modify the attributes of a secondary elastic network interface (ENI), including the name, security group, and description.

Prerequisites

The secondary ENI is in the **Available** state.

Procedure

1.
2.
3.
4. Find the secondary ENI and click **Modify** in the **Actions** column.
5. In the Modify ENI dialog box that appears, modify the name, security group, and description of the ENI.
6. Click **OK**.

2.1.10.4. Bind a secondary ENI to an instance

You can bind a secondary elastic network interface (ENI) to an instance. After the ENI is bound, the instance can process the traffic on the ENI.

Prerequisites

- The secondary ENI that you want to bind is in the **Available** state.
- The instance to which you want to bind the secondary ENI is in the **Running** or **Stopped** state.
- The instance and the secondary ENI belong to the same virtual private cloud (VPC).
- The vSwitch with which the secondary ENI is associated is located within the same zone as the vSwitch to which the instance is connected. An ENI can be bound only to an instance within the same zone. The vSwitches of the ENI and of the instance can be different but must be located within the same zone.

Context

The following limits apply when you bind an ENI to an instance:

- Only secondary ENIs can be manually bound. Primary ENIs share the same lifecycle as instances and cannot be manually bound.
- An ENI can be bound only to a single instance at the same time. Each instance can have one or more bound ENIs. The maximum number of ENIs that can be bound to an instance is determined based on the instance type.

Bind a secondary ENI on the instance details page

To bind multiple secondary ENIs to an instance, you can go to the details page of the instance.

1.
2.
3.

4. Find the instance to which you want to bind a secondary ENI and click the instance ID.
5. Click the **ENIs** tab.
6. Click **Bind NIC**.
7. In the Bind NIC dialog box, select an ENI from the **ENI** drop-down list.
8. Click **OK**.

Bind a secondary ENI on the ENIs page

To bind secondary ENIs to multiple instances, you can go to the ENIs page.

1.
2.
3.
4. Find the secondary ENI that you want to bind and click **Bind** in the **Actions** column.
5. In the Bind dialog box, select an instance and click **OK**.
   In the **Status/Creation Time** column, the state of the secondary ENI changes to **Bound**.

2.1.10.5. Unbind a secondary ENI from an instance

You can unbind a secondary elastic network interface (ENI) from an instance. After the secondary ENI is unbound from the instance, the instance no longer processes the traffic on the ENI.

Prerequisites

- The secondary ENI is in the **Bound** state.
- The instance is in the **Running** or **Stopped** state.

Context

Only secondary ENIs can be unbound. Primary ENIs share the same lifecycle as instances and cannot be unbound.

Procedure

1.
2.
3.
4. Find the secondary ENI and click **Unbind** in the **Actions** column.
5. Click **OK**.

Result

In the **Status/Creation Time** column, the status of the secondary ENI changes to **Available**.

2.1.10.6. Delete a secondary ENI

You can delete a secondary elastic network interface (ENI) that is no longer needed.

Prerequisites

The secondary ENI is in the **Available** state.

Context

You can delete only secondary ENIs. Primary ENIs share the same lifecycle as instances and cannot be deleted.

Procedure

1.
2.
3.
4. Find the secondary ENI and click **Delete** in the **Actions** column.
5. Click **OK**.

2.1.11. Deployment sets

2.1.11.1. Create a deployment set

You can use deployment sets to distribute or aggregate instances involved in your business. You can select hosts, racks, or network switches to improve service availability or network performance.

Procedure

1.
2.
3.
4. Click **Create Deployment Set**.
5. Configure the parameters described in the following table.

Section	Parameter	Required	Description
Region	Organization	Yes	The organization in which to create the deployment set.
	Resource Set	Yes	The resource set in which to create the deployment set.
	Region	Yes	The region in which to create the deployment set.
	Zone	Yes	The zone in which to create the deployment set.
	Deployment Domain	Yes	This parameter determines the valid values of Deployment Granularity.
	Deployment Granularity	Yes	The basic unit that can be scheduled when you deploy instances. Valid values: ○ **Host**: Instances are distributed or aggregated at the host level. ○ **Rack**:Instances are distributed or aggregated at the rack level. ○ **Network Switch**: Instances are distributed or aggregated at the network switch level.

Section	Parameter	Required	Description
Basic Settings	Deployment Policy	No	The dispersion policies are used to improve service availability to avoid impacts on your business when a host, rack, or switch fails. The aggregation policies are used to improve network performance to minimize the access latency between instances. Valid values: ○ **Loose Dispersion** ○ **Strict Dispersion** ○ **Loose Aggregation** ○ **Strict Aggregation**
	Deployment Set Name	Yes	The name of the deployment set. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,). It must start with a letter and cannot start with http:// or https://.
	Description	No	The description of the deployment set. To simplify future management operations, we recommend that you provide an informational description. The description must be 2 to 256 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,). It must start with a letter and cannot start with http:// or https://.

6. Click **Submit**.

2.1.11.2. View deployment sets

You can view the list of deployment sets.

Procedure

1.
2.
3.
4. Select a filter option from the drop-down list, enter the relevant information in the search bar, and then click **Search**.

   You can select multiple filter options to narrow down search results.

Filter option	Description
Deployment Set Name	Enter a deployment set name to search for the deployment set.
Deployment Set ID	Enter a deployment set ID to search for the deployment set.

2.1.11.3. Modify a deployment set

You can modify the name and description of a deployment set.

Procedure

1.
2.
3.
4. Find the deployment set and click **Modify** in the **Actions** column.
5. In the Change Deployment Set dialog box, change the name of the deployment set.
6. Click **OK**.

2.1.11.4. Delete a deployment set

You can delete a deployment set that is no longer needed.

Prerequisites

No instances exist in the deployment set.

Procedure

1.
2.
3.
4. Find the deployment set and click **Delete** in the **Actions** column.
5. Click **OK**.

2.1.12. Dedicated hosts

2.1.12.1. Create a dedicated host

A dedicated host is a cloud host whose physical resources are reserved for the exclusive use of a single tenant. Elastic Compute Service (ECS) instances created on a dedicated host are physically isolated from those created on other hosts. This topic describes how to create a dedicated host.

Procedure

1. Log on to the ECS console.
2. In the left-side navigation pane, click **Dedicated Hosts**.
3.
4. On the **Hosts** tab, click **Create Host**.
5. On the **Create Host** page, configure the parameters described in the following table.

Section	Parameter	Description
Basic Settings	Organization	The organization in which to create the dedicated host.
	Resource Set	The resource set in which to create the dedicated host.
Region and Zone	Region	The region in which to create the dedicated host.
	Zone	The zone in which to create the dedicated host.

Section	Parameter	Description
Instance	Dedicated Host Type	The type of the dedicated host. The dedicated host type determines the instance family and the maximum number of ECS instances that you can deploy on the dedicated host.
	DDH Name	The name of the dedicated host. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,). It must start with a letter and cannot start with http:// or https://.
	Quantity	The number of dedicated hosts that you want to create.
DDH Settings	Allow Automatic Deployment	Specify whether to add the dedicated host to the resource pool for automatic deployment. Valid values: ○ Allow ○ Not Allow
	Automatic Instance Migration upon DDH Failure	Specify whether to fail over the instances deployed on the dedicated host when it fails. Valid values: ○ Enable ○ Disable

6. Click **Submit**.

Result

After the dedicated host is created, you can view it in the Dedicated Host list and create instances on it. For more information about the parameters used to create an ECS instance, see Create an instance.

2.1.12.2. Create a host group

You can group dedicated hosts into host groups for easy management. This topic describes how to create a host group.

Procedure

1. Log on to the ECS console.
2. In the left-side navigation pane, click **Dedicated Hosts**.
3.
4. Click the **Host Groups** tab.
5. Click **Create Host Group**.
6. On the **Create Host Group** page, configure the parameters described in the following table.

Section	Parameter	Required	Description
Basic Settings	Organization	Yes	The organization in which to create the host group.
	Resource Set	Yes	The resource set in which to create the host group.

Section	Parameter	Required	Description
Region and Zone	Region	Yes	The region in which to create the host group.
	Zone	Yes	The zone in which to create the host group.
Instance	Host Group Name	No	The name of the host group. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), hyphens (-), colons (:), and commas (,). It must start with a letter and cannot start with http:// or https://.

7. Click **Submit**.

Result

After the host group is created, you can view it in the host group list.

2.1.12.3. Add dedicated hosts to a host group

After you create a host group, you can add dedicated hosts to the host group for easy management. This topic describes how to add dedicated hosts to a host group.

Procedure

1. Log on to the ECS console.
2. In the left-side navigation pane, click **Dedicated Hosts**.
3.
4. Click the **Host Groups** tab.
5. Find the host group to which you want to add dedicated hosts and click **Add Host** in the **Actions** column.
6. In the **Add Host** panel, select a dedicated host and click **Add Host**.

   To add a new dedicated host, you can click **Create Host** in the **Add Host** panel. For information about the parameters used to create a host, see Create a dedicated host. After the host is created, add it to the host group.

Result

After the dedicated host is added to the host group, you can click the host group name to view the dedicated host in the **Hosts** list.

2.1.13. Install FTP software

2.1.13.1. Overview

File Transfer Protocol (FTP) transfers files between a client and a server by establishing two TCP connections. One is the command link for transferring commands between a client and a server. The other is the data link used to upload or download data. Before uploading files to an instance, you must build an FTP site for the instance.

2.1.13.2. Install and configure vsftp in CentOS

This topic describes how to install and configure vsftp in CentOS to transfer files.

Procedure

1. Install vsftp.

```
yum install vsftpd -y
```

2. Add an FTP account and a directory.
   i. Check the location of the *nologin* file,

      which is usually under the */usr/sbin* or */sbin* directory.
   ii. Create an FTP account.

      Run the following commands to create the */alidata/www/wwwroot* directory and specify this directory as the home directory of the account pwftp. You can also customize the account name and directory.

```
mkdir -p /alidata/www/wwwroot
useradd -d /alidata/www/wwwroot -s /sbin/nologin pwftp
```

   iii. Modify the account password.

```
passwd pwftp
```

   iv. Modify the permissions on the specified directory.

```
chown -R pwftp.pwftp /alidata/www/wwwroot
```

3. Configure vsftp.
   i. Open the vsftp configuration file.

```
vi /etc/vsftpd/vsftpd.conf
```

   ii. Change the value of anonymous_enable from `YES` to `NO`.
   iii. Delete the comment delimiter (`#`) from the following configuration lines:

```
local_enable=YES
        write_enable=YES
        chroot_local_user=YES
```

   iv. Press the Esc key to exit the edit mode, and enter `:wq` to save the modifications and exit.
4. Modify the shell configuration.
   i. Open the shell configuration file.

```
vi /etc/shells
```

   ii. If the file does not contain */usr/sbin/nologin* or */sbin/nologin*, add it to the file.
5. Start vsftp and perform a logon test.
   i. Start vsftp.

```
service vsftpd start
```

   ii. Use the account pwftp to perform an FTP logon test.

      This example uses the directory */alidata/www/wwwroot*.

2.1.13.3. Install vsftp in Ubuntu or Debian

This topic describes how to install and configure vsftp in an instance running Ubuntu or Debian to transfer files.

Procedure

1. Update the software source.

```
apt-get update
```

2. Install vsftp.

```
apt-get install vsftpd -y
```

3. Add an FTP account and a directory.
   i. Check the location of the *nologin* file,

      which is typically under the */usr/sbin* or */sbin* directory.
   ii. Create an FTP account.

      Run the following commands to create the */alidata/www/wwwroot* directory and specify this directory as the home directory of the account pwftp. You can also customize the account name and directory.

```
mkdir -p /alidata/www/wwwroot
useradd -d /alidata/www/wwwroot -s /sbin/nologin pwftp
```

   iii. Modify the account password.

```
passwd pwftp
```

   iv. Modify the permissions on the specified directory.

```
chown -R pwftp.pwftp /alidata/www/wwwroot
```

4. Configure vsftp.
   i. Open the vsftp configuration file.

```
vi /etc/vsftpd.conf
```

   ii. Change the value of anonymous_enable from `YES` to `NO`.
   iii. Delete the comment delimiter (`#`) from the following configuration lines:

```
local_enable=YES
        write_enable=YES
        chroot_local_user=YES
        chroot_list_enable=YES
        chroot_list_file=/etc/vsftpd.chroot_list
```

   iv. Press the Esc key to exit the edit mode, and enter `:wq` to save the modifications and exit.
   v. Open the */etc/vsftpd.chroot_list* file and add the FTP account name to the file. Save the modifications and exit.

      You can follow steps a to d to open and save the file.
5. Modify shell configurations.
   i. Open the shell configuration file.

```
vi /etc/shells
```

   ii. If the file does not contain */usr/sbin/nologin* or */sbin/nologin*, add it to the file.
6. Start vsftp and perform a logon test.
   i. Start vsftp.

```
service vsftpd restart
```

   ii. Use the account pwftp to perform an FTP logon test.

      This example uses the directory */alidata/www/wwwroot*.

2.1.13.4. Build an FTP site in Windows Server 2008

This topic describes how to build an FTP site on an instance running Windows Server 2008.

Prerequisites

You have added the Web Server (IIS) role and installed FTP on an instance.

Procedure

1. Connect to an instance.
2. Choose **Start > Adminstrative Tools > Internet Information Services (IIS) Manager**.
3. Right-click the server name and select **Add FTP Site** from the shortcut menu.
4. Enter an FTP site name and a physical path, and then click **Next**.
5. Set **IP Address** to **All Unassigned** and **SSL** to **No SSL**, and then click **Next**.
6. Set **Authentication** to **Basic**, **Authorization** to **All Users**, and **Permissions** to **Read** and **Write**, and click **Finish**.

Result

Then you can use the administrator account and password to upload and download files through FTP. Make sure that the following conditions are met:

- The port for the FTP site is not in use by other applications, and Windows firewall is not blocking the port.
- The security group of the instance contains a security group rule that allows inbound access to the FTP port.

2.1.13.5. Build an FTP site in Windows Server 2012

This topic describes how to build an FTP site on an instance running Windows Server 2012.

Prerequisites

You have added the Web Server (IIS) role and installed FTP on an instance.

Procedure

1. Connect to an instance.
2. Click the **Server Manager** icon.
3. In the left-side navigation pane, click **IIS**.
4. In the **Server** area, right-click the server name and select **Internet Information Services (IIS) Manager** from the shortcut menu.
5. Right-click the server name and select **Add FTP Site** from the shortcut menu.
6. Enter an FTP site name and a physical path, and then click **Next**.
7. Set **IP Address** to **All Unassigned** and **SSL** to **No SSL**, and then click **Next**.
8. Set **Authentication** to **Basic**, **Authorization** to **All Users**, and **Permissions** to **Read** and **Write**, and click **Finish**.

Result

Then you can use the administrator account and password to upload and download files through FTP. Make sure that the following conditions are met:

- The port for the FTP site is not in use by other applications, and Windows firewall is not blocking the port.
- The security group of the instance contains a security group rule that allows inbound access to the FTP port.

3.Container Service for Kubernetes

3.1. User Guide

3.1.1. Announcements

3.1.1.1. Container Service support for Kubernetes 1.18

Container Service strictly conforms to the terms of the Certified Kubernetes Conformance Program. This topic describes the changes that Container Service has made to support Kubernetes 1.18.

Version upgrades

Container Servcie has upgraded and optimized all of its components to support Kubernetes 1.18.8.

Key component	Version	Description
Kubernetes	1.18.8	Some frequently used API versions are deprecated in Kubernetes 1.18. Before you upgrade a Kubernetes cluster, we recommend that you upgrade the deprecated API versions that are listed in this topic.
Docker	19.03.5 (containerd 1.2.10)	No.
etcd	3.4.3	No.
CoreDNS	1.6.7	No.

Version details

- **Resource changes and deprecation**

  The following APIs are deprecated in Kubernetes 1.18:

  - The APIs apps/v1beta1 and apps/v1beta2 of all the resources are replaced by **apps/v1**.
  - The API extensions/v1beta1 of DaemonSets, Deployments, and ReplicaSets is replaced by **apps/v1**.
  - The API extensions/v1beta1 of Networkpolicies resources is replaced by **networking.k8s.io/v1**.
  - The API extensions/v1beta1 of pod security policies is replaced by **policy/v1beta1**.

  The label that specifies the regions of a node is changed to topology.kubernetes.io/region. The label that specifies the zone of a node is changed to topology.kubernetes.io/zone. We recommend that you update the related configurations for your workloads.

- **Feature upgrades**
  - Server-side Apply Beta 2 is introduced. You can view the relationships between the configuration items of a resource in the metadata.managedFields field of the resource.
  - The Node Local DNS Cache feature is released to improve the DNS availability and performance of your cluster.
  - The Volume Snapshot feature is in public preview and supports operations such as data volume backup, recovery, and scheduled backup.

Container Service upgrades for Kubernetes 1.18.8

In Kubernetes 1.18.8, Container Service enables the following feature in the kubelet configuration file: Users who use raw data volumes can upgrade clusters without the need to drain the nodes.

3.1.1.2. Vulnerability fixed: CVE-2021-1056 in NVIDIA GPU drivers

NVIDIA has reported the CVE-2021-1056 vulnerability, which is related to device isolation and NVIDIA GPU drivers. Elastic GPU Service nodes that are deployed in a Container Service cluster may also be exposed to this vulnerability. This topic describes the background information, affected versions, and fixes of this vulnerability.

Context

The CVE-2021-1056 vulnerability is related to device isolation and NVIDIA GPU drivers. This vulnerability allows an attacker to gain access to all GPU devices on a node by creating character device files in non-privileged containers that run on this node.

For more information about this vulnerability, see CVE-2021-1056.

Affected versions

The affected versions of NVIDIA GPU drivers are listed in the following figure based on the information published on the NVIDIA official website. For more information, see NVIDIA official website.

CVE IDs Addressed	Software Product	Operating System	Driver Branch	Affected Versions	Updated Driver Version
CVE-2021-1052 CVE-2021-1053	GeForce	Linux	R460	All versions prior to 460.32.03	460.32.03
			R450	All versions prior to 450.102.04	450.102.04
	NVIDIA RTX/Quadro, NVS	Linux	R460	All versions prior to 460.32.03	460.32.03
			R450	All versions prior to 450.102.04	450.102.04
	Tesla	Linux	R460	All versions prior to 460.32.03	460.32.03
			R450	All versions prior to 450.102.04	450.102.04
CVE-2021-1056	GeForce	Linux	R460	All versions prior to 460.32.03	460.32.03
			R450	All versions prior to 450.102.04	450.102.04
	NVIDIA RTX/Quadro, NVS	Linux	R460	All versions prior to 460.32.03	460.32.03
			R450	All versions prior to 450.102.04	450.102.04
			R390	All version prior to 390.141	390.141
	Tesla	Linux	R460	All versions prior to 460.32.03	460.32.03
			R450	All versions prior to 450.102.04	450.102.04
			R418	All versions prior to 418.181.07	418.181.07

- If you selected a custom NVIDIA driver or updated an NVIDIA driver, check whether the NVIDIA driver that you installed is affected by this vulnerability based on the preceding figure.
- If you use the NVIDIA driver that is automatically installed by Container Service, you must check whether the Kubernetes version of your cluster is affected by this vulnerability. The following Kubernetes versions are affected by this vulnerability:
  - ACK 1.16.9-aliyun.1. By default, the NVIDIA driver of version 418.87.01 is installed in clusters of this Kubernetes version.
  - ACK 1.18.8-aliyun.1. By default, the NVIDIA driver of version 418.87.01 is installed in clusters of this Kubernetes version.

> **Note** The NVIDIA GPU drivers that are installed by default in clusters of other Kubernetes versions are not affected. The Container Service team of Alibaba Cloud will keep you informed of further CVE content updates and help you fix the vulnerability.

Check the version of the NVIDIA driver on a GPU-accelerated node

Log on to the GPU-accelerated node and run the following command to query the version of the NVIDIA driver.

> **Note** For more information about how to log on to a GPU-accelerated node, see the *Use VNC to connect to and log on to an instance* chapter of the *Elastic Compute Service (ECS) User Guide*.

```
nvidia-smi
```

Expected output:

```
Fri Apr 16 10:58:19 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 418.87.01    Driver Version: 418.87.01    CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:07.0 Off |                    0 |
| N/A   34C    P0    37W / 300W |      0MiB / 16130MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

The output indicates that the version of the NVIDIA driver is 418.87.01.

Fixes

Notice When you upgrade the NVIDIA driver for a node, the node must be restarted. This interrupts services that are deployed on the node.

Upgrade the NVIDIA driver based on the preceding figure.

Note For more information about how to upgrade the NVIDIA driver, see Upgrade the NVIDIA driver on a GPU node.

- If your NVIDIA driver belongs to the R390 branch, upgrade it to version 390.141.
- If your NVIDIA driver belongs to the R418 branch, upgrade it to version 418.181.07.
- If your NVIDIA driver belongs to the R450 branch, upgrade it to version 450.102.04.
- If your NVIDIA driver belongs to the R460 branch, upgrade it to version 460.32.03.

3.1.2. What is Container Service?

Container Service provides high-performance, scalable, and enterprise-class management service for Kubernetes containerized applications throughout the application lifecycle.

Container Service simplifies the deployment and scaling operations on Kubernetes clusters. Integrated with services such as virtualization, storage, network, and security, Container Service aims to provide the optimal cloud environment for Kubernetes containerized applications. Alibaba Cloud is a Kubernetes Certified Service Provider (KCSP). As one of the first services to participate in the Certified Kubernetes Conformance Program, Container Service provides you with professional support and services.

3.1.3. ACK@Edge overview

ACK@Edge is released for commercial use. ACK@Edge is a cloud-managed solution that is provided by Container Service to coordinate cloud and edge computing. This topic describes the background and features of edge Kubernetes clusters.

Overview

With the rapid growth of smart devices connected to the Internet and the advent of 5G and IoT, computing and storage services provided by traditional cloud computing platforms can no longer satisfy the needs of edge devices for time-efficient computing, larger storage capacity, and enhanced computing capacity. Edge Kubernetes clusters are intended for bringing cloud computing to edges (clients). Edge Kubernetes clusters can be created, managed, and maintained in the Container Service console. This is the trend of cloud computing.

An edge Kubernetes cluster is a standard, secure, and highly-available Kubernetes cluster deployed in the cloud. This type of cluster is integrated with features of Alibaba Cloud, such as virtualization, storage, networking, and security. This simplifies the management and maintenance of clusters and allows you to focus on your business development. ACK@Edge provides the following features:

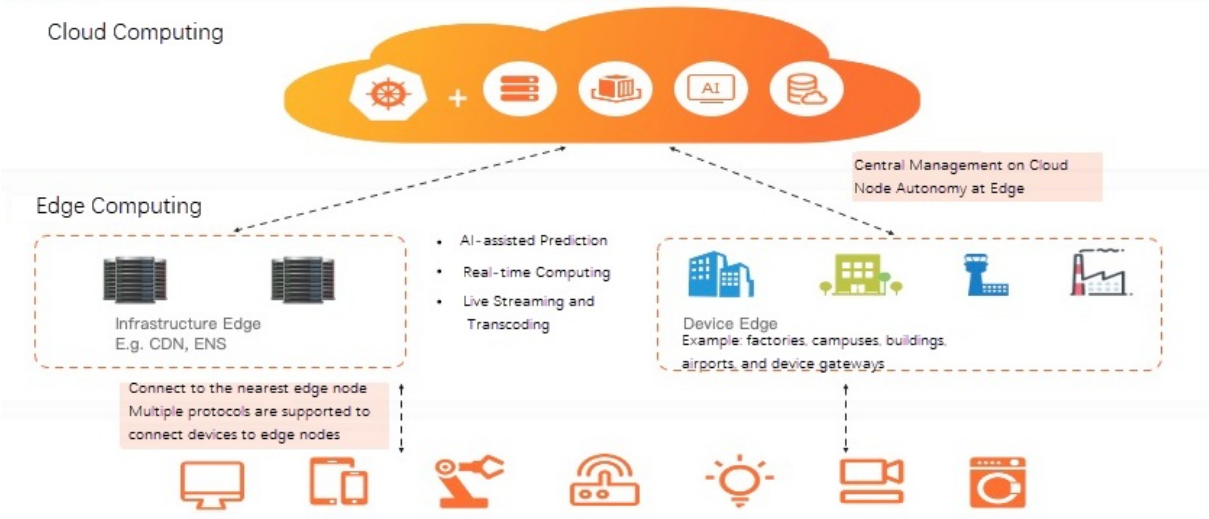
- Allows you to build a cloud-native infrastructure for edge computing with a few clicks.
- Allows you to quickly connect edge computing resources to the cloud. These resources include IoT gateway devices, terminals, Content Delivery Network (CDN) resources, and data centers.
- Applies to diverse scenarios, such as edge intelligence, intelligent buildings, intelligent factories, audio and video live streaming, online education, and CDN.

Edge Kubernetes clusters support features such as node autonomy, cell-based management, and native APIs for the management and maintenance of resources at the edge side. To use these features, you do not need to rewrite the logic of your services. This provides a native and centralized method for application lifecycle management and resource scheduling in edge computing scenarios.

Features

Edge Kubernetes clusters provide the following features to support lifecycle management for containerized applications and resources in edge computing scenarios:

- Allows you to create highly available edge Kubernetes clusters with a few clicks and provides lifecycle management on edge Kubernetes clusters, such as scaling cloud nodes, adding edge nodes to clusters, upgrading, logging, and monitoring. You can perform the preceding operations in the Container Service console.
- Supports access to various heterogeneous resources, such as data centers and IoT devices. Hybrid scheduling of heterogeneous resources is also supported.
- Supports node autonomy and network autonomy to ensure the reliability of edge nodes and services in edge computing scenarios where the network connection is weak.
- Supports reverse tunneling for management and maintenance of edge nodes.



3.1.4. Planning and preparation

Before you start using Container Service, you need to create cloud resources such as VPC networks, VSwitches, disks, and OSS buckets based on your application requirements.

Before you create a Kubernetes cluster, make the following preparations:

- **Create a VPC network** (optional)

  To create a cluster in an existing VPC network, you must create the VPC network and VSwitches in advance.

- **Create a volume** (optional)

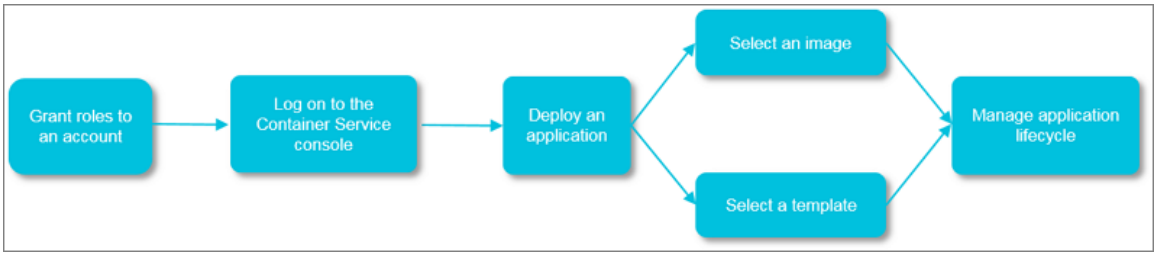
  To create a stateful application with network storage, you must create disks or OSS buckets in advance.

3.1.5. Quick start

3.1.5.1. Procedure

You can perform the following steps to use the Container Service service.

The following diagram shows the procedure to use the Container Service service.



Step 1: Authorize the default role

Authorize the default role of Container Service to perform operations on the resources that belong to the specified organization.

Step 2: Log on to the Container Service console

Log on to the Container Service console. For more information, see Log on to the Container Service console.

Step 3: Create an Container Service cluster

Set the network environment and the number of nodes, and configure node details.

Step 4: Deploy an application by using an image or orchestration template

You can use an existing image or orchestration template, or create a new image or orchestration template.

To create an application that consists of services based on different images, use an orchestration template.

Step 5: Manage the application lifecycle

3.1.5.2. Log on to the Container Service console

You can perform the following steps to log on to the Container Service console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

> **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
>
> - Uppercase or lowercase letters
> - Digits
> - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products > Elastic Computing > Container Service for Kubernetes**.
6. On the Container Service page, select **Access with Authorized Role** or **Access as Administrator**.
   - **Access with Authorized Role**: The system accesses the Container Service console by using an authorized account.
   - **Access as Administrator**: The system accesses the Container Service console as the organization administrator.

3.1.5.3. Create a Kubernetes cluster

This topic describes how to create a Kubernetes cluster in the Container Service console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**. On the Clusters page, click **Create Kubernetes Cluster** in the upper-right corner.
3. On the **Create Kubernetes Cluster** page, set basic configurations for the cluster.

Parameter	Description
Cluster Name	Enter a name for the cluster. The name must be 1 to 63 characters in length, and can contain digits, letters, and hyphens (-). **Note** The cluster name must be unique among clusters that belong to the same Apsara Stack tenant account.

Parameter	Description
Region	Select the region where you want to deploy the Kubernetes cluster.
VPC	You can select a virtual private cloud (VPC) from the drop-down list. ◦ If the specified VPC is already associated with a NAT gateway, the cluster uses this NAT gateway. ◦ Otherwise, the system automatically creates a NAT gateway. If you do not want the system to create a NAT gateway, clear **Configure SNAT for VPC**. **Note** If you disallow the system to automatically create a NAT gateway and want the VPC to access the Internet, you must manually associate the VPC with a NAT gateway or create SNAT rules for the VPC.
VSwitch	Select one or more vSwitches for the cluster. You can select up to three vSwitches that are deployed in different zones.
Kubernetes Version	Select a Kubernetes version.
Container Runtime	You can select Docker or Sandboxed-Container.
MASTER Configuration	Set the Instance Type and System Disk parameters: ◦ Master Node Quantity: You can add three master nodes. ◦ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ◦ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data.
Worker Instance	You can select **Create Instance** or **Add Existing Instance**.

Parameter	Description
WORKER Configuration	If you select **Create Instance** for **Worker Instance**, set the following parameters: ○ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ○ Selected Types: The selected instance types. ○ Quantity: Set the number of worker nodes. ○ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data. ○ Mount Data Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** ■ You can select **Encrypt Disk** to encrypt disks. ■ You can select **Enable Backup** to back up disk data.
Operating System	The CentOS and Alibaba Cloud Linux operating systems are supported.
Password	Set a password that is used to log on to the nodes. **Note** The password must be 8 to 30 characters in length, and must contain at least three of the following types of character: uppercase letters, lowercase letters, digits, and special characters.
Confirm Password	Enter the password again.
Network Plug-in	Flannel and Terway are supported. By default, Flannel is selected.
Pod CIDR Block and Service CIDR (optional)	For more information, see *Network planning* in *VPC User Guide*. **Note** These parameters are available only when you select an **existing** VPC.
Configure SNAT	This parameter is optional. If you clear Configure SNAT for VPC, you must create a NAT gateway or configure SNAT rules for the VPC.

Parameter	Description
Internet Access	Specify whether to expose the API server with an elastic IP address (EIP). The Kubernetes API server provides multiple HTTP-based RESTful APIs that can be used to create, delete, modify, query, and watch resource objects such as pods and Services. ◦ If you select this check box, an EIP is created and attached to an internal-facing Server Load Balancer (SLB) instance. Port 6443 used by the API server is exposed on the master nodes. You can connect to and manage the cluster by using kubeconfig over the Internet. ◦ If you clear this check box, no EIP is created. You can connect to and manage the cluster by using kubeconfig only from within the VPC.
SSH Logon	To enable SSH logon, you must first select Expose API Server with EIP. ◦ If you select Use SSH to Access the Cluster from the Internet, you can access the cluster through SSH. ◦ If you clear Use SSH to Access the Cluster from the Internet, you cannot access the cluster through SSH or kubectl. If you want to access an Elastic Compute Service (ECS) instance in the cluster through SSH, you must manually bind an EIP to the ECS instance and configure security group rules to open SSH port 22.
Security Group	You can select **Create Basic Security Group** or **Create Advanced Security Group**.
Ingress	Specify whether to **Install Ingress Controllers**. By default, **Install Ingress Controllers** is selected.
Log Service	If you enable Log Service, you can select an existing project or create a project. If you select **Enable Log Service**, the Log Service plug-in is automatically installed in the cluster. If you select **Create Ingress Dashboard**, the Ingress dashboard is created to collect and analyze the log of access to Ingresses. By default, **Install node-problem-detector and Create Event Center** is selected.
Monitoring Agents	Specify whether to enable **Prometheus Monitoring**. Prometheus Monitoring provides basic monitoring and alerting for the cluster.
Volume Plug-in	By default, **CSI** is selected.
Deletion Protection	If you select this check box, the cluster cannot be deleted in the console or by calling API operations.
Node Protection	This check box is selected by default to prevent nodes from being deleted in the console or by calling API operations.
Labels	Add labels to the cluster.

4. Complete the advanced settings of the cluster.

Parameter	Description
IP Addresses per Node	The number of IP addresses that can be assigned to a node.
Custom Image	You can select a custom image. After you select a custom image, all nodes in the cluster are deployed by using this image.
Node Port Range	Specify the value of **Node Port Range**.
Taints	Add taints to all worker nodes in the Kubernetes cluster.
Cluster Domain	The default domain name of the cluster is cluster.local. You can specify a custom domain name.
Cluster CA	Specify whether to enable the cluster certification authority (CA) certificate.
User Data	Customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ◦ Run scripts during instance startup. ◦ Import user data as common data to an ECS instance for future reference.

5. Click **Create Cluster** in the upper-right corner of the page.
6. On the **Confirm** page, after all check items are verified, select the terms of service and disclaimer and click **OK** to start the deployment.

Result

After the cluster is created, you can view the created cluster on the **Clusters** page in the Container Service console.

3.1.5.4. Create an application from an orchestration template

Container Service provides orchestration templates that you can use to create applications. You can also modify the templates based on YAML syntax to customize applications.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Context

This topic describes how to use an orchestration template to create an NGINX application that consists of a Deployment and a Service. The Deployment provisions pods for the application and the Service manages access to the pods at the backend.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.

5. On the **Deployments** page, select the namespace and click **Create from Template** in the upper-right corner.
6. Configure the template and click **Create**.
   - **Sample Template**: Container Service provides YAML templates of various resource types to help you quickly deploy resource objects. You can also create a custom template based on YAML syntax to describe the resource that you want to define.
   - **Add Deployment**: This feature allows you to quickly define a YAML template.
   - **Use Existing Template**: You can import an existing template to the configuration page.

   The following NGINX template is based on an orchestration template provided by Container Service. You can use this template to quickly create a Deployment to run an NGINX application

   > **Note** Container Service supports YAML syntax. You can use the `---` symbol to separate multiple resource objects. This allows you to define multiple resource objects in a single template.

```
apiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1
kind: Deployment
metadata:
    name: nginx-deployment
    labels:
      app: nginx
spec:
    replicas: 2
    selector:
      matchLabels:
        app: nginx
    template:
      metadata:
        labels:
          app: nginx
      spec:
        containers:
        - name: nginx
          image: nginx:1.7.9 # replace it with your exactly <image_name:tags>
          ports:
          - containerPort: 80
---
apiVersion: v1     # for versions before 1.8.0 use apps/v1beta1
kind: Service
metadata:
   name: my-service1        #TODO: to specify your service name
   labels:
     app: nginx
spec:
   selector:
     app: nginx             #TODO: change label selector to match your backend pod
   ports:
   - protocol: TCP
     name: http
     port: 30080            #TODO: choose an unique port on each node to avoid port conflict
     targetPort: 80
   type: LoadBalancer        ## In this example, the type is changed from NodePort to LoadBalancer
.
```

7. Click **Create**. A notification that indicates the deployment status appears.

After the application is created, choose **Network > Services** in the left-side navigation pane. On the Services page, you can find that a Service named my-service1 is created for the application. The external endpoint of the Service is also displayed on the page. Click the endpoint in the **External Endpoint** column.

8. You can visit the NGINX welcome page in the browser.

3.1.6. Kubernetes clusters

3.1.6.1. Authorizations

3.1.6.1.1. Assign RBAC roles to a RAM user

This topic describes how to assign role-based access control (RBAC) roles to Resource Access Management (RAM) users. By default, RBAC is enabled for Kubernetes 1.6 and later. RBAC is important for you to manage clusters. You can use RBAC to specify the types of operations that are allowed for specific users based on their roles in an organization.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Authorizations**.
3. On the **RAM Users** tab, select the RAM user to which you want to grant permissions and click **Modify Permissions**.

   > **Note** If you log on to the Container Service console as a RAM user, make sure that the RAM user is assigned the predefined RBAC administrator role or the cluster-admin role.

4. On the **Configure Role-Based Access Control (RBAC)** wizard page, click **Add Permissions** to add cluster-scoped or namespace-scoped permissions and select a predefined or custom RBAC role in the Permission column. You can also click the minus sign (-) to delete permissions. After you add the permissions, click **Next Step**.

   > **Note** For each RAM user, you can assign only one predefined RBAC role but one or more custom RBAC roles to manage the same cluster or namespace.

   The following table describes the permissions that the predefined and custom RBAC roles have on clusters and namespaces.

   Roles and permissions

Role	RBAC permissions on cluster resources
Administrator	Read and write permissions on resources in all namespaces.
O&M Engineer	Read and write permissions on resources in all namespaces and read-only permissions on nodes, persistent volumes (PVs), namespaces, and service quotas within a cluster.

Role	RBAC permissions on cluster resources
Developer	Read and write permissions on resources in a specified namespace or all namespaces.
Restricted User	Read-only permissions on resources in a specified namespace or all namespaces.
Custom	The cluster role that you select for a custom role determines what permissions the custom role has. Before you select a cluster role, make sure that you are aware of the permissions that the cluster role has in case the RAM user is granted excessive permissions.

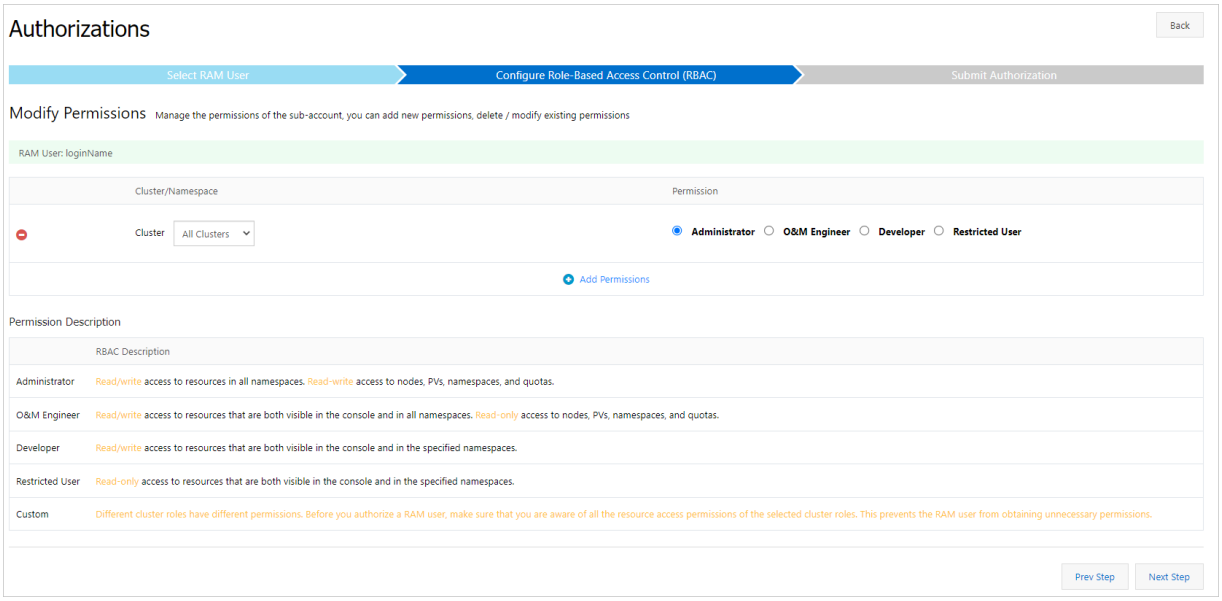
After the authorization is complete, you can use the account of the specified RAM user to log on to the Container Service console. For more information, see Log on to the Container Service console.

Predefined and custom RBAC roles

Container Service provides the following predefined RBAC roles: administrator, O&M engineer, developer, and restricted user. You can use these roles to regulate Container Service access control in most scenarios. In addition, you can use custom roles to customize permissions on clusters.

Container Service provides a set of custom RBAC roles.

> **Note** The cluster-admin role is similar to a super administrator. By default, the cluster-admin role has the permissions to manage all resources within a cluster.



You can log on to a master node of a cluster and run the following command to view the custom RBAC roles that are assigned to the current account:

```
# kubectl get clusterrole
```

```
# kubectl get clusterrole
NAME                                                                   AGE
admin                                                                  13d
alibaba-log-controller                                                 13d
alicloud-disk-controller-runner                                        13d
cluster-admin                                                          13d
cs:admin                                                               13d
edit                                                                   13d
flannel                                                                13d
kube-state-metrics                                                     22h
node-exporter                                                          22h
prometheus-k8s                                                         22h
prometheus-operator                                                    22h
system:aggregate-to-admin                                              13d
....
system:volume-scheduler                                                13d
view                                                                   13d
```

Run the following command to view the details of a role, for example, the cluster-admin role:

```
# kubectl get clusterrole cluster-admin -o yaml
```

Notice After a RAM user is assigned the cluster-admin role, the RAM user has the same permissions as the Alibaba Cloud account to which the RAM user belongs. The RAM user has full control over all resources within the cluster. Proceed with caution.

```
# kubectl get clusterrole cluster-admin  -o yaml
apiVersion: rbac.authorization.k8s.io/v1
kind: ClusterRole
metadata:
  annotations:
    rbac.authorization.kubernetes.io/autoupdate: "true"
  creationTimestamp: 2018-10-12T08:31:15Z
  labels:
    kubernetes.io/bootstrapping: rbac-defaults
  name: cluster-admin
  resourceVersion: "57"
  selfLink: /apis/rbac.authorization.k8s.io/v1/clusterroles/cluster-admin
  uid: 2f29f9c5-cdf9-11e8-84bf-00163e0b2f97
rules:
- apiGroups:
  - '*'
  resources:
  - '*'
  verbs:
  - '*'
- nonResourceURLs:
  - '*'
  verbs:
  - '*'
```

3.1.6.2. Clusters

3.1.6.2.1. Create a Kubernetes cluster

This topic describes how to create a Kubernetes cluster in the Container Service console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**. On the Clusters page, click **Create Kubernetes Cluster** in the upper-right corner.
3. On the **Create Kubernetes Cluster** page, set basic configurations for the cluster.

Parameter	Description
Cluster Name	Enter a name for the cluster. The name must be 1 to 63 characters in length, and can contain digits, letters, and hyphens (-). **Note** The cluster name must be unique among clusters that belong to the same Apsara Stack tenant account.
Region	Select the region where you want to deploy the Kubernetes cluster.
VPC	You can select a virtual private cloud (VPC) from the drop-down list. ◦ If the specified VPC is already associated with a NAT gateway, the cluster uses this NAT gateway. ◦ Otherwise, the system automatically creates a NAT gateway. If you do not want the system to create a NAT gateway, clear **Configure SNAT for VPC**. **Note** If you disallow the system to automatically create a NAT gateway and want the VPC to access the Internet, you must manually associate the VPC with a NAT gateway or create SNAT rules for the VPC.
VSwitch	Select one or more vSwitches for the cluster. You can select up to three vSwitches that are deployed in different zones.
Kubernetes Version	Select a Kubernetes version.
Container Runtime	You can select Docker or Sandboxed-Container.

Parameter	Description
MASTER Configuration	Set the Instance Type and System Disk parameters: ◦ Master Node Quantity: You can add three master nodes. ◦ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ◦ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data.
Worker Instance	You can select **Create Instance** or **Add Existing Instance**.
WORKER Configuration	If you select **Create Instance** for **Worker Instance**, set the following parameters: ◦ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ◦ Selected Types: The selected instance types. ◦ Quantity: Set the number of worker nodes. ◦ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data. ◦ Mount Data Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** ■ You can select **Encrypt Disk** to encrypt disks. ■ You can select **Enable Backup** to back up disk data.
Operating System	The CentOS and Alibaba Cloud Linux operating systems are supported.
Password	Set a password that is used to log on to the nodes. **Note** The password must be 8 to 30 characters in length, and must contain at least three of the following types of character: uppercase letters, lowercase letters, digits, and special characters.
Confirm Password	Enter the password again.

Parameter	Description
Network Plug-in	Flannel and Terway are supported. By default, Flannel is selected.
Pod CIDR Block and Service CIDR (optional)	For more information, see *Network planning* in *VPC User Guide*. **Note** These parameters are available only when you select an **existing** VPC.
Configure SNAT	This parameter is optional. If you clear Configure SNAT for VPC, you must create a NAT gateway or configure SNAT rules for the VPC.
Internet Access	Specify whether to expose the API server with an elastic IP address (EIP). The Kubernetes API server provides multiple HTTP-based RESTful APIs that can be used to create, delete, modify, query, and watch resource objects such as pods and Services. ◦ If you select this check box, an EIP is created and attached to an internal-facing Server Load Balancer (SLB) instance. Port 6443 used by the API server is exposed on the master nodes. You can connect to and manage the cluster by using kubeconfig over the Internet. ◦ If you clear this check box, no EIP is created. You can connect to and manage the cluster by using kubeconfig only from within the VPC.
SSH Logon	To enable SSH logon, you must first select Expose API Server with EIP. ◦ If you select Use SSH to Access the Cluster from the Internet, you can access the cluster through SSH. ◦ If you clear Use SSH to Access the Cluster from the Internet, you cannot access the cluster through SSH or kubectl. If you want to access an Elastic Compute Service (ECS) instance in the cluster through SSH, you must manually bind an EIP to the ECS instance and configure security group rules to open SSH port 22.
Security Group	You can select **Create Basic Security Group** or **Create Advanced Security Group**.
Ingress	Specify whether to **Install Ingress Controllers**. By default, **Install Ingress Controllers** is selected.
Log Service	If you enable Log Service, you can select an existing project or create a project. If you select **Enable Log Service**, the Log Service plug-in is automatically installed in the cluster. If you select **Create Ingress Dashboard**, the Ingress dashboard is created to collect and analyze the log of access to Ingresses. By default, **Install node-problem-detector and Create Event Center** is selected.

Parameter	Description
Monitoring Agents	Specify whether to enable **Prometheus Monitoring**. Prometheus Monitoring provides basic monitoring and alerting for the cluster.
Volume Plug-in	By default, **CSI** is selected.
Deletion Protection	If you select this check box, the cluster cannot be deleted in the console or by calling API operations.
Node Protection	This check box is selected by default to prevent nodes from being deleted in the console or by calling API operations.
Labels	Add labels to the cluster.

4. Complete the advanced settings of the cluster.

Parameter	Description
IP Addresses per Node	The number of IP addresses that can be assigned to a node.
Custom Image	You can select a custom image. After you select a custom image, all nodes in the cluster are deployed by using this image.
Node Port Range	Specify the value of **Node Port Range**.
Taints	Add taints to all worker nodes in the Kubernetes cluster.
Cluster Domain	The default domain name of the cluster is cluster.local. You can specify a custom domain name.
Cluster CA	Specify whether to enable the cluster certification authority (CA) certificate.
User Data	Customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ○ Run scripts during instance startup. ○ Import user data as common data to an ECS instance for future reference.

5. Click **Create Cluster** in the upper-right corner of the page.
6. On the **Confirm** page, after all check items are verified, select the terms of service and disclaimer and click **OK** to start the deployment.

Result

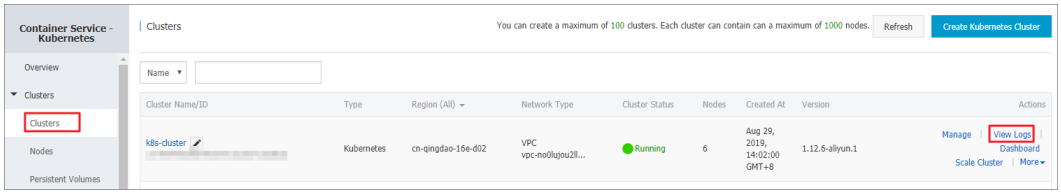
After the cluster is created, you can view the created cluster on the **Clusters** page in the Container Service console.

3.1.6.2.2. View log files of a cluster

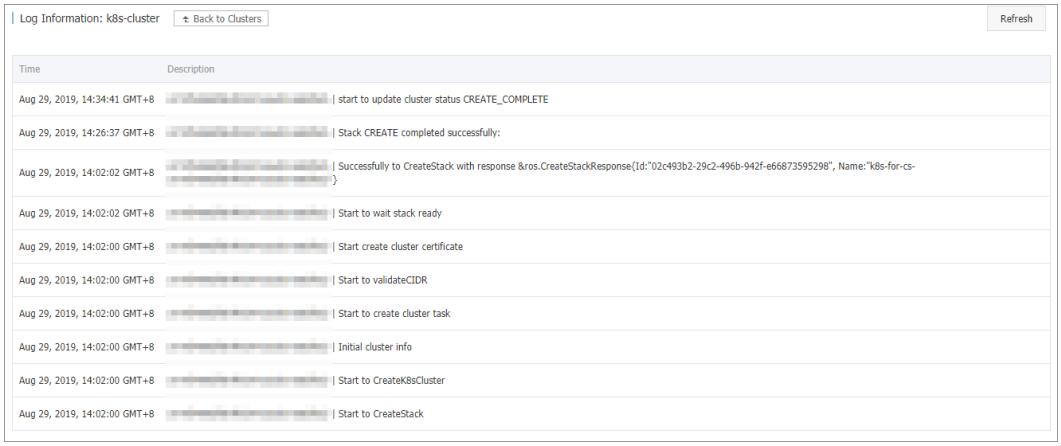
You can view the operation log of a cluster in the Container Service console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. Find the cluster that you want to manage and click **View Logs** in the Actions column.



You can view operations performed on the cluster.



3.1.6.2.3. Connect to a cluster through kubectl

You can use the Kubernetes command line tool, kubectl, to connect to a Kubernetes cluster from a local computer.

Procedure

1. Download the latest kubectl client from the Kubernetes change log page.
2. Install and set up the kubectl client.

   For more information, see Install and set up kubectl.
3. Configure the cluster credentials.

   You can use the `scp` command to securely copy the master node configuration file from the `/etc/kubernetes/kube.conf` directory of the master VM and paste it to the `$HOME/.kube/config` directory of the local computer, where the `kubectl` credentials are expected to be stored.

   ```
   mkdir $HOME/.kube
   scp root@<master-public-ip>:/etc/kubernetes/kube.conf $HOME/.kube/config
   ```

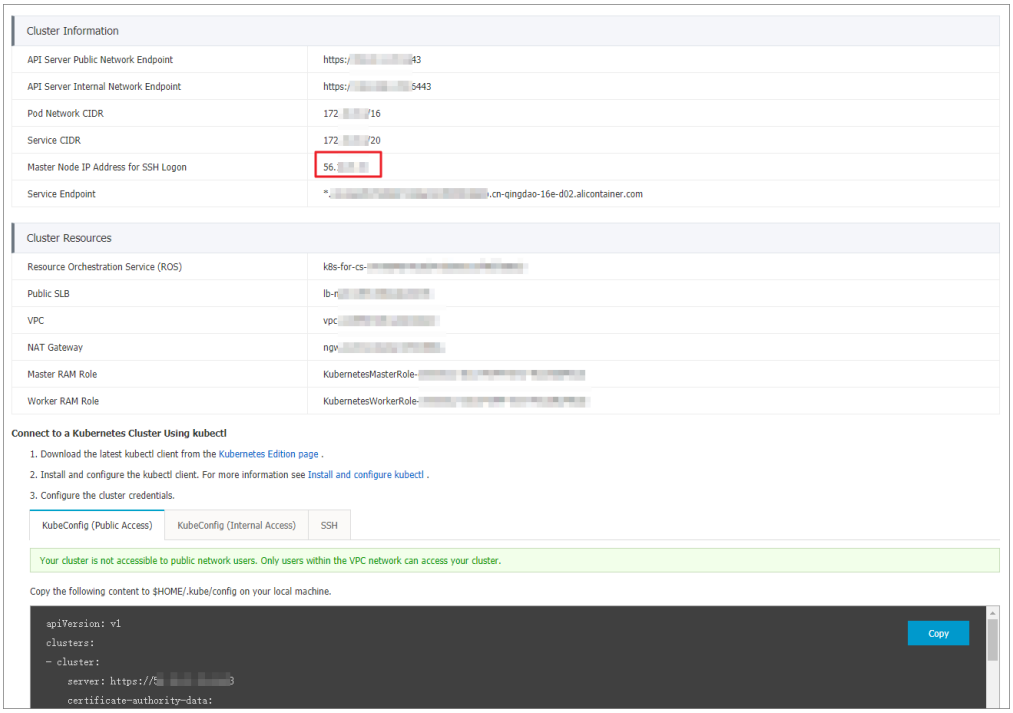
   You can find `master-public-ip` on the cluster details page.

   i. Log on to the Container Service console.

   ii. In the left-side navigation pane, click **Clusters**. The Clusters page appears.

iii. Find the target cluster and click **Manage** in the Actions column.

In the **Cluster Information** section, you can find the master node IP address.



3.1.6.2.4. Connect to a master node by using SSH

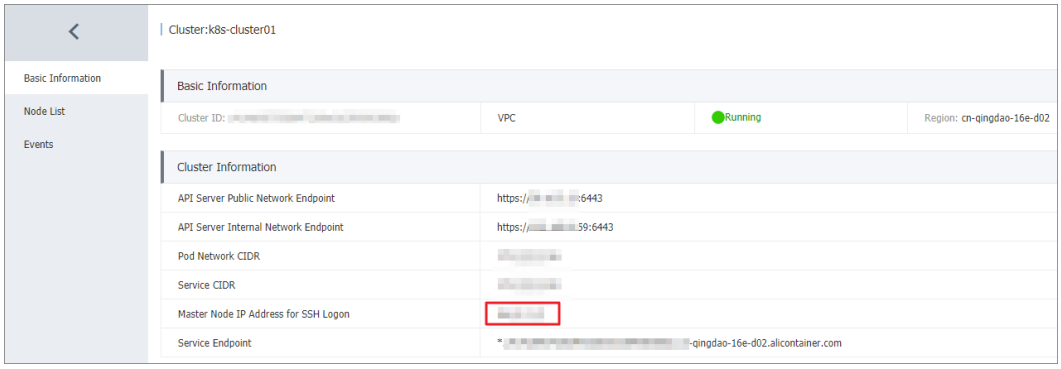
You can access a master node in a cluster by using a Secure Shell (SSH) client.

Prerequisites

- A Kubernetes cluster is created and **Use SSH to Access the Cluster from the Internet** is selected for the cluster. For more information, see Create a Kubernetes cluster.
- The SSH client can connect to the network where the cluster is deployed.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Clusters** > **Clusters** to go to the Clusters page. Find the cluster that you want to manage, and click Manage in the Actions column for the cluster.
3. The Basic Information page appears. In the Cluster Information section, you can find the IP address that is displayed in the **Master Node IP Address for SSH Logon** field.



4. Use SSH to connect to the cluster from an SSH client that has access to the cluster network.
   - If you have a leased line that connects to the cluster network over the Internet, you can use tools such as PuTTY to create an SSH connection.
   - If you have an Elastic Compute Service (ECS) instance that is connected to the Virtual Private Cloud (VPC) network of the cluster, run the following command to create an SSH connection:

     ```
     ssh    root@ssh_ip                    #ssh_ip specifies the IP address of the master node for SSH connection.
     ```

3.1.6.2.5. Expand a cluster

This topic describes how to scale out the worker nodes of a Kubernetes cluster in the Container Service console.

Context

You cannot scale out the master nodes of a Kubernetes cluster.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to expand and choose **More > Expand** in the **Actions** column.
4. On the **Node Pools** page, select the node pool that you want to scale out and click **Scale Out** in the **Actions** column.
5. Go to the Expand page and set the required parameters.

   In this example, the number of worker nodes in the cluster is increased from three to five. The following table describes the required parameters.

Parameter	Description
Nodes to Add	Specify the number of nodes to be added to the cluster.
Region	By default, the region where the cluster is deployed is displayed.
Container Runtime	By default, the container runtime of the cluster is displayed.
VPC	By default, the virtual private cloud (VPC) of the cluster is displayed.
VSwitch	Select one or more vSwitches for the cluster. You can select at most three vSwitches that are deployed in different **zones**.
Instance Type	You can select one or more instance types. For more information, see the *Instance types* topic of *ECS User Guide*.
Selected Types	The selected instance types.
System Disk	Standard SSDs, enhanced SSDs (ESSDs), and ultra disks are supported.
Mount Data Disk	Standard SSDs, ESSDs, and ultra disks are supported.
Operating System	The operating system of the cluster.
Password	◦ **Password**: Enter the password that is used to log on to the nodes. ◦ **Confirm Password**: Enter the password again.

Parameter	Description
ECS Label	You can add labels to the ECS instances.
Node Label	Add labels to nodes.
Taints	Add taints to the worker nodes in the cluster.
Custom Image	You can select a custom image. After you select a custom image, all nodes in the cluster are deployed by using this image.
RDS Whitelist	Set the Relational Database Service (RDS) whitelist. Add the IP addresses of the nodes in the cluster to the RDS whitelist.
User Data	Customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used in the following ways: ○ Run scripts during instance startup. ○ Import user data as normal data to an ECS instance for future reference.

6. Click **Submit**.

What's next

After the cluster is expanded, choose **Nodes > Nodes** in the left-side navigation pane. On the Nodes page, you can find that the number of worker nodes is increased from 3 to 5.

3.1.6.2.6. Renew a certificate

This topic describes how to renew a Kubernetes cluster certificate in the console.

Prerequisites

A Kubernetes cluster is created and the cluster certificate is about to expire.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. Select the cluster for which you want to renew the certificate and click **Update Certificate**. The **Update Certificate** message appears.

   > **Note** The **Update Certificate** button will be displayed two months before your cluster certificate expires.

4. Click **Update** and the **Confirm** page appears.
5. Click **OK**.

Result

- On the **Update Certificate** page, the following message appears: **The certificate has been updated**.
- On the **Clusters** page, the **Update Certificate** button disappears.

3.1.6.2.7. Delete a Kubernetes cluster

This topic describes how to delete a Kubernetes cluster in the Container Service console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to delete and choose **More > Delete** in the **Actions** column.
4. In the **Delete Cluster** dialog box that appears, select the resources that you want to retain, select **I understand the above information and want to delete the specified cluster**, and then click **OK**.

What's next

Resource Orchestration Service (ROS) does not have permissions to delete resources that are manually added to resource created by ROS. For example, if you manually add a vSwitch to a virtual private cloud (VPC) created by ROS, ROS cannot delete the VPC and therefore the cluster cannot be deleted.

Container Service allows you to forcibly delete clusters. If your first attempt to delete a cluster fails, you can forcibly delete the cluster and ROS resource stack. However, you still need to manually release the resources that are manually added.

An error message appears when an attempt to delete a cluster fails.

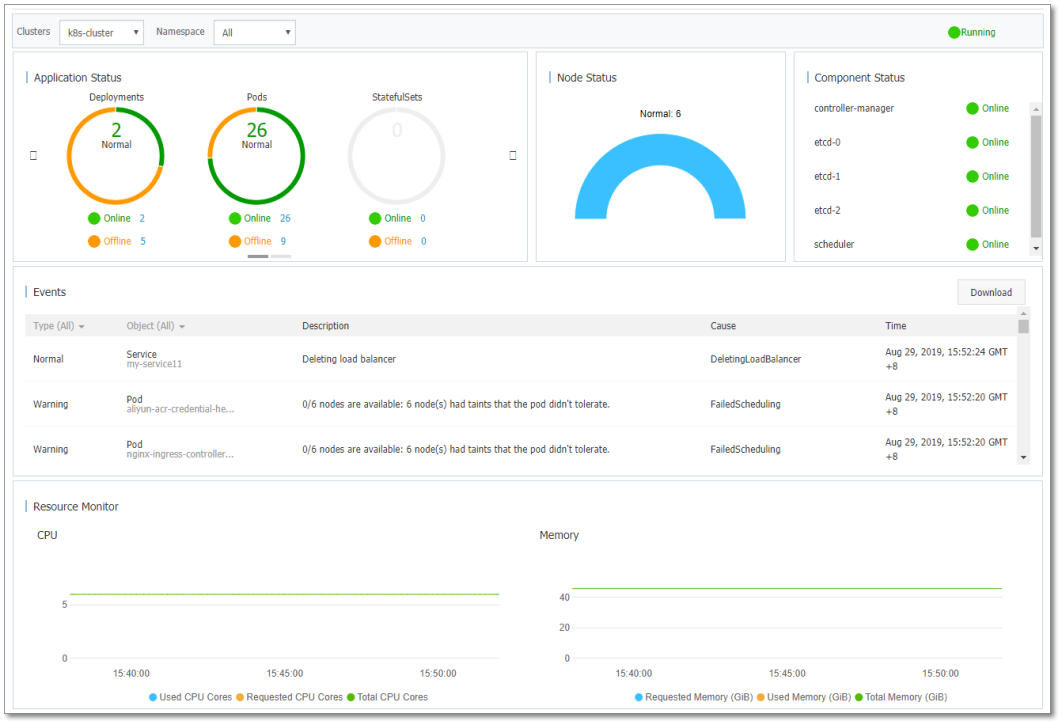
Select the cluster that you failed to delete and choose **More > Delete** in the Actions column. In the dialog box that appears, you can view the resources that are manually added. Select the **Force Delete** check box and click **OK** to delete the cluster and ROS resource stack.

3.1.6.2.8. View cluster overview

The Container Service console provides a cluster overview page. This page displays the information such as application status, component status, and resource monitoring status. This allows you to check the health status of your cluster at your convenience.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Overview**. The Overview page appears.
3. Select the target cluster and namespace. You can view the application status, component status, and resource monitoring charts.
   - **Application Status**: displays the statuses of the deployments, pods, and replica sets that are running in the cluster. Green sections indicate a normal state and yellow sections indicate an exception state.
   - **Node Status**: displays the statuses of the nodes in the cluster.
   - **Component Status**: Components are deployed in the kube-system namespace. Core components are used, such as the scheduler, controller-manager, and etcd.
   - **Events**: displays events such as warnings and errors. If no events are displayed, the cluster is running in the normal state.
   - **Monitoring**: displays CPU and memory monitoring charts. CPU usage is measured in cores or millicores and accurate to three decimal places. A millicore is one thousandth of a core. Memory usage is measured in GiB and accurate to three decimal places. For more information, see Meaning of CPU and Meaning of memory.



3.1.6.3. Nodes

3.1.6.3.1. Add existing nodes to a Kubernetes cluster

Container Service allows you to add an existing Elastic Compute Service (ECS) instance to a Kubernetes cluster. You can add only worker nodes to clusters.

Prerequisites

- A Kubernetes cluster is created. For more information, see Log on to the Container Service console.
- An ECS instance is created. Make sure that the region, zone, organization, project, security group, virtual private cloud (VPC), and operating system settings of the ECS instance are the same as those of the cluster.

Context

- By default, a cluster can contain at most 50 nodes. To increase the quota, submit a ticket.
- The ECS instance that you add must be in the same region and VPC as the cluster.
- The ECS instance must belong to the same Apsara Stack tenant account as the cluster.
- The ECS instance must be running the CentOS operating system.

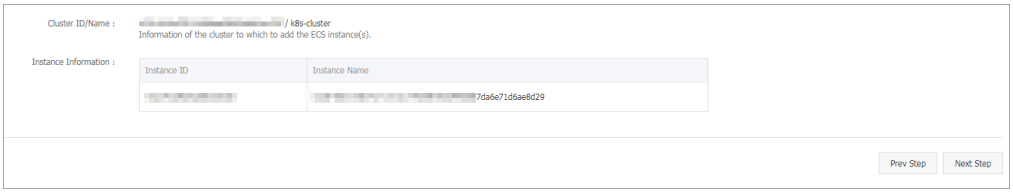
Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. In the upper-right corner of the page, click **Add Existing Node**.
6. On the page that appears, you can manually add existing ECS instances to the cluster.

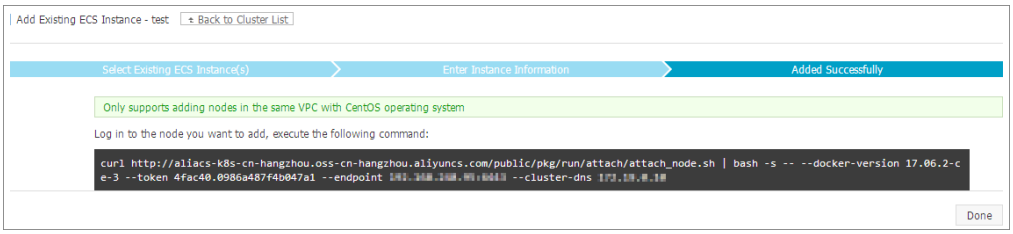
   To manually add an ECS instance, you must obtain the installation command and log on to the ECS instance to run the command. You can add only one ECS instance at a time.

i. Select **Manual**, find and select the ECS instance that you want to add, and then click **Next Step**. You can add only one ECS instance at a time.

ii. Confirm the instance information and click **Next Step**.



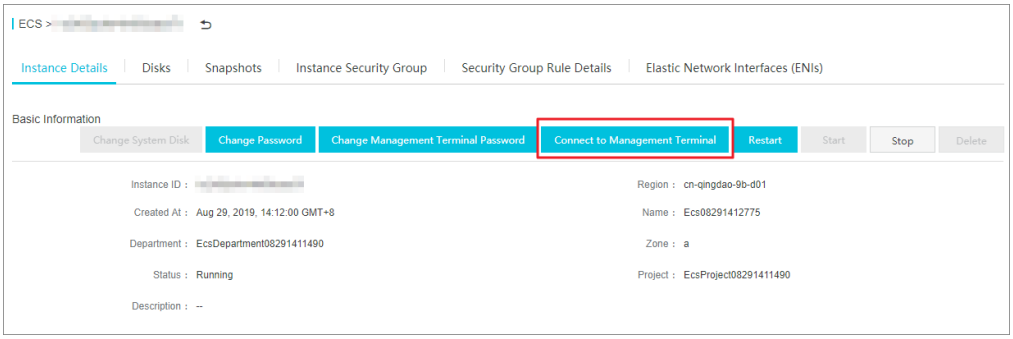
iii. On the Complete wizard page, copy the command.

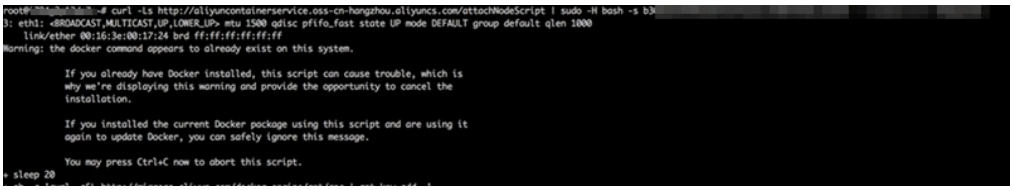


iv. Click **Done**.

v. Go to the Apsara Uni-manager Management Console. In the top navigation bar, choose **Products > Elastic Compute Service**. On the **Instances** page, select the organization and region of the cluster, and then find the ECS instance that you want to add to the cluster.

vi. Click the instance name to go to the Instance Details tab. Click **Connect to VNC**. In the dialog box that appears, enter the VNC password and then click **OK**. After you log on to the instance, paste the copied command and click **OK** to run the script.





vii. After the script is executed, the ECS instance is added to the cluster. You can go to the Clusters page and click the cluster ID to view nodes in the cluster. Check whether the ECS instance has been added to the cluster.

3.1.6.3.2. View nodes

You can view nodes of a Kubernetes cluster by using the Container Service console, kubectl, or Kubernetes Dashboard.

View nodes by using kubectl

> **Note** To view the nodes in a cluster by using kubectl, you must Connect to a Kubernetes cluster through kubectl.

Connect to a cluster by using kubectl and run the following command to view the nodes in the cluster:

```
kubectl get nodes
```

Sample output:

```
$ kubectl get nodes
    NAME                       STATUS    AGE       VERSION
    iz2ze2n6ep53tch701yh9zz    Ready     19m       v1.6.1-2+ed9e3d33a07093
    iz2zeafr762wibijx39e5az    Ready     7m        v1.6.1-2+ed9e3d33a07093
    iz2zeafr762wibijx39e5bz    Ready     7m        v1.6.1-2+ed9e3d33a07093
    iz2zef4dnn9nos8elyr32kz    Ready     14m       v1.6.1-2+ed9e3d33a07093
    iz2zeitvvo8enoreufstkmz    Ready     11m       v1.6.1-2+ed9e3d33a07093
```

View nodes by using the Container Service console

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes** to view the nodes in the cluster.

3.1.6.3.3. Manage node labels

You can manage node labels in the Container Service console. You can add a label to multiple nodes at a time, filter nodes by label, and delete labels.

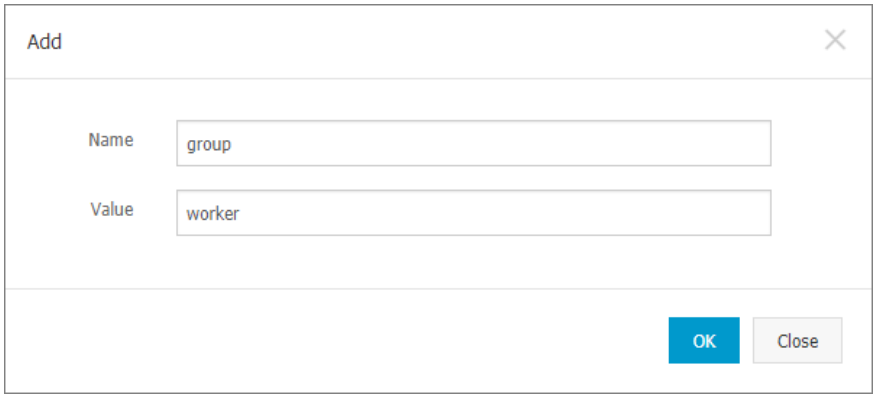
You can use labels to schedule nodes. For more information, see Set node scheduling.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Add a label to multiple nodes at a time

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. On the **Nodes** page, click **Manage Labels and Taints** in the upper-right corner.
6. Select multiple nodes and click **Add Label**.
7. In the dialog box that appears, enter the name and value of the label and click **OK**.



On the Labels tab, you can find that the selected nodes have the same label.

Delete a label

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. On the **Nodes** page, click **Manage Labels and Taints** in the upper-right corner.
6. Select a node, find the label that you want to delete, and then click the icon. In the message that appears, click **Confirm**.

   After the label is deleted, it is removed from the Labels column.

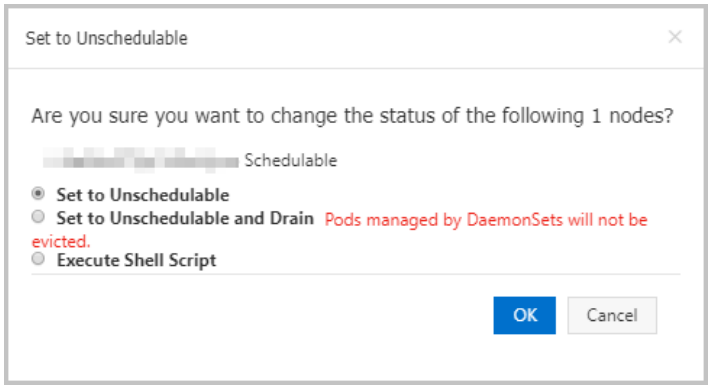
3.1.6.3.4. Set node schedulability

You can mark a node as schedulable or unschedulable in the Container Service console. This allows you to optimize the distribution of the loads on each node. This topic describes how to set node schedulability.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. On the Nodes page, select the node that you want to manage and click **Drain/Set to Unschedulable** in the Actions column.
6. In the dialog box that appears, you can change the status of the node.
   - If you select **Set to Unschedulable**, pods will not be scheduled to this node when you deploy new applications.
   - If you select **Set to Unschedulable and Drain**, pods will not be scheduled to this node when you deploy new applications. Pods on this node will be evicted, except for the pods that are managed by DaemonSets.

     In this example, **Set to Unschedulable** is selected.



7. Click **OK**.

   The status of the node is changed to Unschedulable.

What's next

Pods will not be scheduled to the node when you deploy new applications.

3.1.6.3.5. Remove a node

To restart or release an Elastic Compute Service (ECS) node in a cluster, you must first remove the node from the cluster. This topic describes how to remove a node.

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- You can connect to the Kubernetes cluster by using kubectl. For more information, see Connect to a Kubernetes cluster through kubectl.

Context

- When you remove a node, pods that run on the node are migrated to other nodes. This may cause service interruption. We recommend that you remove nodes during off-peak hours.
- Unknown errors may occur when you remove nodes. Before you remove nodes, we recommend that you back up data on these nodes.
- Nodes remain in the unschedulable state when they are being removed.
- You can remove only worker nodes. You cannot remove master nodes.

Procedure

1. Run the following command to migrate the pods on the node that you want to remove to other nodes.

   > **Note** Make sure that the other nodes have sufficient resources for these pods.

   ```
   kubectl drain node-name
   ```

   > **Note** *node-name* must be in the format of *your-region-name.node-id*.
   >
   > - *your-region-name* specifies the region where the cluster that you want to manage is deployed.
   > - *node-id* specifies the ID of the ECS instance where the node to be removed is deployed. Example: *cn-hanghzou.i-xxx*.

2. Log on to the Container Service console.
3. In the left-side navigation pane, click **Clusters**.

4. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
5. In the left-side navigation pane of the details page, choose **Nodes > Node**.
6. Find the node that you want to remove and choose **More > Remove** in the **Actions** column.

   > **Note** To remove multiple nodes at a time, select the nodes that you want to remove on the **Nodes** page and click **Batch Remove**.

7. (Optional)Set parameters in the Remove Node dialog box.
   - Select **Drain the Node** to migrate the pods on the node to other nodes. If you select this option, make sure that the other nodes have sufficient resources for these pods.
   - Select **Release ECS Instance** to release the ECS instance where the node is deployed.

   > **Note**
   > - Select this option to release only pay-as-you-go ECS instances.
   > - Subscription ECS instances are automatically released after the subscription expires.
   > - If you do not select **Release ECS Instance**, you are still billed for the ECS instance where the node is deployed.

8. In the **Remove Node** message, click **OK**.

3.1.6.3.6. View node resource usage

You can view the resource usage of the nodes in a cluster in the Container Service console.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. On the Nodes page, find the node that you want to manage and choose **More > Details** in the **Actions** column.

   You can view the request rate and usage rate of CPU and memory resources on each node.
   - CPU request rate = sum (The amount of CPU resources requested by all pods on the node) / Total CPU resources of the node
   - CPU usage rate = sum (The amount of CPU resources used by all pods on the node) / Total CPU resources of the node
   - Memory request rate = (The amount of memory resources requested by all pods on the node) / Total memory resources of the node
   - Memory usage rate = sum (The amount of memory resources used by all pods on the node) / Total memory resources of the node

> **Note**
> - You can adjust the workload of a node based on the resource usage. For more information, see Set node scheduling.
> - When the request or usage rate of a node reaches 100%, pods are not scheduled to the node.

3.1.6.3.7. Node pools

3.1.6.3.7.1. Create a node pool

You can use a node pool to manage a set of nodes for a Container Service cluster. For example, you can manage the labels and taints that are added to the nodes in a node pool. This topic describes how to create a node pool in the Container Service console.

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- The Kubernetes version of the cluster must be later than 1.9.

> **Notice**
> - By default, a cluster can contain at most 100 nodes.
> - Before you add an existing Elastic Compute Service (ECS) instance that is deployed in a virtual private cloud (VPC), make sure that an elastic IP address (EIP) is associated with the ECS instance, or a NAT gateway is created for the VPC. In addition, the node must have access to the Internet. Otherwise, you may fail to add the ECS instance.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, click **Create Node Pool**.
6. In the **Create Node Pool** dialog box, set the parameters.

   For more information, see Create a Kubernetes cluster. The following list describes some of the parameters:

   - Name: the name of the node pool.
   - Public IP: If you select Assign a Public IPv4 Address to Each Node, public IPv4 addresses are assigned to the nodes in the node pool. You can connect to the nodes by using the assigned IP addresses.
   - Quantity: Specify the initial number of nodes in the node pool. If you do not want to add nodes to the node pool, set this parameter to 0.
   - ECS Label: Add labels to the ECS instances.
   - Node Label: Add labels to the nodes in the node pool.
   - CPU Policy: Set the CPU policy.
     - None: indicates that the default CPU affinity is used. This is the default policy.
     - Static: allows pods with specific resource characteristics on the node to be granted with enhanced CPU affinity and exclusivity.
   - Custom Node Name: Specify whether to use custom node names.

A custom node name consists of a prefix, an IP substring, and a suffix.

- Both the prefix and suffix can contain one or more parts that are separated by periods (.).These parts can contain lowercase letters, digits, and hyphens (-), and must start and end with a lowercase letter or digit.
- The IP substring length specifies the number of digits to be truncated from the end of the returned node IP address. Valid values: 5 to 12.

For example, if the node IP address is 192.1xx.x.xx, the prefix is aliyun.com, the IP substring length is 5, and the suffix is test, the node name will be aliyun.com00055test.

7. Click **Confirm Order**.
   On the **Node Pools** page, check the **state** of the node pool. If the node pool is in the **Initializing** state, it indicates that the node pool is being created. After the node pool is created, the **state** of the node pool changes to **Active**.

What's next

After the node pool is created, find the node pool on the **Node Pools** page and click **Details** in the **Actions** column to view the details of the node pool.

3.1.6.3.7.2. Scale out a node pool

You can use a node pool to manage multiple nodes in a Kubernetes cluster as a group. For example, you can centrally manage the labels and taints of the nodes in a node pool. This topic describes how to scale out a node pool in the Container Service console.

Prerequisites

A node pool is created. For more information, see Create a node pool.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, select the node pool that you want to manage and click **Scale Out** in the **Actions** column.
6. In the dialog box that appears, set the number of nodes that you want to add to the node pool.
7. (Optional)In the dialog box that appears, click **Modify Node Pool Settings** to modify the configurations of the node pool.

   For more information, see Create a Kubernetes cluster. The following list describes some of the parameters:

   - Public IP: If you select Assign a Public IPv4 Address to Each Node, public IPv4 addresses are assigned to the nodes in the node pool. You can connect to the nodes by using the assigned IP addresses.
   - ECS Label: Add labels to the ECS instances.
   - Node Label: Add labels to the nodes in the node pool.
   - Taints: Add taints to all worker nodes in the Kubernetes cluster.

   > **Note**
   > - If you select **Synchronize Node Labels and Taints**, the added labels and taints are synchronized to both existing and newly added nodes.
   > - If you select **Set New Nodes to Unschedulable**, the nodes are unschedulable when they are added to the cluster.

8. Click **Submit**.
   On the **Node Pools** page, the **state** of the node pool is **Scaling**. This indicates that the scale-out event is in progress. After the scale-out event is completed, the **state** of the node pool changes to **Active**.

What's next

Click **Details** in the Actions column for the node pool. On the **Nodes** tab, you can check the nodes that are added to the node pool.

3.1.6.3.7.3. Schedule an application pod to a specific node pool

Label is an important concept of Kubernetes. Services, Deployments, and pods are associated with each other by labels. You can configure pod scheduling policies based on node labels. This allows you to schedule pods to nodes that have specific labels. This topic describes how to schedule an application pod to a specific node pool.

Procedure

1. Add a label to the nodes in a node pool.

   Container Service allows you to manage a group of cluster nodes by using a node pool. For example, you can centrally manage the labels and taints of the nodes in a node pool. For more information about how to create a node pool, see Create a node pool.

   i. Log on to the Container Service console.
   ii. In the left-side navigation pane, click **Clusters**.
   iii. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
   iv. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
   v. In the upper-right corner of the **Node Pools** page, click **Create Node Pool**.
   vi. In the Create Node Pool dialog box, click **Show Advanced Options** and click  on the right of **Node Label** to add labels to nodes.

   In this example, the pod: nginx label is added.

   You can also click **Scale Out** on the right side of a node pool to update or add labels for the nodes. If automatic scaling is enabled for a node pool, click **Modify** on the right side of the node pool to update or add labels for the nodes.

2. Configure a scheduling policy for an application pod.

   After the preceding step is completed, the pod: nginx label is added to the nodes in the node pool. You can set the nodeSelector or nodeAffinity field in pod configurations to ensure that an application pod is scheduled to nodes with matching labels in a node pool. Perform the following steps:

   - Set nodeSelector.

     nodeSelector is a field in the spec section of pod configurations. Add the pod: nginx label to nodeSelector. Sample template:

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx-deployment-basic
  labels:
    app: nginx
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      nodeSelector:
        pod: nginx      # After you add the label in this field, the application pod can run on
ly on nodes with this label in the node pool.
      containers:
      - name: nginx
        image: nginx:1.7.9
        ports:
        - containerPort: 80
```

- Set nodeAffinity.

  You can also use nodeAffinity to schedule an application pod based on your requirements. nodeAffinity supports the following scheduling policies:

  ```
  - requiredDuringSchedulingIgnoredDuringExecution
  ```

  If this policy is used, a pod can be scheduled only to a node that meets the match rules. If no node meets the match rules, the system retries until a node that meets the rules is found. IgnoreDuringExecution indicates that if the label of the node where the pod is deployed changes and no longer meets the match rules, the pod continues to run on the node.

  ```
  - requiredDuringSchedulingRequiredDuringExecution
  ```

  If this policy is used, the pod can be scheduled only to a node that meets the match rules. If no node meets the rules, the system retries until a node that meets the rules is found. RequiredDuringExecution indicates that if the label of the node where the pod is deployed changes and no longer meets the match rules, the system redeploys the pod to another node that meets the rules.

  ```
  - preferredDuringSchedulingIgnoredDuringExecution
  ```

  If this policy is used, the pod is preferably scheduled to a node that meets the match rules. If no node meets the rules, the system ignores the rules.

  ```
  - preferredDuringSchedulingRequiredDuringExecution
  ```

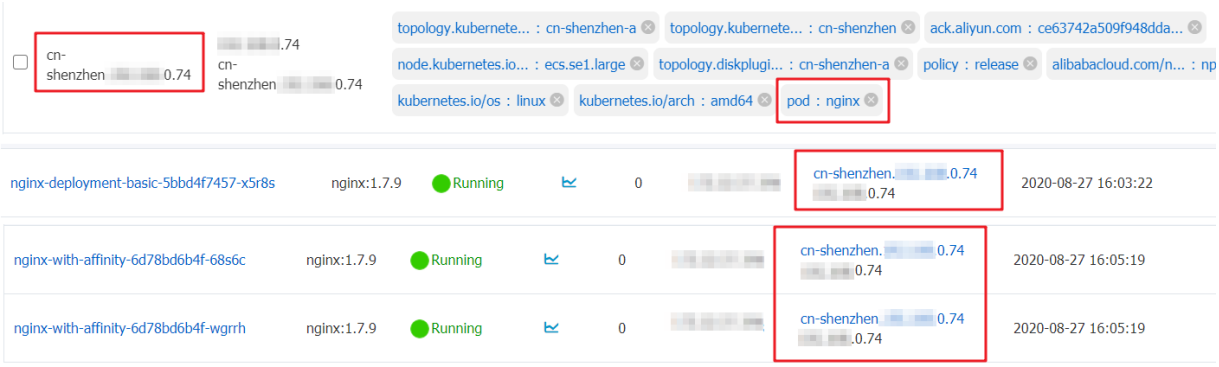
  If this policy is used, the pod is preferably scheduled to a node that meets the match rules. If no node meets the rules, the system ignores the rules. RequiredDuringExecution indicates that if the label of a node where the pod is deployed changes and still meets the match rules, the system reschedules the pod to a node that meets the match rules.

  In the following example, the requiredDuringSchedulingIgnoredDuringExecution policy is used to ensure that the application pod always runs on a node in a specific node pool.

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx-with-affinity
  labels:
    app: nginx-with-affinity
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx-with-affinity
  template:
    metadata:
      labels:
        app: nginx-with-affinity
    spec:
      affinity:
        nodeAffinity:
          requiredDuringSchedulingIgnoredDuringExecution:
            nodeSelectorTerms:
            - matchExpressions:
              - key: pod
                operator: In      # This policy ensures that the application pod can run only o
n a node that has the pod: nginx label.
                values:
                - nginx
      containers:
      - name: nginx-with-affinity
        image: nginx:1.7.9
        ports:
        - containerPort: 80
```

Result

In the preceding example, all application pods are scheduled to the xxx.xxx.0.74 node. This node has the pod: nignx label.



3.1.6.4. Storage

3.1.6.4.1. Overview

In the Container Service console, you can create volumes of other Apsara Stack services, enabling you to create stateful applications and use Apsara Stack disks and OSS to implement persistent storage.

Both static and dynamic volumes are supported. The following table shows how static and dynamic volumes are supported.

Apsara Stack storage	Static volume	Dynamic volume
Apsara Stack disk	You can use a static disk volume through either of the following methods: • Use a volume directly • Use a volume through a PV and PVC	Supported

Apsara Stack NAS	You can use a static NAS volume through either of the following methods:
	• Use a volume through the FlexVolume plug-in
	  ○ Use a volume directly
	  ○ Use a volume through a PV or PVC
	• Use a volume through the Kubernetes NFS driver
		Supported

Apsara Stack OSS	You can use a static OSS volume through either of the following methods: • Use a volume directly • Use a volume through a PV or PVC	Not supported

3.1.6.4.2. Mount disk volumes

You can mount disks as volumes.

Container Service allows you to mount disks as persistent volumes (PVs) in Kubernetes clusters.

Disks can be mounted to Kubernetes clusters as the following volume types:

- Statically provisioned disk volumes

  You can use statically provisioned disk volumes in the following ways:

  - Mount disks as volumes
  - Mount disk volumes by creating a PV and a persistent volume claim (PVC)

- Dynamically provisioned disk volumes

Usage notes

- You can mount a disk only to one pod.
- Before you mount a disk to a pod, you must create the disk and obtain its disk ID.

  The disk must meet the following capacity requirements:

  - If the disk is a basic disk, it must be at least 5 GiB in size.
  - If the disk is an ultra disk, it must be at least 20 GiB in size.
  - If the disk is a standard SSD, it must be at least 20 GiB in size.

- volumeId: the ID of the disk that you want to mount. The value must be the same as those of volumeName and PV Name.
- A disk can be mounted only to a node that is deployed in the same zone as the disk.
- Only pay-as-you-go disks can be mounted. If you change the billing method of an Elastic Compute Service (ECS)

instance in the cluster from pay-as-you-go to subscription, you cannot change the billing method of its disks to subscription. Otherwise, the disks cannot be mounted to the cluster.

Statically provisioned disk volumes

You can mount disks as volumes or by creating PVs and PVCs.

Prerequisites

A disk is created in the ECS console.

- **Mount a disk as a volume**

  Use the following *disk-deploy.yaml* file to create a pod:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: nginx-disk-deploy
spec:
  replicas: 1
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx-flexvolume-disk
        image: nginx
        volumeMounts:
          - name: "d-bp1j17ifxfasvts3tf40"
            mountPath: "/data"
      volumes:
        - name: "d-bp1j17ifxfasvts3tf40"
          flexVolume:
            driver: "alicloud/disk"
            fsType: "ext4"
            options:
              volumeId: "d-bp1j17ifxfasvts3tf40"
```

- **Mount disk volumes by creating a PV and a PVC**

  i. **Create a PV of the disk type**

  You can create a PV of the disk type in the Container Service console or by using a YAML file.

- **Create a PV by using a YAML file**

  Use the following `disk-pv.yaml` file to create a PV:

  > **Note** The PV name must be the same as the disk ID.

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: d-bp1j17ifxfasvts3tf40
  labels:
    failure-domain.beta.kubernetes.io/zone: cn-hangzhou-b
    failure-domain.beta.kubernetes.io/region: cn-hangzhou
spec:
  capacity:
    storage: 20Gi
  storageClassName: disk
  accessModes:
    - ReadWriteOnce
  flexVolume:
    driver: "alicloud/disk"
    fsType: "ext4"
    options:
      volumeId: "d-bp1j17ifxfasvts3tf40"
```

- **Create a PV in the console**

  a. Log on to the Container Service console.

  b. In the left-side navigation pane, click **Clusters**.

  c. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.

  d. In the left-side navigation pane of the details page, choose **Volumes > Persistent Volumes**.

  e. On the **Persistent Volumes** page, click **Create** in the upper-right corner.

  f. In the Create PV dialog box, set the parameters. PV parameters

Parameter	Description
PV Type	In this example, **Cloud Disk** is selected.
Volume Plug-in	Displays the supported storage drivers.
Access Mode	By default, ReadWriteOnce is selected.
Disk ID	Select a disk that is in the same region and zone as your cluster.
File System Type	Select the file system of the disk. Supported file systems are ext4, ext3, xfs, and vfat. Default value: ext4.
Label	Add labels to the PV.

  g. Click **Create**.

ii. **Create a PVC**

Use the following `disk-pvc.yaml` file to create a PVC:

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pvc-disk
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: disk
  resources:
    requests:
      storage: 20Gi
```

iii. **Create a pod**

Use the following *disk-pod.yaml* file to create a pod:

```
apiVersion: v1
kind: Pod
metadata:
  name: "flexvolume-alicloud-example"
spec:
  containers:
    - name: "nginx"
      image: "nginx"
      volumeMounts:
        - name: pvc-disk
          mountPath: "/data"
  volumes:
  - name: pvc-disk
    persistentVolumeClaim:
      claimName: pvc-disk
```

Dynamically provisioned disk volumes

To mount a disk as a dynamically provisioned volume, you must create a StorageClass of the disk type and specify the StorageClass in the storageClassName field of the PVC.

1. **Create a StorageClass**

```
kind: StorageClass
apiVersion: storage.k8s.io/v1beta1
metadata:
  name: alicloud-disk-common-hangzhou-b
provisioner: alicloud/disk
parameters:
  type: cloud_ssd
  regionid: cn-hangzhou
  zoneid: cn-hangzhou-b
```

Required parameters:

- provisioner: Set this parameter to alicloud/disk. This indicates that the Provisioner plug-in is used to create the StorageClass.
- type: Specify the disk type. Valid values: cloud, cloud_efficiency, cloud_ssd, and available. If you set this parameter to available, the system attempts to create a disk in the following order: ultra disk, standard SSD, and basic disk. The system stops trying until a disk is created.
- regionid: Specify the region of the disk.
- zoneid: Specify the zone of the disk.

2. **Create a PVC**

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: disk-common
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: alicloud-disk-common-hangzhou-b
  resources:
    requests:
      storage: 20Gi
---
kind: Pod
apiVersion: v1
metadata:
  name: disk-pod-common
spec:
  containers:
  - name: disk-pod
    image: nginx
    volumeMounts:
      - name: disk-pvc
        mountPath: "/mnt"
  restartPolicy: "Never"
  volumes:
    - name: disk-pvc
      persistentVolumeClaim:
        claimName: disk-common
```

Default options

By default, Kubernetes clusters support the following types of StorageClass:

- alicloud-disk-common: basic disk.
- alicloud-disk-efficiency: ultra disk.
- alicloud-disk-ssd: standard SSD.
- alicloud-disk-available: This option ensures high availability. The system attempts to create an ultra disk first. If no ultra disk is available in the specified zone, the system attempts to create a standard SSD. If no standard SSD is available, the system attempts to create a basic disk.

3. **Create a multi-instance StatefulSet by using a disk**

   We recommend that you use the volumeClaimTemplates parameter. This parameter allows the system to dynamically create PVCs and PVs. PVCs are associated with corresponding PVs.

```
apiVersion: v1
kind: Service
metadata:
  name: nginx
  labels:
    app: nginx
spec:
  ports:
  - port: 80
    name: web
  clusterIP: None
  selector:
    app: nginx
---
apiVersion: apps/v1beta2
kind: StatefulSet
metadata:
  name: web
spec:
  selector:
    matchLabels:
      app: nginx
  serviceName: "nginx"
  replicas: 2
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx
        ports:
        - containerPort: 80
          name: web
        volumeMounts:
        - name: disk-common
          mountPath: /data
  volumeClaimTemplates:
  - metadata:
      name: disk-common
    spec:
      accessModes: [ "ReadWriteOnce" ]
      storageClassName: "alicloud-disk-common"
      resources:
        requests:
          storage: 10Gi
```

3.1.6.4.3. Mount NAS volumes

Container Service allows you to mount Apsara File Storage NAS (NAS) file systems as persistent volumes (PVs) in Container Service clusters.

NAS file systems can be mounted to Kubernetes clusters as the following volume types:

- Statically provisioned NAS volumes

  You can statically provision NAS volumes in the following ways:

- Use the FlexVolume plug-in
  - Directly mount NAS file systems as volumes
  - Mount NAS file systems by creating a PV and a persistent volume claim (PVC)
- Use the Network File System (NFS) driver

- Dynamically provisioned NAS volumes

Prerequisites

A NAS file system is created in the NAS console and a mount target is added. The mount target is used to mount the NAS file system to the Kubernetes cluster. The NAS file system and your cluster are deployed in the same virtual private cloud (VPC).

Statically provisioned NAS volumes

You can use the FlexVolume plug-in provided by Alibaba Cloud or the NFS driver provided by Kubernetes to mount NAS file systems.

Use the FlexVolume plug-in

You can use the FlexVolume plug-in to directly mount a NAS file system as a volume. You can also mount a NAS file system by creating a PV and a PVC.

> **Note**
> - NAS is a shared storage service. You can mount a NAS file system to multiple pods.
> - server: the mount target of the NAS volume.
> - path: the directory to which the NAS volume is mount. You can specify a subdirectory. If the specified subdirectory does not exist, the system automatically creates the subdirectory.
> - vers: the version of the NFS protocol. Version 4.0 is supported.
> - mode: the access permissions on the directory of the NAS file system. If the root directory of the NAS file system is specified, you cannot modify the access permissions. If the NAS file system stores a large amount of data, the mounting operation may be time-consuming or even fail. Therefore, we recommend that you do not set the mode parameter.

Mount a NAS file system as a volume

Use the following `nas-deploy.yaml` file to create a pod:

```
apiVersion: v1
kind: Pod
metadata:
  name: "flexvolume-nas-example"
spec:
  containers:
    - name: "nginx"
      image: "nginx"
      volumeMounts:
        - name: "nas1"
          mountPath: "/data"
  volumes:
    - name: "nas1"
      flexVolume:
        driver: "alicloud/nas"
        options:
          server: "0cd8b4a576-grs79.cn-hangzhou.nas.aliyuncs.com"
          path: "/k8s"
          vers: "4.0"
```

Mount a NAS file system by creating a PV and a PVC

Step 1: Create a PV

You can create a PV in the Container Service console or by using a YAML file.

- Create a PV by using a YAML file.

  Use the following `nas-pv.yaml` file to create a PV:

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv-nas
spec:
  capacity:
    storage: 5Gi
  storageClassName: nas
  accessModes:
    - ReadWriteMany
  flexVolume:
    driver: "alicloud/nas"
    options:
      server: "0cd8b4a576-uih75.cn-hangzhou.nas.aliyuncs.com"
      path: "/k8s"
      vers: "4.0"
```

- Create a PV in the console
  i. Log on to the Container Service console.
  ii. In the left-side navigation pane, click **Clusters**.
  iii. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
  iv. In the left-side navigation pane of the details page, choose **Volumes > Persistent Volumes**.
  v. On the **Persistent Volumes** page, click **Create** in the upper-right corner.
  vi. In the Create PV dialog box, set the parameters.

Parameter	Description
PV Type	Select **NAS**.
Volume Name	Enter a name for the PV. The name must be unique in the cluster.
Volume Plug-in	By default, CSI is selected.
Capacity	Specify the capacity of the PV. The capacity of the PV cannot exceed the capacity of the disk.
Access Mode	By default, ReadWriteMany is selected.
Mount Target Domain Name	The domain name of the mount target that is used to mount the NAS file system to the cluster.

Parameter	Description
Subdirectory	Enter a subdirectory in the NAS file system. The subdirectory must start with a forward slash (/). If you set this parameter, the PV is mounted to the specified subdirectory. ■ If the specified subdirectory does not exist, the system automatically creates the subdirectory in the NAS file system. ■ If you do not set this parameter, the PV is mounted to the root directory of the NAS file system.
Version	The version of the NFS protocol. Version 3 and 4.0 are supported. By default, version 3 is used. We recommend that you use version 3.
Label	Add labels to the PV.

vii. Click **Create**.

Step 2: Create a PVC

Use the following *nas-pvc.yaml* file to create a PVC:

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pvc-nas
spec:
  accessModes:
    - ReadWriteMany
  storageClassName: nas
  resources:
    requests:
      storage: 5Gi
```

Step 3: Create a pod

Use the following *nas-pod.yaml* file to create a pod:

```
apiVersion: v1
kind: Pod
metadata:
  name: "flexvolume-nas-example"
spec:
  containers:
    - name: "nginx"
      image: "nginx"
      volumeMounts:
          - name: pvc-nas
            mountPath: "/data"
  volumes:
  - name: pvc-nas
    persistentVolumeClaim:
        claimName: pvc-nas
```

Use the NFS driver

Step 1: Create a NAS file system

Log on to the NAS console. For more information about how to log on to the NAS console, see the *Create a NAS file system* chapter of the *NAS User Guide*.

> **Note** The NAS file system that you want to create and the Kubernetes cluster must be deployed in the same region.

In this example, the following mount target is used: `055f84ad83-ixxxx.cn-hangzhou.nas.aliyuncs.com`.

Step 2: Create a PV

You can create a PV in the console or by using a YAML file.

- **Create a PV by using a YAML template**

  Use the following *nas-pv.yaml* file to create a PV.

  Run the following command to create a PV that uses the NAS file system:

```
root@master # cat << EOF |kubectl apply -f -
apiVersion: v1
kind: PersistentVolume
metadata:
   name: nas
spec:
   capacity:
     storage: 8Gi
   accessModes:
     - ReadWriteMany
   persistentVolumeReclaimPolicy: Retain
   nfs:
     path: /
     server: 055f84ad83-ixxxx.cn-hangzhou.nas.aliyuncs.com
EOF
```

- **Create a PV by using the console**

  For more information, see Mount a NAS file system by creating a PV and a PVC.

Step 3: Create a PVC

Create a PVC and associate the PVC with the PV that is created in Step 2.

```
root@master # cat << EOF | kubectl apply -f -
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
     name: nasclaim
spec:
     accessModes:
       - ReadWriteMany
     resources:
       requests:
         storage: 8Gi
EOF
```

Step 4: Create a pod

Create an application to use the PV.

```
root@master # cat << EOF |kubectl apply -f -
apiVersion: v1
kind: Pod
metadata:
     name: mypod
spec:
     containers:
       - name: myfrontend
         image: registry.aliyuncs.com/spacexnice/netdia:latest
         volumeMounts:
         - mountPath: "/var/www/html"
           name: mypd
     volumes:
       - name: mypd
         persistentVolumeClaim:
           claimName: nasclaim
EOF
```

The NAS file system is mounted to the application that runs in the pod.

Dynamically provisioned NAS volumes

If you want to dynamically provision NAS volumes, you must install a driver plug-in and configure a NAS mount target.

Install the plug-in

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: alicloud-nas
provisioner: alicloud/nas
---
apiVersion: v1
kind: ServiceAccount
metadata:
  name: alicloud-nas-controller
  namespace: kube-system
---
kind: ClusterRoleBinding
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  name: run-alicloud-nas-controller
subjects:
  - kind: ServiceAccount
    name: alicloud-nas-controller
    namespace: kube-system
roleRef:
  kind: ClusterRole
  name: alicloud-disk-controller-runner
  apiGroup: rbac.authorization.k8s.io
---
kind: Deployment
apiVersion: extensions/v1beta1
metadata:
  name: alicloud-nas-controller
  namespace: kube-system
spec:
  replicas: 1
  strategy:
```

```
    type: Recreate
  template:
    metadata:
      labels:
        app: alicloud-nas-controller
    spec:
      tolerations:
      - effect: NoSchedule
        operator: Exists
        key: node-role.kubernetes.io/master
      - effect: NoSchedule
        operator: Exists
        key: node.cloudprovider.kubernetes.io/uninitialized
      nodeSelector:
         node-role.kubernetes.io/master: ""
      serviceAccount: alicloud-nas-controller
      containers:
        - name: alicloud-nas-controller
          image: registry.cn-hangzhou.aliyuncs.com/acs/alicloud-nas-controller:v1.8.4
          volumeMounts:
          - mountPath: /persistentvolumes
            name: nfs-client-root
          env:
            - name: PROVISIONER_NAME
              value: alicloud/nas
            - name: NFS_SERVER
              value: 0cd8b4a576-mmi32.cn-hangzhou.nas.aliyuncs.com
            - name: NFS_PATH
              value: /
      volumes:
        - name: nfs-client-root
          nfs:
            server: 0cd8b4a576-mmi32.cn-hangzhou.nas.aliyuncs.com
            path: /
```

Dynamically provision a NAS volume

```
apiVersion: apps/v1beta1
kind: StatefulSet
metadata:
  name: web
spec:
  serviceName: "nginx"
  replicas: 2
  volumeClaimTemplates:
  - metadata:
      name: html
    spec:
      accessModes:
        - ReadWriteOnce
      storageClassName: alicloud-nas
      resources:
        requests:
          storage: 2Gi
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:alpine
        volumeMounts:
        - mountPath: "/usr/share/nginx/html/"
          name: html
```

3.1.6.4.4. Mount OSS volumes

You can mount Object Storage Service (OSS) buckets as volumes in Kubernetes clusters.

You can mount OSS buckets in the following ways:

- Mount an OSS bucket as a volume.
- Mount an OSS bucket by creating a PV and a PVC.

Prerequisites

To mount an OSS bucket as a statically provisioned volume, you must create an OSS bucket in the OSS console.

Background

- OSS buckets can be mounted only as statically provisioned volumes.
- An OSS bucket can be shared by multiple pods.
- bucket: Only buckets can be mounted to a Kubernetes cluster. The subdirectories or files in a bucket cannot be mounted to a Kubernetes cluster.
- url: specifies the endpoint of the OSS bucket. The endpoint is the domain name that is used to mount the OSS bucket to the Kubernetes cluster.
- akId: specifies your AccessKey ID.
- akSecret: specifies your AccessKey secret.
- otherOpts: the custom parameters that are used to mount the OSS bucket. The parameters must be in the following format: `-o *** -o ***` .

> **Note** To mount an OSS bucket as a volume, you must create a Secret with your AccessKey pair when you deploy the flexvolume Service.

Mount an OSS bucket as a statically provisioned volume

- **Mount an OSS bucket as a volume**

  Use the following *oss-deploy.yaml* file to create a pod:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: nginx-oss-deploy
spec:
  replicas: 1
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx-flexvolume-oss
        image: nginx
        volumeMounts:
          - name: "oss1"
            mountPath: "/data"
      volumes:
        - name: "oss1"
          flexVolume:
            driver: "alicloud/oss"
            options:
              bucket: "docker"
              url: "oss-cn-hangzhou.aliyuncs.com"
              akId: ***
              akSecret: ***
              otherOpts: "-o max_stat_cache_size=0 -o allow_other"
```

- **Create a static PV and a PVC**

  i. **Create a PV**

  You can create a persistent volume (PV) in the Container Service console or by using a YAML file.

- **Create a PV by using a YAML file**

  Use the following *oss-pv.yaml* file to create a PV:

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv-oss
spec:
  capacity:
    storage: 5Gi
  accessModes:
    - ReadWriteMany
  storageClassName: oss
  flexVolume:
    driver: "alicloud/oss"
    options:
      bucket: "docker"
      url: "oss-cn-hangzhou.aliyuncs.com"
      akId: ***
      akSecret: ***
      otherOpts: "-o max_stat_cache_size=0 -o allow_other"
```

- **Create a PV in the Container Service console**

  a. Log on to the Container Service console.

  b. In the left-side navigation pane, click **Clusters**.

  c. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.

  d. In the left-side navigation pane of the details page, choose **Volumes > Persistent Volumes**.

  e. On the **Persistent Volumes** page, click **Create** in the upper-right corner.

f. In the Create PV dialog box, set the parameters.

Parameter	Description
PV Type	In this example, OSS is selected.
Name	Enter a name for the PV. The name must be unique in the cluster. In this example, pv-oss is used.
Volume Plug-in	By default, CSI is selected.
Capacity	Specify the capacity of the PV.
Access Mode	By default, ReadWriteMany is selected.
AccessKey ID and **AccessKey Secret**	The AccessKey pair that is required to access OSS buckets. To obtain your AccessKey pair, go to the Apsara Uni-manager Management Console, choose **Enterprise > Organizations**, click  on the right side of the organization. Then, click **AccessKey** and copy the AccessKey pair.
Optional Parameters	Enter custom parameters in the format of `-o ** * -o ***`.
Bucket ID	Enter the name of the OSS bucket that you want to mount. Click **Select Bucket**. In the dialog box that appears, select the OSS bucket that you want to mount and click **Select**.
Endpoint	**Internal Endpoint** is recommended.
Label	Add labels to the PV.

g. Click **Create**.

ii. **Create a PVC**

Use the following *oss-pvc.yaml* file to create a persistent volume claim (PVC):

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pvc-oss
spec:
  storageClassName: oss
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 5Gi
```

iii. **Create a pod**

Use the following *oss-pod.yaml* file to create a pod:

```
apiVersion: v1
kind: Pod
metadata:
  name: "flexvolume-oss-example"
spec:
  containers:
    - name: "nginx"
      image: "nginx"
      volumeMounts:
          - name: pvc-oss
            mountPath: "/data"
  volumes:
  - name: pvc-oss
    persistentVolumeClaim:
        claimName: pvc-oss
```

Can I mount an OSS bucket as a dynamically provisioned volume?

No. Dynamically provisioned OSS volumes are not supported.

3.1.6.4.5. Create a PVC

This topic describes how to create a persistent volume claim (PVC).

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- A persistent volume (PV) is created. In this example, a PV created from a disk is used. For more information, see Use Apsara Stack disks.

  By default, PVCs are associated with PVs that have the alicloud-pvname label. This label is added to all PVs that are created in the Container Service console. If a PV does not have this label, manually add the label to the PV before you can associate the PV with a PVC.

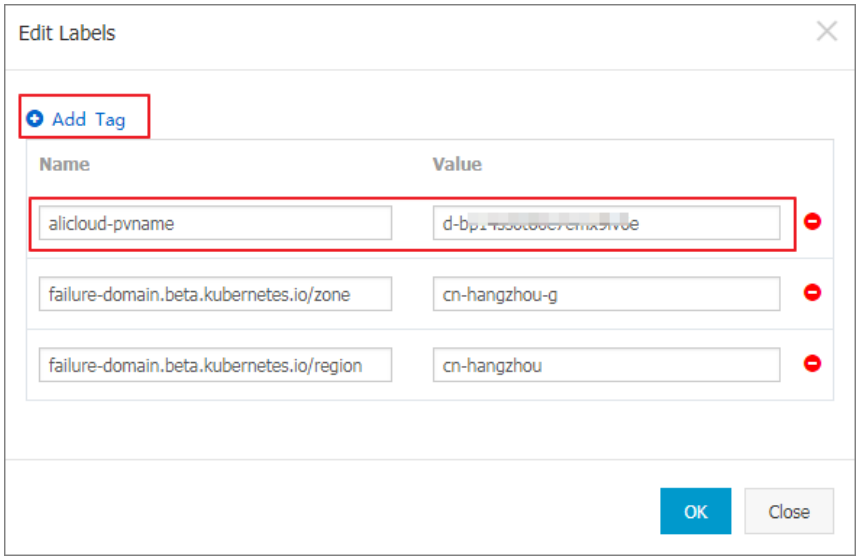
Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Volumes > Persistent Volume Claims**.
5. In the upper-right corner of the **Persistent Volume Claims** page, click **Create**.
6. In the **Create PVC** dialog box, set the parameters and click **Create**.
   - **PVC Type**: The PVC and PV must be of the same type. You can select Cloud Disk, NAS, and OSS.
   - **Name**: Enter the name of the PVC.
   - **Allocation Mode**: **Use StorageClass**, **Existing Volumes**, and **Create Volume** are supported. In this example, **Use StorageClass** or **Existing Volumes** is selected.
   - **Existing Storage Class**: Click **Select**. In the Select Storage Class dialog box, find the Storage Class that you want to use and click **Select** in the Actions column. This parameter is required only when you set Allocation Mode to **Use StorageClass**.
   - **Existing Volumes**: Click **Select PV**. In the Select PV dialog box, find the PV that you want to use and click **Select** in the Actions column. This parameter is required only when you set Allocation Mode to **Existing Volumes**.
   - **Capacity**: Specifies the capacity used by the PVC. The value cannot be larger than the total capacity of the associated PV.

- **Access Mode**: The default value is ReadWriteOnce. This parameter is required only when you set Allocation Mode to **Use StorageClass**.

> **Note** If your cluster has a PV that is not used, but you cannot find it in the **Select PV** dialog box, the reason may be that the PV does not have the alicloud-pvname label.

If you cannot find available PVs, click **Persistent Volumes** in the left-side navigation pane. On the Persistent Volumes page, find the PV that you want to use, click **Manage Labels** in the Actions column, and then add a label to the PV. On the Manage Labels dialog box, set the key of the label to alicloud-pvname and the value to the name of the PV. If the PV is created from a disk, the disk ID is used as the name of the PV.



7. Go to the Persistent Volume Claims page. You can find the newly created PVC in the list.

3.1.6.4.6. Use PVCs

You can use persistent volume claims (PVCs) to mount volumes for applications.

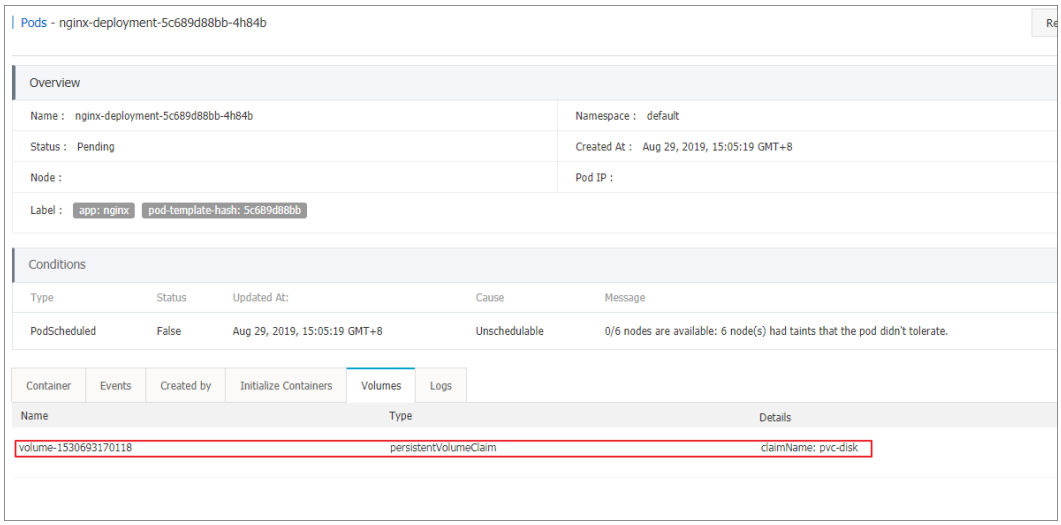
Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- A PVC is created. In this example, a PVC named pvc-disk is created to mount a disk volume. For more information, see Create PVCs.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, click **Create from Image** in the upper-right corner.
6. On the **Basic Information** wizard page, specify the application name, number of replicas, type, and labels, select whether to synchronize the time zone from nodes to containers, and then click **Next**.
7. On the **Container** wizard page, select an image and configure volumes of the cloud storage type. You can select Cloud Disk, Apsara File Storage NAS (NAS), and Object Service Storage (OSS). In this example, select the PVC named pvc-disk and click **Next**. For more information, see Container configurations.
8. On the **Advanced** wizard page, create a Service for the test-nginx application and click **Create**.

9. On the **Complete** wizard page, click View Details to view detailed information about the application. You are redirected to the details page of the test-nginx application.
10. On the **Pods** tab, you can find the pods to which the application belongs. Select a pod and click **View Details** to view detailed information about the pod.
11. On the details page of the pod, click the **Volumes** tab. You can find that the pod is associated with the pvc-disk PVC.



3.1.6.5. Network management

3.1.6.5.1. Set access control for pods

This topic describes how to use network policies to control access between pods.

Prerequisites

You have created a Kubernetes cluster and selected the **Terwaynetwork plug-in**. For more information, see Create a Kubernetes cluster.

Context

You can declare network policies to control access between pods and thus prevent applications from interfering each other.

Procedure

For more information about standard Kubernetes network policies, see Network policies.

1. Create a pod that runs as a server and attach `label run=nginx` to the pod. For more information, see Create an application from an orchestration template.

   The sample YAML file is as follows:

```
apiVersion: v1
kind: Pod
metadata:
  name: server
  labels:
    run: nginx
spec:
  containers:
    - name: nginx
      image: registry.acs.intranet.env22.com/nginx:1.8
```

2. Create a network policy. For more information, see Create an application from an orchestration template.

   The sample YAML file is as follows:

```
kind: NetworkPolicy
apiVersion: networking.k8s.io/v1
metadata:
  name: access-nginx
spec:
  podSelector:
    matchLabels:
      run: nginx # Apply the network policy to pods with the run=nginx label
  ingress:
  - from:
    - podSelector:
        matchLabels:
          access: "true" # Only pods with the access=true label are accessible
```

3. Use the *client.yaml* and *client-label* files to create two pods that run as clients.

   One pod has the required label and the other does not.

   i. Create the *client.yaml* and *client-label* files with the following contents respectively.

```
# This pod has no label
apiVersion: v1
kind: Pod
metadata:
  name: client
spec:
  containers:
    - name: busybox
      image: registry.acs.intranet.env22.com/acs/busybox
      command: ["sh", "-c", "sleep 200000"]
```

```
# This pod has the label
apiVersion: v1
kind: Pod
metadata:
  name: client-label
  labels:
    access: "true"
spec:
  containers:
    - name: busybox
      image: registry.acs.intranet.env22.com/acs/busybox
      command: ["sh", "-c", "sleep 200000"]
```

ii. Run the following commands to create these pods:

```
kubectl apply -f client.yaml
kubectl apply -f client-label.yaml
```

You can see that only the pod with the required label can access the server.

3.1.6.5.2. Set bandwidth limits for pods

This topic describes how to limit the bandwidth of inbound and outbound traffic that flows through a pod.

Prerequisites

You have created a Kubernetes cluster and selected the **Terwaynetwork plug-in**. For more information, see Create a Kubernetes cluster.

Context

Throttling pods helps prevent performance degradation of the host or other workloads when certain pods occupy excessive resources.

Method

You can use the `k8s.aliyun.com/ingress-bandwidth` and `k8s.aliyun.com/egress-bandwidth` annotations for pod throttling.

- `k8s.aliyun.com/ingress-bandwidth`: limits the pod inbound bandwidth.
- `k8s.aliyun.com/egress-bandwidth`: limits the pod outbound bandwidth.
- The bandwidth limit is measured in m and k, which represent Mbit/s and Kbit/s respectively.

Procedure

1. Create a pod that runs as a server in the console. For more information, see Create an application from an orchestration template.

   The sample YAML file is as follows:

   ```
   apiVersion: v1
   kind: Pod
   metadata:
     name: server
     annotations:
       k8s.aliyun.com/ingress-bandwidth: 10m # Set the inbound bandwidth limit to 10 Mbit/s
       k8s.aliyun.com/egress-bandwidth: 10m
   spec:
     containers:
       - name: nginx
         image: registry.acs.intranet.env22.com/nginx:1.8
   ```

2. Run the kubectl exec command to connect to the pod. To verify that pod throttling is effective, run the following commands to create a file on the pod. Assume that the IP address of the pod created in step 1 is 172.16.XX.XX.

   ```
   cd /usr/share/nginx/html
   dd if=/dev/zero of=bigfile bs=1M count=1000
   ```

3. Use the *client-deploy.yaml* file to create a pod that runs as a client.

i. Create the *client-deploy.yaml* file with the following content:

```
apiVersion: v1
kind: Pod
metadata:
  name: client
  annotations:
     k8s.aliyun.com/ingress-bandwidth: 10m # Set the inbound bandwidth limit to 10 Mbit/s
     k8s.aliyun.com/egress-bandwidth: 10m
spec:
  containers:
    - name: busybox
      image: registry.acs.intranet.env22.com/acs/netdia
      command: ["sh", "-c", "sleep 200000"]
```

ii. Run the following command to create the pod:

```
kubectl apply -f client-deploy.yaml
```

4. Run the following command to check whether bandwidth is limited:

```
kubectl exec -it client sh
```

3.1.6.5.3. Work with Terway

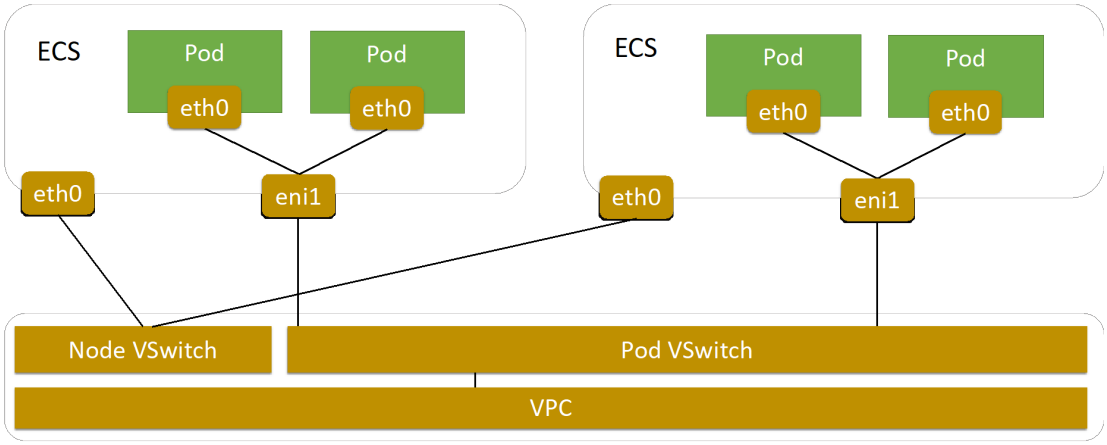
Terway is an open source Container Network Interface (CNI) plug-in developed by Alibaba Cloud. Terway works with Virtual Private Cloud (VPC) and allows you to use standard Kubernetes network policies to regulate how containers communicate with each other. You can use Terway to enable internal communication within a Kubernetes cluster. This topic describes how to use Terway.

Context

Terway is a network plug-in developed by Alibaba Cloud for Container Service. Terway allows you to set up networks for pods by associating Alibaba Cloud elastic network interfaces (ENIs) with the pods. Terway also allows you to use standard Kubernetes network policies to regulate how containers communicate with each other. In addition, Terway is compatible with Calico network polices.

In a cluster that has Terway installed, each pod has a separate network stack and is assigned a separate IP address. Pods on the same Elastic Compute Service (ECS) instance communicate with each other by forwarding packets inside the ECS instance. Pods on different ECS instances communicate with each other through ENIs in the VPC where the ECS instances are deployed. This improves communication efficiency because no tunneling technologies such as Virtual Extensible Local Area Network (VXLAN) are required to encapsulate packets.

How Terway works



Comparison between Flannel and Terway

When you create a Kubernetes cluster, you can choose one of the following network plug-ins:

- Terway: a network plug-in developed by Alibaba Cloud for Container Service. Terway allows you to assign ENIs to containers and use standard Kubernetes network policies to regulate how containers communicate with each other. Terway also supports bandwidth throttling on individual containers. Terway uses flexible IP Address Management (IPAM) policies to allocate IP addresses to containers. This avoids IP address waste. If you do not want to use network policies, you can select Flannel as the network plug-in. Otherwise, we recommend that you select Terway.
- Flannel: an open source CNI plug-in, which is simple and stable. You can use Flannel with VPC of Alibaba Cloud. This allows your clusters and containers to run in high-performance and stable networks. However, Flannel provides only basic features. It does not support standard Kubernetes network policies. For more information, see Flannel.

Item	Terway	Flannel
Performance	The IP address of each pod in a Kubernetes cluster is assigned from the CIDR block of the VPC where the cluster is deployed. Therefore, you do not need to use the Network Address Translation (NAT) service to translate IP addresses. This avoids IP address waste. In addition, each pod in the cluster can use an exclusive ENI.	Flannel works with VPC of Alibaba Cloud. The CIDR block of pods that you specify must be different from that of the VPC where the cluster is deployed. Therefore, the NAT service is required and some IP addresses may be wasted.
Security	Terway supports network policies.	Flannel does not support network policies.
IP address management	Terway allows you to assign IP addresses on demand. You do not have to assign CIDR blocks by node. This avoids IP address waste.	You can only assign CIDR blocks by node. In large-scale clusters, a great number of IP addresses may be wasted.
SLB	Server Load Balancer (SLB) directly forwards network traffic to pods. You can upgrade the pods without service interruption.	SLB forwards network traffic to the NodePort Service. Then, the NodePort Service routes the network traffic to pods.

Considerations

- To use the Terway plug-in, we recommend that you use ECS instances of higher specifications and newer types, such as ECS instances that belong to the g5 or g6 instance family with at least eight CPU cores. For more information, see the *Instance families* chapter of *ECS User Guide*.
- The maximum number of pods that each node supports is based on the number of ENIs assigned to the node.
  - Maximum number of pods supported by each shared ENI = (Number of ENIs supported by each ECS instance - 1) × Number of private IP addresses supported by each ENI
  - Maximum number of pods supported by each exclusive ENI = Number of ENIs supported by each ECS instance - 1

Step 1: Plan CIDR blocks

When you create a Kubernetes cluster, you must specify a VPC, vSwitches, the CIDR block of pods, and the CIDR block of Services. If you want to install the Terway plug-in, you must first create a VPC and two vSwitches in the VPC. The two vSwitches must be created in the same zone.

You can assign the following CIDR blocks for a cluster that uses Terway.

- VPC CIDR block: 192.168.0.0/16
- vSwitch CIDR block: 192.168.0.0/19

- CIDR block of pod vSwitch: 192.168.32.0/19
- Service CIDR block:172.21.0.0/20

> **Note**
> - IP addresses within the CIDR block of the vSwitch are assigned to nodes.
> - IP addresses within the CIDR block of the pod vSwitch are assigned to pods.

The following example describes how to create a VPC and two vSwitches. The CIDR blocks in the preceding section are assigned in this example.

1. Log on to the VPC console.

   > **Note** For more information, see the *Log on to the VPC console* chapter of *VPC*.

2. Create a VPC.
   i. In the top navigation bar, select the region where you want to create the VPC.
   ii. On the **VPCs** page, click **Create VPC**.

iii. On the **Create VPC** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the VPC belongs.
Resource Set	Select the resource set to which the VPC belongs.
Region	Select the region where you want to deploy the VPC.
Sharing Scope	Select the sharing scope of the VPC. ■ **Current Resource Set**: Only the administrator of the current resource set can use the VPC to create resources. ■ **Current Organization and Subordinate Organization**: Only the administrators of the current organization and its subordinate organization can create resources for the shared VPC. ■ **Current Organization**: Only the administrator of the current organization can use create resources for the shared VPC.
VPC Name	Enter a name for the VPC. In this example, vpc_192_168_0_0_16 is used. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`.
IPv4 CIDR Block	Select an IPv4 CIDR block for the VPC. In this example, 192.168.0.0/16 is specified. **Note** After a VPC is created, the IPv4 CIDR block of the VPC cannot be modified.
IPv6 CIDR Block	Specify whether to assign an IPv6 CIDR block. ■ **Do Not Assign**: The system does not assign an IPv6 CIDR block to the VPC. ■ **Assign**: An IPv6 CIDR block is automatically assigned to the VPC. If you set this parameter to Assign, the system automatically creates a free IPv6 gateway for this VPC, and assigns an IPv6 CIDR block with the subnet mask /56, such as 2xx1: db8::/56. By default, IPv6 addresses can be used to communicate within only private networks. If you want to allow an instance assigned with an IPv6 address to access the Internet or be accessed by IPv6 clients over the Internet, you must purchase an Internet bandwidth plan for the IPv6 address. For more information, see the **Activate IPv6 Internet bandwidth** section of the **Manage IPv6 Internet bandwidth** topic of the *IPv6 gateway user guide*.
Description	Enter a description for the VPC. The description must be 2 to 256 characters in length, and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.

3. Create vSwitches.
   i. In the left-side navigation pane, click **VSwitches**.
   ii. Select the region of the VPC in which you want to create a vSwitch.
   iii. On the **VSwitches** page, click **Create VSwitch**.
   iv. On the **Create vSwitch** page, set the following parameters and click **Submit**.

> **Note** Make sure that the two vSwitches are in the same zone.

Parameter	Description
Organization	Select the organization to which the vSwitch belongs.
Resource Set	Select the resource set to which the vSwitch belongs.
Region	Select the region where you want to deploy the vSwitch.
Zone	Select the zone to which the vSwitch belongs. In a VPC, each vSwitch can be deployed in only one zone. You cannot deploy a vSwitch across zones. However, you can deploy cloud resources in vSwitches that belong to different zones to achieve cross-zone disaster recovery. **Note** A cloud instance can be deployed in only one vSwitch.
Sharing Scope	Select the sharing scope of the vSwitch. ■ **Current Resource Set**: Only the administrator of the current resource set can create resources in the shared vSwitch. ■ **Current Organization and Subordinate Organization**: Only the administrators of the current organization and its subordinate organizations can create resources in the shared vSwitch. ■ **Current Organization**: Only the administrator of the current organization can create resources in the shared vSwitch.
VPC	Select the VPC for which you want to create the vSwitch.
Dedicated for Out-of-cloud Physical Machines	Specify whether the vSwitch to be created is dedicated for bare-metal instances. For more information about bare-metal instances, see the **VPC bare-metal instance features** topic in *BMS User Guide*.
vSwitch Name	Enter a name for the vSwitch. In this example, node_switch_192_168_0_0_19 is used. The name must be 2 to 128 characters in length and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.
IPv4 CIDR Block	Specify an IPv4 CIDR block for the vSwitch. In this example, 192.168.0.0/19 is used.
IPv6 CIDR Block	Specify an IPv6 CIDR block for the vSwitch. ■ You must check whether IPv6 is enabled for the specified VPC. If IPv6 is disabled, you cannot assign an IPv6 CIDR block to the vSwitch. ■ If IPv6 is enabled, you can enter a decimal number ranging from 0 to 255 to define the last 8 bits of the IPv6 CIDR block of the vSwitch. For example, if the IPv6 CIDR block of the VPC is 2xx1:db8::/64, specify 255 to define the last 8 bits of the IPv6 CIDR block. In this case, the IPv6 CIDR block of the vSwitch is 2xx1:db8:ff::/64. ff is the hexadecimal value of 255.

Parameter	Description
Description	Enter a description for the vSwitch. The description must be 2 to 256 characters in length and can contain letters, digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). The name must start with a letter and cannot start with `http://` or `https://`.

4. Repeat Step to create a pod vSwitch. Set the name of the pod vSwitch to **pod_switch_192_168_32_0_19** and IPv4 CIDR Block to **192.168.32.0/19**.

Step 2: Set up networks for a cluster that uses Terway

To install Terway in a cluster and set up networks for the cluster, set the following parameters.

> **Note** A Kubernetes cluster is used as an example to show how to set up networks for a cluster that uses Terway as the network plug-in. For more information about how to create a Kubernetes cluster, see Create a Kubernetes cluster.

- **VPC**: Select the VPC created in Step 1: Plan CIDR blocks.
- **VSwitch**: Select the vSwitch created in Step 1: Plan CIDR blocks.
- **Network Plug-in**: Select **Terway**.
- **Pod VSwitch**: Select the pod vSwitch created in Step 1: Plan CIDR blocks.
- **Service CIDR**: Use the default settings.

3.1.6.6. Namespaces

3.1.6.6.1. Create a namespace

You can create a namespace in the Container Service console.

Prerequisites

A Kubernetes cluster is created.

Context

In a Kubernetes cluster, you can use namespaces to create multiple virtual spaces. When a large number of users share a cluster, you can use namespaces to divide different workspaces and allocate cluster resources to different tasks. You can also use resource quotas to allocate resources to each namespace.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, click **Namespaces and Quotas**.
5. On the **Namespace** page, click **Create** in the upper-right corner.
6. In the **Create Namespace** dialog box, configure the namespace.

   Create a namespace

Parameter	Description

Parameter	Description
Name	Enter a name for the namespace. In this example, test is entered. The name must be 1 to 63 characters in length and can contain digits, letters, and hyphens (-). It must start and end with a letter or digit.
Label	**Label**: Add one or more labels to the namespace. Labels are used to identify namespaces. For example, you can label a namespace as one used in the test environment. To add a label, enter a key and a value and click **Add** in the Actions column.

7. Click **OK**.
   You can find the namespace that you created on the Namespace page.

3.1.6.6.2. Set resource quotas and limits

You can set resource quotas and limits for a namespace in the Container Service console.

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- A sample namespace named **test** is created. For more information, see Create a namespace.
- You are connected to a master node of the cluster. For more information, see Connect to a Kubernetes cluster through kubectl.

Context

By default, a running pod uses the CPU and memory resources of a node without limit. This means that pods can compete for computing resources of a cluster. As a result, the pods in a namespace may exhaust all of the computing resources.

Namespaces can be used as virtual clusters to serve multiple purposes. We recommend that you set resource quotas for namespaces.

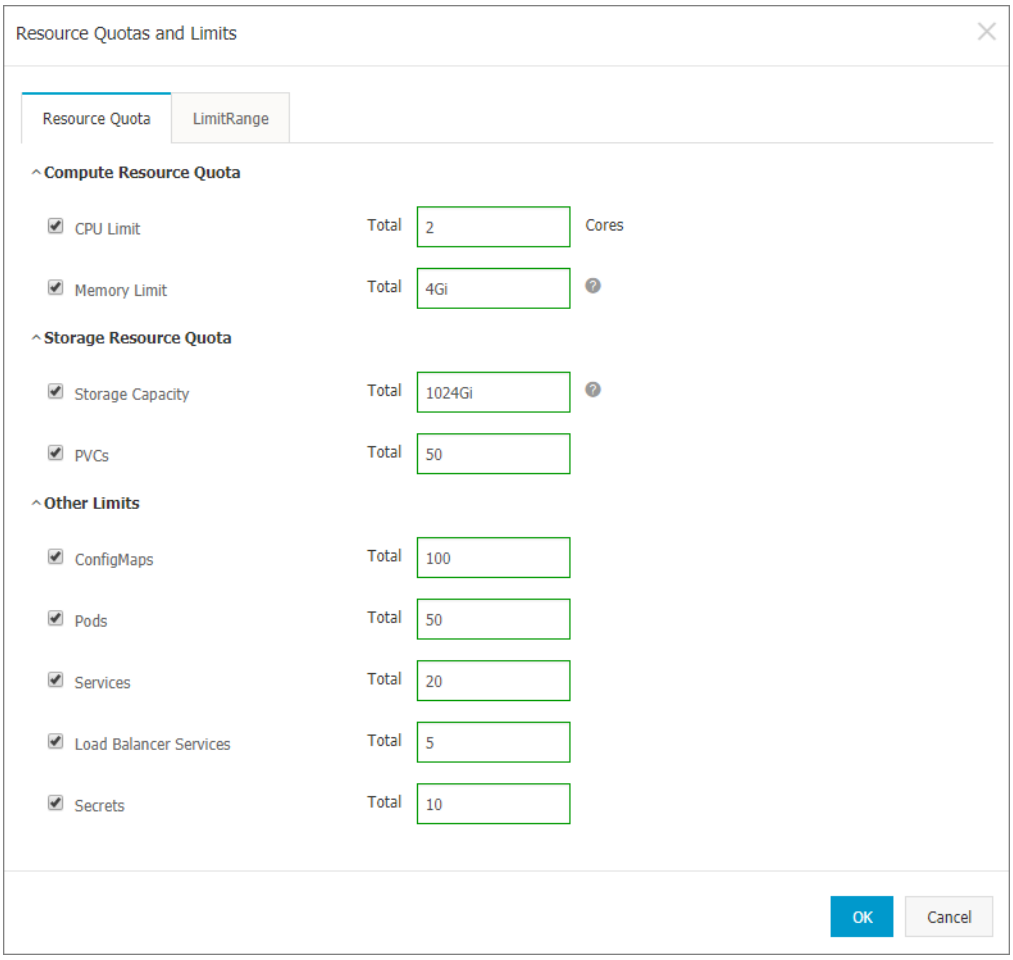
For a namespace, you can set quotas on resources such as CPU, memory, and pod quantity. For more information, see Resource Quotas.

Procedure

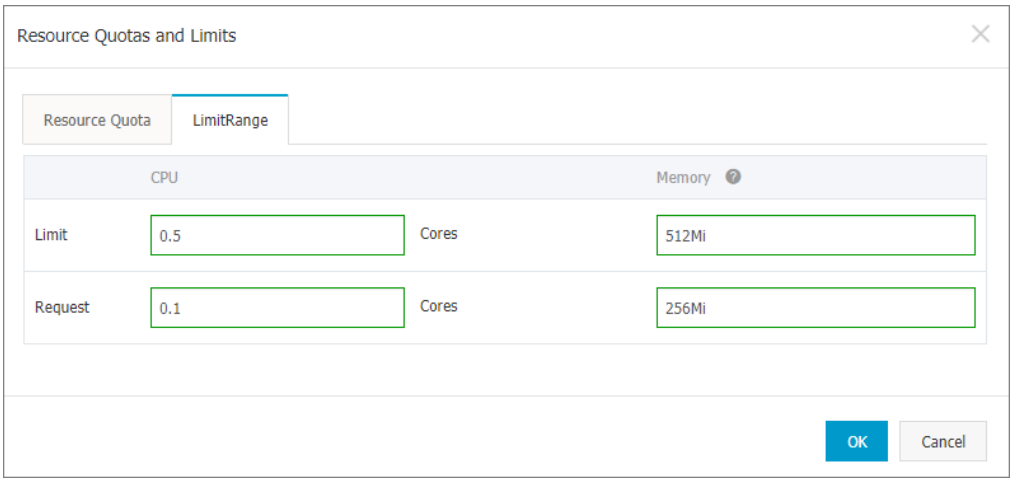
1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, click **Namespaces and Quotas**.
5. Find the namespace that you want to manage and click **Resource Quotas and Limits** in the **Actions** column.
6. In the dialog box that appears, set resource quotas and default resource limits.

> **Note** After you set CPU and memory quotas for a namespace, you must specify CPU and memory limits when you create a pod. You can also set the default resource limits for the namespace. For more information, see Resource Quotas.

i. Set resource quotas for the namespace.



ii. You can set resource limits and resource requests for containers in the namespace. This enables you to control the amount of resources consumed by the containers. For more information, see https://kubernetes.io//memory-default-namespace/.



7. After you set resource quotas and limits, connect to a master node of the cluster and run the following commands to query the resource configurations of the namespace.

```
# kubectl get limitrange,ResourceQuota -n test
NAME AGE
limitrange/limits 8m
NAME AGE
resourcequota/quota 8m
# kubectl describe limitrange/limits resourcequota/quota -n test
Name: limits
Namespace: test
Type Resource Min Max Default Request Default Limit Max Limit/Request Ratio
---- -------- --- --- --------------- ------------- -----------------------
Container cpu - - 100m 500m -
Container memory - - 256Mi 512Mi -
Name: quota
Namespace: test
Resource Used Hard
-------- ---- ----
configmaps 0 100
limits.cpu 0 2
limits.memory 0 4Gi
persistentvolumeclaims 0 50
pods 0 50
requests.storage 0 1Ti
secrets 1 10
services 0 20
services.loadbalancers 0 5
```

3.1.6.6.3. Modify a namespace

You can modify an existing namespace.

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- A sample namespace named **test** is created. For more information, see Create a namespace.

Context

When you modify a namespace, you can add, delete, or modify the labels that are added to the namespace based on your requirements.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, click **Namespaces and Quotas**.
5. Find the namespace that you want to modify and click **Edit** in the **Actions** column.
6. In the **Edit Namespace** dialog box, find the label that you want to modify and click **Edit** to modify the label. For example, you can change the key-value pair of the label to `env:test-V2`. Then, click **Save**.
7. Click **OK**.

   You can find that the labels of the namespace on the Namespace page are updated.

3.1.6.6.4. Delete a namespace

You can delete namespaces that are no longer in use.

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- A sample namespace named **test** is created. For more information, see Create a namespace.

Context

> **Note** When you delete a namespace, all resource objects under the namespace are deleted. Proceed with caution.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, click **Namespaces and Quotas**.
5. Find the namespace that you want to delete and click **Delete** in the **Actions** column.
6. In the message that appears, click **Confirm**.

   Return to the Namespace page. You can find that the namespace is deleted. Resource objects in the namespace are also deleted.

3.1.6.7. Applications

3.1.6.7.1. Create an application from an image

You can use an image to create an NGINX application that is accessible over the Internet.

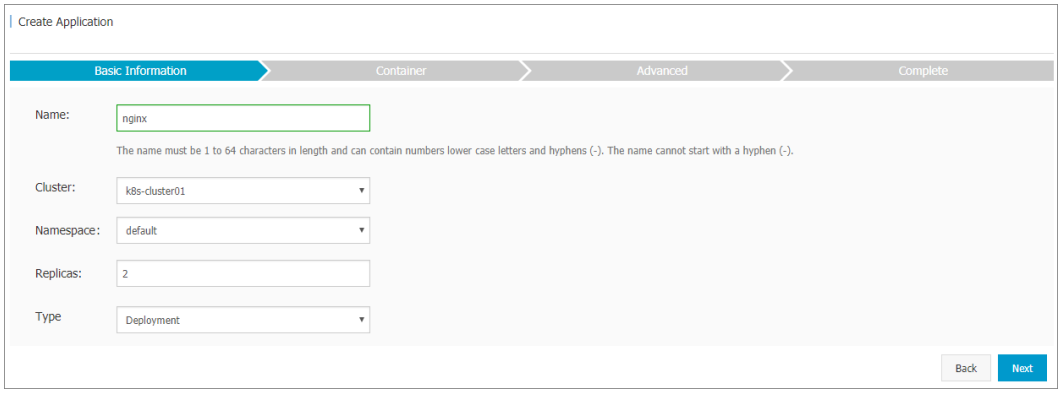
Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- Your Kubernetes cluster is accessible over the Internet.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select the namespace and click **Create from Image** in the upper-right corner.
6. On the page that appears, set the following parameters: **Name**, **Cluster**, **Namespace**, **Replicas**, and **Type**. Then, select **Synchronize Timezone**. Click **Next**.

   If you do not set the **Namespace** parameter, the default namespace is used.



7. Configure containers.

> **Note** You can configure a pod that contains one or more containers for the application.

i. Configure basic settings.

Basic settings of the container

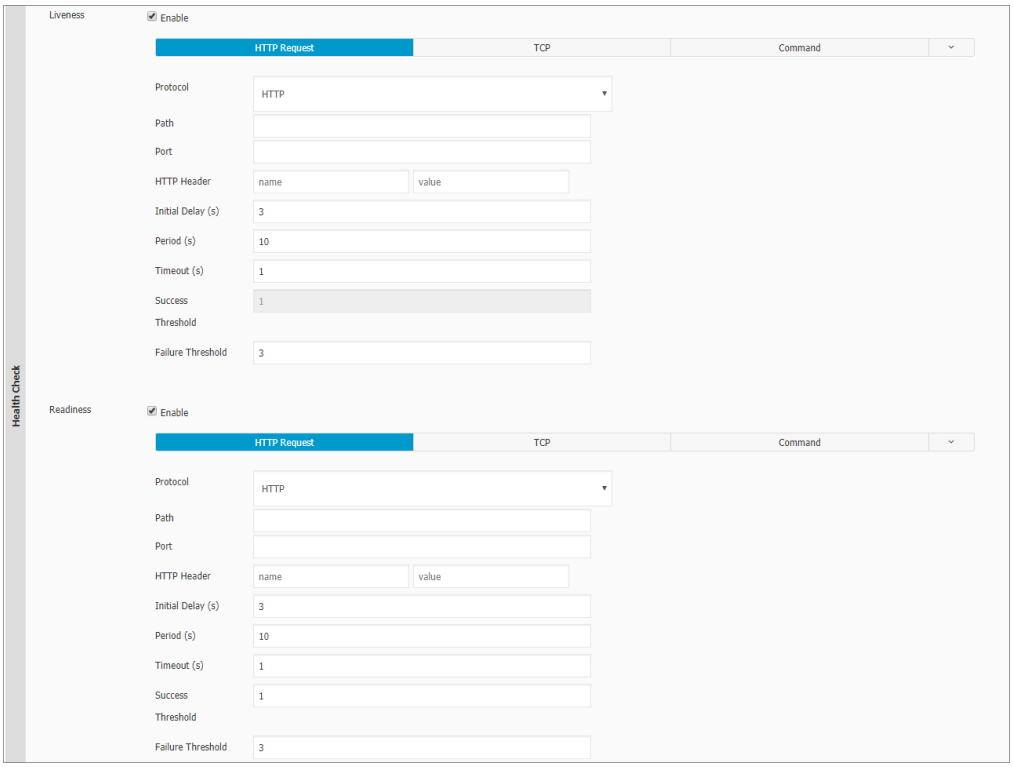
Parameter	Description
Image Name	You can click **Select Image** to select an image in the dialog box that appears and click **OK**. In this example, NGINX is selected. You can also specify an image in a private registry. Use the following format to specify an image in a private registry: `domainname/namespace/imagename` .
Image Version	You can click **Select Image Version** to select a version. If you do not specify the image version, the latest version is used.
Always	Container Service caches images to improve the efficiency of application deployment. When Container Service deploys the application, if the specified image version is the same as the cached image version, the cached image is used. Therefore, when you update the application code, if you do not change the image version for reasons such as to support the upper-layer workloads, the cached image is used to deploy the application. After you select the Always option, Container Service always pulls the latest image from the repository. This ensures that the latest image and code are used to deploy the application.
Set Image Pull Secret	Click Set Image Pull Secret to set the image Secret. The Secret is required if you need to access a private repository.
Resource Limit	The upper limits of CPU and memory resources that can be used by this application. This prevents the application from occupying excessive resources. The unit of CPU resources is Core. The unit of memory is MiB.
Required Resources	The amount of CPU and memory resources that are reserved for this application. These resources are exclusive to the container. This prevents other services or processes from occupying the resources of the application.
Container Start Parameter	**stdin**: Pass stdin to the container. **tty**: Stdin is a TTY. Typically, these two check boxes are both selected. This allows you to associate the terminal (tty) with the standard inputs (stdin) of the container. For example, an interactive program can obtain standard inputs from users and then display the inputs on the terminal.
Init Container	When this check box is selected, the system creates an Init Container that contains useful tools. For more information, see .

ii. (Optional)Set environment variables.

You can set environment variables in key-value pairs for pods. Environment variables are used to apply pod configurations to containers. For more information, see Pod variable.

iii. (Optional)Configure the **Health Check** settings.

Health check settings include liveness and readiness probes. Liveness probes determine when to restart the container. Readiness probes indicate whether the container is ready to accept network traffic. For more information about health checks, see .



Request type	Description

Request type	Description
HTTP	Sends an HTTP GET request to the container. Supported parameters: ■ Protocol: HTTP or HTTPS. ■ Path: the requested path on the server. ■ Port: the container port that you want to open. Enter a port number from 1 to 65535. ■ HTTP Header: the custom headers in the HTTP request. Duplicate headers are allowed. Key-value pairs are supported. ■ Initial Delay (s): The **initialDelaySeconds** field. The time (in seconds) to wait before performing the first probe after the container is started. Default value: 3. ■ Period (s): the **periodSeconds** field. The intervals (in seconds) between two adjacent probes. Default value: 10. Minimum value: 1. ■ Timeout (s): the **timeoutSeconds** field. The time (in seconds) after which a probe times out. Default value: 1. Minimum value: 1. ■ Healthy Threshold: the minimum number of times that an unhealthy container must consecutively pass health checks before it is considered healthy. Default value: 1. Minimum value: 1. For liveness probes, this parameter must be set to 1. ■ Unhealthy Threshold: the minimum number of times that a healthy container must consecutively fail health checks before it is considered unhealthy. Default value: 3. Minimum value: 1.
TCP	Sends a TCP socket to the container. Kubelet attempts to open the socket on the specified port. If the connection can be established, the container is considered healthy. Otherwise, the container is considered unhealthy. You can set the following parameters: ■ Port: the container port that you want to open. Enter a port number from 1 to 65535. ■ Initial Delay (s): the **initialDelaySeconds** field. The time (in seconds) to wait before performing the first probe after the container is started. Default value: 15. ■ Period (s): the **periodSeconds** field. The intervals (in seconds) between two adjacent probes. Default value: 10. Minimum value: 1. ■ Timeout (s): the **timeoutSeconds** field. The time (in seconds) after which a probe times out. Default value: 1. Minimum value: 1. ■ Healthy Threshold: the minimum number of times that an unhealthy container must consecutively pass health checks before it is considered healthy. Default value: 1. Minimum value: 1. For liveness probes, this parameter must be set to 1. ■ Unhealthy Threshold: the minimum number of times that a healthy container must consecutively fail health checks before it is considered unhealthy. Default value: 3. Minimum value: 1.

Request type	Description
Command	Runs a probe command in the container to check the health status of the container. You can set the following parameters: ■ Command: the probe command that is run to check the health status of the container. ■ Initial Delay (s): the **initialDelaySeconds** field. The time (in seconds) to wait before performing the first probe after the container is started. Default value: 5. ■ Period (s): the **periodSeconds** field. The intervals (in seconds) between two adjacent probes. Default value: 10. Minimum value: 1. ■ Timeout (s): the **timeoutSeconds** field. The time (in seconds) after which a probe times out. Default value: 1. Minimum value: 1. ■ Healthy Threshold: the minimum number of times that an unhealthy container must consecutively pass health checks before it is considered healthy. Default value: 1. Minimum value: 1. For liveness probes, this parameter must be set to 1. ■ Unhealthy Threshold: the minimum number of times that a healthy container must consecutively fail health checks before it is considered unhealthy. Default value: 3. Minimum value: 1.

iv. Configure the lifecycle of the container.

You can set the following parameters to configure the lifecycle of the container in the Start, Post Start, and Pre Stop fields. For more information, see https://kubernetes.io/docs/tasks/configure-pod-container/attach-handler-lifecycle-event/.

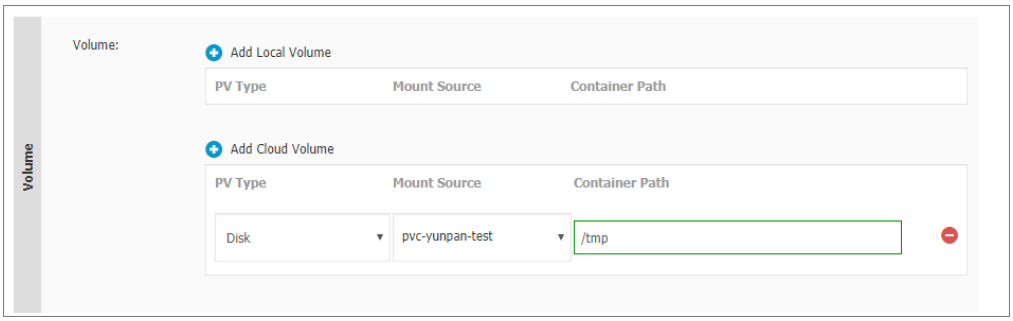
- **Start**: the pre-start command and parameter.
- **Post Start**: the post-start command.
- **Pre Stop**: the pre-stop command.

v. (Optional)Configure volumes.

Local storage and cloud storage are supported.

- **Local Storage**: supports hostPath, ConfigMap, Secret, and emptyDir. The specified volume is mounted to a path in the container. For more information, see Volumes.
- **Cloud Storage**: supports disks, Apsara File Storage NAS (NAS) volumes, and Object Storage Service (OSS) volumes.

In this example, a persistent volume (PV) is created from a disk and mounted to the */tmp* path in the container. Data generated in this path is stored in the disk.



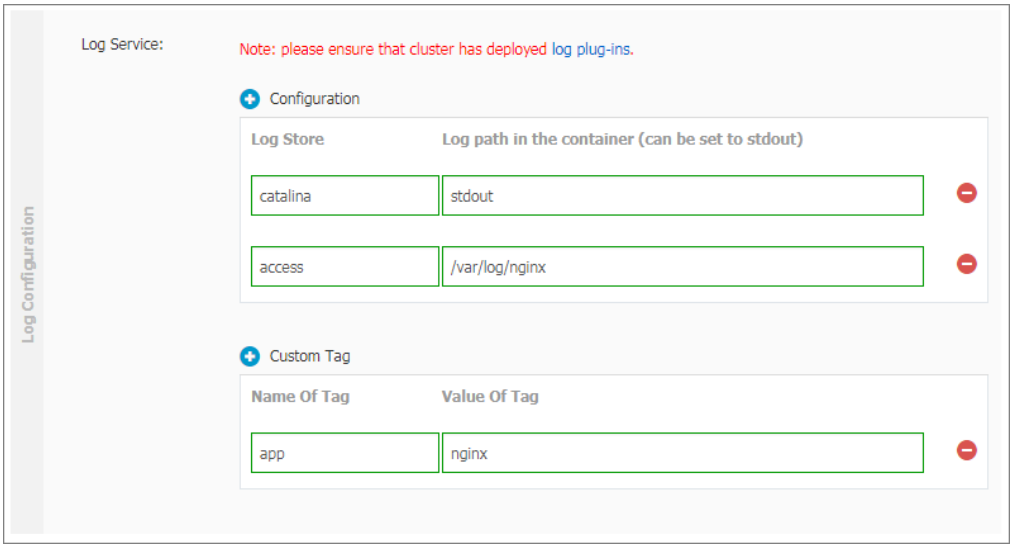
vi. (Optional)Configure **Log Service**. You can specify log collection configurations and custom tags.

> **Note** Make sure that the Log Service agent is installed in the cluster.

You can configure the following log collection settings:

- **Logstore**: Create a Logstore in Log Service to store the collected log data.
- **Log Path in Container**: Specify stdout or a path to collect log data.
  - **stdout**: specifies that the stdout files are collected.
  - **Text Logs**: specifies that the log data in the specified path of the container is collected. In this example, the log data in the */var/log/nginx* path is collected. Wildcard characters can be used in the path.

You can also add custom tags. Tags are added to the log data of the container when the log data is collected. Log data with tags is easier to aggregate and filter.

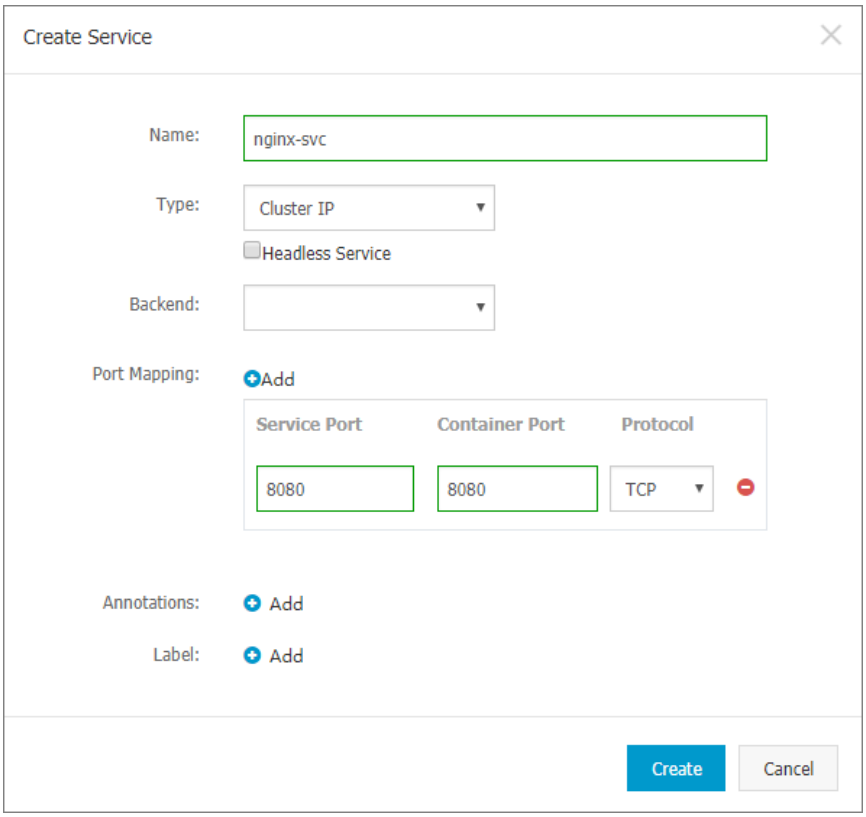


8. After you complete the configurations of the container, click **Next**.
9. Configure advanced settings. Configure the **Access Control** settings.

You can configure the method to expose pods and click **Create**. In this example, a ClusterIP Service and an Ingress are created to enable Internet access to the NGINX application.

> **Note**
>
> You can configure the following access control settings based on your business requirements:
>
> - Internal applications: For applications that run inside the cluster, you can create a ClusterIP Service or a NodePort Service to enable internal communication.
> - External applications: For applications that need to be accessed over the Internet, you can configure access control settings by using one of the following methods:
>   - Create a LoadBalancer Service: When you create a LoadBalancer Service, a Server Load Balancer (SLB) instance is associated with the Service and is created to expose applications to the Internet.
>   - Create a ClusterIP Service or a NordPort Service, and then create an Ingress: When you use this method, an Ingress is created to expose applications to the Internet. For more information, see .

i. To create **Services**, click **Create** in the Access Control section. In the dialog box that appears, set the parameters, and then click **Create**.



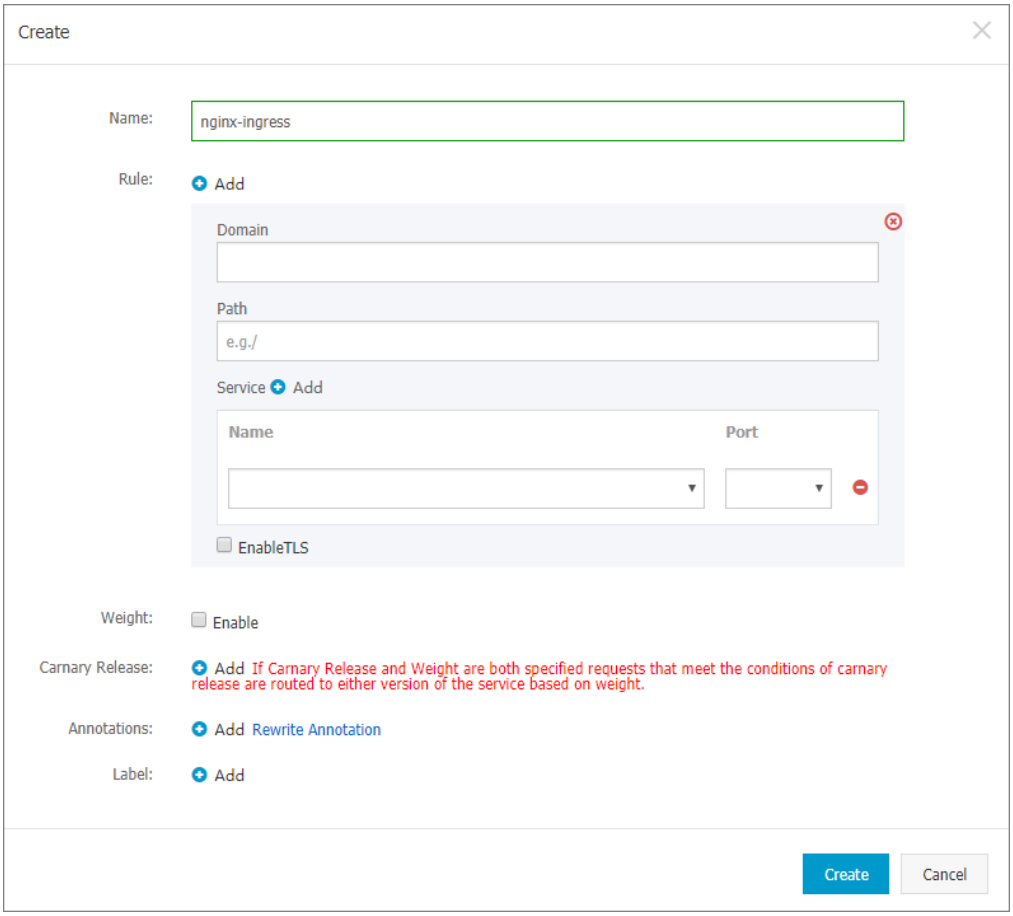
Parameter	Description
Name	Enter a name for the Service. Default value: `applicationname-svc`.

Parameter	Description
Type	Select one from the following three Service types: ■ Cluster IP: creates a ClusterIP Service. This type of Service exposes applications through an internal IP address of the cluster. If you select this type, applications can be accessed only within the cluster. ■ Node Port: creates a NordPort Service. This type of Service exposes applications through the IP address and static port on each node. A NodePort Service can be used to route requests to a ClusterIP Service. The system automatically creates the ClusterIP Service. You can access a NodePort Service from outside the cluster by requesting `<NodeIP>:<NodePort>`. ■ Server Load Balancer: creates a LoadBalancer Service. This type of Service exposes applications through an SLB instance, which supports Internet access or internal access. The SLB instance can route requests to NodePort and ClusterIP Services.
Port Mapping	Specify a Service port and a container port. If the **Type** parameter is set to Node Port, you must specify a node port to avoid port conflicts. TCP and UDP protocols are supported.
Annotations	Add annotations to the Service. SLB parameters are supported. For more information, see Access services by using SLB.
Label	Add one or more labels to the Service. The labels are used to identify the Service.

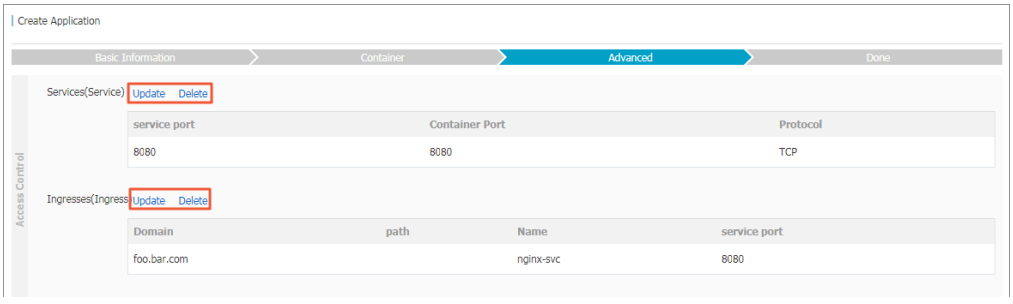
ii. To create **Ingresses**, click **Create** in the Access Control section. Configure Ingress rules in the dialog box that appears, and then click **Create**. For more information about Ingress configuration, see Ingress configurations.

When you create an application from an image, you can create an Ingress for only one Service. In this example, a virtual hostname is specified as the test domain name. You must add a mapping rule to the hosts file for this domain name, as shown in the following code block. In actual scenarios, use a domain name that has obtained an Internet Content Provider (ICP) number.

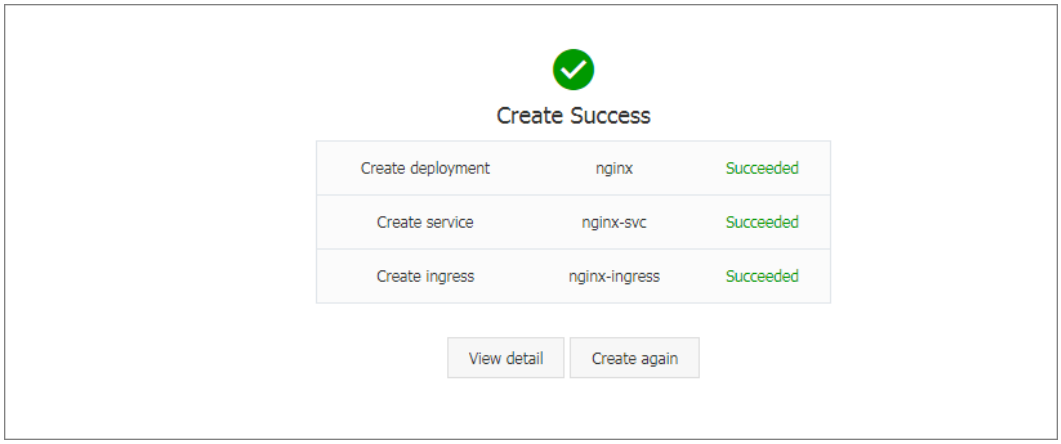
```
101.37.224.146   foo.bar.com    #The IP address of the Ingress.
```



iii. You can find the newly created Service and Ingress in the Access Control section. You can click **Update** or **Delete** to make changes.

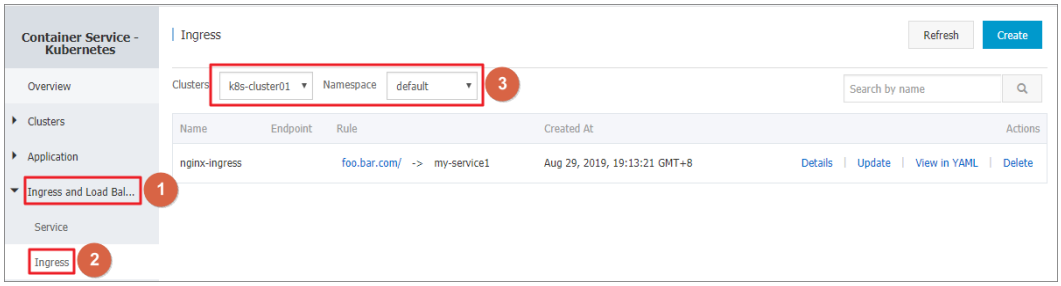


10. Click **Create**.
11. After the application is created, you are redirected to the Complete page, which displays the resource objects under the application. You can click **View Details** to view application details.



The nginx-deployment details page appears.

12. In the left-side navigation pane, choose **Services and Ingresses > Ingresses**. On the Ingresses page, you can find the created Ingress.



13. Enter the test domain in the address bar of your browser and press Enter. The NGINX welcome page appears.

3.1.6.7.2. Create an application from an orchestration template

Container Service provides orchestration templates that you can use to create applications. You can also modify the templates based on YAML syntax to customize applications.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Context

This topic describes how to use an orchestration template to create an NGINX application that consists of a Deployment and a Service. The Deployment provisions pods for the application and the Service manages access to the pods at the backend.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select the namespace and click **Create from Template** in the upper-right corner.
6. Configure the template and click **Create**.
   - **Sample Template**: Container Service provides YAML templates of various resource types to help you quickly deploy resource objects. You can also create a custom template based on YAML syntax to describe the resource that you want to define.
   - **Add Deployment**: This feature allows you to quickly define a YAML template.
   - **Use Existing Template**: You can import an existing template to the configuration page.

   The following NGINX template is based on an orchestration template provided by Container Service. You can use this template to quickly create a Deployment to run an NGINX application

   > **Note** Container Service supports YAML syntax. You can use the `---` symbol to separate multiple resource objects. This allows you to define multiple resource objects in a single template.

```
apiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1
kind: Deployment
metadata:
    name: nginx-deployment
    labels:
      app: nginx
spec:
    replicas: 2
    selector:
      matchLabels:
        app: nginx
    template:
      metadata:
        labels:
          app: nginx
      spec:
        containers:
        - name: nginx
          image: nginx:1.7.9 # replace it with your exactly <image_name:tags>
          ports:
          - containerPort: 80
---
apiVersion: v1     # for versions before 1.8.0 use apps/v1beta1
kind: Service
metadata:
   name: my-service1        #TODO: to specify your service name
   labels:
     app: nginx
spec:
   selector:
     app: nginx             #TODO: change label selector to match your backend pod
   ports:
   - protocol: TCP
     name: http
     port: 30080            #TODO: choose an unique port on each node to avoid port conflict
     targetPort: 80
   type: LoadBalancer        ## In this example, the type is changed from NodePort to LoadBalancer
.
```

7. Click **Create**. A notification that indicates the deployment status appears.

   After the application is created, choose **Network > Services** in the left-side navigation pane. On the Services page, you can find that a Service named my-service1 is created for the application. The external endpoint of the Service is also displayed on the page. Click the endpoint in the **External Endpoint** column.

8. You can visit the NGINX welcome page in the browser.

3.1.6.7.3. Use commands to manage applications

You can use commands to create applications or view application containers.

Prerequisites

Before you use commands on your local host, you have connected to a Kubernetes cluster through kubectl. For more information, see Connect to a Kubernetes cluster through kubectl.

Run a command to create an application

You can use the following command to run a simple container (an NGINX Web server in this example):

```
# kubectl run -it nginx --image=registry.aliyuncs.com/spacexnice/netdia:latest
```

This command creates a service portal for this container. After you specify `--type=LoadBalancer`, an SLB route to the NGINX container is created.

```
# kubectl expose deployment nginx --port=80 --target-port=80 --type=LoadBalancer
```

Run a command to view container information

Run the following command to list all running containers in the default namespace:

```
root@master # kubectl get pods
    NAME                                READY     STATUS    RESTARTS   AGE
    nginx-2721357637-dvwq3              1/1       Running   1          9h
```

3.1.6.7.4. Create a Service

You can create a Service for your application in the Container Service console to provide access to the application.

In Kubernetes, a Service is an abstraction that defines a logical set of pods and a policy by which to access the pods. This pattern is also known as a microservice. A label selector determines whether the set of pods can be accessed by the Service.

Each pod in Kubernetes clusters has its own IP address. However, pods are frequently created and deleted. Therefore, if you directly expose pods to external access, high availability is not ensured. Services decouple the frontend from the backend. The frontend clients do not need to be aware of which backend pods are used. This provides a loosely-coupled microservice architecture.

For more information, see Kubernetes Services.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Step 1: Create a Deployment

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select a namespace and click **Create from Template** in the upper-right corner.
6. Select a sample template or customize a template, and click **Create**.

   In this example, the template of an NGINX Deployment is used.

   ```
   apiVersion: apps/v1 # for versions before 1.8.0 use apps/v1beta1 kind: Deployment metadata: name: nginx-deployment-basic labels: app: nginx spec: replicas: 2 selector: matchLabels: app: nginx template: metadata: labels: app: nginx spec: # nodeSelector: # env: test-team containers: - name: nginx image: nginx:1.7.9 # replace it with your exactly <image_name:tags> ports: - containerPort: 80
   ```

Query the state of the Deployment.

Step 2: Create a Service

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Network > Services**.
5. On the Services page, set **Namespace** and click **Create** in the upper-right corner.
6. In the Create Service dialog box, set the parameters.
   - **Name**: Enter a name for the Service. In this example, nginx-svc is used.
   - **Type**: Select the type of Service. This parameter specifies the Service is accessed. Valid values:
     - Cluster IP: a ClusterIP Service. This type of Service exposes pods through an internal IP address of the cluster. If you select this option, pods can be accessed from only within the cluster. This is the default value.
     - Node Port: a NodePort Service. This type of Service exposes pods by using the IP address and a static port of each node. A NodePort Service can be used to route requests to a ClusterIP Service, which is automatically created by the system. You can access a NodePort Service from outside the cluster by sending requests to `<NodeIP>:<NodePort>`.
     - Server Load Balancer: a LoadBalancer Service. This type of Service exposes pods by using Server Load Balancer (SLB) instances, which support Internet access or internal access. SLB instances can be used to route requests to NodePort and ClusterIP Services.
   - **Backend**: the backend object that you want to associate with the Service. In this example, select nginx-deployment-basic that was created in the preceding step. If you do not associate the Service with a backend object, no Endpoint object is created. You can also manually associate the Service with an Endpoint object. For more information, see Create a Service without selectors.
   - **Port Mapping**: Set the Service port and container port. The container port must be the same as the one that is exposed in the backend pod.
   - **Annotations**: Add one or more annotations to the Service to configure SLB parameters. For example, set the name to `service.beta.kubernetes.io` and the value to `20`. This means that the maximum bandwidth of the Service is 20 Mbit/s. For more information, see Access services by using SLB.
   - **Label**: Add one or more labels to the Service. The labels are used to identify the Service.
7. Click **Create**. After the nginx-svc Service is created, it appears on the Services page.

   On the **Services** page, you can view basic information about the Service. You can also access its external endpoint by using a browser.

3.1.6.7.5. View a Service

You can view details about a Service in the Container Service console.

Context

Each pod in Kubernetes clusters has its own IP address. However, pods are frequently created and deleted. Therefore, it is not practical to directly expose pods to external access. Services decouple the frontend from the backend, which provides a loosely-coupled microservice architecture.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.

4. In the left-side navigation pane of the details page, choose **Network > Services**.
5. On the **Services** page, select the namespace, find the Service that you want to view, and then click **Details** in the Actions column.
   On the details page, you can view detailed information about the Service.

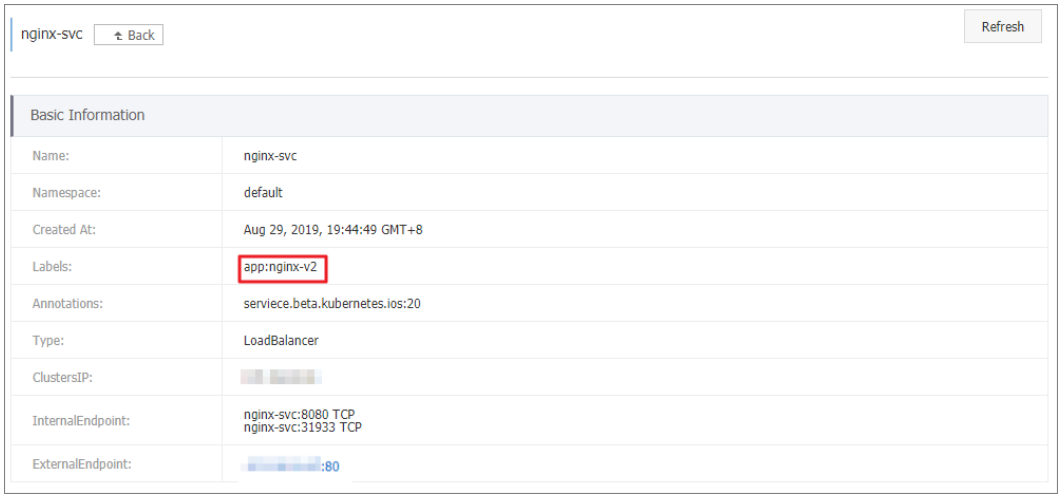
3.1.6.7.6. Update a Service

You can update a Service in the Container Service console. This topic describes how to update a Service.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Network > Services**.
5. On the **Services** page, select the namespace, find the Service that you want to update, and then click **Update** in the Actions column.
6. In the **Update Service** dialog box, modify the configurations and click **Update**.

   In the Service list, find the Service that you updated and click **Details** in the Actions column to view configuration changes. In this example, the labels of the Service are modified.



3.1.6.7.7. Delete a Service

This topic describes how to delete a Service.

Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- A Service is created. For more information, see Create Servcies.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Network > Services**.
5. On the Services page, select the namespace, find the Service that you want to delete, and then click **Delete** in

the Actions column.

6. In the message that appears, click **Confirm**.

3.1.6.7.8. Use a trigger to redeploy an application

You can create a trigger and use it to redeploy an application. This topic describes how to use a trigger.

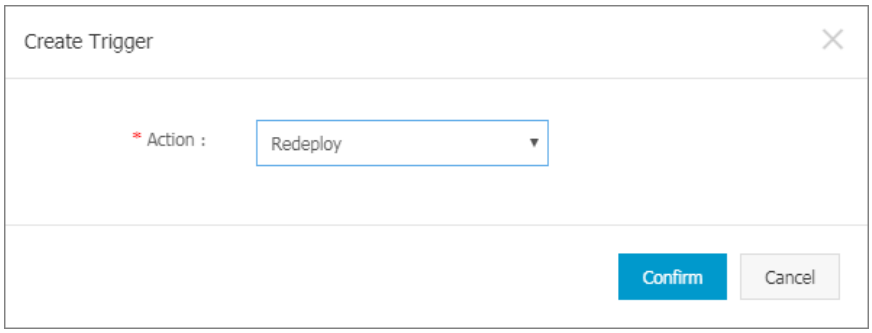
Prerequisites

- A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.
- An application is created. Then, a trigger is created for the application and the application is used to test the trigger. In this example, an NGINX application is created.

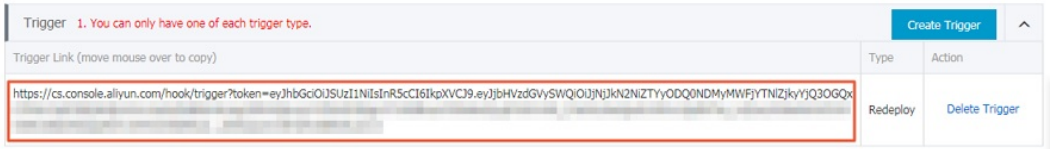
Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Manage** in the **Actions** column of the cluster.
4. In the left-side navigation pane, click Deployments. On the **Deployments** page, find the NGINX application and click **Details** in the Actions column.
5. On the details page of the application, click the **Triggers** tab. Then, click **Create Trigger**.
6. In the dialog box that appears, set Action to **Redeploy** and click **Confirm**.

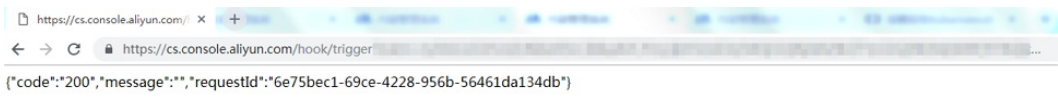
   Note You can create triggers only to redeploy applications.



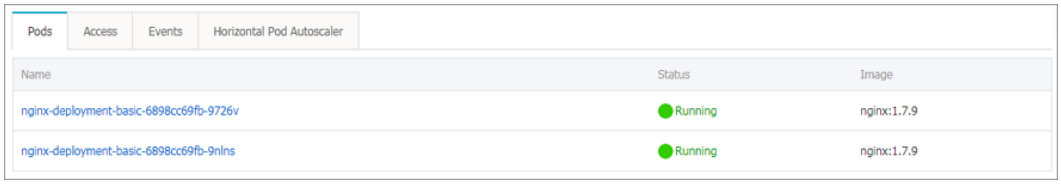
   After the trigger is created, the webhook URL of the trigger appears in the Trigger Link Address column on the **nginx - deployment** page.



7. Copy the webhook URL, enter it into the address bar of your browser, and press Enter. A message that indicates specific information such as the request ID appears.



8. Go to the **nginx - deployment** page. A new pod appears on the **Pods** tab.



After the new pod is deployed, the original pod is automatically deleted.

What's next

You can run triggers by sending GET or POST requests from an external system. For example, you can use the **curl** command-line tool to run triggers.

To run the redeploy trigger, run the following command:

```
curl https://cs.console.aliyun.com/hook/trigger?token=xxxxxxx
```

3.1.6.7.9. View pods

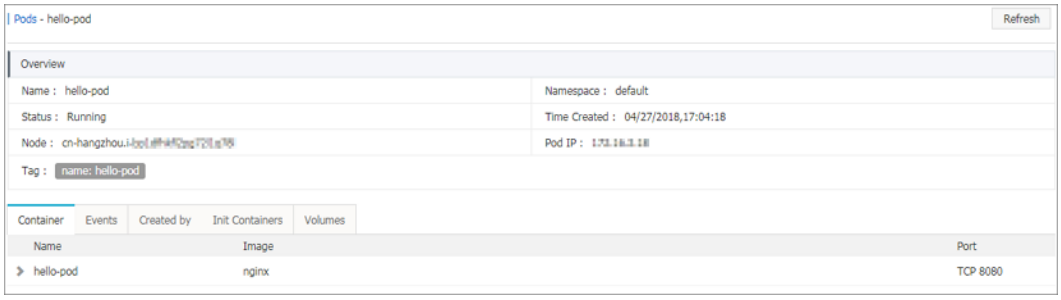
You can view pods in the Container Service console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Pods**.
5. In the left-side navigation pane, choose **Applications > Pods**. The Pods page appears.
6. On the **Pods** page, select the namespace, find the pod that you want to view, and then click **View Details** in the Actions column.

> **Note** You can update or delete pods on the Pods page. We recommend that you use Deployments to manage pods if the pods are created by Deployments.

On the details page of the pod, you can view detailed information about the pod.



3.1.6.7.10. Manage pods

Pods are the smallest deployable units in Kubernetes. A pod runs an instance of an independent application in Kubernetes. Each pod contains one or more containers that are tightly coupled. You can modify pods, view pods, and manually scale the number of pods for an application in the Container Service console.

View pods

View pod details

1. Log on to the Container Service console.

2. In the left-side navigation pane of the Container Service console, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click its name or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Pods**.
5. Find the pod that you want to view and click **View Details** in the Actions column.

   You can view details of pods by using one of the following methods:

   - Method 1: In the left-side navigation pane of the details page, choose **Workloads > Deployments**. Find the application that you want to manage and click its name. On the **Pods** tab, click the name of the pod to view details.
   - Method 2: In the left-side navigation pane of the details page, choose **Network > Services**. Click the name of the Service that you want to manage. On the details page, find and click the name of the application that you want to manage. On the **Pods** tab, click the name of the pod to view details.

   > **Note** On the Pods page, you can modify and delete pods. For pods that are created by using a Deployment, we recommend that you use the Deployment to manage the pods.

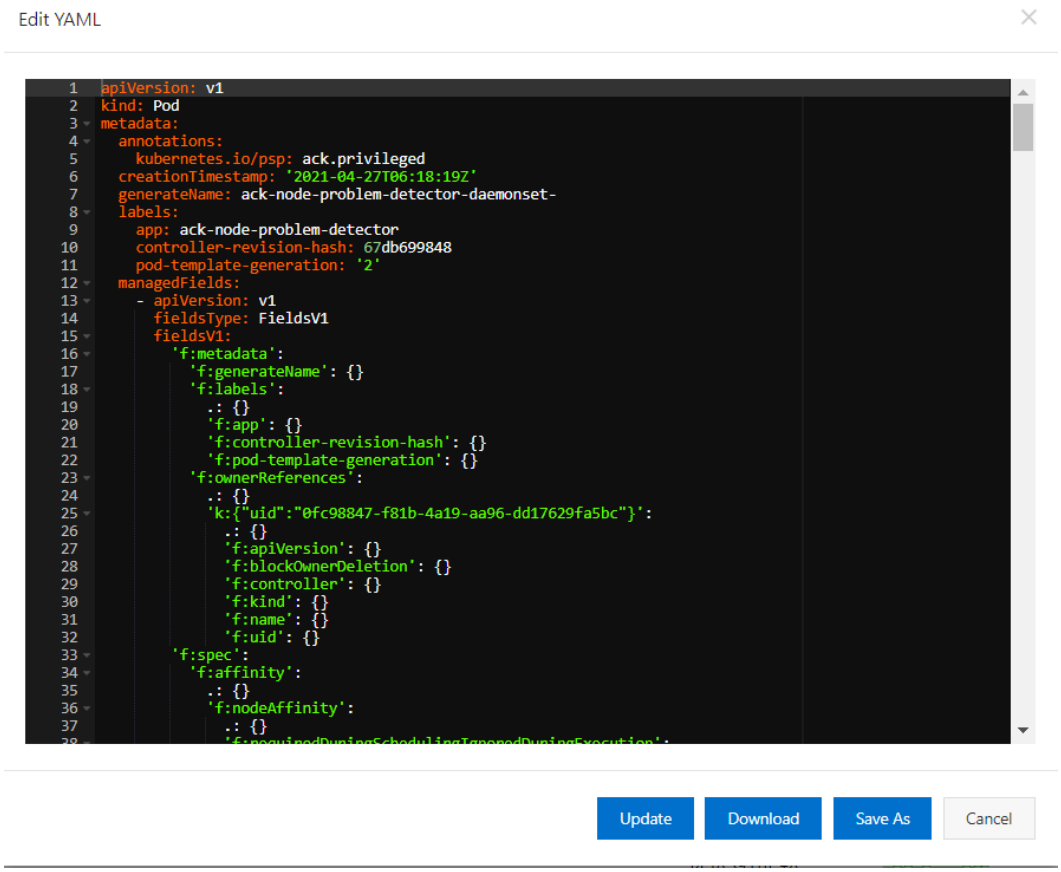
View pod log

You can view the log of a pod by using one of the following methods:

Navigate to the Pods tab, find the pod that you want to manage, and then click **Logs** on the right side of the page to view the log data.

Modify pod configurations

1. Log on to the Container Service console.
2. In the left-side navigation pane of the Container Service console, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click its name or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Pods**.
5. On the **Pods** page, find the pod that you want to manage and click **Edit** on the right side of the page.
6. In the Edit YAML dialog box, modify the configurations and click **Update**.



Manually scale the number of pods for an application

After an application is created, you can scale the number of pods that are provisioned for the application.

1. Log on to the Container Service console.
2. In the left-side navigation pane of the Container Service console, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click its name or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. Select the namespace where the Deployment is deployed, find the Deployment, and then click **Scale** in the **Actions** column.
6. In the Scale dialog box, set Desired Number of Pods to 4 and click **OK**.

> **Note** By default, Deployments are updated based on the rollingUpdate strategy. This ensures that a minimum number of pods are available during the update. You can change this number in the YAML file.

3.1.6.7.11. Schedule pods to specific nodes

You can add labels to nodes and schedule pods to the nodes with specified labels. This topic describes how to schedule a pod to a node with specified labels.

You can add labels to nodes and then configure `nodeSelector` to schedule pods to nodes with specified labels. For more information about how nodeSelector works, see nodeselector.

To meet business requirements, you may want to deploy a controller service on a master node, or deploy services on nodes that use standard SSDs. You can use the following method to schedule pods to nodes with specified labels.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Step 1: Add a label to a node

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. On the **Nodes** page, click **Manage Labels and Taints** in the upper-right corner.
6. Select one or more nodes and then click **Add Label**. In this example, a worker node is selected.
7. In the dialog box that appears, enter the name and value of the label and click **OK**.

   On the Labels tab, you can find the `group:worker` label next to the selected node.

You can also run the following command to add a label to a node: `kubectl label nodes <node-name> <label-key>=<label-value>` .

Step 2: Schedule a pod to the node

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select the namespace and click **Create from Template** in the upper-right corner.
6. On the **Create** page, select a template from the Sample Template drop-down list. In this example, a custom template is selected. Copy the following content to the custom template and click **Create**.

   The following template is used as an example:

```
apiVersion: v1
kind: Pod
metadata:
  labels:
    name: hello-pod
  name: hello-pod
spec:
  containers:
    - image: nginx
      imagePullPolicy: IfNotPresent
      name: hello-pod
      ports:
        - containerPort: 8080
          protocol: TCP
      resources: {}
      securityContext:
        capabilities: {}
        privileged: false
      terminationMessagePath: /dev/termination-log
  dnsPolicy: ClusterFirst
  restartPolicy: Always
  nodeSelector:
    group: worker                          ## This value must be the same as the node label that
is added in Step 1.
 status: {}
```

3.1.6.7.12. Simplify application deployment by using Helm

This topic introduces the basic terms and components of Helm and describes how to use Helm to deploy an Apache Spark-based WordPress application in a Kubernetes cluster.

Prerequisites

- A Kubernetes cluster is created in the Container Service console. For more information, see Create a Kubernetes cluster.

  Tiller is automatically deployed to the cluster when the Kubernetes cluster is created. The Helm CLI is automatically installed on each master node. An Alibaba Cloud chart repository is added to Helm.

- A Kubernetes version that supports Helm is used.

  Only Kubernetes 1.8.4 and later support Helm. If the Kubernetes version of your cluster is 1.8.1, you can **upgrade** the cluster on the Clusters page of the Container Service console.

Context

Application management is the most challenging task in Kubernetes. The Helm project provides a unified method to package software and manage software versions. You can use Helm to simplify application distribution and deployment. App Catalog is integrated with Helm in the Container Service console and provides extended features based on Helm. App Catalog also supports Alibaba Cloud chart repositories to help you accelerate application deployments. You can deploy applications in the Container Service console or by using the Helm CLI.

This topic introduces the basic terms and components of Helm and describes how to use Helm to deploy an Apache Spark-based WordPress application in a Kubernetes cluster.

Basic terms

Helm is an open source project initiated by Deis. Helm can be used to simplify the deployment and management of Kubernetes applications.

Helm serves as a package manager for Kubernetes and allows you to find, share, and use applications built by Kubernetes. Before you use Helm, you must familiarize yourself with the following basic terms:

- Chart: a packaging format used by Helm. Each chart contains the images, dependencies, and resource definitions that are required to run an application. A chart may contain service definitions in a Kubernetes cluster. A Helm chart is similar to a Homebrew formula, an Advanced Package Tool (APT) dpkg, or a Yum rpm.
- Release: an instance of a chart that runs in a Kubernetes cluster. A chart can be installed multiple times in a Kubernetes cluster. After a chart is installed, a new release is created. For example, you can install a MySQL chart. If you want to run two databases in your cluster, you can install the MySQL chart twice. Each time a chart is installed, a release is created with a different name.
- Repository: the storage of charts. Charts are published and stored in repositories.

Helm components

Helm uses a client-server architecture and consists of the following components:

- The Helm CLI is the Helm client that runs on your on-premises machine or on the master nodes of a Kubernetes cluster.
- Tiller is the server-side component and runs in a Kubernetes cluster. Tiller manages the lifecycles of Kubernetes applications.
- A repository is used to store charts. The Helm client can access the index file and packaged charts in a chart repository over HTTP.

Deploy an application in the Container Service console

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > App Catalog**.
3. On the App Catalog page, find and click the WordPress chart to go to the details page of the chart.
4. Click the **Parameters** tab and modify the configurations.

   In this example, a dynamically provisioned disk volume is associated with a persistent volume claim (PVC). For more information, see Use Apsara Stack disks.

   > **Note** You must first provision a disk as a persistent volume (PV). The capacity of the PV cannot be less than the capacity specified in the PVC.

5. In the **Deploy** section, select the cluster in which you want to deploy the application and click **Create**. After the application is deployed, you are redirected to the release page of the application.
6. In the left-side navigation pane, click **Clusters**.
7. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
8. In the left-side navigation pane of the details page, choose **Network > Services**. On the Services page, select the namespace, find the Service that is created for the application, and then check its HTTP and HTTPS external endpoints.
9. Click one of the endpoints to go to the WordPress website where you can publish blog posts.

Deploy an application by using the Helm CLI

After the Helm CLI is automatically installed on the master nodes of the Kubernetes cluster and the required chart repository is added to Helm, you can log on to the master nodes by using SSH. Then, you can deploy applications by using the Helm CLI. For more information, see Connect to a master node through SSH. You can also install and configure the Helm CLI and kubectl on your on-premises machine.

In this example, the Helm CLI and kubectl are installed and configured on your on-premises machine, and then an Apache Spark-based WordPress application is deployed.

1. Install and configure the Helm CLI and kubectl.

i. Install and configure kubectl on your on-premises machine.

For more information, see Connect to a Kubernetes cluster through kubectl.

To view the details of a Kubernetes cluster, run the `kubectl cluster-info` command.

ii. Install Helm on your on-premises machine.

For more information, see Install Helm.

2. Deploy the WordPress application.

In the following example, a WordPress blog website is deployed by using Helm.

i. Run the following command:

```
helm install --name wordpress-test stable/wordpress
```

> **Note** Container Service allows you to mount disks as dynamically provisioned volumes. Before you deploy the application, you must create a disk volume.

The following output is returned:

```
NAME: wordpress-testLAST DEPLOYED: Mon Nov 20 19:01:55 2017NAMESPACE: defaultSTATUS: DEPLOYED.
..
```

ii. Run the following command to query the release and Service that are created for the WordPress application:

```
helm listkubectl get svc
```

iii. Run the following command to view the pods that are provisioned for the WordPress application. You may need to wait until the pods change to the Running state.

```
kubectl get pod
```

iv. Run the following command to obtain the endpoint of the WordPress application:

```
echo http://$(kubectl get svc wordpress-test-wordpress -o jsonpath='{.status.loadBalancer.ingress[0].ip}')
```

You can enter the preceding URL in the address bar of your browser to access the WordPress application.

You can also run the following command based on the chart description to obtain the username and password of the administrator account for the WordPress application:

```
echo Username: userecho Password: $(kubectl get secret --namespace default wordpress-test-wordpress -o jsonpath="{.data.wordpress-password}" | base64 --decode)
```

v. To delete the WordPress application, run the following command:

```
helm delete --purge wordpress-test
```

Use a third-party chart repository

You can use the default Alibaba Cloud chart repository. If a third-party chart repository is accessible from your cluster, you can also use the third-party chart repository. Run the following command to add a third-party chart repository to Helm:

```
helm repo add Repository name Repository URLhelm repo update
```

For more information about Helm commands, see Helm documentation.

References

Helm contributes to the development of Kubernetes. A growing number of software suppliers, such as Bitnami, have provided high-quality charts. For more information about available charts, visit `https://kubeapps.com/`.

3.1.6.8. SLB and Ingress

3.1.6.8.1. Overview

Container Service allows you to flexibly manage load balancing and customize load balancing policies for Kubernetes clusters. Kubernetes clusters provide you with a variety of methods to access containerized applications. They also allow you to use SLB or Ingress to access internal services and implement load balancing.

3.1.6.8.2. Use SLB to access Services

You can access a Service by using Server Load Balancer (SLB).

Procedure

1. Create an NGINX application by using a CLI.

```
root@master # kubectl run nginx --image=registry.aliyuncs.com/acs/netdia:latest
      root@master # kubectl get po
      NAME                                READY     STATUS    RESTARTS   AGE
      nginx-2721357637-dvwq3              1/1       Running   1          6s
```

2. Create a Service for the NGINX application and specify `type=LoadBalancer` to expose the Service to the Internet through an SLB instance.

```
root@master # kubectl expose deployment nginx --port=80 --target-port=80 --type=LoadBalancer
      root@master # kubectl get svc
      NAME                  CLUSTER-IP      EXTERNAL-IP      PORT(S)                       AGE
      nginx                 172.19.XX.XX    101.37.XX.XX     80:31891/TCP                  4s
```

3. Enter `http://101.37.XX.XX` in the address bar of your browser and press Enter to access the nginx Service.

SLB parameters

SLB provides a variety of parameters that you can use to configure features and services such as health check, billing method, and SLB instance type. For more information, see SLB annotations.

Annotations

You can add annotations to use the load balancing features provided by SLB.

Use an existing internal-facing SLB instance

Add two annotations. You must replace your-loadbalancer-id with the ID of your SLB instance.

```
apiVersion: v1
kind: Service
metadata:
  annotations:
    service.beta.kubernetes.io/alicloud-loadbalancer-address-type: intranet
    service.beta.kubernetes.io/alicloud-loadbalancer-id: your-loadbalancer-id
  labels:
    run: nginx
  name: nginx
  namespace: default
spec:
  ports:
  - name: web
    port: 80
    protocol: TCP
    targetPort: 80
  selector:
    run: nginx
  sessionAffinity: None
  type: LoadBalancer
```

Save the preceding code as an slb.svc file and run the `kubectl apply -f slb.svc` command.

Create an HTTPS-based LoadBalancer Service

You must first create a certificate in the SLB console. Then, you can use the certificate ID (cert-id) and the following template to create a LoadBalancer Service and an HTTPS-based SLB instance.

```
apiVersion: v1
kind: Service
metadata:
  annotations:
    service.beta.kubernetes.io/alicloud-loadbalancer-cert-id: your-cert-id
    service.beta.kubernetes.io/alicloud-loadbalancer-protocol-port: "https:443"
  labels:
    run: nginx
  name: nginx
  namespace: default
spec:
  ports:
  - name: web
    port: 443
    protocol: TCP
    targetPort: 443
  selector:
    run: nginx
  sessionAffinity: None
  type: LoadBalancer
```

Note Annotations are case-sensitive.

SLB annotations

Annotation	Description	Default
service.beta.kubernetes.io/alicloud-loadbalancer-protocol-port	The listening port. Separate multiple ports with commas (,). Example: https:443,http:80.	N/A

Annotation	Description	Default
service.beta.kubernetes.io/alicloud-loadbalancer-address-type	The type of the SLB instance. Valid values: internet and intranet.	internet
service.beta.kubernetes.io/alicloud-loadbalancer-slb-network-type	The network type of the SLB instance. Valid values: classic and vpc.	classic
service.beta.kubernetes.io/alicloud-loadbalancer-charge-type	The billing method of the SLB instance. Valid values: paybytraffic and paybybandwidth.	paybybandwidth
service.beta.kubernetes.io/alicloud-loadbalancer-id	The ID of an existing SLB instance. You can set the loadbalancer-id parameter to specify an existing SLB instance. In this case, the existing listeners are overwritten by the SLB instance. The SLB instance is not deleted when the Service is deleted.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-backend-label	The labels that are used to select the nodes to be added as the backend servers of the SLB instance.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-region	The region where the SLB instance is deployed.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-bandwidth	The bandwidth of the SLB instance.	50
service.beta.kubernetes.io/alicloud-loadbalancer-cert-id	The certificate ID. You must first upload the certificate.	""
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-flag	Valid values: on and off.	Default value: off. If TCP is used, do not modify this parameter. The health check feature is automatically enabled for TCP listeners and cannot be disabled.
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-type	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-uri	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-connect-port	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-healthy-threshold	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-unhealthy-threshold	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-interval	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A

Annotation	Description	Default
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-connect-timeout	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A
service.beta.kubernetes.io/alicloud-loadbalancer-health-check-timeout	For more information, see the *CreateLoadBalancerTCPListener* chapter of *SLB Developers Guide*.	N/A

3.1.6.8.3. Configure Ingress monitoring

You can enable the virtual host traffic status (VTS) dashboard to view Ingress monitoring data.

Enable the VTS dashboard by using the CLI

1. Add the following configuration item to the Ingress ConfigMap: `enable-vts-status: "true"` .

```
root@master # kubectl edit configmap nginx-configuration -n kube-system
      configmap "nginx-configuration" edited
```

The following template shows the modified Ingress ConfigMap:

```
apiVersion: v1
data:
  enable-vts-status: "true"# Enables the VTS dashboard.
  proxy-body-size: 20m
kind: ConfigMap
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      {"apiVersion":"v1","data":{"proxy-body-size":"20m"},"kind":"ConfigMap","metadata":{"annotati
ons":{},"labels":{"app":"ingress-nginx"},"name":"nginx-configuration","namespace":"kube-system"}}
  creationTimestamp: 2018-03-20T07:10:18Z
  labels:
    app: ingress-nginx
  name: nginx-configuration
  namespace: kube-system
  selfLink: /api/v1/namespaces/kube-system/configmaps/nginx-configuration
```

2. Verify that the VTS dashboard is enabled for the NGINX Ingress controller.

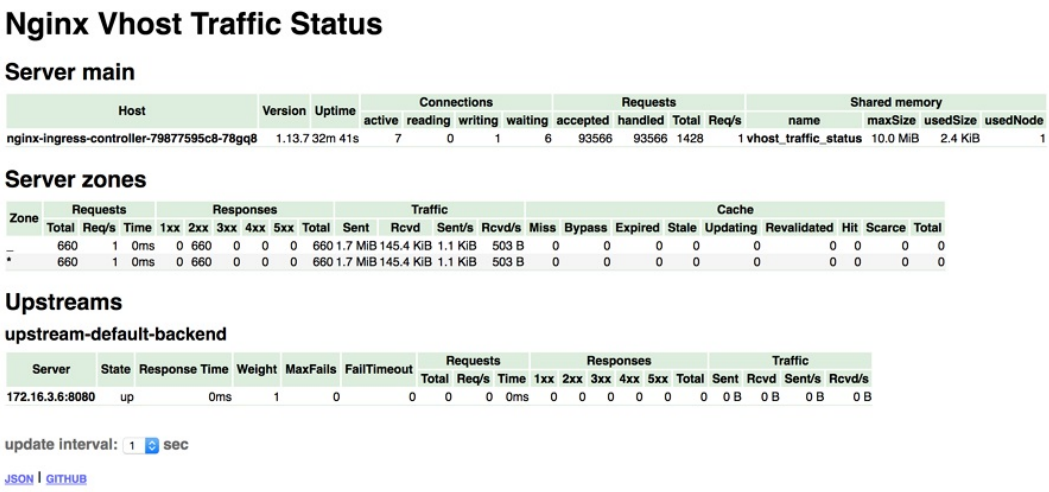
```
root@master # kubectl get pods --selector=app=ingress-nginx -n kube-system
NAME                                       READY     STATUS    RESTARTS   AGE
nginx-ingress-controller-79877595c8-78gq8   1/1       Running   0          1h
root@master # kubectl exec -it nginx-ingress-controller-79877595c8-78gq8 -n kube-system -- cat /et
c/nginx/nginx.conf | grep vhost_traffic_status_display
        vhost_traffic_status_display;
        vhost_traffic_status_display_format html;
```

3. Access the VTS dashboard from an on-premises machine.

> **Note** By default, the VTS port is not exposed due to security concerns. In the following example, port forwarding is used to access the VTS dashboard.

```
root@master # kubectl port-forward nginx-ingress-controller-79877595c8-78gq8 -n kube-system 18080
Forwarding from 127.0.0.1:18080 -> 18080
Handling connection for 18080
```

4. Visit `http://localhost:18080/nginx_status` to access the VTS dashboard.



Nginx Vhost Traffic Status

Server main

Host	Version	Uptime	Connections				Requests				Shared memory			
			active	reading	writing	waiting	accepted	handled	Total	Req/s	name	maxSize	usedSize	usedNode
nginx-ingress-controller-79877595c8-78gq8	1.13.7	32m 41s	7	0	1	6	93566	93566	1428	1	vhost_traffic_status	10.0 MiB	2.4 KiB	1

Server zones

Zone	Requests			Responses					Traffic				Cache							
	Total	Req/s	Time	1xx	2xx	3xx	4xx	5xx	Total	Sent	Rcvd	Sent/s	Rcvd/s	Miss	Bypass	Expired	Stale	Updating	Revalidated	Hit	Scarce	Total
_	660	1	0ms	0	660	0	0	0	660	1.7 MiB	145.4 KiB	1.1 KiB	503 B	0	0	0	0	0	0	0	0	0
*	660	1	0ms	0	660	0	0	0	660	1.7 MiB	145.4 KiB	1.1 KiB	503 B	0	0	0	0	0	0	0	0	0

Upstreams

upstream-default-backend

Server	State	Response Time	Weight	MaxFails	FailTimeout	Requests			Responses						Traffic			
						Total	Req/s	Time	1xx	2xx	3xx	4xx	5xx	Total	Sent	Rcvd	Sent/s	Rcvd/s
172.16.3.6:8080	up	0ms	1	0	0	0	0	0ms	0	0	0	0	0	0	0 B	0 B	0 B	0 B

update interval: 1 sec

JSON | GITHUB

3.1.6.8.4. Ingresses

Kubernetes clusters support Ingress rules. You can define Ingress rules to meet your needs for load balancing.

In Kubernetes clusters, an Ingress is a set of routing rules that authorize external access to Services in clusters. You can use an Ingress to enable Layer 7 load balancing. You can configure an Ingress with URLs, Server Load Balancing (SLB) instances, SSL connections, and name-based virtual hosts to expose Services to external access.

Prerequisites

To test a complex routing scenario, an NGINX application is used in this example. A Deployment and multiple Services are created for the NGINX application. Replace Service names with the actual names.

```
kubectl run nginx --image=registry.cn-hangzhou.aliyuncs.com/acs/netdia:latest
kubectl expose deploy nginx --name=http-svc --port=80 --target-port=80
kubectl expose deploy nginx --name=http-svc1 --port=80 --target-port=80
kubectl expose deploy nginx --name=http-svc2 --port=80 --target-port80
kubectl expose deploy nginx --name=http-svc3 --port=80 --target-port=80
```

Create a simple Ingress

Run the following commands to create a simple Ingress that redirects traffic to the */svc* path to an http-svc Service. `nginx.ingress.kubernetes.io/rewrite-target: /` redirects traffic destined for the */svc* path to the */* path that can be recognized by the backend Service.

```
cat <<EOF | kubectl create -f -
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: simple
  annotations:
    nginx.ingress.kubernetes.io/rewrite-target: /
spec:
  rules:
  - http:
      paths:
      - path: /svc
        backend:
          serviceName: http-svc
          servicePort: 80
EOF
```

```
kubectl get ing
NAME            HOSTS         ADDRESS          PORTS     AGE
simple          *             101.37.19*.***   80        11s
```

You can visit `http://101.37.19*.***/svc` to access the Service of the NGINX application.

Create a simple fanout Ingress that uses multiple domain names

You can create a simple fanout Ingress to route traffic to multiple Services with different domain names. The following example shows the configuration of a simple fanout Ingress:

```
cat <<EOF | kubectl create -f -
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: simple-fanout
spec:
  rules:
  - host: foo.bar.com
    http:
      paths:
      - path: /foo
        backend:
          serviceName: http-svc1
          servicePort: 80
      - path: /bar
        backend:
          serviceName: http-svc2
          servicePort: 80
  - host: foo.example.com
    http:
      paths:
      - path: /film
        backend:
          serviceName: http-svc3
          servicePort: 80
EOF
```

```
kubectl get ing
NAME            HOSTS         ADDRESS          PORTS     AGE
simple-fanout   *             101.37.19*.***   80        11s
```

After the preceding configuration is implemented, you can visit `http://foo.bar.com/foo` to access `http-svc1`, visit `http://foo.bar.com/bar` to access `http-svc2`, and visit `http://foo.example.com/film` to access `http-svc3`.

> **Note**
>
> - In a production environment, you must point the domain name to the returned address `101.37.192.211`.
> - In a test environment, you must add the following mapping rules to the `hosts` file.
>
> ```
> 101.37.19*.*** foo.bar.com
> 101.37.19*.*** foo.example.com
> ```

Create a simple Ingress that uses the default domain name

If you have no domain names, you can create a simple ingress that uses the default domain name provided by Container Service. Then, you can use the default domain name to access Services. The default domain name is in the following format: `*.[cluster-id].[region-id].alicontainer.com`. You can find the default domain name in the basic information of the Kubernetes cluster in the Container Service console.

The following example shows the configuration of a simple Ingress that allows external access to Services through the default domain name:

```
cat <<EOF | kubectl create -f -
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: shared-dns
spec:
  rules:
  - host: foo.[cluster-id].[region-id].alicontainer.com  ## Replace with the default domain name of yo
ur cluster.
    http:
      paths:
      - path: /
        backend:
          serviceName: http-svc1
          servicePort: 80
  - host: bar.[cluster-id].[region-id].alicontainer.com  ## Replace with the default domain name of yo
ur cluster.
    http:
      paths:
      - path: /
        backend:
          serviceName: http-svc2
          servicePort: 80
EOF
```

```
kubectl get ing
NAME            HOSTS         ADDRESS          PORTS     AGE
shared-dns    foo.[cluster-id].[region-id].alicontainer.com,bar.[cluster-id].[region-id].alicontainer.c
om             47.95.16*.***    80        40m
```

You can visit `http://foo.[cluster-id].[region-id].alicontainer.com/` to access Service `http-svc1` and visit `http://bar.[cluster-id].[region-id].alicontainer.com` to access Service `http-svc2`.

Create an Ingress to secure data transmission

Container Service allows you to use multiple types of certificates to reinforce the security of your applications.

1. Prepare a certificate.

   If you do not have a certificate, perform the following steps to generate a test certificate:

   > **Note** The domain name must be the same as the one specified in your Ingress configurations.

   ```
   openssl req -x509 -nodes -days 365 -newkey rsa:2048 -keyout tls.key -out tls.crt -subj "/CN=foo.bar.com/O=foo.bar.com"
   ```

   After you run the preceding command, a certificate file *tls.crt* and a private key file *tls.key* are generated.

   Use the certificate and private key to create a Kubernetes Secret named *foo.bar*. The Secret is referenced when you create the Ingress.

   ```
   kubectl create secret tls foo.bar --key tls.key --cert tls.crt
   ```

2. Create an Ingress to secure data transmission.

   ```
   cat <<EOF | kubectl create -f -
   apiVersion: extensions/v1beta1
   kind: Ingress
   metadata:
     name: tls-fanout
   spec:
     tls:
     - hosts:
       - foo.bar.com
       secretName: foo.bar
     rules:
     - host: foo.bar.com
       http:
         paths:
         - path: /foo
           backend:
             serviceName: http-svc1
             servicePort: 80
         - path: /bar
           backend:
             serviceName: http-svc2
             servicePort: 80
   EOF
   ```

   ```
   kubectl get ing
   NAME            HOSTS         ADDRESS          PORTS     AGE
   tls-fanout      *             101.37.19*.***   80        11s
   ```

3. You must configure the `hosts` file or set a domain name to access the tls Ingress, as described in the **Create a simple fanout Ingress that uses multiple domain names** section.

   You can visit `http://foo.bar.com/foo` to access `http-svc1` and visit `http://foo.bar.com/bar` to access `http-svc2`.

   You can also access the HTTPS Service by using HTTP. By default, an HTTPS Ingress redirects HTTP traffic to HTTPS. Therefore, access to `http://foo.bar.com/foo` is redirected to `https://foo.bar.com/foo`.

Create an Ingress

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Network > Ingresses**.
5. On the **Ingresses** page, select a namespace and click **Create Resources in YAML**.
6. On the **Create** page, select **Custom** from the **Sample Template** drop-down list, copy the following content into the template, and then click **Create**.

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: simple
spec:
  rules:
  - http:
      paths:
      - path: /svc
        backend:
          serviceName: http-svc
          servicePort: 80
```

An Ingress that routes Layer 7 traffic to the `http-svc` Service is created.

3.1.6.8.5. Ingress configurations

Container Service provides Ingress controller components. Integrated with Apsara Server Load Balancer, these components provide Kubernetes clusters with flexible and reliable Ingress service.

An Ingress orchestration template is provided below. When you configure an Ingress through the console, you need to configure annotations and may need to create dependencies. For more information, see Create an ingress through the console, Ingress support, and Kubernetes Ingress. You can also create ConfigMaps to configure Ingresses. For more information, see https://kubernetes.github.io/ingress-nginx/user-guide/nginx-configuration/configmap/.

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  annotations:
    nginx.ingress.kubernetes.io/service-match: 'new-nginx: header("foo", /^bar$/)' #Canary release rul
e. In this example, the request header is used.
    Nginx. ingress. kubernetes. IO/service-weight: 'New-nginx: 50, old-nginx: 50' #The route weight.
  creationTimestamp: null
  generation: 1
  name: nginx-ingress
  selfLink: /apis/extensions/v1beta1/namespaces/default/ingresses/nginx-ingress
spec:
  rules: ##The Ingress rule.
  - host: foo.bar.com
    http:
      paths:
      - backend:
          serviceName: new-nginx
          servicePort: 80
        path: /
      - backend:
          serviceName: old-nginx
          servicePort: 80
        path: /
tls:            ## Enable TLS for secure routing.
  - hosts:
    - *.xxxxxx.cn-hangzhou.alicontainer.com
    - foo.bar.com
    secretName: nginx-ingress-secret     ##The Secret name.
status:
  loadBalancer: {}
```

Annotations

For each Ingress, you can configure its annotations, Ingress controller, and rules, such as the route weight, canary release rule, and rewrite rules. For more information about annotations, see https://kubernetes.github.io/ingress-nginx/user-guide/nginx-configuration/annotations/.

For example, the following rewrite annotation, `nginx.ingress.kubernetes.io/rewrite-target: /`, indicates that */path* is redirected to the root path */*, which can be recognized by the backend service.

Rules

Ingress rules are used to manage external access to the services in the cluster and can be HTTP or HTTPS rules. You can configure the following items in rules: domain name (virtual hostname), URL path, service name, and port.

For each rule, you need to set the following parameters:

- Domain: The test domain or virtual hostname of your service, such as `foo.bar.com`.
- Path: The URL path of your service. Each path is associated with a backend service. Server Load Balancer only forwards traffic to the backend if the incoming request matches the domain and path.
- Service: Specify the service in the form of `service:port`. You also need to specify a route weight for each service. The Ingress routes traffic to the matching service based on the route weight.
  - Name: The name of the backend service.
  - Port: The port of the service.

- Weight: The route weight of the service in the service group.

  > **Note**
  >
  > a. The weight is a percentage value. For example, you can set two services to the same weight of 50%.
  >
  > b. A service group includes services that have the same domain and path defined in the Ingress configuration. If no weight is set for a service, the default value, 100, is used.

Canary release

Container Service supports multiple traffic splitting approaches to suit scenarios such as canary release and A/B testing.

> **Note** Currently, only Ingress controllers of 0.12.0-5 and later versions support traffic splitting.

1. Traffic splitting based on request header
2. Traffic splitting based on cookie
3. Traffic splitting based on query parameter

After canary release is configured, only requests that match certain rules are routed to the corresponding service. If the weight of the corresponding service is lower than 100%, requests that match certain rules are routed to one of the services in the service group based on the weight.

TLS

You can use a Secret that contains a TLS private key and certificate to encrypt the Ingress. This ensures secure routing. The TLS Secret must contain a certificate named tls.crt and a private key named tls.key. For more information about how TLS works, see TLS. For how to create a Secret, see Configure a secure Ingress.

Labels

You can add labels to the Ingress.

3.1.6.8.6. Create an Ingress in the console

The Container Service console is integrated with the Ingress service. You can create an Ingress in the console and manage inbound traffic that is forwarded to different Services to meet your business requirements.

Prerequisites

- A Kubernetes cluster is created and an Ingress controller runs as normal in the cluster. For more information, see Create a Kubernetes cluster.
- You are connected to a master node by using kubectl. For more information, see Connect to a Kubernetes cluster through kubectl.
- Internet access is required when you pull the image from the address specified in this example. You can replace the address with an image address within your cluster. You can also build and push the image to an image repository and then pull the image from the repository when you use the image.

Step 1: Create a Deployment and a Service

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select a namespace and click **Create from Template** in the upper-right corner.

6. On the **Create** page, select a sample template or customize a template and click **Create**.

   In this example, two NGINX applications are created: old-nginx and new-nginx.

   The following template is used to create the old-nginx application:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: old-nginx
spec:
  replicas: 2
  selector:
    matchLabels:
      run: old-nginx
  template:
    metadata:
      labels:
        run: old-nginx
    spec:
      containers:
      - image: registry.cn-hangzhou.aliyuncs.com/xianlu/old-nginx
        imagePullPolicy: Always
        name: old-nginx
        ports:
        - containerPort: 80
          protocol: TCP
      restartPolicy: Always
---
apiVersion: v1
kind: Service
metadata:
  name: old-nginx
spec:
  ports:
  - port: 80
    protocol: TCP
    targetPort: 80
  selector:
    run: old-nginx
  sessionAffinity: None
  type: NodePort
```

   The following template is used to create the new-nginx application:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: new-nginx
spec:
  replicas: 1
  selector:
    matchLabels:
      run: new-nginx
  template:
    metadata:
      labels:
        run: new-nginx
    spec:
      containers:
      - image: registry.cn-hangzhou.aliyuncs.com/xianlu/new-nginx
        imagePullPolicy: Always
        name: new-nginx
        ports:
        - containerPort: 80
          protocol: TCP
      restartPolicy: Always
---
apiVersion: v1
kind: Service
metadata:
  name: new-nginx
spec:
  ports:
  - port: 80
    protocol: TCP
    targetPort: 80
  selector:
    run: new-nginx
  sessionAffinity: None
  type: NodePort
```

7. In the left-side navigation pane of the details page, choose **Network > Services**.

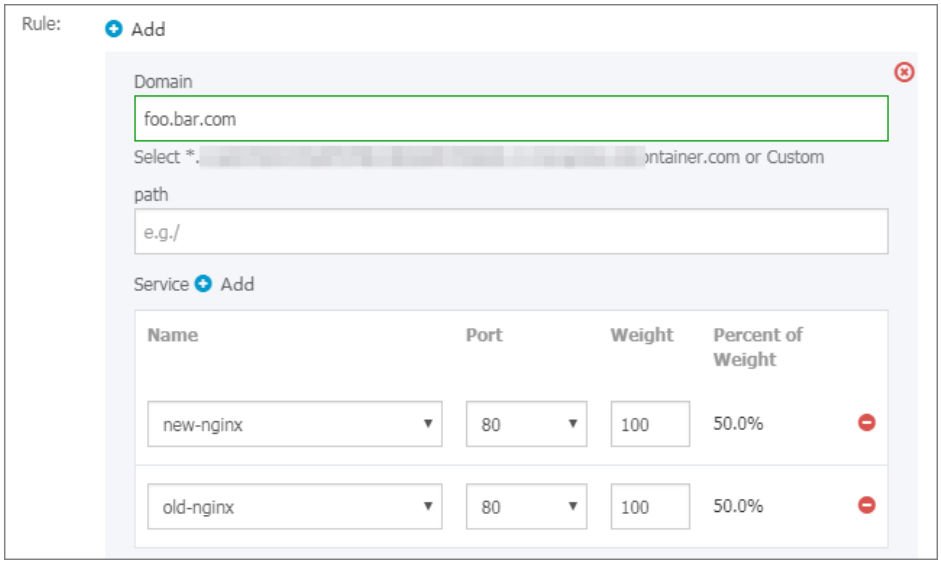
   After the Services are created, you can view the Services on the Services page.

Step 2: Create an Ingress

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. On the details page of the cluster, choose **Network > Ingresses** .
5. On the **Ingresses** page, select a namespace and click **Create** in the upper-right corner.
6. In the dialog box that appears, enter a name for the Ingress. In this example, the Ingress is named nginx-ingress.
7. Configure Ingress rules.

   Ingress rules are used to manage external access to Services in the cluster. Ingress rules can be HTTP or HTTPS rules. You can configure the following items in the rules: domain name (virtual hostname), URL path, Service name, port, and weight. For more information, see Ingress configurations.

   In this example, a complex rule is added to configure Services for the default domain name and virtual hostname of the cluster. Traffic routing is based on domain names.



Create a simple fanout Ingress that uses multiple domain names

In this example, a virtual hostname is used as the test domain name for external access. Route weights are specified for two backend Services and canary release settings are configured for one of the Services. In a production environment, you can use a domain name that has obtained an Internet Content Provider (ICP) number for external access.

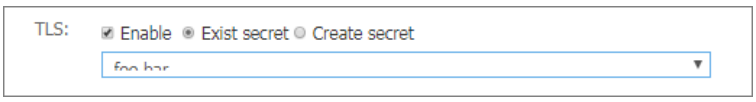
- **Domain**: Enter the test domain name. In this example, the test domain name is `foo.bar.com`.

  You must add the following domain name mapping to the hosts file:

  ```
  118.178.XX.XX foo.bar.com       # The IP address of the Ingress.
  ```

- **Services**: Set the names, paths, port numbers, and weights of the backend Services.
  - Path: Enter the URL of the backend Service. In this example, the root path / is used.
  - Name: In this example, both the old-nginx and new-nginx Services are specified.
  - Port: In this example, port 80 is open.
  - Weight: Set a weight for each backend Service. The weight is a percentage value. The default value is 100. In this example, the weight of each backend Service is set to 50. This means that the two backend Services have the same weight.

8. Configure Transport Layer Security (TLS). Select **EnableTLS** to enable TLS and configure a secure Ingress. For more information, see Configure a secure Ingress.
   - You can use an existing Secret.

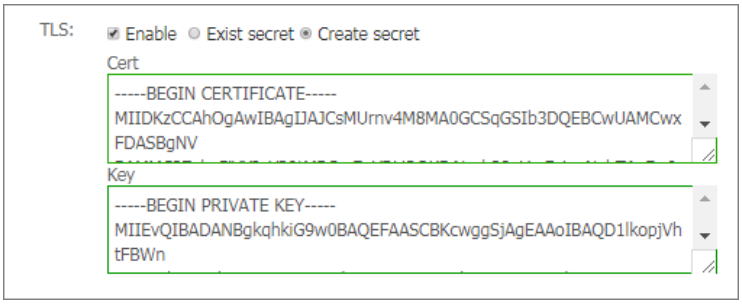
     a. Log on to a master node. Create a file named *tls.key* and another file named *tls.crt*.

     ```
     openssl req -x509 -nodes -days 365 -newkey rsa:2048 -keyout tls.key -out tls.crt -subj "/CN
     =foo.bar.com/O=foo.bar.com"
     ```

     b. Create a Secret.

     ```
     kubectl create secret tls foo.bar --key tls.key --cert tls.crt
     ```

     c. Run the `kubectl get secret` command and verify that the Secret is created. Then, you can select the newly created Secret *foo.bar*.
   - You can also use the TLS private key and certificate to create a Secret.



a. Log on to a master node, and then create a file named *tls.key* and another file named *tls.crt*.

```
openssl req -x509 -nodes -days 365 -newkey rsa:2048 -keyout tls.key -out tls.crt -subj "/CN
=foo.bar.com/O=foo.bar.com"
```

b. Run the `vim tls.key` and `vim tls.crt` commands to obtain the private key and certificate that are generated.

c. Copy the certificate to the Cert field and the private key to the Key field.

9. Configure canary release settings.

> **Note** Only Ingress controllers of 0.12.0-5 and later versions support traffic splitting.

Container Service supports multiple traffic splitting methods. This allows you to select suitable solutions for specific scenarios, such as canary releases and A/B testing:

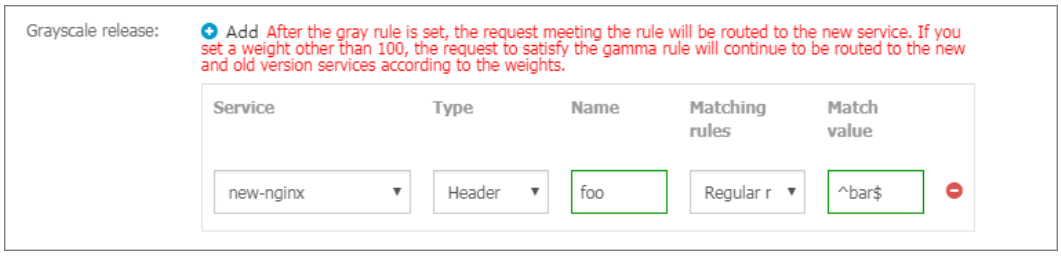
i. Traffic splitting based on request headers

ii. Traffic splitting based on cookies

iii. Traffic splitting based on query parameters

After canary release is configured, only requests that match the specified rules are routed to the new-nginx Service. If the weight of new-nginx is lower than 100%, requests that match the specified rules are routed to the Service based on the Service weight.

In this example, the rule is added to specify that only request headers with headers that match the regular expression `foo=^bar$` are forwarded to new-nginx.



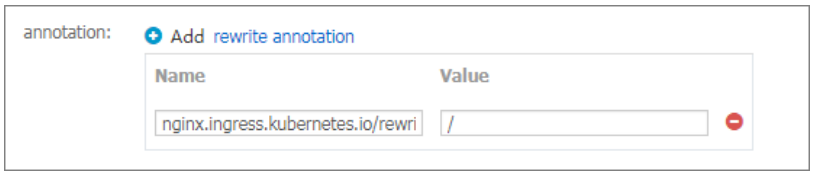
- **Services**: Specify the Services to be accessed.
- **Type**: Select the type of matching rule. Valid values: Header, Cookie, and Query.
- **Name and Match Value**: Specify the names and matching values of custom request fields in key-value pairs.
- **Matching Rule**: Regular expressions and exact matches are supported.

10. Configure annotations.

Click **Rewrite Annotation** and add an annotation to redirect inbound traffic for the Ingress. For example, `nginx.ingress.kubernetes.io/rewrite-target: /` specifies that */path* is redirected to the root path */* that can be recognized by the backend Services.

> **Note** In this example, no path is configured for the backend Services. Therefore, you do not need to configure rewrite annotations. Rewrite annotations allow the Ingress to forward traffic through the root path to the backend Services. This avoids the 404 error that is caused by invalid paths.

You can also click **Add** to enter annotation names and values in key-value pairs. For more information about Ingress annotations, visit https://kubernetes.github.io/ingress-nginx/user-guide/nginx-configuration/annotations/.

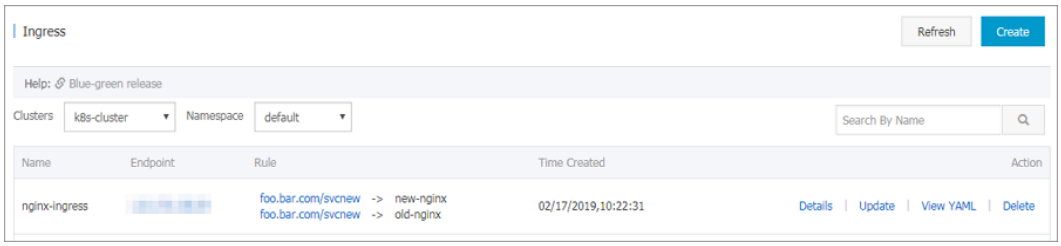


11. Add labels.

    Add labels to describe the characteristics of the Ingress.

12. Click **Create**. You are redirected to the Ingresses page.

    The newly created Ingress appears on the Ingresses page.



13. Click foo.bar.com to visit the NGINX welcome page.

    Click the domain name that points to new-nginx. The old-nginx application page appears.

> **Note** By default, when you enter the domain name in the browser, request headers do not match the regular expression `foo=^bar$`. Therefore, the requests are directed to old-nginx.

14. Log on to a master node by using SSH. Run the following commands to simulate requests with specific headers and check the results:

```
curl -H "Host: foo.bar.com" http://47.107.XX.XX
old
 curl -H "Host: foo.bar.com" http://47.107.XX.XX
old
 curl -H "Host: foo.bar.com" http://47.107.XX.XX                    # Similar to a browser reques
t.
old
 curl -H "Host: foo.bar.com" -H "foo: bar" http://47.107.XX.XX          # Simulate a request with
a specific header. The results are returned based on the weight.
new
 curl -H "Host: foo.bar.com" -H "foo: bar" http://47.107.XX.XX
old
 curl -H "Host: foo.bar.com" -H "foo: bar" http://47.107.XX.XX
old
 curl -H "Host: foo.bar.com" -H "foo: bar" http://47.107.XX.XX
new
```

3.1.6.8.7. Update an Ingress

You can update Ingresses in the Container Service console.

Prerequisites

- A Kubernetes cluster is created and an Ingress controller is running as normal in the cluster. For more information, see Create a Kubernetes cluster.
- An Ingress is created. For more information, see Create an ingress through the console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Network > Ingresses**.
5. On the **Ingresses** page, select the namespace, find the Ingress that you want to update, and then click **Update** in the Actions column.
6. In the dialog box that appears, modify the parameters and click **Update**. In this example, `foo.bar.com` is changed to `test.bar.com`.
   On the Ingresses page, you can verify that the Ingress rule has changed.

3.1.6.8.8. Delete an Ingress

This topic describes how to delete an Ingress.

Prerequisites

- A Kubernetes cluster is created and an Ingress controller is running as normal in the cluster. For more information about how to create a cluster, see Create a Kubernetes cluster.
- An Ingress is created. For more information, see Create an ingress through the console.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Network > Ingresses**.

5. On the **Ingresses** page, select the namespace, find the Ingress that you want to delete, and then click **Delete** in the Actions column.
6. In the message that appears, click **Confirm**.

3.1.6.9. Config maps and secrets

3.1.6.9.1. Create a ConfigMap

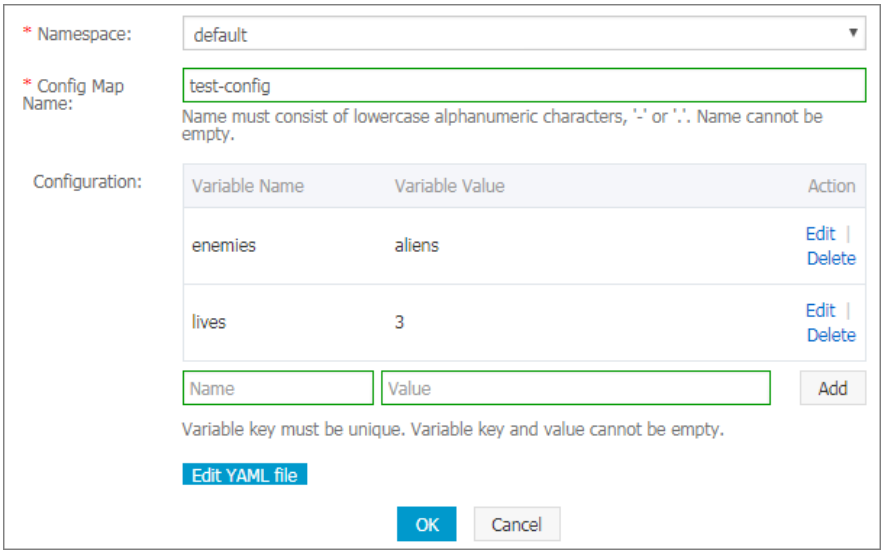
In the Container Service console, you can create a ConfigMap on the ConfigMap page or by using a template. This topic describes how to create a ConfigMap.

Create a ConfigMap on the ConfigMap page

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Configurations > ConfigMaps**.
5. On the **ConfigMap** page, select a namespace and click **Create**.
6. Set the parameters and click **OK**. Create a ConfigMap on the ConfigMap page

Parameter	Description
Clusters	The ID of the cluster that you have selected.
Namespaces	The namespace that you have selected. A ConfigMap is a Kubernetes resource object and must be scoped to a namespace.
ConfigMap Name	The name of the ConfigMap. The name can contain lowercase letters, digits, hyphens (-), and periods (.). This parameter is required. Other resource objects must reference the ConfigMap name to obtain the configuration information.
ConfigMap	Specify **Name** and **Value**, and then click **Add** to add the key-value pair. You can also click **Edit YAML file**, modify the parameters in the dialog box that appears, and then click **OK**.

In this example, two variables named enemies and lives are created. Their values are set to aliens and 3 separately.



7. Click **OK**. You can find the test-config ConfigMap on the ConfigMap page.

You can also click **Browse** and upload a configuration file to create a ConfigMap.

Create a ConfigMap from a template

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select the namespace and click **Create from Template** in the upper-right corner.
6. On the page that appears, set the parameters and click **Create**. Create a ConfigMap from a template

Parameter	Description
Sample Template	Container Service provides YAML templates of various resource types to help you quickly deploy resource objects. You can select *Custom* from the drop-down list and configure your ConfigMap based on YAML syntax. You can also select the *Resource-ConfigMap* template to create a ConfigMap. In the sample template, the ConfigMap is named aliyun-config and contains two variable files: *game.properties* and *ui.properties*. You can modify the ConfigMap based on your needs.
Template	Enter the template content based on YAML syntax. The template can contain multiple resource objects that are separated by `---`.
Add Deployment	This feature allows you to quickly define a YAML template. You can click **Use Existing Template** to import an existing template.

After the deployment is completed, you can find the ConfigMap named *aliyun-config* on the ConfigMap page.

3.1.6.9.2. Use a ConfigMap in a pod

This topic describes how to use a ConfigMap in a pod.

You can use a ConfigMap in a pod in the following scenarios:

- Use a ConfigMap to define environment variables for a pod.
- Use a ConfigMap to set command line parameters.
- Use a ConfigMap in a volume.

For more information, see Configure a pod to use a ConfigMap.

Limits

To use a ConfigMap in a pod, make sure that the ConfigMap and the pod are in the same cluster and namespace.

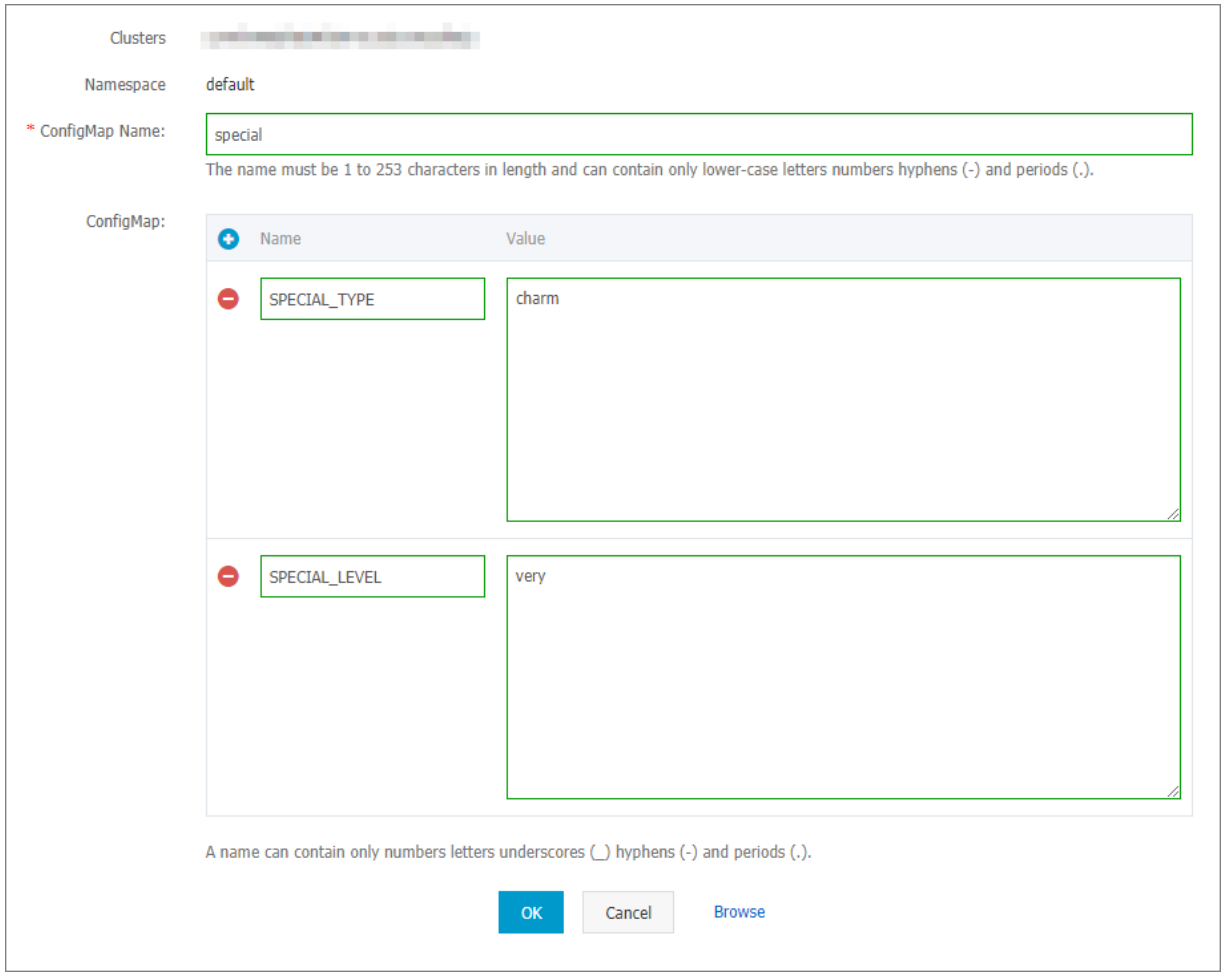
Create a ConfigMap

In this example, a ConfigMap named special_config is created. This ConfigMap consists of two key-value pairs: `SPECIAL_LEVEL: very` and `SPECIAL_TYPE: charm`.

You can use the following YAML template to create the ConfigMap:

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: special-config
  namespace: default
data:
  SPECIAL_LEVEL: very
  SPECIAL_TYPE: charm
```

You can also log on to the Container Service console and choose **Configuration > ConfigMaps** in the left-side navigation pane. You can then click **Create** to create the ConfigMap.



Use a ConfigMap to define one or multiple environment variables for a pod

Use a key-value pair of a ConfigMap to define one environment variable

You can log on to the Container Service console. In the left-side navigation pane, click **Clusters**. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column. In the left-side navigation pane of the details page, choose **Workloads > Deployments**. Click **Create from Template**, select a pod template from the Sample Template drop-down list, and then start the deployment.

You can use the following sample template to create a pod and defines environment variables in the pod. valueFrom is used to reference the value of SPECIAL_LEVEL to define an environment variable.

```
apiVersion: v1
kind: Pod
metadata:
  name: config-pod-1
spec:
  containers:
    - name: test-container
      image: busybox
      command: [ "/bin/sh", "-c", "env" ]
      env:
        - name: SPECIAL_LEVEL_KEY
          valueFrom:                              ##valueFrom is used to reference the value of the Con
figMap to define an environment variable.
            configMapKeyRef:
              name: special-config                ##The name of the referenced ConfigMap.
              key: SPECIAL_LEVEL                  ##The key of the referenced key-value pair.
  restartPolicy: Never
```

To use the values of multiple ConfigMaps to define multiple environment variables, add multiple env parameters to the pod configuration file.

Use all the key-value pairs of a ConfigMap to define multiple environment variables

To define the key-value pairs of a ConfigMap as pod environment variables, you can use the envFrom parameter. The keys in a ConfigMap are used as the names of the environment variables.

The following sample template is used to create a pod:

```
apiVersion: v1
kind: Pod
metadata:
  name: config-pod-2
spec:
  containers:
    - name: test-container
      image: busybox
      command: [ "/bin/sh", "-c", "env" ]
      envFrom:                ##Reference all the key-value pairs of the sepcial-config ConfigMap.
      - configMapRef:
        name: special-config
  restartPolicy: Never
```

Use a ConfigMap to set command line parameters

You can use ConfigMaps to define the commands or parameter values for a container by using the environment variable replacement syntax *$(VAR_NAME)*. The following template is used as an example:

```
apiVersion: v1
kind: Pod
metadata:
  name: config-pod-3
spec:
  containers:
    - name: test-container
      image: busybox
      command: [ "/bin/sh", "-c", "echo $(SPECIAL_LEVEL_KEY) $(SPECIAL_TYPE_KEY)" ]
      env:
        - name: SPECIAL_LEVEL_KEY
          valueFrom:
            configMapKeyRef:
              name: special-config
              key: SPECIAL_LEVEL
        - name: SPECIAL_TYPE_KEY
          valueFrom:
            configMapKeyRef:
              name: special-config
              key: SPECIAL_TYPE
  restartPolicy: Never
```

After you run the pod, the following output is returned:

```
very charm
```

Use a ConfigMap in a volume

You can use a ConfigMap to define volumes. The following sample template specifies a ConfigMap name under volumes. This stores the key-value pairs of the ConfigMap to the path that you specified in the mountPath field. In this example, the path is /etc/config. This generates configuration files that are named after the keys of the ConfigMap. The corresponding values of the ConfigMap are stored in these files.

```
apiVersion: v1
kind: Pod
metadata:
  name: config-pod-4
spec:
  containers:
    - name: test-container
      image: busybox
      command: [ "/bin/sh", "-c", "ls /etc/config/" ]   ##List the files under the directory.
      volumeMounts:
        - name: config-volume
          mountPath: /etc/config
   volumes:
     - name: config-volume
       configMap:
         name: special-config
  restartPolicy: Never
```

After you run the pod, the following output is returned:

```
SPECIAL_TYPE
SPECIAL_LEVEL
```

3.1.6.9.3. Update a ConfigMap

You can use multiple methods to update a ConfigMap.

Considerations

If you update a ConfigMap, the applications that use the ConfigMap are affected.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Configurations > ConfigMaps**.
5. On the **ConfigMap** page, select the namespace, find the ConfigMap that you want to update, and then click **Edit** in the Actions column.
6. In the dialog box that appears, modify the configurations and click **OK**.

3.1.6.9.4. Delete a ConfigMap

You can use multiple methods to delete a ConfigMap.

Considerations

If you delete a ConfigMap, the applications that use this ConfigMap are affected.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Configurations > ConfigMaps**.
5. On the **ConfigMap** page, select the namespace, find the ConfigMap that you want to delete, and then click **Delete** in the **Actions** column.
6. In the message that appears, click **OK**.

3.1.6.9.5. Create a Secret

You can create Secrets for applications in the Container Service console.

Prerequisites

A Kubernetes cluster is created.

Context

We recommend that you use Secrets to store sensitive information in Kubernetes clusters, such as passwords and certificates.

Secrets are classified into the following types:

- Service account: A service account is automatically created by Kubernetes and automatically mounted to the */run/secrets/kubernetes.io/serviceaccount* directory of a pod. The service account provides an identity for the pod to interact with the API server.
- Opaque: This type of secret is encoded in Base64 and used to store sensitive information, such as passwords and certificates.

By default, you can create only Opaque Secrets in the Container Service console. Opaque Secrets store map type data. Therefore, values must be encoded in Base64. You can create Secrets in the Container Service console with a few clicks. Plaintext is automatically encoded in Base64.

You can also create Secrets by using the CLI. For more information, see Kubernetes Secrets.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Configurations > Secrets**.
5. On the **Secrets** page, select the namespace and click **Create** in the upper-right corner.
6. Configure the Secret and click **OK**.

   Note To enter Secret data in plaintext, select **Encode Data Values Using Base64**.

Parameter	Description
Name	Enter a name for the Secret. The name must be 1 to 253 characters in length, and can contain only lowercase letters, digits, hyphens (-), and periods (.).
Namespace	Select the namespace of the Secret.
Type	You can select Opaque, Private Repository Logon Secret, or TLS Certificate.
Opaque	If you set Type to Opaque, configure the following parameters: ○ (Optional)To enter Secret data in plaintext, select **Encode Data Values Using Base64**. ○ Configure the Secret in key-value pairs. Click **+ Add**. Enter the keys and values for the Secret in the **Name** and **Value** fields.
Private Repository Logon Secret	If you set Type to Private Repository Logon Secret, configure the following parameters: ○ Docker Registry URL: Enter the address of the Docker registry where your Secret is stored. ○ Username: Enter the username that is used to log on to the Docker registry. ○ Password: Enter the password that is used to log on to the Docker registry.
TLS Certificate	If you set Type to TLS Certificate, configure the following parameters: ○ Cert: Enter a TLS certificate. ○ Key: Enter the key for the TLS certificate.

You can view the newly created Secret on the Secrets page.

3.1.6.9.6. Modify a Secret

This topic describes how to modify a Secret in the Container Service console.

Prerequisites

- A Kubernetes cluster is created.
- A Secret is created. For more information, see Create a secret.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Configurations > Secrets**.
5. On the Secrets page, select the namespace, find the Secret that you want to modify, and then click **Edit** in the Actions column.
6. In the dialog box that appears, modify the Secret and click **OK**.

3.1.6.9.7. Delete a Secret

This topic describes how to delete a Secret in the Container Service console.

Prerequisites

- A Kubernetes cluster is created.
- A Secret is created. For more information, see Create a secret.

Context

> **Note** Do not delete Secrets that are generated when the cluster is created.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Configurations > Secrets**.
5. On the **Secrets** page, select the namespace, find the Secret that you want to delete, and then click **Delete** in the **Actions** column.
6. In the message that appears, click **OK**.

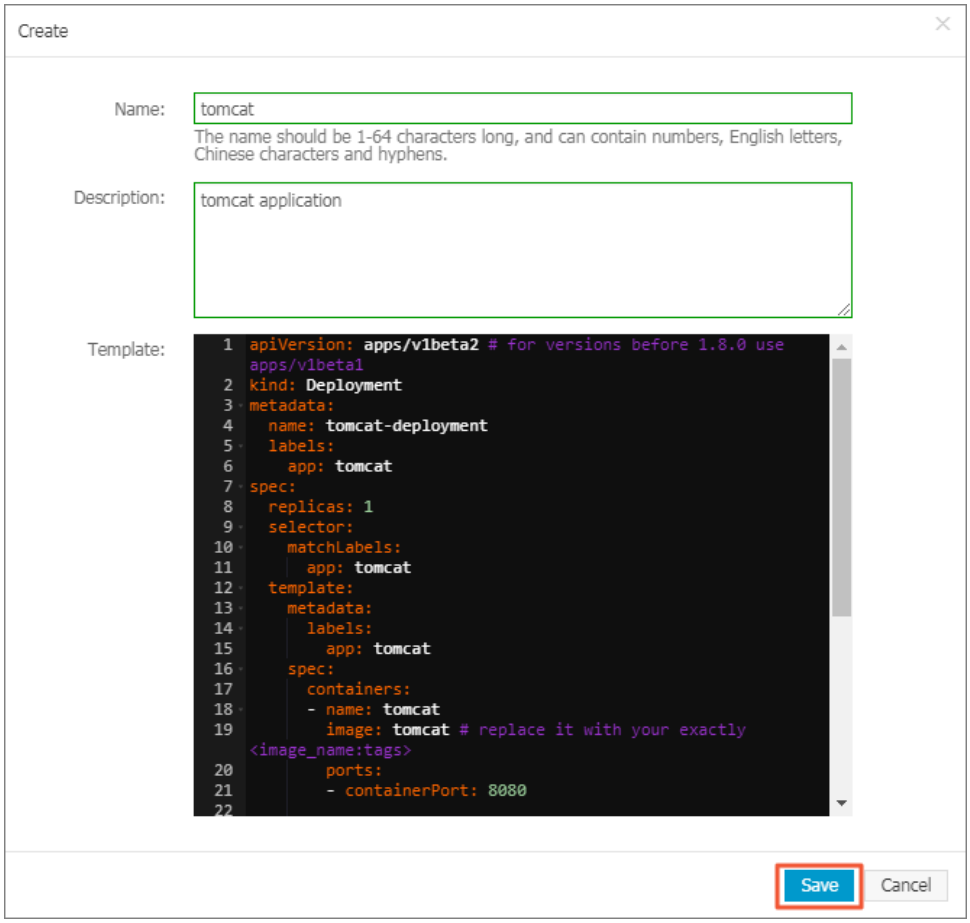
3.1.6.10. Templates

3.1.6.10.1. Create an orchestration template

This topic describes how to use multiple methods to create orchestration templates through the Container Service console.

Procedure

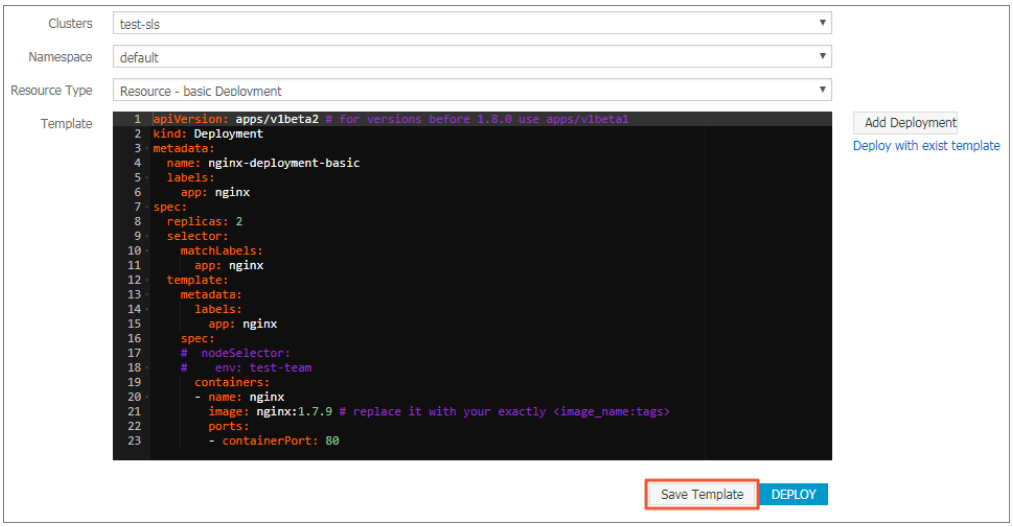
1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > Orchestration Templates** and click **Create** in the upper-right corner.
3. In the dialog box that appears, configure the template, and then click **Save**. This example demonstrates how to create a Tomcat application template that contains a deployment and a service.
   - **Name**: The name of the template.
   - **Description**: Optional. The description of the template.
   - **Template**: Enter the template content based on YAML syntax. The template can contain multiple resource objects that are separated by `---`.



4. After the template is created, you are redirected to the **Templates** page by default. You can find the template on the **My Templates** tab.

5. (Optional)You can also choose **Applications > Deployments** in the left-side navigation pane, and click **Create from Template** to go to the **Create from Template** page. You can modify a built-in template provided by Container Service and save it as a custom template.

i. Select a built-in template and click **Save Template**.



ii. In the dialog box that appears, specify the name, description, and content. Click **Save** to save the template.

> **Note** You can modify the built-in template based on your needs.

iii. In the left-side navigation pane, choose **Marketplace > Orchestration Templates**. You can find the newly created template on the **My Templates** tab.

What's next

You can use the orchestration templates on the **My Templates** tab to quickly create applications.

3.1.6.10.2. Update an orchestration template

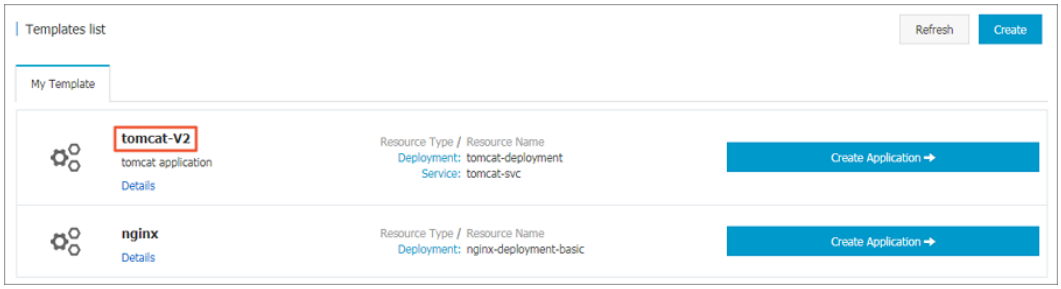
This topic describes how to edit and update an orchestration template.

Prerequisites

You have created an orchestration template. For more information, see Create orchestration templates.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > Orchestration Templates**. The **Templates** page appears. You can view existing templates on the **My Templates** tab.
3. Select the target template and click **Details**.
4. On the template details page, click **Edit** in the upper-right corner.
5. In the dialog box that appears, edit the name, description, and template content, and click **Save**.
6. Go to the **Templates** page. You can view the template that you have updated on the **My Templates** tab.



3.1.6.10.3. Save an orchestration template as a new one

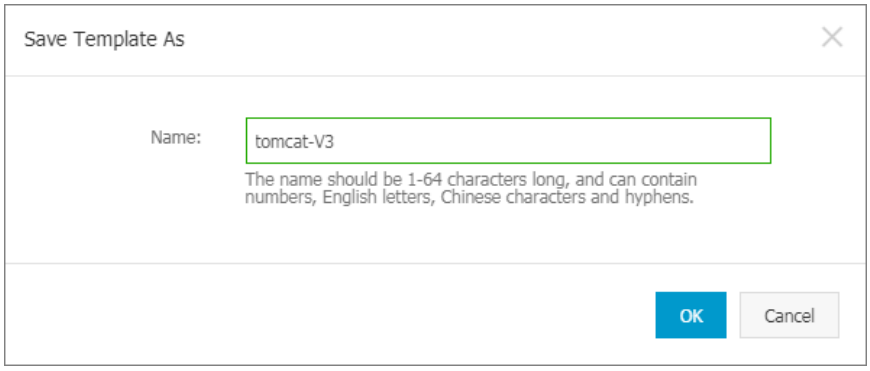
This topic describes how to save an orchestration template as a new one.

Prerequisites

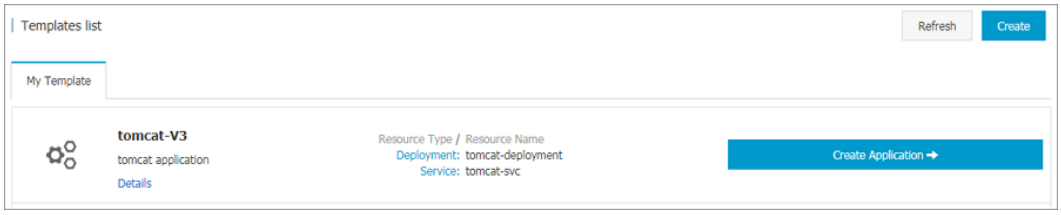
You have created an orchestration template. For more information, see Create orchestration templates.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > Orchestration Templates**. The **Templates** page appears. You can view existing templates on the **My Templates** tab.
3. Select the target template and click **Details**.
4. On the template details page, modify the template and click **Save As** in the upper-right corner.
5. In the dialog box that appears, enter the template name and click **OK**.



6. Go to the **Templates** page. The newly saved template is displayed on the **My Templates** tab.



3.1.6.10.4. Download an orchestration template

This topic describes how to download an orchestration template.

Prerequisites

You have created an orchestration template. For more information, see Create orchestration templates.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > Orchestration Templates**. The **Templates** page appears. You can view existing templates on the **My Templates** tab.
3. Select the target template and click **Details**.
4. On the template details page, click **Download** in the upper-right corner to download the template as a YAML file.

3.1.6.10.5. Delete an orchestration template

Prerequisites

You have created an orchestration template. For more information, see Create orchestration templates.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > Orchestration Templates**. The **Templates** page appears. You can view existing templates on the **My Templates** tab.
3. Select the target template and click **Details**.
4. On the template details page, click **Delete** in the upper-right corner.
5. In the dialog box that appears, click **OK**.

3.1.6.11. Log management

3.1.6.11.1. Use Log Service to collect log data from containers

Container Service is integrated with Log Service. When you create a cluster, you can enable Log Service to collect log data from containers, including standard output (stdout) and text files.

Activate Log Service

To activate Log Service, perform the following steps:

1. Log on to the Apsara Uni-manager Management Console. In the top navigation bar, choose **Products > Log Service** to go to the **Log Service** page.
2. Select the required organization and region.
3. Click **SLS** to go to the Log Service console.

Create a Kubernetes cluster that has Log Service enabled

To create a Kubernetes cluster, perform the following steps:

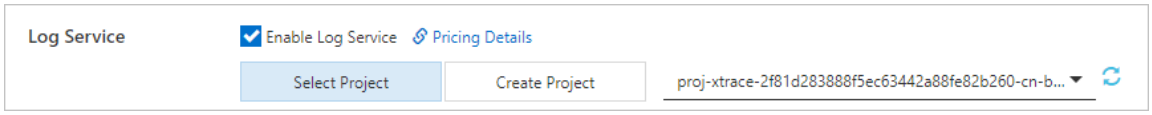
1. Log on to the Container Service console.

   > **Note** The specified organization must be the same as the one that you selected when you activated Log Service. For more information, see Activate Log Service.

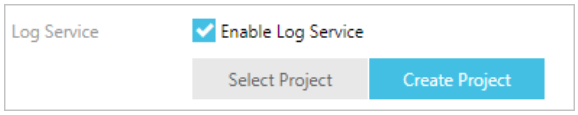
2. In the left-side navigation pane, click **Clusters**.
3. On the Clusters page, click **Create Kubernetes Cluster**. For more information, see Create a Kubernetes cluster.
4. In the lower part of the page, select **Enable Log Service** to install the Log Service component.

When the **Enable Log Service** check box is selected, the system prompts you to create a Log Service project. Select one of the following methods to specify a project in one of the following ways:

- You can select an existing project to manage the collected logs.



- You can click Create Project. Then, a project named `k8s-log-{ClusterID}` is automatically created to manage the collected logs. ClusterID indicates the unique ID of the cluster to be created.



5. Click **Create Cluster** in the upper-right corner of the page.
6. On the **Confirm** page, after all check items are verified, select the terms of service and disclaimer and click **OK** to start the deployment.
   On the Clusters page, you can find the created cluster.

Install the Log Service component in an existing Kubernetes cluster

If you created a Kubernetes cluster and activated Log Service, you can perform the following steps to enable Log Service:

1. Connect to the Kubernetes cluster by using CloudShell.

   For more information, see Connect to a Kubernetes cluster through kubectl.
2. Run the *logtail-dedicated.sh* script to install the Log Service component in the Kubernetes cluster.

```
#!/env/bin/bash
yaml=$(cat <<-END
---
apiVersion: v1
kind: ConfigMap
metadata:
  name: alibaba-log-config-file
  namespace: kube-system
data:
  ilogtail_config.json: |
    {
      "config_server_address" : "http://logtail.$REGION.sls-pub.$INTERNET_DOMAIN",
      "data_server_address" : "http://data.$REGION.sls-pub.$INTERNET_DOMAIN",
      "data_server_list" :
      [
          {
              "cluster" : "$REGION",
              "endpoint" : "data.$REGION.sls-pub.$INTERNET_DOMAIN"
          }
      ],
      "shennong_unix_socket" : false
    }
---
apiVersion: v1
kind: ConfigMap
metadata:
  name: alibaba-log-configuration
  namespace: kube-system
data:
    log-project: "k8s-log-$CLUSTER_ID"
    log-endpoint: "data.$REGION.sls-pub.$INTERNET_DOMAIN"
```

```
    log-machine-group: "k8s-group-$CLUSTER_ID"
    log-config-path: "/etc/ilogtail/conf/apsara/ilogtail_config.json"
    log-ali-uid: "$ALI_UID"
    log-access-id: "" # just use blank string
    log-access-key: "" # just use blank string
---
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: alibaba-log-controller
  namespace: kube-system
  labels:
    k8s-app: alibaba-log-controller
  annotations:
    component.version: "v0.1.3"
    component.revision: "v1"
spec:
  replicas: 1
  template:
    metadata:
      labels:
        k8s-app: alibaba-log-controller
      annotations:
        scheduler.alpha.kubernetes.io/critical-pod: ''
    spec:
      serviceAccountName: alibaba-log-controller
      tolerations:
        - operator: "Exists"
      containers:
      - name: alibaba-log-controller
        image: $IMAGE_REPO_URL/acs/log-controller-$ARCH:v0.1.3.0-527ff4d-aliyun
        resources:
          limits:
            memory: 100Mi
          requests:
            cpu: 50m
            memory: 100Mi
        env:
          - name: "ALICLOUD_LOG_PROJECT"
            valueFrom:
              configMapKeyRef:
                name: alibaba-log-configuration
                key: log-project
          - name: "ALICLOUD_LOG_ENDPOINT"
            valueFrom:
              configMapKeyRef:
                name: alibaba-log-configuration
                key: log-endpoint
          - name: "ALICLOUD_LOG_MACHINE_GROUP"
            valueFrom:
              configMapKeyRef:
                name: alibaba-log-configuration
                key: log-machine-group
          - name: "ALICLOUD_ACS_K8S_FLAG"
            value: "ture"
          - name: "ALICLOUD_ACCESS_KEY_ID"
            valueFrom:
              configMapKeyRef:
                name: alibaba-log-configuration
                key: log-access-id
```

```
          - name: "ALICLOUD_ACCESS_KEY_SECRET"
            valueFrom:
              configMapKeyRef:
                name: alibaba-log-configuration
                key: log-access-key
      nodeSelector:
        beta.kubernetes.io/os: linux
---
apiVersion: apiextensions.k8s.io/v1beta1
kind: CustomResourceDefinition
metadata:
  name: aliyunlogconfigs.log.alibabacloud.com
spec:
  group: log.alibabacloud.com
  version: v1alpha1
  names:
    kind: AliyunLogConfig
    plural: aliyunlogconfigs
  scope: Namespaced
---
apiVersion: rbac.authorization.k8s.io/v1beta1
kind: ClusterRoleBinding
metadata:
  name: alibaba-log-controller
subjects:
- kind: ServiceAccount
  name: alibaba-log-controller
  namespace: kube-system
roleRef:
  kind: ClusterRole
  name: alibaba-log-controller
  apiGroup: rbac.authorization.k8s.io
---
apiVersion: rbac.authorization.k8s.io/v1beta1
kind: ClusterRole
metadata:
  name: alibaba-log-controller
  labels:
    k8s-app: alibaba-log-controller
rules:
- apiGroups: ["log.alibabacloud.com"]
  resources:
  - aliyunlogconfigs
  verbs:
  - update
  - get
  - watch
  - list
- apiGroups: [""]
  resources:
  - configmaps
  verbs:
  - create
  - update
  - get
- apiGroups: [""]
  resources:
  - events
  verbs:
  - create
```

```
  - patch
  - update
---
apiVersion: v1
kind: ServiceAccount
metadata:
  name: alibaba-log-controller
  namespace: kube-system
  labels:
    k8s-app: alibaba-log-controller
---
apiVersion: extensions/v1beta1
kind: DaemonSet
metadata:
  name: logtail-ds
  namespace: kube-system
  labels:
    k8s-app: logtail-ds
  annotations:
    component.version: "v0.16.16"
    component.revision: "v0"
spec:
  updateStrategy:
    type: RollingUpdate
  template:
    metadata:
      labels:
        k8s-app: logtail-ds
      annotations:
        scheduler.alpha.kubernetes.io/critical-pod: ''
    spec:
      tolerations:
        - operator: "Exists"
      containers:
      - name: logtail
        image: $IMAGE_REPO_URL/acs/logtail-$ARCH:v0.16.24.0-c46cd2fe-aliyun
        resources:
          limits:
            memory: 512Mi
          requests:
            cpu: 100m
            memory: 256Mi
        livenessProbe:
          exec:
            command:
            - /etc/init.d/ilogtaild
            - status
          initialDelaySeconds: 30
          periodSeconds: 30
        securityContext:
          privileged: false
        env:
          - name: "ALIYUN_LOGTAIL_CONFIG"
            valueFrom:
              configMapKeyRef:
                name: alibaba-log-configuration
                key: log-config-path
          - name: "ALIYUN_LOGTAIL_USER_ID"
            valueFrom:
              configMapKeyRef:
```

```
              name: alibaba-log-configuration
              key: log-ali-uid
        - name: "ALIYUN_LOGTAIL_USER_DEFINED_ID"
          valueFrom:
            configMapKeyRef:
              name: alibaba-log-configuration
              key: log-machine-group
        - name: "ALICLOUD_LOG_DOCKER_ENV_CONFIG"
          value: "true"
        - name: "ALICLOUD_LOG_ECS_FLAG"
          value: "ture"
        - name: "ALICLOUD_LOG_DEFAULT_PROJECT"
          valueFrom:
            configMapKeyRef:
              name: alibaba-log-configuration
              key: log-project
        - name: "ALICLOUD_LOG_ENDPOINT"
          valueFrom:
            configMapKeyRef:
              name: alibaba-log-configuration
              key: log-endpoint
        - name: "ALICLOUD_LOG_DEFAULT_MACHINE_GROUP"
          valueFrom:
            configMapKeyRef:
              name: alibaba-log-configuration
              key: log-machine-group
        - name: "ALICLOUD_LOG_ACCESS_KEY_ID"
          valueFrom:
            configMapKeyRef:
              name: alibaba-log-configuration
              key: log-access-id
        - name: "ALICLOUD_LOG_ACCESS_KEY_SECRET"
          valueFrom:
            configMapKeyRef:
              name: alibaba-log-configuration
              key: log-access-key
        - name: "ALIYUN_LOG_ENV_TAGS"
          value: "_node_name_|_node_ip_"
        - name: "_node_name_"
          valueFrom:
            fieldRef:
              fieldPath: spec.nodeName
        - name: "_node_ip_"
          valueFrom:
            fieldRef:
              fieldPath: status.hostIP
      volumeMounts:
      - name: sock
        mountPath: /var/run/docker.sock
      - name: root
        mountPath: /logtail_host
        readOnly: true
      - name: alibaba-log-config-file-volume
        mountPath: /etc/ilogtail/conf/apsara
        readOnly: true
    terminationGracePeriodSeconds: 30
    nodeSelector:
      beta.kubernetes.io/os: linux
    volumes:
    - name: sock
```

```
        hostPath:
          path: /var/run/docker.sock
          type: Socket
      - name: root
        hostPath:
          path: /
          type: Directory
      - name: alibaba-log-config-file-volume
        configMap:
          name: alibaba-log-config-file
END
)
echo "$yaml" > logtail.yml
kubectl create -f logtail.yml
```

3. Replace `<your_server_architecture>`, `<your_k8s_cluster_region_id>`, `<your_k8s_cluster_id>`, `<k8s_cluster_domain_suffix>`, `<your_ali_uid>`, and `<your_image_repo_url>` with actual values, and run the following commands. This allows you to set the environment variables and deploy the component.

```
export ARCH=<your_server_architecture>
export REGION=<your_k8s_cluster_region_id>
export CLUSTER_ID=<your_k8s_cluster_id>
export INTERNET_DOMAIN=<k8s_cluster_domain_suffix>
export IMAGE_REPO_URL=<your_image_repo_url>
export ALI_UID=<your_ali_uid>
bash logtail-dedicated.sh // Run the script to install the component.
```

> **Note**
>
> - `<your_server_architecture>`: the server architecture, for example, amd64.
> - `<your_k8s_cluster_region_id>`: the region where the Kubernetes cluster is deployed, for example, cn-qingdao-apsara-d01.
> - `<your_k8s_cluster_id>`: the ID of the Kubernetes cluster.
> - `<k8s_cluster_domain_suffix>`: the domain suffix of the Kubernetes cluster, for example, env28.internet.com.
> - `<your_ali_uid>`: the ID of the Apsara Stack tenant account, for example, 1234074238634394.
> - `<your_image_repo_url>`: the URL of the image repository, for example, registry.cn-hangzhou.aliyuncs.com.

Create an application and configure Log Service

When you create an application in Container Service, you can configure Log Service to collect logs from containers. You can use only YAML templates to configure Log Service.

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click its name or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the cluster details page, choose **Workloads > Deployments**. On the Deployments page, click **Create from Template** in the upper-right corner.
5. Set **Sample Template** to **Custom** and configure the template.

   YAML templates follow the Kubernetes syntax. You can use environment variables to add **collection configurations** and **custom tags**. You must also configure volumeMounts and volumes. The following template is used to create a pod for collecting log data:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  labels:
    app: logtail-test
  name: logtail-test
spec:
  replicas: 1
  template:
    metadata:
      labels:
        app: logtail-test
      name: logtail-test
    spec:
      containers:
      - name: logtail
        image: registry.acs.env28.intranet.com/acs/busybox:latest
        args:
        - ping
        - 127.0.0.1
        env:
        - name: aliyun_logs_log-stdout
          value: stdout
        - name: aliyun_logs_log-varlog
          value: /log/*.log
        - name: aliyun_logs_log_tags
          value: tag1=v1
        volumeMounts:
        - name: volumn-sls
          mountPath: /log
      volumes:
      - name: volumn-sls
        emptyDir: {}
```

- Specify the following configurations in order based on your business requirements:
- Use environment variables to add **collection configurations** and **custom tags**. All environment variables for collection configurations must use the `aliyun_logs_` prefix.
- Add log collection configurations in the following format:

  ```
  - name: aliyun_logs_{Logstore name}
    value: {Log path}
  ```

  In the preceding YAML template, two environment variables are added to the log collection configuration. The environment variable `aliyun_logs_log-stdout` specifies that a Logstore named log-stdout is created to store the stdout collected from containers.

  > **Note** The name of the Logstore cannot contain underscores (_). You can use hyphens (-) instead.

- **Custom tags** must be specified in the following format:

  ```
  - name: aliyun_logs_{Tag name without underscores (_)}_tags
    value: {Tag name }={Tag value}
  ```

  After a tag is added, the tag is automatically appended to the log content that are collected from the container.
- If you specify a log path to collect log files that are not stdout, you must configure volumeMounts.

In the preceding YAML template, the mountPath field in volumeMounts is set to */var/log*. This allows logtail-ds to collect log data from the */var/log/*.log* file.

6. After you modify the YAML template, click **Create** to submit the configurations.

Environment variables

You can specify various environment variables to configure log collection. The following table describes the variables.

Variable	Description	Example	Remarks

aliyun_logs_{key}	• Required.{key} can contain only lowercase letters, digits, and hyphens (-), and cannot contain underscores (_).
	• If the specified aliyun_logs_{key}_logstore does not exist, a Logstore named {key} is created.
	• To collect the stdout of containers, set the value to stdout. You can also set the value to a path inside a container to collect the log files.
		`- name: aliyun_logs_catalina`
		`stdout`
		
		`- name: aliyun_logs_access-log`
		
		`/var/log/nginx/access.log`
			• By default, the Log Service component collects log files in simple mode. In this case, the collected log data not parsed. If you want to parse the log data, we recommend that you change the collection mode in the Log Service console.
			• The value of {key} must be unique in the cluster.

aliyun_logs_{key}_tags	Optional. This variable is used to add tags to log data. The value must be in the following format: {tag-key}={tag-value}.	`- name: aliyun_logs_catalina_tags` `app=catalina`	-
aliyun_logs_{key}_project	Optional. This variable specifies a project in Log Service. By default, the project that you specified when you created the cluster is used.	`- name: aliyun_logs_catalina_project` `my-k8s-project`	The region of the project must be the same as where logtail-ds is deployed.
aliyun_logs_{key}_logstore	Optional. This variable specifies a Logstore in Log Service. By default, the Logstore is named after {key}.	`- name: aliyun_logs_catalina_tags` `my-logstore`	-

Variable	Description	Example	Remarks
aliyun_logs_{key}_shard	Optional. This variable specifies the number of shards in the Logstore. Valid values: 1 to 10. Default value: 2.	- name: aliyun_logs_catalina_shard 4	-
aliyun_logs_{key}_ttl	Optional. This variable specifies the number of days for which log data is retained. Valid values: 1 to 3650. • To permanently retain log data, set the value to 3650. • Default value: 90.	- name: aliyun_logs_catalina_ttl 3650	-
aliyun_logs_{key}_machinegroup	Optional. This variable specifies the machine group of the application. By default, the machine group is the one where logtail-ds is deployed.	- name: aliyun_logs_catalina_machinegroup my-machine-group	-

- Scenario 1: Collect logs from multiple applications and store them in the same Logstore

  In this scenario, set the aliyun_logs_{key}_logstore variable. The following example shows how to collect stdout from two applications and store the outputs in stdout-logstore.

  Configure the following environment variables for Application 1:

```
######### Configure environment variables ############
    - name: aliyun_logs_app1-stdout
      value: stdout
    - name: aliyun_logs_app1-stdout_logstore
      value: stdout-logstore
```

  Configure the following environment variables for Application 2:

```
######### Configure environment variables ############
    - name: aliyun_logs_app2-stdout
      value: stdout
    - name: aliyun_logs_app2-stdout_logstore
      value: stdout-logstore
```

- Scenario 2: Collect logs from different applications and store them in different projects

  In this scenario, perform the following steps:

  i. Create a machine group in each project and set the machine group ID in the following format: k8s-group-{cluster-id}, where {cluster-id} is the ID of the cluster. You can customize the machine group name.

  ii. Specify the project, Logstore, and the created machine group in the environment variables for each application.

```
######### Configure environment variables ###########
    - name: aliyun_logs_app1-stdout
      value: stdout
    - name: aliyun_logs_app1-stdout_project
      value: app1-project
    - name: aliyun_logs_app1-stdout_logstore
      value: app1-logstore
    - name: aliyun_logs_app1-stdout_machinegroup
      value: app1-machine-group
```

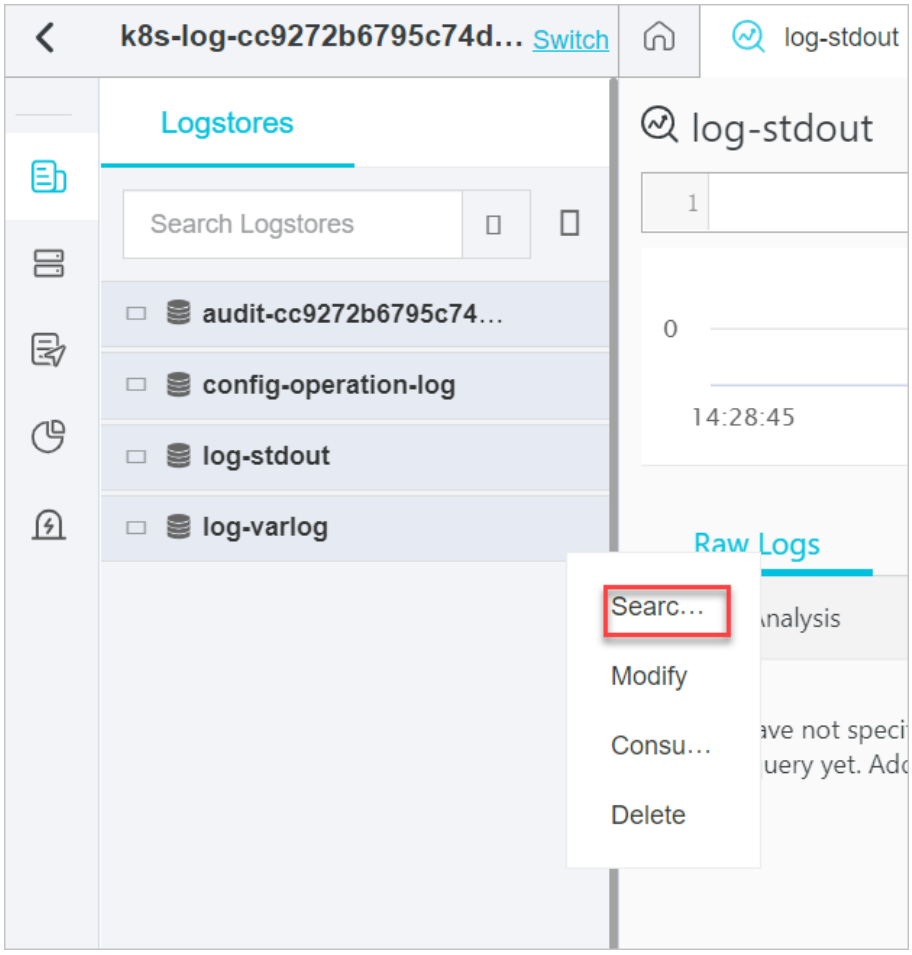
View logs

You can view the container log in the Log Service console.

1. Log on to the Log Service console. For more information, see Activate Log Service.
2. Click the project that is associated with the Kubernetes cluster. By default, the project name is in the format of k8s-log-{Kubernetes cluster ID}.
3. In the Logstore list, find the Logstore that is specified when you configure log collection. Move the pointer over the Logstore name and click the

   icon. Then, click **Search & Analysis**.
4. you can view the stdout and text log files of the container. You can also find that custom tags are appended to the collected log content.



After Log Service is enabled for the application, you can view the logs of containers in the Container Service console. Perform the following steps to view the logs of containers:

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click its name or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Operations > Log Center**.
5. On the **Log Center** page, click the **Application Logs** tab and specify the filter conditions. Then, click **Select Logstore** to view the logs of containers.

3.1.6.11.2. Configure Log4jAppender for Kubernetes and Log Service

This topic describes how to configure a YAML file to export the log of a Container Service cluster to Log Service, without the need to modify the application. An application that can be managed by calling API operations is deployed in the cluster to generate the log for testing purposes.

Prerequisites

- A Container Service cluster is created. For more information, see Create a Kubernetes cluster.
- An AccessKey pair is created or Resource Access Management (RAM) is activated. Make sure that the required permissions are granted. In this example, an AccessKey pair is created.

Context

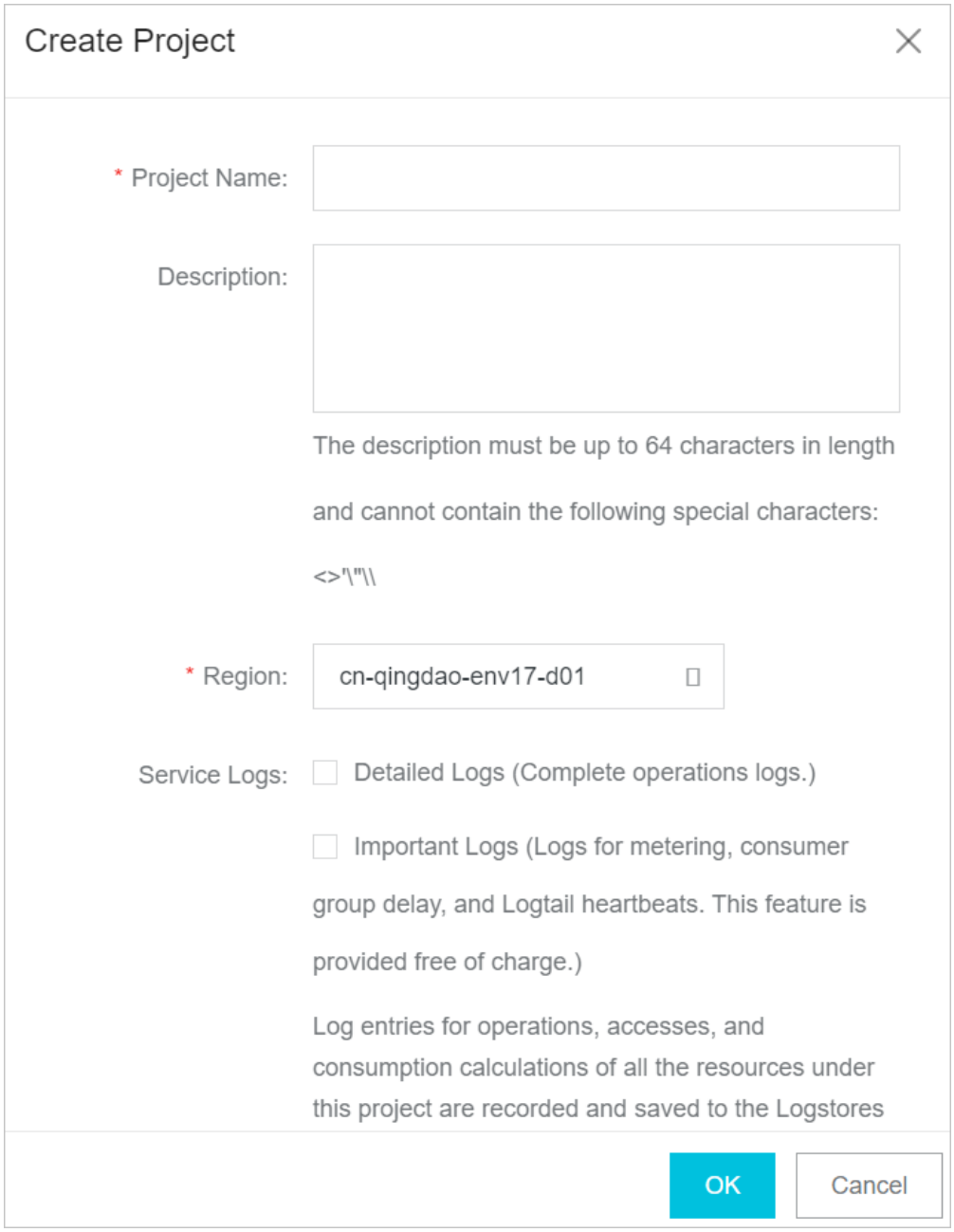
Log4j is an open source project of Apache. Log4j consists of three components: log level, log output destination, and log output format. You can configure Log4jAppender to export log data to the console, a log file, a GUI component, a socket server, an NT event viewer, or a UNIX syslog daemon.

Procedure

1. Configure Log4jAppender in Log Service.
   i. Create a project in Log Service.

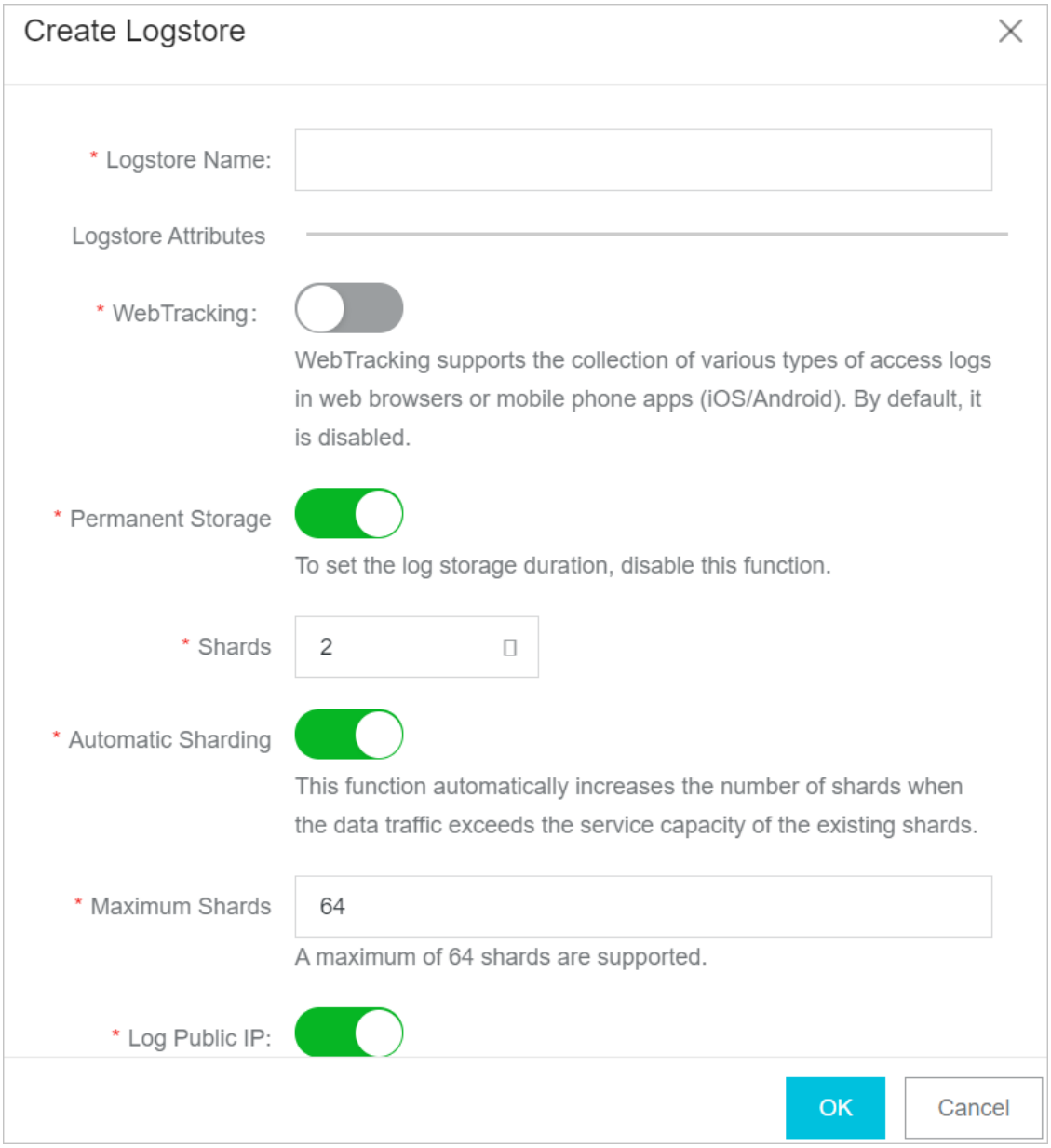
      In this example, a Log Service project named k8s-log4j is created in the same region as your cluster. For more information, see the *Manage projects* chapter of *Log Service User Guide*.

      > **Note** We recommend that you create a Log Service project in the same region as your cluster. When a Log Service project and a cluster are deployed in the same region, the log is transmitted over the internal network. This enables the real-time collection and quick retrieval of log data. This also avoids cross-region transmission, which requires additional bandwidth and time costs.

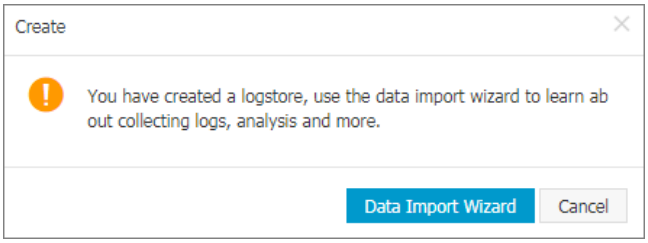

Create Project
* Project Name:
Description:
The description must be up to 64 characters in length and cannot contain the following special characters: <>'\"\\
* Region:
cn-qingdao-env17-d01
Service Logs:
Detailed Logs (Complete operations logs.)
Important Logs (Logs for metering, consumer group delay, and Logtail heartbeats. This feature is provided free of charge.)
Log entries for operations, accesses, and consumption calculations of all the resources under this project are recorded and saved to the Logstores
OK
Cancel

ii. Create a Logstore for the k8s-log4j project.

In this example, a Logstore named k8s-logstore is created. For more information, see the *Manage Logstores* chapter of *Log Service User Guide*.

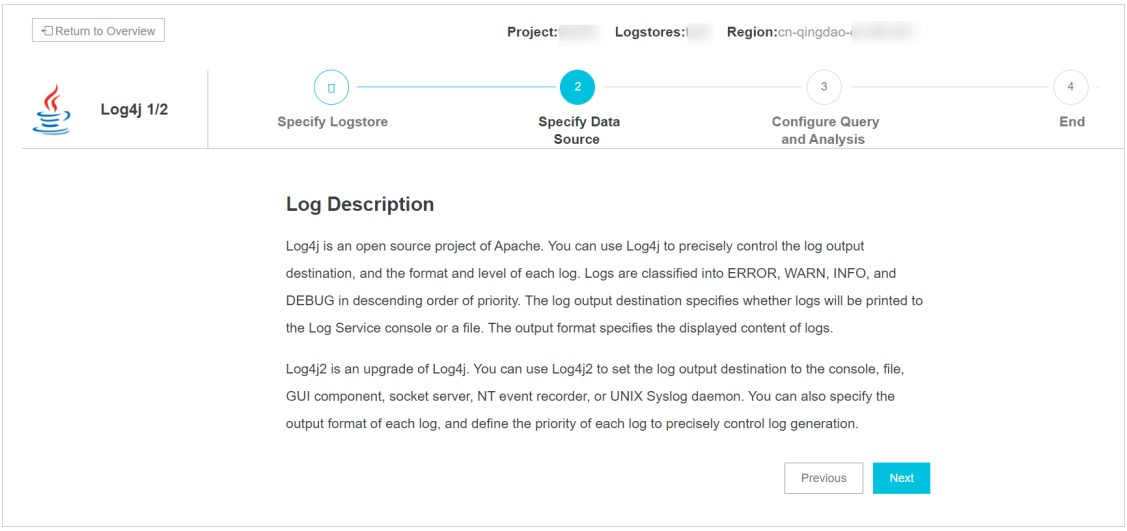


iii. After the k8s-logstore Logstore is created, a message appears, which prompts you to use the Data Import wizard. Click **Data Import Wizard**.



iv. At the top of the **Import Data** dialog box, click the **Custom Code** tab. Then, click Log4j and configure the data import parameters based on the instructions.

In this example, Log4jAppender is configured with the default settings. You can also customize the settings to meet your business requirements.



2. Configure Log4jAppender for the cluster.

In this example, the demo-deployment and demo-Service files are used.

i. Connect to your cluster. For more information, see Connect to a cluster through kubectl.

ii. Obtain the *demo-deployment.yaml* file and configure the `JAVA_OPTS` environment variable.

The following code is a sample template of the *demo-deployment.yaml* file:

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: log4j-appender-demo-spring-boot
  labels:
    app: log4j-appender
spec:
  replicas: 1
  selector:
    matchLabels:
      app: log4j-appender
  template:
    metadata:
      labels:
        app: log4j-appender
    spec:
      containers:
      - name: log4j-appender-demo-spring-boot
        image: registry.cn-hangzhou.aliyuncs.com/jaegertracing/log4j-appender-demo-spring-boot
:0.0.2
        env:
          - name: JAVA_OPTS                ## Take note of this environment variable.
            value: "-Dproject={your_project} -Dlogstore={your_logstore} -Dendpoint={your_endpo
int} -Daccess_key_id={your_access_key_id} -Daccess_key={your_access_key_secret}"
        ports:
        - containerPort: 8080
```

Note The following information is displayed:

- `-Dproject`: the name of your Log Service project. In this example, the name of the project is k8s-log4j.
- `-DlogStore`: the name of your Logstore. In this example, the name of the Logstore is k8s-logstore.
- `-Dendpoint`: the endpoint of Log Service. You must configure the endpoint based on the region of your Log Service project. In this example, the endpoint is cn-hangzhou.log.aliyuncs.com.
- `-Daccess_key_id`: your AccessKey ID.
- `-Daccess_key`: your AccessKey secret.

iii. Run the following command to create a Deployment:

```
kubectl create -f demo-deployment.yaml
```

iv. Obtain the *demo-Service.yaml* file and run the following command to create a Service:

You do not need to modify the configurations in the *demo-Service.yaml* file.

```
kubectl create -f demo-service.yaml
```

3. Generate the log of the Kubernetes cluster log.

You can run the `kubectl get` command to view the deployment status of the related resource objects. After the Deployment and Service are deployed, run the `kubectl get svc` command to check the external IP address of the Service, which is the value of EXTERNAL-IP.

Run the following command:

```
kubectl get svc
```

Output:

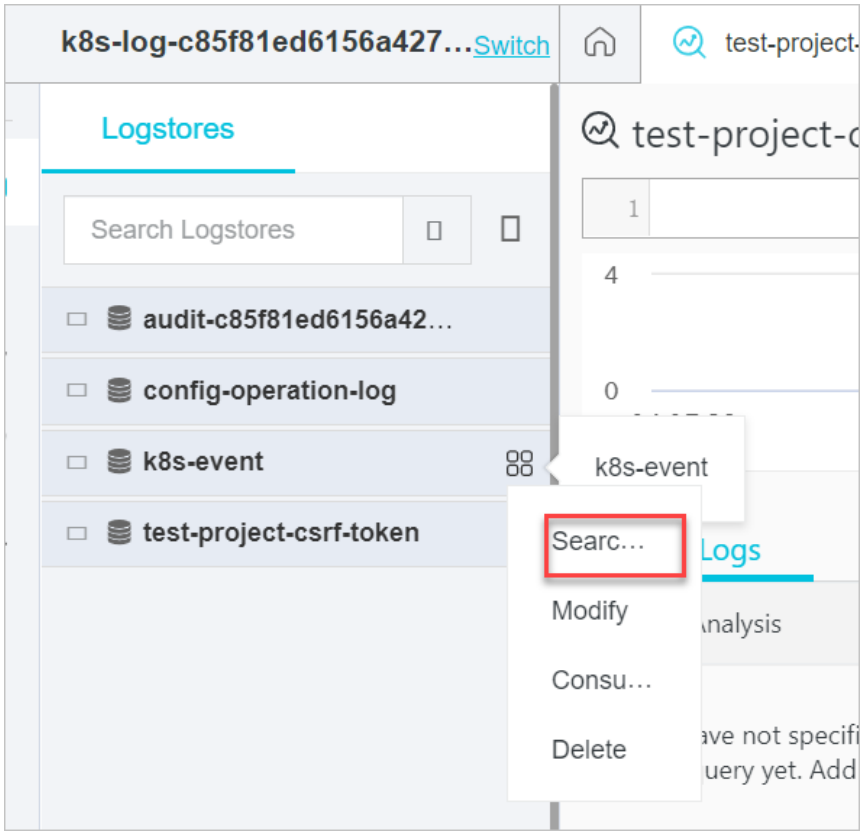
```
NAME                                TYPE           CLUSTER-IP     EXTERNAL-IP      PORT(S)
AGE
log4j-appender-demo-spring-boot-svc   LoadBalancer   172.21.XX.XX   120.55.XXX.XXX   8080:30398/TC
P   1h
```

In this example, you can run the `login` command to generate the log of the Kubernetes cluster log. Replace `K8S_SERVICE_IP` in the command with the value of `EXTERNAL-IP`.

> **Note** For a complete list of API operations, see GitHub log4j-appender-demo.

```
curl http://${K8S_SERVICE_IP}:8080/login?name=bruce
```

4. View the log in Log Service
   i. In the **Projects** list, click your Log Service project.
   ii. Click the icon to the right side of the k8s-logstore Logstore and select **Search & Analysis** to view the log of the Kubernetes cluster.



3.1.6.12. GPU

3.1.6.12.1. Create a dedicated Kubernetes cluster with GPU-accelerated nodes

This topic describes how to create a dedicated Kubernetes cluster with GPU-accelerated nodes.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**. On the Clusters page, click **Create Kubernetes Cluster** in the upper-right corner.
3. On the **Create Kubernetes Cluster** page, configure the basic settings.

Parameter	Description
Cluster Name	Enter a name for the cluster. The name must be 1 to 63 characters in length, and can contain digits, letters, and hyphens (-). **Note** The cluster name must be unique among clusters that belong to the same Alibaba Cloud account.
Region	Select the region where you want to deploy the cluster.
VPC	You can select a virtual private cloud (VPC) from the drop-down list. ◦ If the specified VPC is already associated with a NAT gateway, the cluster uses this NAT gateway. ◦ If the VPC does not have a NAT gateway, the system automatically creates one. If you do not want the system to create a NAT gateway, clear **Configure SNAT for VPC**. **Note** If you disallow the system to automatically create a NAT gateway and want the VPC to access the Internet, you must manually associate the VPC with a NAT gateway or create SNAT rules for the VPC.
VSwitch	Select one or more vSwitches for the cluster. You can select up to three vSwitches that are deployed in different **zones**.
Kubernetes Version	Select a Kubernetes version.
Container Runtime	You can select Docker or Sandboxed-Container.

Parameter	Description
Master Configurations	Set the Instance Type and System Disk parameters: ◦ Master Node Quantity: You can add up to three master nodes. ◦ Instance Type: You can select multiple instance types. For more information, see the *Instance families* chapter of *ECS User Guide*. ◦ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data.
Worker Instance	You can select **Create Instance** or **Add Existing Instance**.
Worker Configurations	If **Worker Instance** is set to **Create Instance**, set the following parameters: ◦ Instance Type: Select GPU-accelerated instance types. For more information, see the *Instance families* chapter of *ECS User Guide*. **Note** Select instance types from instance families whose names start with ecs.gn. ◦ Selected Types: The selected instance types are displayed. ◦ Quantity: Set the number of worker nodes. ◦ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data. ◦ Mount Data Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** ■ You can select **Encrypt Disk** to encrypt disks. ■ You can select **Enable Backup** to back up disk data.
Operating System	The CentOS and Alibaba Cloud Linux operating systems are supported.

Parameter	Description
Password	Set a password that is used to log on to the nodes. **Note** The password must be 8 to 30 characters in length, and must contain at least three of the following types of character: uppercase letters, lowercase letters, digits, and special characters.
Confirm Password	Enter the password again.
Network Plug-in	Flannel and Terway are supported. By default, Flannel is selected.
Pod CIDR Block and Service CIDR	These parameters are optional. For more information, see *Network planning* in *VPC User Guide*. **Note** These parameters are available only when you select an existing VPC.
Configure SNAT	This parameter is optional. If you clear Configure SNAT for VPC, you must create a NAT gateway or configure SNAT rules for the VPC.
Internet access	Specify whether to expose the API server with an elastic IP address (EIP). The Kubernetes API server provides multiple HTTP-based RESTful APIs that can be used to create, delete, modify, query, and watch resource objects such as pods and Services. ◦ If you select this check box, an EIP is created and attached to an internal-facing Server Load Balancer (SLB) instance. Port 6443 used by the API server is exposed on the master nodes. You can connect to and manage the cluster by using kubeconfig files over the Internet. ◦ If you clear this check box, no EIP is created. You can connect to and manage the cluster only by using kubeconfig files from within the VPC.
SSH Logon	To enable SSH logon, you must first select Expose API Server with EIP. ◦ If you select Use SSH to Access the Cluster from the Internet, you can access the cluster by using SSH. ◦ If you clear Use SSH to Access the Cluster from the Internet, you cannot access the cluster by using SSH or kubectl. If you want to access an Elastic Compute Service (ECS) instance in the cluster by using SSH, you must manually bind an EIP to the ECS instance and configure security group rules to open SSH port 22.
Security Group	You can select **Create Basic Security Group** or **Create Advanced Security Group**.

Parameter	Description
Ingress	Specify whether to **Install Ingress Controllers**. By default, **Install Ingress Controllers** is selected.
Log Service	If you enable Log Service, you can select an existing project or create a project. If you select **Enable Log Service**, the Log Service plug-in is automatically installed in the cluster. If you select **Create Ingress Dashboard**, Ingress access logs are collected and displayed on dashboards. By default, **Install node-problem-detector and Create Event Center** is selected.
Monitoring Agents	Select or clear **Enable Prometheus Monitoring**. Prometheus Monitoring provides the basic monitoring of the cluster.
Volume Plug-in	By default, **CSI** is selected.
Deletion Protection	If you select this check box, the cluster cannot be deleted in the console or by calling API operations.
Node Protection	This check box is selected by default to prevent nodes from being deleted in the console or by calling API operations.
Labels	Add labels to the cluster.

4. Configure the advanced settings.

Parameter	Description
IP Addresses per Node	The number of IP addresses that can be assigned to a node.
Custom Image	You can select a custom image. After you select a custom image, all nodes in the cluster are deployed by using this image.
Node Port Range	Specify the value of **Node Port Range**.
Taints	Add taints to all worker nodes in the cluster.
Cluster Domain	The default domain name of the cluster is cluster.local. You can specify a custom domain name.
Cluster CA	Specify whether to enable the cluster certification authority (CA) certificate.
User Data	Customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ○ Run scripts during instance startup. ○ Pass user data as common data into an ECS instance for future reference.

5. Click **Create Cluster** in the upper-right corner of the page.

6. On the **Confirm** page, after all check items are verified, select the terms of service and disclaimer and click **OK** to start the deployment.

3.1.6.12.2. Upgrade the NVIDIA driver on a GPU node

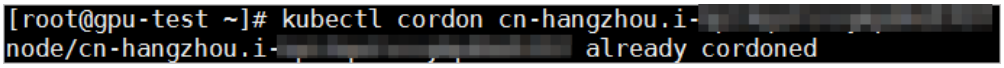
This topic describes how to upgrade the NVIDIA driver on a GPU node when workloads are deployed on the node and when no workload is deployed on the node.

Upgrade the NVIDIA driver on a GPU node where workloads are deployed

1. Connect to a Kubernetes cluster through kubectl.
2. Run the following command to set the target node to unschedulable.

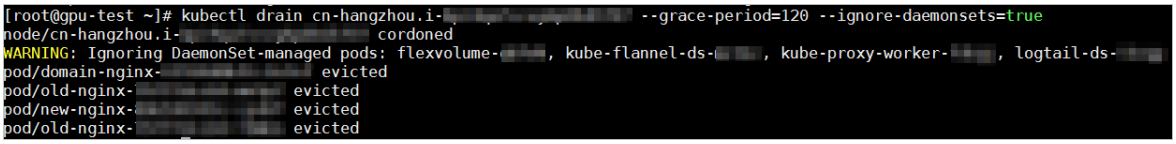
```
kubectl cordon node-name
```

> **Note**
>
> - Currently, you can only upgrade the NVIDIA driver on worker nodes.
> - *node-name* must be in the format of *your-region-name.node-id*.
>   - *your-region-name* represents the region where your cluster is deployed.
>   - *node-id* represents the ID of the ECS instance where the target node is deployed.
>
> You can run the following command to query *node-name*.
>
> ```
> kubectl get node
> ```



3. Run the following command to migrate pods from the target node to other nodes:

```
kubectl drain node-name --grace-period=120 --ignore-daemonsets=true
```



4. Run the following command to log on to the target node:

```
ssh root@xxx.xxx.x.xx
```

5. Run the following command to check the current NVIDIA driver version:

```
nvidia-smi
```

```
[root@ ~]# nvidia-smi
Fri Jan 18 16:44:52 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.111                Driver Version: 384.111                   |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P4            On   | 00000000:00:08.0 Off |                    0 |
| N/A   24C    P8     6W /  75W |      0MiB /  7606MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

6. Run the following commands to uninstall the existing driver:

   > **Note**
   > - If your driver version is *384.111*, perform the following steps.
   > - If your driver version is not *384.111*, download the corresponding driver from the official NVIDIA website first.

   ```
   cd /tmp
   ```

   ```
   curl -O https://cn.download.nvidia.cn/tesla/384.111/NVIDIA-Linux-x86_64-384.111.run
   ```

   ```
   chmod u+x NVIDIA-Linux-x86_64-384.111.run
   ```

   ```
   . /NVIDIA-Linux-x86_64-384.111.run --uninstall -a -s -q
   ```

7. Run the following command to restart the target node:

   ```
   reboot
   ```

8. Download the driver that you want to use from the official NVIDIA website. In this example, version *410.79* is used.
9. Run the following command to install the downloaded driver under the directory where it was saved:

   ```
   sh . /NVIDIA-Linux-x86_64-410.79.run -a -s -q
   ```

10. Run the following commands to configure the driver:

    ```
    nvidia-smi -pm 1 || true
    ```

    ```
    nvidia-smi -acp 0 || true
    ```

11. Run the following commands to update device-plugin:

    ```
    mv /etc/kubernetes/manifests/nvidia-device-plugin.yml /
    ```

    ```
    mv /nvidia-device-plugin.yml /etc/kubernetes/manifests/
    ```

12. Log on to a master node and run the following command to set the target node to schedulable:

    ```
    kubectl uncordon node-name
    ```

Result

Run the following command on a master node to check the NVIDIA driver version on the target node. The driver version is now *410.79*.

> **Note** Replace *node-name* with the target node name.

```
kubectl exec -n kube-system -t nvidia-device-plugin-node-name nvidia-smi
```

```
[root@gpu-test ~]#  kubectl exec -n kube-system -t nvidia-device-plugin-cn-[redacted] nvidia-smi
Mon Jan 21 03:14:48 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.79       Driver Version: 410.79       CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P4            On   | 00000000:00:08.0 Off |                    0 |
| N/A   21C    P8     6W /  75W |      0MiB /  7611MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Upgrade the NVIDIA driver on a GPU node where no workload is deployed

1. Run the following command to log on to the target node:

   ```
   ssh root@xxx.xxx.x.xx
   ```

2. Run the following command to check the current NVIDIA driver version:

   ```
   nvidia-smi
   ```

   ```
   [root@[redacted] ~]# nvidia-smi
   Fri Jan 18 16:44:52 2019
   +-----------------------------------------------------------------------------+
   | NVIDIA-SMI 384.111                Driver Version: 384.111                   |
   |-------------------------------+----------------------+----------------------+
   | GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
   | Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
   |===============================+======================+======================|
   |   0  Tesla P4            On   | 00000000:00:08.0 Off |                    0 |
   | N/A   24C    P8     6W /  75W |      0MiB /  7606MiB |      0%      Default |
   +-------------------------------+----------------------+----------------------+

   +-----------------------------------------------------------------------------+
   | Processes:                                                       GPU Memory |
   |  GPU       PID   Type   Process name                             Usage      |
   |=============================================================================|
   |  No running processes found                                                 |
   +-----------------------------------------------------------------------------+
   ```

3. Run the following commands to uninstall the existing driver:

   > **Note**
   > - If your driver version is *384.111*, perform the following steps.
   > - If your driver version is not *384.111*, download the corresponding driver from the official NVIDIA website first.

```
cd /tmp
```

```
curl -O https://cn.download.nvidia.cn/tesla/384.111/NVIDIA-Linux-x86_64-384.111.run
```

```
chmod u+x NVIDIA-Linux-x86_64-384.111.run
```

```
. /NVIDIA-Linux-x86_64-384.111.run --uninstall -a -s -q
```

4. Run the following command to restart the target node.

```
reboot
```

5. Download the driver that you want to use from the official NVIDIA website. In this example, version *410.79* is used.
6. Run the following command to install the downloaded driver under the directory where it was saved:

```
sh . /NVIDIA-Linux-x86_64-410.79.run -a -s -q
```

7. Run the following commands to configure the driver:

```
nvidia-smi -pm 1 || true
```

```
nvidia-smi -acp 0 || true
```

Result

Run the following command on a master node to check the NVIDIA driver version on the target node. The driver version is now *410.79*.

> **Note** Replace *node-name* with the target node name.

```
kubectl exec -n kube-system -t nvidia-device-plugin-node-name nvidia-smi
```

```
[root@gpu-test ~]#  kubectl exec -n kube-system -t nvidia-device-plugin-cn-[redacted] nvidia-smi
Mon Jan 21 03:14:48 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.79       Driver Version: 410.79       CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P4            On   | 00000000:00:08.0 Off |                    0 |
| N/A   21C    P8     6W /  75W |      0MiB /  7611MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

3.1.6.12.3. Use cGPU to enable GPU sharing and scheduling

You can use the cGPU solution to schedule multiple applications to one GPU and isolate the GPU memory and computing power that are allocated to each application. This topic describes how to use cGPU to enable GPU sharing and scheduling.

Prerequisites

- A dedicated Kubernetes cluster with GPU-accelerated nodes is created. For more information, see Create a dedicated Kubernetes cluster with GPU-accelerated nodes.

- The Docker version used by the nodes is 19.03.5 or later. Docker versions earlier than 19.03.5 support GPU sharing but do not support GPU isolation.

Context

A key requirement of GPU sharing among multiple pods is to isolate the GPU memory and computing power that are allocated to each pod. When you run multiple containers on one GPU, the GPU resources are allocated to each container as requested. However, if one container occupies excessive GPU resources, the performance of the other containers may be affected. To address this issue, many solutions have been developed in the computing industry. Technologies, such as NVIDIA virtual GPU (vGPU), NVIDIA Multi-Process Service (MPS), rCUDA, and vCUDA, all contribute to fine-grained GPU resource allocation.

The cGPU solution uses the server kernel driver that is developed by Alibaba Cloud to provide more efficient use of the underlying drivers of NVIDIA GPUs. cGPU provides the following features:

- High compatibility: cGPU is compatible with standard open source solutions, such as Kubernetes and NVIDIA Docker.
- Ease of use: cGPU provides excellent user experience. To replace a Compute Unified Device Architecture (CUDA) library of an AI application, you do not need to recompile the application or create a new container image.
- Stability: cGPU provides stable underlying operations on NVIDIA GPUs. API operations on CUDA libraries and some private API operations on CUDA Deep Neural Network (cuDNN) are difficult to call.
- Resource isolation: cGPU ensures that the allocated GPU memory and computing power do not affect each other.

cGPU provides a cost-effective, reliable, and user-friendly solution that allows you to enable GPU scheduling and memory isolation.

Step 1: Create a node pool

Create a node pool and add the `cgpu=true` label to the node pool.

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the cluster details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, click **Create Node Pool**.
6. In the **Create Node Pool** dialog box, configure the node pool and click **Confirm Order**.

   For more information, see Create a Kubernetes cluster. The following list describes some of the parameters:

   - Name: the name of the node pool.
   - Quantity: the initial number of nodes in the node pool. If you do not want to add nodes to the node pool, set this parameter to 0.
   - ECS Label: Add labels to the Elastic Compute Service (ECS) instances.
   - Node Label: Click the  icon to add a label. Set the key to cgpu and the value to true.

7. Click **Submit**.

Step 2: Install ack-cgpu

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > App Catalog**.
3. On the **App Catalog** page, search for ack-cgpu and click ack-cgpu after it appears.
4. On the **App Catalog - ack-cgpu** page, select the cluster where you want to install ack-cgpu in the **Deploy** section and click **Create**.

> **Note** The default number of master nodes in a dedicated Kubernetes cluster is three. If the cluster has five master nodes, set the value of `masterCount` to `5`.

5. Verify that ack-cgpu is installed.
   i. In the left-side navigation pane, click **Clusters**.
   ii. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
   iii. In the left-side navigation pane of the details page, choose **Workloads > Pods**.
   iv. At the top of the **Pods** page, set **Namespace** to kube-system. Enter gpushare in the search box and click the search icon. Check the pod status after it appears. If the **state** of the pod is Running or Completed, it indicates that ack-cgpu is installed.

Step 3: Verify GPU sharing and scheduling

1. Create a StatefulSet.
   i. Log on to the Container Service console.
   ii. In the left-side navigation pane, click **Clusters**.
   iii. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
   iv. On the cluster details page, choose **Workloads > StatefulSets**. On the right side of the StatefulSets page, click **Create from Template**.

v. Set **Sample Template** to **Custom**, copy the following code to the template, and then click **Create**.

```
apiVersion: apps/v1
kind: StatefulSet
metadata:
  name: cgpu-test
  labels:
    app: cgpu-test
spec:
  replicas: 3
  serviceName: "cgpu-test"
  podManagementPolicy: "Parallel"
  selector:
    matchLabels:
      app: cgpu-test
  template:
    metadata:
      labels:
        app: cgpu-test
    spec:
      containers:
      - name: cgpu-test
        image: registry.acs.intra.env115.shuguang.com/acs/gpushare-sample:tensorflow-1.5
        command:
          - python
          - cgpu/main.py
        resources:
          limits:
            aliyun.com/gpu-mem: 2 # Apply for 2 GiB of GPU memory.
```

Replace `registry.acs.intra.env115.shuguang.com` in the image address based on your business requirements. You can replace the image repository with that of the image used in Step .

If you want to use ack-cgpu to allocate GPU memory to a pod, you must specify the amount of GPU memory in the `resources` section. In this example, the amount is 2 GiB.

```
// Other settings are omitted.
......
        resources:
          limits:
            aliyun.com/gpu-mem: 2 # Apply for 2 GiB of GPU memory.
```

2. Check the amount of GPU memory allocated to the container.
   i. Log on to the Container Service console.
   ii. In the left-side navigation pane, click **Clusters**.
   iii. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
   iv. In the left-side navigation pane of the details page, choose **Workloads > StatefulSets**.
   v. On the **StatefulSets** page, click the newly created StatefulSet.
   vi. On the **Pods** tab, find the first pod and choose **Terminal > Container: cgpu-test** in the **Actions** column.
   vii. Run the `nvidia-smi -L` command to check the ID of the GPU used by the container.

viii. Run the `nvidia-smi` command to check the amount of GPU memory allocated to the container.

```
[redacted] nvidia-smi
Tue Jun  8 12:21:53 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 418.181.07   Driver Version: 418.181.07   CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P4            On   | 00000000:00:07.0 Off |                    0 |
| N/A   46C    P0    23W /  75W |   1949MiB /  2174MiB |      2%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
```

ix. Perform the preceding steps to check the GPU ID and amount of GPU memory of the second pod.

The results show that both pods have the same GPU ID. Each pod is allocated with 2,174 MiB of GPU memory.

> **Note** Pods have the same GPU ID only when they run on the same node. In this example, the cluster has only one GPU-accelerated node. Therefore, the two pods have to run on the same node.

x. Log on to a master node of the cluster. Run the following command to check the allocation details of the GPU resources of the node:

```
kubectl inspect cgpu
```

Expected output:

```
NAME      IPADDRESS       GPU0(Allocated/Total)  GPU Memory(GiB)
a52120c  192.168.52.120  6/7                    6/7
--------------------------------------------------
Allocated/Total GPU Memory In Cluster:
6/7 (85%)
```

The output shows that the cluster has 7 GiB of GPU memory in total and 6 GiB has been used.

The preceding results indicate that the pods use the same GPU and each pod is allocated with 2,174 MiB of GPU memory. This means that the allocated GPU resources are isolated among pods. The total GPU memory of one GPU is 7,611 MiB. You can run the `nvidia-smi` command on a node to obtain the total GPU memory of a GPU. If the allocated GPU resources are not isolated among pods, the amount of GPU memory allocated to a pod is 7,611 MiB. In this example, the actual amount of GPU memory allocated to a pod is 2,174 MiB, which is equal to 2 GiB as requested by the pod. This means the isolation takes effect.

3.1.6.12.4. GPU scheduling for Kubernetes clusters with GPU-accelerated nodes

This topic describes GPU scheduling for Kubernetes clusters with GPU-accelerated nodes.

Prerequisites

Container Service, Resource Orchestration Service (ROS), and Resource Access Management (RAM) are activated.

> **Note** Container Service uses ROS to deploy applications in Kubernetes clusters. To create a Kubernetes cluster, you must first activate ROS.

Context

Starting from version 1.8, Kubernetes adds support for the following hardware acceleration devices by using device plug-ins: NVIDIA GPUs, InfiniBand devices, and field-programmable gate arrays (FPGAs). GPU solutions developed by the community will be phased out in version 1.10, and removed from the master code in version 1.11. Container Service enables you to use a Kubernetes cluster with GPU-accelerated nodes to run compute-intensive tasks such as machine learning and image processing. You can deploy applications and achieve auto scaling without the need to install NVIDIA drivers or Compute Unified Device Architecture (CUDA) in advance.

The system performs the following operations when a Kubernetes cluster is created:

- Creates Elastic Compute Service (ECS) instances, configures a public key to enable Secure Shell (SSH) logon from master nodes to other nodes, and then configures the Kubernetes cluster through CloudInit.
- Creates a security group that allows access to the virtual private cloud (VPC) over Internet Control Message Protocol (ICMP).
- If you do not specify an existing VPC, the system creates a VPC and a vSwitch and creates SNAT entries for the vSwitch.
- Adds route entries to the VPC.
- Creates a NAT gateway and an elastic IP address (EIP).
- Creates a RAM user and grants it permissions to query, create, and delete ECS instances, and permissions to add and delete disks. The RAM user is also granted full permissions on Server Load Balancer (SLB) instances, CloudMonitor, VPC, Log Service, and Apsara File Storage NAS (NAS). The system also creates an AccessKey pair for the RAM user. The system automatically creates SLB instances, disks, and VPC route entries based on your configuration.
- Creates an internal-facing SLB instance and opens port 6443.
- Creates an Internet-facing SLB instance and open ports 6443, 8443, and 22. If you enable SSH logon when you create the cluster, port 22 is open. Otherwise, port 22 is not open.

Limits

- Kubernetes clusters support only VPCs.
- By default, you can create only a limited amount of cloud resources with each account. You cannot create clusters if the quota on clusters is reached. Make sure that you have sufficient quota before you create a cluster. To request a quota increase, submit a ticket.
  - By default, you can create at most five clusters across all regions with each account. Each cluster can contain at most 40 nodes. To increase the quota of clusters or nodes, submit a ticket.

    > **Note** In a Kubernetes cluster, you can create at most 48 route entries for each VPC. This means that a cluster can contain at most 48 nodes. To increase the quota of nodes, submit a ticket to increase the quota of route entries first.

  - By default, you can create at most 100 security groups with each account.
  - By default, you can create at most 60 pay-as-you-go SLB instances with each account.
  - By default, you can create at most 20 EIPs with each account.
- Limits on ECS instances:

  Only CentOS is supported.

Create a GN5 Kubernetes cluster

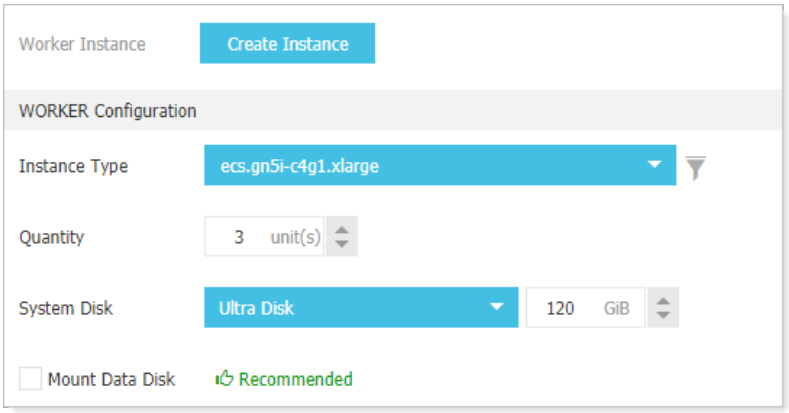
GN5 Kubernetes clusters support only Kubernetes 1.12.6-aliyun.1. Kubernetes 1.11.5 is not supported.

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. In the upper-right corner of the **Clusters** page, click **Create Kubernetes Cluster**.

On the cluster configuration page, set the parameters.

> **Note** To create a cluster with GPU-accelerated nodes, select GPU-accelerated ECS instance types as worker nodes. For more information about other parameters, see Cluster parameters.

4. Configure worker nodes. In this example, worker nodes are used to run GPU computing tasks and the gn5i-c4gi1 instance type is selected.
   i. If you choose to create worker instances, you must set Instance Type and Quantity. In this example, three GPU-accelerated worker nodes are created.



> **Note** We recommend that you use standard SSDs.

   ii. If you choose to add existing instances, you must create GPU-accelerated ECS instances in the region where you want to create the cluster in advance.

5. Specify the other parameters and click **Create Cluster** to start the deployment.

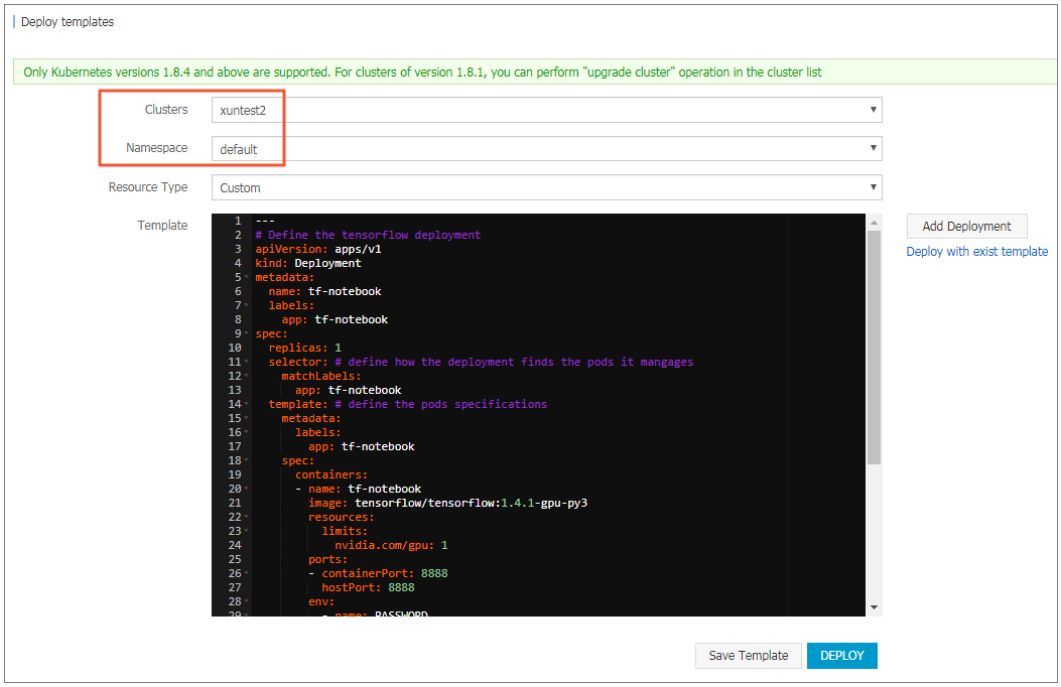
   After the cluster is created, choose **Clusters > Nodes** to go to the Nodes page.

   Select a worker nodes that was configured for the cluster and choose **More > Details** in the Actions column to view the GPU devices that are attached to the node.

Create a GPU experimental environment to run TensorFlow

Jupyter is a standard tool that is used by data scientists to create an experimental environment to run TensorFlow. The following example shows how to deploy a Jupyter application.

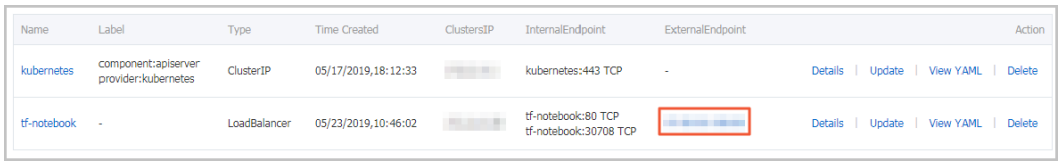
1. Log on to the Container Service console. In the left-side navigation pane, choose **Applications > Deployments** to go to the **Deployments** page.
2. In the upper-right corner of the page, click **Create from Template**.
3. Select the required cluster and namespace. Select a sample template, or set Sample Template to Custom and customize the template in the Template field. Then, click **Create** to create the application.



In this example, the template uses a Deployment and a Service to create a Jupyter application.

```
---
# Define the tensorflow deployment
apiVersion: apps/v1
kind: Deployment
metadata:
  name: tf-notebook
  labels:
    app: tf-notebook
spec:
  replicas: 1
  selector: # define how the deployment finds the pods it manages
    matchLabels:
      app: tf-notebook
  template: # define the pods specifications
    metadata:
      labels:
        app: tf-notebook
    spec:
      containers:
      - name: tf-notebook
        image: tensorflow/tensorflow:1.4.1-gpu-py3
        resources:
          limits:
            nvidia.com/gpu: 1                    #The number of NVIDIA GPUs that are requested b
y the application.
        ports:
        - containerPort: 8888
          hostPort: 8888
        env:
          - name: PASSWORD                       #The password that is used to access the Jupyte
r application. You can modify the password as required.
            value: mypassw0rd
# Define the tensorflow service
---
apiVersion: v1
kind: Service
metadata:
  name: tf-notebook
spec:
  ports:
  - port: 80
    targetPort: 8888
    name: jupyter
  selector:
    app: tf-notebook
  type: LoadBalancer                           #An SLB instance is used to route internal traffic
and perform load balancing.
```

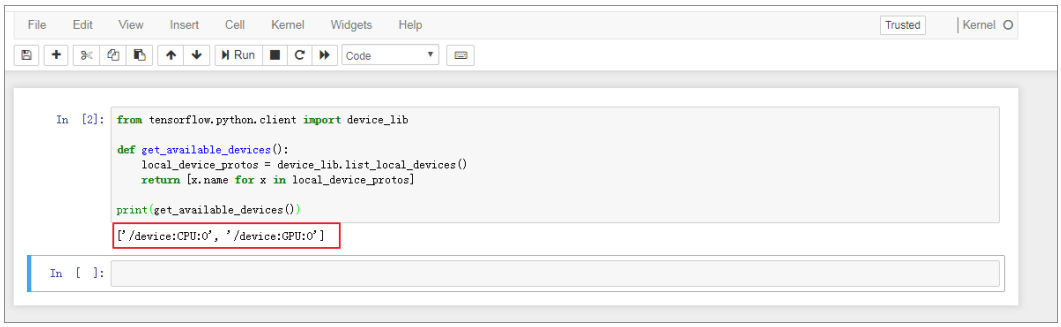
4. In the left-side navigation pane, choose **Ingresses and Load Balancing > Services**. Select the required cluster and namespace, find the tf-notebook Service, and then check its external endpoint.



5. To connect to the Jupyter application, enter `http://EXTERNAL-IP` into the address bar of your browser and enter the password specified in the template.

6. You can run the following program to verify that the Jupyter application is allowed to use GPU resources. The program lists all devices that can be used by TensorFlow:

```
from tensorflow.python.client import device_lib
def get_available_devices():
    local_device_protos = device_lib.list_local_devices()
    return [x.name for x in local_device_protos]
print(get_available_devices())
```



3.1.6.12.5. Use labels to schedule pods to GPU-accelerated nodes

To use Kubernetes clusters for GPU computing, you must schedule pods to GPU-accelerated nodes. Container Service allows you to schedule pods to specific GPU-accelerated nodes by adding labels to the GPU-accelerated nodes.

Context

When Kubernetes deploys nodes with NVIDIA GPUs, the attributes of these GPUs are discovered and exposed as node labels. These labels have the following benefits:

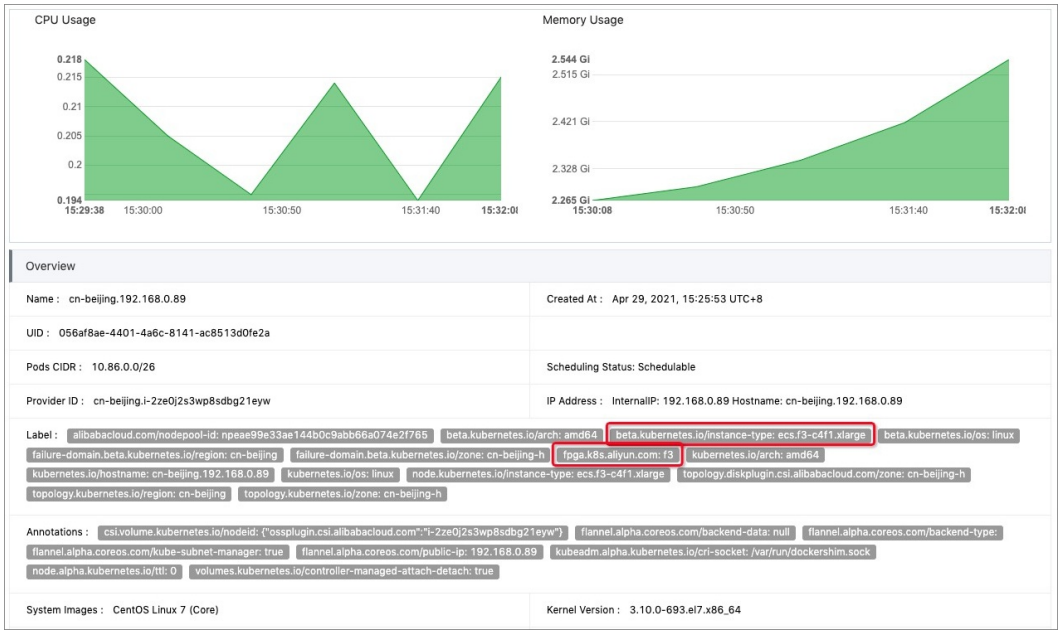
- You can use the labels to filter GPU-accelerated nodes.
- The labels can be used as conditions to schedule pods.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.

   > **Note** In this example, the Kubernetes cluster contains three master nodes, among which two are equipped with GPUs. Record the node IP addresses.

5. On the **Nodes** page, select the node that is equipped with GPUs and choose **More > Details** in the Actions column to view the labels that are added to the node.



You can also log on to a master node and run the following command to view the labels of GPU-accelerated nodes:

```
# kubectl get nodes
NAME                                 STATUS    ROLES     AGE       VERSION
cn-beijing.i-2ze2dy2h9w97v65u****    Ready     master    2d        v1.12.6-aliyun.1
cn-beijing.i-2ze8o1a45qdv5q8a****    Ready     <none>    2d        v1.12.6-aliyun.1            # C
ompare these nodes with the nodes displayed in the console to identify GPU-accelerated nodes.
cn-beijing.i-2ze8o1a45qdv5q8a****    Ready     <none>    2d        v1.12.6-aliyun.1
cn-beijing.i-2ze9xylyn11vop7g****    Ready     master    2d        v1.12.6-aliyun.1
cn-beijing.i-2zed5sw8snjniq6m****    Ready     master    2d        v1.12.6-aliyun.1
cn-beijing.i-2zej9s0zijykp9pw****    Ready     <none>    2d        v1.12.6-aliyun.1
```

Select a GPU-accelerated node and run the following command to query the labels of the node:

```
# kubectl describe node cn-beijing.i-2ze8o1a45qdv5q8a****
Name:               cn-beijing.i-2ze8o1a45qdv5q8a7luz
Roles:              <none>
Labels:             aliyun.accelerator/nvidia_count=1                          #Note
                    aliyun.accelerator/nvidia_mem=12209MiB
                    aliyun.accelerator/nvidia_name=Tesla-M40
                    beta.kubernetes.io/arch=amd64
                    beta.kubernetes.io/instance-type=ecs.gn4-c4g1.xlarge
                    beta.kubernetes.io/os=linux
                    failure-domain.beta.kubernetes.io/region=cn-beijing
                    failure-domain.beta.kubernetes.io/zone=cn-beijing-a
                    kubernetes.io/hostname=cn-beijing.i-2ze8o1a45qdv5q8a****
 ......
```

In this example, the following labels are added to the GPU-accelerated node.

key	value
`aliyun.accelerator/nvidia_count`	The number of GPU cores.
`aliyun.accelerator/nvidia_mem`	The size of the GPU memory. Unit: MiB.

key	value
`aliyun.accelerator/nvidia_name`	The name of the NVIDIA GPU.

GPU-accelerated nodes of the same type have the same GPU name. You can use this label to locate GPU-accelerated nodes.

```
# kubectl get no -l aliyun.accelerator/nvidia_name=Tesla-M40
NAME                                STATUS    ROLES     AGE       VERSION
cn-beijing.i-2ze8o1a45qdv5q8a****   Ready     <none>    2d        v1.12.6-aliyun.1
cn-beijing.i-2ze8o1a45qdv5q8a****   Ready     <none>    2d        v1.12.6-aliyun.1
```

6. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
7. On the **Deployments** page, select the namespace and click **Create from Template** in the upper-right corner.
8. Create a Deployment for a TensorFlow job. The Deployment is used to schedule pods to a GPU-accelerated node.
9. You can also exclude an application from GPU-accelerated nodes. The following example shows how to schedule a pod based on node affinity for an NGINX application. For more information, see the section that describes node affinity in Create an application from an image.

   The following YAML template is used as an example:

```
apiVersion: v1
kind: Pod
metadata:
  name: not-in-gpu-node
spec:
  affinity:
    nodeAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        nodeSelectorTerms:
        - matchExpressions:
          - key: aliyun.accelerator/nvidia_name
            operator: DoesNotExist
  containers:
  - name: not-in-gpu-node
    image: nginx
```

Result

In the left-side navigation pane of the details page, choose **Workloads > Pods**. On the Pods page, select the namespace to view pods created in the namespace. On the Pods page, you can find that the pods in the preceding examples are scheduled to the desired nodes. This means that labels can be used to schedule pods to GPU-accelerated nodes.

3.1.6.12.6. Manually upgrade the kernel of a GPU node in a cluster

This topic describes how to manually upgrade the kernel of a GPU node in a cluster.

Context

The current kernel version is earlier than `3.10.0-957.21.3`.

Procedure

1. Connect to a Kubernetes cluster through kubectl.
2. Run the following command to set the target GPU node to unschedulable. This example uses node cn-beijing.i-2ze19qyi8votgjz12345 as the target node.

```
kubectl cordon cn-beijing.i-2ze19qyi8votgjz12345
node/cn-beijing.i-2ze19qyi8votgjz12345 already cordoned
```

3. Run the following command to drain the target GPU node:

```
# kubectl drain cn-beijing.i-2ze19qyi8votgjz12345 --grace-period=120 --ignore-daemonsets=true
node/cn-beijing.i-2ze19qyi8votgjz12345 cordoned
WARNING: Ignoring DaemonSet-managed pods: flexvolume-9scb4, kube-flannel-ds-r2qmh, kube-proxy-worker-l62sf, logtail-ds-f9vbg
pod/nginx-ingress-controller-78d847fb96-5fkkw evicted
```

4. Uninstall the existing nvidia-driver.

> **Note** This step uninstalls the version 384.111 driver. If your driver version is not 384.111, you need to download a driver from the official NVIDIA website and replace `384.111` with your actual version number.

   i. Log on to the target GPU node and run the `nvidia-smi` command to query the driver version.

```
# nvidia-smi -a | grep 'Driver Version'
Driver Version                      : 384.111
```

   ii. Run the following commands to download the driver installation package:

```
cd /tmp/
curl -O https://cn.download.nvidia.cn/tesla/384.111/NVIDIA-Linux-x86_64-384.111.run
```

> **Note** The installation package is required to uninstall the driver.

   iii. Run the following commands to uninstall the existing nvidia-driver:

```
chmod u+x NVIDIA-Linux-x86_64-384.111.run
. /NVIDIA-Linux-x86_64-384.111.run --uninstall -a -s -q
```

5. Run the following commands to upgrade kernel:

```
yum clean all && yum makecache
yum update kernel  -y
```

6. Run the following command to restart the GPU node:

```
reboot
```

7. Log on to the GPU node and run the following command to install the kernel-devel package.

```
yum install -y kernel-devel-$(uname -r)
```

8. Run the following commands to download the required driver and install it on the target node. In this example, version 410.79 is used.

```
cd /tmp/
curl -O https://cn.download.nvidia.cn/tesla/410.79/NVIDIA-Linux-x86_64-410.79.run
chmod u+x NVIDIA-Linux-x86_64-410.79.run
sh . /NVIDIA-Linux-x86_64-410.79.run -a -s -q
# warm up GPU
nvidia-smi -pm 1 || true
nvidia-smi -acp 0 || true
nvidia-smi --auto-boost-default=0 || true
nvidia-smi --auto-boost-permission=0 || true
nvidia-modprobe -u -c=0 -m || true
```

9. Check the */etc/rc.d/rc.local* file and check whether the following configurations are included. If not, add the following content.

```
nvidia-smi -pm 1 || true
nvidia-smi -acp 0 || true
nvidia-smi --auto-boost-default=0 || true
nvidia-smi --auto-boost-permission=0 || true
nvidia-modprobe -u -c=0 -m || true
```

10. Run the following commands to restart kubelet and Docker.

```
service kubelet stop
service docker restart
service kubelet start
```

11. Run the following command to set the GPU node to schedulable:

```
# kubectl uncordon cn-beijing.i-2ze19qyi8votgjz12345
node/cn-beijing.i-2ze19qyi8votgjz12345 already uncordoned
```

12. Run the following command on the nvidia-device-plugin container to check the driver version:

```
kubectl exec -n kube-system -t nvidia-device-plugin-cn-beijing.i-2ze19qyi8votgjz12345 nvidia-smi
Thu Jan 17 00:33:27 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.79       Driver Version: 410.79       CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  On   | 00000000:00:09.0 Off |                    0 |
| N/A   27C    P0    28W / 250W |      0MiB / 16280MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

3.1.6.13. Auto scaling

3.1.6.13.1. Auto scaling of nodes

Container Service provides the auto scaling component to automatically scale the number of nodes. Regular instances and GPU-accelerated instances can be automatically added to or removed from a Container Service cluster to meet your business requirements. This component supports multiple scaling modes, various instance types, and instances that are deployed across zones. This component is applicable to diverse scenarios.

How it works

The auto scaling model of Kubernetes is different from the traditional scaling model that is based on the resource usage threshold. Developers must understand the differences between the two scaling models before they migrate workloads from traditional data centers or other orchestration systems such as Swarm to Kubernetes.

The traditional scaling model is based on resource usage. For example, if a cluster contains three nodes and the CPU utilization or memory usage of the nodes exceeds the scaling threshold, new nodes are added to the cluster. However, you must consider the following issues when you use the traditional scaling model:

- How do you set a proper scaling threshold and how does the system check whether the threshold is exceeded?

  In a Kubernetes cluster, the resource usage of hot nodes is higher than that of other nodes. If you specify the average resource usage as the scaling threshold, scaling activities may not be promptly triggered. If you specify the lowest node resource usage as the scaling threshold, the newly added nodes may not be used. This causes a waste of resources.

- How are the loads balanced after new nodes are added?

  In Kubernetes, pods are the smallest deployable units for applications. Pods are deployed on different nodes in a Kubernetes cluster. When auto scaling is triggered for a cluster or a node in the cluster, pods with high resource usage are not replicated and the resource limits of these pods are not changed. As a result, the loads cannot be balanced to newly added nodes.

- How do you determine whether scaling activities must be triggered and how are scaling activities performed?

  If scale-in activities are triggered based on resource usage, pods that request large amounts of resources but have low resource usage may be evicted. If the number of these pods is large within a Kubernetes cluster, resources may be exhausted and some pods may fail to be scheduled.

How does the auto scaling model of Kubernetes fix these issues? Kubernetes provides a two-layer scaling model that decouples pod scheduling from resource scaling.

In simple terms, pods are scaled based on resource usage. When pods enter the Pending state due to insufficient resources, a scale-out activity is triggered. After new nodes are added to the cluster, the pending pods are automatically scheduled to the newly added nodes. This way, the loads of the application are balanced. The following section describes the auto scaling model of Kubernetes in detail:

- How is a scale-out activity triggered?

  The cluster-autoscaler component scans for pending pods and then triggers scaling activities. When pods enter the Pending state due to insufficient resources, cluster-autoscaler simulates pod scheduling to decide the scaling group that can provide new nodes to accept the pending pods. If a scaling group meets the requirement, nodes from this scaling group are added to the cluster. In simple terms, a scaling group is treated as a node during the simulation. The instance type of the scaling group specifies the CPU, memory, and GPU resources of the node. The labels and taints of the scaling group are also applied to the node. The node is used to simulate the scheduling of the pending pods. If the pending pods can be scheduled to the node, cluster-autoscaler calculates the number of nodes that need to be added from the scaling group.

- How is a scale-in activity triggered?

  Only nodes that are added by scaling activities can be removed by cluster-autoscaler. This component cannot manage static nodes. Each node is separately evaluated to determine whether the node needs to be removed. If the resource usage of a node drops below the scale-in threshold, a scale-in activity is triggered for the node. In this case, cluster-autoscaler simulates the eviction of all workloads on the node to determine whether the node can be completely drained. cluster-autoscaler does not drain the nodes that contain specific pods, such as non-DaemonSet pods in the kube-system namespace and pods that are controlled by PodDisruptionBudgets (PDBs). A node is drained before it is removed. After pods on the node are evicted to other nodes, the node can be removed.

- How is a scaling group selected from multiple scaling groups that meet the requirements?

Different scaling groups are treated as nodes in different specifications. cluster-autoscaler selects a scaling group based on a scoring policy that is similar to a scheduling policy. Nodes are first filtered by the scheduling policy. Among the filtered nodes, the nodes that conform to policies, such as affinity settings, are selected. If no scheduling policy or affinity settings are configured, cluster-autoscaler selects a scaling group based on the least-waste policy. The least-waste policy selects the scaling group that has the fewest idle resources after simulation. If a scaling group of CPU-accelerated nodes and a scaling group of GPU-accelerated nodes both meet the requirements, the scaling group of CPU-accelerated nodes is selected by default.

- How can the success rate of auto scaling be increased? The success rate of auto scaling depends on the following factors:
  - Whether the scheduling policy is met

    After you configure a scaling group, you must be aware of the pod scheduling policies that the scaling group supports. If you are unaware of the pod scheduling policies, you can simulate a scaling activity by using the node selectors of pending pods and the labels of the scaling group.
  - Whether resources are sufficient

    After the scaling simulation is complete, a scaling group is selected. However, the scaling activity fails if the specified types of Elastic Compute Service (ECS) instances in the scaling group are out of stock. To increase the success rate of auto scaling, you can select different types of instances in more than one zone.
- How can auto scaling be accelerated?
  - Method 1: Perform auto scaling in swift mode. After a scaling group experiences a scale-in activity and a scale-out activity, the swift mode is enabled for the scaling group.
  - Method 2: Use custom images that are created from the base image of Alibaba Cloud Linux 2 (formerly known as Aliyun Linux 2). This ensures that the resources of Infrastructure as a Service (IaaS) are delivered 50% faster.

Considerations

- For each account, the default CPU quota for pay-as-you-go instances in each region is 50 vCPUs. You can add at most 48 custom route entries to each route table of a virtual private cloud (VPC). To request a quota increase, submit a ticket.
- The stock of ECS instances may be insufficient for auto scaling if you specify only one ECS instance type for a scaling group. We recommend that you specify multiple ECS instance types with the same specification for a scaling group. This increases the success rate of auto scaling.
- In swift mode, when a node is shut down and reclaimed, the node stops running and enters the *NotReady* state. When a scale-out activity is triggered, the state of the node changes to *Ready*.
- If a node is shut down and reclaimed in swift mode, you are charged only for the disks. This rule does not apply to nodes that use local disks, such as the instance type of ecs.d1ne.2xlarge, for which you are also charged a computing fee. If the stock of nodes is sufficient, nodes can be launched within a short period of time.
- If elastic IP addresses (EIPs) are associated with pods, we recommend that you do not delete the scaling group or remove ECS instances from the scaling group in the ECS console. Otherwise, these EIPs cannot be automatically released.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the Clusters page, find the cluster that you want to manage and choose **More > Auto Scaling** in the **Actions** column.
4. On the **Configure Auto Scaling** page, set the following parameters and click **Submit**.

Parameter	Description
Cluster	The name of the cluster for which you want to enable auto scaling.

Parameter	Description
Scale-in Threshold	For a scaling group that is managed by cluster-autoscaler, set the value to the ratio of the requested resources per node to the total resources per node. If the actual value is lower than the threshold, the node is removed from the cluster. **Note** In auto scaling, a scale-out activity is automatically triggered based on node scheduling. Therefore, you need to set only scale-in parameters.
GPU Scale-in Threshold	The scale-in threshold for GPU-accelerated instances. If the actual value is lower than the threshold, the node is removed from the cluster.
Defer Scale-in For	The amount of time that the cluster must wait before the cluster scales in. The default value is 10 minutes.
Cooldown	The cooldown period after a scale-in activity is triggered. No scale-in activity is triggered during the cooldown period. The default value is 10 minutes.

5. Click **Create Scaling Group** and specify the type of resource for auto scaling based on your business requirements. Regular instances and GPU-accelerated instances are supported.
6. In the **Auto Scaling Group Configuration** dialog box, set the following parameters.

Parameter	Description
Region	The region where you want to deploy the scaling group. The scaling group and the Kubernetes cluster must be deployed in the same region. You cannot change the region after the scaling group is created.
VPC	The scaling group and the Kubernetes cluster must be deployed in the same VPC.
VSwitch	The vSwitches of the scaling group. You can specify vSwitches of different zones. The vSwitches allocate pod CIDR blocks to the scaling group.

7. Configure worker nodes.

Parameter	Description
Instance Type	The instance types in the scaling group.
Selected Types	The instance types that you select. You can select at most 10 instance types.
System Disk	The system disk of the scaling group.
Mount Data Disk	Specify whether to mount data disks to the scaling group. By default, no data disk is mounted.

Parameter	Description
Instances	The number of instances contained in the scaling group. **Note** ◦ Existing instances in the cluster are excluded. ◦ By default, the minimum number of instances is 0. If you specify one or more instances, the system adds the instances to the scaling group. When a scale-out activity is triggered, the instances in the scaling group are added to the cluster to which the scaling group is bound.
Password	Use a password. ◦ Password: Enter the password that is used to log on to the nodes. ◦ Confirm Password: Enter the password again.
Scaling Mode	You can select **Standard** or **Swift**.
RDS Whitelist	The ApsaraDB RDS instances that can be accessed by the nodes in the scaling group after a scaling activity is triggered.
Label	Labels are automatically added to nodes that are added to the cluster by scale-out activities.
ECS Label	You can add labels to the selected ECS instances.
Taints	After you add taints to a node, Container Service no longer schedules pods to the node.
CPU Policy	Specify the CPU policy. Valid values: ◦ None: indicates that the default CPU affinity is used. This is the default policy. ◦ Static: allows pods with specific resource characteristics on the node to be granted with enhanced CPU affinity and exclusivity.

8. Set advanced options.

Parameter	Description
Custom Security Group	Set a custom security group.
Custom Image	You can select a custom image. Then, all nodes in the Kubernetes cluster are deployed based on the image.
User Data	Customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ◦ Run scripts during instance startup. ◦ Import user data as normal data to an ECS instance for future reference.

9. Click **OK**.

Check the results

In the left-side navigation pane, choose **Applications > Deployments**, select the kube-system namespace. You can find the cluster-autoscaler component. This indicates that the scaling group is created.

FAQ

- Why does the auto scaling component fail to add nodes after a scale-out activity is triggered?

  Check whether the following situations exist:

  - The instance types in the scaling group cannot fulfill the resource request from pods. By default, system components are installed for each node. Therefore, the requested pod resources must be less than the resource capacity of the instance type.
  - The Resource Access Management (RAM) role does not have the permissions to manage the Kubernetes cluster. You must complete the authorization for each Kubernetes cluster that is involved in the scale-out activity.
  - The Kubernetes cluster cannot connect to the Internet. The auto scaling component must call Alibaba Cloud API operations. Therefore, the nodes must have access to the Internet.

- Why does the auto scaling component fail to remove nodes after a scale-in activity is triggered?

  Check whether the following situations exist:

  - The requested resource threshold of each pod is higher than the configured scale-in threshold.
  - Pods that belong to the *kube-system* namespace are running on the node.
  - A scheduling policy forces the pods to run on the current node. Therefore, the pods cannot be scheduled to other nodes.
  - PodDisruptionBudget is set for the pods on the node and the minimum value of PodDisruptionBudget is reached.

  For more information about FAQ, see open source component.

- How does the system choose a scaling group for a scaling activity?

  When pods cannot be scheduled to nodes, the auto scaling component simulates the scheduling of the pods based on the configuration of scaling groups. The configuration includes labels, taints, and instance specifications. If a scaling group can simulate the scheduling of the pods, this scaling group is selected for the scale-out activity. If more than one scaling groups meet the requirements, the system selects the scaling group that has the fewest idle resources after simulation.

3.1.6.13.2. Horizontal pod autoscaling

You can create an application that has Horizontal Pod Autoscaling (HPA) enabled in the Container Service console. HPA can automatically scale container resources for your application. You can also use a YAML file to describe HPA settings.

Create an application that has HPA enabled in the Container Service console

Container Service provided by Alibaba Cloud is integrated with HPA. You can create an application that has HPA enabled in the Container Service console. You can enable HPA when you create an application or after the application is created.

Enable HPA when you create an application

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click the name or click **Details** in the **Actions** column.

4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** tab, click **Create from Image**.
6. On the **Basic Information** wizard page, enter a name for your application, set other required parameters, and then click **Next**.

Parameter	Description
Name	Enter a name for the application.
Replicas	The number of pods that are provisioned for the application. Default value: 2.
Type	The type of the application. You can select **Deployments**, **StatefulSets**, **Jobs**, **Cron Jobs**, or **DaemonSets**.
Label	Add a label to the application. The label is used to identify the application.
Annotations	Add an annotation to the application.
Synchronize Timezone	Specify whether to synchronize the time zone between nodes and containers.

7. On the **Container** wizard page, set the container parameters, select an image, and then configure the required computing resources. Click **Next**. For more information, see Configure the containers.

   Note You must configure the required computing resources for the Deployment. Otherwise, you cannot enable HPA.

8. On the **Advanced** wizard page, find the **Access Control** section, click **Create** on the right side of Services, and then set the parameters. For more information, see Create an application from an image.
9. On the **Advanced** wizard page, select **Enable** for **HPA** and configure the scaling threshold and related settings.
   - **Metric**: Select CPU Usage or Memory Usage. The selected resource type must be the same as the one that you have specified in the Required Resources field.
   - **Condition**: Specify the resource usage threshold. HPA triggers scaling activities when the threshold is exceeded.
   - **Max. Replicas**: Specify the maximum number of pods to which the Deployment can be scaled.
   - **Min. Replicas**: Specify the minimum number of pods that must run for the Deployment.
10. In the lower-right corner of the Advanced wizard page, click **Create**. The application is created with HPA enabled.
    Verify the result
    i. Click **View Details** or choose **Workloads > Deployments**. On the page that appears, click the **name of the created application** or click **Details** in the **Actions** column. Then, click the **Horizontal Pod Autoscaler** tab to view information about the scaling group of the application.
    ii. After the application starts to run, container resources are automatically scaled based on the CPU utilization. You can also check whether HPA is enabled in the staging environment by performing a CPU stress test on the pods of the application. Verify that the pods are automatically scaled within 30 seconds.

Enable HPA after an application is created

This example describes how to enable HPA for a stateless application.

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage, and click the name of the cluster or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, click the name of the application that you want to manage.
6. Click the **Pod Scaling** tab and click **Create**.
7. In the **Create** dialog box, configure the HPA settings. For more information about how to set the parameters, see HPA settings in Step 9.
8. Click **OK**.

Create an application that has HPA enabled by using kubectl

You can also create a Horizontal Pod Autoscaler by using an orchestration template and associate the Horizontal Pod Autoscaler with the Deployment for which you want to enable HPA. Then, you can run **kubectl** commands to enable HPA.

In the following example, HPA is enabled for an NGINX application.

1. Create a file named *nginx.yml* and copy the following content into the file.

   The following code block is a YAML template that is used to create a Deployment:

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx
  labels:
    app: nginx
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.7.9 # replace it with your exactly <image_name:tags>
        ports:
        - containerPort: 80
        resources:
          requests:                         ##To enable HPA, you must set this parameter.
            cpu: 500m
```

2. Run the following command to create an NGINX application:

```
kubectl create -f nginx.yml
```

3. Create a Horizontal Pod Autoscaler.

   Use scaleTargetRef to associate the Horizontal Pod Autoscaler with the Deployment named **nginx**.

```
apiVersion: autoscaling/v2beta1
kind: HorizontalPodAutoscaler
metadata:
  name: nginx-hpa
  namespace: default
spec:
  scaleTargetRef:                              ##Associate the Horizontal Pod Autoscaler with the n
ginx Deployment.
    apiVersion: apps/v1
    kind: Deployment
    name: nginx
  minReplicas: 1
  maxReplicas: 10
  metrics:
  - type: Resource
    resource:
      name: cpu
      targetAverageUtilization: 50
```

Note You must configure the requested resources for the pods of the application. Otherwise, the Horizontal Pod Autoscaler cannot be started.

4. Run the `kubectl describe hpa name` command. The following output is an example of a warning that is returned:

```
Warning  FailedGetResourceMetric       2m (x6 over 4m)  horizontal-pod-autoscaler  missing request
for cpu on container nginx in pod default/nginx-deployment-basic-75675f5897-mqzs7
Warning  FailedComputeMetricsReplicas  2m (x6 over 4m)  horizontal-pod-autoscaler  failed to get c
pu utilization: missing request for cpu on container nginx in pod default/nginx-deployment-basic-7
5675f5
```

5. After the Horizontal Pod Autoscaler is created, run the `kubectl describe hpa name` command.

   If the following output is returned, it indicates that the Horizontal Pod Autoscaler is running as expected:

```
Normal SuccessfulRescale 39s horizontal-pod-autoscaler New size: 1; reason: All metrics below targ
et
```

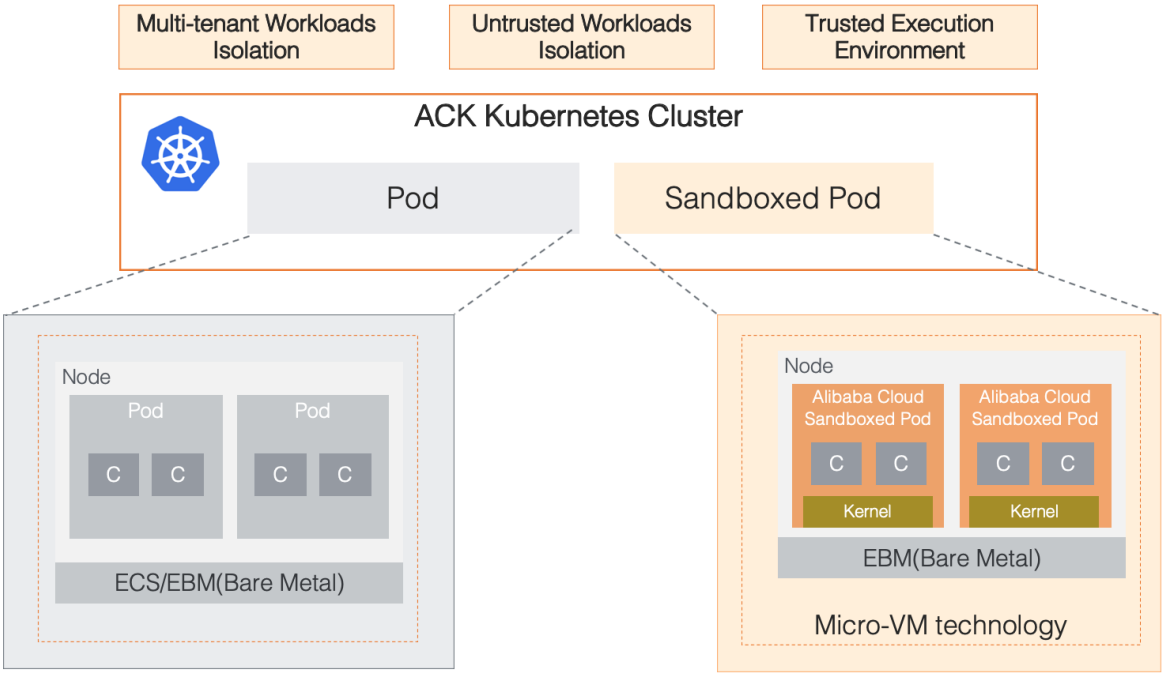
   If the CPU utilization of the NGINX application pod exceeds 50% as specified in the HPA settings, the Horizontal Pod Autoscaler automatically adds pods. If the CPU utilization of the NGINX application pod drops below 50%, the Horizontal Pod Autoscaler automatically removes pods.

3.1.6.14. Sandboxed-containers

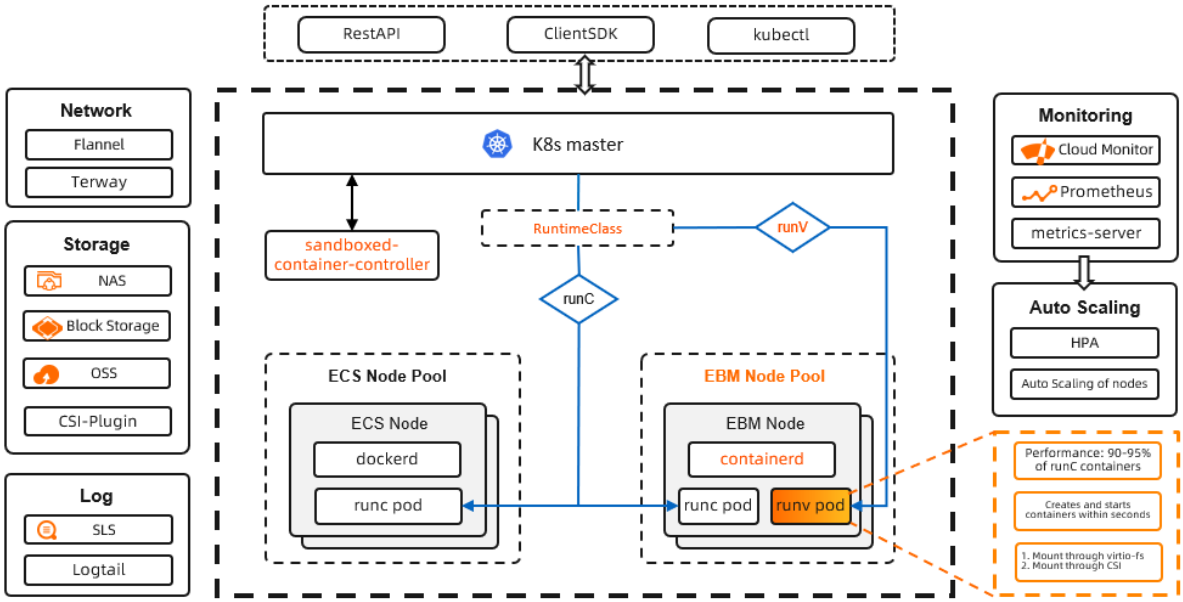
3.1.6.14.1. Overview

Sandboxed-Container is an alternative to the Docker runtime. Sandboxed-Container allows you to run applications in a sandboxed and lightweight virtual machine that has a dedicated kernel. This enhances resource isolation and improves security.

Sandboxed-Container is applicable to scenarios such as untrusted application isolation, fault isolation, performance isolation, and workload isolation among multiple users. Sandboxed-Container provides higher security. Sandboxed-Container has minor impacts on application performance and offers the same user experience as Docker in terms of logging, monitoring, and elastic scaling.



Architecture



Features

Sandboxed-Container is a container-securing runtime that is developed by Alibaba Cloud based on sandboxed and lightweight virtual machines. Compared with Sandboxed-Container V1, Sandboxed-Container V2 maintains the same isolation performance and reduces the pod overhead by 90%. It also allows you to start sandboxed containers 3 times faster and increases the maximum number of pods that can be deployed on a host by 10 times. Sandboxed-Container V2 provides the following key features:

- Strong isolation based on sandboxed and lightweight virtual machines.
- Good compatibility with runC in terms of application management.
- Network Attached Storage (NAS) file systems, disks, and Object Storage Service (OSS) buckets can be mounted both directly and through virtio-fs.

- The same user experience as runC in terms of logging, monitoring, and storage.
- Supports RuntimeClass (runC and runV). For more information, see RuntimeClass.
- Easy to use with minimum requirements on technical skills.
- Higher stability than Kata Containers. For more information about Kata Containers, see Kata Containers.

3.1.6.14.2. Create a Kubernetes cluster that runs sandboxed containers

This topic describes how to create a Kubernetes cluster that runs sandboxed containers in the Container Service console.

1. Log on to the Container Service console.
2. In the left-side navigation pane of the Container Service console, click **Clusters**. On the Clusters page that appears, click **Create Kubernetes Cluster** in the upper-right corner.
3. On the **Create Cluster** page, set basic configurations for the cluster.

Parameter	Description
Cluster Name	Enter a name for the cluster. The name must be 1 to 63 characters in length, and can contain digits, letters, and hyphens (-). **Note** The cluster name must be unique among clusters that belong to the same Alibaba Cloud account.
Region	Select the region where you want to deploy the Kubernetes cluster.
VPC	You can select a virtual private cloud (VPC) from the drop-down list. ○ If the specified VPC is already associated with a NAT gateway, the cluster uses this NAT gateway. ○ Otherwise, the system automatically creates a NAT gateway. If you do not want the system to create a NAT gateway, clear **Configure SNAT for VPC**. **Note** If you disallow the system to automatically create a NAT gateway and want the VPC to access the Internet, you must manually associate the VPC with a NAT gateway or create Source Network Address Translation (SNAT) rules for the VPC.
VSwitch	Select one or more vSwitches for the cluster. You can select up to three vSwitches that are deployed in different zones.
Kubernetes Version	Select a Kubernetes version.
Container Runtime	Select a runtime for the Kubernetes cluster.
Billing Method	Only pay-as-you-go nodes are supported.

Parameter	Description
Master Configurations	Set the Instance Type and System Disk parameters: ◦ Master Node Quantity: You can add up to three master nodes. ◦ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ◦ System Disk: Standard SSDs and ultra disks are supported. **Note** You can select **Enable Backup** to back up disk data.
Worker Instance	By default, **Create Instance** is selected.
Worker Configurations	If **Worker Instance** is set to **Create Instance**, set the following parameters: ◦ Instance Type: Select Elastic Compute Service (ECS) bare metal instance types. ◦ Selected Types: The selected instance types are displayed. ◦ Quantity: Set the number of worker nodes. ◦ System Disk: Standard SSDs and ultra disks are supported. **Note** You can select **Enable Backup** to back up disk data. ◦ Mount Data Disk: Standard SSDs and ultra disks are supported. **Note** You can enable disk encryption and data backup when you mount disks. The disks are used to store the root file systems of containers on the nodes. Therefore, you must mount a disk of at least 200 GiB. We recommend that you mount a disk of at least 1 TiB.
Password	Set a password that is used to log on to the nodes. **Note** The password must be 8 to 30 characters in length, and must contain at least three of the following types of character: uppercase letters, lowercase letters, digits, and special characters.
Confirm Password	Enter the password again.
Network Plug-in	Flannel and Terway are supported. By default, Flannel is selected.
Pod CIDR Block and Service CIDR	These parameters are optional. For more information, see *Network planning* in *VPC User Guide*. **Note** These parameters are available only when you select an existing VPC.
Configure SNAT	This parameter is optional. If you clear Configure SNAT for VPC, you must create a NAT gateway or configure SNAT rules for the VPC.

Parameter	Description
Access to the Internet	Specify whether to expose the API server with an elastic IP address (EIP). The Kubernetes API server provides multiple HTTP-based RESTful APIs that can be used to create, delete, modify, query, and watch resource objects such as pods and Services. ◦ If you select this check box, an EIP is created and attached to an internal-facing Server Load Balancer (SLB) instance. Port 6443 used by the API server is exposed on the master nodes. You can connect to and manage the cluster by using kubeconfig over the Internet. ◦ If you clear this check box, no EIP is created. You can connect to and manage the cluster only by using kubeconfig from within the VPC.
Ingress	Specify whether to install Ingress controllers. By default, **Install Ingress Controller** is selected.
Log Service	If you enable Log Service, you can select an existing project or create a project. If you select **Enable Log Service**, the Log Service plug-in is automatically installed in the cluster. If you select **Create Ingress Dashboard**, Ingress access logs are collected and displayed on dashboards.
Volume Plug-in	By default, CSI is selected.
Deletion Protection	If you select this check box, the cluster cannot be deleted in the console or by calling API operations.
RDS Whitelist	Add the IP addresses of the nodes to the whitelist of the ApsaraDB RDS instance that is allowed to access the Kubernetes cluster. **Note** To enable an ApsaraDB RDS instance to access the Kubernetes cluster, you must deploy the ApsaraDB RDS instance in the same VPC as the Kubernetes cluster.
Node Protection	This check box is selected by default to prevent nodes from being deleted in the console or by calling API operations.
Label	Add labels to the cluster.

4. Complete the advanced settings of the cluster.

Parameter	Description
IP Addresses per Node	The number of IP addresses that is assigned to a node.
Kube-proxy Mode	iptables and IPVS are supported. ◦ iptables is a mature and stable kube-proxy mode. It uses iptables rules to conduct service discovery and load balancing. The performance of this mode is restricted by the size of the Kubernetes cluster. This mode is suitable for Kubernetes clusters that manage a small number of Services. ◦ IPVS is a high-performance kube-proxy mode. It uses Linux Virtual Server (LVS) to conduct service discovery and load balancing. This mode is suitable for Kubernetes clusters that manage a large number of Services. We recommend that you use this mode in scenarios where high-performance load balancing is required.

Parameter	Description
Custom Node Name	Specify whether to use a custom node name.
Node Port Range	Specify the node port range.
Taints	Add taints to all worker nodes in the Kubernetes cluster.
CPU Policy	Specify the CPU policy. Valid values: ○ None: indicates that the default CPU affinity is used. This is the default policy. ○ Static: allows pods with specific resource characteristics on the node to be granted with enhanced CPU affinity and exclusivity.
Cluster Domain	The default domain name of the cluster is cluster.local. You can specify a custom domain name.
Cluster CA	Specify whether to enable the cluster CA certificate.
User Data	You can customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ○ Run user data scripts during instance startup. ○ Import user data as common data to an ECS instance for future reference.

5. Click **Create Cluster** in the upper-right corner of the page.
6. On the **Confirm** page, click **OK** to start the deployment.

Result

After the cluster is created, you can find the cluster on the **Clusters** page in the Container Service console.

3.1.6.14.3. Expand a Container Service cluster that runs sandboxed containers

This topic describes how to scale out the worker nodes in a Container Service cluster that runs sandboxed containers in the Container Service console.

Prerequisites

You cannot scale out the master nodes in a Container Service cluster that runs sandboxed containers.

To expand a Container Service cluster that runs sandboxed containers, you must set the parameters as required in the following table. Otherwise, the added nodes cannot run sandboxed containers.

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Expand** in the **Actions** column.
4. Go to the Expand page and set the required parameters.

   In this example, the number of worker nodes in the Container Service cluster is increased from three to five. The following table describes the required parameters.

Parameter	Description
Cluster Name	By default, the name of the Container Service cluster appears.
Region	The region where the Container Service cluster is deployed.
Container Runtime	By default, Sandboxed-Container appears.
VPC	You can select a virtual private cloud (VPC) from the drop-down list. ○ If the specified VPC is already associated with a NAT gateway, the cluster uses this NAT gateway. ○ Otherwise, the system automatically creates a NAT gateway. If you do not want the system to create a NAT gateway, clear **Configure SNAT for VPC**. **Note** If you disallow the system to automatically create a NAT gateway and want the VPC to access the Internet, you must manually associate the VPC with a NAT gateway or create Source Network Address Translation (SNAT) rules for the VPC.
VSwitch	Select one or more vSwitches for the cluster. You can select up to three vSwitches that are deployed in different zones.
Billing Method	Only pay-as-you-go nodes are supported.
Existing Worker Nodes	The number of existing workers in the Container Service cluster.
Nodes to Add	Set the number of worker nodes to add.
Worker Nodes After Scaling	The number of worker nodes after the scaling.
Instance Type	Select ECS Bare Metal Instance.
Selected Types	The selected instance types are displayed.
System Disk	Standard SSDs and ultra disks are supported. **Note** You can select **Enable Backup** to back up disk data.
Mount Data Disk	Standard SSDs and ultra disks are supported. **Note** You can enable disk encryption and data backup when you mount disks. The disks are used to store the root file systems of containers on the nodes. Therefore, you must mount a disk of at least 200 GiB. We recommend that you mount a disk of at least 1 TiB.
Password	○ **Password**: Enter the password that is used to log on to the nodes. ○ **Confirm Password**: Enter the password again.
RDS Whitelist	Set the Apsara RDS whitelist. Add the IP addresses of the nodes in the cluster to the RDS whitelist.
Label	Add labels to the cluster.

Parameter	Description
Taints	Add taints to all worker nodes in the Kubernetes cluster.
CPU Policy	Specify the CPU policy. Valid values: ○ None: indicates that the default CPU affinity is used. This is the default policy. ○ Static: allows pods with specific resource characteristics on the node to be granted with enhanced CPU affinity and exclusivity.
User Data	You can customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ○ Run user data scripts during instance startup. ○ Import user data as common data to an ECS instance for future reference.

5. Click **Submit**.

What's next

After the Container Service cluster is expanded, go to the details page of the Container Service cluster. In the left-side navigation pane, choose **Clusters > Node Pools**. You can find that the number of worker nodes is increased from 3 to 5.

3.1.6.14.4. Create an application that runs in sandboxed containers

This topic describes how to use an image to create an NGINX application that runs in sandboxed containers. The NGINX application is accessible over the Internet.

Prerequisites

A cluster that contains sandboxed containers is created. For more information, see Create a Kubernetes cluster that supports sandboxed containers.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Workloads > Deployments**.
5. On the **Deployments** page, select the namespace and click **Create from Image** in the upper-right corner.
6. On the **Basic Information** wizard page, specify the basic information of the application and click **Next**.

   Set **Container Runtime** to **runv**. Set the following parameters: **Name**, **Replicas**, **Type**, **Label**, and **Annotations**. Select whether you want to enable **Synchronize Timezone**. The number of replicas specifies the number of pods that are provisioned for in the application.

   > **Note**
   >
   > **Deployments** is selected in this example.

7. Configure containers.

> **Note** In the upper part of the **Container** wizard page, click **Add Container** to add more containers for the application.

The following table describes the parameters that are required to configure the containers.

- General settings

Parameter	Description
Image Name	Click **Select Image**. In the dialog box that appears, select an image and click **OK**. In this example, an NGINX image is selected. You can also enter the address of an image stored in a private registry. The image address must be in the following format: `domainname/namespace/imagename:tag`.
Image Version	■ Click **Select Image Version** and select an image version. If you do not specify an image version, the latest image version is used. ■ You can select the following image pulling policies: ■ **ifNotPresent**: If the image that you want to pull is found in the region where the cluster is deployed, Container Service uses the local image. Otherwise, Container Service pulls the image from the corresponding repository. ■ **Always**: Container Service pulls the image from the repository each time the application is deployed or expanded. ■ **Never**: Container Service uses only images on your on-premise machine. **Note** If you select **Image Pull Policy**, no image pulling policy is applied. ■ To pull the image without a Secret, click **Set Image Pull Secret** to set a Secret for pulling images.
Resource Limit	You can specify an upper limit for the CPU, memory, and ephemeral storage space that the container can consume. This prevents the container from occupying an excessive amount of resources. The CPU resource is measured in milicores (one thousandth of one core). The memory resource is measured in MiB. The ephemeral storage resource is measured in GiB.
Required Resources	The amount of CPU and memory resources that are reserved for this application. These resources are exclusive to the container. This prevents the application from becoming unavailable if other services or processes compete for computing resources.
Container Start Parameter	■ stdin: Pass stdin to the container. ■ tty: Stdin is a TeleTYpewriter (TTY).
Privileged Container	■ If you select Privileged Container, privileged=true is set for the container and the privilege mode is enabled. ■ If you do not select Privileged Container, privileged=false is set for the container and the privilege mode is disabled.
Init Container	If you select Init Container, an init container is created. An init container provides tools to manage pods. For more information, see Init Containers.

- (Optional)Ports

  Configure container ports.

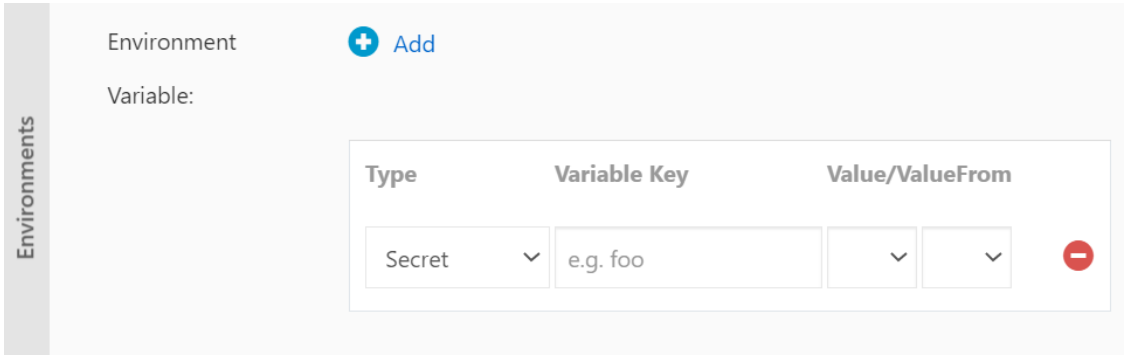
  - Name: Enter a name for the port.
  - Container Port: Enter the container port that you want to open. Enter a port number from 1 to 65535.
  - Protocol: Select TCP or UDP.

- (Optional)Environments

  You can configure environment variables for pods in key-value pairs. Environment variables are used to apply pod configurations to containers. For more information, see Pod variables.

  - Type: Select the type of the environment variable. You can select **Custom**, **ConfigMaps**, **Secrets**, **Value/ValueFrom**, or **ResourceFieldRef**. If you select ConfigMaps or Secret as the type of the environment variable, all values in the selected ConfigMap or Secret are passed to the container environment variables. In this example, Secret is selected.

    Select **Secrets** from the Type drop-down list and select a Secret from the **Value/ValueFrom** drop-down list. All values in the selected Secret are passed to the environment variable.

    In this case, the YAML file that is used to deploy the application contains the settings that reference all data in the selected Secret.

    ```
    envFrom:
      - secretRef:
          name: test
    ```

  - Variable Key: Specify the name of the environment variable.
  - Value/ValueFrom: Specify the value that is referenced by the environment variable.

- (Optional)Health Check

  Health check settings include liveness and readiness probes. Liveness probes determine when to restart the container. Readiness probes determine whether the container is ready to accept network traffic. For more information about health checks, see Configure Liveness, Readiness, and Startup Probes.

Request type	Description

Request type	Description
HTTP	Sends an HTTP GET request to the container. You can configure the following parameters: ■ Protocol: HTTP or HTTPS. ■ Path: the requested path on the server. ■ Port: Enter the container port that you want to open. Enter a port number from 1 to 65535. ■ HTTP Header: Enter the custom headers in the HTTP request. Duplicate headers are allowed. Key-value pairs are supported. ■ Initial Delay (s): the initialDelaySeconds field in the YAML file. This field specifies the time (in seconds) that the system must wait before it can send a probe to the container after the container is started. Default value: 3. ■ Period (s): the periodSeconds field in the YAML file. This field specifies the interval (in seconds) at which probes are performed. Default value: 10. Minimum value: 1. ■ Timeout (s): the timeoutSeconds field in the YAML file. This field specifies the time (in seconds) after which a probe times out. Default value: 1. Minimum value: 1. ■ Healthy Threshold: the minimum number of times that an unhealthy container must consecutively pass health checks before it is considered healthy. Default value: 1. Minimum value: 1. For liveness probes, this parameter must be set to 1. ■ Unhealthy Threshold: the minimum number of times that a healthy container must consecutively fail health checks before it is considered unhealthy. Default value: 3. Minimum value: 1.
TCP	Sends a TCP socket to the container. kubelet attempts to open the socket on the specified port. If the connection can be established, the container is considered healthy. Otherwise, the container is considered unhealthy. You can set the following parameters: ■ Port: Enter the container port that you want to open. Enter a port number from 1 to 65535. ■ Initial Delay (s): the initialDelaySeconds field in the YAML file. This field specifies the time (in seconds) that the system must wait before it can send a probe to the container after the container is started. Default value: 15. ■ Period (s): the periodSeconds field in the YAML file. This field specifies the time interval (in seconds) at which probes are performed. Default value: 10. Minimum value: 1. ■ Timeout (s): the timeoutSeconds field in the YAML file. This field specifies the time (in seconds) after which a probe times out. Default value: 1. Minimum value: 1. ■ Healthy Threshold: the minimum number of times that an unhealthy container must consecutively pass health checks before it is considered healthy. Default value: 1. Minimum value: 1. For liveness probes, this parameter must be set to 1. ■ Unhealthy Threshold: the minimum number of times that a healthy container must consecutively fail health checks before it is considered unhealthy. Default value: 3. Minimum value: 1.

Request type	Description
Command	Runs a probe command in the container to check the health status of the container. You can set the following parameters: ■ Command: the probe command that is run to check the health status of the container. ■ Initial Delay (s): the initialDelaySeconds field in the YAML file. This field specifies the time (in seconds) that the system must wait before it can send a probe to the container after the container is started. Default value: 5. ■ Period (s): the periodSeconds field in the YAML file. This field specifies the interval (in seconds) at which probes are performed. Default value: 10. Minimum value: 1. ■ Timeout (s): the timeoutSeconds field in the YAML file. This field specifies the time (in seconds) after which a probe times out. Default value: 1. Minimum value: 1. ■ Healthy Threshold: the minimum number of times that an unhealthy container must consecutively pass health checks before it is considered healthy. Default value: 1. Minimum value: 1. For liveness probes, this parameter must be set to 1. ■ Unhealthy Threshold: the minimum number of times that a healthy container must consecutively fail health checks before it is considered unhealthy. Default value: 3. Minimum value: 1.

- Lifecycle

  You can set the following parameters to configure the lifecycle of the container: Start, Post Start, and Pre Stop. For more information, see Configure the lifecycle of a container.

  - **Start**: Set the command and parameter that take effect before the container starts.
  - **Post Start**: Set the command that takes effect after the container starts.
  - **Pre Stop**: Set the command that takes effect before the container stops.

- (Optional)Volume

  You can mount local volumes and persistent volume claims (PVCs) to the container.

  - **Add Local Storage**: You can select HostPath, ConfigMap, Secret, and EmptyDir. The source directory or file is mounted to a path in the container. For more information, see Volumes.
  - **Add PVC**: Cloud Storage is supported.

  In this example, a PVC named disk-ssd is mounted to the */tmp* path of the container.

- (Optional)Log

  Configure **Log Service**. You can specify log collection configurations and add custom tags.

  > **Notice** Make sure that the Log Service agent is installed in the cluster.

Parameter	Description
	Logstore: creates a Logstore in Log Service to store collected log data.

Parameter Collection Configuration	Description
	Log Path in Container: specifies stdout or a path to collect log data ■ stdout: specifies that the stdout files are collected. ■ Text Logs: specifies that log data in the specified path of the container is collected. In this example, */var/log/nginx* is specified as the path. Wildcard characters can be used to specify the path.
Custom Tag	You can also add custom tags. Custom tags are added to the log data of the container when the log data is collected. Log data with tags is easier to aggregate and filter.

8. Configure the parameters based on your business requirements and click **Next**.
9. (Optional)Configure advanced settings.
   - Access Control

> **Note**
>
> You can configure the following access control settings based on your business requirements:
>
> - Internal applications: For applications that run inside the cluster, you can create a ClusterIP or NodePort Service to enable internal communication.
> - External applications: For applications that are open to the Internet, you can configure access control by using one of the following methods:
>   - Create a LoadBalancer Service and enable access to your application over the Internet by using a Server Load Balancer (SLB) instance.
>   - Create an Ingress and use the Ingress to expose your application to the Internet. For more information, see Ingress.

You can also specify how the backend pods are exposed to the Internet. In this example, a ClusterIP Service and an Ingress are created to expose the NGINX application to the Internet.

Parameter	Description
Services	Click **Create** on the right side of **Services**. In the Create Service dialog box, set the parameters. Select **Cluster IP**.
Ingresses	Click **Create** on the right side of **Ingresses**. In the Create dialog box, set the parameters. **Note** When you create an application from an image, you can create an Ingress only for one Service. In this example, a virtual hostname is used as the test domain name. You must add the following entry to the hosts file to map the domain name to the IP address of the Ingress. In actual scenarios, use a domain name that has obtained an Internet Content Provider (ICP) number. `101.37.22*.***  foo.bar.com  #The IP address of the Ingress.`

You can find the created Service and Ingress in the **Access Control** section. You can click **Update** or **Delete** to change the configurations.

- Scaling

  Specify whether to enable **HPA** to automatically scale the number of pods based on the CPU and memory usage. This enables the application to run smoothly at different load levels.

  > **Note** To enable HPA, you must configure required resources for the container. Otherwise, HPA does not take effect.

  - **Metric**: Select CPU Usage or Memory Usage. The selected resource type must be the same as the one you have specified in the Required Resources field.
  - **Condition**: Specify the resource usage threshold. HPA triggers scaling events when the threshold is exceeded.
  - **Max. Replicas**: Specify the maximum number of replicated pods to which the application can be scaled.
  - **Min. Replicas**: Specify the minimum number of replicated pods that must run.

- Scheduling

  You can set the following parameters: Update Method, Node Affinity, Pod Affinity, Pod Anti Affinity, and Toleration. For more information, see Affinity and anti-affinity.

  > **Note** Node affinity and pod affinity affect pod scheduling based on node labels and pod labels. You can add node labels and pod labels that are provided by Kubernetes to configure node affinity and pod affinity. You can also add custom labels to nodes and pods, and then configure node affinity and pod affinity based on these custom labels.

Parameter	Description
Update Method	Select Rolling Update or OnDelete. For more information, see Deployments.
Node Affinity	Set **Node Affinity** by adding labels to worker nodes. Node Affinity supports required and preferred rules, and various operators, such as In, NotIn, Exists, DoesNotExist, Gt, and Lt. ■ **Required**: Specify the rules that must be matched for pod scheduling. In the YAML file, these rules are defined by the requiredDuringSchedulingIgnoredDuringExecution field of the nodeAffinity parameter. These rules have the same effect as the `Node Selector` parameter. In this example, pods can be scheduled only to nodes with the specified labels. You can create multiple required rules. However, only one of them must be met. ■ **Preferred**: Specify the rules that are not required to be matched for pod scheduling. Pods are scheduled to a node that matches the preferred rules when multiple nodes match the required rules. In the YAML file, these rules are defined by the preferredDuringSchedulingIgnoredDuringExecution field of the nodeAffinity parameter. In this example, the scheduler attempts to schedule a pod to a node that matches the preferred rules. You can also set weights for preferred rules. If multiple nodes match the rule, the node with the highest weight is preferred. You can create multiple preferred rules. However, all of them must be met before the pod can be scheduled.

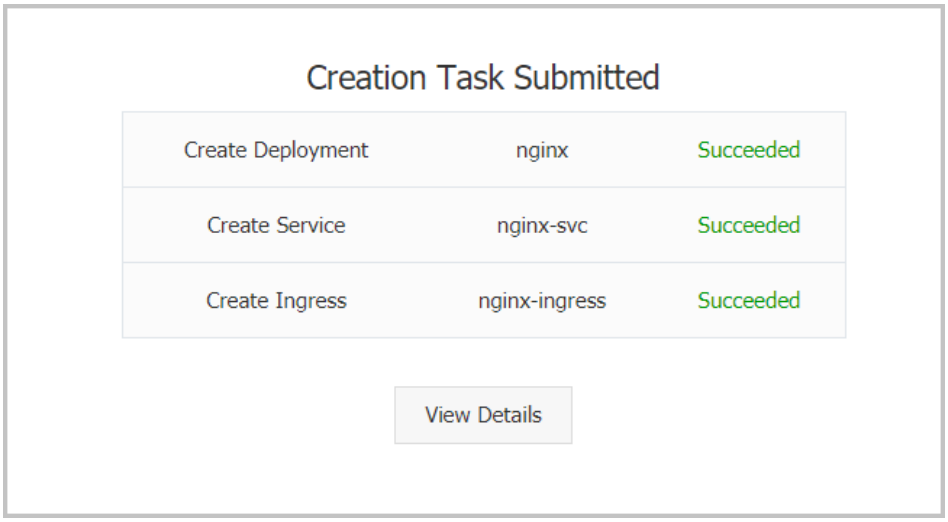
Parameter	Description
Pod Affinity	Pod affinity rules specify how pods are deployed relative to other pods in the same topology domain. For example, you can use pod affinity to deploy services that communicate with each other to the same topological domain, such as a host. This reduces the network latency between these services. Pod affinity enables you to select nodes to which pods can be scheduled based on the labels of other running pods. Pod affinity supports required and preferred rules, and the following operators: `In`, `NotIn`, `Exists`, and `DoesNotExist`. ■ **Required**: Specify rules that must be matched for pod scheduling. In the YAML file, these rules are defined by the requiredDuringSchedulingIgnoredDuringExecution field of the podAffinity parameter. A node must match the required rules before pods can be scheduled to the node. ■ **Namespace**: Specify the namespace to apply the required rule. Pod affinity rules are defined based on the labels that are added to pods and therefore must be scoped to a namespace. ■ **Topological Domain**: Set the topologyKey. This specifies the key for the node label that the system uses to denote the topological domain. For example, if you set the parameter to `kubernetes.io/hostname`, topologies are determined by nodes. If you set the parameter to `beta.kubernetes.io/os`, topologies are determined by the operating systems of nodes. ■ **Selector**: Click Add to add pod labels. ■ **View Applications**: Click **View Applications** and set the namespace and application in the dialog box that appears. You can view the pod labels on the selected application and add the labels as selectors. ■ Required Rules: Specify labels on existing applications, the operator, and the label value. In this example, the required rule specifies that the application to be created is scheduled to a host that runs applications with the `app:nginx` label. ■ **Preferred**: Specify rules that are not required to be matched for pod scheduling. In the YAML file, preferred rules are defined by the preferredDuringSchedulingIgnoredDuringExecution field of the podAffinity parameter. The scheduler attempts to schedule the pod to a node that matches the preferred rules. You can set weights for preferred rules. The other parameters are the same as those of required rules. **Note** **Weight**: Set the weight of a preferred rule to a value from 1 to 100. The scheduler calculates the weight of each node that meets the preferred rule based on an algorithm, and then schedules the pod to the node with the highest weight.

Parameter	Description
Pod Anti Affinity	Pod anti-affinity rules specify that pods are not scheduled to topological domains where pods with matching labels are deployed. Pod anti-affinity rules apply to the following scenarios: ■ Schedule the pods of an application to different topological domains, such as multiple hosts. This allows you to enhance the stability of the service. ■ Grant a pod exclusive access to a node. This enables resource isolation and ensures that no other pod can share the resources of the specified node. ■ Schedule pods of an application to different hosts if the pods may interfere with each other. **Note** The parameters of pod anti-affinity rules are the same as those of pod affinity rules. You can create the rules for different scenarios.
Toleration	Configure toleration rules to allow pods to be scheduled to nodes with matching taints.
Schedule to Virtual Nodes	Specify whether to schedule pods to virtual nodes. This option is unavailable if the cluster does not contain a virtual node.

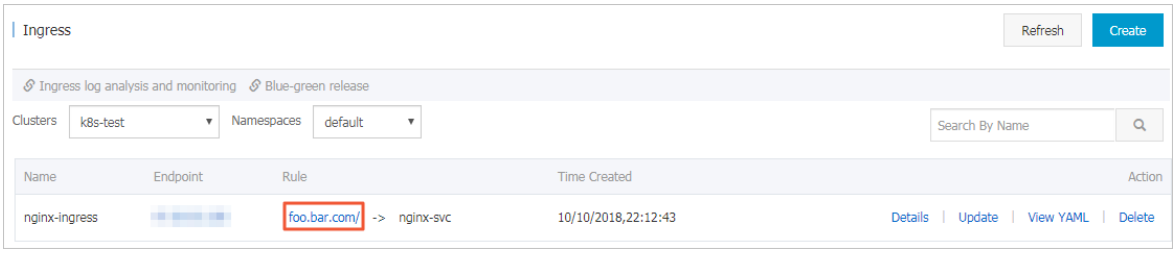
- Labels and Annotations
  - Pod Labels: Add a label to the pod. The label is used to identify the application.
  - Pod Annotations: Add an annotation to the pod.

10. Click **Create**.

    After the application is deployed, you are redirected to the Complete wizard page. The resource objects of the application are displayed. You can find the resource objects under the application and click **View Details** to view application details.



11. In the left-side navigation pane, choose **Ingresses and Load Balancing > Ingresses**.
    The created Ingress rule is displayed on the page.



Result

Enter the test domain in the address bar of your browser and press Enter. The NGINX welcome page appears.

3.1.6.14.5. Configure a Kubernetes cluster that runs both sandboxed and Docker containers

Node pools support multiple types of container runtime. However, nodes in the same node pool must use the same type of container runtime. Container Service allows you to create node pools of different container runtime types for a cluster. This topic describes how to create a node pool that runs sandboxed containers and a node pool that runs Docker containers for a Kubernetes cluster.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

> **Notice** The Kubernetes cluster must meet the following requirements:
>
> - The cluster version must be 1.14.6-aliyun.1 or later.
> - The network plug-in must be Flannel or Terway. Terway must run in One ENI for Multi-Pod mode.
> - The volume plug-in must be CSI-Plugin 1.14.8.39-0d749258-aliyun or later. Flexvolume is not supported.
> - The logtail-ds version must be 0.16.34.2-f6647154-aliyun or later.

Considerations

- By default, a cluster can contain at most 100 nodes.
- Before you add an existing Elastic Compute Service (ECS) instance that is deployed in a virtual private cloud (VPC), make sure that an elastic IP address (EIP) is associated with the ECS instance, or a NAT gateway is created in the VPC. In addition, the nodes that you want to add to the node pool must have access to the Internet. Otherwise, the ECS instance cannot be added.

Create a node pool that runs Docker containers

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.

4. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, click **Create Node Pool** and set the parameters.

   For more information, see Create a Kubernetes cluster. The following table describes the parameters.

Parameter	Description
Name	Enter a name for the node pool.
Container Runtime	Select Docker. This specifies that all containers in the node pool are Docker containers.
Quantity	Specify the initial number of nodes in the node pool. If you do not need to create nodes, set this parameter to 0.
Operating System	The CentOS and Aliyun Linux operating systems are supported.
ECS Label	You can add labels to the ECS instances.
Node Label	You can add labels to the nodes in the cluster.
Custom Resource Group	Specify the resource group of the nodes to be added to the node pool.
Custom Security Group	Select a custom security group.

6. Click **OK**.

Create a node pool that runs sandboxed containers

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, click **Create Node Pool** and set the parameters.

   For more information, see Create a Kubernetes cluster. The following table describes the parameters.

Parameter	Description
Name	Enter a name for the node pool.
Container Runtime	Select Sandboxed-Container. This specifies that all containers in the node pool are sandboxed containers.
Quantity	Specify the initial number of nodes in the node pool. If you do not need to create nodes, set this parameter to 0.
Operating System	Select an operating system. Sandboxed containers support only the Aliyun Linux operating system.
Mount Data Disk	You must mount a disk of at least 200 GiB.
ECS Label	You can add labels to the ECS instances.
Node Label	You can add labels to the nodes in the cluster.

Parameter	Description
Custom Resource Group	Specify the resource group of the nodes to be added to the node pool.
Custom Security Group	Select a custom security group.

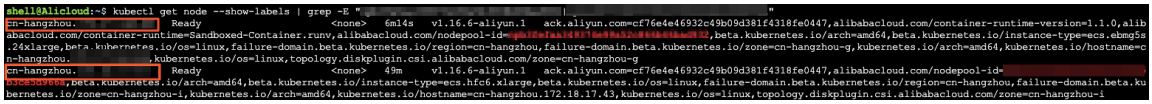
6. Click **OK**.

Result

- After you perform the preceding steps, check the states of the node pools on the **Node Pools** page. If the node pools are in the **Activate** state, the node pools are created.
- You can connect to the Kubernetes cluster and view detailed information about the nodes in the node pools.
  i. On the **Node Pools** page, select a node pool that you have created and click its name. In the **Node Pool Information** section, find and record the **Node Pool ID**.

  ii. Connect to the Kubernetes cluster by using kubectl. For more information, see Connect to a cluster through kubectl.
  iii. Run the following command to query the names of the nodes in a specified node pool:

```
kubectl get node --show-labels | grep -E "${node pool ID}|${node pool ID}"
```



  iv. Run the following command to query detailed information about a specified node:

```
kubectl get node -o wide | grep -E "${node name} {node name}"
```

3.1.6.14.6. How do I select between Docker and Sandboxed-Container?

Containers and images have become the industry standards for software packaging and delivery. Kubernetes has become a standard platform for building, developing, and managing containerized cloud-native applications. An increasing number of enterprises and customers choose to deploy their applications in Container Service. Container Service supports two types of runtime: Docker and Sandboxed-Container. This topic describes the differences between these runtimes in the following aspects: implementations and limits, commonly used commands provided by Docker Engine and Containerd, and deployment architectures. This provides references for you to select between Docker and Sandboxed-Container based on your requirements.

Differences between Docker and Sandboxed-Container in terms of implementations and limits

Item	Docker	Sandboxed-Container V2	Description
Cluster type	All types	All types	N/A

Item	Docker	Sandboxed-Container V2	Description
Node type	• ECS • EBM	EBM	N/A
Node operating system	• CentOS • Alibaba Cloud Linux2	Alibaba Cloud Linux2	• You cannot deploy both Docker and Sandboxed-Container on the same node. • To deploy both Docker and Sandboxed-Container in a cluster, you can create node pools of different runtime types.
Container engine	Docker	Containerd	N/A
Monitoring	Supported	Supported	N/A
Container log collection	Supported	Sidecar: supported. Manual configuration is required.	N/A
Container stdout collection	Supported	Supported	N/A
RuntimeClass	Not supported	Supported (runV)	N/A
Pod scheduling	No configuration is required.	• For Kubernetes V1.14.x, you must add the following configuration to the nodeSelector field: `alibabacloud.com/sandboxed-container: Sandboxed-Container.runv` • For Kubernetes V1.16.x and later, no configuration is required.	N/A
HostNetwork	Supported	Not supported	N/A
exec/logs	Supported	Supported	N/A
Node data disk	N/A	Required. The data disk must be at least 200 GiB.	N/A
Network plug-in	• Flannel • Terway	• Flannel • Terway: supports only the One ENI for Multi-Pod mode.	N/A
Kube-proxy mode	• Iptables • IPVS	• Iptables • IPVS	N/A

Item	Docker	Sandboxed-Container V2	Description
Volume plug-in	CSI Plugin	CSI Plugin	N/A
Container root file system	OverlayFS	VirtioFS	N/A

Differences in the commonly used commands provided by Docker Engine and Containerd

Docker uses Docker Engine for container lifecycle management. Sandboxed-Container uses Containerd for container lifecycle management. These tools support different commands that can be used to manage images and containers. The following table lists the commonly used commands.

Command	Docker	Containerd	
	docker	crictl (recommended)	ctr
Queries containers	docker ps	crictl ps	ctr -n k8s.io c ls
Queries container details	docker inspect	crictl inspect	ctr -n k8s.io c info
Queries container logs	docker logs	crictl logs	N/A
Runs a command in a container	docker exec	crictl exec	N/A
Mounts local standard input, output, and error streams to a running container	docker attach	crictl attach	N/A
Queries resource usage statistics	docker stats	crictl stats	N/A
Creates a container	docker create	crictl create	ctr -n k8s.io c create
Starts one or more containers	docker start	crictl start	ctr -n k8s.io run
Stops one or more containers	docker stop	crictl stop	N/A
Removes one or more containers	docker rm	crictl rm	ctr -n k8s.io c del
Queries images	docker images	crictl images	ctr -n k8s.io i ls
Queries image details	docker inspect	crictl inspecti	N/A
Pulls an image	docker pull	crictl pull	ctr -n k8s.io i pull
Pushes an image	docker push	N/A	ctr -n k8s.io i push
Removes one or more images	docker rmi	crictl rmi	ctr -n k8s.io i rm
Queries pods	N/A	crictl pods	N/A
Queries pod details	N/A	crictl inspectp	N/A

Command	Docker	Containerd	
	docker	crictl (recommended)	ctr
Starts one or more pods	N/A	crictl runp	N/A
Stops one or more pods	N/A	crictl stopp	N/A

Differences between Docker and Sandboxed-Container in terms of deployment architectures

Runtime	Deployment architecture
Docker	`kubelet -> dockerd -> containerd -> containerd-shim -> runC containers`

Sandboxed-Container V1	`kubelet -> (CRI) containerd`
	`\-> containerd-shim -> runC containers`
	`\-> containerd-shim-kata-v2 -> runV sandboxed containers`

Sandboxed-Container V2	`kubelet -> (CRI) containerd`
	`\-> containerd-shim -> runC containers`
	`\-> containerd-shim-rund-v2 -> runV sandboxed containers`

3.1.6.14.7. Benefits of Sandboxed-Container

This topic describes the advantages and application scenarios of Sandboxed-Container and provides a comparison between Sandboxed-Container and open source Kata Containers. This allows you to learn more about the benefits of Sandboxed-Container.

Context

Sandboxed-Container provides an alternative to the Docker runtime environment. It supports the following features:

- Sandboxed-Container allows your applications to run in a sandboxed and lightweight virtual machine. This virtual machine is equipped with a dedicated kernel and provides better isolation and enhanced security.
- Compared with open source Kata Containers, Sandboxed-Container is optimized for storage, networking, and stability.
- You can use Sandboxed-Container to isolate untrusted applications and applications of different tenants for higher security. You can also use Sandboxed-Container to isolate applications with faults and applications with degraded performance. This minimizes the negative impact on your service. In addition, Sandboxed-Container offers the same user experience as Docker in terms of logging, monitoring, and elastic scaling.

Benefits

Compared with Docker, Sandboxed-Container has the following benefits:

- Strong isolation based on sandboxed and lightweight virtual machines.
- Compatibility with runC in terms of application management.

- High performance that corresponds to 90% performance of applications based on runC.
- The same user experience as runC in terms of logging, monitoring, and storage.
- Support for RuntimeClass.
- Easy to use with limited expertise that is required to use virtual machines.
- Higher stability than that provided by Kata Containers.

Comparison between Sandboxed-Container and Kata Containers

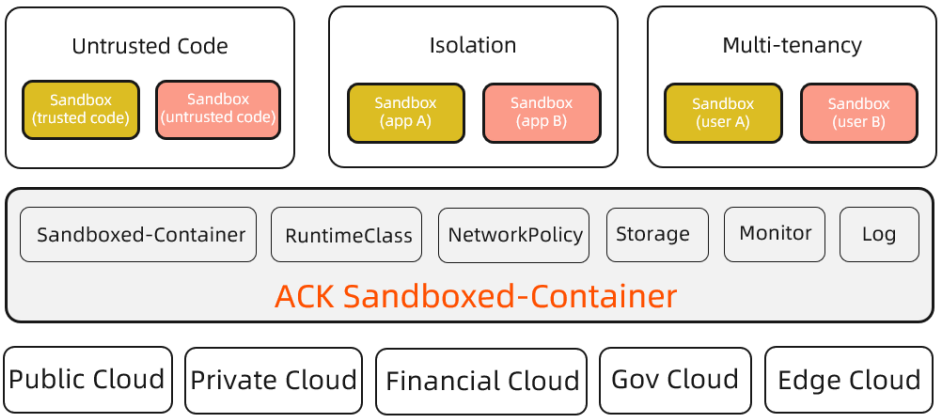
Sandboxed-Container outperforms Kata Containers in the following aspects.

Item	Category	Sandboxed-Container	Kata Containers
Sandbox startup time consumption		About 150 ms	About 500 ms
Root file system		OverlayFS over virtio-fs. Performance: ☆☆☆☆	OverlayFS over 9pfs. Performance: ☆☆
Volume	HostPath	Disks are mounted to Sandboxed-Container over 9pfs. Performance: ☆☆	Disks are mounted to Kata Containers over 9pfs. Performance: ☆☆
	EmptyDir	over VirtioFS	By default, the volume is mounted to Kata Containers over 9pfs.
	Disk	By default, cloud disks are mounted to Sandboxed-Container over virtio-fs. Performance: ☆☆☆☆	Cloud disks are mounted to Kata Containers over 9pfs. Performance: ☆☆
	NAS	By default, Apsara File Storage NAS (NAS) file systems are mounted to Sandboxed-Container over virtio-fs. Performance: ☆☆☆☆	NAS file systems are mounted to Kata Containers over 9pfs. Performance: ☆
Network plug-in		• The Terway network plug-in is used. Its network performance is 20% to 30% higher than Flannel. Terway supports features such as NetworkPolicy. This allows you to define the networking policies for pods. • Flannel	Flannel

Item	Category	Sandboxed-Container	Kata Containers
Monitoring and alerting		• Enhanced monitoring of disks and network conditions for pods that host Sandboxed-Container. • Integrated with Cloud Monitor. This facilitates cluster monitoring and alerting.	Monitoring of disks and network conditions is unavailable for pods that host Sandboxed-Container.
Stability		☆☆☆☆☆	☆☆

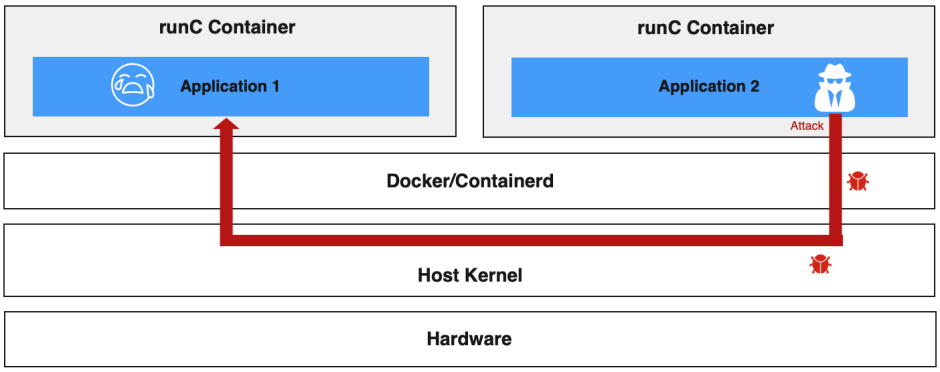
Applicable scenarios of Sandboxed-Container

This section describes the applicable scenarios of Sandboxed-Container.



- Scenario 1: Sandboxed-Container can run untrusted code and applications in isolated containers. This is not supported by containers in runC.

- Security risks of runC



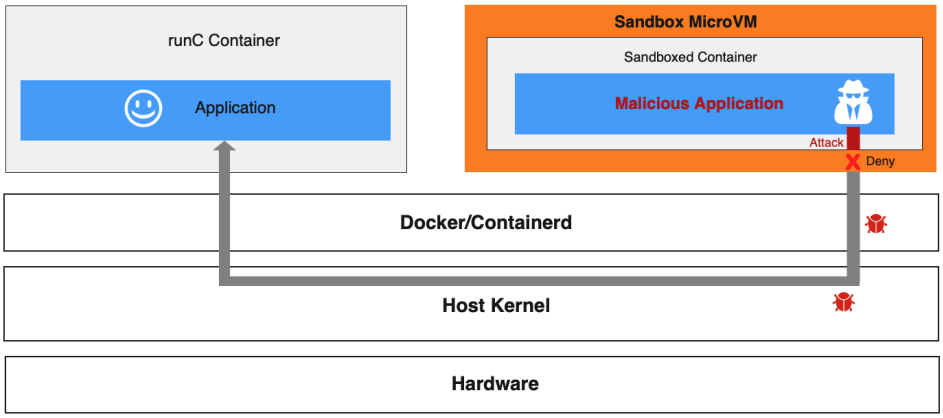
  - runC isolates containers by using namespaces and control groups (cgroups). This exposes containers to security threats.
  - All containers on a node share the host kernel. If a kernel vulnerability is exposed, malicious code may escape to the host and then infiltrate the backend network. Malicious code execution may cause privilege escalation, compromise sensitive data, and destroy system services and other applications.
  - Attackers may also exploit application vulnerabilities to infiltrate the internal network.

  You can implement the following measures to reduce security risks of containers in runC.

  - Seccomp: filters system calls.
  - SElinux: restricts the permissions of container processes, files, and users.
  - Capability: limits the capability of container processes.
  - dockerd rootless mode: forbids users to use root permissions to run the Docker daemon and containers.

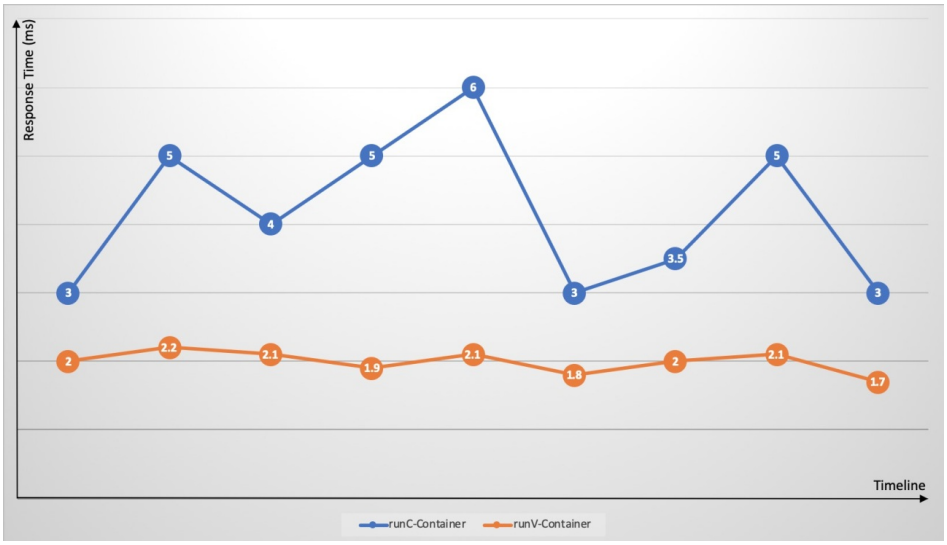
  The preceding measures can enhance the security of containers in runC and reduce attacks on the host kernel by malicious containers. However, container escapes and host kernel vulnerabilities remain unresolved.

- Sandboxed-Container prevents potential risks based on container isolation



  In a Sandboxed-Container runtime environment, applications that have potential security risks are deployed on sandboxed and lightweight virtual machines. Each of the virtual machines has a dedicated guest OS kernel. If a security vulnerability is detected on a guest OS kernel, the attack is limited to one sandbox and does not affect the host kernel or other containers. The Terway network plug-in allows you to define networking policies for pods. This enables system isolation, data isolation, and network isolation for Sandboxed-Containers.

- Scenario 2: Sandboxed-Container resolves common issues of runC containers, such as fault spreading, resource contention, and performance interference.

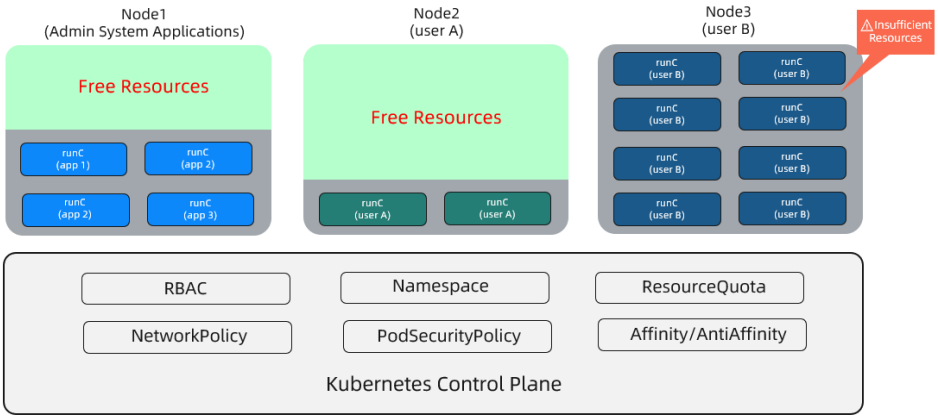


Kubernetes provides easy deployment of different containers on a single node. However, cgroups are not optimized to address resource contention. Resource-intensive applications (such as CPU-intensive, I/O-intensive applications) may compete for the same resources. This causes significant fluctuations in response time and increases the overall response time. Exceptions or faults on an application may spread to the hosting node and disrupt the running of the total cluster. For example, memory leaks and frequent core dumps of an application may overload the node, and exceptions on a container may trigger a host kernel bug that results in complete system failure. Sandboxed-Container addresses the issues that are common with runC containers by using dedicated guest OS kernels and hypervisors. The issues include failure spreading, resource contention, and performance interference.

- Scenario 3: Sandboxed-Container supports multi-tenant services.

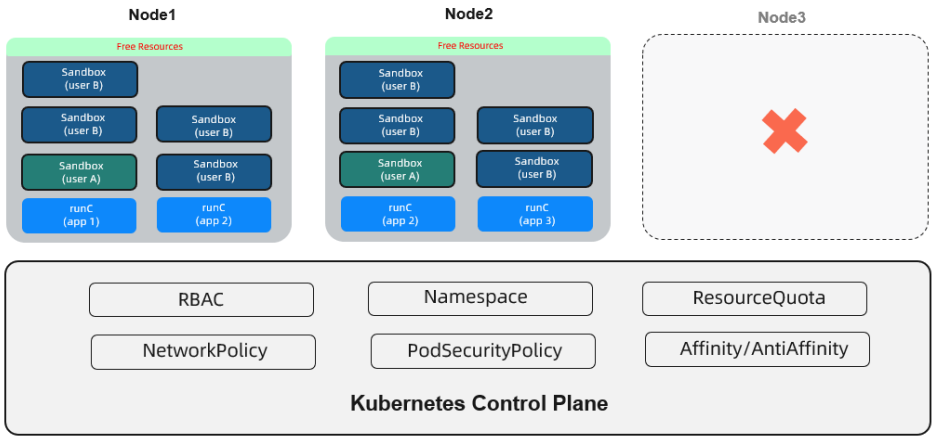
  You may need to isolate the applications of an enterprise that consists of multiple business lines or departments. For example, a financial department requires high security applications. However, other non-security-sensitive applications do not have high security requirements. Containers in runC fail to eliminate the potential risks that arise in untrusted applications. In this scenario, you can implement the following counter measures:

  - Deploy multiple independent single-tenant clusters. For example, deploy financial business and other non-security-sensitive business in different clusters.
  - Deploy a multi-tenant cluster and separate applications of different business lines by namespaces. The resource of a node is exclusive to a single business line. This solution provides data isolation for coordination with the resource quotas and network policies to implement multi-tenant isolation. Compared with multiple single-tenant clusters, this solution focuses on fewer management planes and thus reduces management costs. However, this solution cannot avoid resource waste on nodes. This issue is caused by low resource utilization of some tenants.



Sandboxed-Container allows you to isolate untrusted applications by using sandboxed virtual machines. This prevents the risks of container escapes. This also allows you to deploy different containerized applications on each node. This way, the following benefits are provided:

- Resource scheduling is simplified.
- A node is no longer exclusive to a service. This improves node resource usage and reduces resource fragments and cluster resource costs.
- Sandboxed containers use lightweight virtual machines to provide almost the same performance as containers in runC.



3.1.6.14.8. Differences between runC and runV

This topic describes the differences between runC and Sandboxed-Container (runV) in terms of their performance and pod creation methods. This allows you to better understand and utilize the benefits of sandboxed containers.

Differences between runC and runV

Item	runC	runV
Container engine	Docker and Containerd	Containerd
Node type	Elastic Compute Service (ECS) instances and ECS Bare Metal instances	EBM
Container kernel	Share the host kernel	Dedicated kernel
Container isolation	Cgroups and namespaces	Lightweight virtual machines (VMs)
Rootfs Graph Driver	OverlayFS	DeviceMapper
RootFS I/O throttling	Cgroups	DeviceMapper Block IO Limit **Note** Supported by only Sandboxed-Container V1.
NAS mounting	Not supported	Supported
Disk mounting	Not supported	Supported
Collection of container logs	Logtail directly collects container logs from the host.	logtail sidecar

Item	runC	runV
Pod Overhead	None	• Sandboxed-Container V1: For example, if you set memory: 512 Mi for a pod overhead, it indicates that 512 MiB of memory is allocated to the pod sandbox. Pod overhead refers to the amount of resources consumed by the pod sandbox. In this case, if you set a memory limit of 512 MiB for containers in the pod, the pod will request a total memory of 1,024 MiB. • Sandboxed-Container V2: The memory limit for a pod overhead is calculated based on the following formula: Memory for a pod overhead = 64 MiB + Requested memory of containers in a pod × 2%. If the result is greater than 512 MiB, the value is set to 512 Mi. If the result is smaller than 64 MiB, the value is set to 64 Mi.

Differences in pod creation between runC and runV

You can connect to clusters of Container Service by using kubectl. For more information, see Connect to a cluster through kubectl.

- Create a pod that uses runC
  i. (Optional)Use `runtimeClassName: runc` to set the container runtime to runC.

  > **Note** The preceding command is optional. runC is the default container runtime.

  ii. Run the following commands to create a pod that uses runC:

```
cat <<EOF | kubectl create -f -
apiVersion: v1
kind: Pod
metadata:
  name: busybox-runc
  labels:
    app: busybox-runc
spec:
  containers:
  - name: busybox
    image: registry.cn-hangzhou.aliyuncs.com/acs/busybox:v1.29.2
    command:
    - tail
    - -f
    - /dev/null
    resources:
      limits:
        cpu: 1000m
        memory: 512Mi
      requests:
        cpu: 1000m
        memory: 512Mi
EOF
```

- Create a pod that uses runV

i. Use `runtimeClassName: runv` to set the container runtime to runV.

ii. (Optional)Run the following command to verify that a RuntimeClass object named `runv` exists in the cluster.

```
kubectl get runtimeclass runv -o yaml
```

> **Note** A RuntimeClass object named `runv` is automatically created in a Kubernetes cluster that uses Sandboxed-Container.

iii. Run the following command to create a pod that uses runV:

```
cat <<EOF | kubectl create -f -
apiVersion: v1
kind: Pod
metadata:
  name: busybox-runv
  labels:
    app: busybox-runv
spec:
  runtimeClassName: runv
  nodeSelector:
    alibabacloud.com/container-runtime: Sandboxed-Container.runv
  containers:
  - name: busybox
    image: registry.cn-hangzhou.aliyuncs.com/acs/busybox:v1.29.2
    command:
    - tail
    - -f
    - /dev/null
    resources:
      limits:
        cpu: 1000m
        memory: 512Mi
      requests:
        cpu: 1000m
        memory: 512Mi
EOF
```

> **Notice** If the Kubernetes version is earlier than 1.16, add the following nodeSelector configuration:
>
> ```
> nodeSelector:
>     alibabacloud.com/container-runtime: Sandboxed-Container.runv
> ```

iv. Run the following command to query the pod that you have created: If the output is `runv`, it indicates that the pod is running in a sandbox.

```
kubectl get pod busybox-runv -o jsonpath={.spec.runtimeClassName}
```

v. Run the following command to log on to the pod and query its CPU and memory specifications:

```
kubectl exec -ti pod busybox-runv /bin/sh
/ # cat /proc/meminfo | head -n1
MemTotal:        1130692 kB
/ # cat /proc/cpuinfo | grep processor
processor    : 0
```

The output shows that the number of CPUs is not the same as that of the host. The total memory is the sum of pod memory and pod overhead. Be aware that the total memory is slightly smaller because the system also consumes some memory.

3.1.6.14.9. Compatibility notes

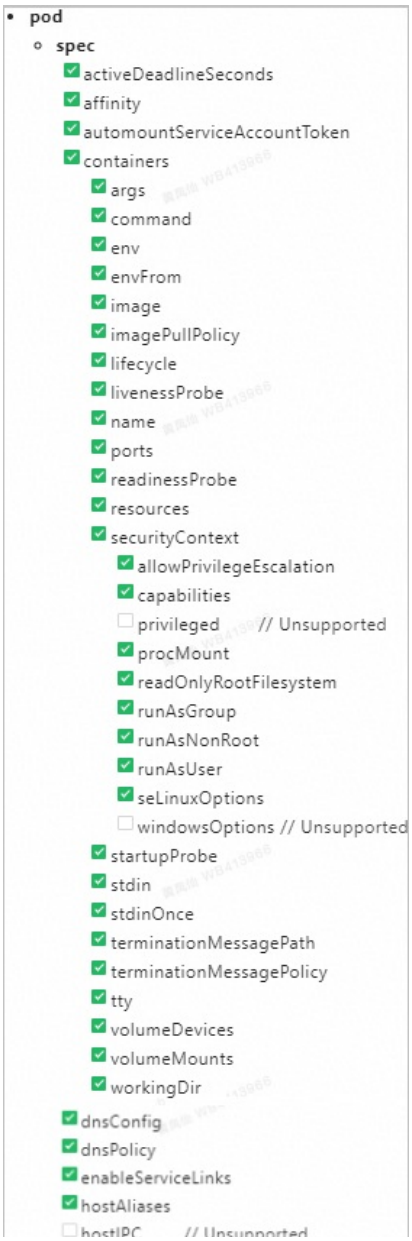
This topic describes the pod fields that are supported by Sandboxed-Container. This allows you to fully use the Sandboxed-Container runtime.

Context

Sandboxed-Container is a new runV container runtime that provides compatibility with runC in terms of pod networking, service networking (ClusterIP and NodePort), and image management. However, Sandboxed-Container does not support all pod fields. To use Sandboxed-Container, you do not need to change your development mode or image packaging method.

Supported pod fields

Sandboxed-Container supports the following pod fields that are marked by ticks:



3.1.6.15. Edge container service

3.1.6.15.1. Create an edge Kubernetes cluster

Edge Kubernetes clusters are intended for bringing cloud computing to edges (clients). Edge Kubernetes clusters can be created, managed, and maintained in the Container Service console. Container Service is a platform built on top of the edge computing infrastructure. It is also integrated with cloud computing and edge computing. This topic describes how to create an edge Kubernetes cluster.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**. On the Clusters page that appears, click **Create Kubernetes Cluster** in the upper-right corner.
3. On the Create Cluster page, click the **Managed Edge Kubernetes** tab and set the cluster parameters.

Parameter	Description
Cluster Name	Enter a name for the cluster. The name must be 1 to 63 characters in length, and can contain digits, letters, and hyphens (-). **Note** The cluster name must be unique among clusters that belong to the same Alibaba Cloud account.
Region	Select the region where you want to deploy the cluster.

Parameter	Description
VPC	You can select a virtual private cloud (VPC) from the drop-down list. ◦ If the specified VPC is already associated with a NAT gateway, the cluster uses this NAT gateway. ◦ Otherwise, the system automatically creates a NAT gateway. If you do not want the system to create a NAT gateway, clear **Configure SNAT for VPC**. **Note** If you disable the system to automatically create a NAT gateway and want the VPC to access the Internet, you must manually associate the VPC with a NAT gateway or create Source Network Address Translation (SNAT) rules for the VPC.
VSwitch	Select one or more vSwitches for the cluster. You can select up to three vSwitches that are deployed in different **zones**.
Kubernetes Version	Select a Kubernetes version.
Master Configurations	Set the Instance Type and System Disk parameters: ◦ Master Node Quantity: You can add up to three master nodes. ◦ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ◦ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data.
Worker Instance	You can select **Create Instance** or **Add Existing Instance**.

Parameter	Description
Worker Configurations	If **Worker Instance** is set to **Create Instance**, set the following parameters: ○ Instance Type: You can select one or more instance types. For more information, see the *Instance types* chapter of *ECS User Guide*. ○ Selected Types: The selected instance types are displayed. ○ Quantity: Set the number of worker nodes. By default, worker nodes are not required in edge Kubernetes clusters. Therefore, the value of this parameter is set to 0. ○ System Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** You can select **Enable Backup** to back up disk data. ○ Mount Data Disk: **SSD Disk** and **Ultra Disk** are supported. **Note** ■ You can select **Encrypt Disk** to encrypt disks. ■ You can select **Enable Backup** to back up disk data.
Operating System	The CentOS and Aliyun Linux operating systems are supported.
Network Plug-in	Edge Kubernetes clusters support only Flannel. You do not need to set this parameter.
Password	Set a password that is used to log on to the nodes. **Note** The password must be 8 to 30 characters in length, and must contain at least three of the following types of character: uppercase letters, lowercase letters, digits, and special characters.
Confirm Password	Enter the password again.
Pod CIDR Block and Service CIDR	These parameters are optional. For more information, see *Network planning* in *VPC User Guide*. **Note** These parameters are available only when you select an existing VPC.

Parameter	Description
Configure SNAT	This parameter is optional. If you clear Configure SNAT for VPC, you must create a NAT gateway or configure SNAT rules for the VPC. **Note** For edge Kubernetes clusters, we recommend that you select Configure SNAT for VPC.
Access to the Internet	Specify whether to expose the API server with an elastic IP address (EIP). The Kubernetes API server provides multiple HTTP-based RESTful APIs that can be used to create, delete, modify, query, and watch resource objects such as pods and Services. **Note** We recommend that you expose the API server with an EIP. ◦ If you select this check box, an EIP is created and associated with an internal-facing Server Load Balancer (SLB) instance. Port 6443 used by the API server is exposed on the master nodes. You can connect to and manage the cluster by using kubeconfig over the Internet. ◦ If you clear this check box, no EIP is created. You can connect to and manage the cluster only by using kubeconfig within the VPC.
SSH Logon	To enable Secure Shell (SSH) logon, you must first select Expose API Server with EIP. ◦ If you select Use SSH to Access the Cluster from the Internet, you can access the cluster through SSH. ◦ If you clear Use SSH to Access the Cluster from the Internet, you cannot access the cluster through SSH or kubectl. If you want to access an Elastic Compute Service (ECS) instance in the cluster through SSH, you must manually bind an elastic IP address (EIP) to the ECS instance and configure security group rules to open SSH port 22.
Log Service	If you enable Log Service, you can select an existing project or create a project. If you select **Enable Log Service**, the Log Service plug-in is automatically installed in the cluster. If you select **Create Ingress Dashboard**, Ingress access logs are collected and displayed on dashboards.
Deletion Protection	If you select this check box, the cluster cannot be deleted in the console or by calling API operations.
Node Protection	This check box is selected by default to prevent nodes from being accidentally deleted in the console or by calling API operations.

4. Complete advanced settings of the cluster.

Parameter	Description
IP Addresses per Node	The number of IP addresses that is assigned to a node. We recommend that you use the default value.
Custom Image	You can select a custom image. After you select a custom image, all nodes in the cluster are deployed by using this image.
Kube-proxy Mode	**iptables** and **IPVS** are supported. ◦ iptables is a tested and stable kube-proxy mode. It uses iptables rules to conduct service discovery and load balancing. The performance of this mode is limited by the size of the Kubernetes cluster. This mode is suitable for Kubernetes clusters that manage a small number of Services. ◦ IPVS is a high-performance kube-proxy mode. It uses Linux Virtual Server (LVS) to conduct service discovery and load balancing. This mode is suitable for Kubernetes clusters that manage a large number of Services. We recommend that you use this mode in scenarios where high-performance load balancing is required.
Node Port Range	Specify the value of **Node Port Range**.
Taints	Add taints to all of the worker nodes in the cluster. We recommend that you do not add additional taints in case the system components cannot be deployed in the edge Kubernetes cluster.
CPU Policy	Specify the CPU policy. Valid values: ◦ None: indicates that the default CPU affinity is used. This is the default policy. ◦ Static: allows pods with specific resource characteristics on the node to be granted with enhanced CPU affinity and exclusivity.
Cluster Domain	The default domain name of the cluster is cluster.local. You can specify a custom domain name.
Cluster CA	Specify whether to enable the cluster CA certificate.
User Data	Customize the startup behaviors of ECS instances and import data to the ECS instances. The user data can be used to perform the following operations: ◦ Run scripts during instance startup. ◦ Import user data as normal data to an ECS instance for future reference.

5. Click **Create Cluster** in the upper-right corner of the page.
6. On the **Confirm** page, after all check items are verified, select the terms of service and disclaimer and click **OK** to start the deployment.

Result

After the cluster is created, you can find the cluster on the **Clusters** page in the Container Service console.

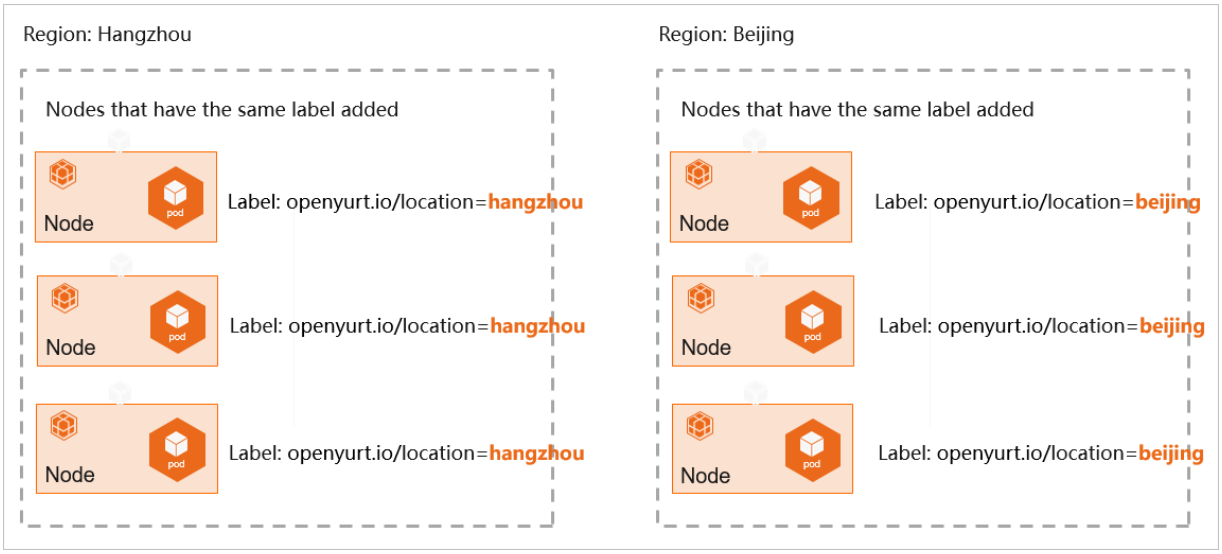
3.1.6.15.2. Edge node pools

3.1.6.15.2.1. Edge node pool overview

In edge computing scenarios, edge container service allows you to abstract nodes as edge node pools based on node attributes. This allows you to control and manage nodes in different regions in a unified manner. This topic describes edge node pools and how edge nodes are managed by edge node pools.

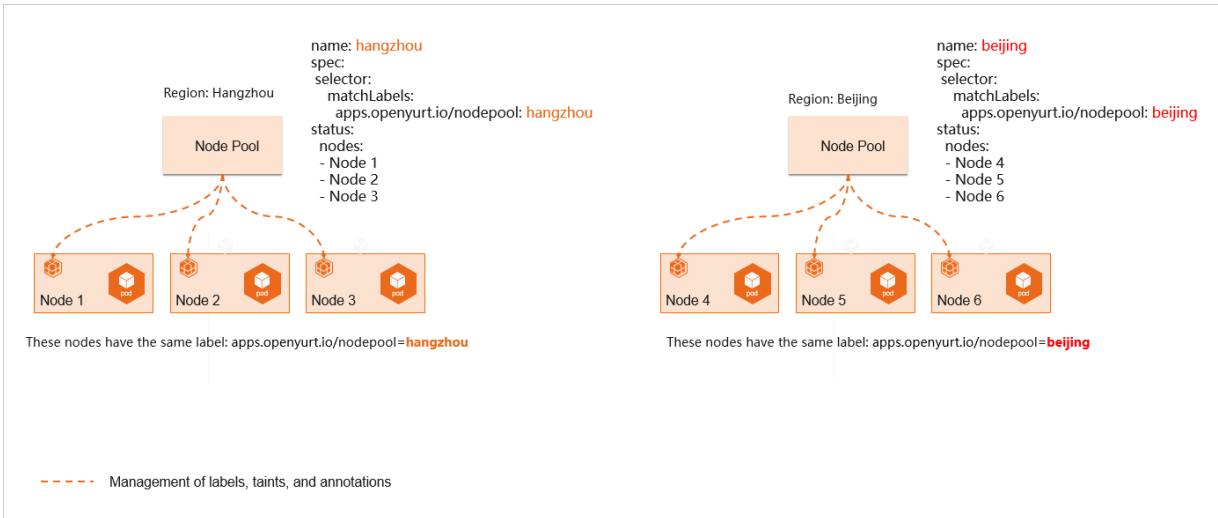
Traditional node management

In edge computing scenarios, edge nodes can be classified by different attributes such as CPU architecture, Internet service provider (ISP), and cloud service provider. Traditionally, labels are used to classify and manage nodes. However, as the numbers of nodes and labels increase, it becomes more complex to manage and maintain nodes. The following figure shows the traditional way of node management.



Edge node pools

Edge node pools allow you to classify nodes from a different dimension. You can centrally manage and maintain edge nodes that are deployed in different regions by using edge node pools, as shown in the following figure.



3.1.6.15.2.2. Create an edge node pool

An edge node pool manages a group of nodes in a cluster. For example, you can centrally manage labels and taints for the nodes in a node pool. This topic describes how to create an edge node pool in the Container Service console.

Prerequisites

- A managed edge Kubernetes cluster is created. For more information, see Create an edge Kubernetes cluster.
- The Kubernetes version of your cluster is 1.18 or later.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane of the Container Service console, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click its name or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, click **Create Edge Node Pool (Beta)** in the upper-right corner of the page.
6. In the **Create Edge Node Pool (Beta)** dialog box, set the parameters and click **Submit**.

Parameter	Description
Name	The name of the edge node pool.
Coordination Network between Cloud and Edge	By default, Basic is selected.
Maximum Nodes	The maximum number of nodes that can be added to the edge node pool.
Node Label	You can add labels to the nodes in the edge node pool.
Taints	You can add taints to the nodes in the edge node pool.

After the edge node pool is created, you can view information about the node pool in the node pool list.

3.1.6.15.2.3. Add nodes to an edge node pool

You can add worker nodes to an edge node pool that you created. Make sure that these worker nodes can communicate with the API server. This topic describes how to add nodes to an edge node pool.

Prerequisites

- An edge node pool is created. For more information, see Create an edge node pool.
- The Kubernetes version of your cluster is 1.18 or later.

> **Notice** You can add only nodes that run CentOS 7.4, CentOS 7.6, or Ubuntu 18.04.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane of the Container Service console, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage. Then, click the name of the cluster or click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Node Pools**.
5. On the **Node Pools** page, find the edge node pool to which you want to add nodes and click **Add Existing Node** in the **Actions** column.

Perform the same steps as you do when you add nodes to an edge Kubernetes cluster. For more information, see Add nodes to an edge Kubernetes cluster.

After you add nodes to the edge node pool, you can click **Details** in the **Actions** column to view the nodes that you added.

3.1.6.15.3. Edge nodes

3.1.6.15.3.1. Add nodes to an edge Kubernetes cluster

You can add worker nodes to an edge Kubernetes cluster in the Container Service console. However, you must make sure that the added nodes can communicate with the Kubernetes API server of the cluster. This topic describes how to add nodes to an edge Kubernetes cluster.

Prerequisites

Create an edge Kubernetes cluster

Context

By default, a cluster can contain at most 50 nodes.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, click the name of the cluster that you want to manage.
4. In the left-side navigation pane of the cluster details page, choose **Nodes > Nodes**. On the page that appears, click **Add Existing Node** in the upper-right corner.
5. On the **Select Existing ECS Instance** wizard page, click **Next Step**.

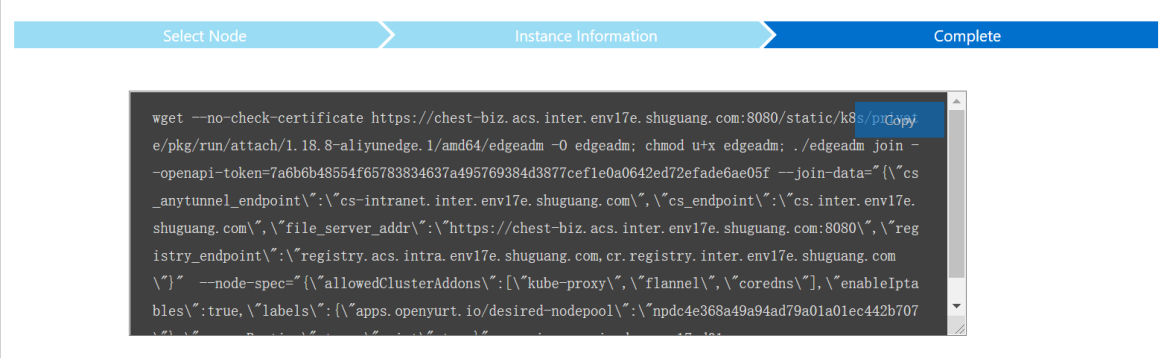
   You can only manually add nodes to edge Kubernetes clusters.
6. On the **Specify Instance Information** wizard page, set the parameters and click **Next Step**.

Parameter	Description	Default
flannelIface	The name of the network interface controller (NIC) that is used by the Flannel plug-in.	The name of the NIC that is specified in the default route entry of the node.
enableIptables	Specifies whether to enable iptables.	false
quiet	Specifies whether to answer all questions with yes when you add nodes.	false
manageRuntime	Specifies whether to use edgeadm to install and manage the runtime.	false

Parameter	Description	Default
nodeNameOverride	The name of the node.	◦ "". This is the default value. This value specifies that the hostname is used as the node name. ◦ "*". This value specifies that a random string that contains six characters is used as the node name. ◦ "*."*.XXX". This value specifies that a random string that is followed by a suffix is used as the node name. The random string contains six characters.
allowedClusterAddons	The list of add-ons to be installed. By default, this parameter is empty. This indicates that no add-on is installed. For a standard edge node, set this parameter to ["kube-proxy","flannel","coredns"].	[]
gpuVersion	Specifies whether the node to be added is a GPU-accelerated node. By default, this parameter is empty. Supported GPU models are Nvidia_Tesla_T4, Nvidia_Tesla_P4, and Nvidia_Tesla_P100.	"". This is the default value. This value specifies that the node to be added is not a GPU-accelerated node.
inDedicatedNetwork	Specifies whether an Express Connect circuit is used to connect to the managed edge Kubernetes cluster.	false
labels	Specifies the labels to be added to the node.	{}
annotations	Specifies the annotations to be added to the node.	{}

Parameter	Description	Default
nodeIface	This parameter specifies the following information: ○ Specifies the node IP address that kubelet retrieves from the specified network interface. If you do not specify this parameter, kubelet attempts to retrieve the node IP address in the following order: ■ Searches */etc/hosts* for the node whose name is the same as the specified hostname. ■ Finds the IP address of the network interface that is specified in the default route entry of the node. ○ Specifies the name of the NIC that is used by the Flannel plug-in. In this case, this parameter is equivalent to the flannelIface parameter. This parameter will soon replace the flannelIface parameter.	""

7. On the **Complete** wizard page, copy the script to the node that you want to add to the edge Kubernetes cluster and click **Done**.



8. Log on to the edge node and execute the script. This way, the node is added to the edge Kubernetes cluster.

3.1.6.15.3.2. Configure node autonomy

If an edge node is autonomous, applications run as expected on the edge node even if the edge node is disconnected from the cloud. This ensures that applications are not removed or migrated to other edge nodes in the case of network errors. This topic describes how to set the autonomy attribute for edge nodes.

Prerequisites

- Create an edge Kubernetes cluster
- Add nodes to an edge Kubernetes cluster

Context

You can enable or disable node autonomy in the Container Service console.

- If an autonomous edge node is disconnected from the cloud, Container Service does not migrate applications on this node to other nodes, and the applications are automatically restored. Node autonomy is applicable to edge computing scenarios where the network connection is weak.
- If a non-autonomous edge node is disconnected from the cloud, the node fails to send heartbeats to the nodes in the cloud. As a result, the state of the node is changed to **Not Ready** and the applications on the node are removed or migrated to other nodes after a specific time period.

Procedure

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Nodes > Nodes**.
5. On the **Nodes** page, find the node that you want to manage and choose **More > Node Autonomy Setting** in the **Actions** column.

   > **Note** The **Node Autonomy Setting** option is available only to edge nodes.

6. In the **Node Autonomy Setting** dialog box, click **OK**.

   > **Note** By default, edge nodes are not autonomous when they are added to the cluster. You can follow the preceding steps to enable or disable node autonomy.

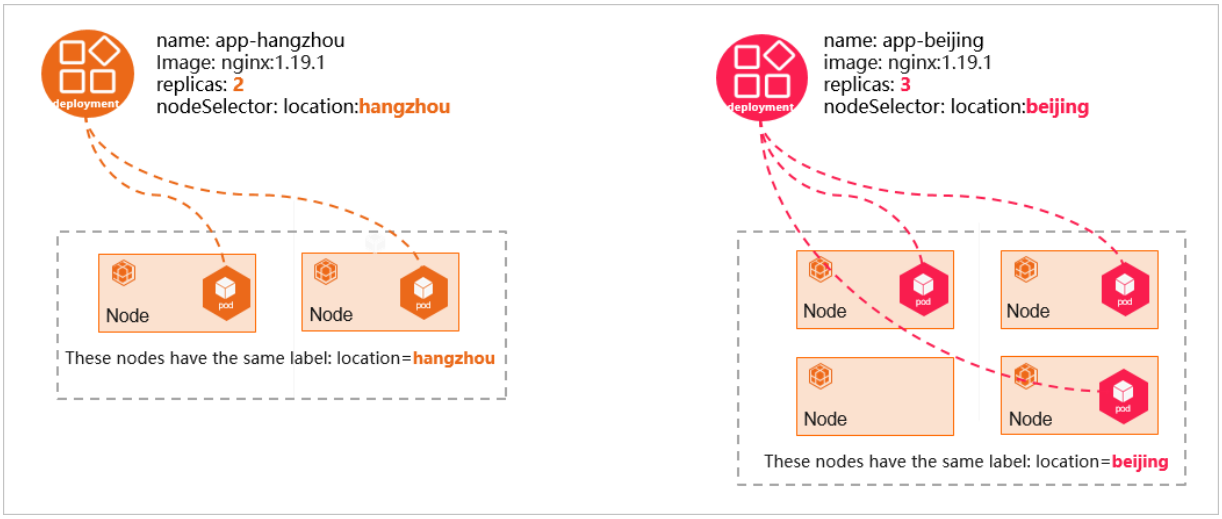
3.1.6.15.4. Cell-based management at the edge

3.1.6.15.4.1. Use the UnitedDeployment controller to deploy applications

In edge computing scenarios, you can use the UnitedDeployment controller to deploy applications to different node pools. This way, you can centrally manage the number of pods and the image version of containers by using node pools. This topic describes how to use the UnitedDeployment controller to deploy applications.

Context

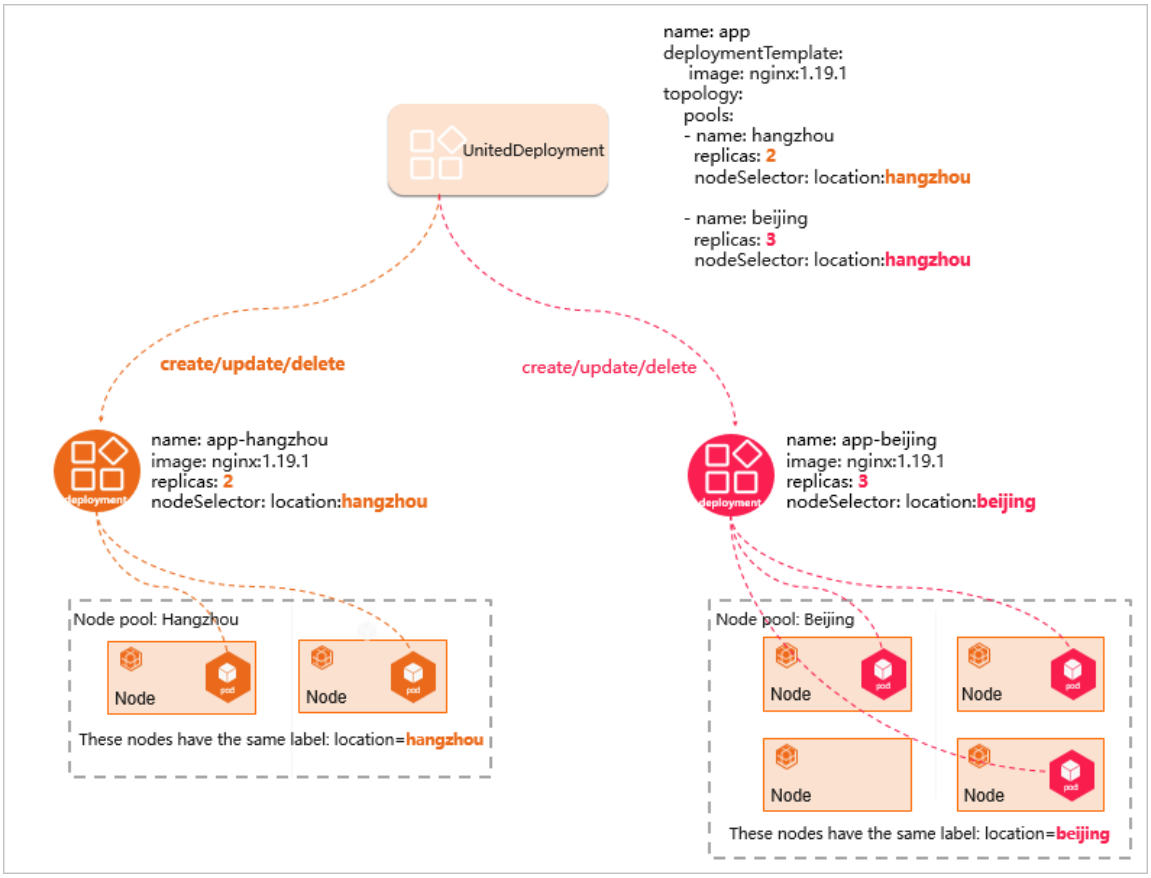
In edge computing scenarios, computing nodes may be deployed across regions, and an application may run on nodes in different regions. A Deployment is used as an example in this topic. Traditionally, you add the same label to the nodes that are deployed in the same region and create multiple Deployments. These Deployments are deployed to nodes in different regions by matching node selectors.



Application management and maintenance become more complex with the increasing number of regions and differentiated requirements for applications in different regions. The following list describes the main challenges:

- When a new image version is released, you must modify the image version for each Deployment.
- You must customize naming conventions to identify Deployments that belong to the same application.
- Deployments that belong to the same application are configured in the same way, except for the name, node selector, and number of replicated pods.

The UnitedDeployment controller is a feature provided for dedicated edge Kubernetes clusters. This feature allows you to centrally manage Deployments from a different dimension. For example, you can create, update, and delete multiple Deployments at a time.



The UnitedDeployment controller provides a template to define applications. This template allows you to deploy workloads in different regions and define the nodes in each region as a node pool. The UnitedDeployment controller supports two types of workload: StatefulSet and Deployment. The UnitedDeployment controller creates Deployments or StatefulSets based on the configurations of node pools. You can specify the number of replicated pods for each type of workload. UnitedDeployment enables automatic management and maintenance of multiple Deployments or StatefulSets within individual node pools. In addition, you can create differentiated configurations for these Deployments or StatefulSets, such as the name, node selector, and number of replicated pods.

Create a UnitedDeployment

Create a UnitedDeployment to deploy Deployments.

The following YAML template is an example:

```
apiVersion: apps.openyurt.io/v1alpha1
kind: UnitedDeployment
metadata:
  name: example
  namespace: default
spec:
  revisionHistoryLimit: 5
  selector:
    matchLabels:
      app: example
  workloadTemplate:
    deploymentTemplate:
      metadata:
        creationTimestamp: null
        labels:
          app: example
      spec:
        selector:
          matchLabels:
            app: example
        template:
          metadata:
            creationTimestamp: null
            labels:
              app: example
          spec:
            containers:
            - image: nginx:1.19.3
              imagePullPolicy: Always
              name: nginx
            dnsPolicy: ClusterFirst
            restartPolicy: Always
  topology:
    pools:
    - name: cloud
      nodeSelectorTerm:
        matchExpressions:
        - key: apps.openyurt.io/nodepool
          operator: In
          values:
          - cloud
      replicas: 2
    - name: edge
      nodeSelectorTerm:
        matchExpressions:
        - key: apps.openyurt.io/nodepool
          operator: In
          values:
          - edge
      replicas: 2
      tolerations:
      - effect: NoSchedule
        key: apps.openyurt.io/taints
        operator: Exists
```

The following table describes the fields in the YAML template.

Field	Description
spec.workloadTemplate	The workload template. Valid values: `deploymentTemplate and statefulSetTemplate`.
spec.topology.pools	Configurations of multiple node pools.
spec.topology.pools[*].name	The name of the node pool.
spec.topology.pools[*].nodeSelectorTerm	Specifies node affinity for the node pool. Set the key to `apps.openyurt.io/nodepool` and the value to the name of the node pool.
spec.topology.pools[*].tolerations	Sets tolerations for the node pool.
spec.topology.pools[*].replicas	The number of pods in each node pool.

Use the UnitedDeployment controller to manage pods

- Upgrade pods: You can modify the `spec.template.workloadTemplate.deploymentTemplate` field to trigger pod upgrades. The UnitedDeployment controller updates the workload template for all node pools. Then, the node pool controller upgrades the pods in the node pools.
- Scale the number of replicated pods for multiple node pools: You can modify the `spec.topology.pools` field to change the number of replicated pods for multiple node pools. Then, the replicated pods in the node pools are scaled based on the configuration.

3.1.6.15.4.2. Configure a Service topology

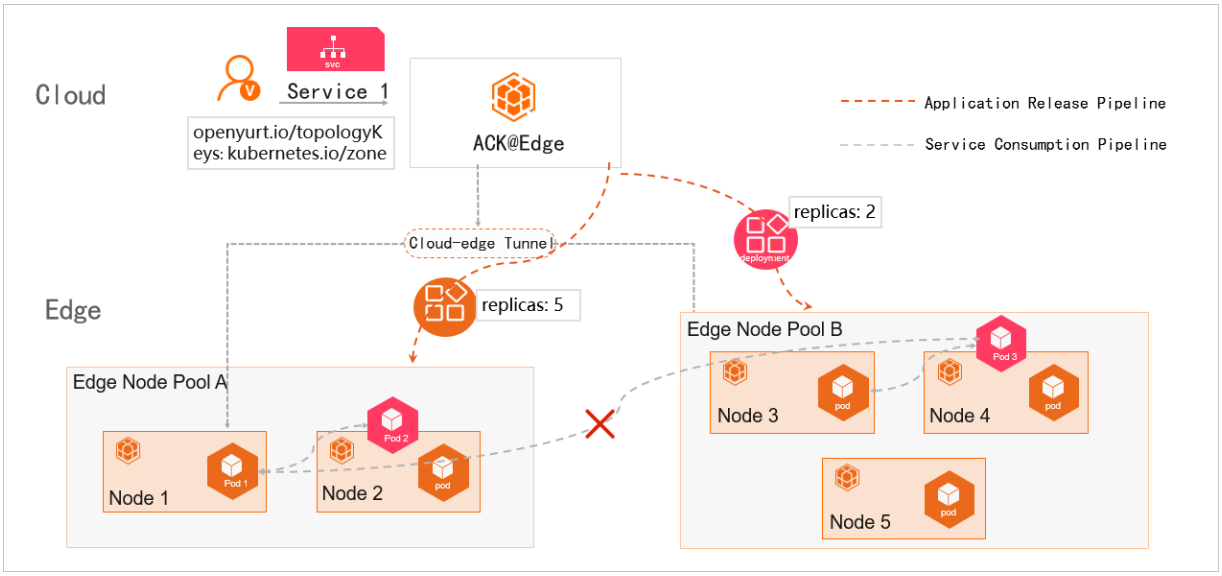
The backend endpoints of Kubernetes-native Services are randomly distributed across nodes. Consequently, when Service requests are distributed to nodes across node groups, these requests may fail to reach the nodes or may not be answered promptly. You can configure a Service topology to expose an application on an edge node only to the current node or nodes in the same edge node pool. This topic describes how a Service topology works and how to configure a Service topology.

Context

In edge computing, edge nodes are classified into groups by zone, region, and other logical attributes, such as CPU architecture, Internet service provider (ISP), or cloud service provider. Nodes in different groups are isolated from each other in one way or another. For example, these nodes may not be able to connect to each other, may not share the same resources, may have heterogeneous resources, and may run applications that are independent of each other.

How a Service topology works

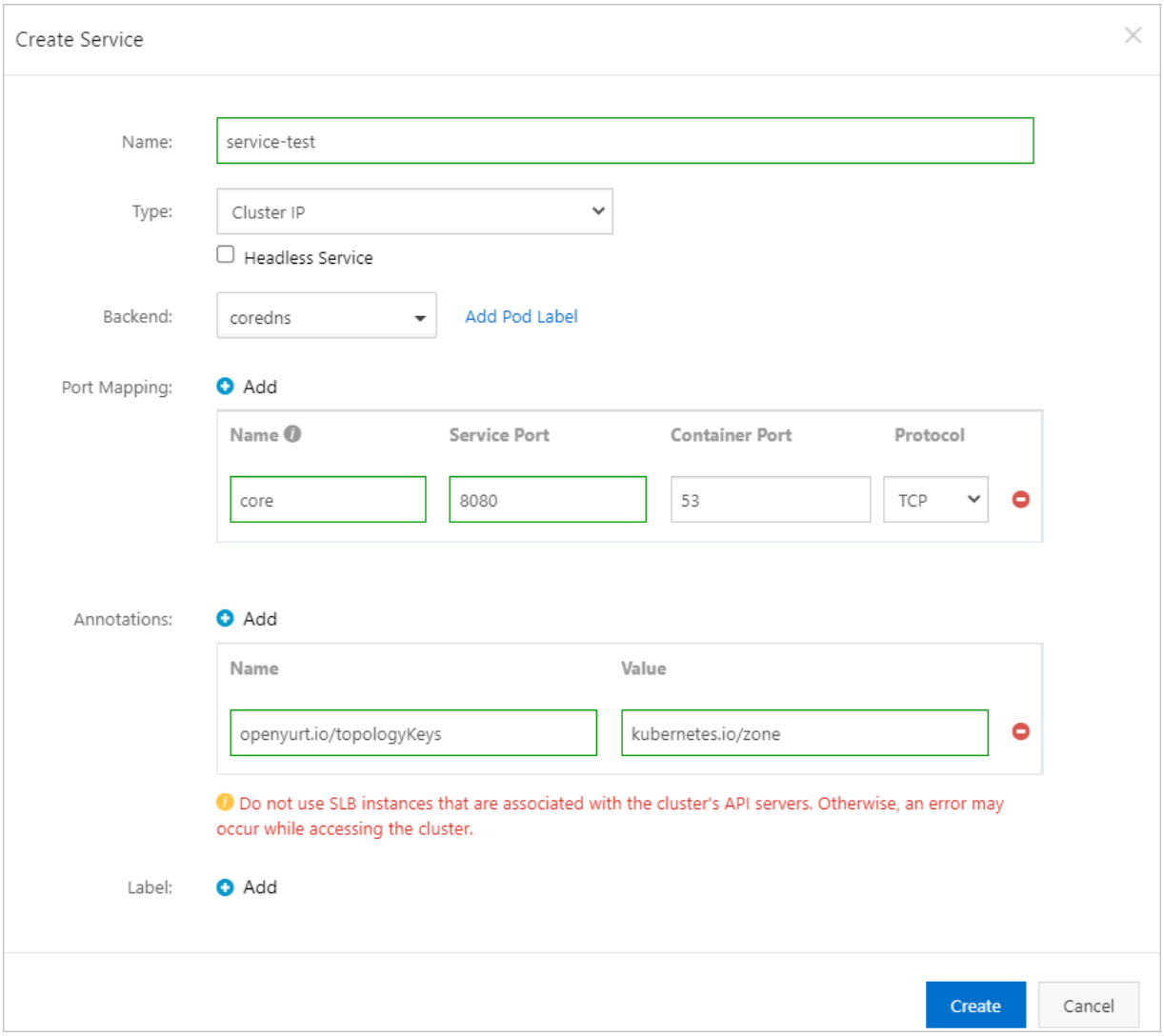
To solve the preceding issues, dedicated edge Kubernetes clusters provide a feature to manage the topology of endpoints of Kubernetes-native Services. You can configure a Service topology to specify how endpoints of a Service are accessed. For example, you can configure a Service topology to expose an application on an edge node only to the current node or nodes in the same edge node pool. The following figure shows how a Service topology works.



- Service 1 is associated with Pod 2 and Pod 3. `annotation:"openyurt.io/topologyKeys: kubernetes.io/zone"` specifies the node pool that is allowed to access Service 1.
- Pod 2 is deployed on Node 2 and Pod 3 is deployed on Node 3. Node 2 belongs to Node Pool A and Node 3 belongs to Node Pool B.
- Pod 3 and Pod 1 do not belong to the same node pool. As a result, when Pod 1 accesses Service 1, the traffic is forwarded only to Pod 2. The traffic is not forwarded to Pod 3.

Method 1: Configure a Service topology in the console

To create a Service that can be accessed only by the node pool where the Service is deployed, you only need to add an annotation to the Service. For example, you can set **Name** to `openyurt.io/topologyKeys` and **Value** to `kubernetes.io/zone`. For more information about how to create a Service, see Create a Service.



Method 2: Configure a Service topology by using a CLI

You can use a CLI to configure a Service topology in the following ways:

- Create a Service that uses the topological domain of a specific node pool. The following code block is an example of the YAML template:

```
apiVersion: v1
kind: Service
metadata:
  annotations:
    openyurt.io/topologyKeys: kubernetes.io/zone
  name: my-service-nodepool
  namespace: default
spec:
  ports:
  - port: 80
    protocol: TCP
    targetPort: 8080
  selector:
    app: nginx
  sessionAffinity: None
  type: ClusterIP
```

- Run the following command to configure the Service topology. The Service uses the topological domain of the specified node pool.

```
kubectl annotate service xxx openyurt.io/topologyKeys='kubernetes.io/zone'
```

Annotations

You can add annotations to a Kubernetes-native Service to configure a Service topology. The annotations are described in the following table.

annotation Key	annotation Value	Description
openyurt.io/topologyKeys	kubernetes.io/hostname	Specifies that the Service can be accessed only by the node where the Service is deployed.
openyurt.io/topologyKeys	kubernetes.io/zone	Specifies that the Service can be accessed only by the nodes in the node pool where the Service is deployed.
None	None	Specifies that access to the Service is unlimited.

3.1.6.15.5. Cloud-edge tunneling

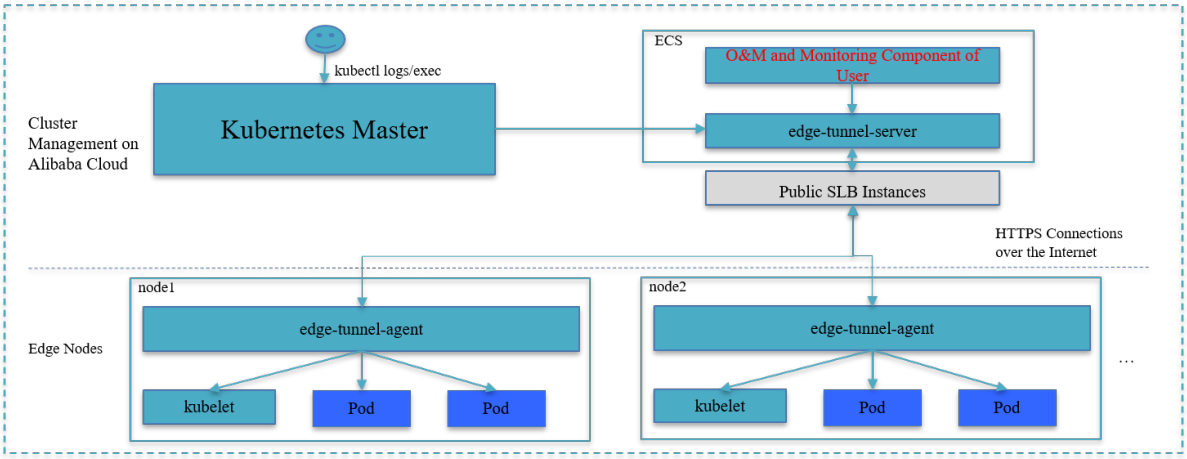
By default, Container Service deploys the **edge-tunnel-server or edge-tunnel-agent** component after you create a cluster. This improves user experience. These components are used to establish tunnels from the cloud to the edge. After the tunnels are established, you can access edge nodes from the cloud. This topic describes the tunneling components provided by managed edge Kubernetes clusters and the features of these components.

Background information

- In a Kubernetes cluster, the controller components in the cloud must run commands in kubelet to manage and maintain edge nodes. The monitoring components must retrieve monitoring data of edge nodes from the cloud. If the edge nodes of a managed edge Kubernetes cluster are deployed in an internal network, you cannot directly access the edge nodes from the cloud.
- The **edge-tunnel-server** component is deployed as a Deployment on nodes in the cloud. The **edge-tunnel-agent** component is deployed as a DaemonSet on each edge node.

Introduction

- edge-tunnel-server is automatically installed on master nodes when you create an edge Kubernetes cluster.
- To establish secure and encrypted tunnels over the Internet, the system creates a Server Load Balancer (SLB) instance for the Service that is created by **edge-tunnel-server**. The **edge-tunnel-agent** component on each edge node establishes a tunnel to the cloud through the SLB instance. The following figure shows how cloud-edge tunneling works.



Note

- When edge nodes are disconnected from the cloud or the network connection is weak, the tunnels may fail to work as normal.
- If you delete or stop the SLB instance through which the tunnels are established, the tunnels cannot work as normal.

3.1.6.16. Use the Kubernetes event center

The event center feature allows you to log Kubernetes events, query events, and configure alerting. This topic describes how to enable and view the Kubernetes event center.

Enable the Kubernetes event center

Enable the Kubernetes event center when a Kubernetes cluster is created

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**. On the page that appears, click **Create Kubernetes Cluster** in the upper-right corner.
3. On the **Create Cluster** page, select **Install node-problem-detector and Create Event Center** in the **Log Service** field. For more information about other parameters, see Create a Kubernetes cluster. Then, click **Create Cluster**.
4. On the **Confirm** dialog box, after all check items are verified, select the terms of service and disclaimer and click **OK** to create the cluster.

Enable the Kubernetes event center in App Catalog

1. Log on to the Container Service console.
2. In the left-side navigation pane, choose **Marketplace > App Catalog**.
3. On the **App Catalog** page, find and click ack-node-problem-detector.
4. On the **App Catalog - ack-node-problem-detector** page, click the **Parameters** tab and set `eventer.sinks.sls.enabled` to `true`.

```
sinks:
  sls:
    enabled: true
    # If you want the monitoring results to be notified by sls, set enabled to true.
    topic: ""
    project:  "k8s-log-cc7640381ac7f4a99971f55a2744a2748"
    # You can view the project information by logging in to the
    # SLS console. Please fill in the name of the project here.
    # eg: your project name is k8s-log-cc18a5f3443dhdss22654da,
    # then you can fill k8s-log-cc18a5f3443dhdss22654da to project label.
    logstore: "k8s-event"
    # You can view the project information by logging in to the
    # SLS console. Please fill the logstore address in here.
```

5. On the **App Catalog - ack-node-problem-detector** page, click the **Description** tab.
6. In the **Deploy** pane, select the cluster and click **Create**.

Access the Kubernetes event center

Access the Kubernetes event center in the Container Service console

1. Log on to the Container Service console.
2. In the left-side navigation pane, click **Clusters**.
3. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
4. In the left-side navigation pane of the details page, choose **Operations > Event Center**.
5. Access the Kubernetes event center on the Event Center page

   > **Note** If the Event Center page does not appear, check whether the Kubernetes event center is enabled.

   - On the **Event Center** page, click the **Events Overview** tab to go to the overview page of the Kubernetes event center.
   - On the **Event Center** page, click the **Cluster Events Query** tab to customize query conditions.
   - On the **Event Center** page, click the **Pod Events** tab to query events of pods.

Access the Kubernetes event center by using Log Service

1. View the cluster ID.
   i. Log on to the Container Service console.
   ii. In the left-side navigation pane, click **Clusters**.
   iii. On the **Clusters** page, find the cluster that you want to manage and click **Details** in the **Actions** column.
   iv. Click the **Basic Information** tab to view the cluster ID.
2. Log on to the Log Service console.
3. In the search box of the Projects section, enter and click k8s-log-<CLUSTER_ID>.

   > **Note** Replace the CLUSTER_ID with the cluster ID that you obtained in Step .

4. In the Logstores section, choose **K8s-event > Visual Dashboards > Kubernetes Event Center V1.2**.
5. On the **Kubernetes Event Center V1.2** page, you can view the trends of WARNING events and ERROR events.

4.Auto Scaling (ESS)

4.1. User Guide

4.1.1. What is Auto Scaling?

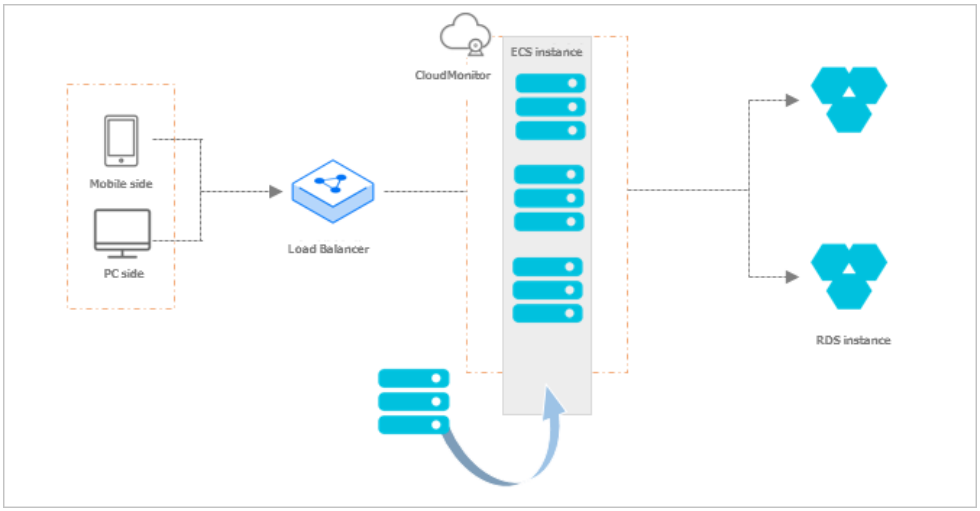
Auto Scaling automatically adjusts your elastic computing resources based on your business requirements and policies that you define.

When demand for services spikes, Auto Scaling automatically scales out Elastic Compute Service (ECS) instances based on your configurations to maintain sufficient computing resources. When demand for services drops, Auto Scaling automatically scales in ECS instances to save costs.

Auto Scaling provides the following features:

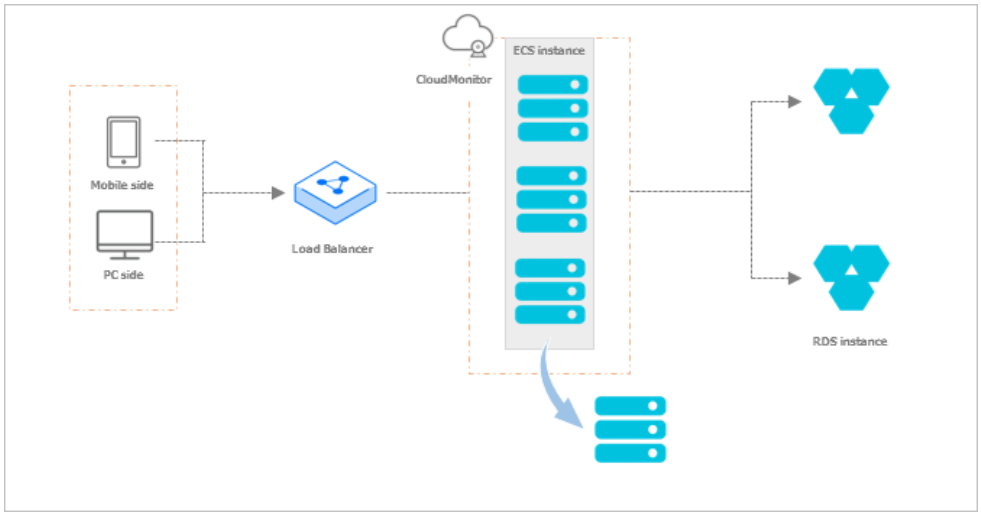
- Scale-out

  When demand for services suddenly grows, Auto Scaling automatically scales out the underlying resources. This ensures that resources are not overloaded and maintains the responsiveness of your servers. For example, if the vCPU utilization of ECS instances exceeds 80%, Auto Scaling scales out ECS resources based on your configurations. During the scale-out event, Auto Scaling automatically creates ECS instances, adds the ECS instances to a scaling group, and then adds the new instances to the backend server groups of the associated Server Load Balancer (SLB) instances and the whitelists of the associated ApsaraDB RDS instances. The following figure shows how a scale-out event is implemented.



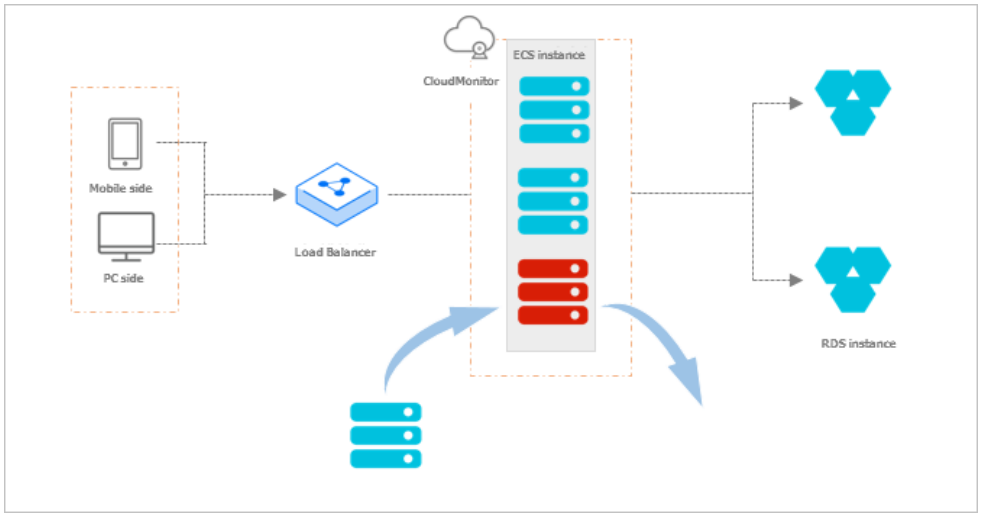
- Scale-in

  When demand for services drops, Auto Scaling automatically releases underlying resources to prevent waste of resources and reduce costs. For example, if the vCPU utilization of ECS instances in a scaling group falls below 30%, Auto Scaling automatically scales in ECS instances based on your configurations. During the scale-in event, Auto Scaling removes ECS instances from the scaling group and also from the backend server groups of the associated SLB instances and the whitelists of the associated ApsaraDB RDS instances. The following figure shows how a scale-in event is implemented.



- Elastic recovery

  If ECS instances in a scaling group are not in the Running state, Auto Scaling considers the instances to be unhealthy. If an ECS instance is considered unhealthy, Auto Scaling automatically releases the instance and creates a new one. This process is called elastic recovery. Elastic recovery ensures that the number of healthy ECS instances in a scaling group does not fall below the minimum number of ECS instances that you specified for the scaling group. The following figure shows how an elastic recovery event is implemented.



4.1.2. Notes

4.1.2.1. Precautions

This topic describes the precautions when you use Auto Scaling (ESS).

Scaling rules

ESS uses scaling rules to scale ECS instances in a scaling group based on the minimum and maximum numbers of ECS instances specified for the scaling group. Assume that a scaling group can contain up to 45 ECS instances. If you configure a scaling rule to increase the number of ECS instances in the scaling group to 50, ESS only increases the number of ECS instances to 45 at most.

Scaling activities

- Only one scaling activity can be executed at a time in a scaling group.
- An ongoing scaling activity cannot be terminated. For example, if a scaling activity is being executed to create

20 ECS instances but only five have been created, you cannot forcibly terminate the scaling activity.

- If some ECS instances fail to be added to a scaling group during a scaling activity, ESS considers that the scaling activity is complete without trying to add the failed instances to the scaling group. ESS rolls back the ECS instances that fails to be added but not the scaling activity. For example, if ESS has created 20 ECS instances for a scaling group, and 19 of the instances are added to SLB instances, only the one ECS instance that failed to be added is automatically released.

Cooldown period

- During the cooldown period, if you manually execute a scaling task, such as a scaling rule or scheduled task, the task is immediately executed without waiting for the cooldown period to expire.
- The cooldown period starts after the last ECS instance is added to or removed from a scaling group during a scaling activity.

4.1.2.2. Manual operations

If you perform a manual operation when an Auto Scaling operation is in progress, the manual operation takes precedence.

Auto Scaling supports manual operations, such as the deletion of automatically created Elastic Compute Service (ECS) instances from the ECS console. The following table describes how Auto Scaling processes manual operations.

Resource	Manual operation type	Processing method
ECS	A user deletes an ECS instance from a scaling group by using the ECS console or calling API operations.	Auto Scaling performs health checks to determine whether the ECS instance is unhealthy. If the instance is unhealthy, Auto Scaling removes it from the scaling group. The internal IP address of the ECS instance is not automatically deleted from the whitelist of the associated ApsaraDB RDS instance. If the total number of ECS instances in the scaling group falls below the lower limit after the ECS instance is removed, the scaling group automatically creates an ECS instance to ensure that the number of instances is at least equal to the lower limit.
ECS	A user revokes the ECS API permissions granted to Auto Scaling.	Auto Scaling rejects all scaling activity requests.
Server Load Balancer (SLB)	A user manually removes an ECS instance from the associated SLB instance by using the SLB console or calling API operations.	Auto Scaling does not automatically detect this operation or handle this exception. The ECS instance remains in the scaling group. When a scale-in activity is triggered, Auto Scaling releases the ECS instance if the instance meets the conditions specified in the removal policy.
SLB	A user manually deletes an SLB instance or disables the health check feature for an SLB instance by using the SLB console or calling API operations.	Auto Scaling does not add ECS instances to scaling groups that are associated with this SLB instance. Auto Scaling removes ECS instances from the scaling groups if a scaling task triggers a scale-in rule or the ECS instances are considered unhealthy after a health check is performed.
SLB	An SLB instance is unavailable due to a system error.	All scaling activities fail, except for instance removal tasks that are manually executed.

Resource	Manual operation type	Processing method
SLB	A user revokes the SLB API permissions granted to Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling groups that are associated with SLB instances.
ApsaraDB RDS	A user manually removes the IP address of an ECS instance from the whitelist of the associated ApsaraDB RDS instance by using the ApsaraDB RDS console or calling API operations.	Auto Scaling does not automatically detect this operation or handle this exception. The ECS instance remains in the scaling group. When a scale-in activity is triggered, Auto Scaling releases the ECS instance if the instance meets the conditions specified in the removal policy.
ApsaraDB RDS	A user manually deletes an ApsaraDB RDS instance by using the ApsaraDB RDS console or calling API operations.	Auto Scaling does not add ECS instances that are associated with the ApsaraDB RDS instance to scaling groups. Auto Scaling removes ECS instances from the scaling groups if a scaling task triggers a scale-in rule or the ECS instances are considered as unhealthy after a health check is performed.
ApsaraDB RDS	An ApsaraDB RDS instance is unavailable due to a system error.	All scaling activities fail, except for instance removal tasks that are manually executed.
ApsaraDB RDS	A user revokes the ApsaraDB RDS API permissions granted to Auto Scaling.	Auto Scaling rejects all scaling activity requests for the scaling groups that are associated with ApsaraDB RDS instances.

4.1.2.3. Limits

This topic describes the limits of Auto Scaling.

- Auto Scaling can automatically scale the number of Elastic Compute Service (ECS) instances in a scaling group, but cannot automatically upgrade or downgrade configurations of the ECS instances, such as vCPUs, memory, and bandwidth.
- Applications deployed on the ECS instances in a scaling group must be stateless and horizontally scalable.
- ECS instances in a scaling group can be automatically released. We recommend that you do not store information such as sessions, application data, or logs on the ECS instances in a scaling group. If you need to store data of the applications deployed on the ECS instances, store status information such as sessions on the independent ECS instances, store application data in ApsaraDB RDS, and store logs in Log Service. For more information, see *What is ApsaraDB RDS?* in *ApsaraDB RDS Product Introduction* and *What is Log Service?* in *Log Service Product Introduction*.
- The following table describes the quantity limits that are applied to a scaling group.

Item	Quota
Scaling configuration	You can create a maximum of 10 scaling configurations in a scaling group.
Scaling rule	You can create a maximum of 50 scaling rules in a scaling group.
ECS instance	You can add a maximum of 1,000 ECS instances to a scaling group.

4.1.2.4. Scaling group status

This topic describes the states of a scaling group in the console and in an API operation.

State in the console	State in an API operation
Creating	Inactive
Created	Inactive
Enabling	Inactive
Enabled	Active
Disabling	Inactive
Disabled	Inactive
Deleting	Deleting

4.1.2.5. Scaling processes

Before you use Auto Scaling, you must understand the processes related to scaling activities.

Automatic scaling of a scaling group

- Automatic scale-out
  i. Check the health status and boundary conditions of the scaling group.
  ii. Assign the activity ID and execute the scaling activity.
  iii. Create ECS instances.
  iv. Modify Total Capacity.
  v. Assign IP addresses to the created ECS instances.
  vi. Add the ECS instances to the whitelist of the associated ApsaraDB RDS instance.
  vii. Start the ECS instances.
  viii. Associate the ECS instances with an SLB instance and set the weight to the SLB weight value that is specified when the scaling configuration is created.
  ix. The cooldown period starts after the scaling activity is complete.
- Automatic scale-in
  i. Check the health status and boundary conditions of the scaling group.
  ii. Assign the activity ID and execute the scaling activity.
  iii. Remove ECS instances from the associated SLB instance.
  iv. Stop the ECS instances.
  v. Remove the ECS instances from the whitelist of the associated ApsaraDB RDS instance.
  vi. Release the ECS instances.
  vii. Modify Total Capacity.
  viii. The cooldown period starts after the scaling activity is complete.

Manually add or remove existing ECS instances

- Manually add instances
  i. Check the health status and boundary conditions of the scaling group, and check the status and type of ECS instances.
  ii. Assign the activity ID and execute the scaling activity.
  iii. Add the ECS instances.

iv. Modify Total Capacity.
v. Add the ECS instances to the whitelist of the associated ApsaraDB RDS instance.
vi. Associate the ECS instances with an SLB instance and set the weight to the SLB weight value that is specified in the active scaling configuration.

> **Note** If you want to manually add an instance to a scaling group, the instance type of the instance must be the same as that specified in the active scaling configuration of the scaling group. Therefore, you must set the weight to the SLB weight value that is specified in the active scaling configuration.

vii. The cooldown period starts after the scaling activity is complete.

- Manually remove instances
  i. Check the health status and boundary conditions of the scaling group.
  ii. Assign the activity ID and execute the scaling activity.
  iii. SLB stops forwarding traffic to ECS instances.
  iv. Remove the ECS instances from SLB after 60 seconds.
  v. Remove the ECS instances from the whitelist of the associated ApsaraDB RDS instance.
  vi. Modify Total Capacity.
  vii. Remove the ECS instances from the scaling group.
  viii. After the scaling activity is complete, the cooldown period starts.

4.1.2.6. Remove unhealthy ECS instances

Before you use ESS, you must understand information about the removal of unhealthy ECS instances.

After an ECS instance is added to a scaling group, ESS checks the status of the instance on a regular basis. If the ECS instance is not in the Running state, ESS removes the ECS instance from the scaling group. The removal method depends on how the ECS instance is added:

- If an ECS instance is automatically created, ESS immediately removes and releases it.
- If an ECS instance is manually added, ESS immediately removes it, but does not stop or release it.

The removal of unhealthy ECS instances is not limited by the MinSize value. After the unhealthy ECS instances are removed, the number of ECS instances (Total Capacity) may fall below the MinSize value. In this case, ESS automatically creates ECS instances based on the difference between the actual instance number and MinSize value to ensure that the total number of ECS instances is equal to the MinSize value.

4.1.2.7. Instance rollback after a failed scaling activity

Before you use ESS, you must understand the mechanism of instance rollback after a failed scaling activity.

If some ECS instances fail to be added to a scaling group during a scaling activity, ESS considers that the scaling activity is complete without trying to add the failed instances to the scaling group. ESS rolls back ECS instances, not the scaling activity.

For example, if a scaling group has created 20 ECS instances, and 19 of the instances are added to SLB instances, only the one ECS instance that failed to be added is automatically released.

4.1.2.8. Instance lifecycle management

Before you use Auto Scaling, we recommend that you understand the instance lifecycle.

Automatically created ECS instances

The ECS instances are automatically created by Auto Scaling based on user-defined scaling configurations and rules.

Auto Scaling manages the entire lifecycle of automatically created ECS instances. Auto Scaling creates ECS instances during scale-out activities, and stops and releases ECS instances during scale-in activities.

Manually added ECS instances

The ECS instances are manually added to scaling groups.

Auto Scaling does not manage the entire lifecycle of manually added ECS instances. These instances are not automatically created by Auto Scaling but are manually added by a user to a scaling group. If the ECS instances are manually or automatically removed from the scaling group, Auto Scaling removes the instances but does not stop or release them.

Instance status

An ECS instance in a scaling group can change to the following status during its lifecycle:

- Pending: The ECS instance is being added to the scaling group. The instance is being created, added as the backend server of the associated SLB instance, or added to the whitelist of the associated ApsaraDB RDS instance.
- InService: The ECS instance is added to the scaling group and is providing services as expected.
- Removing: The ECS instance is being removed from the scaling group.

Instance health status

An ECS instance in a scaling group can change to the following health status:

- Healthy
- Unhealthy

If an ECS instance is not in the Running state, Auto Scaling considers the instance unhealthy and automatically removes it from the scaling group.

- Auto Scaling stops and releases automatically created ECS instances.
- Auto Scaling does not stop or release manually added ECS instances.

4.1.3. Quick start

4.1.3.1. Overview

This topic describes how to get started with Auto Scaling.

You can perform the following steps to get started with Auto Scaling:

1. Create a scaling group

   Configure the parameters for the scaling group, such as Maximum Number of Instances and Minimum Number of Instances.

2. Create a scaling configuration

   Configure the parameters for the scaling configuration, such as Instance Type and Image.

3. Enable a scaling group

   Enable the scaling group for which the scaling configuration is enabled.

4. Create a scaling rule

   Add ECS instances to or remove ECS instances from the scaling group.

5. Create a scheduled task

Create scheduled tasks to add or remove instances at a specified point in time. Auto Scaling executes the scheduled tasks and scaling rules at the specified time. For example, Auto Scaling can trigger a task to execute a specified scaling rule at 08:00.

4.1.3.2. Log on to the Auto Scaling console

This topic describes how to log on to the Auto Scaling console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Elastic Computing > Auto Scaling**.

4.1.3.3. Create a scaling group

This topic describes how to create a scaling group. A scaling group is a group of Elastic Compute Service (ECS) instances that are dynamically scaled based on the preconfigured rules. You can specify the minimum and maximum numbers of ECS instances in a scaling group.

Prerequisites

- A virtual private cloud (VPC) and a vSwitch are created. For more information, see the "Create a VPC" and "Create a vSwitch" topics in *VPC User Guide*.
- To associate a scaling group with Server Load Balancer (SLB) instances, make sure that the following requirements are met:
  - You have one or more SLB instances in the **Running** state. For more information, see *Create an SLB instance* in *SLB User Guide*.
  - The SLB instances and the scaling group are in the same organization, resource set, and region.
  - If the network type is VPC, the SLB instances and the scaling group must be in the same VPC.
  - If the network type of the SLB instances is classic network, the network type of the scaling group is VPC, and the backend server groups of the SLB instances contain VPC-type ECS instances, the ECS instances and the scaling group must be in the same VPC.
  - At least one listener is configured for an SLB instance. For more information, see *Listener overview* in *SLB User Guide*.
  - The health check feature is enabled for the SLB instances. For more information, see *Configure health checks* in *SLB User Guide*.
- Before you associate ApsaraDB RDS instances with a scaling group, make sure that the following requirements are met:
  - You have one or more ApsaraDB RDS instances in the **Running** state. For more information, see *What is Apsara DB RDS?* in *ApsaraDB RDS Product Introduction*.
  - The ApsaraDB RDS instances and the scaling group are in the same organization, resource set, and region.

Procedure

1.
2.
3.
4. In the upper-right corner of the Scaling Groups page, click **Create Scaling Group**.
5. Configure the following parameters for the scaling group.

Parameter	Required	Description
Scaling Group	Yes	The name of the scaling group. The name must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). The name must start with a letter or a digit.
Organization/Resource Group	Yes	The organization and resource set to which the scaling group belongs.
Region	Yes	The ID of the region where the scaling group resides.
Maximum Number of Instances	Yes	The maximum number of instances that the scaling group can contain. To manage costs, specify a value based on your business requirements. Valid values: 0 to 1000.

Parameter	Required	Description
Minimum Number of Instances	Yes	The minimum number of instances that the scaling group must contain. To ensure service availability, specify a value based on your business requirements. When a scaling group is enabled, Auto Scaling automatically creates a certain number of ECS instances in the scaling group to maintain the minimum number. Valid values: 0 to 1000.
Cooldown Time (Seconds)	Yes	The period during which Auto Scaling cannot execute new scaling activities. The cooldown time occurs after Auto Scaling successfully executes a scaling activity. During the cooldown time, Auto Scaling rejects scaling requests triggered by event-triggered tasks from CloudMonitor. Scaling activities triggered by other types of tasks such as manually triggered tasks and scheduled tasks can be immediately executed. The value must be an integer that is greater than or equal to zero. Unit: seconds.
Scale-In Policy	No	The policy that is used to automatically remove the ECS instances from the scaling group. This parameter contains the **Filter First** and **Then Remove From Results** fields. You cannot specify the same values for the two fields. Valid values: ○ **Earliest Instance Created Using Scaling Configuration**: filters instances that are created based on the earliest scaling configuration. ○ **Earliest Created Instance**: filters instances that are added to the scaling group at the earliest point in time. ○ **Most Recently Created Instance**: filters instances that are added to the scaling group at the latest point in time ○ **None**: This value is available only for the **Then Remove From Results** field. This value specifies that Auto Scaling does not filter instances based on the Filter First field. For example, if Auto Scaling filters instances based on the **Earliest Created Instance** value of the Filter First field, you can select only one of the following values for the Then Remove From Results field: ○ **None**: specifies that Auto Scaling does not filter instances based on the Filter First field. ○ **Most Recently Created Instance**: filters instances that are most recently added to the scaling group from instances obtained based on the Filter First field.
VPC	Yes	The ID of the VPC to which the scaling group belongs.
vSwitch	Yes	The ID of the vSwitch with which the scaling group is associated.

Parameter	Required	Description
Associate SLB Instance	No	After you associate SLB instances with the scaling group, ECS instances that are added to the scaling group are automatically added as the backend servers of the SLB instances. You can specify one of the following server groups for the ECS instances: ◦ Default server group: The group of ECS instances that are used to receive requests. If the listener is not configured with a vServer group or a primary/secondary server group, requests are forwarded to the ECS instances in the default server group. ◦ vServer group: If you want to forward different requests to different backend servers or configure domain name- or URL-based routing methods, you can use vServer groups.
Associate RDS Instance	No	After you associate ApsaraDB RDS instances with the scaling group, the internal IP addresses of ECS instances in the scaling group are automatically added to the whitelists of the ApsaraDB RDS instances to allow mutual access between ECS instances and ApsaraDB RDS instances over the internal network.

6. Click **OK**.

Result

The created scaling group is displayed in the scaling group list but is in the **Disabled** state. You need to create scaling configurations to enable the scaling group. For more information, see Create a scaling configuration.

4.1.3.4. Create a scaling configuration

This topic describes how to create a scaling configuration for a scaling group.

Prerequisites

A security group is available in the virtual private cloud (VPC) where the scaling group resides. For more information, see the *Create a security group* topic in *ECS User Guide*.

Context

A limited number of scaling configurations can be created in a scaling group. For more information, see the *Limits* topic in *Auto Scaling Product Introduction*.

Procedure

1.
2.
3.
4.
5.
6. Choose **Create > Create Scaling Configuration**.
7. Configure the following parameters for the scaling configuration.

Section	Parameter	Required	Description
Region	Region	Yes	The region where the Elastic Compute Service (ECS) instance is located.
	Zone	Yes	The zone where the ECS instance is located.
Security Group	Security Group	Yes	The security group to which the ECS instance belongs.
Instance	Instance Type	Yes	The instance type of the ECS instance.
Image	Image Type	Yes	◦ **Public Image**: You can select public images provided by Alibaba Cloud. Public images are licensed to offer a secure and stable operating environment for applications on ECS instances. ◦ **Custom Image**: You can create custom images that you can use to install software or deploy projects that have special requirements. ◦ **Shared Custom Image**: You can select a shared custom image.
Storage	System Disk (GB)	Yes	The type and size of the system disk. The operating system is installed on the system disk. You can select **Ultra Disk** or **Standard SSD**.
	Data Disk (GB)	No	The type and size of the data disk. You can select **Ultra Disk** or **Standard SSD**. You can add a maximum of 16 data disks. The maximum capacity of each data disk is 32 TiB. You can select **Release with Instance** and **Encrypt** for each data disk.
	Password Setting	Yes	Specifies when to set the password. You can select **Set Now** or **Set after Purchase**. If you set Password Setting to Set after Purchase, you need to reset the password in the console after the instance is created. For more information, see the "Change the logon password" topic in *ECS User Guide*.

Section	Parameter	Required	Description
Password	Logon Password	No	The password that is used to log on to the ECS instance. The password must be 8 to 30 characters in length and contain at least three of the following character types: digits, uppercase letters, lowercase letters, and special characters. **Note** The password is used to log on to the operating system and is not the VNC password.
	Confirm Password	No	Enter the logon password again.
Deployment Set	Deployment Set	No	The deployment set to which the instance belongs.
Instance Name	Scaling Configuration	No	The name of the scaling configuration.
	Instance Name	No	The name of the ECS instance.
User Data	User Data	No	The Windows operating system supports batch and PowerShell scripts. Before you perform Base64 encoding of user data, make sure that the first line of the data is included in `[bat]` or `[powershell]`. You can run shell scripts in the Linux operating system for ECS instances.
Quantity	Quantity	No	The number of instances that you want to create.

8. Click **Submit**.

Result

After the scaling configuration is created, the scaling configuration enters the **Disabled** state and is displayed in your scaling configuration list. To enable this scaling configuration, click Select in the Actions column. Then you can use scaling configuration when an ECS instance is created. For more information, see Apply a scaling configuration.

4.1.3.5. Enable a scaling group

This topic describes how to enable a scaling group. You can enable a scaling group to trigger scaling activities.

Prerequisites

- The scaling group is in the **Disabled** state.
- The scaling group has a scaling configuration that is in the **Enabled** state.

Procedure

1.

2.
3.
4. Find the scaling group that you want to enable and click **Enable** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The state of the scaling group is changed from **Disabled** to **Enabled** in the **Status** column.

4.1.3.6. Create a scaling rule

This topic describes how to create a scaling rule. You can create scaling rules to add or remove Elastic Compute Service (ECS) instances. For example, you can add an ECS instance to a scaling group.

Context

- A limited number of scaling rules can be created in a scaling group. For more information, see the "Limits" topic in *Auto Scaling Product Introduction*.
- After a scaling rule is executed, the remaining number of ECS instances in the scaling group may be outside the specified range. In this case, Auto Scaling automatically adjusts the number of ECS instances to ensure that the number of ECS instances in the scaling group is within the specified range.

Procedure

1.
2.
3.
4.
5.
6. Choose **Create > Create Scaling Rule**.
7. Configure the following parameters for the scaling rule.

Parameter	Required	Description
Rule Name	Yes	The name of the scaling rule. The name must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). The name must start with a letter or digit.
Scaling Activity	Yes	The operation that is performed when the scaling rule is triggered. The operations include: ○ **Change to N Units**: When the scaling rule is executed, the number of instances in the scaling group is changed to N. ○ **Add N Units**: When the scaling rule is executed, N instances are added to the scaling group. ○ **Remove N Units**: When the scaling rule is executed, N instances are removed from the scaling group.
Default Cooldown (Seconds)	No	The cooldown period. If this parameter is not configured, the default value is used.

8. In the message that appears, click **OK**.

4.1.3.7. Create a scheduled task

This topic describes how to create a scheduled task to scale computing resources in response to predictable business changes in the future. Scheduled tasks enable the system to automatically obtain sufficient computing resources before business peaks and release idle computing resources after the business peaks.

Context

A scheduled task is preconfigured to execute the specified scaling rule at the specified time. When the specified time arrives, the scheduled task automatically scales computing resources. This allows you to reduce costs and meet business requirements. You can also specify recurring schedules for scheduled tasks if business changes are regular.

If multiple scheduled tasks need to be executed in 1 minute, Auto Scaling executes the most recently created scheduled task.

Procedure

1.
2.
3.
4. In the upper-right corner of the Scheduled Tasks page, click **Create Scheduled Task**.
5. In the dialog box that appears, configure parameters for the scheduled task.

Parameter	Required	Description
Task Name	Yes	The name of the scheduled task. The name must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or a digit.
Description	Yes	The description of the scheduled task.
Organization/Resource Group	Yes	The organization and resource set in which to create the scheduled task.
Start Time	Yes	The time to execute the scheduled task.
Scaling Rules	Yes	The scaling group to be monitored and the scaling rule to be executed.
Retry Interval (Seconds)	No	The period of time during which the system retires to execute the scheduled task. Unit: seconds. If a scaling activity fails to be executed at the specified time, Auto Scaling executes the scheduled task again within the period of time that is specified by the Retry Interval (Seconds) parameter.
Recurrence Settings (Advanced)	No	Specifies whether to execute the scheduled task on a recurring schedule. Select **Recurrence Settings (Advanced)** and set the Recurrence and Expire parameters. The valid values for Recurrence include **Daily**, **Weekly**, and **Monthly**.

6. Click **OK**.

Result

The scheduled task that you created is displayed in the scheduled task list.

4.1.3.8. Create an event-triggered task

This topic describes how to create an event-triggered task associated with monitoring metrics in response to emergent or unpredictable business changes. After you create and enable an event-triggered task, Auto Scaling collects data for the specified metric in real time and triggers an alert when the specified condition is met. Then, Auto Scaling executes the corresponding scaling rule to scale Elastic Compute Service (ECS) instances in the scaling group.

Procedure

1.
2.
3.
4. In the upper-right corner of the Event-Triggered Tasks page, click **Create Alert**.
5. In the dialog box that appears, configure parameters for the event-triggered task.

Parameter	Required	Description
Task Name	Yes	The name of the event-triggered task. It must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or a digit.
Description	No	The description of the event-triggered task.
Organization/Resource Group	Yes	The organization and resource set in which to create the event-triggered task.
Monitoring Metrics/Scaling Rule	Yes	The scaling group to be monitored and the scaling rule to be executed.
Monitoring Type	Yes	**System-Level Monitoring** is selected by default.
Monitoring Metrics	Yes	The metrics that you want to monitor. Valid values: ○ **Average CPU Utilization** ○ **Memory Usage** ○ **Outbound Traffic** ○ **Inbound Traffic** ○ **Average System Load**
Monitoring Period	Yes	The period during which data is aggregated and analyzed. The shorter the period, the higher the frequency that the alert is triggered. Unit: minutes. Valid values: ○ 1 ○ 2 ○ 5 ○ 15

Parameter	Required	Description
Statistic	Yes	The rule that determines whether to trigger an alert. Select **Average**, **Max Capacity**, or **Min Capacity**, and specify a threshold value. For example, to trigger an alert when the CPU utilization exceeds 80%, you can use one of the following methods to specify the trigger condition: ○ Average: An alert is triggered when the average CPU utilization of all ECS instances in the scaling group exceeds 80%. ○ Max Capacity: An alert is triggered when the highest CPU utilization among the ECS instances in the scaling group exceeds 80%. ○ Min Capacity: An alert is triggered when the lowest CPU utilization among the ECS instances in the scaling group exceeds 80%.
Trigger After	Yes	The number of consecutive times that the threshold must be exceeded before the alert is triggered. Valid values: ○ 1 ○ 2 ○ 3 ○ 5

6. Click **OK**.

4.1.4. Scaling groups

4.1.4.1. Create a scaling group

This topic describes how to create a scaling group. A scaling group is a group of Elastic Compute Service (ECS) instances that are dynamically scaled based on the preconfigured rules. You can specify the minimum and maximum numbers of ECS instances in a scaling group.

Prerequisites

- A virtual private cloud (VPC) and a vSwitch are created. For more information, see the "Create a VPC" and "Create a vSwitch" topics in *VPC User Guide*.
- To associate a scaling group with Server Load Balancer (SLB) instances, make sure that the following requirements are met:
  - You have one or more SLB instances in the **Running** state. For more information, see *Create an SLB instance* in *SLB User Guide*.
  - The SLB instances and the scaling group are in the same organization, resource set, and region.
  - If the network type is VPC, the SLB instances and the scaling group must be in the same VPC.
  - If the network type of the SLB instances is classic network, the network type of the scaling group is VPC, and the backend server groups of the SLB instances contain VPC-type ECS instances, the ECS instances and the scaling group must be in the same VPC.
  - At least one listener is configured for an SLB instance. For more information, see *Listener overview* in *SLB User Guide*.
  - The health check feature is enabled for the SLB instances. For more information, see *Configure health checks* in *SLB User Guide*.
- Before you associate ApsaraDB RDS instances with a scaling group, make sure that the following requirements are met:

- You have one or more ApsaraDB RDS instances in the **Running** state. For more information, see *What is Apsara DB RDS?* in *ApsaraDB RDS Product Introduction*.
- The ApsaraDB RDS instances and the scaling group are in the same organization, resource set, and region.

Procedure

1.
2.
3.
4. In the upper-right corner of the Scaling Groups page, click **Create Scaling Group**.
5. Configure the following parameters for the scaling group.

Parameter	Required	Description
Scaling Group	Yes	The name of the scaling group. The name must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). The name must start with a letter or a digit.
Organization/Resource Group	Yes	The organization and resource set to which the scaling group belongs.
Region	Yes	The ID of the region where the scaling group resides.
Maximum Number of Instances	Yes	The maximum number of instances that the scaling group can contain. To manage costs, specify a value based on your business requirements. Valid values: 0 to 1000.
Minimum Number of Instances	Yes	The minimum number of instances that the scaling group must contain. To ensure service availability, specify a value based on your business requirements. When a scaling group is enabled, Auto Scaling automatically creates a certain number of ECS instances in the scaling group to maintain the minimum number. Valid values: 0 to 1000.
Cooldown Time (Seconds)	Yes	The period during which Auto Scaling cannot execute new scaling activities. The cooldown time occurs after Auto Scaling successfully executes a scaling activity. During the cooldown time, Auto Scaling rejects scaling requests triggered by event-triggered tasks from CloudMonitor. Scaling activities triggered by other types of tasks such as manually triggered tasks and scheduled tasks can be immediately executed. The value must be an integer that is greater than or equal to zero. Unit: seconds.

Parameter	Required	Description
Scale-In Policy	No	The policy that is used to automatically remove the ECS instances from the scaling group. This parameter contains the **Filter First** and **Then Remove From Results** fields. You cannot specify the same values for the two fields. Valid values: ◦ **Earliest Instance Created Using Scaling Configuration**: filters instances that are created based on the earliest scaling configuration. ◦ **Earliest Created Instance**: filters instances that are added to the scaling group at the earliest point in time. ◦ **Most Recently Created Instance**: filters instances that are added to the scaling group at the latest point in time ◦ **None**: This value is available only for the **Then Remove From Results** field. This value specifies that Auto Scaling does not filter instances based on the Filter First field. For example, if Auto Scaling filters instances based on the **Earliest Created Instance** value of the Filter First field, you can select only one of the following values for the Then Remove From Results field: ◦ **None**: specifies that Auto Scaling does not filter instances based on the Filter First field. ◦ **Most Recently Created Instance**: filters instances that are most recently added to the scaling group from instances obtained based on the Filter First field.
VPC	Yes	The ID of the VPC to which the scaling group belongs.
vSwitch	Yes	The ID of the vSwitch with which the scaling group is associated.
Associate SLB Instance	No	After you associate SLB instances with the scaling group, ECS instances that are added to the scaling group are automatically added as the backend servers of the SLB instances. You can specify one of the following server groups for the ECS instances: ◦ Default server group: The group of ECS instances that are used to receive requests. If the listener is not configured with a vServer group or a primary/secondary server group, requests are forwarded to the ECS instances in the default server group. ◦ vServer group: If you want to forward different requests to different backend servers or configure domain name- or URL-based routing methods, you can use vServer groups.
Associate RDS Instance	No	After you associate ApsaraDB RDS instances with the scaling group, the internal IP addresses of ECS instances in the scaling group are automatically added to the whitelists of the ApsaraDB RDS instances to allow mutual access between ECS instances and ApsaraDB RDS instances over the internal network.

6. Click **OK**.

Result

The created scaling group is displayed in the scaling group list but is in the **Disabled** state. You need to create scaling configurations to enable the scaling group. For more information, see Create a scaling configuration.

4.1.4.2. Enable a scaling group

This topic describes how to enable a scaling group. You can enable a scaling group to trigger scaling activities.

Prerequisites

- The scaling group is in the **Disabled** state.
- The scaling group has a scaling configuration that is in the **Enabled** state.

Procedure

1.
2.
3.
4. Find the scaling group that you want to enable and click **Enable** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The state of the scaling group is changed from **Disabled** to **Enabled** in the **Status** column.

4.1.4.3. Query scaling groups

This topic describes how to query scaling groups and view the details about a scaling group.

Procedure

1.
2.
3. In the top navigation bar, select an organization, a resource set, and a region.
   The scaling groups that belong to the specified organization, resource set, and region are displayed.
4. Select a filter option, enter the information about a scaling group, and then click **Search**.

   You can select multiple filter options to search for scaling groups.

Filter option	Description
Scaling Group Name	Enter the name of the scaling group that you want to search for.
Scaling Group ID	Enter the ID of the scaling group that you want to search for.

5. Click the name of the scaling group in the **Scaling Group Name/ID** column.
6. View the details about the scaling group.

Page	Description
Basic Information	The configuration information about the scaling group, including the scaling group ID, scaling group name, total instances, minimum number of instances, maximum number of instances, and scale-in policy.

Page	Description
ECS Instances	The details about the ECS instances, including the list of automatically created ECS instances, the list of manually added ECS instances, and the number of ECS instances that are in service.
Scaling Activities	The scaling activities that were executed in the scaling group.
Scaling Configuration	The information about the scaling configuration in the scaling group.
Scaling Rules	The information about the scaling rules.

4.1.4.4. Edit a scaling group

This topic describes how to edit a scaling group. You can edit the parameter settings of a specified scaling group, including the minimum and maximum numbers of ECS instances.

Context

After you change the minimum or maximum number of ECS instances in a scaling group, Auto Scaling automatically creates or removes ECS instances to maintain the number of instances within the valid range.

Procedure

1.
2.
3.
4. Find the scaling group that you want to edit and click **Edit** in the **Actions** column.
5. Edit the parameter settings of the scaling group.

   You can edit only the scaling configuration and other parameters. You cannot change the organization and resource set. For more information about other parameters, see Create a scaling group.
6. Click **OK**.

4.1.4.5. Disable a scaling group

This topic describes how to disable a scaling group.

Prerequisites

- Make sure that the scaling group that you want to disable does not have any scaling activities in progress.
- The scaling group is in the **Enabled** state.

Procedure

1.
2.
3.
4. Find the scaling group that you want to disable and click **Disable** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The state of the scaling group is changed from **Enabled** to **Disabled** in the **Status** column.

4.1.4.6. Delete a scaling group

This topic describes how to delete a scaling group. When you delete a scaling group, Auto Scaling removes and releases ECS instances that are automatically created, removes ECS instances that are manually added, and deletes the scaling configurations and rules in the scaling group. However, Auto Scaling does not delete the scheduled tasks and event-triggered tasks that are associated with the scaling group.

Procedure

1.
2.
3.
4. Find the scaling group that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

4.1.4.7. Query ECS instances

You can query all ECS instances in a scaling group and their states.

Procedure

1.
2.
3.
4.
5.
6. View the details of ECS instances.

Category	Description
Automatically created ECS instances	The ECS instances that are automatically created based on the active scaling configuration when a scaling rule is triggered.
Manually added ECS instances	The ECS instances that are manually added to the specified scaling group.
The number of ECS instances in each state.	The following section describes the states: ○ Total: all ECS instances in the scaling group ○ In Service: the ECS instances that are in normal use ○ On Standby: the ECS instances that are on standby ○ Protected: the ECS instances that are protected ○ Adding: the ECS instances that are being added to the scaling group ○ Removing: the ECS instances that are being removed from the scaling group **Note** The Disabled, Adding:wait, and Suspending states are unavailable.

4.1.4.8. Switch an ECS instance to the Standby state

This topic describes how to switch an ECS instance to the Standby state. Auto Scaling does not perform health checks on or release ECS instances that are in the Standby state.

Context

The following items describe an ECS instance that is switched to the Standby state:

- The ECS instance remains in the Standby state until you manually remove it from the Standby state.
- Auto Scaling stops managing the lifecycle of the ECS instance. You must manually manage the lifecycle of the ECS instance.
- If a scale-in activity is triggered, Auto Scaling does not remove the ECS instance.
- When the ECS instance is stopped or restarted, the health check status of the ECS instance remains unchanged.
- To release the ECS instance, you must first remove it from the scaling group.
- If you delete the scaling group, the ECS instance is automatically removed from the Standby state and is released.
- You can also perform other operations on the ECS instance, such as stop, restart, change the instance type of, or change the operating system of the ECS instance.

Procedure

1.
2.
3.
4.
5.
6.
7. Find the ECS instance that you want to switch to the Standby state and choose **Actions > Switch to Standby** in the **Actions** column.
8. In the message that appears, click **OK**.

4.1.4.9. Remove an ECS instance from the Standby state

This topic describes how to remove an ECS instance from the Standby state. You can reuse instances that are removed from the Standby state.

Context

The following items describe an ECS instance that is removed from the Standby state:

- The ECS instance enters the In Service state.
- When the ECS instance is stopped or restarted, its health status is updated.
- Auto Scaling continues to manage the lifecycle of the ECS instance, and can remove the ECS instance from the scaling group during a scaling activity.

Procedure

1.
2.
3.
4.
5.

6.
7. Find the ECS instance that you want to remove from the Standby state and choose **Actions > Move Out Of Standby** in the **Actions** column.
8. In the message that appears, click **OK**.

4.1.4.10. Switch an ECS instance to the Protected state

This topic describes how to switch an ECS instance to the Protected state. Auto Scaling does not perform health checks on or release ECS instances that are in the Protected state.

Context

The following items describe an ECS instance that is switched to the Protected state:

- The ECS instance remains in the Protected state until you manually remove it from the Protected state.
- If a scale-in activity is triggered, Auto Scaling does not remove the ECS instance. To release the ECS instance, you must remove the ECS instance from the Protected state and then remove it from the scaling group.
- When the ECS instance is stopped or restarted, the health check status of the ECS instance remains unchanged.

Procedure

1.
2.
3.
4.
5.
6.
7. Find the ECS instance that you want to switch to the Protected state and choose **Actions > Switch to Protected** in the **Actions** column.
8. In the message that appears, click **OK**.

4.1.4.11. Remove an ECS instance from the Protected state

This topic describes how to remove an ECS instance from the Protected state. After an ECS instance is removed from the Protected state, Auto Scaling continues to manage the lifecycle of the ECS instance.

Procedure

1.
2.
3.
4.
5.
6.
7. Find the ECS instance that you want to remove from the Protected state and choose **Actions > Move Out Of Protection** in the **Actions** column.
8. In the message that appears, click **OK**.

4.1.5. Scaling configurations

4.1.5.1. Create a scaling configuration

This topic describes how to create a scaling configuration for a scaling group.

Prerequisites

A security group is available in the virtual private cloud (VPC) where the scaling group resides. For more information, see the *Create a security group* topic in *ECS User Guide*.

Context

A limited number of scaling configurations can be created in a scaling group. For more information, see the *Limits* topic in *Auto Scaling Product Introduction*.

Procedure

1.
2.
3.
4.
5.
6. Choose **Create > Create Scaling Configuration**.
7. Configure the following parameters for the scaling configuration.

Section	Parameter	Required	Description
Region	Region	Yes	The region where the Elastic Compute Service (ECS) instance is located.
	Zone	Yes	The zone where the ECS instance is located.
Security Group	Security Group	Yes	The security group to which the ECS instance belongs.
Instance	Instance Type	Yes	The instance type of the ECS instance.
Image	Image Type	Yes	◦ **Public Image**: You can select public images provided by Alibaba Cloud. Public images are licensed to offer a secure and stable operating environment for applications on ECS instances. ◦ **Custom Image**: You can create custom images that you can use to install software or deploy projects that have special requirements. ◦ **Shared Custom Image**: You can select a shared custom image.

Section	Parameter	Required	Description
Storage	System Disk (GB)	Yes	The type and size of the system disk. The operating system is installed on the system disk. You can select **Ultra Disk** or **Standard SSD**.
	Data Disk (GB)	No	The type and size of the data disk. You can select **Ultra Disk** or **Standard SSD**. You can add a maximum of 16 data disks. The maximum capacity of each data disk is 32 TiB. You can select **Release with Instance** and **Encrypt** for each data disk.
Password	Password Setting	Yes	Specifies when to set the password. You can select **Set Now** or **Set after Purchase**. If you set Password Setting to Set after Purchase, you need to reset the password in the console after the instance is created. For more information, see the "Change the logon password" topic in *ECS User Guide*.
	Logon Password	No	The password that is used to log on to the ECS instance. The password must be 8 to 30 characters in length and contain at least three of the following character types: digits, uppercase letters, lowercase letters, and special characters. **Note** The password is used to log on to the operating system and is not the VNC password.
	Confirm Password	No	Enter the logon password again.
Deployment Set	Deployment Set	No	The deployment set to which the instance belongs.
Instance Name	Scaling Configuration	No	The name of the scaling configuration.
	Instance Name	No	The name of the ECS instance.

Section	Parameter	Required	Description
User Data	User Data	No	The Windows operating system supports batch and PowerShell scripts. Before you perform Base64 encoding of user data, make sure that the first line of the data is included in `[bat]` or `[powershell]`. You can run shell scripts in the Linux operating system for ECS instances.
Quantity	Quantity	No	The number of instances that you want to create.

8. Click **Submit**.

Result

After the scaling configuration is created, the scaling configuration enters the **Disabled** state and is displayed in your scaling configuration list. To enable this scaling configuration, click Select in the Actions column. Then you can use scaling configuration when an ECS instance is created. For more information, see Apply a scaling configuration.

4.1.5.2. View scaling configurations

This topic describes how to view scaling configurations.

Procedure

1.
2.
3. In the top navigation bar, select an organization, a resource set, and a region.
   The scaling groups that belong to the specified organization, resource set, and region are displayed.
4.
5.
6. View the scaling configurations.

4.1.5.3. Modify a scaling configuration

This topic describes how to modify a scaling configuration. You can modify the parameter settings of a scaling configuration based on your business requirements.

Procedure

1.
2.
3.
4.
5.
6. Find the scaling configuration that you want to modify and click its name in the **Scaling Configuration Name/ID** column.
7. Modify the parameter settings of the scaling configuration.

   For more information about the parameters of the scaling configuration, see Create a scaling configuration.
8. Click **OK**.

4.1.5.4. Apply a scaling configuration

This topic describes how to apply a scaling configuration. You can create multiple scaling configurations for a scaling group and apply one of the scaling configurations based on your business requirements.

Procedure

1.
2.
3.
4.
5.
6. Find the scaling configuration that you want to apply and click **Apply** in the **Actions** column.

   Only one scaling configuration can be in the **Enabled** state in a scaling group. After a scaling configuration is applied, other scaling configurations are switched to the **Disabled** state.
7. Click **OK**.

Result

The state of the scaling configuration is changed from **Disabled** to **Enabled** in the **Status** column.

4.1.5.5. Delete a scaling configuration

This topic describes how to delete a scaling configuration that you no longer use. After you delete a scaling configuration, the ECS instances that are created from the scaling configuration are not released.

Prerequisites

The scaling configuration is in the **Disabled** state.

Procedure

1.
2.
3.
4.
5.
6. Find the scaling configuration that you want to delete and click **Select Delete** in the **Actions** column.
7. In the message that appears, click **OK**.

4.1.6. Scaling rules

4.1.6.1. Create a scaling rule

This topic describes how to create a scaling rule. You can create scaling rules to add or remove Elastic Compute Service (ECS) instances. For example, you can add an ECS instance to a scaling group.

Context

- A limited number of scaling rules can be created in a scaling group. For more information, see the "Limits" topic in *Auto Scaling Product Introduction*.
- After a scaling rule is executed, the remaining number of ECS instances in the scaling group may be outside the specified range. In this case, Auto Scaling automatically adjusts the number of ECS instances to ensure that the number of ECS instances in the scaling group is within the specified range.

Procedure

1.
2.
3.
4.
5.
6. Choose **Create > Create Scaling Rule**.
7. Configure the following parameters for the scaling rule.

Parameter	Required	Description
Rule Name	Yes	The name of the scaling rule. The name must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). The name must start with a letter or digit.
Scaling Activity	Yes	The operation that is performed when the scaling rule is triggered. The operations include: ○ **Change to N Units**: When the scaling rule is executed, the number of instances in the scaling group is changed to N. ○ **Add N Units**: When the scaling rule is executed, N instances are added to the scaling group. ○ **Remove N Units**: When the scaling rule is executed, N instances are removed from the scaling group.
Default Cooldown (Seconds)	No	The cooldown period. If this parameter is not configured, the default value is used.

8. In the message that appears, click **OK**.

4.1.6.2. View scaling rules

This topic describes how to view scaling rules.

Procedure

1.
2.
3. In the top navigation bar, select an organization, a resource set, and a region.
   The scaling groups that belong to the specified organization, resource set, and region are displayed.
4.
5.
6. View scaling rules.

4.1.6.3. Modify a scaling rule

This topic describes how to modify a scaling rule. You can modify the settings of the following parameters of a scaling rule: Rule Name, Scaling Activity, and Default Cooldown (Seconds).

Procedure

1.
2.
3.
4.
5.
6. Find the scaling rule that you want to modify and click **Edit** in the **Actions** column.
7. Modify the Rule Name, Scaling Activity, and Default Cooldown (Seconds) parameters based on your business requirements.
8. Click **OK**.

4.1.6.4. Delete a scaling rule

This topic describes how to delete a scaling rule that you no longer use.

Procedure

1.
2.
3.
4.
5.
6. Find the scaling rule that you want to delete and click **Delete** in the **Actions** column.
7. In the message that appears, click **OK**.

4.1.7. Scaling tasks

4.1.7.1. Manually execute a scaling rule

This topic describes how to manually execute a scaling rule to add or remove ECS instances.

Prerequisites

- The scaling group for which the scaling rule is created is in the **Enabled** state.
- No scaling activity is in progress in the scaling group for which the scaling rule is created.

Context

After the scaling rule is executed, Auto Scaling automatically adds or removes ECS instances to maintain the number of ECS instances based on the specified maximum and minimum number.

Auto Scaling allows you to manually execute scaling rules. You can also associate scaling rules with scaling tasks when you create scheduled tasks or event-triggered tasks to enable auto scaling. For more information, see Create a scheduled task and Create an event-triggered task.

Procedure

1.
2.
3.
4.
5.
6. Find the scaling rule that you want to execute and click **Run** in the **Actions** column.

7. In the message that appears, click **OK**.

Result

The **Scaling Activities** page appears. You can view the details about your scaling activities.

4.1.7.2. Manually add an ECS instance to a scaling group

This topic describes how to manually add an ECS instance to a scaling group. You can add existing ECS instances to a scaling group to maximize the use of compute resources.

Prerequisites

The ECS instance that you want to add must meet the following conditions:

- The ECS instance and the scaling group to which you want to add the ECS instance share the same region, organization, and resource set.
- The ECS instance is in the **Running** state.
- The ECS instance does not belong to another scaling group.
- The ECS instance and the scaling group are in the same virtual private cloud (VPC).

The scaling group to which you want to add the ECS instance must meet the following conditions:

- The scaling group is in the **Enabled** state.
- No scaling activity is in progress in the scaling group.

Context

- You can manually remove an ECS instance from a scaling group without the need to wait for the cooldown period to expire.
- After the ECS instance is added to a scaling group, the number of instances in the scaling group must be less than or equal to the maximum number of instances. Otherwise, the ECS instance cannot be added.
- The ECS instances that are manually added to a scaling group are not limited by the scaling configurations. The instance types of the manually added instances can be different from the instance type of the scaling configuration that is in the **Enabled** state.

Procedure

1.
2.
3.
4.
5.
6. Click **Add Instance**.
7. Select the ECS instance that you want to add to the scaling group and click **OK**.

Result

The manually added instance is displayed on the **Manually Added** tab.

4.1.7.3. Manually remove an ECS instance

This topic describes how to manually remove an ECS instance that you no longer use from a scaling group.

Prerequisites

The scaling group from which you want to remove the ECS instance must meet the following conditions:

- The scaling group is in the **Enabled** state.

- No scaling activity is in progress in the scaling group.

Context

- You can manually remove the ECS instance from a scaling group without the need to wait for the cooldown period to expire.
- After the ECS instance is removed from a scaling group, the number of instances in the scaling group must be greater than or equal to the minimum number of instances. Otherwise, the ECS instance cannot be removed.

Procedure

1.
2.
3.
4.
5.
6.
7. Select a method to remove the ECS instance.

   Manually added ECS instances can only be removed, but cannot be released.

   - Find the ECS instance that you want to remove from the scaling group and choose **Actions > Remove from Scaling Group** in the **Actions** column.
   - Find the ECS instance that you want to remove and release, and choose **Actions > Remove from Scaling Group and Release** in the **Actions** column.
8. In the message that appears, click **OK**.

4.1.8. Scheduled tasks

4.1.8.1. Create a scheduled task

This topic describes how to create a scheduled task to scale computing resources in response to predictable business changes in the future. Scheduled tasks enable the system to automatically obtain sufficient computing resources before business peaks and release idle computing resources after the business peaks.

Context

A scheduled task is preconfigured to execute the specified scaling rule at the specified time. When the specified time arrives, the scheduled task automatically scales computing resources. This allows you to reduce costs and meet business requirements. You can also specify recurring schedules for scheduled tasks if business changes are regular.

If multiple scheduled tasks need to be executed in 1 minute, Auto Scaling executes the most recently created scheduled task.

Procedure

1.
2.
3.
4. In the upper-right corner of the Scheduled Tasks page, click **Create Scheduled Task**.
5. In the dialog box that appears, configure parameters for the scheduled task.

Parameter	Required	Description

Parameter	Required	Description
Task Name	Yes	The name of the scheduled task. The name must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or a digit.
Description	Yes	The description of the scheduled task.
Organization/Resource Group	Yes	The organization and resource set in which to create the scheduled task.
Start Time	Yes	The time to execute the scheduled task.
Scaling Rules	Yes	The scaling group to be monitored and the scaling rule to be executed.
Retry Interval (Seconds)	No	The period of time during which the system retires to execute the scheduled task. Unit: seconds. If a scaling activity fails to be executed at the specified time, Auto Scaling executes the scheduled task again within the period of time that is specified by the Retry Interval (Seconds) parameter.
Recurrence Settings (Advanced)	No	Specifies whether to execute the scheduled task on a recurring schedule. Select **Recurrence Settings (Advanced)** and set the Recurrence and Expire parameters. The valid values for Recurrence include **Daily**, **Weekly**, and **Monthly**.

6. Click **OK**.

Result

The scheduled task that you created is displayed in the scheduled task list.

4.1.8.2. View scheduled tasks

This topic describes how to view scheduled tasks.

Procedure

1.
2.
3. In the top navigation bar, select an organization, a resource set, and a region.
   The scheduled tasks that correspond to the specified organization, resource set, and region are displayed.
4. Select a filter option, enter the corresponding information, and then click **Search**.

   You can select multiple filter options to narrow down the search results.

Option	Description
Task Name	Enter a task name to search for the scheduled task.
Task ID	Enter a task ID to search for the scheduled task.

5. View the scheduled task list.

4.1.8.3. Modify a scheduled task

This topic describes how to modify a scheduled task. You can modify parameters such as Start Time, Scaling Rules, and Retry Expiry Time for a scheduled task.

Procedure

1.
2.
3.
4. Find the scheduled task that you want to modify and click **Edit** in the **Actions** column.
5. Modify the parameters of the scheduled task.

   You can modify the Recurrence and Expire parameters if you have enabled the Recurrence Settings (Advanced) feature when you create the scheduled task, but the Recurrence Settings (Advanced) feature cannot be disabled. For more information about other parameters of the scheduled task, see Create a scheduled task.
6. Click **OK**.

4.1.8.4. Disable a scheduled task

This topic describes how to disable a scheduled task. You can disable a scheduled task that is no longer needed.

Prerequisites

The scheduled task is in the **Running** state.

Procedure

1.
2.
3.
4. Find the scheduled task that you want to disable and click **Disabled** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The status of the scheduled task is changed from **Running** to **Stop** in the **Status** column.

4.1.8.5. Enable a scheduled task

This topic describes how to enable a scheduled task. You can enable a scheduled task that has been disabled and use it to trigger scaling activities at the specified time point.

Prerequisites

The scheduled task is in the **Stop** state.

Procedure

1.
2.
3.
4. Find the scheduled task that you want to enable and click **Enable** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The status of the scheduled task is changed from **Stop** to **Running** in the **Status** column.

4.1.8.6. Delete a scheduled task

This topic describes how to delete a scheduled task that is no longer needed.

Procedure

1.
2.
3.
4. Find the scheduled task that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

4.1.9. Event-triggered tasks

4.1.9.1. Create an event-triggered task

This topic describes how to create an event-triggered task associated with monitoring metrics in response to emergent or unpredictable business changes. After you create and enable an event-triggered task, Auto Scaling collects data for the specified metric in real time and triggers an alert when the specified condition is met. Then, Auto Scaling executes the corresponding scaling rule to scale Elastic Compute Service (ECS) instances in the scaling group.

Procedure

1.
2.
3.
4. In the upper-right corner of the Event-Triggered Tasks page, click **Create Alert**.
5. In the dialog box that appears, configure parameters for the event-triggered task.

Parameter	Required	Description
Task Name	Yes	The name of the event-triggered task. It must be 2 to 64 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or a digit.
Description	No	The description of the event-triggered task.
Organization/Resource Group	Yes	The organization and resource set in which to create the event-triggered task.
Monitoring Metrics/Scaling Rule	Yes	The scaling group to be monitored and the scaling rule to be executed.
Monitoring Type	Yes	**System-Level Monitoring** is selected by default.
Monitoring Metrics	Yes	The metrics that you want to monitor. Valid values: ○ **Average CPU Utilization** ○ **Memory Usage** ○ **Outbound Traffic** ○ **Inbound Traffic** ○ **Average System Load**

Parameter	Required	Description
Monitoring Period	Yes	The period during which data is aggregated and analyzed. The shorter the period, the higher the frequency that the alert is triggered. Unit: minutes. Valid values: ○ 1 ○ 2 ○ 5 ○ 15
Statistic	Yes	The rule that determines whether to trigger an alert. Select **Average**, **Max Capacity**, or **Min Capacity**, and specify a threshold value. For example, to trigger an alert when the CPU utilization exceeds 80%, you can use one of the following methods to specify the trigger condition: ○ Average: An alert is triggered when the average CPU utilization of all ECS instances in the scaling group exceeds 80%. ○ Max Capacity: An alert is triggered when the highest CPU utilization among the ECS instances in the scaling group exceeds 80%. ○ Min Capacity: An alert is triggered when the lowest CPU utilization among the ECS instances in the scaling group exceeds 80%.
Trigger After	Yes	The number of consecutive times that the threshold must be exceeded before the alert is triggered. Valid values: ○ 1 ○ 2 ○ 3 ○ 5

6. Click **OK**.

4.1.9.2. View event-triggered tasks

This topic describes how to view event-triggered tasks.

Procedure

1.
2.
3. In the top navigation bar, select an organization, a resource set, and a region.
   The event-triggered tasks that correspond to the specific organization, resource set, and region are displayed.
4. Select a filter option, enter the corresponding information, and then click **Search**.

   You can select multiple filter options to narrow down the search results.

Option	Description
Alert Name	Enter an event-triggered task name to search for the event-triggered task.
Scaling Group ID	Enter a scaling group ID to search for the event-triggered task associated with the scaling group.

4.1.9.3. Modify an event-triggered task

This topic describes how to modify an event-triggered task. You can modify parameters such as Scaling Rules, Monitoring Type, and Statistic for an event-triggered task.

Procedure

1.
2.
3.
4. Find the event-triggered task that you want to modify and click **Edit** in the **Actions** column.
5. Modify the parameters of the event-triggered task.

   For more information about other parameters of the scheduled task, see Create an event-triggered task. The following parameters cannot be modified:

   - Organization
   - Resource Group
   - Monitoring Metrics
   - Monitoring Period
6. Click **OK**.

4.1.9.4. Disable an event-triggered task

This topic describes how to disable an event-triggered task. You can disable an event-triggered task if you no longer want to use it to trigger scaling activities.

Prerequisites

The event-triggered task is in the **Normal**, **Alerts**, or **Insufficient Data** state.

Procedure

1.
2.
3.
4. Find the event-triggered task that you want to disable and click **Disable** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The status of the event-triggered task is changed to **Stopped** in the **Status** column.

4.1.9.5. Enable an event-triggered task

This topic describes how to enable an event-triggered task. You can enable an event-triggered task that has been disabled to continue to monitor metrics and trigger scaling activities for a scaling group.

Prerequisites

The event-triggered task is in the **Stopped** state.

Procedure

1.
2.

3.
4. Find the event-triggered task that you want to enable and click **Enable** in the **Actions** column.
5. In the message that appears, click **OK**.

Result

The status of the event-triggered task changes from **Stopped** to **Normal** in the **Status** column.

4.1.9.6. Delete an event-triggered task

This topic describes how to delete an event-triggered task. You can delete an event-triggered task that is no longer needed.

Procedure

1.
2.
3.
4. Find the event-triggered task that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

5.Resource Orchestration Service (ROS)

5.1. User Guide

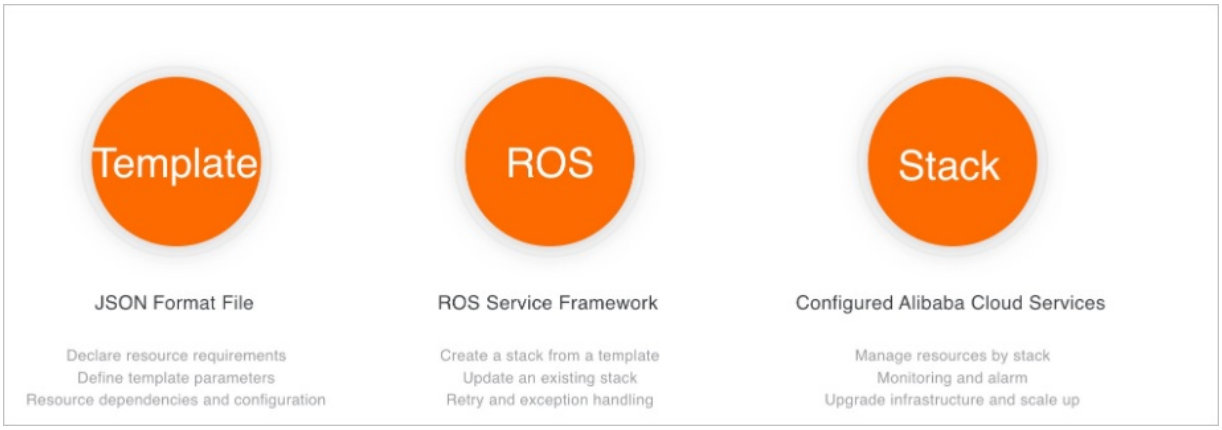
5.1.1. What is ROS?

Resource Orchestration Service (ROS) is an Apsara Stack service that can simplify the management of cloud computing resources. You can create stack templates based on the template specifications defined in ROS. Within a template, you can define cloud computing resources such as Elastic Compute Service (ECS) and ApsaraDB RDS instances, and the dependencies between resources. The ROS engine automatically creates and configures all resources in a stack based on a template, which makes automatic deployment and O&M possible.

An ROS template is a readable, easy-to-create text file. You can directly edit a JSON template or use version control tools such as Apache Subversion (SVN) and Git to manage the template and infrastructure versions. You can use APIs and SDKs to integrate the orchestration capabilities of ROS with your own applications to implement Infrastructure as Code (IaC).

ROS templates are also a standardized way to deliver resources and applications. If you are an independent software vendor (ISV), you can use ROS templates to deliver a holistic system or solution that encompasses cloud resources and applications. ISVs can use this method to integrate Apsara Stack resources with their own software systems for centralized delivery.

ROS manages a group of cloud resources as a single unit called stack. A stack is a group of Apsara Stack resources. You can create, delete, and clone cloud resources by stack.



5.1.2. Log on to the ROS console

This topic describes how to log on to the Resource Orchestration Service (ROS) console.

Prerequisites

- The URL, username, and password of the Apsara Uni-manager Management Console are obtained from the operations administrator before you log on.
- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter the correct username and password.

The first time you log on to the Apsara Uni-manager Management Console, you must change the password of your Apsara Stack tenant account. For security purposes, your password must meet the minimum complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:

- Uppercase and lowercase letters
- Digits
- Special characters including exclamation points (!), at signs (@), number signs (#), dollar signs ($), and percent signs (%)

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products > Elastic Computing > Resource Orchestration Service**.

5.1.3. Manage stacks

5.1.3.1. Create a stack

Resource Orchestration Service (ROS) allows you to use a template to create a group of Apsara Stack resources. This topic describes how to create a stack.

Method 1: (Recommended) Use the Stacks page

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **Stacks**.
3. On the **Stacks** page, click **Create Stack**.
4. In the **Select Template** step, configure **Organization**, **Resource Set**, and **regionId**.
5. Select a value for Specify Template Import Method. Enter the content of your template in the Prepare Template section. Then, click **Next**.
   - **Enter Template Content**: Enter the content of a JSON template in the **Prepare Template** section.
   - **My Templates**: Select an existing template. For more information about how to create a template, see Create a template.
6. In the **Configure Template Parameters** step, configure parameters in the Stacks and Set Parameters sections. Then, click **Next**.

   > **Note** The parameters that you need to configure vary based on the templates. Configure the parameters as instructed in the ROS console.
7. In the **Configure Stack** step, configure Rollback on Failure and Timeout Period. Then, click **Next**.
8. In the **Confirm** step, check the template and stack configurations, and click **Create Stack**.

After the stack is created, the state of the stack is **Created**. You can click the ID of the stack to go to the following tabs: Stack Information, Events, Resources, Outputs, Parameters, and Template.

Method 2: Use an existing template

If you have an existing template, you can perform the following steps to create a stack. For more information about how to create a template, see Create a template.

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **My Templates**.
3. On the **My Templates** page, find the template that you want to use and click **Create** in the **Actions** column.
4. In the **Select Template** step, configure **Organization**, **Resource Set**, and **regionId**. Then, click **Next**.
5. In the **Configure Template Parameters** step, configure parameters in the Stacks and Set Parameters sections. Then, click **Next**.
6. In the **Configure Stack** step, configure Rollback on Failure and Timeout Period. Then, click **Next**.
7. In the **Confirm** step, check the template and stack configurations, and click **Create Stack**.

   After the stack is created, the state of the stack is **Created**. You can click the ID of the stack to go to the following tabs: Stack Information, Events, Resources, Outputs, Parameters, and Template.

5.1.3.2. Update a stack

You can update a stack if you want to modify only the template that is used to create the stack or the stack configurations. The update operation does not change the organization, resource set, or region ID of your stack. This topic describes how to update a stack.

Prerequisites

A stack is created. For more information, see Create a stack.

Procedure

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **Stacks**.
3. On the **Stacks** page, find the stack that you want to update and click **Update** in the **Actions** column.
4. In the **Select Template** step, select a value for Specify Template Import Method. Then, click **Next**.
   - **Enter Template Content**: Enter the content of a JSON template in the **Prepare Template** section.
   - **My Templates**: Select an existing template.
5. In the **Configure Template Parameters** step, modify the parameter settings in the Set Parameter section. Then, click **Next**.

   > **Note** The parameters that you need to configure vary based on the templates. Configure the parameters as instructed in the ROS console.

6. In the **Configure Stack** step, configure Rollback on Failure and Timeout Period based on your business requirements, and click **Next**.
7. In the **Confirm** step, check the template and stack configurations, and click **Create Stack**.

5.1.3.3. Recreate a stack

You can recreate a stack if you want to change your template, stack configurations, and the organization, resource set, and region ID of the stack. This topic describes how to recreate a stack.

Prerequisites

A stack is created. For more information, see Create a stack.

Procedure

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **Stacks**.
3. On the **Stacks** page, find the stack that you want to recreate and click **Re-create** in the **Actions** column.
4. In the **Select Template** step, configure **Organization**, **Resource Set**, and **regionId**.
5. Select a value for Specify Template Import Method. Enter the content of your template in the Prepare Template section. Then, click **Next**.
   - **Enter Template Content**: Enter the content of a JSON template in the **Prepare Template** section.
   - **My Templates**: Select an existing template.
6. In the **Configure Template Parameters** step, modify the parameter settings in the Stacks and Set Parameters sections. Then, click **Next**.

   > Note
   > - The parameters that you need to configure vary based on the templates. Configure the parameters as instructed in the ROS console.
   > - The name must be unique in your Apsara Stack tenant account.

7. In the **Configure Stack** step, configure Rollback on Failure and Timeout Period based on your business requirements and click **Next**.
8. In the **Confirm** step, check the template and stack configurations and click **Create Stack**.

5.1.3.4. Delete a stack

You cannot delete a stack that is in the process of an operation such as create or update.

Prerequisites

A stack is created. For more information, see Create a stack.

Procedure

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **Stacks**.
3. On the **Stacks** page, find the stack that you want to delete and click **Delete** in the **Actions** column.
4. In the dialog box that appears, choose a method to delete the stack.
   - **Retain Resources**: retains the resources in a stack when the stack is deleted.
   - **Release Resources**: releases the resources in a stack when the stack is deleted. Proceed with caution.
5. Click **OK**.

5.1.4. Manage templates

5.1.4.1. Create a template

You can use Resource Orchestration Service (ROS) to create a JSON template, and then use the template to create a stack. This topic describes how to create a template.

Procedure

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **My Templates**.
3. On the **My Templates** page, click **Create Template**.
4. In the **Create Template** dialog box, configure **Template Name**, **Template Description**, **Organization**, **Resource Set**, and **Template Content**.
5. Click **OK**.

5.1.4.2. Edit a template

You can edit the name, description, and content of a template based on your business requirements.

Prerequisites

A template is created. For more information, see Create a template.

Procedure

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **My Templates**.
3. On the **My Templates** page, find the template that you want to edit and click **Edit** in the **Actions** column.
4. In the **Edit Template** dialog box, configure **Template Name**, **Template Description**, and **Template Content**.
5. Click **OK**.

5.1.4.3. Delete a template

You can delete templates that you no longer need.

Prerequisites

A template is created. For more information, see Create a template.

Procedure

1. Log on to the ROS console. For more information, see Log on to the ROS console.
2. In the left-side navigation pane, click **My Templates**.
3. On the **My Templates** page, find the template that you want to delete and click **Delete** in the **Actions** column.
4. In the **Delete** message, click **OK**.

5.1.5. Template syntax

5.1.5.1. Template structure

A template is a UTF-8 encoded JSON file that is used to create stacks. Templates serve as the blueprint for underlying infrastructure and architecture. Templates define the configurations of Apsara Stack resources and dependencies between the resources.

ROS template structure

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Description" : "The description of the template, which is used to provide information such as use scenarios and architecture of the template",
  "Metadata" : {
    // The template metadata that provides information such as the layout for visualizations.
  },
  "Parameters" : {
    // The parameters that you can specify when you create a stack.
  },
  "Mappings" : {
    // The mapping tables. Mapping tables are nested tables.
  },
  "Conditions": {
    // The conditions defined by using internal condition functions. These conditions determine when to create associated resources.
  },
  "Resources" : {
    // The detailed information of resources such as configurations and dependencies.
  },
  "Outputs" : {
    // The outputs that are used to provide information such as resource properties. You can use the Resource Orchestration Service (ROS) console or API to obtain the information.
  }
}
```

ROSTemplateFormatVersion

Required. The template versions supported by ROS. Current version: 2015-09-01.

Description

Optional. The description of the template, which is used to provide information such as use scenarios and architecture of the template.

A detailed description can help users better understand the content of the template.

Metadata

Optional. The metadata of the template, in the JSON format.

Parameters

Optional. The parameters that you can specify when you create a stack. An Elastic Compute Service (ECS) instance type is often defined as a parameter. Parameters have default values. Parameters can improve the flexibility and reusability of the template. When you create a stack, select appropriate specifications.

Mappings

Optional. Mappings are defined as nested mapping tables. You can use Fn::FindInMap to retrieve values corresponding to keys. You can also use parameter values as keys. For example, you can search the region-image mapping table for desired images by region.

Conditions

Optional. The conditions defined by using Fn::And, Fn::Or, Fn::Not, and Fn::Equals. Multiple conditions are separated by commas (,). The system evaluates all conditions in the template before it creates or updates a stack. All resources associated with `true` conditions are created, and all resources associated with `false` conditions are ignored.

Resources

Optional. The detailed information of resources in the stack created based on the template. The information includes resource dependencies and configurations.

Outputs

Optional. The outputs that are used to provide information such as resource properties. You can use the ROS console or API to obtain the information.

5.1.5.2. Parameters

When you create a template, you can use the Parameters section to improve the flexibility and reusability of the template. When you create a stack, you can replace parameter values in the template.

For example, you have a web application requiring a stack that contains one Server Load Balancer (SLB) instance, two Elastic Compute Service (ECS) instances, and one ApsaraDB RDS instance. If the web application has a heavy workload, you can select ECS instances with advanced specifications when you create the stack. Otherwise, you can select ECS instances with basic specifications. The following example shows how to define the InstanceType parameter for an ECS instance:

```
"Parameters" : {
  "InstanceType" : {
    "Type" : "String",
    "AllowedValues":["ecs.t1.small","ecs.s1.medium", "ecs.m1.medium", "ecs.c1.large"],
    "Default": "ecs.t1.small",
    "Label": "ECS instance type",
    "Description" : "The type of the ECS instance that you want to create. Default value: ecs.t1.small
. Valid values: ecs.t1.small, ecs.s1.medium, ecs.m1.medium, and ecs.c1.large."
  }
}
```

You can assign a value to the InstanceType parameter when you create stacks based on templates. If this parameter is not specified, the default value `ecs.t1.small` is used.

The following example shows how to reference the InstanceType parameter when you define a resource:

```
"Webserver" : {
  "Type" : "ALIYUN::ECS::Instance",
  "InstanceType": {
    "Ref": "InstanceType"
  }
}
```

Syntax

Each parameter consists of a name and properties. The parameter name can contain only letters and digits, and must be unique within the template. You can use the `Label` field to define the alias of a parameter.

The following table describes the parameter properties.

Parameter property	Required	Description

Parameter property	Required	Description
Type	Yes	The data type of the parameter. • String: a string value. Example: `"ecs.s1.medium"`. • Number: an integer or a floating-point number. Example: 3.14. • CommaDelimitedList: a set of strings separated by commas (,), which can be indexed by using the Fn::Select function. Example: `"80, foo, bar"`. • Json: a JSON string. Example: `{ "foo": "bar" }`. • Boolean: a Boolean value. Example: `true` or `false`.
Default	No	The default value of the parameter. If you do not specify a value when you create a stack, Resource Orchestration Service (ROS) checks whether a default value is defined in the template. If a default value is found, ROS uses the default value. Otherwise, an error is returned.
AllowedValues	No	The list of one or more valid parameter values.
AllowedPattern	No	The regular expression that is used to check whether the specified parameter value is a string. If the input is not a string, an error is returned.
MaxLength	No	The integer value that determines the longest string allowed for a String-type parameter.
MinLength	No	The integer value that determines the shortest string allowed for a String-type parameter.
MaxValue	No	The numeric value that determines the maximum value allowed for a Number-type parameter.
MinValue	No	The numeric value that determines the minimum value allowed for a Number-type parameter.
NoEcho	No	Specifies whether to mask the parameter value when the GetStack operation is called. If you set this property to `true`, only asterisks (*) are returned.
Description	No	The string that describes the parameter.
ConstraintDescription	No	The description of the parameter constraints.
Label	No	The alias of the parameter, encoded in UTF-8. When you create a web form based on a template, the `Label` value can be mapped to the parameter name.

Parameter property	Required	Description
AssociationProperty	No	The associated resource property. If you specify this parameter property, ROS verifies whether the specified parameter value is valid and provides a list of valid values based on the associated resource property. Valid values: • `ALIYUN::ECS::Instance::ImageId` • `ALIYUN::ECS::Instance::ZoneId` • `ALIYUN::ECS::VPC::VPCId` • `ALIYUN::ECS::VSwitch::VSwitchId` For example, if you set `AssociationProperty` to `ALIYUN::ECS::Instance::ImageId`, ROS verifies whether the specified image ID is valid and lists other valid values in a drop-down list.
Confirm	No	Specifies whether to enter the parameter value for a second time if the NoEcho property is set to `true`. Default value: `false`. **Notice** The Confirm property can be set to `true` only when it is used with a String-type parameter and when the NoEcho parameter is set to `true`.

Examples

In the following example, two parameters are defined in the Parameters section:

- `username`
  - Type: String
  - Valid values:
    - `anonymous`
    - `user-one`
    - `user-two`
  - Length: 6 to 12 characters

  Notice The default value `anonymous` must also meet the length and valid value requirements.
- `password`
  - Type: String
  - Length: 1 to 41 characters
  - The password can contain letters and digits.
  - If you set the NoEcho property to `true`, the GetStack operation does not return parameter values.

```
"Parameters": {
    "username": {
        "Label": "Username"
        "Description": "Enter the username"
        "Default": "anonymous",
        "Type": "String",
        "MinLength": "6",
        "MaxLength": "12",
        "AllowedValues": ["anonymous", "user-one", "user-two"]
    },
    "password": {
        "Label": "Password"
        "NoEcho": "True",
        "Description": "Enter the password"
        "Type": "String",
        "MinLength": "1",
        "MaxLength": "41",
        "AllowedPattern": "[a-zA-Z0-9]*"
    }
}
```

Pseudo parameters

Pseudo parameters are internal parameters provided by the ROS engine. They can be referenced in the same manner as user-defined parameters, and their values are determined when ROS is running. The following pseudo parameters are supported:

- `ALIYUN::StackName` : the name of the stack.
- `ALIYUN::StackId` : the ID of the stack.
- `ALIYUN::Region` : the region where the stack resides.
- A `LIYUN::AccountId` : the account ID of the stack.
- `ALIYUN::NoValue` : specifies whether the specific resource property is deleted when the resource is created or updated.

5.1.5.3. Resources

This topic describes the properties of each resource and dependencies between resources in a stack. A resource can be referenced by other resources and output items.

Syntax

Each resource consists of an ID and a description. All resource descriptions are enclosed in braces {}. Multiple resources are separated by commas (,). The following sample code shows the Resources syntax:

```
"Resources" : {
    "Resource1 ID" : {
        "Type" : "The resource type",
        "Condition": "The condition that specifies whether to create the resource",
        "Properties" : {
            The description of the resource properties
        }
    },
    "Resource2 ID" : {
        "Type" : "The resource type",
        "Condition": "The condition that specifies whether to create the resource",
        "Properties" : {
            The description of the resource properties
        }
    }
}
```

Parameter description:

- The resource ID must be unique within the template. You can use the resource ID to reference the resource in other parts of the template.
- The Type parameter specifies the type of the resource that is being declared. For example, ALIYUN::ECS::Instance indicates that the resource is an Elastic Cloud Service (ECS) instance.
- The Properties section provides additional options that you can specify for a resource. For example, you must specify an image ID for each ECS instance. The image ID is one of the resource properties.

Examples

```
"Resources" : {
  "ECSInstance" : {
    "Type" : "ALIYUN::ECS::Instance",
    "Properties" : {
      "ImageId" : "m-25l0r****"
    }
  }
}
```

If a resource does not need properties to be declared, omit the Properties section of that resource.

Property values can be text strings, string lists, Boolean values, referenced parameters, or return values of functions.

The following example shows how to declare different types of property values:

```
"Properties" : {
    "String" : "string",
    "LiteralList" : [ "value1", "value2" ],
    "Boolean" : "true"
    "ReferenceForOneValue" :  { "Ref" : "ResourceID" } ,
    "FunctionResultWithFunctionParams" : {
        "Fn::Join" : [ "%", [ "Key=", { "Ref" : "SomeParameter" } ] ] }
}
```

DeletionPolicy

The DeletionPolicy parameter specifies whether to retain a resource when its stack is deleted. The following sample code shows how to use the DeletionPolicy parameter to retain an ECS instance when its stack is deleted:

```
"Resources" : {
  "ECSInstance" : {
    "Type" : "ALIYUN::ECS::Instance",
    "Properties" : {
      "ImageId" : "m-25l0r****"
    },
    "DeletionPolicy" : "Retain"
  }
}
```

DependsOn

The DependsOn parameter allows you to create a specific resource after you create its dependent resource. If you specify the DependsOn parameter for a resource, the resource is created only after its dependent resource specified by the DependsOn parameter is created.

In the following example, WebServer is created only after DatabaseServer is created:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "DependsOn": "DatabseServer"
    },
    "DatabseServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId" : "m-25l0r****",
        "InstanceType": "ecs.t1.small"
      }
    }
  }
}
```

Condition

The Condition parameter specifies whether to create the resource. The resource can be created only when the Condition parameter is set to true.

In the following example, WebServer is created only if the condition determined by the MaxAmount parameter is true:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Parameters": {
    "MaxAmount": {
      "Type": "Number",
      "Default": 1
    }
  },
  "Conditions": {
    "CreateWebServer": {"Fn::Not": {"Fn::Equals": [0, {"Ref": "MaxAmount"}]}}
  }
  "Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::InstanceGroup",
      "Condition": "CreateWebServer",
      "Properties": {
        "ImageId" : "m-25l0r****",
        "InstanceType": "ecs.t1.small"
        "MaxAmount": {"Ref": "MaxAmount"}
      }
    },
    "DatabseServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId" : "m-25l0r****",
        "InstanceType": "ecs.t1.small"
      }
    }
  }
}
```

Resource declaration example

The following example shows how to declare a resource:

```
"Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId" : "m-25l0r****",
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "Tags": [{
            "Key": "Department1",
            "Value": "HumanResource"
        },{
            "Key": "Department2",
            "Value": "Finance"
        }
        ]
      }
    },
    "ScalingConfiguration": {
      "Type": "ALIYUN::ESS::ScalingConfiguration",
      "Properties": {
        "ImageId": "ubuntu1404_64_20G_aliaegis_2015****.vhd",
        "InstanceType": "ecs.t1.small",
        "InstanceId": "i-25xhh****",
        "InternetChargeType": "PayByTraffic",
        "InternetMaxBandwidthIn": 1,
        "InternetMaxBandwidthOut": 20,
        "SystemDisk_Category": "cloud",
        "ScalingGroupId": "bwhtvpcBcKYac9fe3vd0****",
        "SecurityGroupId": "sg-25zwc****",
        "DiskMappings": [
            {
                "Size": 10
            },
            {
                "Category": "cloud",
                "Size": 10
            }
        ]
      }
    }
}
```

5.1.5.4. Outputs

The Outputs section is used to define the values returned when the GetStack operation is called. For example, if you define an Elastic Compute Service (ECS) instance ID as an output item, the ECS instance ID is returned when the GetStack operation is called.

Syntax

Each output item consists of an ID and a description. All output descriptions are enclosed in braces {}. Multiple output items are separated by commas (,). Each output item can have multiple values in an array format. The following example shows the Outputs syntax:

```
"Outputs" : {
  "Output1 ID" : {
    "Description": "The description of the output item",
    "Condition": "The condition that specifies whether to provide resource properties",
    "Value": "The output value expression"
  },
  "Output2 ID" : {
    "Description": "The description of the output item",
    "Condition": "The condition that specifies whether to provide resource properties",
    "Value" : [
      "Output value expression 1",
      "Output value expression 2",
      ...
    ]
  }
}
```

- Output ID: the ID of the output item. Duplicate IDs are not allowed within a template.
- Description: optional. The description of the output item.
- Value: required. The value returned when the GetStack operation is called.
- Condition: optional. The condition that specifies whether to create a resource and provide its information. The resource is created and its information is provided only when the specified condition is `true`.

  In the following example, WebServer is created only if the condition determined by the MaxAmount parameter is true:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Parameters": {
    "MaxAmount": {
      "Type": "Number",
      "Default": 1
    }
  },
  "Conditions": {
    "CreateWebServer": {"Fn::Not": {"Fn::Equals": [0, {"Ref": "MaxAmount"}]}}
  }
  "Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::InstanceGroup",
      "Condition": "CreateWebServer",
      "Properties": {
        "ImageId" : "m-2510r****",
        "InstanceType": "ecs.t1.small"
        "MaxAmount": {"Ref": "MaxAmount"}
      }
    }
  }
  "Outputs": {
    "WebServerIP": {
      "Condition": "CreateWebServer",
      "Value": {
        "Fn::GetAtt": ["WebServer", "PublicIps"]
      }
    }
  }
}
```

Examples

The following example contains two output items:

- The InstanceId value of WebServer
- The PublicIp and PrivateIp values of WebServer

```
"Outputs": {
  "InstanceId": {
    "Value" : {"Fn::GetAtt": ["WebServer", "InstanceId"]}
  },
  "PublicIp & PrivateIp": {
    "Value" : [
      {"Fn::GetAtt": ["WebServer", "PublicIp"]},
      {"Fn::GetAtt": ["WebServer", "PrivateIp"]}
    ]
  }
}
```

5.1.5.5. Functions

Resource Orchestration Service (ROS) provides several built-in functions to help you manage stacks. You can use built-in functions to define Resources and Outputs.

Fn::Base64Encode

The Fn::Base64Encode function is used to return the Base64 representation of the input string.

- Declaration

  ```
  "Fn::Base64Encode": "stringToEncode"
  ```

- Parameters

  `stringToEncode`: the string to be encoded in Base64.

- Return value

  The Base64 representation of the input string.

- Examples

  ```
  {"Fn::Base64Encode": "string to encode"}
  ```

  `c3RyaW5nIHRvIGVuY29kZQ==` is returned in this example.

Fn::Base64Decode

The Fn::Base64Decode function is used to return a string decoded from a Base64-encoded string.

- Declaration

  ```
  {"Fn::Base64Decode": "stringToEncode"}
  ```

- Parameters

  `stringToDecode`: the string decoded from the Base64-encoded string.

- Return value

  The string decoded from the Base64-encoded string.

- Examples

  ```
  {"Fn::Base64Decode": "c3RyaW5nIHRvIGVuY29kZQ=="}
  ```

`string to encode` is returned in this example.

Fn::Base64

The Fn::Base64 function returns the Base64 representation of the input string.

- Declaration

  ```
  "Fn::Base64": stringToEncode
  ```

- Parameters

  valueToEncode: the string to be encoded in Base64.

- Return value

  The Base64 representation of the input string.

- Examples

  ```
  "Fn::Base64":"string to encode"
  ```

Fn::FindInMap

The Fn::FindInMap function is used to return the values based on keys in a two-level mapping that is declared in the Mappings section.

- Declaration

  ```
  "Fn::FindInMap": ["MapName", "TopLevelKey", "SecondLevelKey"]
  ```

- Parameters
  - `MapName`: the ID of a mapping declared in the Mappings section that contains keys and values.
  - `TopLevelKey`: the top-level key name. The value is a list of key-value pairs.
  - `SecondLevelKey`: the second-level key name. The value is a string or a number.

- Return value

  The value that is assigned to the SecondLevelKey parameter.

- Examples

  The ImageId property must be specified when you create a WebServer instance. The Mappings section describes the ImageId mappings by region. The Parameters section describes the regions that must be specified by template users. Fn::FindInMap finds the corresponding ImageId mapping in RegionMap based on the region specified by a user, and then finds the corresponding ImageId values in the mapping.
  - MapName can be set to a custom value, which is `RegionMap` in this example.
  - TopLevelKey is set to the region where the stack is created, which is `{ "Ref" : "regionParam" }` in this example.
  - SecondLevelKey is set to the required architecture, which is `32` in this example.

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "regionParam": {
      "Description": "The region where the ECS instance is created",
      "Type": "String",
      "AllowedValues": [
        "hangzhou",
        "beijing"
      ]
    }
  },
  "Mappings": {
    "RegionMap": {
      "hangzhou": {
        "32": "m-25l0rcfjo",
        "64": "m-25l0rcfj1"
      },
      "beijing": {
        "32": "m-25l0rcfj2",
        "64": "m-25l0rcfj3"
      }
    }
  },
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId": {
          "Fn::FindInMap": [
            "RegionMap",
            {"Ref": "regionParam"},
            "32"
          ]
        },
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "Tags": [
          {
            "Key": "key1",
            "Value": "value1"
          },
          {
            "Key": "key2",
            "Value": "value2"
          }
        ]
      }
    }
  }
}
```

- Supported functions
  - Fn::FindInMap
  - Ref

Fn::GetAtt

The Fn::GetAtt function is used to return the value of a property from a resource in a template.

- Declaration

```
"Fn::GetAtt": ["resourceID", "attributeName"]
```

- Parameters
  - `resourceID`: the ID of the resource.
  - `attributeName`: the name of the resource property.
- Return value

  The value of the resource property.
- Examples

  The ImageId property of MyEcsInstance is returned in this example.

```
{"Fn::GetAtt" : ["MyEcsInstance" , "ImageID"]}
```

Fn::Join

The Fn::Join function is used to combine a set of values into a single value that is separated by a specified delimiter.

- Declaration

```
{"Fn::Join": ["delimiter", ["string1", "string2", ... ]]}
```

- Parameters
  - `delimiter`: the value used to divide the string. The delimiter value can be left blank so that all the values are directly combined.
  - `[ "string1", "string2", ... ]`: the list of values that are combined into a string.
- Return value

  The combined string.
- Examples

```
{"Fn::Join": [ ",", ["a", "b", "c"]]}
```

  `"a,b,c"` is returned in this example.
- Supported functions
  - Fn::Base64Encode
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Ref

Fn::Select

The Fn::Select function is used to return a single data element from a list of data elements by using an index.

- Declaration
  - The following example assumes that the list of data elements is an array:

```
"Fn::Select": ["index", ["value1", "value2", ... ]]
```

- The following example assumes that the list of data elements is a mapping table:

```
"Fn::Select": ["index", {"key1": "value1", ... }]
```

- Parameters

  `index`: the index of the object data element. If the list of data elements is an array, the index must be an integer ranging from 0 to N-1, where N indicates the number of elements in the array. If the list of data elements is a mapping table, the index must be a key in the mapping table.

  If the corresponding value of the index cannot be found, the system returns an empty string.

- Return value

  The object data element.

- Examples
  - The following example assumes that the list of data elements is an array:

```
{"Fn::Select": ["1", ["apples", "grapes", "oranges", "mangoes"]]}
```

    `"grapes"` is returned in this example.
  - The following example assumes that the list of data elements is a mapping table:

```
{"Fn::Select": ["key1", {"key1": "grapes", "key2": "mangoes"}]}
```

    `"grapes"` is returned in this example.
  - The following example assumes that the list of data elements is a comma-delimited list:

```
"Parameters": {
  "userParam": {
    "Type": "CommaDelimitedList",
      "Default": "10.0.100.0/24, 10.0.101.0/24, 10.0.102.0/24"
  }
}
"Resources": {
  "resourceID": {
      "Properties": {
          "CidrBlock": {"Fn::Select": ["0", {"Ref": "userParam"}]}
    }
  }
}
```

- Supported functions
  - For the Fn::Select index value, you can use the Ref function.
  - For the Fn::Select list of data elements, you can use the following functions:
    - Fn::Base64Encode
    - Fn::FindInMap
    - Fn::GetAtt
    - Fn::Join
    - Fn::Select
    - Ref

Ref

The Ref function is used to return the value of a specified parameter or resource.

If the specified parameter is a resource ID, the value of the resource is returned. Otherwise, the system returns the value of the specified parameter.

- Declaration

```
"Ref": "logicalName"
```

- Parameters

  `logicalName`: the logical name of the resource or parameter that you want to reference.

- Return value

  The value of the resource or parameter.

- Examples

  The following example demonstrates how to use the Ref function to specify regionParam as the region parameter for RegionMap of WebServer:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "regionParam": {
      "Description": "The region where the ECS instance is created",
      "Type": "String",
      "AllowedValues": [
        "hangzhou",
        "beijing"
      ]
    }
  },
  "Mappings": {
    "RegionMap": {
      "hangzhou": {
        "32": "m-25l0rcfjo",
        "64": "m-25l0rcfj1"
      },
      "beijing": {
        "32": "m-25l0rcfj2",
        "64": "m-25l0rcfj3"
      }
    }
  },
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId": {
          "Fn::FindInMap": [
            "RegionMap",
            {"Ref": "regionParam"},
            "32"
          ]
        },
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "Tags": [
          {
            "Key": "tiantt",
            "Value": "ros"
          },
          {
            "Key": "tiantt1",
            "Value": "ros1"
          }
        ]
      }
    }
  }
}
```

- Supported function

  When you use the Ref function, you cannot use other functions in it at the same time. You must specify a string value for the logical ID of a resource.

Fn::GetAZs

The Fn::GetAZs function is used to return a list of zones for a specified region.

- Declaration

```
"Fn::GetAZs": "region"
```

- Parameters

  `region`: the ID of the region.

- Return value

  The list of zones within the specified region.

- Examples

  The following example demonstrates how to create an Elastic Compute Service (ECS) instance in the first zone of a specified region:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId": "centos7u2_64_40G_cloudinit_2016****.raw",
        "InstanceType": "ecs.n1.tiny",
        "SecurityGroupId": "sg-2zedcm7ep5quses0****",
        "Password": "Ros1****",
        "AllocatePublicIP": true,
        "InternetChargeType": "PayByTraffic",
        "InternetMaxBandwidthIn": 100,
        "InternetMaxBandwidthOut": 100,
        "SystemDiskCategory": "cloud_efficiency",
        "IoOptimized": "optimized",
        "ZoneId": {"Fn::Select": ["0", {"Fn::GetAZs": {"Ref": "ALIYUN::Region"}}]}
      }
    }
  },
  "Outputs": {
    "InstanceId": {
         "Value": {"Fn::GetAtt": ["WebServer","InstanceId"]}
    },
    "PublicIp": {
         "Value": {"Fn::GetAtt": ["WebServer","PublicIp"]}
    }
  }
}
```

- Supported functions
  - Fn::Base64Encode
  - Fn::FindInMap
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Ref

Fn::Replace

The Fn::Replace function is used to replace a specified substring contained in a string with a new substring.

- Declaration

```
{"Fn::Replace": [{"object_key": "object_value"}, "object_string"]}
```

- Parameters
  - `object_key`: the substring to be replaced.
  - `object_value`: the new substring to replace the previous substring.
  - `object_string`: the string whose `object_key` is replaced.
- Return value

  The string after replacement.
- Examples

  The following example demonstrates how to replace "print" with "echo" in the specified script:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId" : "centos_7_2_64_40G_base_2017****.vhd",
        "InstanceType": "ecs.n1.medium",
        "SecurityGroupId": "sg-94q49****",
        "Password": "MytestPassword****",
        "IoOptimized": "optimized",
        "VSwitchId": "vsw-94vdv8****",
        "VpcId": "vpc-949uz****",
        "SystemDiskCategory": "cloud_ssd",
        "UserData": {"Fn::Replace": [{"print": "echo"},
         {"Fn::Join": ["", [
             "#!/bin/sh\n",
             "mkdir ~/test_ros\n",
             "print hello > ~/1.txt\n"
         ]]}]}
      }
    }
  },
  "Outputs": {
    "InstanceId": {
         "Value" : {"Fn::GetAtt": ["WebServer","InstanceId"]}
    },
    "PublicIp": {
         "Value" : {"Fn::GetAtt": ["WebServer","PublicIp"]}
    }
  }
}
```

- Supported functions
  - Fn::Base64Encode
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Ref

Fn::Split

The Fn::Split function is used to split a string into a list of values separated by a specified delimiter and return the list.

- Declaration

```
"Fn::Split": ["delim",  "original_string"]
```

- Parameters
  - `delim`: the specified delimiter, which can be a comma (,), semicolon (;), line break (\n), or indent (\t).
  - `original_string`: the string to be split.
- Return value

  A list of string values.
- Examples
  - The following example assumes that the list of data elements is an array:

```
{"Fn::Split": [";", "foo; bar; achoo"]}
```

    `["foo", " bar", "achoo "]` is returned in this example.
  - The following example demonstrates how to use Fn::Split to split InstanceIds:

```
{
  "Parameters": {
    "InstanceIds": {
      "Type": "String",
      "Default": "instane1_id,instance2_id,instance2_id"
    }
  },
  "Resources": {
    "resourceID": {
      "Type": "ALIYUN::SLB::BackendServerAttachment",
      "Properties": {
        "BackendServerList": {
          "Fn::Split": [
            ",",
            {
              "Ref": "InstanceIds"
            }
          ]
        }
      }
    }
  }
}
```

- Supported functions
  - Fn::Base64Encode
  - Fn::FindInMap
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Fn::Replace
  - Fn::GetAZs
  - Fn::If
  - Ref

Fn::Equals

The Fn::Equals function is used to compare whether two values are equal. If the two values are equal, true is returned. If the two values are not equal, false is returned.

- Declaration

```
{"Fn::Equals": ["value_1", "value_2"]}
```

- Parameters

  `value`: the values to be compared.

- Return value

  true or false.

- Examples

  The following example demonstrates how to use Fn::Equals to define a condition in the Conditions section:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "EnvType": {
      "Default": "pre",
      "Type": "String"
    }
  },
  "Conditions": {
    "TestEqualsCond": {
      "Fn::Equals": [
        "prod",
        {"Ref": "EnvType"}
      ]
    }
  }
}
```

- Supported functions
  - Fn::Or
  - Fn::Not
  - Fn::Equals
  - Fn::FindInMap
  - Fn::And
  - Ref

Fn::And

The Fn::And function is used to represent the AND operator, and must contain at least two conditions. If all the specified conditions are evaluated as true, true is returned. If any condition is evaluated as false, false is returned.

- Declaration

```
{"Fn::And": ["condition", {...}]}
```

- Parameters

  `condition`: the condition to be evaluated.

- Return value

  true or false.

- Examples

The following example demonstrates how to use Fn::And to define a condition in the Conditions section:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Parameters":{
    "EnvType":{
      "Default":"pre",
      "Type":"String"
    }
  },
  "Conditions": {
    "TestEqualsCond": {"Fn::Equals": ["prod", {"Ref": "EnvType"}]},
    "TestAndCond": {"Fn::And": ["TestEqualsCond", {"Fn::Equals": ["pre", {"Ref": "EnvType"}]}]}
  }
}
```

- Supported functions
  - Fn::Or
  - Fn::Not
  - Fn::Equals
  - Fn::FindInMap
  - Fn::And
  - Ref

Fn::Or

The Fn::Or function is used to represent the OR operator, and must contain at least two conditions. If any specified condition is evaluated as true, true is returned. If all the conditions are evaluated as false, false is returned.

- Declaration

```
{"Fn::Or": ["condition", {...}]}
```

- Parameters

  `condition`: the condition to be evaluated.

- Return value

  true or false.

- Examples

  The following example demonstrates how to use Fn::Or to define a condition in the Conditions section:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Parameters":{
    "EnvType":{
      "Default":"pre",
      "Type":"String"
    }
  },
  "Conditions": {
    "TestEqualsCond": {"Fn::Equals": ["prod", {"Ref": "EnvType"}]},
    "TestOrCond": {"Fn::And": ["TestEqualsCond", {"Fn::Equals": ["pre", {"Ref": "EnvType"}]}]}
  }
}
```

- Supported functions
  - Fn::Or

- Fn::Not
- Fn::Equals
- Fn::FindInMap
- Fn::And
- Ref

Fn::Not

The Fn::Not function is used to represent the NOT operator. If a condition is evaluated as false, true is returned. If a condition is evaluated as true, false is returned.

- Declaration

```
{"Fn::Not": "condition"}
```

- Parameters

  `condition`: the condition to be evaluated.

- Return value

  true or false.

- Examples

  The following example demonstrates how to use Fn::Not to define a condition in the Conditions section:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Parameters":{
    "EnvType":{
      "Default":"pre",
      "Type":"String"
    }
  },
  "Conditions": {
    "TestNotCond": {"Fn::Not": {"Fn::Equals": ["pre", {"Ref": "EnvType"}]}}
  }
}
```

- Supported functions
  - Fn::Or
  - Fn::Not
  - Fn::Equals
  - Fn::FindInMap
  - Fn::And
  - Ref

Fn::If

This function returns one of two possible values. If a specified condition is evaluated as true, one value is returned. If the specified condition is evaluated as false, the other value is returned. The property values of Resources and Outputs in templates support the Fn::If function. You can use the `ALIYUN::NoValue` pseudo parameter as the return value to delete the corresponding property.

- Declaration

```
{"Fn::If": ["condition_name", "value_if_true", "value_if_false"]}
```

- Parameters

- `condition_name`: the name of the condition in the Conditions section. A condition is referenced by using the condition name.
- `value_if_true`: If the specified condition is evaluated as true, this value is returned.
- `value_if_false`: If the specified condition is evaluated as false, this value is returned.

- Examples

  The following example demonstrates how to determine whether to create a data disk based on input parameters:

```
{
    "ROSTemplateFormatVersion":"2015-09-01",
    "Parameters":{
        "EnvType":{
            "Default":"pre",
            "Type":"String"
        }
    },
    "Conditions":{
        "CreateDisk":{
            "Fn::Equals":[
                "prod",
                {
                    "Ref":"EnvType"
                }
            ]
        }
    },
    "Resources":{
        "WebServer":{
            "Type":"ALIYUN::ECS::Instance",
            "Properties":{
                "DiskMappings":{
                    "Fn::If":[
                        "CreateDisk",
                        [
                            {
                                "Category":"cloud_efficiency",
                                "DiskName":"FirstDataDiskName",
                                "Size":40
                            },
                            {
                                "Category":"cloud_ssd",
                                "DiskName":"SecondDataDiskName",
                                "Size":40
                            }
                        ],
                        {
                            "Ref":"ALIYUN::NoValue"
                        }
                    ]
                },
                "VpcId":"vpc-2zew9pxh2yirtzqxd****",
                "SystemDiskCategory":"cloud_efficiency",
                "SecurityGroupId":"sg-2zece6wcqriejf1v****",
                "SystemDiskSize":40,
                "ImageId":"centos_6_8_64_40G_base_2017****.vhd",
                "IoOptimized":"optimized",
                "VSwitchId":"vsw-2zed9txvy7h2srqo6****",
                "InstanceType":"ecs.n1.medium"
```

```
            }
        }
    },
    "Outputs":{
        "InstanceId":{
            "Value":{
                "Fn::GetAtt":[
                    "WebServer",
                    "InstanceId"
                ]
            }
        },
        "ZoneId":{
            "Value":{
                "Fn::GetAtt":[
                    "WebServer",
                    "ZoneId"
                ]
            }
        }
    }
}
```

- Supported functions
  - Fn::Or
  - Fn::Not
  - Fn::Equals
  - Fn::FindInMap
  - Fn::And
  - Ref

Fn::ListMerge

The Fn::ListMerge function is used to merge multiple lists into one list.

- Declaration

```
{"Fn::ListMerge": [["list_1_item_1", "list_1_imte_2", ...], ["list_2_item_1", "list_2_imte_2", ...]
]}
```

- Parameters
  - `["list_1_item_1", "list_1_imte_2", ...]`: the first list to merge.
  - `["list_2_item_1", "list_2_imte_2", ...]`: the second list to merge into the first list.
- Examples

  The following example demonstrates how to attach two ECS instance groups to a Server Load Balancer (SLB) instance:

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "LoadBalancer": {
      "Type": "ALIYUN::SLB::LoadBalancer",
      "Properties": {
        "LoadBalancerName": "ros",
        "AddressType": "internet",
        "InternetChargeType": "paybybandwidth",
      }
    },
    "BackendServer1": {
      "Type": "ALIYUN::ECS::InstanceGroup",
      "Properties": {
        "ImageId" : "m-2ze9uqi7wo61hwep****",
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-2ze8yxgempcdsq3****",
        "MaxAmount": 1,
        "MinAmount": 1
      }
    },
    "BackendServer2": {
      "Type": "ALIYUN::ECS::InstanceGroup",
      "Properties": {
        "ImageId" : "m-2ze9uqi7wo61hwep****",
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-2ze8yxgempcdsq3iu****",
        "MaxAmount": 1,
        "MinAmount": 1
      }
    },
    "Attachment": {
      "Type": "ALIYUN::SLB::BackendServerAttachment",
      "Properties": {
        "LoadBalancerId": {"Ref": "LoadBalancer"},
        "BackendServerList": { "Fn::ListMerge": [
            {"Fn::GetAtt": ["BackendServer1", "InstanceIds"]},
            {"Fn::GetAtt": ["BackendServer2", "InstanceIds"]}
          ]
        }
      }
    }
  }
}
```

- Supported functions
  - Fn::Base64Encode
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Ref
  - Fn::If

Fn::GetJsonValue

The Fn::GetJsonValue function is used to resolve a JSON string and obtain its key value from the first layer.

- Declaration

```
{"Fn::GetJsonValue": ["key", "json_string"]}
```

- Parameters
  - `key`: the key value.
  - `json_string`: the specified JSON string to be resolved.
- Examples

  In the following example, the WebServer instance executes UserData and returns a JSON string, and the WebServer2 instance then obtains the corresponding key value from the string.

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId" : "m-2ze45uwova5fedlu****",
        "InstanceType": "ecs.n1.medium",
        "SecurityGroupId": "sg-2ze7pxymaix640qr****",
        "Password": "Wenqiao****",
        "IoOptimized": "optimized",
        "VSwitchId": "vsw-2zei67xd9nhcqxze****",
        "VpcId": "vpc-2zevx9ios1rszqv0a****",
        "SystemDiskCategory": "cloud_ssd",
        "UserData": {"Fn::Join": ["", [
             "#!/bin/sh\n",
             "mkdir ~/test_ros\n",
             "print hello > ~/1.txt\n",
             "Fn::GetAtt": ["WaitConHandle", "CurlCli"],
             "\n",
             "Fn::GetAtt": ["WaitConHandle", "CurlCli"],
             " -d '{\\"id\\" : \\"1\\", \\"data\\": [\\"1111\\", \\"2222\\"]}'\n"
         ]]},
         "PrivateIpAddress": "192.168.XX.XX",
         "HostName": "userdata-1
      }
    },
    "WaitConHandle": {
        "Type": "ALIYUN::ROS::WaitConditionHandle"
    },
    "WaitCondition": {
        "Type": "ALIYUN::ROS::WaitCondition",
        "Properties": {
          "Handle": {"Ref": "WaitConHandle"},
          "Timeout": 900
        }
    },
    "WebServer2": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId" : "m-2ze45uwova5fedlu****",
        "InstanceType": "ecs.n1.medium",
        "SecurityGroupId": "sg-2ze7pxymaix640qr****",
        "Password": "Wenqiao****,
        "IoOptimized": "optimized",
        "VSwitchId": "vsw-2zei67xd9nhcqxzec****",
        "VpcId": "vpc-2zevx9ios1rszqv0a****",
        "SystemDiskCategory": "cloud_ssd",
        "UserData":
```

```
          {"Fn::Join": ["", [
             "#!/bin/sh\n",
             "mkdir ~/test_ros\n",
             "echo hello > ~/1.txt\n",
             "server_1_token=",
             {"Fn::GetJsonValue": ["1", { "Fn::GetAtt": ["WaitCondition", "Data"]}]},
             "\n"
          ]]},
          "PrivateIpAddress": "192.168.XX.XX",
          "HostName": "userdata-2"
       }
     },
   },
   "Outputs": {
     "InstanceId": {
          "Value" : {"Fn::GetAtt": ["WebServer","InstanceId"]}
     },
     "PublicIp": {
          "Value" : {"Fn::GetAtt": ["WebServer","PublicIp"]}
     }
   }
}
```

- Supported functions
  - Fn::Base64Encode
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Ref
  - Fn::If

Fn::MergeMapToList

The Fn::MergeMapToList function is used to merge multiple mappings into a list of mapping elements.

- Declaration

```
{"Fn::MergeMapToList": [{"key_1": ["key_1_item_1", "key_1_item_2", ...]}, {"key_2":["key_2_item_1",
"key_2_item_2", ...]}, ... ]}
```

- Parameters
  - `{"key_1": ["key_1_item_1", "key_1_item_2", ...]}` : the first mapping to merge. The `"key_1"` value must be a list. `"key_1"` is the key for each mapping in the list of merged mappings. The "key_1" value is `"key_1_item_1"` for the first merged mapping and `"key_1_item_2"` for the second merged mapping. All values follow the same format. The length of the final list of merged mappings is the length of the longest list `"key_x"` from all mappings being merged. If a `"key_y"` list is shorter, the last element of the list is repeated until the list is the longest.
  - `{"key_2": ["key_2_item_1", "key_2_item_2", ...]}` : the second mapping to merge into the first mapping. The `"key_2"` value must be a list. `"key_2"` is the key for each mapping in the merged list. The "key_2" value is `"key_2_item_1"` for the first merged mapping and `"key_2_item_2"` for the second merged mapping. All values follow the same format.
- Examples

- The following example demonstrates how to merge three mappings. The length of the list based on the key values for each mapping is the same.

```
{
   "Fn::MergeMapToList": [
      {"key_1": ["kye_1_item_1", "kye_1_item_2"]},
      {"key_2": ["kye_2_item_1", "kye_2_item_2"]},
      {"key_3": ["kye_3_item_1", "kye_3_item_2"]}
   ]
}
```

The following code shows the merged result:

```
[
   {
      "key_1": "kye_1_item_1",
      "key_2": "kye_2_item_1",
      "key_3": "kye_3_item_1"
    },
    {
      "key_1": "kye_1_item_2",
      "key_2": "kye_2_item_2",
      "key_3": "kye_3_item_2"
    }
]
```

- The length of the list based on the key values for each mapping varies in the following example:

```
{
   "Fn::MergeMapToList": [
      {"key_1": ["kye_1_item_1", "kye_1_item_2"]},
      {"key_2": ["kye_2_item_1", "kye_2_item_2", "key_2_item_3"]},
      {"key_3": ["kye_3_item_1", "kye_3_item_2"]}
   ]
}
```

The following code shows the merged result:

```
[
   {
      "key_1": "kye_1_item_1",
      "key_2": "kye_2_item_1",
      "key_3": "kye_3_item_1"
    },
    {
      "key_1": "kye_1_item_2",
      "key_2": "kye_2_item_2",
      "key_3": "kye_3_item_2"
    },
    {
      "key_1": "kye_1_item_2",
      "key_2": "kye_2_item_3",
      "key_3": "kye_3_item_2"
    }
]
```

- In the following template example, all instances created in WebServer are added to the vServer group of an SLB instance:

```
{
    "ROSTemplateFormatVersion": "2015-09-01",
    "Resources": {
        "WebServer": {
          "Type": "ALIYUN::ECS::InstanceGroupClone",
          "Properties": {
              "SourceInstanceId": "i-xxxxx",
               "Password": "Hello****",
               "MinAmount": 1,
               "MaxAmount": 1
          }
        },
        "CreateVServerGroup": {
            "Type": "ALIYUN::SLB::VServerGroup",
            "Properties": {
                "LoadBalancerId": "lb-****",
                "VServerGroupName": "VServerGroup-****",
                "BackendServers": {
                   "Fn::MergeMapToList": [
                       {"Port": [6666, 9090, 8080]},
                       {"ServerId": {"Fn::GetAtt": ["WebServer", "InstanceIds"]}},
                       {"Weight": [20, 100]}
                   ]
                }
            }
        }
    }
}
```

- Supported functions
  - Fn::Base64Encode
  - Fn::GetAtt
  - Fn::Join
  - Fn::Select
  - Ref
  - Fn::If
  - Fn::ListMerge
  - Fn::GetJsonValue

Fn::Avg

The Fn::Avg function is used to return the average value of a set of numbers.

- Declaration

```
{"Fn::Avg": [ndigits, [number1, number2, ... ]]}
```

- Parameters
  - `ndigits`: the number of decimal places to display. This parameter value must be an integer.
  - `[ number1, number2, ... ]`: the set of numbers for which the average value will be calculated. Each element in the group must be a number or a string that can be converted into a number.
- Return value

  The average value of the set of numbers.

- Examples

```
{ "Fn::Avg": [ 1, [1, 2, 6.0] ] }
{ "Fn::Avg": [ 1, ['1', '2', '6.0'] ] }
```

3.0 is returned in this example.

- Supported functions
  - Fn::GetAtt
  - Ref

Fn::SelectMapList

The Fn::SelectMapList function is used to return a list of map elements.

- Declaration

```
{"Fn::SelectMapList": ["key2", [{"key1": "value1-1", "key3": "value1-3"}, {"key1": "value2-1", "key2": "value2-2"}, {"key1": "value3-1", "key2": "value3-2"}, ...] ] }
```

- Parameters
  - `key2`: the key to be queried in the map.
  - `[{ "key1": "value1-1", "key3": "value1-3" }, ... ]`: the list of maps.
- Return value

  A list of key values for all maps in the map list.
- Examples

```
{
 "Fn::SelectMapList": [
   "key2",
   [
      {"key1": "value1-1", "key3": "value1-3"},
      {"key1": "value2-1", "key2": "value2-2"},
      {"key1": "value3-1", "key2": "value3-2"}
   ]
 ]
}
```

`["value2-2","value3-2"]` is returned in this example.

Fn::Add

The Fn::Add function is used to sum the values of parameters.

- Declaration

```
{"Fn::Add": [{"Product": "ROS"}, {"Fn": "Add"}]}
```

- Parameters
  - The parameters must be arranged as a list.
  - The parameters in the list can be of the Number, List, or Dictionary type. All the parameters must be of the same type. The list must contain at least two parameters.
- Return value

  If the parameter values are numbers, sum the parameter values. If the parameter values are lists, concatenate the values. If the parameter values are dictionaries, merge the values. If the two parameters have the same key, overwrite the former parameter value with the latter.
- Examples

```
{
 "Fn::Add": [
   {"Product": "ROS"},
   {"Fn": "Add"}
 ]
}
```

`{"Fn":"Add","Product":"ROS"}` is returned in this example.

5.1.5.6. Mappings

The Mappings section is a key-value mapping table. When mappings are used in resource and output definitions, use Fn::FindInMap to find their values by specifying corresponding keys.

Syntax

A mapping consists of key-value pairs, where both the keys and values can be strings or numbers. Multiple mappings are separated by commas (,). Each mapping name must be unique. Mappings must be pure data and cannot parse functions.

Examples

The following example shows a correct mapping definition:

```
"Mappings": {
    "ValidMapping": {
        "TestKey1": {"TestValu1": "value1"},
        "TestKey2": {"TestValu2": "value2"},
        1234567890: {"TestValu3": "value3"},
        "TestKey4": {"TestValu4": 1234}
    }
}
```

The following example shows an incorrect mapping definition:

```
"Mappings": {
    "InvalidMapping1": {
        "ValueList": ["foo", "bar"],
          "ValueString": "baz"
    },
    "InvalidMapping2": ["foo", {"bar" : "baz"}],
    "InvalidMapping3": "foobar"
}
```

The following example shows how to use Fn::FindInMap to find the return value:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "regionParam": {
      "Description": "The region where the ECS instance is created",
      "Type": "String",
      "AllowedValues": [
        "hangzhou",
        "beijing"
      ]
    }
  },
  "Mappings": {
    "RegionMap": {
      "hangzhou": {
        "32": "m-25l0rcfjo",
        "64": "m-25l0rcfj1"
      },
      "beijing": {
        "32": "m-25l0rcfj2",
        "64": "m-25l0rcfj3"
      }
    }
  },
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId": {
          "Fn::FindInMap": [
            "RegionMap",
            {
              "Ref": "regionParam"
            },
            "32"
          ]
        },
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "Tags": [
          {
            "Key": "Department1",
            "Value": "HumanResource"
          },
          {
            "Key": "Department2",
            "Value": "Finance"
          }
        ]
      }
    }
  }
}
```

5.1.5.7. Conditions

Condition bodies are defined by Fn::And, Fn::Or, Fn::Not, and Fn::Equals operators. These operators, along with the parameters that you specify when you create or update a stack, are used to evaluate each condition. You can reference other conditions, parameters, and mappings in your condition. Conditions are used in resource and output definitions to establish dependencies. Use Fn::If or Condition in resource and output definitions to implement conditions.

Syntax

Each condition consists of a condition name and a condition body. The condition name is a string. The condition body starts with Fn::And, Fn::Or, Fn::Not, or Fn::Equals. You can reference other conditions in your condition. Separate multiple conditions with comma (,). Each condition name must be unique.

The following functions can be used, but not as the outermost functions:

"Fn::Select", "Fn::Join", "Fn::Split", "Fn::Replace","Fn::Base64Encode", "Fn::Base64Decode", "Fn::MemberListToMap","Fn::If", "Fn::ListMerge", "Fn::GetJsonValue", "Fn::MergeMapToList","Fn::SelectMapList", "Fn::Add", "Fn::Avg", "Fn::Str", "Fn::Calculate","Ref"(parameter references only), and "Fn:: FindInMap".

Examples

- The following example shows how to define conditions:

```
"Conditions" : {
       "DevEnv": {"Fn::Equals": ["Dev", {"Ref": "EnvType"}]},
       "UTEnv": {"Fn::Equals": ["UT", {"Ref": "EnvType"}]},
       "PREEnv": {"Fn::Not": {"Fn::Or": ["DevEnv", "UTEnv"]}},
       "ProdEnv": {"Fn::And": [{"Fn::Equals": ["Prod", {"Ref": "EnvType"}]}, "PREEnv"]}
}
```

- The following example shows how to use conditions in a resource definition.

  In this example, a condition is used to determine whether to create a data disk and an Object Storage Service (OSS) bucket for an Elastic Compute Service (ECS) instance based on the EnvType value.

```
{
    "ROSTemplateFormatVersion":"2015-09-01",
    "Parameters":{
        "EnvType":{
            "Default":"pre",
            "Type":"String"
        }
    },
    "Conditions":{
        "CreateProdRes":{
            "Fn::Equals":[
                "prod",
                {
                    "Ref":"EnvType"
                }
            ]
        }
    },
    "Resources":{
        "WebServer":{
            "Type":"ALIYUN::ECS::Instance",
            "Properties":{
                "DiskMappings":{
                    "Fn::If":[
                        "CreateProdRes",
                        [
                            {
                                "Category":"cloud_efficiency",
```

```
                                "DiskName":"FirstDataDiskName",
                                "Size":40
                            },
                            {
                                "Category":"cloud_ssd",
                                "DiskName":"SecondDataDiskName",
                                "Size":40
                            }
                        ],
                        {
                            "Ref":"ALIYUN::NoValue"
                        }
                    ]
                },
                "VpcId":"vpc-2zew9pxh2yirtzqxd****",
                "SystemDiskCategory":"cloud_efficiency",
                "SecurityGroupId":"sg-2zece6wcqriejf1v****",
                "SystemDiskSize":40,
                "ImageId":"centos_6_8_64_40G_base_2017****.vhd",
                "IoOptimized":"optimized",
                "VSwitchId":"vsw-2zed9txvy7h2srqo6****",
                "InstanceType":"ecs.n1.medium"
            }
        },
        "OssBucket": {
            "Type": "ALIYUN::OSS::Bucket",
            "Condition": "CreateProdRes",
            "Properties": {
                "AccessControl": "private",
                "BucketName": "myprodbucket"
            }
        }
    },
    "Outputs":{
        "InstanceId":{
            "Value":{
                "Fn::GetAtt":[
                    "WebServer",
                    "InstanceId"
                ]
            }
        },
        "OssDomain":{
            "Condition": "CreateProdRes",
            "Value":{
                "Fn::GetAtt":[
                    "OssBucket",
                    "DomainName"
                ]
            }
        }
    }
}
```

5.1.6. Resource types

5.1.6.1. ECS

5.1.6.1.1. ALIYUN::ECS::AutoSnapshotPolicy

ALIYUN::ECS::AutoSnapshotPolicy is used to create an automatic snapshot policy.

Statement

```
{
  "Type" : "ALIYUN::ECS::AutoSnapshotPolicy",
  "Properties" : {
    "TimePoints" : String,
    "RepeatWeekdays" : String,
    "RetentionDays" : Integer,
    "DiskIds" : String,
    "AutoSnapshotPolicyName" : String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
TimePoints	List	Retained	Yes	The points in time at which automatic snapshots are created. Unit: hours.	Value range:[0, 23], represents 24 time points from 00:00 to 23:00. For example: `1` indicating 01:00. To schedule multiple automatic snapshot creation tasks in a day, you can set the TimePoints parameter as an array. • The maximum number of time points allowed is 24. • Use one format for multiple time points like [0, 1,... 23]. Separate time points with commas (,).

Parameter	Type	Required	Editable	Description	Constraint
RepeatWeekdays	List	Retained	Yes	The days of a week on which automatic snapshots are created.	Value range:[1, 7], `1` indicates Monday. To schedule multiple automatic snapshot creation tasks in a week, you can set the RepeatWeekdays parameter as an array. • You can specify up to 7 days over a one week period. • Use one format for multiple time points like [1, 2,... 7]. Separate the time points with commas (,).
RetentionDays	Integer	Retained	Yes	The number of days for which you want to retain automatic snapshots.	Default value: -1. Valid values: • -1: The automatic snapshots are retained indefinitely. • [1, 65536]: The automatic snapshots are retained for the specified number of days. Default value: -1.

Parameter	Type	Required	Editable	Description	Constraint
DiskIds	List	Retained	Yes	The ID of the destination disk. When you want to apply the automatic snapshot policy to multiple disks, you can set the diskids "d-zzzzzzzzz"]. Separate multiple disk IDs with commas (,).	None
AutoSnapshotPolicyName	String	Yes	True	The name of the automatic snapshot policy.	• The name must be 2 to 128 characters in length • It can contain letters, digits, colons (:), underscores (_), and hyphens (-). • It cannot start with http:// or https://. This parameter is empty by default.

Response parameters

Fn::GetAtt

AutoSnapshotPolicyId: the ID of the automatic snapshot policy.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "AutoSnapshotPolicy": {
      "Type": "ALIYUN::ECS::AutoSnapshotPolicy",
      "Properties": {
        "TimePoints": ["0"],
        "RepeatWeekdays": ["1"],
        "RetentionDays": 10,
        "DiskIds": ["<DiskId>"],
        "AutoSnapshotPolicyName": "MyAutoSnapshotPolicy"
      }
    }
  }
}
```

5.1.6.1.2. ALIYUN::ECS::BandwidthPackage

ALIYUN::ECS::BandwidthPackage is used to create a service plan for a NAT gateway.

Syntax

```
{
  "Type": "ALIYUN::ECS::BandwidthPackage",
  "Properties": {
    "Description": String,
    "NatGatewayId": String,
    "ZoneId": String,
    "BandwidthPackageName": String,
    "Bandwidth": Integer,
    "IpCount": Integer
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
NatGatewayId	String	Yes	No	The ID of the NAT gateway to which you want to bind the service plan.	None
Bandwidth	Integer	Yes	No	The bandwidth.	Valid values: 5 to 5000. Unit: Mbit/s. Default value: 5.
IpCount	Integer	Yes	No	The number of public IP addresses assigned to the NAT gateway.	Valid values: 1 to 5.
Description	String	No	No	The description of the service plan.	The description must be 2 to 256 characters in length.
ZoneId	String	No	No	The ID of the zone where the NAT gateway resides.	None

Property	Type	Required	Editable	Description	Constraint
BandwidthPackageName	String	No	No	The name of the service plan.	The name must be 2 to 128 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter.

Response parameters

Fn::GetAtt

- BandwidthPackageId: the ID of the service plan.
- BandwidthPackageIps: all IP addresses included in the service plan.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "BandwidthPackage": {
      "Type": "ALIYUN::ECS::BandwidthPackage",
      "Properties": {
        "BandwidthPackageName": "pkg_2",
        "Description": "my_bandwidth",
        "NatGatewayId": "ngw-h1xox****",
        "IpCount": 2,
        "Bandwidth": 5,
        "ZoneId": "cn-beijing-c"
      }
    }
  },
  "Outputs": {
    "BandwidthPackageId": {
         "Value": {"Fn::GetAttr": ["BandwidthPackage","BandwidthPackageId"]}
    },
    "BandwidthPackageIps": {
         "Value": {"Fn::GetAttr": ["BandwidthPackage","BandwidthPackageIps"]}
    }
  }
}
```

5.1.6.1.3. ALIYUN::ECS::Command

ALIYUN::ECS::Command is used to create a Cloud Assistant command.

Statement

```
{
  "Type": "ALIYUN::ECS::Command",
  "Properties": {
    "Name": String,
    "WorkingDir": String,
    "CommandContent": String,
    "Timeout": Integer,
    "Type": String,
    "Description": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
Name	String	Yes	True	The name of the command, which supports all character sets. The name can be up to 30 characters in length.	None
WorkingDir	String	Yes	True	The working directory on the ECS instance where the command will be run.	None
CommandContent	String	Yes	Released	The Base64-encoded content of the command. When you specify request parameters `Type` you must also specify this parameter. The parameter value must be Base64-encoded and cannot exceed 16 KB in size after encoding.	None
Timeout	String	No.	True	The timeout period that is specified for the command to run on ECS instances. Unit: seconds. If the command fails to run within the specified period, the command execution will time out and the process will be forcibly terminated. Default value: 3600.	None

Parameter	Type	Required	Editable	Description	Constraint
Type	String	No	No	The command type. Valid values: • RunBatScript: Creates a Bat script for a Windows instance. • RunPowerShellScript: Create a PowerShell script to run on a Windows instance. • RunShellScript: Creates a Shell script for Linux-based instances.	None
Description	String	Yes	True	The description of the command, which supports all character sets. The description can be up to 100 characters in length.	None

Response parameters

Fn::GetAtt

CommandId: the ID of the command.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "WorkingDir": {
      "Type": "String",
      "Description": "The path where command will be executed in the instance."
    },
    "CommandContent": {
      "Type": "String",
      "Description": "The content of command. Content requires base64 encoding. Maximum size support 1
6KB."
    },
    "Type": {
      "Type": "String",
      "Description": "The type of command."
    },
    "Description": {
      "Type": "String",
      "Description": "The description of command."
    },
    "Timeout": {
      "Type": "Number",
      "Description": "Total timeout when the command is executed in the instance. Input the time unit
as second. Default is 3600s."
    },
    "Name": {
      "Type": "String",
      "Description": "The name of command."
    }
  }
```

```
  },
  "Resources": {
    "Command": {
      "Type": "ALIYUN::ECS::Command",
      "Properties": {
        "WorkingDir": {
          "Ref": "WorkingDir"
        },
        "CommandContent": {
          "Ref": "CommandContent"
        },
        "Type": {
          "Ref": "Type"
        },
        "Description": {
          "Ref": "Description"
        },
        "Timeout": {
          "Ref": "Timeout"
        },
        "Name": {
          "Ref": "Name"
        }
      }
    }
  },
  "Outputs": {
    "CommandId": {
      "Description": "The id of command created.",
      "Value": {
        "Fn::GetAtt": [
          "Command",
          "CommandId"
        ]
      }
    }
  }
}
```

5.1.6.1.4. ALIYUN::ECS::CustomImage

ALIYUN::ECS::CustomImage is used to create a custom image.

Statement

```
{
  "Type": "ALIYUN::ECS::CustomImage",
  "Properties": {
    "Description": String,
    "InstanceId": String,
    "ImageName": String,
    "ImageVersion": String,
    "SnapshotId": String,
    "Tag": List,
    "ResourceGroupId": String,
    "Platform": String,
    "DiskDeviceMapping": List,
    "Architecture": String
  }
}
```

Properties

Parameter	Type	Required or Not	Editable	Description	Constraint
Description	String	Yes	Released	The description of the image.	The description can be up to 256 characters in length. This parameter is empty by default. It cannot start with http:// or https://.
InstanceId	String	Yes	Released	The ID of the ECS instance.	If this parameter is specified, an ECS instance will be used to create the custom image.
ImageName	String	Yes	Released	The name of the image.	The name must be 2 to 128 characters in length and can contain letters, digits, underscores (_), and hyphens(-). It must start with a letter but cannot start with http:// or https://.
ImageVersion	String	Yes	Released	The image version.	The image version must be 1 to 40 characters in length.
SnapshotId	String	Yes	Released	The ID of the snapshot.	• If this parameter is specified, a snapshot will be used to create the custom image. • If both this parameter and the InstanceId parameter are specified, this parameter will be ignored and an instance will be used to create the custom image.

Parameter	Type	Required or Not	Editable	Description	Constraint
Tags	List	Erased	Released	The tags of the image.	None
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the custom image belongs.	None
Platform	String	Yes	Released	If you specify a data disk snapshot to be used to create the system disk of the custom image, you must use the Platform parameter to determine the release version of the operating system for the system disk.	None
DiskDeviceMapping	List	Erased	Released	The mappings between images and snapshots.	None
Architecture	String	Yes	Released	If you specify a data disk snapshot to be used to create the system disk of the custom image, you must use the Architecture parameter to determine the architecture of the system disk. Default value: x86_64.	Valid values: • i386 • x86_64

Tag syntax

```
"Tag": [
  {
    "Key": String,
    "Value": String
  }
]
```

Tag properties

Parameter	Type	Required or Not	Editable	Description	Constraint

Parameter	Type	Required or Not	Editable	Description	Constraint
Key	String	Yes	Released	The tag key of the image.	The tag key cannot be a null string. The key can be up to 64 characters in length. It cannot start with aliyun or acs: and cannot contain http:// or https://.
Value	String	Yes	Released	The tag value of the image.	The tag value can be an empty string. The value can be up to 128 characters in length. It cannot start with aliyun or acs: and cannot contain http:// or https://.

DiskDeviceMapping

```
"DiskDeviceMapping": [
  {
    "Device": String,
    "SnapshotId": String,
    "Size": Integer,
    "DiskType": String
  }
]
```

DiskDeviceMapping properties

Parameter	Type	Required or Not	Editable	Description	Constraint
Device	String	Yes	Released	The device name of disk N in the custom image.	The system allocates a device name in alphabetical order from /dev/xvda to /dev/xvdz.
SnapshotId	String	Yes	Released	The ID of the snapshot that is used to create the custom image.	None

Parameter	Type	Required or Not	Editable	Description	Constraint
Size	String	Optional	Released	The size of disk N. Unit: GiB.	Valid values: 5 to 2000. • The default value is the size of the snapshot specified by the DiskDeviceMapping.N.SnapshotId parameter. • If the DiskDeviceMapping.N.SnapshotId parameter is not specified, the default disk size is 5 GiB. • The disk size must be greater than or equal to the size of the snapshot specified by the DiskDeviceMapping.N.SnapshotId parameter.
DiskType	String	Yes	Released	The type of disk N in the custom image. You can specify this parameter to create the system disk of the custom image from a data disk snapshot. If you do not specify this parameter, the disk type is determined by the corresponding snapshot.	Valid values: • system: indicates a system disk. • data: indicates a data disk.

Response parameters

Fn::GetAtt

ImageId: the ID of the custom image.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::InstanceGroup",
      "Properties": {
        "VpcId": "vpc-2zevx9ios1rszqv0a****",
        "MinAmount": 1,
        "SecurityGroupId": "sg-2ze7pxymaix640qr****",
        "ImageId": {
          "Ref": "CustomImage"
        },
        "IoOptimized": "optimized",
        "SystemDisk_Description": "SystemDisk.Description",
        "SystemDisk_DiskName": "SystemDisk.DiskName",
        "SystemDisk_Category": "cloud_ssd",
        "VSwitchId": "vsw-2zei67xd9nhcqxzec****",
        "Password": "Wenqiao****",
        "InstanceType": "ecs.n1.medium",
        "MaxAmount": 1
      }
    },
    "CustomImage": {
      "Type": "ALIYUN::ECS::CustomImage",
      "Properties": {
        "InstanceId": "i-2zefq1f3ynnrr89q****",
        "SnapshotId": "s-2ze0ibk1pvak4mw6****",
        "ImageName": "image-test-****",
        "ImageVersion": "verison-6-1"
      }
    }
  },
  "Outputs": {
    "CustomImage": {
      "Value": {
        "Fn::GetAtt": [
          "CustomImage",
          "ImageId"
        ]
      }
    },
    "InstanceIds": {
      "Value": {
        "Fn::GetAtt": [
          "WebServer",
          "InstanceIds"
        ]
      }
    }
  }
}
```

5.1.6.1.5. ALIYUN::ECS::DedicatedHost

ALIYUN::ECS::DedicatedHost is used to create a dedicated host.

Statement

```
{
  "Type": "ALIYUN::ECS::DedicatedHost",
  "Properties": {
    "DedicatedHostType": String,
    "DedicatedHostName": String,
    "AutoReleaseTime": String,
    "Description": String,
    "AutoPlacement": String,
    "Tags": List,
    "ActionOnMaintenance": String,
    "NetworkAttributesSlbUdpTimeout": Integer,
    "ChargeType": String,
    "ResourceGroupId": String,
    "ZoneId": String,
    "NetworkAttributesUdpTimeout": Integer,
    "Quantity": Integer
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
DedicatedHostType	String	No	No	The dedicated host type.	None
DedicatedHostName	String	Yes	Released	The name of the dedicated host.	• The name must be 2 to 128 characters in length and can contain letters, digits, colons (:), underscores (_), and hyphens (-). • Must start with an uppercase or lowercase letter, and cannot start with `http://` or `https://` the beginning. • It can contain digits, colons (:), underscores (_), and hyphens (-).

Parameter	Type	Required	Editable	Description	Constraint
AutoReleaseTime	String	Yes	Released	The time scheduled for the dedicated host to be automatically released. If you do not specify the AutoReleaseTime parameter, the dedicated host will not be automatically released. • The minimum release time must be at least 30 minutes after the current time. • The maximum release time must be at most three years from the current time. • If the value of `ss` is not `00`, the start time is automatically rounded down to the nearest minute based on the value of `mm`.	None
Description	String	Yes	Released	The description of the dedicated host.	None
ZoneId	String	Yes	Released	The ID of the zone where the dedicated host resides. This parameter is empty by default. If this parameter is not specified, the system will automatically select a zone.	None
ChargeType	String	Yes	Released	The billing method of the dedicated host.	Valid values: PostPaid and pay-as-you-go.

Parameter	Type	Required	Editable	Description	Constraint
AutoPlacement	String	Yes	Released	Specifies whether to add the dedicated host to the resource pool for automatic deployment. If you do not specify a DedicatedHostId when you create an instance on a DDH, Alibaba Cloud automatically selects a DDH from the resource pool to host the instance.	Valid values: • on • off **Note** If you do not specify this parameter, the dedicated host is added to the automatic deployment resource pool. If you do not want to add the dedicated host to the resource pool for automatic deployment, set the value to off.
Tags	List	Erased	Released	The custom tags of the instance.	A maximum of 20 tags are supported. The format is as follows: `[{"Key": "tagKey", "Value": "tagValue"},{"Key": "tagKey2", "Value": "tagValue2"}]`.

Parameter	Type	Required	Editable	Description	Constraint
ActionOnMaintenance	String	Yes	Released	The method used to migrate the instances on the DDH when the DDH fails or needs to be repaired online.	Valid valus: • Migrate: specifies that the instances are migrated to another physical server and restarted. • Stop: specifies that all the instances on the DDH are stopped. If the DDH cannot be repaired, the instances are migrated to another physical server and restarted. The default value is "Migrate" for a dedicated host and "Stop" for a local disk.
NetworkAttributesSlbUdpTimeout	String	Optional	Released	The timeout period for a UDP session.	Valid values: 15 to 310. Unit: seconds.
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the dedicated host belongs.	None
NetworkAttributesUdpTimeout	String	Optional	Released	The timeout period for UDP sessions that users can access for cloud services running on the dedicated host.	Valid values: 15 to 310. Unit: seconds.
Quantity	String	Optional	Released	The number of DDHs that you want to create this time.	Valid values: 1 to 100. Default value: 1

Response parameters

Fn::GetAtt

- OrderId: the ID of the order.
- DedicatedHostIds: the list of host IDs.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "AutoRenewPeriod": {
      "Type": "Number",
      "Description": "The time period of auto renew. When the parameter InstanceChargeType is PrePaid,
it will take effect.It could be 1, 2, 3, 6, 12\. Default value is 1.",
```

```
      "AllowedValues": [
        1,
        2,
        3,
        6,
        12
      ],
      "Default": 1
    },
    "Description": {
      "Type": "String",
      "Description": "The description of host."
    },
    "ZoneId": {
      "Type": "String",
      "Description": "The zone to create the host."
    },
    "DedicatedHostName": {
      "Type": "String",
      "Description": "The name of the dedicated host, [2, 128] English or Chinese characters. It must begin with an uppercase/lowercase letter or a Chinese character, and may contain numbers, '_' or '-'. It cannot begin with http:// or https://."
    },
    "ChargeType": {
      "Type": "String",
      "Description": "Instance Charge type, allowed value: Prepaid and Postpaid. If specified Prepaid, please ensure you have sufficient balance in your account. Or instance creation will be failure. Default value is Postpaid.",
      "AllowedValues": [
        "PrePaid",
        "PostPaid"
      ],
      "Default": "PostPaid"
    },
    "AutoRenew": {
      "Type": "String",
      "Description": "Whether renew the fee automatically? When the parameter InstanceChargeType is PrePaid, it will take effect. Range of value:True: automatic renewal.False: no automatic renewal. Default value is False.",
      "AllowedValues": [
        "True",
        "False"
      ],
      "Default": "False"
    },
    "Period": {
      "Type": "Number",
      "Description": "Prepaid time period. Unit is month, it could be from 1 to 9 or 12, 24, 36, 48, 60\. Default value is 1.",
      "AllowedValues": [
        1,
        2,
        3,
        4,
        5,
        6,
        7,
        8,
        9,
        12,
```

```
        24,
        36,
        48,
        60
      ],
      "Default": 1
    },
    "DedicatedHostType": {
      "Type": "String",
      "Description": "The instance type of host."
    },
    "PeriodUnit": {
      "Type": "String",
      "Description": "Unit of prepaid time period, it could be Week/Month. Default value is Month.",
      "AllowedValues": [
        "Week",
        "Month"
      ],
      "Default": "Month"
    },
    "AutoReleaseTime": {
      "Type": "String",
      "Description": "Auto release time for created host, Follow ISO8601 standard using UTC time. form
at is 'yyyy-MM-ddTHH:mm:ssZ'. Not bigger than 3 years from this day onwards"
    }
  },
  "Resources": {
    "Host": {
      "Type": "ALIYUN::ECS::DedicatedHost",
      "Properties": {
          "Description": {
          "Ref": "Description"
        },
        "ZoneId": {
          "Ref": "ZoneId"
        },
        "DedicatedHostName": {
          "Ref": "DedicatedHostName"
        },
        "ChargeType": {
          "Ref": "ChargeType"
        },
          "DedicatedHostType": {
          "Ref": "DedicatedHostType"
        },
        "PeriodUnit": {
          "Ref": "PeriodUnit"
        },
        "AutoReleaseTime": {
          "Ref": "AutoReleaseTime"
        }
      }
    }
  },
  "Outputs": {
    "OrderId": {
      "Description": "The order id list of created instance.",
      "Value": {
        "Fn::GetAtt": [
          "Host",
```

```
            "OrderId"
          ]
        }
      },
      "DedicatedHostIds": {
        "Description": "The host id list of created hosts",
        "Value": {
          "Fn::GetAtt": [
            "Host",
            "DedicatedHostIds"
          ]
        }
      }
    }
}
```

5.1.6.1.6. ALIYUN::ECS::Disk

ALIYUN::ECS::Disk is used to create an ECS Disk.

Statement

```
{
  "Type": "ALIYUN::ECS::Disk",
  "Properties": {
    "DiskName": String,
    "Description": String,
    "Tags": List,
    "AutoSnapshotPolicyId": String,
    "Encrypted": Boolean,
    "ZoneId": String,
    "ResourceGroupId": String,
    "SnapshotId": String,
    "DiskCategory": String,
    "PerformanceLevel": String,
    "DeleteAutoSnapshot": Boolean,
    "Size": Integer
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None
ZoneId	String	No	No	The ID of the zone where the instance resides.	None

Parameter	Type	Required	Editable	Description	Constraint
DiskName	String	Yes	Released	The name of the disk.	• The name must be 2 to 128 characters in length • And can contain letters, digits, periods, underscores (_), and hyphens (-). • It cannot start with http:// or https://. • The disk name will be displayed in the ECS console.
Description	String	Yes	Released	The description of the disk.	• The description must be 2 to 256 characters in length. • Cannot `http://` or `https://` the beginning. • The disk description will be displayed in the ECS console.
Tags	List	Erased	Released	The custom tags of the instance.	Up to four tags are supported. Example values: `[{"Key": "tagKey", "Value": "tagValue"}, {"Key": "tagKey2", "Value": "tagValue2"}]`.
DiskCategory	String	Yes	Released	The type of the data disk.	Value range • cloud: indicates a basic disk. • cloud_efficiency: indicates an ultra disk. • cloud_ssd: indicates a standard SSD. • cloud_essd: enhanced SSD (ESSD) Default value: cloud.

Parameter	Type	Required	Editable	Description	Constraint
SnapshotId	String	Yes	Released	The ID of the snapshot used to create the data disk.	• If both this parameter and 'Size' are specified, the value of this parameter prevails. • The actual size of the created disk is the size of the specified snapshot. • Snapshots created on or before July 15, 2013 cannot be used to create disks.
PerformanceLevel	String	Yes	Released	Specifies the performance level of an ESSD when you create the ESSD.	Default value: PL1. Valid values: • PL1: A single enhanced SSD delivers up to 50,000 random read/write IOPS. • PL2: A single ESSD delivers up to 100,000 random read/write IOPS. • PL3: maximum random read /write IOPS of 100,000 per disk.
Size	String	Optional	Released	The size of the disk. Unit: GiB. The value of this parameter must be equal to or greater than the capacity of the specified snapshot.	Valid values: • cloud: 5 to 2000 • cloud_efficiency: 20 to 32768 • cloud_ssd: 20 to 32768 • cloud_essd: 20 to 32768
AutoSnapshotPolicyId	String	Yes	Released	The ID of each automatic snapshot policy.	None
Encrypted	Boolean	Erased	Released	Specifies whether to encrypt the disk.	Valid values: • true • false Default value: false.

Parameter	Type	Required	Editable	Description	Constraint
DeleteAutoSnapshot	Boolean	Erased	Released	Specifies whether to delete the automatic snapshots of the disk when the disk is released.	Valid values: • true • false Default value: true.

Tags syntax

```
"Tags" : [
  {
    "Value" : String,
    "Key" : String
  }
]
```

Tags properties

Parameter	Type	Required	Editable
Key	String	No	No
Value	String	Yes	Released

Response parameters

Fn::GetAtt

- DiskId: the ID of the disk.
- Status: The Status of the disk.

Sample request

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "DataDisk": {
      "Type": "ALIYUN::ECS::Disk",
      "Properties": {
        "Size": 10,
        "ZoneId": "cn-beijing-a",
        "DiskName": "DataDisk",
        "Description": "ECSDataDisk"
      }
    }
  },
  "Outputs": {
    "DiskId": {
         "Value" : {"Fn::GetAtt": ["DataDisk","DiskId"]}
    },
    "Status": {
         "Value" : {"Fn::GetAtt": ["DataDisk","Status"]}
    }
  }
}
```

5.1.6.1.7. ALIYUN::ECS::DiskAttachment

ALIYUN::ECS::DiskAttachment is used to attach an ECS disk.

Statement

```
{
   "Type" : "ALIYUN::ECS::DiskAttachment",
   "Properties" : {
      "DiskId" : String,
      "InstanceId" : String,
      "Device" : String,
      "DeleteWithInstance" : String
   }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
InstanceId	String	No	No	The ID of the instance.	None
DiskId	String	No	No	The ID of the disk.	The disk and the ECS instance must belong to the same zone.

Parameter	Type	Required	Editable	Description	Constraint
Device	String	Yes	Released	The name of the disk.	If you do not set this parameter, the system will automatically allocate a device name in alphabetical order from /dev/xvdb to /dev/xvdz.
DeleteWithInstance	Boolean	Erased	Released	Specifies whether the disk is to be released together with the instance.	Valid values: • true: The disk will be released when the instance is released. • false: The disk will be retained when the instance is released.

Response parameters

Fn::GetAtt

- DiskId: the ID of the disk.
- Status: The Status of the disk.
- The name of the Device: disk.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "DiskAttachment": {
      "Type": "ALIYUN::ECS::DiskAttachment",
      "Properties": {
        "InstanceId": {
          "Ref": "InstanceId"
        },
        "Device": {
          "Ref": "Device"
        },
        "DeleteWithInstance": {
          "Ref": "DeleteWithInstance"
        },
        "DiskId": {
          "Ref": "DiskId"
        }
      }
    }
  },
  "Parameters": {
    "InstanceId": {
```

```
      "Type": "String",
      "Description": "The ID of the instance to attach the disk."
    },
    "Device": {
      "Type": "String",
      "Description": "The device where the volume is exposed on the instance. The device name could be
/dev/xvd[a-z]. If this parameter is not specified, the default value will be used."
    },
    "DeleteWithInstance": {
      "Type": "Boolean",
      "Description": "If this parameter is set to true, the disk will be deleted while the instance is
deleted. If this parameter is set to false, the disk will be retained after the instance is deleted.",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    },
    "DiskId": {
      "Type": "String",
      "Description": "The ID of the disk to be attached."
    }
  },
  "Outputs": {
    "Status": {
      "Description": "The disk status now.",
      "Value": {
        "Fn::GetAtt": [
          "DiskAttachment",
          "Status"
        ]
      }
    },
    "Device": {
      "Description": "The device where the volume is exposed on the ECS instance.",
      "Value": {
        "Fn::GetAtt": [
          "DiskAttachment",
          "Device"
        ]
      }
    },
    "DiskId": {
      "Description": "The ID of the created disk.",
      "Value": {
        "Fn::GetAtt": [
          "DiskAttachment",
          "DiskId"
        ]
      }
    }
  }
}
```

5.1.6.1.8. ALIYUN::ECS::ForwardEntry

ALIYUN::ECS::ForwardEntry is used to configure the DNAT table of a NAT Gateway.

Statement

```
{
  "Type": "ALIYUN::ECS::ForwardEntry",
  "Properties": {
    "ExternalIp": String,
    "ExternalPort": String,
    "ForwardTableId": String,
    "InternalIp": String,
    "IpProtocol": String,
    "InternalPort": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ExternalIp	String	No	No	The public IP address of the NAT gateway.	It must be an IP address that is included in the shared NAT Gateway of the bandwidth plan to which the DNAT table belongs.
ExternalPort	String	No	No	The public port number.	Valid values: 1 to 65535.
ForwardTableId	String	No	No	The ID of the DNAT table.	None
InternalIp	String	No	No	The destination IP address to which the request is forwarded.	This IP address is a private IP address.
IpProtocol	String	No	No	The type of the protocol.	Valid values: TCP, UDP, and Any.
InternalPort	String	No	No	The destination private port.	Valid values: 1 to 65535.

Response parameters

Fn::GetAtt

ForwardEntryId: the ID of each entry in the DNAT table.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ForwardEntry": {
      "Type": "ALIYUN::ECS::ForwardEntry",
      "Properties": {
        "ForwardTableId": "my_forwardtable",
        "ExternalIp": "101.201.XX.XX",
        "ExternalPort": "8080",
        "IpProtocol": "TCP",
        "InternalIp": "10.2.XX.XX",
        "InternalPort": "80"
      }
    }
  },
  "Outputs": {
    "ForwardEntryId": {
      "Value" : {"Fn::GetAttr": ["ForwardEntry","ForwardEntryId"]}
    }
  }
}
```

5.1.6.1.9. ALIYUN::ECS::Instance

ALIYUN::ECS::Instance is used to create an ECS instance.

Statement

```
{
  "Type": "ALIYUN::ECS::Instance",
  "Properties": {
    "RamRoleName": String,
    "IoOptimized": String,
    "PrivateIpAddress": String,
    "KeyPairName": String,
    "SystemDiskDiskName": String,
    "Description": String,
    "Tags": List,
    "HostName": String,
    "ImageId": String,
    "ResourceGroupId": String,
    "VSwitchId": String,
    "Password": String,
    "InstanceType": String,
    "SystemDiskCategory": String,
    "UserData": String,
    "SystemDiskSize": Number,
    "ZoneId": String,
    "VpcId": String,
    "InstanceName": String,
    "InternetMaxBandwidthIn": Integer,
    "DeletionProtection": Boolean,
    "DeploymentSetId": String,
    "SecurityGroupId": String,
    "HpcClusterId": String,
    "SystemDiskDescription": String,
    "DiskMappings": List
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None

Parameter	Type	Required	Editable	Description	Constraint
ImageId	String	No	Yes	The ID of the image used to start the ECS instance. You can use a public image, a custom image, or an Alibaba Cloud Marketplace image.	When editing a template, you can specify the image type and version or only the image type. ROS automatically selects an appropriate public image ID. You can use the wildcard character (*) to represent part of an image ID. Take all Ubuntu public images provided by Alibaba Cloud as an example. You can use one of the following methods to specify the public image ID for the ECS instance: • If you enter ubuntu, the system matches it with the following ID: ubuntu16_0402_64_20G_alibase_20170818.vhd • If you enter ubuntu_14, the system matches it with the following ID: ubuntu_14_0405_64_20G_alibase_20170824.vhd • If you enter ubuntu*14*32, the system matches it with the following ID: ubuntu_14_0405_32_40G_alibase_20170711.vhd • If you enter ubuntu_16_0402_32, the system matches it with the following ID: ubuntu_16_0402_32_40G_alibase_20170711.vhd
InstanceType	String	No	No	The type of the ECS instance.	None
SecurityGroupId	String	No	No	The ID of the security group to which the created instance will belong.	None
Description	String	Yes	Released	The description of created instances.	The description can be up to 256 characters in length.

Parameter	Type	Required	Editable	Description	Constraint
InstanceName	String	Yes	Released	The name of a created instance.	The name can be up to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-).
Password	String	Yes	Released	The password used to log on to the ECS instance.	The characters in length is 8 to 30. And must contain at least one of the following character types: uppercase letters, lowercase letters, digits, and special character. Special characters include ( ) ` ~ ! @ # $ % ^ & * - + = | { } [ ] : ; ' < > , . ? / - / - If you specify the Password parameter in the API request, use HTTPS to secure the API and protect your password.
HostName	String	Yes	Released	The hostname of the instance.	The password must be at least 2 characters in length. It cannot And hyphens (-) cannot start or end the hostname and cannot be used consecutively. On Windows, the hostname can be up to 15 characters in length and can contain letters, digits, and hyphens (-). It cannot contain periods (.) and cannot be composed of only digits. On other OSs such as Linux, the hostname can contain a maximum of 30 characters, including periods (.), each segment can contain uppercase or lowercase letters, digits, and hyphens (-).
PrivateIpAddress	String	Yes	Released	The private IP address of an ECS instance in a VPC. The specified IP address must not be used by other instances in the VPC.	None

Parameter	Type	Required	Editable	Description	Constraint
InternetMaxBandwidthIn	String	Optional	Released	The maximum inbound bandwidth from the Internet.	Valid values: 1 to 100. Default value: 100. Unit: Mbit/s.
IoOptimized	String	Yes	Released	Specifies whether an I/O optimized instance is created.	Valid values: • none (non-I/O optimized) • optimized Default value: none.
DiskMappings	List	Erased	Released	The data disks to be attached to the instance.	A maximum of 16 disks can be attached to each instance.
SystemDiskCategory	String	Yes	Released	The type of the system disk.	Valid values: • cloud • cloud_efficiency • cloud_ssd • ephemeral_ssd
SystemDiskDescription	String	Yes	Released	The description of the ECS instance system disk.	None
SystemDiskDiskName	String	Yes	Released	The name of the ECS instance system disk.	None
SystemDiskSize	Number	No.	True	The size of the system disk. Unit: GB.	Valid values: 40 to 500. If a custom image is used to create a system disk, make sure that the size of the system disk is greater than that of the custom image.
Tags	List	Erased	Released	The custom tags of the instance.	A maximum of 20 tags are supported. The format is as follows: `[{&ldquo;Key”:”tagKey”,”Value”:”tagValue”}, {&ldquo;Key”:”tagKey2”,”Value”:”tagValue2”}]`.
UserData	String	Yes	Released	The user data that you provide when you create ECS instances.	The user data can be up to 16KB in size. You do not need to use Base64. you must use special characters. `\` Escape.
ZoneId	String	Yes	Released	The ID of the zone where the instance resides.	None

Parameter	Type	Required	Editable	Description	Constraint
HpcClusterId	String	Yes	Released	The ID of the HPC cluster to which the ECS instance belongs.	None
VpcId	String	Yes	Released	The ID of the VPC to which the ECS instance belongs.	None
VSwitchId	String	Yes	Released	The ID of the VSwitch for the ECS instance.	None
KeyPairName	String	Yes	Released	The name of the key pair that is used to connect to created ECS instances.	• For Windows-based instances, this parameter is empty by default. • In the Linux, if this parameter is specified, the Password content is still set to the instance, but the Password logon method is disabled by default. The key pair is used to verify the logon.
RamRoleName	String	Yes	Released	The RAM role name of the instance. You can call the ListRoles operation to query the role name.	None
DeletionProtection	Boolean	Erased	Released	The release protection property of created instances. It specifies whether the instances can be released from the ECS console or through the DeleteInstance operation.	Valid values: • True • False
DeploymentSetId	String	Yes	True	Deployment Set ID.	None

DiskMappings syntax

```
"DiskMappings": [
  {
    "Category": String,
    "DiskName": String,
    "Description": String,
    "Device": String,
    "SnapshotId": String,
    "Size": String
  }
]
```

DiskMappings properties

Parameter	Type	Required	Editable	Description	Constraint
Size	String	No	No	The size of data disk N. Unit: GB.	None
Category	String	Yes	Released	The type of the data disk.	Valid values: • cloud • cloud_efficiency • cloud_ssd • ephemeral_ssd Default value: cloud.
DiskName	String	Yes	Released	The name of data disk N.	The name can be up to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-).
Description	String	Yes	Released	The description of data disk N.	Valid values: 2 to 256. Default value: Null.
Device	String	Yes	Released	The device name of the data disk.	If you do not specify this parameter, the system automatically allocates a device name in alphabetical order from /dev/xvdb to /dev/xvdz.
SnapshotId	String	Yes	Released	The ID of the snapshot used to create the data disk.	None

Tags syntax

```
"Tags": [
  {
    "Value": String,
    "Key": String
  }
]
```

Tags properties

Parameter	Type	Required	Editable	Description	Constraint
Key	String	No	No	None	None
Value	String	Yes	Released	None	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- InstanceId: the ID of the instance. An ID is a globally unique identifier (GUID) generated by the system for an instance.
- PrivateIp: The private IP address of the instance in a VPC. This parameter takes effect only when the NetworkType parameter is set to VPC.
- InnerIp: The private IP address of the instance in a Classic network. This parameter takes effect only when the NetworkType parameter is set to Classic.
- PublicIp: the public IP address of the instance in a Classic network. This parameter takes effect only when the NetworkType parameter is set to Classic.
- ZoneId: the zone ID.
- HostName: the hostname of the instance.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId": "m-25l0rc****",
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "Tags": [{
            "Key": "tiantt",
            "Value": "ros"
        },{
            "Key": "tiantt1",
            "Value": "ros1"
        }
        ]
      }
    }
  },
  "Outputs": {
    "InstanceId": {
         "Value": {"get_attr": ["WebServer","InstanceId"]}
    },
    "PublicIp": {
         "Value": {"get_attr": ["WebServer","PublicIp"]}
    }
  }
}
```

5.1.6.1.10. ALIYUN::ECS::InstanceClone

ALIYUN::ECS::InstanceClone is used to clone an ECS instance.

Statement

```
{
  "Type": "ALIYUN::ECS::InstanceClone",
  "Properties": {
    "DeletionProtection": Boolean,
    "DiskMappings": List,
    "LoadBalancerIdToAttach": String,
    "Description": String,
    "BackendServerWeight": Integer,
    "Tags": List,
    "SecurityGroupId": String,
    "RamRoleName": String,
    "ImageId": String,
    "ResourceGroupId": String,
    "SpotPriceLimit": String,
    "InstanceChargeType": String,
    "SourceInstanceId": String,
    "Period": Number,
    "SpotStrategy": String,
    "Password": String,
    "InstanceName": String,
    "InternetMaxBandwidthIn": Integer,
    "ZoneId": String,
    "KeyPairName": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None
SourceInstanceId	String	No	No	The ID of the ECS instance to be cloned.	The clone operation clones the specified instance, including its instance type, image, bandwidth billing method, bandwidth limit, and network type. If the source ECS instance belongs to multiple security groups, the cloned instance is added only to the first of these security groups.
BackendServerWeight	String	Optional	Released	The weight of the ECS instance in the Server Load Balancer instance.	Value range:[0, 100]. Default value: 100.
LoadBalancerIdToAttach	String	Yes	Released	The ID of the SLB instance to which the ECS instance is to be attached.	None
Description	String	Yes	Released	The description of created instances.	The description can be up to 256 characters in length.

Parameter	Type	Required	Editable	Description	Constraint
ImageId	String	Yes	True	The ID of the image used to start created instances. You can use a public image, a custom image, or an Alibaba Cloud Marketplace image.	When editing a template, you can specify the image type and version or only the image type. ROS automatically selects an appropriate public image ID. You can use the wildcard character (*) to represent part of an image ID. Take all Ubuntu public images provided by Alibaba Cloud as an example. You can use one of the following methods to specify the public image ID for the ECS instance: If you enter ubuntu, the system matches it with the following ID: ubuntu16_0402_64_20G_alibase_20170818.vhd If you enter ubuntu_14, the system matches it with the following ID: ubuntu_14_0405_64_20G_alibase_20170824.vhd If you enter ubuntu*14*32, the system matches it with the following ID: ubuntu_14_0405_32_40G_alibase_20170711.vhd If you enter ubuntu_16_0402_32, the system matches it with the following ID: ubuntu_16_0402_32_40G_alibase_20170711.vhd
SecurityGroupId	String	Yes	Released	The ID of the security group to which the created instance will belong.	None
InstanceName	String	Yes	Released	The name of a created instance.	The name can be up to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-).

Parameter	Type	Required	Editable	Description	Constraint
Password	String	Yes	Released	The password used to log on to the ECS instance.	The password must be 8 to 30 characters in length. The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include ( ) ` ~ ! @ # $ % ^ & * - + = | { } [ ] : ; < > , . ? / If you specify the password parameter in the API request, use HTTPS to secure the API and protect your password.
DiskMappings	List	Erased	Released	The disks to be attached to created instances.	A maximum of 16 disks can be attached to each instance.
Tags	List	Erased	Released	The custom tags of the instance.	A maximum of 20 tags can be specified in the [{"Key":"tagKey","Value":"tagValue" },{ "Key":"tagKey2","Value":"tagValue 2"}].
ZoneId	String	Yes	Released	The ID of the zone where the instance resides.	None
InstanceChargeType	String	Yes	Released	The billing method of the new ECS instance.	Valid values: PrePaid and PostPaid. Default value: Postpaid. If you set this parameter to Prepaid, make sure that you have sufficient balance in your account. Otherwise, the instance fails to be created.
Period	Number	Erased	Released	The subscription period of the new ECS instance. This parameter is required when the InstanceChargeType parameter is set to PrePaid. This parameter is ignored when the InstanceChargeType parameter is set to PostPaid.	Valid values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 24, and 36. Unit: month.

Parameter	Type	Required	Editable	Description	Constraint
KeyPairName	String	Yes	Released	The name of the key pair that is used to connect to created ECS instances.	For Windows-based instances, this parameter is empty by default. In the Linux, if this parameter is specified, the Password content is still set to the instance, but the Password logon method is disabled by default. The key pair is used to verify the logon.
RamRoleName	String	Yes	Released	The RAM role name of the instance. You can call the ListRoles operation to query the role name.	None
SpotPriceLimit	String	Yes	Released	The maximum hourly price of the instance.	Parameter SpotStrategy this parameter takes effect only when the value is SpotWithPriceLimit.
SpotStrategy	String	Yes	Released	The bidding policy for the pay-as-you-go instance.	This parameter is valid only when the InstanceChargeType parameter is set to PostPaid. Valid values: NoSpot: applies to regular pay-as-you-go instances. SpotWithPriceLimit: applies to preemptible instances with a maximum hourly price. SpotAsPriceGo: applies to pay-as-you-go instances priced at the market price at the time of purchase. Default value: NoSpot.
InternetMaxBandwidthIn	String	Optional	Released	The maximum inbound bandwidth from the Internet. Unit: Mbit/s.	Valid values: 1 to 200.
DeletionProtection	Boolean	Erased	Released	Specifies whether to enable instance release protection in the console or by calling an API operation. (DeleteInstance) Release instances.	Valid values: true and false.

DiskMappings syntax

```
"DiskMappings": [
  {
    "Category": String,
    "DiskName": String,
    "Description": String,
    "Device": String,
    "SnapshotId": String,
    "Size": String
  }
]
```

DiskMappings properties

Parameter	Type	Required	Editable	Description	Constraint
Size	String	No	No	The size of data disk N. Unit: GB.	None
Category	String	Yes	Released	The type of the data disk.	Valid values: cloud, cloud_efficiency, cloud_ssd, and ephemeral_ssd Default value: cloud.
DiskName	String	Yes	Released	Disk name.	The name can be up to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-).
Description	String	Yes	Released	The description of data disk N.	The description must be 2 to 256 characters in length. This parameter is empty by default.
Device	String	Yes	Released	The device name of the data disk.	If you do not specify this parameter, the system automatically allocates a device name in alphabetical order from /dev/xvdb to /dev/xvdz.
SnapshotId	String	Yes	Released	The ID of the snapshot used to create the data disk.	None

Tags syntax

```
"Tags": [
  {
    "Value": String,
    "Key": String
  }
]
```

Tags properties

Parameter	Type	Required	Editable	Description	Constraint
Key	String	No	No	The tag key, which cannot be an empty string. It can be up to 64 characters in length, cannot start with acs: or aliyun, and cannot contain http:// or https://.	None
Value	String	Yes	Released	The tag value, which can be an empty string. It can be up to 128 characters in length and cannot start with acs: or aliyun. It cannot contain http:// or https://.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- InstanceId: the ID of the instance. An ID is a globally unique identifier (GUID) generated by the system for an instance.
- PrivateIp: The private IP address of the instance in a VPC. This parameter takes effect only when the NetworkType parameter is set to VPC.
- InnerIp: The private IP address of the instance in a Classic network. This parameter takes effect only when the NetworkType parameter is set to Classic.
- PublicIp: the public IP address of the instance in a Classic network. This parameter takes effect only when the NetworkType parameter is set to Classic.
- ZoneId: the zone ID.
- HostName: the hostname of the instance.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::InstanceClone",
      "Properties": {
        "SourceInstanceId": "i-25zsk****",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "DiskMappings": [
            {"Size": 10, "Category": "cloud"},
            {"Size": 10, "Category": "cloud", "SnapshotId": "s-25wsw****"}
        ]
      }
    }
  },
  "Outputs": {
    "InstanceId": {
         "Value": {"get_attr": ["WebServer","InstanceId"]}
    },
    "PublicIp": {
         "Value": {"get_attr": ["WebServer","PublicIp"]}
    }
  }
}
```

5.1.6.1.11. ALIYUN::ECS::InstanceGroup

ALIYUN::ECS::InstanceGroup is used to create an ECS instance group.

Syntax

```
{
  "Type": "ALIYUN::ECS::InstanceGroup",
  "Properties": {
    "SystemDiskAutoSnapshotPolicyId": String,
    "DedicatedHostId": String,
    "LaunchTemplateName": String,
    "RamRoleName": String,
    "IoOptimized": String,
    "PrivateIpAddress": String,
    "KeyPairName": String,
    "SystemDiskDiskName": String,
    "Description": String,
    "Tags": List,
    "HostName": String,
    "ImageId": String,
    "ResourceGroupId": String,
    "VSwitchId": String,
    "EniMappings": List,
    "Password": String,
    "InstanceType": String,
    "MaxAmount": Integer,
    "AutoReleaseTime": String,
    "SystemDiskCategory": String,
    "UserData": String,
    "LaunchTemplateId": String,
    "LaunchTemplateVersion": String,
    "SystemDiskSize": Number,
    "ZoneId": String,
    "VpcId": String,
    "InternetMaxBandwidthIn": Integer,
    "DeletionProtection": Boolean,
    "DeploymentSetId": String,
    "Ipv6AddressCount": Integer,
    "SecurityGroupId": String,
    "HpcClusterId": String,
    "SystemDiskDescription": String,
    "Ipv6Addresses": List,
    "NetworkType": String,
    "DiskMappings": List,
    "SystemDiskPerformanceLevel": String
  }
}
```

Properties

Attribute	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	No	Yes	The ID of the resource group to which a created instance belongs.	None.

Attribute	Type	Required	Editable	Description	Constraint
HpcClusterId	String	No	Yes	The ID of the HPC cluster to which a created instance belongs.	None.
MaxAmount	Integer	Supported	Yes	The maximum number of ECS instances that can be created at a time.	Valid values: 1 to 100. The MaxAmount parameter must be set to a value greater than or equal to the value of MinAmount.
MinAmount	String	Yes	Yes	The minimum number of ECS instances that can be created at a time.	Valid values: 1 to 100. The MinAmount parameter must be set to a value less than or equal to the value of MaxAmount.
Description	String	No	Yes	The description of created instances.	The description can be up to 256 characters in length.
InstanceType	String	Yes	Yes	The type of the ECS instance.	None.
ImageId	String	Yes	Yes	The ID of the image used to start an ECS instance. You can use a public image, a custom image, or an Alibaba Cloud Marketplace image.	You can specify a partial public image ID instead of providing the complete ID. The following example shows how to use a CNAME record: • Ubuntu is specified and ubuntu_16_0402_64_20G_alibase_20170818.vhd are matched. • If ubuntu1432 is specified, ubuntu_14_0405_32_40G_alibase_20170711.vhd is matched.
SecurityGroupId	String	No	No	The ID of the security group to which created instances belong.	None.

Attribute	Type	Required	Editable	Description	Constraint
InstanceName	String	No	No	The name of an instance.	The names can be up to 128 characters in length It can contain English letters, Chinese characters, digits, underscores (_), periods (.), and hyphens (-). By `name_prefix[begin_number,bits]name_suffix` format to specify different instance name for each ECS instance.
Password	String	No	Yes	The password used to log on to created ECS instances.	• The password must be 8 to 30 characters in length • and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. • The following special characters are supported `:( ) '~ ! @ # $ % ^ & * - + = | { } [ ] : ; ' < > , . ? /` If you specify the Password parameter in the API request, use HTTPS to secure the API and protect your password.
HostName	String	No	No	The hostname of created ECS instances.	The hostname must contain at least two characters in length. The periods and hyphens (-) cannot start or end the hostname and cannot be used together.
AutoReleaseTime	String	No	No	The time scheduled for created ECS instances to be automatically released.	The time format must comply with ISO8601 specifications, for example, `"yyyy-MM-ddTHH:mm:ssZ"`. The maximum release time must be within three years from the current time.

Attribute	Type	Required	Editable	Description	Constraint
PrivateIpAddress	String	No	No	The private IP address of an ECS instance in a VPC.	The specified IP address must not be used by other instances in the VPC.
DiskMappings	List	No	Yes	The data disks to be attached to created instances.	None.
InternetMaxBandwidthIn	Integer	No	No	The maximum inbound bandwidth from the Internet.	Unit: Mbit/s. Valid values: 1 to 100 Default value: 100.
IoOptimized	String	No	No	Specifies whether the created instances are I/O optimized.	Valid values: • none (non-I/O optimized) • optimized Default value: none.
SystemDiskCategory	String	No	Yes	The category of the system disk.	Valid values: cloud: basic disk cloud_efficiency: the ultra disk cloud_ssd: standard SSDs cloud_essd: enhanced SSD ephemeral_ssd: local SSDs
SystemDiskDescription	String	No	Yes	The description of the ECS instance system disk.	None.
SystemDiskDiskName	String	No	Yes	The name of the ECS instance system disk.	None.
SystemDiskSize	Number	No	Yes	The size of the system disk.	Valid values: 40 to 500. If a custom image is used to create a system disk, make sure that the size of the system disk is greater than that of the custom image.

Attribute	Type	Required	Editable	Description	Constraint
Tags	List	No	Yes	The custom tags of a created instance.	A maximum of 20 tags are supported. The format is as follows: `[{"Key":"tagKey","Value":"tagValue"},{"Key":"tagKey2","Value":"tagValue2"}]`.
UserData	String	No	Yes	The user data that you provide when you create ECS instances.	The user data can be up to 16 KB in size. You do not need to use Base64 for transcoding. Special characters need to be escaped.
ZoneId	String	No	No	The zone ID of the disk.	None.
VpcId	String	No	No	The ID of the VPC.	None.
VSwitchId	String	No	No	The ID of the VSwitch for the ECS instance.	None.
KeyPairName	String	No	Yes	The name of the key pair that is used to connect to created ECS instances.	For Windows-based ECS instances, this parameter is ignored. and it is empty by default. For Linux-based ECS instances, the Password parameter still takes effect if this parameter is specified. However, logon by password is disabled, and the KeyPairName value is used.
RamRoleName	String	No	Yes	The name of the instance RAM role.	You can call the ListRoles operation to query the role name.
DedicatedHostId	String	No	No	The ID of the dedicated host.	None.
LaunchTemplateName	String	No	Yes	The name of the launch template for the instance.	None.

Attribute	Type	Required	Editable	Description	Constraint
EniMappings	List	No	Yes	The elastic network interfaces (ENIs) to be attached to created instances.	Only one ENI can be attached to each instance.
LaunchTemplateId	String	No	Yes	The ID of the launch template.	None.
LaunchTemplateVersion	String	No	Yes	The version of the launch template.	If you do not specify a version, the default version is used.
NetworkType	String	No	No	The network type of created ECS instances.	Valid values: • vpc • classic Default value: classic.
DeletionProtection	Boolean	No	No	The release protection attribute of the instance. It specifies whether you can use the ECS console or call the DeleteInstance operation to release the instance.	Valid values: • true • false
DeploymentSetId	String	No	Yes	Deployment Set ID.	None.

DiskMappings syntax

```
"DiskMappings": [
  {
    "Category": String,
    "DiskName": String,
    "Description": String,
    "Device": String,
    "SnapshotId": String,
    "Size": String,
    "Encrypted": String,
    "KMSKeyId": String,
    "PerformanceLevel": String,
    "AutoSnapshotPolicyId": String
  }
]
```

DiskMappings properties

Attribute	Type	Required	Editable	Description	Constraint
Size	String	Yes	No	The size of a data disk.	Unit: GB.
Category	String	No	No	The type of the data disk.	Valid values: • cloud • cloud_efficiency • cloud_ssd • cloud_essd • ephemeral_ssd For I/O optimized instances, the default value is cloud_efficiency. For non-I/O optimized instances, the default value is cloud.
DiskName	String	No	No	The name of data disk N.	The name can be up to 128 characters in length. It can contain English letters, Chinese characters, digits, underscores (_), periods (.), and hyphens (-).
Description	String	No	No	The description of data disk N.	The description must be 2 to 256 characters in length. The description cannot start with `http://` or `https://`.
Device	String	No	No	The device name of the data disk.	The system allocates a device name in alphabetical order from /dev/xvda to /dev/xvdz.
SnapshotId	String	No	No	The ID of the snapshot.	None.
Encrypted	Boolean	No	No	Specifies whether to encrypt the data disk.	Valid values: • true • false Default value: false
KMSKeyId	String	No	No	The ID of the KMS key corresponding to the data disk.	None.
AutoSnapshotPolicyId	String	No	Yes	The ID of the automatic snapshot policy.	None.

Attribute	Type	Required	Editable	Description	Constraint
PerformanceLevel	String	No	No	The performance level of the enhanced SSD used as the data disk.	• (Default): Maximum random read /write IOPS of 50,000 per disk • PL2: A single enhanced SSD delivers up to 100,000 random read/write IOPS. • PL3: A single enhanced SSD delivers up to 1,000,000 random read/write IOPS.

Tags syntax

```
"Tags": [
  {
    "Value": String,
    "Key": String
  }
]
```

Tags properties

Attribute	Type	Required	Editable	Description	Constraint
Key	String	Yes	No	The key of tag N.	It must be 1 to 128 characters in length, and cannot start with `aliyun` and `acs:` beginning, cannot contain `http://` or `https://` .
Value	String	No	No	The value of tag N.	It must be 0 to 128 characters in length and cannot start with `aliyun` and `acs:` beginning, cannot contain `http://` or `https://` .

EniMappings syntax

```
"EniMappings": [
  {
    "SecurityGroupId": String,
    "VSwitchId": String,
    "Description": String,
    "NetworkInterfaceName": String,
    "PrimaryIpAddress": String
  }
]
```

EniMappings properties

Attribute	Type	Required	Editable	Description	Constraint
SecurityGroupId	String	Yes	Yes	The ID of the security group to which an instance belongs.	The security group and the instance must be in the same VPC.
VSwitchId	String	Yes	No	The ID of the VSwitch to which the instance is connected.	None.
Description	String	No	Yes	The description of the ENI.	It can contain 2 to 256 English letters or Chinese character. It cannot start with `http://` or `https://` the beginning.
NetworkInterfaceName	String	No	Yes	The ENI name.	• The name must be 2 to 128 characters in length • Must start with an uppercase or lowercase letter, and cannot start with `http://` or `https://` the beginning. • It must start with a letter and cannot start with http:// or https://.
PrimaryIpAddress	String	No	No	The primary private IP address of ENI N.	The specified IP address must be available within the CIDR block of the VSwitch. If this parameter is not specified, an available IP address in the VSwitch CIDR block will be selected at random.

Return value

Fn::GetAtt

- InstanceIds: the IDs of created instances in the ECS instance group. An ID is a system-generated globally unique identifier (GUID) for an instance.
- PrivateIps: the list of private IP addresses of instances in a VPC. This parameter takes effect only when the NetworkType parameter is set to VPC. For example, a Json-formatted Array: `["172.16.XX.XX", "172.16.XX.XX",&hellip; "172.16.XX.XX"]` the maximum number of IP addresses that can be specified. Separate multiple IP addresses with commas (,).
- InnerIps: the list of private IP addresses of instances in a classic network. This parameter takes effect only when

the NetworkType parameter is set to Classic. For example, a Json-formatted Array: `["10.1.XX.XX", "10.1.XX.XX",&hellip; "10.1.XX.XX"]` the maximum number of IP addresses that can be specified. Separate multiple IP addresses with commas (,).

- PublicIps: the list of public IP addresses of instances in a classic network. This parameter takes effect only when the NetworkType parameter is set to Classic. For example, a Json-formatted Array: `["42.1.XX.XX", "42.1.XX.XX",&hellip; "42.1.XX.XX"]` the maximum number of IP addresses that can be specified. Separate multiple IP addresses with commas (,).
- HostNames: the list of hostnames of all instances.
- OrderId: the list of order IDs of all instances.
- ZoneIds: the IDs of the zones where created instances reside.
- RelatedOrderIds: the list of related order IDs of created ECS instances.

Examples

```
{
  "ROSTemplateFormatVersion":"2015-09-01",
  "Resources":{
    "WebServer": {
      "Type": "ALIYUN::ECS::InstanceGroup",
      "Properties": {
        "ImageId":"m-25l0r****",
        "InstanceType": "ecs.t1.small",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "MaxAmount":1,
        "MinAmount":1,
        "Tags": [{
            "Key": "tiantt",
            "Value": "ros"
        },{
            "Key": "tiantt1",
            "Value": "ros1"
        }
        ]
      }
    }
  },
  "Outputs": {
    "InstanceIds": {
         "Value":{"get_attr": ["WebServer","InstanceIds"]}
    },
    "PublicIps": {
         "Value":{"get_attr": ["WebServer","PublicIps"]}
    }
  }
}
```

5.1.6.1.12. ALIYUN::ECS::InstanceGroupClone

ALIYUN::ECS::InstanceGroupClone is used to clone an ECS instance group.

Statement

```
{
  "Type": "ALIYUN::ECS::InstanceGroupClone",
  "Properties": {
    "BackendServerWeight": Integer,
    "DiskMappings": List,
    "LaunchTemplateName": String,
    "SpotPriceLimit": String,
    "ResourceGroupId": String,
    "KeyPairName": String,
    "SystemDiskDiskName": String,
    "PeriodUnit": String,
    "Description": String,
    "Tags": List,
    "ImageId": String,
    "SpotStrategy": String,
    "SourceInstanceId": String,
    "EniMappings": List,
    "Password": String,
    "MaxAmount": Integer,
    "AutoReleaseTime": String,
    "SystemDiskCategory": String,
    "LoadBalancerIdToAttach": String,
    "LaunchTemplateId": String,
    "LaunchTemplateVersion": String,
    "ZoneId": String,
    "InstanceName": String,
    "InternetMaxBandwidthIn": Integer,
    "DeletionProtection": Boolean,
    "DeploymentSetId": String,
    "SecurityGroupId": String,
    "RamRoleName": String,
    "HpcClusterId": String,
    "SystemDiskDescription": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the enterprise resource group to which a created instance belongs.	None
HpcClusterId	String	Yes	True	The ID of the E-HPC cluster to which a created instance belongs.	None
SourceInstanceId	String	No	No	The ID of the ECS instance to be cloned.	The clone operation clones the specified instance, including its instance type, image, bandwidth limit, and network type. If the source ECS instance belongs to multiple security groups, the cloned instance is added only to the first of these security groups.

Parameter	Type	Required	Editable	Description	Constraint
MaxAmount	Integer	Retained	Yes	The maximum number of ECS instances to be created.	Valid values: 1 to 100. The MaxAmount parameter must be set to a value greater than or equal to the value of the MinAmount parameter.
MinAmount	String	No	Yes	The minimum number of ECS instances to be created.	Valid values: 1 to 100. The MinAmount parameter must be set to a value less than or equal to the value of the MaxAmount parameter.
BackendServerWeight	String	Optional	Released	The weight assigned to the ECS instance in the Server Load Balancer instance.	Valid values: 0 to 100. Default value: 100.
LoadBalancerIdToAttach	String	Yes	Released	The ID of the SLB instance to which the ECS instance is to be attached.	None
Description	String	Yes	Released	The description of created instances.	The description can be up to 256 characters in length.

Parameter	Type	Required	Editable	Description	Constraint
ImageId	String	Yes	True	The ID of the image used to start created instances. You can use a public image, a custom image, or an Alibaba Cloud Marketplace image.	You can specify a partial public image ID instead of providing the complete ID. When editing a template used to deploy an ECS instance, you can specify the image type and version or only the image type. ROS automatically selects an appropriate public image ID. You can use the wildcard character (*) to represent part of an image ID. You can use one of the following methods to specify the public image ID for the ECS instance: • If you set the parameter to ubuntu, ubuntu_16_0402_64_20G_alibase_20170818.vhd. • If this parameter is set to ubuntu_14, ubuntu_14_0405_64_20G_alibase_20170824.vhd is returned. • Specify: ubuntu1432, which will eventually match: ubuntu_14_0405_32_40G_alibase_20170711.vhd • Specify: ubuntu_16_0402_32, which will eventually match: ubuntu_16_0402_32_40G_alibase_20170711.vhd
SecurityGroupId	String	Yes	Released	The ID of the security group to which created instances belong.	None
InstanceName	String	Yes	Released	The name of a created instance.	The name can be up to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-).
Password	String	Yes	Released	The password used to log on to the ECS instance.	The password The password must be 8 to 30 characters in length and must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special character. Special characters include ( ) ` ~ ! @ # $ % ^ & * - + = | { } [ ] : ; ' < > , . ? / If the Password parameter is specified, you must use HTTPS to call the operation to ensure that the Password remains confidential.
DiskMappings	List	Erased	Released	The data disks to be attached to the instance.	A maximum of 16 disks can be attached to each instance.
Tags	List	Erased	Released	The custom tags of the instance.	You can specify a maximum of 20 tags. The format is as follows: `[{"Key":"tagKey","Value":"tagValue"},{"Key":"tagKey2","Value":"tagValue2"}]`.

Parameter	Type	Required	Editable	Description	Constraint
ZoneId	String	Yes	Released	The ID of the zone where the instance resides.	None
KeyPairName	String	Yes	Released	The name of the key pair that is used to connect to the ECS instance. For Windows-based ECS instances, this parameter is ignored. Default value: empty. For Linux-based instances, the Password parameter still takes effect if this parameter is specified. However, logon by Password is disabled, and the KeyPairName value is used.	None
RamRoleName	String	Yes	Released	The RAM role name of the instance.	You can use RAM API ListRoles you can call this operation to query the RAM role name of an instance.
SpotPriceLimit	String	Yes	Released	The maximum hourly price of the instance.	This parameter supports up to three decimal places. Parameter SpotStrategy this parameter takes effect only when the value is SpotWithPriceLimit.
SystemDiskDiskName	String	Yes	True	The name of the system disk of created instances.	The name must be 2 to 128 characters in length Must start with an uppercase or lowercase letter, and cannot start with `http://` or `https://` and can contain digits, colons (:), underscores (_), and hyphens (-).
PeriodUnit	String	Yes	True	The billing cycle for created ECS instances.	Valid values: • Week. 1, 2, 3, and 4} when the value of the PeriodUnit parameter is Week. AutoRenewPeriod values are 1, 2, "3"}. • Month 1, 2, 3, 4, 5, 6, and 7 when the PeriodUnit parameter is set to Month "8", "9", "12", "24", "36", "48", "60"},AutoRenewPeriod can be {"1", "2", "3", "6", "12"}. Default value: Month.
EniMappings	List	No.	True	The elastic network interfaces (ENIs) to be attached to a created instance.	Only a single ENI can be attached to each instance.

Parameter	Type	Required	Editable	Description	Constraint
AutoReleaseTime	String	Yes	Released	The time scheduled for a created ECS instance to be automatically released. Specify the time in the ISO 8601 standard in the YYYY-MM-DDThh:mmZ format. in the yyyy-MM-ddTHH:mm:ssZ format. The time must be in UTC. • If the value of seconds (ss) is a value other than 00, the start time is automatically rounded down to the nearest minute based on the value of mm. • The minimum release time must be at least 30 minutes later than the current time. • The maximum release time must be at most three years from the current time. If you do not specify the AutoReleaseTime it indicates that the auto release feature is disabled and the ECS instance will not be automatically released.	None

Parameter	Type	Required	Editable	Description	Constraint
SystemDiskCategory	String	Yes	True	The type of the system disk.	Valid values: • cloud: basic disk • cloud_efficiency: indicates an ultra disk. • cloud_ssd: indicates a standard SSD. • ephemeral_ssd: local SSD. • cloud_essd: indicates an enhanced SSD (ESSD). ESSDs are still in public preview and only available in some regions. For phased-out instance types that are not I/O optimized, the default value is cloud. For other instances, the default value is cloud_efficiency.
LaunchTemplateName	String	Yes	True	The name of the launch template.	None
LaunchTemplateVersion	String	Yes	True	The version of the launch template. If you do not specify this parameter, the default version is used.	None
InternetMaxBandwidthIn	String	Optional	Released	The maximum inbound bandwidth from the Internet. Unit: Mbit/s.	Valid values: 1 to 200. Default value: 200.
LaunchTemplateId	String	Yes	True	The ID of the launch template.	None
SystemDiskDescription	String	Yes	True	The description of the system disk.	The description must be 2 to 256 characters in length and cannot start with http:// or https://. This parameter is empty by default.
DeletionProtection	Boolean	Erased	Released	The release protection attribute of the instance. Specifies whether the ECS console or API DeleteInstance) to release the instance.	Valid values: • true • false
DeploymentSetId	String	Yes	True	Deployment Set ID.	None

DiskMappings syntax

```
"DiskMappings": [
  {
    "Category": String,
    "DiskName": String,
    "Description": String,
    "Encrypted": String,
    "KMSKeyId": String,
    "Device": String,
    "SnapshotId": String,
    "Size": String
  }
]
```

DiskMappings properties

Parameter	Type	Required	Editable	Description	Constraint
Encrypted	String	Yes	Released	Specifies whether to encrypt the data disk.	Valid values: • true • false Default value :false
KMSKeyId	String	Yes	Released	The KMS key ID for data disk N.	None
Size	String	No	No	The size of data disk N. Unit: GB.	None
Category	String	Yes	Released	The type of the data disk.	Valid values: • cloud • cloud_efficiency • cloud_ssd • ephemeral_ssdDefault
DiskName	String	Yes	Released	The name of data disk N.	The name can be up to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-).
Description	String	Yes	Released	The description of data disk N.	Valid values: 2 to 256. Default value: Null.
Device	String	Yes	Released	The device name of the data disk attached to an ECS instance.	The system allocates a device name in alphabetical order from /dev/xvda to /dev/xvdz.
SnapshotId	String	Yes	Released	Create a data disk by using a snapshot.	None

EniMappings syntax

```
"EniMappings": [
  {
    "SecurityGroupId": String,
    "VSwitchId": String,
    "Description": String,
    "NetworkInterfaceName": String,
    "PrimaryIpAddress": String
  }
]
```

EniMappings properties

Parameter	Type	Required	Editable	Description	Constraint
SecurityGroupId	String	No	Yes	The ID of the security group to which the ENI belongs.	None
VSwitchId	String	No	No	The ID of the VSwitch to which the ENI belongs.	None
Description	String	Yes	True	The description of the ENI.	It can contain 2 to 256 English letters or Chinese character. It cannot start with `http://` and `https://` the beginning.
NetworkInterface Name	String	Yes	True	The name of the ENI.	None
PrimaryIpAddress	String	Yes	Released	The primary IP address of the ENI.	None

Tags syntax

```
"Tags": [
  {
    "Value": String,
    "Key": String
  }
]
```

Tags properties

Parameter	Type	Required	Editable	Description	Constraint
Key	String	No	No	None	None

Parameter	Type	Required	Editable	Description	Constraint
Value	String	Yes	Released	None	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- InstanceIds: the IDs of instances in the ECS instance group. An ID is a globally unique identifier (GUID) generated by the system for an instance.
- PrivateIps: the private IP addresses of VPC-type instances. This parameter is valid only when the NetworkType parameter is set to VPC. The parameter value is a JSON-formatted array, containing up to 100 IP addresses separated by commas (,). Example: ["172.16.XX.XX", "172.16.XX.XX", ... "172.16.XX.XX"].
- InnerIps: the private IP addresses of instances in the classic network. This parameter is valid only when the NetworkType parameter is set to Classic. The parameter value is a JSON-formatted array, containing up to 100 IP addresses separated by commas (,). Example: ["10.1.XX.XX", "10.1.XX.XX", ... "10.1.XX.XX"].
- PublicIps: the public IP addresses of instances in the classic network. This parameter is applicable only when the NetworkType parameter is set is Classic. The parameter value is a JSON-formatted array, containing up to 100 IP addresses separated by commas (,). Example: ["42.1.XX.XX", "42.1.XX.XX", ... "42.1.XX.XX"].
- HostNames: the hostnames of all instances. The parameter value is a JSON-formatted array. Example: ["host1", "host2", ... "host3"].
- OrderId: the order IDs of all instances.
- ZoneIds: the IDs of the zones where created instances reside.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "WebServer": {
      "Type": "ALIYUN::ECS::InstanceGroupClone",
      "Properties": {
        "SourceInstanceId": "i-25zsk****",
        "ImageId": "m-25l0r****",
        "SecurityGroupId": "sg-25zwc****",
        "ZoneId": "cn-beijing-b",
        "MaxAmount": 1,
        "MinAmount": 1
      }
    }
  },
  "Outputs": {
    "InstanceIds": {
      "Value": {"get_attr": ["WebServer","InstanceIds"]}
    },
    "PublicIps": {
      "Value": {"get_attr": ["WebServer","PublicIps"]}
    }
  }
}
```

5.1.6.1.13. ALIYUN::ECS::Invocation

ALIYUN::ECS::Invocation is used to invoke a Cloud Assistant command for one or more ECS instances.

Statement

```
{
  "Type": "ALIYUN::ECS::Invocation",
  "Properties": {
    "Timed": Boolean,
    "Frequency": String,
    "CommandId": String,
    "InstanceIds": List
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
Timed	Boolean	Erased	Released	Specifies whether to invoke the command on a periodic basis. Default value: false.	None
Frequency	String	Yes	Released	The frequency at which the command is invoked.	None

Parameter	Type	Required	Editable	Description	Constraint
CommandId	String	No	No	The ID of the script.	None
InstanceIds	List	Yes	No	The IDs of the instances for which you want to invoke the command. A maximum of 20 instance IDs can be specified.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

The execution ID of the InvokeId: command.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Timed": {
      "Type": "Boolean",
      "Description": "Whether it is timed execution. Default is False.",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    },
    "Frequency": {
      "Type": "String",
      "Description": "The frequency of timing execution (the shortest frequency is performed every 1 minute). It is mandatory when Timing is True.The value rule follows the rules of the cron expression. "
    },
    "CommandId": {
      "Type": "String",
      "Description": "The id of command."
    },
    "InstanceIds": {
      "Type": "CommaDelimitedList",
      "Description": "The instance id list. Select up to 20 instances at a time.Instances selected network type must be VPC network, status must be running"
    }
  },
  "Resources": {
    "Invocation": {
      "Type": "ALIYUN::ECS::Invocation",
      "Properties": {
        "Timed": {
          "Ref": "Timed"
        },
        "Frequency": {
          "Ref": "Frequency"
```

```
            "Ref": "Frequency"
          },
          "CommandId": {
            "Ref": "CommandId"
          },
          "InstanceIds": {
            "Fn::Split": [
              ",",
              {
                "Ref": "InstanceIds"
              },
              {
                "Ref": "InstanceIds"
              }
            ]
          }
        }
      }
    },
    "Outputs": {
      "InvokeId": {
        "Description": "The id of command execution.",
        "Value": {
          "Fn::GetAtt": [
            "Invocation",
            "InvokeId"
          ]
        }
      }
    }
}
```

5.1.6.1.14. ALIYUN::ECS::JoinSecurityGroup

ALIYUN::ECS::JoinSecurityGroup is used to add one or more ECS instances to a specified security group.

Syntax

```
{
  "Type": "ALIYUN::ECS::JoinSecurityGroup",
  "Properties": {
    "InstanceId": String,
    "InstanceIdList": List,
    "SecurityGroupId": String,
    "NetworkInterfaceList": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
SecurityGroupId	String	Yes	No	The ID of the security group.	None

Property	Type	Required	Editable	Description	Constraint
InstanceId	String	No	No	The ID of the ECS instance to be added to the security group.	None
InstanceIdList	List	No	Yes	The IDs of the ECS instances to be added to the security group.	None
NetworkInterfaceList	List	No	Yes	The IDs of the elastic network interfaces (ENIs).	None

Response parameters

Fn::GetAtt

None

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SG": {
      "Type": "ALIYUN::ECS::JoinSecurityGroup",
      "Properties": {
        "SecurityGroupId": "sg-m5eagh7rzys2z8sa****",
        "InstanceIdList": [
          "i-m5e505h9bgsio0wy****",
          "i-m5e505hio0wyjc6r****"
        ]
      }
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Resources:
  SG:
    Type: ALIYUN::ECS::JoinSecurityGroup
    Properties:
      SecurityGroupId: sg-m5eagh7rzys2z8sa****
      InstanceIdList:
      - i-m5e505h9bgsio0wy****
      - i-m5e505hio0wyjc6r****
```

5.1.6.1.15. ALIYUN::ECS::LaunchTemplate

ALIYUN::ECS::LaunchTemplate is used to create an ECS instance launch template.

Syntax

```
{
  "Type": "ALIYUN::ECS::LaunchTemplate",
  "Properties": {
    "LaunchTemplateName": String,
    "VersionDescription": String,
    "ImageId": String,
    "InstanceType": String,
    "SecurityGroupId": String,
    "NetworkType": String,
    "VSwitchId": String,
    "InstanceName": String,
    "Description": String,
    "InternetMaxBandwidthIn": Integer,
    "InternetMaxBandwidthOut": Integer,
    "HostName": String,
    "ZoneId": String,
    "SystemDiskCategory": String,
    "SystemDiskSize": Number,
    "SystemDiskDiskName": String,
    "SystemDiskDescription": String,
    "IoOptimized": String,
    "InternetChargeType": String,
    "UserData": String,
    "KeyPairName": String,
    "RamRoleName": String,
    "AutoReleaseTime": String,
    "SpotStrategy": String,
    "SpotPriceLimit": String,
    "SecurityEnhancementStrategy": String,
    "DiskMappings": List,
    "NetworkInterfaces": List,
    "Tags": List,
    "TemplateTags": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
LaunchTemplateName	String	Yes	No	The name of the instance launch template.	• The name must be 2 to 128 characters in length. • It must start with a letter but cannot start with `http://` or `https://`. • It can contain letters, digits, colons (:), underscores (_), and hyphens (-).

Property	Type	Required	Editable	Description	Constraint
VersionDescription	String	No	No	The description of the version of the instance launch template.	• The description must be 2 to 128 characters in length. • It must start with a letter but cannot start with `http://` or `https://`.
ImageId	String	No	No	The ID of the image.	None
InstanceType	String	No	No	The type of the instance.	None
SecurityGroupId	String	No	No	The ID of the security group.	None
NetworkType	String	No	No	The network type of the instance.	Valid values: • classic • vpc
VSwitchId	String	No	No	The ID of the VSwitch. You must specify this parameter when you create an instance in a VPC.	None
InstanceName	String	No	No	The name of the instance.	• The name must be 2 to 128 characters in length. • It must start with a letter but cannot start with `http://` or `https://`.
Description	String	No	No	The description of the instance.	• The description must be 2 to 128 characters in length. • It must start with a letter but cannot start with `http://` or `https://`.
InternetMaxBandwidthIn	Integer	No	No	Maximum inbound bandwidth from the Internet.	Valid values: 1 to 200. Unit: Mbit/s.
InternetMaxBandwidthOut	Integer	No	No	Maximum outbound bandwidth to the Internet.	Valid values: 0 to 100. Unit: Mbit/s.

Property	Type	Required	Editable	Description	Constraint
HostName	String	No	No	The hostname of the instance.	The name cannot start or end with a period (.) or a hyphen (-). It cannot contain consecutive periods (.) or hyphens (-). • For Windows instances: ◦ The name must be 2 to 15 characters in length and can contain letters, digits, and hyphens (-). ◦ It cannot only contain digits. • For other instances such as Linux instances: ◦ The name must be 2 to 64 characters in length and can contain letters, digits, and hyphens (-).
ZoneId	String	No	No	The ID of the zone where the instance resides.	None
SystemDiskCategory	String	No	No	The category of the system disk.	Valid values: • cloud: the basic disk • cloud_efficiency: the ultra disk • cloud_ssd: the standard SSD • ephemeral_ssd: the local SSD
SystemDiskSize	Number	No	No	The size of the system disk.	Valid values: 20 to 500. Unit: GiB.
SystemDiskDiskName	String	No	No	The name of the system disk.	• The name must be 2 to 128 characters in length and can contain letters, digits, colons (:), underscores (_), and hyphens (-). • It must start with a letter but cannot start with `http://` or `https://`.

Property	Type	Required	Editable	Description	Constraint
SystemDiskDescription	String	No	No	The description of the system disk.	The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.
IoOptimized	String	No	No	Specifies whether the instance is I/O optimized.	Valid values: • none • optimized
InternetChargeType	String	No	No	The billing method for network usage.	Valid values: • PayByBandwidth • PayByTraffic
UserData	String	No	No	The user data of the instance.	The data must be encoded in Base64. The maximum size of the raw data is 16 KB.
KeyPairName	String	No	No	The AccessKey pair name.	• For Windows instances, this parameter is ignored and is empty by default. The Password parameter takes effect even if the KeyPairName parameter is specified. • For Linux instances, the username and password authentication method is disabled by default.
RamRoleName	String	No	No	The RAM role name of the instance.	None
AutoReleaseTime	String	No	No	The time scheduled for the instance to be automatically released.	Specify the time in the ISO 8601 standard in the `yyyy-MM-ddTHH:mm:ssZ` format. The time must be in UTC.

Property	Type	Required	Editable	Description	Constraint
SpotStrategy	String	No	No	The preemption policy for pay-as-you-go instances.	This parameter takes effect only when the InstanceChargeType parameter is set to PostPaid. Valid values: • NoSpot: The instance is created as a regular pay-as-you-go instance. • SpotWithPriceLimit: The instance to be created is a preemptible instance with a user-defined maximum hourly price. • SpotAsPriceGo: The instance to be created is a preemptible instance whose price is based on the market price at the time of purchase.
SpotPriceLimit	String	No	No	The maximum hourly price of the instance.	A maximum of three decimal places can be specified.
SecurityEnhancementStrategy	String	No	No	Specifies whether to enable security hardening.	Valid values: • Active: enables security hardening. • Deactive: disables security hardening.
DiskMappings	List	No	No	The list of data disks.	A maximum of 16 data disks can be specified.
NetworkInterfaces	List	No	No	The list of elastic network interfaces (ENIs).	A maximum of eight ENIs can be specified.
Tags	List	No	No	The tags of the instance, security group, disks, and ENIs.	A maximum of 20 tags can be specified.
TemplateTags	List	No	No	The tags of the launch template.	A maximum of 20 tags can be specified.

DiskMappings syntax

```
"DiskMappings": [
  {
    "Category": String,
    "DiskName": String,
    "Description": String,
    "SnapshotId": String,
    "Size": String,
    "Encrypted": String,
    "DeleteWithInstance": String
  }
]
```

DiskMappings properties

Property	Type	Required	Editable	Description	Constraint
Category	String	No	No	The category of the data disk.	Valid values: • cloud: the basic disk • cloud_efficiency: the ultra disk • cloud_ssd: the standard SSD • ephemeral_ssd: the local SSD
DiskName	String	No	No	The name of the data disk.	• The name must be 2 to 128 characters in length. • It must start with a letter but cannot start with `http://` or `https://`. • It can contain letters, digits, colons (:), underscores (_), and hyphens (-).
Description	String	No	No	The description of the data disk.	The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.
SnapshotId	String	No	No	The ID of the snapshot used to create the data disk.	None

Property	Type	Required	Editable	Description	Constraint
Size	String	No	No	The size of the system disk.	• Valid values when the Category parameter is set to cloud: 5 to 2000. • Valid values when the Category parameter is set to cloud_efficiency: 20 to 32768. • Valid values when the Category parameter is set to cloud_ssd: 20 to 32768. • Valid values when the Category parameter is set to ephemeral_ssd: 5 to 800. Unit: GiB.
Encrypted	Boolean	No	No	Specifies whether to encrypt the data disk.	None
DeleteWithInstance	Boolean	No	No	Specifies whether to release the data disk when the instance is released.	None

NetworkInterfaces syntax

```
"NetworkInterfaces": [
  {
    "PrimaryIpAddress": String,
    "VSwitchId": String,
    "SecurityGroupId": String,
    "NetworkInterfaceName": String,
    "Description": String
  }
]
```

NetworkInterfaces properties

Property	Type	Required	Editable	Description	Constraint
PrimaryIpAddress	String	No	No	The primary private IP address of the ENI.	None
VSwitchId	String	No	No	The ID of the VSwitch to which the ENI belongs.	None
SecurityGroupId	String	No	No	The ID of the security group to which the ENI belongs.	None
NetworkInterfaceName	String	No	No	The name of the ENI.	None

Property	Type	Required	Editable	Description	Constraint
Description	String	No	No	The description of the ENI.	The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

Tags syntax

```
"Tags": [
  {
    "Value": String,
    "Key": String
  }
]
```

Tags properties

Property	Type	Required	Editable	Description	Constraint
key	String	Yes	No	None	None
value	String	No	No	None	None

TemplateTags syntax

```
"TemplateTags": [
  {
    "Value": String,
    "Key": String
  }
]
```

TemplateTags properties

Property	Type	Required	Editable	Description	Constraint
key	String	Yes	No	None	None
value	String	No	No	None	None

Response parameters

Fn::GetAtt

- LaunchTemplateId: the ID of the instance launch template.
- LaunchTemplateName: the name of the instance launch template.
- DefaultVersionNumber: the default version number of the instance launch template.
- LatestVersionNumber: the latest version number of the instance launch template.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
```

```
  "Template1": {
    "Type": "ALIYUN::ECS::LaunchTemplate",
    "Properties": {
      "LaunchTemplateName": "MyTemplate",
      "VersionDescription": "Launch template with all properties set",
      "ImageId": "m-2ze9uqi7wo61hwep****",
      "InstanceType": "ecs.n4.small",
      "SecurityGroupId": "sg-2ze8yxgempcdsq3i****",
      "NetworkType": "vpc",
      "VSwitchId": "vsw-2zei67xd9nhcqxzec****",
      "InstanceName": "InstanceName",
      "Description": "Description of template",
      "InternetMaxBandwidthIn": 100,
      "InternetMaxBandwidthOut": 200,
      "HostName": "tttinfo",
      "ZoneId": "cn-beijing-a",
      "SystemDiskCategory": "cloud_ssd",
      "SystemDiskSize": "40",
      "SystemDiskDiskName": "TheSystemDiskName",
      "SystemDiskDescription": "The system disk description",
      "IoOptimized": "optimized",
      "InternetChargeType": "PayByBandwidth",
      "UserData": "dGhpcyBpcyBhIHVzZXIgZGF0YSBleG1h****",
      "KeyPairName": "ThisIsKeyPair",
      "RamRoleName": "ThisIsRamRole",
      "AutoReleaseTime": "2050-10-01T00:00:00Z",
      "SpotStrategy": "SpotWithPriceLimit",
      "SpotPriceLimit": "100.001",
      "SecurityEnhancementStrategy": "Active",
      "DiskMappings": [
        {
          "Category": "cloud_ssd",
          "Size": 40,
          "SnapshotId": "s-2ze1fr2bipove27b****",
          "Encrypted": true,
          "DiskName": "dataDisk1",
          "Description": "I am data disk 1",
          "DeleteWithInstance": true
    },
    {
          "Category": "cloud_efficiency",
          "Size": 20,
          "SnapshotId": "s-2ze4k0w8b33mlsqu****",
          "Encrypted": false,
          "DiskName": "dataDisk2",
          "Description": "I am data disk 2",
          "DeleteWithInstance": true
      }
    ],
    "NetworkInterfaces": [
      {
          "PrimaryIpAddress": "10.10.1.1",
          "VSwitchId": "vsw-2zetgeiqlemyok9z5****",
          "SecurityGroupId": "sg-2ze8yxgempcdsq3i****",
          "NetworkInterfaceName": "my-eni1",
          "Description": "My eni 1"
       },
    ],
    "Tags": [
        {
```

```
            "Key": "key1",
            "Value": "value1"
          },
          {
            "Key": "key2",
            "Value": "value2"
          }
        ],
    "TemplateTags": [
      {
           "Key": "templateKey1",
           "Value": "templateValue1"
            },
          {
           "Key": "templateKey2",
           "Value": "templateValue2"
          }
      ]
      }
    }
  },
  "Outputs": {
      "LaunchTemplateId": {
          "Value": {"Fn::GetAtt": ["Template1","LaunchTemplateId"]}
    },
   "LaunchTemplateName": {
        "Value": {"Fn::GetAtt": ["Template1","LaunchTemplateName"]}
    },
   "DefaultVersionNumber": {
          "Value": {"Fn::GetAtt": ["Template1","DefaultVersionNumber"]}
    },
   "LatestVersionNumber": {
          "Value": {"Fn::GetAtt": ["Template1","LatestVersionNumber"]}
    }
  }
}
```

5.1.6.1.16. ALIYUN::ECS::NatGateway

ALIYUN::ECS::NatGateway is used to create a NAT Gateway for a VPC.

Statement

```
{
  "Type": "ALIYUN::ECS::NatGateway",
  "Properties": {
    "DeletionProtection": Boolean,
    "VpcId": String,
    "Description": String,
    "NatGatewayName": String,
    "VSwitchId": String,
    "DeletionForce": Boolean,
    "Spec": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
VpcId	String	Yse	No	The ID of the VPC that you want to create NAT Gateway.	None
VSwitchId	String	Yse	No	The ID of the vSwitch in the specified VPC.	None
Description	String	Erased	Released	The description of the NAT Gateway.	The description must be 2 to 256 characters in length.
NatGatewayName	String	Erased	Released	The name of the NAT Gateway.	The name must be 2 to 128 characters in length and can contain letters, digits, underscores (_), periods (.), and hyphens (-). It must start with a letter.
Spec	String	Erased	Released	The type of the NAT Gateway.	Valid values: Small,Middle, and Large.
DeletionProtection	Boolean	Erased	Released	Indicates whether deletion protection is enabled. Default value: false.	None
DeletionForce	Boolean	Erased	Released	Specifies whether to forcibly delete SNAT and DNAT entries in the Gateway and unbind EIP from the NAT gateway. Default value: false.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- ForwardTableId: the ID of the port forwarding table.
- The ID of the SNatTableId:SNat source network address translation table.
- NatGatewayId: the unique ID of the Nat gateway.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "NatGateway": {
      "Type": "ALIYUN::ECS::NatGateway",
      "Properties": {
        "NatGatewayName": "nat_gateway_1",
        "Description": "my nat gateway",
        "VpcId": "vpc-25o8s****",
        "VSwitchId": "vsw-25rc1****",
        "Spec": "Small"
      }
    }
  },
  "Outputs": {
    "NatGatewayId": {
         "Value": {"Fn::GetAttr": ["NatGateway","NatGatewayId"]}
    },
    "ForwardTableId": {
         "Value": {"Fn::GetAttr": ["NatGateway","ForwardTableId"]}
    },
    "SNatTableId": {
         "Value": {"Fn::GetAttr": ["NatGateway", "SNatTableId"]}
    }
  }
}
```

5.1.6.1.17. ALIYUN::ECS::NetworkInterface

ALIYUN::ECS::NetworkInterface is used to create an elastic network interface (ENI).

Statement

```
{
  "Type": "ALIYUN::ECS::NetworkInterface",
  "Properties": {
    "Description": String,
    "SecurityGroupId": String,
    "PrimaryIpAddress": String,
    "ResourceGroupId": String,
    "VSwitchId": String,
    "NetworkInterfaceName": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None

Parameter	Type	Required	Editable	Description	Constraint
SecurityGroupId	String	No	Yes	The ID of the security group to which the instance belongs. The security group and the instance must be in the same VPC.	None
VSwitchId	String	No	No	The ID of the VSwitch in the VPC.	None
Description	String	Yes	True	The description of the ENI. It can contain 2 to 256 English letters or Chinese character. It cannot start with `http://` and `https://` the beginning. This parameter is empty by default.	None

Parameter	Type	Required	Editable	Description	Constraint
NetworkInterfaceName	String	Yes	True	The name of the ENI. The name must be 2 to 128 characters in length Must start with an uppercase or lowercase letter, and cannot start with `http://` and `https://` the beginning. It can contain letters, digits, colons (:), underscores (_), and hyphens (-). Default value: null.	None
PrimaryIpAddress	String	Yes	Released	The primary private IP address of the ENI. The specified IP address must be available within the CIDR block of the VSwitch. If this parameter is not specified, an available IP address in the VSwitch CIDR block is assigned at random.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- NetworkInterfaceId: the ID of the ENI.
- The MAC address of the MacAddress: Elastic Network Interface.

- The private IP address of the PrivateIpAddress: Elastic Network Interface.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Description": {
      "Type": "String",
      "Description": "Description of your ENI. It is a string of [2, 256] English or Chinese characters."
    },
    "SecurityGroupId": {
      "Type": "String",
      "Description": "The ID of the security group that the ENI joins. The security group and the ENI must be in a same VPC."
    },
    "VSwitchId": {
      "Type": "String",
      "Description": "VSwitch ID of the specified VPC. Specifies the switch ID for the VPC."
    },
    "NetworkInterfaceName": {
      "Type": "String",
      "Description": "Name of your ENI. It is a string of [2, 128]  Chinese or English characters. It must begin with a letter and can contain numbers, underscores (_), colons (:), or hyphens (-)."
    },
    "PrimaryIpAddress": {
      "Type": "String",
      "Description": "The primary private IP address of the ENI.  The specified IP address must have the same Host ID as the VSwitch. If no IP addresses are specified, a random network ID is assigned for the ENI."
    }
  },
  "Resources": {
    "EniInstance": {
      "Type": "ALIYUN::ECS::NetworkInterface",
      "Properties": {
        "Description": {
          "Ref": "Description"
        },
        "SecurityGroupId": {
          "Ref": "SecurityGroupId"
        },
        "VSwitchId": {
          "Ref": "VSwitchId"
        },
        "NetworkInterfaceName": {
          "Ref": "NetworkInterfaceName"
        },
        "PrimaryIpAddress": {
          "Ref": "PrimaryIpAddress"
        }
      }
    }
  },
  "Outputs": {
    "NetworkInterfaceId": {
      "Description": "ID of your Network Interface.",
      "Value": {
        "Fn::GetAtt": [
```

```
            "EniInstance",
            "NetworkInterfaceId"
          ]
        }
      }
    }
}
```

5.1.6.1.18. ALIYUN::ECS::NetworkInterfaceAttachment

ALIYUN::ECS::NetworkInterfaceAttachment is used to attach an elastic network interface (ENI) to an instance in a VPC.

Statement

```
{
  "Type": "ALIYUN::ECS::NetworkInterfaceAttachment",
  "Properties": {
    "InstanceId": String,
    "NetworkInterfaceId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
InstanceId	String	No	No	The ID of the RDS instance.	None
NetworkInterfaceId	String	No	No	The IDs of the elastic network interfaces (ENIs).	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

NetworkInterfaceId: the ID of the ENI.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "InstanceId": {
      "Type": "String",
      "Description": "ECS instance id"
    },
    "NetworkInterfaceId": {
      "Type": "String",
      "Description": "Network interface id"
    }
  },
  "Resources": {
    "EniAttachment": {
      "Type": "ALIYUN::ECS::NetworkInterfaceAttachment",
      "Properties": {
        "InstanceId": {
          "Ref": "InstanceId"
        },
        "NetworkInterfaceId": {
          "Ref": "NetworkInterfaceId"
        }
      }
    }
  },
  "Outputs": {
    "NetworkInterfaceId": {
      "Description": "ID of your Network Interface.",
      "Value": {
        "Fn::GetAtt": [
          "EniAttachment",
          "NetworkInterfaceId"
        ]
      }
    }
  }
}
```

5.1.6.1.19. ALIYUN::ECS::NetworkInterfacePermission

ALIYUN::ECS::NetworkInterfacePermission is used to grant an account the permission to attach an elastic network interface (ENI) to an instance.

Syntax

```
{
  "Type": "ALIYUN::ECS::NetworkInterfacePermission",
  "Properties": {
    "NetworkInterfaceId": String,
    "AccountId": String,
    "Permission": String
  }
}
```

Properties

Name	Type	Required	Editable	Description	Validity
NetworkInterfaceId	String	Yes	No	The ID of the ENI.	None
AccountId	String	Yes	No	The ID of the account.	None
Permission	String	Yes	No	The permission granted to the account.	None

Response parameters

Fn::GetAtt

- NetworkInterfacePermissionId: the ID of the ENI permission.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "AccountId": {
      "Type": "String",
      "Description": "the account id"
    },
    "Permission": {
      "Type": "String",
      "Description": "the permission",
      "Default": "InstanceAttach"
    },
    "NetworkInterfaceId": {
      "Type": "String",
      "Description": "Network interface id"
    }
  },
  "Resources": {
    "EniPermission": {
      "Type": "ALIYUN::ECS::NetworkInterfacePermission",
      "Properties": {
        "AccountId": {
          "Ref": "AccountId"
        },
        "Permission": {
          "Ref": "Permission"
        },
        "NetworkInterfaceId": {
          "Ref": "NetworkInterfaceId"
        }
      }
    }
  },
  "Outputs": {
    "NetworkInterfacePermissionId": {
      "Description": "the network interface permission id",
      "Value": {
        "Fn::GetAtt": [
          "EniPermission",
          "NetworkInterfacePermissionId"
        ]
      }
    }
  }
}
```

5.1.6.1.20. ALIYUN::ECS::Route

ALIYUN::ECS::Route is used to create a custom route.

Syntax

```
{
  "Type": "ALIYUN::ECS::Route",
  "Properties": {
    "DestinationCidrBlock": String,
    "RouteTableId": String,
    "NextHopId": String,
    "NextHopType": String,
    "RouteId": String,
    "NextHopList": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
DestinationCidrBlock	String	Yes	No	The destination Classless Inter-Domain Routing (CIDR) block of the route entry.	None
RouteTableId	String	Yes	No	The ID of the route table.	None
NextHopId	String	No	No	The ID of the next-hop instance of the route entry.	The route is a non-ECMP route.
RouteId	String	Yes	No	The ID of the route.	None
NextHopType	String	No	No	The type of the next hop.	Default value: Instance. Valid values: • Instance • Tunnel • HaVip • RouterInterface

Property	Type	Required	Editable	Description	Constraint
NextHopList	List	No	No	The list of next hops of the route entry.	You must specify the NextHopType and NextHopId parameters to specify the next hops. • If you specify the NextHopList parameter, the route is an ECMP route. The list contains two to four next hops of the ECMP route entry. **Note** The NextHopList parameter can be specified only when the route entry belongs to a VRouter. In addition, the next hops must be the router interfaces pointing to the connected VBRs. • If you do not specify the NextHopList parameter, the route is a non-ECMP route.

NextHopList syntax

```
"NextHopList": [
  {
    "NextHopId": String,
    "NextHopType": String
  }
]
```

NextHopList properties

Property	Type	Required	Editable	Description	Constraint
NextHopId	String	Yes	No	The ID of the next-hop instance of the route entry.	None
NextHopType	String	No	No	The type of the next hop.	Default value: RouterInterface. Valid values: • Instance • Tunnel • HaVip • RouterInterface

Response parameters

Fn::GetAtt

None

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ECSRoute": {
      "Type": "ALIYUN::ECS::Route",
      "Properties": {
 "RouteId": "vrt-25mz0****",
        "RouteTableId": "vtb-25oud****",
        "DestinationCidrBlock": "172.16.XX.XX/24",
        "NextHopId": "i-25xzy****"
      }
    }
  }
}
```

5.1.6.1.21. ALIYUN::ECS::SNatEntry

ALIYUN::ECS::SNatEntry is used to configure a NAT Gateway table in a source network address translation.

Statement

```
{
  "Type": "ALIYUN::ECS::SNatEntry",
  "Properties": {
    "SNatTableId": String,
    "SNatIp": String,
    "SourceVSwitchId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
SNatTableId	String	Retained	Yes	The ID source network address translation table.	None
SNatIp	String	Retained	Yes	The public IP address used to source network address translation.	The public IP address must be NAT Gateway in the bandwidth plan. It cannot exist in both the forwarding table and the SNAT table.

Parameter	Type	Required	Editable	Description	Constraint
SourceVSwitchId	String	Retained	Yes	The ID of the VSwitch that accesses the Internet through the SNAT function of NAT Gateway.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

SNatEntryId: the table entry ID in the source network address translation table.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SNatEntry": {
      "Type": "ALIYUN::ECS::SNatEntry",
      "Properties": {
        "SNatTableId": "stb-3er41****",
        "SourceVSwitchId": "vsw-25rc1****",
        "SNatIp": "101.201.XX.XX"
      }
    }
  },
  "Outputs": {
    "SNatEntryId": {
         "Value": {"Fn::GetAttr": ["SNatEntry","SNatEntryId"]}
    }
}
```

5.1.6.1.22. ALIYUN::ECS::SecurityGroup

ALIYUN::ECS::SecurityGroup is used to create a security group.

Statement

```
{
  "Type": "ALIYUN::ECS::SecurityGroup",
  "Properties": {
    "VpcId": String,
    "Description": String,
    "SecurityGroupName": String,
    "Tags": List,
    "SecurityGroupEgress": List,
    "SecurityGroupIngress": List,
    "ResourceGroupId": String,
    "SecurityGroupType": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None
VpcId	String	Yes	Released	The ID of the VPC.	None
Description	String	Yes	Released	The description of the new security group.	The description must be 2 to 256 characters in length.
Tags	List	Erased	Released	The tags of the security group.	A maximum of 20 tags can be specified.
SecurityGroupName	String	Yes	Released	The name of the new security group.	Default value: empty. • The name must be 2 to 128 characters in length • Must start with an uppercase or lowercase letter, and cannot start with `http://` and `https://` the beginning. • It must start with a letter and cannot start with http:// or https://.
SecurityGroupEgress	List	Erased	Released	The outbound access rules of the security group.	None
SecurityGroupIngress	List	Erased	Released	The inbound access rules of the security group.	None

Parameter	Type	Required	Editable	Description	Constraint
SecurityGroupType	String	Yes	Released	The type of the new security group.	Valid values: • normal (basic security group) • enterprise (Advanced Security Group)

Tags syntax

```
"Tags": [
  {
    "Value" : String,
    "Key" : String
  }
]
```

Tags properties

Parameter	Type	Required	Editable	Description	Constraint
Key	String	No	No	None	None
Value	String	Yes	Released	None	None

SecurityGroupEgress syntax

```
"SecurityGroupEgress": [
  {
    "Description": String,
    "PortRange": String,
    "SecurityGroupId": String,
    "NicType": String,
    "Priority": Integer,
    "DestGroupId": String,
    "DestCidrIp": String,
    "Policy": String,
    "IpProtocol": String,
    "DestGroupOwnerAccount": String,
    "DestGroupOwnerId": String,
    "Ipv6DestCidrIp": String
  }
]
```

SecurityGroupEgress properties

Parameter	Type	Required	Editable	Description	Constraint
Description	String	Yes	Released	The description of the security group rule.	The description must be 1 to 512 characters in length.

Parameter	Type	Required	Editable	Description	Constraint
DestGroupOwnerId	String	Yes	Released	The ID of the Alibaba Cloud account that owns the destination security group. This parameter is used to grant the current security group access to security groups in another Alibaba Cloud account.	If neither the DestGroupOwnerId parameter nor the DestGroupOwnerAccount parameter is specified, the current security group is granted access to other security groups in the same Alibaba Cloud account. If the DestCidrIp parameter is specified, the DestGroupOwnerId parameter is ignored.
IpProtocol	String	No	No	The Internet protocol over which the listener will forward requests.	Valid values: • TCP • udp • icmp • GRE • All A value of all specifies that all the four protocols are supported.

Parameter	Type	Required	Editable	Description	Constraint
PortRange	String	Yes	Released	The range of port numbers corresponding to the Internet protocol.	The range of destination ports corresponding to the transport layer protocol. Valid values: • When the IpProtocol parameter is set to tcp or udp, the port number range is 1 to 65535. Separate the starting port and the ending port with a forward slash (/). Correct example: 1/200. Incorrect example: 200/1. • When the IpProtocol parameter is set to icmp, the port number range is -1/-1, indicating that all ports are available. • When the IpProtocol parameter is set to gre, the port number range is -1/-1, indicating that all ports are available. • When the IpProtocol parameter is set to all, the port number range is -1/-1.

Parameter	Type	Required	Editable	Description	Constraint
SecurityGroupId	String	Yes	Released	The ID of the security group for which to create the outbound access rules.	None
NicType	String	Yes	Released	The network type of the instance. Valid values:	Valid values: • internet • intranet Default value: internet.
Priority	String	Optional	Released	The priority of the authorization policy.	Valid values: 1 to 100. Default value: 1
DestGroupId	String	Yes	Released	The ID of the destination security group within the same region.	You must specify either the DestGroupId parameter or the DestCidrIp parameter. If both parameters are specified, the system authorizes the destination CIDR block specified by the DestCidrIp parameter. If the DestGroupId parameter is specified, but the DestCidrIp parameter is not, you must set the NicType parameter to intranet.

Parameter	Type	Required	Editable	Description	Constraint
DestCidrIp	String	Yes	Released	The source IPv4 CIDR block.	The value must be in CIDR format. The default value is 0.0.0.0/0, indicating that access from any IP addresses is allowed. Examples of other supported formats include 10.159.XX.XX/12. Only IPv4 addresses are supported.
Policy	String	Yes	Released	The authorization policy.	Valid values: • accept: grants access • drop: denies access Default value: accept.
DestGroupOwnerAccount	String	Yes	Released	The Alibaba Cloud account of the destination security group when you grant security group permissions across accounts.	None
Ipv6DestCidrIp	String	Yes	Released	The destination IPv6 CIDR block.	IPv6 addresses in CIDR format are supported. You can only specify the IP addresses for ECS instances in VPCs.

SecurityGroupIngress syntax

```
"SecurityGroupIngress": [
  {
    "SourceGroupOwnerId": String,
    "SourceGroupOwnerAccount": String,
    "Description": String,
    "PortRange": String,
    "SecurityGroupId": String,
    "NicType": String,
    "Ipv6SourceCidrIp": String,
    "Priority": Integer,
    "SourceGroupId": String,
    "Policy": String,
    "IpProtocol": String,
    "SourcePortRange": String,
    "SourceCidrIp": String
  }
]
```

SecurityGroupIngress properties

Parameter	Type	Required	Editable	Description	Constraint
SourceGroupOwnerId	String	Yes	Released	The ID of the Alibaba Cloud account that owns the source security group.	None
Description	String	Yes	Released	The description of the security group rule.	The description must be 1 to 512 characters in length.
IpProtocol	String	No	No	The Internet protocol over which the listener will forward requests.	Valid values: tcp, udp, icmp, gre, and all. A value of all specifies that all the four protocols are supported.

Parameter	Type	Required	Editable	Description	Constraint
PortRange	String	Yes	Released	The range of port numbers corresponding to the Internet protocol.	The range of destination ports corresponding to the transport layer protocol. Valid values: • When the IpProtocol parameter is set to tcp or udp, the port number range is 1 to 65535. Separate the starting port and the ending port with a forward slash (/). Correct example: 1/200. Incorrect example: 200/1. • When the IpProtocol parameter is set to icmp, the port number range is -1/-1, indicating that all ports are available. • When the IpProtocol parameter is set to gre, the port number range is -1/-1, indicating that all ports are available. • When the IpProtocol parameter is set to all, the port number range is -1/-1, indicating that all ports are available.

Parameter	Type	Required	Editable	Description	Constraint
SourceGroupId	String	Yes	Released	The ID of the source security group within the same region.	You must specify either the SourceGroupId parameter or the SourceCidrIp parameter. If both parameters are specified, the system authorizes the source CIDR block specified by the SourceCidrIp parameter. If the SourceGroupId parameter is specified, but the SourceCidrIp parameter is not, you must set the NicType parameter to intranet.
SecurityGroupId	String	Yes	Released	The ID of the security group for which you want to create the inbound access rule.	None
NicType	String	Yes	Released	The network type of the instance. Valid values:	Valid values: • internet • intranet Default value: internet.
SourceGroupOwnerAccount	String	Yes	Released	The Alibaba Cloud account of the destination security group when you grant security group permissions across accounts.	None
Priority	String	Optional	Released	The priority of the authorization policy.	Valid values: 1 to 100. Default value: 1

Parameter	Type	Required	Editable	Description	Constraint
SourceCidrIp	String	Yes	Released	The source IPv4 CIDR block.	The value must be in CIDR format. The default value is 0.0.0.0/0, indicating that access from any IP addresses is allowed. Examples of other supported formats include 10.159.XX.XX/12. Only IPv4 CIDR blocks are supported.
Policy	String	Yes	Released	The authorization policy.	Valid values: • accept: accepts the request. • drop: access is denied. Default value: accept.

Parameter	Type	Required	Editable	Description	Constraint
SourcePortRange	String	Yes	Released	The range of source ports relevant to transport layer protocols.	Valid values: • When the IpProtocol parameter is set to tcp or udp, the port number range is 1 to 65535. Separate the starting port and the ending port with a forward slash (/). Correct example: 1/200. Incorrect example: 200/1. • When the IpProtocol parameter is set to icmp, the port number range is -1/-1, indicating that all values are valid. • When the IpProtocol parameter is set to gre, the port number range is -1/-1, indicating that all ports are available. • The IpProtocol value is all:-1/-1.
Ipv6SourceCidrIp	String	Yes	Released	The source IPv6 CIDR block. IPv6 addresses in CIDR format are supported.	You can only specify the IP addresses of ECS instances in VPCs.

Response parameters

Fn::GetAtt

SecurityGroupId: the ID of the new security group.

Sample request

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "SG": {
      "Type": "ALIYUN::ECS::SecurityGroup",
      "Properties": {
        "SecurityGroupName": {
          "Ref": "SecurityGroupName"
        },
        "SecurityGroupIngress": [
          {
            "SourceCidrIp": "0.0.0.0/0",
            "IpProtocol": "all",
            "NicType": "internet",
            "PortRange": "-1/-1",
            "Priority": 1
          },
          {
            "SourceCidrIp": "0.0.0.0/0",
            "IpProtocol": "all",
            "NicType": "intranet",
            "PortRange": "-1/-1",
            "Priority": 1
          }
        ],
        "SecurityGroupEgress": [
          {
            "IpProtocol": "all",
            "DestCidrIp": "0.0.0.0/0",
            "NicType": "internet",
            "PortRange": "-1/-1",
            "Priority": 1
          },
          {
            "IpProtocol": "all",
            "DestCidrIp": "0.0.0.0/0",
            "NicType": "intranet",
            "PortRange": "-1/-1",
            "Priority": 1
          }
        ],
        "VpcId": {
          "Ref": "Vpc"
        }
      }
    }
  },
  "Outputs": {
    "SecurityGroupId": {
      "Value" : {"Fn::GetAtt": ["SG","SecurityGroupId"]}
    }
  }
}
```

5.1.6.1.23. ALIYUN::ECS::SecurityGroupClone

ALIYUN::ECS::SecurityGroupClone is used to clone a security group.

Syntax

```
{
  "Type": "ALIYUN::ECS::SecurityGroupClone",
  "Properties": {
    "DestinationRegionId": String,
    "VpcId": String,
    "Description": String,
    "SecurityGroupName": String,
    "SourceSecurityGroupId": String,
    "ResourceGroupId": String,
    "NetworkType": String,
    "SecurityGroupType": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	No	No	The ID of the resource group to which the instance belongs.	None
SourceSecurityGroupId	String	Yes	No	The ID of the source security group.	Only applicable security group rules are copied to the new security group. The security group rules are selected based on the network type of the new security group.
NetworkType	String	No	No	The network type of the new security group.	Set the value to Classic.
VpcId	String	No	No	The ID of the VPC to which the new security group belongs.	The NetworkType parameter is ignored if both the VpcId and NetworkType parameters are specified.
Description	String	No	No	The description of the new security group.	The description must be 2 to 256 characters in length. It cannot start with http:// or https://.

Property	Type	Required	Editable	Description	Constraint
SecurityGroupName	String	No	No	The name of the new security group.	This parameter is empty by default. The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), and hyphens (-). It must start with a letter and cannot start with http:// or https://.
DestinationRegionId	String	No	No	The ID of the destination region where the new security group resides.	Default value: CURRENT.
SecurityGroupType	String	No	No	The type of the new security group.	Valid values: normal and enterprise. A value of normal specifies a basic security group. A value of enterprise specifies an advanced security group.

Response parameters

Fn::GetAtt

SecurityGroupId: the ID of the new security group.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SecurityGroupClone": {
      "Type": "ALIYUN::ECS::SecurityGroupClone",
      "Properties": {
        "SourceSecurityGroupId": {
          "Ref": "SourceSecurityGroupId"
        },
        "VpcId": {
          "Ref": "VpcId"
        },
        "Description": {
          "Ref": "Description"
        },
        "SecurityGroupName": {
```

```
          "Ref": "SecurityGroupName"
        },
        "DestinationRegionId": {
          "Ref": "DestinationRegionId"
        },
        "NetworkType": {
          "Ref": "NetworkType"
        }
      }
    }
  },
  "Parameters": {
    "SourceSecurityGroupId": {
      "Type": "String",
      "Description": "Source security group ID is used to copy properties to clone new security group. If the NetworkType and VpcId is not specified, the same security group will be cloned. If NetworkType or VpcId is specified, only proper security group rules will be cloned."
    },
    "VpcId": {
      "Type": "String",
      "Description": "Physical ID of the VPC."
    },
    "Description": {
      "Type": "String",
      "Description": "Description of the security group, [2, 256] characters. Do not fill or empty, the default is empty."
    },
    "SecurityGroupName": {
      "Type": "String",
      "Description": "Display name of the security group, [2, 128] English or Chinese characters, must start with a letter or Chinese in size, can contain numbers, '_' or '.', '-'"
    },
    "DestinationRegionId": {
      "Default": "CURRENT",
      "Type": "String",
      "Description": "Clone security group to the specified region. Default to current region."
    },
    "NetworkType": {
      "Type": "String",
      "Description": "Clone new security group as classic network type. If the VpcId is specified, the value will be ignored.",
      "AllowedValues": [
        "Classic"
      ]
    }
  },
  "Outputs": {
    "SecurityGroupId": {
      "Description": "Generated security group id of new security group.",
      "Value": {
        "Fn::GetAtt": [
          "SecurityGroupClone",
          "SecurityGroupId"
        ]
      }
    }
  }
}
```

5.1.6.1.24. ALIYUN::ECS::SecurityGroupEgress

ALIYUN::ECS::SecurityGroupEgress is used to create an outbound access rule for a security group.

Syntax

```
{
  "Type": "ALIYUN::ECS::SecurityGroupEgress",
  "Properties": {
    "SecurityGroupId": String,
    "IpProtocol": String,
    "PortRange": String,
    "DestGroupId": String,
    "DestGroupOwnerAccount": String,
    "DestCidrIp": String,
    "Policy": String,
    "Priority": String,
    "NicType": String,
    "Ipv6DestCidrIp": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
IpProtocol	String	Yes	No	The transport layer protocol.	Valid values: • tcp • udp • icmp • gre • all **Note** The value all indicates that all the four protocols are supported.

Property	Type	Required	Editable	Description	Constraint
PortRange	String	Yes	No	The range of destination port numbers corresponding to the transport layer protocol.	• Valid values when IpProtocol is set to tcp or udp: 1 to 65535. Separate the starting and ending port numbers with a forward slash (/). Correct example: 1/200. Incorrect example: 200/1. • Set the value to -1/-1 when IpProtocol is set to icmp. • Set the value to -1/-1 when IpProtocol is set to gre. • Set the value to -1/-1 when IpProtocol is set to all.
SecurityGroupId	String	No	No	The ID of the source security group.	None
NicType	String	No	No	The type of the network interface controller (NIC).	Default value: internet. Valid values: • internet • intranet If the DestGroupId parameter is specified, but the DestCidrIp parameter is not, set the value to intranet.
Priority	Integer	No	No	The priority of the security group rule.	Valid values: 1 to 100 Default value: 1.

Property	Type	Required	Editable	Description	Constraint
DestGroupId	String	No	No	The ID of the destination security group for which you want to set access permissions.	You must specify at least one of the DestGroupId and DestCidrIp parameters. If the DestGroupId parameter is specified, but the DestCidrIp parameter is not, set the NicType value to intranet. If both the DestGroupId and DestCidrIp parameters are specified, the DestCidrIp parameter takes precedence.
DestCidrIp	String	No	No	The destination CIDR block.	Only IPv4 CIDR blocks are supported.
Policy	String	No	No	The authorization policy.	Default value: accept. Valid values: • accept: allows access. • drop: denies access.
DestGroupOwnerAccount	String	No	No	The Alibaba Cloud account that manages the destination security group when you set a security group rule across accounts.	If you specify neither of the DestGroupOwnerAccount parameter nor the DestGroupOwnerId parameter, the access permission is configured on another security group managed by your account. If you specify the DestCidrIp parameter, the DestGroupOwnerAccount parameter is ignored.

Property	Type	Required	Editable	Description	Constraint
Description	String	No	Yes	The description of the security group rule.	The description must be 1 to 512 characters in length.
DestGroupOwnerId	String	No	No	The ID of the Alibaba Cloud account that manages the destination security group when you set a security group rule across accounts.	If you specify neither of the DestGroupOwnerAccount parameter nor the DestGroupOwnerId parameter, the access permission is configured on another security group managed by your account. If you specify the DestCidrIp parameter, the DestGroupOwnerId parameter is ignored.
Ipv6DestCidrIp	String	No	No	The destination IPv6 CIDR block.	IPv6 addresses in the CIDR format are supported. You can specify only the IP addresses of ECS instances in VPCs.

Response parameters

Fn::GetAtt

None

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SG": {
      "Type": "ALIYUN::ECS::SecurityGroupEgress",
      "Properties": {
        "SecurityGroupId": "sg-25bow****",
        "IpProtocol": "tcp",
        "PortRange": "65535/65535",
        "DestCidrIp": "0.0.0.0/0"
      }
    }
  }
}
```

5.1.6.1.25. ALIYUN::ECS::SecurityGroupIngress

ALIYUN::ECS::SecurityGroupIngress is used to create an inbound access rule for a security group.

Syntax

```
{
  "Type": "ALIYUN::ECS::SecurityGroupIngress",
  "Properties": {
    "SourceGroupOwnerId": String,
    "Description": String,
    "PortRange": String,
    "SecurityGroupId": String,
    "NicType": String,
    "Ipv6SourceCidrIp": String,
    "Priority": Integer,
    "SourceGroupId": String,
    "Policy": String,
    "IpProtocol": String,
    "SourcePortRange": String,
    "SourceCidrIp": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
IpProtocol	String	Yes	No	The Internet protocol.	Valid values: tcp, udp, icmp, gre, and all. A value of all specifies that all the four protocols are supported.

Property	Type	Required	Editable	Description	Constraint
PortRange	String	Yes	No	The range of destination ports relevant to transport layer protocols.	• When the IpProtocol parameter is set to tcp or udp, the port number range is 1 to 65535. Separate the starting port and the ending port with a forward slash (/). Correct example: 1/200. Incorrect example: 200/1. • When the IpProtocol parameter is set to icmp, the port number range is -1/-1. • When the IpProtocol parameter is set to gre, the port number range is -1/-1. • When the IpProtocol parameter is set to all, the port number range is -1/-1.

Property	Type	Required	Editable	Description	Constraint
SourceGroupId	String	No	No	The ID of the source security group for which you want to set access permissions.	You must specify at least one of the SourceGroupId and SourceCidrIp parameters. If the SourceGroupId parameter is specified, but the SourceCidrIp parameter is not, the NicType parameter must be set to intranet. If both the SourceGroupId and SourceCidrIp parameters are specified, the SourceCidrIp value is used by default.
SecurityGroupId	String	No	No	The ID of the security group for which you want to create the inbound access rule.	None
NicType	String	No	No	The network type of the instance.	Valid values: • internet • intranet Default value: internet.
SourceGroupOwnerAccount	String	No	No	The Alibaba Cloud account that manages the source security group when you set a security group rule across accounts.	If neither the SourceGroupOwnerAccount parameter nor the SourceGroupOwnerId parameter is specified, the access permission is configured for another security group managed by your account. If the SourceCidrIp parameter is specified, this parameter is ignored.

Property	Type	Required	Editable	Description	Constraint
Priority	Integer	No	No	The priority of the security group rule.	Valid values: 1 to 100. Default value: 1.
SourceCidrIp	String	No	No	The source IPv4 CIDR block.	Only IPv4 CIDR blocks are supported.
Policy	String	No	No	The access control policy.	Valid values: • accept: grants access. • drop: denies access. Default value: accept.
SourceGroupOwnerId	String	No	No	The ID of the Alibaba Cloud account that manages the source security group when you set a security group rule across accounts.	If neither the SourceGroupOwnerId parameter nor the SourceGroupOwnerAccount parameter is specified, the access permission is configured for another security group managed by your account. If the SourceCidrIp parameter is specified, this parameter is ignored.
Description	String	No	Yes	The description of the security group rule.	The description must be 1 to 512 characters in length.

Property	Type	Required	Editable	Description	Constraint
SourcePortRange	String	No	No	The range of source ports relevant to transport layer protocols.	• When the IpProtocol parameter is set to tcp or udp, the port number range is 1 to 65535. Separate the starting port and the ending port with a forward slash (/). Correct example: 1/200. Incorrect example: 200/1. • When the IpProtocol parameter is set to icmp, the port number range is -1/-1. • When the IpProtocol parameter is set to gre, the port number range is -1/-1. • When the IpProtocol parameter is set to all, the port number range is -1/-1.
Ipv6SourceCidrIp	String	No	No	The range of source IPv6 addresses.	CIDR blocks and IPv6 addresses are supported. You can only specify the IP addresses of ECS instances in VPCs.

Response parameters

Fn::GetAtt

None

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SG": {
      "Type": "ALIYUN::ECS::SecurityGroupIngress",
      "Properties": {
        "SecurityGroupId": "sg-25bow****",
        "IpProtocol": "tcp",
        "PortRange": "65535/65535",
        "SourceCidrIp": "0.0.0.0/0"
      }
    }
  }
}
```

5.1.6.1.26. ALIYUN::ECS::Snapshot

ALIYUN::ECS::Snapshot is used to create a disk Snapshot.

Statement

```
{
  "Type": "ALIYUN::ECS::Snapshot",
  "Properties": {
    "SnapshotName": String,
    "Timeout": Integer,
    "Description": String,
    "DiskId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
DiskId	String	No	No	The ID of the disk for which you want to create the snapshot.	None

Parameter	Type	Required	Editable	Description	Constraint
SnapshotName	String	Yes	Released	The name of the snapshot.	It must be 2 to 128 characters in length. And can contain letters, digits, underscores (_), and hyphens (-). It cannot start with auto. Snapshot names starting with auto are reserved for automatic snapshots. It cannot start with `http://` or `https://`.
Timeout	String	Optional	Released	The timeout period that is specified for the snapshot creation request.	If this parameter is set, the timeout period to create a resource stack is prolonged. If the snapshot is not created within the specified time period, the entire resource stack fails to be created. You must set the timeout period according to the disk size and data amount. Valid values: 200 to 1440. Unit: minute. The default value is 200 minutes.
Description	String	Yes	Released	The description of the snapshot.	The length must be 2 to 256 characters in length. This parameter is empty by default. It cannot start with `http://` or `https://`.

Response parameters

Fn::GetAtt

SnapshotId: the ID of the snapshot.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Snapshot": {
      "Type": "ALIYUN::ECS::Snapshot",
      "Properties": {
        "DiskId": "d-2zedgvuvu8cylvrd****"
      }
    }
  },
  "Outputs": {
    "SnapshotId": {
      "Value": {
        "Fn::GetAtt": [
          "Snapshot",
          "SnapshotId"
        ]
      }
    }
  }
}
```

5.1.6.1.27. ALIYUN::ECS::SSHKeyPair

ALIYUN::ECS::SSHKeyPair is used to create or import an SSH key pair to an ECS instance.

Statement

```
{
  "Type": "ALIYUN::ECS::SSHKeyPair",
  "Properties": {
    "ResourceGroupId": String,
    "KeyPairName": String,
    "PublicKeyBody": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None

Parameter	Type	Required	Editable	Description	Constraint
KeyPairName	String	No	No	The globally unique name of the SSH key pair.	The name must be 2 to 128 characters in length and can contain letters, digits, periods (.), underscores (_), and hyphens (-). It cannot start with `http://` or `https://`.
PublicKeyBody	String	Yes	Released	Specifies the SSH public key to import.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- KeyPairFingerPrint: the fingerprint of the key pair. The message-digest algorithm 5 (MD5) is used based on the public key fingerprint format defined in RFC 4716.
- PrivateKeyBody: the private key of the key pair. An unencrypted RSA private key must be encoded using PEM and must be in the PKCS#8 format. The private key of a key pair can only be obtained at the time of its creation. If you import an existing public key, no private key information will be available.
- KeyPairName: the globally unique name of the SSH key pair.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SSHKeyPair": {
      "Type": "ALIYUN::ECS::SSHKeyPair",
      "Properties": {
        "KeyPairName": "ssh_key_pair_v1"
      }
    }
  },
  "Outputs": {
    "KeyPairName": {
      "Value": {
        "Fn::GetAtt": [
          "SSHKeyPair",
          "KeyPairName"
        ]
      }
    },
    "PrivateKeyBody": {
      "Value": {
        "Fn::GetAtt": [
          "SSHKeyPair",
          "PrivateKeyBody"
        ]
      }
    },
    "KeyPairFingerPrint": {
      "Value": {
        "Fn::GetAtt": [
          "SSHKeyPair",
          "KeyPairFingerPrint"
        ]
      }
    }
  }
}
```

5.1.6.1.28. ALIYUN::ECS::SSHKeyPairAttachment

ALIYUN::ECS::SSHKeyPairAttachment is used to bind an SSH key pair to an ECS instance.

Statement

```
{
  "Type": "ALIYUN::ECS::SSHKeyPairAttachment",
  "Properties": {
    "InstanceIds": List,
    "KeyPairName": String
  }
}
```

Properties

Parameter	Type	Required or Not	Editable	Description	Constraint
InstanceIds	List	Retained	Yes	The IDs of the ECS instances with which you want to associate the EIP.	Separate the IDs with a comma (,). Only Linux instances are supported.
KeyPairName	String	No	No	The name of the SSH key pair.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

None

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "SSHKeyPairAttachment": {
      "Type": "ALIYUN::ECS::SSHKeyPairAttachment",
      "Properties": {
        "KeyPairName": "ssh_key_pair_v1",
        "InstanceIds": [
          'I-2zeiofnh20hj**** has been added * ',
          'I-2zebt3kfvxm2**** has two records *'
        ]
      }
    }
  }
}
```

5.1.6.1.29. ALIYUN::ECS::VPC

ALIYUN::ECS::VPC is used to create a VPC.

Statement

```
{
  "Type": "ALIYUN::ECS::VPC",
  "Properties": {
    "Description": String,
    "Ipv6CidrBlock": String,
    "EnableIpv6": Boolean,
    "ResourceGroupId": String,
    "VpcName": String,
    "CidrBlock": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the instance belongs.	None
VpcName	String	Yes	True	The name of the VPC.	• The name must be 2 to 128 characters in length • Must start with english letters or starts with a Chinese character. • It cannot start with `http://` or `https://`. • It must start with a letter and cannot start with http:// or https://.
CidrBlock	String	Yes	True	The CIDR block of the VPC.	Valid values: 10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16.
Description	String	Yes	True	The description of the VPC.	The description must be 2 to 256 characters in length. It cannot start with `http://` or `https://`.

Parameter	Type	Required	Editable	Description	Constraint
Ipv6CidrBlock	String	Yes	Released	The IPv6 CIDR block of the VPC.	None
EnableIpv6	Boolean	No.	True	Specifies whether to enable an IPv6 CIDR block.	Valid values: • true • false Default value: false.

Response parameters

Fn::GetAtt

- VpcId: The VPC ID allocated by the system.
- VRouterId: the ID of the vRouter.
- RouteTableId: the ID of the routing table.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "EcsVpc": {
      "Type": "ALIYUN::ECS::VPC",
      "Properties": {
        "CidrBlock": "172.16.0.0/12",
        "VpcName": "vpc-test-del"
      }
    }
  },
  "Outputs": {
    "VpcId": {
      "Value": {
        "Fn::GetAtt": [
          "EcsVpc",
          "VpcId"
        ]
      }
    },
    "VRouterId": {
      "Value": {
        "Fn::GetAtt": [
          "EcsVpc",
          "VRouterId"
        ]
      }
    },
    "RouteTableId": {
      "Value": {
        "Fn::GetAtt": [
          "EcsVpc",
          "RouteTableId"
        ]
      }
    }
  }
}
```

5.1.6.1.30. ALIYUN::ECS::VSwitch

ALIYUN::ECS::VSwitch is used to create a VSwitch.

Statement

```
{
  "Type": "ALIYUN::ECS::VSwitch",
  "Properties": {
    "VSwitchName": String,
    "VpcId": String,
    "Description": String,
    "Ipv6CidrBlock": Integer,
    "ZoneId": String,
    "CidrBlock": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
VpcId	String	No	No	The ID of the VPC where a vSwitch is to be created	None
ZoneId	String	No	No	The ID of the zone where the instance resides.	None
VSwitchName	String	Yes	True	The name of the VSwitch.	• The name must be 2 to 128 characters in length • It must start with a letter. • Cannot `http://` or `https://` the beginning. • It must start with a letter and cannot start with http:// or https://.
CidrBlock	String	No	No	The CIDR block of the VSwitch.	The VSwitch CIDR block must be a subset of the CIDR block assigned to the VPC where the VSwitch resides and not be used by other VSwitches.
Description	String	Yes	True	The description of the vSwitch.	The description must be 2 to 256 characters in length. It cannot start with `http://` or `https://`.
Ipv6CidrBlock	String	Optional	Released	The IPv6 CIDR block of the VSwitch. You can customize the last eight bits of the IPv6 CIDR block.	Valid values: 0 to 255. The value is a decimal integer. By default, the prefix of the IPv6 CIDR block of the VSwitch is set to /64.

Response parameters

Fn::GetAtt

- VSwitchId: indicates the vSwitch ID allocated by the system.
- CidrBlock: the IPv4 CIDR block of the vSwitch.
- Ipv6CidrBlock: the IPv6 CIDR block of the vSwitch.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "VpcName": {
      "Type": "String"
    },
    "VSwitch1CidrBlock": {
      "Type": "String",
      "Default": "172.16.100.0/24"
    },
    "VSwitch2CidrBlock": {
        "Type": "String",
        "Default": "172.16.80.0/24"
    }
  },
  "Resources": {
    "EcsVpc": {
      "Type": "ALIYUN::ECS::VPC",
      "Properties": {
        "CidrBlock": "172.16.0.0/12",
        "VpcName": {"Ref": "VpcName"},
      },
    },
    "VSwitch1": {
      "Type": "ALIYUN::ECS::VSwitch",
      "Properties": {
        "ZoneId": "cn-beijing-a",
        "CidrBlock": {"Ref": "VSwitch1CidrBlock"},
        "VpcId": { "Fn::GetAtt": [ "EcsVpc", "VpcId" ] },
        "VSwitchName": "create_vpc_vswitch_sg1"
      }
    },
    "VSwitch2": {
      "Type": "ALIYUN::ECS::VSwitch",
      "Properties": {
        "ZoneId": "cn-beijing-a",
        "CidrBlock": {"Ref": "VSwitch2CidrBlock"},
        "VpcId": { "Fn::GetAtt": [ "EcsVpc", "VpcId" ] },
        "VSwitchName": "create_vpc_vswitch_sg2"
      }
    },
    "SG_VSwitch1": {
      "Type": "ALIYUN::ECS::SecurityGroup",
      "Properties": {
        "SecurityGroupName": "app_mall",
        "Description": "this is created by heat",
        "VpcId": { "Fn::GetAtt": [ "EcsVpc", "VpcId" ] }
      },
      "Outputs": {
        "SecurityGroupId": {
```

```
              "Value": {"get_attr": ["SG_VSwitch1","SecurityGroupId"]}
          }
        }
      },
      "SG_VSwitch1_InRule": {
        "Type": "ALIYUN::ECS::SecurityGroupIngress",
        "Properties": {
          "SecurityGroupId": { "Fn::GetAtt": [ "SG_VSwitch1", "SecurityGroupId" ] },
          "IpProtocol": "tcp",
          "PortRange": "1/65535",
          "SourceCidrIp": {"Ref": "VSwitch2CidrBlock"}
        }
      },
      "SG_VSwitch1_OutRule": {
        "Type": "ALIYUN::ECS::SecurityGroupEgress",
        "Properties": {
          "SecurityGroupId": { "Fn::GetAtt": [ "SG_VSwitch1", "SecurityGroupId" ] },
          "IpProtocol": "tcp",
          "PortRange": "1/65535",
          "DestCidrIp": {"Ref": "VSwitch2CidrBlock"}
        }
      },
      "SG_VSwitch2": {
        "Type": "ALIYUN::ECS::SecurityGroup",
        "Properties": {
          "SecurityGroupName": "app_mall",
          "Description": "this is created by heat",
          "VpcId": { "Fn::GetAtt": [ "EcsVpc", "VpcId" ] }
        },
      },
      "SG_VSwitch2_InRule": {
        "Type": "ALIYUN::ECS::SecurityGroupIngress",
        "Properties": {
          "SecurityGroupId": { "Fn::GetAtt": [ "SG_VSwitch2", "SecurityGroupId" ] },
          "IpProtocol": "tcp",
          "PortRange": "1/65535",
          "SourceCidrIp": {"Ref": "VSwitch1CidrBlock"}
        }
      },
      "SG_VSwitch2_OutRule": {
        "Type": "ALIYUN::ECS::SecurityGroupEgress",
        "Properties": {
          "SecurityGroupId": { "Fn::GetAtt": [ "SG_VSwitch2", "SecurityGroupId" ] },
          "IpProtocol": "tcp",
          "PortRange": "1/65535",
          "DestCidrIp": {"Ref": "VSwitch1CidrBlock"}
        }
      }
    }
}
```

5.1.6.2. ESS

5.1.6.2.1. ALIYUN::ESS::AlarmTask

ALIYUN::ESS::AlarmTask is used to create a metric-based alarm task.

Syntax

```
{
  "Type": "ALIYUN::ESS::AlarmTask",
  "Properties": {
    "Statistics": String,
    "Name": String,
    "EvaluationCount": Integer,
    "Period": Integer,
    "MetricType": String,
    "ComparisonOperator": String,
    "Dimensions": List,
    "ScalingGroupId": String,
    "AlarmAction": List,
    "Threshold": Number,
    "MetricName": String,
    "GroupId": Integer,
    "Description": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
Statistics	String	No	No	The method used to calculate monitoring data. The statistics must be appropriate for the metric chosen.	Valid values: Average, Minimum, and Maximum. Default value: Average.
Name	String	No	Yes	The name of the alarm rule.	None
EvaluationCount	Integer	No	No	The number of consecutive times that the threshold must be exceeded before an alarm is triggered.	Default value: 3. Minimum value: 1.
Period	Integer	No	No	The metric query period, which must be appropriate for the metric chosen. Unit: seconds.	Valid values: 60, 120, 300, and 900. Default value: 300.
MetricType	String	No	No	The metric type.	Valid values: system and custom.
ComparisonOperator	String	No	No	The alarm comparison operator used to define a condition in the alarm rule.	Valid values: <=, <, >, and > =.
Dimensions	List	No	No	The list of instances associated with the alarm rule.	You must include at least one instance in the list.

Property	Type	Required	Editable	Description	Constraint
ScalingGroupId	String	Yes	No	The ID of the scaling group.	None
AlarmAction	List	Yes	Yes	The list of alarm actions.	You must include one to five alarm actions in the list.
Threshold	Number	Yes	No	The alarm threshold, which must be a numeric value.	None
MetricName	String	Yes	No	The metric name of a service. For more information, see the metrics defined for each service.	None
GroupId	Integer	No	No	The group ID.	None
Description	String	No	Yes	The description of the alarm task.	None

Dimensions syntax

```
"Dimensions": [
  {
    "DimensionKey": String,
    "DimensionValue": String
  }
]
```

Dimensions properties

Property	Type	Required	Editable	Description	Constraint
DimensionValue	String	Yes	No	None	None
DimensionKey	String	Yes	No	None	None

Response parameters

Fn::GetAtt

AlarmTaskId: the ID of the alarm task.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "ComparisonOperator": {
      "Type": "String",
      "Description": "Comparison Operator",
      "AllowedValues": [
        ">=",
        "<=",
        ">",
```

```
        "<"
      ]
    },
    "Description": {
      "Type": "String",
      "Description": "Description"
    },
    "ScalingGroupId": {
      "Type": "String",
      "Description": "The ID of the scaling group."
    },
    "MetricType": {
      "Type": "String",
      "Description": "Metric Type",
      "AllowedValues": [
        "system",
        "custom"
      ]
    },
    "EvaluationCount": {
      "Type": "Number",
      "Description": "Evaluation Count",
      "MinValue": 1
    },
    "Period": {
      "Type": "Number",
      "Description": "Period",
      "AllowedValues": [
        60,
        120,
        300,
        900
      ]
    },
    "Dimensions": {
      "Type": "CommaDelimitedList",
      "Description": "Dimensions",
      "MinLength": 1
    },
    "Statistics": {
      "Type": "String",
      "Description": "Statistics",
      "AllowedValues": [
        "Average",
        "Minimum",
        "Maximum"
      ]
    },
    "Name": {
      "Type": "String",
      "Description": "Name"
    },
    "GroupId": {
      "Type": "Number",
      "Description": "Group Id"
    },
    "MetricName": {
      "Type": "String",
      "Description": "Metric Name"
    },
```

```
    "AlarmAction": {
      "Type": "CommaDelimitedList",
      "Description": "Alarm Actions",
      "MinLength": 1,
      "MaxLength": 5
    },
    "Threshold": {
      "Type": "Number",
      "Description": "Threshold"
    }
  },
  "Resources": {
    "AlarmTask": {
      "Type": "ALIYUN::ESS::AlarmTask",
      "Properties": {
        "ComparisonOperator": {
          "Ref": "ComparisonOperator"
        },
        "Description": {
          "Ref": "Description"
        },
        "ScalingGroupId": {
          "Ref": "ScalingGroupId"
        },
        "MetricType": {
          "Ref": "MetricType"
        },
        "EvaluationCount": {
          "Ref": "EvaluationCount"
        },
        "Period": {
          "Ref": "Period"
        },
        "Dimensions": {
          "Fn::Split": [
            ",",
            {
              "Ref": "Dimensions"
            },
            {
              "Ref": "Dimensions"
            }
          ]
        },
        "Statistics": {
          "Ref": "Statistics"
        },
        "Name": {
          "Ref": "Name"
        },
        "GroupId": {
          "Ref": "GroupId"
        },
        "MetricName": {
          "Ref": "MetricName"
        },
        "AlarmAction": {
          "Fn::Split": [
            ",",
            {
```

```
              "Ref": "AlarmAction"
            },
            {
              "Ref": "AlarmAction"
            }
          ]
        },
        "Threshold": {
          "Ref": "Threshold"
        }
      }
    }
  },
  "Outputs": {
    "AlarmTaskId": {
      "Description": "The alarm task ID",
      "Value": {
        "Fn::GetAtt": [
          "AlarmTask",
          "AlarmTaskId"
        ]
      }
    }
  }
}
```

5.1.6.2.2. ALIYUN::ESS::AlarmTaskEnable

ALIYUN::ESS::AlarmTaskEnable is used to start an alarm task. You can call this operation to enable alarm tasks when the task is stopped.

Statement

```
{
  "Type": "ALIYUN::ESS::AlarmTaskEnable",
  "Properties": {
    "AlarmTaskId": String,
    "Enable": Boolean
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
AlarmTaskId	String	No	No	The ID of the monitoring task.	None

Parameter	Type	Required	Editable	Description	Constraint
Enable	String	Retained	Yes	Specifies whether to enable the alarm task.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

None

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Enable": {
      "Type": "Boolean",
      "Description": "Enable alarm task or not",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    },
    "AlarmTaskId": {
      "Type": "String",
      "Description": "The id of alarm task."
    }
  },
  "Resources": {
    "AlarmTaskEnable": {
      "Type": "ALIYUN::ESS::AlarmTaskEnable",
      "Properties": {
        "Enable": {
          "Ref": "Enable"
        },
        "AlarmTaskId": {
          "Ref": "AlarmTaskId"
        }
      }
    }
  },
  "Outputs": {}
}
```

5.1.6.2.3. ALIYUN::ESS::LifecycleHook

ALIYUN::ESS::LifecycleHook is used to create a lifecycle hook for a scaling group.

Syntax

```
{
  "Type": "ALIYUN::ESS::LifecycleHook",
  "Properties": {
    "LifecycleHookName": String,
    "NotificationArn": String,
    "HeartbeatTimeout": Integer,
    "NotificationMetadata": String,
    "ScalingGroupId": String,
    "DefaultResult": String,
    "LifecycleTransition": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
LifecycleHookName	String	No	Yes	The name of the lifecycle hook. Each lifecycle hook name must be unique within a scaling group.	The name must be 2 to 40 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or digit. The default name is the ID of the lifecycle hook.
NotificationArn	String	No	Yes	The Alibaba Cloud Resource Name (ARN) of the notification target that Auto Scaling uses to notify you when an instance is in the transition state for the lifecycle hook.	This target can be either an MNS queue or an MNS topic. The format of the parameter value is acs:ess:{region}:{account-id}:{resource-relative-id}. • `region`: the region where the scaling group resides. • `account-id`: the ID of the Apsara Stack tenant account. Examples: • MNS queue: acs:ess:{region}:{account-id}:queue/{queuename} • MNS topic: acs:ess:{region}:{account-id}:topic/{topicname}

Property	Type	Required	Editable	Description	Constraint
HeartbeatTimeout	Integer	No	Yes	The waiting period before the lifecycle hook times out. When the lifecycle hook times out, the scaling group performs the action specified by the DefaultResult parameter. Unit: seconds.	Valid values: 30 to 21600. Default value: 600.
NotificationMetadata	String	No	Yes	The fixed string to include when Auto Scaling sends a notification about the wait state of a scaling activity. Auto Scaling sends the specified `NotificationMetadata` parameter value along with the notification message so that you can easily categorize notifications. The `NotificationMetadata` parameter is valid only after you set the `NotificationArn` parameter.	The parameter value cannot exceed 128 characters in length.
ScalingGroupId	String	Yes	No	The ID of the scaling group.	None
DefaultResult	String	No	Yes	The action that the scaling group takes when the lifecycle hook times out. If the scaling group has multiple lifecycle hooks and one of them is terminated when the `DefaultResult` parameter is set to ABANDON during a `scale-in` event, the remaining lifecycle hooks in the same scaling group will also be terminated. Otherwise, the scaling activity will proceed normally after the waiting period expires and continue with the action specified by the DefaultResult parameter.	Valid values: • CONTINUE: The scaling group continues the scale-in or scale-out event. • ABANDON: The scaling group releases the created ECS instances if the scaling activity type is scale-out or removes the ECS instances to be scaled in if the scaling activity type is scale-in. Default value: CONTINUE.

Property	Type	Required	Editable	Description	Constraint
LifecycleTransition	String	Yes	Yes	The type of scaling activity to which the lifecycle hook applies.	Valid values: • SCALE_OUT: scale-out events of the scaling group. • SCALE_IN: scale-in events of the scaling group.

Response parameters

Fn::GetAtt

LifecycleHookId: the ID of the lifecycle hook.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "LifecycleHookName": {
      "Type": "String",
      "Description": "The name of the lifecycle hook. Each name must be unique within a scaling group. The name must be 2 to 40 characters in length and can contain letters, numbers, Chinese characters, and special characters including underscores (_), hyphens (-) and periods (.).\nDefault value: Lifecycle Hook ID",
      "AllowedPattern": "^[a-zA-Z0-9\\u4e00-\\u9fa5][-_.a-zA-Z0-9\\u4e00-\\u9fa5]{1,63}$"
    },
    "NotificationArn": {
      "Type": "String",
      "Description": "The Alibaba Cloud Resource Name (ARN) of the notification target that Auto Scaling will use to notify you when an instance is in the transition state for the lifecycle hook. This target can be either an MNS queue or an MNS topic. The format of the parameter value is acs:ess:{region}:{account-id}:{resource-relative-id}.\nregion: the region to which the scaling group locates\naccount-id: Alibaba Cloud ID\nFor example:\nMNS queue: acs:ess:{region}:{account-id}:queue/{queuename}\nMNS topic: acs:ess:{region}:{account-id}:topic/{topicname}",
      "AllowedPattern": "^acs:ess:([a-zA-Z0-9-]+):(\\d+):(queue|topic)/([a-zA-Z0-9][a-zA-Z0-9-]{0,255})$",
      "MaxLength": 300
    },
    "ScalingGroupId": {
      "Type": "String",
      "Description": "The ID of the scaling group."
    },
    "LifecycleTransition": {
      "Type": "String",
      "Description": "The scaling activities to which lifecycle hooks apply Value range:\n SCALE_OUT: scale-out event\n SCALE_IN: scale-in event",
      "AllowedValues": [
        "SCALE_OUT",
        "SCALE_IN"
      ]
    },
    "HeartbeatTimeout": {
      "Type": "Number",
      "Description": "The time, in seconds, that can elapse before the lifecycle hook times out. If the lifecycle hook times out, the scaling group performs the default action (DefaultResult). The range is from 30 to 21,600 seconds. The default value is 600 seconds.\nYou can prevent the lifecycle hook fro
```

```
m timing out by calling the RecordLifecycleActionHeartbeat operation. You can also terminate the lifec
ycle action by calling the CompleteLifecycleAction operation.",
      "MinValue": 30,
      "MaxValue": 21600
    },
    "NotificationMetadata": {
      "Type": "String",
      "Description": "The fixed string that you want to include when Auto Scaling sends a message abou
t the wait state of the scaling activity to the notification target. The length of the parameter can b
e up to 128 characters. Auto Scaling will send the specified NotificationMetadata parameter along with
the notification message so that you can easily categorize your notifications. The NotificationMetadat
a parameter will only take effect after you specify the NotificationArn parameter.",
      "MaxLength": 128
    },
    "DefaultResult": {
      "Type": "String",
      "Description": "The action that the scaling group takes when the lifecycle hook times out. Value
range:\n CONTINUE: the scaling group continues with the scale-in or scale-out process.\n ABANDON: the
scaling group stops any remaining action of the scale-in or scale-out event.\nDefault value: CONTINUE\
nIf the scaling group has multiple lifecycle hooks and one of them is terminated by the DefaultResult=
ABANDON parameter during a scale-in event (SCALE_IN), the remaining lifecycle hooks under the same sca
ling group will also be terminated. Otherwise, the action following the wait state is the next action,
as specified in the parameter DefaultResult, after the last lifecycle event under the same scaling gro
up.",
      "AllowedValues": [
        "CONTINUE",
        "ABANDON"
      ]
    }
  },
  "Resources": {
    "LifecycleHook": {
      "Type": "ALIYUN::ESS::LifecycleHook",
      "Properties": {
        "LifecycleHookName": {
          "Ref": "LifecycleHookName"
        },
        "NotificationArn": {
          "Ref": "NotificationArn"
        },
        "ScalingGroupId": {
          "Ref": "ScalingGroupId"
        },
        "LifecycleTransition": {
          "Ref": "LifecycleTransition"
        },
        "HeartbeatTimeout": {
          "Ref": "HeartbeatTimeout"
        },
        "NotificationMetadata": {
          "Ref": "NotificationMetadata"
        },
        "DefaultResult": {
          "Ref": "DefaultResult"
        }
      }
    }
  },
  "Outputs": {
    "LifecycleHookId": {
```

```
      "Description": "The lifecycle hook ID",
      "Value": {
        "Fn::GetAtt": [
          "LifecycleHook",
          "LifecycleHookId"
        ]
      }
    }
  }
}
```

5.1.6.2.4. ALIYUN::ESS::ScalingConfiguration

ALIYUN::ESS::ScalingConfiguration is used to create a scaling configuration for a scaling group.

Statement

```
{
  "Type": "ALIYUN::ESS::ScalingConfiguration",
  "Properties": {
    "PasswordInherit": Boolean,
    "DiskMappings": List,
    "RamRoleName": String,
    "IoOptimized": String,
    "InternetChargeType": String,
    "KeyPairName": String,
    "InstanceId": String,
    "InstanceTypes": List,
    "ImageId": String,
    "ResourceGroupId": String,
    "SpotStrategy": String,
    "InstanceType": String,
    "SystemDiskCategory": String,
    "SystemDiskSize": Integer,
    "SystemDiskAutoSnapshotPolicyId": String,
    "InternetMaxBandwidthOut": Integer,
    "InstanceName": String,
    "InternetMaxBandwidthIn": Integer,
    "ScalingConfigurationName": String,
    "UserData": String,
    "DeploymentSetId": String,
    "SecurityGroupId": String,
    "SpotPriceLimit": Number,
    "HpcClusterId": String,
    "ScalingGroupId": String,
    "SpotPriceLimitForInstanceType": Map,
    "TagList": List
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	True	The ID of the resource group to which the instance belongs.	None

Parameter	Type	Required	Editable	Description	Constraint
DeploymentSetId	String	Yes	Released	The ID of the deployment set.	None
HpcClusterId	String	Yes	Released	The ID of the E-HPC cluster to which the instance belongs.	None
ScalingGroupId	String	No	No	The ID of the scaling group to which the scaling configuration belongs.	None
DiskMappings	List	No.	True	The disks to be attached to created instances.	A maximum of 16 disks can be attached to each instance.
InternetChargeType	String	Yes	True	The billing method for Internet usage.	Valid values: • PayByBandwidth • PayByTraffic: pay-by-traffic Default value: PayByTraffic
InternetMaxBandwidthIn	String	Optional	Released	The maximum inbound bandwidth from the Internet.	Unit: Mbit/s. Valid values: 1 to 100. Default value: 100
InternetMaxBandwidthOut	String	No.	True	The maximum outbound bandwidth to the Internet.	Valid values: • Pay-by-bandwidth: 0 to 100. Default value: 0. • Pay-by-data-transfer: 1 to 200. This parameter is required. Unit: Mbit/s.
InstanceId	String	Yes	Released	The instance ID of the scaling configuration.	None

Parameter	Type	Required	Editable	Description	Constraint
SystemDiskCategory	String	Yes	True	The category of the system disk.	Valid values: • cloud: indicates a basic disk. • cloud_efficiency: indicates an ultra disk. • cloud_ssd: indicates a standard SSD. • ephemeral_ssd: indicates a local SSD. • cloud_essd: enhanced SSD (ESSD) Default value: cloud for Generation I instance types that are not I/O optimized, default value: cloud_efficiency.
ImageId	String	Yes	True	The ID of the image used to start the instance. You can use a public image, a custom image, or an Alibaba Cloud Marketplace image.	None
InstanceType	String	Yes	True	The specification of the instance.	None
SecurityGroupId	String	Yes	True	The ID of the security group to which the instance belongs.	None
IoOptimized	String	Yes	True	Specifies whether the created instances are I/O optimized.	Valid values: • none (non-I/O optimized) • optimized Default value: none.

Parameter	Type	Required	Editable	Description	Constraint
ScalingConfigurationName	String	Yes	True	The name of the scaling configuration.	• The name must be 2 to 64 characters in length. It can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or digit. • The name of the scaling configuration must be unique within a scaling group. • If this parameter is not specified, the value of scalingconfigurationid is used.
KeyPairName	String	Yes	True	The name of the key pair that is bound to the instance.	• This parameter is ignored if the instance type is Windows and the default value is null. • If the instance type is Linux, password logon is disabled by default.
RamRoleName	String	Yes	True	The RAM role name of the instance.	You can use RAM API ListRoles you can call this operation to query the RAM role name of an instance.
SystemDiskSize	String	No.	True	The size of the system disk. Unit: GB.	Valid values: 40 to 500. Unit: GB. If a custom image is used to create a system disk, the system disk size must be larger than the size of the custom image.
UserData	String	Yes	True	The user data that you pass when you create the instance.	The user can encode up to 16KB in size. You do not need to perform Base64 encoding. Special characters must be escaped with a backslash (\).

Parameter	Type	Required	Editable	Description	Constraint
InstanceTypes	List	No.	True	The instance types from which ECS instances can be created. If you specify InstanceTypes, InstanceType is invalid.	Up to 10 instance types can be configured in a scaling configuration. The priority of each instance type is decreased in the order of its list elements. Auto Scaling creates instances in order of priority. If an instance of the highest priority type cannot be created, Auto Scaling will create an instance of the next highest priority type.
PasswordInherit	Boolean	No.	True	Specifies whether to use the preconfigured password of the specified image.	To use this parameter, ensure that a password is configured for the specified image.
TagList	List	No.	True	The tags of the instance.	Tags must be specified as key-value pairs. You can specify a maximum of five Tag groups in the format of `{"key1": "value1", "key2": "value2", ... "key5": "value5"}`. The key can contain a maximum of 64 characters. `aliyun`, `http://` or `https://` the beginning. If you use tags, the key cannot be an empty string. The value must be 0 to 128 characters in length.
SpotStrategy	String	Yes	True	The preemption policy for pay-as-you-go instances.	Valid values: • NoSpot (pay-as-you-go instance) • SpotWithPriceLimit (a preemptible instance with a maximum price) • SpotAsPriceGo (the SpotAsPriceGo parameter that is set automatically based on the actual market price.) Default value: NoSpot.
InstanceName	String	Yes	True	The name of the instance created based on the current scaling configuration.	None

Parameter	Type	Required	Editable	Description	Constraint
SpotPriceLimit	Number	No.	True	The maximum hourly price of the instance.	A maximum of three decimal places can be specified. This parameter takes effect only when the SpotStrategy parameter is set to SpotWithPriceLimit. The value of this parameter can be overwritten by the value of the SpotPriceLimitForInstanceType parameter.
SpotPriceLimitForInstanceType	Map	No.	True	Preemptible instance type and bid of the instance.	The format is `{"<instance_type_1>": <price_limit_1>, ..., {"<instance_type_10>": <price_limit_10>}`. This parameter takes effect only when the SpotStrategy parameter is set to SpotWithPriceLimit. You can set up to 10 instance groups and prices.
SystemDiskAutoSnapshotPolicyId	String	Yes	True	The ID of the automatic snapshot policy applied to the data disk.	None

DiskMappings syntax

```
"DiskMappings": [
  {
    "Category": String,
    "Device": String,
    "SnapshotId": String,
    "Size": String,
    "Encrypted": String,
    "KMSKeyId": String,
    "Description": String,
    "DiskName": String
  }
]
```

DiskMappings properties

Parameter	Type	Required	Editable	Description	Constraint

Parameter	Type	Required	Editable	Description	Constraint
Size	String	No	No	The size of the data disk.	Valid values: • cloud: 5 to 2000 • cloud_efficiency: 20 to 32768 • cloud_ssd: 20 to 32768 • cloud_essd: 20 to 32768. • ephemeral_ssd: 5 to 800 The value of this parameter must be greater than or equal to that of the snapshot specified by SnapshotId. Unit: GiB.
Category	String	Yes	Released	The type of the data disk.	Valid values: cloud, cloud_efficiency, cloud_ssd, ephemeral_ssd, and cloud_essd. For I/O optimized instances, the default value is cloud_efficiency. For non-I/O optimized instances, the default value is cloud.
DiskName	String	Yes	Released	The name of the data disk.	The name must be 2 to 128 characters in length It can contain letters, digits, colons (:), underscores (_), and hyphens (-). It must start with a letter and cannot start with `http://` or `https://` the beginning.
Description	String	Yes	Released	The description of the data disk.	The description must be 2 to 256 characters in length. Cannot `http://` or `https://` the beginning.
Device	String	Yes	Released	The device name of the data disk.	By default, the system automatically assigns a value for this parameter when the ECS instance is created. The value starts from /dev/xvdb and ends at /dev/xvdz.

Parameter	Type	Required	Editable	Description	Constraint
SnapshotId	String	Yes	Released	The ID of the snapshot used to create the data disk.	If this parameter is specified, the Size parameter will be ignored, and the Size of the created disk will be the Size of the specified snapshot. If the snapshot was created on or before July 15, 2013, calling the snapshot is denied and InvalidSnapshot.TooOld is displayed in the response parameter.
Encrypted	String	Yes	Released	Specifies whether to encrypt the data disk.	Default value: false.
KMSKeyId	String	Yes	Released	The KMS key ID for data disk N.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

ScalingConfigurationId: the ID of the scaling configuration. This ID is a globally unique identifier (GUID) generated by the system.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ScalingConfiguration": {
      "Type": "ALIYUN::ESS::ScalingConfiguration",
      "Properties": {
        "ImageId": "ubuntu1404_64_20G_aliaegis_2015****.vhd",
        "InstanceType": "ecs.t1.small",
        "InstanceId": "i-25xhh****",
        "InternetChargeType": "PayByTraffic",
        "InternetMaxBandwidthIn": 1,
        "InternetMaxBandwidthOut": 20,
        "SystemDisk_Category": "cloud",
        "ScalingGroupId": "bwhtvpcBcKYac9fe3vd0****",
        "SecurityGroupId": "sg-25zwc****",
        "DiskMappings": [
          {
            "Size": 10
          },
          {
            "Category": "cloud",
            "Size": 10
          }
        ]
      }
    }
  },
  "Outputs": {
    "ScalingConfiguration": {
      "Value": {"get_attr": ["ScalingConfigurationId"]}
    }
  }
}
```

5.1.6.2.5. ALIYUN::ESS::ScalingGroup

ALIYUN::ESS::ScalingGroup is used to create a scaling group. A scaling group is a group of ECS instances that are dynamically scaled based on the configured scenario. A scaling group does not take effect immediately after it is created. You must use ALIYUN::ESS::ScalingGroupEnable to enable the scaling group to trigger scaling rules and execute scaling activities.

Syntax

```
{
  "Type": "ALIYUN::ESS::ScalingGroup",
  "Properties": {
    "MultiAZPolicy": String,
    "DesiredCapacity": Integer,
    "NotificationConfigurations": List,
    "ProtectedInstances": List,
    "LaunchTemplateId": String,
    "LaunchTemplateVersion": String,
    "ScalingGroupName": String,
    "VSwitchIds": List,
    "DefaultCooldown": Integer,
    "MinSize": Integer,
    "GroupDeletionProtection": Boolean,
    "MaxSize": Integer,
    "InstanceId": String,
    "VSwitchId": String,
    "LoadBalancerIds": List,
    "StandbyInstances": List,
    "RemovalPolicys": List,
    "HealthCheckType": String,
    "DBInstanceIds": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
MinSize	Integer	Yes	Yes	The minimum number of ECS instances in the scaling group.	Valid values: 0 to 1000. When the number of ECS instances in the scaling group is less than the MinSize value, Auto Scaling automatically creates ECS instances until the number of instances is equal to the MinSize value.

Property	Type	Required	Editable	Description	Constraint
MaxSize	Integer	Yes	Yes	The maximum number of ECS instances in the scaling group.	Valid values: 0 to 1000. When the number of ECS instances in the scaling group is greater than the MaxSize value, Auto Scaling removes ECS instances from the scaling group until the number of instances is equal to the MaxSize value.
ScalingGroupName	String	No	Yes	The display name of the scaling group.	• The name must be 2 to 40 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). • It must start with an uppercase letter, lowercase letter, or digit. • The name must be unique to an Alibaba Cloud account in a region. The default value is the ID of the scaling group.
LaunchTemplateId	String	No	Yes	The ID of the instance launch template from which the scaling group obtains launch configurations.	None

Property	Type	Required	Editable	Description	Constraint
LaunchTemplateVersion	String	No	Yes	The version of the instance launch template.	Valid values: • The fixed template version number. • Default: The default template version is always used. • Latest: The latest template version is always used.
RemovalPolicys	List	No	Yes	The list of one or more policies that are used to remove ECS instances from the scaling group.	Default value: OldestScalingConfiguration or OldestInstance. Valid values: • OldestInstance: removes the ECS instance that is added to the scaling group at the earliest point in time. • NewestInstance: removes the ECS instance that is added to the scaling group at the latest point in time. • OldestScalingConfiguration: removes the ECS instance that is created based on the earliest scaling configuration.
VSwitchId	String	No	No	The ID of the vSwitch.	None

Property	Type	Required	Editable	Description	Constraint
LoadBalancerIds	List	No	Yes	The ID of the Server Load Balancer (SLB) instance.	This value can be a JSON array that contains up to five SLB instance IDs. Separate multiple IDs with commas (,).
DefaultCooldown	Integer	No	Yes	The cooldown time after a scaling activity (adding or removing ECS instances) is executed.	• Valid values: 0 to 86400. • Unit: seconds. • Default value: 300. During the cooldown time, the scaling group executes only scaling activities that are triggered by Cloud Monitor event-triggered tasks.
DBInstanceIds	List	No	Yes	The list of one or more ApsaraDB RDS instance IDs.	This value can be a JSON array that contains up to eight ApsaraDB RDS instance IDs. Separate multiple IDs with commas (,).

Property	Type	Required	Editable	Description	Constraint
VSwitchIds	List	No	No	The list of one or more vSwitch IDs.	You can specify a maximum of five vSwitch IDs. If you specify this parameter, the VSwitchId parameter is ignored. vSwitches are sorted in descending order of priority. When an ECS instance cannot be created in the zone where the vSwitch with the highest priority resides, the system uses the vSwitch with the next highest priority to create the ECS instance.
MultiAZPolicy	String	No	No	The ECS instance scaling policy for the multi-zone scaling group.	Valid values: • PRIORITY: ECS instances are scaled based on the specified vSwitch. When an ECS instance cannot be created in the zone where the vSwitch with the highest priority resides, the system uses the vSwitch with the next highest priority to create the ECS instance. • BALANCE: ECS instances are distributed evenly in multiple zones specified in the scaling group. • COST_OPTIMIZ

Property	Type	Required	Editable	Description	Constraint
					ED: ECS instances are created based on the unit price of vCPUs, from low to high. Preemptible instances are created first when preemptible instance types are specified for the scaling configuration. Pay-as-you-go instances are automatically created when no preemptible instances are available due to issues such as insufficient ECS resources.
NotificationConfigurations	List	No	Yes	The notification configurations for event and resource changes.	None
ProtectedInstances	List	No	Yes	The number of protected ECS instances in the scaling group.	Maximum value: 1000.
StandbyInstances	List	No	Yes	The number of ECS instances that are in the standby state in the scaling group.	Maximum value: 1000.
HealthCheckType	String	No	Yes	The health check type.	Valid values: • ECS • NONE

Property	Type	Required	Editable	Description	Constraint
GroupDeletionProtection	Boolean	No	Yes	Specifies whether to enable deletion protection for the scaling group.	Default value: false. Valid values: • true: enables deletion protection for the scaling group. In this case, you cannot delete the scaling group. • false: disables deletion protection for the scaling group.
DesiredCapacity	Integer	No	Yes	The expected number of ECS instances in the scaling group. The scaling group automatically keeps the number of ECS instances at the expected value.	The number of ECS instances must be greater than the MinSize value and less than the MaxSize value.
InstanceId	String	No	No	The ID of the ECS instance from which the scaling group obtains configuration information to create scaling configurations.	None

Response parameters

Fn::GetAtt

ScalingGroupId: the ID of the scaling group. This ID is a globally unique identifier (GUID) that is generated by the system.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ScalingGroup": {
      "Type": "ALIYUN::ESS::ScalingGroup",
      "Properties": {
        "MaxSize": 1,
        "MinSize": 1,
        # "ScalingGroupName": "HeatCreatedR****",
        # "DefaultCooldown": 500,
        # "RemovalPolicy_1": "",
        # "RemovalPolicy_2": "",
      }
    }
  },
  "Outputs": {
    "ScalingGroup": {
         "Value": {"Fn::GetAtt": ["ScalingGroup", "ScalingGroupId"]}
    }
  }
}
```

5.1.6.2.6. ALIYUN::ESS::ScalingGroupEnable

ALIYUN::ESS::ScalingGroupEnable is used to enable a scaling group.

Syntax

```
{
  "Type": "ALIYUN::ESS::ScalingGroupEnable",
  "Properties": {
    "ScalingConfigurationId": String,
    "ScalingRuleArisExecuteVersion": Integer,
    "ScalingRuleAris": List,
    "ScalingGroupId": String,
    "RemoveInstanceIds": List,
    "InstanceIds": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ScalingGroupId	String	Yes	No	The ID of the scaling group.	None
ScalingConfigurationId	String	No	No	The ID of the scaling configuration to be activated in the scaling group.	None

Property	Type	Required	Editable	Description	Constraint
InstanceIds	List	No	Yes	The IDs of ECS instances to be added to the enabled scaling group.	A maximum of 20 instance IDs can be specified.
ScalingRuleArisExecuteVersion	Integer	No	Yes	The version of the identifier for the scaling rule to be executed. If you change this property, all scaling rules specified by ScalingRuleAris will be executed once.	Minimum value: 0.
ScalingRuleAris	List	No	Yes	The unique identifiers of scaling rules in the scaling group. Invalid unique identifiers are not displayed in the query results and no errors are reported.	A maximum of 10 scaling rule identifiers can be specified.
RemoveInstanceIds	List	No	Yes	The IDs of ECS instances to be deleted.	A maximum of 1,000 instance IDs can be specified.

Response parameters

Fn::GetAtt

- LifecycleState: the status of the scaling group.
- ScalingInstances: the instances that are automatically created in the scaling group.
- ScalingGroupId: the ID of the scaling group.
- ScalingRuleArisExecuteResultInstancesRemoved: the instances that are removed from the scaling group by executing the scaling rules specified by ScalingRuleAris.
- ScalingRuleArisExecuteResultNumberOfAddedInstances: the number of instances that are added to the scaling group by executing the scaling rules specified by ScalingRuleAris.
- ScalingInstanceDetails: the instance scaling details.
- ScalingRuleArisExecuteErrorInfo: the error information about the execution of the scaling rules specified by ScalingRuleAris.
- ScalingRuleArisExecuteResultInstancesAdded: the instances that are added to the scaling group by executing the scaling rules specified by ScalingRuleAris.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ScalingGroupEnable": {
      "Type": "ALIYUN::ESS::ScalingGroupEnable",
      "Properties": {
        "ScalingGroupId": "r0HUqbJ411cc2eQw8bU****",
        "ScalingConfigurationId": "bJlLfdexm77Ldsyptmel****",
        "InstanceIds": "",
      }
    }
  },
  "Outputs": {
    "ScalingGroupEnable": {
      "Value": {"Fn::GetAtt": ["ScalingGroupEnable", "LifecycleState"]}
    }
  }
}
```

5.1.6.2.7. ALIYUN::ESS::ScalingRule

ALIYUN::ESS::ScalingRule is used to create a scaling rule.

Syntax

```
{
  "Type": "ALIYUN::ESS::ScalingRule",
  "Properties": {
    "AdjustmentValue": Integer,
    "Cooldown": Integer,
    "ScalingGroupId": String,
    "AdjustmentType": String,
    "ScalingRuleName": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
AdjustmentValue	Integer	No	Yes	The number of ECS instances to add or release when scaling occurs. The number of ECS instances to be adjusted in a single scaling activity cannot exceed 500.	Valid values in different adjustment modes: • QuantityChangeInCapacity: -500 to 500. • PercentChangeInCapacity: -100 to 10000. • TotalCapacity: 0 to 1000.
Cooldown	Integer	No	Yes	The cooldown period of the scaling rule. Unit: seconds.	Valid values: 0 to 86400. This parameter is empty by default.

Property	Type	Required	Editable	Description	Constraint
ScalingGroupId	String	Yes	No	The ID of the scaling group to which the scaling rule belongs.	None
AdjustmentType	String	Yes	Yes	The adjustment mode of the scaling rule.	Valid values: • QuantityChangeInCapacity: adds or removes a specified number of ECS instances. • PercentChangeInCapacity: adds or removes a specified proportion of ECS instances. • TotalCapacity: adds or removes ECS instances to ensure that the current scaling group has a specified number of ECS instances.
ScalingRuleName	String	No	Yes	The display name of the scaling rule.	The name must be 2 to 40 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or digit. The name of a scaling rule must be unique within the scaling group that it belongs to. The default value is the ID of the scaling rule.

Response parameters

Fn::GetAtt

- ScalingRuleAri: the unique identifier of the scaling rule.
- ScalingRuleId: the ID of the scaling rule. It is a globally unique identifier (GUID) generated by the system.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ScalingRule": {
      "Type": "ALIYUN::ESS::ScalingRule",
      "Properties": {
        "ScalingRuleName": {
          "Ref": "ScalingRuleName"
        },
        "Cooldown": {
          "Ref": "Cooldown"
        },
        "ScalingGroupId": {
```

```
          "Ref": "ScalingGroupId"
        },
        "AdjustmentType": {
          "Ref": "AdjustmentType"
        },
        "AdjustmentValue": {
          "Ref": "AdjustmentValue"
        }
      }
    }
  },
  "Parameters": {
    "ScalingRuleName": {
      "AllowedPattern": "^[a-zA-Z0-9\\u4e00-\\u9fa5][-_.a-zA-Z0-9\\u4e00-\\u9fa5]{1,63}$",
      "Type": "String",
      "Description": "Name shown for the scaling group, which is a string containing 2 to 40 English or Chinese characters. It must begin with a number, a letter (case-insensitive) or a Chinese character and can contain numbers, \"_\", \"-\" or \". \". The account name in the same scaling group is unique in the same region. If this parameter value is not specified, the default value is ScalingRuleId."
    },
    "Cooldown": {
      "Type": "Number",
      "Description": "Cool-down time of a scaling rule. Value range: [0, 86,400], in seconds. The default value is empty.",
      "MaxValue": 86400,
      "MinValue": 0
    },
    "ScalingGroupId": {
      "Type": "String",
      "Description": "ID of the scaling group of a scaling rule."
    },
    "AdjustmentType": {
      "Type": "String",
      "Description": "Adjustment mode of a scaling rule. Optional values:\n- QuantityChangeInCapacity: It is used to increase or decrease a specified number of ECS instances.\n- PercentChangeInCapacity: It is used to increase or decrease a specified proportion of ECS instances.\n- TotalCapacity: It is used to adjust the quantity of ECS instances in the current scaling group to a specified value.",
      "AllowedValues": [
        "QuantityChangeInCapacity",
        "PercentChangeInCapacity",
        "TotalCapacity"
      ]
    },
    "AdjustmentValue": {
      "Type": "Number",
      "Description": "Adjusted value of a scaling rule. Value range:\n- QuantityChangeInCapacity: [-500, 500]\n- PercentChangeInCapacity: [-100, 10000]\n- TotalCapacity: [0, 1000]",
      "MaxValue": 10000,
      "MinValue": -500
    }
  },
  "Outputs": {
    "ScalingRuleAri": {
      "Description": "Unique identifier of a scaling rule.",
      "Value": {
        "Fn::GetAtt": [
          "ScalingRule",
          "ScalingRuleAri"
        ]
      }
```

```
    },
    "ScalingRuleId": {
      "Description": "ID of a scaling rule, generated by the system and globally unique.",
      "Value": {
        "Fn::GetAtt": [
          "ScalingRule",
          "ScalingRuleId"
        ]
      }
    }
  }
}
```

5.1.6.2.8. ALIYUN::ESS::ScheduledTask

ALIYUN::ESS::ScheduledTask is used to create a scheduled task based on input parameters.

Syntax

```
{
  "Type": "ALIYUN::ESS::ScheduledTask",
  "Properties": {
    "TaskEnabled": Boolean,
    "Description": String,
    "ScheduledTaskName": String,
    "LaunchExpirationTime": Integer,
    "LaunchTime": String,
    "RecurrenceEndTime": String,
    "RecurrenceType": String,
    "RecurrenceValue": String,
    "ScheduledAction": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
TaskEnabled	Boolean	No	Yes	Specifies whether to start the scheduled task. • true: starts the scheduled task. • false: stops the scheduled task. Default value: true.	None
Description	String	No	Yes	The description of the scheduled task.	The description must be 2 to 200 characters in length.

Property	Type	Required	Editable	Description	Constraint
ScheduledTaskName	String	No	Yes	The display name of the scheduled task.	The name must be 2 to 40 characters in length and can contain letters, digits, underscores (_), hyphens (-), and periods (.). It must start with a letter or digit. This parameter must be unique in a region and under an Apsara Stack tenant account. The default value is the ID of the scheduled scaling task.
LaunchExpirationTime	Integer	No	Yes	The time period during which a failed scheduled task is retried. Unit: seconds. Default value: 600.	Valid values: 0 to 21600.
LaunchTime	String	Yes	Yes	The time at which the scheduled task is triggered. Specify the time in the ISO 8601 standard in the YYYY-MM-DDThh:mmZ format. The time must be in UTC. If the RecurrenceType parameter is specified, the task is executed each day at the time specified by LaunchTime. If the RecurrenceType parameter is not specified, the task is only executed once at the date and time specified by LaunchTime. You cannot enter a point in time later than 90 days from the date of scheduled task creation or modification.	None

Property	Type	Required	Editable	Description	Constraint
RecurrenceEndTime	String	No	Yes	The end time after which the scheduled task will not be repeated. Specify the time in the ISO 8601 standard in the YYYY-MM-DDThh:mmZ format. The time must be in UTC. You cannot enter a point in time later than 90 days from the date of scheduled task creation or modification. If you set RecurrenceEndTime, you must also set both RecurrenceType and RecurrenceValue.	None
RecurrenceType	String	No	Yes	The interval that the scheduled task is repeated at.	Valid values: • Daily: The scheduled task is executed once every specified number of days. • Weekly: The scheduled task is executed on each specified day of a week. • Monthly: The scheduled task is executed on each specified day of a month. • Cron: The scheduled task is executed based on the specified Cron expression. If you set RecurrenceType, you must also set both RecurrenceEndTime and RecurrenceValue.

Property	Type	Required	Editable	Description	Constraint
RecurrenceValue	String	No	Yes	Specifies how often the scheduled task recurs.	• Daily: indicates the interval of days that the scheduled task is repeated on. You can enter a single value ranging from 1 to 31. • Weekly: indicates which days of the week that the scheduled task is repeated on. You can enter multiple values separated by commas (,). The values 0 to 6 correspond to the days of the week in sequence from Sunday to Saturday. • Monthly: indicates which days of the month that the scheduled task is repeated on. You can enter two values ranging from 1 to 31. The format is A-B. B must be greater than or equal to A. • Cron: indicates a user-defined Cron expression that the scheduled task is repeated on. A Cron expression is written in UTC time and consists of five fields: minute, hour, day of month (date), month, and day of week. The expression can contain wildcard characters including commas (,), question marks (?), hyphens (-), asterisks (*), number signs (#), forward slashes (/), and the L and W characters. If you set RecurrenceValue, you must also set both RecurrenceEndTime and RecurrenceType.
ScheduledAction	String	Yes	Yes	The operations to be performed when the scheduled task is triggered. When you set this parameter, you must also enter the unique identifier of the scaling rule.	The parameter value can be up to 200 characters in length.

Response parameters

Fn::GetAtt

ScheduledTaskId: the ID of the scheduled task. This ID is a globally unique identifier (GUID) generated by the system.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "ScheduledTask": {
      "Type": "ALIYUN::ESS::ScheduledTask",
      "Properties": {
        "TaskEnabled": "true",
        "Description": "scheduledtask",
        "ScheduledTaskName": "task1",
        "LaunchTime": "2014-08-17T16:52Z",
        "RecurrenceEndTime": "2014-08-17T16:55Z",
        "RecurrenceType": "Daily",
        "RecurrenceValue": "1",
        "ScheduledAction": "ari:acs:ess:cn-qingdao:1344371:scalingRule/cCBpdYdQuBe2cUxOdu6piOk"
      }
    }
  },
  "Outputs": {
    "ScheduledTaskId": {
      "Value": {
        "FN::GetAtt": [
          "ScheduledTask",
          "ScheduledTaskId"
        ]
      }
    }
  }
}
```

5.1.6.3. NAS

5.1.6.3.1. ALIYUN::NAS::AccessGroup

ALIYUN::NAS::AccessGroup is used to create a permission group.

Syntax

```
{
  "Type": "ALIYUN::NAS::AccessGroup",
  "Properties": {
    "AccessGroupType": String,
    "AccessGroupName": String,
    "Description": String
  }
}
```

Properties

Name	Type	Required	Editable	Description	Validity
AccessGroupType	String	Yes	No	The type of the permission group.	Valid values: VPC and Classic.
AccessGroupName	String	Yes	No	The name of the permission group.	The name must be 3 to 64 characters in length. It must start with a letter, digit, or hyphen (-).
Description	String	No	Yes	The description of the permission group. The description is the same as the permission group name by default.	None

Response parameters

FN::GetAtt

AccessGroupName: the name of the permission group.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "AccessGroup": {
      "Type": "ALIYUN::NAS::AccessGroup",
      "Properties": {
        "AccessGroupType": {
          "Ref": "AccessGroupType"
        },
        "AccessGroupName": {
          "Ref": "AccessGroupName"
        },
        "Description": {
          "Ref": "Description"
        }
      }
    }
  },
  "Parameters": {
    "AccessGroupType": {
      "Type": "String",
      "Description": "Permission group type, including the Vpc and Classic types",
      "AllowedValues": ["Vpc", "Classic"]
    },
    "AccessGroupName": {
      "AllowedPattern": "^[a-zA-Z0-9-]{3,64}$",
      "Type": "String",
      "Description": "Permission group name"
    },
    "Description": {
      "Type": "String",
      "Description": "Permission group description. It is the same as the permission group name by default."
    }
  },
  "Outputs": {
    "AccessGroupName": {
      "Description": "Permission group name",
      "Value": {
        "Fn::GetAtt": ["AccessGroup", "AccessGroupName"]
      }
    }
  }
}
```

5.1.6.3.2. ALIYUN::NAS::AccessRule

ALIYUN::NAS::AccessRule is used to create a permission rule.

Syntax

```
{
  "Type": "ALIYUN::NAS::AccessRule",
  "Properties": {
    "Priority": Integer,
    "UserAccessType": String,
    "AccessGroupName": String,
    "SourceCidrIp": String,
    "RWAccessType": String
  }
}
```

Properties

Name	Type	Required	Editable	Description	Validity
Priority	Integer	No	Yes	The user access priority.	Valid values: 1 to 100. Default value: 1.
UserAccessType	String	No	Yes	The user access type.	Valid values: no_squash, root_squash, and all_squash. Default value: no_squash.
AccessGroupName	String	Yes	No	The name of the permission group.	None
SourceCidrIp	String	Yes	Yes	The authorized IP address or CIDR block.	None
RWAccessType	String	No	Yes	The read/write permission type.	Valid values: RDWR and RDONLY. Default value: RDWR.

Response parameters

FN::GetAtt

AccessRuleId: the ID of the permission rule.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "AccessRule": {
      "Type": "ALIYUN::NAS::AccessRule",
      "Properties": {
        "Priority": {
          "Ref": "Priority"
        },
        "RWAccessType ":{
          "Ref": "RWAccessType"
        },
        "UserAccessType": {
```

```
          "Ref": "UserAccessType"
        },
        "SourceCidrIp": {
          "Ref": "SourceCidrIp"
        },
        "AccessGroupName": {
          "Ref": "AccessGroupName"
        }
      }
    }
  },
  "Parameters": {
    "Priority": {
      "Default": 1,
      "Type": "Number",
      "Description": "Priority level. Range: 1-100. Default value: 1",
      "MaxValue": 100,
      "MinValue": 1
    },
    "RWAccessType": {
      "Default": "RDWR",
      "Type": "String",
      "Description": "Read-write permission type: RDWR (default), RDONLY",
      "AllowedValues": ["RDWR", "RDONLY"]
    },
    "UserAccessType": {
      "Default": "no_squash",
      "Type": "String",
      "Description": "User permission type: no_squash (default), root_squash, all_squash",
      "AllowedValues": ["no_squash", "root_squash", "all_squash"]
    },
    "SourceCidrIp": {
      "Type": "String",
      "Description": "Address or address segment"
    },
    "AccessGroupName": {
      "Type": "String",
      "Description": "Permission group name"
    }
  },
  "Outputs": {
    "AccessRuleId": {
      "Description": "Rule serial number",
      "Value": {
        "Fn::GetAtt": ["AccessRule", "AccessRuleId"]
      }
    }
  }
}
```

5.1.6.3.3. ALIYUN::NAS::FileSystem

ALIYUN::NAS::FileSystem is used to create a file system.

Syntax

```
{
  "Type": "ALIYUN::NAS::FileSystem",
  "Properties": {
    "SnapshotId": String,
    "Description": String,
    "StorageType": String,
    "DeletionForce": Boolean,
    "EncryptType": Integer,
    "VpcId": String,
    "ZoneId": String,
    "Capacity": Integer,
    "Tags": List,
    "ProtocolType": String,
    "FileSystemType": String,
    "Bandwidth": Integer,
    "VSwitchId": String,
    "Duration": Integer,
    "ChargeType": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ProtocolType	String	Yes	No	The protocol type.	Valid values: • NFS • SMB

StorageType	String	Yes	No	The storage type.	• Valid values when FileSystemType is set to standard:
					  ◦ Performance
					  ◦ Capacity
					• Valid values when FileSystemType is set to extreme or cpfs:
					  ◦ standard
					  ◦ advance

DeletionForce	Boolean	No	Yes	Specifies whether to forcibly delete all mount targets on the file system and then delete the file system.	Default value: false. Valid values: • true • false
Description	String	No	Yes	The description of the file system.	The description must be 2 to 128 characters in length and can contain letters, digits, colons(:), underscores (_), and hyphens (-). It must start with a letter and cannot start with `http://` or `https://`.

Property	Type	Required	Editable	Description	Constraint
ZoneId	String	No	No	The ID of the zone.	None
Tags	List	No	No	The list of one or more tags of the file system.	You can bind up to 20 tags to each file system. For more information, see Tags properties.
SnapshotId	String	No	No	The ID of the snapshot.	You can specify this parameter to create the NAS file system from the specified snapshot. This parameter takes effect only for Extreme NAS. **Note** If you create a file system from a snapshot, the version of the file system is the same as the version of the source file system of the snapshot. If the version of the source file system of a snapshot is 1 and you want to create a file system of version 2, perform the following operations: Create a file system (File System 1) from the snapshot. Create another file system (File System 2) whose version is 2. Migrate data from File System 1 to File System 2. Switch your business from File System 1 to File System 2.

Property	Type	Required	Editable	Description	Constraint
EncryptType	Integer	No	No	Specifies whether to encrypt the file system. You can use keys that are hosted by Key Management Service (KMS) to encrypt data that is stored in a file system. Data is automatically decrypted when you access encrypted data.	This parameter takes effect only when the FileSystemType parameter is set to standard or extreme. Valid values: • 0: The file system is not encrypted. • 1: The file system is encrypted.
Capacity	Integer	No	No	The capacity of the file system.	This parameter takes effect only when the FileSystemType parameter is set to standard or extreme. • Valid values when FileSystemType is set to extreme: 100 to 262144. • Valid values when FileSystemType is set to cpfs: 2048 to 512000. Unit: GB.
FileSystemType	String	No	No	The type of the file system.	Default value: standard. Valid values: • standard • extreme • cpfs
VpcId	String	No	No	The ID of the VPC. If you specify the VpcId and VSwitchId parameters, a default mount target is pre-configured when the file system is created.	This parameter is required when the FileSystemType parameter is set to cpfs.
Bandwidth	Integer	No	No	The maximum throughput of the file system.	This parameter is required when the FileSystemType parameter is set to cpfs. The value of this parameter is determined by the Capacity parameter. For more information, see the CPFS buy page. Unit: Mbit/s.

Property	Type	Required	Editable	Description	Constraint
VSwitchId	String	No	No	The ID of the vSwitch. If you specify the VpcId and VSwitchId parameters, a default mount target is pre-configured when the file system is created.	This parameter is required when the FileSystemType parameter is set to cpfs.
Duration	Integer	No	No	The subscription period of the file system.	This parameter is valid and required only when the ChargeType parameter is set to Subscription. If you do not renew a subscription file system before it expires, the file system is released. Valid values: • 1 • 2 • 3 • 6 • 12 • 36 Unit: months.
ChargeType	String	No	Yes	The billing method of the file system.	Valid values: • PayAsYouGo • Subscription

Tags syntax

```
"Tags": [
  {
    "Key": String,
    "Value": String
  }
]
```

Tags properties

Property	Type	Required	Editable	Description	Constraint
Key	String	Yes	No	The key of the tag.	The tag key must be 1 to 128 characters in length and cannot contain `http://` or `https://`. It cannot start with `acs:` or `aliyun`.

Property	Type	Required	Editable	Description	Constraint
Value	String	No	No	The value of the tag.	The tag value must be 0 to 128 characters in length and cannot contain `http://` or `https://`. It cannot start with `acs:` or `aliyun`.

Response parameters

Fn::GetAtt

FileSystemId: the ID of the file system.

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Description": {
      "Type": "String",
      "Description": "File system description (space characters are not allowed)"
    },
    "StorageType": {
      "Type": "String",
      "Description": "The file system type. Currently includes the Performance type and the Capacity type"
    },
    "ZoneId": {
      "Type": "String",
      "Description": "Zone ID."
    },
    "VSwitchId": {
      "Type": "String",
      "Description": "VSwitch ID."
    },
    "Duration": {
      "Type": "Number",
      "Description": "The period of subscription in months. Required and valid when ChargeType is Subscription.\nWhen the annual and monthly subscription instance expires without renewal, the instance will automatically expire and be released."
    },
    "SnapshotId": {
      "Type": "String",
      "Description": "Snapshot ID."
    },
    "DeletionForce": {
      "Type": "Boolean",
      "Description": "Whether delete all mount targets on the file system and then delete file system. Default is false",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ],
      "Default": false
```

```
    },
    "EncryptType": {
      "Type": "Number",
      "Description": "Specifies whether to encrypt data. You can use keys that are hosted by Key Management Service (KMS) to encrypt data stored on a file system. Data is automatically decrypted when you access encrypted data. Valid values:\n0: specifies that no encryption is applied to data on the file system.\n1: specifies that encryption is applied to data on the file system."
    },
    "VpcId": {
      "Type": "String",
      "Description": "Vpc ID."
    },
    "Capacity": {
      "Type": "Number",
      "Description": "File system capacity, the unit is GB. Required and valid when FileSystemType=extreme or cpfs."
    },
    "ProtocolType": {
      "Type": "String",
      "Description": "Type of protocol used. Currently includes the NFS type and the SMB type",
      "AllowedValues": [
        "NFS",
        "SMB"
      ]
    },
    "ChargeType": {
      "Type": "String",
      "Description": "Type of payment:\nPayAsYouGo (pay as you go)\nSubscription",
      "AllowedValues": [
        "PayAsYouGo",
        "Subscription"
      ]
    },
    "FileSystemType": {
      "Type": "String",
      "Description": "File system type. Allowed values: standard, extreme, cpfs"
    },
    "Bandwidth": {
      "Type": "Number",
      "Description": "Maximum file system throughput, unit is MB/s. Required and valid only when FileSystemType=cpfs."
    },
    "Tags": {
      "Type": "Json",
      "Description": "Tags to attach to filesystem. Max support 20 tags to add during create filesystem. Each tag with two properties Key and Value, and Key is required.",
      "MaxLength": 20
    }
  },
  "Resources": {
    "FileSystem": {
      "Type": "ALIYUN::NAS::FileSystem",
      "Properties": {
        "Description": {
          "Ref": "Description"
        },
        "StorageType": {
          "Ref": "StorageType"
        },
        "ZoneId": {
```

```
          "Ref": "ZoneId"
        },
        "VSwitchId": {
          "Ref": "VSwitchId"
        },
        "Duration": {
          "Ref": "Duration"
        },
        "SnapshotId": {
          "Ref": "SnapshotId"
        },
        "DeletionForce": {
          "Ref": "DeletionForce"
        },
        "EncryptType": {
          "Ref": "EncryptType"
        },
        "VpcId": {
          "Ref": "VpcId"
        },
        "Capacity": {
          "Ref": "Capacity"
        },
        "ProtocolType": {
          "Ref": "ProtocolType"
        },
        "ChargeType": {
          "Ref": "ChargeType"
        },
        "FileSystemType": {
          "Ref": "FileSystemType"
        },
        "Bandwidth": {
          "Ref": "Bandwidth"
        },
        "Tags": {
          "Ref": "Tags"
        }
      }
    }
  },
  "Outputs": {
    "FileSystemId": {
      "Description": "ID of the file system created",
      "Value": {
        "Fn::GetAtt": [
          "FileSystem",
          "FileSystemId"
        ]
      }
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Parameters:
  Description:
    Type: String
```

```
    Type: String
    Description: File system description (space characters are not allowed)
  StorageType:
    Type: String
    Description: >-
      The file system type. Currently includes the Performance type and the
      Capacity type
  ZoneId:
    Type: String
    Description: Zone ID.
  VSwitchId:
    Type: String
    Description: VSwitch ID.
  Duration:
    Type: Number
    Description: >-
      The period of subscription in months. Required and valid when ChargeType
      is Subscription.
      When the annual and monthly subscription instance expires without renewal,
      the instance will automatically expire and be released.
  SnapshotId:
    Type: String
    Description: Snapshot ID.
  DeletionForce:
    Type: Boolean
    Description: >-
      Whether delete all mount targets on the file system and then delete file
      system. Default is false
    AllowedValues:
      - 'True'
      - 'true'
      - 'False'
      - 'false'
    Default: false
  EncryptType:
    Type: Number
    Description: >-
      Specifies whether to encrypt data. You can use keys that are hosted by Key
      Management Service (KMS) to encrypt data stored on a file system. Data is
      automatically decrypted when you access encrypted data. Valid values:
      0: specifies that no encryption is applied to data on the file system.
      1: specifies that encryption is applied to data on the file system.
  VpcId:
    Type: String
    Description: Vpc ID.
  Capacity:
    Type: Number
    Description: >-
      File system capacity, the unit is GB. Required and valid when
      FileSystemType=extreme or cpfs.
  ProtocolType:
    Type: String
    Description: Type of protocol used. Currently includes the NFS type and the SMB type
    AllowedValues:
      - NFS
      - SMB
  ChargeType:
    Type: String
    Description: |-
      Type of payment:
```

```
      PayAsYouGo (pay as you go)
      Subscription
    AllowedValues:
      - PayAsYouGo
      - Subscription
  FileSystemType:
    Type: String
    Description: 'File system type. Allowed values: standard, extreme, cpfs'
  Bandwidth:
    Type: Number
    Description: >-
      Maximum file system throughput, unit is MB/s. Required and valid only when
      FileSystemType=cpfs.
  Tags:
    Type: Json
    Description: >-
      Tags to attach to filesystem. Max support 20 tags to add during create
      filesystem. Each tag with two properties Key and Value, and Key is
      required.
    MaxLength: 20
Resources:
  FileSystem:
    Type: 'ALIYUN::NAS::FileSystem'
    Properties:
      Description:
        Ref: Description
      StorageType:
        Ref: StorageType
      ZoneId:
        Ref: ZoneId
      VSwitchId:
        Ref: VSwitchId
      Duration:
        Ref: Duration
      SnapshotId:
        Ref: SnapshotId
      DeletionForce:
        Ref: DeletionForce
      EncryptType:
        Ref: EncryptType
      VpcId:
        Ref: VpcId
      Capacity:
        Ref: Capacity
      ProtocolType:
        Ref: ProtocolType
      ChargeType:
        Ref: ChargeType
      FileSystemType:
        Ref: FileSystemType
      Bandwidth:
        Ref: Bandwidth
      Tags:
        Ref: Tags
Outputs:
  FileSystemId:
    Description: ID of the file system created
    Value:
      'Fn::GetAtt':
        - FileSystem
```

```
      - FileSystemId
```

5.1.6.3.4. ALIYUN::NAS::MountTarget

ALIYUN::NAS::MountTarget is used to create a mount point.

Syntax

```
{
  "Type": "ALIYUN::NAS::MountTarget",
  "Properties": {
    "Status": String,
    "VpcId": String,
    "FileSystemId": String,
    "VSwitchId": String,
    "NetworkType": String,
    "AccessGroupName": String
  }
}
```

Properties

Name	Type	Required	Editable	Description	Validity
Status	String	No	Yes	The status of a file system.	Valid values: Active and Inactive.
VpcId	String	No	No	The ID of the VPC to which the file system belongs.	None
FileSystemId	String	Yes	No	The ID of the file system.	None
VSwitchId	String	No	No	The ID of the vSwitch in the VPC.	None
NetworkType	String	Yes	No	The network type.	Valid values: VPC and Classic.
AccessGroupName	String	Yes	Yes	The permission group name.	None

Response parameters

FN::GetAtt

MountTargetDomain: the domain name of the mount point.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "MountTarget": {
      "Type": "ALIYUN::NAS::MountTarget",
```

```
      "Properties": {
        "Status": {
          "Ref": "Status"
        },
        "VpcId": {
          "Ref": "VpcId"
        },
        "FileSystemId": {
          "Ref": "FileSystemId"
        },
        "VSwitchId": {
          "Ref": "VSwitchId"
        },
        "NetworkType": {
          "Ref": "NetworkType"
        },
        "AccessGroupName": {
          "Ref": "AccessGroupName"
        }
      }
    }
  },
  "Parameters": {
    "Status": {
      "Type": "String",
      "Description": "Status, including Active and Inactive",
      "AllowedValues": ["Active", "Inactive"]
    },
    "VpcId": {
      "Type": "String",
      "Description": "VPC network ID"
    },
    "FileSystemId": {
      "Type": "String",
      "Description": "File system ID"
    },
    "VSwitchId": {
      "Type": "String",
      "Description": "vSwitch ID."
    },
    "NetworkType": {
      "Type": "String",
      "Description": "Network type, including Vpc and Classic networks.",
      "AllowedValues": ["Vpc", "Classic"]
    },
    "AccessGroupName": {
      "Type": "String",
      "Description": "Permission group name"
    }
  },
  "Outputs": {
    "MountTargetDomain": {
      "Description": "Mount point domain name",
      "Value": {
        "Fn::GetAtt": ["MountTarget", "MountTargetDomain"]
      }
    }
  }
}
```

5.1.6.4. OSS

5.1.6.4.1. ALIYUN::OSS::Bucket

ALIYUN::OSS::Bucket is used to create an OSS bucket.

Syntax

```
{
  "Type": "ALIYUN::OSS::Bucket",
  "Properties": {
    "AccessControl": String,
    "RefererConfiguration": Map,
    "ServerSideEncryptionConfiguration": Map,
    "CORSConfiguration": Map,
    "Tags": Map,
    "LoggingConfiguration": Map,
    "LifecycleConfiguration": Map,
    "StorageClass": String,
    "DeletionForce": Boolean,
    "WebsiteConfiguration": Map,
    "Policy": Map,
    "BucketName": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
BucketName	String	Yes	No	The name of the bucket.	• The name must be 3 to 63 characters in length and can contain lowercase letters, digits, and hyphens (-). • It must start and end with a lowercase letter or digit.
AccessControl	String	No	No	The access control policy.	Valid values: private, public-read, and public-read-write.
CORSConfiguration	Map	No	No	The configuration of cross-origin resource sharing for objects in the bucket.	None
LifecycleConfiguration	Map	No	No	The lifecycle configuration for objects in the bucket.	None
LoggingConfiguration	Map	No	No	The logging configuration.	None

Property	Type	Required	Editable	Description	Constraint
RefererConfiguration	Map	No	No	The hotlinking protection configuration.	None
DeletionForce	Boolean	No	No	Specifies whether to forcibly delete objects from an OSS bucket	Valid values: • true • false Default value: false.
WebsiteConfiguration	Map	No	No	The information used to configure the bucket as a static website.	None
ServerSideEncryptionConfiguration	Map	No	No	The server-side encryption rules.	None
Tags	Map	No	No	The tags of the bucket. Tags exist as key-value pairs.	• A maximum of 20 tags can be specified. • A tag key must be 1 to 64 bytes in length and cannot start with `http://`, `https://`, or `Aliyun`. • A tag value can be up to 128 bytes in length and must be encoded in UTF-8.
StorageClass	String	No	No	The type of the bucket.	Valid values: Standard, IA, and Archive.
Policy	Map	No	No	The bucket policy configuration.	None

CORSConfiguration syntax

```
"CORSConfiguration": {
  "CORSRule": [
    {
      "AllowedHeader": String,
      "AllowedMethod": List,
      "AllowedOrigin": List,
      "ExposeHeader": List,
      "MaxAgeSeconds": Integer
    }
  ]
}
```

CORSConfiguration properties

Property	Type	Required	Editable	Description	Constraint
CORSRule	List	No	No	The rules that define cross-origin resource sharing of objects in the bucket.	None
AllowedHeader	String	No	No	The allowed cross-origin request headers.	Valid values: *, Cache-Control, Content-Language, Content-Type, Expires, Last-Modified, and Pragma.
AllowedMethod	List	No	No	The allowed cross-origin request methods.	Valid values: *, GET, PUT, POST, DELETE, and HEAD.
AllowedOrigin	List	No	No	The origins from which cross-origin requests are allowed.	None
ExposeHeader	List	No	No	The response headers for allowed access requests from applications.	Asterisks (*) cannot be used as wildcard characters.
MaxAgeSeconds	Integer	No	No	The period of time that the browser can cache the response of a preflight (OPTIONS) request to a specific resource.	None

LifecycleConfiguration syntax

```
"LifecycleConfiguration": {
  "Rule": [
    {
      "ID": String,
      "Prefix": String,
      "Status": String,
      "Expiration": Map,
      "AbortMultipartUpload": Map
    }
  ]
}
```

LifecycleConfiguration properties

Property	Type	Required	Editable	Description	Constraint
Rule	List	No	No	The lifecycle rule.	None
ID	String	No	No	The unique ID of the rule.	The ID can be up to 255 characters in length. When this parameter is empty or not specified, OSS generates a unique rule ID.
Prefix	String	No	No	The prefix to which the rule applies.	The rule takes effect only on objects that have a matching prefix.
Status	String	No	No	Specifies whether to enable or disable the rule.	Valid values: Enable and Disable.
Expiration	Map	No	No	The expiration attributes of the rule for the specified object.	None
AbortMultipartUpload	Map	No	No	The expiration attributes of the multipart upload tasks that are not complete.	None

Expiration syntax

```
"Expiration":{
  "Days": Number,
  "CreatedBeforeDate": String
}
```

Expiration properties

Property	Type	Required	Editable	Description	Constraint
Days	Number	No	No	The number of days since the object was last modified after which the rule will take effect.	When the number of days since the object was last modified exceeds the specified number of days, the object is deleted. If you set the Days parameter to 30, objects that were last modified on January 1, 2016 are deleted by the backend application on January 31, 2016.
CreatedBeforeDate	String	No	No	The date before which the rule takes effect.	Specify the time in the ISO 8601 standard. The time must be UTC 00:00. Example: 2002-10-11T00:00:00.000Z.

AbortMultipartUpload syntax

```
"AbortMultipartUpload": {
  "CreatedBeforeDate": String,
  "Days": Number
}
```

AbortMultipartUpload properties

Property	Type	Required	Editable	Description	Constraint
Days	Number	No	No	The number of days since the object was last modified after which the rule will take effect.	When the number of days since the object was last modified exceeds the specified number of days, the object is deleted. If you set the Days parameter to 30, objects that were last modified on January 1, 2016 are deleted by the backend application on January 31, 2016.
CreatedBeforeDate	String	No	No	The date before which the rule takes effect.	Specify the time in the ISO 8601 standard. The time must be UTC 00:00. Example: 2002-10-11T00:00:00.000Z.

LoggingConfiguration syntax

```
"LoggingConfiguration": {
  "TargetBucket": String,
  "TargetPrefix": String
}
```

LoggingConfiguration properties

Property	Type	Required	Editable	Description	Constraint
TargetBucket	String	No	No	The storage space for storing access logs.	None
TargetPrefix	String	No	No	The prefix of the names of saved access log files.	None

WebsiteConfiguration syntax

```
"WebsiteConfiguration":{
  "IndexDocument": String,
  "ErrorDocument": String
}
```

WebsiteConfiguration properties

Property	Type	Required	Editable	Description	Constraint
IndexDocument	String	No	No	The default homepage for a static website.	None
ErrorDocument	String	No	No	The default error page for a static website.	None

RefererConfiguration syntax

```
"RefererConfiguration":{
  "AllowEmptyReferer": String,
  "RefererList": List
}
```

RefererConfiguration properties

Property	Type	Required	Editable	Description	Constraint
AllowEmptyReferer	String	No	No	Specifies whether the Referer field can be left empty in an access request.	None

Property	Type	Required	Editable	Description	Constraint
RefererList	List	No	No	The referer whitelist. OSS allows requests whose Referer field values are in the referer whitelist.	None

ServerSideEncryptionConfiguration syntax

```
"ServerSideEncryptionConfiguration":{
  "KMSMasterKeyID": String,
  "SSEAlgorithm": String
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
KMSMasterKeyID	String	No	No	The ID of the customer master key.	The key ID is required only when the SSEAlgorithm value is KMS and the specified key is used for encryption.
SSEAlgorithm	String	Yes	No	The default server-side encryption method.	Valid values: KMS and AES256.

Response parameters

Fn::GetAtt

- Name: the bucket name, which must be globally unique.
- DomainName: the public domain name of the specified bucket.
- InternalDomainName: the internal domain name of the specified bucket.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Bucket": {
      "Type": "ALIYUN::OSS::Bucket",
      "Properties": {
        "AccessControl": "private",
        "BucketName": "rostest",
        "WebsiteConfiguration":{
            "IndexDocument": "index1.html",
            "ErrorDocument": "error404.html"
        },
        "LoggingConfiguration": {
            "TargetBucket": "cos-mirror",
            "TargetPrefix": "test404"
        },
        "CORSConfiguration": {
            "CORSRule": [{
                "AllowedHeader": ["*"],
                "AllowedMethod": ["GET", "PUT"],
                "AllowedOrigin": ["*"],
                "ExposeHeader": ["Date"],
                "MaxAgeSeconds": 3600
            }]
        },
        "LifecycleConfiguration": {
            "Rule": [{
                "ID": "deleteRule",
                "Prefix": "test/",
                "Status": "Enabled",
                "Expiration":{
                    "Days": 2
                },
                "AbortMultipartUpload":{
                    "CreatedBeforeDate": "2014-10-11T00:00:00.000Z"
                }
            }]
        },
        "RefererConfiguration": {
            "AllowEmptyReferer": true,
            "RefererList": ["http://www.aliyun.com", "https://www.?.aliyuncs.com"]
        }
      }
    }
  },
  "Outputs": {
    "Name": {
         "Value": {"Fn::GetAtt": ["Bucket","Name"]}
    },
    "DomainName": {
         "Value": {"Fn::GetAtt": ["Bucket","DomainName"]}
    }
  }
}
```

5.1.6.5. RDS

5.1.6.5.1. ALIYUN::RDS::Account

ALIYUN::RDS::Account is used to create a database management Account.

Statement

```
{
  "Type": "ALIYUN::RDS::Account",
  "Properties": {
    "AccountDescription": String,
    "DBInstanceId": String,
    "AccountPassword": String,
    "AccountType": String,
    "AccountName": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
AccountDescription	String	Yes	True	The description of the account.	The name must be 2 to 256 characters in length. It can contain digits, letters, underscores (_), and hyphens (-); but must start with a letter.
DBInstanceId	String	No	No	The ID of the RDS instance.	None
AccountPassword	String	No	No	The password of the database account.	The password must be 8 to 32 characters in length.
AccountType	String	Yes	Released	The type of the database account.	Valid values: • Normal: indicates a standard account. • Super: indicates a privileged account. Default value: Normal.

Parameter	Type	Required	Editable	Description	Constraint
AccountName	String	No	No	The name of the database account.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

AccountName: the name of the database account.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Account": {
      "Type": "ALIYUN::RDS::Account",
      "Properties": {
        "AccountDescription": {
          "Ref": "AccountDescription"
        },
        "DBInstanceId": [
          "Ref": "DBInstanceId"
        },
        "AccountPassword": {
          "Ref": "AccountPassword"
        },
        "AccountType": {
          "Ref": "AccountType"
        },
        "AccountName": {
          "Ref": "AccountName"
        }
      }
    }
  },
  "Parameters": {
    "AccountDescription": {
      "Type": "String",
      "Description": "Account remarks.\nIt cannot begin with http:// or https://.\nIt must start with a Chinese character or English letter.\nIt can include Chinese and English characters/letters, underscores (_), hyphens (-), and digits.\nThe length may be 2-256 characters."
    },
    "DBInstanceId": [
      "Type": "String",
      "Description": "RDS instance ID."
    },
    "AccountPassword": {
      "MinLength": 8,
      "Type": "String",
      "Description": "The account password for the database instance. It may consist of letters, digit
```

```
s, or underlines, with a length of 8 to 32 characters.",
      "MaxLength": 32
    },
    "AccountType": {
      "Default": "Normal",
      "Type": "String",
      "Description": "Privilege type of account.\nNormal: Common privilege.\nSuper: High privilege. And the default value is Normal.\nThis parameter is valid for MySQL 5.5/5.6 only.\nMySQL 5.7, SQL Server 2012/2016, PostgreSQL, and PPAS each can have only one initial account. Other accounts are created by the initial account that has logged on to the database.",
      "AllowedValues": ["Normal", "Super"]
    },
    "AccountName": {
      "Type": "String",
      "Description": "Account name, which must be unique and meet the following requirements:\nStart with a letter;\nConsist of lower-case letters, digits, and underscores (_);\nContain no more than 16 characters.\nFor other invalid characters, see Forbidden keywords table."
    }
  },
  "Outputs": {
    "AccountName": {
      "Description": "Account name",
      "Value": {
        "Fn::GetAtt": ["Account", "AccountName"]
      }
    }
  }
}
```

5.1.6.5.2. ALIYUN::RDS::AccountPrivilege

ALIYUN::RDS::AccountPrivilege is used to grant database access permissions to accounts.

Statement

```
{
  "Type": "ALIYUN::RDS::AccountPrivilege",
  "Properties": {
    "AccountPrivilege": String,
    "DBInstanceId": String,
    "DBName": String,
    "AccountName": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint

Parameter	Type	Required	Editable	Description	Constraint
AccountPrivilege	String	No	Yes	The permissions of the database account.	Valid values: • ReadWrite: has read and write permissions on the database. • ReadOnly: The account has read-only permission on the database. • DDLOnly: The account can run only data definition language (DDL) commands in the database. This is applicable to MySQL and MariaDB. • DMLOnly: The account can run only data manipulation language (DML) commands in the database. This is applicable to MySQL and MariaDB. • DBOwner: The account has full permissions on the database. This is applicable to SQL Server.
DBInstanceId	String	No	No	The ID of the RDS instance.	None
DBName	String	No	No	The name of the database.	None

Parameter	Type	Required	Editable	Description	Constraint
AccountName	String	No	No	The name of the account.	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

None

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "AccountPrivilege": {
      "Type": "ALIYUN::RDS::AccountPrivilege",
      "Properties": {
        "AccountPrivilege": {
          "Ref": "AccountPrivilege"
        },
        "DBInstanceId": [
          "Ref": "DBInstanceId"
        },
        "DBName": {
          "Ref": "DBName"
        },
        "AccountName": {
          "Ref": "AccountName"
        }
      }
    }
  },
  "Parameters": {
    "AccountPrivilege": {
      "Type": "String",
      "Description": "RDS account privilege",
      "AllowedValues": ["ReadOnly", "ReadWrite", "DDLOnly", "DMLOnly", "DBOwner"]
    },
    "DBInstanceId": [
      "Type": "String",
      "Description": "RDS instance ID."
    },
    "DBName": {
      "Type": "String",
      "Description": "RDS database name"
    },
    "AccountName": {
      "Type": "String",
      "Description": "RDS account name."
    }
  },
  "Outputs": {}
}
```

5.1.6.5.3. ALIYUN::RDS::DBInstance

ALIYUN::RDS::DBInstance is used to create an ApsaraDB RDS instance.

Syntax

```
{
  "Type": "ALIYUN::RDS::DBInstance",
  "Properties": {
    "Engine": String,
    "MultiAZ": Boolean,
    "VpcId": String,
    "DBMappings": List,
    "DBInstanceDescription": String,
    "ConnectionMode": String,
    "MasterUsername": String,
    "MasterUserPassword": String,
    "ZoneId": String,
    "DBInstanceNetType": String,
    "DBInstanceStorage": Integer,
    "VSwitchId": String,
    "AllocatePublicConnection": Boolean,
    "EngineVersion": String,
    "PreferredBackupTime": String,
    "DBInstanceClass": String,
    "SecurityIPList": String,
    "BackupRetentionPeriod": Integer,
    "PrivateIpAddress": String,
    "PreferredBackupPeriod": List,
    "PeriodType": String,
    "PayType": String,
    "Period": Integer,
    "ResourceGroupId": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	No	No	The ID of the resource group.	None
Engine	String	Yes	No	The database engine that the instance runs.	Valid values: • MySQL • SQLServer • PostgreSQL • PPAS

Property	Type	Required	Editable	Description	Constraint
DBInstanceStorage	Integer	Yes	Yes	The storage capacity of the instance.	• Valid values when Engine is set to MySQL: 5 to 1000. • Valid values when Engine is set to SQLServer: 10 to 1000. • Valid values when Engine is set to PostgreSQL: 5 to 2000. • Valid values when Engine is set to PPAS: 5 to 2000. Unit: GB. **Note** This value must be in 5 GB increments.
EngineVersion	String	Yes	No	The version of the database engine.	• Valid values when Engine is set to MySQL: 5.5, 5.6, 5.7, and 8.0. • Set the value to 2008r2 when Engine is set to SQLServer. • Set the value to 9.4 when Engine is set to PostgreSQL. • Set the value to 9.3 when Engine is set to PPAS.
DBInstanceClass	String	Yes	Yes	The instance type.	Valid values: • rds.mys2.large • rds.mss1.large • rds.pg.s1.small
SecurityIPList	String	Yes	Yes	The whitelist of IP addresses that are allowed to access all databases in the instance.	• Separate multiple IP addresses with commas (,). Each IP address in the whitelist must be unique. A maximum of 1,000 IP addresses can be specified. • The 0.0.0.0/0 format is supported. You can specify IP addresses in the 10.23.XX.XX format and CIDR blocks in the 10.23.XX.XX/24 format. In 10.23.XX.XX/24, /24 indicates the length of the prefix in the CIDR block, and the prefix length can range from 1 to 32. 0.0.0.0/0 indicates that no access restriction is applied.
MultiAZ	Boolean	No	No	Specifies whether the instance can be deployed across multiple zones.	None
VpcId	String	No	No	The ID of the VPC.	None

Property	Type	Required	Editable	Description	Constraint
DBMappings	List	No	No	The list of one or more databases to be created in the instance.	None
DBInstanceDescription	String	No	No	The description of the instance.	• The description must be 2 to 256 characters in length and can contain letters, digits, underscores (_), and hyphens (-). • It must start with a letter and cannot start with `http://` or `https://`.
ConnectionMode	String	No	No	The connection mode of the instance.	Valid values: • Performance: standard connection mode • Safty: safe connection mode If you do not specify this parameter, the system assigns a connection mode.
MasterUsername	String	No	No	The name of the database account.	The name must be unique. The name can be up to 16 characters in length and can contain letters, digits, and underscores (_).
MasterUserPassword	String	No	No	The password of the database account.	The password must be 6 to 32 characters in length and can contain letters, digits, and underscores (_).
ZoneId	String	No	No	The zone ID of the instance.	None
DBInstanceNetType	String	No	No	The network type of the instance.	Default value: Intranet. Valid values: • Internet • Intranet
VSwitchId	String	No	No	The ID of the vSwitch in the specified VPC.	None
AllocatePublicConnection	Boolean	No	No	Specifies whether to apply for a public endpoint for the instance.	None

Property	Type	Required	Editable	Description	Constraint
PreferredBackupTime	String	No	No	The backup window.	• Specify the window in the `HH:mmZ-HH:mmZ` format. • Valid values: 00:00Z-01:00Z, 01:00Z-02:00Z, 02:00Z-03:00Z, 03:00Z-04:00Z, 04:00Z-05:00Z, 05:00Z-06:00Z, 06:00Z-07:00Z, 07:00Z-08:00Z, 08:00Z-09:00Z, 09:00Z-10:00Z, 10:00Z-11:00Z, 11:00Z-12:00Z, 12:00Z-13:00Z, 13:00Z-14:00Z, 14:00Z-15:00Z, 15:00Z-16:00Z, 16:00Z-17:00Z, 17:00Z-18:00Z, 18:00Z-19:00Z, 19:00Z-20:00Z, 20:00Z-21:00Z, 21:00Z-22:00Z, 22:00Z-23:00Z, and 23:00Z-24:00Z.
BackupRetentionPeriod	Number	No	No	The number of days for which backup files can be retained.	Valid values: 7 to 30. Unit: days. Default value: 7.
PrivateIpAddress	String	No	No	The private IP address within the CIDR block of the vSwitch.	If you do not specify this parameter, the system allocates a private IP address.
PreferredBackupPeriod	List	No	No	The backup cycle.	Valid values: • Monday • Tuesday • Wednesday • Thursday • Friday • Saturday • Sunday
MasterUserType	String	No	No	The type of the database account.	Default value: Normal. Valid values: • Normal • Super
Tags	Map	No	Yes	The list of one or more tags. Each tag consists of a tag key and a tag value.	• The tag key is required and the tag value is optional. • Format example: `{"key1":"value1","key2":""}`.
PeriodType	String	No	No	The unit of the subscription period.	Default value: Month. Valid values: • Month • Year

Property	Type	Required	Editable	Description	Constraint
PayType	String	No	No	The billing method of the instance.	Valid values: • PostPaid: pay-as-you-go • PrePaid: subscription
Period	Integer	No	No	The subscription period of the instance.	• Valid values when PeriodType is set to Year: 1, 2, and 3. • Valid values when PeriodType is set to Month: 1, 2, 3, 4, 5, 6, 7, 8, and 9.

DBMappings syntax

```
"DBMappings": [
  {
    "DBDescription": String,
    "CharacterSetName": String,
    "DBName": String
  }
]
```

DBMappings properties

Property	Type	Required	Editable	Description	Constraint

CharacterSetName	String	Yes	No	The character set.	• Valid values when Engine is set to MySQL:
					  ○ utf8
					  ○ gbk
					  ○ latin1
					  ○ utf8mb4 (applicable to versions 5.5 and 5.6)
					• Valid values when Engine is set to SQLServer:
					  ○ Chinese_PRC_CI_AS
					  ○ Chinese_PRC_CS_AS
					  ○ SQL_Latin1_General_CP1_CI_AS
					  ○ SQL_Latin1_General_CP1_CS_AS
					  ○ Chinese_PRC_BIN

DBName	String	Yes	No	The name of the database.	The name must be unique. The name can be up to 64 characters in length and can contain letters, digits, and underscores (_). It must start with a letter.

Property	Type	Required	Editable	Description	Constraint
DBDescription	String	No	No	The description of the database.	• The description must be 2 to 256 characters in length and can contain letters, digits, underscores (_), and hyphens (-). • It must start with a letter and cannot start with `http://` or `https://`.

Response parameters

Fn::GetAtt

- DBInstanceId: the ID of the instance.
- InnerPort: the internal port of the instance.
- InnerIPAddress: the internal IP address of the instance.
- InnerConnectionString: the internal endpoint of the instance.
- PublicPort: the public port of the instance.
- PublicConnectionString: the public endpoint of the instance.
- PublicIPAddress: the public IP address of the instance.

Examples

The following example demonstrates how to create an ApsaraDB RDS instance in the classic network:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Database": {
      "Type": "ALIYUN::RDS::DBInstance",
      "Properties": {
        "Engine": "MySQL",
        "EngineVersion": "5.6",
        "DBInstanceClass":"rds.mysql.t1.small",
        "DBInstanceStorage": 10,
        "DBInstanceNetType": "Internet",
        "SecurityIPList": "0.0.0.0/0",
        "MasterUsername": "A****",
        "DBMappings":[{
            "DBName": "hope",
            "CharacterSetName": "utf8"
        }]
      }
    }
  },
  "Outputs": {
    "DBInstanceId": {
      "Value": {"get_attr": ["DBInstanceId"]}
    },
    "PublicConnectionString": {
      "Value": {"get_attr": ["ConnectionString"]}
    },
    "PublicPort": {
      "Value": {"get_attr": ["Port"]}
    }
  }
}
```

The following example demonstrates how to create an ApsaraDB RDS instance in a VPC:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Database": {
      "Type": "ALIYUN::RDS::DBInstance",
      "Properties": {
        "Engine":"MySQL",
        "EngineVersion":"5.6",
        "DBInstanceClass":"rds.mys2.small",
        "DBInstanceStorage":"10",
        "DBInstanceNetType":"Intranet",
        "SecurityIPList": "0.0.0.0/0",
        "VSwitchId": "ttt",
        "VpcId": "myvp****"
      }
    }
  },
  "Outputs": {
    "DBInstanceId": {
      "Value": {"get_attr": ["DBInstanceId"]}
    },
    "InnerConnectionString": {
      "Value": {"get_attr": ["ConnectionString"]}
    },
    "InnerPort": {
      "Value": {"get_attr": ["Port"]}
    }
  }
}
```

5.1.6.5.4. ALIYUN::RDS::DBInstanceParameterGroup

ALIYUN::RDS::DBInstanceParameterGroup is used to modify parameters of an ApsaraDB RDS instance.

Syntax

```
{
  "Type": "ALIYUN::RDS::DBInstanceParameterGroup",
  "Properties": {
    "Forcerestart": String,
    "DBInstanceId": String,
    "Parameters": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
DBInstanceId	String	Yes	No	The ID of the ApsaraDB RDS instance.	None

Property	Type	Required	Editable	Description	Constraint
Parameters	List	Yes	No	The list of one or more parameters of the instance.	The parameters and their values must be arranged in the JSON format. The parameter values must be of the string type. Example: {"auto_increment_increment":"1","character_set_client":"utf8"}.
Forcerestart	String	No	No	Specifies whether to forcibly restart the instance.	Default value: false. Valid values: • true: The system forcibly restarts the instance. • false: The system does not forcibly restart the instance.

Response parameters

Fn::GetAtt

None

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Database": {
      "Type": "ALIYUN::RDS::DBInstance",
      "Properties": {
        "Engine": "MySQL",
        "EngineVersion": "5.6",
        "DBInstanceClass": "rds.mys2.small",
        "DBInstanceStorage": "10",
        "DBInstanceNetType": "Intranet",
        "SecurityIPList": "0.0.0.0/0"
      }
    },
    "DatabaseConfig": {
      "Type": "ALIYUN::RDS::DBInstanceParameterGroup",
      "Properties": {
        "DBInstanceId": {
          "Ref": "Database"
        },
        "Parameters": [
          {
            "Key": "auto_increment_increment",
            "Value": "xxx"
          }
        ]
      }
    }
  },
  "Outputs": {
    "DBInstanceId": {
      "Value": {
        "Fn::GetAtt": [
          "Database",
          "DBInstanceId"
        ]
      }
    }
  }
}
```

5.1.6.5.5. ALIYUN::RDS::DBInstanceSecurityIps

ALIYUN::RDS::DBInstanceSecurityIps is used to modify the instance whitelist.

Statement

```
{
  "Type": "ALIYUN::RDS::DBInstanceSecurityIps",
  "Properties": {
    "DBInstanceId": String,
    "DBInstanceIPArrayName": String,
    "DBInstanceIPArrayAttribute": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
DBInstanceId	String	No	No	The ID of the RDS instance.	None
DBInstanceIPArrayAttribute	String	No	Yes	The attribute of the IP address whitelist.	The console does not display groups labeled with hidden.
DBInstanceIPArrayName	String	Yes	Released	The name of the IP address whitelist.	The name can contain only lowercase letters and underscores (_). Default value: Default.

Response parameters

Fn::GetAtt

SecurityIps: the IP address whitelist after the modification.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "DBInstanceSecurityIps": {
      "Type": "ALIYUN::RDS::DBInstanceSecurityIps",
      "Properties": {
        "DBInstanceIPArrayName": {
          "Ref": "DBInstanceIPArrayName"
        },
        "DBInstanceId": [
          "Ref": "DBInstanceId"
        },
        "DBInstanceIPArrayAttribute": {
          "Ref": "DBInstanceIPArrayAttribute"
        }
      }
    }
  },
  "Parameters": {
    "DBInstanceIPArrayName": {
      "Type": "String",
      "Description": "Group name of the security ips, only support lower characters and '_'. Advice use a new group name avoid effect your database system. If the properties is not specified, it will set to default group, please be careful."
    },
    "DBInstanceId": [
      "Type": "String",
      "Description": "Database instance id to update security ips."
    },
    "DBInstanceIPArrayAttribute": {
      "Type": "String",
      "Description": "Security ips to add or remove."
    }
  },
  "Outputs": {
    "SecurityIps": {
      "Description": "The security ips of selected database instance.",
      "Value": {
        "Fn::GetAtt": [
          "DBInstanceSecurityIps",
          "SecurityIps"
        ]
      }
    }
  }
}
```

5.1.6.5.6. ALIYUN::RDS::PrepayDBInstance

ALIYUN::RDS::PrepayDBInstance is used to create a subscription ApsaraDB RDS instance.

Syntax

```
{
  "Type": "ALIYUN::RDS::PrepayDBInstance",
  "Properties": {
    "DBMappings": List,
    "CouponCode": String,
    "MasterUsername": String,
    "PeriodType": String,
    "PayType": String,
    "DBInstanceNetType": String,
    "MasterUserType": String,
    "AutoRenew": Boolean,
    "PreferredBackupTime": String,
    "PrivateIpAddress": String,
    "Engine": String,
    "MultiAZ": Boolean,
    "VpcId": String,
    "ConnectionMode": String,
    "ResourceGroupId": String,
    "VSwitchId": String,
    "BackupRetentionPeriod": Number,
    "Quantity": Number,
    "CommodityCode": String,
    "ZoneId": String,
    "AutoPay": Boolean,
    "EngineVersion": String,
    "DBInstanceClass": String,
    "PreferredBackupPeriod": List,
    "DBInstanceStorage": Integer,
    "DBInstanceDescription": String,
    "Tags": Map,
    "Period": Number,
    "MasterUserPassword": String,
    "AllocatePublicConnection": Boolean
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	No	No	The ID of the resource group.	None
DBMappings	List	No	No	The list of one or more databases to be created in the instance.	None
CouponCode	String	No	No	None	None
MasterUsername	String	No	No	The name of the database account.	The name must be unique. The name can be up to 16 characters in length and can contain lowercase letters, digits, and underscores (_). It must start with a lowercase letter.

Property	Type	Required	Editable	Description	Constraint
PeriodType	String	Yes	No	The unit of the subscription period.	Default value: Month. Valid values: • Year • Month
DBInstanceNetType	String	No	No	The network type of the instance.	Default value: Intranet. Valid values: • Internet • Intranet
MasterUserType	String	No	No	The type of the database account.	Valid values: • Normal • Master
PreferredBackupTime	String	No	No	The backup window.	Specify the window in the HH:mmZ-HH:mmZ format. Value values: 00:00Z-01:00Z, 01:00Z-02:00Z, 02:00Z-03:00Z, 03:00Z-04:00Z, 04:00Z-05:00Z, 05:00Z-06:00Z, 06:00Z-07:00Z, 07:00Z-08:00Z, 08:00Z-09:00Z, 09:00Z-10:00Z, 10:00Z-11:00Z, 11:00Z-12:00Z, 12:00Z-13:00Z, 13:00Z-14:00Z, 14:00Z-15:00Z, 15:00Z-16:00Z, 16:00Z-17:00Z, 17:00Z-18:00Z, 18:00Z-19:00Z, 19:00Z-20:00Z, 20:00Z-21:00Z, 21:00Z-22:00Z, 22:00Z-23:00Z, and 23:00Z-24:00Z.
PrivateIpAddress	String	No	No	The private IP address with the CIDR block of the specified vSwitch.	If you do not specify this parameter, the system allocates a private IP address.
Engine	String	Yes	No	The database engine that the instance runs.	Valid values: • MySQL • SQLServer • PostgreSQL • PPAS

Property	Type	Required	Editable	Description	Constraint
MultiAZ	Boolean	No	No	Specifies whether the instance can be deployed across multiple zones.	None
VpcId	String	No	No	The ID of the VPC.	None
ConnectionMode	String	No	No	The connection mode of the instance.	Default value: Safty. Valid values: • Performance: the standard mode. • Safty: the database proxy mode. If you do not specify this parameter, the system assigns a connection mode.
AutoRenew	Boolean	No	No	Specifies whether to enable automatic renewal for the instance.	Valid values: • True • False
VSwitchId	String	No	No	The ID of the vSwitch in the specified VPC.	None
BackupRetentionPeriod	Number	No	No	The number of days for which backup files can be retained.	None
Quantity	Number	No	No	The number of instances to be created.	Valid values: 1 to 99. Default value: 1.
CommodityCode	String	Yes	No	The commodity code.	Valid values: • rds • bards • rords
ZoneId	String	No	No	The zone ID of the instance.	None
EngineVersion	String	Yes	No	The version of the database engine.	• Valid values when Engine is set to MySQL: 5.5 and 5.6. • Set the value to 2008r2 when Engine is set to SQLServer. • Set the value to 9.4 when Engine is set to PostgreSQL. • Set the value to 9.3 when Engine is set to PPAS.

Property	Type	Required	Editable	Description	Constraint
DBInstanceClass	String	Yes	Yes	The instance type.	Examples: rds.mys2.large, rds.mss1.large, and rds.pg.s1.small.
PreferredBackupPeriod	List	No	No	The backup cycle.	Valid values: • Monday • Tuesday • Wednesday • Thursday • Friday • Saturday • Sunday
DBInstanceStorage	Integer	Yes	Yes	The storage capacity of the instance.	• Valid values when Engine is set to MySQL: 5 to 1000. • Valid values when Engine is set to SQLServer: 10 to 1000. • Valid values when Engine is set to PostgreSQL or PPAS: 5 to 2000. Unit: GB. This value must be in 5 GB increments.
DBInstanceDescription	String	No	No	The description of the instance.	The description must be 2 to 256 characters in length and can contain letters, digits, underscores (_), and hyphens (-). It must start with a letter and cannot start with `http://` or `https://`.
Tags	map	No	Yes	The tags of the instance.	None
Period	Number	Yes	No	The subscription period of the instance.	• Valid values when PeriodType is set to Month: 1, 2, 3, 4, 5, 6, 7, 8, and 9. • Valid values when PeriodType is set to Year: 1, 2, and 3.
MasterUserPassword	String	No	No	The password of the database account.	The password must be 6 to 32 characters in length and can contain letters, digits, and underscores (_).

Property	Type	Required	Editable	Description	Constraint
AllocatePublicConnection	Boolean	No	No	Specifies whether to apply for a public endpoint for the instance.	None
PayType	String	No	No	The billing method of the instance.	Valid values: • Postpaid: pay-as-you-go • Prepaid: subscription
AutoPay	Boolean	No	No	Specifies whether to enable automatic payment for the instance.	Default value: False. Valid values: • True • False

DBMappings syntax

```
"DBMappings": [
  {
    "DBDescription": String,
    "CharacterSetName": String,
    "DBName": String
  }
]
```

DBMappings properties

Property	Type	Required	Editable	Description	Constraint
DBDescription	String	No	No	The description of the database.	The description must be 2 to 256 characters in length and can contain letters, digits, underscores (_), and hyphens (-). It must start with a letter and cannot start with `http://` or `https://`.
CharacterSetName	String	Yes	No	The character set.	• Valid values when Engine is set to MySQL: utf8, gbk, latin1, and utf8mb4 (applicable to versions 5.5 and 5.6). • Valid values when Engine is set to SQLServer: Chinese_PRC_CI_AS, Chinese_PRC_CS_AS, SQL_Latin1_General_CP1_CI_AS, SQL_Latin1_General_CP1_CS_AS, and Chinese_PRC_BIN.

Property	Type	Required	Editable	Description	Constraint
DBName	String	Yes	No	The name of the database.	The name must be unique. It can be up to 64 characters in length and can contain lowercase letters, digits, and underscores (_). It must start with a lowercase letter.

Response parameters

Fn::GetAtt

- InnerPort: the internal port of the instance.
- OrderId: the order ID of the instance.
- PublicConnectionString: the public endpoint of the instance.
- InnerIPAddress: the internal IP address of the instance.
- DBInstanceId: the ID of the instance.
- PublicIPAddress: the public IP address of the instance.
- PublicPort: the public port of the instance.
- InnerConnectionString: the internal endpoint of the instance.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "PeriodType": {
      "Type": "String",
      "Description": "Charge period for created instances.",
      "AllowedValues": [
        "Month",
        "Year"
      ],
      "Default": "Month"
    },
    "PrivateIpAddress": {
      "Type": "String",
      "Description": "The private ip for created instance."
    },
    "DBInstanceNetType": {
      "Type": "String",
      "Description": "Database instance net type, default is Intranet.Internet for public access, Intranet for private access.",
      "AllowedValues": [
        "Internet",
        "Intranet"
      ],
      "Default": "Intranet"
    },
    "AutoRenew": {
      "Type": "Boolean",
      "Description": "Auto renew the prepay instance. If the period type is by year, it will renew by year, else it will renew by month.",
      "AllowedValues": [
        "True",
        "true",
        "False",
```

```
        "false"
      ],
      "Default": false
    },
    "PreferredBackupPeriod": {
      "Type": "CommaDelimitedList",
      "Description": "Automate backups cycle if automated backups are enabled.",
      "AllowedValues": [
        "Monday",
        "Tuesday",
        "Wednesday",
        "Thursday",
        "Friday",
        "Saturday",
        "Sunday"
      ]
    },
    "DBInstanceStorage": {
      "Type": "Number",
      "Description": "Database instance storage size. mysql is [5,1000]. sql server 2008r2 is [10,1000
], sql server 2012/2012_web/2016-web is [20,1000]. PostgreSQL and PPAS is [5,2000]. Increased every 5
GB, Unit in GB"
    },
    "CommodityCode": {
      "Type": "String",
      "Description": "The CommodityCode of the order.",
      "AllowedValues": [
        "rds",
        "bards",
        "rords"
      ],
      "Default": "rds"
    },
    "DBMappings": {
      "Type": "CommaDelimitedList",
      "Description": "Database mappings to attach to db instance."
    },
    "MultiAZ": {
      "Type": "Boolean",
      "Description": "Specifies if the database instance is a multiple Availability Zone deployment. "
,
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ],
      "Default": false
    },
    "Engine": {
      "Type": "String",
      "Description": "Database instance engine type. Support MySQL/SQLServer/PostgreSQL/PPAS now.",
      "AllowedValues": [
        "MySQL",
        "SQLServer",
        "PostgreSQL",
        "PPAS"
      ]
    },
    "DBInstanceDescription": {
```

```
      "Type": "String",
      "Description": "Description of created database instance."
    },
    "Tags": {
      "Type": "Json",
      "Description": "The tags of an instance.\nYou should input the information of the tag with the format of the Key-Value, such as {\"key1\":\"value1\",\"key2\":\"value2\", ... \"key5\":\"value5\"}.\nAt most 5 tags can be specified.\nKey\nIt can be up to 64 characters in length.\nCannot begin with aliyun.\nCannot begin with http:// or https://.\nCannot be a null string.\nValue\nIt can be up to 128 characters in length.\nCannot begin with aliyun.\nCannot begin with http:// or https://.\nCan be a null string."
    },
    "EngineVersion": {
      "Type": "String",
      "Description": "Database instance version of the relative engine type.Support MySQL: 5.5/5.6/5.7; SQLServer: 2008r2, 20012, 2012_web, 2012_std_ha, 2012_ent_ha, 2016_web, 2016_std_ha, 2016_ent_ha; PostgreSQL:9.4; PPAS: 9.3.",
      "AllowedValues": [
        "5.5",
        "5.6",
        "5.7",
        "2008r2",
        "2012",
        "2012_web",
        "2012_std_ha",
        "2012_ent_ha",
        "2016_web",
        "2016_std_ha",
        "2016_ent_ha",
        "9.4",
        "9.3"
      ]
    },
    "ZoneId": {
      "Type": "String",
      "Description": "selected zone to create database instance. You cannot set the ZoneId parameter if the MultiAZ parameter is set to true."
    },
    "DBInstanceClass": {
      "Type": "String",
      "Description": "Database instance type. Refer the RDS database instance type reference, such as 'rds.mys2.large', 'rds.mss1.large', 'rds.pg.s1.small' etc"
    },
    "AllocatePublicConnection": {
      "Type": "Boolean",
      "Description": "If true, allocate public connection automate.",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    },
    "PreferredBackupTime": {
      "Type": "String",
      "Description": "The daily time range during which automated backups are created if automated backups are enabled.",
      "AllowedValues": [
        "00:00Z-01:00Z",
        "01:00Z-02:00Z",
```

```
        "02:00Z-03:00Z",
        "03:00Z-04:00Z",
        "04:00Z-05:00Z",
        "05:00Z-06:00Z",
        "06:00Z-07:00Z",
        "07:00Z-08:00Z",
        "08:00Z-09:00Z",
        "09:00Z-10:00Z",
        "10:00Z-11:00Z",
        "11:00Z-12:00Z",
        "12:00Z-13:00Z",
        "13:00Z-14:00Z",
        "14:00Z-15:00Z",
        "15:00Z-16:00Z",
        "16:00Z-17:00Z",
        "17:00Z-18:00Z",
        "18:00Z-19:00Z",
        "19:00Z-20:00Z",
        "20:00Z-21:00Z",
        "21:00Z-22:00Z",
        "22:00Z-23:00Z",
        "23:00Z-24:00Z"
      ]
    },
    "VSwitchId": {
      "Type": "String",
      "Description": "The vSwitch id of created instance. For VPC network, the property is required."
    },
    "Quantity": {
      "Type": "Number",
      "Description": "The number of instance to be created, default is 1, max number is 99",
      "MinValue": 1,
      "MaxValue": 99,
      "Default": 1
    },
    "Period": {
      "Type": "Number",
      "Description": "Prepaid time period. While choose by pay by month, it could be from 1 to 9. Whil
e choose pay by year, it could be from 1 to 3.",
      "MinValue": 1,
      "MaxValue": 9,
      "Default": 1
    },
    "MasterUserPassword": {
      "Type": "String",
      "Description": "The master password for the database instance. ",
      "MinLength": 8,
      "MaxLength": 32
    },
    "CouponCode": {
      "Type": "String",
      "Description": "The coupon code of the order."
    },
    "MasterUserType": {
      "Type": "String",
      "Description": "Privilege type of account.\n Normal: Common privilege. \n Super: High privilege.
And the default value is Normal.This parameter is valid for MySQL 5.5/5.6 only. MySQL 5.7, SQL Server
2012/2016, PostgreSQL, and PPAS each can have only one initial account. \nOther accounts are created b
y the initial account that has logged on to the database.",
      "AllowedValues": [
```

```
        "Normal",
        "Super"
      ],
      "Default": "Normal"
    },
    "VpcId": {
      "Type": "String",
      "Description": "The VPC id of created database instance. For VPC network, the property is requir
ed."
    },
    "MasterUsername": {
      "Type": "String",
      "Description": "The master user name for the database instance. "
    },
    "ConnectionMode": {
      "Type": "String",
      "Description": "Connection Mode for database instance,support 'Performance' and 'Safty' mode. De
fault is RDS system assigns. ",
      "AllowedValues": [
        "Performance",
        "Safty"
      ]
    },
    "BackupRetentionPeriod": {
      "Type": "Number",
      "Description": "The number of days for which automatic DB backups are retained.",
      "MinValue": 7,
      "MaxValue": 30,
      "Default": 7
    }
  },
  "Resources": {
    "PrepayDBInstance": {
      "Type": "ALIYUN::RDS::PrepayDBInstance",
      "Properties": {
        "PeriodType": {
          "Ref": "PeriodType"
        },
        "PrivateIpAddress": {
          "Ref": "PrivateIpAddress"
        },
        "DBInstanceNetType": {
          "Ref": "DBInstanceNetType"
        },
        "AutoRenew": {
          "Ref": "AutoRenew"
        },
        "PreferredBackupPeriod": {
          "Fn::Split": [
            ",",
            {
              "Ref": "PreferredBackupPeriod"
            },
            {
              "Ref": "PreferredBackupPeriod"
            }
          ]
        },
        "DBInstanceStorage": {
          "Ref": "DBInstanceStorage"
```

```
        "Ref": "DBInstanceStorage"
      },
      "CommodityCode": {
        "Ref": "CommodityCode"
      },
      "DBMappings": {
        "Fn::Split": [
          ",",
          {
            "Ref": "DBMappings"
          },
          {
            "Ref": "DBMappings"
          }
        ]
      },
      "MultiAZ": {
        "Ref": "MultiAZ"
      },
      "Engine": {
        "Ref": "Engine"
      },
      "DBInstanceDescription": {
        "Ref": "DBInstanceDescription"
      },
      "Tags": {
        "Ref": "Tags"
      },
      "EngineVersion": {
        "Ref": "EngineVersion"
      },
      "ZoneId": {
        "Ref": "ZoneId"
      },
      "DBInstanceClass": {
        "Ref": "DBInstanceClass"
      },
      "AllocatePublicConnection": {
        "Ref": "AllocatePublicConnection"
      },
      "PreferredBackupTime": {
        "Ref": "PreferredBackupTime"
      },
      "VSwitchId": {
        "Ref": "VSwitchId"
      },
      "Quantity": {
        "Ref": "Quantity"
      },
      "Period": {
        "Ref": "Period"
      },
      "MasterUserPassword": {
        "Ref": "MasterUserPassword"
      },
      "CouponCode": {
        "Ref": "CouponCode"
      },
      "MasterUserType": {
        "Ref": "MasterUserType"
      }
```

```
        },
        "VpcId": {
          "Ref": "VpcId"
        },
        "MasterUsername": {
          "Ref": "MasterUsername"
        },
        "ConnectionMode": {
          "Ref": "ConnectionMode"
        },
        "BackupRetentionPeriod": {
          "Ref": "BackupRetentionPeriod"
        }
      }
    }
  },
  "Outputs": {
    "InnerConnectionString": {
      "Description": "DB instance connection url by Intranet.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "InnerConnectionString"
        ]
      }
    },
    "DBInstanceId": {
      "Description": "The instance id of created database instance.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "DBInstanceId"
        ]
      }
    },
    "InnerIPAddress": {
      "Description": "IP Address for created DB instance of Intranet.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "InnerIPAddress"
        ]
      }
    },
    "PublicConnectionString": {
      "Description": "DB instance connection url by Internet.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "PublicConnectionString"
        ]
      }
    },
    "PublicIPAddress": {
      "Description": "IP Address for created DB instance of Internet.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "PublicIPAddress"
        ]
      }
```

```
      }
    },
    "OrderId": {
      "Description": "The order id list of created instance.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "OrderId"
        ]
      }
    },
    "PublicPort": {
      "Description": "Internet port of created DB instance.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "PublicPort"
        ]
      }
    },
    "InnerPort": {
      "Description": "Intranet port of created DB instance.",
      "Value": {
        "Fn::GetAtt": [
          "PrepayDBInstance",
          "InnerPort"
        ]
      }
    }
  }
}
```

5.1.6.6. ROS

5.1.6.6.1. ALIYUN::ROS::WaitCondition

ALIYUN::ROS::WaitCondition is used to create an instance to process UserData messages.

Statement

```
{
  "Type": "ALIYUN::ROS::WaitCondition",
  "Properties": {
    "Count": Number,
    "Handle": String,
    "Timeout": Number
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint

Parameter	Type	Required	Editable	Description	Constraint
Handle	String	No	No	Reference ALIYUN::ROS::WaitConditionHandle.	None
Timeout	Number	Yes	No	The length of time to wait for UserData messages.	Valid values: 1 to 43200. Unit: seconds.
Count	Number	No.	True	The total number of messages to be received.	None

Response parameters

Fn::GetAtt

- Data: A JSON-serialized dictionary that contains the signal Data after the most recent stack creation or update.
- LastData: a JSON-serialized dictionary that contains the signal data before the most recent stack update.
- JoinedErrorData: a string consisting of the ErrorData signal data.
- JoinedLastErrorData: a string consisting of the LastErrorData signal data.
- ErrorData: a JSON-serialized dictionary that contains the error signal data after the most recent stack creation or update.
- Lasterprotodata: a JSON-serialized dictionary that contains the error signal data before the most recent stack update.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "WaitCondition": {
      "Type": "ALIYUN::ROS::WaitCondition",
      "Properties": {
        "Handle": {
          "Ref": "WaitConHandle"
        },
        "Timeout": 5,
        "Count": 2
      }
    },
    "WaitConHandle": {
      "Type": "ALIYUN::ROS::WaitConditionHandle"
    }
  },
  "Outputs": {
    "CurlCli": {
      "Value": {
        "Fn::GetAtt": [
          "WaitConHandle",
          "CurlCli"
        ]
      }
    },
    "Data": {
      "Value": {
        "Fn::GetAtt": [
          "WaitCondition",
          "Data"
        ]
      }
    }
  }
}
```

5.1.6.6.2. ALIYUN::ROS::WaitConditionHandle

ALIYUN::ROS::WaitConditionHandle is used to create an instance that sends and receives messages during UserData execution.

Statement

```
{
  "Type": "ALIYUN::ROS::WaitConditionHandle",
  "Properties": {
    "Count": Integer,
    "Mode": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
Count	Integer	No.	True	The total number of messages to be received.	Default value: -1.
Mode	String	Yes	True	If you set this parameter to Increment, all previous signals will be updated before they are deleted. If you set this parameter to Full, no previous signals will be deleted unless the Count parameter is specified.	Valid values: • Increment • Full Default value: Full.

Response parameters

Fn::GetAtt

- CurlCli: A curl Command is generated by the resource. You can use the command to send the UserData execution result or status to Resource Orchestration Service.
- WindowsCurlCli: provides Windows with cURL CLI command prefixes and sends a message indicating that the execution is completed or failed. Windows does not support the curl command. Therefore, you must install curl.exe and add it to PATH. You can add `--data-binary "{\"status\": \" success \"}` to indicate success, or by adding `--data-binary "{\"status\": \" failure \"}` to indicate failure.
- PowerShellCurlCli: provides PowerShell with cURL CLI command prefixes and sends a message indicating that the execution is completed or failed. Because this cmdlet was introduced in PowerShell 3.0, make sure that the PowerShell version meets this constraint. By `$PSVersionTable.PSVersion` displays the version. You can add `-Body '{"status": "success "}` to indicate success, or by adding `-Body '{"status": "failure "}` to indicate failure.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Mode": {
      "Type": "String",
      "Description": "If set to Increment, all old signals will be deleted before update. In this mode
, WaitCondition.Count should reference an incremental value instead of a full value, such as ScalingGr
oupEnable.ScalingRuleArisExecuteResultNumberOfAddedInstances.\n\nIf set to Full, no old signal will be
deleted unless Count is set. In this mode, WaitCondition.Count should reference a full value, such as
the same value with InstanceGroup.MaxAmount. It is recommended to use this mode with Count.\n\nDefault
to Full.",
      "AllowedValues": [
        "Increment",
        "Full"
      ],
      "Default": "Full"
    },
    "Count": {
```

```
      "Type": "Number",
      "Description": "There are 3 preconditions that make Count taking effect:\n1.Mode is set to Full.
\n2.Count >= 0.\n3.The id of signal is not specified. If so, it will be a self-increasing integer star
ted from 1. For example, the id of the first signal is 1, the id of the second signal is 2, and so on.
\n\nIf Count takes effect, signals with id > Count will be deleted before update.\nThe default value i
s -1, which means no effect.\nIt is recommended to quote the same value with WaitCondition.Count.",
      "Default": -1
    }
  },
  "Resources": {
    "WaitConditionHandle": {
      "Type": "ALIYUN::ROS::WaitConditionHandle",
      "Properties": {
        "Mode": {
          "Ref": "Mode"
        },
        "Count": {
          "Ref": "Count"
        }
      }
    }
  },
  "Outputs": {
    "CurlCli": {
      "Description": "Convenience attribute, provides curl CLI command prefix, which can be used for s
ignalling handle completion or failure.  You can signal success by adding --data-binary '{\"status\":
\"SUCCESS\"}' , or signal failure by adding --data-binary '{\"status\": \"FAILURE\"}'",
      "Value": {
        "Fn::GetAtt": [
          "WaitConditionHandle",
          "CurlCli"
        ]
      }
    },
    "WindowsCurlCli": {
      "Description": "Convenience attribute, provides curl CLI command prefix for Windows, which can b
e used for signalling handle completion or failure. As Windows does not support curl command, you need
to install curl.exe and add it to PATH first. You can signal success by adding --data-binary \"{\\\"st
atus\\\": \\\"SUCCESS\\\"}\" , or signal failure by adding --data-binary \"{\\\"status\\\": \\\"FAILUR
E\\\"}\" ",
      "Value": {
        "Fn::GetAtt": [
          "WaitConditionHandle",
          "WindowsCurlCli"
        ]
      }
    },
    "PowerShellCurlCli": {
      "Description": "Convenience attribute, provides curl CLI command prefix for PowerShell, which ca
n be used for signalling handle completion or failure. As this cmdlet was introduced in PowerShell 3.0
, ensure the version of PowerShell satisfies the constraint. (Show the version via $PSVersionTable.PSV
ersion.) You can signal success by adding -Body '{\"status\": \"SUCCESS\"}' , or signal failure by add
ing -Body '{\"status\": \"FAILURE\"}' ",
      "Value": {
        "Fn::GetAtt": [
          "WaitConditionHandle",
          "PowerShellCurlCli"
        ]
      }
    }
```

```
    }
}
```

5.1.6.6.3. ALIYUN::ROS::Stack

ALIYUN::ROS::Stack is used to create a nested stack. You can have a maximum of five nested levels.

ALIYUN::ROS::Stack is used in a top-level template to nest stacks as resources.

You can add output values from a nested stack contained within the template. You can use Fn::GetAtt together with the logical name of the nested stack and the output name in the Outputs.NestedStackOutputName format.

> **Note** We recommend that you run an update to the Nested stack from the parent stack.

When you apply a template change to update a top-level stack, ROS updates the top-level stack and initiates an update to its nested stacks. Resource orchestration service (ROS) updates resources that have been modified in the nested stack, but does not update resources that have not been modified in the nested stack.

Statement

```
{
  "Type": "ALIYUN::ROS::Stack",
  "Properties": {
    "TemplateURL": String,
    "TimeoutMins": Number,
    "Parameters": Map
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint

Parameter	Type	Required	Editable	Description	Constraint
TemplateURL	String	No	Yes	The URL of the file containing the template body. The template file can be up to 524,288 bytes in size. The URL must point to a template located on the http or https Web server or Alibaba Cloud OSS bucket. For example: `oss://ros/template/demo` , `oss://ros/template/demo?RegionId=cn-hangzhou` . If the region of the OSS bucket is not specified, the RegionId of the stack is used.	The URL can be up to 1,024 bytes in length.
TimeoutMins	Number	No.	True	The length of time that ROS will wait for the nested stack to be created or updated.	Unit: minutes. Default value: 60.

Parameter	Type	Required	Editable	Description	Constraint
Parameters	Map	No.	True	A set of value pairs that represent the parameters passed to ROS when this Nested stack is created. Each parameter has a name corresponding to a parameter defined in the embedded template and the value to which you want to set the parameter. This parameter is required if the nested stack needs input parameters.	None

Response parameters

Fn::GetAtt

You can use the following code to obtain the output of the nested stack:

```
{
  "Fn::GetAtt": [
    "<nested_stack>",
    "Outputs.<nested_stack_output_name>"
  ]
}
```

When you use `Ref` to reference resources in a nested stack, the Alibaba Cloud Resource Name (ARN) of the nested stack is returned. Example: `arn:acs:ros::cn-hangzhou:12345****:stacks/test-nested-stack-Demo-jzkyq7mn2ykj/e71c1e04-1a57-46fc-b9a4-cf7ce0d3****`

Examples

- The following code provides an example of how to create a VPC, a VSwitch, and a security group in a nested stack and save the output results to the *oss://ros/template/vpc.txt* directory:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Description": "One VPC, VSwitch, security group.",
  "Parameters": {
    "ZoneId": {
      "Type": "String",
      "Description": "The available zone"
    },
    "SecurityGroupName": {
      "Type": "String",
      "Description": "The security group name",
      "Default": "my-sg-name"
```

```
    },
    "VpcName": {
      "Type": "String",
      "Description": "The VPC name",
      "MinLength": 2,
      "MaxLength": 128,
      "ConstraintDescription": "[2, 128] English or Chinese letters",
      "Default": "my-vpc-name"
    },
    "VpcCidrBlock": {
      "Type": "String",
      "AllowedValues": [
        "192.168.0.0/16",
        "172.16.0.0/12",
        "10.0.0.0/8"
      ],
      "Default": "10.0.0.0/8"
    },
    "VSwitchCidrBlock": {
      "Type": "String",
      "Description": "The VSwitch subnet which must be within VPC",
      "Default": "10.0.10.0/24"
    },
    "UpdateVersion": {
      "Type": "Number",
      "Default": 0
    }
  },
  "Resources": {
    "Vpc": {
      "Type": "ALIYUN::ECS::VPC",
      "Properties": {
        "CidrBlock": {
          "Ref": "VpcCidrBlock"
        },
        "VpcName": {
          "Ref": "VpcName"
        }
      }
    },
    "VSwitch": {
      "Type": "ALIYUN::ECS::VSwitch",
      "Properties": {
        "CidrBlock": {
          "Ref": "VSwitchCidrBlock"
        },
        "ZoneId": {
          "Ref": "ZoneId"
        },
        "VpcId": {
          "Fn::GetAtt": [
            "Vpc",
            "VpcId"
          ]
        }
      }
    },
    "SecurityGroup": {
      "Type": "ALIYUN::ECS::SecurityGroup",
      "Properties": {
```

```
        "SecurityGroupName": {
          "Ref": "SecurityGroupName"
        },
        "VpcId": {
          "Ref": "Vpc"
        }
      }
    },
    "WaitConditionHandle": {
      "Type": "ALIYUN::ROS::WaitConditionHandle",
      "Properties": {
        "UpdateVersion": {
          "Ref": "UpdateVersion"
        }
      }
    }
  },
  "Outputs": {
    "SecurityGroupId": {
      "Value": {
        "Fn::GetAtt": [
          "SecurityGroup",
          "SecurityGroupId"
        ]
      }
    },
    "VpcId": {
      "Value": {
        "Fn::GetAtt": [
          "Vpc",
          "VpcId"
        ]
      }
    },
    "VSwitchId": {
      "Value": {
        "Fn::GetAtt": [
          "VSwitch",
          "VSwitchId"
        ]
      }
    }
  }
}
```

- The following code provides an example of a top-level stack:

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Description": "One ECS instance.",
  "Parameters": {
    "ImageId": {
      "Default": "centos_7",
      "Type": "String",
      "Description": "Image Id, represents the image resource to startup the ECS instance"
    },
    "InstanceType": {
      "Type": "String",
      "Description": "The ECS instance type,",
      "Default": "ecs.xn4.small"
```

```
    },
    "ZoneId": {
      "Type": "String",
      "Description": "The available zone "
    },
    "InstanceChargeType": {
      "Type": "String",
      "AllowedValues": [
        "PrePaid",
        "PostPaid"
      ],
      "Default": "PostPaid",
      "Description": "The instance charge type"
    },
    "SecurityGroupName": {
      "Type": "String",
      "Description": "The security group name",
      "Default": "my-sg-name"
    },
    "NetworkInterfaceName": {
      "Type": "String",
      "Description": "The Network interface name",
      "Default": "my-eni-name"
    },
    "VpcName": {
      "Type": "String",
      "Description": "The VPC name",
      "MinLength": 2,
      "MaxLength": 128,
      "ConstraintDescription": "[2, 128] English or Chinese letters",
      "Default": "my-vpc-name"
    },
    "IoOptimized": {
      "AllowedValues": [
        "none",
        "optimized"
      ],
      "Description": "IO optimized, optimized is for the IO optimized instance type",
      "Type": "String",
      "Default": "optimized"
    },
    "SystemDiskCategory": {
      "AllowedValues": [
        "cloud",
        "cloud_efficiency",
        "cloud_ssd"
      ],
      "Description": "System disk category: average cloud disk(cloud), efficient cloud disk(cloud_e
fficiency) or SSD cloud disk(cloud_ssd)",
      "Type": "String",
      "Default": "cloud_ssd"
    },
    "VpcCidrBlock": {
      "Type": "String",
      "AllowedValues": [
        "192.168.0.0/16",
        "172.16.0.0/12",
        "10.0.0.0/8"
      ],
      "Default": "10.0.0.0/8"
```

```
    },
    "VSwitchCidrBlock": {
      "Type": "String",
      "Description": "The VSwitch subnet which must be within VPC",
      "Default": "10.0.10.0/24"
    },
    "UpdateVersion": {
      "Type": "Number",
      "Default": 0
    }
  },
  "Resources": {
    "NetworkStack": {
      "Type": "ALIYUN::ROS::Stack",
      "Properties": {
        "TemplateURL": "oss://ros/template/vpc.txt",
        "TimeoutMins": 5,
        "Parameters": {
          "ZoneId": {
            "Ref": "ZoneId"
          },
          "SecurityGroupName": {
            "Ref": "SecurityGroupName"
          },
          "VpcName": {
            "Ref": "VpcName"
          },
          "VpcCidrBlock": {
            "Ref": "VpcCidrBlock"
          },
          "VSwitchCidrBlock": {
            "Ref": "VSwitchCidrBlock"
          },
          "UpdateVersion": {
            "Ref": "UpdateVersion"
          }
        }
      }
    },
    "WebServer": {
      "Type": "ALIYUN::ECS::Instance",
      "Properties": {
        "ImageId": {
          "Ref": "ImageId"
        },
        "InstanceType": {
          "Ref": "InstanceType"
        },
        "InstanceChargeType": {
          "Ref": "InstanceChargeType"
        },
        "SecurityGroupId": {
          "Fn::GetAtt": [
            "NetworkStack",
            "Outputs.SecurityGroupId"
          ]
        },
        "VpcId": {
          "Fn::GetAtt": [
            "NetworkStack",
```

```
            "Outputs.VpcId"
          ]
        },
        "VSwitchId": {
          "Fn::GetAtt": [
            "NetworkStack",
            "Outputs.VSwitchId"
          ]
        },
        "IoOptimized": {
          "Ref": "IoOptimized"
        },
        "ZoneId": {
          "Ref": "ZoneId"
        },
        "SystemDisk_Category": {
          "Ref": "SystemDiskCategory"
        },
        "DiskMappings": [
          {
            "Category": "cloud_ssd",
            "Size": 20
          }
        ]
      }
    }
  },
  "Outputs": {
    "InstanceId": {
      "Value": {
        "Fn::GetAtt": [
          "WebServer",
          "InstanceId"
        ]
      }
    },
    "PublicIp": {
      "Value": {
        "Fn::GetAtt": [
          "WebServer",
          "PublicIp"
        ]
      }
    },
    "SecurityGroupId": {
      "Value": {
        "Fn::GetAtt": [
          "NetworkStack",
          "Outputs.SecurityGroupId"
        ]
      }
    },
    "VpcId": {
      "Value": {
        "Fn::GetAtt": [
          "NetworkStack",
          "Outputs.VpcId"
        ]
      }
    },
```

```
    "VSwitchId": {
      "Value": {
        "Fn::GetAtt": [
          "NetworkStack",
          "Outputs.VSwitchId"
        ]
      }
    },
    "NetworkStackArn": {
      "Value": {
        "Ref": "NetworkStack"
      }
    }
  }
}
```

5.1.6.7. SLB

5.1.6.7.1. ALIYUN::SLB::AccessControl

ALIYUN::SLB::AccessControl is used to create an access control list (ACL).

Syntax

```
{
  "Type": "ALIYUN::SLB::AccessControl",
  "Properties": {
    "AddressIPVersion": String,
    "AclName": String,
    "AclEntrys": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
AddressIPVersion	String	No	No	The Internet protocol version.	Valid values: ipv4 and ipv6.
AclName	String	Yes	Yes	The name of the ACL.	None
AclEntrys	List	No	No	The list of ACL entries.	A list can contain up to 50 ACL entries.

AclEntrys syntax

```
"AclEntrys": [
  {
    "comment": String,
    "entry": String
  }
]
```

AclEntrys properties

Property	Type	Required	Editable	Description	Constraint
comment	String	No	No	The comments on ACL entries.	None
entry	String	Yes	No	The authorized IP addresses or CIDR blocks.	None

Response parameters

Fn::GetAtt

AclId: the ID of the ACL.

Examples

Resource usage example

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "AccessControl": {
      "Type": "ALIYUN::SLB::AccessControl",
      "Properties": {
        "AddressIPVersion": {
          "Ref": "AddressIPVersion"
        },
        "AclName": {
          "Ref": "AclName"
        },
        "AclEntrys": {
          "Fn::Split": [",", {
            "Ref": "AclEntrys"
          }, {
            "Ref": "AclEntrys"
          }]
        }
      }
    }
  },
  "Parameters": {
    "AddressIPVersion": {
      "Type": "String",
      "Description": "IP version. Could be \"ipv4\" or \"ipv6\".",
      "AllowedValues": ["ipv4", "ipv6"]
    },
    "AclName": {
      "Type": "String",
      "Description": "The name of the access control list."
    },
    "AclEntrys": {
      "Type": "CommaDelimitedList",
      "Description": "A list of acl entrys. Each entry can be IP addresses or CIDR blocks. Max length:
50.",
      "MaxLength": 50
    }
  },
  "Outputs": {
    "AclId": {
      "Description": "The ID of the access control list.",
      "Value": {
        "Fn::GetAtt": ["AccessControl", "AclId"]
      }
    }
  }
}
```

Example of combined use of SLB-related resources

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "LoadBalancer": {
      "Type": "ALIYUN::SLB::LoadBalancer",
      "Properties": {
        "LoadBalancerName": "slb-with-listener-and-acl",
        "AddressType": "internet",
```

```
        "InternetChargeType": "paybybandwidth",
        "Bandwidth": 10,
        "VpcId": "vpc-xxxxxxxxxxxxxxxxxxx",
        "VSwitchId": "vsw-xxxxxxxxxxxxxxxxxxx"
      }
    },
    "ACL": {
      "Type": "ALIYUN::SLB::AccessControl",
      "Properties": {
        "AclName": "acl-for-listener",
        "AddressIPVersion": "ipv4",
        "AclEntrys": [
          {
            "entry": "192.168.x.x"
          },
          {
            "entry": "10.0.x.x/24",
            "comment": "just comment"
          }
        ]
      }
    },
    "CreateListener": {
      "Type": "ALIYUN::SLB::Listener",
      "Properties": {
        "LoadBalancerId": {
          "Ref": "LoadBalancer"
        },
        "ListenerPort": "80",
        "BackendServerPort": 8080,
        "Bandwidth": 1,
        "Protocol": "http",
        "HealthCheck": {
          "HealthyThreshold": 3,
          "UnhealthyThreshold": 3,
          "Interval": 2,
          "Timeout": 5
        },
        "Scheduler": "wrr",
        "RequestTimeout": 179,
        "IdleTimeout": 59,
        "AclId": {
          "Ref": "ACL"
        },
        "AclStatus": "on",
        "AclType": "white"
      }
    }
  },
  "Outputs": {
    "LoadBalanceDetails": {
      "Value": {
        "Fn::GetAtt": [
          "LoadBalancerId",
          "Listeners"
        ]
      }
    }
  }
}
```

5.1.6.7.2. ALIYUN::SLB::BackendServerAttachment

ALIYUN::SLB::BackendServerAttachment is used to add backend servers.

Statement

```
{
  "Type": "ALIYUN::SLB::BackendServerAttachment",
  "Properties": {
    "LoadBalancerId": String,
    "BackendServers": List,
    "BackendServerList": List,
    "BackendServerWeightList": List
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
LoadBalancerId	String	No	No	The unique ID of the SLB instance.	None
BackendServerList	List	No.	True	The list of backend servers to add.	You can call this operation with LoadBalancerId and BackendServerWeightList. Separate ECS instance IDs with commas (,). This parameter is ignored when the BackendServers parameter is specified.

Parameter	Type	Required	Editable	Description	Constraint
BackendServerWeightList	List	No.	True	The weights of the ECS instances in the BackendServerList, which are specified in order.	If this parameter is not specified, the weight of all ECS instances included in the BackendServerList is 100. When the BackendServerWeightList length is less than BackendServerList, the last value in the BackendServerWeightList is used to weight the remaining ECS instances in the BackendServerList.
BackendServers	List	No.	True	The list of backend servers to add.	Only backend servers in the running state can be attached to the SLB instance.

BackendServers syntax

```
"BackendServers": [
  {
    "ServerId" : String,
    "Weight" : Integer
  }
]
```

BackendServers properties

Parameter	Type	Required	Editable	Description	Constraint
ServerId	String	No	Yes	The ID of the ECS instance that acts as a backend server.	The ECS instance must be in the Running state.
Weight	Integer	Retained	Yes	The weight of the ECS instance in the SLB instance.	Valid values: 0 to 100. Default value: 100.

Response parameters

Fn::GetAtt

- BackendServers: the backend servers added to the SLB instance.

- LoadBalancerId: the ID of the SLB instance.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Attachment2": {
      "Type": "ALIYUN::SLB::BackendServerAttachment",
      "Properties": {
        "LoadBalancerId": "15187200816-cn-beijing-btc-****",
        "BackendServerList": [
          "i-25o0m****",
          "i-25zsk****"
        ],
        "BackendServerWeightList": [
          "20",
          "100"
        ]
      }
    }
  }
}
```

5.1.6.7.3. ALIYUN::SLB::BackendServerToVServerGroupAddition

ALIYUN::SLB::BackendServerToVServerGroupAddition is used to add backend servers to an existing VServer group.

Statement

```
{
  "Type": "ALIYUN::SLB::BackendServerToVServerGroupAddition",
  "Properties": {
    "BackendServers": List,
    "VServerGroupId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
VServerGroupId	String	No	No	The ID of the VServer group.	None
BackendServers	List	Retained	Yes	The list of ECS instances to be added.	None

BackendServers syntax

```
"BackendServers": [
  {
    "ServerId": String,
    "Port": Integer,
    "Weight": Integer
  }
]
```

BackendServers properties

Parameter	Type	Required	Editable	Description	Constraint
ServerId	String	No	Yes	The ID of the ECS instance that acts as a backend server.	None
Port	Integer	Retained	Yes	The ECS port number that is listened to in the server load balancer instance.	Valid values: 1 to 65535.
Weight	Integer	Retained	Yes	The weight of the ECS instance to be attached to the SLB instance.	Valid values: 0 to 100.

Response parameters

Fn::GetAtt

VServerGroupId: the ID of the VServer group.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "AttachVServerGroup": {
      "Type": "ALIYUN::SLB::BackendServerToVServerGroupAddition",
      "Properties": {
        "VServerGroupId": "sg-2zenh4ndwrqg14yt0****",
        "BackendServers": [
          {
            "ServerId": "i-25zsk****",
            "Weight": 20,
            "Port": 8080
          },
          {
            "ServerId": "i-25zsk****",
            "Weight": 100,
            "Port": 8081
          }
        ]
      }
    }
  }
}
```

5.1.6.7.4. ALIYUN::SLB::Certificate

ALIYUN::SLB::Certificate is used to upload a certificate to an SLB instance. Server certificates and CA certificates are supported.

> **Notice**
> - You can upload only one CA certificate at a time ("CertificateType": "CA ").
> - You can upload only one server certificate and the corresponding private key at a time ("CertificateType": "Server ").

Syntax

```
{
  "Type": "ALIYUN::SLB::Certificate",
  "Properties": {
    "CertificateName": String,
    "Certificate": String,
    "AliCloudCertificateName": String,
    "PrivateKey": String,
    "ResourceGroupId": String,
    "CertificateType": String,
    "AliCloudCertificateId": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint

Property	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	No	No	The ID of the resource group.	None
CertificateName	String	No	Yes	The name of the certificate.	None
Certificate	String	Yes	No	The public key of the certificate.	None
AliCloudCertificateName	String	No	No	The name of the Alibaba Cloud certificate.	None
PrivateKey	String	No	No	The server private key that you want to upload.	None
AliCloudCertificateId	String	No	No	The ID of the Alibaba Cloud certificate.	This parameter is required if you use a certificate from Alibaba Cloud SSL Certificates Service.
CertificateType	String	No	No	The type of the certificate.	Valid values: Server and CA.

Response parameters

Fn::GetAtt

- CertificateId: the ID of the certificate.
- Fingerprint: the fingerprint of the certificate.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "CertificateType": {
      "Type": "String",
      "Description": "The type of the certificate.",
      "AllowedValues": [
        "Server",
        "CA"
      ],
      "Default": "Server"
    },
    "AliCloudCertificateName": {
      "Type": "String",
      "Description": "The name of the Alibaba Cloud certificate."
    },
    "PrivateKey": {
      "Type": "String",
      "Description": "The private key."
    }
```

```
    },
    "CertificateName": {
      "Type": "String",
      "Description": "The name of the certificate."
    },
    "Certificate": {
      "Type": "String",
      "Description": "The content of the certificate public key."
    },
    "AliCloudCertificateId": {
      "Type": "String",
      "Description": "The ID of the Alibaba Cloud certificate."
    }
  },
  "Resources": {
    "Certificate": {
      "Type": "ALIYUN::SLB::Certificate",
      "Properties": {
        "CertificateType": {
          "Ref": "CertificateType"
        },
        "AliCloudCertificateName": {
          "Ref": "AliCloudCertificateName"
        },
        "PrivateKey": {
          "Ref": "PrivateKey"
        },
        "CertificateName": {
          "Ref": "CertificateName"
        },
        "Certificate": {
          "Ref": "Certificate"
        },
        "AliCloudCertificateId": {
          "Ref": "AliCloudCertificateId"
        }
      }
    }
  },
  "Outputs": {
    "Fingerprint": {
      "Description": "The fingerprint of the certificate.",
      "Value": {
        "Fn::GetAtt": [
          "Certificate",
          "Fingerprint"
        ]
      }
    },
    "CertificateId": {
      "Description": "The ID of the certificate.",
      "Value": {
        "Fn::GetAtt": [
          "Certificate",
          "CertificateId"
        ]
      }
    }
  }
}
```

5.1.6.7.5. ALIYUN::SLB::DomainExtension

ALIYUN::SLB::DomainExtension is used to create a domain extension for an SLB instance.

Statement

```
{
  "Type": "ALIYUN::SLB::DomainExtension",
  "Properties": {
    "Domain": String,
    "ListenerPort": Integer,
    "ServerCertificateId": String,
    "LoadBalancerId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
Domain	String	No	No	The custom domain name.	None
ListenerPort	Integer	Yes	No	The frontend port used by the HTTPS listener of the SLB instance.	Valid values: 1 to 65535.
ServerCertificateId	String	No	Yes	The ID of the certificate corresponding to the domain name.	None
LoadBalancerId	String	No	No	The ID of the SLB instance.	None

Response parameters

Fn::GetAtt

- DomainExtensionId: the ID of the created domain extension.
- ListenerPort: The frontend port used by the SLB instance.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "DomainExtension": {
      "Type": "ALIYUN::SLB::DomainExtension",
      "Properties": {
        "Domain": "*.example.com",
        "ListenerPort": "443",
        "ServerCertificateId": "123157908552****_166f8204689_1714763408_70998****",
        "LoadBalancerId": "lb-bp1o94dp5i6earr9g****"
      }
    }
  },
  "Outputs": {
    "DomainExtensionId": {
      "Value": {
        "Fn::GetAtt": [
          "DomainExtension",
          "DomainExtensionId"
        ]
      }
    },
    "ListenerPort": {
      "Value": {
        "Fn::GetAtt": [
          "DomainExtension",
          "ListenerPort"
        ]
      }
    }
  }
}
```

5.1.6.7.6. ALIYUN::SLB::Listener

ALIYUN::SLB::Listener is used to create a listener for an SLB instance.

Statement

```
{
  "Type": "ALIYUN::SLB::Listener",
  "Properties": {
    "MasterSlaveServerGroupId": String,
    "AclStatus": String,
    "Protocol": String,
    "AclId": String,
    "ServerCertificateId": String,
    "HealthCheck": Map,
    "RequestTimeout": Integer,
    "IdleTimeout": Integer,
    "ListenerPort": Integer,
    "HttpConfig": Map,
    "Bandwidth": Integer,
    "AclType": String,
    "BackendServerPort": Integer,
    "Scheduler": String,
    "LoadBalancerId": String,
    "CACertificateId": String,
    "Persistence": Map,
    "VServerGroupId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
MasterSlaveServerGroupId	String	Yes	Released	The ID of the active/standby server group.	None
AclStatus	String	Yes	Released	Specifies whether to enable access control on the listener.	Valid values: • on • off Default value: off.
AclId	String	Yes	Released	The ID of the access control list (ACL) to which the listener is bound. This parameter is required when the AclStatus parameter is set to on.	None
				The type of the ACL. Valid values: white and black. • white: specifies the ACL as a whitelist. Only	

Parameter	Type	Required	Editable	Description	Constraint
AclType	String	Yes	Released	requests from the IP addresses or CIDR blocks specified in the ACL are forwarded. Whitelists are applicable to scenarios where you want an application to only be accessed from specific IP addresses. Configuring a whitelist poses risks to your services. After a whitelist is configured, only the IP addresses specified in the whitelist are able to access the SLB listener. If a whitelist is enabled without any IP addresses specified, the SLB listener will not forward any requests. • black: specifies the ACL as a blacklist. Requests from the IP addresses or CIDR blocks specified in the ACL are not forwarded. Blacklists are applicable to scenarios where you want an application to only be denied access from specific IP addresses.	Valid values: • White • Black

Parameter	Type	Required	Editable	Description	Constraint
				If a blacklist is enabled without any IP addresses specified, the SLB listener will forward all requests. This parameter is required when the AclStatus parameter is set to on.	
Protocol	String	No	No	The Internet protocol over which the listener will forward requests.	Valid values: • http • https • tcp • udp
ListenerPort	Integer	Yes	No	The frontend port used by the SLB instance.	Valid values: 1 to 65535.

Parameter	Type	Required	Editable	Description	Constraint
Bandwidth	Integer	Yes	No	The peak bandwidth of the listener. Unit: Mbit/s.	• Valid values: -1 and 1 to 1000. • For an SLB instance that is connected to the Internet and billed by fixed bandwidth, this parameter cannot be set to -1, and the sum of peak bandwidth values assigned to different listeners cannot exceed the Bandwidth value specified when the SLB instance is created. For an SLB instance that is connected to the Internet and billed by traffic, this parameter can be set to -1. Unit: Mbit/s.
BackendServerPort	Integer	Yes	No	The backend port used by the SLB instance.	Valid values: 1 to 65535.
LoadBalancerId	String	No	No	The ID of the SLB instance.	None
HealthCheck	Map	Erased	Released	The health check settings of the listener.	None
Persistence	Map	Erased	Released	The persistence properties.	None

Parameter	Type	Required	Editable	Description	Constraint
Scheduler	String	Yes	Released	The algorithm used to direct traffic to individual servers.	Valid values: • wrr • wlc Default value: wrr
CACertificateId	String	Yes	Released	The ID of the CA certificate.	Only valid for. HTTPS
ServerCertificateId	String	Yes	Released	The ID of the server certificate.	This parameter is required and valid only for HTTPS listeners.
VServerGroupId	String	Yes	Released	The ID of the VServer group.	None
RequestTimeout	String	Optional	Released	The request timeout period. Unit: seconds.	Valid values: 1 to 180.
IdleTimeout	String	Optional	Released	The idle connection timeout period. Unit: seconds.	Valid values: 1 to 60.
HttpConfig	Map	Erased	Released	The HTTP configurations.	None

HealthCheck syntax

```
"HealthCheck": {
  "Domain": String,
  "Interval": Integer,
  "URI": String,
  "HttpCode": String,
  "HealthyThreshold": Integer,
  "Timeout": Integer,
  "UnhealthyThreshold": Integer,
  "Port": Integer
}
```

HealthCheck properties

Parameter	Type	Required	Editable	Description	Constraint

Parameter	Type	Required	Editable	Description	Constraint
Domain	String	Yes	Released	The domain name used for health checks.	• The value can be $_ip, a custom string, or an empty string. • A custom string must be 1 to 80 characters in length and can contain only letters, digits, hyphens (-), and periods (.). • When this parameter is set to $_ip or left empty, the SLB instance uses the private IP addresses of backend servers as the domain names for health checks.
Interval	String	Optional	Released	The time interval between consecutive health checks. Unit: seconds.	Valid values: 1 to 5. Unit: seconds.

Parameter	Type	Required	Editable	Description	Constraint
URI	String	Yes	Released	The URI used for health checks.	• The URI must be 1 to 80 characters in length • and can contain letters, digits, hyphens (-), forward slashes (/), periods (.), percent signs (%), question marks (?), number signs (#), and ampersands (&). It must start with a forward slash (/).
HttpCode	String	Yes	Released	The HTTP status code that indicates a positive health status of the backend servers.	• Valid values: http_2xx, http_3xx, http_4xx, and http_5xx. • Separate multiple HTTP status codes with commas (,). Default value: http_2xx
HealthyThreshold	String	Optional	Released	The threshold used to determine that the backend servers are healthy. This value indicates the number of consecutive successful health checks required before the health status of a backend server can be changed from fail to success.	Valid values: 1 to 10.

Parameter	Type	Required	Editable	Description	Constraint
Timeout	String	Optional	Released	The length of time to wait for the response from a health check. Unit: seconds.	Valid values: 1 to 50. **Notice** This parameter is valid only when its value is greater than or equal to that of the Interval parameter. Otherwise, this parameter will be overridden by the Interval value.
UnhealthyThreshold	String	Optional	Released	The threshold used to determine that the backend servers are unhealthy. This value indicates the number of consecutive failed health checks required before the health status of a backend server can be changed from success to fail.	Valid values: 1 to 10.
Port	String	Optional	Released	The port used for health checks.	Valid values: 0 to 65535.

Persistence syntax

```
"Persistence": {
  "PersistenceTimeout": Integer,
  "CookieTimeout": Integer,
  "XForwardedFor": String,
  "Cookie": String,
  "StickySession": String,
  "StickySessionType": String
}
```

Persistence properties

Parameter	Type	Required	Editable	Description	Constraint
StickySession	String	No	No	Specifies whether to enable session persistence.	Valid values: • on • off
PersistenceTimeout	String	Optional	Released	The maximum duration that the client can be connected to the server. Unit: seconds.	Valid values: 0 to 1000. The default value is 0, which indicates that connection persistence is disabled. Unit: seconds.
CookieTimeout	String	Optional	Released	The maximum duration the cookie can be retained before it expires. Unit: seconds.	This parameter is required when the StickySession parameter is set to on and the StickySessionType parameter is set to insert. Valid values: 1 to 86400. Unit: seconds.
XForwardedFor	String	Yes	Released	Specifies whether to use the X-Forwarded-For header field to obtain the real IP address of the client.	Valid values: • on • off Default value: on

Parameter	Type	Required	Editable	Description	Constraint
Cookie	String	Yes	Released	The cookie configured on the backend server.	• The parameter value must be 1 to 200 characters in length and follow the RFC 2965 standard. It can contain only ASCII characters. It cannot contain commas (,), semicolons (;), or spaces, and cannot start with a dollar sign ($). • This parameter is required when the StickySession parameter is set to on and the StickySession Type parameter is set to server.

Parameter	Type	Required	Editable	Description	Constraint
StickySessionType	String	Yes	Released	The method for processing cookies.	• Valid values: insert and server. • When this parameter is set to insert, SLB adds a cookie to the first response from the backend server. Then, the next request contains the cookie and the listener distributes the request to the same backend server. When this parameter is set to server, SLB overwrites the original cookie when a new cookie is set. The next time the client carries the new cookie to access SLB, the listener distributes the request to the previously recorded backend server. • This parameter is required when the StickySession parameter is set to on. This parameter is ignored when the StickySession parameter is set to off.

HttpConfig syntax

```
"HttpConfig": {
  "ForwardPort": Integer,
  "ListenerForward": String
}
```

HttpConfig properties

Parameter	Type	Required	Editable	Description	Constraint
ForwardPort	String	Optional	Released	The port used to redirect HTTP requests to HTTPS.	Valid values: 1 to 65535. Default value: 443.
ListenerForward	String	No	No	Specifies whether to enable HTTP-to-HTTPS redirection.	Valid values: • on • off Default value: off.

Response parameters

Fn::GetAtt

- LoadBalancerId: the unique ID of the SLB instance.
- ListenerPortsAndProtocol: an array consisting of the ports and protocols used by the SLB listener.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "LoadBalancer": {
      "Type": "ALIYUN::SLB::LoadBalancer",
      "Properties": {
        "LoadBalancerName": "createdByHeat",
        "AddressType": "internet",
        "InternetChargeType": "paybybandwidth"
      }
    },
    "CreateListener": {
      "Type": "ALIYUN::SLB::Listener",
      "Properties": {
        "LoadBalancerId": {"Ref": "LoadBalancer"},
        "ListenerPort": "8094",
        "BackendServerPort": 8080,
        "Bandwidth": 1,
        "Protocol": "http",
        "HealthCheck": {
          "HealthyThreshold": 3,
          "UnhealthyThreshold": 3,
          "Interval": 2,
          "Timeout": 5,
          "HttpCode": "http_2xx,http_3xx,http_4xx,http_5xx"
        },
        "Scheduler": "wrr",
          "Persistence": {
          "PersistenceTimeout": 1,
          "XForwardedFor": "on",
          "StickySession": "on",
          "StickySessionType": "insert",
          "CookieTimeout": 10,
          "Cookie": "1"
        }
      }
    }
  },
  "Outputs": {
  "LoadBalanceDetails": {
     "Value": {"Fn::GetAtt": ["LoadBalancer", "LoadBalancerId"]}
  }
  }
}
```

5.1.6.7.7. ALIYUN::SLB::LoadBalancer

ALIYUN::SLB::LoadBalancer is used to create an SLB instance.

Statement

```
{
  "Type": "ALIYUN::SLB::LoadBalancer",
  "Properties": {
    "DeletionProtection": Boolean,
    "AddressType": String,
    "Tags": List,
    "InternetChargeType": String,
    "Bandwidth": Integer,
    "SlaveZoneId": String,
    "ResourceGroupId": String,
    "AutoPay": Boolean,
    "VpcId": String,
    "PricingCycle": String,
    "LoadBalancerName": String,
    "Duration": Number,
    "VSwitchId": String,
    "LoadBalancerSpec": String,
    "MasterZoneId": String,
    "PayType": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the RDS instance belongs.	None
DeletionProtection	Boolean	Erased	Released	Specifies whether to enable deletion protection to prevent the SLB instance from being deleted by mistake.	Valid values: • true • false
VpcId	String	Yes	Released	The ID of the VPC.	None
SlaveZoneId	String	Yes	Released	The ID of the secondary zone to which the SLB instance belongs.	None

Parameter	Type	Required	Editable	Description	Constraint
Bandwidth	String	Optional	Released	The peak bandwidth of SLB instances that are connected to the Internet and billed by fixed bandwidth.	For SLB instances that are connected to the Internet and billed by fixed Bandwidth, this parameter is valid only when the Bandwidth parameter of the SLB Listener is specified. For Internet instances whose billing type is to pay by traffic, we recommend that you set the peak Bandwidth through the Listener parameter. In this case, this parameter is ignored. Valid values: 1 to 1000. Unit: Mbps. Default value: 1 VPC-type instances are billed by traffic.
AddressType	String	Yes	Released	The address type of the SLB instance.	Valid values: • internet • intranet Default value: internet.
VSwitchId	String	Yes	Released	The ID of the vSwitch in the VPC.	None

Parameter	Type	Required	Editable	Description	Constraint
LoadBalancerName	String	Yes	Released	The name of the SLB instance to be created.	A custom string. The name must be 1 to 80 characters in length and can contain letters, digits, hyphens (-), forward slashes (/), periods (.), and underscores (_). If this parameter is not specified, the system assigns a value.
InternetChargeType	String	Yes	Released	The billing method of SLB instances that are connected to the Internet.	Valid values: • paybybandwidth • paybytraffic Default value: paybytraffic.
MasterZoneId	String	Yes	Released	The ID of the primary zone to which the SLB instance belongs.	None
Tags	List	Erased	Released	The tags to be attached to the SLB instance. The tags are listed in JSON format. Each tag consists of a TagKey and a TagValue.	A maximum of five tags can be attached to an SLB instance.
LoadBalancerSpec	String	Yes	Released	The type of the SLB instance.	Valid values: • slb.s1.small • slb.s2.small • slb.s2.medium • slb.s3.small • slb.s3.medium • slb.s3.large The available types vary by region.

Parameter	Type	Required	Editable	Description	Constraint
AutoPay	Boolean	Erased	Released	Specifies whether to automatically pay for subscription SLB instances that are connected to the Internet.	Valid values: • true • false Default value: false. **Note** This parameter is valid only on the China site (aliyun.com).
PayType	String	Yes	Released	The billing method of the SLB instance.	Valid values: • PayOnDemand • PrePay
PricingCycle	String	Yes	Released	The billing cycle of subscription SLB instances that are connected to the Internet.	Valid values: • month • year **Note** This parameter is valid only on the China site (aliyun.com).

Parameter	Type	Required	Editable	Description	Constraint
Duration	Number	Erased	Released	The subscription period of subscription SLB instances that are connected to the Internet.	Valid values: • Valid values when the PricingCycle parameter is set to month: 1 to 9. • Valid values when the PricingCycle parameter is set to year: 1 to 3. **Note** This parameter is valid only on the China site (aliyun.com).

Tags syntax

```
"Tags": [
  {
    "Value": String ,
    "Key": String
  }
]
```

Tags properties

Property	Type	Required or Not	Editable	Description	Constraint
Key	String	No	No	None	None
Value	String	Yes	Released	None	Valid values: 1025 to 10000. You cannot use the following commonly used port numbers: 2222, 4500, 4510, 4560, 7505, 9000, 9001, and 9002.

Response parameters

Fn::GetAtt

- LoadBalancerId: the unique ID of the SLB instance.
- NetworkType: the network type of the SLB instance, which can be vpc or classic.

- AddressType: the address type of the SLB instance, which can be intranet or internet.
- IpAddress: the IP address of the SLB instance.
- OrderId: the ID of the order.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "CreateLoadBalance": {
      "Type": "ALIYUN::SLB::LoadBalancer",
      "Properties": {
        "LoadBalancerName": "createdByHeat",
        "AddressType": "internet",
        "InternetChargeType": "paybybandwidth",
      }
    }
  },
  "Outputs": {
    "LoadBalanceDetails": {
      "Value": {
        "Fn::GetAtt": ["CreateLoadBalance", "LoadBalancerId"]}
    }
  }
}
```

5.1.6.7.8. ALIYUN::SLB::LoadBalancerClone

ALIYUN::SLB::LoadBalancerClone is used to clone an SLB instance.

Syntax

```
{
  "Type": "ALIYUN::SLB::LoadBalancerClone",
  "Properties": {
    "Tags": List,
    "ResourceGroupId": String,
    "VSwitchId": String,
    "LoadBalancerName": String,
    "SourceLoadBalancerId": String,
    "TagsPolicy": String,
    "BackendServersPolicy": String,
    "BackendServers": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	No	No	The ID of the resource group.	None

Property	Type	Required	Editable	Description	Constraint
VSwitchId	String	No	No	The ID of the VSwitch.	The VSwitch must exist in the VPC to which the source SLB instance belongs. If the parameter is not specified, the VSwitch to which the source SLB instance belongs is used.
SourceLoadBalancerId	String	Yes	No	The ID of the SLB instance to be cloned.	None

Property	Type	Required	Editable	Description	Constraint
BackendServersPolicy	String	No	No	The clone policy. The ECS instances listened by the new SLB instance and the weight of each ECS instance are specified in the policy.	Valid values: • clone: The ECS instances listened by the source SLB instance and the ECS instance weights are cloned to the new SLB instance. • empty: No ECS instances are attached to the new SLB instance. • append: The ECS instances listened by the source SLB instance and the ECS instance weights are cloned to the new SLB instance. New ECS instances with specified weights are also attached to the new SLB instance. • replace: New ECS instances with specified weights are attached to the new SLB instance. However, the ECS instances listened by the source SLB instance and the ECS instance weights are not cloned to the new SLB instance. Default value: clone.

Property	Type	Required	Editable	Description	Constraint
BackendServers	List	No	Yes	The new ECS instances to be listened.	None
LoadBalancerName	String	No	No	The name of the new SLB instance.	You can set the name to any string. The name must be 1 to 80 characters in length and can contain letters, digits, hyphens (-), forward slashes (/), periods (.), and underscores (_).
Tags	List	No	Yes	The tags of the SLB instance.	Tags must be specified as key-value pairs. A maximum of five tags can be specified.
TagsPolicy	String	No	No	The policy of the tags.	Valid values: • clone: The tags of the source SLB instance are used. • empty: No tags are configured. • append: The tags of the source SLB instance are reserved while new tags are added. • replace: The tags of the source SLB instance are deleted while new tags are added. Default value: empty.

BackendServers syntax

```
"BackendServers": [
  {
    "ServerId": String,
    "Weight": Integer
  }
]
```

BackendServers properties

Property	Type	Required	Editable	Description	Constraint
ServerId	String	Yes	Yes	The ID of the ECS instance.	The ECS instances must be in the running state.
Weight	Integer	Yes	Yes	The weight of the ECS instance to be attached to the SLB instance.	Valid values: 0 to 100. Default value: 100.

Response parameters

Fn::GetAtt

LoadBalancerId: the ID of the new SLB instance.

Examples

```
{
  "ROSTemplateFormatVersion" : "2015-09-01",
  "Resources" : {
    "CloneLoadBalance": {
      "Type": "ALIYUN::SLB::LoadBalancerClone",
      "Properties": {
        "SourceLoadBalancerId": "150ebed5f06-cn-beijing-btc-***",
        "LoadBalancerName": "rosnew",
        "BackendServersPolicy": "replace",
        "BackendServers": [
            {
                "ServerId": "i-25zsk****",
                "Weight": 20
            }
        ]
      }
    }
  },
  "Outputs": {
    "LoadBalanceDetails": {
         "Value" : {"Fn::GetAtt": ["CloneLoadBalance", "LoadBalancerId"]}
    }
  }
}
```

5.1.6.7.9. ALIYUN::SLB::MasterSlaveServerGroup

ALIYUN::SLB::MasterSlaveServerGroup is used to create a primary/secondary server group.

> **Notice** A primary/secondary server group contains only two ECS instances: a primary server and a secondary server.

Syntax

```
{
  "Type": "ALIYUN::SLB::MasterSlaveServerGroup",
  "Properties": {
    "MasterSlaveServerGroupName": String,
    "MasterSlaveBackendServers": List,
    "LoadBalancerId": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
MasterSlaveServerGroupName	String	No	No	The name of the primary/secondary server group.	None
MasterSlaveBackendServers	List	Yes	No	The list of backend servers in the primary/secondary server group.	A primary/secondary server group can contain a maximum of two backend servers. If you do not specify this parameter, an empty primary/secondary server group is created.
LoadBalancerId	String	Yes	No	The ID of the SLB instance.	None

MasterSlaveBackendServers syntax

```
"MasterSlaveBackendServers": [
  {
    "ServerId": String,
    "Port": Integer,
    "Weight": Integer,
    "ServerType": String
  }
]
```

MasterSlaveBackendServers properties

Property	Type	Required	Editable	Description	Constraint

Property	Type	Required	Editable	Description	Constraint
ServerId	String	Yes	No	The ID of the ECS instance or Elastic Network Interface (ENI) to be added.	None
ServerType	String	No	No	The type of the server.	Default value: Master. Valid values: • Master • Slave
Port	Integer	Yes	No	The port number used by the backend server.	Valid values: 1 to 65535.
Weight	Integer	Yes	No	The weight of the backend server.	Valid values: 0 to 100.

Response parameters

Fn::GetAtt

MasterSlaveServerGroupId: the ID of the primary/secondary server group.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "MasterSlaveServerGroup": {
      "Type": "ALIYUN::SLB::MasterSlaveServerGroup",
      "Properties": {
        "MasterSlaveServerGroupName": "Group1",
        "MasterSlaveBackendServers": [
          {
            "ServerId": "vm****",
            "Port": "80",
            "Weight": "100",
            "ServerType": "Master"
          },
          {
            "ServerId": "vm****",
            "Port": "90",
            "Weight": "100",
            "ServerType": "Slave"
          }
        ],
        "LoadBalancerId": "lb-bp1hv944r69al4j9j****"
      }
    }
  },
  "Outputs": {
    "MasterSlaveServerGroupId": {
      "Value": {
        "Fn::GetAtt": [
          "MasterSlaveServerGroup",
          "MasterSlaveServerGroupId"
        ]
      }
    }
  }
}
```

5.1.6.7.10. ALIYUN::SLB::Rule

ALIYUN::SLB::Rule is used to add forwarding rules for a specified HTTP or HTTPS listener.

Statement

```
{
  "Type": "ALIYUN::SLB::Rule",
  "Properties": {
    "ListenerPort": Integer,
    "RuleList": List,
    "LoadBalancerId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ListenerPort	Integer	Yes	No	The frontend listener port used by the SLB instance.	Valid values: 1 to 65,535.
RuleList	List	Yes	No	The list of forwarding rules to be added.	A maximum of 10 forwarding rules can be added at a time. Each forwarding rule contains the following parameters: RuleName, Domain, Url, and VServerGroupId. You must specify at least one of the following parameters: Domain and URL. The combination of Domain and URL must be unique in a listener.
LoadBalancerId	String	No	No	The IDs of SLB instances.	None

RuleList syntax

```
"RuleList": [
  {
    "Url": String,
    "Domain": String,
    "VServerGroupId": String,
    "RuleName": String
  }
]
```

RuleList properties

Parameter	Type	Required	Editable	Description	Constraint

Parameter	Type	Required	Editable	Description	Constraint
Url	String	Yes	Released	The request URL.	The name must be 1 to 80 characters in length and can contain letters, numbers, and special characters. `-/. percent signs (%), question marks (?), #&` .
Domain	String	Yes	Released	The request domain name associated with the forwarding rule.	None
VServerGroupId	String	No	No	The ID of the destination VServer group specified in the forwarding rule.	None
RuleName	String	No	No	The name of the forwarding rule.	The name must be 1 to 40 characters in length and can contain letters, numbers, and special characters. `-/. _` . Forwarding rule names must be unique within each listener.

Response parameters

Fn::GetAtt

Rules: the list of forwarding rules.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Rule": {
      "Type": "ALIYUN::SLB::Rule",
      "Properties": {
        "ListenerPort": {
          "Ref": "ListenerPort"
        },
        "RuleList": {
          "Fn::Split": [",", {
            "Ref": "RuleList"
          }, {
            "Ref": "RuleList"
          }]
        },
        "LoadBalancerId": {
          "Ref": "LoadBalancerId"
        }
      }
    }
  },
  "Parameters": {
    "ListenerPort": {
      "Type": "Number",
      "Description": "The front-end HTTPS listener port of the Server Load Balancer instance. Valid va
lue:\n1-65535",
      "MaxValue": 65535,
      "MinValue": 1
    },
    "RuleList": {
      "MinLength": 1,
      "Type": "CommaDelimitedList",
      "Description": "The forwarding rules to add.",
      "MaxLength": 10
    },
    "LoadBalancerId": {
      "Type": "String",
      "Description": "The ID of Server Load Balancer instance."
    }
  },
  "Outputs": {
    "Rules": {
      "Description": "A list of forwarding rules. Each element of rules contains \"RuleId\".",
      "Value": {
        "Fn::GetAtt": ["Rule", "Rules"]
      }
    }
  }
}
```

5.1.6.7.11. ALIYUN::SLB::VServerGroup

ALIYUN::SLB::VServerGroup is used to create a VServer group and adds backend servers to the SLB instance.

Syntax

```
{
    "Type" : "ALIYUN::SLB::VServerGroup",
    "Properties" : {
        "VServerGroupName" : String,
        "BackendServers" : List,
        "LoadBalancerId" : String
    }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
VServerGroupName	String	Yes	No	The name of the VServer group.	None
BackendServers	List	Yes	Yes	The list of ECS instances to be added.	A list can contain up to 20 instances.
LoadBalancerId	String	Yes	No	The ID of the SLB instance.	None

BackendServers syntax

```
"BackendServers" : [
    {
        "ServerId" : String,
        "Port" : Integer,
        "Weight" : Integer
    }
]
```

BackendServers properties

Property	Type	Required	Editable	Description	Constraint
ServerId	String	Yes	Yes	The ID of the ECS instance.	None
Port	Integer	Yes	Yes	The backend port used by the SLB instance.	Valid values: 1 to 65535.
Weight	Integer	Yes	Yes	The weight of the ECS instance to be attached to the SLB instance.	Valid values: 0 to 100.

Response parameters

Fn::GetAtt

- VServerGroupId: the ID of the VServer group.
- BackendServers: the list of backend servers added to the SLB instance

Examples

```
{
    "ROSTemplateFormatVersion": "2015-09-01",
    "Resources": {
        "CreateVServerGroup": {
            "Type": "ALIYUN::SLB::VServerGroup",
            "Properties": {
                "LoadBalancerId": "lb-2zenh4ndwrqg14yt0****",
                "VServerGroupName": "VServerGroup-****",
                "BackendServers": [
                    {
                        "ServerId": "i-25zsk****",
                        "Weight": 20,
                        "Port": 8080
                    },
                    {
                        "ServerId": "i-25zsk****",
                        "Weight": 100,
                        "Port": 8081
                    }
                ]
            }
        }
    },
    "Outputs": {
        "VServerGroupId": {
            "Value" : {"Fn::GetAttr": ["CreateVServerGroup",  "VServerGroupId"]
         }
      }
   }
}
```

5.1.6.8. SLS

5.1.6.8.1. ALIYUN::SLS::Index

ALIYUN::SLS::Index is used to create an index for a specified Logstore.

Syntax

```
{
  "Type": "ALIYUN::SLS::Index",
  "Properties": {
    "ProjectName": String,
    "FullTextIndex": Map,
    "LogstoreName": String,
    "KeyIndices": List,
    "LogReduce": Boolean
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ProjectName	String	Yes	No	The name of the Log Service project.	The name must be 3 to 36 characters in length and can contain lowercase letters, digits, hyphens (-), and underscores (_). It must start with a lowercase letter or digit.
FullTextIndex	Map	Yes	Yes	The full-text index configurations.	For more information, see FullTextIndex properties.
LogstoreName	String	Yes	No	The name of the Logstore.	None
KeyIndices	List	No	Yes	The field index configurations.	You must specify at least one of the FullTextIndex and KeyIndices parameters. For more information, see Properties.
LogReduce	Boolean	No	Yes	Specifies whether to enable LogReduce.	Default value: false. Valid values: • true • false

FullTextIndex syntax

```
"FullTextIndex": {
  "CaseSensitive": Boolean,
  "Delimiter": String,
  "IncludeChinese": Boolean,
  "Enable": Boolean
}
```

FullTextIndex properties

Property	Type	Required	Editable	Description	Constraint

Property	Type	Required	Editable	Description	Constraint
Enable	Boolean	Yes	Yes	Specifies whether to enable full-text indexing.	Default value: true. Valid values: • true • false
CaseSensitive	Boolean	No	Yes	Specifies whether the field is case-sensitive.	Default value: false. Valid values: • true • false
Delimiter	String	No	Yes	The delimiter that is used to separate keywords.	Valid values: `,'";=()[]{}? @&<>/:\n\t\r`
IncludeChinese	Boolean	No	Yes	Specifies whether to support Chinese word segmentation.	Default value: false. Valid values: • true • false

KeyIndices syntax

```
"KeyIndices": [
  {
    "Name": String,
    "EnableAnalytics": Boolean,
    "Delimiter": String,
    "CaseSensitive": Boolean,
    "JsonKeyIndices": List,
    "Alias": String,
    "IncludeChinese": String,
    "Type": String
  }
]
```

KeyIndices properties

Property	Type	Required	Editable	Description	Constraint
Name	String	Yes	Yes	The name of the field.	You can use a nested name that is separated with periods (.). Example: k1.k2.k3.

Property	Type	Required	Editable	Description	Constraint
EnableAnalytics	Boolean	No	Yes	Specifies whether to enable statistical analysis on the field.	Default value: true. Valid values: • true • false
Delimiter	String	No	Yes	The delimiter that is used to separate keywords.	Valid values: `,'";=()[]{}? @&<>/:\n\t\r`
CaseSensitive	Boolean	No	Yes	Specifies whether the field is case-sensitive.	Default value: false. Valid values: • true • false This parameter takes effect only when the Type parameter is set to text or json.
JsonKeyIndices	List	No	Yes	The JSON index configurations. Format: `[{"key1": "value1", "key2": "value2", ...}]`.	Supported keys are Name, Alias, Type, and EnableAnalytics. For more information, see JsonKeyIndices properties.
Alias	String	No	Yes	The alias of the field.	None
IncludeChinese	Boolean	No	Yes	Specifies whether to support Chinese word segmentation.	Default value: false. Valid values: • true • false This parameter takes effect only when the Type parameter is set to text.

Property	Type	Required	Editable	Description	Constraint
Type	String	Yes	Yes	The type of the field.	Default value: text. Valid values: • text • long • double • json

JsonKeyIndices syntax

```
"JsonKeyIndices": [
  {
    "Type": String,
    "Alias": String,
    "EnableAnalytics": Boolean,
    "Name": String
  }
]
```

JsonKeyIndices properties

Property	Type	Required	Editable	Description	Constraint
Name	String	Yes	Yes	The name of the field.	None
EnableAnalytics	Boolean	No	Yes	Specifies whether to enable statistical analysis on the field.	Valid values: • true • false
Alias	String	No	Yes	The alias of the field.	None
Type	String	Yes	Yes	The type of the field.	None

Response parameters

Fn::GetAtt

None

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Index": {
      "Type": "ALIYUN::SLS::Index",
      "Properties": {
        "ProjectName": {
```

```
          "Ref": "ProjectName"
        },
        "FullTextIndex": {
          "Ref": "FullTextIndex"
        },
        "LogstoreName": {
          "Ref": "LogstoreName"
        },
        "KeyIndices": {
          "Ref": "KeyIndices"
        },
        "LogReduce": {
          "Ref": "LogReduce"
        }
      }
    }
  },
  "Parameters": {
    "ProjectName": {
      "Type": "String",
      "Description": "Project name:1. Only supports lowercase letters, numbers, hyphens (-) and unders
cores (_).2. Must start and end with lowercase letters and numbers.3. The name length is 3-63 characte
rs.",
      "MinLength": 3,
      "MaxLength": 63
    },
    "FullTextIndex": {
      "Type": "Map",
      "Description": "Full-text indexing configuration.Full-text indexing and key indexing must have a
t least one enabled."
    },
    "LogstoreName": {
      "Type": "String",
      "Description": "Logstore name:1. Only supports lowercase letters, numbers, hyphens (-) and under
scores (_).2. Must start and end with lowercase letters and numbers.3. The name length is 3-63 charact
ers.",
      "MinLength": 3,
      "MaxLength": 63
    },
    "KeyIndices": {
      "Type": "List",
      "Description": "Key index configurations.Full-text indexing and key indexing must have at least
one enabled."
    },
    "LogReduce": {
      "Default": false,
      "Type": "Boolean",
      "Description": "Whether to enable log reduce. Default to false.",
      "AllowedValues": [
        true,
        false
      ]
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Resources:
  Index:
    Type: ALIYUN::SLS::Index
    Properties:
      ProjectName:
        Ref: ProjectName
      FullTextIndex:
        Ref: FullTextIndex
      LogstoreName:
        Ref: LogstoreName
      KeyIndices:
        Ref: KeyIndices
      LogReduce:
        Ref: LogReduce
Parameters:
  ProjectName:
    MinLength: 3
    MaxLength: 63
    Type: String
    Description: Project name:1. Only supports lowercase letters, numbers, hyphens
      (-) and underscores (_).2. Must start and end with lowercase letters and numbers.3.
      The name length is 3-63 characters.
  FullTextIndex:
    Type: Map
    Description: Full-text indexing configuration.Full-text indexing and key indexing
      must have at least one enabled.
  LogstoreName:
    MinLength: 3
    MaxLength: 63
    Type: String
    Description: Logstore name:1. Only supports lowercase letters, numbers, hyphens
      (-) and underscores (_).2. Must start and end with lowercase letters and numbers.3.
      The name length is 3-63 characters.
  KeyIndices:
    Type: List
    Description: Key index configurations.Full-text indexing and key indexing must
      have at least one enabled.
  LogReduce:
    Default: false
    Type: Boolean
    Description: Whether to enable log reduce. Default to false.
    AllowedValues:
    - true
    - false
```

5.1.6.8.2. ALIYUN::SLS::Logstore

ALIYUN::SLS::Logstore is used to create a Logstore in a Log Service project.

Syntax

```
{
  "Type": "ALIYUN::SLS::Logstore",
  "Properties": {
    "ProjectName": String,
    "ShardCount": Integer,
    "AutoSplit": Boolean,
    "MaxSplitShard": Integer,
    "LogstoreName": String,
    "AppendMeta": Boolean,
    "TTL": Integer,
    "EnableTracking": Boolean,
    "PreserveStorage": Boolean
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ProjectName	String	Yes	No	The name of the Log Service project.	• The name must be 3 to 36 characters in length. • It can contain lowercase letters, digits, hyphens (-), and underscores (_). • It must start and end with a lowercase letter or digit.
ShardCount	Integer	No	Yes	The number of shards.	Valid values: 1 to 100. Default value: 2.
MaxSplitShard	Integer	No	Yes	The maximum number of shards during automatic splitting.	Valid values: 1 to 64. This parameter is required when the AutoSplit parameter is set to true.

Property	Type	Required	Editable	Description	Constraint
LogstoreName	String	Yes	No	The name of the Logstore.	The name must be unique in a project. • The name must be 3 to 36 characters in length. • It can contain lowercase letters, digits, hyphens (-), and underscores (_). • It must start and end with a lowercase letter or digit.
AutoSplit	Boolean	No	Yes	Specifies whether to automatically split shards.	Valid values: • true • false Default value: false.
TTL	Integer	No	Yes	The data retention period.	Valid values: 1 to 3600. Default value: 30. Unit: days.
EnableTracking	Boolean	No	Yes	Specifies whether to enable WebTracking.	WebTracking can collect access information about web browsers, iOS applications, and Android applications. Valid values: • true • false Default value: false.

Property	Type	Required	Editable	Description	Constraint
PreserveStorage	Boolean	No	Yes	Specifies whether to permanently preserve logs.	Valid values: • true • false Default value: false. If this parameter is set to true, the TTL parameter does not take effect.
AppendMeta	Boolean	No	Yes	Specifies whether to add the public IP address of the client and the log arrival time after the log is received.	Valid values: • true • false Default value: false.

Response parameters

Fn::GetAtt

LogstoreName: the name of the Logstore.

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Logstore": {
      "Type": "ALIYUN::SLS::Logstore",
      "Properties": {
        "ProjectName": {
          "Ref": "ProjectName"
        },
        "TTL": {
          "Ref": "TTL"
        },
        "AutoSplit": {
          "Ref": "AutoSplit"
        },
        "MaxSplitShard": {
          "Ref": "MaxSplitShard"
        },
        "LogstoreName": {
          "Ref": "LogstoreName"
        },
        "AppendMeta": {
          "Ref": "AppendMeta"
        },
        "ShardCount": {
          "Ref": "ShardCount"
        },
```

```
        "EnableTracking": {
          "Ref": "EnableTracking"
        },
        "PreserveStorage": {
          "Ref": "PreserveStorage"
        }
      }
    }
  },
  "Parameters": {
    "ProjectName": {
      "MinLength": 3,
      "Type": "String",
      "Description": "Project name: 1. Only supports lowercase letters, numbers, hyphens (-) and under
scores (_). 2. Must start and end with lowercase letters and numbers. 3. The name length is 3-63 chara
cters.",
      "MaxLength": 63
    },
    "TTL": {
      "Default": 30,
      "Type": "Number",
      "Description": "The lifecycle of log in the logstore in days. Allowed Values: 1-3600, default to
30.",
      "MaxValue": 3600,
      "MinValue": 1
    },
    "AutoSplit": {
      "Default": false,
      "Type": "Boolean",
      "Description": "Whether to automatically split the shard. Default to false.",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    },
    "MaxSplitShard": {
      "Type": "Number",
      "Description": "The maximum number of shards when splitting automatically. Must be specified if
AutoSplit is set to true. Allowed Values: 1-64.",
      "MaxValue": 64,
      "MinValue": 1
    },
    "LogstoreName": {
      "MinLength": 3,
      "Type": "String",
      "Description": "Logstore name: 1. Only supports lowercase letters, numbers, hyphens (-) and unde
rscores (_). 2. Must start and end with lowercase letters and numbers. 3. The name length is 3-63 char
acters.",
      "MaxLength": 63
    },
    "AppendMeta": {
      "Default": false,
      "Type": "Boolean",
      "Description": "Whether to add client external network IP and log arrival time after receiving t
he log. Default to false.",
      "AllowedValues": [
        "True",
        "true",
```

```
        "False",
        "false"
      ]
    },
    "ShardCount": {
      "Default": 2,
      "Type": "Number",
      "Description": "The number of Shards. Allowed Values: 1-100, default to 2.",
      "MaxValue": 100,
      "MinValue": 1
    },
    "EnableTracking": {
      "Default": false,
      "Type": "Boolean",
      "Description": "Whether to enable WebTracking, which supports fast capture of various browsers a
nd iOS/Android/APP access information. Default to false.",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    },
    "PreserveStorage": {
      "Default": false,
      "Type": "Boolean",
      "Description": "Whether to keep the log permanently. If set to true, TTL will be ignored. Defaul
t to false.",
      "AllowedValues": [
        "True",
        "true",
        "False",
        "false"
      ]
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Resources:
  Logstore:
    Type: ALIYUN::SLS::Logstore
    Properties:
      ProjectName:
        Ref: ProjectName
      TTL:
        Ref: TTL
      AutoSplit:
        Ref: AutoSplit
      MaxSplitShard:
        Ref: MaxSplitShard
      LogstoreName:
        Ref: LogstoreName
      AppendMeta:
        Ref: AppendMeta
      ShardCount:
        Ref: ShardCount
```

```
      EnableTracking:
        Ref: EnableTracking
      PreserveStorage:
        Ref: PreserveStorage
Parameters:
  ProjectName:
    MinLength: 3
    Type: String
    Description: Project name: 1. Only supports lowercase letters, numbers, hyphens
      (-) and underscores (_). 2. Must start and end with lowercase letters and numbers. 3.
      The name length is 3-63 characters.
    MaxLength: 63
  TTL:
    Default: 30
    Type: Number
    Description: 'The lifecycle of log in the logstore in days. Allowed Values: 1-3600,
      default to 30.'
    MaxValue: 3600
    MinValue: 1
  AutoSplit:
    Default: false
    Type: Boolean
    Description: Whether to automatically split the shard. Default to false.
    AllowedValues:
    - 'True'
    - 'true'
    - 'False'
    - 'false'
  MaxSplitShard:
    Type: Number
    Description: 'The maximum number of shards when splitting automatically. Must
      be specified if AutoSplit is set to true. Allowed Values: 1-64.'
    MaxValue: 64
    MinValue: 1
  LogstoreName:
    MinLength: 3
    Type: String
    Description: Logstore name: 1. Only supports lowercase letters, numbers, hyphens
      (-) and underscores (_). 2. Must start and end with lowercase letters and numbers. 3.
      The name length is 3-63 characters.
    MaxLength: 63
  AppendMeta:
    Default: false
    Type: Boolean
    Description: Whether to add client external network IP and log arrival time after
      receiving the log. Default to false.
    AllowedValues:
    - 'True'
    - 'true'
    - 'False'
    - 'false'
  ShardCount:
    Default: 2
    Type: Number
    Description: 'The number of Shards. Allowed Values: 1-100, default to 2.'
    MaxValue: 100
    MinValue: 1
  EnableTracking:
    Default: false
    Type: Boolean
```

```
    Description: Whether to enable WebTracking, which supports fast capture of various
      browsers and iOS/Android/APP access information. Default to false.
    AllowedValues:
    - 'True'
    - 'true'
    - 'False'
    - 'false'
  PreserveStorage:
    Default: false
    Type: Boolean
    Description: Whether to keep the log permanently. If set to true, TTL will be ignored. Default
      to false.
    AllowedValues:
    - 'True'
    - 'true'
    - 'False'
    - 'false'
```

5.1.6.8.3. ALIYUN::SLS::LogtailConfig

ALIYUN::SLS::LogtailConfig is used to configure Logtail parameters for data collection.

Syntax

```
{
  "Type": "ALIYUN::SLS::LogtailConfig",
  "Properties": {
    "ProjectName": String,
    "LogtailConfigName": String,
    "LogstoreName": String,
    "RawConfigData": Map,
    "CloneFrom": Map
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
ProjectName	String	Yes	No	The name of the Log Service project.	None
LogtailConfigName	String	Yes	No	The name of the Logtail configuration file.	The configuration file name must be unique in a project. The name must be 2 to 128 characters in length and can contain lowercase letters, digits, hyphens (-), and underscores (_). It must start and end with a lowercase letter or digit.
LogstoreName	String	Yes	No	The name of the Logstore.	None

Property	Type	Required	Editable	Description	Constraint
RawConfigData	Map	No	Yes	The raw configuration data.	The format is the same as that of the response from the GetConfig operation of Log Service. If CloneFrom and RawConfigData are both specified, the data of LogtailConfig and RawConfigData is merged. In this case, configName, outputType, and outputDetail of RawConfigData are ignored. Sample value: <code>{"configName": "test-logtail-config", "createTime": 1574843554, "inputDetail": {"acceptNoEnoughKeys": false, "adjustTimezone": false, "advanced": { "force_multiconfig": false }, "autoExtend": true, "delayAlarmBytes": 0, "delaySkipBytes": 0, "discardNonUtf8": false, "discardUnmatch": false, "dockerExcludeEnv": {}, "dockerExcludeLabel": {}, "dockerFile": false, "dockerIncludeEnv": {}, "dockerIncludeLabel": {}, "enableRawLog": false, "enableTag": false, "fileEncoding": "utf8", "filePattern": "test.log*", "filterKey": [], "filterRegex": [], "key": ["time", "logger", "level", "request_id", "user_id", "region_id", "content"], "localStorage": true,</code>

Property	Type	Required	Editable	Description	Constraint
					"logPath": "/var/log/test", "logTimezone": "", "logType": "delimiter_log", "maxDepth": 100, "maxSendRate": -1, "mergeType": "topic", "preserve": true, "preserveDepth": 1, "priority": 0, "quote": "\u0001", "sendRateExpire": 0, "sensitive_keys": [], "separator": ",,,", "shardHashKey": [], "tailExisted": false, "timeFormat": "", "timeKey": "", "topicFormat": "none" }, "inputType": "file", "lastModifyTime": 1574843554, "logSample": "2019-11-27 10:48:23,160,,,MAIN,,,INFO,,,98DCC51D-BE5D-49C7-B3FD-37B2BAEFB296,,,123456789,,,cn-hangzhou,,,this is a simple test.", "outputDetail": { "endpoint": "cn-hangzhou-intranet.log.aliyuncs.com", "logstoreName": "test-logstore", "region": "cn-hangzhou"}, "outputType": "LogService"} .
CloneFrom	Map	No	Yes	The configurations for cloning LogtailConfig data of another Log Service project.	You must specify one of the CloneFrom and LogtailConfig parameters. For more information, see CloneFrom properties.

CloneFrom syntax

```
"CloneFrom": {
  "ProjectName": String,
  "LogtailConfigName": String
}
```

CloneFrom properties

Property	Type	Required	Editable	Description	Constraint
ProjectName	String	Yes	Yes	The name of the Log Service project.	None
LogtailConfigName	String	Yes	Yes	The name of the Logtail configuration file.	None

Response parameters

Fn::GetAtt

- Endpoint: the endpoint.
- AppliedMachineGroups: a list of machines configured for log collection.
- LogtailConfigName: the name of the Logtail configuration file.

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "LogtailConfig": {
      "Type": "ALIYUN::SLS::LogtailConfig",
      "Properties": {
        "ProjectName": {
          "Ref": "ProjectName"
        },
        "LogtailConfigName": {
          "Ref": "LogtailConfigName"
        },
        "LogstoreName": {
          "Ref": "LogstoreName"
        },
        "RawConfigData": {
          "Ref": "RawConfigData"
        },
        "CloneFrom": {
          "Ref": "CloneFrom"
        }
      }
    }
  },
  "Parameters": {
    "ProjectName": {
      "MinLength": 3,
      "Type": "String",
```

```
      "Description": "Project name: 1. Only supports lowercase letters, numbers, hyphens (-) and under
scores (_). 2. Must start and end with lowercase letters and numbers. 3. The name length is 3-63 chara
cters.",
      "MaxLength": 63
    },
    "LogtailConfigName": {
      "MinLength": 3,
      "Type": "String",
      "Description": "Logtail config name: 1. Only supports lowercase letters, numbers, hyphens (-) an
d underscores (_). 2. Must start and end with lowercase letters and numbers. 3. The name length is 3-6
3 characters.",
      "MaxLength": 63
    },
    "LogstoreName": {
      "MinLength": 3,
      "Type": "String",
      "Description": "Logstore name: 1. Only supports lowercase letters, numbers, hyphens (-) and unde
rscores (_). 2. Must start and end with lowercase letters and numbers. 3. The name length is 3-63 char
acters.",
      "MaxLength": 63
    },
    "RawConfigData": {
      "Type": "Json",
      "Description": "The format is the same as the response of SLS API GetConfig. Either CloneFrom or
RawConfigData must be specified. If CloneFrom and RawConfigData are both specified, logtail config dat
a will be merged from both with RawConfigData first. configName, outputType, outputDetail in data will
be ignored.For example:\n{\"configName\": \"test-logtail-config\",\"createTime\": 1574843554,\"inputDe
tail\": { \"acceptNoEnoughKeys\": false, \"adjustTimezone\": false, \"advanced\": {   \"force_multicon
fig\": false }, \"autoExtend\": true, \"delayAlarmBytes\": 0, \"delaySkipBytes\": 0, \"discardNonUtf8\
": false, \"discardUnmatch\": false, \"dockerExcludeEnv\": {}, \"dockerExcludeLabel\": {}, \"dockerFil
e\": false, \"dockerIncludeEnv\": {}, \"dockerIncludeLabel\": {}, \"enableRawLog\": false, \"enableTag
\": false, \"fileEncoding\": \"utf8\", \"filePattern\": \"test.log*\", \"filterKey\": [], \"filterRege
x\": [], \"key\": [   \"time\",   \"logger\",   \"level\",   \"request_id\",   \"user_id\",   \"region
_id\",   \"content\" ], \"localStorage\": true, \"logPath\": \"/var/log/test\", \"logTimezone\": \"\",
\"logType\": \"delimiter_log\", \"maxDepth\": 100, \"maxSendRate\": -1, \"mergeType\": \"topic\", \"pr
eserve\": true, \"preserveDepth\": 1, \"priority\": 0, \"quote\": \"\\u0001\", \"sendRateExpire\": 0,
\"sensitive_keys\": [], \"separator\": \",,,\", \"shardHashKey\": [], \"tailExisted\": false, \"timeFo
rmat\": \"\", \"timeKey\": \"\", \"topicFormat\": \"none\"}, \"inputType\": \"file\", \"lastModifyTime
\": 1574843554, \"logSample\": \"2019-11-27 10:48:23,160,,,MAIN,,,INFO,,,98DCC51D-BE5D-49C7-B3FD-37B2B
AEFB296,,,123456789,,,cn-hangzhou,,,this is a simple test.\", \"outputDetail\": {\"endpoint\": \"cn-ha
ngzhou-intranet.log.aliyuncs.com\", \"logstoreName\": \"test-logstore\",\"region\": \"cn-hangzhou\"},
\"outputType\": \"LogService\"}"
    },
    "CloneFrom": {
      "Type": "Json",
      "Description": "Clone logtail config data from existing logtail config. Either CloneFrom or RawC
onfigData must be specified. If CloneFrom and RawConfigData are both specified, logtail config data wi
ll be merged from both with RawConfigData first."
    }
  },
  "Outputs": {
    "LogtailConfigName": {
      "Description": "Logtail config name.",
      "Value": {
        "Fn::GetAtt": [
          "LogtailConfig",
          "LogtailConfigName"
        ]
      }
    },
```

```
    "Endpoint": {
      "Description": "Endpoint address.",
      "Value": {
        "Fn::GetAtt": [
          "LogtailConfig",
          "Endpoint"
        ]
      }
    },
    "AppliedMachineGroups": {
      "Description": "Applied machine groups.",
      "Value": {
        "Fn::GetAtt": [
          "LogtailConfig",
          "AppliedMachineGroups"
        ]
      }
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Resources:
  LogtailConfig:
    Type: ALIYUN::SLS::LogtailConfig
    Properties:
      ProjectName:
        Ref: ProjectName
      LogtailConfigName:
        Ref: LogtailConfigName
      LogstoreName:
        Ref: LogstoreName
      RawConfigData:
        Ref: RawConfigData
      CloneFrom:
        Ref: CloneFrom
Parameters:
  ProjectName:
    MinLength: 3
    Type: String
    Description: 'Project name: 1. Only supports lowercase letters, numbers, hyphens
      (-) and underscores (_). 2. Must start and end with lowercase letters and numbers.
      3. The name length is 3-63 characters.'
    MaxLength: 63
  LogtailConfigName:
    MinLength: 3
    Type: String
    Description: 'Logtail config name: 1. Only supports lowercase letters, numbers,
      hyphens (-) and underscores (_). 2. Must start and end with lowercase letters
      and numbers. 3. The name length is 3-63 characters.'
    MaxLength: 63
  LogstoreName:
    MinLength: 3
    Type: String
    Description: 'Logstore name: 1. Only supports lowercase letters, numbers, hyphens
      (-) and underscores (_). 2. Must start and end with lowercase letters and numbers.
      3. The name length is 3-63 characters.'
    MaxLength: 63
```

```
    MaxLength: 65
  RawConfigData:
    Type: Json
    Description: |-
      The format is the same as the response of SLS API GetConfig. Either CloneFrom or RawConfigData m
ust be specified. If CloneFrom and RawConfigData are both specified, logtail config data will be merge
d from both with RawConfigData first. configName, outputType, outputDetail in data will be ignored.For
example:
      {"configName": "test-logtail-config","createTime": 1574843554,"inputDetail": { "acceptNoEnoughKe
ys": false, "adjustTimezone": false, "advanced": {   "force_multiconfig": false }, "autoExtend": true,
"delayAlarmBytes": 0, "delaySkipBytes": 0, "discardNonUtf8": false, "discardUnmatch": false, "dockerEx
cludeEnv": {}, "dockerExcludeLabel": {}, "dockerFile": false, "dockerIncludeEnv": {}, "dockerIncludeLa
bel": {}, "enableRawLog": false, "enableTag": false, "fileEncoding": "utf8", "filePattern": "test.log*
", "filterKey": [], "filterRegex": [], "key": [   "time",   "logger",   "level",   "request_id",   "us
er_id",   "region_id",   "content" ], "localStorage": true, "logPath": "/var/log/test", "logTimezone":
"", "logType": "delimiter_log", "maxDepth": 100, "maxSendRate": -1, "mergeType": "topic", "preserve":
true, "preserveDepth": 1, "priority": 0, "quote": "\u0001", "sendRateExpire": 0, "sensitive_keys": [],
"separator": ",,,", "shardHashKey": [], "tailExisted": false, "timeFormat": "", "timeKey": "", "topicF
ormat": "none"}, "inputType": "file", "lastModifyTime": 1574843554, "logSample": "2019-11-27 10:48:23,
160,,,MAIN,,,INFO,,,98DCC51D-BE5D-49C7-B3FD-37B2BAEFB296,,,123456789,,,cn-hangzhou,,,this is a simple
test.", "outputDetail": {"endpoint": "cn-hangzhou-intranet.log.aliyuncs.com", "logstoreName": "test-lo
gstore","region": "cn-hangzhou"}, "outputType": "LogService"}
  CloneFrom:
    Type: Json
    Description: Clone logtail config data from existing logtail config. Either CloneFrom
      or RawConfigData must be specified. If CloneFrom and RawConfigData are both
      specified, logtail config data will be merged from both with RawConfigData first.
Outputs:
  Endpoint:
    Description: Endpoint address.
    Value:
      Fn::GetAtt:
      - LogtailConfig
      - Endpoint
  AppliedMachineGroups:
    Description: Applied machine groups.
    Value:
      Fn::GetAtt:
      - LogtailConfig
      - AppliedMachineGroups
  LogtailConfigName:
    Description: Logtail config name.
    Value:
      Fn::GetAtt:
      - LogtailConfig
      - LogtailConfigName
```

To view more examples, visit SLS.json and SLS.yml. In the examples, the ALIYUN::SLS::Project, ALIYUN::SLS::Logstore, ALIYUN::SLS::Index, ALIYUN::SLS::LogtailConfig, ALIYUN::SLS::MachineGroup, ALIYUN::SLS::ApplyConfigToMachineGroup, ALIYUN::ApiGateway::LogConfig, ALIYUN::SLS::Savedsearch, and ALIYUN::SLS::Alert resource types are involved.

5.1.6.8.4. ALIYUN::SLS::Savedsearch

ALIYUN::SLS::Savedsearch is used to save search results as a saved search.

Syntax

```
{
  "Type": "ALIYUN::SLS::Savedsearch",
  "Properties": {
    "Project": String,
    "Detail": Map
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
Project	String	Yes	No	The name of the Log Service project.	None
Detail	Map	Yes	Yes	The details of the query.	For more information, see Detail properties.

Detail syntax

```
"Detail": {
  "SearchQuery": String,
  "Logstore": String,
  "DisplayName": String,
  "SavedsearchName": String,
  "Topic": String
}
```

Detail properties

Property	Type	Required	Editable	Description	Constraint
SearchQuery	String	Yes	Yes	The query statement.	None
Logstore	String	Yes	Yes	The Logstore in which the query is performed.	None
DisplayName	String	No	Yes	The display name that you specify for the saved search.	The name must be 1 to 63 characters in length.
SavedsearchName	String	Yes	No	The name of the saved search, which is generated by the system in the project.	None
Topic	String	Yes	Yes	The log topic that is used to classify logs.	None

Response parameters

Fn::GetAtt

SavedsearchName: the name of the saved search.

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Project": {
      "Type": "String",
      "Description": "Project name"
    },
    "Detail": {
      "Type": "Json",
      "Description": ""
    }
  },
  "Resources": {
    "SavedSearch": {
      "Type": "ALIYUN::SLS::Savedsearch",
      "Properties": {
        "Project": {
          "Ref": "Project"
        },
        "Detail": {
          "Ref": "Detail"
        }
      }
    }
  },
  "Outputs": {
    "SavedsearchName": {
      "Description": "Savedsearch name.",
      "Value": {
        "Fn::GetAtt": [
          "SavedSearch",
          "SavedsearchName"
        ]
      }
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Parameters:
  Project:
    Type: String
    Description: Project name
  Detail:
    Type: Json
    Description: ''
Resources:
  SavedSearch:
    Type: 'ALIYUN::SLS::Savedsearch'
    Properties:
      Project:
        Ref: Project
      Detail:
        Ref: Detail
Outputs:
  SavedsearchName:
    Description: Savedsearch name.
    Value:
      'Fn::GetAtt':
        - SavedSearch
        - SavedsearchName
```

5.1.6.8.5. ALIYUN::SLS::Project

ALIYUN::SLS::Project is used to create a Log Service project.

Syntax

```
{
  "Type": "ALIYUN::SLS::Project",
  "Properties": {
    "Name": String,
    "Description": String,
    "Tags": List
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
Name	String	Yes	No	The name of the project.	The name must be 3 to 36 characters in length and can contain lowercase letters, digits, and hyphens (-). It must start and end with a lowercase letter or digit.
Description	String	No	No	The description of the project.	The description can be up to 64 characters in length and cannot contain the following special characters: `< > ' \ "`.

Property	Type	Required	Editable	Description	Constraint
Tags	List	No	No	The list of one or more tags of the project.	A maximum of 20 tags can be specified. For more information, see Tags properties.

Tags syntax

```
"Tags": [
  {
    "Key": String,
    "Value": String
  }
]
```

Tags properties

Property	Type	Required	Editable	Description	Constraint
Key	String	Yes	No	The key of the tag.	The tag key must be 1 to 128 characters in length and cannot contain `http://` or `https://`. It cannot start with `acs:` or `aliyun`.
Value	String	No	No	The value of the tag.	The tag value must be 0 to 128 characters in length and cannot contain `http://` or `https://`. It cannot start with `acs:` or `aliyun`.

Response parameters

Fn::GetAtt

Name: The name of the Log Service project.

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Parameters": {
    "Description": {
      "Type": "String",
      "Description": "Project description: <>'\"\\ is not supported, up to 64 characters.",
      "MaxLength": 64
    },
    "Tags": {
      "Type": "Json",
      "Description": "Tags to attach to project. Max support 20 tags to add during create project. Each tag with two properties Key and Value, and Key is required.",
      "MaxLength": 20
    },
    "Name": {
      "Type": "String",
      "Description": "Project name:\n1. Only supports lowercase letters, numbers, hyphens (-) and underscores (_).\n2. Must start and end with lowercase letters and numbers.\n3. The name length is 3-63 characters.",
      "AllowedPattern": "^[a-zA-Z0-9_-]+$",
      "MinLength": 3,
      "MaxLength": 63
    }
  },
  "Resources": {
    "Project": {
      "Type": "ALIYUN::SLS::Project",
      "Properties": {
        "Description": {
          "Ref": "Description"
        },
        "Tags": {
          "Ref": "Tags"
        },
        "Name": {
          "Ref": "Name"
        }
      }
    }
  },
  "Outputs": {
    "Name": {
      "Description": "Project name.",
      "Value": {
        "Fn::GetAtt": [
          "Project",
          "Name"
        ]
      }
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Parameters:
  Description:
    Type: String
    Description: 'Project description: <>''"\ is not supported, up to 64 characters.'
    MaxLength: 64
  Tags:
    Type: Json
    Description: >-
      Tags to attach to project. Max support 20 tags to add during create
      project. Each tag with two properties Key and Value, and Key is required.
    MaxLength: 20
  Name:
    Type: String
    Description: >-
      Project name:
      1. Only supports lowercase letters, numbers, hyphens (-) and underscores
      (_).
      2. Must start and end with lowercase letters and numbers.
      3. The name length is 3-63 characters.
    AllowedPattern: '^[a-zA-Z0-9_-]+$'
    MinLength: 3
    MaxLength: 63
Resources:
  Project:
    Type: 'ALIYUN::SLS::Project'
    Properties:
      Description:
        Ref: Description
      Tags:
        Ref: Tags
      Name:
        Ref: Name
Outputs:
  Name:
    Description: Project name.
    Value:
      'Fn::GetAtt':
        - Project
        - Name
```

To view more examples, visit SLS.json and SLS.yml. The following resource types are involved in these examples:

- ALIYUN::SLS::Project
- ALIYUN::SLS::Logstore
- ALIYUN::SLS::Index
- ALIYUN::SLS::LogtailConfig
- ALIYUN::SLS::MachineGroup
- ALIYUN::SLS::ApplyConfigToMachineGroup
- ALIYUN::ApiGateway::LogConfig
- ALIYUN::SLS::Savedsearch
- ALIYUN::SLS::Alert

5.1.6.9. VPC

5.1.6.9.1. ALIYUN::VPC::EIP

ALIYUN::VPC::EIP is used to apply for an Elastic IP address.

Statement

```
{
  "Type": "ALIYUN::VPC::EIP",
  "Properties": {
    "Isp": String,
    "Period": Number,
    "ResourceGroupId": String,
    "AutoPay": Boolean,
    "InstanceChargeType": String,
    "PricingCycle": String,
    "InternetChargeType": String,
    "Bandwidth": Number
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
ResourceGroupId	String	Yes	Released	The ID of the resource group to which the RDS instance belongs.	None
Bandwidth	Number	Erased	Released	The network bandwidth. Unit: Mbit/s.	If this parameter is not specified, the default value 5Mbps is used.
InternetChargeType	String	Yes	Released	The billing method for network usage. Default value: PayByBandwidth.	Valid values: • PayByBandwidth: pay-by-bandwidth. • PayByTraffic Default value: PayByBandwidth.
InstanceChargeType	String	Yes	Released	The billing method of the Elastic IP address. Default value: Postpaid.	Valid values: • Prepaid • Postpaid: pay-as-you-go Default value: PostPaid.

Parameter	Type	Required	Editable	Description	Constraint
PricingCycle	String	Yes	Released	The billing cycle of the subscription. Default value: Month.	Valid values: • Month: paid by month. • Year Default value: Month. **Note** This parameter is required when InstanceChargeType is set to Prepaid.
Period	Number	Erased	Released	The subscription period.	Valid values: • If pay by month is selected, the billing method can be a fee of 1 to 9. • If pay-as-you-go is selected, the payment can be in the range of 1 to 3. Default value: 1 **Note** This parameter is required when InstanceChargeType is set to Prepaid.
AutoPay	Boolean	Erased	Released	Specifies whether to enable automatic payment.	Valid values: • false: Automatic payment is disabled. After an order is generated, you must go to the Order Center to make the payment. • true: Automatic payment is enabled. Payments are automatically made. **Note** This parameter is required when InstanceChargeType is set to Prepaid.
Isp	String	Yes	Released	The ISP tag used for Finance Cloud. This parameter takes effect only when your region is set to China (Hangzhou).	This parameter is ignored if you are not a Finance Cloud user.

Response parameters

Fn::GetAtt

- EipAddress: the allocated Elastic IP address.
- AllocationId: the ID of the instance that the Elastic IP address is allocated to.
- OrderId: The order ID that is returned when you set the InstanceChargeType parameter to Prepaid.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Eip": {
      "Type": "ALIYUN::VPC::EIP",
      "Properties": {
        "InternetChargeType": "PayByTraffic",
          "Bandwidth": 200
      }
    }
  },
  "Outputs": {
    "EipAddress": {
      "Value" : {"Fn::GetAtt": ["Eip", "EipAddress"]}
    },
    "AllocationId": {
      "Value" : {"Fn::GetAtt": ["Eip", "AllocationId"]}
    },
    "OrderId": {
      "Value" : {"Fn::GetAtt": ["Eip", "OrderId"]}
    }
  }
}
```

5.1.6.9.2. ALIYUN::VPC::EIPAssociation

ALIYUN::VPC::EIPAssociation is used to associate an Elastic IP address with a cloud service instance.

Statement

```
{
  "Type": "ALIYUN::VPC::EIPAssociation",
  "Properties": {
    "AllocationId": String,
    "InstanceId": String,
    "PrivateIpAddress": String,
    "Mode": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
AllocationId	String	No	Yes	The ID of the Elastic IP address.	None

Parameter	Type	Required	Editable	Description	Constraint
InstanceId	String	No	Yes	The ID of the cloud service instance.	The following instance types are supported: • VPC-connected ECS instances • VPC-connected SLB instances • NAT gateways • HAVIP • Elastic network interfaces
PrivateIpAddress	String	Yes	True	The private IP address in the CIDR block of the VSwitch.	None
Mode	String	Yes	True	The association mode.	Valid values: • NAT • MULTI_BINDED

Response parameters

Fn::GetAtt

- EipAddress: The allocated Elastic IP address.
- AllocationId: The ID of the instance to which the Elastic IP address is allocated.

Examples

`JSON` format

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "Eip": {
      "Type": "ALIYUN::VPC::EIP",
      "Properties": {
        "InternetChargeType": "PayByTraffic",
          "Bandwidth": 200
      }
    },
    "EipAssociation": {
      "Type": "ALIYUN::VPC::EIPAssociation",
      "Properties": {
        "InstanceId": "<LoadBalancerId>",
        "InstanceType": "EcsInstance",
        "AllocationId": {
          "Fn::GetAtt": ["Eip", "AllocationId"]
        }
      }
    }
  },
  "Outputs": {
    "EipAddress": {
      "Value" : {"Fn::GetAtt": ["EipAssociation", "EipAddress"]}
    },
    "AllocationId": {
      "Value" : {"Fn::GetAtt": ["EipAssociation", "AllocationId"]}
    }
  }
}
```

`YAML` format

```
ROSTemplateFormatVersion: '2015-09-01'
Resources:
  Eip:
    Type: ALIYUN::VPC::EIP
    Properties:
      InternetChargeType: PayByTraffic
      Bandwidth: 200
  EipAssociation:
    Type: ALIYUN::VPC::EIPAssociation
    Properties:
      InstanceId: "<LoadBalancerId>"
      InstanceType: EcsInstance
      AllocationId:
        Fn::GetAtt:
        - Eip
        - AllocationId
Outputs:
  EipAddress:
    Value:
      Fn::GetAtt:
      - EipAssociation
      - EipAddress
  AllocationId:
    Value:
      Fn::GetAtt:
      - EipAssociation
      - AllocationId
```

5.1.6.9.3. ALIYUN::VPC::PeeringRouterInterfaceBinding

ALIYUN::VPC::PeeringRouterInterfaceBinding is used to associate two router interfaces to be interconnected.

Statement

```
{
  "Type": "ALIYUN::VPC::PeeringRouterInterfaceBinding",
  "Properties": {
    "OppositeRouterId": String,
    "OppositeInterfaceId": String,
    "OppositeInterfaceOwnerId": String,
    "RouterInterfaceId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
RouterInterfaceId	String	No	No	The ID of the router interface.	None
OppositeInterfaceId	String	No	No	The ID of the peer router interface.	None

Parameter	Type	Required	Editable	Description	Constraint
OppositeRouterId	String	Yes	Released	The ID of the router to which the peer router interface belongs.	None
OppositeInterfaceOwnerId	String	Yes	Released	The ID of the owner of the peer router interface.	None

Response parameters

Fn::GetAtt

RouterInterfaceId: the ID of the vRouter.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "InitiatorRouterInterfaceBinding": {
      "Type": "ALIYUN::VPC::PeeringRouterInterfaceBinding",
      "Properties": {
        "RouterInterfaceId": "ri-2zedgo0ih64g1me29****",
        "OppositeInterfaceId": "ri-2zex1tkyym98pjaor****",
        "OppositeRouterId": "vrt-2zexb35tzoriu0286****"
      }
    }
  }
}
```

5.1.6.9.4. ALIYUN::VPC::PeeringRouterInterfaceConnection

ALIYUN::VPC::PeeringRouterInterfaceConnection is used to initiate a router interface connection.

Statement

```
{
  "Type": "ALIYUN::VPC::PeeringRouterInterfaceConnection",
  "Properties": {
    "OppositeInterfaceId": String,
    "RouterInterfaceId": String
  }
}
```

Properties

Parameter	Type	Required	Editable	Description	Constraint
OppositeInterfaceId	String	No	No	The ID of the acceptor router interface.	None

Parameter	Type	Required	Editable	Description	Constraint
RouterInterfaceId	String	No	No	The ID of the router interface to initiate the connection.	None

Response parameters

Fn::GetAtt

- OppositeInterfaceId: the ID of the acceptor router interface.
- RouterInterfaceId: the ID of the router interface that initiates the connection.

Sample request

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "InitiatorRouterInterfaceBinding": {
      "Type": "ALIYUN::VPC::PeeringRouterInterfaceConnection",
      "Properties": {
        "RouterInterfaceId": "ri-2zedgo0ih64g1me29****",
        "OppositeInterfaceId": "ri-2ze4k5n2aeardu8cy****"
      }
    }
  }
}
```

5.1.6.9.5. ALIYUN::VPC::RouterInterface

ALIYUN::VPC::RouterInterface is used to create a router interface.

Syntax

```
{
  "Type": "ALIYUN::VPC::RouterInterface",
  "Properties": {
    "OppositeRegionId": String,
    "Description": String,
    "HealthCheckSourceIp": String,
    "RouterType": String,
    "AccessPointId": String,
    "RouterId": String,
    "Role": String,
    "OppositeInterfaceOwnerId": String,
    "OppositeAccessPointId": String,
    "HealthCheckTargetIp": String,
    "OppositeRouterId": String,
    "Spec": String,
    "OppositeRouterType": String,
    "Name": String,
    "PricingCycle": String,
    "Period": Number,
    "AutoPay": Boolean,
    "InstanceChargeType": String
  }
}
```

Properties

Property	Type	Required	Editable	Description	Constraint
RouterId	String	Yes	No	The ID of the router	None
Role	String	Yes	No	The role of the router interface.	• When RouterType is set to VBR, set the value to InitiatingSide. • When OppositeRouterType is set to VBR, set the value to AcceptingSide.
RouterType	String	No	No	The type of the router to which the router interface belongs.	Valid values: • VRouter • VBR
AccessPointId	String	No	No	The ID of the access point of the router interface.	• This parameter is required when RouterType is set to VBR. The access point ID cannot be modified after the router interface is created. • This parameter is not required when RouterType is set to VRouter.

Property	Type	Required	Editable	Description	Constraint
Spec	String	No	No	The specifications of the router interface.	The following list includes available specifications and the corresponding bandwidth values: • Mini.2: 2 Mbit/s • Mini.5: 5 Mbit/s • Small.1: 10 Mbit/s • Small.2: 20 Mbit/s • Small.5: 50 Mbit/s • Middle.1: 100 Mbit/s • Middle.2: 200 Mbit/s • Middle.5: 500 Mbit/s • Large.1: 1,000 Mbit/s • Large.2: 2,000 Mbit/s • Large.5: 5,000 Mbit/s • Xlarge.1: 10,000 Mbit/s Note • This parameter is required when Role is set to InitiatingSide. • The value Negative is used by default when Role is set to Accepting Side.
OppositeRegionId	String	No	No	The region ID of the peer router interface.	None
OppositeInterfaceOwnerId	String	No	No	The ID of the owner of the peer router interface.	The default value is the ID of the current user.
OppositeRouterId	String	No	No	The ID of the router to which the peer router interface belongs.	None

Property	Type	Required	Editable	Description	Constraint
OppositeRouterType	String	No	No	The type of the router to which the peer router interface belongs.	Valid values: • When RouterType is set to VBR, set the value to VRouter. • VBR
OppositeAccessPointId	String	No	No	The ID of the access point of the peer router interface.	• When OppositeRouterType is set to VBR, this parameter is required . The access point ID cannot be modified after the router interface is created. • When OppositeRouterType is set to VRouter, this parameter is not required. • When OppositeRouterType is set to VBR, the VBR specified by the OppositeRouterId parameter must be in the access point specified by the OppositeAccessPointId parameter.
Description	String	No	No	The description of the router interface.	The description must be 2 to 256 characters in length. It cannot start with `http://` or `https://` . The parameter is empty by default.
Name	String	No	No	The display name of the router interface.	• The name must be 2 to 128 characters in length and can contain letters, digits, periods(.), underscores (_), and hyphens (-). • It must start with a letter but cannot start with `http://` or `https://` .

Property	Type	Required	Editable	Description	Constraint
HealthCheckSourceIp	String	No	No	The source IP address of health check packets used in leased line disaster recovery and ECMP scenarios.	This parameter is valid only for VRouter interfaces with a peer router interface on a VBR. It must be an unused IP address in the VPC where the local VRouter is located. The HealthCheckSourceIp and HealthCheckTargetIp parameters must either both be specified or both left unspecified.
HealthCheckTargetIp	String	No	No	The destination IP address of health check packets used in leased line disaster recovery and ECMP scenarios.	This parameter is valid only for VRouter interfaces with a peer router interface on a VBR. Typically, you can use the IP address of a customer premises equipment (CPE) on the user side of the leased line, which is the IP address of the peer gateway on the VBR where the peer router interface is located. You can also specify another IP address on the user side of the leased line as the destination IP address. The HealthCheckSourceIp and HealthCheckTargetIp parameters must either both be specified or both left unspecified.

Property	Type	Required	Editable	Description	Constraint
PricingCycle	String	No	No	The billing cycle of the subscription.	Valid values: • Month • Year
Period	Number	No	No	The subscription duration.	• Valid values when the PricingCycle parameter is set to Month: 1 to 9. • Valid values when the PricingCycle parameter is set to Year: 1 to 3.
AutoPay	Boolean	No	No	Specifies whether to enable automatic payment.	Default value: false. Valid values: • true • false
InstanceChargeType	String	No	No	The billing method of the instance.	Valid values: • Postpaid: pay-as-you-go • Prepaid: subscription

Response parameters

Fn::GetAtt

RouterInterfaceId: the ID of the router interface.

Examples

```
{
  "ROSTemplateFormatVersion": "2015-09-01",
  "Resources": {
    "RouterInterface": {
      "Type": "ALIYUN::VPC::RouterInterface",
      "Properties": {
        "Name": "RouterInterface_1",
        "Description": "VPC initiator RouterInterface",
        "RouterId": "vrt-2ze2i147e5n0bicoe****",
        "Role": "AcceptingSide",
        "OppositeRegionId": "cn-beijing",
        "HealthCheckSourceIp": "10.0.XX.XX",
        "HealthCheckTargetIp": "192.168.XX.XX"
      }
    }
  },
  "Outputs": {
    "RouterInterfaceId": {
      "Value": {"Fn::GetAtt": ["RouterInterface", "RouterInterfaceId"]}
    }
  }
}
```

6.Object Storage Service (OSS)

6.1. User Guide

6.1.1. What is OSS?

Object Storage Service (OSS) is a secure, cost-effective, and highly reliable cloud storage service provided by Alibaba Cloud.

Compared with user-created server storage, OSS has outstanding advantages in reliability, security, cost-effectiveness, and data processing capabilities. OSS enables you to store and retrieve a variety of unstructured data objects, such as text, images, audios, and videos over networks anytime.

OSS is an object storage service based on key-value pairs. Files uploaded to OSS are stored as objects in buckets. You can obtain the content of an object based on the object key.

In OSS, you can perform the following operations:

- Create a bucket and upload objects to the bucket.
- Obtain an object URL from OSS to share or download the object.
- Modify the attributes or metadata of a bucket or an object. You can also configure the access control list (ACL) of the bucket or the object.
- Perform basic and advanced operations in the OSS console.
- Perform basic and advanced operations by using OSS SDKs or calling RESTful API operations in your application.

6.1.2. Usage notes

Before you use OSS, you must understand the following content:

To allow other users to use all or part of OSS features, you must create RAM users and grant permissions to the users by configuring RAM policies.

Before you use OSS, you must also understand the following limits.

Item	Limit
Bucket	• You can create up to 100 buckets. • After a bucket is created, its name and region cannot be modified.
Upload objects	• Objects larger than 5 GB cannot be uploaded by using the following modes: console upload, simple upload, form upload, or append upload. To upload an object that is larger than 5 GB, you must use multipart upload. The size of an object uploaded by using multipart upload cannot exceed 48.8 TB. • If you upload an object that has the same name of an existing object in OSS, the new object will overwrite the existing object.
Delete objects	• Deleted objects cannot be recovered. • You can delete up to 100 objects at a time in the OSS console. To delete more than 100 objects at a time, you must call an API operation or use an SDK.
Lifecycle	You can configure up to 1,000 lifecycle rules for each bucket.

6.1.3. Quick start

6.1.3.1. Log on to the OSS console

This topic describes how to log on to the Object Storage Service (OSS) console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation pane, choose **Products > Object Storage Service**.

6.1.3.2. Create buckets

Objects uploaded to OSS are stored in a bucket. You must create a bucket before you upload objects to OSS.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click **Create Bucket**.

3. On the **Create OSS Bucket** page, configure parameters.

   The following table describes the parameters that you can configure.

Parameter	Description
Organization	Select an organization from the drop-down list for the bucket.
Resource Set	Select a resource set from the drop-down list for the bucket.
Region	Select a region from the drop-down list for the bucket. **Note** ◦ The region of a bucket cannot be changed after the bucket is created. ◦ If you want to access OSS from your ECS instance through the internal network, select the same region where your ECS instance is deployed.
Cluster	Select a cluster for the bucket.
Bucket Name	Enter the name of the bucket. **Note** ◦ The bucket name must comply with the naming conventions. ◦ The bucket name must be globally unique among all existing buckets in OSS. ◦ The bucket name cannot be changed after the bucket is created.
Storage Class	Set the value to **Standard**. Only Standard is supported.
Bucket Capacity	Specify the capacity of the bucket. Valid values: 0 to 2000000. Unit: TB or GB.
Access Control List (ACL)	Set the ACL of the bucket. You can select the following options: ◦ **Private**: Only the owner or authorized users of this bucket can read and write objects in the bucket. Other users, including anonymous users cannot access objects in the bucket without authorization. ◦ **Public Read**: Only the owner or authorized users of this bucket can read and write objects in the bucket. Other users, including anonymous users can only read objects in the bucket. ◦ **Public Read/Write**: All users, including anonymous users can read and write objects in the bucket. Fees incurred by such operations are paid by the owner of the bucket. Exercise caution when you configure this option. **Note** You can modify the ACL of a bucket after the bucket is created. For more information, see Modify bucket ACLs.

Parameter	Description
Server-Side Encryption	Configure server-side encryption for the bucket. You can select the following options: ○ **None**: Server-side encryption is not performed. ○ **AES256**: AES256 is used to encrypt each object in the bucket using different data keys. Customer master keys (CMKs) used to encrypt the data keys are rotated regularly. ○ **KMS**: CMKs managed by KMS are sued to encrypt objects in the bucket.
Encryption Algorithm	You can configure this parameter when you select **KMS** for **Server-Side Encryption**.
Key ID	You can configure this parameter when you select **KMS** for **Server-Side encryption**. OSS uses the specified CMK to encrypt objects in the bucket. **Note** To select a CMK ID for server-side encryption, you must create the CMK in the KMS console.

4. Click **Submit**.

6.1.3.3. Upload objects

After you create a bucket, you can upload objects to it.

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

You can upload an object of any format to a bucket. You can use the OSS console to upload an object up to 5 GB in size. To upload an object larger than 5 GB, use OSS SDKs or call an API operation.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. On the bucket details page that appears, click **Files**.
4. Click **Upload**.
5. In the **Upload** panel, set the parameters described in the following table.

Parameter	Description
Upload To	Set the directory to which you want to upload objects. ○ **Current**: Objects are uploaded to the current directory. ○ **Specified**: Objects are uploaded to the specified directory. You must enter the directory name. If the specified directory does not exist, OSS automatically creates the specified directory and uploads the object to the directory.

Parameter	Description
File ACL	Set the access control list (ACL) of the object to upload. Default value: **Inherited from Bucket**. ○ **Inherited from Bucket**: The ACL of uploaded objects is the same as that of the bucket. ○ **Private**: Only the owner or authorized users can read and write objects in the bucket. Other users, which includes anonymous users, cannot access the objects in the bucket without authorization. ○ **Public Read**: Only the bucket owner can perform write operations on objects in the bucket. Other users, which includes anonymous users, can perform only read operations on objects in the bucket. ○ **Public Read/Write**: All users, which includes anonymous users, can read and write objects in the bucket.
Upload	Drag one or more files to upload to this section, or click **Upload** to select one or more files to upload. **Notice** ○ When the object to upload has the same name as an existing object in the bucket, the existing object is overwritten. ○ If you upload a directory, only the files in the directory are uploaded, and the files are stored in the same directory in the bucket. ○ The name of an uploaded object must comply with the following conventions: ■ The name must be encoded in UTF-8. ■ The name is case-sensitive. ■ The name must be 1 to 1,023 bytes in length. ■ The name cannot start with a forward slash (/) or backslash (\).

6. In the **Upload Tasks** panel, wait until the upload task is completed.

   During the upload process, you can click **Cancel All** to cancel the task. After the task is completed, click **Removed** to remove the task.

   Notice Do not refresh or close the **Upload Tasks** panel when objects are being uploaded. Otherwise, the upload tasks are interrupted.

6.1.3.4. Obtain object URLs

You can obtain the URL of an uploaded object in Object Storage Service (OSS) and share the URL with other users to preview or download the object.

Prerequisites

An object is uploaded to the bucket. For more information, see Upload objects.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.

3. Click the **Files** tab.
4. Obtain object URLs
   - Obtain the URL of a single object.

     Click the name of the object whose URL you want to obtain, or click **View Details** in the Actions column that corresponds to the object. In the **View Details** panel, click **Copy File URL**.
   - Batch export URL lists

     Select the objects that you want to share. Choose **Batch Operation > Export URL List**.

6.1.4. Buckets

6.1.4.1. View bucket information

You can view the detailed information about created buckets in the Object Storage Service (OSS) console.

Prerequisites

A bucket is created. For more information, see Create buckets.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. On the **Overview** tab, you can view the information about the bucket, which includes Organization and Resource Set, Domain Names, and Basic Settings.

6.1.4.2. Delete buckets

You can delete a bucket in the Object Storage Service (OSS) console.

Prerequisites

All objects and parts in the bucket are deleted. For more information, see Delete objects and Manage parts.

> **Warning** Deleted objects, parts, and buckets cannot be recovered. Exercise caution when you delete objects, parts, and buckets.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. In the upper-right corner, click **Delete Bucket**.
4. In the message that appears, click **OK**.

6.1.4.3. Modify bucket ACLs

Object Storage Service (OSS) provides access control list (ACL) to control access to buckets. By default, the ACL of a bucket is private. You can modify the ACL of a bucket after the bucket is created.

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

You can set the ACL of a bucket to one of the following values:

- Private: Only the owner or authorized users of this bucket can read and write objects in the bucket. Other users, including anonymous users, cannot access the objects in the bucket without authorization.
- Public Read: Only the owner or authorized users of this bucket can write objects in the bucket. Other users, including anonymous users, can only read the objects in the bucket.
- Public Read/Write: Any users, including anonymous users, can read and write the objects in the bucket.

> **Warning** If you set the ACL of a bucket to Public Read or Public Read/Write, other users can read the data in the bucket without authentication, which may result in security risks. To ensure the security of your data, we recommend that you set the ACL of your bucket to private.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. Click the **Basic Settings** tab. Find the **Access Control List (ACL)** section.
4. Click **Configure**. Modify the bucket ACL.
5. Click **Save**.

6.1.4.4. Configure static website hosting

You can configure static website hosting for a bucket in the Object Storage Service (OSS) console so that users can access the website by using the domain name of the bucket.

Prerequisites

A bucket is created. For more information, see Create buckets.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. Click the **Basic Settings** tab. Find the **Static Pages** section.
4. Click **Configure** and then set the parameters described in the following table.

Parameter	Description
Default Homepage	Specify an index page that functions similar to index.html. Only HTML objects can be specified as the index page. If you do not specify this parameter, static website hosting is disabled.
Default 404 Page	Set the default 404 page that is displayed when the requested resource does not exist. Only HTML, JPG, PNG, BMP, or WebP objects in the root directory of the bucket can be set to the default 404 page. If you do not specify this parameter, Default 404 Page is disabled.

5. Click **Save**.

6.1.4.5. Configure hotlink protection

You can configure hotlink protection for a bucket in the Object Storage Service (OSS) console to prevent data in your bucket from being accessed by unauthorized domain names.

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

The hotlink protection feature allows you to configure a Referer whitelist for a bucket. This way, only requests from domain names included in the Referer whitelist can access your data in the bucket. OSS allows you to configure Referer whitelists based on the Referer header field in HTTP or HTTPS requests.

After hotlink protection is configured for a bucket, OSS verifies requests to objects in the bucket only when the requests are initiated by using signed URLs or from anonymous users. Requests that contain the Authorization field in the header are not verified.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. Click the **Basic Settings** tab. Find the **Hotlink Protection** section.
4. Click **Configure** and configure the following parameters:
   - Enter domain names or IP addresses in the **Referer Whitelist** field. Separate multiple Referers by using line feed. You can use asterisks (*) and question marks (?) as wildcards. Examples:
     - If you add www.example.com to the Referer whitelist, requests sent from URLs that start with *www.example.com*, such as *www.example.com/123* and *www.example.com.cn* are allowed.
     - If you add *www.example.com/ to the Referer whitelist, requests sent from *http://www.example.com/* and *https://www.example.com/* are allowed.
     - An asterisk (*) can be used as a wildcard to indicate zero or more characters. For example, if you add *.example.com to the Referer whitelist, requests sent from URLs such as *help.example.com* and *www.example.com* are allowed.
     - A question mark (?) can be used as a wildcard to indicate a single character. For example, if you add example?.com to the Referer whitelist, requests sent from URLs such as *examplea.com* and *exampleb.com* are allowed.
     - You can add domain names or IP addresses that include a port number, such as *www.example.com:8080* and *10.10.10.10:8080*, to the Referer whitelist.
   - Select whether to turn on **Allow Empty Referer** to allow requests in which the Referer field is empty.

     An HTTP or HTTPS request with an empty Referer field indicates that the request does not contain the Referer field or the value of the Referer field is empty.

     If you do not allow empty Referers fields, only HTTP or HTTPS requests which include an allowed Referer field can access the objects in the bucket.

   > **Note** By default, if you use the bucket endpoint to preview an MP4 object, the browser sends a request that contains the Referer field and a request that does not contain the Referer field at the same time. Therefore, to allow access to the MP4 objects in your bucket, you must not only add the bucket endpoint to the Referer whitelist but also allow empty Referer fields. To preview a non-MP4 object by using the bucket domain name, you need only to allow empty Referer fields.

5. Click **Save**.

6.1.4.6. Configure logging

When you access Object Storage Service (OSS), a large number of access logs are generated. You can use the logging feature to store OSS access logs in a specified bucket.

Prerequisites

A bucket is created. For more information, see Create buckets.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. Click the **Basic Settings** tab. Find the **Logging** section.
4. Click **Configure**. Turn on the **Logging** switch. Select **Destination Bucket** and set **Log Prefix**.
   - **Destination Bucket**: Select the bucket in which you want to store access logs from the drop-down list. You must be the owner of the selected bucket, and the selected bucket must be in the same region as the bucket for which logging is enabled.
   - **Log Prefix**: Enter the prefix and directory where the access logs are stored. If you specify *log/<TargetPrefix>* as the prefix, access logs are stored in the *log/* directory.
5. Click **Save**.

6.1.4.7. Configure CORS

You can configure cross-origin resource sharing (CORS) in the Object Storage Service (OSS) console to enable cross-origin access.

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

OSS provides CORS over HTML5 to implement cross-origin access. When OSS receives a cross-origin request (or an OPTIONS request) for a bucket, OSS reads the CORS rules of the bucket and checks the relevant permissions of the request. OSS matches the request with the rules one by one. When OSS finds the first match, OSS returns a corresponding header in the response. If no match is found, OSS does not include any CORS header in the response.

Procedure

1. Log on to the OSS console.
2. Click Buckets. On the Buckets page that appears, click the name of the bucket to which you want to upload objects.
3. Click the **Basic Settings** tab. In the **Cross-Origin Resource Sharing (CORS)** section, click **Configure**.
4. On the page that appears, click **Create Rule**. Then, in the **Create Rule** panel, configure the parameters in the following table.

Parameter	Required	Description

Parameter	Required	Description
Sources	Yes	Specify the sources from which you want to allow cross-origin requests. Take note of the following items when you configure the sources: ○ You can configure multiple rules for sources. Separate multiple rules with line feeds. ○ The domain names must include the protocol name, such as HTTP or HTTPS. ○ Asterisks (*) are supported as wildcards. Each rule can contain up to one asterisk (*). ○ A domain name must include the port number if the domain name does not use the default port. Example: https://www.example.com:8080. The following examples show how to configure domain names: ○ To match a specified domain name, enter the full domain name. Example: https://www.example.com. ○ Use an asterisk (*) as a wildcard in the domain name to match second-level domains. Example: https://*.example.com. ○ Enter only an asterisk (*) as the wildcard to match all domain names.
Allowed Methods	Yes	Select the cross-origin request methods that are allowed.
Allowed Headers	No	Specify the response headers for the allowed cross-origin requests. Take note of the following rules when you configure the allowed response headers: ○ This parameter is in the key:value format and case-insensitive. Example: content-type:text/plain. ○ You can configure multiple rules for allowed headers. Separate multiple rules with new lines. ○ Each rule can contain up to one asterisk (*) as the wildcard. We recommend that you set this parameter to an asterisk (*) if you do not have special requirements.
Exposed Headers	No	Specify the response headers for allowed access requests from applications, such as an XMLHttpRequest object in JavaScript. Exposed headers cannot include asterisks (*).
Cache Timeout (Seconds)	No	Specify the period of time within which the browser can cache the response for an OPTIONS preflight request to a specific resource. Unit: seconds.

Note You can configure up to 10 CORS rules for each bucket.

5. Click **OK**.

6.1.4.8. Configure lifecycle rules

You can configure a lifecycle rule for a bucket to regularly delete expired objects and parts to save storage costs.

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

Take note of the following items when you configure lifecycle rules for a bucket:

- After a lifecycle rule is configured, it is loaded within 24 hours and takes effect within 24 hours after it is loaded. Check the configurations of a rule before you save the rule.
- Objects that are deleted based on lifecycle rules cannot be recovered. Configure lifecycle rules based on your requirements.
- You can configure up to 100 lifecycle rules for each bucket in the Object Storage Service (OSS) console and up to 1,000 lifecycle rules for each bucket by using ossutil.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure lifecycle rules.
3. Click the **Basic Settings** tab. Find the **Lifecycle** section. Click **Configure**.
4. Click **Create Rule**. In the **Create Rule** panel that appears, configure the parameters described in the following table.

Parameter	Description
Status	Specify the status of the lifecycle rule. Valid values: **Enabled** and **Disabled**.
Applied To	Select objects to which the rule applies. You can select **Files with Specified Prefix** or **Whole Bucket**. Files with Specified Prefix indicates that this rule applies to objects whose names contain a specified prefix. Whole Bucket indicates that this rule applies to all objects in the bucket. **Note** If you select **Files with Specified Prefix**, you can configure different lifecycle rules for objects whose names contain different prefixes. If you select **Whole Bucket**, only one lifecycle rule can be configured for the bucket.
Prefix	If you set **Applied To** to **Files with Specified Prefix**, you must specify the prefix of the objects to which the rule applies. For example, if you want that the rule applies to objects whose names start with img, enter *img*.
File Lifecycle	Configure rules for objects to specify when objects expire. You can set File Lifecycle to **Validity Period (Days)**, **Expiration Date**, or **Disabled**. If you select **Disabled**, the configurations of File Lifecycle do not take effect.

Parameter	Description
Delete	Specify when objects expire based on **Validity Period (Days)** or **Expiration Date** that you set for **File Lifecycle**. After objects expire, the objects are deleted. ○ **Validity Period (Days)**: Specify the number of days to retain objects after they are last modified. Select the check box next to Delete and specify a number such as N. The objects expire N days after they are last modified. Then, the objects are deleted the next day after they expire. For example, if you specify the number as 30, objects that are last modified on January 1,2019 are deleted on February 1, 2019. ○ **Expiration Date**: Specify the expiration date. Select the check box next to Delete and set an expiration date. The objects that are last modified before this date expire and are deleted. For example, if you set Expiration Date to January 1, 2019, objects that are last modified before January 1, 2019 are deleted.
Part Lifecycle	Specify the operations to perform on expired parts. You can set Part Lifecycle to **Validity Period (Days)**, **Expiration Date**, or **Disabled**. If you select **Disabled**, the configurations of Part Lifecycle do not take effect. **Notice** You must configure at least one of **File Lifecycle** and **Part Lifecycle**.
Delete Parts	Specify when parts that match the rule expire based on **Validity Period (Days)** or **Expiration Data** that you set for **Part Lifecycle**. Expired parts are deleted. You can configure this parameter in the same way as you configure the Delete parameter in Clear Policy.

5. Click **OK**.

6.1.4.9. Configure storage quota

If the capacity of a bucket reaches the specified storage quota, write operations such as PutObject, MultipartUpload, CopyObject, PostObject, and AppendObject cannot be performed on the bucket. This topic describes how to configure the storage quota of a bucket in Object Storage Service (OSS).

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

Take note of the following items when you configure the storage quota of a bucket:

- Before you configure the storage quota of a bucket, make sure that the quota does not limit your business because write operations cannot be performed if the bucket capacity reaches the specified quota.
- In general, it takes about an hour for OSS to determine whether the bucket capacity exceeds the storage quota. In some cases, it can take longer than one hour.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure storage quota.
3. Click the **Basic Settings** tab and find the **Storage Quota** section.
4. Click **Configure**.
5. On the **Modify OSS Bucket** page, modify the storage quota of the bucket.

- Units: TB or GB.
- Valid values: -1 to 2000000

  The default value is -1, which indicates that the bucket capacity is not limited.

6. Click **Submit**.

   After you submitted, you can click **Back to Console** in the pop-up dialog box to go back to the **Overview** page.

6.1.4.10. Configure cluster-disaster recovery

In cluster-disaster recovery mode, buckets with the same name are replicated. Cluster-based disaster recovery is automatically enabled based on configurations made when the cluster is created. In other words, after a primary bucket is created, a secondary bucket with the same name is automatically created. Information stored in the primary bucket is automatically synchronized to the secondary bucket. By default, Cluster-disaster Recovery is turned on for buckets that are created by using the Object Storage Service (OSS) console.

Prerequisites

A bucket is created. For more information about how to create a bucket, see Create buckets.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click Buckets. On the Buckets page, click the name of the bucket to go to the bucket details page.
3. On the bucket details page, click the **Basic Settings** tab. Find the **Cluster-disaster Recovery** section.
4. Click **Configure**. Turn on or turn off **Cluster-disaster Recovery**.
5. Click **Save**.

6.1.4.11. Bucket tagging

Object Storage Service (OSS) allows you to configure bucket tagging to classify and manage buckets. For example, you can use this feature to list buckets that have specific tags and configure access control lists (ACLs) for buckets that have specific tags.

Context

The bucket tagging feature uses a key-value pair to identify a bucket. You can add tags to buckets that are used for different purposes and manage the buckets by tags.

- Only the bucket owner or authorized Resource Access Management (RAM) users can configure tagging for the bucket. Otherwise, 403 Forbidden is returned with the AccessDenied error code.
- You can configure up to 20 tags for a bucket.
- The tag key is required. The tag key can be up to 64 Bytes in length and cannot start with `http://`, `https://`, or `Aliyun`.
- The tag value is optional. The tag value can be up to 128 bytes in length. You can leave the parameter empty.
- The key and value of a tag must be encoded in UTF-8.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure tagging.
3. Choose **Basic Settings > Bucket Tagging**.
4. Click **Configure**.

5. Add tags to the bucket based on the naming conventions. You can click the + icon to add multiple tags to a bucket.
6. Click **Save**.

6.1.4.12. Configure server-side encryption

OSS supports server-side encryption. When you upload an object to a bucket for which server-side encryption is enabled, OSS encrypts the object and stores the encrypted object. When you download the encrypted object from OSS, OSS automatically decrypts the object and returns the decrypted object to you. A header is added in the response to indicate that the object is encrypted on the OSS server.

Context

OSS supports the following encryption methods:

- Server-side encryption by using KMS (SSE-KMS)

  OSS uses the default customer master key (CMK) managed by KMS or a specified CMK to encrypt objects. The CMK is managed by KMS to ensure confidentiality, integrity, and availability at minimal costs.

- Server-side encryption by using OSS-managed keys (SSE-OSS)

  OSS uses data keys to encrypt objects and manages the data keys. In addition, OSS uses master keys that are regularly rotated to encrypt data keys.

You can enable server-side encryption in the OSS console by using one of the following methods:

- Method 1: Enable server-side encryption when you create a bucket
- Method 2: Enable server-side encryption on the Basic Settings tab

Method 1: Enable server-side encryption when you create a bucket

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click **Create Bucket**.
3. On the **Create OSS Bucket** page, set parameters.

   You can set the following parameters to configure server-side encryption for the bucket.

   - **Server-Side Encryption**: Specify the encryption methods.
     - **None**: Server-side encryption is not performed.
     - **AES256**: AES256 is used to encrypt each object in the bucket by using different data keys. The CMKs used to encrypt the data keys are rotated regularly.
     - **SM4**: SM4 is used to encrypt each object in the bucket by using different data keys. The CMKs used to encrypt the data keys are rotated regularly.
     - **KMS**: CMKs managed by KMS are used to encrypt objects in the bucket.
   - **Encryption Algorithm**: This parameter can be configured when you select **KMS** for **Server-Side Encryption**. You can select **SM4** or **AES256**.
   - **Key ID**: This parameter can be configured when you select **KMS** for **Server-Side Encryption**. OSS uses the specified CMK to encrypt objects in the bucket.

     > **Note** To select a CMK ID for server-side encryption, you must create the CMK in the KMS console.

4. Click **Submit**.

Method 2: Enable server-side encryption on the Basic Settings tab

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which

you want to configure server-side encryption.

3. Click the **Basic Settings** tab. Find the **Server-side Encryption** section.
4. Click **Configure** and set the following parameters:
   - **Encryption Method**: Specify the encryption method.
     - **None**: Server-side encryption is not performed.
     - **OSS-Managed**: Keys managed by OSS are used to encrypt your data.
     - **KMS**: CMKs managed by KMS are used to encrypt objects in the bucket.
   - **Encryption Algorithm**: You can select **SM4** or **AES256**.
     - **AES256**: AES256 is used to encrypt each object in the bucket by using different data keys. CMKs used to encrypt the data keys are rotated regularly.
     - **SM4**: SM4 is used to encrypt each object in the bucket by using different data keys. CMKs used to encrypt the data keys are rotated regularly.
   - **CMK**: This parameter can be configured when you select **KMS** for **Encryption Method**. OSS uses the specified CMK to encrypt objects in the bucket.

     Note To select a CMK ID for server-side encryption, you must create the CMK in the KMS console.
5. Click **Save**.

Notice The configurations of the default encryption method for a bucket do not affect the encryption configurations of existing objects within the bucket.

6.1.4.13. Bind a bucket to a VPC network

You can bind your bucket to a specified virtual private cloud (VPC) network to allow only requests from IP addresses within the VPC network to access your bucket.

Prerequisites

A VPC network is created. For more information, see the "Create a VPC" chapter of *VPC User Guide*.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket that you want to bind to the VPC network.
3. Click the **Overview** tab. Click **Bind VPC** in the **VPC Info** section.
4. On the **Bind VPC** page, select the VPC network that you create.

   You can also click **Create VPC** to create a new VPC network.
5. Click **Submit**.

6.1.4.14. Configure CRR

Cross-region replication (CRR) allows you to perform automatic and asynchronous (near real-time) replication on objects across buckets that are located in different Object Storage Service (OSS) regions. If you enable CRR, operations such as the creation, overwriting, and deletion of objects can be synchronized from the source bucket to the destination bucket.

Prerequisites

The source bucket and destination bucket are created. For more information, see Create buckets.

Context

CRR meets the requirements of geo-disaster recovery or data replication. Objects in the destination bucket are extra duplicates of objects in the source bucket. They have the same names, content, and metadata, such as the created time, owner, user metadata, and object access control list (ACL).

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure CRR.
3. On the bucket details page, click the **Basic Settings** tab. Find the **Cross-Region Replication** section.
4. Click **Enable**. In the **Cross-Region Replication** panel, configure the parameters described in the following table.

Parameter	Description
Source Region	The region in which the current bucket is located.
Source Bucket	The name of the current bucket.
Destination Region	Select the region in which the destination bucket is located. The source and destination buckets for CRR must be located in different regions. Data cannot be synchronized between buckets that are located within the same region.
Destination Bucket	Select the destination bucket to which data is synchronized. The source bucket and destination bucket specified in a CRR rule are not allowed to synchronize data with other buckets. For example, if you configure a CRR rule to synchronize data from Bucket A to Bucket B, Bucket A and Bucket B are not allowed to synchronize data with other buckets.
Applied To	Select the source data that you want to synchronize. ○ **All Files in Source Bucket**: All objects within the source bucket are synchronized to the destination bucket. ○ **Files with Specified Prefix**: Only objects whose names contain one of the specified prefixes are synchronized to the destination bucket. For example, if you have a directory named *management* in the root directory of a bucket and want to synchronize objects in a subdirectory named *abc* in *management*, you can enter the prefix *management/abc*. You can specify up to 10 prefixes.
Operations	Select the synchronization policy. ○ **Add/Change**: Only the added or updated data is synchronized from the source bucket to the destination bucket. ○ **Add/Delete/Change**: All changes to data including the creation, modification, and deletion of objects are synchronized from the source bucket to the destination bucket.

Parameter	Description
Replicate Historical Data	Specify whether to synchronize historical data that exists before you enable CRR for the source bucket. ○ **Yes**: OSS synchronizes historical data to the destination bucket. **Notice** When historical data is synchronized, objects in the destination bucket, which have the same name as objects from the source bucket, may be overwritten. ○ **No**: Only objects that are uploaded or updated after CRR is enabled are synchronized to the destination bucket.

5. Click **OK**.

Note
- It takes 3 to 5 minutes for a CRR rule to take effect after the rule is configured. Synchronization information is displayed after the source data is synchronized to the destination bucket.
- In CRR, data is asynchronously replicated in near real-time. It may take a few minutes to several hours to replicate data to the destination bucket based on the data amount.

6.1.4.15. Configure cross-cloud replication

You can use the cross-cloud replication feature to synchronize Object Storage Service (OSS) data between two clouds. This topic describes how to configure cross-cloud replication.

Step 1: Obtain the parameters of the destination cloud.

Before you configure cross-cloud replication, you must obtain the required parameters of the destination cloud.

1. Log on to the Apsara Uni-manager Operations Console of the destination cloud.

   For more information about how to log on to the Apsara Uni-manager Operations console, see *Log on to the Apsara Uni-manager Operations console* in *Operations and Maintenance Guide*.
2. In the left-side navigation pane, choose **Product Management > Products**.
3. Click **OSS O&M**.
4. In the left-side navigation pane, choose **Service O&M - OSS > Synchronization Management > Cross-Cloud Synchronization**.
5. In the upper-right corner, select the destination cluster from the **Cluster** drop-down list, and then click **View Parameters of the Current Cloud**.

   Record the information displayed in the **Parameters of the Current Cloud** dialog box.

Step 2: Configure cross-cloud synchronization for the source cloud.

After you obtain the required parameters of the destination cloud, you must configure cross-cloud synchronization for the source cloud in the Apsara Uni-manage Operations Console.

1. Log on to the Apsara Uni-manager Operations Console of the source cloud.
2. In the left-side navigation pane, choose **Product Management > Products**.
3. Click **OSS O&M**.
4. In the left-side navigation pane, choose **Service O&M - OSS > Synchronization Management > Cross-Cloud Synchronization**.
5. In the upper-right corner, click **Create**. In the **Create Cross-Cloud Synchronization Task** dialog box, add

the obtained parameters of the destination cloud.

6. Click **Submit**.

   Wait a few minutes until the cross-synchronization configurations take effect.

Step 3: Configure cross-cloud replication in the Apsara Uni-manager Management Console of the source cloud.

After the cross-cloud synchronization configurations take effect, you must configure cross-cloud replication in the Apsara Uni-manager Management Console.

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure cross-cloud replication.
3. On the bucket details page, click the **Basic Settings** tab. In the **Cross-Cloud Replication** section, click **Enable**.
4. In the **Cross-Cloud Replication** panel, configure the parameters described in the following table.

Parameter	Description
Source Region	The region in which the source bucket is located.
Source Bucket	The name of the source bucket.
Destination Cloud	Enter the name of the destination cloud obtained in Step 1.
Destination Cloud Address	Enter the value of Location of the destination cloud obtained in Step 1.
Destination Bucket	Enter the name of the destination bucket.
Applied To	Select the objects that you want to synchronize. ◦ **All Files in Source Bucket**: All objects within the source bucket are synchronized to the destination bucket. ◦ **Files with Specified Prefix**: Only objects whose names contain one of the specified prefixes are synchronized to the destination bucket. Click **Add**. You can add up to 10 prefixes.
Operations	Select a synchronization policy. ◦ **Add/Change**: Only newly added and changed data is synchronized from the source bucket to the destination bucket. ◦ **Add/Delete/Change**: All changes to data including creation, modification, and deletion of objects are synchronized from the source bucket to the destination bucket.

Parameter	Description
Replicate Historical Data	Specify whether to synchronize historical data that is generated before you enable cross-cloud replication. ○ **Yes**: Historical data is synchronized to the destination bucket. ○ **No**: Only objects that are uploaded or updated after cross-cloud replication is enabled are synchronized to the destination bucket.

5. Click **OK** to save your settings.

6.1.4.16. IMG

6.1.4.16.1. Configure image styles

You can encapsulate multiple Image Processing (IMG) parameters in a style and perform complex IMG operations by using the style.

Context

Up to 50 styles can be created for a bucket. These styles can be used only for image objects in the bucket.

Create a style

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to create styles.
3. Click the **Image Processing (IMG)** tab. On the page that appears, click **Create Rule**.
4. In the **Create Rule** panel, configure the style.

   You can use **Basic Edit** or **Advanced Edit** to create a style:

   ○ **Basic Edit**: You can use the IMG features by using the graphical user interface (GUI). For example, resize an image, add a watermark, and modify the image format.

   ○ **Advanced Edit**: You can use the API code to edit the IMG features that you want to use to process an image. The format is: `image/action1,parame_value1/action2,parame_value2/...` .

   Example: `image/resize,p_63/quality,q_90` indicates that the image is scaled down to 63% of the source image, and then the relative quality of the image is set to 90%.

   > **Note** If you want to add image and text watermarks to images at the same time by using a style, use **Advanced Edit** to create the style.

5. Click **OK**.

   After you create a style, you can click **Export** to export the style to a specified local path.

Apply styles

You can perform the following steps to use a created style to process an image object in the current bucket:

1. On the Overview page, click the **Files** tab.
2. Click the name of the image that you want to process.
3. In the **View Details** panel, select an image style from the **Image Style** drop-down list.

   You can view the processed image in the **View Details** panel. Right-click the image and click **Save As** to save

the image to your local disk.

Simplify IMG URLs that include style parameters

IMG URLs that include style parameters generally include the access URL of the image to process, the style parameter and the name of the style. Example: `https://image-demo.oss-cn-qd-ase-d01-a.mytest-inc.com/example.jpg?x-oss-process=style/small`. You can replace `?x-oss-process=style/` with customized delimiters to simplify the IMG URL. For example, if you specify the delimiter as an exclamation point (!), the IMG URL can be simplified to `https://image-demo.oss-cn-qd-ase-d01-a.mytest-inc.com/example.jpg!small`.

1. In the Buckets page, click the **Image Processing (IMG)** tab.
2. Click **Access Settings**.
3. On the **Access Settings** panel, select one of the **Delimiters**.

   Only hyphens (-), underscores (_), forward slashes (/), and exclamation points (?) can be used as delimiters.
4. Click **OK**.

6.1.4.16.2. Configure source image protection

Object Storage Service (OSS) provides the source image protection feature to protect your images from being used by unauthorized anonymous requesters. After you enable source image protection for your bucket, anonymous requesters can access original images in the bucket only by adding style parameters in the requests or by using signed URLs.

Context

You can use the following methods to access the protected original images:

- Use URLs that contain style parameters in the following format: `https://BucketName.Endpoint/ObjectName?x-oss-process=style/StyleName`.
- Use signed URLs in the following format: `https://BucketName.Endpoint/ObjectName?Signature`.

Procedure

1. Log on to the OSS console.
2. Click the **Image Processing (IMG)** tab. On the page that appears, click **Access Settings**.
3. In the **Access Settings** panel, turn on **Protect Source Image File** and configure the parameters described in the following table.

Parameter	Description
Protected File Extensions	Select a file suffix from the **Protected File Extensions** drop-down list. All objects in the bucket that match the specified suffix are protected.
Delimiters	After you select a custom delimiter, you can use the delimiter to replace ?x-oss-process=style/ to simplify the IMG URL. OSS supports the following delimiters: hyphens (-), underscores (_), forward slashes (/), and exclamation points (!). Click the check box before the delimiter that you want to select. For example, if you set the delimiter to an exclamation point (!), the IMG URL can be simplified to the following format: `http(s)//:BucketName.Endpoint/ObjcetName!StyleName`.

4. Click **OK**.

6.1.5. Objects

6.1.5.1. Search for objects

You can search for objects whose names contain specific prefixes in buckets or folders in the OSS console.

Prerequisites

Object are uploaded to the bucket. For more information, see Upload objects.

Context

When you search for objects based on a prefix, search strings are case-sensitive and cannot contain forward slashes (/). You can search for objects only in the root folder of the current bucket or in the current folder. Subfolders and objects stored in subfolders cannot be searched.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket in which the objects that you want to search for are stored.
3. Click the **Files** tab.
4. Search for objects.
   - Search for objects or folders within the root folder of the bucket

     In the upper-right corner, enter the prefix to search in the search box and press Enter or click the icon to search for related objects. Objects and subfolders whose names contain the specified prefix within the root folder of the bucket are displayed.
   - Search for objects or subfolders within a specified folder

     Click the folder in which the objects or subfolders that you want to search for are stored. In the upper-right corner, enter the prefix to search in the search box and press Enter or click the icon to search for related objects. Objects and subfolders whose names contain the specified prefix within the current folder are displayed.

6.1.5.2. Configure object ACLs

You can configure the ACL of an object in the OSS console to control access to the object.

Prerequisites

An object is uploaded to the bucket. For more information, see Upload objects.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket that contains the object whose ACL you want to configure.
3. In the left-side navigation pane, click **Files**.
4. Click the name of the object whose ACL you want to configure. In the **View Details** panel, click **Set ACL** on the right side of **File ACL**.

   You can also choose **More > Set ACL** in the Actions column corresponding to the object whose ACL you want to configure.
5. In the **Set ACL** panel, configure the ACL of the object.

   You can set the ACL of the object to one of the following values:

- **Inherited from Bucket**: The ACL of the object is the same as that of the bucket.
- **Private**: Only the owner or authorized users of this bucket can read and write objects in the bucket. Other users, including anonymous users cannot access the objects in the bucket without authorization.
- **Public Read**: Only the owner or authorized users of this bucket can write objects in the bucket. Other users, including anonymous users can only read objects in the bucket.
- **Public Read/Write**: All users, including anonymous users can read and write objects in the bucket. Fees incurred by such operations are charged to the owner of the bucket. Exercise caution when you set the object ACL to this value.

6. Click **OK**.

6.1.5.3. Create folders

You can use the OSS console to create and simulate basic features of folders in Windows. This topic describes how to create a folder by using the OSS console.

Prerequisites

A bucket is created. For more information, see Create buckets.

Context

OSS does not use a hierarchical structure for objects, but instead uses a flat structure. All elements are stored in buckets as objects. To facilitate object grouping and to simplify management, the OSS console displays objects whose names end with a forward slash (/) as folders. These objects can be uploaded and downloaded. You can use OSS folders in the OSS console in the same manner as you use folders in Windows.

> **Note** The OSS console displays objects whose names end with a forward slash (/) as folders, regardless of whether these objects contain data. The objects can only be downloaded by calling an API operation or by using OSS SDKs.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket in which you want to create folders.
3. Click the **Files** tab. On the page that appears, click **Create Folder**.
4. In the **Create Folder** panel, enter the folder name.

   The folder name must comply with the following conventions:

   - The name can contain only UTF-8 characters and cannot contain emojis.
   - The name cannot start with a forward slash (/) or backslash (\). The name cannot contain consecutive forward slashes (/). You can use forward slashes (/) in a folder name to quickly create a subfolder. For example, when you create a folder named *example/test/*, the folder named *example/* is created in the root folder of the bucket and the subfolder named *test/* is created in the example/ folder.
   - The name cannot be two consecutive periods (`..`).
   - The folder name must be 1 to 254 characters in length.

5. Click **OK**.

6.1.5.4. Configure bucket policies to authorize other users to access OSS resources

You can configure bucket policies to authorize other users to access specified Object Storage Service (OSS) resources.

Context

You can configure multiple bucket policies for a bucket. However, the total size of the policies cannot exceed 16 KB.

Method 1: Configure bucket policies by using the GUI

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure bucket policies.
3. On the bucket details page that appears, click **Files**. On the page that appears, click **Authorize**.
4. On the **GUI** tab, click **Authorize**.
5. In the **Authorize** panel, configure the parameters and then click **OK**. The following table describes the parameters that you can configure.

Parameter	Description
Applied To	Select the resources that you want to authorize other users to access. ○ **Whole Bucket**: The authorization policy applies to all resources in the whole bucket. ○ **Specified Resource**: The authorization policy applies only to specified resources in the bucket. You can configure multiple bucket policies for specific resources in a bucket. ■ Configure a bucket policy for a directory To configure a bucket policy to authorize users to access all subdirectories and objects within a directory, add an asterisk (*) at the end of the directory name. For example, to authorize users to access all subdirectories and objects within a directory named abc, enter *abc/**. ■ Configure a bucket policy for a specific object To configure a bucket policy to authorize users to access a specific object, enter the full path of the object that excludes the bucket name. For example, to authorize users to access an object named myphoto.png in the abc directory, enter *abc/myphoto.png*.

Parameter	Description
Accounts	Select the type of accounts that you want to authorize. ◦ **Anonymous Accounts (*)**: Select this option if you want to authorize all users to access the specified resources. ◦ **Other Accounts**: Select this option if you want to authorize other Apsara Stack tenant accounts, Resource Access Management (RAM) users, or users that use a temporary token generated by Security Token Service (STS) to access the specified resources. ■ To authorize other Apsara Stack tenant accounts or RAM users to access the specified resources, enter the UIDs of the accounts or RAM users. ■ To authorize users that use a temporary token generated by STS to access the specified resources, enter the user and role information in the following format: `arn:sts::{RoleOwnerUid}:assumed-role/{RoleName}/{RoleSessionName}`. For example, the role used to generate a user that uses a temporary token generated by STS is testrole, the UID of the Apsara Stack tenant account that owns the role is 12345, and the RoleSessionName that is specified when the temporary user is generated is testsession. In this case, enter `arn:sts::12345:assumed-role/testrole/testsession`. To authorize all users that use a temporary token generated by STS to access the specified resources, use asterisks (*) as wildcards. For example, enter `arn:sts::*:*/*/*`. **Notice** If an authorized user is a user that uses a temporary token generated by STS, the user cannot access the specified resources on the OSS console.
Authorized Operation	You can use the following methods to specify authorized operations: **Basic Settings** and **Advanced Settings**. ◦ Basic Settings If you select this method, you can configure the following permissions based on your requirements. You can move the pointer over the icon on the right side of each permission to view the actions that correspond to the permission option. ■ **Read Only**: Authorized users can view, list, and download the specified resources. ■ **Read/Write**: Authorized users can read data from and write data to the specified resources. ■ **Any Operation**: Authorized users can perform all operations on the specified resources. ■ **None**: Authorized users cannot perform operations on the specified resources. **Notice** If multiple bucket policies are configured for a user, the user has all the permissions defined in these policies. However, a policy in which Authorized Operation is set to **None** takes precedence over other policies. For example, if you configure a policy to grant the **Read Only** permission to a user, and then configure another policy to grant the **Read/Write** permission to the same user, the permission of the user is **Read/Write**. If you configure a third policy to grant the **None** permission to the user, the permission of the user is **None**. ◦ Advanced Settings If you select this method, you must configure the following parameters: ■ **Effect**: Select Allow or Deny. ■ **Action**: Specify the operation that you want to allow or deny. You can specify any action that is supported by OSS.

Parameter	Description
Conditions	Optional. You can configure this parameter in both Basic Settings and Advanced Settings to specify the conditions that users must meet to access the specified OSS resources. ○ **Access Method**: Select HTTPS or HTTP. ○ **IP =**: Specify the IP addresses or Classless Inter-Domain Routing (CIDR) blocks that can be used to access the specified OSS resources. Separate multiple IP addresses with commas (,). ○ **IP ≠**: Specify IP addresses or CIDR blocks that cannot be used to access OSS resources. Separate multiple IP addresses with commas (,).

6. Click **OK**.

Method 2: Configure bucket policies by specifying policy syntax

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket for which you want to configure bucket policies.
3. On the bucket details page that appears, click **Files**. On the page that appears, click **Authorize**.
4. On the **Syntax** tab, click **Edit**.

   You can specify policy syntax based on your requirements to manage fine-grained permissions. The following examples describe the bucket policies configured by the resource owner whose UID is `174649585760xxxx` in different scenarios:

   ○ Example 1: Allow anonymous users to list all objects in a bucket named examplebucket.

```
{
    "Statement": [
        {
            "Action": [
                "oss:ListObjects",
                "oss:ListObjectVersions"
            ],
            "Effect": "Allow",
            "Principal": [
                "*"
            ],
            "Resource": [
                "acs:oss:*:174649585760xxxx:examplebucket"
            ]
        },
    ],
    "Version": "1"
}
```

   ○ Example 2: Forbid anonymous users whose IP addresses are not in the CIDR block `192.168.0.0/16` from performing operations on a bucket named examplebucket.

```
{
    "Version": "1",
    "Statement": [
        {
            "Effect": "Deny",
            "Action": "oss:*",
            "Principal": [
                "*"
            ],
            "Resource": [
                "acs:oss:*:174649585760xxxx:examplebucket"
            ],
            "Condition":{
                "NotIpAddress": {
                    "acs:SourceIp": ["192.168.0.0/16"]
                }
            }
        }
    ]
}
```

- Example 3: Allow a RAM user whose UID is `20214760404935xxxx` only to read the `hangzhou/2020` and `hangzhou/2015` directories in a bucket named examplebucket.

```
{
    "Statement": [
        {
            "Action": [
                "oss:GetObject",
                "oss:GetObjectAcl",
                "oss:GetObjectVersion",
                "oss:GetObjectVersionAcl"
            ],
            "Effect": "Allow",
            "Principal": [
                "20214760404935xxxx"
            ],
            "Resource": [
                "acs:oss:*:174649585760xxxx:examplebucket/hanghzou/2020/*",
                "acs:oss:*:174649585760xxxx:examplebucket/hangzhou/2015/*"
            ]
        },
        {
            "Action": [
                "oss:ListObjects",
                "oss:ListObjectVersions"
            ],
            "Condition": {
                "StringLike": {
                    "oss:Prefix": [
                        "hanghzou/2020/*",
                        "hangzhou/2015/*"
                    ]
                }
            },
            "Effect": "Allow",
            "Principal": [
                "20214760404935xxxx"
            ],
            "Resource": [
                "acs:oss:*:174649585760xxxx:examplebucket"
            ]
        }
    ],
    "Version": "1"
}
```

5. Click **Save**.

Access authorized OSS resources

After you configure a bucket policy for a bucket, you can use the following methods to access the resources specified in the policy:

- Object URL (only for authorized anonymous users)

  Anonymous users can enter the URL of an object specified in the policy in a browser to access the object. The URL of the object consists of the default domain name of the bucket and the path of the object. Example: `http://mybucket.oss-cn-beijing.aliyuncs.com/file/myphoto.png`.

- OSS console

  Log on to the OSS console. In the left-side navigation pane, click the + icon next to **My OSS Paths**. In the Add Path panel, add the region in which the bucket is located and object path specified in the bucket policy. For more information, see Add OSS paths.

6.1.5.5. Delete objects

You can delete uploaded objects in the OSS console when they are no longer needed.

Context

You can delete a single object or batch delete multiple objects. You can batch delete up to 100 objects. To delete specific objects or batch delete more than 100 objects, we recommend that you use API operations or OSS SDKs.

> **Notice** Deleted objects cannot be recovered. Exercise caution when you delete objects.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket in which the objects you want to delete are stored.
3. In the left-side navigation pane, click **Files**.
4. Select one or more objects that you want to delete in the object list, and then choose **Batch Operation > Delete**.

   You can also choose **More > Completely Delete** in the Actions column corresponding to the object you want to delete.
5. In the dialog box that appears, click **OK**.

6.1.5.6. Manage parts

When you use multipart upload to upload an object, the object is split into several parts. After all of the parts are uploaded to the OSS server, you can call CompleteMultipartUpload to combine the parts into a complete object.

Context

You can also configure lifecycle rules to clear parts that are not needed on a regular basis. For more information, see Manage lifecycle rules.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation pane, click **Buckets**. On the Buckets page, click the name of the bucket in which the parts you want to delete are stored.
3. Click the **Files** tab. On the page that appears, click **Parts**.
4. In the **Parts** panel, delete the parts.
   - To delete all parts in the bucket, select all parts and then click **Delete All**.
   - To delete specific parts in the bucket, select these parts and then click **Delete**.
5. In the dialog box that appears, click **OK**.

6.1.6. Create single tunnels

You can create single tunnels between OSS and a virtual private cloud (VPC) to access OSS resources from the VPC.

Prerequisites

A VPC and a vSwitch are created.

For more information, see the *Create a VPC* and *Create a vSwitch* topics in *VPC User Guide*.

Procedure

1. Log on to the OSS console.
2. In the left-side navigation panel, click **Create Single Tunnel**.
3. Click **Create**.
4. On the **Create Single Tunnel** page, configure the parameters described in the following table.

Parameter	Required	Description
Organization	Yes	Select the organization of the VPC from which you want to access OSS resources.
Resource Set	Yes	After you select an organization, the resource set is automatically selected based on the organization.
Region	Yes	After you select an organization, a region is automatically selected based on the organization.
Description	No	Enter the description of the single tunnel you want to create. The description cannot exceed 180 characters in length.
VPC	Yes	Select the VPC that you created. You can also click **Create VPC** to create a VPC.
vSwitch	Yes	Select the vSwitch that you created. You can also click **Create vSwitch** to create a vSwitch.

5. Click **Submit**.

6.1.7. Add OSS paths

You can add the paths of OSS resources in the console for quicker access.

Prerequisites

A bucket is created. For more information about how to create a bucket, see Create buckets.

Procedure

1. Log on to the OSS console.
2. Click the + icon on the right side of **My OSS Paths**.
3. In the **Add Authorized OSS Path** panel, add a path.

   You can configure the following parameters to add a path.

   - **Region**: Select the region of the bucket in the path that you want to add.
   - **File Path**: Add the path of the resource that you want to access. The path is in the `bucket/object-prefix` format. For example, if the OSS resource that you want to access is the root folder of a bucket named *example*, set File Path to *example*. If the OSS resource that you want to access is the *test* folder in the root folder of the bucket named *example*, set File Path to *example/test/*.

7.Apsara File Storage NAS

7.1. User Guide

7.1.1. What is NAS?

Apsara File Storage NAS is a cloud service that provides a file storage solution for compute nodes. The compute nodes include Elastic Compute Service (ECS) instances, Elastic High-Performance Computing (E-HPC) instances, and Container Service for Kubernetes (ACK) clusters.

NAS provides standard file access protocols. You can use the distributed file storage solution that provides shared access, scalability, high reliability, and high performance without the need to modify the configurations of the applications. You can mount a NAS file system on multiple compute nodes at a time. This reduces a large number of costs in data transmission and synchronization.

You can perform the following operations:

- Create a NAS file system and mount target.
- Create a permission group for the file system and add rules to the permission group. This allows access from specific IP addresses or CIDR blocks to the file system. This also allows you to grant different levels of access permissions to the IP addresses or CIDR blocks.
- Mount the file system on compute nodes. The compute nodes include ECS instances and ACK clusters. NAS allows you to access the file system by using the Network File System (NFS) and Server Message Block (SMB) protocols. You can also call POSIX-based APIs to access the file system.
- Manage file systems, mount targets, and permission groups in the NAS console.
- Call NAS API operations to manage file systems.

7.1.2. Precautions

Before you use NAS, you must familiarize yourself with the following limits.

Limits on file systems

- Maximum number of files in a single file system: 1 billion.
- Maximum name length: 255 bytes.
- Maximum size of a single file: 32 TB.
- Maximum directory depth: 1,000 levels deep.
- Maximum capacity of a single file system: 10 PB for NAS Capacity and 1 PB for NAS Performance.
- Maximum number of compute nodes on which you can mount a single file system: 10,000. Note: The file system allows simultaneous access from the 10,000 compute nodes.
- Maximum size of a protocol packet: 4 MB.
- Maximum number of Change Notify requests: 512.

Limits on NFS clients

Limits on the usage of NFS clients are listed as follows.

- You can open a maximum of 32,768 files at a time on an NFS client. Files in the list folder and its subfolders are not counted as part of the total number of open files.
- Each unique mount on an NFS client can acquire a maximum of 8,192 locks across a maximum of 256 unique file or process pairs. For example, a single process can acquire one or more locks on 256 separate files, or 8 processes can each acquire one or more locks on 32 files.
- We recommend that you do not use an NFS client in a Windows environment to access an NFS file system.

Limits on SMB clients

Each file or folder can be opened a maximum of 8,192 times in parallel across compute nodes that each have a file system mounted and users that share access to each of these file systems. This represents a maximum of 8,192 active file handlers for each file system. A maximum of 65,536 active file handlers can exist on a file system.

Limits on the NFS protocol

- NAS supports the NFSv3 and NFSv4 protocols.
- NFSv4.0 does not support the following attributes: FATTR4_MIMETYPE, FATTR4_QUOTA_AVAIL_HARD, FATTR4_QUOTA_AVAIL_SOFT, FATTR4_QUOTA_USED, FATTR4_TIME_BACKUP, and FATTR4_TIME_CREATE. If one of the preceding attributes is applied to a file system, an NFS4ERR_ATTRNOTSUPP error appears on a client that has the file system mounted.
- NFSv4.1 does not support the following attributes: FATTR4_DIR_NOTIF_DELAY, FATTR4_DIR_NOTIF_DELAY, FATTR4_DACL, FATTR4_SACL, FATTR4_CHANGE_POLICY, FATTR4_FS_STATUS, FATTR4_LAYOUT_HINT, FATTR4_LAYOUT_TYPES, FATTR4_LAYOUT_ALIGNMENT, FATTR4_FS_LOCATIONS_INFO, FATTR4_MDSTHRESHOLD, FATTR4_RETENTION_GET, FATTR4_RETENTION_SET, FATTR4_RETENTEVT_GET, FATTR4_RETENTEVT_SET, FATTR4_RETENTION_HOLD, FATTR4_MODE_SET_MASKED, FATTR4_FS_CHARSET_CAP. If one of the preceding attributes is applied to a file system, an NFS4ERR_ATTRNOTSUPP error appears on a client that has the file system mounted.
- NFSv4 does not support the following operations: OP_DELEGPURGE, OP_DELEGRETURN, and NFS4_OP_OPENATTR. If one of the preceding operations is applied to a file system, an NFS4ERR_ATTRNOTSUPP error appears on a client that has the file system mounted.
- NFSv4 does not support delegations.
- The following issues are related to user IDs (UIDs) and group IDs (GIDs):
  - On Linux, mappings between UIDs or GIDs and usernames or group names are defined in configurations files. For NFSv3 file systems, if the mapping between an ID and a name is defined in a configuration file, the name is displayed. If no mapping can be found for a UID or GID, the UID or GID is displayed.
  - For NFSv4 file systems, the usernames and group names are displayed as nobody for all files if the version of a Linux kernel is earlier than 3.0. If the kernel version is later than 3.0, the rule used by NFSv3 file systems applies to display files.

> **Notice** If a file or directory is stored on an NFSv4 file system and the Linux kernel version is earlier than 3.0, we recommend that you do not use the chown or chgrp command. If you use either one of the commands, the UID and GID of the file or directory will change to nobody.

Limits on the SMB protocol

- NAS supports protocols including SMB 2.1 or later and operating systems including Windows 7, and Windows Server 2008 R2 or later. However, NAS does not support Windows Vista, or Windows Server 2008 or earlier. Compared with SMB 2.1 or later, SMB 1.0 has lower performance and functionality. Furthermore, Windows products that support SMB 1.0 are no longer offered or supported.
- Extended file attributes and client-side caching based on leases.
- Input/output control (IOCTL) or file system control (FSCTL) operations, such as creating sparse files, compressing files, retrieving NIC status, and creating reparse points.
- Alternate data streams.
- Identity authentication provided by Active Directory (AD) or Lightweight Directory Access Protocol (LDAP).
- Several features provided by SMB 3.0 or later, such as SMB Direct, SMB Multichannel, SMB Directory Leasing, and persistent handles.
- Access control lists (ACLs) on files or directories.

7.1.3. Quick start

7.1.3.1. Log on to the NAS console

This topic describes how to log on to the Apsara File Storage NAS console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, click **Products**. In the **Storage** section, click **Apsara File Storage NAS**.

7.1.3.2. Create a file system

This topic describes how to create a file system in the Apsara File Storage NAS console.

Context

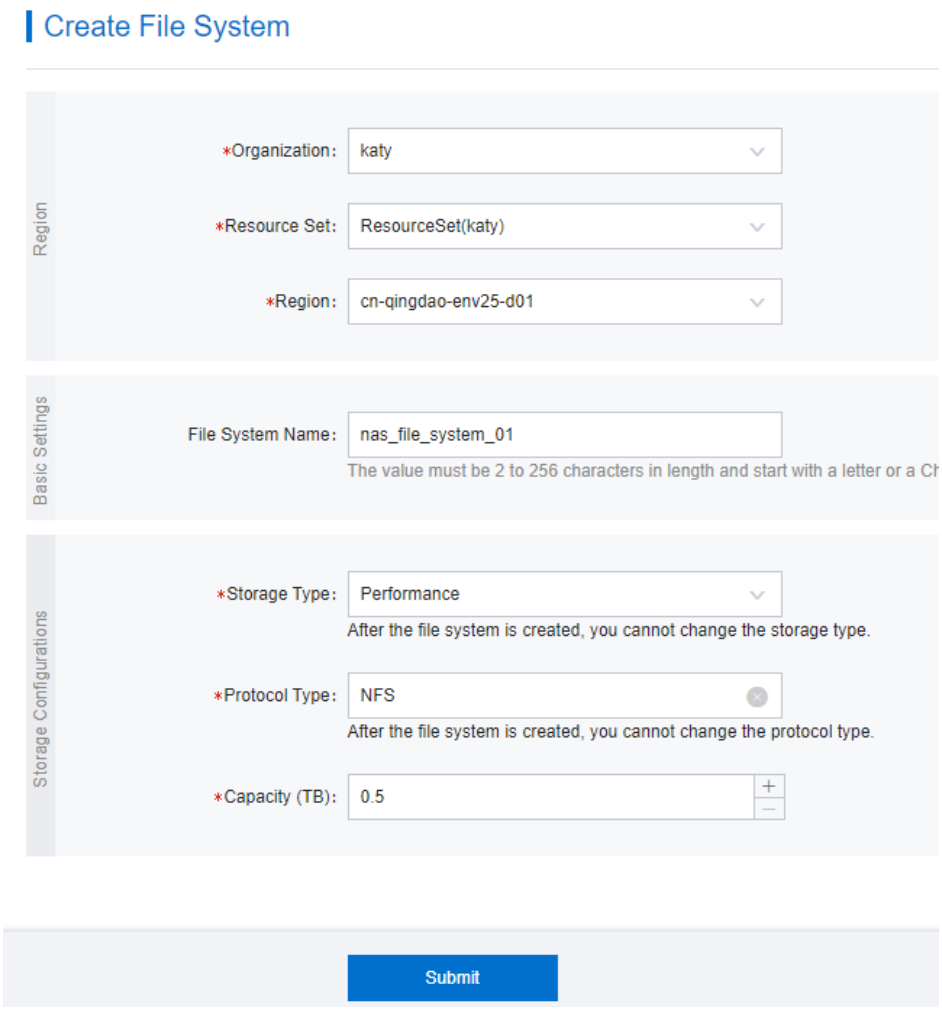
Before you create a file system, you must note the following limitations:

- You can use an Alibaba Cloud account to create a maximum of 1,000 file systems.
- The maximum capacity of a NAS Performance file system is 1 PB. The maximum capacity of a NAS Capacity file system is 10 PB.

If you want to increase the maximum storage capacity, we recommend that you contact Alibaba Cloud Technical Support.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List** and click **Create File System**.
3. In the **Create File System** dialog box, set the required parameters.



The following table lists the required parameters.

Parameter	Description
Region	Select a region where you need to create a file system.
Organization	Select an organization from the drop-down list for the instance.
Resource Set	Select a resource set from the drop-down list for the instance.
File System Name	The name of the file system. The name must be 2 o 256 characters in length and can contain letters, digits, and special characters. These special characters include underscores (_) and hyphens (-). The name must start with a letter and cannot start with http:// or https://.

Parameter	Description
Storage Type	The storage type. Select **Performance** or **Capacity** based on your business requirements. The maximum capacity of an NAS Performance file system is 1 PB. The maximum capacity of an NAS Capacity file system is 10 PB.
Protocol Type	The protocol type. Select **NFS** or **SMB** based on your business requirements. We recommend that you mount Network File System (NFS) file systems on Linux clients and Server Message Block (SMB) file systems on Windows clients.
Capacity (TB)	The capacity of the file system. ◦ The capacity of an NAS Performance file system ranges from 0.5 TB to 1024 TB. ◦ The capacity of an NAS Capacity file system ranges from 0.5 TB to 10240 TB.

4. Click **OK** to complete the creation.

7.1.3.3. Create a permission group and add rules

This topic describes how to create a permission group and add rules to the permission group in the Apsara File Storage NAS console.

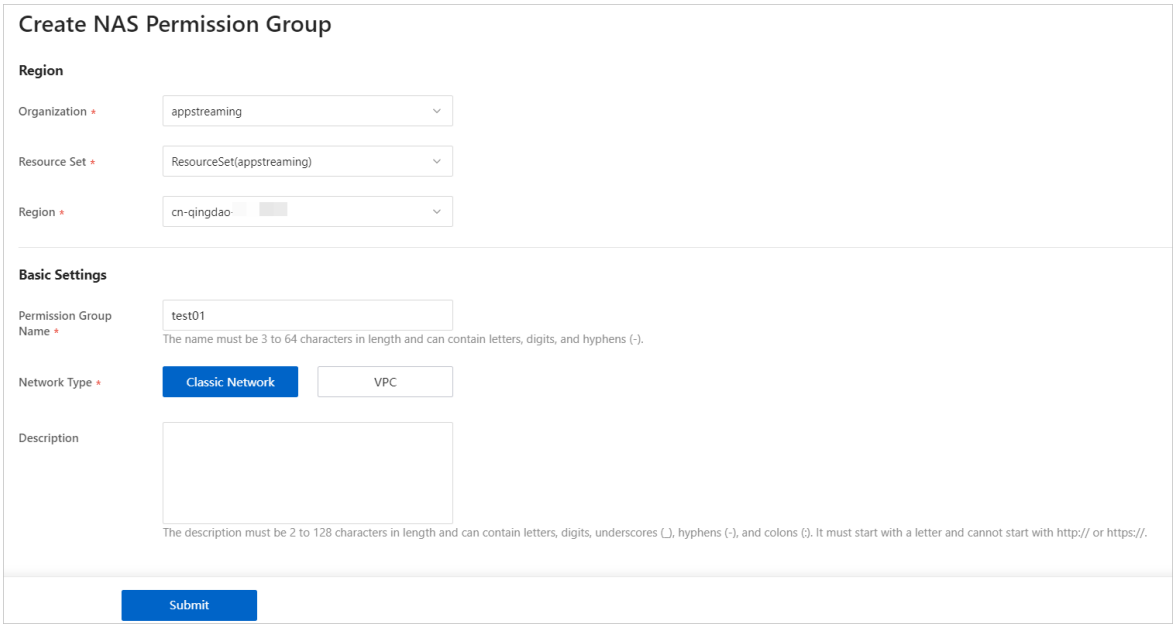
Context

In NAS, each permission group represents a whitelist. You can add rules to a permission group to allow access to a file system from specific IP addresses or CIDR blocks. You can also grant different access permissions to different IP addresses or CIDR blocks.

> **Note** You can use an Alibaba Cloud account to create a maximum of 100 permission groups. If you want to increase the limit, we recommend that you contact Alibaba Cloud Technical Support.

Creates a permission group

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > Permission Group** and click **Create Permission Group**.
3. In the **Create Permission Group** dialog box, specify the required parameters.



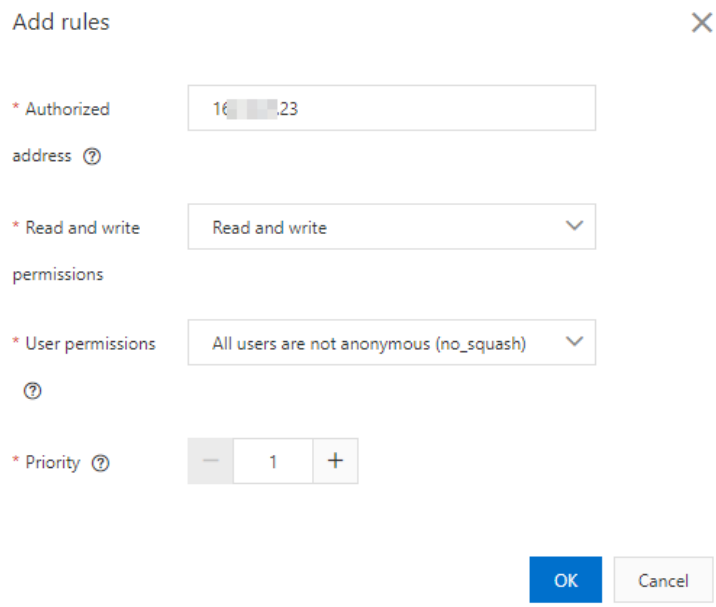
The following table lists the required parameters.

Parameter	Description
Organization	The organization to which the permission group belongs.
Resource Set	The resource set to which the permission group belongs.
Region	The region where you want to create the permission group.
Name	The name of the permission group. The name must be 3 to 64 characters in length and can contain letters, digits, and hypens (-).
Network Type	The network type. Select **Classic Network** or **VPC** based on your business requirements.

4. Click **OK** to complete the creation of the permission group.

Create a rule

1. Log on to the Apsara File Storage NAS console.
2. On the **Permission Group** page, find the target permission group and click **Manage**.
3. Click **Add Rule**.
4. In the **Add Rule** dialog box that appears, specify the required parameters.



The following table lists the required parameters.

Parameter	Description
Authorization Address	Specifies the authorized object to which the rule applies. You can specify an IP address or CIDR block. Only IP addresses are available for permission groups of the classic network type.
Read/Write Permission	Specifies whether to allow read-only or read/write access to the file system from the authorized object. Valid values: **Read-only** and **Read/Write**.
User Permission	Specifies whether to limit a Linux user's access to a file system. ◦ **Do not limit root users (no_squash)**: allows access to a file system from root users. ◦ **Limit root users (root_squash)**: denies access to a file system from root users. All root users are treated as nobody users. ◦ **Limit all users (all_squash)**: denies access to a file system from all users including root users. All users are treated as nobody users. The nobody user is created by default on Linux. The user has only the most basic permissions and can access only the open content of servers. This feature offers high security.
Priority	When multiple rules are applied to an authorized object, the rule with the highest priority takes effect. Valid values: 1 to 100, in which 1 is the highest priority.

5. Click **OK** to complete the creation of the rule.

7.1.3.4. Add a mount target

This topic describes how to add a mount target. After an Apsara File Storage NAS file system is created, you must add a mount target to the file system. Then, you can use the mount target to mount the file system on compute nodes. These compute nodes include Elastic Compute Service (ECS) instances and Alibaba Cloud Container Service for Kubernetes (ACK) nodes.

Prerequisites

- A file system is created. For more information, see Create a file system.
- A permission group and a rule are created. For more information, see Create a permission group and add rules.

Context

A mount target is an endpoint that resides in a VPC or classic network. Each mount target corresponds to a file system. Mount targets of the VPC and classic network types are available for NAS file systems.

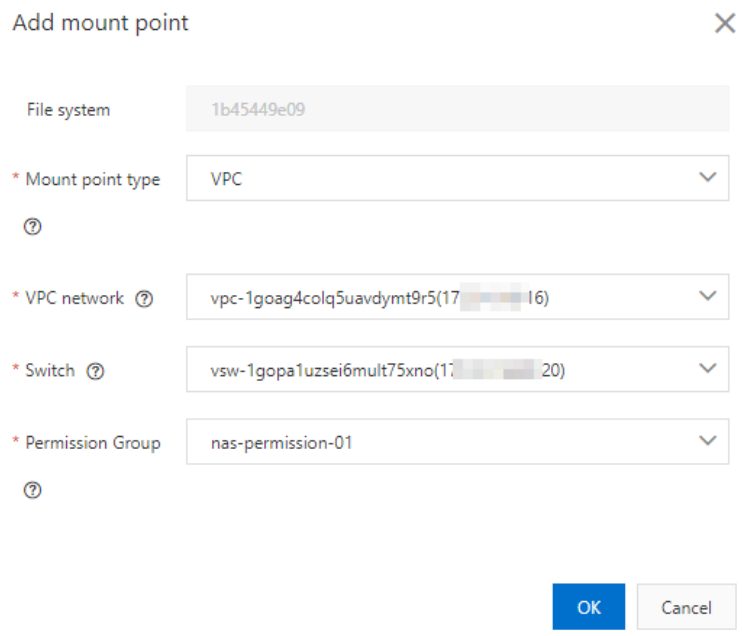
> **Note** You use a mount target to mount a file system on multiple compute nodes for shared access. These compute nodes include ECS instances and ACK nodes.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.
3. Find the target file system and click **Manage**.
4. On the **Mount Target** tab, click **Add Mount Target**.
5. In the **Add Mount Target** dialog box that appears, specify the required parameters.

   Mount Target Type: includes **VPC** and **Classic Network**.

   > **Note** Mount targets of the classic network type allow access only from ECS instances that belong to the same Alibaba Cloud account as the mount targets.



   - If you want to create a mount target of the VPC type, specify the following parameters.

Parameter	Description
VPC	The VPC. **Note** The VPC you specify must be the same as the VPC where the compute nodes reside. These compute nodes include ECS instances and ACK nodes.

Parameter	Description
VSwitch	The VSwitch.
Permission Group	The permission group.

- If you want to create a mount target of the classic network type, specify the following parameters.

Parameter	Description
Permission Group	The permission group.

6. Click **OK** to complete the configuration.

7.1.3.5. Mount an NFS file system

This topic describes how to mount a Network File System (NFS) file system. Before you mount a file system, you must create the file system and a mount target for the file system. Then, you can use the mount target to mount the file system on compute nodes. These compute nodes include Elastic Compute Service (ECS) instances and Alibaba Cloud Container Service for Kubernetes (ACK) nodes.

Prerequisites

- A file system is created. For more information, see Create a file system.
- A permission group and a rule are created. For more information, see Create a permission group and add rules.
- A mount target is created. This topic takes a mount target of the VPC type as an example. For more information, see Add a mount target.
  - If you create a mount target of the VPC type for a file system, you can mount the file system only on ECS instances that reside in the same VPC as the mount target. You can specify a permission group for the mount target. Then, you can add several rules to the permission group. The authorization address of a rule must match the IP range of the VPC that hosts the ECS instances.
  - If you create a mount target of the classic network type for a file system, you can mount the file system only on ECS instances that belong to the same Alibaba Cloud account as the mount target. You can specify a permission group for the mount target. Then, you can add several rules to the permission group. The authorization address of a rule must match the IP range of the private network that hosts the ECS instances.
- A compute node is created. This topic takes a Linux ECS instance as an example.

Step 1: Install an NFS client

Before you mount an NFS file system on a Linux ECS instance, you must install an NFS client. If an NFS client is instaned, skip this step.

1. Log on to the Linux ECS instance. For more information, see the **Quick start > Connect to an ECS instance** topic of the *ECS User Guide* .
2. Install the NFS client.
   - If CentOS, RHEL, or Aliyun Linux runs on the ECS instance, use the following command to install the NFS client.

     ```
     sudo yum install nfs-utils
     ```

   - If Ubuntu or Debian runs on the ECS instance, use the following commands to install the NFS client.

     ```
     sudo apt-get update
     ```

     ```
     sudo apt-get install nfs-common
     ```

Step 2: Mount an NFS file system

1. Log on to the Linux ECS instance. For more information, see the **Quick start > Connect to an ECS instance** topic of the *ECS User Guide* .
2. Mount the NFS file system.

   Use the following command to mount the NFS file system. In the command, replace file-system-id.region.nas.aliyuncs.com:/ /mnt with a value that is specific to your environment.

   - To mount an NFSv4 file system, use the following command.

     ```
     sudo mount -t nfs -o vers=4.0,minorversion=0,rsize=1048576,wsize=1048576,hard,timeo=600,retrans=2,noresvport file-system-id.region.nas.aliyuncs.com:/ /mnt
     ```

   - To mount an NFSv3 file system, use the following command.

     ```
     sudo mount -t nfs -o vers=3,nolock,proto=tcp,rsize=1048576,wsize=1048576,hard,timeo=600,retrans=2,noresvport file-system-id.region.nas.aliyuncs.com:/ /mnt
     ```

     Mount parameters

Parameter	Description
file-system-id.region.nas.aliyuncs.com:/ /mnt	Specifies the mount target of the NAS file system, the forward slash (/) following the mount target specifies the root directory of the NAS file system, and /mnt specifies a local directory that resides on the Linux ECS instance. You must replace the example values based on your business requirements. ■ The mount target, for example, file-system-id.region.nas.aliyuncs.com. To obtain information about a mount target, follow these steps. Log on to the NAS console, find the target system, click **Manage** next to the file system to go to the Details page. The Details page shows information about the mount target. ■ The directory of the NAS file system: specifies the root directory (/) or a subdirectory (/sub1). If a subdirectory is specified, make sure that the subdirectory exists. ■ The local directory on which you want to mount a file system: specifies the root directory (/) or a subdirectory (/mnt) of a system such as Linux. If a subdirectory is specified, make sure that the subdirectory exists.
vers	The version of the file system. Only NFSv3 and NFSv4 are available.
Mount option	When you mount a file system, multiple mount options are available. Separate multiple mount options with commas (,). For more information, see the following Mount options. **Note** When you specify mount options, take note of the following items. ■ To avoid a decrease in performance, we recommend that you specify the maximum value (1048576) for both the rsize mount option and the wsize mount option. ■ If you need to modify the timeo mount option, we recommend that you specify a minimum of 150 for the mount option. The timeo mount option is measured in deciseconds (tenths of a second). For example, a value of 150 indicates 15 seconds. ■ To avoid data inconsistency, we recommend that you do not use the soft mount option. Use caution with the soft mount option. ■ We recommend that you use the default values for other mount options. For example, a decrease in performance may occur due to changes in some mount options. These mount options include the size of the read or write buffer or the use of attribute caching.

Mount options

Option	Description
rsize	Specifies the maximum number of bytes in each read request that the NFS client can receive. Recommended value: 1048576.
wsize	Specifies the maximum number of bytes in each write request that the NFS client can send. Recommended value: 1048576.
hard	Specifies that applications must stop accessing a file system when the file system is unavailable, and wait until the file system is available. We recommended that you use the hard mount option.
timeo	Specifies the time in deciseconds (tenths of a seconds) that the NFS client waits before it retries an NFS request. Recommended value: 600.
retrans	Specifies the number of times the NFS client retries a request. Recommended value: 2.
noresvport	Specifies that the NFS client uses a different TCP source port for a new network connection to ensure data integrity. We recommend that you use the noresvport mount option.

3. Use the `mount -l` command to view the mount result.

   The following figure shows an example of a successful mount.

   After a file system is mounted, you can use the `df -h` command to view the size of the file system.

4. After you mount an NAS file system on an ECS instance, you can access the file system from the ECS instance.

   You can access the file system in the same way you access a local directory. The following figure shows an example.

7.1.3.6. Mount an SMB file system

This topic describes how to mount a Server Message Block (SMB) file system. Before you mount a file system, you must create the file system and a mount target for the file system. Then, you can use the mount target to mount the file system on compute nodes such as Elastic Compute Service (ECS) instances.

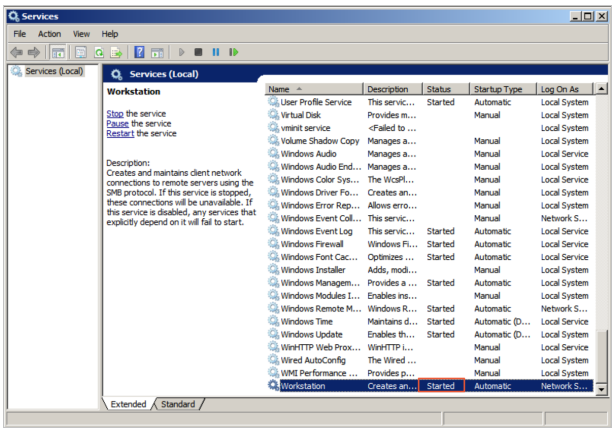
Prerequisites

1. A file system is created. For more information, see Create a file system.
2. A permission group and a rule are created. For more information, see Create a permission group and add rules.
3. A mount target is created. This topic takes a mount target of the Classic Network type as an example. For more information, see Add a mount target.
   - If you create a mount target of the VPC type for a file system, you can mount the file system only on ECS instances that reside in the same VPC as the mount target. You can specify a permission group for the mount target. Then, you can add several rules to the permission group. The authorization address of a rule must match the IP range of the VPC that hosts the ECS instances.
   - If you create a mount target of the Classic Network type for a file system, you can mount the file system only on ECS instances that belong to the same Alibaba Cloud account as the mount target. You can specify

a permission group for the mount target. Then, you can add several rules to the permission group. The authorization address of a rule must match the IP range of the private network that hosts the ECS instances.

4. An ECS instance is created. This topic takes a Windows ECS instance as an example.
5. The following Windows services are started:
   - Workstation
     a. Choose **All Programs > Accessories > Run**, or press `Win+R` and enter `services.msc` to open the Services console.
     b. Find the Workstation service and ensure that the service is **Running** and the startup type is **Automatic**.

     The default state for the Workstation service is Running.

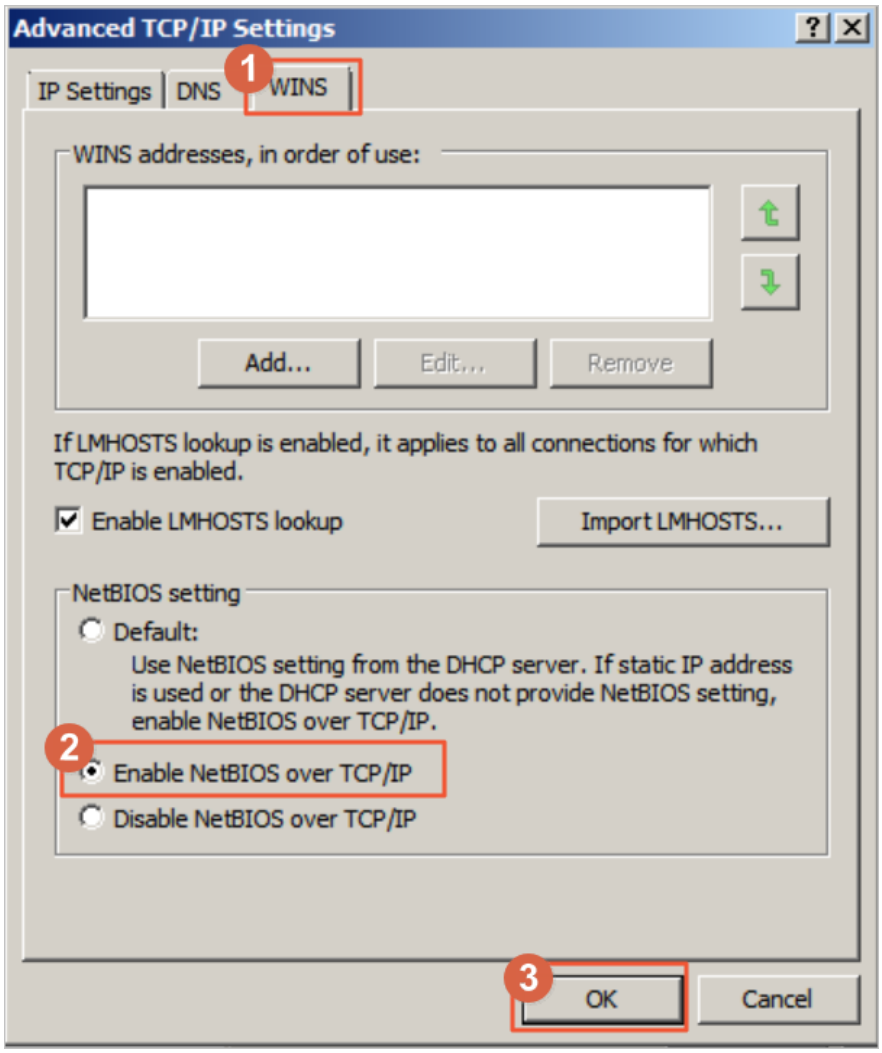


   - TCP/IP NetBIOS Helper

     Follow these steps to start the TCP/IP NetBIOS Helper service:

     a. Double-click **Network and Sharing Center** and right-click the active network connection.
     b. Click **Properties** to open the Local Area Network Properties dialog box. Double-click **Internet Protocol Version 4 (TCP/IPv4)** to open the Internet Protocol Version 4 (TCP/IPv4) Properties dialog box, and then click **Advanced**.

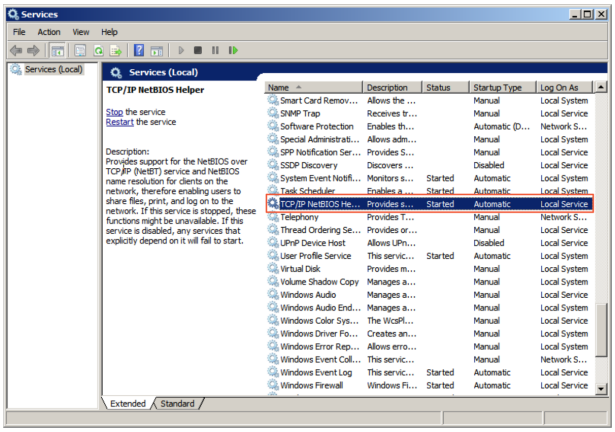
c. In the **Advanced TCP/IP Settings** dialog box, choose **WINS > Enable NetBIOS over TCP/IP**.



d. Choose **All Programs > Accessories > Run**, or press `Win+R` and enter `services.msc` to open the Services console.

e. Find the TCP/IP NetBIOS Helper service and make sure that the service is **Running** and the startup type is **Automatic**.

The default state for the TCP/IP NetBIOS Helper service is Running.



Procedure

1. Log on to a Windows ECS instance. For more information, see the **Quick start > Connect to an ECS instance** topic of the *ECS User Guide*.
2. Open the command prompt and use the following command to mount the file system.

   ```
   net use D: \\file-system-id.region.nas.aliyuncs.com\myshare
   ```

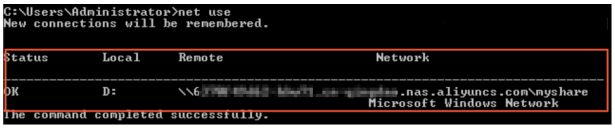
   The syntax of the command is `net use <the letter of a local drive> \\<the domain name of a mount target>\myshare`.

   - The letter of a local drive: specifies a drive on which you need to mount a file system. You can specify the target drive based on your business requirements.

     > **Note** The target drive must be different from any existing drives.

   - The domain name of the mount target: When you create a mount target for a file system, the domain name of the mount target is automatically generated. You can replace the domain name based on your business requirements. To obtain the mount target of the file system, follow these steps. Log on to the NAS console, find the target file system, and click **Manage** to go to the Details page.
   - myshare: specifies the name of an SMB share. You cannot change the name.
3. Use the `net use` command to check mount results.

   The following figure shows an example of a successful mount.

4. After you mount an NAS file system on an ECS instance, you can access the file system from the ECS instance.

7.1.4. File systems

7.1.4.1. View the details of a file system

This topic describes how to view the details of a file system. The details include the information of the file system and the attached mount targets. You can view the details of a file system in the Apsara File Storage NAS console.

Prerequisites

A file system is created. For more information, see Create a file system.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.
3. Find the file system and click **Management** to go to the details page of the file system.

   The details page consists of the following sections:

   - **Basic Information**: In this section, the information of the file system is displayed. The information includes the ID, region, protocol type, and storage specifications.
   - **Mounting Use**: In this section, the list of mount targets attached to the file system is displayed. You can manage the mount targets and view the clients on which the file system is mounted.
   - **Quota Management**: In this section, the quota status of each directory in the file system is displayed. You can add, edit, and delete quotas for each directory.

7.1.4.2. Delete a file system

This topic describes how to delete a file system in the Apsara File Storage NAS console.

Prerequisites

A file system is created. For more information, see Create a file system.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.
3. Find the target file system, and click **Manage**.

   > **Note**
   > - Before you can delete a file system, you must remove all mount targets from the file system.
   > - Use caution when you delete a file system. After a file system is deleted, the data on the file system cannot be restored. We recommend that you ensure that all data is backed up.

4. In the **Delete File System** dialog box, click **OK** to complete the deletion.

7.1.4.3. Scale up a file system

If the used space of an Apsara File Storage NAS file system reaches the configured capacity, data can no longer be written to the file system. To ensure the availability of the NAS service, we recommend that you scale up the file system before the used space reaches the configured capacity. This topic describes how to scale up a file system in the NAS console.

Prerequisites

A file system is created. For more information, see Create a file system.

> **Notice** A file system can be scaled up but cannot be scaled down. A NAS Performance file system can be scaled up to a maximum of 1 PB. A NAS Capacity file system can be scaled up to a maximum of 10 PB.

Procedure

1.
2. In the left-side navigation pane, choose **File System > File System List**.
3. Find the file system and click **Upgrade** in the **Operations** column.
4. In the **Upgrade** dialog box, enter the scaled capacity of the file system in the **New Total Capacity** field.
5. Click **OK**.

7.1.5. Mount targets

7.1.5.1. View mount targets

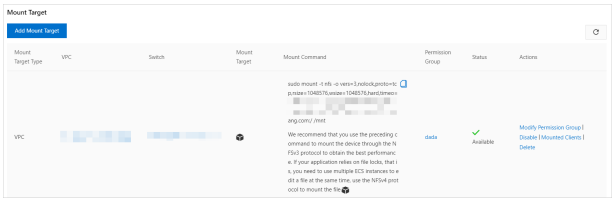
This topic describes how to view mount targets in the Apsara File Storage NAS console.

Context

- A file system is created. For more information, see Create a file system.
- A permission group and a rule are created. For more information, see Create a permission group and add rules.
- A mount target is created. For more information, see Add a mount target.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.
3. Find the target file system and click **Manage**.
4. In the **Mount Target** section, view mount targets in the file system.



7.1.5.2. Enable or disable a mount target

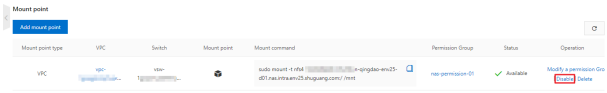
This topic describes how to enable or disable a mount target. You can control access to a mount target from clients by enabling or disabling the mount target.

Prerequisites

- A file system is created. For more information, see Create a file system.
- A permission group and a rule are created. For more information, see Create a permission group and add rules.
- A mount target is created. For more information, see Add a mount target.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.
3. Find the target file system and click **Manage**.
4. After you find the mount target that you want to disable or enable, you can perform the following operations:
   - Disable the mount target. Click **Disable**. In the Disable Mount Target dialog box, click **OK** to deny access to the mount target from clients.
   - Enable the mount target. Click **Enable**. In the Enable Mount Target dialog box, click **OK** to allow access to the mount target from clients.



7.1.5.3. Delete a mount target

This topic describes how to delete a mount target in the Apsara File Storage NAS console.

Prerequisites

- A file system is created. For more information, see Create a file system.
- A permission group and a rule are created. For more information, see Create a permission group and add rules.
- A mount target is created. For more information, see Add a mount target.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.

3. Find the target file system and click **Manage**.
4. Find the mount target that you want to delete and click **Delete**.

> **Note** Use caution when you delete a mount target. After you delete a mount target, the mount target cannot be restored.

5. In the Delete Mount Target dialog box, click **OK**.

7.1.5.4. Modify the permission group of a mount target

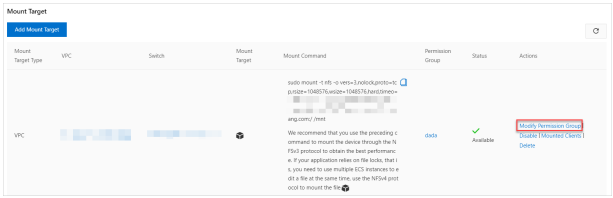
You must add a permission group to each mount target. You can modify the permission group of a mount target in the Apsara File Storage NAS console.

Prerequisites

- A file system is created. For more information, see Create a file system.
- A permission group and a rule are created. For more information, see Create a permission group and add rules.
- A mount target is created. For more information, see Add a mount target.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > File System List**.
3. Find the target file system, and click **Manage**.
4. Find the mount target that you want to modify and click **Modify Permission Group**.



5. In the **Modify Permission Group** dialog box, change the permission group and click **OK**.

7.1.6. Permission groups

7.1.6.1. View permission groups

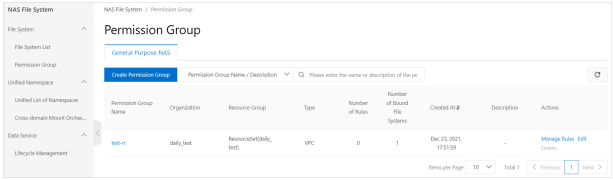
This topic describes how to view permission groups in the Apsara File Storage console.

Prerequisites

A permission group is created. For more information, see Create a permission group and add rules.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > Permission Group** and view permission groups in the region.



7.1.6.2. Delete a permission group

This topic describes how to delete a permission group in the Apsara File Storage NAS console.

Prerequisites

A permission group is created. For more information, see Create a permission group and add rules.

Operating system

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > Permission Group**.
3. Find the target permission group and click **Delete**.

   > **Note** Permission groups in use cannot be deleted. Before you can delete a permission group, you must remove the permission group from the linked mount target.

4. In the Delete Permission Group dialog box, click **OK**.

7.1.6.3. Manage permission group rules

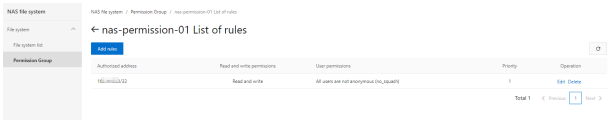
This topic describes how to manage permission group rules in the Apsara File Storage NAS console. The management includes viewing the details of rules, modifying rules, and deleting rules.

Prerequisites

A permission group and a permission group rule are created. For more information, see Create a permission group and add rules.

Procedure

1. Log on to the Apsara File Storage NAS console.
2. Choose **NAS > Permission Group**.
3. Find the target permission group and click **Manage**.
4. On the Rules page, you can perform the following operations:
   - View all rules for the permission group.

   - Modify a rule. Find the target rule and click **Edit** to modify the details of the rule. The details include the **authorization address**, **read/write permissions**, **user permission**, and **priority**.
   - Delete a rule. Find the target rule and click **Delete**. In the Delete Rule dialog box, click **OK**.

7.1.7. Manage quotas

This topic describes how to use the Alibaba Cloud quota_tool tool to manage quotas on Elastic Compute Service (ECS) instances that have Apsara File Storage NAS file systems mounted. You can configure, view, and cancel quotas on these ECS instances.

Prerequisites

An NFS file system of the NAS Capacity or NAS Performance type is mounted on an ECS instance. For more information, see Mount an NFS file system.

Context

NAS allows you to view and manage directory-level quotas. Directory-level quotas specify the maximum number of files in each directory and the maximum storage space for these files.

From the perspective of the application scope, quotas are sorted into quotas for all users and quotas for a single user or group. Quotas for all users specify the maximum storage space for files that all users can create in a directory. Quotas for a single user or group specify the maximum storage space for files that a user or group can create in a directory.

From the perspective of the restriction level, quotas are sorted into statistical quotas and restriction quotas. Statistical quotas collect only the usage of storage space. You can query and view statistical data. Restriction quotas specify the maximum capacity of storage space for files that you can create in a directory. If the limit is exceeded, you may fail to create a file or subdirectory, append data to a file, or perform other operations.

Notice

- Only statistical quotas are available.
- NAS performs asynchronous calculation for quotas at the backend. When you use the quota_tool tool to query statistical data about quotas, the process requires a period of time to complete. In most cases, the time period is about 5 to 15 minutes.

Configure quotas

This topic uses the */mnt* directory as an example.

1. Log on to an Elastic Compute Service (ECS) instance by using a root account.

   You can use the quota_tool tool on an ECS instance that has a NAS file system mounted. You must run the tool with the root permissions. The following describes how to use the quota_tool tool on the ECS instance.

2. Use the following command to download the quota_tool tool.

   ```
   wget https://nasimport.oss-cn-shanghai.aliyuncs.com/quota_tool_v1.0 -O quota_tool
   ```

3. Use the following command to grant the execute permission to the quota_tool tool.

   ```
   sudo chmod a+x quota_tool
   ```

4. Configure quotas.

   Note For each file system, you can configure quotas only for a maximum of 10 directories.

   The syntax of the command that you use to configure quotas is `sudo ./quota_tool set --dir [DIR] [OPTION]`.

Parameter	Description
--dir [DIR]	Specifies the directory for which you want to configure quotas. For example, --dir /mnt/data/.

Parameter	Description
OPTION	Specifies the required options. Note When you specify options, you must follow these rules: 1. the --accounting option is required. 2. One of the --alluser, --uid, and --gid options is required. . ○ --accounting: specifies a statistical quota. ○ --alluser: specifies a directory-level quota for all users. ○ --uid: specifies the UID of a user. For example, --uid 505 indicates the quota is configured only for the user whose UID is 505. ○ --gid: specifies the GID of a group. For example, --gid 1000 indicates the quota is configured only for the group whose GID is 1000.

The following examples describe how to configure quotas.

- Use the following command to configure a statistical quota for the */mnt/data/* directory to limit the total number of files that reside in a directory.

```
sudo ./quota_tool set --dir /mnt/data/ --accounting --alluser
```

- Use the following command to configure a statistical quota for the */mnt/data/* directory to limit the total number of files that can be created by the user whose UID is 505.

```
sudo ./quota_tool set --dir /mnt/data/ --accounting --uid 505
```

Query quotas

After you configure a quota for an NAS directory, you can query statistical data about the quota for the directory.

1. Log on to an Elastic Compute Service (ECS) instance by using a root account.
2. Use the following command to query quotas.

```
sudo ./quota_tool get --dir /mnt/data/ --all
```

In the preceding command, the --all parameter is optional. If you specify the parameter, statistical data about all quotas that are configured for the file system returns.

Note

- The first time you query a quota, a state called Initializing appears. After the **Initializing** process is complete, you can query the quota and a result showing **success** appears. The duration of the initialization process is based on the number of files and subdirectories in a directory.
- After the initialization process is complete, you can query quotas daily. A delay of 5 to 10 minutes may occur before the expected FileCountReal and SizeReal appear. This occurs due to the asynchronous calculation for quotas at the backend.

The following table lists parameters that are included in a response in the JSON format.

Parameter	Description
Path	Indicates a directory for which you query a quota.
Report	Includes all information about a quota that is specified for a directory, for example, UID and GID.
ReportStatus	The state for the query of a quota.
FileCountLimit	Indicates the limit for the number of files. A value of Empty indicates no limit.
FileCountReal	Indicates the total number of files including subdirectories, files, and special files that reside in a directory.
QuotaType	Accounting indicates a statistical quota and Force indicates a restriction quota.
Uid	Indicates the UID of a user. A value of All indicates all users.
Gid	Indicates the GID of a group. A value of All indicates all groups.
SizeLimit	Indicates the maximum capacity of files that reside in a directory. A value of Empty indicates no limit.
SizeReal	Indicates the total capacity of files that reside in a directory.

Cancel quotas

You can cancel a quota.

1. Log on to an ECS instance.
2. Use the following command to cancel quotas.

   The syntax of the command that you can use to cancel quotas is `sudo ./quota_tool cancel --dir [DIR] [OPTION]`.

Parameter	Description
--dir [DIR]	Specifies the directory for which you want to cancel quotas, for example, -dir /mnt/data/.

Parameter	Description
OPTION	Specifies the required options. **Note** When you specify the OPTION parameter, one of the --alluser, --uid, and --gid options is required. ○ --alluser: specifies a directory-level quota for all users. ○ --uid: specifies the UID of a user. For example, --uid 505 indicates that the quota is canceled for the user whose UID is 505. ○ --gid: specifies the GID of a group. For example, --gid 505 indicates that the quota is canceled for the group whose GID is 505.

The following examples describes how to cancel quotas.

- If you have configured a quota for the */mnt/data/* directory, use the following command to cancel the quota for the user whose UID is 100.

```
sudo ./quota_tool cancel --dir /mnt/data/ --uid 100
```

- If you have configured a quota for the */mnt/data/* directory, use the following command to cancel the quota for all users.

```
sudo ./quota_tool cancel --dir /mnt/data/ --alluser
```

7.1.8. Unified namespace

This topic describes how to create a unified namespace and create a mount target for the unified namespace in the Apsara File Storage NAS console. The topic also describes how to add, remove, and modify file systems in a namespace, view namespace details, and enable the cross-domain mount orchestration feature.

Features

A unified namespace allows you to mount multiple file systems in a NAS cluster by using a single domain name. You do not need to maintain multiple mount targets and mount directories.

You can create a mount target for a unified namespace. You can use a unified namespace to manage multiple file systems the same way you manage a single file system.

A unified namespace contains a virtual root directory in which file systems are the first-level subdirectories. After you add a file system to a unified namespace, you can still mount the file system by using the mount targets of the file system.

Limits

A unified namespace has the following limits:

- You can add a maximum of 1,000 file systems to each unified namespace.
- The mapping name of a file system in a unified namespace cannot exceed 255 characters in length. The name can contain only letters, digits, and the following special characters:

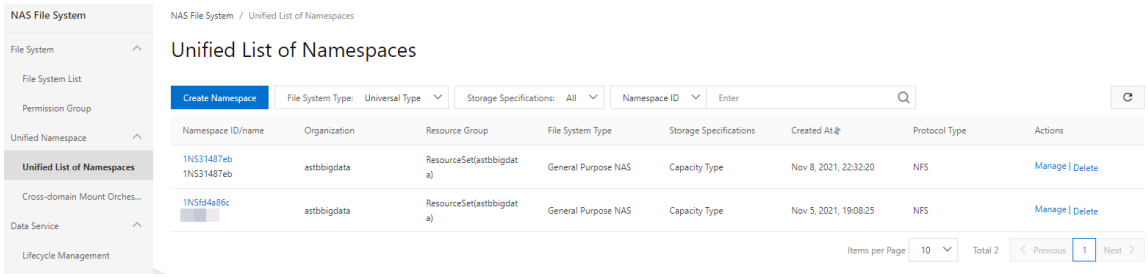
  .+-_()<>@#
- To enable a cross-domain mount orchestration, the mount directory name of a unified namespace cannot exceed 255 characters. The name can contain only letters, digits, and the following special characters:

  .+-_()<>@#
- You can create a maximum of two mount targets for each unified namespace.

- You can create a maximum of 20 namespaces in each region.
- You can mount a unified namespace only by using the NFSv3 protocol.
- The file systems that are added to a namespace must belong to the same Alibaba Cloud account and cluster as the namespace. The storage type, protocol type, and encryption type of the file systems must be the same.
- The mapping name of a file system must be unique in each unified namespace.
- File systems can be mapped only to first-level subdirectories in a unified namespace. You cannot modify the access permissions, owner, or access control lists (ACLs).

Basic features of a unified namespace

- Create a unified namespace and create a mount target for the unified namespace.
  i. Log on to the Apsara File Storage NAS console.
  ii. In the left-side navigation pane, choose **Unified Namespace > Unified List of Namespaces**.
  iii. On the **Unified List of Namespaces** page, click **Create Namespace**. You can then create a namespace and a mount target for the namespace as prompted.



> **Note** We recommend that you use different CIDR blocks if you create mount targets in virtual private clouds (VPCs) for unified namespaces in different regions. This identifies CIDR blocks when you mount unified namespaces across regions. For example, you can use `192.168.0.0/16` for Region 1 and `172.16.0.0/16` for Region 2. For more information, see Cross-domain mount orchestration.

- Add a file system to the unified namespace.

  After you create the unified namespace, you can add a file system and set the mapping name of the file system.

  > **Note** The mapping name is the name of the virtual directory for the file system.

- Remove the file system from the unified namespace.

  You can remove the existing file system from the unified namespace.

  > **Note** The file system is not deleted but removed from the file system list of the namespace.

- Modify the mapping name of the file system.

  You can modify the mapping name of the file system in the unified namespace.

  > **Note** If a symbolic or hard link exists between a file system and another file system, a connection failure may occur when you modify the mapping name of the file system.

- View the details of the unified namespace.

  The details of the namespace are divided into three sections:

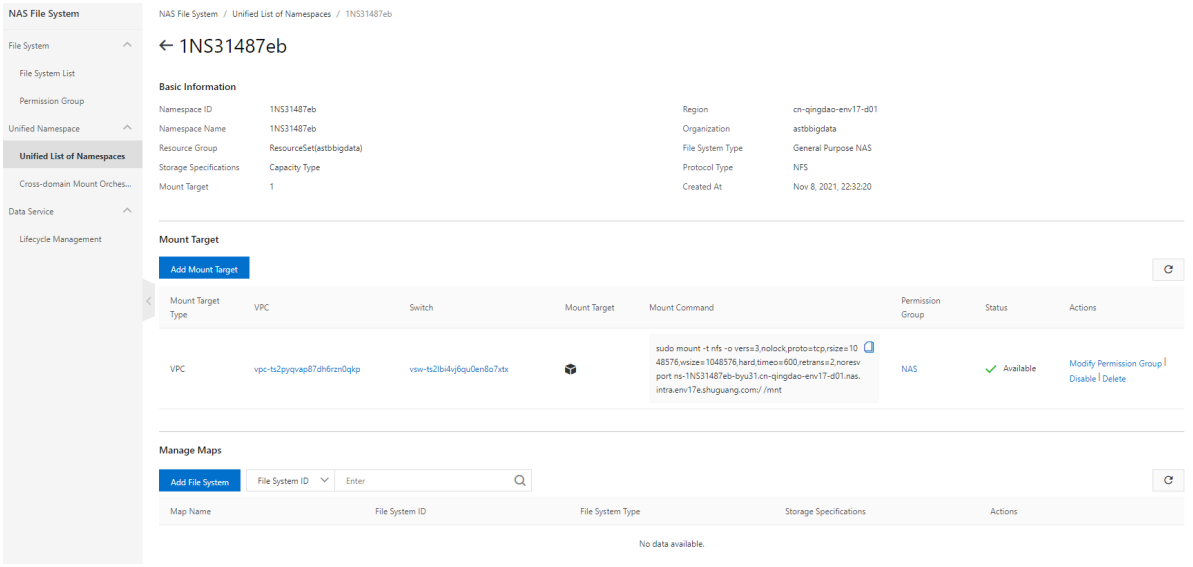
  - Properties

- Mount target list

  > **Note** You can create or delete mount targets.

- File system list

  > **Note** You can add or remove file systems.



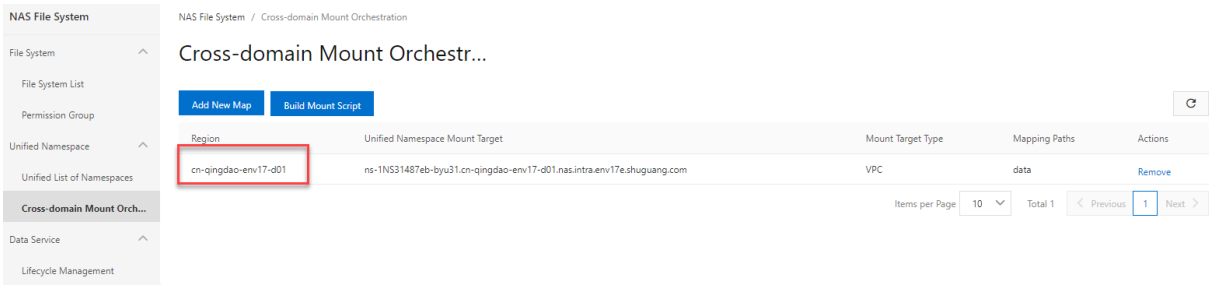
Cross-domain mount orchestration

In traditional solutions, you can add a file system to a namespace only if the file system and namespace reside in the same region. To add file systems to namespaces across regions, NAS provides the cross-domain mount orchestration feature. To use this feature, perform the following operations:

- Create a unified namespace and a mount target for the namespace.
- Map the unified namespace to the local directory tree of a client by using the cross-domain mount orchestration feature.
- After the orchestration is complete, specify the root directory to generate an automatic mount script.

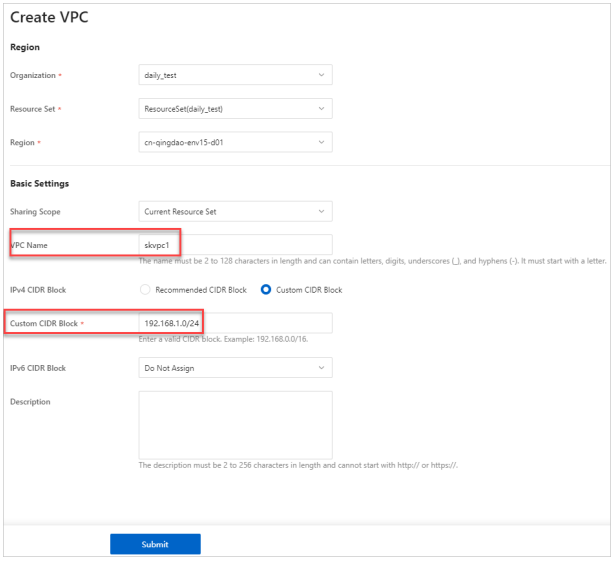
> **Notice** The specified mount target and mapping path cannot be modified after the map is created. You can remove the map. However, if you create the map again, the specified mount target can only be attached to the original mapping path.

To mount the namespace on a client, you can run the automatic mount script on the client. This allows the client to access file systems across regions. The format of the local path on the client for each unified namespace is `<Specified root directory>/<Mapping path of the namespace>`.



To enable a cross-domain mount orchestration for different VPCs that reside in different regions, you must establish VPC connections. The following procedure describes how to establish VPC connections and enable a cross-domain mount by using an example.

1. Create two VPCs.
   i. Log on to the Apsara File Storage NAS console.
   ii. In the left-side navigation pane, choose **Unified Namespace > Cross-domain Mount Orchestration**.
   iii. In the top navigation bar, choose **Products > Networking > Virtual Private Cloud**.
   iv. On the **Create VPC** page, create two VPCs. One VPC is for Region 1 and the other VPC is for Region 2.
   - Create a VPC named skvpc1 `192.168.1.0/24` for Region 1.
   - Create a VPC named skvpc2 `192.168.2.0/24` for Region 2.



2. Configure an Express Connect circuit across regions.

   Configure an Express Connect circuit between the VPCs of the two regions.

   Configure an Express Connect circuit in Region 1 to connect skvpc1 and skvpc2.

   i. Configure the route table for skvpc1.

   Add the CIDR block `192.168.2.0/24` of skvpc2 to the route table of skvpc1.

   a. In the left-side navigation pane, choose **VPCs > Route Tables**.
   b. On the **Route Tables** page, find the instance ID of skvpc1 and click **Manage** in the Actions column.
   c. On the Route Table page, click **Add Route Entry** and set the required parameters. Click **Create VPC-to-VPC Connection** to configure an Express Connect circuit between skvpc1 and skvpc2.

> **Note**
> - Specify the CIDR block of skvpc2 for **Destination CIDR Block**.
> - Select **Router Interface (To VPC)** from the **Next Hop Type** drop-down list.
> - If you configure an Express Connect circuit across regions for the first time, no VPCs are available.

ii. Create a VPC-to-VPC connection

Click Create VPC-to-VPC Connection to go to the **Create Peering Connection** page. Specify the source VPC ID, destination VPC ID, and bandwidth based on your business requirements.

After the VPC-to-VPC connection is created, you are redirected to the VPC-to-VPC page. If the initiator and acceptor are in the Activated state, the connection between skvpc1 and skvpc2 is established.

iii. View the VPC-to-VPC connection.

Return to the **Add Route Entry** page. Click the Refresh icon. The VPC-to-VPC connection that you have created appears in the VPC list.

Configure an Express Connect circuit in Region 2 to connect skvpc1 and skvpc2.

Perform the preceding procedure. In the Add Route Entry panel, select the VPC-to-VPC connection from the VPC list and click **OK**.

The route entry is then added to skvpc2 in Region 2. You can select the VPC-to-VPC connection from the VPC list because you have created a VPC-to-VPC connection. `vpc-6b xxxx42t` is the ID of skvpc1.

3. Add rules for NAS permission groups.

In the top navigation bar, choose **Products > Storage > Apsara File Storage NAS** . In the left-side navigation pane, choose **Permission Group**. On the Permission Group page, find the permission group for which you want to add rules, or create a permission group. Then, click **Add Rules**. In the dialog box that appears, set the authorized address and read and write permissions.

4. Creates a mount target.

In Region 2, you can create a mount target that resides in skvpc2 for the specified unified namespace.

In Region 1, you can create a mount target that resides in skvpc1 for the specified unified namespace.

5. Create an Elastic Compute Service (ECS) instance.

For example, if you select skvpc2 when you create an ECS instance, you can mount file systems on the ECS instance without the need to establish a VPC connection.

> **Note** In most cases, each ECS instance uses an independent VPC. You do not need to establish connections between VPCs that are used by mount targets of namespaces in two regions. However, you must establish connections between the VPC that is used by mount targets of the namespace and the VPC that is used by the ECS instance.

7.1.9. Lifecycle Management

This topic describes how to use the lifecycle management feature in the Apsara File Storage NAS console. The topic also describes how to configure lifecycle management policies and dump cold data to an Infrequent Access (IA) storage medium based on the policies. In the NAS console, you can create, view, and modify lifecycle management policies. You can also query the usage of General-purpose NAS storage and IA storage media.

Prerequisites

A Network File System (NFS) file system of the Capacity or Performance type is mounted on an Elastic Compute Service (ECS) instance. For more information, see Mount an NFS file system.

Context

The lifecycle management feature of NAS allows you to manage hot data and cold data in a tiered manner. You can configure lifecycle management policies and dump infrequently accessed data to an IA storage medium based on the policies to save your storage costs. You can still access the data in the IA storage medium the same way you access the data in a NAS file system.

Limits

The lifecycle management feature supports only NFS file systems. Server Message Block (SMB) file systems and file systems whose data is encrypted are not supported.

Usage notes

In Apsara Stack, the lifecycle management feature allows you to dump cold data from a NAS file system to a specified Object Storage Service (OSS) bucket. However, you must be the owner of the specified NAS file system and OSS bucket.

- You cannot delete the OSS bucket before you retrieve the cold data to the NAS file system. Otherwise, data loss may occur and the associated NAS cluster may become unavailable.
- You cannot revoke permissions that you have granted to NAS by using Resource Access Management (RAM).
- You must minimize the permissions on the OSS bucket to prevent data leaks.
- We recommend that you use an independent OSS bucket for the lifecycle management feature of NAS. This prevents cold data from being accidentally deleted if you store the cold data with other business data in the same OSS bucket.

Preparations

1. Authorize NAS by using RAM.

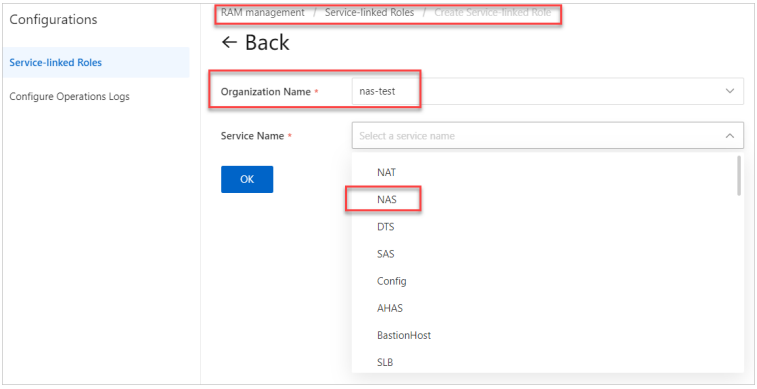
   To use the lifecycle management feature, you must grant access permissions on OSS to NAS. NAS can then read data from and write data to specified OSS buckets.

   > **Notice** When you create a RAM role, you must specify an organization where the file system resides. In this example, the specified organization is bms.
   >
   > You need to grant permissions to NAS only once for an organization.

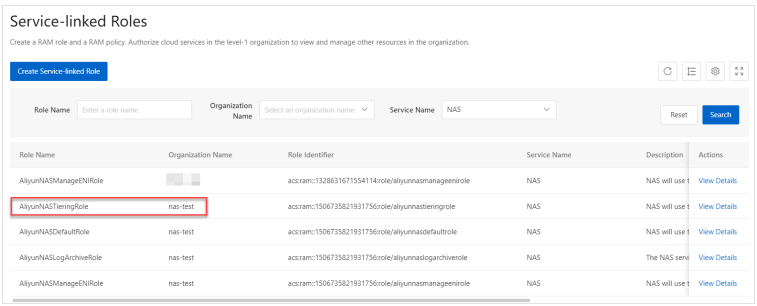
   i. Log on to the Apsara Uni-manager Management Console.

   ii. In the top navigation bar, choose **Configurations > RAM Service Linked Role**.

   iii. On the **RAM Roles** page, click **Create RAM Role** to create a RAM role and grant access permissions on OSS to the role.



   iv. View the authorization status of bms in the role list to confirm that AliyunNASTieringRole is in the list.

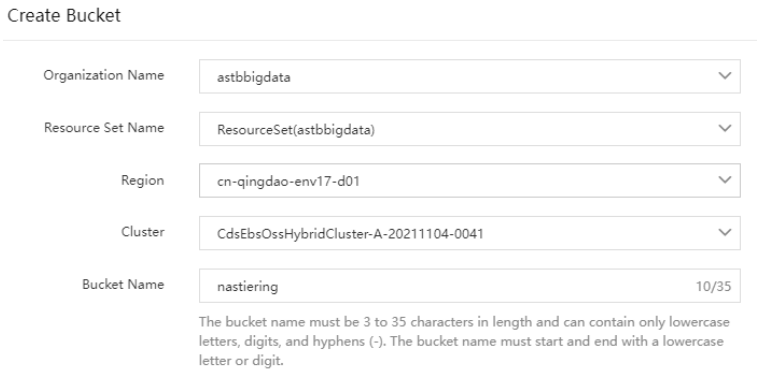


2. Create an OSS bucket.

   You can use an existing OSS bucket or create an OSS bucket. NAS allows you to configure multiple OSS buckets for each file system.

   > **Note** If you use an existing OSS bucket, the following requirements must be met:
   > - The OSS bucket is in the first cluster of the region. The domain name of the first cluster starts with oss- whereas the domain names of the other clusters start with oss*xxxx*.
   > - The OSS bucket and the file system belong to the same organization.

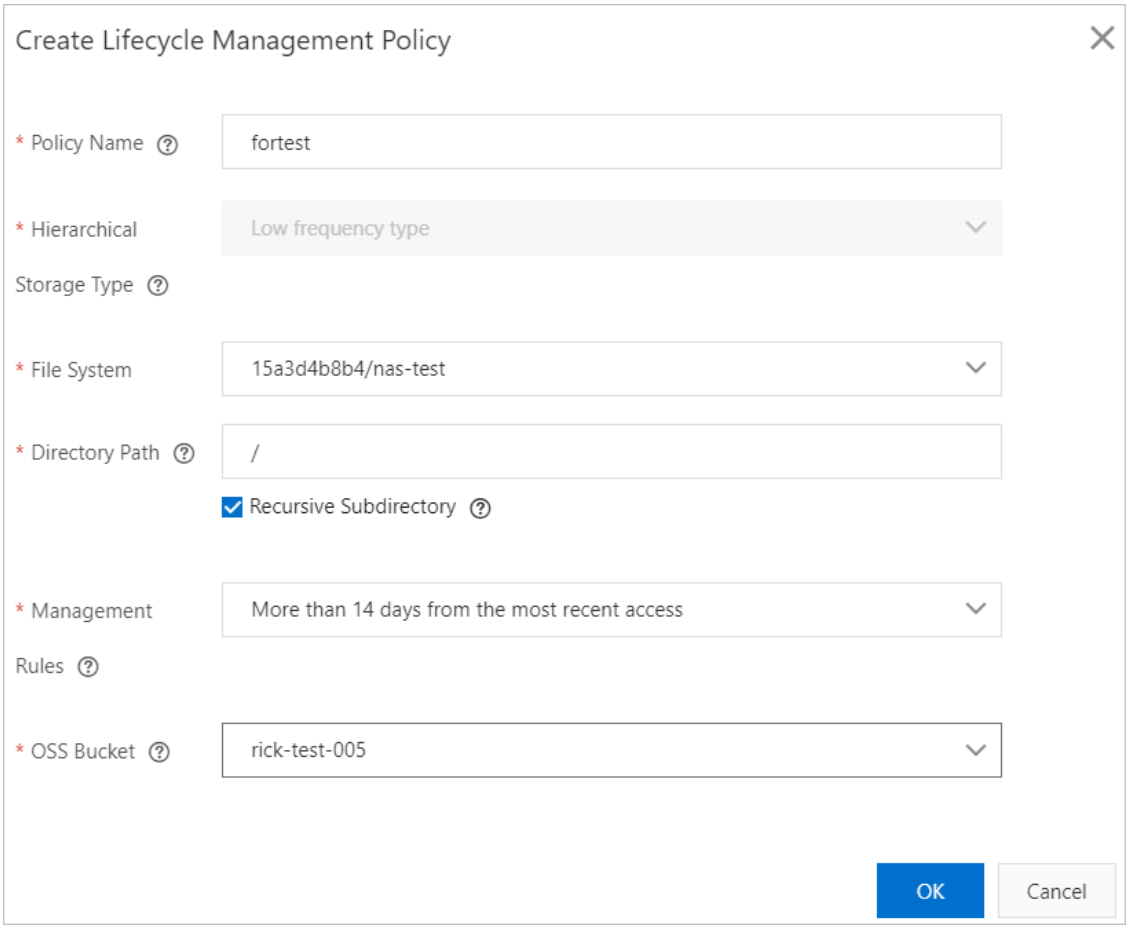
   Log on to the Apsara Uni-manager Operations console. In the top navigation bar, choose **Products > Object Storage Service** . In the left-side navigation pane, choose **+ > Create Bucket**. On the Create OSS Bucket page, set the parameters and create an OSS bucket that is used to store cold data. In this example, an OSS bucket named nastiering is created.

   > **Note** When you create an OSS bucket, the following requirements must be met:
   > - **Storage Class** is selected as **Standard** and the first cluster in the region is selected for the OSS cluster. The domain name of the first cluster starts with oss- whereas the domain names of the other clusters start with oss*xxxx*.
   > - The OSS bucket and the file system belong to the same organization.

Procedure

1. Create a lifecycle management policy.
   i. Log on to the Apsara Uni-manager Management Console.
   ii. In the top navigation bar, choose **Products > Apsara File Storage NAS**.
   iii. In the left-side navigation pane, click **Lifecycle Management**. On the page that appears, click Create Policy to create a lifecycle management policy.



After you create a lifecycle management policy for a directory in the file system, NAS dumps the cold data that meets the rule of the policy to the IA storage medium. When you create a lifecycle management policy, you can set the following parameters:

- Policy Name: the name of the policy. The name must be unique within different policies.
- File System: the file system for which the lifecycle management policy is configured.
- Directory Path: the directory path on the file system. The path must start with a forward slash (/).

  You can enter a forward slash (/) to indicate the root directory.

  If you select **Recursive Subdirectory**, all subdirectories in the directory are recurred.
- Management Rules: the pre-configured rules of the policy. You can select a rule to dump files that have not been accessed for more than 14, 30, 60, or 90 days to an IA storage medium.
- OSS Bucket: the OSS bucket to which cold data is dumped.

  > **Note** Cold data in a file system can be dumped to only one OSS bucket.
  >
  > Only the OSS buckets of the organization where the file system resides are displayed in the OSS Bucket list.

2. View the lifecycle management policy.

   On the **Lifecycle Management** page, you can view the lifecycle management policy that you have created. You can also filter the policies by file system ID.

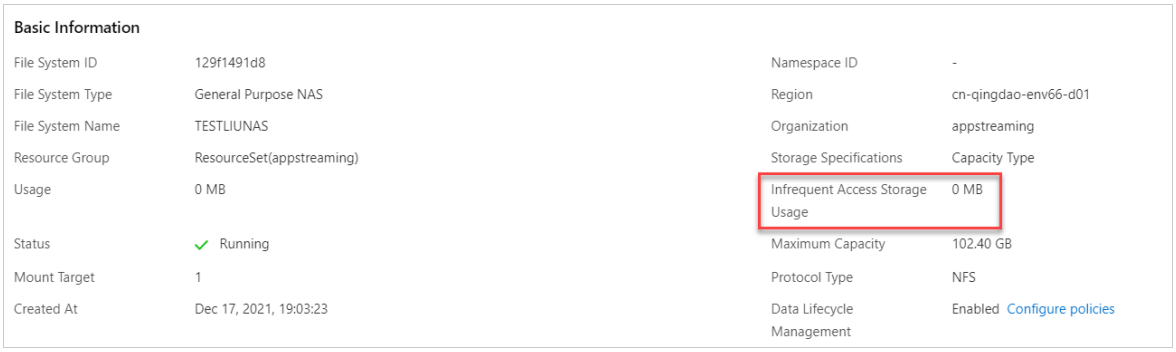
3. Modify the lifecycle management policy.

   On the **Lifecycle Management** page, you can modify the lifecycle management policy. You can modify the following parameters:

- Recursive Subdirectory
- Management Rules

4. Query the usage of the General-purpose NAS storage and IA storage medium.

   In the left-side navigation pane, click **File System List**. You can query the usage of General-purpose NAS storage and IA storage medium within the file system for which you have configured the lifecycle management policy.



7.1.10. Directory-level ACLs that grant the read and write access

7.1.10.1. Overview

Apsara File Storage NAS supports NFSv4 access control lists (ACLs) and Portable Operating System Interface (POSIX) ACLs. This topic describes POSIX ACLs and NFSv4 ACLs. It also lists precautions for using these ACLs.

Access control and user management are important for enterprise-level users who want to share files between different users and groups by using a shared file system. To control access to different files and directories, you can grant users and groups different types of access. NAS provides Network File System (NFS) ACLs to allow you to meet specific requirements. An ACL consists of one or more access control entries (ACEs) that each grant a user or group one or more permissions to access a file or directory.

The NFSv3 protocol includes the extended support for POSIX ACLs. POSIX ACLs extend the support for access control over file mode creation masks. You can grant permissions to specific users and groups besides users of the owner, group, and other classes. Permissions can also be inherited from parent objects. For more information, see acl - Linux man page.

The NFSv4 protocol includes extended support for NFSv4 ACLs that provide more fine-grained access control than POSIX ACLs do. For more information, see nfs4_acl - Linux man page.

You can mount an NFSv3 file system that has NFSv4 ACLs applied. These NFSv4 ACLs will then be converted into POSIX ACLs. You can also mount an NFSv4 file system that has POSIX ACLs applied. These POSIX ACLs will then be converted into NFSv4 ACLs. If you use NFS ACLs, we recommend that you mount NFSv4 file systems and control access by using NFSv4 ACLs rather than file mode creation masks or POSIX ACLs. This is because: NFSv4 ACLs and POSIX ACLs are not fully compatible. The interaction between ACLs and file mode creation masks is not in an ideal state. The file systems that are mounted by using the NFSv3 do not support locks. For more information about NFS ACL features, see Features.

Precautions for using POSIX ACLs

- We recommend that you use the default inheritance method that allows a subdirectory or file to inherit the same ACL from the parent directory. This allows you to avoid configuring another ACL when you create a new file or subdirectory in the parent directory.
- We recommend that you retain a minimum number of ACEs because a file system needs to scan all ACEs each time it performs permission verification. Abuse of ACLs may diminish the performance of file systems.

- Use caution when you configure ACLs by using the recursive method (`setfacl -R`). Large amounts of metadata are produced when you perform a recursive operation on a directory that contains a large number of files and subdirectories. This may affect your businesses.
- Before you configure ACLs, we recommend that you manage groups and related permissions. For example, you can add a user to one or more groups. If you want to add, remove, or modify permissions for a user, we recommend that you move the user to a group that has the required permissions. You do not need to modify the ACL of a group as long as the structure of groups remains unchanged. We recommend that you configure ACLs for groups rather than single users. This provides a simple and effective time-saving method to control access and ensure the better organization of permissions.
- You can apply a POSIX ACL to multiple objects that resides on different clients. In such cases, you must ensure that the ACL you apply to each object is the same. Apsara File Storage NAS stores user IDs (UIDs) and group IDs (GIDs) at the backend. You must ensure that the mappings between a username or group name and a UID or GID are the same.
- We recommend that you grant the least permissions to the other class because all users have the permissions that are granted to the other class. A potential security vulnerability may be exposed if the other class has more permissions than any ACE.
- We recommended that you configure the least permissions for the other class. Before creating files or directories, you can use the `umask 777` command to configure the file mode creation mask. This command sets the file mode creation mask to 000 when the mask is used as a parameter to create a new file or directory. This ensures that the newly created file or directory has the least permissions. For more information, see umask and the default file mode creation mask.
- After you enable POSIX ACLs, the semantics of the other class for the POSIX ACL are equal to the semantics of the EVERYONE@ principal. The semantics of the other class for the file mode creation mask are also equal to the semantics of the EVERYONE@ principal. When a system performs permission verification, the system treats the other class the same as the EVERYONE@ principal.

Precautions for using NFSv4 ACLs

- Use UIDs or GIDs such as UID 1001 to configure ACLs.
- We recommend that you do not configure the file mode creation mask after you configure an NFSv4 ACL.
- The `nfs4_setfacl` command provides -a, -x, -m, and other options. You can use these options to add, remove, or modify ACEs. However, we recommend that you use `nfs4_setfacl -e <file>` the command to edit an ACL in an interactive mode.
- We recommend that you configure the least permissions for the EVERYONE@ principal because NFSv4 ACLs only support allow rather than deny ACEs. A potential security vulnerability may be exposed if the EVERYONE@ principal has more permissions than other ACEs.
- NFSv4 ACLs have fine-grained permissions. In most cases, it is unnecessary to subdivide permissions at such a fine-grained level. For example, if you have the write access (w) to a file but do not have the append-only (a) access, an error may occur when you write data to the file. The same issue occurs for a directory. To avoid unexpected permission errors, we recommend that you specify a capital w (W) as a parameter when you use the `nfs4_setfacl` command to configure an ACL. The `nfs4_setfacl` command converts W to a full write access permission. For a file, W is expanded to wadT. For a directory, W is expanded to wadTD.
- We recommend that you use the default inheritance method that allows a subdirectory or file to inherit the same ACL from the parent directory. This allows you to avoid configuring another ACL when you create a new file or subdirectory in the parent directory.
- We recommend that you retain a minimum number of ACEs because a file system needs to scan all ACEs each time it performs permission verification. Abuse of ACLs may diminish the performance of file systems.
- Use caution when you configure ACLs by using the recursive method (`nfs4_setfacl -R`). Large amounts of metadata are generated when you perform a recursive operation on a directory that contains a large number of files and subdirectories. This may affect your businesses.
- Before you configure ACLs, we recommend that you manage groups and related permissions. For example, you can add a user to one or more groups. If you want to add, remove, or modify permissions for a user, we recommend that you move the user to a group that has the required permissions. You do not need to modify the

ACL of a group as long as the structure of groups remains unchanged. We recommend that you configure ACLs for groups rather than single users. This provides a simple and effective time-saving method to control access and ensure the better organization of permissions.

7.1.10.2. Features

This topic describes the features of NFSv4 access control lists (ACLs) and POSIX ACLs.

Features of Apsara File Storage NAS NFSv4 ACLs

- Only access control entries (ACEs) of the allow type are supported. ACEs of the following types are not supported: deny, audit, and alarm.

  Deny ACEs increase the complexity of access control. In most cases, complexity leads to confusion and increases potential security risks. As agreed by the industry, we recommend that you avoid using deny ACEs. For more information about why deny ACEs are not recommended, see FAQ.

  Audit and alarm ACEs are not available for NFS file systems. Instead, you can audit file systems and configure alerts based on auditing results in the NAS console.

- If no ACL is specified for a file or a directory, the default ACL that corresponds to the predefined file mode creation mask is applied.

  ```
  touch file
  ```

  ```
  [root@vbox test]# ls -l file
  -rw-r--r--. 1 root root 0 May  6 14:27 file
  ```

  ```
  [root@vbox test]# nfs4_getfacl file
  # file: file
  A::OWNER@:rwatTnNcCy
  A::GROUP@:rtncy
  A::EVERYONE@:rtncy
  ```

- An ACL is an ordered list that contains and deduplicates ACEs. This scheme ensures that permissions defined in an ACL are clear and informative.

  If you apply both a new ACE and an existing ACE to the same object and the existing ACE is inherited from the parent object, the permissions of the new ACE override the permissions of the existing ACE. For example:

  - In most cases, ACEs that include the following principals are queued in sequence at the beginning of an ACL: OWNER@, GROUP@, and EVERYONE@. These ACEs take precedence over other ACEs.

    ```
    [root@vbox test]# nfs4_getfacl file
    # file: file
    A::OWNER@:rwaxtTnNcCy
    A::GROUP@:rtncy
    A::EVERYONE@:rtncy
    A::1001:rwaxtTNcCy
    ```

  - Add an ACE of the read and write permissions to the following ACL for a user principal named 1009. The ACE is placed after the ACE that is defined for a user principal named 1001 based on the predefined order.

    ```
    [root@vbox test]# nfs4_setfacl -a A::1009:X file
    [root@vbox test]# nfs4_getfacl file
    # file: file
    A::OWNER@:rwaxtTcCy
    A::GROUP@:rwatcy
    A::EVERYONE@:tcy
    A::1001:rwaxtTNcCy
    A::1009:xtcy
    ```

- Add a new ACE that includes the execute permission to the ACL for the user principal named 1009. The system automatically merges the execute permission into the existing ACE for the 1009 user principal.

```
[root@vbox test]# nfs4_setfacl -a A::1009:W file
[root@vbox test]# nfs4_getfacl file
# file: file
A::OWNER@:rwaxtTcCy
A::GROUP@:rwatcy
A::EVERYONE@:tcy
A::1001:rwaxtTNcCy
A::1009:waxtTncCy
```

- Add the f and d inheritance flags to an ACE that includes a user principal named 1009. Then, the system splits the ACE into two ACEs. One ACE has an extra inheritance flag named i specified, which indicates an inherit-only ACE. The other ACE only applies to the file object without inheritance flags. If the inheritance type of an existing ACE matches the type for one of the two ACEs, the system combines the existing ACE with the ACE out of the two ACEs. The two matching ACEs are combined into one ACE.

```
[root@vbox test]# nfs4_setfacl -a A:fd:1009:R file
[root@vbox test]# nfs4_getfacl file
# file: file
A::OWNER@:rwaxtTcCy
A::GROUP@:rwatcy
A::EVERYONE@:tcy
A::1001:rwaxtTNcCy
A::1009:rwaxtTNcCy
A:fdi:1009:r
```

- All ACEs can be inherited.
  i. For example, the OWNER@ principal has the write access, the GROUP@ principal has the read access, and the EVERYONE@ has no access to the dir directory.

```
[root@vbox nfs]# nfs4_getfacl dir
# file: dir
A::OWNER@:rwaDxtTnNcCy
A::GROUP@:rxtcy
A::EVERYONE@:tncy
```

  ii. Add an ACE that grants a user principal named 1000 the read, write, and execute access to the dir directory. The f and d inheritance flags are also specified for the ACE.

```
[root@vbox nfs]# nfs4_setfacl -a A:fd:1000:rwx dir
[root@vbox nfs]# nfs4_getfacl dir
# file: dir
A::OWNER@:rwaDxtTcCy
A::GROUP@:rxtcy
A::EVERYONE@:tcy
A::1000:rwx
A:fdi:1000:rwx
```

  iii. When you create a file or subdirectory in the dir directory, the file or the subdirectory automatically inherits all ACEs from the dir directory.

```
[root@vbox nfs]# touch dir/file
[root@vbox nfs]# nfs4_getfacl dir/file
# file: dir/file
A::OWNER@:rwatTcCy
A::GROUP@:rwatcy
A::EVERYONE@:rwatcy
A::1000:rwx
```

```
[root@vbox nfs]# mkdir dir/subdir
[root@vbox nfs]# nfs4_getfacl dir/subdir
# file: dir/subdir
A::OWNER@:rwaDxtTcCy
A::GROUP@:rwaDxtcy
A::EVERYONE@:rwaDxtcy
A:fdi:1000:rwx
```

Note

- We recommend that you grant the least privileges to the EVERYONE@ principal. Before you perform the following steps, we recommend that you run the `umask 777` command. This command ensures that no access to a file or directory is granted when the file or directory is created. For more information, see Why doesn't umask change execute permissions on files?
- When Linux calls functions to create files or directory, the predefined file mode creation mask is used as a request parameter. You can obtain the final ACL for a child object from the overlap of the inherited ACL (parent to child) and the file mode creation mask, as specified in the RFC7530 standard. When you modify the group bits of a file mode creation mask based on the standard, permissions included in an ACL for each group must be less than or equal to permissions defined in group bits. However, this scheme results in an invalid inheritance for groups. For example, you create a file and the file attempts to inherit A:RWX from a parent object. However, the predefined file mode creation mask sets the group bits to R. The final permission for the file becomes A:R. In actual practice, we recommend that you only modify file mode creation masks for ACLs that include the following principals: OWNER@, GROUP@, and EVERYONE@. This prevents against potential issues and ensures that semantics are clear. To remove permissions for a group, we recommend that you remove the ACE that relates to the group.

- You need to manage mappings between usernames or group names and user IDs (UIDs) or group IDs (GIDs) across multiple independent instances.

  NAS NFS adopts IP security groups rather than usernames to authenticate users. When you configure NFSv4 ACLs, UIDs or GIDs that are included in ACEs are stored in Linux. When you print an ACL for an object in a shell, Linux automatically loads the */etc/passwd* file and converts UIDs or GIDs into usernames or group names. You need to manage mappings between usernames or group names and UIDs or GIDs across multiple instances. You must ensure a username or group name is mapped to its UID or GID.

- NFSv4 ACLs can be printed by using extended attributes.

```
[root@vbox nfs]# getfattr -n system.nfs4_acl file
# file: file
system.nfs4_acl=0sAAAABgAAAAAAAAAAABYBhwAAAAZPV05FUkAAAAAAAAAAAAAAAABIAhwAAAAZHUk9VUEAAAAAAAAAAAAAAAA
BIAhwAAAAlFVkVSWU9ORUAAAAAAAAAAAAAAAAAAAAAEAAAAEMTAwMAAAAAAAAAALAAAAAwAAAAQxMDAwAAAAAAAAAEAAFgGQAAAA
BTEwMDAxAAAA
```

- Tools such as cp are supported for migrating NFSv4 ACLs.

  NAS allows you to migrate NFSv4 ACLs by using the cp, tar, and rsync tools. For more information, see How to preserve NFS v4 ACLs via extended attributes when copying file.

  The following `cp --preserve=xattr file1 file2` command makes a copy of the file1 file as the file2 file while making a copy of the ACL of the file1 file for the file2 file. The `cp -ar dir1 dir2` command makes a copy of the dir1 directory as the dir2 directory while making a copy of the ACL of the dir1 directory for the dir2 directory.

  Note You may fail to migrate NFSv4 ACLs if the version of the rsync tool is earlier than 3.1.2.

```
[root@vbox nfs]# nfs4_getfacl file1
# file: file1
A::OWNER@:rwatTcCy
A::GROUP@:rwatcy
A::EVERYONE@:rwatcy
A::1000:rtncy
[root@vbox nfs]# cp --preserve=xattr file1 file2
```

```
[root@vbox nfs]# nfs4_getfacl file2
# file: file2
A::OWNER@:rwatTcCy
A::GROUP@:rwatcy
A::EVERYONE@:rwatcy
A::1000:rtncy
[root@vbox nfs]# cp -ar dir1 dir2
```

- Interaction between NFSv4 ACLs and file mode creation masks is supported. The modification for the ACL of an object may change the file mode creation mask of the object. The modification for the file mode creation mask of an object may change the ACL of the object.

  For example, the file mode creation mask of the file object is 0666.

  ```
  [root@vbox nfs]# ls -l file
  -rw-rw-rw-. 1 root root 0 May  3  2019 file
  [root@vbox nfs]# nfs4_getfacl file
  # file: file
  A::OWNER@:rwatTcCy
  A::GROUP@:rwatcy
  A::EVERYONE@:rwatcy
  ```

  - If you add the execute permission to the file mode creation mask by modifying the owner bits, the execute permission is also added to the ACE that includes the OWNER@ principal.

    ```
    [root@vbox nfs]# chmod u+x file
    [root@vbox nfs]# ls -l file
    -rwxrw-rw-. 1 root root 0 May  3  2019 file
    [root@vbox nfs]# nfs4_getfacl file
    # file: file
    A::OWNER@:rwaxtTcCy
    A::GROUP@:rwatcy
    A::EVERYONE@:rwatcy
    ```

  - If you add the execute permission to an ACE that includes the GROUP@ principal, the execute permission is also added to the related file mode creation mask.

    ```
    [root@vbox nfs]# nfs4_setfacl -a A::GROUP@:x file
    [root@vbox nfs]# ls -l file
    -rwxrwxrw-. 1 root root 0 May  3  2019 file
    ```

> **Note**
> - In the interaction between ACLs and file mode creation masks, the EVERYONE@ principal is equal to the others class. When you modify the others class, the change also applies to the EVERYONE@ principal. This operation results in a slight impact on the semantics of permissions. For example, the current file mode creation mask is 177. After you run the `chmod o+r` command, all users that include the file owner and group members have the read permission. This occurs because the read permission is added to the related ACE that includes the EVERYONE@ principal. If no change is applied to the default file mode creation mask, the owner and group classes still have no read permission after you run the chmod o+r command.
> - If no change is applied to NFSv4 ACLs, the others class of the file mode creation mask keeps the same semantics. If an NFSv4 ACL is changed, the semantics of the others class change to the semantics of the EVERYONE@ principal and the latest semantics remain. We recommend that you do not use file mode creation masks after using NFSv4 ACLs.

- Interaction between NFSv4 ACLs and POSIX ACLs is supported.

  You can mount NFSv3 file systems that have NFSv4 ACLs applied. These NFSv4 ACLs will then be converted into POSIX ACLs. You can also mount NFSv4 file systems that have POSIX ACLs applied. These POSIX ACLs will then be converted into NFSv4 ACLs.

  > **Note** The semantics of POSIX ACLs are different from the semantics of NFSv4 ACLs. For example, the inheritance rules that apply to POSIX ACLs do not differentiate files and directories. NFSv4 ACLs have more permissions than POSIX ACLs, which have only read, write, and execute permissions. We recommend that you use either NFSv4 ACLs or POSIX ACLs to prevent against potential issues.

  For example, you configure an NFSv4 ACL for the dir0 directory. The permissions are listed as follows.

```
[root@vbox test] sudo nfs4_getfacl dir0
A::OWNER@:tTnNcCy
A::GROUP@:tncy
A::EVERYONE@:tncy
A:fdi:EVERYONE@:tncy
A:fdi:OWNER@:tTnNcCy
A:fdi:GROUP@:tncy
A:g:19064:rxtncy
A:g:19065:rwaDxtTnNcCy
A:fdig:19064:rxtncy
A:fdig:19065:rwaDxtTnNcCy
```

  You configure a POSIX ACL for the dir0 directory. The permissions are listed as follows.

```
[root@vbox test] sudo getfacl dir0
user::---
group::---
group:players:r-x
group:adminis:rwx
mask::rwx
other::---
default:user::---
default:group::---
default:group:players:r-x
default:group:adminis:rwx
default:mask::rwx
default:other::---
```

  For example, you configure an NFSv4 ACL for the dir0/file file. The permissions are listed as follows.

```
[root@vbox test] sudo nfs4_getfacl dir0/file
A::OWNER@:tTnNcCy
A::GROUP@:tncy
A::EVERYONE@:tncy
A:g:19064:rxtncy
A:g:19065:rwaxtTnNcCy
```

For example, you configure a POSIX ACL for the dir0/file file. The permissions are listed as follows.

```
[root@vbox test] sudo getfacl dir0/file
user::---
group::---
group:players:r-x
group:adminis:rwx
mask::rwx
other::---
```

- The number of NFSv4 ACLs is limited.

  NAS supports a maximum of 100,000 ACLs that are different from one another in each file system. Each ACL contains a maximum of 500 ACEs.

  > **Note** We recommend that you do not abuse ACLs and ACEs. This reduces the time and resources consumed for verifying permissions.

Features of NAS POSIX ACLs

- Permissions that are specified for the other class apply to all users.

  Everyone includes the owner, group, and users that are related to each ACE. The other class is equal to the EVERYONE@ principal of an NFSv4 ACL.

  > **Note** We recommend that you grant the least permissions to the other class in all cases.

  For example, the following ACL is configured for the *myfile* file. Although the ACE contains a user named alice who does not have the write permission, the write permission propagates to the ACE because the permission is specified for the other class.

```
[root@vbox 3]# getfacl myfile
# file: myfile
# owner: root
# group: root
user::rw-
user:alice:r--
group::r--
mask::r--
other::rw-
```

- Permissions that are configured by ACLs will not be changed after you run the `chmod` command.

  > **Note** We recommend that you avoid modifying the file mode creation mask of a file that has a POSIX ACL applied. You can configure permissions for the file by modifying the POSIX ACL.

  i. For example, an ACE that grants the players group the read and write access to the *myfile* file.

```
[root@vbox 3]# getfacl myfile
# file: myfile
# owner: root
# group: root
user::rw-
user:player:rw-
group::rw-
group:players:rw-
mask::rw-
other::---
```

ii. The `chmod g-w myfile` or `chmod u-w myfile` command does not change the permissions that are granted to the player user and the players group, which is different from the POSIX ACL standard. However, this ensures that permissions that are granted by POSIX ACLs to non-reserved users are the same after you modify permissions by using file mode creation masks. The non-reserved users include all users except for the users of the owner, group, and other classes.

```
[root@vbox 3]# getfacl myfile
# file: myfile
# owner: root
# group: root
user::r--
user:player:rw-
group::r--
group:players:rw-
mask::rw-
other::---
```

- If the execute permission is not granted to the group and other classes of an ACL, the ACL has no execute permission.

  The rule is predefined in Linux. The execute action is allowed by the backend of NAS. However, to make the execute permission in the ACL effective, you must grant the execute permission to the group or other class.

  For example, if the group and other classes do not have the execute access to the *myfile* file, the player user cannot execute the file.

  ```
  [root@vbox 3]# getfacl myfile
  # file: myfile
  # owner: root
  # group: root
  user::rw-
  user:player:r-x
  group::r--
  mask::r-x
  other::r--
  ```

  If you grant the execute permission to the group class, the execute permission also propagates to the player user.

  ```
  [root@vbox 3]# getfacl myfile
  # file: myfile
  # owner: root
  # group: root
  user::rw-
  user:player:r-x
  group::r-x
  mask::r-x
  other::r--
  ```

- If you configure inheritable NFSv4 ACLs for directories, these settings may not conform to the POSIX ACL

standard when these directories reside in NFSv3 file systems.

Inheritance rules that apply to files are different from those that apply to directories in NFSv4 ACLs. The same inheritance rules apply to both files and directories in POSIX ACLs.

> **Note** We recommend that you apply either NFS4 ACLs or POSIX ACLs to an NFS file system to prevent against potential issues.

- File mode creation masks cannot be modified.

  The file mode creation mask of a POSIX ACL is yielded by the combination and interaction of permissions from all users and groups. The mask has no practical meaning and cannot be changed.

- You need to manage mappings between usernames or group names and user IDs (UIDs) or group IDs (GIDs) across multiple instances.

  Apsara File Storage NAS NFS adopts IP security groups rather than usernames to authenticate users. When you configure POSIX ACLs, UIDs or GIDs that are included in ACEs are stored in Linux. When you print an ACL for an object in a shell, Linux automatically loads the */etc/passwd* file and converts UIDs or GIDs into actual usernames or group names. You need to manage mappings between usernames or group names and UIDs or GIDs across multiple instances. You must ensure a username or group name is mapped to its related UID or GID.

- POSIX ACLs can be printed by using extended attributes.

```
[root@vbox nfs]# getfattr -n system.posix_acl_access file
# file: file
system.posix_acl_access=0sAgAAAAEAAAD/////AgAFACAEAAAEAAAA/////xAABQD/////IAABAP////8=
```

- POSIX ACLs can be migrated by using tools such as cp.

  NAS allows you to migrate POSIX ACLs by using the cp, tar, and rsync tools. For more information, see How to preserve NFS v4 ACLs via extended attributes when copying file.

  The following `cp --preserve=xattr file1 file2` command makes a copy of the file1 file as the file2 file while making a copy of the ACL of the file1 file for the file2 file. The `cp -ar dir1 dir2` command makes a copy of the dir1 directory as the dir2 directory while making a copy of the ACL of the dir1 directory for the dir2 directory.

  > **Note** You may fail to migrate POSIX ACLs if the version of the rsync tool is earlier than 3.1.2.

```
[root@vbox nfs]# getfacl file1
user::---
user:player:r-x
group::---
mask::r-x
other::--x
[root@vbox nfs]# cp --preserve=xattr file1 file2
```

```
[root@vbox nfs]# getfacl file2
# file: file2
user::---
user:player:r-x
group::---
mask::r-x
other::--x
[root@vbox nfs]# cp -ar dir1 dir2
```

- The number of POSIX ACLs is limited.

  NAS supports a maximum of 100,000 ACLs that are different from one another in each file system. Each ACL contains a maximum of 500 ACEs.

> **Note** We recommend that you do not abuse ACLs and ACEs. This reduces the time and resources consumed for verifying permissions.

FAQ

Why are deny ACEs not supported?

- The position of an ACE that resides in an ACL is important.

  The sequence for ACEs that reside in an NFSv4 ACL is random. A deny ACE may be placed in any position of an NFSv4 ACL. For example, an ACL contains two ACEs: A::Alice:r and D::Alice:r. The position of the ACEs determines whether the user named Alice has the write permission.

  > **Note** When you configure an ACL, you must consider the position of each ACE.

- The number of ACEs in an ACL experiences a sharp increase.

  You may have difficulties to combine and deduplicate ACEs in an ACL because the sequencing for ACEs is not mandatory. The number of ACEs may increase up to tens or hundreds over a long period of time. To manage the final permissions that are produced by these ACEs, you need to check each ACE. The process to check is strenuous and time-consuming.

- The interactions between file mode creation masks and ACLs become more complex after deny ACEs are applied because deny features do not exist in file mode creation masks.
  - If deny ACEs are available, you may need to add several ACEs to an ACL when the file mode creation mask is changed. For example, if you change the file mode creation mask to -rw-rw-rw, you need to add the following ACEs to an ACL. You must add the ACEs in sequence at the beginning of the ACL.

    ```
    A::OWNER@:rw
    D::OWNER@:x
    A::GROUP@:rw
    D::GROUP@:x
    A::EVERYONE@:rw
    D::EVERYONE@:x
    ```

  - If deny ACEs are unavailable, you can sequence and deduplicate ACEs. You do not need to differentiate the EVERYONE@ principal and the other class. You can modify an ACL with ease when the file mode creation mask is changed. In such cases, you only need to find ACEs that contain the OWNER@, GROUP@, and EVERYONE@ principals and modify these ACEs as follows.

    ```
    A::OWNER@:rw
    A::GROUP@:rw
    A::EVERYONE@:rw
    ```

- Conversions between NFSv4 ACLs and POSIX ACLs are not supported in some cases.

  POSIX ACLs do not support deny ACEs. If deny ACEs are included in an NFSv4 ACL, you cannot convert the ACL into a POSIX ACL.

7.1.10.3. Use POSIX ACLs to control access

This topic describes how to configure Portable Operating System Interface (POSIX) access control lists (ACLs). You can use POSIX ACLs to control access to files and directories that reside in an NFSv3 file system.

Prerequisites

An NFSv3 file system is mounted. For more information, see Mount an NFS file system.

Commands

Before you configure POSIX ACLs, we recommend that you familiarize yourself with the related commands.

Command	Description
getfacl <filename>	Shows the ACL that applies to the specified file.
setfacl -m g::w <filename>	Grants the owing group the write access.
setfacl -m u:player:w <filename>	Grants the player user the write access.
setfacl -m g:players:rwx <filename>	Grants the players group the read, write, and execute access.
setfacl -x g:players <filename>	Removes permissions from the players group
getfacl file1 | setfacl --set-file=- file2	Copies the ACL for the *file1* file to the *file2* file.
setfacl -b file1	Removes all extended ACEs from the *file1* file. The base ACEs of the owner, group, and others are retained.
setfacl -k file1	Removes all default ACEs from the *file1* file.
nfs4_setfacl -R -m g:players:rw dir	Grants the players group the read and write access to files and subdirectories in the *dir* directory.
setfacl -d -m g:players:rw dir1	Grants the players group the read and write access to the newly created files and subdirectories in the *dir1* directory.

Procedure

To control access to files and directories by configuring NFS ACLs, follow these steps.

1. Create users and groups.

   In this example, the following users are created: player, admini, and anonym. The following groups are created: players and adminis. The player user is added to the players group and the admini user is added to the adminis group.

   ```
   sudo useradd player
   sudo groupadd players
   sudo usermod -g players player
   sudo useradd admini
   sudo groupadd adminis
   sudo usermod -g adminis admini
   sudo useradd anonym
   ```

2. Configure POSIX ACLs to control access to files and directories.

   Use the following commands to complete the operations: create a directory named *dir0* and grant the players group the read-only access, the adminis group the read, write, and execute permissions, and the others class no access to all the files in the *dir0* directory.

```
sudo umask 777
sudo mkdir dir0
sudo setfacl -m g:players:r-x dir0
sudo setfacl -m g:adminis:rwx dir0
sudo setfacl -m u::--- dir0
sudo setfacl -m g::--x dir0
sudo setfacl -m o::--- dir0
sudo setfacl -d -m g:players:r-x dir0
sudo setfacl -d -m g:adminis:rwx dir0
sudo setfacl -d -m u::--- dir0
sudo setfacl -d -m g::--x dir0
sudo setfacl -d -m o::--- dir0
```

Use the `sudo getfacl dir0` command to verify the result after the configuration is complete.

```
# file: dir0
# owner: root
# group: root
user::---
group::--x
group:players:r-x
group:adminis:rwx
mask::rwx
other::---
default:user::---
default:group::--x
default:group:players:r-x
default:group:adminis:rwx
default:mask::rwx
default:other::---
```

3. Verify the ACL configuration.
   i. Verify that the admini user has read and write access to the dir0/file file.

```
[root@vbox test] sudo su admini -c 'touch dir0/file'
[root@vbox test] sudo su admini -c 'echo 123 > dir0/file'
```

   ii. Use the following command to verify the read-only access of the player user.

```
[root@vbox test] sudo su player -c 'touch dir0/file'
touch: cannot touch ‘dir0/file’: Permission denied
[root@vbox test] sudo su player -c 'cat dir0/file'
123
[root@vbox test] sudo su player -c 'echo 456 >> dir0/file'
bash: dir0/file: Permission denied
[root@vbox test] sudo su player -c 'getfacl dir0/file'
# file: dir0/file
# owner: admini
# group: adminis
user::---
group::---
group:players:r-x
group:adminis:rwx
mask::rwx
other::---
```

iii. Verify that the anonym user does not have access to the dir0/file file.

```
[root@vbox test] sudo su anonym -c 'ls dir0'
ls: cannot open directory dir0: Permission denied
[root@vbox test] sudo su anonym -c 'cat dir0/file'
cat: dir0/file: Permission denied
[root@vbox test] sudo su anonym -c 'getfacl dir0/file'
getfacl: dir0/file: Permission denied
```

Related operations

If you want to remove user permissions, use the following method.

When you use NFSv4 ACLs, we recommend that you sort each user into different groups. This allows you to configure permissions for a group rather than a separate user. To disable access to an object from a user, you can remove the user from a group that has access to the object. For example, the following commands remove the admini user from the adminis group and add the user to the adminis2 group.

```
[root@vbox test] sudo groupadd adminis2
[root@vbox test] sudo usermod -g adminis2 admini
[root@vbox test] id admini
uid=1057(admini) gid=1057(admini) groups=1061(adminis2)
[root@vbox test] sudo su admini -c 'ls dir0'
ls: cannot open directory dir0: Permission denied
[root@vbox test] sudo su admini -c 'cat dir0/file'
cat: dir0/file: Permission denied
[root@vbox test] sudo su admini -c 'getfacl dir0/file'
getfacl: dir0/file: Permission denied
```

7.1.10.4. Use NFSv4 ACLs to control access

This topic describes how to configure NFSv4 access control lists (ACLs) and apply these ACLs to NFSv4 file systems to control access to files or directories in the file systems.

Prerequisites

An NFSv4 file system is mounted. For more information, see Mount an NFS file system.

Context

You can mount an NFSv4 file system on a Linux Elastic Compute Service (ECS) instance and install the Linux-specific nfs4-acl-tools tool on the instance. After the tool is installed, you can use the standard nfs4_getfacl and nfs4_setfacl tools to configure NFSv4 ACLs.

Commands

Before you configure NFSv4 ACLs, we recommend that you get familiar with the related commands.

Command	Description
`nfs4_getfacl <filename>`	Views the access permissions for a specified file.
`nfs4_setfacl -a A::GROUP@:W <filename>`	Adds an access control entry (ACE) that grants the write permissions on a specified file to the GROUP@ principal.
`nfs4_setfacl -a A::1000:W <filename>`	Adds an ACE that grants the write permissions on a specified file to a user principal named 1000.
`nfs4_setfacl -a A:g:10001:W <filename>`	Adds an ACE that grants the write permissions on a specified file to a user principal named 10001.

Command	Description
`nfs4_setfacl -e <filename>`	Edits an ACL in an interactive mode.
`nfs4_getfacl <filename> > saved_acl.txt`	Saves a list of permissions for a specified file as a TXT file.
`nfs4_setfacl -S saved_acl.txt <filename>`	Configures permissions for a specified file by using a TXT file that includes a list of permissions.
`nfs4_setfacl -m A::1001:rwaxTNcCy A::1001:rxtcy file1`	Modifies the permission of an ACE that applies to the *file1* file.
`nfs4_getfacl file1 | nfs4_setfacl -S - file2`	Copies the permissions for the *file1* file to the *file2* file.
`nfs4_getfacl file1 | grep @ | nfs4_setfacl -S - file1`	Deletes all ACEs that apply to the *file1* file except for ACEs that include the following principals: OWNER@, GROUP@, and EVERYONE@.
`nfs4_setfacl -R -a A:g:10001:rW dir`	Adds an ACE that grants the read and write permissions on files and subdirectories in the *dir* directory to a group principal named 10001.
`find dir -type f -exec sh -c 'for ace in $(nfs4_getfacl \{} | grep "^A.*\:1005\:"); do nfs4_setfacl -x $ace \{}; done' \;`	Deletes ACEs that grant a user principal named 1005 access to files in the *dir* directory.
`nfs4_setfacl -a A:fdg:10001:rW dir1`	Adds an ACE that grants the read and write permissions on new files and subdirectories in the *dir1* directory to a group principal named 10001.
`nfs4_setfacl -a A:fg:10001:rx dir1`	Adds an ACE that grants the read, write, and execute permissions on all newly created files in the *dir1* directory to a group principal named 10001.

Procedure

To control access to files or directories by configuring NFSv4 ACLs, perform the following steps.

1. Create users and groups.

   In this example, the following users are created: player, admini, and anonym. The following groups are created: players and adminis. The player user is added to the players group and the admini user is added to the adminis group.

   ```
   sudo useradd player
   sudo groupadd players
   sudo usermod -g players player
   sudo useradd admini
   sudo groupadd adminis
   sudo usermod -g adminis admini
   sudo useradd anonym
   ```

2. Install the related tools to configure NFSv4 ACLs.

   If you have installed these tools, skip this step.

   ```
   sudo yum -y install nfs4-acl-tools
   ```

3. Obtain the group IDs of the players and adminis groups.

Open the */etc/group* file. The group IDs of the players and adminis groups are displayed.

```
players:x:19064:player
adminis:x:19065:admini
```

4. Configure NFSv4 ACLs for files and directories.

   Run the following commands to complete the operations: create a directory named *dir0* and add ACEs that grant the read-only permissions on all files in the *dir0* directory to the players group, grant the read, write, and execute permissions to the adminis group, and do not grant permissions to other users.

```
sudo umask 777
sudo mkdir dir0
sudo nfs4_setfacl -a A:fdg:19064:RX dir0
sudo nfs4_setfacl -a A:fdg:19065:RWX dir0
sudo nfs4_setfacl -a A:fdg:OWNER@: dir0
sudo nfs4_setfacl -a A:fdg:GROUP@: dir0
sudo nfs4_setfacl -a A:fdg:EVERYONE@: dir0
```

   Run the `sudo nfs4_getfacl dir0` command to verify the configuration.

```
A::OWNER@:tTnNcCy
A::GROUP@:tncy
A::EVERYONE@:tncy
A:fdi:EVERYONE@:tncy
A:fdi:OWNER@:tTnNcCy
A:fdi:GROUP@:tncy
A:g:19064:rxtncy
A:g:19065:rwaDxtTnNcCy
A:fdig:19064:rxtncy
A:fdig:19065:rwaDxtTnNcCy
```

5. Verify the configuration of the ACL.

   i. Run the following commands to verify that the admini user has the read and write permissions.

```
[root@vbox test] sudo su admini -c 'touch dir0/file'
[root@vbox test] sudo su admini -c 'echo 123 > dir0/file'
```

   ii. Run the following command to verify that the player user has the read-only permissions.

```
[root@vbox test] sudo su player -c 'touch dir0/file'
touch: cannot touch ‘dir0/file’: Permission denied
[root@vbox test] sudo su player -c 'echo 456 >> dir0/file'
bash: dir0/file: Permission denied
[root@vbox test] sudo su player -c 'cat dir0/file'
123
[root@vbox test] sudo su player -c 'nfs4_getfacl dir0/file'
A::OWNER@:tTnNcCy
A::GROUP@:tncy
A::EVERYONE@:tncy
A:g:19064:rxtncy
A:g:19065:rwaxtTnNcCy
```

iii. Run the following command to verify that the anonym user does not have permissions on the /dir0/file file.

```
[root@vbox test] sudo su anonym -c 'ls dir0'
ls: cannot open directory dir0: Permission denied
[root@vbox test] sudo su anonym -c 'cat dir0/file'
cat: dir0/file: Permission denied
[root@vbox test] sudo su anonym -c 'nfs4_getfacl dir0/file'
Invalid filename: di
```

What to do next

If you want to remove user permissions, use the following method.

When you use NFSv4 ACLs, we recommend that you sort each user into different groups. This allows you to configure permissions for a group rather than a separate user. To disable access to an object from a user, you can remove the user from a group that has access to the object. For example, use the following commands to remove the admini user from the adminis group and add the user to the adminis2 group:

```
[root@vbox test] sudo groupadd adminis2
[root@vbox test] sudo usermod -g adminis2 admini
[root@vbox test] id admini
uid=1057(admini) gid=1057(admini) groups=1054(adminis2)
[root@vbox test] sudo su admini -c 'ls dir0'
ls: cannot open directory dir0: Permission denied
[root@vbox test] sudo su admini -c 'cat dir0/file'
cat: dir0/file: Permission denied
[root@vbox test] sudo su admini -c 'nfs4_getfacl dir0/file'
Invalid filename: dir0/file
```

8.Tablestore

8.1. User Guide

8.1.1. What is Tablestore?

Tablestore is a NoSQL database service independently developed by Alibaba Cloud. Tablestore is a proprietary software program that is certified by the relevant authorities in China. Tablestore is built on the Apsara system of Alibaba Cloud, and can store large amounts of structured data and allow real-time access to these data.

Tablestore provides the following features:

- Offers schema-free data storage. You do not need to define attribute columns before you use them. Table-level changes are not required to add or delete attribute columns. You can configure the time to live (TTL) parameter for a table to manage the lifecycle of data. The expired data is deleted from the table.
- Adopts the triplicate technology to keep three copies of data on three servers across three different racks. A cluster can support single storage type instances (SSD only) or mixed storage type instances (SSD and HDD) to meet different budget and performance requirements.
- Adopts a fully redundant architecture that prevents single points of failure (SPOFs). Tablestore supports smooth online upgrades, hot cluster upgrades, and automatic data migration, which enable you to dynamically add or remove nodes for maintenance without incurring service interruptions. The concurrent read and write throughput and storage capacity can be linearly scaled. Each cluster can have at least 500 hosts.
- Supports highly concurrent read and write operations. Concurrent read and write capabilities can be scaled out as the number of hosts increases. The read and write performance is indirectly related to the amount of data in a single table.
- Supports identity authentication and multi-tenancy. Comprehensive access control and isolation mechanisms are provided to safeguard your data. VPC and access over HTTPS are supported. Provides multiple authentication and authorization mechanisms so that you can define access permissions on individual tables and operations.

8.1.2. Precautions

Before you use Tablestore, you need to take note of the following precautions and limits.

The following table describes the limits for Tablestore. A part of the limits indicate the maximum allowable values rather than the suggested values. To ensure better performance, set the table scheme and data size in a single row based on actual conditions, and adjust the following configurations.

Item	Limit	Description
The number of instances under an Apsara Stack tenant account	1024	To raise the limit, contact the technical support personnel.
The number of tables in an instance	1024	To raise the limit, contact the technical support personnel.
The length of an instance name	3 to 16 bytes	The instance name can contain uppercase and lowercase letters, digits, and hyphens (-). It must start with a letter and cannot end with a hyphen (-).
The length of a table name	1 to 255 bytes	The table name can contain uppercase and lowercase letters, digits, and underscores (_). It must start with a letter or underscore (_).

Item	Limit	Description
The length of a column name	1 to 255 bytes	The column name can contain uppercase and lowercase letters, digits, and underscores (_). It must start with a letter or underscore (_).
The number of columns in a primary key	1 to 4	A primary key can contain one to four primary key columns.
The size of the value in a string type primary key column	1 KB	The size of the value in a STRING primary key column cannot exceed 1 KB.
The size of the value in a STRING attribute column	2 MB	The size of the value in a STRING attribute column cannot exceed 2 MB.
The size of the value in a BINARY primary key column	1 KB	The size of the value in a BINARY primary key column cannot exceed 1 KB.
The size of the value in a BINARY attribute column	2 MB	The size of the value in a BINARY attribute column cannot exceed 2 MB.
The number of attribute columns in a single row	Unlimited	A single row can contain an unlimited number of attribute columns.
The number of attribute columns written by one request	1,024	During a PutRow, UpdateRow, or BatchWriteRow operation, the number of attribute columns written in a row cannot exceed 1,024.
The data size of a row	Unlimited	The total size of all column names and column values for a row is unlimited.

8.1.3. Quick start

8.1.3.1. Log on to the Tablestore console

This topic shows how to log on to the Tablestore console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

> **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
>
> - Uppercase or lowercase letters
> - Digits
> - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products > Tablestore**.

8.1.3.2. Create an instance

An instance is a logical entity in Tablestore and is used to manage tables. An instance is the basic unit of the resource management system of Tablestore. Tablestore controls application access and implements resource measurement at the instance level. This topic describes how to create an instance.

Procedure

1. Log on to the Tablestore console.
2. On the **Overview** tab, click **Create Instance**.

   > **Note** You can create different instances to manage the associated tables for different business, or create different instances for development, testing, and production environments of the same business. By default, Tablestore allows you to create up to 1,024 instances and up to 1,024 tables in each instance that belongs to an Apsara Stack tenant account.
3. On the **Create Instance** page, configure parameters.

Parameter	Description
Region	Select a region from the drop-down list for the instance.
Organization	Select an organization from the drop-down list for the instance.
Resource Set	Select a resource set from the drop-down list for the instance.

Parameter	Description
Instance Name	Enter a name for the instance. Instance naming conventions: The name must be 3 to 16 characters in length and can contain only letters, digits, and hyphens (-). It must start with a letter and cannot start with case-insensitive string `ali` or `ots`.
Description	Enter a description for the instance.
Instance Type	Select an instance type from the drop-down list for the instance. Tablestore provides high-performance instances and capacity instances. The instance types vary based on the type of cluster you deploy.

4. Click **Submit**.
5. In the **Submitted.** dialog box, click **Back to Console**.

   On the **Overview** tab, you can view the created instance.

   After the instance is created, you can perform the following operations on the instance:

   - Click the name of the instance or click Manage Instance in the **Actions** column that corresponds to the instance. On the **Instance Management** page, click each tab to perform various operations.
     - On the **Instance Details** tab, you can view the Instance Access URL, Basic Information, and Tables sections.
     - On the **Instance Monitoring** tab, you can view monitoring data by using time ranges, metric categories, and operations.
     - On the **Network Management** tab, you can bind or unbind virtual private clouds (VPCs) and view the list of VPCs.
   - Click **Release** in the **Actions** column to release an instance.

     Notice
     - Before you release an instance, ensure that all tables are deleted, and VPCs are unbound from instances.
     - To create an instance when you release an existing instance, ensure that the name of the instance you want to create is different from that of the existing instance to avoid conflicts.

8.1.3.3. Create tables

This topic describes how to create a table in the Tablestore console.

Procedure

1.
2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. On the **Instance Details** tab, click **Create Table**.

   Note You can create a maximum of 1,024 tables in each instance.

4. In the **Create Table** dialog box, set **Table Name** and **Primary Key**.

   The following table describes the parameters you can configure.

Parameter	Description
Table Name	The name of the table. This name is used to uniquely identify a table in an instance. The name must be 1 to 255 bytes in length and can contain letters, digits, and underscores (_). The name must start with a letter or an underscore (_).
Primary Key	One or more primary key columns in the table that uniquely identify each record in the table. Enter a primary key name and select a data type. Click **Add a Primary Key** to add a primary key column. You can add one to four primary key columns. By default, the first primary key column is the partition key. The configurations and order of primary key columns cannot be modified after the table is created. **Note** ◦ In Tablestore, only a primary key column can be used as an auto-increment primary key column. Partition keys cannot be used as auto-increment primary key columns. ◦ After a primary key column is set to an auto-increment primary key column, Tablestore automatically generates a value for the auto-increment primary key column when you write a row of data. You do not need to specify a value for the auto-increment primary key column. The values of auto-increment primary key columns are incremental and unique within the rows that share the same partition key. ◦ Naming conventions of primary key columns: The name must be 1 to 255 bytes in length and can contain letters, digits, and underscores (_). The name must start with a letter or underscore (_) ◦ Data types supported by primary key columns are **String**, **Integer**, and **Binary**.

5. Optional. Configure advanced parameters.

   If you need to configure parameters such as Time to Live and Max Versions, perform this operation.

   i. Turn on **Advanced Settings**.

ii. Configure advanced parameters.

The following table describes the advanced parameters you can configure.

Parameter	Description
Time to Live	The period for which data in the table can be retained. When the retention period exceeds the Time to Live (TTL) value, the system deletes the expired data. The minimum TTL value is 86,400 seconds (one day). A value of -1 indicates that data never expires.
Max Versions	The maximum number of versions of data that can be retained for an attribute column. When the versions of data in an attribute column exceed the Max Versions value, the system deletes the earliest versions of data to keep the maximum number of versions equal to the Max Versions value. Valid values: 1 to 10.
Max Version Offset	The difference between the version number and the data written time must be within the value of Max Version Offset. Otherwise, an error occurs when the data is written. Unit: seconds. The valid version range for attribute columns is calculated based on the following formula: Valid version range = **[Data written time - Max version offset value, Data written time + Max version offset value)**.
Reserved Read Throughput	You can set this parameter only for high-performance instances. The read and write throughput that is allocated and reserved for the table. Valid values: integers from 0 to 5000. When the specified reserved read and write throughput is **0**, Tablestore does not reserve related resources for the table.
Reserved Write Throughput	

6. Optional. Create secondary indexes.

If you need to create secondary indexes, perform this operation.

i. Turn on **Create Secondary Index**.

ii. Click the **+ Add** button in the Pre-defined Column section. Enter the name of the pre-defined column and select a data type from the drop-down list.

- This operation is performed to create a predefined column for the base table. Tablestore uses a schema-free model. You can write an unlimited number of columns to a row and do not need to specify a fixed number of predefined columns in a schema. When you create a table, you can also predefine columns and specify their data types.
- You can add up to 14 predefined columns. To deleted the predefined column you add, click the icon on the left of the corresponding predefined column.
- The name of a predefined column must be 1 to 255 bytes in length and can contain letters, digits, and underscores (_). The name must start with a letter or underscore (_).
- The data types of predefined columns include STRING, INTEGER, BINARY, FLOAT, and BOOLEAN.

iii. Click **Add Secondary Index**. Enter Index Name and set Primary Key and Pre-defined Column for the index table.

- The name of an index table must be 1 to 255 bytes in length and can contain letters, digits, and underscores (_). The name must start with a letter or underscore (_).
- You can set the primary key of the index table to the primary key or predefined columns of the base table.
- Pre-defined Column is optional. You can set the predefined columns of the index table to only the predefined columns of the base table.

7. Click **OK**.

After a table is created, you can view the table in the **Table List** section. If the created table is not displayed in the list of tables, click the icon to refresh the list of tables.

After a table is created, you can perform the following operations on the table:

- Click the name of the table or click Details in the Actions column. On the **Manage Table** page, you can perform the following operations:
  - On the **Details** tab, you can view the description of the table and the primary key columns list, and modify the attributes of the table.
  - On the **Data Editor** tab, you can insert or update data, query data, view data details, and delete multiple data at a time.
- Click the icon in the Actions column corresponding to a table and choose **Delete** from the shortcut menu. Click OK in the Delete Table dialog box. The table is deleted.

> **Notice** If you delete a table, the table and the data in the table are permanently deleted from Tablestore and cannot be recovered. Exercise caution when you perform this operation.

8.1.3.4. Read and write data in the console

After a data table is created, you can use the Tablestore console to perform read and write operations on the data in the data table.

Write data

1.
2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of the required table, and click the **Data Editor** tab. You can also click **Data Editor** in the Actions column corresponding to the table.
4. On the **Data Editor** tab, click **Insert**.
5. In the **Insert** dialog box, set Primary Key Value. Click **Add Column**. Set **Name**, **Type**, **Value**, and **Version**.

   By default, **System Time** is selected, which indicates that the current system time is used as the version number of the data. You can also clear **System Time** and enter the version number of the data.
6. Click **OK**.

   Rows that contain the written data are displayed on the **Query** tab.

Update data

You can update data in the attribute columns of a row.

1.

2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of the required table, and click the **Data Editor** tab. You can also click **Data Editor** in the Actions column corresponding to the table.
4. On the **Query** tab, select the row that you want to update. Click **Update**.
5. In the **Update** dialog box, modify the type and value of a primary key column, add or remove attribute columns, or update or delete data from attribute columns.
   - You can click **Add Column** to add an attribute column. You can also click the icon to delete an attribute column.
   - In the first Actions column, if you select **Update**, you can modify data in attribute columns. If you select **Delete**, you can delete data of the selected version. If you select **Delete All**, you can delete all versions of the data.
6. Click **OK**.

Query data

In the Tablestore console, you can query data in a single row (GetRow) or data within a specified range (RangeQuery).

To query data in a single row, perform the following operations:

1.
2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of the required table, and click the **Data Editor** tab. You can also click **Data Editor** in the Actions column corresponding to the table.
4. On the **Data Editor** tab, click **Search**.
5. Set filter conditions.
   i. In the **Search** dialog box, set Modes to **GetRow**.
   ii. By default, the system returns all columns. To return specific attribute columns, disable **All Columns**. Enter the names of the attribute columns to return. Separate the names of the attribute columns with commas (,).
   iii. Set **Primary Key Columns**.
   The integrity and accuracy of the primary key values affect the query results.
   iv. Set **Max Versions** to specify the maximum number of versions to return.

   > **Note** You can specify a maximum of 20 versions in the console. This limit does not apply when you use Tablestore SDKs.
6. Click **OK**.
   Data that meets the filter conditions is displayed on the **Data Editor** tab.

To query data within a specified range, perform the following steps:

1.
2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of the required table, and click the **Data Editor** tab. You can also click **Data Editor** in the Actions column corresponding to the table.
4. On the **Data Editor** tab, click **Search**.
5. Set filter conditions.

i. In the **Search** dialog box, set Modes to **Range Search**.

ii. By default, the system returns all columns. To return specific attribute columns, disable **All Columns**. Enter the names of the attribute columns to return. Separate the names of the attribute columns with commas (,).

iii. Set Start Primary Key Column and End Primary Key Column.

You can set Start Primary Key Column to **Min Value** or **Custom** and End Primary Key Column to **Max Value** or **Custom**. If you select **Custom**, enter a custom value.

> **Note**
> - The value in the first primary key column takes priority when the range query mode is used. When the start primary key column value and the end primary key column value are the same in the first primary key column, the system uses the value in the second primary key column to perform queries. The query rules for subsequent primary keys are the same as those for the first two primary keys.
> - The Custom range is a left-open and right-closed interval.

iv. Set **Max Versions** to specify the maximum number of versions to return.

> **Note** You can specify a maximum of 20 versions in the console. This limit does not apply when you use Tablestore SDKs.

v. Set Sequence to **Forward Search** or **Backward Search**.

6. Click **OK**.

Data that meets the filter conditions is displayed based on the specified order on the **Data Editor** tab.

Delete data

You can delete data that you no longer need.

1.
2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of the required table, and click the **Data Editor** tab. You can also click **Data Editor** in the Actions column corresponding to the table.
4. On the **Query** tab, select the row that you want to delete. Click **Delete**.
5. In the **Delete** message, click **OK**.

8.1.3.5. Bind a VPC to a Tablestore instance

After you bind a VPC to a Tablestore instance, you can access the Tablestore instance from the ECS instances in the VPC in the same region.

Prerequisites

- A VPC that is within the same region as the Tablestore instance is created.
- After the VPC is created, create an ECS instance in the VPC.

Procedure

1. Log on to the Tablestore console.
2. On the **Overview** page, click the name of the required instance or click **Manage Instance** in the Actions column corresponding to the instance.
3. Click the **Network Management** tab.

4. On the **Network Management** tab, click **Bind VPC**.
5. In the **Bind VPC** dialog box, select a VPC and switch, enter **Instance VPC Name**.

   The name of a VPC can contain only letters and digits and must start with a letter. The name must be 3 to 16 bytes in length.
6. Click **OK**.

   After the VPC is bound to the instance, you can view the information of the VPC in the **VPC List** on the **Network Management** tab. You can use the VPC address to access the Tablestore instance from the ECS instances in the VPC.

   After you bind a VPC, you can perform the following operations:

   - Click **VPC Instance List** in the Actions column to view the VPC instances list, which contains the instance name, instance VPC name, and VPC domain name.
   - Click **Unbind** in the Actions column to unbind the VPC from the instance. After the VPC is unbound, the ECS instance in the VPC can no longer access the Tablestore instance by using the VPC address. To access the Tablestore instance from the ECS instance, you must bind the VPC to the Tablestore instance again.

8.1.3.6. Use Tunnel Service

After the Stream feature is enabled, you can create tunnels for the data table to consume historical and incremental data in the data table.

Background information

Tunnel Service is built on the Tablestore API to provide tunnels that are used to consume data in full, incremental, and differential modes. You can create full, incremental, and differential tunnels and consume distributed data through these tunnels in real time.

Enable Stream

After the Stream feature is enabled, the system periodically deletes expired Stream operation logs that have been stored longer than the specified period.

1. Log on to the Tablestore console.
2. On the **Overview** page, click the name of an instance or click **Manage Instance** in the Actions column that corresponds to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of a data table and then click the Tunnels tab. You can also click ⋮ in the Actions column that corresponds to the data table and select **Tunnels**.
4. On the **Tunnels** tab, find the **Stream Information** section and click **Enabled**.
5. In the **Enable Stream** dialog box, set **Log Expiration Time**.

   > **Note**
   > - **Log Expiration Time** specifies the duration after which Stream operation logs expire.
   > - The unit of **Log Expiration Time** is hour. The value must be a non-zero integer and cannot be changed after it is specified. The maximum value can be set to 168 hours.
6. Click **Enabled**.

Create tunnels

After you create a tunnel for a data table, you can use the tunnel to consume historical and incremental data in the data table.

1. Log on to the Tablestore console.

2. On the **Overview** page, click the name of an instance or click **Manage Instance** in the Actions column that corresponds to the instance.
3. In the **Tables** section of the **Instance Details** tab, click the name of a data table and then click the Tunnels tab. You can also click ⋮ in the Actions column that corresponds to the data table and select **Tunnels**.
4. On the **Tunnels** tab, click **Create Tunnel**.
5. In the **Create Tunnel** dialog box, set **Tunnel Name** and **Type**.

   Tunnel Service provides three types of tunnels to consume distributed data in real time, including **Incremental**, **Full**, and **Differential**.
6. Click **OK**.

   After you create a tunnel, you can view the information about the tunnel on the Tunnels tab.

Preview data types in a tunnel

Simulate data consumption by writing or deleting data and preview data types in the tunnel.

1. For more information about how to write data or delete data in the console, see Read and write data in the console.
2. Preview the data types in the tunnel.
   i. Log on to the Tablestore console.
   ii. On the **Overview** page, click the name of an instance or click **Manage Instance** in the Actions column that corresponds to the instance.
   iii. In the **Tables** section of the **Instance Details** tab, click the name of a data table and then click the Tunnels tab. You can also click ⋮ in the Actions column that corresponds to the data table and select **Tunnels**.
   iv. On the **Tunnels** tab, click **Show Channels** in the Actions column that corresponds to the tunnel.
   v. In the channel list, click **View Simulated Export Records** in the Actions column that corresponds to a channel.

Enable data consumption for a tunnel

Obtain the ID of the created tunnel. Use the Tablestore SDK in any programming language for Tunnel Service to enable data consumption for the tunnel.

1. Obtain the ID of the created tunnel.
   i. Log on to the Tablestore console.
   ii. On the **Overview** page, click the name of an instance or click **Manage Instance** in the Actions column that corresponds to the instance.
   iii. In the **Tables** section of the **Instance Details** tab, click the name of a data table and then click the Tunnels tab. You can also click ⋮ in the Actions column that corresponds to the data table and select **Tunnels**.
   iv. On the **Tunnels** tab, copy the ID of the created tunnel.
2. Use the Tablestore SDK in any programming language for Tunnel Service to enable data consumption for the tunnel.

```
// Customize the data consumption callback to implement the IChannelProcessor operation. Specify t
he process and shutdown methods.
private static class SimpleProcessor implements IChannelProcessor {
    @Override
    public void process(ProcessRecordsInput input) {
        System.out.println("Default record processor, would print records count");
        System.out.println(
            String.format("Process %d records, NextToken: %s", input.getRecords().size(), input.ge
tNextToken()));
        try {
            // Mock Record Process.
            Thread.sleep(1000);
        } catch (InterruptedException e) {
            e.printStackTrace();
        }
    }
    @Override
    public void shutdown() {
        System.out.println("Mock shutdown");
    }
}
// Configure advanced parameters in TunnelWorkerConfig.
TunnelWorkerConfig config = new TunnelWorkerConfig(new SimpleProcessor());
// Configure TunnelWorkers to start automatic data processing.
TunnelWorker worker = new TunnelWorker($tunnelId, tunnelClient, config);
try {
    worker.connectAndWorking();
} catch (Exception e) {
    e.printStackTrace();
    worker.shutdown();
    tunnelClient.shutdown();
}
```

9.ApsaraDB RDS for MySQL

9.1. User Guide (RDS for MySQL)

9.1.1. What is ApsaraDB RDS?

ApsaraDB RDS is a stable, reliable, and scalable online database service. Based on the distributed file system and high-performance storage, ApsaraDB RDS provides a set of solutions for disaster recovery, backup, restoration, monitoring, and migration.

ApsaraDB RDS supports four database engines, which are MySQL, SQL Server, PolarDB, and PostgreSQL. You can create database instances based on these engines to meet your business requirements.

ApsaraDB RDS for MySQL

ApsaraDB RDS for MySQL is developed based on AliSQL and provides excellent performance. ApsaraDB RDS for MySQL is a tried and tested solution that handled the high-volume concurrent traffic during Double 11. ApsaraDB RDS for MySQL supports deployment with mixed x86 and ARM clusters. It integrates basic features such as whitelist configuration, backup and restoration, Transparent Data Encryption (TDE), data migration, and management for instances, accounts, and databases. ApsaraDB RDS for MySQL also provides the following advanced features:

- **Read-only instance:** In scenarios where ApsaraDB RDS for MySQL handles a small number of write requests but a large number of read requests, you can create read-only instances to scale up the reading capability and increase the application throughput.
- **Read/write splitting:** The read/write splitting feature provides a read/write splitting endpoint. This endpoint enables an automatic link for the primary instance and all of its read-only instances. An application can connect to the read/write splitting endpoint to read and write data. Write requests are distributed to the primary instance and read requests are distributed to read-only instances based on their weights. To scale up the reading capability of the system, you need only to add more read-only instances.

9.1.2. Log on to the ApsaraDB RDS console

This topic describes how to log on to the ApsaraDB RDS console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

> **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
>
> - Uppercase or lowercase letters
> - Digits
> - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products > Database Services > ApsaraDB RDS**.

9.1.3. Quick start

9.1.3.1. Limits

To ensure instance stability and security, ApsaraDB RDS for MySQL has some service limits, as listed in the following table.

Operation	Description
Instance parameters	Instance parameters can be modified by using the ApsaraDB RDS console or API operations. Due to security and stability considerations, only specific parameters can be modified.
Root permissions of databases	The root or system administrator permissions are not provided.
Database backup	• Logical backup can be performed by using the command line interface (CLI) or graphical user interface (GUI). • Physical backup can be performed only by using the ApsaraDB RDS console or API operations.
Database restoration	• Logical restoration can be performed by using the CLI or GUI. • Physical restoration can be performed only by using the ApsaraDB RDS console or API operations.

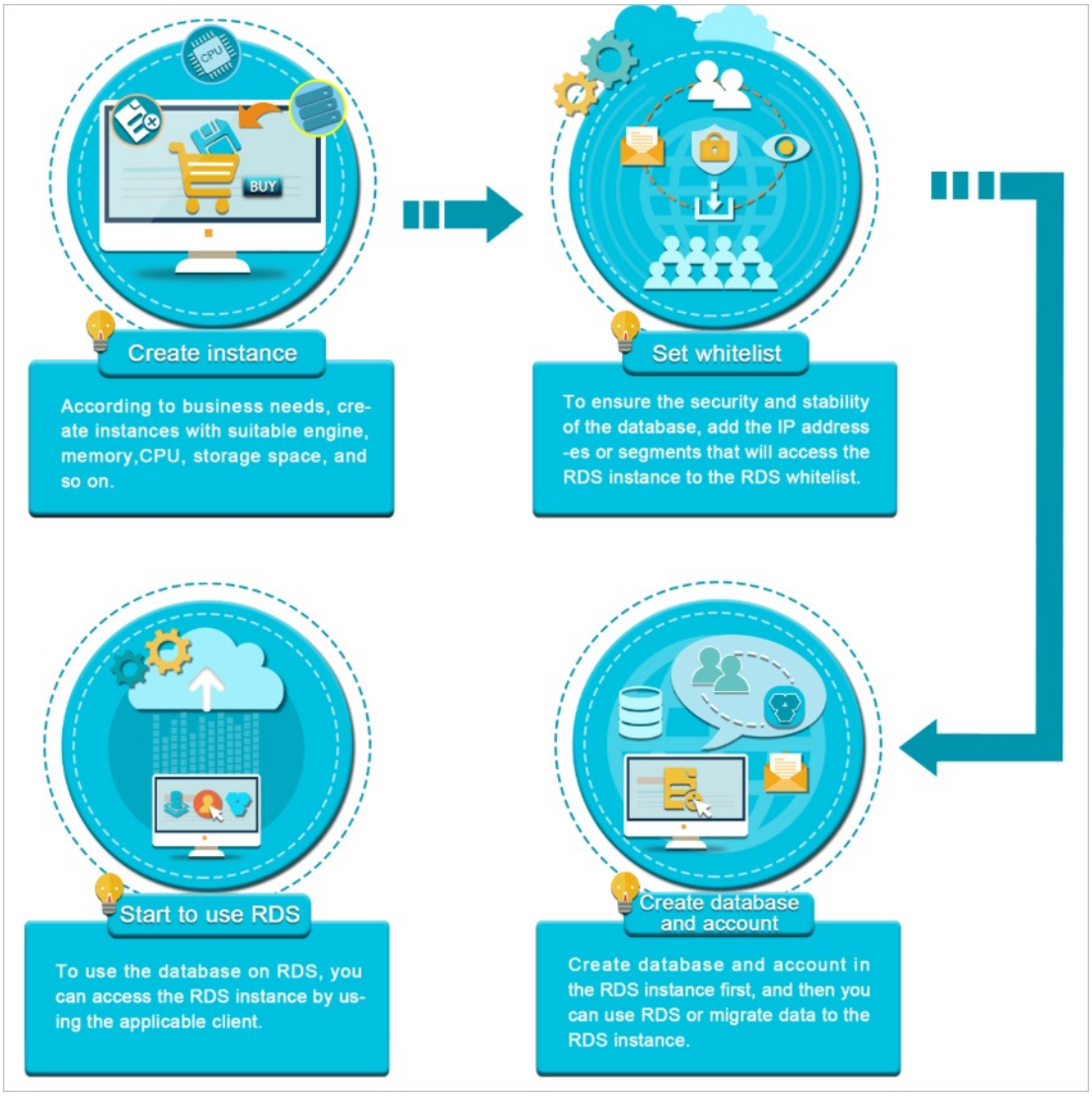
Operation	Description
ApsaraDB RDS for MySQL storage engine	Only InnoDB is supported. • To ensure performance and security, we recommend that you use the InnoDB storage engine. • The TokuDB engine is not supported. Percona no longer provides support for TokuDB, which leads to bugs that cannot be fixed and business losses in extreme cases. • The MyISAM engine is not supported. Due to the inherent shortcomings of the MyISAM engine, some data may be lost. Only some existing instances use the MyISAM engine. MyISAM engine tables in newly created instances are automatically converted to InnoDB engine tables. • The Memory engine is not supported. Newly created Memory tables are automatically converted into InnoDB tables.
Database replication	ApsaraDB RDS for MySQL provides dual-node clusters based on a primary/secondary replication architecture. The secondary instances in this replication architecture are hidden and cannot be accessed directly.
Instance restart	Instances must be restarted by using the ApsaraDB RDS console or API operations.
Account and database management	ApsaraDB RDS for MySQL manages accounts and databases by using the ApsaraDB RDS console. ApsaraDB RDS for MySQL also allows you to create a privileged account to manage users, passwords, and databases.
Standard account	• Authorization is not allowed. • The ApsaraDB RDS console allows you to manage accounts and databases. • Instances that support standard accounts also support privileged accounts.
Privileged account	• Authorization is allowed to standard accounts. • The ApsaraDB RDS console does not provide interfaces to manage accounts or databases. These operations can be performed only by using code or DMS. • The privileged account cannot be reverted to a standard account.

9.1.3.2. Procedure

ApsaraDB RDS quick start covers the following operations: creating an instance, configuring a whitelist, creating a database, creating an account, and connecting to the instance. This topic describes how to use ApsaraDB RDS and provides all the necessary information to create an ApsaraDB RDS instance. ApsaraDB RDS for MySQL is used in the example.

Typically, after an instance is created, you must perform several operations to make the instance ready for use, as shown in Quick start flowchart.

Quick start flowchart



- Create an instance

  An instance is a virtual database server on which you can create and manage multiple databases.

- Configure a whitelist

  After you create an ApsaraDB RDS instance, you must configure its whitelist to allow access from external devices.

  Whitelists make your ApsaraDB RDS instance more secure. We recommend that you maintain whitelists on a regular basis. The whitelist configuration process does not affect the normal operations of the ApsaraDB RDS instance.

- Create a database and Create an account

  Before you use a database, you must first create the database and an account in the ApsaraDB RDS instance.

- Connect to an ApsaraDB RDS for MySQL instance

  After you create an ApsaraDB RDS instance, configure a whitelist, and create a database and an account, you can connect to the instance by using a database client.

9.1.3.3. Create an instance

This topic describes how to create one or more instances in the ApsaraDB RDS console.

Prerequisites

An Apsara Stack tenant account is created.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.
3. Configure the following parameters.

Section	Parameter	Description
Basic Settings	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	Region	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.
	Deployment Method	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
	Quantity	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.

Section	Parameter	Description
Specifications	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **MySQL**.
	Chip Architecture	The chip architecture of the host on which the instance is deployed. **Note** If you do not have permissions to select this option, contact the operations administrator to authorize your account.
	Engine Version	The version of the database engine. Valid values: ○ 8.0 ○ 5.7 ○ 5.6
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Type	The storage type of the instance. Select Local SSD.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, which includes the space to store data, system files, binlog files, and transaction files. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Network Type	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. **Note** Instances that use standard SSDs cannot be deployed in the classic network. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security.

Section	Parameter	Description
Network	VPC	The VPC in which you want to create the instance. Note When Network Type is set to VPC, you must specify this parameter.
	vSwitch	The vSwitch in the VPC. Note When Network Type is set to VPC, you must specify this parameter.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click Submit.

9.1.3.4. Initialization settings

9.1.3.4.1. Configure an IP address whitelist for an ApsaraDB RDS instance

After you create an ApsaraDB RDS instance, you must add the IP addresses or CIDR blocks that are used for database access to the IP address whitelist of the instance to ensure database security and reliability.

Context

IP address whitelists make your ApsaraDB RDS instance more secure and do not interrupt the operations of your ApsaraDB RDS instance when you configure whitelists. We recommend that you maintain your IP address whitelists on a regular basis.

To configure a whitelist, you can perform the following operations:

- Configure an IP address whitelist: Add IP addresses to allow them to connect to the ApsaraDB RDS instance.
- Configure an Elastic Compute Service (ECS) security group: Add an ECS security group for the ApsaraDB RDS instance to allow ECS instances in the group to connect to the ApsaraDB RDS instance.

Precautions

- The default IP address whitelist can be modified or cleared, but cannot be deleted.
- You can add up to 1,000 IP addresses or CIDR blocks to a whitelist. If you want to add a large number of IP addresses, we recommend that you merge them into CIDR blocks, such as 192.168.1.0/24.

Introduction to IPv6

IPv4 addresses are widely used, but the limited number of IPv4 addresses restricts the development of the Internet. Compared with IPv4 addresses, IPv6 addresses are more sufficient and allow more types of devices to access the Internet. ApsaraDB RDS supports both IPv4 and IPv6 addresses.

The following table describes the differences between IPv4 and IPv6.

Item	IPv4	IPv6
Address length	32 bits (4 bytes)	128 bits (16 bytes)

Item	IPv4	IPv6
Number of addresses	2^32	2^128
Address format	xxx.xxx.xxx.xxx Where xxx is a decimal number that can range from 0 to 255. Each x is a decimal integer, and leading zeros can be omitted. Example: 192.168.1.1	xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx Where each x is a hexadecimal number, and leading zeros can be omitted. You can use a double colon (::) once in an IPv6 address to indicate a series of zeros. Example: CDDC:0000:0000:0000:8475:1111:3900:2020
Address Resolution Protocol (ARP)	Uses broadcast ARP Request frames to resolve an IP address to a link layer address.	Uses multicast neighbor solicitation messages to resolve an IP address to a link layer address.
Security	Implements a security mechanism based on applications and cannot provide protections at the IP layer.	Supports packet fragmentation to ensure data confidentiality and integrity and provides security at the IP layer.
LAN connection	Connects to LANs by using network interfaces.	Can work with Ethernet adapters and is supported over virtual Ethernet networks between logical partitions.
Address type	• Unicast address • Multicast address • Broadcast address	• Unicast address • Multicast address • Anycast address

Create an IP address whitelist

Each IP address whitelist of an ApsaraDB RDS instance can contain **IPv4** or **IPv6** addresses. By default, the system provides an IP address whitelist of the **IPv4** type. If you want an IP address whitelist of the **IPv6** type, manually create one.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. On the **Whitelist Settings** tab, click **Create Whitelist**. In the dialog box that appears, configure the following parameters.

Parameter	Description
Whitelist Name	The name of the IP address whitelist. **Note** ◦ The name can contain lowercase letters, digits, and underscores (_). ◦ The name must start with a lowercase letter and end with a lowercase letter or digit. ◦ The name must be 2 to 32 characters in length.

Parameter	Description
IP Type	The IP type of the IP address whitelist. Valid values: ○ **IPv4** ○ **IPv6** **Note** For more information about the differences between IPv4 and IPv6, see the "Introduction to IPv6" section of this topic.
IP Addresses	The IP addresses that are allowed to access the instance.

Configure an IP address whitelist

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. On the **Whitelist Settings** tab, click **Edit** corresponding to an IP address whitelist.

   Note If you want to connect an ECS instance to an ApsaraDB RDS instance by using an internal endpoint, you must make sure that the two instances are in the same region and have the same network type. Otherwise, the connection fails.

5. In the **Edit Whitelist** dialog box, enter the IP addresses or CIDR blocks that are allowed to access the instance and click **OK**.

Note

- Limits for **IPv4** addresses:
  - You must separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.

    Supported formats are `0.0.0.0/0`, IP addresses such as `10.23.12.24`, or CIDR blocks such as `10.23.12.24/24`. `/24` indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits.
  - If an IP address whitelist is empty or contains `0.0.0.0/0`, all IP addresses can access the ApsaraDB RDS instance. This may cause security risks to the instance. Proceed with caution.
- Limits for **IPv6** addresses:
  - You must separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.

    Supported formats are `::`, IP addresses such as `0:0:0:0:0:0:0:1`, or CIDR blocks such as `0:0:0:0:0:0:0:1/24`. `/24` indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 128 bits.
  - If an IP address whitelist is empty or contains only `::`, all IP addresses can access the ApsaraDB RDS instance. This may cause security risks to the instance. Proceed with caution.
- You cannot specify both **IPv4** and **IPv6** addresses in a single IP address whitelist. If you want to specify both IPv4 and IPv6 addresses, specify them in separate IP address whitelists.
- If you click **Add Internal IP Addresses of ECS Instances**, the IP addresses of all the ECS instances that are created within your Apsara Stack tenant account appear. Then, you can select the IP addresses and add them to an IP address whitelist.

9.1.3.4.2. Create an account

After you create an ApsaraDB RDS instance and configure its IP address whitelist, you must create a database and an account on the instance. This topic describes how to create privileged and standard accounts.

Context

ApsaraDB RDS for MySQL supports two types of database accounts: privileged and standard. You can manage all your accounts and databases in the ApsaraDB RDS console. For more information about permissions that can be granted to each type of account, see Account permissions.

Account type	Description
Privileged account	• You can create and manage privileged accounts by using the ApsaraDB RDS console or API operations. • You can create only a single privileged account on each ApsaraDB RDS instance. The privileged account can be used to manage all standard accounts and databases on the instance. • A privileged account allows you to manage permissions to a fine-grained level. For example, you can grant each standard account the permissions to query specific tables. • A privileged account has the permissions to disconnect all standard accounts on the instance.

Account type	Description
Standard account	• You can create and manage standard accounts by using the ApsaraDB RDS console, API operations, or SQL statements. • You can create up to 500 standard accounts on an instance. • You must manually grant standard accounts the specific database permissions. • You cannot use a standard account to create, manage, or disconnect other accounts from databases.

Account type	Maximum number of databases	Maximum number of tables	Maximum number of accounts
Privileged account	Unlimited	< 200,000	Varies based on the engine parameter settings of the instance.
Standard account	500	< 200,000	Varies based on the engine parameter settings of the instance.

Create a privileged account

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts** to go to the **Accounts** page.
4. On the **Accounts** tab, click **Create Account**.
5. On the **Create Account** page, configure the following parameters.

Parameter	Description
Database Account	Enter the name of the account. The account name must meet the following requirements: ◦ The name is 1 to 16 characters in length. ◦ The name starts with a lowercase letter and ends with a lowercase letter or digit. ◦ The name contains lowercase letters, digits, and underscores (_).
Account Type	Select Privileged Account.
Password	Enter the password of the account. The password must meet the following requirements: ◦ The password is 8 to 32 characters in length. ◦ The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ◦ Special characters include ! @ # $ % ^ & * ( ) _ + - =
Re-enter Password	Enter the password of the account again.
Description	Optional. Enter information about the account to facilitate subsequent management. The description can be up to 256 characters in length.

6. Click **Create**.

Reset the permissions of a privileged account

If an issue occurs on the privileged account, you can enter the password of the privileged account to reset permissions. For example, you can reset the permissions if the permissions are unexpectedly revoked.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts** to go to the **Accounts** page.
4. On the **Accounts** tab, find the privileged account and click **Reset Permissions** in the **Actions** column.
5. On the **Initialize Account** page, enter the password of the privileged account and click **OK**.

Create a standard account

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts** to go to the **Accounts** page.
4. On the **Accounts** tab, click **Create Account**.
5. On the **Create Account** page, configure the following parameters.

Parameter	Description
Database Account	Enter the name of the account. The account name must meet the following requirements: ○ The name is 1 to 16 characters in length. ○ The name starts with a lowercase letter and ends with a lowercase letter or digit. ○ The name contains lowercase letters, digits, and underscores (_).
Account Type	Select Standard Account.
Authorized Databases	Select one or more databases on which you want to grant permissions to the account. You can also leave this parameter empty at this time and authorize databases after the account is created. i. Select one or more databases from the Unauthorized Databases section and click **Add** to add them to the Authorized Databases section. ii. In the Authorized Databases section, select the **Read/Write**, **Read-only**, **DDL Only**, or **DML Only** permissions on each authorized database. If you want to grant the same permissions on multiple databases to the account, click the button in the upper-right corner of the section. The button may appear as **Set All to Read/Write**.
Password	Enter the password of the account. The password must meet the following requirements: ○ The password is 8 to 32 characters in length. ○ The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include ! @ # $ % ^ & * ( ) _ + - =
Re-enter Password	Enter the password of the account again.
Description	Optional. Enter information about the account to facilitate subsequent management. The description can be up to 256 characters in length.

6. Click **Create**.

Account permissions

Account type	Authorization type	Permission				
Privileged account	None	SELECT	INSERT	UPDATE	DELETE	CREATE
		DROP	RELOAD	PROCESS	REFERENCES	INDEX
		ALTER	CREATE TEMPORARY TABLES	LOCK TABLES	EXECUTE	REPLICATION SLAVE
		REPLICATION CLIENT	CREATE VIEW	SHOW VIEW	CREATE ROUTINE	ALTER ROUTINE
		CREATE USER	EVENT	TRIGGER	None	None
Standard account	**Read-only**	SELECT	LOCK TABLES	SHOW VIEW	PROCESS	REPLICATION SLAVE
		REPLICATION CLIENT	None	None	None	None
	Read/write	SELECT	INSERT	UPDATE	DELETE	CREATE
		DROP	REFERENCES	INDEX	ALTER	CREATE TEMPORARY TABLES
		LOCK TABLES	EXECUTE	CREATE VIEW	SHOW VIEW	CREATE ROUTINE
		ALTER ROUTINE	EVENT	TRIGGER	PROCESS	REPLICATION SLAVE
		REPLICATION CLIENT	None	None	None	None
	DDL-only	CREATE	DROP	INDEX	ALTER	CREATE TEMPORARY TABLES
		LOCK TABLES	CREATE VIEW	SHOW VIEW	CREATE ROUTINE	ALTER ROUTINE
		PROCESS	REPLICATION SLAVE	REPLICATION CLIENT	None	None
	DML-only	SELECT	INSERT	UPDATE	DELETE	CREATE TEMPORARY TABLES
		LOCK TABLES	EXECUTE	SHOW VIEW	EVENT	TRIGGER
		PROCESS	REPLICATION SLAVE	REPLICATION CLIENT	None	None

9.1.3.4.3. Create a database

After you create an ApsaraDB RDS instance and configure its IP address whitelist, you must create a database and an account on the instance.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Databases**.
4. Click **Create Database**. On the page that appears, configure the following parameters.

Parameter	Description
Database Name	○ The name must be 1 to 64 characters in length. ○ The name must start with a letter and end with a letter or a digit. ○ The name can contain lowercase letters, digits, underscores (_), and hyphens (-). ○ The name must be unique within the instance.
Supported Character Sets	Select **utf8**, **gbk**, **latin1**, **utf8mb4**, or **all**. If you want to use other character sets, select **all**, and then select the required character set from the list.
Description	Optional. Enter information about the database to facilitate subsequent management. The description must be 2 to 256 characters in length.

5. Click **Create**.

9.1.3.5. Connect to an ApsaraDB RDS for MySQL instance

After you complete the initial configuration of your ApsaraDB RDS for MySQL instance, you can connect to it from an Elastic Compute Service (ECS) instance or an on-premises client.

Context

After you perform operations such as Create an instance, Configure a whitelist, and Create an account, you can use a general database client or configure the endpoint, port number, and account information in an application to connect to the ApsaraDB RDS for MySQL instance.

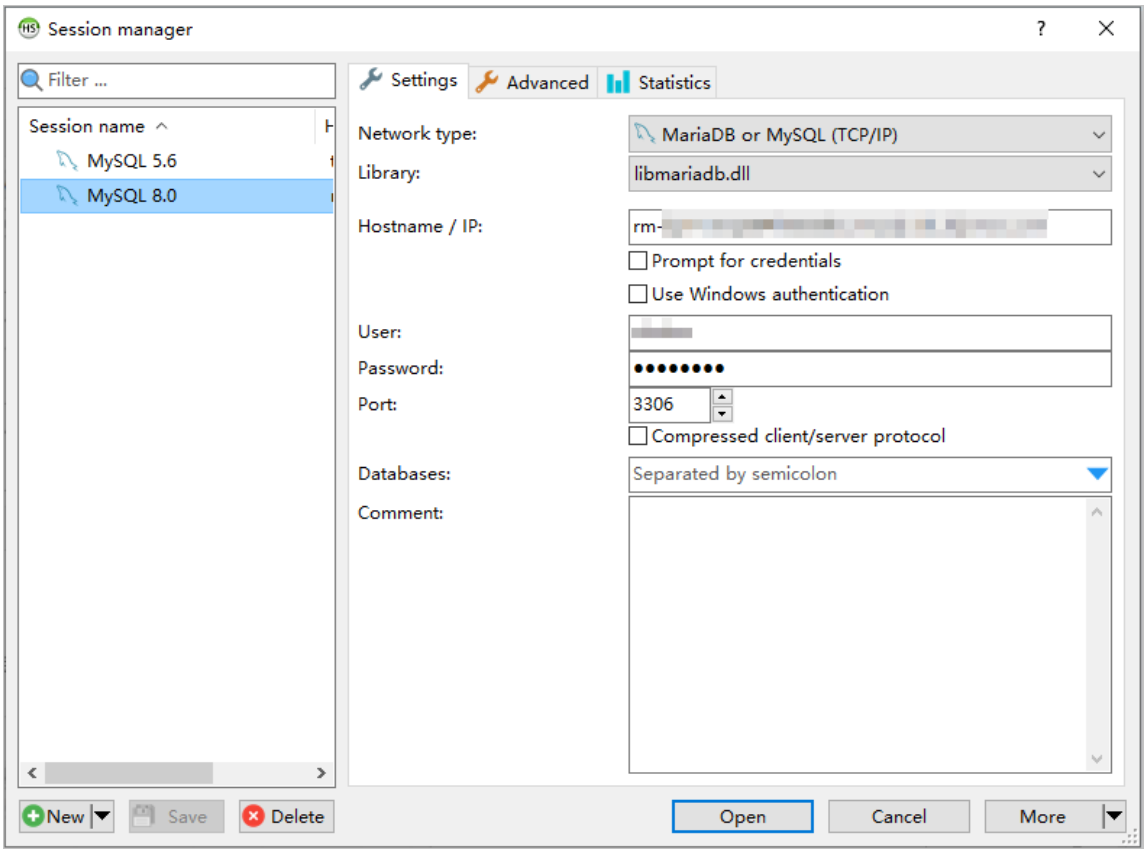
If you need to connect an ECS instance to an ApsaraDB RDS instance, you must make sure that both instances are in the classic network or in the same virtual private cloud (VPC), and the IP address of the ECS instance is added to an IP address whitelist of the ApsaraDB RDS instance.

Connect to an instance from a client

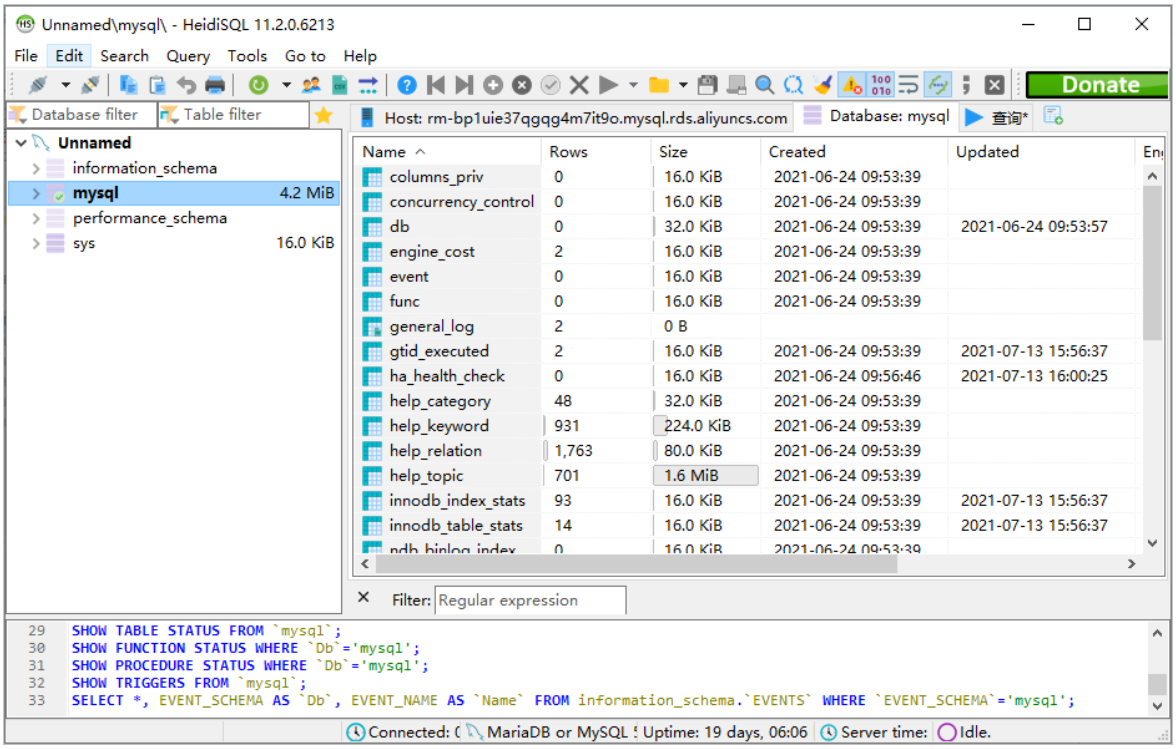
ApsaraDB RDS for MySQL is fully compatible with open source MySQL. You can connect to an ApsaraDB RDS instance from a database client by using a method similar to the method that is used to connect to an open source MySQL database. In the following example, the HeidiSQL client is used.

1. Start the HeidiSQL client.
2. In the lower-left corner of the Session manager dialog box, click **New**.
3. Enter information about the ApsaraDB RDS instance to which you want to connect. The following table describes the required parameters.

Parameter	Description
Network type	Select the network type of the ApsaraDB RDS instance to which you want to connect. For example, select **MariaDB or MySQL (TCP/IP)**.
Hostname / IP	Enter the internal or public endpoint of the ApsaraDB RDS instance. ○ If your client is deployed on an ECS instance that is in the same region and has the same network type as the ApsaraDB RDS instance, use the internal endpoint. For example, if the ECS and ApsaraDB RDS instances are both in a VPC located in the China (Hangzhou) region, you can use the internal endpoint of the ApsaraDB RDS instance to create a secure connection. ○ In other scenarios, use the public endpoint. To view the internal and public endpoints and port numbers of the ApsaraDB RDS instance, perform the following operations: i. Log on to the ApsaraDB for RDS console. ii. Find the ApsaraDB RDS instance to which you want to connect and click its ID. iii. In the **Basic Information** section, view the internal and public endpoints and port numbers of the instance.
User	Enter the name of the account used to connect to the ApsaraDB RDS instance.
Password	Enter the password of the account used to connect to the ApsaraDB RDS instance.
Port	If you connect to the instance over an internal network, enter the internal port number of the instance. If you connect to the instance over the Internet, enter the public port number of the instance.



4. Click **Open**. If the connection information is correct, you can connect to the instance.



9.1.4. Instances

9.1.4.1. Create an instance

This topic describes how to create one or more instances in the ApsaraDB RDS console.

Prerequisites

An Apsara Stack tenant account is created.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.
3. Configure the following parameters.

Section	Parameter	Description
Basic Settings	**Organization**	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
	Region	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.

Section	Parameter	Description
Region	**Deployment Method**	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
Specifications	**Quantity**	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.
	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **MySQL**.
	Chip Architecture	The chip architecture of the host on which the instance is deployed. **Note** If you do not have permissions to select this option, contact the operations administrator to authorize your account.
	Engine Version	The version of the database engine. Valid values: ○ 8.0 ○ 5.7 ○ 5.6
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Type	The storage type of the instance. Select Local SSD.

Section	Parameter	Description
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, which includes the space to store data, system files, binlog files, and transaction files. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
Network	**Network Type**	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. **Note** Instances that use standard SSDs cannot be deployed in the classic network. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security.
	VPC	The VPC in which you want to create the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch in the VPC. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click **Submit**.

9.1.4.2. View basic information of an instance

This topic describes how to view the details of an ApsaraDB RDS instance, such as its basic information, internal network connection information, status, and configurations.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. Use one of the following methods to go to the **Basic Information** page of an instance:
   - On the **Instances** page, click the ID of an instance to go to the **Basic Information** page.
   - On the **Instances** page, click **Manage** in the **Actions** column corresponding to an instance to go to the **Basic Information** page.

9.1.4.3. Restart an instance

This topic describes how to manually restart an ApsaraDB RDS instance. This applies if the number of connections exceeds the specified threshold or if an instance has performance issues.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click **Restart Instance** in the upper-right corner.

   > **Note** When you restart an instance, applications are disconnected from the instance. We recommend that you make appropriate service arrangements before you restart an instance. Proceed with caution.

4. In the Restart Instance message, click **Confirm**.

9.1.4.4. Change the specifications of an instance

This topic describes how to change the specifications of your instance, such as the instance type and storage capacity, if the specifications do not meet the requirements of your application.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the upper-right corner of the **Configuration Information** section, click **Change Specifications**.
4. On the **Change Specifications** page, set **Edition**, **Instance Type**, and **Storage Capacity**.
5. Click **Submit**.

9.1.4.5. Set a maintenance window

This topic describes how to set a maintenance window for an ApsaraDB RDS instance.

Context

To ensure the stability of ApsaraDB RDS instances, the backend system performs maintenance of the instances at irregular intervals. The default maintenance window is from 02:00 to 06:00. You can set the maintenance window to the off-peak period of your business to avoid impact on business.

Precautions

- To ensure stability of the maintenance process, the instance changes to the **Maintaining Instance** state before the maintenance window. When the instance is in this state, access to data in the database and query operations such as performance monitoring are not affected. However, except for account and database management and IP address whitelist configuration, modification operations such as upgrade, downgrade, and restart are temporarily unavailable.
- You may encounter a network interruption during the maintenance window. Make sure that your application is configured to automatically reconnect to the instance.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.

3. In the **Configuration Information** section, click **Configure** to the right of **Maintenance Window**.
4. Select a maintenance window and click **Save**.

> **Note** The maintenance window is displayed in UTC+8.

9.1.4.6. Change the data replication mode

You can set the data replication mode between primary and secondary ApsaraDB RDS instances to improve database availability.

Context

ApsaraDB RDS supports the following data replication modes:

- Semi-sync

  After an application-initiated update is complete on the primary instance, logs are synchronized to all secondary instances. This transaction is considered committed after at least one secondary instance has received the logs, regardless of whether the secondary instance finishes executing the updates specified in the logs.

  If the secondary instances are unavailable or a network exception occurs between the primary and secondary instances, semi-synchronous replication degrades to the asynchronous mode.

- Asynchronous

  When your application initiates a request to add, delete, or modify data, the primary instance responds to your application immediately after it completes the operation. At the same time, the primary instance starts to asynchronously replicate data to its secondary instances. During asynchronous data replication, the unavailability of secondary instances does not affect the operations on the primary instance. Data remains consistent even if the primary instance is unavailable.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Service Availability**.
4. In the upper-right corner of the **Availability Information** section, click **Change Data Replication Mode**.
5. In the dialog box that appears, select a data replication mode and click **OK**.

9.1.4.7. Release an instance

This topic describes how to manually release an instance.

Precautions

- Only instances in the running state can be manually released.
- After an instance is released, the instance data is immediately deleted. We recommend that you back up your data before you release an instance.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the **Actions** column corresponding to the instance you want to release, choose **More > Release Instance**.
4. In the message that appears, click **OK**.

9.1.4.8. Update the minor version of an instance

ApsaraDB RDS for MySQL supports automatic and manual updates of the minor version. These updates increase performance, provide new features, and fix known issues.

Introduction

By default, ApsaraDB RDS for MySQL automatically updates the minor version. You can log on to the ApsaraDB RDS console, go to the **Basic Information** page of your ApsaraDB RDS instance, and then view the current **Minor Version Upgrade Mode** in the Configuration Information section.

- **Auto**: When a new minor version is released, the system automatically updates the minor version of your instance during the specified maintenance window. For more information, see Set a maintenance window.
- **Manual**: You can manually update the minor version on the **Basic Information** page. For more information, see Manually update the minor version.

Precautions

- When you update the minor engine version of your ApsaraDB RDS instance, a network interruption of about 30 seconds may occur. We recommend that you update the minor engine version during off-peak hours or make sure that your application is configured to automatically reconnect to the instance.
- After you update the minor engine version of your ApsaraDB RDS instance, you cannot downgrade the instance version.
- After you upgrade the specifications of your ApsaraDB RDS instance, ApsaraDB RDS updates the instance to the latest minor engine version.

Change the minor version update mode

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the **Configuration Information** section of the Basic Information page, click **Configure** to the right of **Minor Version Upgrade Mode**.
4. In the dialog box that appears,select **Auto** or **Manual** and click **OK**.

> **Note** By default, the minor version update mode is set to **Auto**.

Manually update the minor version

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the **Configuration Information** section of the page, click **Upgrade Minor Engine Version**.

> **Note** The **Upgrade Minor Engine Version** button is displayed only when a new minor version is available.

4. In the dialog box that appears, specify the update time and click **OK**.

FAQ

- Q: After I updated the minor version of my ApsaraDB RDS instance, the MySQL version remains unchanged. Why?

  A: The minor version that you updated is the minor engine version of ApsaraDB RDS, but not the MySQL version. To view the minor version of your instance, you can execute the `show variables like '%rds_release_date%'` statement.

- Q: When an update takes effect, is my instance updated only to the next minor version?

  A: No, when an update takes effect, your instance is updated to the latest minor version.

9.1.4.9. Modify parameters of an instance

This topic describes how to view and modify the values of some parameters and query parameter modification records in the console.

Precautions

- To ensure instance stability, you can select specific parameters to modify in the ApsaraDB RDS console.
- When you modify parameters on the **Editable Parameters** tab, you can refer to the **Value Range** column corresponding to each parameter.
- After some parameters are modified, you must restart your ApsaraDB RDS instance for the changes to take effect. You can refer to the **Force Restart** column on the **Editable Parameters** tab. We recommend that you modify the parameters of an instance during off-peak hours and make sure that your applications are configured to automatically reconnect to your instance.

Modify parameters

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Parameters**.
4. Perform the following operations:

   Export the parameter settings of the ApsaraDB RDS instance to your computer.

   On the Editable Parameters tab, click **Export Parameters**. The parameter settings of the ApsaraDB RDS instance are exported as a TXT file to your computer.

   Modify and import the parameter settings.

   i. After you modify parameters in the exported parameter file, click **Import Parameters** and copy the parameter settings to the field.

   ii. Click **OK**.

   iii. In the upper-right corner of the page, click **Apply Changes**.

   > **Note**
   > - If the new parameter value takes effect only after you restart your instance, the system prompts you to restart the ApsaraDB RDS instance. We recommend that you restart the ApsaraDB RDS instance during off-peak hours and make sure that your applications are configured to automatically reconnect to your instance.
   > - Before the new parameter values are applied, you can click **Cancel Changes** to cancel the modification.

   Modify a single parameter.

   i. On the **Editable Parameters** tab, find the parameter that you want to modify and click the icon in the **Actual Value** column.

   ii. Enter a new value based on the value range that is displayed.

   iii. Click **Confirm**.

iv. In the upper-right corner of the page, click **Apply Changes**.

> **Note**
>
> - If the new parameter value takes effect only after you restart your instance, the system prompts you to restart the ApsaraDB RDS instance. We recommend that you restart the ApsaraDB RDS instance during off-peak hours and make sure that your applications are configured to automatically reconnect to your instance.
> - Before the new parameter value is applied, you can click **Cancel Changes** to cancel the modification.

View the parameter modification history

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Parameters**.
4. On the Parameters page, click the **Edit History** tab.
5. Select a time range and click **Search**.

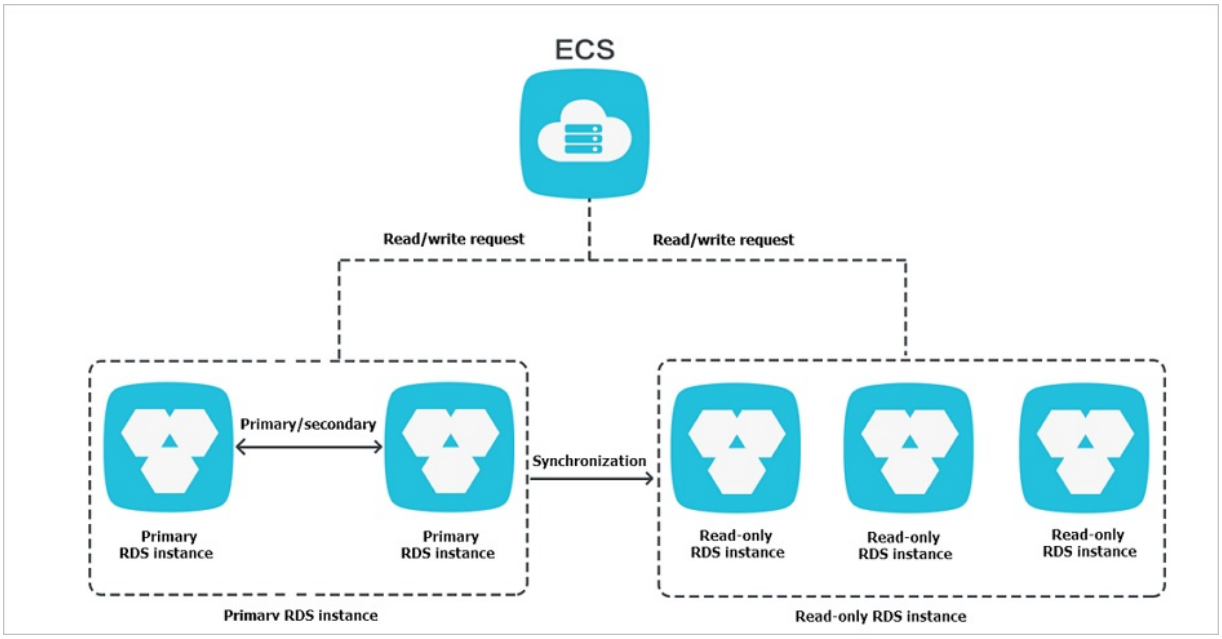
9.1.4.10. Read-only instances

9.1.4.10.1. Overview of read-only instances

ApsaraDB RDS for MySQL allows you to create read-only instances. In scenarios where an instance has a small number of write requests but a large number of read requests, you can create read-only instances to distribute database access loads away from the primary instance. This topic describes the features and limits of read-only instances.

To scale up the reading capability and distribute database access loads, you can create one or more read-only instances in a region. Read-only instances can increase the application throughput when a large amount of data is being read.

A read-only instance with a single physical node and no backup node uses the native replication capability of MySQL to synchronize changes from the primary instance to all its read-only instances. Read-only instances must be in the same region as the primary instance but do not have to be in the same zone as the primary instance. The following figure shows the topology of read-only instances.



Read-only instances have the following features:

- Specifications of a read-only instance can be different from those of the primary instance and can be changed at any time. This facilitates elastic scaling.
- Read-only instances do not require account or database maintenance. Account and database information is synchronized from the primary instance.
- The whitelists of read-only instances can be independently configured.
- System performance monitoring is provided.

  ApsaraDB RDS provides up to 20 system performance monitoring views, including those for disk capacity, IOPS, connections, CPU utilization, and network traffic. You can view the load of instances.
- ApsaraDB RDS provides a variety of optimization recommendations, such as storage engine check, primary key check, large table check, and check for excessive indexes and missing indexes. You can optimize your databases based on the optimization recommendations and specific applications.

9.1.4.10.2. Create a read-only instance

You can create read-only instances of different specifications based on your business requirements.

Precautions

- A maximum of five read-only instances can be created for a primary instance.
- Backup settings and temporary backup are not supported.
- Instance restoration is not supported.
- Data migration to read-only instances is not supported.
- Database creation and deletion are not supported.
- Account creation, deletion, authorization, and password changes are not supported.
- After a read-only instance is created, you cannot restore data by directly overwriting the primary instance with a backup set.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the **Distributed by Instance Role** section on the right side of the **Basic Information** page, click **Create**

Read-only Instance.

4. On the **Create Read-only RDS Instance** page, configure the read-only instance parameters.

Section	Parameter	Description
Region	**Region**	The region in which you want to create the read-only instance.
Specifications	**Database Engine**	The database engine of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	Engine Version	The version of the database engine, which is the same as that of the primary instance and cannot be changed.
	Edition	Set the value to **Read-only**.
	Instance Type	The instance type of the read-only instance. The instance type of the read-only instance can be different from that of the primary instance, and can be changed at any time to facilitate flexible upgrade and downgrade.
	Storage Capacity	The storage capacity of the read-only instance. To ensure sufficient I/O throughput for data synchronization, we recommend that you select at least the same instance type and storage capacity as the primary instance for the read-only instance. Valid values: 20 to 6000. Unit: GB. The value is in 1 GB increments.
Network Type	**Network Type**	The network type of the read-only instance. This must be the same as that of the primary instance and cannot be changed.
	VPC	The VPC in which you want to create the read-only instance.
	vSwitch	The vSwitch in the VPC.

5. Click **Submit**.

9.1.4.10.3. View details of read-only instances

This topic describes how to view details of read-only instances. You can go to the Basic Information page of a read-only instance from the Instances page or from the read-only instance list of the primary instance. Read-only instances are managed in the same manner as primary instances. The Basic Information page shows the management operations that can be performed.

View details of a read-only instance from the Instances page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click the ID of a read-only instance. The **Basic Information** page appears.

> **Note** In the instance list, **Instance Role** of read-only instances is displayed as Read-only Instance.

View details of a read-only instance from the Basic Information page of the primary instance

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. On the **Basic Information** page, move the pointer over the number below **Read-only Instances** in the **Distributed by Instance Role** section. The ID of the read-only instance is displayed.
4. Click the ID of the read-only instance to go to the Basic Information page of the read-only instance.

9.1.5. Accounts

9.1.5.1. Create an account

After you create an ApsaraDB RDS instance and configure its IP address whitelist, you must create a database and an account on the instance. This topic describes how to create privileged and standard accounts.

Context

ApsaraDB RDS for MySQL supports two types of database accounts: privileged and standard. You can manage all your accounts and databases in the ApsaraDB RDS console. For more information about permissions that can be granted to each type of account, see Account permissions.

Account type	Description
Privileged account	• You can create and manage privileged accounts by using the ApsaraDB RDS console or API operations. • You can create only a single privileged account on each ApsaraDB RDS instance. The privileged account can be used to manage all standard accounts and databases on the instance. • A privileged account allows you to manage permissions to a fine-grained level. For example, you can grant each standard account the permissions to query specific tables. • A privileged account has the permissions to disconnect all standard accounts on the instance.
Standard account	• You can create and manage standard accounts by using the ApsaraDB RDS console, API operations, or SQL statements. • You can create up to 500 standard accounts on an instance. • You must manually grant standard accounts the specific database permissions. • You cannot use a standard account to create, manage, or disconnect other accounts from databases.

Account type	Maximum number of databases	Maximum number of tables	Maximum number of accounts
Privileged account	Unlimited	< 200,000	Varies based on the engine parameter settings of the instance.
Standard account	500	< 200,000	Varies based on the engine parameter settings of the instance.

Create a privileged account

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts** to go to the **Accounts** page.
4. On the **Accounts** tab, click **Create Account**.
5. On the **Create Account** page, configure the following parameters.

Parameter	Description
Database Account	Enter the name of the account. The account name must meet the following requirements: ○ The name is 1 to 16 characters in length. ○ The name starts with a lowercase letter and ends with a lowercase letter or digit. ○ The name contains lowercase letters, digits, and underscores (_).
Account Type	Select Privileged Account.
Password	Enter the password of the account. The password must meet the following requirements: ○ The password is 8 to 32 characters in length. ○ The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include ! @ # $ % ^ & * ( ) _ + - =
Re-enter Password	Enter the password of the account again.
Description	Optional. Enter information about the account to facilitate subsequent management. The description can be up to 256 characters in length.

6. Click **Create**.

Reset the permissions of a privileged account

If an issue occurs on the privileged account, you can enter the password of the privileged account to reset permissions. For example, you can reset the permissions if the permissions are unexpectedly revoked.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts** to go to the **Accounts** page.
4. On the **Accounts** tab, find the privileged account and click **Reset Permissions** in the **Actions** column.
5. On the **Initialize Account** page, enter the password of the privileged account and click **OK**.

Create a standard account

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts** to go to the **Accounts** page.
4. On the **Accounts** tab, click **Create Account**.
5. On the **Create Account** page, configure the following parameters.

Parameter	Description
Database Account	Enter the name of the account. The account name must meet the following requirements: ○ The name is 1 to 16 characters in length. ○ The name starts with a lowercase letter and ends with a lowercase letter or digit. ○ The name contains lowercase letters, digits, and underscores (_).

Parameter	Description
Account Type	Select Standard Account.
Authorized Databases	Select one or more databases on which you want to grant permissions to the account. You can also leave this parameter empty at this time and authorize databases after the account is created. i. Select one or more databases from the Unauthorized Databases section and click **Add** to add them to the Authorized Databases section. ii. In the Authorized Databases section, select the **Read/Write**, **Read-only**, **DDL Only**, or **DML Only** permissions on each authorized database. If you want to grant the same permissions on multiple databases to the account, click the button in the upper-right corner of the section. The button may appear as **Set All to Read/Write**.
Password	Enter the password of the account. The password must meet the following requirements: ○ The password is 8 to 32 characters in length. ○ The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include ! @ # $ % ^ & * ( ) _ + - =
Re-enter Password	Enter the password of the account again.
Description	Optional. Enter information about the account to facilitate subsequent management. The description can be up to 256 characters in length.

6. Click **Create**.

Account permissions

Account type	Authorization type	Permission				
Privileged account	None	SELECT	INSERT	UPDATE	DELETE	CREATE
		DROP	RELOAD	PROCESS	REFERENCES	INDEX
		ALTER	CREATE TEMPORARY TABLES	LOCK TABLES	EXECUTE	REPLICATION SLAVE
		REPLICATION CLIENT	CREATE VIEW	SHOW VIEW	CREATE ROUTINE	ALTER ROUTINE
		CREATE USER	EVENT	TRIGGER	None	None

Account type	Authorization type	Permission				
Standard account	Read-only	SELECT	LOCK TABLES	SHOW VIEW	PROCESS	REPLICATION SLAVE
		REPLICATION CLIENT	None	None	None	None
	Read/write	SELECT	INSERT	UPDATE	DELETE	CREATE
		DROP	REFERENCES	INDEX	ALTER	CREATE TEMPORARY TABLES
		LOCK TABLES	EXECUTE	CREATE VIEW	SHOW VIEW	CREATE ROUTINE
		ALTER ROUTINE	EVENT	TRIGGER	PROCESS	REPLICATION SLAVE
		REPLICATION CLIENT	None	None	None	None
	DDL-only	CREATE	DROP	INDEX	ALTER	CREATE TEMPORARY TABLES
		LOCK TABLES	CREATE VIEW	SHOW VIEW	CREATE ROUTINE	ALTER ROUTINE
		PROCESS	REPLICATION SLAVE	REPLICATION CLIENT	None	None
	DML-only	SELECT	INSERT	UPDATE	DELETE	CREATE TEMPORARY TABLES
		LOCK TABLES	EXECUTE	SHOW VIEW	EVENT	TRIGGER
		PROCESS	REPLICATION SLAVE	REPLICATION CLIENT	None	None

9.1.5.2. Reset the password

You can use the ApsaraDB RDS console to reset the password of your database account.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts**.
4. Find an account and click **Reset Password** in the **Actions** column.

5. In the dialog box that appears, enter and confirm the new password, and then click **OK**.

> **Note** The password must meet the following requirements:
> - The password is 8 to 32 characters in length.
> - The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.
> - Special characters include
>
> ! @ # $ % ^ & * ( ) _ + - =

9.1.5.3. Edit account permissions

You can edit the account permissions of your ApsaraDB RDS instance at any time.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts**.
4. Find an account and click **Edit Permissions** in the **Actions** column.

> **Note** You can edit the permissions of a standard account. The permissions of privileged accounts can only be reset to the default settings and cannot be changed to a specific set of permissions.

5. Configure the following parameters.

Parameter	Description
Authorized Databases	In the **Unauthorized Databases** section, select a database and click **Add** to authorize the database. In the **Authorized Databases** section, select a database and click **Remove** to remove the permissions from the database.
Permission	You can set permissions on each database in the Authorized Database section. You can also click the button such as **Set All to Read/Write** in the upper-right corner to set the permissions of the account on all authorized databases. ○ **Read-only**: grants the account read-only permissions on databases. ○ **Read/Write**: grants the account read and write permissions on databases. ○ **DDL Only**: grants the account DDL permissions on databases. ○ **DML Only**: grants the account DML permissions on databases.

6. Click **OK**.

9.1.5.4. Delete an account

You can delete a database account in the ApsaraDB RDS console.

Prerequisites

You can use the console to delete privileged and standard accounts that are no longer used.

Procedure

1. Log on to the ApsaraDB for RDS console.

2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts**.
4. Find the account that you want to delete and click **Delete** in the Actions column.
5. In the message that appears, click **Confirm**.

> **Note** Accounts in the **Processing** state cannot be deleted.

9.1.6. Databases

9.1.6.1. Create a database

After you create an ApsaraDB RDS instance and configure its IP address whitelist, you must create a database and an account on the instance.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Databases**.
4. Click **Create Database**. On the page that appears, configure the following parameters.

Parameter	Description
Database Name	◦ The name must be 1 to 64 characters in length. ◦ The name must start with a letter and end with a letter or a digit. ◦ The name can contain lowercase letters, digits, underscores (_), and hyphens (-). ◦ The name must be unique within the instance.
Supported Character Sets	Select **utf8**, **gbk**, **latin1**, **utf8mb4**, or **all**. If you want to use other character sets, select **all**, and then select the required character set from the list.
Description	Optional. Enter information about the database to facilitate subsequent management. The description must be 2 to 256 characters in length.

5. Click **Create**.

9.1.6.2. Delete a database

You can delete databases that are no longer used in the ApsaraDB RDS console.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Databases**.
4. Find the database that you want to delete and click **Delete** in the **Actions** column.
5. In the Delete Database message, click **Confirm**.

9.1.7. Database connection

9.1.7.1. Change the endpoint and port number of an instance

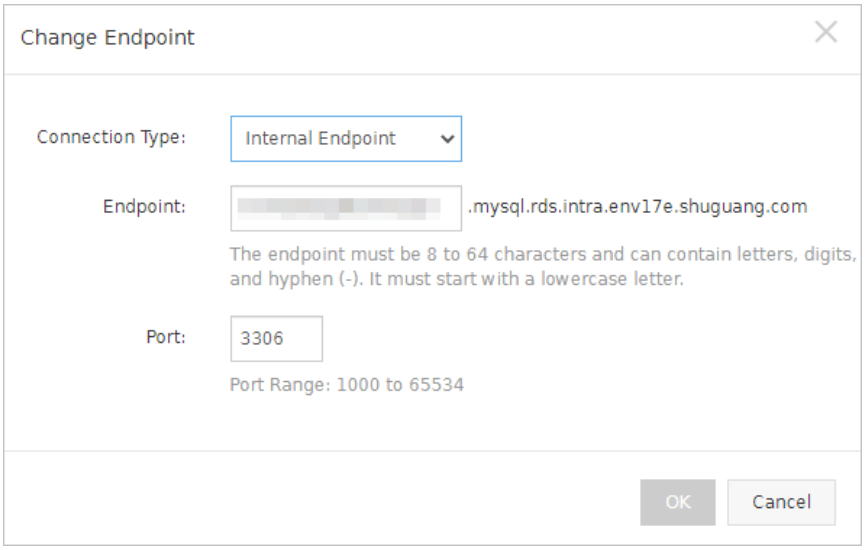
This topic describes how to view and change the endpoint and port number of an ApsaraDB RDS instance.

View the endpoint and port number

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the **Basic Information** section, view the internal and public endpoints and port numbers.

Change the endpoint and port number

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. In the upper-right corner of the Database Connection section, click **Change Endpoint**.
5. In the dialog box that appears, set Connection Type, Endpoint, and Port, and then click **OK**.



> **Note**
> - The prefix of an endpoint must be 8 to 64 characters in length and can contain letters, digits, and hyphens (-). It must start with a lowercase letter.
> - The port number must be a value within the range of 1000 to 65534.

9.1.7.2. Apply for and release an internal endpoint or a public endpoint for an instance

ApsaraDB RDS supports two types of endpoints: internal endpoints and public endpoints. The default type of the endpoint used to connect to an ApsaraDB RDS instance is determined by the network connection type selected when you create the instance. This topic describes how to apply for and release an internal endpoint or a public endpoint for an ApsaraDB RDS instance.

Apply for an internal endpoint or a public endpoint

If you set **Connection Type** to **Internet** when you create an ApsaraDB RDS instance, the database system assigns a public endpoint to the instance and you can apply for an internal endpoint. Otherwise, the database system assigns an internal endpoint to the instance, and you can apply for a public endpoint.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. Apply for an internal endpoint or a public endpoint:
   - To apply for a public endpoint, click **Apply for Public Endpoint**.
   - To apply for en internal endpoint, click **Apply for Internal Endpoint**.
5. In the message that appears, click **OK**.

Release an internal endpoint or a public endpoint

If an endpoint is no longer needed, you can release the endpoint to ensure instance security.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. Release an internal endpoint or a public endpoint:
   - To release a public endpoint, click **Release Public Endpoint**.
   - To release an internal endpoint, click **Release Internal Endpoint**.
5. In the message that appears, click **OK**.

FAQ

- Q: Can I change the endpoints and port numbers of my ApsaraDB RDS instance?

  A: No, you cannot change the endpoints of your ApsaraDB RDS instance. You can change the prefixes of the endpoints. You can also change the port numbers of your instance. For more information, see Change the endpoint and port number of an instance.

- Q: Can I configure the endpoints of my ApsaraDB RDS instances to static IP addresses?

  A: No, you cannot configure the endpoints of your ApsaraDB RDS instance to static IP addresses. Both primary/secondary switchovers and specification changes may cause changes to the IP addresses. We recommend that you connect to your instance by using an endpoint. This allows you to minimize impacts on your workloads and eliminates the need to modify configuration data on your application.

- Q: How do I connect to my ApsaraDB RDS instance by using a public endpoint?

  A: You can connect to your ApsaraDB RDS instance from an ECS instance or a database client. For more information, see Connect to an ApsaraDB RDS for MySQL instance.

9.1.7.3. Use DMS to log on to an ApsaraDB RDS instance

This topic describes how to use Data Management (DMS) to log on to an ApsaraDB RDS instance.

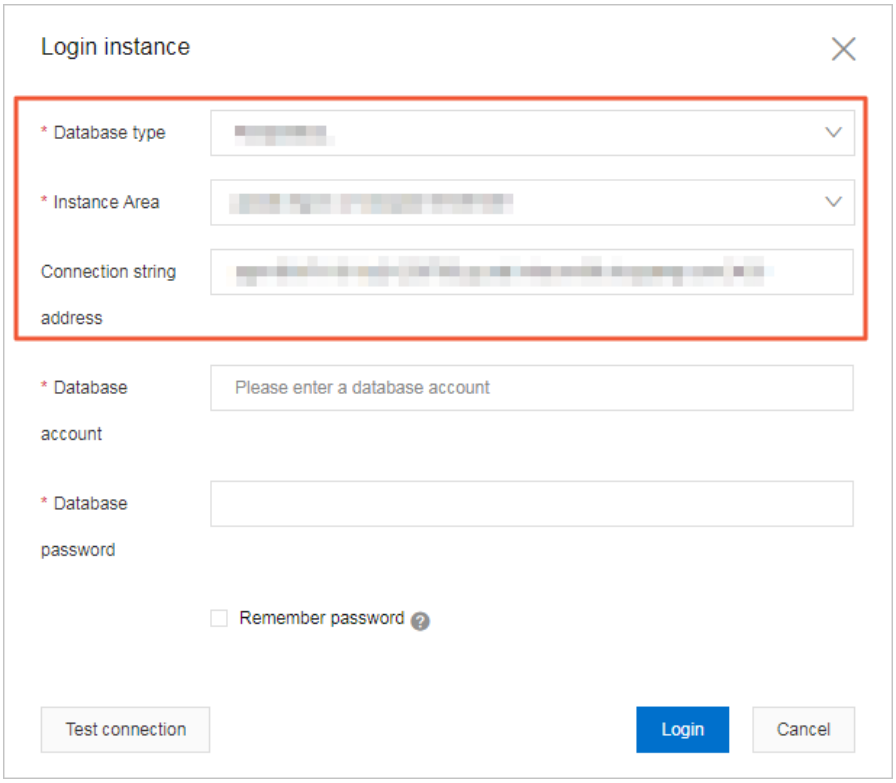
Prerequisites

An IP address whitelist is configured. For more information about how to configure an IP address whitelist, see Configure a whitelist.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.

3. Click **Log On to DB** in the upper-right corner of the page.
4. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

5. Click **Login**.

> **Note** If you want the browser to remember the password, select **Remember password** before you click **Login**.

9.1.7.4. Configure the hybrid access solution for an instance

This topic describes how to configure the hybrid access solution for an ApsaraDB RDS instance. This solution allows you to retain both the classic network endpoint and Virtual Private Cloud (VPC) endpoint of your ApsaraDB RDS instance. This way, you can migrate your ApsaraDB RDS instance from the classic network to a VPC without network interruptions.

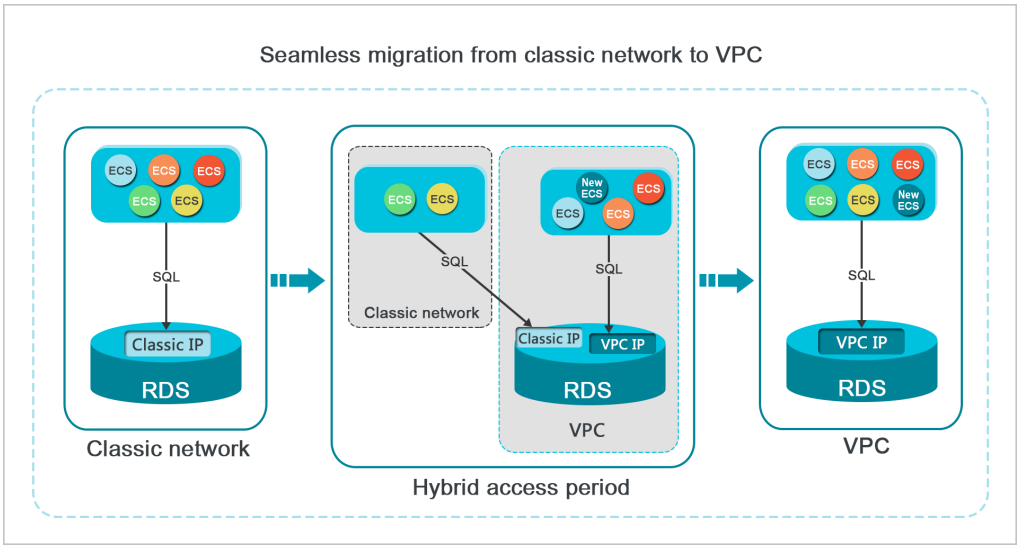
Background information

When you migrate your ApsaraDB RDS instance from the classic network to a VPC, the internal classic network endpoint of the instance changes to the internal VPC endpoint. In this case, the endpoint remains unchanged, but the IP address that is bound to the endpoint changes. This change causes a network interruption of 30 seconds or less, and Elastic Compute Service (ECS) instances located in the classic network can no longer connect to your ApsaraDB RDS instance over an internal network. To facilitate a smooth migration, ApsaraDB RDS provides the hybrid access solution.

Hybrid access refers to the ability of your ApsaraDB RDS instance to be connected by both ECS instances located in the classic network and ECS instances located in a VPC. During the validity period of the hybrid access solution, ApsaraDB RDS retains the internal classic network endpoint and generates an internal VPC endpoint. This prevents network interruptions when you migrate your instance from the classic network to a VPC.

For security and performance purposes, we recommend that you use only the internal VPC endpoint. You must specify a validity period for the hybrid access solution. When the validity period expires, ApsaraDB RDS releases the internal classic network endpoint and applications are unable to use the endpoint to connect to your instance. Therefore, you must add the internal VPC endpoint to your applications before the validity period expires. This ensures a smooth migration and prevents interruptions to your workloads.

For example, assume that a company uses the hybrid access solution to migrate its ApsaraDB RDS instance from the classic network to a VPC. During the validity period of the hybrid access solution, some applications use the internal VPC endpoint to connect to the ApsaraDB RDS instance, whereas the others continue to use the internal classic network endpoint to connect to the instance. When all applications of the company can use the internal VPC endpoint to connect to the instance, the internal classic network endpoint can be released. The following figure demonstrates the scenario.



Limits

During the validity period of the hybrid access solution, your ApsaraDB RDS instance has the following limits:

- The network type of the instance cannot be changed to classic network.
- The instance cannot be migrated to another zone.

Prerequisites

- The ApsaraDB RDS instance resides in the classic network.
- Available VPCs and vSwitches exist in the zone where the ApsaraDB RDS instance resides.
- Your ApsaraDB RDS instance provides an internal endpoint. If no internal endpoint exists, you must apply for one. For more information, see Apply for and release an internal endpoint or a public endpoint for an instance.

Change the network type from classic network to VPC

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. On the Instance Connection tab, click **Switch to VPC**.
5. In the Switch to VPC dialog box, select a VPC and a vSwitch and specify whether to retain the classic network endpoint.

   Clear or select the **Reserve Original Classic Network Endpoint** check box based on the details described in the following table.

Action	Description
Clear the Reserve Original Classic Network Endpoint check box	The classic network endpoint is not retained and changes to a VPC endpoint. When you change the network type from classic network to VPC, a network interruption of 30 seconds occurs. When this occurs, ECS instances located in the classic network are disconnected from your ApsaraDB RDS instance.
Select the Reserve Original Classic Network Endpoint check box	The classic network endpoint is retained, and a new VPC endpoint is generated. In this case, your ApsaraDB RDS instance runs in hybrid access mode. Both ECS instances located in the classic network and ECS instances located in the selected VPC can access your ApsaraDB RDS instance over an internal network. When you change the network type from classic network to VPC, no network interruptions occur. ECS instances located in the classic network are still connected with your ApsaraDB RDS instance until the classic network endpoint expires. Specify the expiration date of the classic network endpoint. Before the classic network endpoint expires, you must add the new VPC endpoint to your applications that run on the ECS instances located in the selected VPC. This allows ApsaraDB RDS to migrate your workloads to the selected VPC without network interruptions.

6. Add the internal IP addresses of ECS instances located in the selected VPC to an IP address whitelist of the VPC network type. This allows the ECS instances to connect to your ApsaraDB RDS instance over an internal network. If no IP address whitelists of the VPC network type are available, create one. For more information, see Configure a whitelist.

Change the expiration date of the internal classic network endpoint

During the validity period of the hybrid access solution, you can change the expiration date of the classic network endpoint based on your business requirements. The expiration date is immediately recalculated starting from the day when you make the change. For example, assume that the classic network endpoint is configured to expire on August 18, 2017. On August 15, 2017, you increase the validity period of the classic network endpoint by 14 days. In this case, ApsaraDB RDS releases the classic network endpoint on August 29, 2017.

To change the expiration date, perform the following operations:

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. Click **Change Expiration Time**.
5. In the **Change Expiration Time** dialog box, select an expiration date and click **OK**.

9.1.7.5. Change the network type of an instance

This topic describes how to change the network type of an ApsaraDB RDS instance between classic network and Virtual Private Cloud (VPC).

Context

- Classic network: ApsaraDB RDS instances in the classic network are not isolated. Unauthorized access to these instances can be blocked only by IP address whitelists.
- VPC: Each VPC is an isolated virtual network. We recommend that you select the VPC type because it is more secure than the classic network.

  You can configure route tables, CIDR blocks, and gateways in a VPC. To smoothly migrate applications to the cloud, you can use the leased line or VPN method to create a virtual data center that consists of your data center and a VPC.

Change the network type from VPC to classic network

Precautions

- After you change the network type from VPC to classic network, the internal endpoint of your ApsaraDB RDS instance remains unchanged. However, the IP address that is associated with the internal endpoint changes.
- After you change the network type from VPC to classic network, Elastic Compute Service (ECS) instances located in the same VPC as your ApsaraDB RDS instance can no longer connect to your ApsaraDB RDS instance by using the internal endpoint. You must update the endpoint for the applications deployed on the ECS instances.
- When you change the network type, a network interruption of 30 seconds may occur. To avoid business interruption, change the network type during off-peak hours or make sure that your application is configured to automatically reconnect to the instance.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. In the upper-right corner of the Database Connection section, click **Switch to Classic Network**.
5. In the message that appears, click **OK**.

Change the network type from classic network to VPC

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. In the upper-right corner of the Database Connection section, click **Switch to VPC**.
5. In the Switch to VPC dialog box, select a VPC and a vSwitch, and then select or clear **Reserve Original Classic Network Endpoint**. Click OK. For more information about **Reserve Original Classic Network Endpoint**, see Hybrid access from both the classic network and VPCs.

9.1.7.6. Change the VPC and vSwitch for an instance

This topic describes how to change the virtual private cloud (VPC) and vSwitch for an ApsaraDB RDS instance.

Prerequisites

The instance is deployed in a VPC.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.

4. In the upper-right corner of the Database Connection section, click **Switch vSwitch**.
5. Select a VPC and a vSwitch, and then click **OK**.
6. In the message that appears, click **OK**.

> **Note**
> - A network interruption of 30 seconds may occur when you switch the VPC and vSwitch of an ApsaraDB RDS instance. Make sure that your application is configured to automatically reconnect to the instance.
> - We recommend that you clear the cache immediately after the instance is switched to a new VPC and vSwitch. Otherwise, data can be read but not written.

9.1.8. Database proxy

9.1.8.1. Configure dedicated proxy

This topic describes the dedicated proxy feature provided by ApsaraDB RDS. The dedicated proxy feature provides advanced features such as read/write splitting, connection pooling, and transaction splitting.

Context

The dedicated proxy feature uses dedicated computing resources. This feature has the following benefits:

- A unified proxy endpoint is provided to connect to all the dedicated proxies that are enabled on your ApsaraDB RDS instance. This reduces maintenance costs by eliminating the need to update the endpoints on your application. The proxy endpoint remains valid until you release the dedicated proxies. For example, you may enable read/write splitting during peak hours, and then release read-only instances and disable read/write splitting after peak hours. In these cases, you do not need to update the endpoints on your application because the proxy endpoint remains connected.
- Dedicated proxies serve your ApsaraDB RDS instance and its read-only instances exclusively. You do not need to compete with other users for resources. This ensures service stability.
- Dedicated proxies are scalable. You can add dedicated proxies based on your business requirements to handle more workloads.

Limits

- Dedicated proxies do not support Secure Sockets Layer (SSL) encryption.
- Dedicated proxies do not support compression protocols.

Precautions

- When you change the specifications of your ApsaraDB RDS instance or its read-only instances, a network interruption may occur.
- If you connect your application to the proxy endpoint, all requests that are encapsulated in transactions are routed to your ApsaraDB RDS instance. This applies when the transaction splitting feature is not enabled.
- If a proxy endpoint is used to implement read/write splitting, read consistency cannot be ensured for the requests that are not encapsulated in transactions. If you require read consistency for these requests, you can encapsulate these requests in transactions.
- If a proxy endpoint is used for connection, the `SHOW PROCESSLIST` statement returns a result set for each query. The result set consists of the query results from the primary and read-only instances.
- If you execute multi-statements or stored procedures, the read/write splitting feature is disabled and all subsequent requests over the current connection are routed to the primary ApsaraDB RDS instance. To enable the read/write splitting feature again, you must close the current connection and establish a new connection.
- The dedicated proxy feature supports the `/*FORCE_MASTER*/` and `/*FORCE_SLAVE*/` hints. However,

requests that contain hints have the highest route priorities and are not constrained by consistency or transaction limits. Before you use these hints, you must check whether these hints are suitable for your workloads. A hint cannot contain statements that change environment variables. An example is `/*FORCE_SLAVE*/ set names utf8;`. Otherwise, an error may occur in the subsequent procedure.

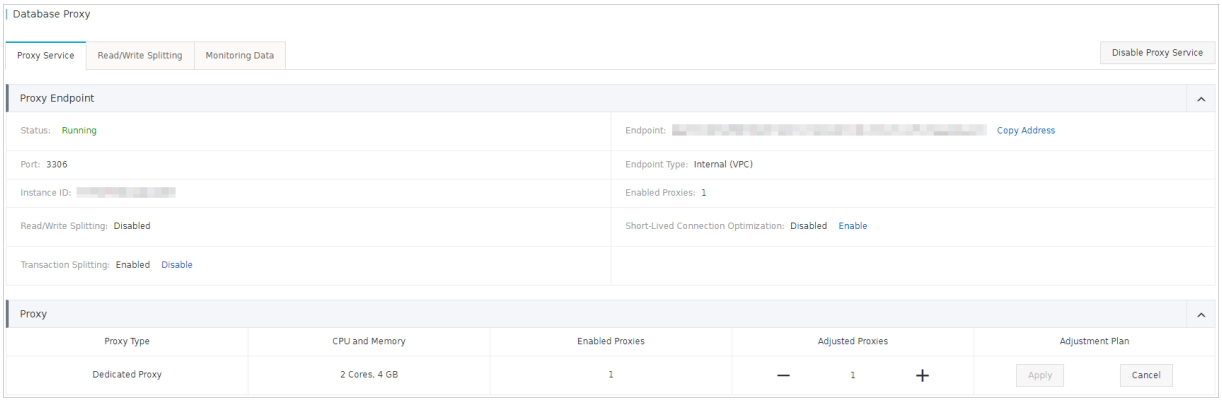
- After you enable the dedicated proxy feature, each connection is replicated to the primary ApsaraDB RDS instance and all of its read-only instances in compliance with the 1:N connection model. We recommend that you specify the same connection specifications for these instances. If these instances have different connection specifications, the number of connections allowed depends on the lowest connection specifications among these instances.
- If you create or restart a read-only instance after you enable the dedicated proxy feature, only the requests sent over new connections are routed to the new or restarted read-only instance.
- The **max_prepared_stmt_count** parameter must be set to the same value for the primary ApsaraDB RDS instance and all of its read-only instances.

Enable the dedicated proxy feature

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. Click **Enable now**.

Overview of the Proxy Service tab

When the dedicated proxy feature is enabled, you can use the generated proxy endpoint to implement features such as read/write splitting, connection pooling, and transaction splitting.



Section	Parameter	Description
	Instance ID	The ID of the primary ApsaraDB RDS instance.
	Enabled Proxies	The number of enabled dedicated proxies. You can enable more dedicated proxies to increase the maximum number of requests that can be processed. After public preview ends, you must pay for enabled proxies.
	Read/Write Splitting	Specifies whether to enable the read/write splitting feature for the proxy endpoint. For more information, see Read/write splitting.

Section	Parameter	Description
Proxy Endpoint	Short-Lived Connection Optimization	The type of connection pool for the proxy endpoint. This feature is suitable for scenarios where PHP short-lived connections are established. For more information, see Short-lived connection optimization. **Note** You can click **Enable** or **Disable** to the right of Short-Lived Connection Optimization to enable or disable this feature.
	Transaction Splitting	Specifies whether to enable the transaction splitting feature for the proxy endpoint. For more information, see Transaction splitting. **Note** You can click **Enable** or **Disable** to the right of Transaction Splitting to enable or disable this feature.
	Endpoint	The proxy endpoint that is generated when the dedicated proxy feature is enabled. This endpoint connects to all the dedicated proxies that are enabled on the ApsaraDB RDS instance. The read/write splitting feature is also bound to this endpoint. **Note** You can click **Copy Address** to the right of Endpoint to copy the endpoint.
	Port	The port that is used to connect to the proxy endpoint.
	Endpoint Type	The network type of the proxy endpoint.
Proxy	Proxy Type	The type of proxy that is enabled. Only **Dedicated Proxy** is supported.
	CPU and Memory	The CPU and memory specifications of the dedicated proxies. Only 2 Cores, 4 GB is supported.
	Enabled Proxies	The number of dedicated proxies that are enabled on the primary ApsaraDB RDS instance. Up to 60 dedicated proxies are supported. **Note** We recommend that you use the following formula to determine the number of dedicated proxies to specify: (Total number of CPU cores of your ApsaraDB RDS instance and its read-only instances)/8, rounded up to the nearest integer. For example, if your ApsaraDB RDS instance has 8 CPU cores and its read-only instances have 4 CPU cores, the recommended number of dedicated proxies is 2, as calculated in the following formula: ⌈(8 + 4)/8⌉ = 2.

Adjust the number of dedicated proxies

Note When you adjust the number of dedicated proxies, a network interruption may occur. Make sure that your application is configured to automatically reconnect to the instance.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. In the **Proxy** section of the Proxy Service tab, change the number in the **Adjusted Proxies** column and click **Apply** in the **Adjustment Plan** column.
5. In the Configure Proxy Resources dialog box, select **Migrate Immediately** to apply the change. You can also select **Next Maintenance Period** to set a maintenance window for the change to take effect. Click **OK**.

View the monitoring data of dedicated proxies

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. Click the **Monitoring Data** tab.
5. Select a time range to view the **CPU Utilization** metric within that time range.

Disable the dedicated proxy feature

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. In the upper-right corner of the page, click **Disable Proxy Service**.
5. In the message that appears, click **OK**.

9.1.8.2. Configure short-lived connection optimization

This topic describes the short-lived connection optimization feature provided by ApsaraDB RDS in its dedicated proxy feature.

Prerequisites

The database proxy feature is enabled for your ApsaraDB RDS instance. For more information, see Dedicated proxy.

Context

The short-lived connection optimization feature is used to reduce workloads on the ApsaraDB RDS instance caused by frequent short-lived connections. When a client is disconnected, the system checks whether the closed connection is idle. If the connection is considered idle, the dedicated proxy retains the connection in the connection pool for a short period of time. When the client initiates a request for access to your instance again, the dedicated proxy searches the connection pool for an idle connection that matches the request. The connection pool is matched based on the values of the user, clientip, and dbname fields in the request. If the dedicated proxy finds an idle connection that matches the request, it reuses the matched idle connection. If no idle connection can be matched, a new connection is established with your instance to reduce database connection overheads.

> **Note** The short-lived connection optimization feature does not reduce the number of concurrent connections with the instance. It decreases the frequency at which connections are established between an application and your instance to reduce overheads of the primary MySQL thread and improve efficiency to process business requests. However, idle connections in the connection pool still occupy the database threads for a short period of time.

Precautions

You cannot configure different permissions for different source IP addresses by using the same account. Otherwise, errors may occur when connections in the connection pool are reused. For example, if a user account has permissions on database_a when its source IP address is 192.168.1.1 but does not have permissions on database_a when its source IP address is 192.168.1.2, the short-lived connection optimization feature may encounter permission errors.

Enable short-lived connection optimization

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. On the **Proxy Service** tab, click **Enable** to the right of **Short-Lived Connection Optimization**.

9.1.8.3. Configure transaction splitting

This topic describes the transaction splitting feature provided by the dedicated proxy of ApsaraDB RDS. This feature identifies and distributes read requests initiated before write requests within a transaction to read-only instances. This reduces workloads on the primary instance.

Prerequisites

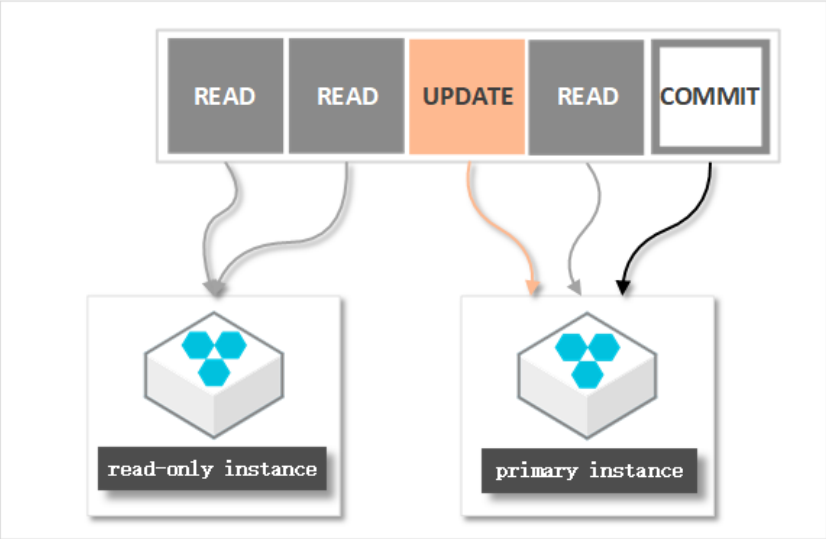
The dedicated proxy feature is enabled for your ApsaraDB RDS instance. For more information, see Dedicated proxy.

Context

By default, the dedicated proxy sends all requests in transactions to the primary instance to ensure the correctness of the transactions. If the framework encapsulates all requests in transactions, the primary instance becomes heavily loaded. In this case, you can enable the transaction splitting feature.

When transaction splitting is enabled and the default isolation level READ COMMITTED is used, the ApsaraDB RDS instance starts a transaction only for write requests when autocommit is disabled (set autocommit=0). Read requests that arrive before the transaction is started are distributed to read-only instances by the load balancer.

> **Note**
>
> - Explicit transactions do not support splitting, such as transactions started by using the BEGIN or START statement.
> - After you enable the transaction splitting feature, global consistency cannot be ensured. Before you enable this feature, we recommend that you evaluate whether this feature is suitable for your workloads.



Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. On the **Proxy Service** tab, click **Enable** to the right of **Transaction Splitting**.

> Note
> - When you no longer need transaction splitting, you can click **Disable** to the right of **Transaction Splitting**.
> - The operation to enable or disable transaction splitting takes effect only on new connections.

9.1.8.4. Read/write splitting

9.1.8.4.1. Enable read/write splitting

This topic describes the read/write splitting feature. This feature allows ApsaraDB RDS to route read and write requests to the primary and read-only instances based on the dedicated proxy endpoint (also called read/write splitting endpoint).

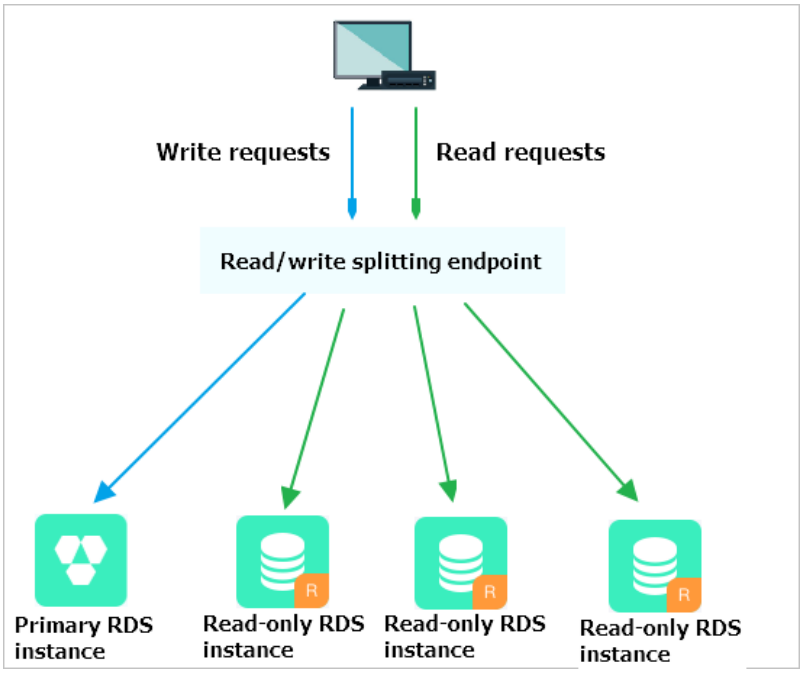
Prerequisites

- The database proxy or dedicated proxy feature is enabled. For more information, see Enable the dedicated proxy feature.
- At least one read-only instance is created. For more information about how to create a read-only instance, see Create a read-only instance.

Context

If your database system processes a large number of read requests and a small number of write requests, a single primary ApsaraDB RDS instance may fail to efficiently process the read requests. This may interrupt your workloads. In this case, you can create one or more read-only ApsaraDB RDS instances to offload read requests from the primary instance and increase the read capability of your database system. For more information, see Create a read-only instance.

After you create read-only instances, you can enable read/write splitting. In this case, a read/write splitting endpoint is provided. After you add the endpoint to your application, write requests are routed to the primary instance and read requests are routed to the read-only instances.



Differences between the read/write splitting endpoint and the internal and public endpoints

After you enable read/write splitting and add the read/write splitting endpoint to your application, all requests are first routed to this endpoint, and then to the primary and read-only instances based on the request types and the read weights of these instances.

If the internal or public endpoint of the primary instance is added to your application, all requests are routed to the primary instance. To implement read/write splitting, you must add the endpoints and read weights of the primary and read-only instances to your application.

Logic to route requests

- The following requests are routed only to the primary instance:
  - Requests for DML statements such as INSERT, UPDATE, DELETE, and SELECT FOR UPDATE
  - Requests for DDL statements used to perform operations such as creating databases or tables, deleting databases or tables, and changing schemas or permissions
  - All requests that are encapsulated in transactions
  - Requests for user-defined functions
  - Requests for stored procedures
  - Requests for EXECUTE statements
  - Requests for multi-statements
  - Requests that involve temporary tables
  - Requests for SELECT last_insert_id() statements
  - All requests to query or modify user environment variables
  - Requests for SHOW PROCESSLIST statements
  - All requests for KILL statements in SQL (not KILL commands in Linux)
- The following requests are routed to the primary instance or its read-only instances:
  - Read requests that are not encapsulated in transactions

- Requests for COM_STMT_EXECUTE statements

- The following requests are routed to all the primary and read-only instances:
  - All requests to modify system environment variables
  - Requests for USE statements
  - Requests for COM_STMT_PREPARE statements
  - Requests for COM_CHANGE_USER, COM_QUIT, and COM_SET_OPTION statements

Benefits

- Easier maintenance by using a unified endpoint

  If you do not enable the read/write splitting feature, you must add the endpoints of the primary and read-only instances to your application. This way, your database system routes write requests to the primary instance and read requests to the read-only instances.

  If you enable the read/write splitting feature, you can use a dedicated proxy endpoint to implement read/write splitting. After your application is connected to this endpoint, your database system routes read and write requests to the primary and read-only instances based on the read weights of these instances. This reduces maintenance costs.

  You can also create read-only instances to improve the read capability of your database system. You do not need to modify the configuration data on your application.

- Higher performance and lower maintenance costs by using a native link

  You can build your own proxy layer on the cloud to implement read/write splitting. In this case, data needs to be parsed and forwarded by multiple components before the data reaches your database system. As a result, response latencies increase. The read/write splitting feature is built in the ApsaraDB RDS ecosystem and can efficiently reduce response latencies, increase processing speeds, and reduce maintenance costs.

- Ideal in various use scenarios based on configurable read weights and thresholds

  You can specify the read weights of the primary and read-only instances. You can also specify the latency threshold for data replication to the read-only instances.

- High availability based on instance-level health checks

  The read/write splitting feature enables ApsaraDB RDS to actively check the health status of the primary and read-only instances. If a read-only instance unexpectedly breaks down or its data replication latency exceeds the specified threshold, ApsaraDB RDS stops routing read requests to the instance. ApsaraDB RDS redirects the read requests that are destined for the faulty read-only instance to healthy instances in your database system. This ensures service availability in the event of faults on individual read-only instances. After the faulty read-only instance is recovered, ApsaraDB RDS resumes routing read requests to the instance.

  > **Note** To avoid single points of failure (SPOFs), we recommend that you create at least two read-only instances.

Precautions

- When you change the specifications of your ApsaraDB RDS instance or its read-only instances, a network interruption may occur.
- After you create a read-only instance, only the requests over new connections can be routed to the read-only instance.
- The dedicated proxy endpoint does not support SSL encryption.
- The dedicated proxy endpoint does not support compression.
- If a dedicated proxy endpoint is used to connect to your database system, all the requests that are encapsulated in transactions are routed to the primary instance.
- If a dedicated proxy endpoint is used to implement read/write splitting, the read consistency of the requests that are not encapsulated in transactions cannot be ensured. If you require read consistency for these requests,

you can encapsulate these requests in transactions.

- If a dedicated proxy endpoint is used for connection, the `SHOW PROCESSLIST` statement returns a result set for each query. The result set consists of the query results from the primary and read-only instances.
- If the short-lived connection optimization feature is enabled, the `SHOW PROCESSLIST` statement may return idle connections.
- If you execute multi-statements or stored procedures, the read/write splitting feature is disabled and all the subsequent requests over the current connection are routed to the primary ApsaraDB RDS instance. To enable the read/write splitting feature again, you must close the current connection and establish a new connection.
- The dedicated proxy feature supports the `/*FORCE_MASTER*/` and `/*FORCE_SLAVE*/` hints. However, requests that contain hints have the highest route priorities and are not constrained by consistency or transaction limits. Before you use these hints, you must check whether these hints are suitable for your workloads. A hint cannot contain statements that change environment variables. An example is `/*FORCE_SLAVE*/ set names utf8;`. Otherwise, an error may occur in the subsequent procedure.

Prerequisites

A read-only instance is created for the primary instance. For more information, see Create a read-only instance.

Enable read/write splitting

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. On the **Read/Write Splitting** tab, click **Enable now**.
5. Configure the following parameters.

Parameter	Description
Latency Threshold	The maximum latency that is allowed for data replication from the primary instance to its read-only instances. If the latency of data replication to a read-only instance exceeds the specified threshold, ApsaraDB RDS stops routing read requests to the instance. This applies even if the instance has a high read weight. Valid values: 0 to 7200. Unit: seconds. The read-only instances may replicate data from the primary instance at a specific latency due to SQL statement execution limits. We recommend that you set this parameter to a value that is greater than or equal to 30.
Read Weight Distribution	The read weight of each instance in your database system. A higher read weight indicates more read requests to process. For example, assume that your primary instance is attached with three read-only instances, and the read weights of the primary and read-only instances are 0, 100, 200, and 200. In this case, your primary instance processes only write requests, and the three read-only instances process all of the read requests at the 1:2:2 ratio. ◦ **Automatic Distribution**: Your database system assigns a read weight to each instance based on the instance specifications. After you create a read-only instance, your database system assigns a read weight to the read-only instance and adds the read-only instance to the read/write splitting link. ◦ **Customized Distribution**: You must manually specify the read weight of each instance. Valid values: 0 to 10000. After you create a read-only instance, ApsaraDB RDS sets the read weight of the read-only instance to 0. You must manually modify the read weight of the created read-only instance.

6. Click **OK**.

9.1.8.4.2. Configure read/write splitting

This topic describes how to configure the latency threshold and specify read weights for an ApsaraDB RDS instance in the ApsaraDB RDS console.

Prerequisites

Read/write splitting is enabled. For more information, see Enable read/write splitting.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. On the **Read/Write Splitting** tab, click **Configure Read/Write Splitting**.
5. Configure the following parameters.

Parameter	Description
Latency Threshold	The maximum latency that is allowed for data replication from the primary instance to its read-only instances. If the latency of data replication to a read-only instance exceeds the specified threshold, ApsaraDB RDS stops routing read requests to the instance. This applies even if the instance has a high read weight. Valid values: 0 to 7200. Unit: seconds. The read-only instances may replicate data from the primary instance at a specific latency due to SQL statement execution limits. We recommend that you set this parameter to a value that is greater than or equal to 30.
Read Weight Distribution	The read weight of each instance in your database system. A higher read weight indicates more read requests to process. For example, assume that your primary instance is attached with three read-only instances, and the read weights of the primary and read-only instances are 0, 100, 200, and 200. In this case, your primary instance processes only write requests, and the three read-only instances process all of the read requests at the 1:2:2 ratio. ◦ **Automatic Distribution**: Your database system assigns a read weight to each instance based on the instance specifications. After you create a read-only instance, your database system assigns a read weight to the read-only instance and adds the read-only instance to the read/write splitting link. ◦ **Customized Distribution**: You must manually specify the read weight of each instance. Valid values: 0 to 10000. After you create a read-only instance, ApsaraDB RDS sets the read weight of the read-only instance to 0. You must manually modify the read weight of the created read-only instance.

6. Click **OK**.

9.1.8.4.3. Disable read/write splitting

This topic describes how to disable the read/write splitting feature of an ApsaraDB RDS instance in the ApsaraDB RDS console.

Prerequisites

Read/write splitting is enabled. For more information, see Enable read/write splitting.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Proxy**.
4. On the **Read/Write Splitting** tab, click **Disable Read/Write Splitting**.

5. In the message that appears, click **Confirm**.

9.1.9. Monitoring and alerts

9.1.9.1. View resource and engine monitoring data

The ApsaraDB RDS console provides a variety of performance metrics to monitor the status of your instances.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Monitoring and Alerts**.
4. On the **Monitoring and Alerts** page, select **Resource Monitoring** or **Engine Monitoring**, and select a time range to view the corresponding monitoring data. The following table describes the metrics.

Monitoring type	Metric	Description
Resource Monitoring	**Disk Space (MB)**	The disk space usage of the instance. It consists of the following items: ○ Instance size ○ Data usage ○ Log size ○ Temporary file size ○ Other system file size Unit: MB.
	IOPS (Input/Output Operations per Second)	The number of input/output operations per second (IOPS) of the instance.
	Total Connections	The number of active connections to the instance and the total number of connections to the instance.
	CPU Utilization and Memory Usage (%)	The CPU utilization and memory usage of the instance. These metrics do not include the CPU utilization and memory usage for the operating system.
	Network Traffic (KB)	The inbound and outbound traffic of the instance per second. Unit: KB.
	Transactions per Second (TPS)/Queries per Second (QPS)	The average number of transactions per second (TPS) and the average number of SQL statements executed per second.
	InnoDB Buffer Pool Read Hit Ratio, Usage Ratio, and Dirty Block Ratio (%)	The read hit ratio, usage ratio, and dirty block ratio of the InnoDB buffer pool.
	InnoDB Read/Write Volume (KB)	The amount of data that InnoDB reads and writes per second. Unit: KB.
	InnoDB Buffer Pool Read/Write Frequency	The number of read and write operations that InnoDB performs per second.

Monitoring type	Metric	Description
Engine Monitoring	**InnoDB Log Read/Write/fsync**	The average frequency of physical writes to log files per second by InnoDB, the frequency of log write requests, and the average frequency of fsync writes to log files.
	Temporary Tables Automatically Created on Hard Disk when MySQL Statements Are Executed	The number of temporary tables that are automatically created on the hard disk when the database executes SQL statements.
	MySQL_COMDML	The number of SQL statements that the database executes per second. The following SQL statements are included: ○ Insert ○ Delete ○ Insert_Select ○ Replace ○ Replace_Select ○ Select ○ Update
	MySQL_RowDML	The numbers of operations that InnoDB performs per second. The following items are included: ○ The number of physical writes to log files per second ○ The number of rows that are read, updated, deleted, and inserted from InnoDB tables per second
	MyISAM Read/Write Frequency	The numbers of operations that MyISAM performs per second. The following items are included: ○ The number of MyISAM reads and writes from the buffer pool per second ○ The number of MyISAM reads and writes from the hard disk per second
	MyISAM Key Buffer Read/Write/Usage Ratio (%)	The read hit ratio, write hit ratio, and usage of the MyISAM key buffer per second.

9.1.9.2. Set a monitoring frequency

This topic describes how to set a monitoring frequency for an ApsaraDB RDS instance.

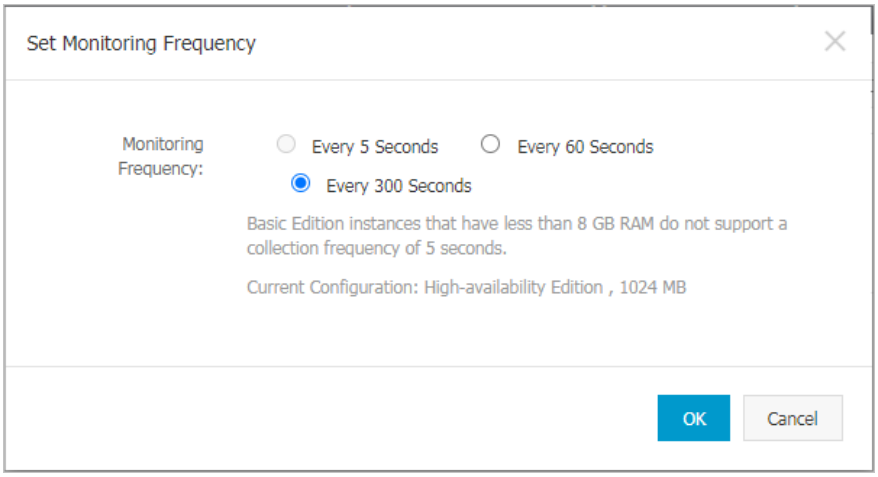
Context

ApsaraDB RDS provides the following monitoring frequencies:

- Every 5 seconds for the first seven days. After the seven days, performance metrics are monitored every 60 seconds.
- Every 60 seconds.
- Every 300 seconds.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Monitoring and Alerts**.
4. On the **Monitoring** tab, click **Set Monitoring Frequency**.
5. In the **Set Monitoring Frequency** dialog box, select a new monitoring frequency.



> **Note** If an ApsaraDB RDS instance runs the RDS Basic Edition or its memory capacity is less than 8 GB, the **Every 5 Seconds** monitoring frequency is not supported.

6. Click **OK**.

9.1.10. Data security

9.1.10.1. Configure an IP address whitelist for an ApsaraDB RDS instance

After you create an ApsaraDB RDS instance, you must add the IP addresses or CIDR blocks that are used for database access to the IP address whitelist of the instance to ensure database security and reliability.

Context

IP address whitelists make your ApsaraDB RDS instance more secure and do not interrupt the operations of your ApsaraDB RDS instance when you configure whitelists. We recommend that you maintain your IP address whitelists on a regular basis.

To configure a whitelist, you can perform the following operations:

- Configure an IP address whitelist: Add IP addresses to allow them to connect to the ApsaraDB RDS instance.
- Configure an Elastic Compute Service (ECS) security group: Add an ECS security group for the ApsaraDB RDS instance to allow ECS instances in the group to connect to the ApsaraDB RDS instance.

Precautions

- The default IP address whitelist can be modified or cleared, but cannot be deleted.
- You can add up to 1,000 IP addresses or CIDR blocks to a whitelist. If you want to add a large number of IP addresses, we recommend that you merge them into CIDR blocks, such as 192.168.1.0/24.

Introduction to IPv6

IPv4 addresses are widely used, but the limited number of IPv4 addresses restricts the development of the Internet. Compared with IPv4 addresses, IPv6 addresses are more sufficient and allow more types of devices to access the Internet. ApsaraDB RDS supports both IPv4 and IPv6 addresses.

The following table describes the differences between IPv4 and IPv6.

Item	IPv4	IPv6
Address length	32 bits (4 bytes)	128 bits (16 bytes)
Number of addresses	2^32	2^128
Address format	xxx.xxx.xxx.xxx Where xxx is a decimal number that can range from 0 to 255. Each x is a decimal integer, and leading zeros can be omitted. Example: 192.168.1.1	xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx Where each x is a hexadecimal number, and leading zeros can be omitted. You can use a double colon (::) once in an IPv6 address to indicate a series of zeros. Example: CDDC:0000:0000:0000:8475:1111:3900:2020
Address Resolution Protocol (ARP)	Uses broadcast ARP Request frames to resolve an IP address to a link layer address.	Uses multicast neighbor solicitation messages to resolve an IP address to a link layer address.
Security	Implements a security mechanism based on applications and cannot provide protections at the IP layer.	Supports packet fragmentation to ensure data confidentiality and integrity and provides security at the IP layer.
LAN connection	Connects to LANs by using network interfaces.	Can work with Ethernet adapters and is supported over virtual Ethernet networks between logical partitions.
Address type	• Unicast address • Multicast address • Broadcast address	• Unicast address • Multicast address • Anycast address

Create an IP address whitelist

Each IP address whitelist of an ApsaraDB RDS instance can contain **IPv4** or **IPv6** addresses. By default, the system provides an IP address whitelist of the **IPv4** type. If you want an IP address whitelist of the **IPv6** type, manually create one.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. On the **Whitelist Settings** tab, click **Create Whitelist**. In the dialog box that appears, configure the following parameters.

Parameter	Description

Parameter	Description
Whitelist Name	The name of the IP address whitelist. **Note** ◦ The name can contain lowercase letters, digits, and underscores (_). ◦ The name must start with a lowercase letter and end with a lowercase letter or digit. ◦ The name must be 2 to 32 characters in length.
IP Type	The IP type of the IP address whitelist. Valid values: ◦ **IPv4** ◦ **IPv6** **Note** For more information about the differences between IPv4 and IPv6, see the "Introduction to IPv6" section of this topic.
IP Addresses	The IP addresses that are allowed to access the instance.

Configure an IP address whitelist

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. On the **Whitelist Settings** tab, click **Edit** corresponding to an IP address whitelist.

   Note If you want to connect an ECS instance to an ApsaraDB RDS instance by using an internal endpoint, you must make sure that the two instances are in the same region and have the same network type. Otherwise, the connection fails.

5. In the **Edit Whitelist** dialog box, enter the IP addresses or CIDR blocks that are allowed to access the instance and click **OK**.

> **Note**
>
> - Limits for **IPv4** addresses:
>   - You must separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.
>
>   Supported formats are `0.0.0.0/0`, IP addresses such as `10.23.12.24`, or CIDR blocks such as `10.23.12.24/24`. `/24` indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits.
>   - If an IP address whitelist is empty or contains `0.0.0.0/0`, all IP addresses can access the ApsaraDB RDS instance. This may cause security risks to the instance. Proceed with caution.
> - Limits for **IPv6** addresses:
>   - You must separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.
>
>   Supported formats are `::`, IP addresses such as `0:0:0:0:0:0:0:1`, or CIDR blocks such as `0:0:0:0:0:0:0:1/24`. `/24` indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 128 bits.
>   - If an IP address whitelist is empty or contains only `::`, all IP addresses can access the ApsaraDB RDS instance. This may cause security risks to the instance. Proceed with caution.
> - You cannot specify both **IPv4** and **IPv6** addresses in a single IP address whitelist. If you want to specify both IPv4 and IPv6 addresses, specify them in separate IP address whitelists.
> - If you click **Add Internal IP Addresses of ECS Instances**, the IP addresses of all the ECS instances that are created within your Apsara Stack tenant account appear. Then, you can select the IP addresses and add them to an IP address whitelist.

9.1.10.2. Configure SSL encryption

This topic describes how to enhance endpoint security. You can enable Secure Sockets Layer (SSL) encryption and install SSL certificates that are issued by certificate authorities (CAs) to the required application services. SSL is used at the transport layer to encrypt network connections and enhance the security and integrity of communication data. However, SSL increases the response time.

Prerequisites

Your ApsaraDB RDS instance runs one of the following MySQL versions and RDS editions:

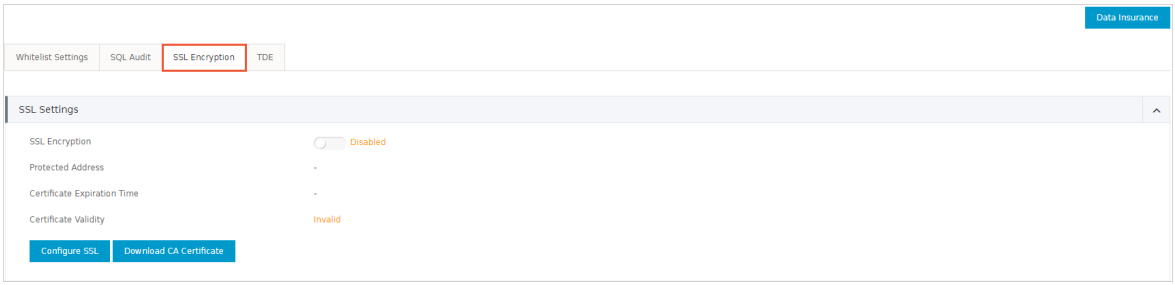
- MySQL 8.0 on RDS High-availability Edition (with local SSDs)
- MySQL 5.7 on RDS High-availability Edition (with local SSDs)
- MySQL 5.6 on RDS High-availability Edition (with local SSDs)

Precautions

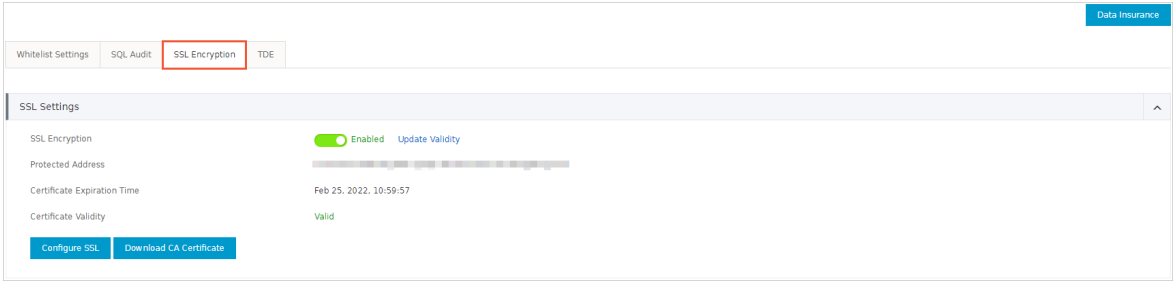
- An SSL CA certificate is valid for one year. You must update the validity period of the SSL CA certificate in your application or client within one year. Otherwise, your application or client that uses encrypted network connections cannot connect to the ApsaraDB RDS instance.
- SSL encryption may cause a significant increase in CPU utilization. We recommend that you enable SSL encryption only when you want to encrypt connections from the Internet. In most cases, connections that use an internal endpoint do not require SSL encryption.
- Read/write splitting endpoints do not support SSL encryption.
- If you disable SSL encryption, the ApsaraDB RDS instance restarts. Proceed with caution.

Enable SSL encryption

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **SSL Encryption** tab.



5. In the **SSL Settings** section, turn on **SSL Encryption**.
6. In the **Configure SSL** dialog box, select the endpoint for which you want to enable SSL encryption and click **OK**.
7. Click **Download CA Certificate** to download the SSL CA certificate files in a compressed package.



The downloaded package contains the following files:

- P7B file: contains the server CA certificate that can be imported into a Windows operating system.
- PEM file: contains the server CA certificate that can be imported into an operating system other than Windows or an application that is not Windows-based.
- JKS file: contains the server CA certificate that is stored in a Java-supported truststore. You can use the file to import the CA certificate chain into a Java-based application. The default password is apsaradb.

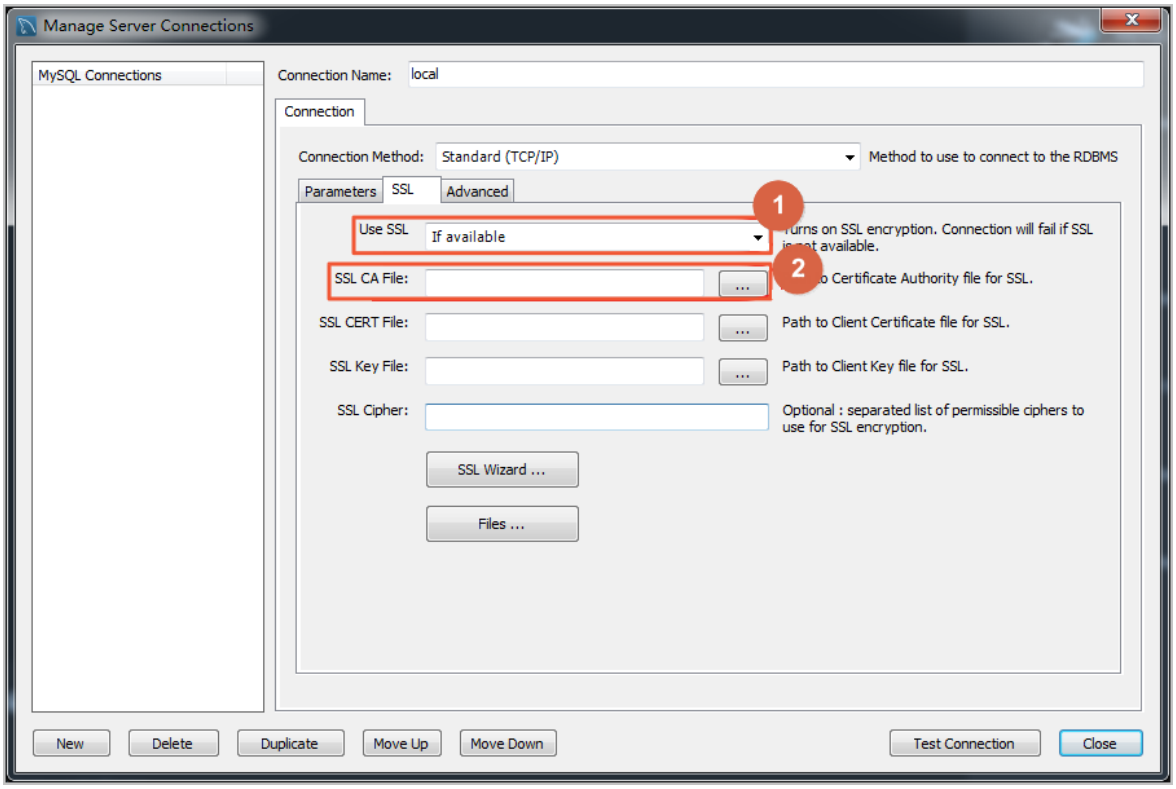
> **Note** When the JKS file is used in Java, you must modify the default JDK security configuration in JDK 7 and JDK 8. Open the `/jre/lib/security/java.security` file on the host where your application resides and modify the following configurations:
>
> ```
> jdk.tls.disabledAlgorithms=SSLv3, RC4, DH keySize < 224
> jdk.certpath.disabledAlgorithms=MD2, RSA keySize < 1024
> ```
>
> If you do not modify the JDK security configuration, the following error is reported. Similar errors are also caused by the Java security configuration.
>
> ```
> javax.net.ssl.SSLHandshakeException: DHPublicKey does not comply to algorithm constraints
> ```

Configure an SSL CA certificate

After you enable SSL encryption, configure the SSL CA certificate on your application or client before they can connect to the ApsaraDB RDS instance. This section describes how to configure an SSL CA certificate. MySQL Workbench and Navicat are used in the example. If you are using other applications or clients, see the related instructions.

Configure a certificate on MySQL Workbench

1. Start MySQL Workbench.
2. Choose **Database > Manage Connections**.
3. In the **Connection** section, click the **SSL** tab and configure the following parameters.



①: Enable **Use SSL**.

②: Import the SSL CA certificate file.

Configure a certificate on Navicat

1. Start Navicat.
2. Right-click the database and select **Edit Connection**.
3. Click the **SSL** tab. Select the path of the PEM-formatted CA certificate, as shown in the following figure.
4. Click **OK**.

> **Note** If the `connection is being used` error is reported, the previous session is still connected. Restart Navicat.

5. Double-click the database to test whether the database is connected.

Update the validity period of an SSL CA certificate

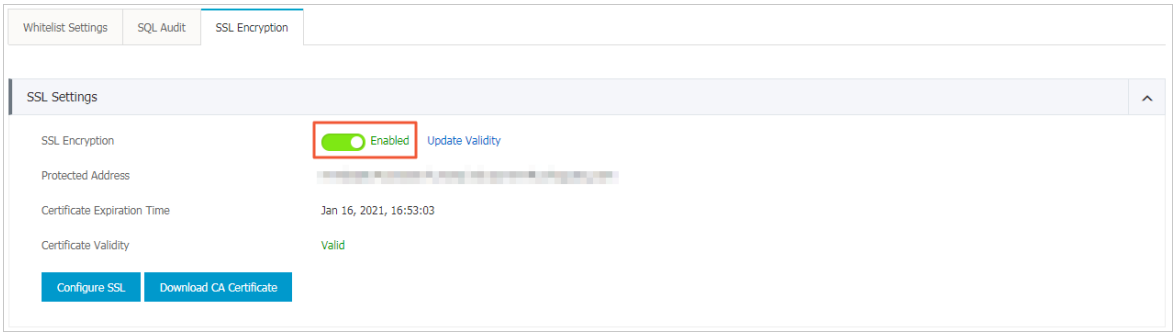
> **Note** **Update Validity** causes the ApsaraDB RDS instance to restart. Proceed with caution.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **SSL Encryption** tab.
5. Click **Update Validity**.

Disable SSL encryption

> **Note**
>
> - If you disable SSL encryption, the ApsaraDB RDS instance restarts. To reduce the impact on your business, the system triggers a primary/secondary switchover. We recommend that you disable SSL encryption during off-peak hours.
> - After you disable SSL encryption, access performance increases, but security decreases. We recommend that you disable SSL encryption only in secure environments.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **SSL Encryption** tab.
5. In the **SSL Settings** section, turn off **SSL Encryption**. In the message that appears, click **OK**.



9.1.10.3. Configure TDE

This topic describes how to configure Transparent Data Encryption (TDE) for an ApsaraDB RDS instance. TDE encrypts and decrypts data files in real time. It encrypts data files when they are written to disks, and decrypts data files when they are loaded to the memory from disks. TDE does not increase the size of data files. You can use TDE without the need to make changes to applications.

Prerequisites

- Your ApsaraDB RDS instance runs the RDS High-availability Edition with local SSDs.
- Key Management Service (KMS) is activated. If KMS is not activated, you can activate it when you enable TDE.

Context

The key used for TDE is created and managed by KMS. ApsaraDB RDS does not provide the key or certificates that are required for encryption. For specific zones, you can use the keys that are automatically generated by Apsara Stack, or you can use your own key materials to generate data keys and authorize your ApsaraDB RDS instance to use these keys.

Precautions

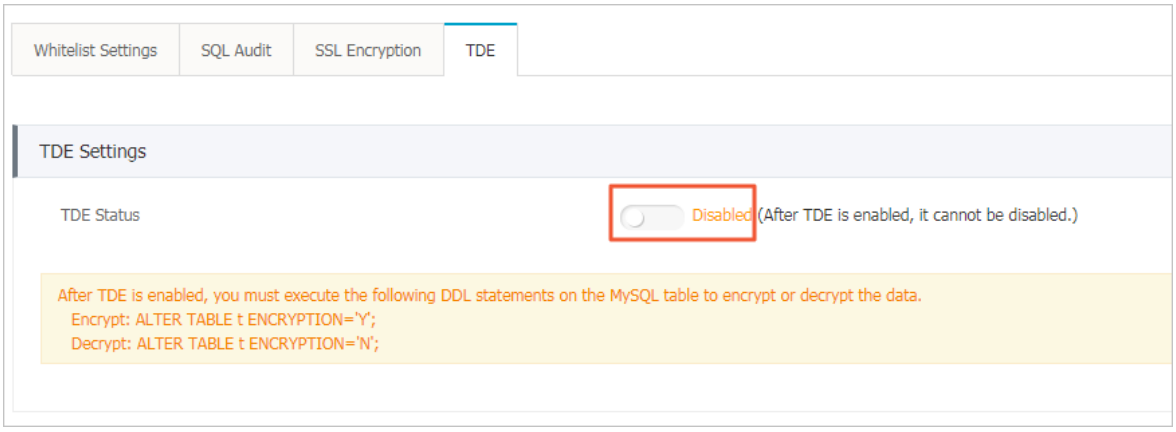
- When TDE is being enabled, your ApsaraDB RDS instance is restarted and all services are disconnected. Make appropriate service arrangements before you enable TDE. Proceed with caution.
- You cannot disable TDE after it is enabled.
- You cannot change the key used for encryption after TDE is enabled.
- If you want to restore the data to your computer after TDE is enabled, you must decrypt data on your ApsaraDB RDS instance. For more information, see the "Decrypt a table" section of this topic.
- After TDE is enabled, CPU utilization significantly increases.

- If you use an existing custom key for encryption, take note of the following items:
  - If you disable the key, configure a plan to delete the key, or delete the key material, the key becomes unavailable.
  - If you revoke the key that is authorized for an ApsaraDB RDS instance, the instance becomes unavailable after it is restarted.
  - You must use an Apsara Stack tenant account or an account that has the AliyunSTSAssumeRoleAccess permission.

> **Note** For more information, see topics about key management in *Key Management Service User Guide*.

Use a key that is automatically generated by Apsara Stack

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **TDE** tab.
5. In the **TDE Settings** section, turn on **TDE Status**.

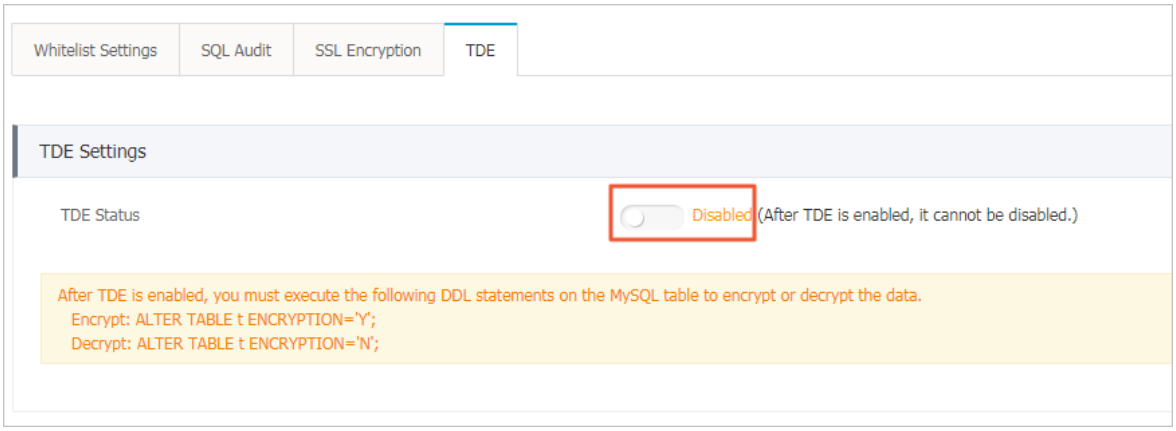


6. In the dialog box that appears, select **Use an Automatically Generated Key** and click **OK**.

> **Note** If the instance runs MySQL 5.7 on RDS High-availability Edition, you can select one of the following encryption methods:
>
> - SM4 encryption
> - AES_256_CBC encryption

Use an existing custom key

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **TDE** tab.
5. In the **TDE Settings** section, turn on **TDE Status**.



6. In the dialog box that appears, select **Use an Existing Custom Key** and click **OK**.

> **Note** If you do not have a custom key, click **create a key** to go to the KMS console and import the key materials. For more information, see Create a key in *Key Management Service User Guide*.

Encrypt a table

Log on to the database and execute one of the following statements to encrypt a table:

- MySQL 5.6

```
alter table <tablename> engine=innodb,block_format=encrypted;
```

- MySQL 5.7 or MySQL 8.0

```
alter table <tablename> encryption='Y';
```

Decrypt a table

Execute one of the following statements to decrypt a table that is encrypted by using TDE:

- MySQL 5.6

```
alter table <tablename> engine=innodb,block_format=default;
```

- MySQL 5.7 or MySQL 8.0

```
alter table <tablename> encryption='N';
```

FAQ

- Q: After I enable TDE, can I still use common database tools such as Navicat?

  A: Yes, after you enable TDE, you can still use common database tools such as Navicat.

- Q: After I enable TDE, why is my data still in plaintext?

  A: After you enable TDE, your data is stored in ciphertext. However, when the data is queried, it is decrypted and loaded into memory as plaintext. A: TDE encrypts backup files to prevent data leaks. Before you restore the data of your ApsaraDB RDS instance from an encrypted backup file to your computer, you must decrypt the file. For more information, see the "Decrypt a table" section of this topic.

9.1.10.4. Configure SQL audit

You can use the SQL audit feature to audit SQL executions and view their details. The SQL audit feature does not affect instance performance.

Context

> **Note** You cannot view the logs that are generated before you enable SQL audit.

You can view the incremental data of your ApsaraDB RDS for MySQL instance in SQL audit logs or binlogs. However, these two methods differ in the following aspects:

- SQL audit logs are similar to audit logs in MySQL and record all DML and DDL operations by using network protocol analysis. SQL audit does not parse the actual parameter values. Therefore, a small amount of information may be lost if a large number of SQL statements are executed to query data. The incremental data obtained by using this method may be inaccurate.
- Binlogs record all add, delete, and modify operations and the incremental data used for data restoration. Binlogs are temporarily stored in your ApsaraDB RDS instance after they are generated. The system transfers full binlog files to Object Storage Service (OSS) on a regular basis. OSS then stores the files for seven days. However, a binlog file cannot be transferred if data is being written to it. Such binlog files cannot be uploaded to OSS after you click **Upload Binlogs** on the **Backup and Restoration** page. Binlogs are not generated in real time, but you can obtain accurate incremental data from them.

Precautions

- By default, SQL audit is enabled.
- SQL audit logs are retained for 30 days.
- Log files exported from SQL audit are retained for two days. The system clears files that are retained for longer than two days.

Disable SQL audit

> **Note** If SQL audit is disabled, all SQL audit logs are deleted. We recommend that you export and store audit logs to your computer before you disable SQL audit.

You can disable SQL audit to avoid charges when you do not need it. To disable SQL audit, perform the following operations:

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **SQL Audit** tab.
5. Click **Export File** to export and store the SQL audit content to your computer.
6. After the file is exported, click **Disable SQL Audit**.
7. In the message that appears, click **Confirm**.

Enable SQL audit

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Data Security**.
4. Click the **SQL Audit** tab.
5. Click **Enable SQL Audit**.
6. In the message that appears, click **Confirm**.

   After SQL audit is enabled, you can query SQL information based on conditions such as the time range, database, user, and keyword.

9.1.11. Service availability

9.1.11.1. Switch workloads over between primary and secondary ApsaraDB RDS instances

This topic describes how to switch workloads over between a primary ApsaraDB RDS instance and its secondary instance. ApsaraDB RDS supports both manual switchover and automatic switchover. After a switchover is complete, the primary ApsaraDB RDS instance becomes the secondary instance.

Context

- Automatic switchover: By default, the automatic switchover feature is enabled. If the primary ApsaraDB RDS instance becomes faulty, ApsaraDB RDS automatically switches workloads over to the secondary instance.
- Manual switchover: You can manually switch workloads over between the primary and secondary ApsaraDB RDS instances even when the automatic switchover feature is enabled.

> **Note** Data is synchronized in real time between the primary and secondary ApsaraDB RDS instances. You can access only the primary instance. The secondary instance serves only as a standby and does not allow external access.

Precautions

- You may encounter a network interruption during a switchover. Make sure that your application is configured to automatically reconnect to the instance.
- If the primary ApsaraDB RDS instance is attached with read-only instances, the read-only instances need to re-establish the connections that are used for data replication and synchronize incremental data after a switchover. As a result, data on the read-only instances shows latencies of a few minutes.

Perform a manual switchover

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Service Availability**.
4. Click **Switch Primary/Secondary Instance** on the right side of the page.

   > **Note** You may encounter a network interruption during a switchover. Make sure that your application is configured to automatically reconnect to the instance.

5. In the dialog box that appears, click **OK**.

   > **Note** In the dialog box, you can also select **Switch Within Maintenance Window** and click **OK**. Then, the system performs the primary/secondary switchover within the maintenance window. For more information about how to set the maintenance window, see Set a maintenance window. You can also click **Change** on the right to change the maintenance window.

FAQ

Q: Can I connect to secondary instances?

A: No, you cannot connect to secondary instances. You can connect only to primary instances. Secondary instances serve only as a standby and do not allow external access.

9.1.11.2. Change the data replication mode

You can set the data replication mode between primary and secondary ApsaraDB RDS instances to improve database availability.

Prerequisites

Your ApsaraDB RDS instance runs the RDS High-availability Edition.

Data replication modes

- Semi-synchronous

  After an update that is initialized by your application is complete on the primary instance, the log is synchronized to all the secondary instances. After the secondary instances receive the log, the update transaction is considered committed. Your database system does not need to wait for the log to be replayed.

  If the secondary instances are unavailable or a network exception occurs between the primary and secondary instances, semi-synchronous replication degrades to the asynchronous mode.

- Asynchronous

  When your application initiates a request to add, delete, or modify data, the primary instance responds to your application immediately after it completes the operation. At the same time, the primary instance starts to asynchronously replicate data to its secondary instances. During asynchronous data replication, the unavailability of secondary instances does not affect the operations on the primary instance. Data remains consistent even if the primary instance is unavailable.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Service Availability**.
4. Click **Change Data Replication Mode** on the right side of the page.
5. In the dialog box that appears, select a data replication mode and click **OK**.

FAQ

Q: Which data replication mode is recommended?

A: You can select a data replication mode based on your business requirements. If you require quick responses, we recommend that you select the asynchronous mode. In other scenarios, you can select the semi-synchronous mode.

9.1.12. Database backup and restoration

9.1.12.1. Configure automatic backup

Automatic backup of ApsaraDB RDS supports full physical backups. ApsaraDB RDS automatically backs up data based on pre-configured policies. This topic describes how to configure a policy for automatic backup.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Backup and Restoration**.
4. Click the **Backup Settings** tab.
5. Click **Edit**. In the dialog box that appears, configure the following parameters.

> **Note** To ensure data security, the system compares the new backup cycle and time with the original settings and selects the most recent point in time to back up the data. Therefore, the next backup may still be performed based on the original backup cycle and time. For example, assume that the scheduled backup cycle and time is set to 19:00-20:00 every Wednesday. If you modify the backup cycle and time to 19:00-20:00 every Thursday before the scheduled backup occurs, the system still backs up data at the original scheduled time of 19:00-20:00 of Wednesday.

Parameter	Description
Data Retention Period	The number of days for which data backup files are retained. Valid values: 7 to 730. Default value: 7.
Backup Cycle	The backup cycle. You can select one or more days within a week.
Backup Time	A period of time within a day. Unit: hours. We recommend that you back up data during off-peak hours.
Log Backup	Specifies whether to enable log backup. **Notice** If you disable log backup, all the log backup files are deleted, and you cannot restore data to a specific point in time.
Log Retention Period	The number of days for which log backup files are retained. Valid values: 7 to 730. Default value: 7.
OSS Dump Status	Specifies whether to enable Object Storage Service (OSS) dump. When OSS dump is enabled, new backup files are automatically dumped to a specific OSS bucket. For more information, see Create a bucket in *Object Storage Service User Guide*.
OSS Dumped Data	The type of backup files that are dumped to an OSS bucket. Valid values: ○ **Data Backup** ○ **Log Backup**
OSS Bucket	The OSS bucket to which backup files are dumped. The OSS bucket must belong to the same Apsara Stack tenant account as your ApsaraDB RDS instance. You must make sure that an OSS bucket is created. For more information, see Create a bucket in *Object Storage Service User Guide*.
Restore Individual Database/Table	Specifies whether to enable restoration for individual databases or tables. By default, this features is enabled and cannot be disabled. **Note** This feature is available only for ApsaraDB RDS instances that run MySQL 5.6 and 5.7 on High-availability Edition.

6. Click **OK**.

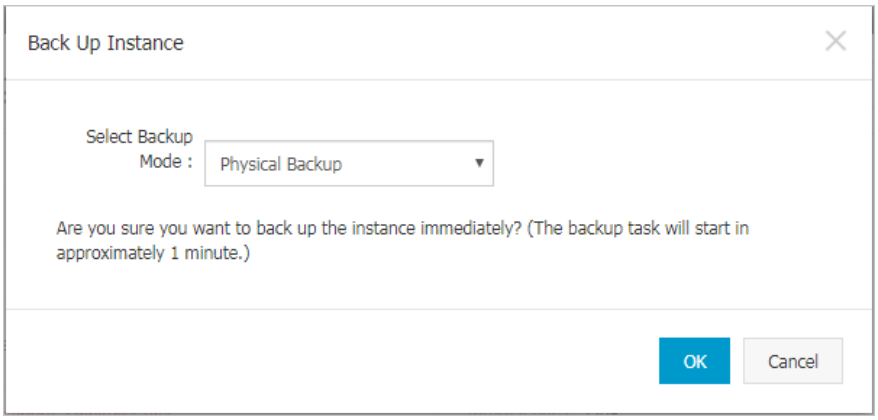
9.1.12.2. Manually back up an instance

Manual backup of ApsaraDB RDS supports both full physical and logical backups. This topic describes how to manually back up an ApsaraDB RDS instance.

Procedure

1. Log on to the ApsaraDB for RDS console.

2. On the **Instances** page, find the target instance.
3. Click **Back Up Instance** in the upper-right corner.



4. In the dialog box that appears, set the backup mode and backup policy and then click **OK**.

> **Note** The following backup modes are available:
>
> - Physical backup: directly backs up all files in all databases.
> - Logical backup: extracts data from the databases by using SQL statements and backs up the data in the text format. If you select logical backup, you must select one of the following backup policies:
>   - Instance Backup: backs up the entire instance.
>   - Single-Database Backup: backs up one of the databases on the instance.

9.1.12.3. Download data and log backup files

This topic describes how to download unencrypted data and log backup files in the ApsaraDB RDS console to archive the files and restore data to an on-premises database.

Limits

Database engine	Download of data backup files	Download of log backup files
MySQL 5.6 on RDS High-availability Edition	Supported	Supported
MySQL 5.7 on RDS High-availability Edition or Enterprise Edition	Supported	Supported
MySQL 8.0 on RDS High-availability Edition	Supported	Supported

Procedure

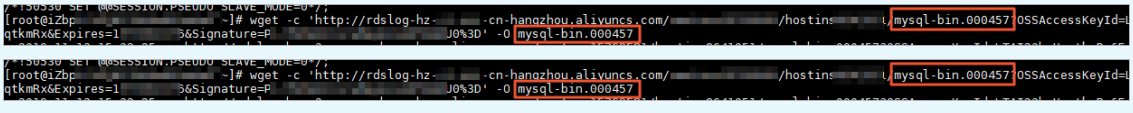
1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Backup and Restoration**.
4. Click the Data Backup or Log Backup tab.
   - To download data backup files, click the **Data Backup** tab.
   - To download log backup files, click the **Log Backup** tab.
5. Select a time range to which you want to restore the instance.
6. Find the data or log backup file that you want to download, and click **Download** in the **Actions** column.

> **Note**
> - If the Download button is unavailable, see the "Limits" section of this topic.
> - If you want to use a data backup file to restore data, select the backup file that is the closest to the time for restoration.
> - If you want to use a log backup file to restore data to an on-premises database, take note of the following items:
>   - The instance No. of the log backup file must be the same as that of the data backup file.
>   - The start time of the log backup file must be later than the end time of the selected data backup file and earlier than the point in time to which you want to restore the data of your instance.

7. In the message that appears, click **Download**.

Download method	Description
Download	Use a browser to download the backup file.
Copy Internal URL	Copy the internal URL to download the file. If your Elastic Compute Service (ECS) and ApsaraDB RDS instances reside within the same region, you can log on to the ECS instance and use the internal URL to download the file. This method is fast and secure.
Copy Public URL	Copy the public URL to download the file. If you want to use other tools to download the file, use the public URL.

> **Note** If you use a Linux operating system, you can run the following command to download the file:
>
> ```
> wget -c '<The URL used to download the backup file>' -O <The name of the backup file>
> ```
>
> - The -c option enables resumable download.
> - The -O option saves the downloaded file by using the specified name. We recommend that you use the file name contained in the download URL.
> - If the URL contains more than one parameter, enclose the download URL in a pair of single quotation marks (').



9.1.12.4. Upload binlogs

Context

Context

This topic describes how to upload binlog files to Object Storage Service (OSS).

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Backup and Restoration** to go to the **Backup and Restoration** page.
4. In the upper-right corner of the page, click **Upload Binlogs**.
5. In the message that appears, click **Confirm**.

9.1.12.5. Restore data to a new instance (formerly known as cloning an instance)

A cloned instance is a new instance that has the same content as the primary instance, including data and settings. This feature allows you to restore data of the primary instance or create multiple instances that are the same as the primary instance.

Prerequisites

Before you clone an instance, make sure that the following requirements are met:

- The primary instance is in the running state.
- The primary instance does not have ongoing migration tasks.
- Data backup and log backup are enabled.
- The primary instance has at least one completed backup set before you clone the instance by backup set.

Context

You can specify a backup set or a point in time within the backup retention period to clone an instance.

> Note
>
> - A cloned instance copies only the content of the primary instance, but not the content of read-only instances. The copied data includes database information, account information, and instance settings such as whitelist settings, backup settings, parameter settings, and alert threshold settings.
> - The database engine of a cloned instance must be the same as that of the primary instance. Other settings can be different, such as the instance edition, zone, network type, instance type, and storage capacity. If you want to restore the data of a primary instance, we recommend that you select a higher instance type and more storage capacity than those of the primary instance. This can speed up the data restoration process.
> - The account type of a cloned instance must be the same as that of the primary instance. The account password of the cloned instance can be changed.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Backup and Restoration**.
4. In the backup list, find a backup and click **Restore** in the Actions column.
5. In the dialog box that appears, select **Restore Database** and click **OK**.
6. On the **Restore Instance** page, configure the following parameters.

Section	Parameter	Description
Region	**Region**	The region in which the cloned instance resides.
Restore Database	**Restore Mode**	The data restore mode of the primary instance. Valid values: ○ **By Time** ○ **By Backup Set**
	Restore Time	The point in time to which you want to restore the primary instance. **Note** When **Restore Mode** is set to **By Time**, you must specify this parameter.
	Backup Set	The backup set for restoration. **Note** When **Restore Mode** is set to **By Backup Set**, you must specify this parameter.
Specifications	**Instance Name**	The name of the cloned instance.
	Database Engine	The database engine of the cloned instance, which cannot be modified.
	Engine Version	The engine version of the cloned instance, which cannot be modified.
	Edition	The RDS edition of the cloned instance. The actual values are displayed in the console.
	Storage Type	The storage type of the cloned instance. The actual values are displayed in the console.
	Instance Type	The instance type of the cloned instance. **Note** We recommend that you select a higher instance type and more storage capacity than those of the primary instance. This can speed up the data restoration process.
	Storage Capacity	The storage capacity of the cloned instance, which includes the space to store data, system files, binlog files, and transaction files. The available storage capacity is displayed in the console. **Note** AsparaDB RDS instances with local SSDs in the dedicated instance family occupy exclusive resources. The storage capacities are determined based on instance types.

Section	Parameter	Description
Network Type	Network Type	The network type of the cloned instance. ApsaraDB RDS instances support the following network types: ◦ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ◦ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways within a VPC. We recommend that you select VPC for improved security.
	VPC	The VPC in which the cloned instance resides. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch in the VPC. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.

7. Click **Submit**.

9.1.13. CloudDBA

9.1.13.1. Introduction to CloudDBA

CloudDBA is a cloud service for database self-detection, self-repair, self-optimization, self-maintenance, and self-security based on machine learning and expertise. CloudDBA helps you ensure stable, secure, and efficient databases without worrying about the management complexity and service failures caused by manual operations.

Features

In ApsaraDB RDS for MySQL, CloudDBA provides the following features:

- Diagnostics

  You can diagnose your instance and view the visualized diagnostic results.

- Autonomy center

  You can configure automatic detection for exceptions on core metrics. When an exception is detected, the system performs diagnostics on sessions, SQL statements, and the database capacity, provides optimization suggestions, and then performs automatic optimization if the related permissions have been granted.

- Instance sessions

  You can view sessions, check session statistics, analyze SQL statements, and optimize the execution of SQL statements.

- Real-time monitoring

  You can view the real-time information of your instance, such as the queries per second (QPS), transactions per second (TPS), number of connections, and network traffic.

- Storage analysis

You can view the storage overview, trends, exceptions, tablespaces, and data spaces.

- Deadlock analysis

  You can view and analyze the last deadlock in a database.

- Dashboard

  You can view and compare performance trends, customize monitoring dashboards, check exceptions, and view instance topologies.

- Slow query logs

  You can view the trends and statistics of slow queries.

- Diagnostic reports

  You can generate diagnostic reports or view automatically generated reports about instance health, alerts, and slow queries.

9.1.13.2. Diagnostics

In ApsaraDB RDS for MySQL, CloudDBA provides the diagnostics feature. This feature diagnoses your ApsaraDB RDS for MySQL instance and visualizes the results.

Navigate to the Diagnostics tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Diagnostics** tab.

> **Note** For more information, see Diagnostics in *Database Autonomy Service User Guide*.

9.1.13.3. Autonomy center

In ApsaraDB RDS for MySQL, CloudDBA provides the autonomy center feature. When an exception on core metrics is detected by CloudDBA, the system performs diagnostics on sessions, SQL statements, and the database capacity to identify the causes. CloudDBA also provides optimization and mitigation suggestions. If the related permissions have been granted, CloudDBA automatically performs the optimization and mitigation operations.

Navigate to the Autonomy Center tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Autonomy Center** tab.

> **Note** For more information, see Diagnostics in *Database Autonomy Service User Guide*.

9.1.13.4. Session management

In ApsaraDB RDS for MySQL, CloudDBA provides the session management feature. This feature allows you to view and manage the sessions of an instance.

Navigate to the Session Management page

1. Log on to the ApsaraDB for RDS console.

2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Session Management** tab.

> **Note** For more information, see Instance sessions in *Database Autonomy Service User Guide*.

9.1.13.5. Real-time monitoring

In ApsaraDB RDS for MySQL, CloudDBA provides the real-time monitoring feature. This feature allows you to view the real-time performance of your ApsaraDB RDS for MySQL instance.

Navigate to the Real-time Monitoring tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Real-time Monitoring** tab.

> **Note** For more information, see Real-time monitoring in *Database Autonomy Service User Guide*.

9.1.13.6. Storage analysis

In ApsaraDB RDS for MySQL, CloudDBA provides the storage analysis feature. This feature allows you to check and solve storage exceptions in a timely manner to ensure database stability.

Context

You can use the storage analysis feature of CloudDBA to view the disk space usage of your ApsaraDB RDS for MySQL instance and the number of remaining days when disk space is available. It also provides information about the space usage, fragmentation, and exception diagnostic results of a table.

Navigate to the Storage Analysis tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Storage Analysis** tab.

> **Note** For more information, see Storage analysis in *Database Autonomy Service User Guide*.

9.1.13.7. Deadlock analysis

In ApsaraDB RDS for MySQL, CloudDBA provides the deadlock analysis feature. This feature allows you to view and analyze the last deadlock in a database.

Navigate to the Deadlock Analysis page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Deadlock Analysis** tab.

> **Note** For more information, see Deadlock analysis in *Database Autonomy Service User Guide*.

9.1.13.8. Dashboard

In ApsaraDB RDS for MySQL, CloudDBA provides the dashboard feature. This feature allows you to view performance trends in specific ranges, compare performance trends, and customize charts to view performance trends.

Navigate to the Dashboard page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Dashboard**.

> **Note** For more information, see Dashboard in *Database Autonomy Service User Guide*.

9.1.13.9. Slow query logs

In ApsaraDB RDS for MySQL, CloudDBA provides the slow query logs feature. This feature allows you to view the trends and execution details of slow queries and obtain optimization suggestions for your ApsaraDB RDS for MySQL instance.

Navigate to the Slow Query Logs page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Slow Query Logs**.

> **Note** For more information, see Slow query logs in *Database Autonomy Service User Guide*.

9.1.13.10. Diagnostic reports

In ApsaraDB RDS for MySQL, CloudDBA provides the diagnostic reports feature. This feature allows you to create and view diagnostic reports.

Navigate to the Diagnostic Reports page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostic Reports**.

> **Note** For more information, see Diagnostic reports in *Database Autonomy Service User Guide*.

9.1.14. Manage logs

All ApsaraDB RDS instances support log management. You can query the details of error and slow query logs of an ApsaraDB RDS instance by using the ApsaraDB RDS console. The logs help you perform troubleshooting.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.

3. In the left-side navigation pane, click **Logs**.
4. On the **Logs** page, click the **Error Logs**, **Slow Query Logs**, **Slow Query Log Summary**, or **Primary/Secondary Switching Logs** tab, select a time range, and then click **Search**.

Log type	Description
Error Logs	Records database running errors that occurred within the latest month.
Slow Query Logs	Records SQL statements within the last month that took longer than one second to execute. Duplicated SQL statements are removed. **Note** Slow query logs in the ApsaraDB RDS console are updated once every minute. However, you can query real-time slow query logs from the mysql.slow_log table.
Slow Query Log Summary	Records and analyzes SQL statements within the last month that took longer than one second to execute. Analysis reports of slow query logs are provided.
Primary/Secondary Switching Logs	Records the primary/secondary instance switching logs. This feature is available for ApsaraDB RDS instances that run MySQL on RDS High-availability Edition.

9.1.15. Use mysqldump to migrate MySQL data

This topic describes how to use mysqldump to migrate data from an on-premises database to an ApsaraDB RDS for MySQL instance.

Prerequisites

An ECS instance is created.

Context

mysqldump is easy to use but requires extensive downtime. This tool is suitable for scenarios where the amount of data is small or extensive downtime is allowed.

ApsaraDB RDS for MySQL is fully compatible with the native database service. The procedure of migrating data from the original database to an ApsaraDB RDS for MySQL instance is similar to that of migrating data from one MySQL server to another.

Before you migrate data, you must create an account that is used to migrate data from the on-premises MySQL database. You must grant the read and write permissions on the on-premises MySQL databases to the account.

Procedure

1. Run the following command to create a migration account for the on-premises database:

   ```
   CREATE USER 'username'@'host' IDENTIFIED BY 'password';
   ```

   Parameter description:

   - username: the name of the account to be created.
   - host: the host from which the account is authorized to log on to the on-premises MySQL database. If you want to allow access from a local host, set this parameter to localhost. If you want to allow access from all hosts, set this parameter to a percent sign (%).
   - password: the password of the account.

   For example, you can run the following command to create an account with the username William and the password Changme123. The account is authorized to log on to the on-premises MySQL database from all hosts.

```
CREATE USER 'William'@'%' IDENTIFIED BY 'Changme123';
```

2. Run the following command to grant permissions to the migration account in the on-premises database:

```
GRANT SELECT ON databasename.tablename TO 'username'@'host' WITH GRANT OPTION; GRANT REPLICATION SLAVE ON databasename.tablename TO 'username'@'host' WITH GRANT OPTION;  GRANT REPLICATION SLAVE ON databasename.tablename TO 'username'@'host' WITH GRANT OPTION;
```

Parameter description:

- privileges: the operation permissions granted to the account, such as SELECT, INSERT, and UPDATE. To authorize the account to perform all operations, set this parameter to ALL.
- databasename: the name of the on-premises MySQL database. If you want to grant all database permissions to the account, set this parameter to an asterisk (*).
- tablename: the name of the table whose data you want to migrate. If you want to grant all table permissions to the account, set this parameter to an asterisk (*).
- username: the name of the account.
- host: the host from which the account is authorized to log on to the on-premises MySQL database. If you want to allow access from a local host, set this parameter to localhost. If you want to allow access from all hosts, set this parameter to a percent sign (%).
- WITH GRANT OPTION: authorizes the account to use the GRANT statement. This parameter is optional.

For example, you can execute the following statement to grant all permissions on tables and databases to the William account. The account is authorized to log on to the database from all hosts.

```
GRANT ALL ON *. * TO 'William'@'%';
```

3. Use the data export tool of mysqldump to export data from the database as a data file.

> **Notice** Do not update data during data export. In this step, only data is exported. Stored procedures, triggers, and functions are not exported.

```
mysqldump -h localIp -u userName -p --opt --default-character-set=utf8 --hex-blob dbName --skip-triggers > /tmp/dbName.sql
```

Parameter description:

- localIp: the IP address of the host where the on-premises MySQL database resides.
- userName: the account that is used to migrate data from the on-premises MySQL database.
- dbName: the name of the on-premises MySQL database.
- /tmp/dbName.sql: the name of the exported data file.

4. Use mysqldump to export stored procedures, triggers, and functions.

> **Notice** Skip this step if no stored procedures, triggers, or functions are used in the database. When stored procedures, triggers, and functions are exported, you must remove the DEFINER to ensure compatibility with ApsaraDB RDS for MySQL.

```
mysqldump -h localIp -u userName -p --opt --default-character-set=utf8 --hex-blob dbName -R | sed -e 's/DEFINER[ ]*=[ ]*[^*]*\*/\*/' > /tmp/triggerProcedure.sql
```

Parameter description:

- localIp: the IP address of the host where the on-premises MySQL database resides.
- userName: the account that is used to migrate data from the on-premises MySQL database.
- dbName: the name of the on-premises MySQL database.
- /tmp/triggerProcedure.sql: the name of the exported stored procedure file.

5. Upload the data file and stored procedure file to the ECS instance.

In this example, the files are uploaded to the following paths:

```
/tmp/dbName.sql
```

```
/tmp/triggerProcedure.sql
```

6. Log on to the ECS console and import both the data file and the stored procedure file to the destination ApsaraDB RDS for MySQL instance.

   > **Note** For information about how to log on to the ECS instance, see topics in the **Connect to an instance** section of ECS User Guide.

   ```
   mysql -h intranet4example.mysql.rds.aliyuncs.com -u userName -p dbName < /tmp/dbName.sql
   ```

   ```
   mysql -h intranet4example.mysql.rds.aliyuncs.com -u userName -p dbName < /tmp/triggerProcedure.sql
   ```

   Parameter description:

   - intranet4example.mysql.rds.aliyuncs.com: the endpoint of the ApsaraDB RDS for MySQL instance. An internal endpoint is used in this example.
   - userName: the migration account of the ApsaraDB RDS for MySQL database.
   - dbName: the name of the on-premises MySQL database from which you want to import data.
   - /tmp/dbName.sql: the name of the data file that you want to import.
   - /tmp/triggerProcedure.sql: the name of the stored procedure file that you want to import.

10.ApsaraDB RDS for SQL Server

10.1. User Guide(RDS SQL Server)

10.1.1. What is ApsaraDB RDS?

ApsaraDB RDS is a stable, reliable, and scalable online database service. Based on the distributed file system and high-performance storage, ApsaraDB RDS provides a set of solutions for disaster recovery, backup, restoration, monitoring, and migration.

ApsaraDB RDS supports four database engines, which are MySQL, SQL Server, PolarDB, and PostgreSQL. You can create database instances based on these engines to meet your business requirements. This topic describes the SQL Server engine.

ApsaraDB RDS for SQL Server

ApsaraDB RDS for SQL Server provides strong support for a variety of enterprise applications under the high-availability architecture. ApsaraDB RDS for SQL Server can also restore data to a specific point in time, which reduces costs.

ApsaraDB RDS for SQL Server provides basic features such as whitelist configuration, backup and restoration, transparent data encryption, data migration, and management for instances, accounts, and databases.

10.1.2. Log on to the ApsaraDB RDS console

This topic describes how to log on to the ApsaraDB RDS console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.

b. Enter the account and password again as in Step 2 and click **Log On**.

c. Enter a six-digit MFA verification code and click **Authenticate**.

- You have enabled MFA and bound an MFA device.

  Enter a six-digit MFA authentication code and click **Authenticate**.

> **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Database Services > ApsaraDB RDS**.

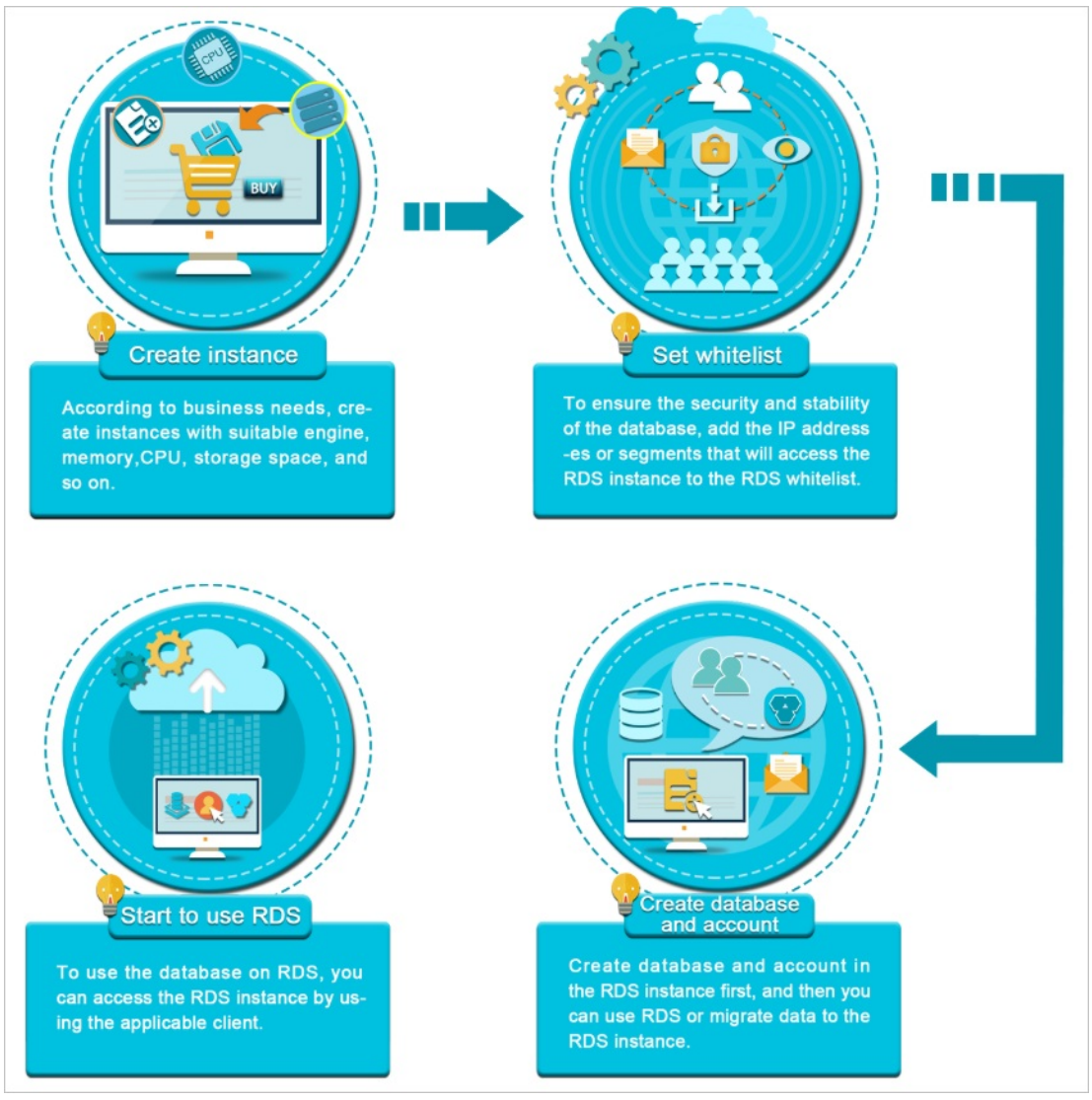
10.1.3. Quick Start

10.1.3.1. Procedure

ApsaraDB RDS quick start covers the following topics: creating an ApsaraDB RDS instance, configuring an IP address whitelist, creating a database, creating an account, and connecting to the instance.

The following figure shows the operations that you must perform before you use an ApsaraDB RDS instance.

Quick start flowchart



10.1.3.2. Create an instance

This topic describes how to create an instance in the ApsaraDB RDS console.

Prerequisites

An Apsara Stack tenant account is created.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.
3. Configure the following parameters.

Section	Parameter	Description
Basic Settings	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	Region	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.
	Deployment Method	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
	Quantity	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.

Section	Parameter	Description
Specifications	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **SQLServer**.
	Engine Version	The version of the database engine. Valid values: ○ **2012_ent_ha**: SQL Server 2012 EE ○ **2012_std_ha**: SQL Server 2012 SE ○ **2016_ent_ha**: SQL Server 2016 EE ○ **2016_std_ha**: SQL Server 2016 SE ○ **2017_ent_ha**: SQL Server 2017 EE
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Type	The storage type of the instance. The storage type is automatically set to **Standard SSD**.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, which includes the space to store data, system files, binlog files, and transaction files. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
Network	Network Type	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security.
	VPC	The VPC in which you want to create the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.

Section	Parameter	Description
	vSwitch	The vSwitch in the VPC. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click **Submit**.

10.1.3.3. Configure an IP address whitelist for an ApsaraDB RDS instance

After you create an ApsaraDB RDS instance, you must add the IP addresses or CIDR blocks that are used for database access to the IP address whitelist of the instance to ensure database security and reliability.

Context

IP address whitelists make your ApsaraDB RDS instance more secure and do not interrupt the operations of your ApsaraDB RDS instance when you configure whitelists. We recommend that you maintain your IP address whitelists on a regular basis.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. On the **Whitelist Settings** tab, click **Edit** corresponding to the **default** whitelist.

   Note If you want to connect an ECS instance to an ApsaraDB RDS instance by using an internal endpoint, you must make sure that the two instances are in the same region and have the same network type. Otherwise, the connection fails.

6. In the **Edit Whitelist** dialog box, enter the IP addresses or CIDR blocks that are allowed to access the instance and click **OK**.

> **Note**
>
> - Limits for IP address whitelists:
>   - You must separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.
>
>   Supported formats are `0.0.0.0/0`, IP addresses such as `10.23.12.24`, or CIDR blocks such as `10.23.12.24/24`. `/24` indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits.
>   - If an IP address whitelist is empty or contains only `0.0.0.0/0`, all IP addresses can access the ApsaraDB RDS instance. This may cause security risks to the instance. Proceed with caution.
> - If you click **Add Internal IP Addresses of ECS Instances**, the IP addresses of all the Elastic Compute Service (ECS) instances that are created within your Apsara Stack tenant account appear. Then, you can select the IP addresses and add them to an IP address whitelist.

10.1.3.4. Connect to an instance

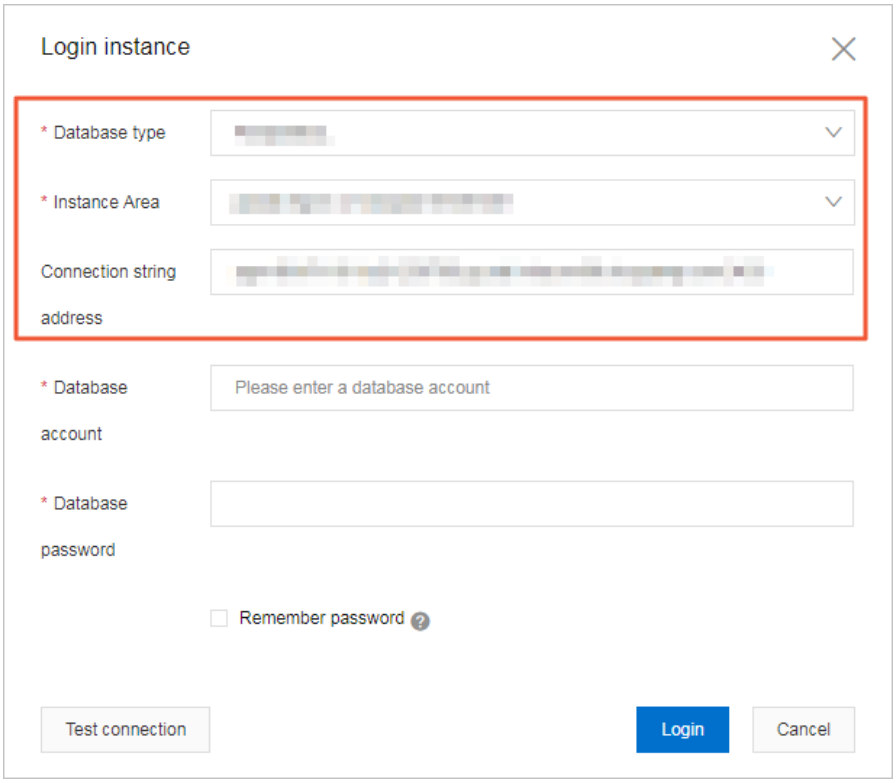
This topic describes how to use Data Management (DMS) to connect to an ApsaraDB RDS instance.

Prerequisites

- A database is created. For more information, see Create a database.
- A database account is created. For more information, see Create an account.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**.

> **Note**
> - If you want the browser to remember the password, select **Remember password** and click **Login**.
> - If you cannot connect to the instance, check the IP address whitelist settings. For more information, see Configure a whitelist.

10.1.3.5. Create an account

This topic describes how to create an account on an ApsaraDB RDS for SQL Server instance.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. On the right side of the page, click **Create Account**.
6. Enter the information of the account that you want to create.

Parameter	Description
Database Account	Enter the name of the account. The name must be 2 to 16 characters in length and can contain lowercase letters, digits, and underscores (_). It must start with a lowercase letter and end with a lowercase letter or a digit.
Account Type	○ **Privileged Account**: You can select the **Privileged Account** option only if you create an account on your ApsaraDB RDS instance for the first time. Each ApsaraDB RDS instance can have only a single privileged account. The privileged account of an ApsaraDB RDS instance cannot be deleted. ○ **Standard Account**: You can select the **Standard Account** option only after a privileged account is created on your ApsaraDB RDS instance. Each ApsaraDB RDS instance can have more than one standard account. You must manually grant the permissions on databases to each standard account.
Authorized Databases (available only for standard accounts)	Select the authorized databases of the account when the **Standard Account** type is selected. If no databases are created, you can leave this parameter empty. You can perform the following steps to grant permissions on more than one database to the account: i. In the **Unauthorized Databases** section, select the databases on which you want to grant permissions to the account. ii. Click **Add** to add the selected databases to the **Authorized Databases** section. iii. In the Authorized Databases section, specify the permissions that the account is granted on each authorized database. The permissions can be **Read/Write**, **Read-only**, or **Owner**. You can also click **Set All to Read/Write**, **Set All to Read-only**, or **Set All to Owner** to set the permissions of the account on all authorized databases. **Note** ■ The account is authorized to create tables, delete tables, and modify schemas in a database only when it has the **Owner** permission on the database. ■ The account has permissions on all databases and does not require authorization if you select the **Privileged Account** type.
Password	Enter the password of the account. The password must meet the following requirements: ○ The password is 8 to 32 characters in length. ○ The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include `! @ # $ % ^ & * ( ) _ + - =`
Re-enter Password	Enter the password of the account again.

Parameter	Description
Description	Enter a description that helps identify the account. The description can be up to 256 characters in length.

7. Click **Create**.

10.1.3.6. Create a database

This topic describes how to create a database on an ApsaraDB RDS for SQL Server instance in the ApsaraDB RDS console.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Databases**.
5. In the upper-right corner of the page, click **Create Database**.
6. Configure the parameters for the database that you want to create.

Parameter	Description
Database Name	Enter the name of the database. The name must be 2 to 64 characters in length. It can contain lowercase letters, digits, underscores (_), and hyphens (-). It must start with a lowercase letter and end with a lowercase letter or digit.
Supported Character Sets	Select the character set that is supported by the database. You can also select **all** and then select a character set from the drop-down list that appears.
Description	Enter a description of the database to facilitate subsequent management. The description can be up to 256 characters in length.

7. Click **Create**.

10.1.4. Instances

10.1.4.1. Create an instance

This topic describes how to create an instance in the ApsaraDB RDS console.

Prerequisites

An Apsara Stack tenant account is created.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.
3. Configure the following parameters.

Section	Parameter	Description
Basic Settings	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	Region	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.
	Deployment Method	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
Specifications	Quantity	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.
	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **SQLServer**.

Section	Parameter	Description
	Engine Version	The version of the database engine. Valid values: ○ **2012_ent_ha**: SQL Server 2012 EE ○ **2012_std_ha**: SQL Server 2012 SE ○ **2016_ent_ha**: SQL Server 2016 EE ○ **2016_std_ha**: SQL Server 2016 SE ○ **2017_ent_ha**: SQL Server 2017 EE
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Type	The storage type of the instance. The storage type is automatically set to **Standard SSD**.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, which includes the space to store data, system files, binlog files, and transaction files. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
Network	**Network Type**	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security.
	VPC	The VPC in which you want to create the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch in the VPC. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click **Submit**.

10.1.4.2. View basic information of an instance

This topic describes how to view the details of an ApsaraDB RDS instance, such as its basic information, internal network connection information, status, and configurations.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. Use one of the following methods to go to the **Basic Information** page of an instance:
   - On the **Instances** page, click the ID of an instance to go to the **Basic Information** page.
   - On the **Instances** page, click **Manage** in the **Actions** column corresponding to an instance to go to the **Basic Information** page.

10.1.4.3. Restart an instance

This topic describes how to manually restart an ApsaraDB RDS instance. This applies if the number of connections exceeds the specified threshold or if an instance has performance issues.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click **Restart Instance** in the upper-right corner.

   > **Note** When you restart an instance, applications are disconnected from the instance. We recommend that you make appropriate service arrangements before you restart an instance. Proceed with caution.

4. In the Restart Instance message, click **Confirm**.

10.1.4.4. Change the specifications of an instance

This topic describes how to change specifications such as the instance type and storage space if they do not meet the requirements of your application. When the specification changes take effect, a 30-second network interruption may occur. Business operations that involve databases, accounts, and networks are interrupted. We recommend that you change the specifications during off-peak hours or make sure that your applications are configured with automatic reconnection policies.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the **Configuration Information** section, click **Change Specifications**.
5. On the **Change Specifications** page, specify **Instance Type** and **Storage Capacity**.
6. After you configure the preceding parameters, click **Submit**.

10.1.4.5. Set a maintenance window

This topic describes how to set the maintenance window of an ApsaraDB RDS for SQL Server instance. The backend system performs maintenance on the ApsaraDB RDS instance during the maintenance window. This ensures the stability of the ApsaraDB RDS instance. The default maintenance window is from 02:00 (UTC+8) to 06:00 (UTC+8). We recommend that you set the maintenance window to off-peak hours of your business to avoid impacts on your business.

Context

- An instance enters the **Maintaining Instance** state before the maintenance window to ensure stability during the maintenance process. When the instance is in this state, access to data in the database and query operations such as performance monitoring are not affected. However, except for account and database management and IP address whitelist configuration, modification operations such as upgrade, downgrade, and restart are temporarily unavailable.
- During the maintenance window, one or two network interruptions may occur. Make sure that your applications are configured with automatic reconnection policies.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the **Configuration Information** section, click **Configure** to the right of **Maintenance Window**.
5. Select a maintenance window and click **Save**.

> **Note** The maintenance window is displayed in UTC+8.

10.1.4.6. Configure primary/secondary switchover

ApsaraDB RDS provides the primary/secondary switchover feature to ensure the high availability of databases. The primary/secondary switchover is performed when the primary instance becomes unavailable. You can also manually switch your business to the secondary instance.

Prerequisites

The instance is in the **Running** state.

Context

An ApsaraDB RDS for SQL Server instance has a secondary instance. Data is synchronized in real time between the primary and secondary instances. You can access only the primary instance. The secondary instance serves only as a backup and does not allow external access. If the primary instance cannot be accessed, your workloads are automatically switched over to the secondary instance. After the switchover, the primary instance becomes the secondary instance.

> **Notice**
> - You may encounter a network interruption during a switchover. Make sure that your application is configured to automatically reconnect to the instance.
> - During a switchover, a 1-minute data quality protection mechanism is enabled for data synchronization. If the primary and secondary database states are incorrect or if the latency for data synchronization exceeds 1 minute due to SQL Server errors, the HA system does not automatically perform the primary/secondary switchover. You must determine whether to perform the switchover.
> - If an instance is intermittently unavailable due to excessive mirroring event waits, the switchover is not performed. The instance automatically becomes available again.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic**

Information page.

4. In the left-side navigation pane, click **Service Availability**.
5. In the **Availability Information** section, click **Switch Primary/Secondary Instance**.
6. In the dialog box that appears, click **OK**.

Result

After the switchover is complete, the original primary instance becomes the secondary instance for the next primary/secondary switchover.

10.1.4.7. Release an instance

This topic describes how to manually release an instance.

Context

- Only instances in the running state can be manually released.
- After an instance is released, the instance data is immediately deleted. We recommend that you back up your data before you release an instance.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. Find the instance that you want to release and choose **More > Release Instance**.
3. In the **Release Instance** message, click **Confirm**.

10.1.4.8. Read-only instances

10.1.4.8.1. Overview of read-only ApsaraDB RDS for SQL Server instances

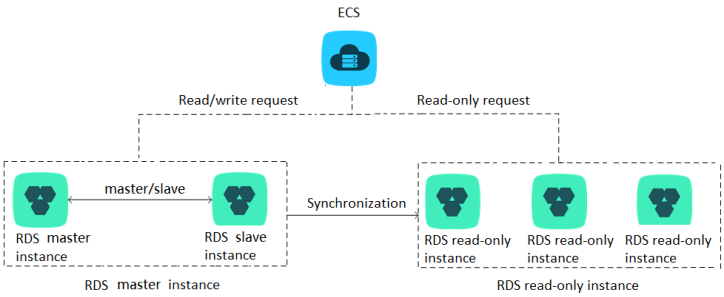
This topic provides an overview of read-only ApsaraDB RDS for SQL Server instances. If a large number of read requests overwhelm the primary instance, your business may be interrupted. In this case, you can create one or more read-only instances to offload read requests from the primary instance. This scales the read capability of your database system and increases the throughput of your application.

Overview

When a read-only instance is created, the data is replicated from the secondary instance. The data is consistent with that of the primary instance. Data updates of the primary instance are synchronized to all read-only instances.

> **Note**
> - Only ApsaraDB RDS instances that run SQL Server 2017 EE support read-only instances.
> - Each read-only instance works in a single-node architecture, where no instances are provided as backups.

The following figure shows the topology of read-only instances.



Features

- The specifications of a read-only instance can differ from the specifications of the primary instance, and can be changed at any time. We recommend that you select specifications of a read-only instance that are higher than or equal to those of the primary instance. If the specifications of a read-only instance are lower than those of the primary instance, the read-only instance may have high latency or workloads.
- Read-only instances do not require database or account maintenance, because their database and account information is synchronized with the primary instance.
- A read-only instance automatically replicates the IP address whitelists of the primary instance. However, the IP address whitelists for the read-only instance are independent of those of the primary instance. For information about how to modify the whitelists of a read-only instance, see Configure a whitelist.
- You can monitor up to 20 system performance metrics, such as the disk capacity, input/output operations per second (IOPS), number of connections, CPU utilization, and network traffic.

Limits

- You can create up to seven read-only instances.
- You cannot configure backup policies or manually create backups for read-only instances, because these are already configured or created on the primary instance.
- You cannot create a temporary instance by using a backup set or from a point in time. In addition, you cannot overwrite a read-only instance by using a backup set.
- After a read-only instance is created, you cannot use a data backup file to restore it in overwrite mode.
- You cannot migrate data to read-only RDS instances.
- You cannot create or delete databases on read-only instances.
- You cannot create or delete accounts, authorize accounts, or change the passwords of accounts on read-only instances.

FAQ

Can I manage the accounts created on the primary instance from its read-only instances?

No, although accounts created on the primary instance are replicated to its read-only instances, you cannot manage the accounts on the read-only instances. The accounts have only read permissions on the read-only instances.

10.1.4.8.2. Create a read-only ApsaraDB RDS for SQL Server instance

This topic describes how to create a read-only instance for your primary ApsaraDB RDS for SQL server instance. This allows your database system to process a large number of read requests and increases the throughput of your application. Each read-only ApsaraDB RDS instance is a replica of the primary instance. Data updates on the primary instance are synchronized to all the read-only instances.

Prerequisites

The primary instance runs SQL Server 2017 EE.

Precautions

- You can create read-only instances for the primary ApsaraDB RDS instance. However, you cannot convert existing ApsaraDB RDS instances into read-only instances.
- While you create a read-only instance, the system replicates data from a secondary instance. Therefore, the operation of your primary instance is not interrupted.
- You can create up to seven read-only instances.
- For more information about read-only ApsaraDB RDS instances, see Overview of read-only ApsaraDB RDS for SQL Server instances.

Create a read-only instance

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the **Distributed by Instance Role** section on the right side of the page, click **Create Read-only Instance**.
5. Configure the following parameters and click **Submit**.

Section	Parameter	Description
Region	**Region**	The region in which you want to create the instance.
Specifications	**Database Engine**	The database engine of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	Engine Version	The engine version of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	Edition	Set the value to **Read-only**.
	Instance Type	The instance type of the read-only instance. The instance type of the read-only instance can be different from that of the primary instance, and can be changed at any time to facilitate flexible upgrade and downgrade. For more information, see Instance types in *ApsaraDB RDS Product Information*. **Note** To ensure sufficient I/O throughput for data synchronization, we recommend that you select at least the same instance type as the primary instance for read-only instances.
	Storage Capacity	The storage space of the read-only instance. To ensure sufficient I/O throughput for data synchronization, we recommend that you select at least the same instance type and storage space as the primary instance for the read-only instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.

Section	Parameter	Description
Network Type	Network Type	The network type of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	VPC	Select a VPC if the network type is set to VPC.
	vSwitch	Select a vSwitch if the network type is set to VPC.

10.1.4.8.3. View details of read-only instances

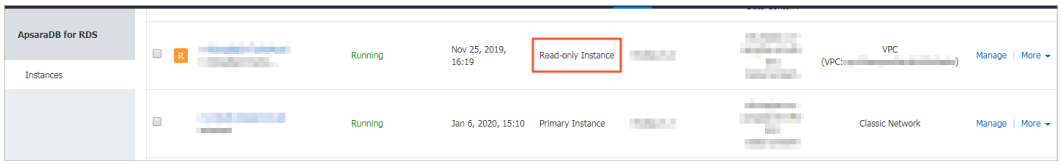
This topic describes how to view details of read-only instances. You can go to the Basic Information page of a read-only instance from the Instances page or the read-only instance list of the primary instance. Read-only instances are managed in the same way as primary instances. The read-only instance management page shows the management operations that can be performed.

View instance details by using a read-only instance

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click the ID of a read-only instance. The **Basic Information** page appears.

   In the instance list, Instance Role of read-only instances is displayed as Read-only Instance, as shown in View a read-only instance.

   View a read-only instance



View instance details by using the primary instance

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. On the **Basic Information** page, move the pointer over the number below **Read-only Instance** in the **Distributed by Instance Role** section. The ID of the read-only instance is displayed.
5. Click the ID of the read-only instance to go to the read-only instance management page.

10.1.5. Accounts

10.1.5.1. Create an account

This topic describes how to create an account on an ApsaraDB RDS for SQL Server instance.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.

3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. On the right side of the page, click **Create Account**.
6. Enter the information of the account that you want to create.

Parameter	Description
Database Account	Enter the name of the account. The name must be 2 to 16 characters in length and can contain lowercase letters, digits, and underscores (_). It must start with a lowercase letter and end with a lowercase letter or a digit.
Account Type	○ **Privileged Account**: You can select the **Privileged Account** option only if you create an account on your ApsaraDB RDS instance for the first time. Each ApsaraDB RDS instance can have only a single privileged account. The privileged account of an ApsaraDB RDS instance cannot be deleted. ○ **Standard Account**: You can select the **Standard Account** option only after a privileged account is created on your ApsaraDB RDS instance. Each ApsaraDB RDS instance can have more than one standard account. You must manually grant the permissions on databases to each standard account.
Authorized Databases (available only for standard accounts)	Select the authorized databases of the account when the **Standard Account** type is selected. If no databases are created, you can leave this parameter empty. You can perform the following steps to grant permissions on more than one database to the account: i. In the **Unauthorized Databases** section, select the databases on which you want to grant permissions to the account. ii. Click **Add** to add the selected databases to the **Authorized Databases** section. iii. In the Authorized Databases section, specify the permissions that the account is granted on each authorized database. The permissions can be **Read/Write**, **Read-only**, or **Owner**. You can also click **Set All to Read/Write**, **Set All to Read-only**, or **Set All to Owner** to set the permissions of the account on all authorized databases. **Note** ■ The account is authorized to create tables, delete tables, and modify schemas in a database only when it has the **Owner** permission on the database. ■ The account has permissions on all databases and does not require authorization if you select the **Privileged Account** type.
Password	Enter the password of the account. The password must meet the following requirements: ○ The password is 8 to 32 characters in length. ○ The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include `! @ # $ % ^ & * ( ) _ + - =`
Re-enter Password	Enter the password of the account again.
Description	Enter a description that helps identify the account. The description can be up to 256 characters in length.

7. Click **Create**.

10.1.5.2. Reset the password

You can use the ApsaraDB RDS console to reset the password of your database account.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts**.
4. Find an account and click **Reset Password** in the **Actions** column.
5. In the dialog box that appears, enter and confirm the new password, and then click **OK**.

> **Note** The password must meet the following requirements:
>
> - The password is 8 to 32 characters in length.
> - The password contains at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.
> - Special characters include
>
>   ! @ # $ % ^ & * ( ) _ + - =

10.1.6. Databases

10.1.6.1. Create a database

This topic describes how to create a database on an ApsaraDB RDS for SQL Server instance.

Terms

- Instance: a virtualized database server on which you can create and manage more than one database.
- Database: a set of data that is stored in an organized manner and can be shared by a number of users. A database provides the minimal redundancy and is independent of applications. In simple words, a database is a data warehouse that is used to store data.
- Character set: a collection of letters, special characters, and encoding rules that are used in a database.

Prerequisites

An ApsaraDB RDS for SQL Server instance is created. For more information, see Create an instance.

Procedure

For more information, see Create a database.

10.1.6.2. Delete a database

This topic describes how to delete a database from an ApsaraDB RDS for SQL Server instance. You can delete a database by using the ApsaraDB RDS console or an SQL statement.

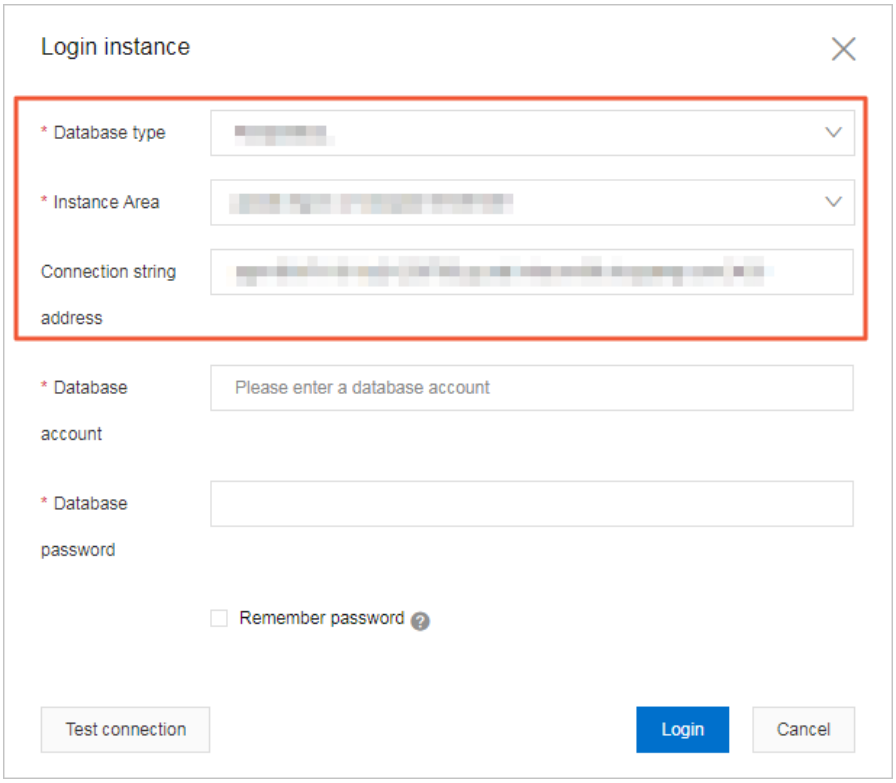
Use the console to delete a database

1. Log on to the ApsaraDB for RDS console.

2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Databases**.
5. Find the database that you want to delete and click **Delete** in the **Actions** column.
6. In the message that appears, click **Confirm**.

Execute an SQL statement to delete a database

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.

Parameter	Description
Database password	The password of the account used to connect to the database.

6. Click **Login**.

> **Note**
> - If you want the browser to remember the password, select **Remember password** and click **Login**.
> - If you cannot connect to the instance, check the IP address whitelist settings. For more information, see Configure a whitelist.

7. The **SQLConsole** page appears after you log on to the instance. Execute a statement in the following format to delete a database:

```
drop database <database name>;
```

> **Note** If the instance runs SQL Server 2012 or later on RDS High-availability Edition, you can also use the following stored procedure. This stored procedure deletes the specified database, removes the associated image, and closes the connection to the database.
>
> ```
> EXEC sp_rds_drop_database 'database name'
> ```

8. Click **execute**.

10.1.6.3. Change the character set collation and the time zone of system databases

This topic describes how to change the character set collation and the time zone of system databases. System databases include master, msdb, tempdb, and model.

Prerequisites

- The instance runs SQL Server 2012, 2016, or 2017.
- No database other than system databases exists on the instance.

> **Note** If you have just deleted databases from the instance, the deletion task may be pending in the secondary instance. Before you change the character set collation and the time zone, make sure that the primary and secondary instances do not contain databases.

Precautions

- The default character set collation is Chinese_PRC_CI_AS.
- The default time zone is China Standard Time.
- You can view the available character set collations and time zones in the console.
- The instance is in the unavailable state during the change process. It takes about 1 minute to change the time zone, and 2 to 10 minutes to change the character set collation.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic**

Information page.

4. In the left-side navigation pane, click **Databases**.
5. On the Databases page, click **Change Character Set Collation and Time Zone**.

> **Note** If you fail to find this button on the page, make sure that the requirements in Prerequisites are met.

6. In the dialog box that appears, select **Time Zone**, **Character Set Collation**, or both of them, and click **OK**.

UTC offsets of time zones

Time zone	UTC offset	Description
Afghanistan Standard Time	(UTC+04:30)	Kabul
Alaskan Standard Time	(UTC-09:00)	Alaska
Arabian Standard Time	(UTC+04:00)	Abu Dhabi, Muscat
Atlantic Standard Time	(UTC-04:00)	Atlantic Time (Canada)
AUS Central Standard Time	(UTC+09:30)	Darwin
AUS Eastern Standard Time	(UTC+10:00)	Canberra, Melbourne, Sydney
Belarus Standard Time	(UTC+03:00)	Minsk
Canada Central Standard Time	(UTC-06:00)	Saskatchewan
Cape Verde Standard Time	(UTC-01:00)	Cabo Verde Is.
Cen. Australia Standard Time	(UTC+09:30)	Adelaide
Central America Standard Time	(UTC-06:00)	Central America
Central Asia Standard Time	(UTC+06:00)	Astana
Central Brazilian Standard Time	(UTC-04:00)	Cuiaba
Central Europe Standard Time	(UTC+01:00)	Belgrade, Bratislava, Budapest, Ljubljana, Prague
Central European Standard Time	(UTC+01:00)	Sarajevo, Skopje, Warsaw, Zagreb
Central Pacific Standard Time	(UTC+11:00)	Solomon Islands, New Caledonia
Central Standard Time	(UTC-06:00)	Central Time (US and Canada)
Central Standard Time (Mexico)	(UTC-06:00)	Guadalajara, Mexico City, Monterrey
China Standard Time	(UTC+08:00)	Beijing, Chongqing, Hong Kong, Urumqi
E. Africa Standard Time	(UTC+03:00)	Nairobi
E. Australia Standard Time	(UTC+10:00)	Brisbane
E. Europe Standard Time	(UTC+02:00)	Chisinau

Time zone	UTC offset	Description
E. South America Standard Time	(UTC-03:00)	Brasilia
Eastern Standard Time	(UTC-05:00)	Eastern Time (US and Canada)
Georgian Standard Time	(UTC+04:00)	Tbilisi
GMT Standard Time	(UTC)	Dublin, Edinburgh, Lisbon, London
Greenland Standard Time	(UTC-03:00)	Greenland
Greenwich Standard Time	(UTC)	Monrovia, Reykjavik
GTB Standard Time	(UTC+02:00)	Athens, Bucharest
Hawaiian Standard Time	(UTC-10:00)	Hawaii
India Standard Time	(UTC+05:30)	Chennai, Kolkata, Mumbai, New Delhi
Jordan Standard Time	(UTC+02:00)	Amman
Korea Standard Time	(UTC+09:00)	Seoul
Middle East Standard Time	(UTC+02:00)	Beirut
Mountain Standard Time	(UTC-07:00)	Mountain Time (US and Canada)
Mountain Standard Time (Mexico)	(UTC-07:00)	Chihuahua, La Paz, Mazatlan
US Mountain Standard Time	(UTC-07:00)	Arizona
New Zealand Standard Time	(UTC+12:00)	Auckland, Wellington
Newfoundland Standard Time	(UTC-03:30)	Newfoundland
Pacific SA Standard Time	(UTC-03:00)	Santiago
Pacific Standard Time	(UTC-08:00)	Pacific Time (US and Canada)
Pacific Standard Time (Mexico)	(UTC-08:00)	Baja California
Russian Standard Time	(UTC+03:00)	Moscow, St. Petersburg, Volgograd
SA Pacific Standard Time	(UTC-05:00)	Bogota, Lima, Quito, Rio Branco
SE Asia Standard Time	(UTC+07:00)	Bangkok, Hanoi, Jakarta
China Standard Time	(UTC+08:00)	Kuala Lumpur, Singapore
Tokyo Standard Time	(UTC+09:00)	Osaka, Sapporo, Tokyo
US Eastern Standard Time	(UTC-05:00)	Indiana (East)
UTC	UTC	Coordinated Universal Time
UTC-02	(UTC-02:00)	Coordinated Universal Time-02
UTC-08	(UTC-08:00)	Coordinated Universal Time-08

Time zone	UTC offset	Description
UTC-09	(UTC-09:00)	Coordinated Universal Time-09
UTC-11	(UTC-11:00)	Coordinated Universal Time-11
UTC+12	(UTC+12:00)	Coordinated Universal Time+12
W. Australia Standard Time	(UTC+08:00)	Perth
W. Central Africa Standard Time	(UTC+01:00)	West Central Africa
W. Europe Standard Time	(UTC+01:00)	Amsterdam, Berlin, Bern, Rome, Stockholm, Vienna

10.1.7. Database connection

10.1.7.1. Change a vSwitch

This topic describes how to change a vSwitch for an ApsaraDB RDS instance that is deployed in a VPC.

Prerequisites

The instance is deployed in a VPC.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. In the upper-right corner of the Database Connection section, click **Switch vSwitch**.
5. In the dialog box that appears, select a vSwitch and click **OK**.
6. In the message that appears, click **Switch**.

> **Note**
>
> - You may encounter a network interruption of about 30 seconds during the change process. Make sure that your application is configured to automatically reconnect to the instance.
> - We recommend that you clear the cache immediately after the instance is switched to a new VPC and vSwitch. Otherwise, data can only be read but cannot be written.

10.1.7.2. Change the endpoint and port number of an instance

This topic describes how to view and change the endpoint and port number of an ApsaraDB RDS instance.

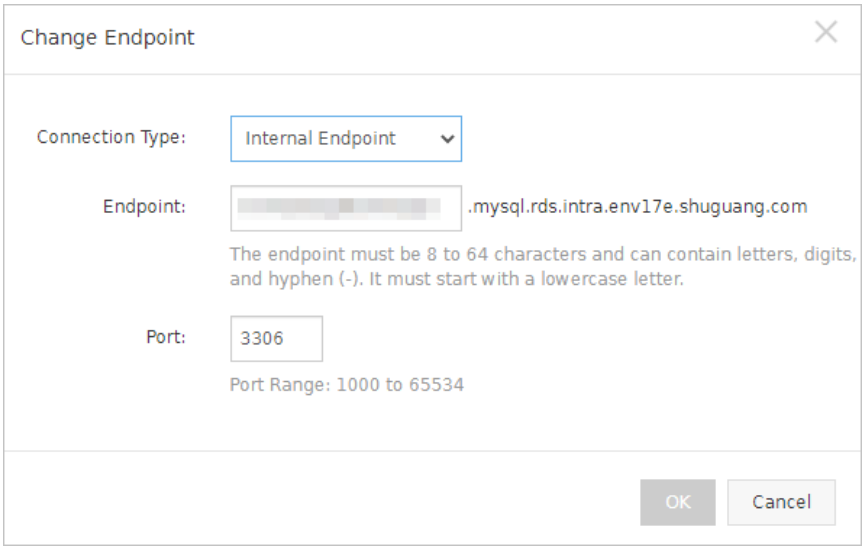
View the endpoint and port number

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the **Basic Information** section, view the internal and public endpoints and port numbers.

Change the endpoint and port number

1. Log on to the ApsaraDB for RDS console.

2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. In the upper-right corner of the Database Connection section, click **Change Endpoint**.
5. In the dialog box that appears, set Connection Type, Endpoint, and Port, and then click **OK**.



> **Note**
> - The prefix of an endpoint must be 8 to 64 characters in length and can contain letters, digits, and hyphens (-). It must start with a lowercase letter.
> - The port number must be a value within the range of 1000 to 65534.

10.1.7.3. Apply for and release an internal endpoint or a public endpoint for an instance

ApsaraDB RDS supports two types of endpoints: internal endpoints and public endpoints. The default type of the endpoint used to connect to an ApsaraDB RDS instance is determined by the network connection type selected when you create the instance. This topic describes how to apply for and release an internal endpoint or a public endpoint for an ApsaraDB RDS instance.

Apply for an internal endpoint or a public endpoint

If you set **Connection Type** to **Internet** when you create an ApsaraDB RDS instance, the database system assigns a public endpoint to the instance, and you can apply for an internal endpoint. Otherwise, the database system assigns an internal endpoint to the instance, and you can apply for a public endpoint.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. Apply for an internal endpoint or a public endpoint:
   - To apply for a public endpoint, click **Apply for Public Endpoint**.
   - To apply for an internal endpoint, click **Apply for Internal Endpoint**.
5. In the message that appears, click **OK**.

Release an internal endpoint or a public endpoint

If an endpoint is no longer needed, you can release the endpoint to ensure instance security.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Database Connection**.
4. Release an internal endpoint or a public endpoint:
   - To release a public endpoint, click **Release Public Endpoint**.
   - To release an internal endpoint, click **Release Internal Endpoint**.
5. In the message that appears, click **OK**.

FAQ

- Q: Can I change the endpoints and port numbers of my ApsaraDB RDS instance?

  A: No, you cannot change the endpoints of your ApsaraDB RDS instance. You can change the prefixes of the endpoints. You can also change the port numbers of your instance. For more information, see Change the endpoint and port number of an instance.

- Q: Can I configure the endpoints of my ApsaraDB RDS instances to static IP addresses?

  A: No, you cannot configure the endpoints of your ApsaraDB RDS instance to static IP addresses. Both primary/secondary switchovers and specification changes may cause changes to the IP addresses. Therefore, we recommend that you connect to your instance by using an endpoint. This allows you to minimize the impact on your workloads and eliminates the need to modify the configuration data on your application.

10.1.7.4. Connect to an instance

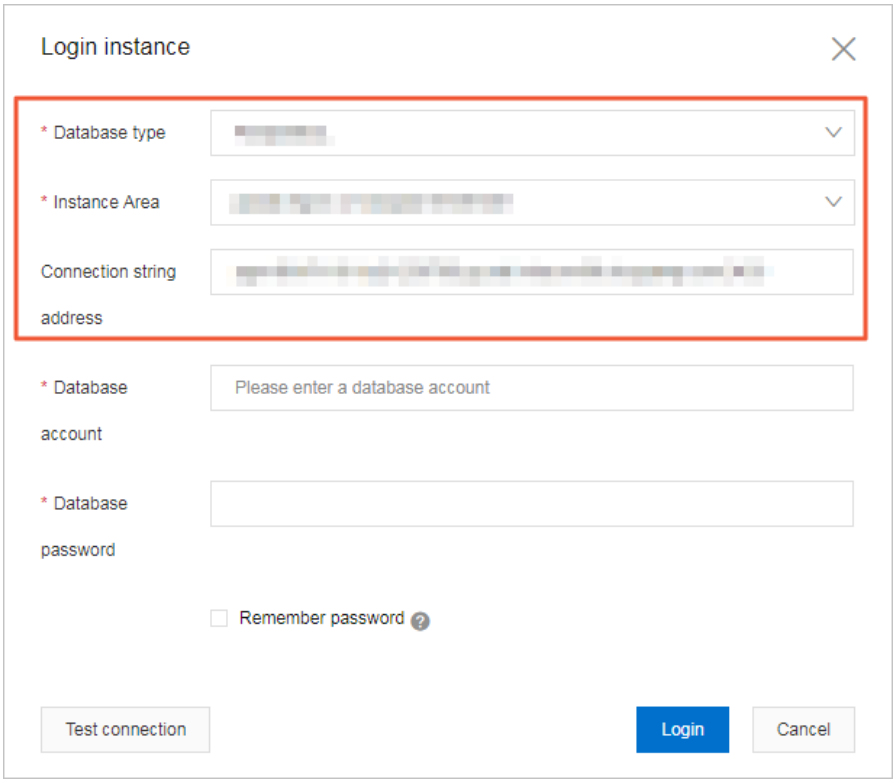
This topic describes how to use Data Management (DMS) to connect to an ApsaraDB RDS instance.

Prerequisites

- A database is created. For more information, see Create a database.
- A database account is created. For more information, see Create an account.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**.

> **Note**
> - If you want the browser to remember the password, select **Remember password** and click **Login**.
> - If you cannot connect to the instance, check the IP address whitelist settings. For more information, see Configure a whitelist.

10.1.8. Monitoring and alerting

10.1.8.1. Set a monitoring frequency

This topic describes how to set the monitoring frequency of an ApsaraDB RDS for SQL Server instance.

Context

ApsaraDB RDS provides the following monitoring frequencies:

- Every 60 seconds
- Every 300 seconds

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Monitoring and Alerts**.
5. On the **Resource Monitoring** tab, click **Set Monitoring Frequency**.
6. In the **Set Monitoring Frequency** dialog box, select the required monitoring frequency.
7. Click **OK**.

10.1.8.2. View resource and engine monitoring data

The ApsaraDB RDS console provides a variety of performance metrics to monitor the status of your instances.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Monitoring and Alerts**.
5. On the **Monitoring and Alerts** page, select **Resource Monitoring** or **Engine Monitoring**, and select a time range to view the corresponding monitoring data. The following table describes the metrics.

Monitoring type	Metric	Description
Resource Monitoring	Disk Space (unit: MB)	The disk usage of the instance, which includes the following items: ○ Instance Size ○ Data Usage ○ Log Size ○ Temporary File Size ○ Other System File Size
	IOPS	The number of input/output operations per second (IOPS) for the instance.
	Total Connections	The total number of current connections of the instance.
	MSSQL Instance CPU Utilization (percentage in the operating system)	The CPU utilization of the instance. This includes the CPU utilization for the operating system. Unit: %.
	SQLServer Average Input/Output Traffic	The inbound and outbound traffic of the instance per second. Unit: KB.

Monitoring type	Metric	Description
Engine Monitoring	Average Transaction Frequency	The number of transactions processed per second.
	Average QPS	The number of SQL statements executed per second.
	Buffer Hit Ratio (%)	The read hit ratio of the buffer pool.
	Page Write Frequency at Check Point	The number of checkpoints written to pages per second.
	Login Frequency	The number of logons to the instance per second.
	Average Frequency of Whole Table Scans	The number of full table scans per second.
	SQL Compilations per Second	The number of SQL statements compiled per second.
	Lock Timeout Times	The number of lock timeouts on the instance per second.
	Deadlock Frequency	The number of deadlocks on the instance per second.
	Lock Wait Frequency	The number of lock waits on the instance per second.

10.1.9. Data security

10.1.9.1. Configure an IP address whitelist for an ApsaraDB RDS instance

After you create an ApsaraDB RDS instance, you must add the IP addresses or CIDR blocks that are used for database access to the IP address whitelist of the instance to ensure database security and reliability.

Context

IP address whitelists make your ApsaraDB RDS instance more secure and do not interrupt the operations of your ApsaraDB RDS instance when you configure whitelists. We recommend that you maintain your IP address whitelists on a regular basis.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. On the **Whitelist Settings** tab, click **Edit** corresponding to the **default** whitelist.

   > **Note** If you want to connect an ECS instance to an ApsaraDB RDS instance by using an internal endpoint, you must make sure that the two instances are in the same region and have the same network type. Otherwise, the connection fails.

6. In the **Edit Whitelist** dialog box, enter the IP addresses or CIDR blocks that are allowed to access the instance

and click **OK**.

> **Note**
>
> - Limits for IP address whitelists:
>   - You must separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.
>
>   Supported formats are `0.0.0.0/0`, IP addresses such as `10.23.12.24`, or CIDR blocks such as `10.23.12.24/24`. `/24` indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits.
>   - If an IP address whitelist is empty or contains only `0.0.0.0/0`, all IP addresses can access the ApsaraDB RDS instance. This may cause security risks to the instance. Proceed with caution.
> - If you click **Add Internal IP Addresses of ECS Instances**, the IP addresses of all the Elastic Compute Service (ECS) instances that are created within your Apsara Stack tenant account appear. Then, you can select the IP addresses and add them to an IP address whitelist.

10.1.9.2. Configure SSL encryption

This topic describes how to enhance endpoint security. You can enable Secure Sockets Layer (SSL) encryption and install SSL certificates that are issued by certificate authorities (CAs) to the required application services. SSL is used at the transport layer to encrypt network connections and enhance the security and integrity of communication data. However, SSL increases the response time.

Precautions

- An SSL CA certificate is valid for one year. You must update the validity period of the SSL CA certificate in your application or client within one year. Otherwise, your application or client that uses encrypted network connections cannot connect to the ApsaraDB RDS instance.
- SSL encryption may cause a significant increase in CPU utilization. We recommend that you enable SSL encryption only when you want to encrypt connections from the Internet. In most cases, connections that use an internal endpoint do not require SSL encryption.
- SSL encryption cannot be disabled after it is enabled. Proceed with caution.

Enable SSL encryption

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. Click the **SSL Encryption** tab.
6. In the SSL Settings section, turn on **SSL Encryption**.
7. In the **Configure SSL** dialog box, select the endpoint for which you want to enable SSL encryption and click **OK**.
8. Click **Download CA Certificate** to download the SSL CA certificate files in a compressed package.

   The downloaded package contains the following files:

   - P7B file: contains the server CA certificate that can be imported into a Windows operating system.
   - PEM file: contains the server CA certificate that can be imported into an operating system rather than Windows or an application that is not Windows-based.

- JKS file: contains the server CA certificate that is stored in a Java-supported truststore. You can use the file to import the CA certificate chain into a Java-based application. The default password is apsaradb.

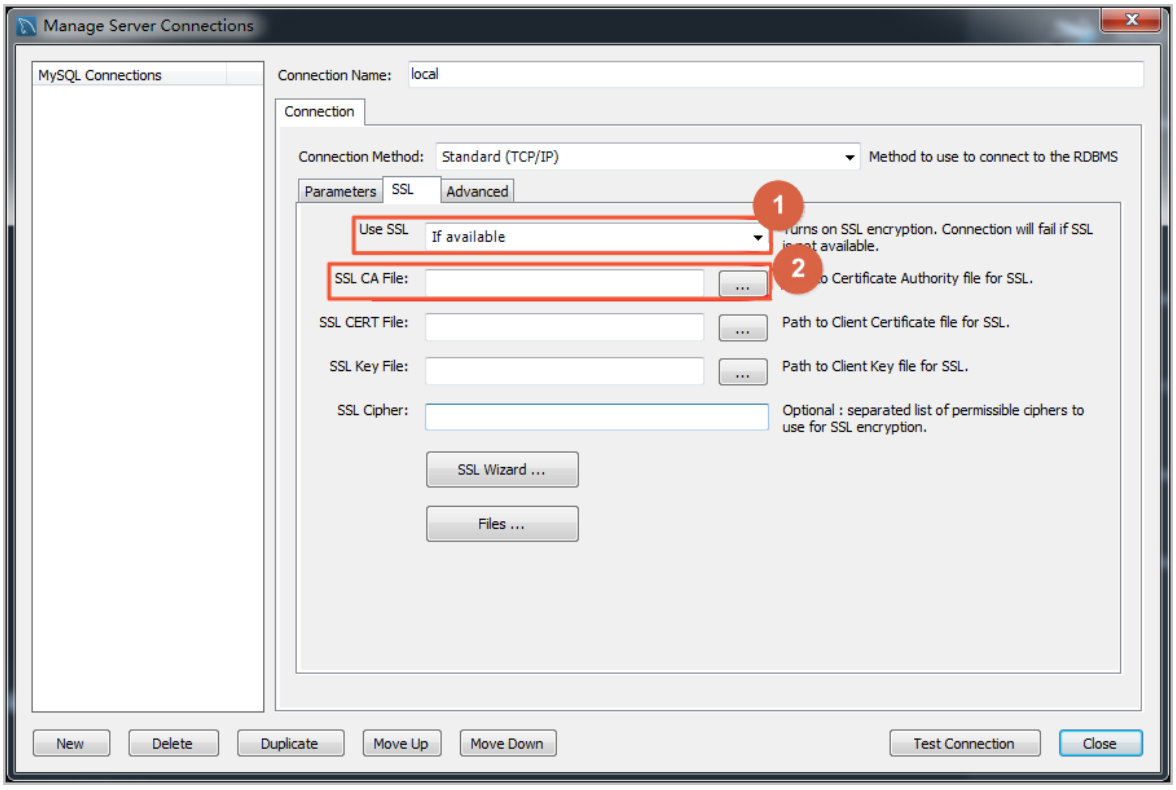
> **Note** When the JKS file is used in Java, you must modify the default JDK security configuration in JDK 7 and JDK 8. Open the `/jre/lib/security/java.security` file on the host where your application resides, and modify the following configurations:
>
> ```
> jdk.tls.disabledAlgorithms=SSLv3, RC4, DH keySize < 224
> jdk.certpath.dis
> abledAlgorithms=MD2, RSA keySize < 1024
> ```
>
> If you do not modify the JDK security configuration, the following error is reported. Similar errors are also caused by the Java security configuration.
>
> ```
> javax.net.ssl.SSLHandshakeException: DHPublicKey does not comply to algorithm constraints
> ```

Configure an SSL CA certificate

After you enable SSL encryption, configure the SSL CA certificate on your application or client before they can connect to the ApsaraDB RDS instance. This section describes how to configure an SSL CA certificate. MySQL Workbench and Navicat are used in the example. If you are using other applications or clients, see the related instructions.

Configure a certificate on MySQL Workbench

1. Start MySQL Workbench.
2. Choose **Database > Manage Connections**.
3. Enable **Use SSL** and import the SSL CA certificate file.



Configure a certificate on Navicat

1. Start Navicat.

2. Right-click the database and select **Edit Connection**.
3. Click the **SSL** tab. Select the path of the PEM-formatted CA certificate, as shown in the following figure.
4. Click **OK**.

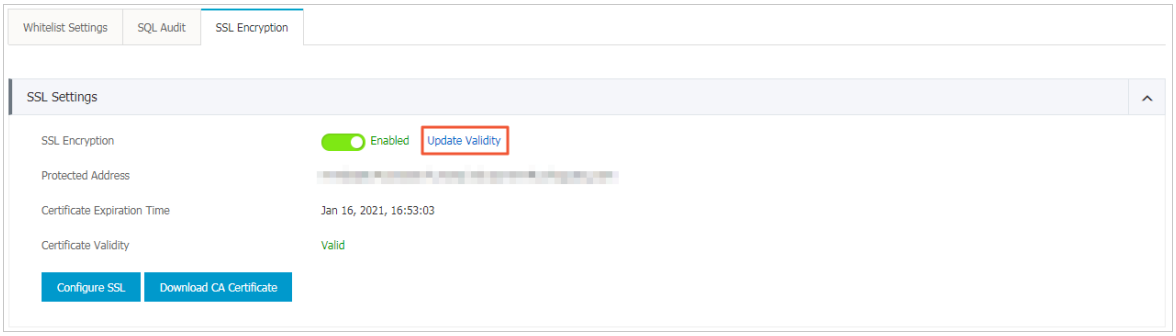
> **Note** If the `connection is being used` error is reported, the previous session is still connected. Restart Navicat.

5. Double-click the database to test whether the database is connected.

Update the validity period of an SSL CA certificate

> **Note**
> - **Update Validity** causes the ApsaraDB RDS instance to restart. Proceed with caution.
> - After you perform the **Update Validity period** operation, you must download and configure the SSL CA certificate again.

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. Click the **SSL Encryption** tab.
6. Click **Update Validity**.



7. In the message that appears, click **OK**.

10.1.9.3. Configure TDE

This topic describes how to configure Transparent Data Encryption (TDE) for your ApsaraDB RDS for SQL Server instance. TDE allows your ApsaraDB RDS instance to encrypt the data that will be written into the disk and decrypt the data that will be read from the disk to the memory. TDE does not increase the sizes of data files. When you use TDE, you do not need to modify the application that uses the ApsaraDB RDS instance.

Precautions

- Instance-level TDE can be enabled but cannot be disabled. Database-level TDE can be enabled or disabled.
- The keys used for data encryption are generated and managed by Key Management Service (KMS). ApsaraDB RDS does not provide the keys or certificates used for data encryption. If you want to restore data to your computer after TDE is enabled, you must decrypt the data on your ApsaraDB RDS instance. For more information, see Decrypt data.
- TDE increases CPU utilization.

Prerequisites

- Your ApsaraDB RDS instance runs SQL Server EE.
- KMS is activated. If KMS is not activated, you can activate it as prompted when you enable TDE.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. Click the **TDE** tab.
6. Turn on **TDE Status**.
7. In the dialog box that appears, click **Confirm**.

   > **Note** If you have not enabled KMS, you are prompted to do so when you enable TDE. After you enable KMS, you can turn on **TDE Status** to enable TDE.

8. Click **Configure TDE**. In the Database TDE Settings dialog box, select the databases you want to encrypt from the Unselected Databases list, click the icon to add them to the **Selected Databases** list, and then click **OK**.

Decrypt data

If you want to decrypt a database that is encrypted by using TDE, you need only to remove the database from the Selected Databases section in the **Database TDE Settings** dialog box.

10.1.10. Service availability

10.1.10.1. Switch workloads over between primary and secondary ApsaraDB RDS instances

This topic describes how to switch workloads over between a primary ApsaraDB RDS instance and its secondary instance. ApsaraDB RDS supports both manual switchover and automatic switchover. After a switchover is complete, the primary ApsaraDB RDS instance becomes the secondary instance.

Context

- Automatic switchover: By default, the automatic switchover feature is enabled. If the primary ApsaraDB RDS instance becomes faulty, ApsaraDB RDS automatically switches workloads over to the secondary instance.
- Manual switchover: You can manually switch workloads over between the primary and secondary ApsaraDB RDS instances even when the automatic switchover feature is enabled.

> **Note** Data is synchronized in real time between the primary and secondary ApsaraDB RDS instances. You can access only the primary instance. The secondary instance serves only as a standby and does not allow external access.

Precautions

- You may encounter a network interruption during a switchover. Make sure that your application is configured to automatically reconnect to the instance.

- If the primary ApsaraDB RDS instance is attached with read-only instances, the read-only instances need to re-establish the connections that are used for data replication and synchronize incremental data after a switchover. As a result, data on the read-only instances shows latencies of a few minutes.

Perform a manual switchover

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Service Availability**.
4. Click **Switch Primary/Secondary Instance** on the right side of the page.

   > **Note** You may encounter a network interruption during a switchover. Make sure that your application is configured to automatically reconnect to the instance.

5. In the dialog box that appears, click **OK**.

   > **Note** In the dialog box, you can also select **Switch Within Maintenance Window** and click **OK**. Then, the system performs the primary/secondary switchover within the maintenance window. For more information about how to set the maintenance time, see Set a maintenance window. You can also click **Change** on the right to change the maintenance window.

FAQ

Q: Can I connect to secondary instances?

A: No, you cannot connect to secondary instances. You can connect only to primary instances. Secondary instances serve only as a standby and do not allow external access.

10.1.11. Database backup and restoration

10.1.11.1. Configure an automatic backup policy

Automatic backup supports full physical backups. ApsaraDB RDS automatically backs up data based on pre-configured policies. This topic describes how to configure a policy for automatic backup.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Backup and Restoration**.
5. On the **Backup and Restoration** page, click the **Backup Settings** tab.
6. Click **Edit**.
7. In the dialog box that appears, configure the automatic backup policy.

Parameter	Description
Data Retention Period	The number of days for which you want to retain data backup files. Valid values: 7 to 730. Default value: 7.

Parameter	Description
Backup Cycle	The cycle based on which you want to create a backup. You can select one or more days within a week. **Note** For data security purposes, we recommend that you back up your ApsaraDB RDS instance at least twice a week.
Backup Time	The period of time for which you want to back up data. Unit: hours.
Backup Frequency	The frequency at which you want to back up logs. The following options are available: ○ Same as Data Backup ○ Every 30 Minutes The total size of log backup files remains the same regardless of the backup frequency.

8. Click **OK**.

10.1.11.2. Manually back up an instance

This topic describes how to manually back up an ApsaraDB RDS instance.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. On the **Basic Information** page, click **Back Up Instance** in the upper-right corner.
5. In the **Back Up Instance** dialog box, select **Automatic Backup** or **Full Backup** from the **Select Backup Mode** drop-down list.

   Note ApsaraDB RDS supports the following backup methods:
   - **Automatic Backup**: After you select Automatic Backup, the system immediately performs an incremental or full backup based on the instance.
   - **Full Backup**: After you select Full Backup, the system immediately performs a full backup.

6. Click **OK**.

Result

After the backup is complete, you can view the backup task on the **Data Backup** tab of the **Backup and Restoration** page.

10.1.11.3. Shrink transaction logs

ApsaraDB RDS for SQL Server allows you to shrink transaction logs to reduce the log file size.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Backup and Restoration**.
5. In the upper-right corner of the page, click **Shrink Transaction Log**. In the message that appears, click **OK**.

> **Note** The shrinkage takes about 20 minutes to complete. ApsaraDB RDS for SQL Server shrinks transaction logs during each backup.

10.1.12. Migrate full backup data to ApsaraDB RDS for SQL Server

This topic describes how to migrate full backup files of an on-premises database from Object Storage Service (OSS) to ApsaraDB RDS for SQL Server.

Prerequisites

- Your ApsaraDB for RDS instance has sufficient storage space. If the space is insufficient, you must increase it before you migrate data to the instance.
- The destination database on your ApsaraDB for RDS instance has a different name from the on-premises database.
- A privileged account is created on your ApsaraDB for RDS instance. For more information, see Create an account.
- An Object Storage Service (OSS) bucket is created in the region where your ApsaraDB for RDS instance is created. For more information, see Create buckets in the *OSS User Guide*.
- The DBCC CHECKDB statement is executed, and the execution result indicates that no allocation or consistency errors occur.

> **Note** If no allocation or consistency errors occur, the following execution result is returned:

```
...
CHECKDB found 0 allocation errors and 0 consistency errors in database 'xxx'.
DBCC execution completed. If DBCC printed error messages, contact your system administrator.
```

Precautions

- Full backup files cannot be migrated to an ApsaraDB for RDS instance of an earlier SQL Server version. For example, if the on-premises database runs SQL Server 2016 and your ApsaraDB for RDS instance runs SQL Server 2012, you cannot migrate full backup files of the on-premises database to your ApsaraDB for RDS instance.
- Differential or log backup files are not supported.
- The names of full backup files cannot contain special characters, such as `at signs (@) and vertical bars (|)`. If the file names contain special characters, the migration fails.
- After the service account of your ApsaraDB for RDS instance is granted the access permission on the OSS bucket, the system creates a role named **AliyunRDSImportRole** in RAM. Do not modify or delete this role. Otherwise, you cannot download full backup files when you migrate data to your ApsaraDB for RDS instance. In this case, you must re-authorize the service account of your ApsaraDB for RDS instance.
- Before the migration is complete, do not delete the backup files from the OSS bucket. Otherwise, the migration fails.

- The names of backup files can be suffixed only with bak, diff, trn, or log. If you do not use the script in this topic to generate a backup file, you must name the backup file by using one of the following suffixes:
  - bak: indicates a full backup file.
  - diff: indicates a differential backup file.
  - trn or log: indicates a log backup file.

Back up the on-premises database

> **Note** Before you perform a full backup, stop writing data to the on-premises database. The data written during the backup process is not backed up.

1. Download the backup script. Double-click the backup script to open it by using the Microsoft SQL Server Management Studio (SSMS) client.
2. Configure the following parameters.

Parameter	Description
@backup_databases_list	The databases that you want to back up. Separate them with semicolons (;) or commas (,).
@backup_type	The backup type. Valid values: ○ FULL: full backup ○ DIFF: differential backup ○ LOG: log backup
@backup_folder	The directory in which you want to store the backup files on your computer. If the specified directory does not exist, the system creates a directory.
@is_run	Specifies whether to perform a backup. Valid values: ○ 1: performs a backup. ○ 0: performs no backup but a check.

3. Run the backup script.

Upload full backup files to the OSS bucket

After the on-premises database is backed up, you must upload full backup files to the OSS bucket. You can use one of the following methods:

- Use the OSS console

  If the size of backup files is smaller than 5 GB, you can upload the files in the OSS console. For more information, see Upload objects in the *OSS User Guide*.

- Call an OSS API operation

  You can call an OSS API operation to upload the full backup files in resumable mode. For more information, see Multipart upload-relevant operations in the *OSS Developer Guide*.

Create a migration task

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Backup and Restoration**.
4. In the upper-right corner of the page, click **Migrate OSS Backup Data to RDS**.
5. Click **Next** twice until the Import Data step appears.

6. Configure the following parameters.

Parameter	Description
Database Name	Enter the name of the destination database on your ApsaraDB for RDS instance. **Note** The name of the database must meet the requirements of SQL Server.
OSS Bucket	Select the OSS bucket that stores the backup files.
OSS Subfolder Name	Enter the name of the OSS subfolder that stores the backup files.
OSS File	Click the search icon to search for backup files by using the prefix-based fuzzy match. The system displays the name, size, and update time of each backup file. Select the backup file that you want to migrate to your ApsaraDB for RDS instance.
Cloud Migration Method	**One-time Full Backup File Migration**: uploads full backup data to your ApsaraDB for RDS instance. Select this option if you want to migrate only a single full backup file.

7. Click **OK**.

Wait for the migration task to complete. You can click **Refresh** to view the latest status of the migration task. If the migration fails, fix the error based on the message displayed in the Task Description column. For more information, see Common errors.

View the migration task

In the left-side navigation pane, click **Backup and Restoration**. Click the **Backup Data Upload History** tab. The system displays the migration tasks in the last week.

Common errors

Each record of a migration task contains a task description, which helps you identify the error cause and fix the error. The following list describes common errors:

- A database with the same name as the on-premises database exists on your ApsaraDB for RDS instance.
  - Error message: The database (xxx) is already exist on RDS, please backup and drop it, then try again.
  - Cause: The on-premises database is named the same as an existing database on your ApsaraDB for RDS instance. For data security purposes, ApsaraDB RDS for SQL Server does not allow such a database to be migrated.
  - Solution: If you need to overwrite the database in your ApsaraDB for RDS instance with the on-premises database, you must back up the database, delete it from your ApsaraDB for RDS instance, and then migrate the on-premises database to your ApsaraDB for RDS instance.
- A differential backup file is used.
  - Error message: Backup set (xxx.bak) is a Database Differential backup, we only accept a FULL Backup.
  - Cause: The file that you uploaded is a differential backup file, but not a full backup file. The migration solution for full backup data supports only full backup files.
- A log backup file is used.
  - Error message: Backup set (xxx.trn) is a Transaction Log backup, we only accept a FULL Backup.
  - Cause: The file that you uploaded is a log backup file, but not a full backup file. The migration solution for full backup data supports only full backup files.
- The backup file fails the verification.
  - Error message: Failed to verify xxx.bak, backup file was corrupted or newer edition than RDS.

  - Cause: The backup file is damaged, or the on-premises database runs an SQL Server version later than your ApsaraDB for RDS instance. For example, if the on-premises database runs SQL Server 2016 and your ApsaraDB for RDS instance runs SQL Server 2012, the error message is returned.
  - Solution: If the backup file is damaged, perform a full backup on the on-premises database again. If the database engine version does not meet the requirements, select an ApsaraDB for RDS instance that runs the same version as or a later version than the on-premises database.

- DBCC CHECKDB fails to be executed.
  - Error message: DBCC checkdb failed.
  - Cause: Allocation or consistency errors occurred in the on-premises database.
  - Solution: Execute the following statement in the on-premises database.

    > **Note** Data loss may occur when you use this statement to fix errors.

    ```
    DBCC CHECKDB (DBName, REPAIR_ALLOW_DATA_LOSS) WITH NO_INFOMSGS, ALL_ERRORMSGS
    ```

- The remaining storage space of your ApsaraDB for RDS instance is insufficient. (1)
  - Error message: Not Enough Disk Space for restoring, space left (xxx MB) < needed (xxx MB).
  - Cause: The remaining storage space of your ApsaraDB for RDS instance does not meet the migration requirements.
  - Solution: Increase the storage space of your ApsaraDB for RDS instance.

- The remaining storage space of your ApsaraDB for RDS instance is insufficient. (2)
  - Error message: Not Enough Disk Space, space left xxx MB < bak file xxx MB.
  - Cause: The remaining storage space of your ApsaraDB for RDS instance is smaller than the size of the backup file.
  - Solution: Increase the storage space of your ApsaraDB for RDS instance.

- No privileged account exists.
  - Error message: Your RDS doesn't have any init account yet, please create one and grant permissions on RDS console to this migrated database (XXX).
  - Cause: No privileged account is created on your ApsaraDB for RDS instance, and the database permissions are not granted to accounts. However, when this error message is returned, the backup file has been restored to your ApsaraDB for RDS instance, and the migration task is successful.
  - Solution: Create a privileged account. For more information, see Create an account.

11.ApsaraDB RDS for PostgreSQL

11.1. User Guide(RDS PostgreSQL)

11.1.1. What is ApsaraDB RDS?

ApsaraDB RDS is a stable, reliable, and scalable online database service. Based on the distributed file system and high-performance storage, ApsaraDB RDS allows you to perform database operations and maintenance with its set of solutions for disaster recovery, backup, restoration, monitoring, and migration.

ApsaraDB RDS supports four database engines, which are MySQL, SQL Server, PolarDB, and PostgreSQL. You can create database instances based on these database engines to meet your business requirements. This topic describes the PostgreSQL engine.

ApsaraDB RDS for PostgreSQL

ApsaraDB RDS for PostgreSQL is the most advanced open source database. It is fully compatible with SQL and supports a diverse range of data formats such as JSON, IP, and geometric data. In addition to features such as transactions, subqueries, multi-version concurrency control (MVCC), and data integrity check, ApsaraDB RDS for PostgreSQL integrates a series of features including high availability, backup, and restoration to ease operation and maintenance loads.

11.1.2. Limits on ApsaraDB RDS for PostgreSQL

Before you use ApsaraDB RDS for PostgreSQL, you must understand its limits and take the necessary precautions.

The following table describes the limits on ApsaraDB RDS for PostgreSQL.

Operation	Limit
Root permissions of databases	Superuser permissions are not provided.
Database replication	ApsaraDB RDS for PostgreSQL provides a primary/secondary replication architecture except in the Basic Edition. The secondary instances in the architecture are hidden and cannot be accessed by your applications.
Instance restart	Instances must be restarted by using the ApsaraDB RDS console or API operations.

11.1.3. Log on to the ApsaraDB RDS console

This topic describes how to log on to the ApsaraDB RDS console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

> **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
>
> - Uppercase or lowercase letters
> - Digits
> - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products > Database Services > ApsaraDB RDS**.

11.1.4. Quick Start

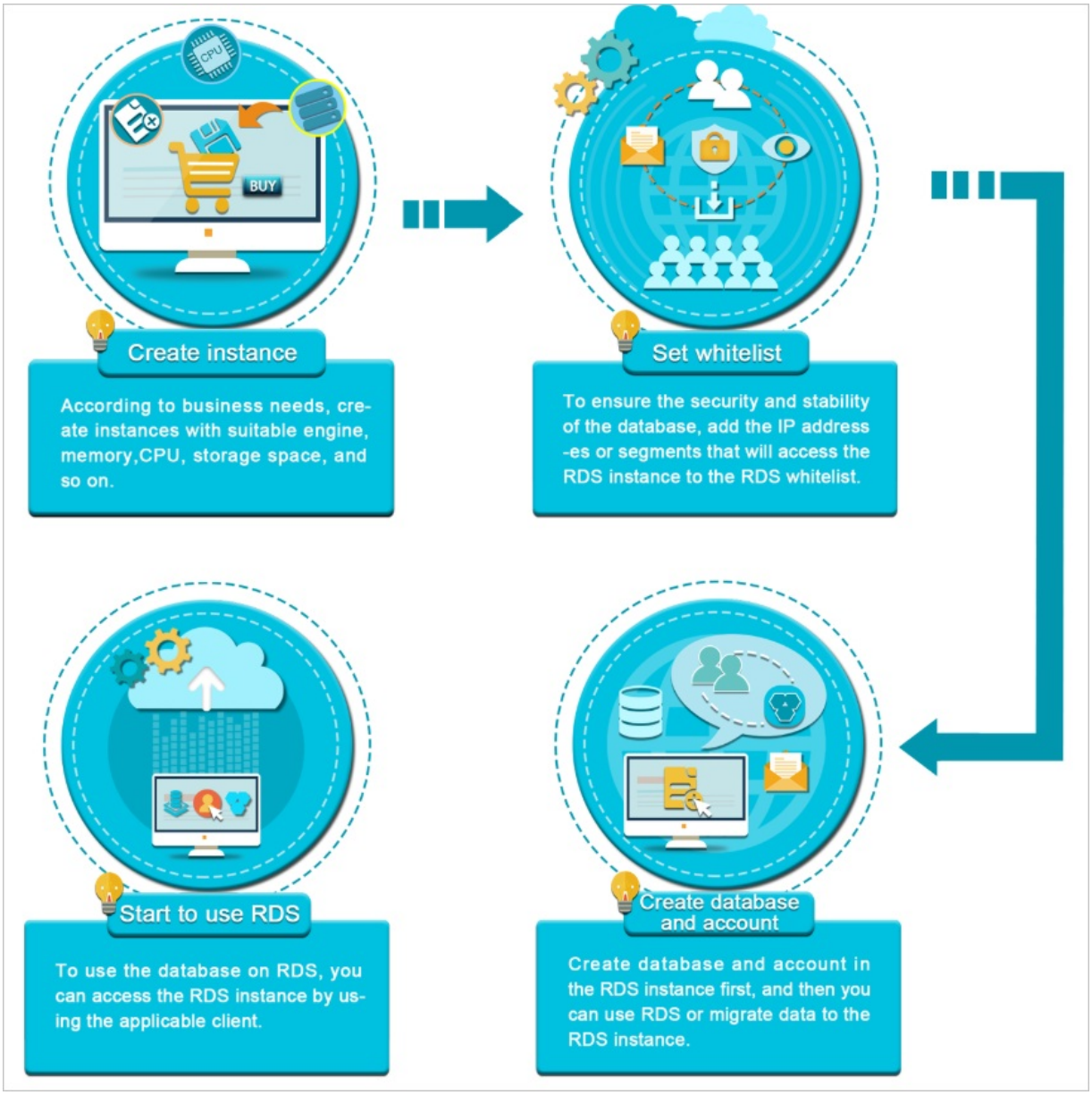
11.1.4.1. Procedure

ApsaraDB RDS quick start covers the following topics: creating an ApsaraDB RDS instance, configuring a whitelist, creating a database, creating an account, and connecting to the instance.

Flowchart for an ApsaraDB RDS instance

If you are using ApsaraDB RDS for the first time, you can start with Limits.

The following figure shows the operations that you must perform before you use an ApsaraDB RDS instance.



11.1.4.2. Create an instance

This topic describes how to create one or more ApsaraDB RDS for PostgreSQL instances in the ApsaraDB RDS console.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.
3. Configure the following parameters.

Section	Parameter	Description
Basic	**Organization**	The organization to which the instance belongs.

Settings Section	Parameter	Description
	Resource Set	The resource set to which the instance belongs.
Region	**Region**	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.
	Deployment Method	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
Specifications	**Quantity**	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.
	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **PostgreSQL**.
	Engine Version	The version of the database engine. Valid values: ○ 9.4 ○ 10.0 ○ 11.0 ○ 12.0

Section	Parameter	Description
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Type	The storage type of the instance. Local and standard SSDs are supported. **Note** Instances that run PostgreSQL 9.4 support only local SSDs.
	Encrypted	Specifies whether to encrypt the standard SSD. This parameter is available only when **Storage Type** is set to **Standard SSD**. If you select Encrypted, you must specify the **Encryption Key** parameter. If you do not have a key, you must first create one in Key Management Service (KMS). For more information, see Create a CMK in *Key Management Service User Guide*.
	Encryption Key	The key that is used to encrypt the standard SSD. This parameter is available only when you select **Encrypted**.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, which includes the space to store data, system files, binlog files, and transaction files. Valid values: 20 to 600. Unit: GB. The value must be in 1 GB increments.
Network	**Network Type**	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security. **Note** ■ If you configure multi-zone deployment, you must create vSwitches for the zones of primary and secondary instances in the specified VPC. ■ If you select VPC, you must specify a VPC and a vSwitch.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click **Submit**.

11.1.4.3. Configure an IP address whitelist

This topic describes how to configure a whitelist for an ApsaraDB RDS instance. Only entities that are listed in a whitelist can access your ApsaraDB RDS instance.

Context

Whitelists make your ApsaraDB RDS instance more secure and do not interrupt the operations of your ApsaraDB RDS instance when you configure whitelists. We recommend that you perform maintenance on your whitelists on a regular basis.

To configure a whitelist, perform the following operations:

- Configure a whitelist: Add IP addresses to allow them to connect to the ApsaraDB RDS instance.

  > **Note** The IP address whitelist labeled default contains only the default IP address 0.0.0.0/0, which allows all entities to access your ApsaraDB RDS instance.

- Configure an ECS security group: Add an ECS security group for the ApsaraDB RDS instance to allow ECS instances in the group to connect to the ApsaraDB RDS instance.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. On the **Whitelist Settings** tab, click **Edit** corresponding to the **default** whitelist.

   > **Note** You can also click **Create Whitelist** to create a whitelist.

6. In the **Edit Whitelist** dialog box, enter the IP addresses or CIDR blocks used to access the instance and click **OK**. The following section describes the rules:
   - If you enter the CIDR block 10.10.10.0/24 in the IP Addresses field, all IP addresses in the 10.10.10.X format can access your ApsaraDB RDS instance.
   - If you enter more than one IP address or CIDR block, you must separate them with commas (,). Do not add spaces before or after the commas. Example: 192.168.0.1,172.16.213.9.
   - If you click **Add Internal IP Addresses of ECS Instances**, the IP addresses of all ECS instances created within your Alibaba Cloud account are displayed. You can select the required IP addresses to add them to the IP address whitelist.

11.1.4.4. Create a database and an account

Before you start to use ApsaraDB RDS, you must create databases and accounts on an ApsaraDB RDS instance. This topic describes how to create a database and an account on an ApsaraDB RDS for PostgreSQL instance.

Account types

ApsaraDB RDS for PostgreSQL instances support two types of accounts: privileged accounts and standard accounts. The following table describes these account types.

Account type	Description

Account type	Description
Privileged account	• You can create and manage privileged accounts only by using the ApsaraDB RDS console or API operations. • If your ApsaraDB RDS instance uses local SSDs, you can create only a single privileged account. If your ApsaraDB RDS instance uses standard or enhanced SSDs, you can create more than one privileged account. A privileged account allows you to manage all the standard accounts and databases that are created on your ApsaraDB RDS instance. • A privileged account has more permissions that allow you to manage your ApsaraDB RDS instance at more fine-grained levels. For example, you can grant the query permissions on different tables to different users. • A privileged account has the permissions to disconnect accounts that are created on your ApsaraDB RDS instance.
Standard account	• You can create and manage standard accounts by using the ApsaraDB RDS console, API operations, or SQL statements. • You can create more than one standard account on your ApsaraDB RDS instance. • You must grant the permissions on specific databases to a standard account. • A standard account does not have the permissions to create, manage, or disconnect other accounts on your ApsaraDB RDS instance.

Precautions

- If your ApsaraDB RDS instance uses local SSDs, you can create one privileged account in the ApsaraDB RDS console. After the privileged account is created, it cannot be deleted. You can also create and manage more than one standard account by using SQL statements.
- If your ApsaraDB RDS instance uses standard or enhanced SSDs, you can create more than one privileged account and standard account in the ApsaraDB RDS console. You can also create and manage more than one standard account by using SQL statements.
- To migrate data from an on-premises database to your ApsaraDB RDS instance, you must create a database and an account on the ApsaraDB RDS instance. Make sure that the created database has the same properties as the on-premises database. Also make sure that the created account has the same permissions on the created database as the account that is authorized to manage the on-premises database.
- Follow the least privilege principle to create accounts and grant them read-only permissions or read and write permissions on databases. If necessary, you can create more than one account and grant them only the permissions on specific databases. If an account does not need to write data to a database, grant only the read-only permissions on that database to the account.
- For security purposes, we recommend that you specify strong passwords for the accounts on your ApsaraDB RDS instance and change the passwords on a regular basis.

Create a privileged account on an instance that uses local SSDs

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. On the Accounts page, click **Create Privileged Account** and configure the following parameters.

Parameter	Description

Parameter	Description
Database Account	○ The name of the account must be 2 to 16 characters in length. ○ The name can contain lowercase letters, digits, and underscores (_). ○ The name must start with a letter and end with a letter or digit.
Password	○ The password of the account must be 8 to 32 characters in length. ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include ! @#$%^&*()_+-=
Re-enter Password	Enter the password of the account again.

6. Click **Create**.

Create a privileged or standard account on an instance that uses standard or enhanced SSDs

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. On the Accounts page, click **Create Account** and configure the following parameters.

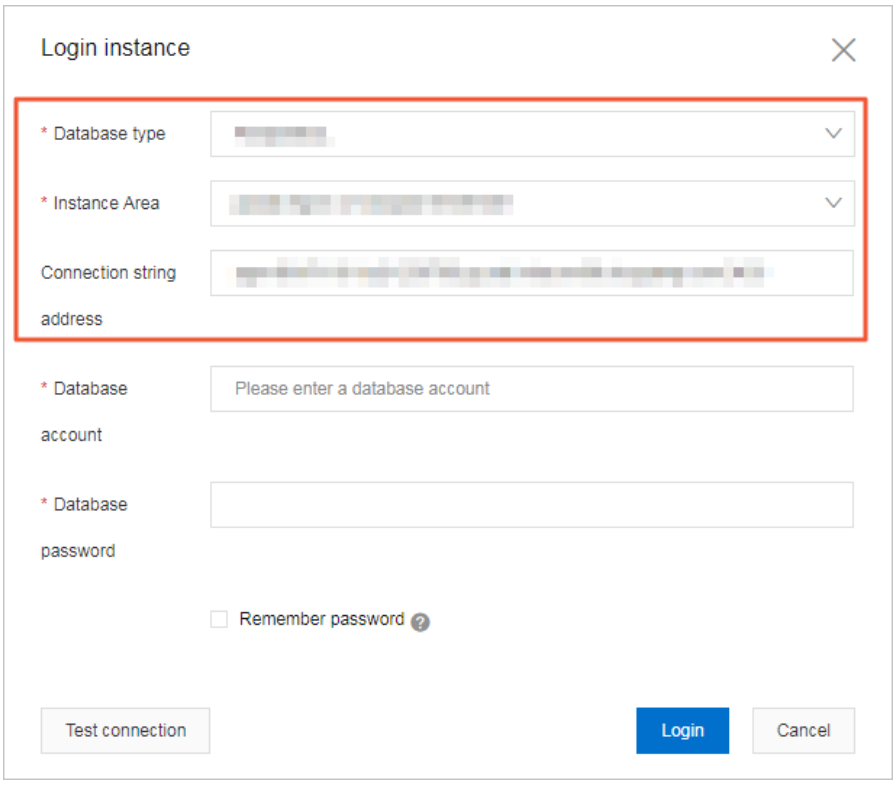
Parameter	Description
Database Account	○ The name of the account must be 2 to 16 characters in length. ○ The name can contain lowercase letters, digits, and underscores (_). ○ The name must start with a letter and end with a letter or digit.
Account Type	Select **Privileged Account** or **Standard Account**.
Password	○ The password of the account must be 8 to 32 characters in length. ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ Special characters include ! @#$%^&*()_+-=
Re-enter Password	Enter the password of the account again.
Description	This parameter is optional. You can enter relevant description to make the instance identifiable. The description can be up to 256 characters in length.

6. Click **Create**.

Create a database and a standard account

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.

3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**. If you want the browser to remember the password, select **Remember password** before you click **Login**.

> **Note** If a message prompts you that the connection fails, the problem may be caused by an improperly configured whitelist. Reconfigure the whitelist in the console. For more information, see Configure an IP address whitelist.

7. The **SQLConsole** page appears after you log on to the instance. Execute a statement in the following format to create a database:

```
CREATE DATABASE name
 [ [ WITH ] [ OWNER [=] user_name ]
        [ TEMPLATE [=] template ]
        [ ENCODING [=] encoding ]
        [ LC_COLLATE [=] lc_collate ]
        [ LC_CTYPE [=] lc_ctype ]
        [ TABLESPACE [=] tablespace_name ]
        [ CONNECTION LIMIT [=] connlimit ] ]
```

For example, if you want to create a database named test, execute the following statement:

```
create database test;
```

8. Click **execute**.
9. In the SQL window, execute a statement in the following format to create a standard account:

```
CREATE USER name [ [ WITH ] option [ ... ] ]
where option can be:
   SUPERUSER | NOSUPERUSER
 | CREATEDB | NOCREATEDB
 | CREATEROLE | NOCREATEROLE
 | CREATEUSER | NOCREATEUSER
 | INHERIT | NOINHERIT
 | LOGIN | NOLOGIN
 | REPLICATION | NOREPLICATION
 | CONNECTION LIMIT connlimit
 | [ ENCRYPTED | UNENCRYPTED ] PASSWORD 'password'
 | VALID UNTIL 'timestamp'
 | IN ROLE role_name [, ...]
 | IN GROUP role_name [, ...]
 | ROLE role_name [, ...]
 | ADMIN role_name [, ...]
 | USER role_name [, ...]
 | SYSID uid
```

For example, if you want to create a user account named test2 whose password is 123456, execute the following statement:

```
create user test2 password '123456';
```

10. Click **execute**.

11.1.4.5. Connect to an ApsaraDB RDS for PostgreSQL instance

This topic describes how to use Data Management (DMS) or the pgAdmin 4 client to connect to an ApsaraDB RDS instance.

Context

You can log on to DMS from the ApsaraDB RDS console and then connect to an ApsaraDB RDS instance.

DMS is a data management service that integrates data, schema, and server management, access security, BI charts, data trends, data tracking, and performance optimization. DMS can be used to manage relational and non-relational databases, such as MySQL, SQL Server, PostgreSQL, MongoDB, and Redis. It can also be used to manage Linux servers.

You can also use a client to connect to an ApsaraDB RDS instance. ApsaraDB RDS for PostgreSQL is fully compatible with PostgreSQL. You can connect to an ApsaraDB RDS for PostgreSQL instance in a similar manner as you would connect to an open source PostgreSQL instance. In this topic, the pgAdmin 4 client is used to connect to an ApsaraDB RDS instance.

Use DMS to connect to an ApsaraDB RDS instance

For more information about how to use DMS to connect to an ApsaraDB RDS instance, see Log on to an ApsaraDB for RDS instance by using DMS.

Use the pgAdmin 4 client to connect to an ApsaraDB RDS instance

1. Add the IP address of the pgAdmin client to an IP address whitelist of the ApsaraDB RDS instance. For more information about how to configure a whitelist, see Configure an IP address whitelist.
2. Start the pgAdmin 4 client.

   > **Note** For information about how to download the pgAdmin 4 client, visit pgAdmin 4 (Windows).

3. Right-click **Servers** and choose **Create > Server**, as shown in the following figure.
4. On the **General** tab of the **Create - Server** dialog box, enter the name of the server, as shown in the following figure.
5. Click the **Connection** tab and enter the information of the instance, as shown in the following figure.

Parameter	Description
Host name/address	The internal endpoint of the ApsaraDB RDS instance. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
Port	The internal port number that is used to connect to the ApsaraDB RDS instance. For more information about how to view the internal port number, see View and modify the internal endpoint and port number.
Username	The name of the privileged account on the ApsaraDB RDS instance. For more information about how to obtain a privileged account, see Create a database and an account.
Password	The password of the privileged account of the ApsaraDB RDS instance.

6. Click **Save**.
7. If the connection information is correct, choose **Servers > Server Name > Databases > postgres**. If the following page appears, the connection is established.

   > **Notice** The postgres database is the default system database of the ApsaraDB RDS instance. Do not perform operations on this database.

11.1.5. Instances

11.1.5.1. Create an instance

This topic describes how to create one or more ApsaraDB RDS for PostgreSQL instances in the ApsaraDB RDS console.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.

3. Configure the following parameters.

Section	Parameter	Description
Basic Settings	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	Region	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.
	Deployment Method	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
	Quantity	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.
	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **PostgreSQL**.

Section	Parameter	Description
Specifications	**Engine Version**	The version of the database engine. Valid values: ○ 9.4 ○ 10.0 ○ 11.0 ○ 12.0
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Type	The storage type of the instance. Local and standard SSDs are supported. **Note** Instances that run PostgreSQL 9.4 support only local SSDs.
	Encrypted	Specifies whether to encrypt the standard SSD. This parameter is available only when **Storage Type** is set to **Standard SSD**. If you select Encrypted, you must specify the **Encryption Key** parameter. If you do not have a key, you must first create one in Key Management Service (KMS). For more information, see Create a CMK in *Key Management Service User Guide*.
	Encryption Key	The key that is used to encrypt the standard SSD. This parameter is available only when you select **Encrypted**.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, which includes the space to store data, system files, binlog files, and transaction files. Valid values: 20 to 600. Unit: GB. The value must be in 1 GB increments.

Section	Parameter	Description
Network	Network Type	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security. Note ■ If you configure multi-zone deployment, you must create vSwitches for the zones of primary and secondary instances in the specified VPC. ■ If you select VPC, you must specify a VPC and a vSwitch.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click **Submit**.

11.1.5.2. Create an ApsaraDB RDS for PostgreSQL instance that uses standard SSDs

Cloud disks are block-level data storage products provided by Alibaba Cloud for Elastic Compute Service (ECS). They provide low latency, high performance, durability, and reliability. This topic describes how to create one or more instances that use standard SSDs in the ApsaraDB RDS console.

Prerequisites

An instance that runs PostgreSQL 10.0 or later can be created.

Context

An ApsaraDB RDS instance with standard SSDs uses a distributed triplicate mechanism to ensure 99.9999999% data reliability. If service disruptions occur within a zone due to hardware faults, data in that zone is copied to an unaffected disk in another zone to ensure data availability.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, click **Create Instance** in the upper-right corner.
3. Configure the following parameters.

Section	Parameter	Description
Basic Settings	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	Region	The region in which you want to create the instance. Services in different regions cannot communicate over an internal network. After the instance is created, the region cannot be changed.
	Zone of Primary Node	The zone where the primary instance is deployed.
	Deployment Method	Specifies whether to deploy the primary and secondary instances in separate zones. ApsaraDB RDS supports **Multi-zone Deployment** and **Single-zone Deployment**. If you select **Multi-zone Deployment**, you must configure **Zone of Secondary Node**.
	Zone of Secondary Node	The zone where the secondary instance is deployed. This parameter is available only when **Deployment Method** is set to **Multi-zone Deployment**. **Note** If you select the same zone for both the primary and secondary instances, the deployment is equivalent to single-zone deployment.
Specifications	Quantity	The number of ApsaraDB RDS instances that you want to create. Default value: 1.
	Instance Name	The name of the instance. ○ The name must be 2 to 64 characters in length. ○ The name must start with a letter. ○ The name can contain letters, digits, and the following special characters: _ - : ○ The name cannot start with http:// or https://.
	Connection Type	The connection type of the instance. ApsaraDB RDS instances support the following connection types: ○ **Internet**: ApsaraDB RDS instances of this connection type can be connected over the Internet. ○ **Internal Network**: ApsaraDB RDS instances of this connection type can be connected over an internal network. **Note** The value of this parameter cannot be changed after the instance is created. Proceed with caution.
	Database Engine	The database engine of the instance. Select **PostgreSQL**.
	Engine Version	The version of the database engine. Set the value to **10.0** or a later version number.
	Edition	The edition of the instance. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.

Section	Parameter	Description
	Storage Type	The storage type of the instance. Select **Standard SSD**.
	Encrypted	Specifies whether to encrypt the standard SSD. If you select Encrypted, you must specify the **Key** parameter. If you do not have a key, you must first create one in Key Management Service (KMS). For more information, see Configure data encryption. **Note** Disk encryption provides maximum protection for your data with minimal impact on your business or applications. Both the snapshots generated from encrypted disks and the disks created from those snapshots are automatically encrypted.
	Encryption Key	The key that is used to encrypt the standard SSD. This parameter is available only when you select **Encrypted**.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and the input/output operations per second (IOPS). The actual values are displayed in the console. For more information, see Instance types in *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the cloned instance, which includes the space to store data, system files, binlog files, and transaction files. Valid values: 20 to 600. Unit: GB. The value must be in 1 GB increments.
Network	**Network Type**	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways within a VPC. We recommend that you select VPC for improved security. **Note** ■ If you configure multi-zone deployment, you must create vSwitches for the zones of primary and secondary instances in the specified VPC. ■ If you select VPC, you must specify a VPC and a vSwitch.
	IP Address Whitelist	The IP addresses that are allowed to connect to the instance.

4. Click **Submit**.

11.1.5.3. View basic information of an instance

This topic describes how to view the details of an ApsaraDB RDS instance, such as basic information, internal network connection information, status, and configurations.

Procedure

1. Log on to the ApsaraDB RDS console.
2. Use one of the following methods to go to the **Basic Information** page of an instance:
   - On the **Instances** page, click the ID of the instance to go to the **Basic Information** page.
   - On the **Instances** page, find the instance and click **Manage** in the corresponding **Actions** column. The **Basic Information** page appears.

11.1.5.4. Restart an instance

This topic describes how to manually restart an ApsaraDB RDS instance. This applies if the number of connections exceeds the specified threshold or if an instance has performance issues.

Prerequisites

The instance is in the **Running** state.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the upper-right corner of the page, click **Restart Instance**.

   > **Note** When you restart an instance, applications are disconnected from the instance. We recommend that you make appropriate service arrangements before you restart an instance. Proceed with caution.

5. In the message that appears, click **Confirm**.

11.1.5.5. Change the specifications of an instance

This topic describes how to change the specifications of an ApsaraDB RDS instance. You can upgrade or downgrade an ApsaraDB RDS instance to meet your business needs.

Prerequisites

The instance is in the **Running** state and is not in the backing up or restoring state.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the **Configure Information** section of the **Basic Information** page, click **Change Specifications**.
5. On the **Change Specifications** page, set **Edition**, **Instance Type**, and **Storage Capacity**.
6. Click **Submit**.

11.1.5.6. Set a maintenance window

This topic describes how to set a maintenance window for an ApsaraDB RDS instance.

Context

The backend system performs maintenance on the ApsaraDB RDS instances during the maintenance window. This

ensures the stability of the ApsaraDB RDS instance. The default maintenance window is from 02:00 (UTC+8) to 06:00 (UTC+8). We recommend that you set the maintenance window to off-peak hours of your business to avoid impacts on your business.

Precautions

- An instance enters the **Maintaining Instance** state before the maintenance window to ensure stability during the maintenance process. When the instance is in this state, access to data in the database and query operations such as performance monitoring are not affected. However, except for account and database management and IP address whitelist configuration, modification operations such as upgrade, downgrade, and restart are temporarily unavailable.
- During the maintenance window, one or two network interruptions may occur. Make sure that your applications are configured with automatic reconnection policies.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the **Configuration Information** section, click **Configure** to the right of **Maintenance Window**.
5. Select a maintenance window and click **Save**.

> **Note** The maintenance window is displayed in UTC+8.

11.1.5.7. Configure primary/secondary switchover

ApsaraDB RDS provides the primary/secondary switchover feature to ensure the high availability of databases. The primary/secondary switchover is performed when the primary instance becomes unavailable. You can also manually switch your business to the secondary instance. This topic describes how to manually switch over services between a primary instance and its secondary instance.

Context

Data is synchronized in real time between the primary and secondary instances. You can access only the primary instance. The secondary instance serves only as a backup instance and does not allow external access. After the switchover, the original primary instance becomes the secondary instance.

> **Note** Network interruptions may occur during a switchover. Make sure that your applications are configured with automatic reconnection policies.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Service Availability**.
5. In the **Availability Information** section, click **Switch Primary/Secondary Instance**.
6. In the **Switch Primary/Secondary Instance** message, click **OK**.

> **Note**
> - During the switchover, operations such as managing databases and accounts and changing network types cannot be performed. Therefore, we recommend that you select Switch Within Maintenance Window.
> - For more information about how to set a maintenance window, see Set a maintenance window.

11.1.5.8. Release an instance

This topic describes how to manually release an instance.

Precautions

- Only instances in the running state can be released.
- After an instance is released, the instance data is immediately deleted. We recommend that you back up your data before you release an instance.
- When you release a primary instance, all of its read-only instances are also released.

Procedure

1. Log on to the ApsaraDB RDS console.
2. Find the instance that you want to release and choose **More > Release Instance** in the Actions column.
3. In the **Release Instance** message, click **Confirm**.

11.1.5.9. Modify parameters of an instance

This topic describes how to view and modify the values of some parameters and query parameter modification records in the console.

Precautions

- To ensure instance stability, you can modify only specific parameters in the ApsaraDB RDS console.
- When you modify parameters on the **Editable Parameters** tab, refer to the **Value Range** column corresponding to each parameter.
- After specific parameters are modified, you must restart your instance for the changes to take effect. The necessity of restart is displayed in the **Force Restart** column on the **Editable Parameters** tab. We recommend that you modify the parameters of an instance during off-peak hours and make sure that your applications are configured with automatic reconnection policies.

Modify parameters

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Parameters**.
5. Perform the following operations:

   Export the parameter settings of the instance to your computer.

   On the Editable Parameters tab, click **Export Parameters**. The parameter settings of the instance are exported as a TXT file to your computer.

   Modify and import the parameter settings.

i. After you modify parameters in the exported parameter file, click **Import Parameters** and copy the parameter settings to the field.
ii. Click **OK**.
iii. In the upper-right corner of the page, click **Apply Changes**.

> **Note**
> - If the new parameter value takes effect only after an instance restart, the system prompts you to restart the instance. We recommend that you restart the instance during off-peak hours and make sure that your applications are configured with automatic reconnection policies.
> - Before the new parameter values are applied, you can click **Cancel Changes** to cancel them.

Modify a single parameter.

i. On the **Editable Parameters** tab, find the parameter that you want to modify and click the icon in the **Actual Value** column.
ii. Enter a new value based on the prompted value range.
iii. Click **Confirm**.
iv. In the upper-right corner of the page, click **Apply Changes**.

> **Note**
> - If the new parameter value takes effect only after an instance restart, the system prompts you to restart the instance. We recommend that you restart the instance during off-peak hours and make sure that your applications are configured with automatic reconnection policies.
> - Before the new parameter value is applied, you can click **Cancel Changes** to cancel it.

View the parameter modification history

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Parameters**.
5. On the page that appears, click the **Edit History** tab.
6. Select a time range and click **Search**.

11.1.5.10. Read-only instances

11.1.5.10.1. Overview of read-only ApsaraDB RDS for PostgreSQL instances

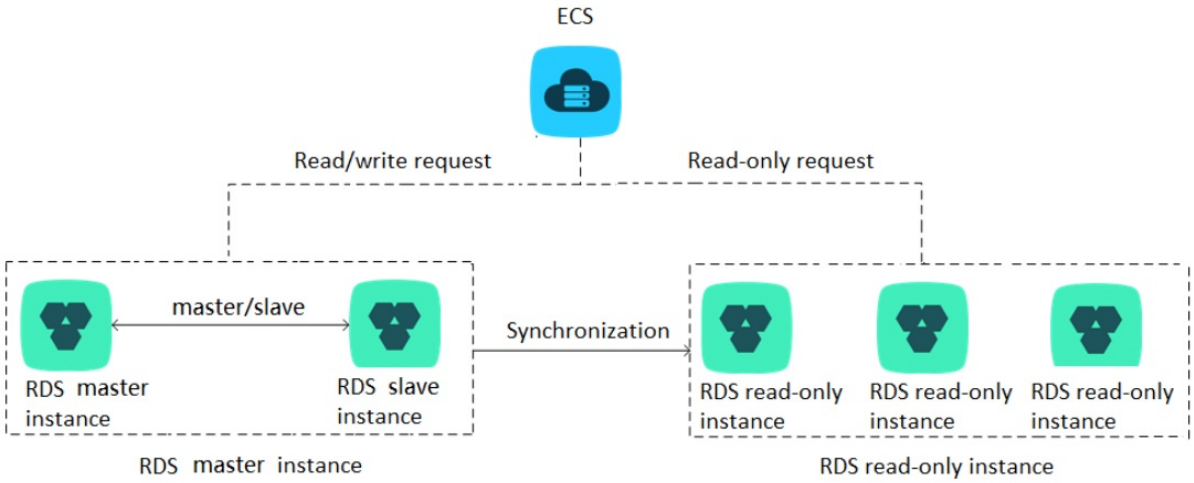
This topic provides an overview of read-only ApsaraDB RDS for PostgreSQL instances. If a large number of read requests overwhelm the primary instance, your business may be interrupted. In this case, you can create one or more read-only instances to offload read requests from the primary instance. This scales the read capability of your database system and increases the throughput of your application.

Overview

When a read-only instance is created, the data is replicated from the secondary instance. The data is consistent with that of the primary instance. Data updates of the primary instance are automatically synchronized to all read-only instances immediately after the primary instance completes operations.

> **Note**
> - Only ApsaraDB RDS instances that run PostgreSQL 10.0 support read-only instances.
> - The specifications of the primary instance must have at least eight CPU cores and 32 GB of memory.
> - Each read-only instance works in a single-node architecture, where no instances are provided as backups.

The following figure shows the topology of read-only instances.



Features

- Region and zone: Read-only instances reside within the same region as the primary instance, but possibly in different zones.
- Specifications and storage space: The specifications and storage space of read-only instances cannot be lower than those of the primary instance.
- Network type: The network type of a read-only instance can differ from that of the primary instance.
- Account and database management: Read-only instances do not require database or account maintenance, because their database and account information is synchronized with the primary instance.
- IP address whitelist: A read-only instance automatically replicates the IP address whitelists of the primary instance. However, the IP address whitelists for the read-only instance are independent of those of the primary instance. For information about how to modify the IP address whitelists of a read-only instance, see Configure an IP address whitelist.
- Monitoring and alerts: You can monitor system performance metrics, such as the disk capacity, IOPS, number of connections, and CPU utilization.

Limits

- Number of read-only instances: A maximum of five read-only instances can be created on a primary instance.
- Instance backup: Read-only instances do not support backup settings or manual backups because backups have been configured or created on the primary instance.
- Data migration: You cannot migrate data to read-only instances.
- Database management: You cannot create or delete databases on read-only instances.
- Account management: You cannot create or delete accounts, authorize accounts, or change the passwords of accounts on read-only instances.

FAQ

Q: Can I manage accounts created on the primary instance from its read-only instances?

A: No, although accounts created on the primary instance are replicated to its read-only instances, you cannot manage the accounts on the read-only instances. The accounts have only read permissions on the read-only instances.

11.1.5.10.2. Create a read-only ApsaraDB RDS for PostgreSQL instance

This topic describes how to create a read-only instance for your primary ApsaraDB RDS for PostgreSQL instance. This allows your database system to process a large number of read requests and increases the throughput of your application. The data on each read-only instance is a copy of that of the primary instance. Data updates to the primary instance are synchronized to all of its read-only instances.

Prerequisites

- The primary instance runs PostgreSQL 10.0.
- The specifications of the primary instance must have at least eight CPU cores and 32 GB of memory.

Precautions

- You can create read-only instances only for your primary instance. You cannot change existing instances to read-only instances.
- When you create a read-only instance, the system replicates data from a secondary instance. Therefore, operations on your primary instance are not interrupted.
- A read-only instance does not inherit the parameter settings of the primary instance. The system generates default parameter settings for the read-only instance, and you can modify the settings in the ApsaraDB RDS console.
- The instance type and storage capacity of a read-only instance cannot be lower than that of the primary instance.
- You can create up to five read-only instances.

Create a read-only instance

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the Distributed by Instance Role section of the Basic Information page, click **Create Read-only Instance**.
5. On the Create Read-only Instance page, configure parameters and click **Submit**. The following table describes the parameters.

Section	Parameter	Description
Region	Region	The region where the instance is deployed.
	Database Engine	The database engine of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	Engine Version	The engine version of the read-only instance, which is the same as that of the primary instance and cannot be changed.

Section	Parameter	Description
Specifications	Edition	The edition of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	Instance Type	The instance type of the read-only instance. The instance type of the read-only instance can be different from that of the primary instance, and can be changed at any time to facilitate flexible upgrade and downgrade. To ensure sufficient I/O throughput for data synchronization, we recommend that you select at least the same instance type as the primary instance for the read-only instance.
	Storage Capacity	The storage capacity of the read-only instance. To ensure sufficient I/O throughput for data synchronization, we recommend that you select at least the same storage capacity as the primary instance for the read-only instance.
Network Type	Network Type	The network type of the read-only instance, which is the same as that of the primary instance and cannot be changed.
	VPC	Select a VPC if the network type is set to VPC.
	vSwitch	Select a vSwitch if the network type is set to VPC.

11.1.5.10.3. View a read-only ApsaraDB RDS for PostgreSQL instance

This topic describes how to view details of a read-only ApsaraDB RDS for PostgreSQL instance. You can go to the Basic Information page of a read-only instance from the Instances page or the read-only instance list of the primary instance. Read-only instances are managed in the same manner as primary instances. The Basic Information page shows the management operations that can be performed.

View instance details of a read-only instance by using its ID

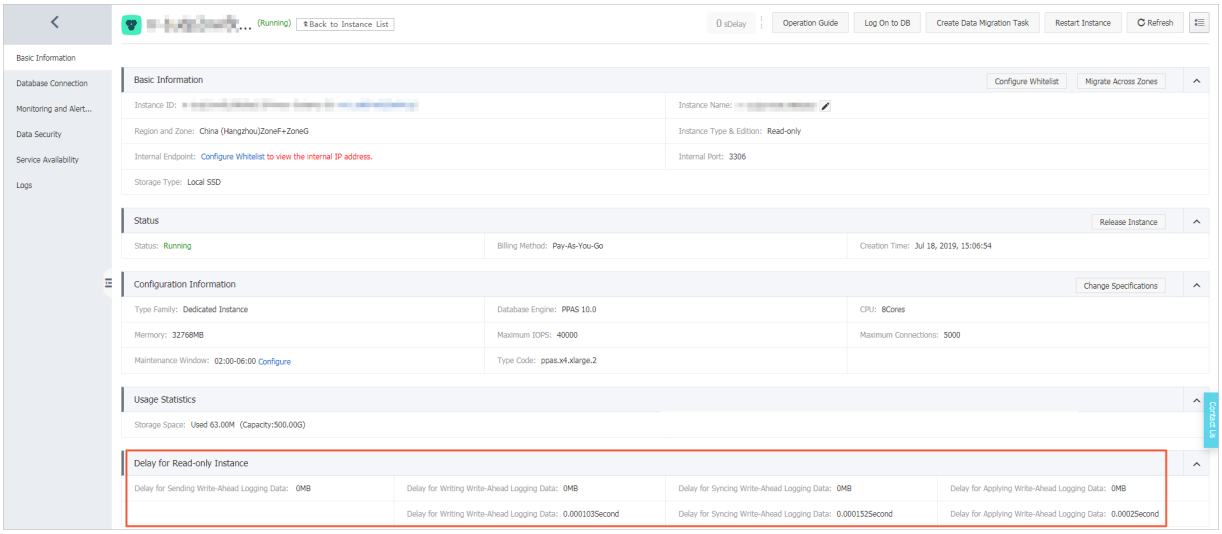
1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the instance that you want to view.
3. Click the ID of the instance or click **Manage** in the corresponding Actions column to go to the Basic Information page.

View details of a read-only instance by using the primary instance

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. On the **Basic Information** page, move the pointer over the number below **Read-only Instance** in the **Distributed by Instance Role** section. The ID of the read-only instance is displayed.
5. Click the ID of the read-only instance to go to the Basic Information page of the read-only instance.

View the latency of a read-only instance

When a read-only instance synchronizes data from its primary RDS instance, latency may occur. You can navigate to the Basic Information page of the read-only instance to view the latency of data synchronization to the instance.



11.1.6. Database connection

11.1.6.1. Connect to an ApsaraDB RDS for PostgreSQL instance

This topic describes how to use Data Management (DMS) or the pgAdmin 4 client to connect to an ApsaraDB RDS instance.

Context

You can log on to DMS from the ApsaraDB RDS console and then connect to an ApsaraDB RDS instance.

DMS is a data management service that integrates data, schema, and server management, access security, BI charts, data trends, data tracking, and performance optimization. DMS can be used to manage relational and non-relational databases, such as MySQL, SQL Server, PostgreSQL, MongoDB, and Redis. It can also be used to manage Linux servers.

You can also use a client to connect to an ApsaraDB RDS instance. ApsaraDB RDS for PostgreSQL is fully compatible with PostgreSQL. You can connect to an ApsaraDB RDS for PostgreSQL instance in a similar manner as you would connect to an open source PostgreSQL instance. In this topic, the pgAdmin 4 client is used to connect to an ApsaraDB RDS instance.

Use DMS to connect to an ApsaraDB RDS instance

For more information about how to use DMS to connect to an ApsaraDB RDS instance, see Log on to an ApsaraDB for RDS instance by using DMS.

Use the pgAdmin 4 client to connect to an ApsaraDB RDS instance

1. Add the IP address of the pgAdmin client to an IP address whitelist of the ApsaraDB RDS instance. For more information about how to configure a whitelist, see Configure an IP address whitelist.
2. Start the pgAdmin 4 client.

   > **Note** For information about how to download the pgAdmin 4 client, visit pgAdmin 4 (Windows).

3. Right-click **Servers** and choose **Create > Server**, as shown in the following figure.
4. On the **General** tab of the **Create - Server** dialog box, enter the name of the server, as shown in the following figure.

5. Click the **Connection** tab and enter the information of the instance, as shown in the following figure.

Parameter	Description
Host name/address	The internal endpoint of the ApsaraDB RDS instance. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
Port	The internal port number that is used to connect to the ApsaraDB RDS instance. For more information about how to view the internal port number, see View and modify the internal endpoint and port number.
Username	The name of the privileged account on the ApsaraDB RDS instance. For more information about how to obtain a privileged account, see Create a database and an account.
Password	The password of the privileged account of the ApsaraDB RDS instance.

6. Click **Save**.
7. If the connection information is correct, choose **Servers > Server Name > Databases > postgres**. If the following page appears, the connection is established.

> **Notice** The postgres database is the default system database of the ApsaraDB RDS instance. Do not perform operations on this database.

11.1.6.2. Use DMS to log on to an ApsaraDB RDS instance

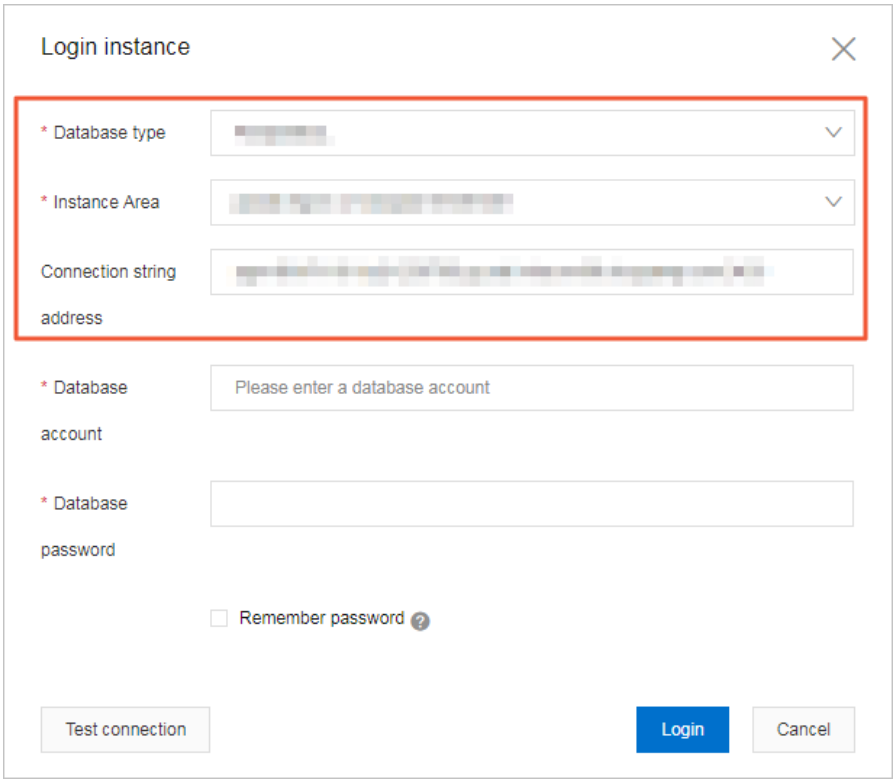
This topic describes how to use Data Management (DMS) to log on to an ApsaraDB RDS instance.

Prerequisites

The IP address whitelist is configured. For more information about how to configure an IP address whitelist, see Configure an IP address whitelist.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**.

> **Note** If you want the browser to remember the password, select **Remember password** before you click **Login**.

11.1.6.3. View and modify the internal endpoint and port number

You must use the internal endpoint and port number to access an ApsaraDB RDS instance. This topic describes how to view and modify the internal endpoint and port number of an ApsaraDB RDS instance in the ApsaraDB RDS console.

View the internal endpoint and port number

1. Log on to the ApsaraDB RDS console.

2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the **Basic Information** section, view the internal endpoint and port number of the instance.

Modify the internal endpoint and port number

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Database Connection**.
5. On the right side of the page, click **Change Endpoint**.
6. In the dialog box that appears, set **Connection Type** to **Internal Endpoint**.
7. Modify the endpoint prefix and port number and click **OK**.

FAQ

- Q: Do I need to modify the endpoint or port number in my application after I modify the endpoint or port number of an instance?

  A: Yes, you must modify the endpoint or port number in the application after you modify them. Otherwise, the application cannot connect to databases of the instance.
- Q: Does the modification of the endpoint take effect immediately? Do I need to restart the instance?

  A: No, you do not need to restart the instance. The modification takes effect immediately.

11.1.7. Accounts

11.1.7.1. Create an account

Before you start to use ApsaraDB RDS, you must create an account on an ApsaraDB RDS instance. This topic describes how to create an account on an ApsaraDB RDS for PostgreSQL instance.

Account types

ApsaraDB RDS for PostgreSQL instances support two types of accounts: privileged accounts and standard accounts. The following table describes these account types.

Account type	Description
Privileged account	• You can create and manage privileged accounts only by using the ApsaraDB RDS console or the API. • If your ApsaraDB RDS instance uses local SSDs, you can create only a single privileged account. If your ApsaraDB RDS instance uses standard or enhanced SSDs, you can create more than one privileged account. A privileged account allows you to manage all the standard accounts and databases that are created on your ApsaraDB RDS instance. • A privileged account has more permissions that allow you to manage your ApsaraDB RDS instance at more fine-grained levels. For example, you can grant the query permissions on different tables to different users. • A privileged account has the permissions to disconnect accounts that are created on your ApsaraDB RDS instance.

Account type	Description
Standard account	• You can create and manage standard accounts by using the ApsaraDB RDS console, API operations, or SQL statements. • You can create more than one standard account on your ApsaraDB RDS instance. • You must grant the permissions on specific databases to a standard account. • A standard account does not have the permissions to create, manage, or disconnect other accounts on your ApsaraDB RDS instance.

Precautions

- If your ApsaraDB RDS instance uses local SSDs, you can create one privileged account in the ApsaraDB RDS console. After the privileged account is created, it cannot be deleted. You can also create and manage more than one standard account by using SQL statements.
- If your ApsaraDB RDS instance uses standard or enhanced SSDs, you can create more than one privileged account and standard account in the ApsaraDB RDS console. You can also create and manage more than one standard account by using SQL statements.
- To migrate data from an on-premises database to your ApsaraDB RDS instance, you must create a database and an account on the ApsaraDB RDS instance. Make sure that the created database has the same properties as the on-premises database. Also make sure that the created account has the same permissions on the created database as the account that is authorized to manage the on-premises database.
- Follow the least privilege principle to create accounts and grant them read-only permissions or read and write permissions on databases. If necessary, you can create more than one account and grant them only the permissions on specific databases. If an account does not need to write data to a database, grant only the read-only permissions on that database to the account.
- For security purposes, we recommend that you specify strong passwords for the accounts on your ApsaraDB RDS instance and change the passwords on a regular basis.

Create a privileged account on an instance that uses local SSDs

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. On the Accounts page, click **Create Privileged Account** and configure the following parameters.

Parameter	Description
Database Account	◦ The name of the account must be 2 to 16 characters in length. ◦ The name can contain lowercase letters, digits, and underscores (_). ◦ The name must start with a letter and end with a letter or digit.
Password	◦ The password of the account must be 8 to 32 characters in length. ◦ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.characters. ◦ Special characters include ! @#$%^&*()_+-=
Re-enter Password	Enter the password of the account again.

6. Click **Create**.

Create a privileged or standard account on an instance that uses standard or enhanced SSDs

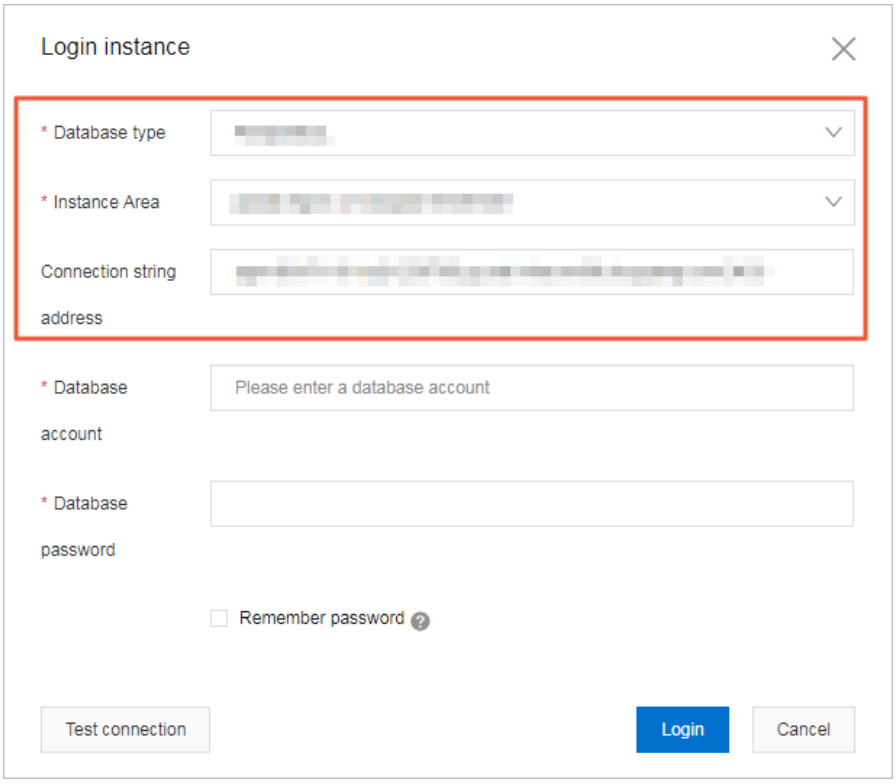
1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. On the Accounts page, click **Create Account** and configure the following parameters.

Parameter	Description
Database Account	○ The name of the account must be 2 to 16 characters in length. ○ The name can contain lowercase letters, digits, and underscores (_). ○ The name must start with a letter and end with a letter or digit.
Account Type	Select **Privileged Account** or **Standard Account**.
Password	○ The password of the account must be 8 to 32 characters in length. ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.characters. ○ Special characters include ! @#$%^&*()_+-=
Re-enter Password	Enter the password of the account again.
Description	This parameter is optional. You can enter relevant description to make the instance identifiable. The description can be up to 256 characters in length.

6. Click **Create**.

Create a standard account on an instance that uses local SSDs

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**. If you want the browser to remember the password, select **Remember password** before you click **Login**.

> **Note** If a message prompts you that the connection fails, the problem may be caused by an improperly configured whitelist. Reconfigure the whitelist in the console. For more information, see Configure an IP address whitelist.

7. The **SQLConsole** page appears after you log on to the instance. Execute a statement in the following format to create a standard account:

```
CREATE USER name [ [ WITH ] option [ ... ] ]
where option can be:
    SUPERUSER | NOSUPERUSER
  | CREATEDB | NOCREATEDB
  | CREATEROLE | NOCREATEROLE
  | CREATEUSER | NOCREATEUSER
  | INHERIT | NOINHERIT
  | LOGIN | NOLOGIN
  | REPLICATION | NOREPLICATION
  | CONNECTION LIMIT connlimit
  | [ ENCRYPTED | UNENCRYPTED ] PASSWORD 'password'
  | VALID UNTIL 'timestamp'
  | IN ROLE role_name [, ...]
  | IN GROUP role_name [, ...]
  | ROLE role_name [, ...]
  | ADMIN role_name [, ...]
  | USER role_name [, ...]
  | SYSID uid
```

For example, if you want to create a user account named test2 whose password is 123456, execute the following statement:

```
create user test2 password '123456';
```

8. Click **execute**.

11.1.7.2. Reset the password

This topic describes how to use the ApsaraDB RDS console to reset the password of your database account if you forget the password.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. In the **Actions** column corresponding to the account, click **Reset Password**.
6. In the dialog box that appears, enter a new password and click **OK**.

> **Note** The password must meet the following requirements:
>
> - The password must be 8 to 32 characters in length.
> - The password must contain at least three of the following characters: uppercase letters, lowercase letters, digits, and special characters.
> - Special characters include ! @ # $ % ^ & * ( ) _ + - =

11.1.7.3. Lock an account

You can lock a database account in the ApsaraDB RDS console to make the account unavailable.

Procedure

1. Log on to the ApsaraDB RDS console.

2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Accounts**.
5. Find the account that you want to lock and click **Lock** in the **Actions** column.
6. To unlock the account, click **Unlock** in the **Actions** column.

11.1.7.4. Delete an account

You can delete a database account in the ApsaraDB RDS console.

Prerequisites

You can use the console to delete privileged and standard accounts that are no longer used.

Procedure

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, click **Accounts**.
4. Find the account that you want to delete and click **Delete** in the Actions column.
5. In the message that appears, click **Confirm**.

> **Note** Accounts in the **Processing** state cannot be deleted.

11.1.8. Databases

11.1.8.1. Create a database

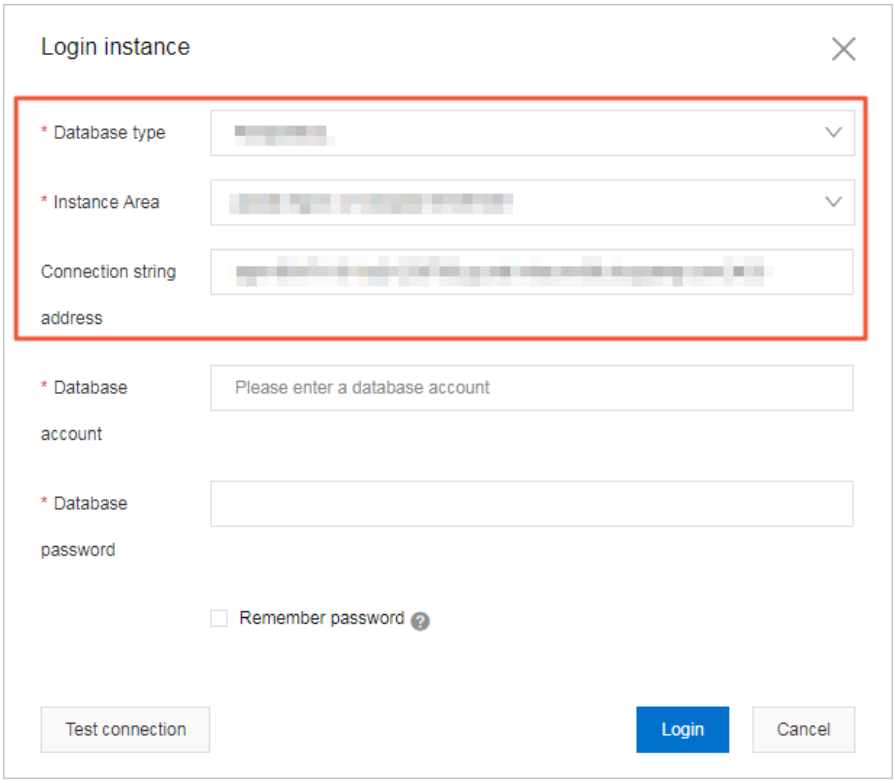
Before you start to use ApsaraDB RDS, you must create a database on an ApsaraDB RDS instance. This topic describes how to create a database on an ApsaraDB RDS for PostgreSQL instance.

Prerequisites

- An ApsaraDB RDS for PostgreSQL instance is created. For more information, see Create an instance.
- An account is created. For more information, see Create an account.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**. If you want the browser to remember the password, select **Remember password** before you click **Login**.

   > **Note** If a message prompts you that the connection fails, the problem may be caused by an improperly configured whitelist. Reconfigure the whitelist in the console. For more information, see Configure an IP address whitelist.

7. The **SQLConsole** page appears after you log on to the instance. Execute a statement in the following format to create a database:

   ```
   CREATE DATABASE name;
   ```

   For example, if you want to create a database named test, execute the following statement:

   ```
   create database test;
   ```

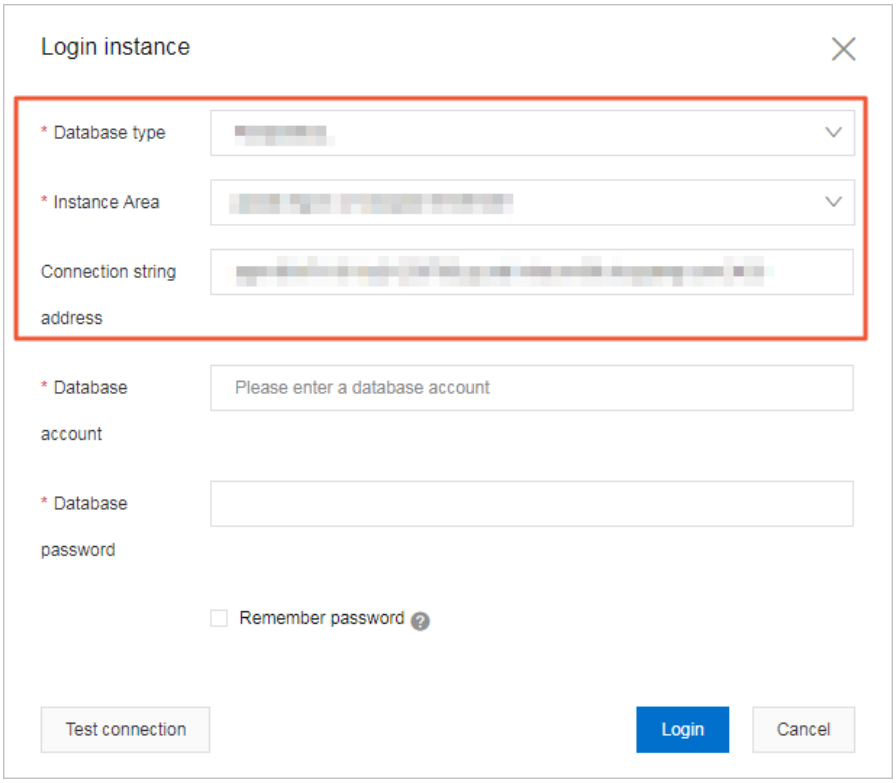
8. Click **execute**.

11.1.8.2. Delete a database

This topic describes how to delete a database in the ApsaraDB RDS for PostgreSQL console.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. Click **Log On to DB** in the upper-right corner of the page.
5. In the **Login instance** dialog box of the **DMS** console, check values of **Database type**, **Instance Area**, and **Connection string address**. If the information is correct, enter **Database account** and **Database password**, as shown in the following figure.



Parameter	Description
Database type	The engine of the database. By default, the engine of the database to be connected is displayed.
Instance Area	The region where the instance is deployed. By default, the region of the current instance is displayed.
Connection string address	The endpoint of the instance. By default, the endpoint of the current instance is displayed.
Database account	The account of the database to be connected.
Database password	The password of the account used to connect to the database.

6. Click **Login**. If you want the browser to remember the password, select **Remember password** before you click

Login.

> **Note** If a message prompts you that the connection fails, the problem may be caused by an improperly configured whitelist. Reconfigure the whitelist in the console. For more information, see Configure an IP address whitelist.

7. The **SQLConsole** page appears after you log on to the instance. Execute the following statement to delete a database:

```
drop database <database name>;
```

For example, if you want to delete a database named test2, execute the following statement:

```
drop database test2;
```

8. Click **execute**.

11.1.9. Networks, VPCs, and vSwitches

11.1.9.1. Change the VPC and vSwitch for an ApsaraDB RDS for PostgreSQL instance

This topic describes how to change the virtual private cloud (VPC) and vSwitch for an ApsaraDB RDS for PostgreSQL instance.

Prerequisites

The ApsaraDB RDS for PostgreSQL instance resides in a VPC.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Database Connection**.
5. In the upper-right corner of the Database Connection section, click **Switch vSwitch**.
6. Select a VPC and a vSwitch, and then click **OK**.

> **Note** If you want to create a VPC and a vSwitch, you can click **go to the VPC console**.

7. In the message that appears, click **Switch**.

> **Note**
> - You may encounter a network interruption of about 30 seconds during the change process. Make sure that your application is configured to automatically reconnect to the instance.
> - We recommend that you clear the cache immediately after the VPC and vSwitch are changed. Otherwise, data can be read but not written.

11.1.9.2. Change the network type of an ApsaraDB RDS for PostgreSQL instance

This topic describes how to change the network type of an ApsaraDB RDS for PostgreSQL instance between classic network and Virtual Private Cloud (VPC).

Prerequisites

The ApsaraDB RDS instance uses local SSDs.

Context

- Classic network: ApsaraDB RDS instances in the classic network are not isolated. You can block unauthorized access only by configuring IP address whitelists on these instances.
- VPC: Each VPC is an isolated network. We recommend that you use the VPC network type because it provides a higher security level.

  You can configure route tables, CIDR blocks, and gateways within a VPC. To smoothly migrate applications to the cloud, you can use the leased line or VPN method to create a virtual data center that consists of your self-managed data center and a VPC.

Change the network type from VPC to classic network

Precautions

- The ApsaraDB RDS instance must be in a VPC.
- After you change the network type from VPC to classic network, the internal endpoint of the ApsaraDB RDS instance remains unchanged. However, the IP address that is associated with the internal endpoint changes.
- After you change the network type from VPC to classic network, you cannot connect Elastic Compute Service (ECS) instances deployed in VPCs to the ApsaraDB RDS instance by using the internal endpoint. You must update the endpoint for the applications deployed on the ECS instances.
- You may encounter a network interruption of about 30 seconds during the change process. To avoid business interruptions, we recommend that you change the network type during off-peak hours or make sure that your application is configured to automatically reconnect to the instance.

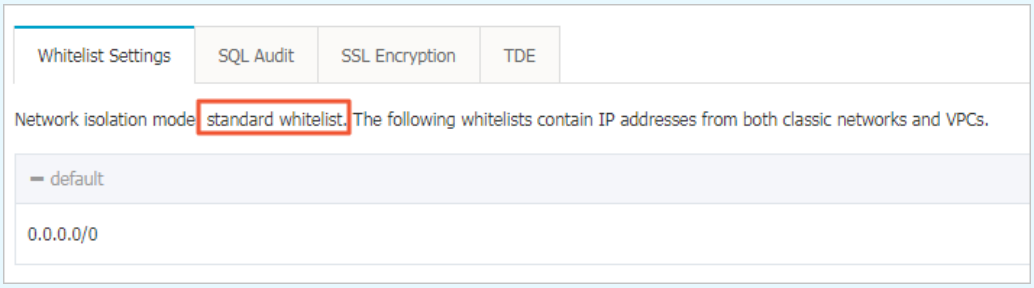
1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Database Connection**.
5. In the upper-right corner of the Database Connection section, click **Switch to Classic Network**.
6. In the message that appears, click **OK**.

   > **Note** After the network type is changed to classic network, only ECS instances within the classic network can connect to the ApsaraDB RDS instance by using the internal endpoint. You must configure the internal endpoint for the ECS instances.

7. Configure an IP address whitelist to allow ECS instances within the classic network to connect to the ApsaraDB RDS instance by using the internal endpoint.

Note

- If the network isolation mode of the ApsaraDB RDS instance is standard whitelist, add the internal IP addresses of the ECS instances to a whitelist of your ApsaraDB RDS instance.



- If the network isolation mode of the ApsaraDB RDS instance is enhanced whitelist, add the internal IP addresses of the ECS instances to a whitelist of the classic network type. If no whitelists of the classic network type are available, create a whitelist. For more information about the enhanced whitelist mode, see Switch to the enhanced whitelist mode.

Change the network type from classic network to VPC

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Database Connection**.
5. In the upper-right corner of the Database Connection section, click **Switch to VPC**.
6. In the Switch to VPC dialog box, select a VPC and a vSwitch and specify whether to retain the classic network endpoint.

> **Note**
>
> - Select a VPC. We recommend that you select the VPC where your ECS instances are deployed. Otherwise, the ECS instances cannot communicate with the ApsaraDB RDS instance over the internal network.
> - Select a vSwitch. If no vSwitches are available in the selected VPC, create one in the same zone where the ApsaraDB RDS instance is deployed. For more information, see Create a vSwitch in *Virtual Private Cloud User Guide*.
> - Clear or select the **Reserve Original Classic Network Endpoint** check box.
>   - **Clear the Reserve Original Classic Network Endpoint check box**
>
>   The classic network endpoint is not retained and changes to a VPC endpoint.
>
>   When you change the network type from classic network to VPC, a network interruption of 30 seconds occurs. In this case, ECS instances located in the classic network are disconnected from your ApsaraDB RDS instance.
>   - **Select the Reserve Original Classic Network Endpoint check box**
>
>   The classic network endpoint is retained, and a new VPC endpoint is generated. In this case, your ApsaraDB RDS instance runs in hybrid access mode. Both ECS instances located in the classic network and ECS instances located in the selected VPC can access your ApsaraDB RDS instance over an internal network. You must set **Expiration Time (Important)** to **14 Days Later**, **30 Days Later**, **60 Days Later**, or **120 Days Later** for the classic network. You can also modify the expiration date after the network type is changed. For more information, see Configure hybrid access from both the classic network and VPCs.
>
>   When you change the network type from classic network to VPC, no network interruptions occur. ECS instances located in the classic network are still connected with your ApsaraDB RDS instance until the classic network endpoint expires.
>
>   Before the classic network endpoint expires, you must add the new VPC endpoint to your applications that run on the ECS instances located in the selected VPC. This allows ApsaraDB RDS to migrate your workloads to the selected VPC without network interruptions. Seven days before the classic network endpoint expires, the system sends a text message to the phone number bound to your Apsara Stack tenant account every day.
>
>   For more information, see Hybrid access from both the classic network and VPCs.

7. Add the internal IP addresses of ECS instances located in the selected VPC to an IP address whitelist of the VPC network type. This allows the ECS instances to connect to your ApsaraDB RDS instance over an internal network. If no IP address whitelists of the VPC network type are available, create one.

> **Note**
>
> - If you retain the classic network endpoint, add the VPC endpoint to the ECS instances before the classic network endpoint expires.
> - If you do not retain the classic network endpoint, connections between ECS instances in the classic network and the ApsaraDB RDS instance over the internal network are interrupted. You must add the VPC endpoint to ECS instances in the VPC immediately after the network type is changed.

11.1.9.3. Configure hybrid access from both the classic network and VPCs

This topic describes how to use the hybrid access solution of ApsaraDB RDS to change the network type of an instance from classic network to Virtual Private Cloud (VPC) without network interruptions.

Prerequisites

- The ApsaraDB RDS instance uses local SSDs.
- The ApsaraDB RDS instance is deployed in the classic network.
- Available VPCs and vSwitches exist in the zone where the ApsaraDB RDS instance is deployed.

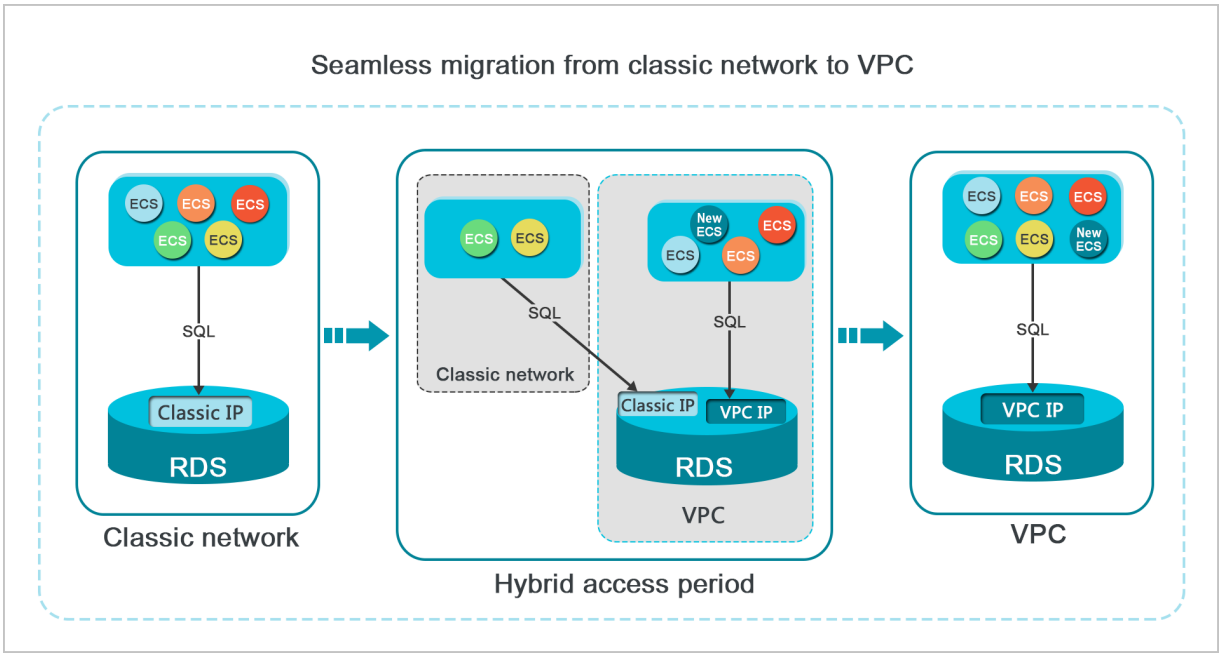
Context

In the past, when you change the network type of an ApsaraDB RDS instance from classic network to VPC, the internal endpoint of the ApsaraDB RDS instance would remain the same but the IP address bound to the endpoint would change to the corresponding IP address in the VPC. This change would cause a 30-second network interruption, and ECS instances within the classic network would not be able to access the ApsaraDB RDS instance by using the internal endpoint within this period. To smoothly change the network type, ApsaraDB RDS provides the hybrid access solution.

Hybrid access refers to the ability of an ApsaraDB RDS instance to be accessed by ECS instances in both the classic network and VPCs. During the hybrid access period, the ApsaraDB RDS instance reserves the original internal endpoint of the classic network and adds the internal endpoint of VPCs. This prevents network interruptions during the network type switchover.

For better security and performance, we recommend that you use the internal endpoint of VPCs. Hybrid access is available for a limited period of time. The internal endpoint of the classic network is released when the hybrid access period expires. In that case, your applications cannot access the ApsaraDB RDS database by using the internal endpoint of the classic network. You must configure the internal endpoint of VPCs in all your applications during the hybrid access period. This ensures smooth network switchover and minimizes the impact on your services.

For example, your company wants to use the hybrid access solution to change the network type from classic network to VPC. During the hybrid access period, some applications can access the database by using the internal endpoint of VPCs, and the other applications can access the database by using the original internal endpoint of the classic network. When all the applications access the database by using the internal endpoint of VPCs, the internal endpoint of the classic network can be released. The following figure illustrates the scenario.



Limits

During the hybrid access period, the instance has the following limits:

- The network type of your instance cannot be changed to classic network.
- Your instance cannot be migrated to another zone.

Change the network type from classic network to VPC

For more information, see Change the network type from classic network to VPC.

Change the expiration time for the original internal endpoint of the classic network

During the period in which your instance can be accessed over the classic network or VPCs, you can specify the expiration time for the endpoint of the classic network. The setting takes effect immediately. For example, if the endpoint of the classic network is about to expire on August 18, 2017 and you change the expiration time to 14 days later on August 15, 2017, the endpoint of the classic network is released on August 29, 2017.

To change the expiration time, perform the following steps:

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Database Connection**.
5. On the **Instance Connection** tab, click **Change Expiration Time**.
6. In the **Change Expiration Time** dialog box, select an expiration time and click **OK**.

11.1.10. Monitoring

11.1.10.1. View monitored resources

ApsaraDB RDS provides a wide range of performance metrics. This topic describes how to view resource monitoring data in the ApsaraDB RDS console.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Monitoring and Alerts**.
5. On the **Monitoring** tab, select a time range to query the corresponding monitoring data. The following table lists the specific metrics.

Metric	Description
CPU Utilization	The CPU utilization of the instance. Unit: %.
Memory Usage	The memory usage of the instance. Unit: %.
IOPS	The number of input and output operations that are performed per second.
Disk Space Used	The used disk space of the instance. Unit: MB.
Received Traffic	The inbound and outbound bandwidths of the instance.
Data Disk Usage	The data disk usage of the instance. Unit: %.

> **Note** You can click **Refresh** in the upper-right corner of the **Monitoring** tab to refresh the monitoring information.

11.1.11. Data security

11.1.11.1. Switch to the enhanced whitelist mode

This topic describes how to switch from the standard whitelist mode to the enhanced whitelist mode for an ApsaraDB RDS instance. The enhanced whitelist mode provides higher security.

Network isolation modes

ApsaraDB RDS instances support the following network isolation modes:

- Standard whitelist mode

  IP addresses from both the classic network and virtual private clouds (VPCs) are added to the same IP address whitelist. The standard whitelist mode may incur security risks. Therefore, we recommend that you switch the network isolation mode to enhanced whitelist.

- Enhanced whitelist mode

  An enhanced IP address whitelist can contain only the IP addresses from the classic network or VPCs. When you create an enhanced IP address whitelist, you must specify its network type.

Changes after you switch to the enhanced whitelist mode

- If your ApsaraDB RDS instance resides in a VPC, an IP address whitelist of the VPC network type is automatically created. The new IP address whitelist contains all IP addresses that are replicated from the original IP address whitelists.
- If your ApsaraDB RDS instance resides in the classic network, an IP address whitelist of the classic network type is automatically created. The new IP address whitelist contains all IP addresses that are replicated from the original IP address whitelists.
- If your ApsaraDB RDS instance runs in hybrid access mode, two identical IP address whitelists are created: an IP address whitelist of the VPC network type and an IP address whitelist of the classic network type. Both the new IP address whitelists contain all IP addresses that are replicated from the original IP address whitelists. For more information, see Hybrid access from both the classic network and VPCs.

> **Note** After you switch to the enhanced whitelist mode, the IP addresses that come from Elastic Compute Service (ECS) security groups remain unchanged.

Precautions

- You can switch from the standard whitelist mode to the enhanced whitelist mode for ApsaraDB RDS instances that use local SSDs, but not the other way around.
- In enhanced whitelist mode, an IP address whitelist of the classic network type can also be used to allow access over the Internet. If you want to access your ApsaraDB RDS instance from a host over the Internet, you must add the public IP address of the host to an IP address whitelist of the classic network type.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.

5. On the **Whitelist Settings** tab, click **Switch to Enhanced Whitelist (Recommended)**.
6. In the message that appears, click **Confirm**.

11.1.11.2. Configure an IP address whitelist

This topic describes how to configure a whitelist for an ApsaraDB RDS instance. Only entities that are listed in a whitelist can access your ApsaraDB RDS instance.

Context

Whitelists make your ApsaraDB RDS instance more secure and do not interrupt the operations of your ApsaraDB RDS instance when you configure whitelists. We recommend that you perform maintenance on your whitelists on a regular basis.

To configure a whitelist, perform the following operations:

- Configure a whitelist: Add IP addresses to allow them to connect to the ApsaraDB RDS instance.

  > **Note** The IP address whitelist labeled default contains only the default IP address 0.0.0.0/0, which allows all entities to access your ApsaraDB RDS instance.

- Configure an ECS security group: Add an ECS security group for the ApsaraDB RDS instance to allow ECS instances in the group to connect to the ApsaraDB RDS instance.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. On the **Whitelist Settings** tab, click **Edit** corresponding to the **default** whitelist.

   > **Note** You can also click **Create Whitelist** to create a whitelist.

6. In the **Edit Whitelist** dialog box, enter the IP addresses or CIDR blocks used to access the instance and click **OK**. The following section describes the rules:
   - If you enter the CIDR block 10.10.10.0/24 in the IP Addresses field, all IP addresses in the 10.10.10.X format can access your ApsaraDB RDS instance.
   - If you enter more than one IP address or CIDR block, you must separate them with commas (,). Do not add spaces before or after the commas. Example: 192.168.0.1,172.16.213.9.
   - If you click **Add Internal IP Addresses of ECS Instances**, the IP addresses of all ECS instances created within your Alibaba Cloud account are displayed. You can select the required IP addresses to add them to the IP address whitelist.

11.1.11.3. Configure SSL encryption

This topic describes how to configure SSL encryption for an ApsaraDB RDS instance.

Prerequisites

The ApsaraDB RDS instance uses standard SSDs.

Precautions

- After SSL encryption is enabled, SSL is used to encrypt all the data that is transmitted over an internal network

or the Internet. SSL encryption protects the data in transit from being leaked.

- After SSL encryption is enabled, you must close the existing connection and establish a new one to bring SSL encryption into effect.

Enable SSL encryption

SSL 3.0 has been upgraded by the Internet Engineering Task Force (IETF) to Transport Layer Security (TLS), but the term SSL encryption is still commonly used in the communications industry. Therefore, SSL encryption is used in this topic to refer to TLS encryption.

> **Note** ApsaraDB RDS supports TLS 1.0, 1.1, and 1.2.

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Parameters**.
5. On the **Editable Parameters** tab, find the ssl parameter and click the icon in the **Actual Value** column.
6. Change the value of ssl to **ON** and wait for the system to enable SSL encryption.

> **Note** After SSL encryption is enabled, you must set the SSL mode to **Prefer** when you log on from your client.

Disable SSL encryption

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Parameters**.
5. On the **Editable Parameters** tab, find the ssl parameter and click the icon in the **Actual Value** column.
6. Change the value of ssl to **OFF** and wait for the system to disable SSL encryption.

11.1.11.4. Configure data encryption

This topic describes how to configure data encryption for an ApsaraDB RDS instance that uses standard or enhanced SSDs. The disk encryption feature maximizes the protection for your data and eliminates the need to modify business or application configurations. ApsaraDB RDS automatically applies disk encryption to both the snapshots that are generated from the encrypted SSDs and the SSDs that are created from those snapshots.

Prerequisites

The storage type of the instance is standard SSD.

Configure disk encryption

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

> **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
>
> - Uppercase or lowercase letters
> - Digits
> - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. In the top navigation bar, choose **Products** > Security > Key Management Service.
5. On the Keys page, click **Create Key**.
6. Configure the following parameters.

Section	Parameter	Description
Region	**Organization**	The organization to which the key belongs.
	Resource Set	The resource set to which the key belongs.
	Region	The region to which the key belongs.
Basic Settings	**Key Type**	KMS supports the following key types: ○ Symmetric keys: ■ Aliyun_AES_256 ■ Aliyun_SM4 ○ Asymmetric keys: ■ RSA_2048 ■ EC_P256 ■ EC_P256K ■ EC_SM2
	Key Purpose	ENCRYPT/DECRYPT: The purpose of the CMK is to encrypt or decrypt data.
	Protection Level	○ SOFTWARE: Use a software module to protect the CMK. ○ HSM: Host the CMK in a hardware security module (HSM). Managed HSM uses the HSM as dedicated hardware to safeguard the CMK.
	Alias	The identifier of the CMK. For more information, see Use aliases in *KMS User Guide*.
	Description	The description of the CMK.

Section	Parameter	Description
Advanced Settings	Rotation Period	Specifies whether to enable automatic rotation. If you choose to enable automatic rotation, you must select a rotation period. For more information about rotation, see Key rotation in *KMS User Guide*. Valid values: ○ 30 Days ○ 90 Days ○ 180 Days ○ 365 Days ○ Custom: Customize a period that ranges from 7 to 730 days. **Note** You can specify this parameter only if Key Type is set to Aliyun_AES_256 or Aliyun_SM4.
	Key Material Source	The source of key material. ○ Key Management Service: Use KMS to generate key material. ○ External: Manually import external key material. **Note** If **Rotation Period** is set to **Enable**, the **External** option is unavailable.

7. Click **Submit**.
8. Create an ApsaraDB RDS instance with disk encryption enabled. For more information, see Create an ApsaraDB RDS for PostgreSQL instance that uses standard or enhanced SSDs.

11.1.12. Logs and audit

11.1.12.1. Configure SQL audit

This topic describes how to configure the SQL audit feature to audit SQL executions and check the details. SQL audit does not affect instance performance.

Precautions

- SQL audit does not affect instance performance.
- SQL audit logs are retained for 30 days.
- Log files exported from SQL audit are retained for two days. The system deletes files that are retained for longer than two days.
- SQL audit is disabled by default. You must manually enable it.
- You cannot view logs that are generated before SQL audit is enabled.

Enable SQL audit

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. Click the **SQL Audit** tab.

6. Click **Enable SQL Audit** or **Enable now**.
7. In the message that appears, click **Confirm**.

> **Note** After SQL audit is enabled, you can query SQL information based on conditions such as the time, database, user, and keyword.

Disable SQL audit

You can disable SQL audit when it is no longer needed. To disable SQL audit, perform the following steps:

> **Notice** After SQL audit is disabled, all SQL audit logs are deleted. We recommend that you export and store audit logs to your computer before you disable SQL audit.

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Data Security**.
5. Click the **SQL Audit** tab. Click **Export File**.

> **Note** If more than 1 million SQL audit logs meet the filter conditions you specify, only 1 million logs can be exported. SQL audit logs are exported at a speed of 900 entries per second. It takes about 20 minutes to export 1 million SQL audit logs.

6. Click **Files**. Find a file and click Download in the **Action** column to download the file to your computer.
7. Click **Disable SQL Audit**.
8. In the message that appears, click **Confirm**.

11.1.12.2. Manage logs

You can view logs for errors, slow queries, and primary/secondary instance switching for ApsaraDB RDS for PostgreSQL instances in the ApsaraDB RDS console or by executing SQL statements. These logs help you troubleshoot errors. This topic describes how to manage logs in the console.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Logs**.
5. On the **Logs** page, click the **Error Logs**, **Slow Query Logs**, or **Primary/Secondary Switching Logs** tab, select a time range, and then click **Search**.

Tab	Description
Error Logs	Records database running errors that occurred within the last month.
Slow Query Logs	Records SQL statements within the last month that took longer than one second to execute. Duplicated SQL statements are removed.

Tab	Description
Primary/Secondary Switching Logs	Records switchovers between the primary and secondary instances within the last month.

11.1.13. Backup

11.1.13.1. Back up an ApsaraDB RDS for PostgreSQL instance

This topic describes how to back up an ApsaraDB RDS for PostgreSQL instance. You can configure a backup policy that is used to automatically back up your ApsaraDB RDS instance. If you do not configure a backup policy, the default backup policy is used. You can also manually back up your ApsaraDB RDS instance.

Precautions

- Do not execute data definition language (DDL) statements during a backup. These statements trigger locks on tables, and the backup may fail as a result of the locks.
- We recommend that you back up your ApsaraDB RDS instance during off-peak hours.
- If your ApsaraDB RDS instance has a large amount of data, a backup may require a long period of time.
- Backup files are retained for a specific retention period. Before the specific retention period elapses, we recommend that you download the required backup files to your computer.

Backup description

ApsaraDB RDS for PostgreSQL allows you to perform full physical backup and back up archived redo log files of databases.

Configure a backup policy for automatic backups

ApsaraDB RDS can automatically back up your instance based on the backup policy that you specify.

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Backup and Restoration**.
5. On the **Backup and Restoration** page, click the **Backup Settings** tab and click **Edit**.
6. In the dialog box that appears, configure the following parameters and click **OK**. The following table lists the parameters.

Parameter	Description
Data Retention Period	The number of days for which you want to retain data backup files. Valid values: 7 to 730. Unit: days. Default value: 7.
Backup Cycle	The cycle to create backups. You can select one or more days of the week. **Note** To ensure data security, we recommend that you back up your ApsaraDB RDS instance at least twice a week.
Backup Time	The period of time for which you want to back up data. Unit: hours.

Parameter	Description
Log Backup	Specifies whether to enable the log backup feature. **Notice** If you disable this feature, all log backup files are deleted and your instance cannot be restored to previous points in time.
Log Retention Period	○ The period of time for which you want to retain log backup files. Valid values: 7 to 730. Unit: days. Default value: 7. ○ The log retention period must be less than or equal to the data retention period.
OSS Dump Status	Specifies whether to enable Object Storage Service (OSS) dump. When OSS dump is enabled, new backup files are automatically dumped to a specific OSS bucket. Valid values: ○ Enabled ○ Disabled
OSS Dumped Data	The type of backup files that are dumped to an OSS bucket. You can select multiple values. Valid values: ○ Data Backup ○ Log Backup
OSS Bucket	The OSS bucket to which you want to dump backup files.

Manually back up your ApsaraDB RDS instance

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the upper-right corner of the page, click **Back Up Instance**. The **Back Up Instance** dialog box appears.
5. Select the backup mode and backup policy, and click **OK**.

Note The backup mode is **Full Backup** and the backup policy is **Instance Backup**.

What's next

You can click the icon in the upper-right corner of the page to view the task progress displayed in the **Task Progress** list.

11.1.13.2. Download data and log backup files

This topic describes how to download unencrypted data and log backup files in the ApsaraDB RDS console to archive the files and restore data to an on-premises database.

Procedure

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.

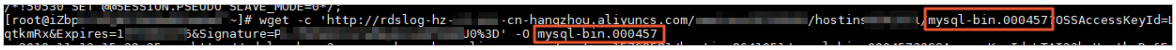
4. In the left-side navigation pane, click **Backup and Restoration** to go to the **Backup and Restoration** page.
5. Click the Data Backup or Archived Logs tab.
   - To download data backup files, click the **Data Backup** tab.
   - To download log files, click the **Archived Logs** tab.
6. Select a time range to which you want to restore the instance.
7. Find the data backup or log file that you want to download and click **Download** in the **Actions** column.

> **Note**
> - If you want to use a data backup file to restore data, select the backup file that is the closest to the time for restoration.
> - If you want to use a log file to restore data to an on-premises database, take note of the following items:
>   - The instance No. of the log file must be the same as that of the data backup file.
>   - The start time of the log file must be later than the data backup time and earlier than the time for restoration.

8. In the message that appears, select a download method.

Download method	Description
Download	Download the file by using the public endpoint.
Copy Internal Endpoint	Copy the internal endpoint to download the file. If your ECS and ApsaraDB RDS instances are deployed within the same region, you can log on to the ECS instance and use the internal endpoint to download the file. This method is fast and secure.
Copy Public Endpoint	Copy the public endpoint to download the file. If you want to use other tools to download the file, use the public endpoint.

> **Note** If you use a Linux operating system, you can run the following command to download the file:
>
> ```
> wget -c '<Public endpoint of the backup file, which is the download URL>' -O <File name>
> ```
>
> - The -c option enables resumable download.
> - The -O option saves the downloaded file by using a specified name. We recommend that you use the file name contained in the download URL.
> - If the URL contains more than one parameter, enclose the download URL in a pair of single quotation marks (').



11.1.13.3. Create a logical backup for an ApsaraDB RDS for PostgreSQL instance

This topic describes how to use pg_dump to create a logical backup for an ApsaraDB RDS for PostgreSQL instance and export the backup file to your computer.

Context

The pg_dump utility provided with PostgreSQL is used to back up individual databases. For more information, visit pg_dump.

In this example, an ApsaraDB RDS for PostgreSQL instance that runs Linux 7 and PostgreSQL 10 is used.

Prerequisites

- The IP address of your ECS instance or host is added to a whitelist of the ApsaraDB RDS for PostgreSQL instance. For more information, see Configure an IP address whitelist.
- Your ECS instance or host runs the same version of PostgreSQL as the ApsaraDB RDS for PostgreSQL instance.

Precautions

We recommend that you use the privileged account of the ApsaraDB RDS for PostgreSQL instance to ensure that you have all the required permissions.

Back up a database

1. Log on to your ECS instance or host. Then, run the following command to back up a database from the ApsaraDB RDS for PostgreSQL instance:

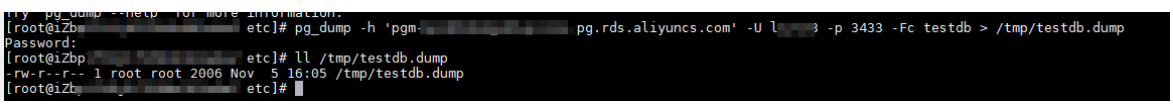
```
pg_dump -h '<hostname>' -U <username> -p <port> -Fc <dbname> > <dumpdir>
```

Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances use the VPC network type, make sure that both instances are deployed within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
-Fc	The output file format. `-Fc` specifies the custom format, which is ideal when you use pg_restore to import logical backup files and restore databases. For more information, visit pg_dump.
dbname	The name of the database that you want to back up.
dumpdir	The directory and name of the logical backup file to export.

Example:

```
pg_dump -h 'pgm-bpxxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -Fc testdb > /tmp/testdb.dump
```

2. When `Password:` appears, enter the password of the privileged account of the ApsaraDB RDS for PostgreSQL instance and press the Enter key.



Back up one or more tables

1. Log on to your ECS instance or host. Then, run the following command to back up one or more tables from a database in the ApsaraDB RDS for PostgreSQL instance:

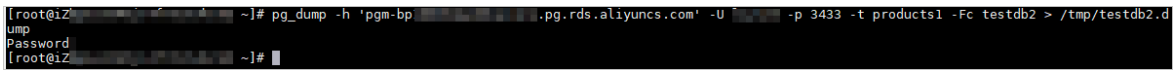
```
pg_dump -h '<hostname>' -U <username> -p <port> -t <table> -Fc <dbname> > <dumpdir>
```

Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances use the VPC network type, make sure that both instances are deployed within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
table	The name of the table that you want to back up. You can use `-t <table>` to specify more than one table.
-Fc	The output file format. `-Fc` specifies the custom format, which is ideal when you use pg_restore to import logical backup files and restore databases. For more information, visit pg_dump.
dbname	The name of the database that you want to back up.
dumpdir	The directory and name of the logical backup file to export.

Example:

```
pg_dump -h 'pgm-bpxxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -t products1 -Fc testdb2 > /tmp/testdb2.dump
```

2. When `Password:` appears, enter the password of the privileged account of the ApsaraDB RDS for PostgreSQL instance and press the Enter key.



Back up a database with one or more tables excluded

1. Log on to your ECS instance or host. Then, run the following command to back up a database from the ApsaraDB RDS instance with one or more tables excluded:

```
pg_dump -h '<hostname>' -U <username> -p <port> -T <table> -Fc <dbname> > <dumpdir>
```

Parameter	Description

Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances use the VPC network type, make sure that both instances are deployed within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
table	The name of the table that you want to exclude. You can use `-T <table>` to specify more than one table.
-Fc	The output file format. `-Fc` specifies the custom format, which is ideal when you use pg_restore to import logical backup files and restore databases. For more information, visit pg_dump.
dbname	The name of the database that you want to back up.
dumpdir	The directory and name of the logical backup file to export.

Example:

```
pg_dump -h 'pgm-bpxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -T products1 -Fc testdb2 > /tmp/testdb2.dump
```

2. When `Password:` appears, enter the password of the privileged account of the ApsaraDB RDS for PostgreSQL instance and press the Enter key.

Back up the schema of a database with data excluded

1. Log on to your ECS instance or host. Then, run the following command to back up the schema of a database from the ApsaraDB RDS for PostgreSQL instance:

```
pg_dump -h '<hostname>' -U <username> -p <port> -s -Fc <dbname> > <dumpdir>
```

Parameter	Description

Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances use the VPC network type, make sure that both instances are deployed within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
-s	Specifies whether to back up only the schema of the database. The data of the database is not backed up. For more information, visit pg_dump.
-Fc	The output file format. `-Fc` specifies the custom format, which is ideal when you use pg_restore to import logical backup files and restore databases. For more information, visit pg_dump.
dbname	The name of the database that you want to back up.
dumpdir	The directory and name of the logical backup file to export.

Example:

```
pg_dump -h 'pgm-bpxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -s -Fc testdb2 > /tmp/testdb2.dump
```

2. When `Password:` appears, enter the password of the privileged account of the ApsaraDB RDS for PostgreSQL instance and press the Enter key.

11.1.13.4. Create a full backup of an ApsaraDB RDS for PostgreSQL instance

This topic describes how to use the pg_basebackup utility provided by open source PostgreSQL to create a full backup of your ApsaraDB RDS for PostgreSQL instance and export the backup files to your computer.

Prerequisites

- The IP address of your ECS instance or host is added to a whitelist of your ApsaraDB RDS for PostgreSQL instance. For more information, see Configure an IP address whitelist.

- Your ECS instance or host runs the same version of PostgreSQL as the ApsaraDB RDS for PostgreSQL instance.

Context

pg_basebackup backs up all data of a PostgreSQL instance. Backup files can be used for point-in-time recovery. For more information, visit pg_basebackup.

In this example, CentOS 7 is used to create a full backup.

Precautions

We recommend that you use the privileged account of the ApsaraDB RDS for PostgreSQL instance to ensure that you have all the required permissions.

Procedure

> **Note** pg_basebackup cannot back up a single database or database object. For more information about how to back up a single database or database object, see Create a logical backup for an ApsaraDB RDS for PostgreSQL instance.

1. Log on to your ECS instance or host. Then, run the following command to back up a database from the ApsaraDB RDS for PostgreSQL instance:

   ```
   pg_basebackup -Ft -Pv -Xf -z -D <backupdir> -Z5 -h '<hostname>' -p <port> -U <username> -W
   ```

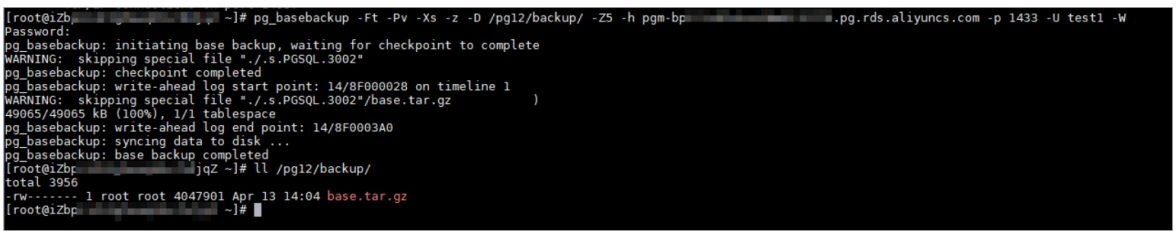
   The following table describes the parameters in this command. For more information, visit pg_basebackup.

Parameter	Description
backupdir	The directory of backup files that are exported. The system automatically creates this directory. However, if this directory already exists and is not empty, the system reports an error.
hostname	The internal endpoint of the ApsaraDB RDS for PostgreSQL instance. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
username	A username of the ApsaraDB RDS for PostgreSQL instance.

   Example:

   ```
   pg_basebackup -Ft -Pv -Xf -z -D /pg12/backup1/ -Z5 -h pgm-bpxxxxx.pg.rds.aliyuncs.com -p 1433 -U test1 -W
   ```

2. When `Password:` appears, enter the password of the username of the ApsaraDB RDS for PostgreSQL instance and press the Enter key.

11.1.14. Restoration

11.1.14.1. Restore data of an ApsaraDB RDS for PostgreSQL instance

This topic describes how to use the backup data of an ApsaraDB RDS for PostgreSQL instance to restore data.

Precautions

- The new instance must have the same whitelist, backup, and parameter settings as the original instance.
- The new instance must have the same data and account information as the backup set or instance at the time point.

Prerequisites

The original instance must meet the following requirements:

- The original instance is in the Running state and is not locked.
- The original instance does not have ongoing migration tasks.
- If you want to restore data to a point in time, the log backup feature is enabled for the original instance.
- If you want to restore an instance from a backup set, the original instance has at least one backup set.

Restore data of an ApsaraDB RDS for PostgreSQL instance

1. Log on to the ApsaraDB RDS console.
2. On the **Instances** page, find the target instance.
3. Click the instance ID or click **Manage** in the Actions column corresponding to the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, click **Backup and Restoration**.
5. In the upper-right corner of the page, click **Restore Database (Previously Clone Database)**.
6. Configure the following parameters.

Section	Parameter	Description
Region	**Region**	The region where the instance is deployed.
Restore Database	**Restore Mode**	◦ **By Time**: You can restore data to a point in time within the retention period of the log backup. For more information about how to view or change the retention period of log backups, see Back up an ApsaraDB RDS for PostgreSQL instance. ◦ **By Backup Set** **Note** The **By Time** option appears only when the log backup feature is enabled.
	Restore Time	The time to which the database is restored. This parameter is displayed when you set **Restore Mode** to **By Time**.
	Backup Set	The backup set used to restore the database. This parameter is displayed when you set **Restore Mode** to **By Backup Set**.
	Instance Name	The name of the instance.

Section	Parameter	Description
Specifications	**Database Engine**	The engine of the database. The value of this parameter is set to **PostgreSQL** and cannot be changed.
	Engine Version	The version of the database engine. The value of this parameter is set to the engine version of the current instance and cannot be changed.
	Edition	The edition of the instance.
	Storage Type	The storage type of the instance. The value of this parameter is set to the storage type of the current instance and cannot be changed.
	Instance Type	The instance type of the instance. Memory size determines the maximum number of connections and IOPS. The actual values are displayed in the console. For more information, see Instance types in *Instance types* of *ApsaraDB RDS Product Introduction*.
	Storage Capacity	The storage capacity of the instance, including the space to store data, system files, binlog files, and transaction files. Valid values: 20 to 600. Unit: GB. The value must be in 1 GB increments.
Network Type	**Network Type**	The network type of the instance. ApsaraDB RDS instances support the following network types: ○ **Classic Network**: Cloud services on a classic network are not isolated from each other. Unauthorized access to a cloud service is blocked only by the security group or whitelist policy of the service. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security. **Note** If you set the network type to **VPC**, you must also select a **VPC** and a **vSwitch**.

7. Click **Submit**.

11.1.14.2. Restore data from a logical backup file

This topic describes how to restore data from a logical backup file to an ApsaraDB RDS for PostgreSQL instance or an on-premises PostgreSQL database.

Context

A logical backup file is used to restore a small amount of data, such as data in a table. For a large amount of data, we recommend that you restore it from a full physical backup file to a new ApsaraDB RDS instance and then use Data Transmission Service (DTS) to migrate data to the original ApsaraDB RDS instance.

Prerequisites

Data in the ApsaraDB RDS for PostgreSQL instance is logically backed up. For more information, see Create a logical backup for an ApsaraDB RDS for PostgreSQL instance.

Precautions

- We recommend that you do not restore data to the default postgres database.
- When you restore the data of a table, the system does not restore the database objects on which the table depends. The restoration may fail.

Restore the data of a database

1. Log on to the ECS instance or on-premises host that houses the logical backup file and run the following command to restore the data of a database:

```
pg_restore -h '<hostname>' -U <username> -p <port> -d <dbname> <dumpdir>
```

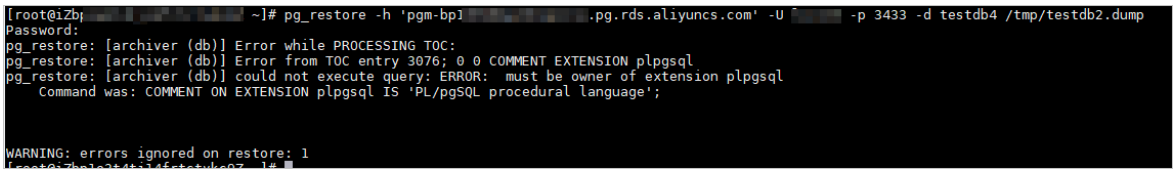
Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances are of the VPC network type, make sure that both instances reside within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
dbname	The name of the database whose data you want to restore.
dumpdir	The directory and name of the logical backup file to use.

Example:

```
pg_restore -h 'pgm-bpxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -d testdb2 /tmp/testdb.dump
```

2. When `Password:` appears, enter the password of the privileged account of your ApsaraDB RDS instance and press the Enter key.

Note You can ignore alerts generated by the embedded plpgsql plug-in.



Restore the data of a table

1. Log on to the ECS instance or on-premises host that houses the logical backup file and run the following command to restore the data of a table:

```
pg_restore -h '<hostname>' -U <username> -p <port> -d <dbname> -t <table> -c <dumpdir>
```

Parameter	Description

Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances are of the VPC network type, make sure that both instances reside within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
dbname	The name of the database whose data you want to restore.
table	The name of the table whose data you want to restore.
-c	`-c`: specifies to delete the database objects on which the table depends before data restoration. For more information, visit pg_restore.
dumpdir	The directory and name of the logical backup file to use.

Example:

```
pg_restore -h 'pgm-bpxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -d testdb2 -t products -c /tmp/testdb.dump
```

2. When `Password:` appears, enter the password of the privileged account of your ApsaraDB RDS instance and press the Enter key.

Restore the schema of a database with data excluded

1. Log on to the ECS instance or on-premises host that houses the logical backup file and run the following command to restore only the schema of a database:

```
pg_restore -h '<hostname>' -U <username> -p <port> -d <dbname> -s <dumpdir>
```

Parameter	Description

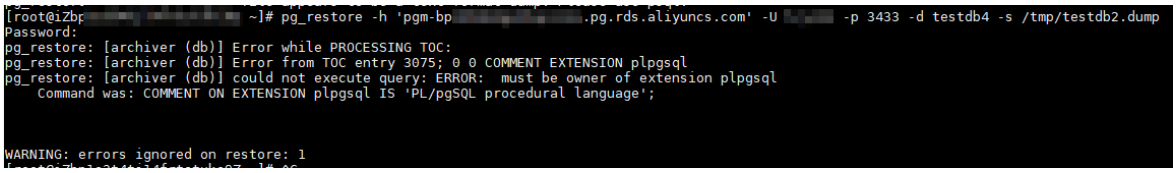
Parameter	Description
hostname	The endpoint of the ApsaraDB RDS for PostgreSQL instance. **Note** If your ECS instance connects to the ApsaraDB RDS for PostgreSQL instance by using an internal endpoint, make sure that the ECS and ApsaraDB RDS instances have the same network type. If both instances are of the VPC network type, make sure that both instances reside within the same VPC. For more information about how to view the internal endpoint, see View and modify the internal endpoint and port number.
username	The username of the privileged account of the ApsaraDB RDS for PostgreSQL instance.
port	The port number of the ApsaraDB RDS for PostgreSQL instance.
dbname	The name of the database whose schema you want to restore.
-s	`-s` : specifies to restore only the schema of the database. The data of the database is not restored. For more information, visit pg_restore.
dumpdir	The directory and name of the logical backup file to use.

Example:

```
pg_restore -h 'pgm-bpxxxxx.pg.rds.aliyuncs.com' -U test123 -p 3433 -d testdb4 -s /tmp/testdb2.dump
```

2. When `Password:` appears, enter the password of the privileged account of your ApsaraDB RDS instance and press the Enter key.

Note You can ignore alerts generated by the embedded plpgsql plug-in.



11.1.15. CloudDBA

11.1.15.1. Introduction to CloudDBA

CloudDBA is a cloud service for database self-detection, self-repair, self-optimization, self-maintenance, and self-security based on machine learning and expertise. CloudDBA helps you ensure stable, secure, and efficient databases without worrying about the management complexity and service failures caused by manual operations.

Features

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the following features:

- Diagnostics

  You can diagnose your instance and view the visualized diagnostic results.

- Instance sessions

  You can view sessions, check session statistics, analyze SQL statements, and optimize the execution of SQL statements.

- Real-time monitoring

  You can view the real-time information of your instance, such as the queries per second (QPS), transactions per second (TPS), number of connections, and network traffic.

- Storage analysis

  You can view the storage overview, trends, exceptions, tablespaces, and data spaces.

- Dashboard

  You can view and compare performance trends, customize monitoring dashboards, check exceptions, and view instance topologies.

- Slow query logs

  You can view the trends and statistics of slow queries.

11.1.15.2. Diagnostics

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the diagnostics feature. This feature diagnoses your ApsaraDB RDS for PostgreSQL instance and visualizes the results.

Navigate to the Diagnostics tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Diagnostics** tab.

> **Note** For more information, see Diagnostics in *Database Autonomy Service User Guide*.

11.1.15.3. Session management

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the session management feature. This feature allows you to view and manage the sessions of an instance.

Navigate to the Session Management tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Session Management** tab.

> **Note** For more information, see Instance sessions in *Database Autonomy Service User Guide*.

11.1.15.4. Real-time monitoring

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the real-time monitoring feature. This feature allows you to view the real-time performance of your ApsaraDB RDS for PostgreSQL instance.

Navigate to the Real-time Monitoring tab

1. Log on to the ApsaraDB for RDS console.

2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Real-time Monitoring** tab.

> **Note** For more information, see Real-time monitoring in *Database Autonomy Service User Guide*.

11.1.15.5. Storage analysis

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the storage analysis feature. This feature allows you to check and solve storage exceptions in a timely manner to ensure database stability.

Context

You can use the storage analysis feature of CloudDBA to view the disk space usage of your ApsaraDB RDS for PostgreSQL instance and the number of remaining days when disk space is available. It also provides information about the space usage, fragmentation, and exception diagnostic results of a table.

Navigate to the Storage Analysis tab

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Diagnostics**.
4. Click the **Storage Analysis** tab.

> **Note** For more information, see Storage analysis in *Database Autonomy Service User Guide*.

11.1.15.6. Dashboard

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the dashboard feature. This feature allows you to view performance trends in specific ranges, compare performance trends, and customize charts to view performance trends.

Navigate to the Dashboard page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Dashboard**.

> **Note** For more information, see Dashboard in *Database Autonomy Service User Guide*.

11.1.15.7. Slow query logs

In ApsaraDB RDS for PostgreSQL, CloudDBA provides the slow query logs feature. This feature allows you to view the trends and execution details of slow queries and obtain optimization suggestions for your ApsaraDB RDS for PostgreSQL instance.

Navigate to the Slow Query Logs page

1. Log on to the ApsaraDB for RDS console.
2. On the **Instances** page, find the target instance.
3. In the left-side navigation pane, choose **CloudDBA > Slow Query Logs**.

> **Note** For more information, see Slow query logs in *Database Autonomy Service User Guide*.

11.1.16. Plug-ins

11.1.16.1. Plug-ins supported

This topic describes the plug-ins that are supported by ApsaraDB RDS for PostgreSQL and their available versions.

PostgreSQL 12

Plug-in	Version
btree_gin	1.3
btree_gist	1.5
citext	1.6
cube	1.4
dblink	1.2
dict_int	1
earthdistance	1.1
fuzzystrmatch	1.1
hstore	1.6
intagg	1.1
intarray	1.2
isn	1.2
ltree	1.1
pg_buffercache	1.3
pg_prewarm	1.2
pg_stat_statements	1.7
pg_trgm	1.4
pgcrypto	1.3
pgrowlocks	1.2
pgstattuple	1.5
postgres_fdw	1
sslinfo	1.2
tablefunc	1

Plug-in	Version
unaccent	1.1
plpgsql	1
plperl	1
pg_roaringbitmap	0.5.0
rdkit	3.8
mysql_fdw	1.1
ganos_geometry_sfcgal	3.0
ganos_geometry_topology	3.0
ganos_geometry	3.0
ganos_networking	3.0
ganos_pointcloud_geometry	3.0
ganos_pointcloud	3.0
ganos_raster	3.0
ganos_spatialref	3.0
ganos_trajectory	3.0
ganos_tiger_geocoder	3.0
ganos_address_standardizer	3.0
ganos_address_standardizer_data_us	3.0
wal2json	2.0
hll	2.14
plproxy	2.9.0
tsm_system_rows	1.0
tsm_system_time	1.0
smlar	1.0
tds_fdw	1.0
bigm	1.2
timescaledb	1.7.1

PostgreSQL 11

Plug-in	Version
plpgsql	1
pg_stat_statements	1.6
btree_gin	1.3
btree_gist	1.5
citext	1.5
cube	1.4
rum	1.3
dblink	1.2
dict_int	1
earthdistance	1.1
hstore	1.5
intagg	1.1
intarray	1.2
isn	1.2
ltree	1.1
pgcrypto	1.3
pgrowlocks	1.2
pg_prewarm	1.2
pg_trgm	1.4
postgres_fdw	1
sslinfo	1.2
tablefunc	1
timescaledb	1.7.1
unaccent	1.1
fuzzystrmatch	1.1
pgstattuple	1.5
pg_buffercache	1.3
zhparser	1
pg_pathman	1.5

Plug-in	Version
plperl	1
orafce	3.8
pg_concurrency_control	1
varbitx	1
postgis	2.5.1
pgrouting	2.6.2
postgis_sfcgal	2.5.1
postgis_topology	2.5.1
address_standardizer	2.5.1
address_standardizer_data_us	2.5.1
ogr_fdw	1
ganos_pointcloud	3.0
ganos_spatialref	3.0
log_fdw	1.0
wal2json	2.2
PL/v8	2.3.13
pg_cron	1.1
pase	0.0.1
hll	2.14
oss_fdw	1.1
tds_fdw	2.0.1
plproxy	2.9.0
tsm_system_rows	1.0
tsm_system_time	1.0
smlar	1.0
zombodb	4.0
bigm	1.2

PostgreSQL 10

Plug-in	Version
pg_stat_statements	1.6
btree_gin	1.2
btree_gist	1.5
chkpass	1
citext	1.4
cube	1.2
dblink	1.2
dict_int	1
earthdistance	1.1
hstore	1.4
intagg	1.1
intarray	1.2
isn	1.1
ltree	1.1
pgcrypto	1.3
pgrowlocks	1.2
pg_prewarm	1.1
pg_trgm	1.3
postgres_fdw	1
sslinfo	1.2
tablefunc	1
unaccent	1.1
postgis_sfcgal	2.5.1
postgis_topology	2.5.1
fuzzystrmatch	1.1
postgis_tiger_geocoder	2.5.1
address_standardizer	2.5.1
address_standardizer_data_us	2.5.1
ogr_fdw	1

Plug-in	Version
plperl	1
plv8	1.4.2
plls	1.4.2
plcoffee	1.4.2
uuid-ossp	1.1
zhparser	1
pgrouting	2.6.2
pg_hint_plan	1.3.0
pgstattuple	1.5
oss_fdw	1.1
ali_decoding	0.0.1
varbitx	1
pg_buffercache	1.3
q3c	1.5.0
pg_sphere	1
smlar	1
rum	1.3
pg_pathman	1.5
aggs_for_arrays	1.3.1
mysql_fdw	1
orafce	3.6
plproxy	2.8.0
pg_concurrency_control	1
postgis	2.5.1
ganos_geometry_sfcgal	2.2
ganos_geometry_topology	2.2
ganos_geometry	2.2
ganos_networking	2.2
ganos_pointcloud_geometry	2.2

Plug-in	Version
ganos_pointcloud	2.2
ganos_raster	2.2
ganos_spatialref	2.2
ganos_trajectory	2.2
ganos_tiger_geocoder	2.2
ganos_address_standardizer	2.2
ganos_address_standardizer_data_us	2.2

PostgreSQL 9.4

Plug-in	Version
plpgsql	1
pg_stat_statements	1.2
btree_gin	1
btree_gist	1
chkpass	1
citext	1
cube	1
dblink	1.1
dict_int	1
earthdistance	1
hstore	1.3
intagg	1
intarray	1
isn	1
ltree	1
pgcrypto	1.1
pgrowlocks	1.1
pg_prewarm	1
pg_trgm	1.1
postgres_fdw	1

Plug-in	Version
sslinfo	1
tablefunc	1
tsearch2	1
unaccent	1
postgis	2.2.8
postgis_topology	2.2.8
fuzzystrmatch	1
postgis_tiger_geocoder	2.2.8
plperl	1
pltcl	1
plv8	1.4.2
plls	1.4.2
plcoffee	1.4.2
uuid-ossp	1
zhparser	1
pgrouting	2.0.0
rdkit	3.4
pg_hint_plan	1.1.3
pgstattuple	1.2
oss_fdw	1.1
jsonbx	1
ali_decoding	0.0.1
varbitx	1
pg_buffercache	1
smlar	1
pg_sphere	1
q3c	1.5.0
pg_awr	1
imgsmlr	1

Plug-in	Version
orafce	3.6
pg_concurrency_control	1

11.1.16.2. Use mysql_fdw to read data from and write data to a MySQL database

This topic describes how to use the mysql_fdw plug-in of ApsaraDB RDS for PostgreSQL to read data from and write data to a database on an ApsaraDB RDS for MySQL instance or a self-managed MySQL database.

Prerequisites

- The instance runs PostgreSQL 10.
- Communication between your ApsaraDB RDS for PostgreSQL instance and the MySQL database is normal.

Context

PostgreSQL 9.6 and later support parallel computing. PostgreSQL 11 can use joins on up to a billion data records to complete queries in seconds. A number of users prefer to use PostgreSQL to build small-sized data warehouses and process highly concurrent access requests. PostgreSQL 13 is under development. It will support columnar storage engines that further improve analysis capabilities.

The mysql_fdw plug-in establishes a connection to synchronize data from a MySQL database to your ApsaraDB RDS for PostgreSQL instance.

Procedure

1. Log on to a database of your ApsaraDB RDS for PostgreSQL instance. For more information, see Connect to an ApsaraDB RDS for PostgreSQL instance.
2. Create the mysql_fdw plug-in.

   ```
   create extension mysql_fdw;
   ```

3. Define a MySQL server.

   ```
   CREATE SERVER <Name of the MySQL server>
   FOREIGN DATA WRAPPER mysql_fdw
   OPTIONS (host '<Endpoint used to connect to the MySQL server>', port '<Port used to connect to the
   MySQL server>');
   ```

   Example:

   ```
   CREATE SERVER mysql_server
   FOREIGN DATA WRAPPER mysql_fdw
   OPTIONS (host 'rm-xxx.mysql.rds.aliyuncs.com', port '3306');
   ```

4. Map the MySQL server to an account created on your ApsaraDB RDS for PostgreSQL instance. Then, the account can be used to access data in the MySQL database on the MySQL server.

   ```
   CREATE USER MAPPING FOR <Username of the account to which the MySQL server is mapped>
   SERVER <Name of the MySQL server>
   OPTIONS (username '<Username used to log on to the MySQL database>', password '<Password used to l
   og on to the MySQL database>');
   ```

   Example:

```
CREATE USER MAPPING FOR pgtest
SERVER mysql_server
OPTIONS (username 'mysqltest', password 'Test1234!') ;
```

5. Create a foreign MySQL table by using the account that you mapped to the MySQL server in the previous step.

> **Note** The field names in the foreign MySQL table must be the same as those in the table of the MySQL database. You can choose to create only the fields you want to query. For example, if the table in the MySQL database contains the ID, NAME, and AGE fields, you can create only the ID and NAME fields in the foreign MySQL table.

```
CREATE FOREIGN TABLE <Name of the foreign MySQL table> (<Name of Field 1> <Data type of Field 1>,<Name of Field 2> <Data type of Field 2>...) server <Name of the MySQL server> options (dbname '<Name of the MySQL database>', table_name '<Name of the table in the MySQL database>');
```

Example:

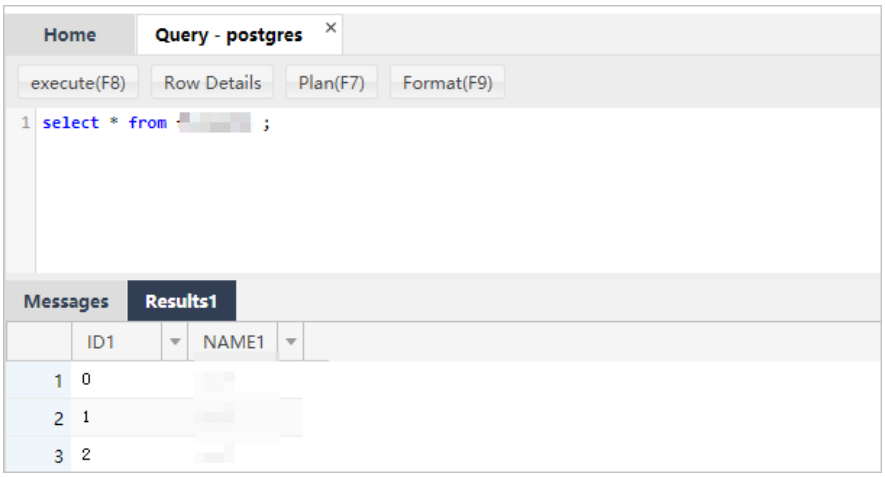
```
CREATE FOREIGN TABLE ft_test (id1 int, name1 text) server mysql_server options (dbname 'test123', table_name 'test');
```

What to do next

You can use the foreign MySQL table to test the performance of read and write operations on the MySQL database.

> **Note** Data can be written to the table in the MySQL database only when the table is assigned a primary key. If the table is not assigned a primary key, the following error is reported:
>
> ```
> ERROR:  first column of remote table must be unique for INSERT/UPDATE/DELETE operation.
> ```

```
select * from ft_test ;
insert into ft_test values (2,'abc');
insert into ft_test select generate_series(3,100),'abc';
select count(*) from ft_test ;
```



Run `postgres=> explain verbose select count(*) from ft_test;` to find out how the requests sent from your ApsaraDB RDS for PostgreSQL instance are executed to query data from the MySQL database. Command output:

```
                                QUERY PLAN
-------------------------------------------------------------------------------
 Aggregate  (cost=1027.50..1027.51 rows=1 width=8)
   Output: count(*)
   ->  Foreign Scan on public.ft_test  (cost=25.00..1025.00 rows=1000 width=0)
         Output: id, info
         Remote server startup cost: 25
         Remote query: SELECT NULL FROM `test123`.`test`
(6 rows)
```

11.1.16.3. Use oss_fdw to read and write foreign data files

This topic describes how to use the oss_fdw plug-in to load data between Object Storage Service (OSS) and PostgreSQL or PPAS databases.

oss_fdw parameters

The oss_fdw plug-in uses a method similar to other Foreign Data Wrapper (FDW) interfaces to encapsulate foreign data stored in OSS. You can use oss_fdw to read data stored in OSS. This process is similar to reading data tables. oss_fdw provides unique parameters to connect and parse file data in OSS.

> **Note**
>
> - oss_fdw can read and write files of the following types in OSS: TXT and CSV files as well as GZIP-compressed TXT and CSV files.
> - The value of each parameter must be enclosed in double quotation marks (") and cannot contain unnecessary spaces.

CREATE SERVER parameters

- ossendpoint: the endpoint used to access OSS over the internal network, also known as the host.
- id oss: the AccessKey ID of the OSS account.
- key oss: the AccessKey secret of the OSS account.
- bucket: the bucket where the data you want to access is stored. You must create an OSS account before you specify this parameter.

The following fault tolerance parameters can be used for data import and export. If network connectivity is poor, you can adjust these parameters to ensure successful import and export.

- oss_connect_timeout: the connection timeout period. Default value: 10. Unit: seconds.
- oss_dns_cache_timeout: the DNS timeout period. Default value: 60. Unit: seconds.
- oss_speed_limit: the minimum data transmission rate. Default value: 1024. Unit: bytes/s.
- oss_speed_time: the maximum waiting period during which the data transmission rate is lower than the minimum value. Default value: 15. Unit: seconds.

If the default values of oss_speed_limit and oss_speed_time are used, a timeout error occurs when the transmission rate is lower than 1,024 bytes/s for 15 consecutive seconds.

CREATE FOREIGN TABLE parameters

- filepath: a file name that contains a path in OSS.
  - The file name specified by this parameter contains the directory name but not the bucket name.
  - This parameter matches multiple files in the corresponding path in OSS. You can load multiple files to a database.

  - You can import files that adhere to the filepath or filepath.x format to a database. The values of x must be consecutive numbers starting from 1.

    For example, among the files named filepath, filepath.1, filepath.2, filepath.3, and filepath.5, the first four files are matched and imported. The filepath.5 file is not imported.

- dir: the virtual file directory in OSS.
  - The specified directory must end with a forward slash (/).
  - All files (excluding subfolders and files in subfolders) in the virtual file directory specified by dir are matched and imported to a database.
- prefix: the prefix of the path name corresponding to the data file. The prefix does not support regular expressions. The prefix, filepath, and dir parameters are mutually exclusive. Therefore, only one of them can be specified at a time.
- format: the file format, which can only be CSV.
- encoding: the file data encoding format. It supports common PostgreSQL encoding formats, such as UTF-8.
- parse_errors: the fault-tolerant parsing mode. If an error occurs during the parsing process, the entire row of data is ignored.
- delimiter: the string used to delimit columns.
- quote: the quote character for files.
- escape: the escape character for files.
- null: sets the column matching the specified string to null. For example, null 'test' is used to set the value of the 'test' column to null.
- force_not_null: sets the value of a column to a non-null value. For example, force_not_null 'id' is used to set the value of the 'id' column to empty strings.
- compressiontype: the format of the files to be read or written in OSS.
  - none: The files are uncompressed. This is the default value.
  - gzip: The files are compressed in the GZIP format.
- compressionlevel: the degree to which data files written to OSS are compressed. Valid values: 1 to 9. Default value: 6.

> **Note**
> - You must specify filepath and dir in the OPTIONS parameter.
> - You must specify filepath or dir.
> - The export mode can only be dir.

Export mode parameters for CREATE FOREIGN TABLE

- oss_flush_block_size: the buffer size for the data written to OSS at a time. Default value: 32. Valid values: 1 to 128. Unit: MB.
- oss_file_max_size: the maximum size of a data file allowed to be written to OSS. If a data file reaches the maximum size, the remaining data is written to another data file. Default value: 1024. Valid values: 8 to 4000. Unit: MB.
- num_parallel_worker: the maximum number of threads that are allowed to run in parallel to compress the data written to OSS. Valid values: 1 to 8. Default value: 3.

Auxiliary functions

FUNCTION oss_fdw_list_file (relname text, schema text DEFAULT 'public')

- This function obtains the name and size of the OSS file that a foreign table matches.
- The file size is measured in bytes.

The following result is returned after `select * from oss_fdw_list_file('t_oss');` is executed:

```
              name              |   size
--------------------------------+-----------
 oss_test/test.gz.1  | 739698350
 oss_test/test.gz.2  | 739413041
 oss_test/test.gz.3  | 739562048
(3 rows)
```

Auxiliary features

oss_fdw.rds_read_one_file: In read mode, this feature is used to specify a file to match the foreign table. The foreign table matches only the specified file during data import.

Example: set oss_fdw.rds_read_one_file = 'oss_test/example16.csv.1';

The following result is returned after `set oss_fdw.rds_read_one_file = 'oss_test/test.gz.2';` and `select * from oss_fdw_list_file('t_oss');` are executed:

```
              name              |   size
--------------------------------+-----------
  oss_test/test.gz.2  | 739413041
(1 rows)
```

oss_fdw example

```
# Create the plug-in for a PostgreSQL database.
create extension oss_fdw; -- For a PPAS database, execute select rds_manage_extension('create','oss_fdw');
# Create a server.
CREATE SERVER ossserver FOREIGN DATA WRAPPER oss_fdw OPTIONS
     (host 'oss-cn-hangzhou.aliyuncs.com', id 'xxx', key 'xxx', bucket 'mybucket');
# Create an OSS foreign table.
CREATE FOREIGN TABLE ossexample
    (date text, time text, open float,
     high float, low float, volume int)
     SERVER ossserver
     OPTIONS ( filepath 'osstest/example.csv', delimiter ',' ,
         format 'csv', encoding 'utf8', PARSE_ERRORS '100');
# Create a table named example to which to import data.
create table example
        (date text, time text, open float,
         high float, low float, volume int);
# Load data from ossexample to example.
insert into example select * from ossexample;
# Result
# oss_fdw estimates the file size in OSS and formulates a query plan.
explain insert into example select * from ossexample;
                            QUERY PLAN
------------------------------------------------------------------------
 Insert on example  (cost=0.00..1.60 rows=6 width=92)
   ->  Foreign Scan on ossexample  (cost=0.00..1.60 rows=6 width=92)
         Foreign OssFile: osstest/example.csv.0
         Foreign OssFile Size: 728
(4 rows)
# Write the data in the example table to OSS.
insert into ossexample select * from example;
explain insert into ossexample select * from example;
                           QUERY PLAN
-----------------------------------------------------------------
 Insert on ossexample  (cost=0.00..16.60 rows=660 width=92)
   ->  Seq Scan on example  (cost=0.00..16.60 rows=660 width=92)
(2 rows)
```

Additional considerations

- oss_fdw is a foreign table plug-in developed based on the PostgreSQL FOREIGN TABLE framework.
- The data import performance varies based on the PostgreSQL cluster resources (CPU, I/O, and memory) and OSS.
- To ensure data import performance, the ApsaraDB RDS for PostgreSQL instance must be in the same region as the OSS bucket.

ID and key encryption

If the id and key parameters for CREATE SERVER are not encrypted, the `select * from pg_foreign_server` statement execution result displays the information. Your AccessKey ID and AccessKey secret are exposed. You can use symmetric encryption to hide your AccessKey ID and AccessKey secret. Use different AccessKey pairs for different instances to further protect your information. However, to avoid incompatibility with earlier versions, do not add data types as you would in Greenplum.

Encrypted information:

```
postgres=# select * from pg_foreign_server ;
  srvname  | srvowner | srvfdw | srvtype | srvversion | srvacl |
srvoptions
-----------+----------+--------+---------+------------+--------+-----------------------------------
-----------------------------------------------------------------------------------------------
---------------------------------
 ossserver |       10 |  16390 |         |            |        | {host=oss-cn-hangzhou-zmf.aliyuncs.co
m,id=MD5xxxxxxxx,key=MD5xxxxxxxx,bucket=067862}
```

The encrypted information is preceded by the MD5 hash value. The remainder of the total length divided by 8 is 3. Therefore, encryption is not performed again when the exported data is imported. You cannot create an AccessKey pair that is preceded by MD5.

11.1.17. Use Pgpool for read/write splitting in ApsaraDB RDS for PostgreSQL

This topic describes how to use the Pgpool tool of PostgreSQL installed on an ECS instance to implement read/write splitting for your primary and read-only ApsaraDB RDS for PostgreSQL instances.

Context

If you do not use Pgpool to ensure high availability, Pgpool is stateless. The decrease in performance can be ignored. Additionally, Pgpool supports horizontal scaling of your database system. You can use Pgpool and the high availability architecture of ApsaraDB RDS for PostgreSQL to implement read/write splitting.

Set up a test environment

If you have purchased a primary ApsaraDB RDS instance that runs PostgreSQL 10 and have attached read-only instances to the primary instance, you need only to install Pgpool . For more information, see Create an instance and Create a read-only ApsaraDB RDS for PostgreSQL instance. After you install Pgpool, go to Configure Pgpool.

1. Run the `vi /etc/sysctl.conf` command to open the sysctl.conf file. Modify the following configurations:

```
# add by digoal.zhou
fs.aio-max-nr = 1048576
fs.file-max = 76724600
# Optional. Set the kernel.core_pattern parameter to /data01/corefiles/core_%e_%u_%t_%s.%p.

# The /data01/corefiles directory that is used to store core dumps is created with the 777 permiss
ion before testing. If a symbolic link is used, change the directory to 777.
kernel.sem = 4096 2147483647 2147483646 512000
# Specify the semaphore. You can run the ipcs -l or -u command to obtain the semaphore count. Each
group of 16 processes requires a semaphore with a count of 17.
kernel.shmall = 107374182
# Specify the total size of shared memory segments. Recommended value: 80% of the memory capacity.
Unit: pages.
kernel.shmmax = 274877906944
# Specify the maximum size of a single shared memory segment. Recommended value: 50% of the memory
capacity. Unit: bytes. In PostgreSQL versions later than 9.2, the use of shared memory significant
ly drops.
kernel.shmmni = 819200
# Specify the total number of shared memory segments that can be generated. At least two shared me
mory segments must be generated within each PostgreSQL cluster.
net.core.netdev_max_backlog = 10000
net.core.rmem_default = 262144
# The default setting of the socket receive buffer in bytes.
net.core.rmem_max = 4194304
# The maximum receive socket buffer size in bytes
```

```
net.core.wmem_default = 262144
# The default setting (in bytes) of the socket send buffer.
net.core.wmem_max = 4194304
# The maximum send socket buffer size in bytes.
net.core.somaxconn = 4096
net.ipv4.tcp_max_syn_backlog = 4096
net.ipv4.tcp_keepalive_intvl = 20
net.ipv4.tcp_keepalive_probes = 3
net.ipv4.tcp_keepalive_time = 60
net.ipv4.tcp_mem = 8388608 12582912 16777216
net.ipv4.tcp_fin_timeout = 5
net.ipv4.tcp_synack_retries = 2
net.ipv4.tcp_syncookies = 1
# Enable SYN cookies. If an SYN waiting queue overflows, you can enable SYN cookies to defend agai
nst a small number of SYN attacks.
net.ipv4.tcp_timestamps = 1
# Reduce the time after which a network socket enters the TIME-WAIT state.
net.ipv4.tcp_tw_recycle = 0
# If you set this parameter to 1 to enable the recycle function, network sockets in the TIME-WAIT
state over TCP connections are recycled. However, if network address translation (NAT) is used, TC
P connections may fail. We recommend that you set this parameter to 0 on the database server.
net.ipv4.tcp_tw_reuse = 1
# Enable the reuse function. This function enables network sockets in the TIME-WAIT state to be re
used over new TCP connections.
net.ipv4.tcp_max_tw_buckets = 262144
net.ipv4.tcp_rmem = 8192 87380 16777216
net.ipv4.tcp_wmem = 8192 65536 16777216
net.nf_conntrack_max = 1200000
net.netfilter.nf_conntrack_max = 1200000
vm.dirty_background_bytes = 409600000
# If the size of dirty pages reaches the specified limit, a background scheduling process (for exa
mple, pdflush) is invoked to flush the dirty pages to disks. These are the pages that are generate
d n seconds earlier. The value of n is calculated by using the following formula: n = Value of the
dirty_expire_centisecs parameter/100.
# The default limit is 10% of the memory capacity. If the memory capacity is large, we recommend t
hat you specify the limit in bytes.
vm.dirty_expire_centisecs = 3000
# Specify the maximum period to retain dirty pages. Dirty pages are flushed to disks after the tim
e period specified by this parameter elapses. The value 3000 indicates 30 seconds.
vm.dirty_ratio = 95
# The processes that users call to write data onto disks must actively flush dirty pages to disks.
This applies when the background scheduling process to flush dirty pages is slow and the size of d
irty pages exceeds 95% of the memory capacity. These processes include fsync and fdatasync.

# Set this parameter properly to prevent user-called processes from flushing dirty pages to disks.
This allows you to create multiple ApsaraDB RDS instances on a single server and use control group
s to limit the input/output operations per second (IOPS) per instance.
vm.dirty_writeback_centisecs = 100
# Specify the time interval at which the background scheduling process (such as pdflush) flushes d
irty pages to disks. The value 100 indicates 1 second.
vm.swappiness = 0
# Disable the swap function.
vm.mmap_min_addr = 65536
vm.overcommit_memory = 0
# Specify whether you can allocate more memory space than the physical host has available. If you
set this parameter to 1, the system always considers the available memory space sufficient. If the
memory capacity provided in the test environment is low, we recommend that you set this parameter
to 1.
vm.overcommit_ratio = 90
# Specify the memory capacity that can be allocated when the overcommit_memory parameter is set to
```

```
2.
vm.swappiness = 0
# Disable the swap function.
vm.zone_reclaim_mode = 0
# Disable non-uniform memory access (NUMA). You can also disable NUMA in the vmlinux file.

net.ipv4.ip_local_port_range = 40000 65535
# Specify the range of TCP or UDP port numbers for the physical host to allocate.
fs.nr_open=20480000
# Specify the maximum number of file handles that a single process can open.
# Take note of the following parameters:
#vm.extra_free_kbytes = 4096000   # If the physical host provides a low memory capacity, do not sp
ecify a large value such as 4096000. If you specify a large value, the physical host may not start
.
#vm.min_free_kbytes = 6291456    # We recommend that you increase the value of the vm.min_free_kby
tes parameter by 1 GB for every 32 GB of memory.
# If the physical host does not provide much memory, we recommend that you do not configure vm.ext
ra_free_kbytes and vm.min_free_kbytes.
# vm.nr_hugepages = 66536
# If the size of the shared buffer exceeds 64 GB, we recommend that you use huge pages. You can sp
ecify the page size by setting the Hugepagesize parameter in the /proc/meminfo file.

#vm.lowmem_reserve_ratio = 1 1 1
# If the memory capacity exceeds 64 GB, we recommend that you set this parameter. Otherwise, we re
commend that you retain the default value 256 256 32.
```

2. Run the `vi /etc/security/limits.conf` command to open the limits.conf file. Modify the following configurations:

```
* soft    nofile  1024000
* hard    nofile  1024000
* soft    nproc   unlimited
* hard    nproc   unlimited
* soft    core    unlimited
* hard    core    unlimited
* soft    memlock unlimited
* hard    memlock unlimited
# Comment out the other parameters in the limits.conf file.
# Comment out the /etc/security/limits.d/20-nproc.conf file.
```

3. Run the following commands to open the rc.local file:

```
chmod +x /etc/rc.local
vi /etc/rc.local
```

Modify the following configurations to disable transparent huge pages, configure huge pages, and start PostgreSQL:

```
# Disable transparent huge pages.
if test -f /sys/kernel/mm/transparent_hugepage/enabled; then
   echo never > /sys/kernel/mm/transparent_hugepage/enabled
fi
# Configure huge pages for two instances. Each instance has a shared buffer of 16 GB.
sysctl -w vm.nr_hugepages=17000
# Start the two instances.
su - postgres -c "pg_ctl start -D /data01/pg12_3389/pg_root"
su - postgres -c "pg_ctl start -D /data01/pg12_8002/pg_root"
```

4. Create a file system.

> **Warning** If you use a new disk, you must verify that the new disk belongs to the vdb partition instead of the vda partition. If the new disk belongs to the vda partition, data may be deleted from the new disk.

```
parted -a optimal -s /dev/vdb mklabel gpt mkpart primary 1MiB 100%FREE
mkfs.ext4 /dev/vdb1 -m 0 -O extent,uninit_bg -E lazy_itable_init=1 -b 4096 -T largefile -L vdb1
vi /etc/fstab
LABEL=vdb1 /data01 ext4 defaults,noatime,nodiratime,nodelalloc,barrier=0,data=writeback 0 0
mkdir /data01
mount -a
```

5. Start the irqbalance command line tool.

```
systemctl status irqbalance
systemctl enable irqbalance
systemctl start irqbalance
systemctl status irqbalance
```

6. Install PostgreSQL 10 and Pgpool.

```
yum install -y https://dl.fedoraproject.org/pub/epel/epel-release-latest-7.noarch.rpm
yum install -y https://download.postgresql.org/pub/repos/yum/reporpms/EL-7-x86_64/pgdg-redhat-repo
-latest.noarch.rpm
yum search all postgresql
yum search all pgpool
yum install -y postgresql12*
yum install -y pgpool-II-12-extensions
```

7. Initialize the data directory of your database system.

```
mkdir /data01/pg12_3389
chown postgres:postgres /data01/pg12_3389
```

8. Configure environment variables for the postgres user.

```
su - postgres
vi .bash_profile
```

Append the following parameters to the environment variables:

```
export PS1="$USER@`/bin/hostname -s`-> "
export PGPORT=3389
export PGDATA=/data01/pg12_$PGPORT/pg_root
export LANG=en_US.utf8
export PGHOME=/usr/pgsql-12
export LD_LIBRARY_PATH=$PGHOME/lib:/lib64:/usr/lib64:/usr/local/lib64:/lib:/usr/lib:/usr/local/lib
:$LD_LIBRARY_PATH
export DATE=`date +"%Y%m%d%H%M"`
export PATH=$PGHOME/bin:$PATH:.
export MANPATH=$PGHOME/share/man:$MANPATH
export PGHOST=$PGDATA
export PGUSER=postgres
export PGDATABASE=db1
alias rm='rm -i'
alias ll='ls -lh'
unalias vi
```

9. Initialize your primary ApsaraDB RDS instance.

```
initdb -D $PGDATA -U postgres -E UTF8 --lc-collate=C --lc-ctype=en_US.utf8
```

10. Modify the postgresql.conf file.

```
listen_addresses = '0.0.0.0'
port = 3389
max_connections = 1500
superuser_reserved_connections = 13
unix_socket_directories = '., /var/run/postgresql, /tmp'
tcp_keepalives_idle = 60
tcp_keepalives_interval = 10
tcp_keepalives_count = 10
shared_buffers = 16GB
huge_pages = on
work_mem = 8MB
maintenance_work_mem = 1GB
dynamic_shared_memory_type = posix
vacuum_cost_delay = 0
bgwriter_delay = 10ms
bgwriter_lru_maxpages = 1000
bgwriter_lru_multiplier = 10.0
bgwriter_flush_after = 512kB
effective_io_concurrency = 0
max_worker_processes = 128
max_parallel_maintenance_workers = 3
max_parallel_workers_per_gather = 4
parallel_leader_participation = off
max_parallel_workers = 8
backend_flush_after = 256
wal_level = replica
synchronous_commit = off
full_page_writes = on
wal_compression = on
wal_buffers = 16MB
wal_writer_delay = 10ms
wal_writer_flush_after = 1MB
checkpoint_timeout = 15min
max_wal_size = 64GB
min_wal_size = 8GB
checkpoint_completion_target = 0.2
checkpoint_flush_after = 256kB
random_page_cost = 1.1
effective_cache_size = 48GB
log_destination = 'csvlog'
logging_collector = on
log_directory = 'log'
log_filename = 'postgresql-%a.log'
log_truncate_on_rotation = on
log_rotation_age = 1d
log_rotation_size = 0
log_min_duration_statement = 1s
log_checkpoints = on
log_connections = on
log_disconnections = on
log_line_prefix = '%m [%p] '
log_statement = 'ddl'
log_timezone = 'Asia/Shanghai'
autovacuum = on
log_autovacuum_min_duration = 0
autovacuum_vacuum_scale_factor = 0.1
autovacuum_analyze_scale_factor = 0.05
autovacuum_freeze_max_age = 800000000
```

```
autovacuum_multixact_freeze_max_age = 900000000
autovacuum_vacuum_cost_delay = 0
vacuum_freeze_table_age = 750000000
vacuum_multixact_freeze_table_age = 750000000
datestyle = 'iso, mdy'
timezone = 'Asia/Shanghai'
lc_messages = 'en_US.utf8'
lc_monetary = 'en_US.utf8'
lc_numeric = 'en_US.utf8'
lc_time = 'en_US.utf8'
default_text_search_config = 'pg_catalog.english'
```

11. Modify the pg_hba.conf file.

> **Note** Pgpool-II is installed on the same ECS instance as the database server where PostgreSQL resides. If you specify the 127.0.0.1 IP address in the pg_hba.conf file, you must enter the correct password to ensure a successful logon.

```
# "local" is for Unix domain socket connections only
local   all             all                                     trust
# IPv4 local connections:
host    all             all             127.0.0.1/32            md5
# IPv6 local connections:
host    all             all             ::1/128                 trust
# Allow replication connections from localhost, by a user with the
# replication privilege.
local   replication     all                                     trust
host    replication     all             127.0.0.1/32            trust
host    replication     all             ::1/128                 trust
host db123 digoal 0.0.0.0/0 md5
```

12. Execute a statement in the database to create a user authorized with streaming replication permissions. Example:

```
create role rep123 login replication encrypted password 'xxxxxxx';
```

13. Execute statements in the database to create a user and authorize it to manage your ApsaraDB RDS instances. Example:

```
create role digoal login encrypted password 'xxxxxxx';
create database db123 owner digoal;
```

14. Create a user who is authorized to check the health heartbeats between Pgpool and your read-only ApsaraDB RDS instances. With the parameters of Pgpool properly configured, this user can check the write-ahead logging (WAL) replay latency on each read-only ApsaraDB RDS instance. Example:

```
create role nobody login encrypted password 'xxxxxxx';
```

Create a secondary ApsaraDB RDS instance

To simplify the test procedure, perform the following steps to create a secondary ApsaraDB RDS instance on the same ECS instance as your primary ApsaraDB RDS instance:

1. Use the pg_basebackup tool to create a secondary ApsaraDB RDS instance.

```
pg_basebackup -D /data01/pg12_8002/pg_root -F p --checkpoint=fast -P -h 127.0.0.1 -p 3389 -U rep12
3
```

2. Run the following commands to open the postgresql.conf file of the secondary ApsaraDB RDS instance:

```
cd /data01/pg12_8002/pg_root
vi postgresql.conf
```

Modify the following configurations:

```
# The secondary ApsaraDB RDS instance has the following configurations different from the primary
ApsaraDB RDS instance:
port = 8002
primary_conninfo = 'hostaddr=127.0.0.1 port=3389 user=rep123' # You do not need to set the passwor
d. This is because trust relationships are configured on the primary ApsaraDB RDS instance.
hot_standby = on
wal_receiver_status_interval = 1s
wal_receiver_timeout = 10s
recovery_target_timeline = 'latest'
```

3. Configure the standby.signal file of the secondary ApsaraDB RDS instance.

```
cd /data01/pg12_8002/pg_root
touch standby.signal
```

4. Execute the `SELECT * FROM pg_stat_replication ;` statement in the database to check whether data is properly synchronized between the primary and secondary ApsaraDB RDS instances. The following output is returned:

```
-[ RECORD 1 ]----+------------------------------
pid              | 21065
usesysid         | 10
usename          | postgres
application_name | walreceiver
client_addr      | 127.0.0.1
client_hostname  |
client_port      | 47064
backend_start    | 2020-02-29 00:26:28.485427+08
backend_xmin     |
state            | streaming
sent_lsn         | 0/52000060
write_lsn        | 0/52000060
flush_lsn        | 0/52000060
replay_lsn       | 0/52000060
write_lag        |
flush_lag        |
replay_lag       |
sync_priority    | 0
sync_state       | async
reply_time       | 2020-02-29 01:32:40.635183+08
```

Configure Pgpool

1. Query the location where Pgpool is installed.

```
rpm -qa|grep pgpool
pgpool-II-12-extensions-4.1.1-1.rhel7.x86_64
pgpool-II-12-4.1.1-1.rhel7.x86_64
rpm -ql pgpool-II-12-4.1.1
```

2. Run the following commands to open the pgpool.conf file:

```
cd /etc/pgpool-II-12/
cp pgpool.conf.sample-stream pgpool.conf
vi pgpool.conf
```

Modify the following configurations:

```
listen_addresses = '0.0.0.0'
port = 8001
socket_dir = '/tmp'
reserved_connections = 0
pcp_listen_addresses = ''
pcp_port = 9898
pcp_socket_dir = '/tmp'
# - Backend Connection Settings -
backend_hostname0 = '127.0.0.1'
                                    # Host name or IP address to connect to for backend 0
backend_port0 = 3389
                                    # Port number for backend 0
backend_weight0 = 1
                                    # Weight for backend 0 (only in load balancing mode)
backend_data_directory0 = '/data01/pg12_3389/pg_root'
                                    # Data directory for backend 0
backend_flag0 = 'ALWAYS_MASTER'
                                    # Controls various backend behavior
                                    # ALLOW_TO_FAILOVER, DISALLOW_TO_FAILOVER
                                    # or ALWAYS_MASTER
backend_application_name0 = 'server0'
                                    # walsender's application_name, used for "show pool_nodes" comm
and
backend_hostname1 = '127.0.0.1'
backend_port1 = 8002
backend_weight1 = 1
backend_data_directory1 = '/data01/pg12_8002/pg_root'
backend_flag1 = 'DISALLOW_TO_FAILOVER'
backend_application_name1 = 'server1'
# - Authentication -
enable_pool_hba = on
                                    # Use pool_hba.conf for client authentication
pool_passwd = 'pool_passwd'
                                    # File name of pool_passwd for md5 authentication.
                                    # "" disables pool_passwd.
                                    # (change requires restart)
allow_clear_text_frontend_auth = off
                                    # Allow Pgpool-II to use clear text password authentication
                                    # with clients, when pool_passwd does not
                                    # contain the user password
# - Concurrent session and pool size -
num_init_children = 128
                                    # Number of concurrent sessions allowed
                                    # (change requires restart)
max_pool = 4
                                    # Number of connection pool caches per connection
                                    # (change requires restart)
# - Life time -
child_life_time = 300
                                    # Pool exits after being idle for this many seconds
child_max_connections = 0
                                    # Pool exits after receiving that many connections
                                    # 0 means no exit
connection_life_time = 0
                                    # Connection to backend closes after being idle for this many s
econds
                                    # 0 means no close
client_idle_limit = 0
```

```
                                    # Client is disconnected after being idle for that many seconds

                                    # (even inside an explicit transactions!)
                                    # 0 means no disconnection
#------------------------------------------------------------------------------
# LOGS
#------------------------------------------------------------------------------
# - Where to log -
log_destination = 'syslog'
                                    # Where to log
                                    # Valid values are combinations of stderr,
                                    # and syslog. Default to stderr.
log_connections = on
                                    # Log connections
log_standby_delay = 'if_over_threshold'
                                    # Log standby delay
                                    # Valid values are combinations of always,
                                    # if_over_threshold, none
#------------------------------------------------------------------------------
# FILE LOCATIONS
#------------------------------------------------------------------------------
pid_file_name = '/var/run/pgpool-II-12/pgpool.pid'
                                    # PID file name
                                    # Can be specified as relative to the"
                                    # location of pgpool.conf file or
                                    # as an absolute path
                                    # (change requires restart)
logdir = '/tmp'
                                    # Directory of pgPool status file
                                    # (change requires restart)
#------------------------------------------------------------------------------
# CONNECTION POOLING
#------------------------------------------------------------------------------
connection_cache = on
                                    # Activate connection pools
                                    # (change requires restart)
                                    # Semicolon separated list of queries
                                    # to be issued at the end of a session
                                    # The default is for 8.3 and later
reset_query_list = 'ABORT; DISCARD ALL'
#------------------------------------------------------------------------------
# LOAD BALANCING MODE
#------------------------------------------------------------------------------
load_balance_mode = on
                                    # Activate load balancing mode
                                    # (change requires restart)
ignore_leading_white_space = on
                                    # Ignore leading white spaces of each query
white_function_list = ''
                                    # Comma separated list of function names
                                    # that don't write to database
                                    # Regexp are accepted
black_function_list = 'currval,lastval,nextval,setval'
                                    # Comma separated list of function names
                                    # that write to database
                                    # Regexp are accepted
black_query_pattern_list = ''
                                    # Semicolon separated list of query patterns
                                    # that should be sent to primary node
                                    # Regexp are accepted
```

```
                                    # valid for streaming replicaton mode only.
database_redirect_preference_list = ''
                                    # comma separated list of pairs of database and node id.
                                    # example: postgres:primary,mydb[0-4]:1,mydb[5-9]:2'
                                    # valid for streaming replicaton mode only.
app_name_redirect_preference_list = ''
                                    # comma separated list of pairs of app name and node id.
                                    # example: 'psql:primary,myapp[0-4]:1,myapp[5-9]:standby'
                                    # valid for streaming replicaton mode only.
allow_sql_comments = off
                                    # if on, ignore SQL comments when judging if load balance or
                                    # query cache is possible.
                                    # If off, SQL comments effectively prevent the judgment
                                    # (pre 3.4 behavior).
disable_load_balance_on_write = 'transaction'
                                    # Load balance behavior when write query is issued
                                    # in an explicit transaction.
                                    # Note that any query not in an explicit transaction
                                    # is not affected by the parameter.
                                    # 'transaction' (the default): if a write query is issued,
                                    # subsequent read queries will not be load balanced
                                    # until the transaction ends.
                                    # 'trans_transaction': if a write query is issued,
                                    # subsequent read queries in an explicit transaction
                                    # will not be load balanced until the session ends.
                                    # 'always': if a write query is issued, read queries will
                                    # not be load balanced until the session ends.
statement_level_load_balance = off
                                    # Enables statement level load balancing
#------------------------------------------------------------------------------
# MASTER/SLAVE MODE
#------------------------------------------------------------------------------
master_slave_mode = on
                                    # Activate master/slave mode
                                    # (change requires restart)
master_slave_sub_mode = 'stream'
                                    # Master/slave sub mode
                                    # Valid values are combinations stream, slony
                                    # or logical. Default is stream.
                                    # (change requires restart)
# - Streaming -
sr_check_period = 3
                                    # Streaming replication check period
                                    # Disabled (0) by default
sr_check_user = 'nobody'
                                    # Streaming replication check user
                                    # This is neccessary even if you disable streaming
                                    # replication delay check by sr_check_period = 0
sr_check_password = ''
                                    # Password for streaming replication check user
                                    # Leaving it empty will make Pgpool-II to first look for the
                                    # Password in pool_passwd file before using the empty password

sr_check_database = 'postgres'
                                    # Database name for streaming replication check
delay_threshold = 512000
                                    # Threshold before not dispatching query to standby node
                                    # Unit is in bytes
                                    # Disabled (0) by default
#------------------------------------------------------------------------------
```

```
# HEALTH CHECK GLOBAL PARAMETERS
#------------------------------------------------------------------------------
health_check_period = 5
                                    # Health check period
                                    # Disabled (0) by default
health_check_timeout = 10
                                    # Health check timeout
                                    # 0 means no timeout
health_check_user = 'nobody'
                                    # Health check user
health_check_password = ''
                                    # Password for health check user
                                    # Leaving it empty will make Pgpool-II to first look for the
                                    # Password in pool_passwd file before using the empty password

health_check_database = ''
                                    # Database name for health check. If '', tries 'postgres' frist
,
health_check_max_retries = 60
                                    # Maximum number of times to retry a failed health check before
giving up.
health_check_retry_delay = 1
                                    # Amount of time to wait (in seconds) between retries.
connect_timeout = 10000
                                    # Timeout value in milliseconds before giving up to connect to
backend.
                                    # Default is 10000 ms (10 second). Flaky network user may want
to increase
                                    # the value. 0 means no timeout.
                                    # Note that this value is not only used for health check,
                                    # but also for ordinary conection to backend.
#------------------------------------------------------------------------------
# FAILOVER AND FAILBACK
#------------------------------------------------------------------------------
failover_on_backend_error = off
                                    # Initiates failover when reading/writing to the
                                    # backend communication socket fails
                                    # If set to off, pgpool will report an
                                    # error and disconnect the session.
relcache_expire = 0  # After the configuration file is restructured, we recommend that you set thi
s parameter to 1, reload the configuration file, and then set this parameter to 0 again. You can a
lso set this parameter to a specific point in time.
                                    # Life time of relation cache in seconds.
                                    # 0 means no cache expiration(the default).
                                    # The relation cache is used for cache the
                                    # query result against PostgreSQL system
                                    # catalog to obtain various information
                                    # including table structures or if it's a
                                    # temporary table or not. The cache is
                                    # maintained in a pgpool child local memory
                                    # and being kept as long as it survives.
                                    # If someone modify the table by using
                                    # ALTER TABLE or some such, the relcache is
                                    # not consistent anymore.
                                    # For this purpose, cache_expiration
                                    # controls the life time of the cache.
relcache_size = 8192
                                    # Number of relation cache
                                    # entry. If you see frequently:
                                    # "pool_search_relcache: cache replacement happend"
```

```
                                    # in the pgpool log, you might want to increate this number.
```

3. Run the `cd /etc/pgpool-II-12` command to configure the pool_passwd file.

> **Note** If you connect to your ApsaraDB RDS instances by using Pgpool, you must configure the pool_passwd file. This is because Pgpool supports the authentication protocol of PostgreSQL.

```
# Run the following command:
#pg_md5 --md5auth --username=username password
# Generate the passwords of the digoal and nobody users. The passwords are automatically written i
nto the pool_passwd file.
pg_md5 --md5auth --username=digoal "xxxxxxx"
pg_md5 --md5auth --username=nobody "xxxxxxx"
```

4. Use the system to automatically generate the pool_passwd file.

```
cd /etc/pgpool-II-12
cat pool_passwd
```

5. Run the following commands to configure the pgpool_hba file:

```
cd /etc/pgpool-II-12
cp pool_hba.conf.sample pool_hba.conf
vi pool_hba.conf
```

Configure the following parameters:

```
host all all 0.0.0.0/0 md5
```

6. Configure the pcp.conf file.

> **Note** The pcp.conf file is used to manage the users and passwords of Pgpool. It is not related to the users and passwords of your ApsaraDB RDS instances.

```
cd /etc/pgpool-II-12
# pg_md5 abc  # In this command, you set the password to abc and encrypt it by using the MD5 encry
ption algorithm.
900150983cd24fb0d6963f7d28e17f72
cp pcp.conf.sample pcp.conf
vi pcp.conf
# USERID:MD5PASSWD
manage:900150983cd24fb0d6963f7d28e17f72  # In this command, the manage user is used to manage PCP.
```

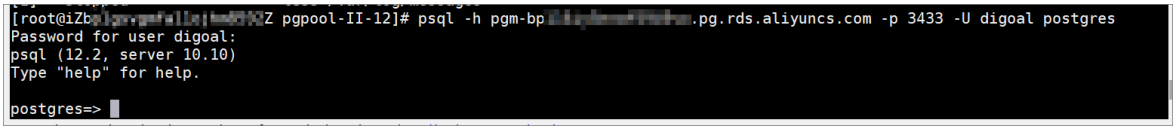
7. Start Pgpool.

```
cd /etc/pgpool-II-12
pgpool -f ./pgpool.conf -a ./pool_hba.conf -F ./pcp.conf
```

> **Note** If you want to view the logs of Pgpool, run the following command:
>
> ```
> less /var/log/messages
> ```

8. Use Pgpool to connect to your ApsaraDB RDS instances.

```
psql -h 127.0.0.1 -p 8001 -U digoal postgres
```



FAQ

- Q: How do I test whether read/write splitting is enabled?

  A: You can connect to your ApsaraDB RDS instances by using Pgpool and call the pg_is_in_recovery() function. Then, close the connection, establish a connection again, and call the pg_is_in_recovery() function again. If you receive a value of false and then a value of true, Pgpool routes requests to your primary ApsaraDB RDS instance and then to your read-only ApsaraDB RDS instances, and read/write splitting is enabled.

- Q: Does Pgpool increase the latency?

  A: Pgpool increases the latency slightly. In the test environment you set up in this topic, the latency increases by about 0.12 milliseconds.

- Q: How does Pgpool check the latency and health on my read-only ApsaraDB RDS instances?
  - A: If the WAL replay latency on a read-only ApsaraDB RDS instance exceeds the specified limit, Pgpool stops routing SQL requests to the read-only instance. Pgpool resumes routing SQL requests to the read-only instance only after it detects that the WAL replay latency on the read-only instance falls below the specified limit.

    > **Note** Connect to your primary ApsaraDB RDS instance and query the location where the current WAL data record is written. This location is referred to as log sequence number (LSN) 1. Then, connect to a read-only ApsaraDB RDS instance and query the location where the current WAL data record is replayed. This location is referred to as LSN 2. You can obtain the number of bytes between LSN 1 and LSN 2. This number indicates the latency.

  - Pgpool monitors the health of your read-only ApsaraDB RDS instances. If a read-only instance is unhealthy, Pgpool stops routing requests to the read-only instance.

- Q: How do I stop Pgpool and reload the configuration of Pgpool?

  A: Run the `pgpool --help` command to obtain more information about the commands used in Pgpool. Example:

  ```
  cd /etc/pgpool-II-12
  pgpool -f ./pgpool.conf -m fast stop
  ```

- Q: How do I configure Pgpool if more than one read-only ApsaraDB RDS instance is attached to my primary ApsaraDB RDS instance?

  A: Add the configurations of all the attached read-only ApsaraDB RDS instances to the pgpool.conf file. Example:

  ```
  backend_hostname1 = 'xx.xx.xxx.xx'
  backend_port1 = 8002
  backend_weight1 = 1
  backend_data_directory1 = '/data01/pg12_8002/pg_root'
  backend_flag1 = 'DISALLOW_TO_FAILOVER'
  backend_application_name1 = 'server1'
  backend_hostname2 = 'xx.xx.xx.xx'
  backend_port1 = 8002
  backend_weight1 = 1
  backend_data_directory1 = '/data01/pg12_8002/pg_root'
  backend_flag1 = 'DISALLOW_TO_FAILOVER'
  backend_application_name1 = 'server1'
  ```

- Q: How do I use pcp commands to view the status of my read-only ApsaraDB RDS instances?

A: To obtain the status of your read-only ApsaraDB RDS instances by using pcp commands, run the following command:

```
# pcp_node_info -U manage -h /tmp -p 9898 -n 1 -v
Password: # Enter the password.
Hostname               : 127.0.0.1
Port                   : 8002
Status                 : 2
Weight                 : 0.500000
Status Name            : up
Role                   : standby
Replication Delay      : 0
Replication State      :
Replication Sync State :
Last Status Change     : 2020-02-29 00:20:29
```

- Q: Which listening ports are used by Pgpool for read/write splitting?

  A: The following listening ports are used by Pgpool for read/write splitting:

  - Primary ApsaraDB RDS instance: Port 3389
  - Secondary ApsaraDB RDS instance: Port 8002
  - Pgpool: Port 8001
  - PCP: Port 9898

11.1.18. Use ShardingSphere to develop ApsaraDB RDS for PostgreSQL

ShardingSphere is an open source ecosystem that consists of a set of distributed database middleware solutions.

Prerequisites

All PostgreSQL versions used with ApsaraDB RDS support ShardingSphere.

Context

ApsaraDB RDS for PostgreSQL supports database-integrated sharding plug-ins (such as Citus, Postgres-XC, and AntDB) and massively parallel processing (MPP) products. It also supports sharding middleware products that are similar to those widely used in MySQL, such as ShardingSphere.

ShardingSphere is suitable for services that run in databases with thorough, well-organized logical sharding. It offers the following features:

- Data sharding
  - Database and table sharding
  - Read/write splitting
  - Sharding strategy customization
  - Decentralized distributed primary key
- Distributed transaction
  - Unified transaction API
  - XA transaction
  - BASE transaction
- Database orchestration
  - Dynamic configuration
  - Orchestration and governance

- Data encryption
- Tracing and observability
- Elastic scaling out (planning)

For more information, visit the ShardingSphere documentation.

ShardingSphere products

ShardingSphere includes three independent products. You can choose the product that best suits your business requirements. The following table describes these products.

Parameter	Sharding-JDBC	Sharding-Proxy	Sharding-Sidecar
Supported database engine	All JDBC-compatible database engines such as MySQL, PostgreSQL, Oracle, and SQL Server	MySQL and PostgreSQL	MySQL and PostgreSQL
Connections consumed	High	Low	High
Supported heterogeneous language	Java	All	All
Performance	Low consumption	Moderate consumption	Low consumption
Decentralized	Yes	No	Yes
Stateless API	No	Yes	No

Prepare configuration templates

1. On your ECS instance, run the following command to go to the directory where configuration templates are stored. The directory is under the root directory in this example.

   ```
   cd /root/apache-shardingsphere-incubating-4.0.0-sharding-proxy-bin/conf
   ```

2. Run the `ll` command to view all files stored in the directory: Command output:

   ```
   total 24
   -rw-r--r-- 1 501 games 3019 Jul 30  2019 config-encrypt.yaml
   -rw-r--r-- 1 501 games 3582 Apr 22  2019 config-master_slave.yaml
   -rw-r--r-- 1 501 games 4278 Apr 22  2019 config-sharding.yaml
   -rw-r--r-- 1 501 games 1918 Jul 30  2019 server.yaml
   ```

   Note
   - config-encrypt.yaml: the data encryption configuration file.
   - config-master_slave.yaml: the read/write splitting configuration file.
   - config-sharding.yaml: the data sharding configuration file.
   - server.yaml: the common configuration file.

3. Modify the configuration files.

Note For more information about the configuration files, visit the ShardingSphere documentation. In this example, the data sharding and common configuration files are used.

- Example of a data sharding configuration file:

```
schemaName: # The name of the logical data source.
dataSources: # The configuration of the data source. You can configure more than one da
ta source by using the data_source_name variable.
  <data_source_name>: # You do not need to configure a database connection pool. This i
s different in Sharding-JDBC.
    url: # The URL used to connect to your database.
    username: # The username used to log on to the database.
    password: # The password used to log on to the database.
    connectionTimeoutMilliseconds: 30000 # The connection timeout period in millisecond
s.
    idleTimeoutMilliseconds: 60000 # The idle-connection reclaiming timeout period in m
illiseconds.
    maxLifetimeMilliseconds: 1800000 # The maximum connection time to live (TTL) in mil
liseconds.
    maxPoolSize: 65 # The maximum number of connections allowed.
shardingRule: # You do not need to configure a sharding rule, because it is the same in
Sharding-JDBC.
```

- Example of a common configuration file:

```
Proxy properties
# You do not need to configure proxy properties that are the same in Sharding-JDBC
props:
  acceptor.size: # The number of worker threads that receive requests from the client.
The default number is equal to the number of CPU cores multiplied by 2.
  proxy.transaction.type: # The type of transaction processed by the proxy. Valid value
s: LOCAL | XA | BASE. Default value: LOCAL. Value XA specifies to use Atomikos as the t
ransaction manager. Value BASE specifies to copy the .jar package that implements the S
hardingTransactionManager operation to the lib directory.
  proxy.opentracing.enabled: # Specifies whether to enable link tracing. Link tracing i
s disabled by default.
  check.table.metadata.enabled: # Specifies whether to check the consistency of metadat
a among sharding tables during startup. Default value: false.
  proxy.frontend.flush.threshold: # The number of packets returned in a batch during a
complex query.
Permission verification
This part of the configuration is used to verify your permissions when you attempt to l
og on to Sharding-Proxy. After you configure the username, password, and authorized dat
abases, you must use the correct username and password to log on to Sharding-Proxy from
the authorized databases.
authentication:
  users:
    root: # The username of the root user.
      password: root# The password of the root user.
    sharding: # The username of the sharding user.
      password: sharding# The password of the sharding user.
      authorizedSchemas: sharding_db, masterslave_db # The databases in which the speci
fied user is authorized. If you want to specify more than one database, separate them w
ith commas (,). You are granted the permissions of the root user by default. This way,
you can access all databases.
```

Set up a test environment

- On your ECS instance, install Java.

```
yum install -y java
```

- Configure an ApsaraDB RDS instance that runs PostgreSQL 10.
  - Create an account with username r1.
  - Set the password of the account to "PW123321!".
  - Create the following databases whose owners are user r1: db0, db1, db2, and db3.
  - Add the IP address of your ECS instance to an IP address whitelist of the ApsaraDB RDS for PostgreSQL instance.

> **Note**
> - For more information about how to create an ApsaraDB RDS for PostgreSQL instance, database, and account, see Create an instance and Create a database and an account.
> - For more information about how to configure an IP address whitelist, see Configure an IP address whitelist.

- Run `vi /root/apache-shardingsphere-incubating-4.0.0-sharding-proxy-bin/conf/server.yaml` to configure the following common configuration file:

```
  authentication:
  users:
    r1:
      password: PW123321!
      authorizedSchemas: db0,db1,db2,db3
props:
  executor.size: 16
  sql.show: false
```

Test horizontal sharding

1. Run `vi /root/apache-shardingsphere-incubating-4.0.0-sharding-proxy-bin/conf/config-sharding.yaml` to modify the following data sharding configuration file:

```
schemaName: sdb
dataSources:
  db0:
    url: jdbc:postgresql://pgm-bpxxxxx.pg.rds.aliyuncs.com:1433/db0
    username: r1
    password: PW123321!
    connectionTimeoutMilliseconds: 30000
    idleTimeoutMilliseconds: 60000
    maxLifetimeMilliseconds: 1800000
    maxPoolSize: 65
  db1:
    url: jdbc:postgresql://pgm-bpxxxxx.pg.rds.aliyuncs.com:1433/db1
    username: r1
    password: PW123321!
    connectionTimeoutMilliseconds: 30000
    idleTimeoutMilliseconds: 60000
    maxLifetimeMilliseconds: 1800000
    maxPoolSize: 65
  db2:
    url: jdbc:postgresql://pgm-bpxxxxx.pg.rds.aliyuncs.com:1433/db2
    username: r1
    password: PW123321!
    connectionTimeoutMilliseconds: 30000
    idleTimeoutMilliseconds: 60000
    maxLifetimeMilliseconds: 1800000
```

```
    maxPoolSize: 65
  db3:
    url: jdbc:postgresql://pgm-bpxxxxx.pg.rds.aliyuncs.com:1433/db3
    username: r1
    password: PW123321!
    connectionTimeoutMilliseconds: 30000
    idleTimeoutMilliseconds: 60000
    maxLifetimeMilliseconds: 1800000
    maxPoolSize: 65
shardingRule:
  tables:
    t_order:
      actualDataNodes: db${0..3}.t_order${0..7}
      databaseStrategy:
        inline:
          shardingColumn: user_id
          algorithmExpression: db${user_id % 4}
      tableStrategy:
        inline:
          shardingColumn: order_id
          algorithmExpression: t_order${order_id % 8}
      keyGenerator:
        type: SNOWFLAKE
        column: order_id
    t_order_item:
      actualDataNodes: db${0..3}.t_order_item${0..7}
      databaseStrategy:
        inline:
          shardingColumn: user_id
          algorithmExpression: db${user_id % 4}
      tableStrategy:
        inline:
          shardingColumn: order_id
          algorithmExpression: t_order_item${order_id % 8}
      keyGenerator:
        type: SNOWFLAKE
        column: order_item_id
  bindingTables:
    - t_order,t_order_item
  defaultTableStrategy:
    none:
```

2. Start ShardingSphere and listen to Port 8001.

```
cd /root/apache-shardingsphere-incubating-4.0.0-sharding-proxy-bin/bin/
./start.sh 8001
```

3. Connect to the destination database.

```
psql -h 127.0.0.1 -p 8001 -U r1 sdb
```

4. Create a table.

```
create table t_order(order_id int8 primary key, user_id int8, info text, c1 int, crt_time timestam
p);
create table t_order_item(order_item_id int8 primary key, order_id int8, user_id int8, info text,
c1 int, c2 int, c3 int, c4 int, c5 int, crt_time timestamp);
```

> **Note** When you create a table, the system creates horizontal shards in the destination database based on the sharding strategy that you specify.

FAQ

- If you want to view the SQL parsing and routing statements used in ShardingSphere, run `vi /root/apache-shardingsphere-incubating-4.0.0-sharding-proxy-bin/conf/server.yaml`.

```
  authentication:
  users:
    r1:
      password: PW123321!
      authorizedSchemas: db0,db1,db2,db3
props:
  executor.size: 16
  sql.show: true # Specifies to log parsed SQL statements.
```

- If you want to test writes and queries, run the following commands:

```
insert into t_order (user_id, info, c1, crt_time) values (0,'a',1,now());
insert into t_order (user_id, info, c1, crt_time) values (1,'b',2,now());
insert into t_order (user_id, info, c1, crt_time) values (2,'c',3,now());
insert into t_order (user_id, info, c1, crt_time) values (3,'c',4,now());
select * from t_order;
```

  The following result is returned in this example:

```
      order_id      | user_id | info | c1 |          crt_time
--------------------+---------+------+----+----------------------------
 433352561047633921 |       0 | a    |  1 | 2020-02-09 19:48:21.856555
 433352585668198400 |       1 | b    |  2 | 2020-02-09 19:48:27.726815
 433352610813050881 |       2 | c    |  3 | 2020-02-09 19:48:33.721754
 433352628370407424 |       3 | c    |  4 | 2020-02-09 19:48:37.907683
(4 rows)
```

- If you want to view ShardingSphere logs, run the following command:

```
/root/apache-shardingsphere-incubating-4.0.0-sharding-proxy-bin/logs/stdout.log
```

- If you want to use pgbench for stress testing, run the following commands:

```
vi test.sql
\set user_id random(1,100000000)
\set order_id random(1,2000000000)
\set order_item_id random(1,2000000000)
insert into t_order (user_id, order_id, info, c1 , crt_time) values (:user_id, :order_id,random()::
text, random()*1000, now()) on conflict (order_id) do update set info=excluded.info,c1=excluded.c1,
crt_time=excluded.crt_time;
insert into t_order_item (order_item_id, user_id, order_id, info, c1,c2,c3,c4,c5,crt_time) values (
:order_item_id, :user_id,:order_id,random()::text, random()*1000,random()*1000,random()*1000,random
()*1000,random()*1000, now()) on conflict(order_item_id) do nothing;
pgbench -M simple -n -r -P 1 -f ./test.sql -c 24 -j 24  -h 127.0.0.1 -p 8001 -U r1 sdb -T 120
progress: 1.0 s, 1100.9 tps, lat 21.266 ms stddev 6.349
progress: 2.0 s, 1253.0 tps, lat 18.779 ms stddev 7.913
progress: 3.0 s, 1219.0 tps, lat 20.083 ms stddev 13.212
```

12.Cloud Native Distributed Database PolarDB-X

12.1. User Guide (1.0)

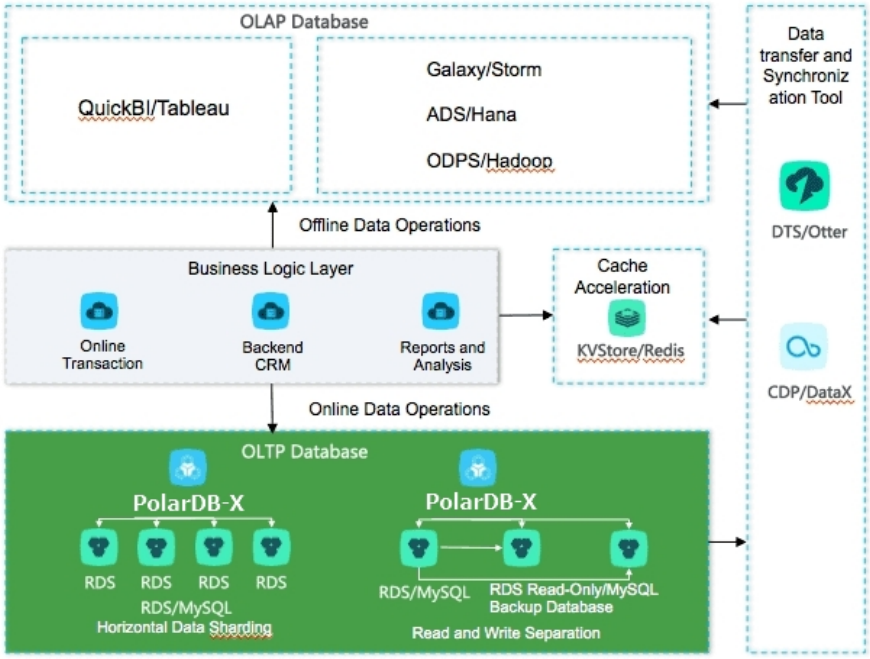
12.1.1. What is PolarDB-X?

is a database product that is developed by Alibaba Group and focuses on scaling out single-instance relational databases. This service is compatible with former Distributed Relational Database Service (DRDS).

Compatible with the MySQL communication protocols, supports most MySQL data manipulation language (DML) and data definition language (DDL) syntax. It provides the core capabilities and features of distributed databases, such as sharding, smooth scale-out, service upgrade and downgrade, and transparent read/write splitting. PolarDB-X is lightweight (stateless), flexible, stable, and efficient, and provides you with O&M capabilities throughout the lifecycle of distributed databases.

is used for operations on large-scale online data. PolarDB-X maximizes the operation efficiency by partitioning data in specific business scenarios. This meets the requirements of online business on relational databases in an effective way.

Figure of the architecture



Fixed issues

- Capacity bottlenecks of single-instance databases: As the data volume and access increase, traditional single-instance databases encounter great challenges that cannot be solved by hardware upgrades in a complete way. In distributed database solutions, multiple instances work in a joint way. This resolves the bottlenecks of data storage capacity and access volumes in an effective way.
- Difficult scale-out of relational databases: Due to the inherent attributes of distributed databases, you can change the shards where data is stored through smooth data migration. This way, the dynamic scale-out of relational databases is achieved.

12.1.2. Quick start

This topic describes how to get started with .

A instance is physically a distributed cluster that consists of multiple server nodes and underlying storage instances. A database is a logical concept and contains only metadata. Specific data is stored in the physical databases of the underlying storage instances. To get started with , perform the following steps:

1. Create a PolarDB-X instance.
2. Create a database.

   To create a database in a instance, you must select one or more ApsaraDB RDS for MySQL instances as the data storage nodes. If no ApsaraDB RDS for MySQL instance exists, create one first. For more information about how to create an ApsaraDB RDS for MySQL instance, see **User Guide > Instance management > Create an instance** in the *ApsaraDB RDS documentation*.
3. After a database is created, you must create tables in the database. This is similar to creating tables in a common standalone database. However, the syntax is different in the expression of data partitioning information in the table creation statement of . For more information about how to create a table, see Table creation syntax.

12.1.3. Log on to the PolarDB-X console

This topic describes how to log on to the console by using Google Chrome.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %
3. Click **Log On**.
4. In the top navigation bar, choose **Products > Database Services > Distributed Relational Database Service**.

12.1.4. Instance management

12.1.4.1. Create an instance

Before you use , create an instance. This topic describes how to create a instance.

1. Log on to the PolarDB-X console.

2. In the upper-right corner of the page, click **Create Instance**.
3. On the **Create PolarDB-X Instance** page, configure the parameters as needed.

   Parameters for creating an instance describes the parameters.

   Parameters for creating an instance

Category	Parameter	Description
Region	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
	Region	The region in which the instance is deployed. The instances in different regions cannot communicate with each other over an internal network. After an instance is created, you cannot change the region for the instance.
	Zone	The zone in which the instance is deployed.
Basic Settings	Instance Type	The type of the instance. Select an instance type from the options that are available on the page.
	Edition	The edition of the instance. Valid values: ○ Standard ○ Enterprise ○ Starter
	Specifications	The specifications of the instance. The specifications vary based on instance editions. Select the instance specifications from the options that are available on the page.
Network Type	Network Type	The network type of the instance. Valid values: ○ **Classic Network**: Cloud services on the classic network are not isolated from each other. Unauthorized access to a cloud service is blocked only by the security group or based on the whitelist policy of the service. ○ **VPC**: You can create a virtual private cloud (VPC) to build an isolated network environment on Apsara Stack. You can customize route tables, IP address ranges, and gateways in a VPC. We recommend that you select a VPC for improved security. If you set Network Type to VPC, you must set **VPC** and **vSwitch**. **Note** The instance and the Elastic Compute Service (ECS) instance to connect must use the same network type. Otherwise, they cannot communicate with each other over an internal network.
	VPC	Select a VPC. **Note** If you set **Network Type** to **VPC**, specify this parameter.
	vSwitch	Select a vSwitch in the VPC. **Note** If you set **Network Type** to **VPC**, specify this parameter.

4. Click **Submit**.

   After the instance is created, the instance appears in the instance list and the status of the instance changes to **Running**. The instance name is a unique identifier of an instance. You can identify a instance based on this unique identifier.

12.1.4.2. Change instance specifications

When you use , you can change the specifications of a instance as needed.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the **Common Actions** section or the **Configuration Information** section, click **Upgrade** or **Downgrade** to go to the Change Specifications page.

   > **Note** Alternatively, on the **DRDS Instance Management** page, choose **More > Downgrade** in the **Actions** column of the instance.

5. On the Change Specifications page, set **Instance Edition** and **Instance Specifications**, and click **Submit**.

   After a few minutes, you can view the new specifications of the PolarDB-X instance in the instance list.

   > **Note** Specifications downgrade leads to transient disconnections between applications and within a short period of time. Make sure that your applications can be automatically reconnected.

12.1.4.3. Create a non-integrated PolarDB-X instance

This topic describes how to create a non-integrated instance.

Context

instances are divided into integrated instances and non-integrated instances. The following list describes the differences between integrated instances and non-integrated instances:

- Integrated instances are instances that are integrated with the resource pool of ApsaraDB RDS for MySQL. You can manage your dedicated ApsaraDB RDS for MySQL instances in the console.
- Non-integrated instances are the ApsaraDB RDS for MySQL instances that are purchased in the ApsaraDB RDS console and added to the console. ApsaraDB RDS for MySQL instances must be managed in the ApsaraDB RDS for MySQL console.

is integrated with ApsaraDB RDS in and Apsara Stack Enterprise Edition V3.11 hot-feature and later. By default, all instances created in the PolarDB-X console are integrated instances. Apsara Stack provides an entry point that allows you to create non-integrated instances.

Operation notes

When you create a non-integrated PolarDB-X instance, you may need to pull an image. In this case, if an error such as a pull timeout occurs, you can contact Alibaba Cloud technical support.

Procedure

Contact Alibaba Cloud O&M personnel for assistance.

12.1.4.4. Read-only PolarDB-X instances

12.1.4.4.1. Overview

Read-only instances are an extension and supplement to primary instances and are compatible with SQL query syntax of primary instances.

Features

Read-only and primary instances can share the same replica of data. You can perform complex data query and analytics on read-only or primary ApsaraDB RDS for MySQL instances. Multiple instance types are provided to handle highly concurrent access requests and reduce the RT for complex queries. Read-only PolarDB-X instances provide resource isolation, which alleviates the load pressure on the primary PolarDB-X instances and reduces the link complexity of the business architecture. No additional data synchronization operations are required, and the O&M and budget costs are reduced.

Instance types

Concurrent read-only instances: For high-concurrency and high-traffic simple queries or offline data extraction, resource isolation helps you handle highly concurrent queries. This way, the stability of online business links is ensured.

> **Note** For the business for which primary instances are used, concurrent read-only instances can be used in the following scenarios:
>
> - High-concurrency and high-traffic simple queries are performed.
> - Data is extracted offline.

Limits

- Primary and read-only instances must be in the same region, but they can be in different zones.
- A read-only instance must belong to a primary instance. Before you create a read-only PolarDB-X instance, you must create a primary PolarDB-X instance. After you create a database on the primary instance, the database is replicated to the read-only instance. If you delete the database from the primary instance, the corresponding database on the read-only instance is also deleted.
- You are not allowed to migrate data to read-only instances.
- You are not allowed to create databases in or delete databases from read-only instances.
- Read-only instances cannot be cloned.
- Read-only instances support DDL statements but do not support DML statements for data modification.

12.1.4.4.2. Create a read-only PolarDB-X instance

This topic describes how to create a read-only PolarDB-X instance.

Procedure

1. Log on to the PolarDB-X console.
2. In the instance list, find the primary PolarDB-X instance for which you want to create a read-only instance.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. Click **Read-only Instance** on the right of the primary instance.
5. Specify Region, Basic Settings, and Network Type. Then, click **Submit**.

   Parameters for creating a read-only PolarDB-X instance

Section	Parameter	Description
Region	**Region**	The region where the instance is deployed. PolarDB-X instances in different regions cannot communicate with each other. After a PolarDB-X instance is created, the region cannot be changed.
	Zone	The zone where the PolarDB-X instance is deployed.
Basic Settings	**Instance Type**	The type of the read-only PolarDB-X instance. Select an instance type from the options that are available on the page.
	Instance Edition	The edition of the instance. Valid values: ○ Starter ○ Standard ○ Enterprise
	Instance Specifications	The specifications of the instance. The specifications vary based on instance editions. Select the instance specifications from the options that are available on the page.
	Describe	The description of the read-only PolarDB-X instance. We recommend that you provide an informative description to simplify future management operations.
Network Type	**Network Type**	The network type of the PolarDB-X instance. instances support the following network types: ○ **Classic Network**: Cloud services in the classic network are not isolated from each other. Unauthorized access to a cloud service can be blocked only by the security group or the whitelist policy of the service. ○ **VPC**: You can create a virtual private cloud (VPC) to build an isolated network environment on Apsara Stack. You can customize route tables, IP address ranges, and gateways in a VPC. We recommend that you select VPC for higher security. Select VPC for Network Type and then specify **VPC** and **vSwitch**. **Note** Make sure that the instance has the same network type as the Elastic Compute Service (ECS) instance to which you want to connect. If the PolarDB-X and ECS instances have different network types, they cannot communicate over an internal network.

6. It takes several minutes to create the instance. Wait until the instance is created. After the instance is created, it appears in the instance list in the console.

12.1.4.4.3. Manage a read-only PolarDB-X instance

Read-only instances are managed in a similar way as primary instances. However, databases cannot be created or deleted on the read-only instance management page. Databases on read-only instances are created when databases on primary instances are created. Databases on read-only instances are deleted when databases on primary instances are deleted. In the console, you can go to the read-only instance management page in two ways.

Manage a read-only PolarDB-X instance by its ID

1. Log on to the PolarDB-X console.
2. On the **PolarDB-X Instance Management** page, find the read-only instance that you want to manage.
3. Click the instance ID or choose **More > Manage** from the Actions column of the read-only instance that you

want to manage to access the **Basic Information** page.

Manage a read-only PolarDB-X instance by the ID of its primary instance

1.
2. On the **DRDS Instance Management** page, find the primary instance to which the read-only instance belongs.
3. Click the ID of the primary instance or choose **More > Manage** from the Actions column of the primary instance to access the **Basic Information** page.
4. On the **Basic Information** page, move the pointer over the number of read-only instances in the **Related Instances** section to view the ID of the read-only PolarDB-X instance.
5. Click the ID of the read-only PolarDB-X instance. The **Basic Information** page of the read-only instance appears.

12.1.4.4.4. Release a read-only PolarDB-X instance

If you no longer need a read-only instance, you can release it.

Prerequisites

The read-only instance must be in the **Running** state.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. In the PolarDB-X instance list, find the target instance, and choose **More > Release** from the Actions column.

   Notice You cannot recover the PolarDB-X instances that have been released. Exercise caution when you perform this operation.

4. In the **Release DRDS Instance** dialog box, click **OK**.

12.1.4.5. Restart a PolarDB-X instance

This topic describes how to restart a instance.

Prerequisites

The PolarDB-X instance must be in the **Running** state.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. Click **Restart Instance** in the upper-right corner.
5. In the **Restart Instance** dialog box, click **OK**.

   Notice Restarting a PolarDB-X instance terminates all its connections. Make appropriate service arrangements before you restart a PolarDB-X instance. Exercise caution when you perform this operation.

12.1.4.6. Release a PolarDB-X instance

This topic describes how to release a running PolarDB-X instance in the console.

Prerequisites

- All databases on the PolarDB-X instance have been deleted.
- The PolarDB-X instance must be in the Running state.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. In the PolarDB-X instance list, find the target instance, and choose **More > Release** from the Actions column.
4. In the **Release DRDS Instance** dialog box, click **OK**.

> **Warning** After the PolarDB-X instance is released, data is not deleted from its attached ApsaraDB RDS for MySQL instances. However, a released PolarDB-X instance cannot be restored. Exercise caution when you perform this operation.

12.1.4.7. Recover data

12.1.4.7.1. Backup and restoration

allows you to back up data of instances and databases, and restore them by using the backup data. Instances can be automatically and manually backed up. PolarDB-X provides fast backup and consistent backup. Existing backup sets are used to restore data in instances. Data is restored to the new and ApsaraDB RDS for MySQL instances based on the existing backup sets.

Considerations

- By default, the automatic backup policy of is disabled. You must manually enable it.
- The log backup capability of depends on underlying ApsaraDB RDS for MySQL instances. Therefore, the log backup policy configured in the console is automatically synchronized to all the underlying ApsaraDB RDS for MySQL instances. After the policy is configured, do not modify it in the ApsaraDB RDS for MySQL console. Otherwise, relevant data backup sets become invalid.
- The backup and restoration feature of depends on log backup. We recommend that you enable the log backup policy by default to prevent backup sets from becoming invalid.
- Data definition language (DDL) operations cannot be performed during the backup process. Otherwise, instance backup and restoration fail.
- During the backup, make sure that the underlying ApsaraDB RDS for MySQL instances for the instance are normal to prevent backup failures.
- Consistent backup and restoration is supported only in 5.3.8 and later.
- All the tables must have primary keys to ensure data accuracy during consistent backup and restoration.
- During consistent backup, distributed transactions on instances are locked for seconds. During the locking period, non-transactional SQL statements and non-distributed transactions can be run as expected. However, the commitment of distributed transactions is blocked and the response time (RT) for executing SQL statements may have millisecond-level jitters. We recommend that you perform consistent backup during off-peak hours.
- Due to changes in the inventory of and ApsaraDB RDS for MySQL, automatically adjusts the instance type and zone during instance restoration. We recommend that you confirm and adjust the instance type and zone after the instance restoration to avoid business disruption.

Backup methods

In different scenarios, provides fast backup and consistent backup and the related data restoration capabilities. The following table compares the two backup methods.

Backup method	Scenario	Advantage	Disadvantage
Fast backup	Applies to routine backup and restoration scenarios.	• It provides fast data backup and restoration. • It supports restoration at a point in time based on backup sets. • It supports all the instance versions.	It ensures data consistency only within a single ApsaraDB RDS for MySQL instance in sharding scenarios. However, it does not ensure global data consistency.
Consistent backup	It applies to backup and restoration for the financial industry and online core transactions that require high data consistency.	It ensures global data consistency in sharding scenarios.	• It features slow backup and restoration. • It supports data restoration by backup set but does not support restoration by time. • It is supported only in 5.3.8 and later. • During data backup, distributed transactions on instances are locked for seconds. During the locking period, the response time (RT) for executing SQL statements may have millisecond-level jitters. Therefore, we recommend that you perform consistent backup during off-peak hours.

12.1.4.7.2. Configure an automatic backup policy

provides the automatic backup feature. This topic describes how to configure an automatic backup policy.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Backup and Recovery**.
5. Click the **Backup Policy** tab and then click **Edit**.
6. In the **Backup Policy** dialog box, set parameters based on the business requirements, and click **OK**.

> **Notice** instances cannot back up logs. The configured log backup policy is applied to all the underlying ApsaraDB RDS for MySQL instances.

12.1.4.7.3. Configure local logs

You can use local logs and the backup and restoration feature or the SQL flashback feature of to accurately restore an instance or a database to the desired point in time. This topic describes how to configure local logs.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Backup and Recovery**.
5. Click the **Local Log Settings** tab and then click **Edit**.
6. In the **Local Binlog Settings** dialog box, specify the parameters based on the business requirements and click **OK**.

> **Notice** The local log settings are applied to all the underlying ApsaraDB RDS for MySQL instances.

12.1.4.7.4. Manual backup

In addition to automatic backup, also provides manual backup, which allows you to back up data at any time. This topic describes how to manually back up instances and databases.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Backup and Recovery**.
5. On the page that appears, click **Data Backup** on the right.
6. In the dialog box that appears, set Backup Method and Backup Level.
   - Backup Method can be set to **Fast Backup** or **Consistent Backup**. For more information about the differences between the two methods, see Backup methods.

     > **Notice** If you select Consistent Backup, distributed transactions are locked for seconds and the RT may have sub-second-level jitters. Therefore, we recommend that you perform consistent backup during off-peak hours.

   - Backup Level can be set to **Instance Backup** or **Database Backup**. You can select **Instance Backup** to back up the entire instance, or select **Database Backup** to back up a database as needed.
7. Click **OK**.

12.1.4.7.5. Restore data

You can use the data restoration feature of to restore an instance or a database to the time when the backup is created. You can perform this operation anytime. This topic describes how to restore the data of an instance or a database to a specific point in time.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the

Basic Information page.

4. In the left-side navigation pane, choose **Data Recovery > Backup and Recovery**.
5. On the page that appears, click **Data Recovery (Original Clone Instance)** on the right.
6. Select a restoration method.
   - **By Time**: Restore data to the selected point in time. You must specify **Restoration Time** and **Recovery Level**.

     > **Note** Only non-integrated instances support recovery **By Time**. For information about how to create a non-integrated instance, see Create a non-integrated PolarDB-X instance.

   - **By Backup Set**: Restore data from the selected backup file.

     > **Note** You can also click **Recover** in the Actions column of the backup set to restore data **by backup set**, as shown in the following figure.

7. Click **Precheck** to check whether a valid backup set is available for data restoration. If the precheck fails, the data cannot be restored.
8. Click **Enable** to go to the **Confirm Order** page.
9. Confirm the order details and click **Enable** to restore the data. You can view the data restoration progress in **Task Progress** in the upper-right corner of the page.

12.1.4.7.6. SQL flashback

12.1.4.7.6.1. Overview

provides the SQL flashback feature to recover data of particular rows.

When you mistakenly run an SQL statement such as INSERT, UPDATE, or DELETE on , provide the relevant SQL information to match the event in the binary log file and generate the corresponding recovery file. You can download the file and recover data as needed. SQL flashback automatically chooses **fuzzy match** or **exact match** to locate lost data caused by the error. For more information, see Exact match and fuzzy match and Rollback SQL statements and original SQL statements.

Features

- Easy-to-use: SQL flashback allows you to retrieve the lost data by entering required information about the corresponding SQL statement.
- Fast and lightweight: Regardless of the backup policy of ApsaraDB RDS for MySQL instances, you only need to enable log backup before an SQL statement error occurs.
- Flexible recovery: Rollback SQL statements and original SQL statements are available for different scenarios.
- Exact match: SQL flashback supports exact match of data about the corresponding SQL statement, which improves precision of data recovery.

Limits

- SQL flashback depends on the binary log retention time and the log backup feature of ApsaraDB RDS for MySQL must be enabled. Binary log files can be retained only for a certain period. Use SQL flashback to generate files for recovery as soon as possible when an error occurs.
- The recovery files generated by SQL flashback are retained for seven days by default, and you need to download these files as soon as possible.
- The following conditions must be met for SQL flashback exact match:
  - The instance version is 5.3.4-15378085 or later.
  - The version of the ApsaraDB RDS for MySQL instance used by the database is 5.6 or later.

- SQL flashback exact match is enabled before the error SQL statement is executed.
- The TRACE_ID information for the error SQL statement is provided.

- To ensure the precision of data recovery, the exact match feature is enabled by default for the database created in a instance of 5.3.4-15378085 or later. After this feature is enabled, SQL execution information is included in the binary log file by default, which requires more storage space for ApsaraDB RDS for MySQL instances. If you need to use the exact match feature, we recommend that you upgrade before enabling the feature. For more information, see Enable exact match.

12.1.4.7.6.2. Generate a restoration file

This topic describes how to generate a restoration file in the PolarDB-X console.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > SQL Flashback**. The **SQL Flashback** page appears.
5. On the **SQL Flashback** page, enter the basic information about an incorrectly executed SQL statement, including Database, Time Range, Table Name, TRACE_ID, and SQL Statement Type. The following table describes the parameters.

Parameter	Description
Database	The database where the SQL statement was incorrectly executed.
Time Range	The time range during which the SQL statement was incorrectly executed. The beginning of the time range is earlier than the time when the SQL statement starts to be executed. The end of the time range is later than the time when the SQL statement execution ends. To ensure efficient restoration, we recommend that you limit the time range to 5 minutes.
Table Name	The name of the table on which the SQL statement was incorrectly executed. This parameter is optional.
TRACE_ID	The unique TRACE_ID that allocates for each executed SQL statement. You can obtain the TRACE_ID of the incorrectly executed SQL statement by using the SQL audit feature of .
SQL Statement Type	The type of the incorrectly executed SQL statement. Valid values: ○ INSERT ○ UPDATE ○ DELETE

6. Click **Precheck**. The system checks whether a binary log file exists within the specified time range. For more information about binary log files, see Configure local logs.

> **Note**
> - If no binary log file exists within the specified time range, the precheck fails and the system cannot restore the data for you.
> - If a binary log file exists within the specified time range, the precheck is successful and you can proceed to the next step.

7. Set SQL Statement Type for Recovery to **Rollback SQL** or **Original SQL Statement**. For more information about the differences between the two types, see Rollback SQL statements and original SQL statements.
8. Click **Generate SQL** to generate an SQL flashback task. The status of the SQL flashback tasks that are running on the current instance appears in the lower part of the page.

What's next

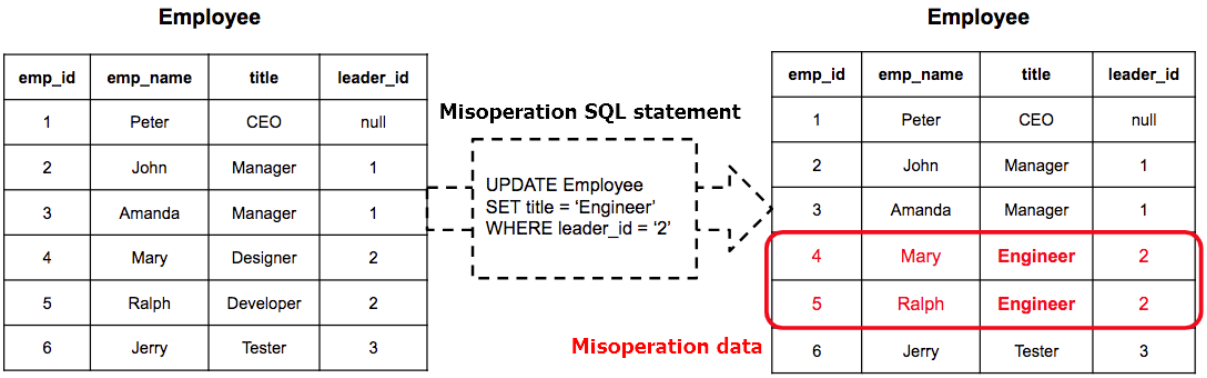
After an SQL flashback task is complete, the task information, such as the exact match status and the number of recovered rows, appears. You can click **Download** in the Actions column of the SQL flashback task to download the corresponding restoration file.

> **Notice** By default, the restoration file is retained for seven days. Download the file at the earliest opportunity.

12.1.4.7.6.3. Rollback SQL statements and original SQL statements

To support different business scenarios, SQL flashback provides rollback SQL statements and original SQL statements. Before generating an SQL statement for recovering data, you must select a corresponding recovery method based on your scenario.

Recovery methods



Recovery method	Description	Example
Rollback SQL statement	Traverses the events in the binary log file in reverse order to reverse the INSERT, UPDATE, and DELETE events. • The reverse of INSERT is equivalent to DELETE. • The reverse of DELETE is equivalent to INSERT. • The reverse of UPDATE is equivalent to the value before UPDATE.	UPDATE Employee SET title = 'Developer' WHERE emp_id = '5' UPDATE Employee SET title = 'Designer' WHERE emp_id = '4'

Recovery method	Description	Example
Original SQL statement	Traverses the events in the binary log file in order to mirror all records of the INSERT, UPDATE, and DELETE events. • An INSERT mirror is equivalent to INSERT. • A DELETE mirror is equivalent to INSERT. • An UPDATE mirror is equivalent to the value before INSERT.	INSERT INTO Employee(emp_id,emp_name,title,leader_id) values('4','Mary','Designer','2') INSERT INTO Employee(emp_id,emp_name,title,leader_id) values('5','Ralph','Developer','2')

12.1.4.7.6.4. Exact match and fuzzy match

SQL flashback supports **exact match** and **fuzzy match** for binary log events. You do not need to select a match policy. SQL flashback automatically detects and selects the optimal match policy, and notifies you when the flashback task is completed.

Match policy	Description	Advantage	Disadvantage
Exact match	The system performs exact match on the event of a mistaken SQL statement in the binary log file and generates a restoration file.	The restoration file contains only data that is deleted or modified due to the execution of the mistaken SQL statement. The file can be directly used to ensure the precision and efficiency of data restoration.	The following requirements must be met: • The instance version is 5.3.4-15378085 or later. • The version of the ApsaraDB RDS for MySQL instance used by databases is 5.6 or later. • Exact match of SQL flashback is enabled before mistaken SQL statements are executed. • The TRACE_ID of the mistaken SQL statement is provided.
Fuzzy match	The system matches the events in the binary log file based on the information about the mistaken SQL statement, including the time range, table name, and SQL statement type. Then, the system generates a restoration file.	Fuzzy match is supported on all instances, regardless of the instance version or parameter settings.	Data that is deleted or modified by the mistaken SQL statement cannot be exactly matched. The restoration file contains data changes that are made by other business SQL operations. You must filter the required data.

Enable exact match

> **Note** By default, fuzzy match is enabled.

1. Log on to the console, and go to the Parameter Settings page of the specified instance. For more information,

see Set parameters.

2. Change the value of **ENABLE_SQL_FLASHBACK_EXACT_MATCH** to **ON**.

12.1.4.7.7. Table recycle bin

12.1.4.7.7.1. Overview

The table recycle bin of allows you to recover mistakenly deleted tables.

After the table recycle bin is enabled for your database, the tables that are deleted by using the DROP TABLE statement are moved to the recycle bin and are no longer visible to you. After the tables are moved to the recycle bin for two hours, they are automatically cleared and cannot be recovered. You can view, recover, and clear the deleted tables in the recycle bin.

Limits and notes

- The table recycle bin feature is only supported by 5.3.3-1670435 and later. For more information, see View the instance version.
- The table recycle bin is disabled for your database by default. For more information about how to enable it, see Enable the table recycle bin.
- The table recycle bin of does not support the recovery of tables deleted by the TRUNCATE TABLE command.
- Tables in the recycle bin still occupy the storage space of ApsaraDB RDS for MySQL before they are automatically cleared. To release the storage space as soon as possible, you can access the recycle bin to manually delete them.

12.1.4.7.7.2. Enable the table recycle bin

This topic describes how to enable the table recycle bin.

Prerequisites

An ApsaraDB RDS for MySQL database has been created in the instance. For more information, see Create a database.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Table Recycle Bin**. The **Table Recycle Bin** page appears.
5. At the top of the **Table Recycle Bin** page, click the tab of the database for which the table recycle bin needs to be enabled.
6. Click **Enabled**.
7. In the dialog box that appears, click **OK**.

12.1.4.7.7.3. Restore tables

This topic describes how to restore your tables from the table recycle bin.

Procedure

1. Log on to the PolarDB-X console.

2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Table Recycle Bin**. The **Table Recycle Bin** page appears.
5. In the upper part of the **Table Recycle Bin** page, click the tab of the database for which you want to restore a table.
6. Click **Recover** in the **Actions** column of the table that you want to restore.

12.1.4.7.7.4. Delete tables

This topic describes how to delete unnecessary tables from the table recycle bin.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Table Recycle Bin**. The **Table Recycle Bin** page appears.
5. In the upper part of the **Table Recycle Bin** page, click the tab of the database in which you want to delete a table.
6. Click **Delete** in the **Actions** column of the table that you want to delete.

> **Note** To clear all tables from the table recycle bin, click **Empty Recycle Bin** on the tab of the corresponding database.

12.1.4.7.7.5. Disable the table recycle bin feature

If you no longer need the table recycle bin feature, you can disable it. This topic describes how to disable the table recycle bin feature.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Data Recovery > Table Recycle Bin**. The **Table Recycle Bin** page appears.
5. In the upper part of the **Table Recycle Bin** page, click the tab of the database for which you want to disable the table recycle bin feature.
6. Click **Disable** to disable the table recycle bin feature for the database.

12.1.4.8. Diagnostics and optimization

12.1.4.8.1. Query details about slow SQL queries

If the execution period for an SQL statement exceeds 1 second, defines the statement as a slow SQL statement.

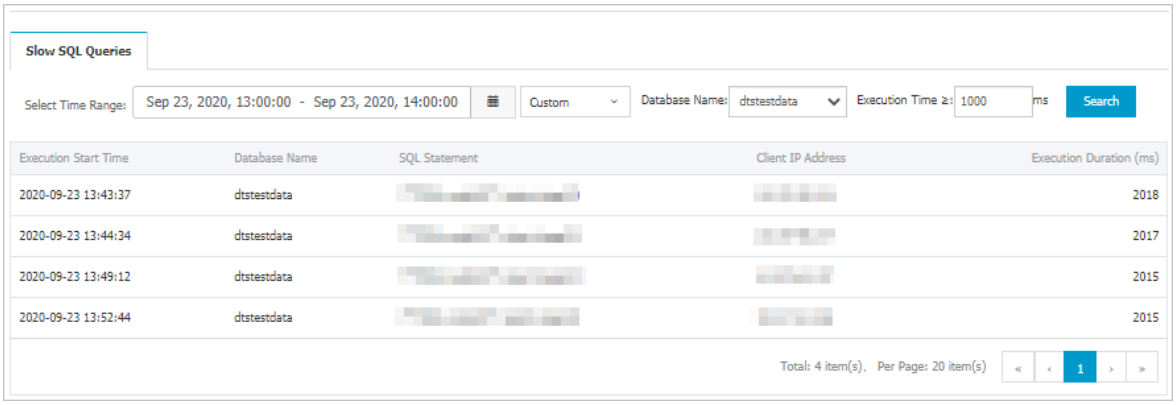
Considerations

Each instance stores the details about a maximum of 5,000 slow SQL queries. The details about extra slow SQL queries are deleted in a rolling way.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, click **Slow SQL Queries** under **Optimization and Diagnostics**.
5. Specify Select Time Range, Database Name, and Execution Time, and then click **Search** to view the information about slow SQL queries.

> **Note** You can view slow SQL details for up to seven days at a time.



12.1.4.8.2. Specify parameters

allows you to specify parameters for instances and databases. You can view and modify parameter values in the console based on your business requirements.

> **Note** You cannot specify parameters for read-only PolarDB-X instances.

Procedure

1. Log on to the PolarDB-X console.
2. In the instance list, find the PolarDB-X instance for which you want to specify parameters.
3. Click the instance ID or choose **More > Manage** from the Actions column of the instance to go to the **Basic Information** page.
4. In the left-side navigation pane, choose **Diagnostics and Optimization > Parameter Settings**. Click the Instance or Database tab to view parameters that you can modify for instances and databases. For more information about the parameters, see Parameter description.
5. Click the icon next to the parameter that you want to modify, enter the expected value, and then click **OK**.
6. Click **Submit** in the upper-right corner to submit the new value.

> **Note** To undo parameter modification, click **Cancel** in the upper-right corner.

Parameter description

Parameter	Level	Description
Slow SQL threshold	Instance	The threshold for slow SQL statements. The SQL statements whose execution duration exceeds this threshold are recorded in logical slow SQL logs.
Logical idle connection timeout	Instance	The logical time-out period of the idle connection between user applications and .Unit: millisecond.
Maximum packet size	Instance	The maximum size of a packet between user applications and .Unit: byte.
Limit on the instance memory pool size	Instance	The maximum size of the memory pool for an instance. If the memory usage on an instance exceeds the value, an error is reported and the query ends.
Whether to prohibit full-table deletion and update	Database	Specifies whether to disable full-table deletion or update.
Enable the recycle bin	Database	Specifies whether to enable the recycle bin for storing deleted logical tables of .
Temporary table size	Database	The size of the temporary table that is used during distributed queries in .Unit: row.
Number of unioned tables	Database	The maximum number of table shards that can be combined by the JOIN statement when you query multiple table shards in a database.
SQL query timeout period	Database	The time-out period of SQL statements for interaction between and ApsaraDB RDS for MySQL.Unit: millisecond. The value 0 indicates the time-out period is not limited.
SQL exact flashback switch	Database	Specifies whether to support exact match of SQL flashback. By default, exact match is disabled. After exact match is enabled, the information about query execution is added to the binary log file that is used by the database.
Enable logical INFORMATION_SCHEMA query	Database	Specifies whether to enable the logical INFORMATION_SCHEMA query. When this feature is enabled, it does not depend on the shadow database, and the aggregation results of logical databases and tables are returned. When this feature is disabled,it depends on the shadow database, and the information of physical databases and tables are returned.
Time period to start transaction log cleanup	Database	The period during which transaction log cleanup starts at a random time.
Limit on database-level memory pool size	Database	The maximum size of the database-level memory pool. When the memory usage of a database exceeds this value, an error is reported and the query terminates. The value -1 indicates that the maximum size of the database-level memory pool is not limited.
Limit on query-level memory pool size	Database	The maximum size of the query-level memory pool. When the memory usage of a query exceeds this value, an error is reported and the query terminates. The value -1 indicates that the maximum size of the query-level memory pool is not limited.

Parameter	Level	Description
Enable CBO	Database	Specifies whether to enable the cost-based optimizer (CBO) that provides features such as Join Reorder and Hash Join.
Default parallelism	Database	Specifies whether to enable parallel query, and the degree of parallelism after parallel query is enabled. This parameter takes effect only when CBO is enabled. The value -1 indicates that the degree of parallelism is automatically selected. The value 0 indicates that parallel query is disabled.
Whether to enable the asynchronous DDL engine	Database	Specifies whether to enable the asynchronous data definition language (DDL) engine. If you disable the engine, the original execution logic of the DDL engine is still used.
Enable the pure asynchronous mode after the asynchronous DDL engine is enabled	Database	Specifies whether to enable the pure asynchronous mode when the asynchronous DDL engine is enabled. • Enabled: The submitted DDL statement is executed immediately after the client connects to the instance. You can execute only asynchronous DDL management statements to view the execution status. • Disabled: The client still interacts with the instance in synchronous mode. The instance does not respond to the client until the client finishes executing the submitted DDL statement.
Maximum number of physical tables allowed to be created in a single physical database	Database	The maximum number of table shards that can be created in a database shard. If you create a logical table whose data is partitioned into table shards in a database shard, the number of table shards cannot exceed this value.
Aggregate statistics for INFORMATION_SCHEMA.TABLES	Database	Specifies whether to aggregate statistics for INFORMATION_SCHEMA.TABLES queries. By default, the statistics are not aggregated. This ensures high performance.
Maximum number of physical database shard connections	Database	The maximum number of connections between a database and a single ApsaraDB RDS for MySQL database shard.
Minimum number of physical database shard connections	Database	The minimum number of connections between a database and a single ApsaraDB RDS for MySQL database shard.
Physical idle connection timeout	Database	The time-out period for an idle connection between a database and an ApsaraDB RDS for MySQL database.Unit: minute.

12.1.4.9. SQL audit and analysis

12.1.4.9.1. Overview of SQL audit and analysis

collaborates with Log Service to enable the SQL audit and analysis feature. This feature allows you to audit historical SQL statements. You can also diagnose and analyze the execution status, performance metrics, and security risks of SQL statements in real time. You can enable the SQL audit and analysis feature in the console.

Benefits

- Ease of use: The SQL audit and analysis feature is easy to configure. You can use this feature to audit and

analyze SQL query logs in real time.

- Lossless performance: The SQL audit and analysis feature pulls SQL query logs from nodes and uploads the logs to Log Service in real time. This does not compromise the instance performance.
- Historical issue tracking: You can import historical SQL query logs to track issues.
- Real-time analysis: PolarDB-X collaborates with Log Service to enable the SQL audit and analysis feature and the out-of-the-box report center. The SQL audit and analysis feature allows you to analyze SQL queries in real time. The report center allows you to create custom reports and perform drill-down analysis. This way, you can gain a clear profile of your database, including the running status, performance, and security risks.
- Real-time alerts: You can configure custom quasi-real-time alerts based on specific metrics. This ensures quick response to exceptions in critical business.

Limits

- Before you use the SQL audit and analysis feature, enable Log Service.
- By default, audit logs for SQL queries are retained for 30 days. You can also modify the log retention period as needed.
- We recommend that you do not randomly delete or modify the default settings of the projects, Logstores, indexes, or dashboards in Log Service. Log Service updates and upgrades the SQL audit and analysis feature from time to time. The indexes and default reports of the dedicated Logstores are also automatically updated.
- A single SQL statement can be up to 5 MB.

Scenarios

- Troubleshoot problematic SQL queries

  After the SQL audit and analysis feature is enabled, you can retrieve SQL query logs to identify and troubleshoot problematic SQL queries. For example, you can execute the following statement to check whether a DROP operation is performed:

  ```
  sql_type: Drop
  ```

  The query result contains information about the time when the SQL query was performed, the user who performed the query, and the IP address of the client on which the SQL query was performed.

- Analyze costly SQL templates

  In most applications, SQL statements are dynamically generated based on templates. The parameter values in templates may vary based on SQL statements. The real-time analysis feature of Log Service allows you to obtain a list of costly SQL statements in a database.

  For example, you can perform the following queries:

  ```
  SELECT sql_code as "SQL template ID",
  round(total_time * 1.0 /sum(total_time) over() * 100,2) as "Execution time ratio (%)",
  execute_times as "Number of execution times",
  round(avg_time) as "Average execution time",
  round(avg_rows) as "Average number of affected rows",
  CASE WHEN length(sql) > 200 THEN  concat(substr(sql, 1, 200), '......') ELSE trim(lpad(sql, 200, '
  ')) end as "Sample SQL statement" FROM (SELECT sql_code, count(1) as execute_times,
  sum(response_time) as total_time,
  avg(response_time) as avg_time,
  avg(affect_rows) as avg_rows,
  arbitrary(sql) as sql FROM log GROUP BY sql_code) ORDER BY "Execution time ratio (%)" desc limit 10
  ```

  The query result contains the SQL template ID, ratio of the response time of the template-based SQL statement to the response time of all SQL statements, number of execution times, average execution time, average number of affected rows, and sample SQL statement. You can identify and optimize the most costly SQL template in the application based on the query result.

- Collect log statistics

collaborates with Log Service to enable the SQL audit and analysis feature and the out-of-the-box report center. This way, you can diagnose and analyze your database in real time, including the running status, performance, and potential security risks. The report center include **Operation Center**, **Performance Center**, and **Security Center**.

12.1.4.9.2. Enable SQL audit and analysis

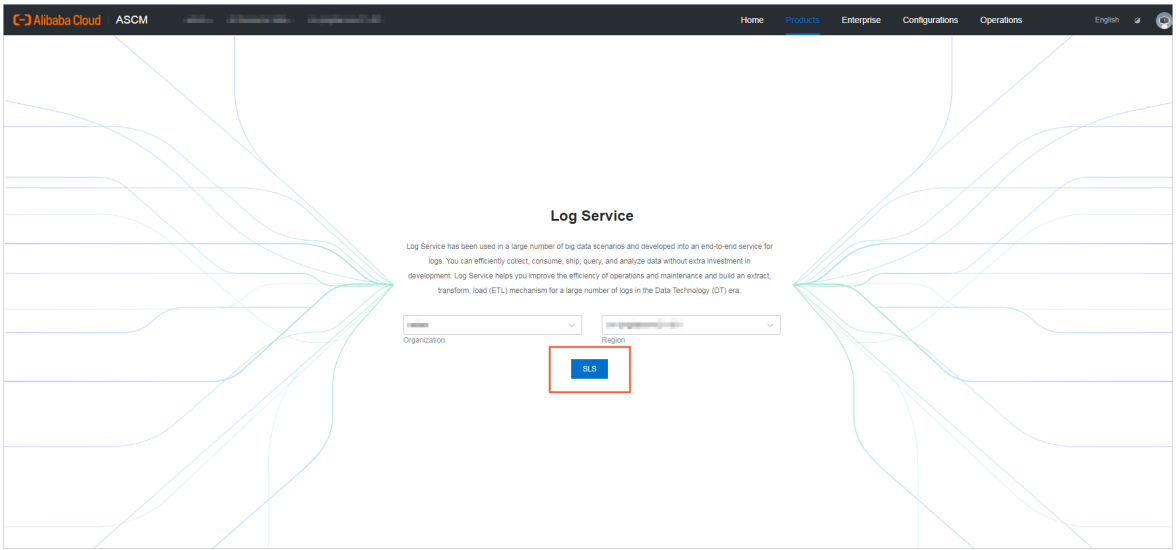
By default, the SQL audit and analysis feature is disabled. You can enable this feature in the console. By default, you can audit and analyze the log data that is generated after the SQL audit and analysis feature is enabled. You can also import part of historical data.

Procedure

1. Log on to the Log Service console. For more information, see *Log Service* documentation by choosing**User guide > Quick start > Log on to the Log Service console**.
2. Select the organization to which the instance belongs.

   > **Note** Use the same account to log on to Log Service and the instance.

3. Click **SLS** to go to the Log Service page.



4. Log on to the PolarDB-X console.
5. Find the target instance in the instance list.
6. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
7. In the left-side navigation pane, choose **Diagnosis and optimization > SQL audit and analysis**.
8. In the left-side section, select the database for which you want to enable the SQL audit and analysis feature.
9. On the right side of the SQL Audit and Analysis page, turn on the switch next to **SQL Audit Log Status of Current Database**.

   > **Note** On the left side of the SQL Audit and Analysis page, you can also turn on the **SQL Audit and Analysis** switch next to the database that you want to manage.

10. Determine whether to import historical data.

> **Note** By default, you can audit and analyze the log data that is generated after the SQL audit and analysis feature is enabled. If the SQL audit and analysis feature is disabled and the data in the database is modified, you can import historical data when you enable the SQL audit and analysis feature. This way, you can audit and analyze historical logs. This helps you track data tampering. dynamically checks the range of historical data that can be imported based on the log storage on the instance node. PolarDB-X allows you to import logs that are generated within the last seven days.
>
> - If you need to import historical data, turn on the **Import Historical Data or Not** switch, specify the start time and end time for operation tracking, and then click **Enable**.
> - If you do not need to import historical data, click **Enable**.

What's next

Every time you use the SQL audit and analysis feature, perform the preceding steps again.

12.1.4.9.3. Log fields

This topic describes the log fields that are used for SQL audit and analysis.

Field	Description	Supported version
__topic__	The topic of the log entry. The value follows the format `drds_audit_log_{instance_id}_{db_name}`, for example, `drds_audit_log_drdsxyzabcd_demo_drds_db`.	All versions
instance_id	The ID of the instance.	All versions
db_name	The name of the database in .	All versions
user	The username that is used to execute the SQL statement.	All versions
client_ip	The IP address of the client that connects to the instance.	All versions
client_port	The port that is used by the client to connect to the instance.	All versions
sql	The SQL statement that is executed.	All versions
trace_id	The ID that is used to trace the executed SQL statement. If you want to trace a transaction that is executed, use a value that consists of a trace ID, a hyphen, and a number, for example, `drdsabcdxyz-1` and `drdsabcdxyz-2`.	All versions
sql_code	The hash value of the template-based SQL statement.	All versions
hint	The hint in the executed SQL statement.	All versions

Field	Description	Supported version
table_name	The name of the table to query. Separate multiple tables with commas (,).	All versions
sql_type	The type of the SQL statement. Valid values: Select, Insert, Update, Delete, Set, Alter, Create, Drop, Truncate, Replace, and Other.	All versions
sql_type_detail	The name of the SQL parser.	All versions
sql_time	The start time for executing the SQL statement. The value follows the format `yyyy-MM-dd HH:mm:ss.SSS`.	All versions
response_time	The response time. Unit: ms.	V5.3.4-15378085 or later
affect_rows	The number of rows returned by the SQL statement. If you execute the INSERT, DELETE, or UPDATE statement, this field indicates the number of affected rows.	V5.3.4-15378085 or later
fail	The execution status of the SQL statement. Valid values: • 0: The SQL statement is executed. • 1: The SQL statement fails to be executed.	V5.3.4-15378085 or later

12.1.4.9.4. Log analysis

The SQL audit and analysis feature is based on Log Service (SLS) and provides powerful log analytics capabilities. This topic describes SQL statements for log analysis in common scenarios and provides relevant examples.

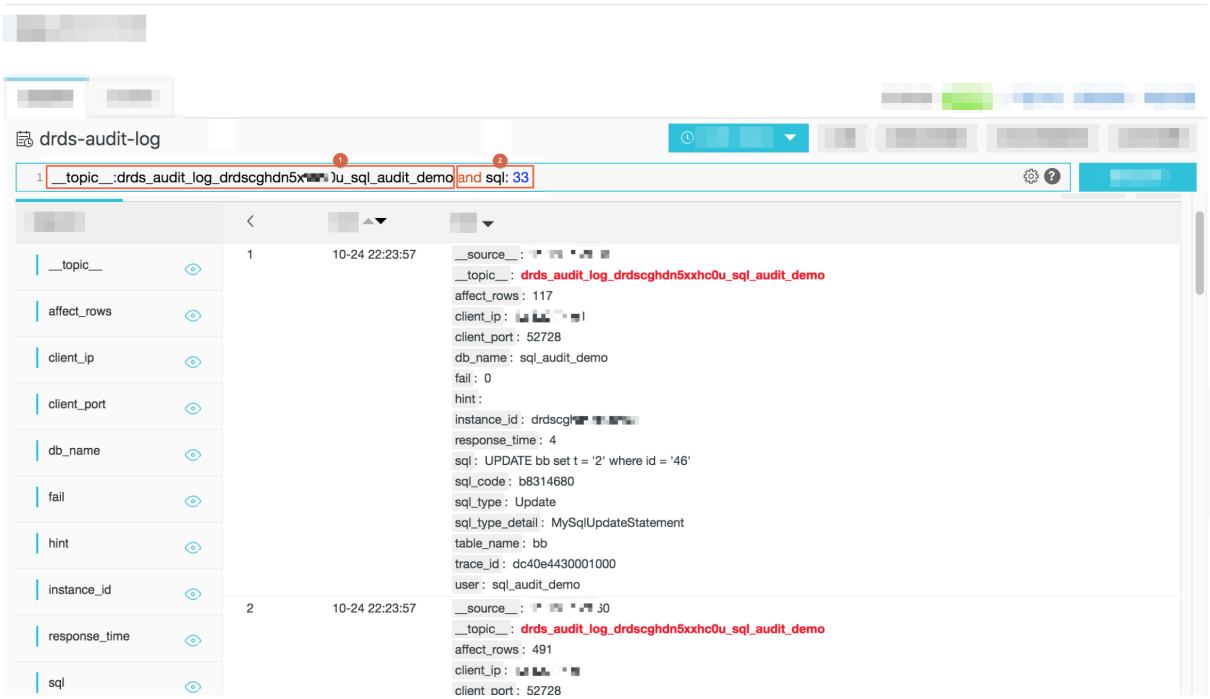
After the SQL audit and analysis feature is enabled, you can perform audit and analysis on SQL log files by using the query and analysis syntax of SLS on the SQL Audit and Analysis page. Based on the query and analysis syntax of SLS, you can find problematic SQL statements on the Log Analysis tab and analyze the SQL statement execution status, performance metrics, and security issues of . For more information about the query and analysis syntax of SLS, see *Log Service User Guide > Query and Analysis > Query Syntax and Functions > Query Syntax*.

Precautions

All the audit logs of databases in the same region are written to the same Logstore in SLS. Therefore, by default, the SQL Audit and Analysis page provides the filter conditions based on __topic__, to ensure that the searched SQL log files are from . Therefore, all the statements provided in this topic must be used after the existing filter conditions.

An example is shown in the following figure:

- The ① part is the default filter condition.
- The ② part is the additional filter condition.



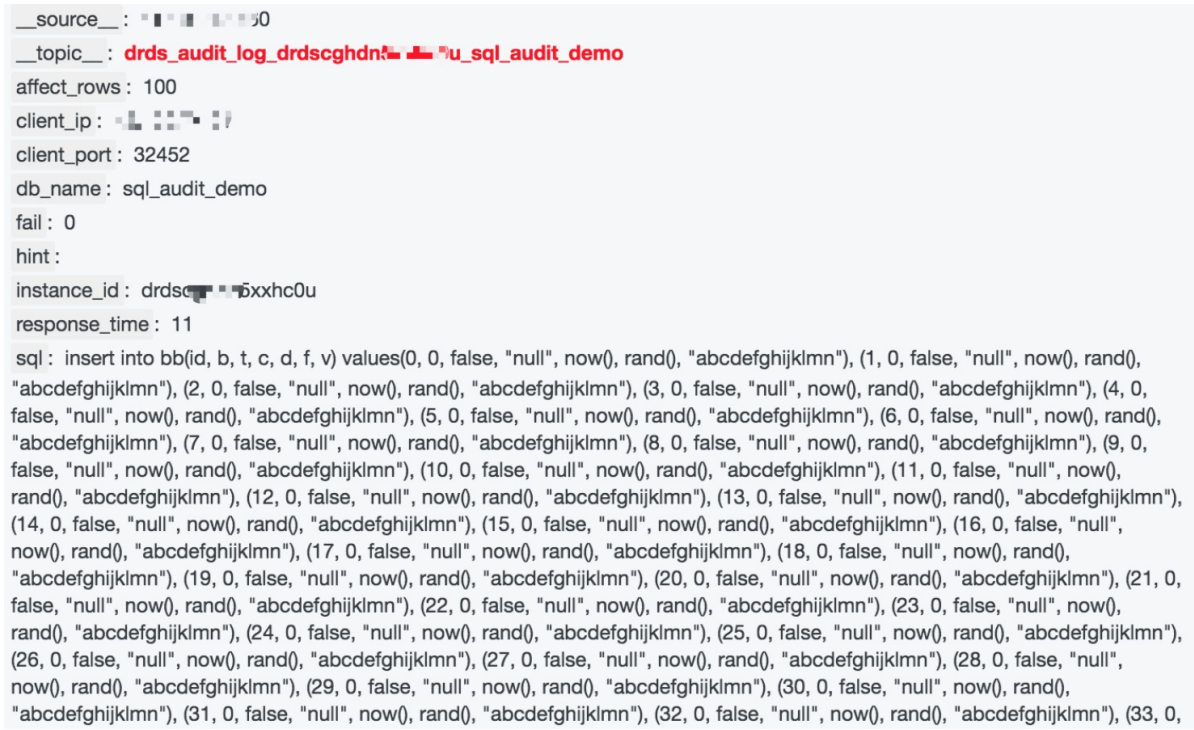
Find problematic SQL statements

- Fuzzy search

  For example, to query SQL statements that contain the "34" keyword, enter the following content in the search box:

  ```
  and sql: 34
  ```

  The result is shown in the following figure.

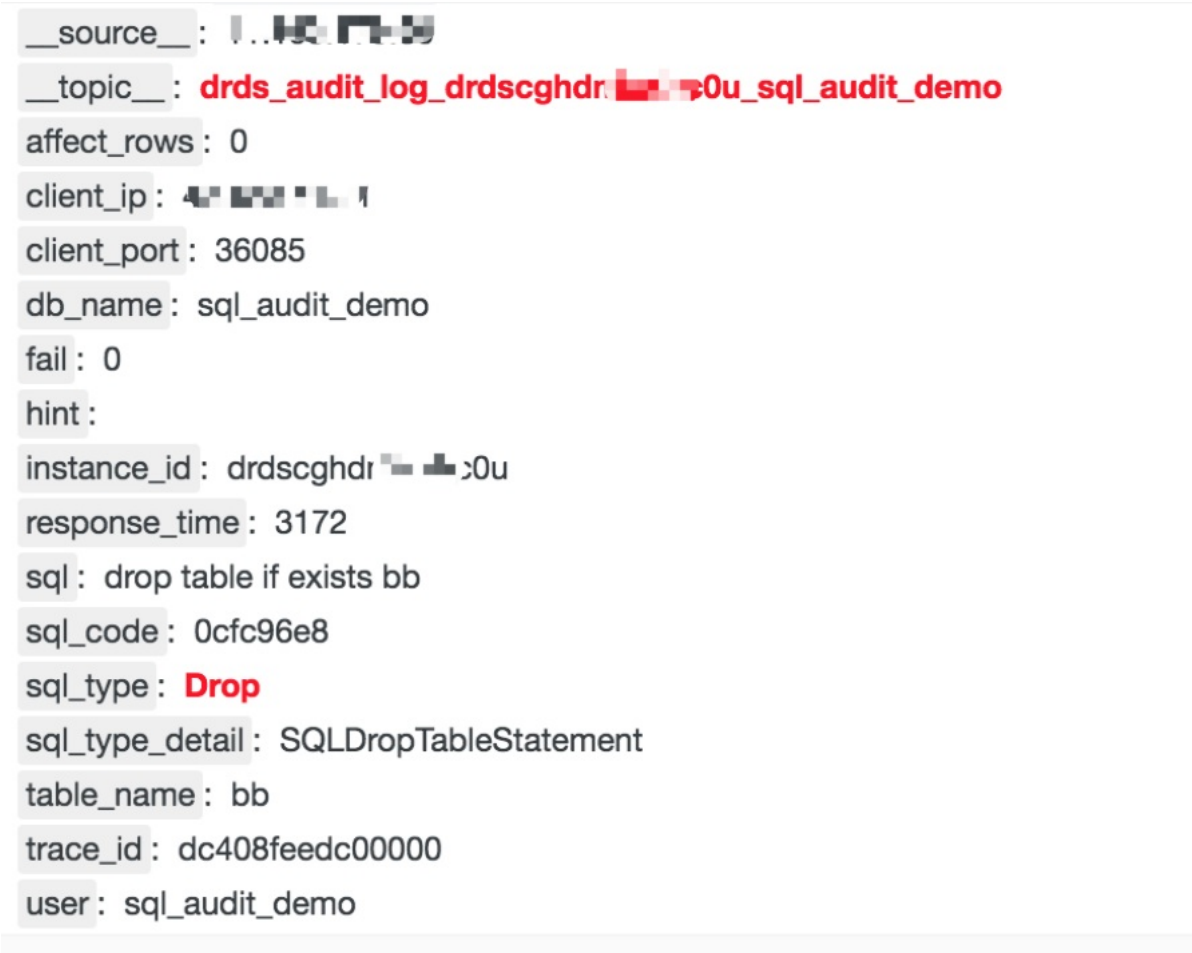
- Field search

Based on built-in index fields, the SQL audit and analysis feature also supports field-based search.

For example, to query SQL statements of the Drop type, execute the following statement:

```
and sql_type:Drop
```

The result is shown in the following figure.



- Multi-condition search

  You can use the "and" and "or" keywords to perform a multi-condition search.

  For example, you can query the delete operation on the rows whose id is 34:

  ```
  and sql:34 and sql_type: Delete
  ```

- Search based on numeric comparison

  affect_rows and response_time in the index filed are numeric values and support comparison operators.

  For example, you can query the SQL INSERT statements whose response_time is greater than 1s.

  ```
  and response_time  > 1507 and sql_type: Insert
  ```

  For example, you can query the SQL statement that deletes more than 100 rows of data:

  ```
  and affect_rows  > 100 and sql_type: Delete
  ```

Analysis of the SQL statement execution status

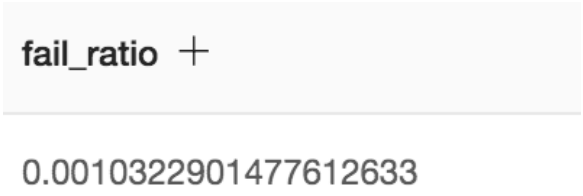
This section introduces the statements used to query the SQL statement execution status in .

- Failure rate of SQL statement execution

  Execute the following statement to query the failure rate of SQL statement execution:

  ```
  | SELECT sum(case when fail = 1 then 1 else 0 end) * 1.0 / count(1) as fail_ratio
  ```

  The result is shown in the following figure.

  If your business is sensitive to the error rate of SQL statement execution, you can customize the alert information based on the query result. Click **Save as Alert** in the upper-right corner of the page.

  In the alert settings shown in the preceding figure, the number of log entries that have an error rate of SQL statement execution greater than 0.01 within 15 minutes is checked within every 15 minutes. You can also customize alerts as needed.

- Total number of rows returned by SELECT statements

  Execute the following statement to query the cumulative number of rows queried by SELECT statements:

  ```
    and sql_type: Select | SELECT sum(affect_rows)
  ```

- SQL statement type distribution

  Execute the following statement to query the SQL statement type distribution:

  ```
    | SELECT  sql_type, count(sql) as times GROUP BY sql_type
  ```

- IP address distribution of SQL independent users

  Execute the following statement to query the distribution of IP addresses of independent users who execute SQL statements:

  ```
  | SELECT  user, client_ip, count(sql) as times GROUP BY user, client_ip
  ```

SQL performance analysis

This section describes typical SQL statements for SQL performance analysis.

- Average response time of SELECT statements

  Execute the following statement to query the average response time of SELECT statements:

  ```
  and sql_type: Select | SELECT avg(response_time)
  ```

- Distribution of SQL statement response time

  Execute the following statement to query the distribution of SQL statement response time:

  ```
   and response_time > 0 | select case when response_time <=10 then '<=10 ms when response_time > 10
  and response_time <= 100 then '10~100 ms when response_time > 100 and response_time <= 1000 then '1
  00 ms ~ 1s 'When response_time > 1000 and response_time <= 10000 then '1s ~ 10s' when response_time
  > 10000 and response_time <= 60000 then '10s ~ 1 min'> 1 min' end as latency_type, count(1) as cnt
  group by latency_type order by latency_type DESC
  ```

  The preceding query shows the distribution of SQL statement execution time based on a given time range. You can adjust the time range to obtain finer-grained results.

- Top 50 slow SQL statements

Execute the following statement to query slow SQL statements:

```
 | SELECT date_format(from_unixtime(__time__), '%m/%d %H:%i:%s') as time, user, client_ip, client_port, sql_type, affect_rows, response_time, sql ORDER BY  response_time desc LIMIT 50
```

The following figure shows the result, which includes the SQL statement execution time, user name, IP address, port number, SQL statement type, number of affected rows, response time, and text of SQL statements.

time	user	client_ip	client_port	sql_type	affect_rows	response_time	sql
09/28 14:04:05	sql_audit_demo	[redacted]	[redacted]7	Drop	0	9583	drop table if exists bb
09/28 14:04:05	sql_audit_demo	[redacted]	[redacted]7	Drop	0	9583	drop table if exists bb
09/28 14:04:05	sql_audit_demo	[redacted]	[redacted]477	Drop	0	9583	drop table if exists bb
09/27 17:38:18	sql_audit_demo	[redacted]	[redacted]073	Drop	0	7200	drop table if exists bb

- Top 10 costly SQL templates

  In most applications, SQL statements are dynamically generated based on several templates, and only the parameters are different. You can find, analyze, and optimize the costly SQL templates based on template IDs. Enter the following query statement:

  ```
  | SELECT sql_code as "SQL template ID", round(total_time * 1.0 /sum(total_time) over() * 100, 2) as "response time share (%)" ,execute_times as "number of executions", round(avg_time) as "average response time", round(avg_rows) as "average number of affected rows", CASE WHEN length(sql) > 200 THEN concat(substr(sql, 1, 200℃, '......') ELSE trim(lpad(sql, 200,'hour) end as "sample SQL" FROM (SELECT sql_code, count(1) as execute_times, sum(response_time) as total_time, avg(response_time) as avg_time, avg(affect_rows) as avg_rows, arbitrary(sql) as sql FROM log GROUP BY sql_code) ORDER BY "execution time share (%)" desc limit 10
  ```

  The statistics include the SQL template ID, percentage of response time of the statement generated from the template in the total response time of SQL statements, number of executions, average response time, average number of affected rows, and sample SQL statement. For better display effect, each page displays 200 entries. In the preceding query result, statements are ranked by the response time share. However, you can rank the statements by the average response time or the number of executions to troubleshoot relevant issues.

- Average transaction response time

  For SQL statements within the same transaction, the preset trace_id field prefixes are the same, and the suffixes are '-' followed by sequence numbers. trace_id of non-transactional SQL statements does not contain '-'. Based on this, you can analyze the performance of transactions.

  > **Note** Transaction analysis is less efficient than other query operations because it involves prefix matching.

  For example, execute the following statement to query the average response time of transactions:

  ```
   | SELECT  sum(response_time) / COUNT(DISTINCT substr(trace_id, 1, strpos(trace_id, '-') - 1)) where strpos(trace_id, '-') > 0
  ```

- Top 10 slow transactions

  You can query the list of slow transactions by response time of transactions. Use the following statement:

  ```
  | SELECT substr(trace_id, 1, strpos(trace_id, '-') - 1) as "transaction ID" , sum(response_time) as "response time" where strpos(trace_id, '-') > 0 GROUP BY substr(trace_id, 1, strpos(trace_id, '-') - 1) ORDER BY "response time" DESC LIMIT 10
  ```

Based on this, you can use the transaction ID to search for all the SQL statements under the transaction and analyze the specific causes of slow execution. Use the following statement:

```
and trace_id: db3226a20402000*
```

- Top 10 transactions with batch operations

  Based on the number of rows affected by SQL statements in a transaction, you can obtain the list of transactions that contain batch operations. Use the following statement:

```
| SELECT substr(trace_id, 1, strpos(trace_id, '-') - 1) as "transaction ID" , sum(affect_rows) as "number of affected rows" where strpos(trace_id, '-') > 0 GROUP BY substr(trace_id, 1, strpos(trace_id, '-') - 1) ORDER BY "number of affected rows" DESC LIMIT 10
```

SQL security analysis

This section provides typical query statements for SQL security analysis.

- Distribution of types of failed SQL statements

```
and fail > 0 | select sql_type, count(1) as "number of errors" group by sql_type
```

- High-risk SQL statements

  High-risk SQL statements are of the Drop or Truncate type. You can also add more conditions as needed.

```
and sql_type: Drop OR  sql_type: Truncate
```

- SQL batch delete events

```
and affect_rows > 100 and sql_type: Delete | SELECT date_format(from_unixtime(__time__), '%m/%d %H:%i:%s') as time, user, client_ip, client_port, affect_rows, sql ORDER BY  affect_rows desc LIMIT 50
```

12.1.4.9.5. Log reports

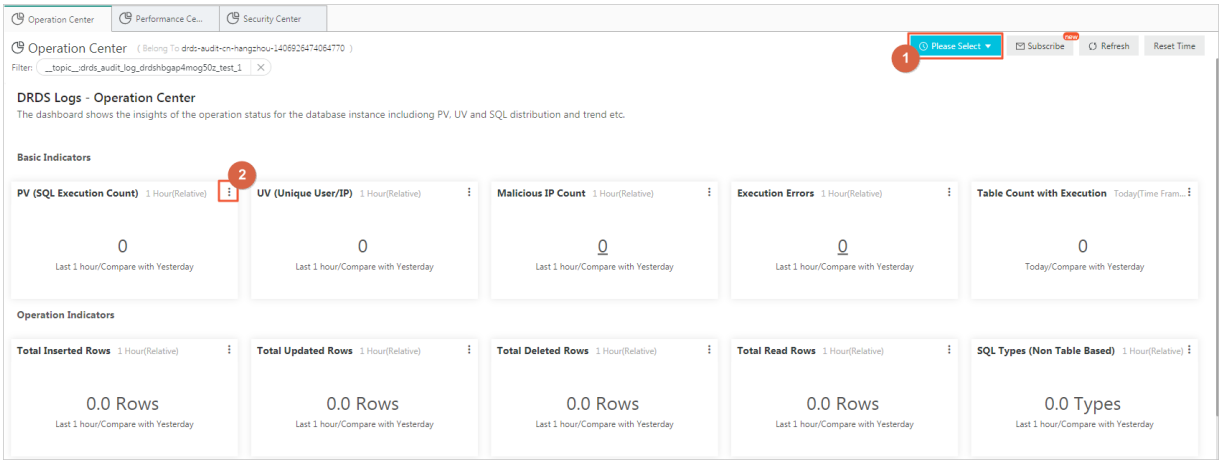
collaborates with Log Service to enable the SQL audit and analysis feature and the out-of-the-box report center. The report center provides the modules Operation Center, Performance Center, and Security Center. This way, you can gain a clear profile of your database, including the running status, performance metrics, and potential security risks.

View reports

After you enable the SQL audit and analysis feature, click the **Log Reports** tab. Then, you can view the reports that are provided by Log Service. You can view information on the **Operation Center**, **Performance Center**, and **Security Center** tabs. For information about how to enable the SQL audit and analysis feature, see Enable SQL audit and analysis.

> **Note**
>
> - If databases are deployed in a region, Log Service stores all the audit logs of the databases in a Logstore. When you view the reports of a database, Log Service automatically adds a filter condition that follows the format __topic__:drds_audit_log_Instance ID_Database name. This indicates that you view the data of the database. For example, Log Service adds the filter condition `drds_audit_log_drdsxyzabcd_demo_drds_db`.
> - If the version of your instance is earlier than 5.3.4-15378085, some fields are missing in the SQL query logs. For more information about the description of log fields, see Log fields. The Log Reports tab provides only a simplified version of Operation Center. If you need to use a full version, upgrade the instance to the latest version.

Statistics in the charts on the Log Reports tab are generated during different periods of time. You can modify the time range as needed. You can modify the time range for all charts or for a single chart.



- Click the time picker (position ① in the figure). In the widget that appears, you can modify the time range for all the charts.
- Click the time picker for a single chart (position ② in the figure). In the widget that appears, you can modify the time range for the chart.

Operation Center

Operation Center displays information about SQL statements that are executed in databases, including the metrics, distribution, and trends.

Chart	Type	Default time range	Description
PV (SQL Execution Count)	Single value	1 Hour (Relative)	The number of execution times for SQL statements.
UV (Unique User/IP)	Single value	1 Hour (Relative)	The number of unique users and IP addresses.
Malicious IP Count	Single value	1 Hour (Relative)	The number of malicious IP addresses. For more information about the definition of malicious IP addresses, see Security check functions.
Execution Errors	Single value	1 Hour (Relative)	The number of SQL statements that have execution errors.
Table Count with Execution	Single value	1 Hour (Relative)	The total number of tables that are queried by SQL statements.
Total Inserted Rows	Single value	1 Hour (Relative)	The total number of rows that are inserted by INSERT operations.
Total Updated Rows	Single value	1 Hour (Relative)	The total number of rows that are updated by UPDATE operations.

Chart	Type	Default time range	Description
Total Deleted Rows	Single value	1 Hour (Relative)	The total number of rows that are deleted by DELETE operations.
Total Read Rows	Single value	1 Hour (Relative)	The total number of rows that are returned by SELECT operations.
SQL Types (Non Table Based)	Single value	1 Hour (Relative)	The types of SQL statements that are used for non-table operations, for example, SHOW VARIABLES LIKE.
SQL Execution Trend	Column chart	1 Hour (Relative)	The execution trend of SQL statements and the distribution trend of error SQL statements.
Operated Tables	Flow diagram	1 Hour (Relative)	The distribution of tables that are queried by SQL statements.
SQL Statement Type	Flow diagram	1 Hour (Relative)	The distribution of SQL statement types by time.
User Distribution	Pie chart	1 Hour (Relative)	The distribution of users who execute SQL statements.
SQL Type Distribution	Area chart	1 Hour (Relative)	The ratio of each SQL statement type by time range.
Tables with Most Operations (Top 50)	Table	1 Hour (Relative)	The list of top 50 tables on which most operations are performed, including the table names and the number of operations. The operations include READ, DELETE, UPDATE, and INSERT operations.
SQL Type (World)	Map	1 Hour (Relative)	The distribution of IP addresses of the clients on which SQL statements are executed. The IP addresses are distributed on the world map.
SQL Type (China)	Map	1 Hour (Relative)	The distribution of IP addresses of the clients on which SQL statements are executed. The IP addresses are distributed on the map of China.

Performance Center

Performance Center displays performance metrics, the distribution of slow and fast SQL queries, and the distribution and sources of costly SQL queries in databases.

Chart	Type	Default time range	Description
Peak SQL Execution Traffic	Single value	1 Hour (Relative)	The maximum number of SQL statements that are executed per second.
Peak Select Traffic	Single value	1 Hour (Relative)	The maximum number of rows that are returned by SELECT statements per second.
Peak Insert Traffic	Single value	1 Hour (Relative)	The maximum number of rows that are inserted by INSERT statements per second.
Peak Update Traffic	Single value	1 Hour (Relative)	The maximum number of rows that are updated by UPDATE statements per second.
Peak Delete Traffic	Single value	1 Hour (Relative)	The maximum number of rows that are deleted by DELETE statements per second.
Average Response Time	Single value	1 Hour (Relative)	The average response time of SQL statements.
Select SQL	Single value	1 Hour (Relative)	The average number of SELECT statements that are executed per second.
Insert SQL	Single value	1 Hour (Relative)	The average number of INSERT statements that are executed per second.
Update SQL	Single value	1 Hour (Relative)	The average number of UPDATE statements that are executed per second.
Delete SQL	Single value	1 Hour (Relative)	The average number of DELETE statements that are executed per second.
Select/Update Traffic Trend	Line chart	1 Hour (Relative)	The distribution of rows that are affected by the SELECT and UPDATE statements by time.
SQL Execution Time Distribution	Pie chart	1 Hour (Relative)	The distribution of the execution time of SQL statements.
Slow SQL Table Distribution	Pie chart	1 Hour (Relative)	The distribution of tables that are queried by slow SQL queries whose response time exceeds 1s.

Chart	Type	Default time range	Description
Slow SQL User Distribution	Pie chart	1 Hour (Relative)	The distribution of users who perform slow SQL queries whose response time exceeds 1s.
Slow SQL Type Distribution	Pie chart	1 Hour (Relative)	The distribution of the types of slow SQL queries whose response time exceeds 1s.
Slow SQL (Top 50)	Table	1 Hour (Relative)	The table that lists slow SQL queries whose response time exceeds 1s. The following information is included: the time, client, response time, instance, database, table, user, number of affected rows, SQL type, and SQL text.
SQL Template Execution Time Top 20	Table	1 Hour (Relative)	The execution status of SQL statements by SQL template. The table contains the SQL template ID, response time ratio, number of execution times, average response time, average number of affected rows, and sample SQL statement.
Transaction Affected Rows Top 20	Table	1 Hour (Relative)	The table that lists top 20 transactions by the number of affected rows. The table contains the transaction ID and number of affected rows.
Transaction Executed Time Top 20	Table	1 Hour (Relative)	The table that lists top 20 transactions by response time. The table contains the transaction ID and number of affected rows.

Security Center

Security Center displays failed SQL statements and malicious SQL statements that are executed in databases. You can also view information about batch DELETE and UPDATE events, including the event details, distribution, and trends.

Chart	Type	Default time range	Description
Error Count	Single value	1 Hour (Relative)	The number of execution times for failed SQL statements.

Chart	Type	Default time range	Description
Batch Delete Events	Single value	1 Hour (Relative)	The number of SQL statements for batch DELETE events that affect more than 100 rows.
Batch Update Events	Single value	1 Hour (Relative)	The number of SQL statements for batch UPDATE events that affect more than 100 rows.
Malicious SQL Executions	Single value	1 Hour (Relative)	The number of execution times for malicious SQL statements, for example, DROP and TRUNCATE statements.
Malicious IP Count	Single value	1 Hour (Relative)	The number of malicious IP addresses. For more information about the definition of malicious IP addresses, see Security check functions.
Error Distribution	Area chart	1 Hour (Relative)	The distribution of the types of failed SQL statements.
Distribution of Client with Errors	Map	1 Hour (Relative)	The distribution of the clients on which SQL statements fail to be executed. The clients are distributed on the map of China.
Client with Most Errors	Table	1 Hour (Relative)	The table that lists the clients on which SQL statements fail to be executed. The following information is included: the client IP addresses, the number of errors, the types of failed SQL statements, and sample failed SQL statements.
Malicious SQL Executions	Table	1 Hour (Relative)	The table that lists the malicious SQL statements. The following information is included: the time, IP address, SQL statement, instance ID, database, table, and user.

Chart	Type	Default time range	Description
Batch Delete Events (Top 50)	Table	1 Hour (Relative)	The table that lists top 50 SQL statements for batch DELETE events. The following information is included: the earliest execution time, most recent execution time, instance ID, database, table, number of execution times, average number of deleted rows, average response time, and sample SQL statement.
Batch Update Events (Top 50)	Table	1 Hour (Relative)	The table that lists top 50 SQL statements for batch UPDATE events. The following information is included: the earliest execution time, most recent execution time, instance ID, database, table, number of execution times, average number of updated rows, average response time, and sample SQL statement.

12.1.4.10. Monitoring and alerts

12.1.4.10.1. Monitor instances

provides the monitoring feature to help you monitor the status of your instances. You can view the detailed information about monitoring metrics of instances, databases, or storage nodes in the console. This topic describes how to view performance monitoring data of a instance on the **Instance Monitoring** page in the console.

Monitoring metrics of instances

Metric category	Metric	Parameter	Description
Resource	**CPU**	`CPU Utilization (%)`	The average CPU utilization of the compute node in the instance.
	Memory	`Memory Usage (%)`	The memory usage of the compute node in the instance. Memory usage fluctuations are normal.
	Network	`Inbound Traffic (Kbps)`	The total inbound traffic of the compute node in the instance. Unit: kbit/s. Inbound network traffic is generated when a storage node returns data to the compute node.

Metric category	Metric	Parameter	Description
		`Outbound Traffic (Kbps)`	The total outbound traffic of the compute node in the instance. Unit: kbit/s. Outbound network traffic is generated when the compute node sends physical SQL statements to the storage node or the compute node returns data to applications.
Engine	**Logical QPS**	`Logical QPS`	The total number of SQL statements that are processed by the instance per second.
	Physical QPS	`Physical QPS`	The total number of SQL operations that are sent from the compute node in the instance to the storage node per second.
	Logical RT	`Logical RT (ms)`	The average response time (RT) for processing each SQL statement by the instance.
	Physical RT	`Physical RT (ms)`	The average RT for processing SQL statements that are sent from the compute node in the instance to the storage node.
	Connections	`Connections`	The total number of connections established between an application and the instance.
	Active Threads	`Active Threads`	The number of threads that are used to execute SQL statements in the instance.

Note

- The data collection cycle is 1 minute for the metrics of the **Resource** category and 5 seconds for the metrics of the **Engine** category in the preceding table.
- You can view the monitoring data that is generated within seven days.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. On the **Basic Information** page, choose **Monitoring and Alerts > Instance Monitoring** in the left-side navigation pane.
5. On the **Instance Monitoring** page, specify the following parameters.

Parameter	Description
Monitoring Metric	Select **Resource** or **Engine**.
Monitoring Metric	Select the monitoring metric that you want to view. For more information about monitoring metrics, see Monitoring metrics of instances.
Query Time	You can specify the query time as 1 hour, 6 hours, 12 hours, one day, or one week. You can also customize a query time range. The minimum time range is 1 minute and the maximum time range is one week.

12.1.4.10.2. Monitor databases

provides the monitoring feature to help you monitor the status of your instances. You can view the detailed information about monitoring metrics of instances, databases, or storage nodes in the console. This topic describes how to view the performance monitoring data of a single database in a instance on the **Database Monitoring** page.

Monitoring metrics of databases

Monitoring metric	Parameter	Description
QPS	`QPS`	The total number of SQL statements that are processed by the database in the instance per second.
Connections	`Connections`	The total number of connections between an application and the database in the instance.
Active Threads	`Active Threads`	The number of threads that are used by the database in the instance to execute SQL statements.

> **Note**
> - The data collection cycle is 1 minute for each metric in the preceding table.
> - You can view the monitoring data that is generated within seven days.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. On the **Basic Information** page, choose **Monitoring and Alerts > Database Monitoring** in the left-side navigation pane.
5. On the **Database Monitoring** page, specify the following parameters.

Parameter	Description
Database	Select the database for which you want to view monitoring metrics from the **Database** drop-down list.
Data Indexes	The monitoring metrics of the database. Select the monitoring metric that you want to view. For more information about monitoring metrics, see Monitoring metrics of databases.
Query Time	You can specify the query time as 1 hour, 6 hours, 12 hours, one day, or one week. You can also customize a query time range. The minimum time range is 1 minute and the maximum time range is one week.

12.1.4.10.3. Monitor storage nodes

provides the monitoring feature to help you monitor the status of your instances. You can view the detailed information about monitoring metrics of instances, databases, or storage nodes in the console. This section describes how to view the performance monitoring data of storage nodes in a instance on the **Storage Monitoring** page.

Monitoring metrics of storage nodes

Monitoring category	Monitoring metric	Parameter	Description
Resource	CPU and Memory	CPU Utilization and Memory Usage (%)	The CPU utilization and memory usage of the storage node in the instance.
	Disk Space	Disk Space (MB)	The usage of the total space, data space, log space, temporary space, and system space of the storage node in the instance. Unit: MB.
	IOPS	IOPS	The input/output operations per second (IOPS) of the storage node in the instance.
	Connections	Total Current Connections	The number of active connections and the total number of connections of the storage node in the instance.
	Network Traffic	Network Traffic (KB)	The inbound and outbound traffic of the storage node per second in the instance. Unit: KB.
	TPS/QPS	TPS (Transactions per Second)/QPS (SQL Queries per Second)	The average number of times that SQL statements are executed per second and the average number of transactions that run per second on the storage node in the instance.
	InnoDB Cache	InnoDB Cache Read Hit Ratio, Usage, and Dirty Ratio (%)	The read hit ratio, usage, and dirty ratio of the InnoDB buffer pool on the storage node in the instance.
	InnoDB Read/Write	InnoDB Reads/Writes (KB)	The average amount of data read by InnoDB per second and the average amount of data written by InnoDB per second on the storage node in the instance. Unit: KB.
	Cached Requests	InnoDB Cached Requests	The average number of reads from the InnoDB buffer pool per second and the average number of writes to the InnoDB buffer pool per second on the storage node in the instance.
	InnoDB Log	InnoDB Log Read/Write/fsync	The average number of log write requests per second, the average number of physical writes to log files per second, and the average number of times that the fsync operation is performed for log files per second on the storage node in the instance.

Monitoring category	Monitoring metric	Parameter	Description
Engine	**Temporary Tables**	`Temporary Tables Automatically Created on Hard Disk When MySQL Statements Are Executed`	The number of temporary tables that are automatically created on hard disks during statement execution on the storage node in the instance.
	COMDML	`MySQL_COMDML`	The average number of times that DELETE statements are executed per second, the average number of times that INSERT statements are executed per second, the average number of times that INSERT_SELECT statements are executed per second, the average number of times that REPLACE statements are executed per second, the average number of times that REPLACE_SELECT statements are executed per second, the average number of times that SELECT statements are executed per second, and the average number of times that UPDATE statements are executed per second on the storage node in the instance.
	RowDML	`MySQL_RowDML`	The average number of rows read from InnoDB tables per second, the average number of rows updated in InnoDB tables per second, the average number of rows deleted from InnoDB tables per second, the average number of rows inserted into InnoDB tables per second, and the average number of physical writes to log files per second on the storage node in the instance.
	MyISAM Read/Write	`MyISAM Reads and Writes`	The average number of reads from the buffer pool by MyISAM per second, the average number of writes to the buffer pool by MyISAM per second, the average number of reads from hard disks by MyISAM per second, and the average number of writes to hard disks by MyISAM per second on the storage node in the instance.
	MyISAM Key	`MyISAM Key Buffer Read Ratio/Write Ratio/Usage (%)`	The average usage, read hit ratio, and write hit ratio of the MyISAM key buffer per second on the storage node in the instance.

Note

- The data collection cycle of the preceding **Resource** and **Engine** items is 1 minute.
- You can view the monitoring data that is generated within seven days.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.

4. On the **Basic Information** page, choose **Monitoring and Alerts > Storage Monitoring** in the left-side navigation pane.
5. On the **Storage Monitoring** page, specify the following parameters.

Parameter	Description
Database	Select the database for which you want to view monitoring metrics from the **Database** drop-down list.
Monitoring Metric	Select **Resource** or **Engine**.
Monitoring Metric	The monitoring metrics of the storage resources. Select the monitoring metric that you want to view. For more information about monitoring metrics, see Monitoring metrics of storage nodes.
Query Time	You can specify the query time as 1 hour, 6 hours, 12 hours, one day, or one week. You can also customize a query time range. The minimum time range is 1 minute and the maximum time range is one week.

12.1.4.10.4. Prevent performance problems

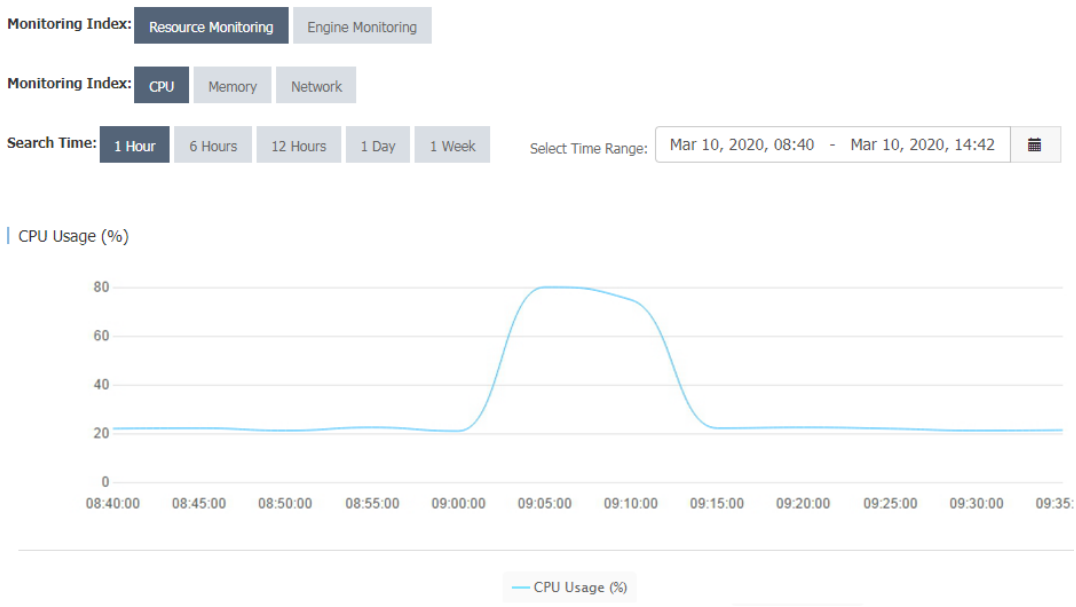
12.1.4.10.4.1. PolarDB-X CPU utilization

This topic describes the CPU utilization metric.

Performance metrics often fluctuate with system traffic, as shown in the following two common scenarios:

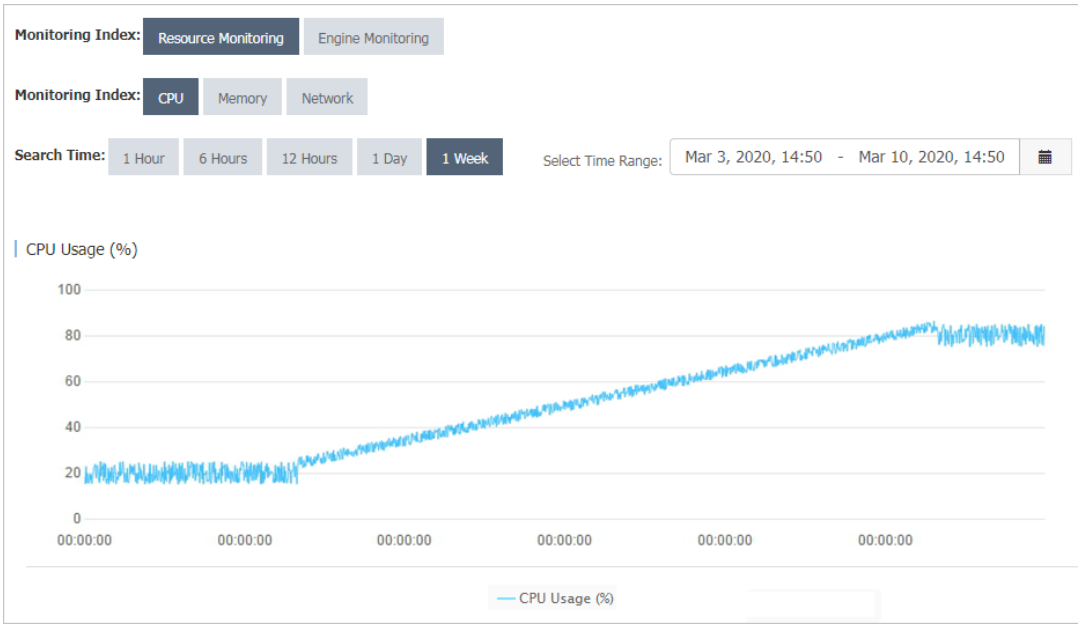
- An application has a shopping spree activity at 09:00 every morning. Therefore, the traffic of the system increases significantly at this time point. As shown by the monitoring data, the CPU utilization of the instance increased from 20% to about 80% from 09:00. The peak lasts about 10 minutes.

  CPU utilization-1



- An application has increasing traffic. Therefore, the system traffic keeps increasing until it reaches a stable level. The monitored CPU utilization of the instance also reflects this change.

  CPU utilization-2



When the load on the instance changes with the business, you must take note of the changes in metrics. If the CPU utilization exceeds the threshold, you must upgrade the instance specifications to alleviate the performance pressure.

You can set alert rules for instances in the console. When the average CPU utilization exceeds the preset threshold, the system sends text messages to the corresponding contacts. You can set the CPU utilization threshold as needed. We recommend that you set it to 80%.

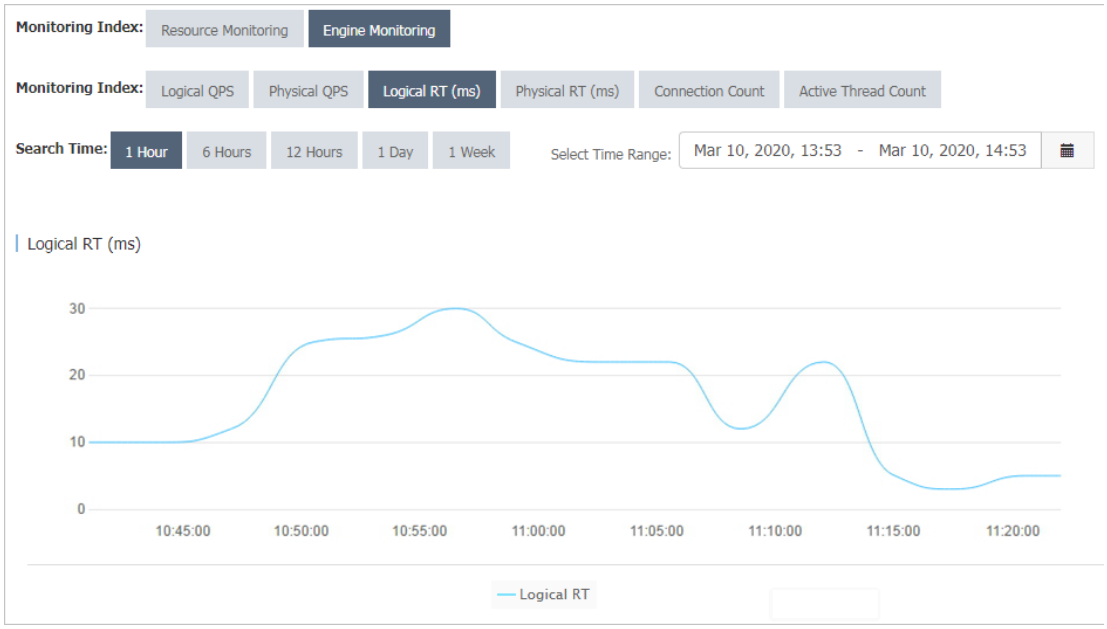
12.1.4.10.4.2. Logical RT and physical RT

This topic describes two metrics: logical RT and physical RT.

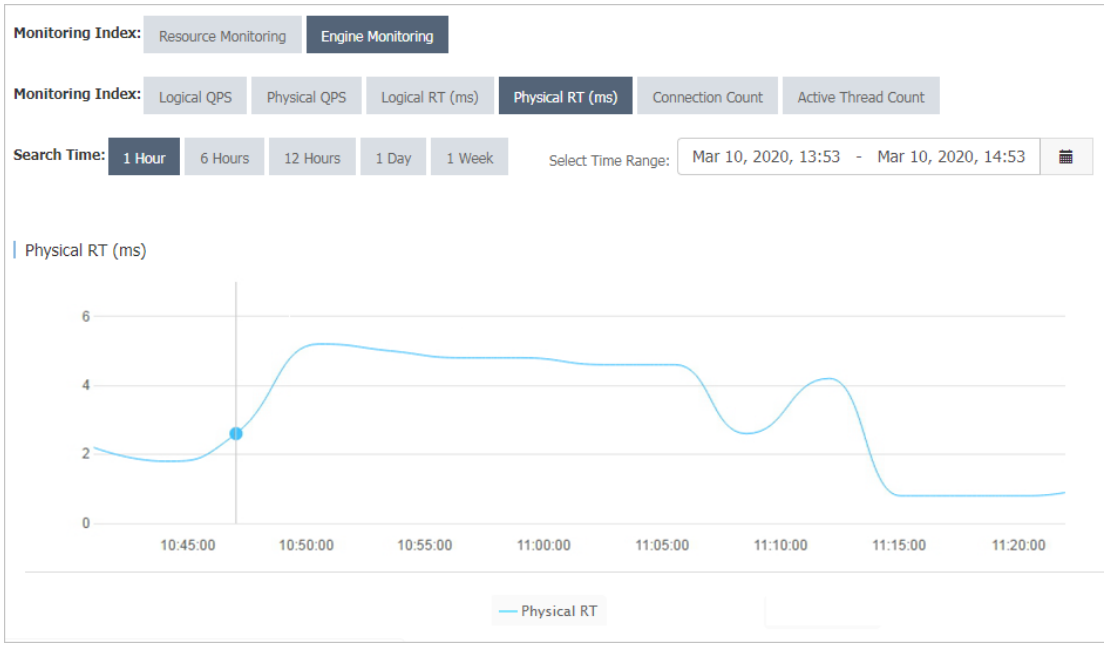
Logical RT is the time from when a instance receives a logical SQL statement to when it returns data to an application. Physical RT is the time from when a instance routes a physical SQL statement to an ApsaraDB RDS for MySQL instance to when it receives the data returned by the ApsaraDB RDS for MySQL instance.

If a logical SQL statement is partitioned into one or more physical SQL statements, the logical RT is greater than or equal to the physical RT. Ideally, performs only a few operations on the data returned by ApsaraDB RDS for MySQL. Therefore, logical RT is slightly longer than physical RT. Under special circumstances, physical SQL queries are fast executed, whereas logical SQL queries are slowly executed. The following figure shows the trends of the logical RT and physical RT.

Logical RT



Physical RT



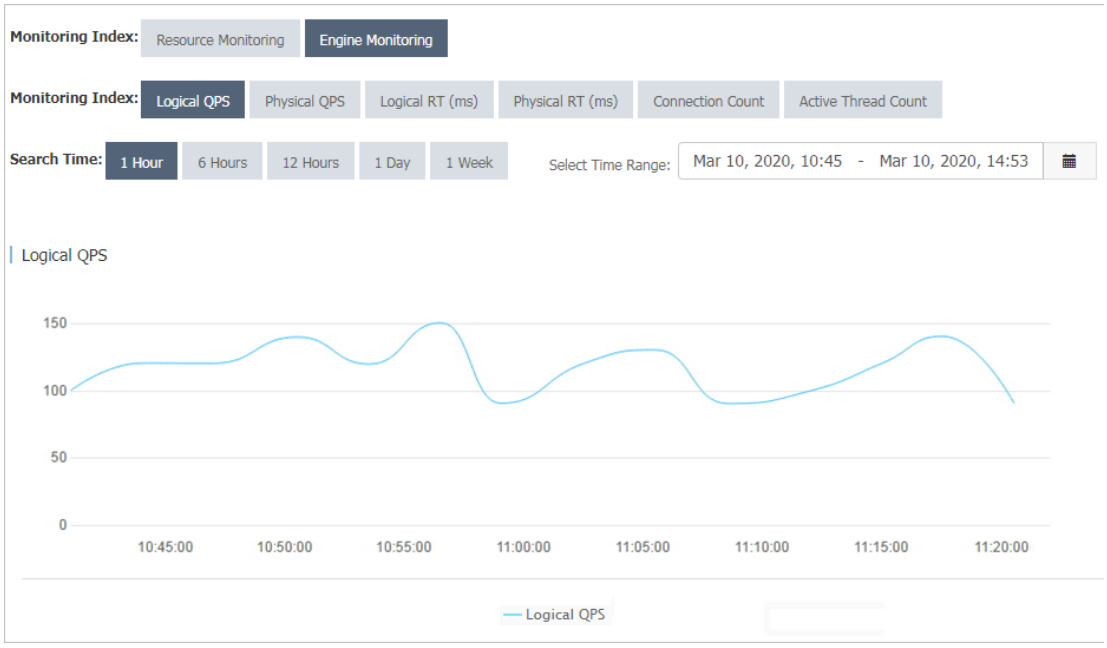
As shown in the preceding figures, the change trends of logical RT and physical RT in the two monitoring charts are basically the same. The logical RT fluctuates between 10 ms and 20 ms, and the physical RT fluctuates between 2 ms and 5 ms. This means that the instance has a heavy load. To alleviate the load pressure, you can upgrade the instance configuration. If both the logical RT and physical RT are high, you can upgrade the ApsaraDB RDS for MySQL configuration or optimize SQL statements on the ApsaraDB RDS for MySQL instance.

12.1.4.10.4.3. Logical QPS and physical QPS

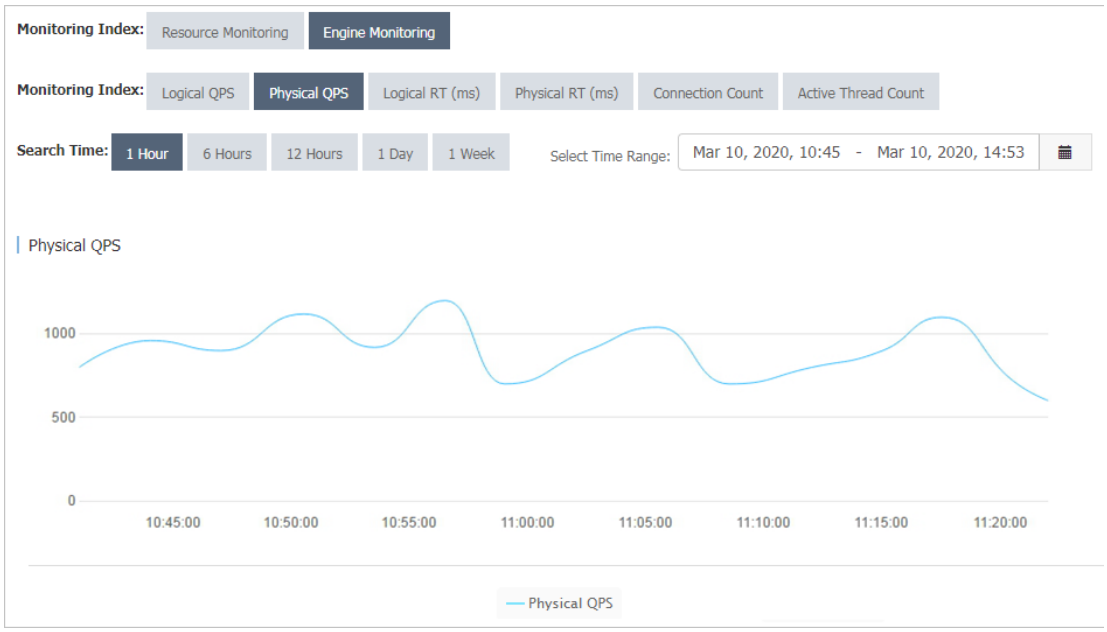
This topic describes two metrics: logical QPS and physical QPS.

As shown by the monitoring data in the following figures, the logical QPS and physical QPS have the same trends. However, the logical QPS and physical QPS have a large difference in value and are in a specific proportion.

Logical QPS



Physical QPS



As shown in the preceding figures, logical QPS fluctuates between 80 and 150, and physical QPS fluctuates between 700 and 1,200.

The reason is that generates physical SQL statements based on logical SQL statements. The ratio of logical SQL statements to physical SQL statements is not necessarily 1:1. For example, a logical table is created with the following statement:

```
CREATE TABLE drds_user
(id int,
name varchar(30))
dbpartition by hash(id);
```

When the query condition contains the database shard key, routes the logical SQL statement to the ApsaraDB RDS for MySQL instance for execution. The following execution plan shows that the number of physical SQL statements is 1:

```
mysql> explain select name from drds_user where id = 1;
+---------------------------+-------------------------------------------------------------------------
+--------+
| GROUP_NAME                | SQL
| PARAMS |
+---------------------------+-------------------------------------------------------------------------
+--------+
|     SANGUAN_BSQT_0001_RDS | select `drds_user`.`name` from `drds_user` where (`drds_user`.`id` = 1)
| {}     |
+---------------------------+-------------------------------------------------------------------------
+--------+
```

When the query does not contain the database shard key, partitions the logical SQL statement into multiple physical SQL statements. The following execution plan shows that the number of physical SQL statements is 8:

```
mysql> explain select name from drds_user where name = 'LiLei';
+----------- ---------------+-------------------------------------------------------------------------
--------+--------+
| GROUP_NAME                | SQL
| PARAMS |
+---------------------------+-------------------------------------------------------------------------
--------+--------+
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
| SANGUAN_BSQT_0001_RDS     | select `drds_user`.`name` from `drds_user` where (`drds_user`.`name` = '
LiLei') | {}     |
+---------------------------+-------------------------------------------------------------------------
--------+--------+
8 rows in set (0.06 sec)
```

Logical or physical QPS indicates the total number of logical or physical SQL statements processed per unit of time. When most SQL statements in the system contain the shard key, the ratio of logical QPS to physical QPS is close to 1:1. If the difference between the logical and physical QPS is too large, many SQL statements of the current application do not contain the shard key. In this case, check the SQL statements of the application to improve performance.

12.1.4.10.4.4. High memory usage

This topic describes the memory usage metric.

The overly high memory usage of a instance is mostly caused by the large number of SQL queries in your application and the overlarge result set that is returned. If the memory usage of your instance remains at about 100%, perform the Restart a PolarDB-X instance operations to locate and optimize the slow SQL queries of your application.

12.1.4.11. View the instance version

This topic describes two methods that you can use to view the version of a instance.

View the instance version in the console

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the **Configuration Information** section, view the value of **Current Version**.

View the instance version by using the *version()* function

Use the MySQL command line to connect to and execute the `SELECT version()` statement to view the version of the instance. For example, the following result is returned after you execute the `selectversion();` statement:

```
+----------------------------+
| VERSION()                  |
+----------------------------+
| 5.6.29-TDDL-5.1.28-1320920  |
+----------------------------+
1 row in set (0.00 sec)
```

In the preceding example, 5.1.28-1320920 is the version of the instance.

12.1.5. Private ApsaraDB RDS for MySQL instance management

12.1.5.1. Overview

The middleware architecture consists of instances at the compute layer and private ApsaraDB RDS for MySQL instances at the storage layer. Horizontal partitioning is implemented based on multiple mounted ApsaraDB RDS for MySQL instances for database sharding and table sharding. Private ApsaraDB RDS for MySQL instances at the storage layer and instances at the compute layer can closely collaborate. This way, provides the better performance and more stable service access link than before.

O&M management

You can Log on to the PolarDB-X console to perform the following operations on a private ApsaraDB RDS for MySQL instance:

- Change the specifications of an instance
- Add Read-only Instances
- View the monitoring information about a private ApsaraDB RDS for MySQL instance.

12.1.5.2. Change the specifications of an instance

This topic describes how to change the specifications of a private ApsaraDB RDS for MySQL instance.

Prerequisites

A private ApsaraDB RDS for MySQL instance is created. For more information, see Create a database.

Procedure

1. Log on to the PolarDB-X console.

2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Private RDS management**.
5. Find the instance for which you want to change the specifications and click **Change Specifications** in the **Actions** column.
6. In the **Change Specifications** dialog box, change the settings of **Instance Specifications** and **Storage Capacity** based on your business requirements.
7. Click **OK**.

12.1.5.3. Add Read-only Instances

This topic describes how to add read-only instances on the Private RDS Instances page.

Prerequisites

A private ApsaraDB RDS for MySQL instance is created. For more information, see Create a database.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Private RDS Instances**.
5. Find the primary instance to which you want to add read-only instances and click **Add Read-only Instances** in the **Actions** column.
6. In the **Add Read-only Instances** dialog box, select **Instance Specifications** and **Storage Capacity** based on your business requirements.
7. Click **OK**.

12.1.5.4. View the monitoring information about a private ApsaraDB RDS for MySQL instance.

This topic describes how to view the monitoring metrics of a private ApsaraDB RDS for MySQL instance.

Context

provides the monitoring feature to help you monitor the status of your private ApsaraDB RDS for MySQL instances. You can view the detailed information about monitoring metrics of a specified instance in the console.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Private RDS management**.
5. Find the private ApsaraDB RDS for MySQL instance that you want to view and click **Monitoring** in the **Actions** column.
6. On the **Storage Monitoring** page under **Monitoring and Alerts**, you can view the details of each

monitoring metric. For more information, see Monitor storage nodes.

12.1.6. Account management

12.1.6.1. Basic concepts

This topic introduces the basic concepts of the account and permission system.

The usage of the account and permission system in is the same as in MySQL. PolarDB-X supports statements such as `GRANT`, `REVOKE`, `SHOW GRANTS`, `CREATE USER`, `DROP USER`, and `SET PASSWORD`. PolarDB-X allows you to grant permissions at the database and table levels, but does not allow you to grant permissions at the global or column level.

For more information about the MySQL account and permission system, see MySQL documentation.

Notice Accounts that you create by using the `CREATE USER` statement in a instance exist only in the instance. The accounts will not be synchronized to the backend ApsaraDB RDS for MySQL instances.

Accounts

An account is a combination of a username and a hostname in the username@'host' format. Accounts that have the same username but different hostnames are different accounts. For example, lily@30.9.73.96 and lily@30.9.73.100 are two different accounts, and their passwords and permissions may be different.

After a database is created in the console, the system automatically creates two system accounts for the database: the administrator account and the read-only account. These two accounts are built-in accounts. You cannot delete them or modify their permissions.

- The administrator account name is the same as the database name. For example, if the database name is easydb, the administrator account name is also easydb.
- The read-only account name is the database name suffixed with _RO. For example, if the database name is easydb, the read-only account name is easydb_RO.

Assume that the dreamdb and andordb databases are available. Based on the preceding rules, the dreamdb database contains the administrator account named dreamdb and the read-only account named dreamdb_RO. The andordb database contains the administrator account named andordb and the read-only account named andordb_RO.

Account rules

- An administrator account has all permissions.
- You can use only an administrator account to create accounts and grant permissions. Other accounts can only be created and granted permissions by the administrator account.
- An administrator account is bound to a database and does not have permissions on other databases. The administrator account can access only the bound database, but cannot grant permissions on other databases to an account. For example, the easydb administrator account can connect only to the easydb database, and can grant only the permissions on the easydb database or tables in the easydb database to an account.
- A read-only account has only the SELECT permission.

Username rules

- Usernames are not case-sensitive.
- A username must be 4 to 20 characters in length.
- A username must start with a letter.
- A username can contain uppercase letters, lowercase letters, and digits.

Password rules

- A password must be 6 to 20 characters in length.
- A password can contain uppercase letters, lowercase letters, digits, and the following special characters:

  @#$%^&+=

Hostname matching rules

- A hostname must be an IP address. It can contain underscores (_) and percent signs (%). An underscore (_) represents a single character and a percent sign (%) represents zero or more characters. Hostnames that contain wildcards must be quoted with single quotation marks ('), for example, `lily@'30.9. %.%'` and `david@'%'`.
- If two usernames in the system can be used to log on to the database, the username with the longest prefix (the longest IP segment excluding wildcards) prevails. For example, the `david@'30.9.12_.234'` and `david@'30.9.1%.234'` accounts are available in the system. If you use the david username to log on to a database from the 30.9.127.234 host, the `david@'30.9.12_.234'` account is used.
- When you activate the VPC service, the IP address of the host changes. To avoid invalid configurations in the account and permission system, set the hostname to `'%'` to match all IP addresses.

Permission levels

The following list describes the support for permissions of different levels:

- Global permissions (not supported)
- Database-level permissions (supported)
- Table-level permissions (supported)
- Column-level permissions (not supported)
- Subprogram-level permissions (not supported)

Permissions

Eight table-associated basic permissions are supported: `CREATE`, `DROP`, `ALTER`, `INDEX`, `INSERT`, `DELETE`, `UPDATE`, and `SELECT`.

- The `TRUNCATE` statement requires the table-level DROP permission.
- The `REPLACE` statement requires the table-level INSERT and DELETE permissions.
- The `CREATE INDEX` and `DROP INDEX` statements require the table-level INDEX permission.
- The `CREATE SEQUENCE` statement requires the database-level CREATE permission.
- The `DROP SEQUENCE` statement requires the database-level DROP permission.
- The `ALTER SEQUENCE` statement requires the database-level ALTER permission.
- The `INSERT ON DUPLICATE UPDATE` statement requires the table-level INSERT and UPDATE permissions.

Permission rules

- Permissions are bound to an account (`username@'host'`) instead of a username (`username`).
- An error occurs if the table does not exist during authorization.
- The database permissions are listed by level in descending order: global permissions (not supported), database-level permissions, table-level permissions, and column-level permissions (not supported). A granted higher-level permission overwrites a lower-level permission. If you remove the higher-level permission, the lower-level permission is also removed.
- USAGE authorization is not supported.

12.1.6.2. Create an account

This topic describes how to create a account by using the console and SQL statements.

Prerequisites

You have created or added a database. For more information, see Create a database.

Create an account by using the console

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Account Management**.
5. On the **Accounts** page, click **Create Account** in the upper-right corner.
6. Set the following parameters.

Parameter	Description
Database Account	Enter a name for the account. The name of the account must meet the following requirements: ○ The name must be 4 to 20 characters in length. ○ It must start with a letter and end with a letter or digit. ○ The name can contain uppercase letters, lowercase letters, digits, and underscores (_).
New Password	Enter a password for the account. The password of the account must meet the following requirements: ○ The password must be 8 to 32 characters in length. ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. ○ The password can contain the following special characters: ! @#$%^&*()_+-=
Confirm New Password	Enter the password again.
Authorize Databases	You can grant permissions on one or more databases to the account. i. From the Databases section, select one or more databases. Then, click **Add** to add them to the Authorized Databases section. ii. In the Authorized Databases section, select **Read/Write**, **Read-only**, **DDL Only**, or **DML Only** for the specified database. **Note** You can also grant permissions on multiple added databases by clicking **Set All to Read-only**, **Set All to DDL Only**, **Set All to DML Only**, or **Set All to Read/Write** in the upper-right corner of the Authorized Databases section. The buttons in the upper-right corner change after you click them. For example, after you click **Set All to Read-only**, this button is changed to **Set All to DDL Only**.

7. Click **OK**.

Create an account by using the command line

Syntax

```
CREATE USER user_specification [, user_specification] ...
  user_specification: user [ auth_option ]
  auth_option: IDENTIFIED BY 'auth_string'
```

Examples

Create an account whose name is lily and password is 123456. The account can be used to log on to the databases only from 30.xx.xx.96.

```
CREATE USER lily@30.xx.xx.96 IDENTIFIED BY '123456';
```

Create an account that is named david and has no password. The account can be used to log on to the databases from all hosts.

```
CREATE USER david@'%';
```

12.1.6.3. Reset the password

When you use , you can reset the password of your database account by using the console or the command line.

> **Note**
> - Accounts that have root permissions cannot be deleted or modified.
> - For data security, we recommend that you change your password on a regular basis.

Reset the password by using the console

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Account Management**.
5. Find the account for which you want to reset the password, and click **Reset Password** in the Actions column.
6. In the **Reset Account Password** dialog box, enter the new password in the **New Password** field and enter the password again in the **Confirm New Password** field.

   > **Note** The password must meet the following requirements:
   > - The password must be 8 to 32 characters in length.
   > - The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.
   > - The password can contain the following special characters:
   >
   >   ! @#$%^&*()_+-

7. After you confirm that the password is correct, click **OK**.

Reset the password by using the command line

Syntax

```
SET PASSWORD FOR user = password_option
password_option: {
    PASSWORD('auth_string')
}
```

Examples

Change the password of the lily@30.xx.xx.96 account to 123456.

```
SET PASSWORD FOR lily@30.xx.xx.96 = PASSWORD('123456')
```

12.1.6.4. Modify the permissions of an account

You can modify the account permissions of your instances at any time.

Considerations

- The permissions of a privileged account cannot be modified.
- In the PolarDB-X console, you can grant only DML, DDL, read-only, and read and write permissions to standard accounts. If you need more fine-grained authorization, use the command line.

Modify account permissions in the console

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Account Management**.
5. Find the account whose permissions you want to modify, and click **Modify Permission** in the Actions column.
6. In the **Modify Permissions** dialog box, **grant** or **remove** the permissions on one or more databases to or from the account.
   - Add databases:

     From the Databases section, select one or more databases. Then, click **Add** to add them to the Authorized Databases section on the right.
   - Remove databases:

     From the Authorized Databases section, select one or more databases. Then, click **Remove** to move them to the Databases section on the left.
   - Modify permissions on added databases:

     In the Authorized Databases section, select **Read/Write**, **Read-only**, **DDL Only** or **DML Only** for the specified database.

     > **Note** You can also grant permissions on multiple added databases by clicking **Set All to Read-only**, **Set All to DDL Only**, **Set All to DML Only**, or **Set All to Read/Write** in the upper-right corner of the Authorized Databases section.
     >
     > The buttons in the upper-right corner change after you click them. For example, after you click **Set All to Read-only**, this button is changed to **Set All to DDL Only**.
7. After the configuration is completed, click **OK**.

GRANT statement

Syntax

```
GRANT
    priv_type[, priv_type] ...
    ON priv_level
    TO user_specification [, user_specification] ...
    [WITH GRANT OPTION]
priv_level: {
  | db_name.*
  | db_name.tbl_name
  | tbl_name
}
user_specification:
    user [ auth_option ]
auth_option: {
    IDENTIFIED BY 'auth_string'
}
```

Notice

- If the account specified in the GRANT statement does not exist and no IDENTIFIED BY information is provided, an error message is returned. The error message indicates that the account does not exist.
- If the account specified in the GRANT statement does not exist but the IDENTIFIED BY information is provided, the account is created and granted with the specified permissions.

For example, create an account that has the username david. The account can be used to log on to the easydb database from all hosts and has all the permissions on the easydb database.

Method 1: Create an account. Then, grant permissions to the account.

```
CREATE USER david@'%' IDENTIFIED BY 'your#password';
GRANT ALL PRIVILEGES ON easydb.* to david@'%';
```

Method 2: Create an account and grant permissions to the account by executing only one statement.

```
GRANT ALL PRIVILEGES ON easydb.* to david@'%' IDENTIFIED BY 'your#password';
```

Create an account that has the username hanson. The account can be used to log on to the easydb database from all hosts and has all the permissions on the easydb.employees table.

```
GRANT ALL PRIVILEGES ON easydb.employees to hanson@'%' IDENTIFIED BY 'your#password';
```

Create an account that has the username hanson. The account can be used to log on to the easydb database from only 192.xx.xx.10 and has the INSERT and SELECT permissions on the easydb.emp table.

```
GRANT INSERT,SELECT ON easydb.emp to hanson@'192.xx.xx.10' IDENTIFIED BY 'your#password';
```

Create a read-only account that has the username actro. The account can be used to log on to the easydb database from all hosts.

```
GRANT SELECT ON easydb.* to actro@'%' IDENTIFIED BY 'your#password';
```

REVOKE statement

Syntax

- Delete specific permissions from an account: Delete the permissions at a specific level from an account. The permission level is specified by priv_level.

```
REVOKE
    priv_type
      [, priv_type] ...
    ON priv_level
    FROM user [, user] ...
```

- Delete all permissions from an account: Delete all permissions at the database and table levels from an account.

```
REVOKE ALL PRIVILEGES, GRANT OPTION
    FROM user [, user] ...
```

Examples

Delete the CREATE, DROP, and INDEX permissions on the easydb.emp table from the hanson@'%' account.

```
REVOKE CREATE,DROP,INDEX ON easydb.emp FROM hanson@'%';
```

Delete all permissions from the lily@30.xx.xx.96 account.

```
REVOKE ALL PRIVILEGES,GRANT OPTION FROM lily@30.xx.xx.96;
```

Notice GRANT OPTION must be added to the preceding statement for compatibility with MySQL.

SHOW GRANTS statement

Syntax

```
SHOW GRANTS[FOR user@host];
```

Query all permissions:

```
SHOW GRANTS;
```

Query the permissions of an account:

```
SHOW GRANTS FOR user@host;
```

12.1.6.5. Delete an account

You can delete an account in the Cloud Native Distributed Database PolarDB-X (PolarDB-X) console or by using the command line.

Delete an account in the PolarDB-X console

Note You can delete only standard accounts that are created in the console.

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Account Management**.

5. Find the target account and click **Delete**.
6. In the **Delete Account** dialog box, click **OK**.

Delete an account by using the command line

Use the following syntax rule:

```
DROP USER user [, user] ...
```

For example:

Delete the lily@30.xx.xx.96 account.

```
DROP USER lily@30.xx.xx.96;
```

12.1.7. Database management

12.1.7.1. Create a database

After you create a instance, you must create a database that is based on one or more ApsaraDB RDS for MySQL instances.

Prerequisites

- An ApsaraDB RDS for MySQL instance in the same department of is created.
- The Resource Access Management (RAM) permissions are granted. For more information, see the *RAM Management* topic in the *Apsara Uni-manager User Guide*.

Context

provides two partition modes to help you create a database: **horizontal splitting** and **vertical splitting**. The two modes have the following features:

- The horizontal splitting mode allows you to scale out a database to linearly improve the overall storage capacity and concurrent throughput of the database.
- The vertical splitting mode allows you to import multiple existing databases across multiple ApsaraDB RDS for MySQL instances in batches. In addition, this mode provides capabilities of union queries and write transactions across databases on multiple ApsaraDB RDS for MySQL instances.

> **Note** The following list describes the supported partition modes of instances:
> - integrated instances support only Horizontal splitting.
> - non-integrated instances support both Horizontal splitting and Vertical splitting.

Horizontal splitting

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. On the **Basic Information** page, click **Create Database** in the upper-right corner.
5. In the **Enter Basic Information** step, enter the following information.

Parameter	Description
Partition Mode	Select **Horizontal Partitioning**.
Database Name	Enter a database name. The name must meet the following requirements: ◦ It must be 2 to 24 characters in length. ◦ It must start with a lowercase letter and end with a lowercase letter or digit. ◦ It can contain lowercase letters, digits, and underscores (_). ◦ It must be unique in the PolarDB-X instance.
Character Set	Select utf8, gbk, latin1, or utf8mb4.
DRDS Link Password	Set the connection password for the database. The password must meet the following requirements: ◦ It must be 8 to 30 characters in length. ◦ It must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and underscores (_).
Confirm Password	Enter the password again.

6. Click **Next**.
7. In the **Select RDS Instance** step, click the **Buy New RDS Instance** or **Use Existing RDS Instance** tab.

   Buy New RDS Instance:

   i. Click the **Buy New RDS Instance** tab.

   ii. Set **Storage Type**, **Series**, **Instance Specifications**, **Storage Capacity**, **Availability Zone**, and **Quantity**.

   iii. Click **Next**.

   Use Existing RDS Instance:

   i. Click the **Use Existing RDS Instance** tab.

   ii. In the left-side section, select the ApsaraDB RDS for MySQL instances to be added.

   iii. Click > to move the selected instances to the **Selected RDS Instances** section on the right.

   iv. Click **Next**.
8. After the precheck is passed in the **Precheck** step, click **Next**.

   > **Note** If the precheck fails, fix the bug as prompted.
9. In the **Preview** step, click **Next**. Wait until the database is created.

Vertical splitting

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. On the **Basic Information** page, click **Create Database** in the upper-right corner.
5. In the **Enter Basic Information** step, enter the following information.

Parameter	Description
Partition Mode	Select **Vertical Partitioning**.
Add Database	Select the database that you want to add. Perform the following steps: i. Select the specified ApsaraDB RDS for MySQL instance in the upper part of the page. ii. In the **RDS Database** section, select one or more databases and move them to the **Added Database** section on the right. **Note** To add a database from another ApsaraDB RDS for MySQL instance, repeat the preceding steps.
Character Set	Select utf8, gbk, latin1, or utf8mb4.
Account Type	Select **Create Account** or **Select Existing Account**.
Database Account	Enter a database name. The name must meet the following requirements: ○ It must be 2 to 20 characters in length. ○ It must start with a lowercase letter and end with a lowercase letter or digit. ○ It can contain uppercase letters, lowercase letters, digits, and underscores (_). ○ It must be unique in the PolarDB-X instance.
DRDS Link Password	Set the connection password for the database. The password must meet the following requirements: ○ It must be 8 to 30 characters in length. ○ It must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and underscores (_).
Confirm Password	Enter the password again.

6. Click **Next**.
7. After the precheck is passed in the **Precheck** step, click **Next**.

   Note If the precheck fails, fix the bug as prompted.

8. In the **Preview** step, click **Next**.
9. Wait until the database is created.

12.1.7.2. View a database

After the database is created, you can view the basic information of the database on the Database Management page in the console.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.

4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. Find the target database, and click **Manage** in the Actions column. The **Basic Information** page of the database appears.

> **Note** On the **Basic Information** page, you can delete the database or reset the password.

What's next

is fully compatible with the MySQL protocol. You can use **Command Line URL** on the MySQL client to connect to the instance and enter the user name and password to log on to the database. When you use the MySQL client, note the following points:

> **Note**
> - MySQL clients of some earlier versions have limits on the user name length, which cannot exceed 16 characters. The database name and user name are the same. If the database name exceeds 16 characters in length, an error is reported.
> - When you use the MySQL client, you must add the `-c` parameter to the HINT command. In , an annotation is used to implement HINT. If the `-c` parameter is not added, the annotation is lost and the HINT of is lost.

12.1.7.3. Storage management

This topic describes how to configure the read/write splitting feature of instances on the Storage Management page in Databases. You can perform operations and maintenance (O&M) operations on ApsaraDB RDS for MySQL instances, such as upgrading and downgrading configurations, adding read-only instances, and specifying the read policy.

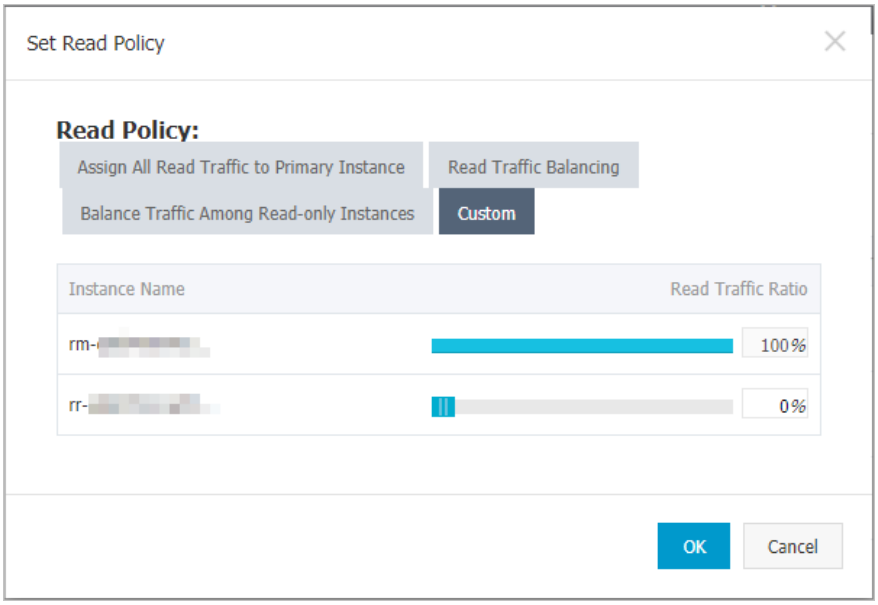
Prerequisites

A read-only instance is added to your ApsaraDB RDS for MySQL instance. For more information, see Add read-only instances.

> **Note** You can create an ApsaraDB RDS for MySQL read-only instance for a non-integrated instance in the ApsaraDB RDS for MySQL console. For more information, see **User Guide (ApsaraDB RDS for MySQL) > Instance Management > Read-only Instances** in *ApsaraDB RDS*.

Specify the read policy

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. In the left-side navigation pane, click **Storage Management**.
6. Find the primary ApsaraDB RDS for MySQL instance and click **Set Read Policy** in the Actions column.



provides the following three read policies:

- **Assign All Read Traffic to Primary Instance**: All the read traffic is sent to the primary RDS instance.
- **Balance Traffic Among Read-only Instances**: The traffic is evenly distributed to the primary ApsaraDB RDS for MySQL instance and the read-only ApsaraDB RDS for MySQL instances.
- **Read Traffic Balancing**: All the read traffic is sent to the read-only ApsaraDB RDS for MySQL instance.

You can also drag the slider next to a read-only instance to customize the read/write ratio.

> **Note**
>
> - On the Storage Management page, the read/write ratio is specified by read-only ApsaraDB RDS for MySQL instance. If a instance has multiple ApsaraDB RDS for MySQL instances, you must specify the read/write ratio for each ApsaraDB RDS for MySQL instance.
> - If you need to release a read-only ApsaraDB RDS for MySQL instance, set **Read Traffic Ratio** to 0 for the instance in the **Set Read Policy** dialog box. Otherwise, the traffic is continuously sent to the read-only instance. As a result, the instance fails to be released.

Add read-only instances

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. In the left-side navigation pane, click **Storage Management**.
6. Find the primary ApsaraDB RDS for MySQL instance that you want to manage and click **Add Read-only Instances** in the Actions column.
7. In the **Add Read-only Instances** dialog box, specify **Instance Specifications** and **Storage Capacity** for the read-only ApsaraDB RDS for MySQL instance, and click **OK**.

Change the specifications of an instance

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.

3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. In the left-side navigation pane, click **Storage Management**.
6. Find the ApsaraDB RDS for MySQL instance for which you want to change the specifications.
   - Primary ApsaraDB RDS for MySQL instance: Find the primary ApsaraDB RDS for MySQL instance and choose **More > Change Specifications** in the Actions column.
   - Read-only ApsaraDB RDS for MySQL instance: Find the read-only ApsaraDB RDS for MySQL instance and click **Change Specifications** in the Actions column.
7. In the **Change Specifications** dialog box, specify **Instance Specifications** and **Storage Capacity** for the ApsaraDB RDS for MySQL instance, and click **OK**.

12.1.7.4. Smooth scale-out

Physical bottlenecks already exist in the underlying storage of the logical database. In this case, you must perform horizontal scaling to solve the physical bottlenecks. For example, when the remaining disk space is close to 30%, you can smoothly scale out the database to improve database performance. The smooth scale-out process is divided into four steps: configuration, migration, switchover, and cleanup.

Configure a scale-out task

> **Note** In smooth scale-out, ApsaraDB RDS for MySQL instances are added, and some source database shards are migrated to the new ApsaraDB RDS for MySQL instances. This way, the overall data storage capacity is increased, and the number of requests that a single ApsaraDB RDS for MySQL instance needs to process is reduced.

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. Find the database that you want to smoothly scale out, and click **Manage** in the Actions column. The **Basic Information** page of the database appears.
6. In the left-side navigation pane, choose **Configuration and Management > Scale-out Management**.
7. In the upper-right corner of the **Scale-out Management** page, click **Scale Out**.
8. Select **Smooth Scale-out** and click **Next**.
9. After the precheck is passed in the **Precheck** step, click **Next**.

   > **Note** If the precheck fails, fix the bug as prompted.

10. In the **Select RDS Instance** step, click the **Buy New RDS Instance** or **Use Existing RDS Instance** tab.

    > **Note** For non-integrated instances, you can select only Use Existing RDS Instance.

    Buy new RDS instance:

    i. Click the **Buy New RDS Instance** tab.

    ii. Set **Storage Type**, **Series**, **Instance Specifications**, **Storage Capacity**, **Availability Zone**, and **Quantity**.

    iii. Click **Next**.

Use existing RDS instance:

i. Click the **Use Existing RDS Instance** tab.

ii. In the left-side section, select the ApsaraDB RDS for MySQL instances to be added.

iii. Click > to move the selected instances to the **Selected RDS Instances** section on the right.

iv. Click **Next**.

11. In the **Preview** step, click **Start Scale-out**.

> **Note** By default, the console evenly distributes the physical database shards to the ApsaraDB RDS for MySQL instances that you added. You can also manually add or delete physical database shards to or from the new ApsaraDB RDS for MySQL instances.

12. Click the icon in the upper-right corner to view the progress of the scale-out task.

Migration

Some physical database shards are migrated during smooth scale-out.

The migration does not modify the data in the database or affect online services. Before the switchover, you can cancel the smooth scale-out operation by rolling back the scale-out.

> **Note**
> - This is because before the switchover, the current scale-out operation does not have a real impact on the existing data in the database.
> - During scale-out, the binary log files of the source RDS instance are not cleaned. This may result in insufficient disk space. Therefore, you must reserve sufficient disk space on the source ApsaraDB RDS for MySQL instance. We recommend that you reserve more than 30% of the disk space. If the disk space is not sufficient, you can submit a ticket to expand the storage space of the ApsaraDB RDS for MySQL instance.
> - To reduce the load of read operations on the source ApsaraDB RDS for MySQL instance, perform scale-out when the load on the source ApsaraDB RDS for MySQL instance is low.
> - During the scale-out, do not submit data definition language (DDL) tasks in the console or connect to the instance to execute DDL statements. Otherwise, the scale-out task may fail.
> - Before the scale-out, make sure that all the tables in the source database have primary keys.

After historical data and incremental data are migrated, the migration progress reaches 100%. Then, you can **switch** the read and write traffic to the new ApsaraDB RDS for MySQL instance or **roll back** the scale-out.

Switchover

The switchover task switches the read and write traffic to the new ApsaraDB RDS for MySQL instance. The entire process takes 3 to 5 minutes. During the switchover, the service is not affected except for one or two transient connections. Perform the switchover during off-peak hours.

1. In the upper-right corner of the **Basic Information** page, click the icon. The **Task Progress** dialog box appears.
2. In the **Task Progress** dialog box, click **Switch Over** and click **OK**.

   During the switchover, a switchover task is generated and appears in the task progress.
3. After the switchover is complete, the **Clean Up** button appears in the **Task Progress** dialog box.

Cleanup

In this step, the migrated database shards are deleted from the source ApsaraDB RDS for MySQL instance.

1. After the switchover is complete, click **Clean Up** next to the task.
2. Click **OK**. A cleanup task appears in the Task Progress dialog box.

> **Note**
> - The cleanup task is an asynchronous task. You can view the execution status in the Task Progress dialog box.
> - After the cleanup task is complete, the smooth scale-out process ends. The new ApsaraDB RDS for MySQL instance becomes the storage node of the logical database.
> - You can implement smooth scale-out by migrating physical database shards. If the number of database shards exceeds the capacity of a single ApsaraDB RDS for MySQL instance, no further scale-out is allowed. In this case, you can submit a ticket to apply for increasing the number of database shards and scaling out the database. In this case, data is calculated again based on the hash algorithm to reallocate the data.
> - The cleanup task deletes database shards that are no longer used after the current scale-out. You can back up the database shards before you run the cleanup task.
> - The cleanup operation increases the load on databases. We recommend that you perform this operation during off-peak hours.

12.1.7.5. Set the IP address whitelist

allows you to configure the IP address whitelist to block unauthorized access requests. This topic describes how to configure the IP address whitelist.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. Find the target database, and click **Manage** in the Actions column. The **Basic Information** page of the database appears.
6. On the **Basic Information** page of the database, choose **Data Security > Whitelist Settings** in the left-side navigation pane.
7. On the Whitelist Settings page, click **Manually Modify**.
8. Enter the IP addresses that are allowed to access the database, and click **OK**.

Note
- The following formats are supported in the whitelist:
  - Single IP addresses, such as 192.168.1.1.
  - IP addresses in CIDR format, such as 192.168.1.1/24.
  - IP addresses that include an asterisk (*) as a wildcard, such as 192.168.1. *. This example indicates that hosts with an IP address in the range from 192.168.1.1 to 192.168.1.254 are allowed to access the database.
  - IP range, such as 192.168.1.1-192.168.1.254.
- If you need to add multiple IP addresses or IP ranges, separate them with commas (,). Do not use spaces before and after the commas, for example, 192.168.0.1,172.16.213.9.

12.1.7.6. Delete a database

This topic describes how to delete a database in the Cloud Native Distributed Database PolarDB-X (PolarDB-X) console.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.
4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. Find the target database and click **Delete**.

   Warning You cannot recover databases that have been deleted. Exercise caution when you perform this operation.
6. In the **Delete Database** dialog box, click **OK**.

12.1.7.7. Fix database shard connections

This topic describes how to manually fix database shard connections when a instance cannot access an ApsaraDB RDS for MySQL instance.

Context

When you use a instance, you must access the mounted ApsaraDB RDS for MySQL instances. If the network configuration of a connected ApsaraDB RDS for MySQL instance changes, the network connection between the instance and the ApsaraDB RDS for MySQL instance is broken. For example, the network configuration changes if the zone is switched or the network type is changed from the classic network to VPC. Therefore, the instance cannot access the ApsaraDB RDS for MySQL instance. In this case, you must manually fix the database shard connection in the console. This way, you can recover the network connection between the instance and the ApsaraDB RDS for MySQL instance.

Procedure

1. Log on to the PolarDB-X console.
2. Find the target instance in the instance list.
3. Click the target instance ID or choose **More > Manage** from the Actions column of the instance to access the **Basic Information** page.

4. In the left-side navigation pane, choose **Configuration and Management > Database Management**.
5. Find the target database, and click **Manage** in the Actions column. The **Basic Information** page of the database appears.
6. In the **Shortcuts** section, click **Fix Database Shard Connections**.
7. In the message that appears, click **OK**.

12.1.8. Custom control commands

provides a series of auxiliary SQL commands to help you conveniently use .

12.1.8.1. Overview

provides auxiliary statements for you to use and maintain .

Take note of the following syntax description: The identifier provided by the user is in `[]` and optional content is in `()`. In addition, this section is applicable to the current version. If some statements are unavailable, the reason is that the version you use is earlier than required.

12.1.8.2. SHOW HELP statement

This topic describes how to use the SHOW HELP statement to view all the auxiliary SQL statements of and their descriptions.

Execute the `SHOW HELP;` statement. The following response returned:

```
+----------------------------------------+-------------------------------------------------------+-
-----------------------------------------+
| STATEMENT                              | DESCRIPTION                                           |
EXAMPLE                                  |
+----------------------------------------+-------------------------------------------------------+-
-----------------------------------------+
| show rule                              | Report all table rule                                 |
|
| show rule from TABLE                   | Report table rule                                     |
show rule from user                      |
| show full rule from TABLE              | Report table full rule                                |
show full rule from user                 |
| show topology from TABLE               | Report table physical topology                        |
show topology from user                  |
| show partitions from TABLE             | Report table dbPartition or tbPartition columns       |
show partitions from user                |
| show broadcasts                        | Report all broadcast tables                           |
|
| show datasources                       | Report all partition db threadPool info               |
|
| show node                              | Report master/slave read status                       |
|
| show slow                              | Report top 100 slow sql                               |
|
| show physical_slow                     | Report top 100 physical slow sql                      |
|
| clear slow                             | Clear slow data                                       |
|
| trace SQL                              | Start trace sql, use show trace to print profiling data |
trace select count(*) from user; show trace |
| show trace                             | Report sql execute profiling info                     |
|
| explain SQL                            | Report sql plan info                                  |
explain select count(*) from user        |
| explain detail SQL                     | Report sql detail plan info                           |
explain detail select count(*) from user |
| explain execute SQL                    | Report sql on physical db plan info                   |
explain execute select count(*) from user |
| show sequences                         | Report all sequences status                           |
|
| create sequence NAME [start with COUNT] | Create sequence                                       |
create sequence test start with 0        |
| alter sequence NAME [start with COUNT] | Alter sequence                                        |
alter sequence test start with 100000    |
| drop sequence NAME                     | Drop sequence                                         |
drop sequence test                       |
+----------------------------------------+-------------------------------------------------------+-
-----------------------------------------+
20 rows in set (0.00 sec)
```

12.1.8.3. Statements for viewing rules and node topologies

This topic describes the statements that are used to view rules and node topologies, and provides examples of the statements.

SHOW RULE [FROM tablename]

You must take note of the following usage notes:

- `SHOW RULE`: You can execute this statement to view the partitioning information of each logical table in a database.
- `SHOW RULE FROM tablename`: You can execute this statement to view the partitioning information of a specified logical table in a database.

The following section describes the important columns:

- BROADCAST: indicates whether the table is a broadcast table. 0 indicates No and 1 indicates Yes.
- DB_PARTITION_KEY: indicates the database shard key. If no database shards exist, the parameter value is empty.
- DB_PARTITION_POLICY: indicates the database sharding policy. Valid values are hash and date policies such as YYYYMM, YYYYDD, and YYYYWEEK.
- DB_PARTITION_COUNT: indicates the number of database shards.
- TB_PARTITION_KEY: indicates the table shard key. If no table shards exist, the parameter value is empty.
- TB_PARTITION_POLICY: indicates the table sharding policy. Valid values are hash and date policies such as MM, DD, MMDD, and WEEK.
- TB_PARTITION_COUNT: indicates the number of table shards.

Execute the `SHOW RULE;` statement. The following response is returned:

```
+------+--------------+-----------+-----------------+---------------------+---------------------+-----
------------+---------------------+--------------------+
| ID   | TABLE_NAME   | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT | TB_P
ARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+--------------+-----------+-----------------+---------------------+---------------------+-----
------------+---------------------+--------------------+
|    0 | dept_manager |         0 |                 | NULL                | 1                   |
| NULL                | 1                   |
|    1 | emp          |         0 | emp_no          | hash                | 8                   | id
| hash                | 2                   |
|    2 | example      |         0 | shard_key       | hash                | 8                   |
| NULL                | 1                   |
+------+--------------+-----------+-----------------+---------------------+---------------------+-----
------------+---------------------+--------------------+
3 rows in set (0.01 sec)
```

SHOW FULL RULE [FROM tablename]

You can execute this SQL statement to view the sharding rules of logical tables in a database. This statement shows more detailed information than the SHOW RULE statement.

The following section describes the important columns:

- BROADCAST: indicates whether the table is a broadcast table. 0 indicates No and 1 indicates Yes.
- JOIN_GROUP: a reserved field. Currently, it is meaningless.
- ALLOW_FULL_TABLE_SCAN: indicates whether to allow data query when no table shard key is specified for database or table sharding. If this parameter is set to True, each physical table is scanned to find data that meets the condition. This is a full table scan.
- DB_NAME_PATTERN: The digit 0 inside a pair of braces ({}) in DB_NAME_PATTERN is a placeholder. When the SQL statement is executed, this value is replaced by the value of DB_RULES_STR. The number of digits remains unchanged. For example, if the value of DB_NAME_PATTERN is SEQ_{0000}_RDS and the value of DB_RULES_STR is [1,2,3,4], four DB_NAME values are generated: SEQ_0001_RDS, SEQ_0002_RDS, SEQ_0003_RDS, and SEQ_0004_RDS.
- DB_RULES_STR: indicates the database sharding rule.
- TB_NAME_PATTERN: The digit 0 inside a pair of braces ({}) in TB_NAME_PATTERN is a placeholder. When the SQL statement is executed, this value is replaced by the value of TB_RULES_STR. The number of digits remains unchanged. For example, if the value of TB_NAME_PATTERN is table_{00} and the value of TB_RULES_STR is [1,2,3,4,5,6,7,8], eight tables are generated: table_01, table_02, table_03, table_04, table_05, table_06,

table_07, and table_08.

- TB_RULES_STR: indicates the table sharding rule.
- PARTITION_KEYS: indicates the database and table shard keys. When both database sharding and table sharding are performed, the database shard key is placed before the table shard key.
- DEFAULT_DB_INDEX: indicates the database shard in which a single-database non-partitioned table is stored.

Execute the `SHOW FULL RULE;` statement. The following response is returned:

```
+------+--------------+-----------+------------+----------------------+------------------------------
----------------------+-----------------------------------------------------------+-----------------+-------
------------------------------+----------------+-----------------------------------------------------+
| ID   | TABLE_NAME   | BROADCAST | JOIN_GROUP | ALLOW_FULL_TABLE_SCAN | DB_NAME_PATTERN
| DB_RULES_STR                                        | TB_NAME_PATTERN | TB_RULES_STR
| PARTITION_KEYS | DEFAULT_DB_INDEX                                   |
+------+--------------+-----------+------------+----------------------+------------------------------
----------------------+-----------------------------------------------------------+-----------------+-------
------------------------------+----------------+-----------------------------------------------------+
|    0 | dept_manager |         0 | NULL       |                    0 | SEQ_TEST_1487767780814RGKKSEQ
_TEST_WNJG_0000_RDS   | NULL                                                      | dept_manager    | NULL
| NULL           | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0000_RDS |
|    1 | emp          |         0 | NULL       |                    1 | SEQ_TEST_1487767780814RGKKSEQ
_TEST_WNJG_{0000}_RDS | ((#emp_no,1,8#).longValue().abs() % 8)              | emp_{0}         | ((#id,
1,2#).longValue().abs() % 2) | emp_no      id | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0000_RDS |
|    2 | example      |         0 | NULL       |                    1 | SEQ_TEST_1487767780814RGKKSEQ
_TEST_WNJG_{0000}_RDS | ((#shard_key,1,8#).longValue().abs() % 8).intdiv(1) | example         | NULL
| shard_key      | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0000_RDS |
+------+--------------+-----------+------------+----------------------+------------------------------
----------------------+-----------------------------------------------------------+-----------------+-------
------------------------------+----------------+-----------------------------------------------------+
3 rows in set (0.01 sec)
```

SHOW TOPOLOGY FROM tablename

You can execute this SQL statement to view the topology of a specified logical table. The information contains the database shards to which data in the logical table is partitioned and the table shards in each database shard.

Execute the `SHOW TOPOLOGY FROM emp;` statement. The following response is returned:

```
+------+-----------------------------------------------------+------------+
| ID   | GROUP_NAME                                          | TABLE_NAME |
+------+-----------------------------------------------------+------------+
|    0 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0000_RDS | emp_0      |
|    1 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0000_RDS | emp_1      |
|    2 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0001_RDS | emp_0      |
|    3 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0001_RDS | emp_1      |
|    4 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0002_RDS | emp_0      |
|    5 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0002_RDS | emp_1      |
|    6 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0003_RDS | emp_0      |
|    7 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0003_RDS | emp_1      |
|    8 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0004_RDS | emp_0      |
|    9 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0004_RDS | emp_1      |
|   10 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0005_RDS | emp_0      |
|   11 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0005_RDS | emp_1      |
|   12 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0006_RDS | emp_0      |
|   13 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0006_RDS | emp_1      |
|   14 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0007_RDS | emp_0      |
|   15 | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0007_RDS | emp_1      |
+------+-----------------------------------------------------+------------+
16 rows in set (0.01 sec)
```

SHOW PARTITIONS FROM tablename

You can execute this SQL statement to view the database shard key and table shard key, which are separated with commas (,). If the returned results contain two values, both database sharding and table sharding are performed. The first value is the database shard key and the second value is the table shard key. If only one value is returned, only database sharding is performed. This value is the database shard key.

Execute the `SHOW PARTITIONS FROM emp;` statement. The following response is returned:

```
+-----------+
| KEYS      |
+-----------+
| emp_no,id |
+-----------+
1 row in set (0.00 sec)
```

SHOW BROADCASTS

You can execute this SQL statement to view broadcast tables.

Execute the `SHOW BROADCASTS;` statement. The following response is returned:

```
+------+------------+
| ID   | TABLE_NAME |
+------+------------+
|    0 | brd2       |
|    1 | brd_tbl    |
+------+------------+
2 rows in set (0.01 sec)
```

SHOW DATASOURCES

You can execute this SQL statement to view the information about the underlying storage, including the database name, database group name, connection URL, username, storage type, read and write weights, and connection pool information.

The following section describes the important columns:

- SCHEMA: indicates the database name.
- GROUP: indicates the database group name. Grouping aims to manage multiple groups of databases that have identical data. For example, the databases can be the primary and secondary databases after data replication that is implemented by ApsaraDB RDS for MySQL. It is used for read/write splitting and primary/secondary switchovers.
- URL: indicates the connection information of the underlying ApsaraDB RDS for MySQL instances.
- TYPE: indicates the type of the underlying storage. Currently, only ApsaraDB RDS for MySQL instances are supported.
- READ_WEIGHT: indicates the read weight of the database. When the primary ApsaraDB RDS for MySQL instance is under a heavy load of read requests, you can use the read/write splitting feature of to distribute the read traffic. This way, the read pressure on the primary instance can be reduced. automatically identifies the read and write traffic. It directs the write traffic to the primary ApsaraDB RDS for MySQL instance and the read traffic to all ApsaraDB RDS for MySQL instances based on the specified weight.
- WRITE_WEIGHT: indicates the write weight. For more information, see READ_WEIGHT.

Execute the `SHOW DATASOURCES;` statement. The following response is returned:

```
+------+----------------------------+-------------------------------------------------+----------------
--------------------------------+-----------------------------------------------------------------------
--------------+-----------+-------+------+------+------+--------------+----------+--------------+----
----------+-------------------------------------------------+-------------+--------------+
| ID   | SCHEMA                     | NAME                                              | GROUP
| URL                                                                                 | USER      | TYPE
| INIT | MIN  | MAX  | IDLE_TIMEOUT | MAX_WAIT | ACTIVE_COUNT | POOLING_COUNT | ATOM
| READ_WEIGHT | WRITE_WEIGHT |
+------+----------------------------+-------------------------------------------------+----------------
--------------------------------+-----------------------------------------------------------------------
--------------+-----------+-------+------+------+------+--------------+----------+--------------+----
----------+-------------------------------------------------+-------------+--------------+
|    0 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0000_iiab_1 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0000_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0000 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0000_iiab | 10          | 10           |
|    1 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0001_iiab_2 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0001_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0001 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0001_iiab | 10          | 10           |
|    2 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0002_iiab_3 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0002_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0002 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0002_iiab | 10          | 10           |
|    3 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0003_iiab_4 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0003_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0003 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0003_iiab | 10          | 10           |
|    4 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0004_iiab_5 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0004_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0004 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0004_iiab | 10          | 10           |
|    5 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0005_iiab_6 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0005_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0005 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0005_iiab | 10          | 10           |
|    6 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0006_iiab_7 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0006_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0006 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0006_iiab | 10          | 10           |
|    7 | seq_test_1487767780814rgkk | rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0007_iiab_8 | SEQ_TEST_148776
7780814RGKKSEQ_TEST_WNJG_0007_RDS | jdbc:mysql://rds1ur80kcv8g3t6p3ol.mysql.rds.aliyuncs.com:3306/seq_
test_wnjg_0007 | jnkinsea0 | mysql | 0    | 24   | 72   | 15           | 5000     | 0            | 1
| rds1ur80kcv8g3t6p3ol_seq_test_wnjg_0007_iiab | 10          | 10           |
+------+----------------------------+-------------------------------------------------+----------------
--------------------------------+-----------------------------------------------------------------------
--------------+-----------+-------+------+------+------+--------------+----------+--------------+----
----------+-------------------------------------------------+-------------+--------------+
8 rows in set (0.01 sec)
```

SHOW NODE

You can execute this SQL statement to view the accumulative number of read and write operations and accumulative read and write weights of a physical database.

The following section describes the important columns:

- MASTER_READ_COUNT: indicates the accumulative number of read-only queries processed by the primary ApsaraDB RDS for MySQL instance.

- SLAVE_READ_COUNT: indicates the accumulative number of read-only queries processed by the secondary ApsaraDB RDS for MySQL instances.
- MASTER_READ_PERCENT: indicates the actual percentage of read-only queries processed by the primary ApsaraDB RDS for MySQL instance, instead of the specified percentage.
- SLAVE_READ_PERCENT: indicates the actual percentage of read-only queries processed by the secondary ApsaraDB RDS for MySQL instances, instead of the specified percentage.

> **Note**
> - Read-only queries in transactions are sent to the primary ApsaraDB RDS for MySQL instance.
> - The `MASTER_READ_PERCENT` and `SLAVE_READ_PERCENT` fields indicate the accumulative historical data. After the ratio between the read weight and the write weight is changed, these values do not immediately reflect the latest ratio, which appears after a long period of time.

Execute the `SHOW NODE;` statement. The following response is returned:

```
+------+-------------------------------------------------+------------------+------------------+----
----------------+--------------------+
| ID   | NAME                                            | MASTER_READ_COUNT | SLAVE_READ_COUNT | MAS
TER_READ_PERCENT | SLAVE_READ_PERCENT |
+------+-------------------------------------------------+------------------+------------------+----
----------------+--------------------+
| 0    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0000_RDS |               12 |                0 | 100
%                | 0%                 |
| 1    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0001_RDS |                0 |                0 | 0%
| 0%                 |
| 2    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0002_RDS |                0 |                0 | 0%
| 0%                 |
| 3    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0003_RDS |                0 |                0 | 0%
| 0%                 |
| 4    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0004_RDS |                0 |                0 | 0%
| 0%                 |
| 5    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0005_RDS |                0 |                0 | 0%
| 0%                 |
| 6    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0006_RDS |                0 |                0 | 0%
| 0%                 |
| 7    | SEQ_TEST_1487767780814RGKKSEQ_TEST_WNJG_0007_RDS |                0 |                0 | 0%
| 0%                 |
+------+-------------------------------------------------+------------------+------------------+----
----------------+--------------------+
8 rows in set (0.01 sec)
```

12.1.8.4. Statements for SQL optimization

This topic describes statements for SQL optimization and provides examples of the statements.

SHOW [FULL] SLOW [WHERE expr] [limit expr]

SQL statements that take more than one second to execute are slow SQL statements. Logical slow SQL statements are slow SQL statements sent from an application to a instance.

- `SHOW SLOW`: You can execute this statement to view the top 100 logical slow SQL statements. These SQL statements are recorded since the instance is started or the last time when the `CLEAR SLOW` statement is executed.

> **Note** The recorded top 100 slow SQL statements are cached in the system. When the PolarDB-X instance is restarted or the `CLEAR SLOW` statement is executed, these statements will be discarded.

- `SHOW FULL SLOW` : You can execute this SQL statement to view all the logical slow SQL statements since the PolarDB-X instance is started. These SQL statements are recorded and persisted to the built-in database of the instance. The upper limit for the number of records is specified in the specifications of the PolarDB-X instance. The instance deletes earlier slow SQL statements when the disk space is insufficient. If the specifications of the PolarDB-X instance include 4 cores and 4 GB of memory, a maximum of 10,000 slow SQL statements can be recorded, including logical slow and physical slow SQL statements. If the specifications of the PolarDB-X instance include 8 cores and 8 GB of memory, a maximum of 20,000 slow SQL statements can be recorded, including logical slow and physical slow SQL statements. The same rule applies to other instance specifications.

The following section describes the important columns:

- HOST: the IP address of the host from which the SQL statement is sent.
- START_TIME: the time when the SQL statement starts to be executed.
- EXECUTE_TIME: the execution duration of the SQL statement.
- AFFECT_ROW: For DML statements, this parameter indicates the number of affected rows. For query statements, this parameter indicates the number of returned records.

Execute the `SHOW SLOW WHERE execute_time > 1000 limit 1;` statement. The following response is returned:

```
+-----------+---------------------+--------------+------------+-----------+
| HOST      | START_TIME          | EXECUTE_TIME | AFFECT_ROW | SQL       |
+-----------+---------------------+--------------+------------+-----------+
| 127.0.0.1 | 2016-03-16 13:02:57 |         2785 |          7 | show rule |
+-----------+---------------------+--------------+------------+-----------+
1 row in set (0.02 sec)
```

SHOW [FULL] PHYSICAL_SLOW [WHERE expr] [limit expr]

SQL statements that take more than one second to execute are slow SQL statements. Logical slow SQL statements are slow SQL statements sent from an application to a instance.

- `SHOW SLOW` : You can execute this statement to view the top 100 logical slow SQL statements. These SQL statements are recorded since the instance is started or the last time when the `CLEAR SLOW` statement is executed.

> **Note** The recorded top 100 slow SQL statements are cached in the system. When the PolarDB-X instance is restarted or the `CLEAR SLOW` statement is executed, these statements will be discarded.

- `SHOW FULL SLOW` : You can execute this SQL statement to view all the logical slow SQL statements since the PolarDB-X instance is started. These SQL statements are recorded and persisted to the built-in database of the instance. The upper limit for the number of records is specified in the specifications of the PolarDB-X instance. The instance deletes earlier slow SQL statements when the disk space is insufficient. If the specifications of the PolarDB-X instance include 4 cores and 4 GB of memory, a maximum of 10,000 slow SQL statements can be recorded, including logical slow and physical slow SQL statements. If the specifications of the PolarDB-X instance include 8 cores and 8 GB of memory, a maximum of 20,000 slow SQL statements can be recorded, including logical slow and physical slow SQL statements. The same rule applies to other instance specifications.

The following section describes the important columns:

- GROUP_NAME: the name of the group to which the database that executes the SQL statement belongs.
- START_TIME: the time when the SQL statement starts to be executed.
- EXECUTE_TIME: the execution duration of the SQL statement.
- AFFECT_ROW: For DML statements, this parameter indicates the number of affected rows. For query statements, this parameter indicates the number of returned records.

Execute the `SHOW PHYSICAL_SLOW;` statement. The following response is returned:

```
+----------------+-----------------------------------+--------------------+--------------+-----------
-------+-------------------------+-----------------------+------------+-----------------+
| GROUP_NAME     | DBKEY_NAME                        | START_TIME         | EXECUTE_TIME | SQL_EXECUT
E_TIME | GETLOCK_CONNECTION_TIME | CREATE_CONNECTION_TIME | AFFECT_ROW | SQL             |
+----------------+-----------------------------------+--------------------+--------------+-----------
-------+-------------------------+-----------------------+------------+-----------------+
| TDDL5_00_GROUP | db218249098_sqa_zmf_tddl5_00_3309 | 2016-03-16 13:05:38 |         1057 |
1011 |                       0 |                     0 |          1 | select sleep(1) |
+----------------+-----------------------------------+--------------------+--------------+-----------
-------+-------------------------+-----------------------+------------+-----------------+
1 row in set (0.01 sec)
```

CLEAR SLOW

You can execute this SQL statement to clear the top 100 logical slow SQL statements and the top 100 physical slow SQL statements. These statements are recorded since the instance is started or the last time when the `CLEAR SLOW` statement is executed.

> **Note** Both `SHOW SLOW` and `SHOW PHYSICAL_SLOW` can be executed to show the top 100 slow SQL statements. If the `CLEAR SLOW` statement has not been executed for a long time, the SQL statements might be recorded a long time ago. Therefore, we recommend that you execute the `CLEAR SLOW` statement after statements for SQL optimization are executed. After the system runs for a while, check the optimized results of slow SQL statements.

Execute the `CLEAR SLOW;` statement. The following response is returned:

```
Query OK, 0 rows affected (0.00 sec)
```

EXPLAIN SQL

You can execute this SQL statement to view the execution plan of a specified SQL statement in . Note that this SQL statement is not actually executed.

Examples

You can execute this SQL statement to view the execution plan of the `SELECT * FROM doctest` statement. The data of the doctest table is partitioned into database shards based on the id column. Based on the execution plan, the SQL statement will be routed to each database shard for execution, and the execution results will be aggregated.

Execute the `EXPLAIN SELECT * FROM doctest;` statement. The following response is returned:

```
+-----------------------------------------------+-------------------------------------+---------+
| GROUP_NAME                                    | SQL                                 | PARAMS |
+-----------------------------------------------+-------------------------------------+---------+
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0000_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0001_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0002_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0003_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0004_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0005_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0006_RDS | select `doctest`.`id` from `doctest` | {}     |
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0007_RDS | select `doctest`.`id` from `doctest` | {}     |
+-----------------------------------------------+-------------------------------------+---------+
8 rows in set (0.00 sec)
```

You can execute this SQL statement to view the execution plan of the `SELECT * FROM doctest WHERE id = 1` statement. The data of the doctest table is partitioned into database shards based on the id column. Based on the execution plan, the PolarDB-X instance will calculate a specified database shard based on the shard key, which is id. Then, the PolarDB-X instance will directly route the SQL statement to the database shard and aggregate the execution results.

Execute the `EXPLAIN SELECT * FROM doctest WHERE id = 1;` statement. The following response is returned:

```
+-----------------------------------------------+-----------------------------------------------------
------------+---------+
| GROUP_NAME                                    | SQL
| PARAMS |
+-----------------------------------------------+-----------------------------------------------------
------------+---------+
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0001_RDS | select `doctest`.`id` from `doctest` where (`doctes
t`.`id` = 1) | {}     |
+-----------------------------------------------+-----------------------------------------------------
------------+---------+
1 row in set (0.01 sec)
```

EXPLAIN DETAIL SQL

You can execute this SQL statement to view the execution plan of a specified SQL statement in . Note that this SQL statement is not actually executed.

Execute the `EXPLAIN DETAIL SELECT * FROM doctest WHERE id = 1;` statement. The following response is returned:

```
+-----------------------------------------------+-----------------------------------------------------
------------------------------------------------------------------------------------------------------
-----------------------------------------------+---------+
| GROUP_NAME                                    | SQL
| PARAMS |
+-----------------------------------------------+-----------------------------------------------------
------------------------------------------------------------------------------------------------------
-----------------------------------------------+---------+
| DOCTEST_1488704345426RCUPDOCTEST_CAET_0001_RDS | Query from doctest as doctest
    keyFilter:doctest.id = 1
    queryConcurrency:SEQUENTIAL
    columns:[doctest.id]
    tableName:doctest
    executeOn:DOCTEST_1488704345426RCUPDOCTEST_CAET_0001_RDS
 | NULL   |
+-----------------------------------------------+-----------------------------------------------------
------------------------------------------------------------------------------------------------------
-----------------------------------------------+---------+
1 row in set (0.02 sec)
```

EXPLAIN EXECUTE SQL

You can execute this SQL statement to view the execution plan of a specified SQL statement on an underlying ApsaraDB RDS for MySQL instance. This statement is equivalent to the EXPLAIN statement in MySQL.

Execute the `EXPLAIN EXECUTE SELECT * FROM tddl_mgr_log limit 1;` statement. The following response is returned:

```
+----+-------------+--------------+------+---------------+------+---------+------+------+-------+
| id | select_type | table        | type | possible_keys | key  | key_len | ref  | rows | Extra |
+----+-------------+--------------+------+---------------+------+---------+------+------+-------+
|  1 | SIMPLE      | tddl_mgr_log | ALL  | NULL          | NULL | NULL    | NULL |    1 | NULL  |
+----+-------------+--------------+------+---------------+------+---------+------+------+-------+
1 row in set (0.07 sec)
```

TRACE SQL and SHOW TRACE

You can execute these SQL statements to view the execution results of an SQL statement. Note that you must use the TRACE SQL statement and the SHOW TRACE statement together. The difference between the TRACE SQL statement and the EXPLAIN SQL statement is that the TRACE SQL statement is actually executed.

For example, you can execute these statements to view the execution results of the SELECT 1 statement.

Execute the `TRACE SELECT 1;` statement. The following response is returned:

```
+---+
| 1 |
+---+
| 1 |
+---+
1 row in set (0.03 sec)
```

Execute the `SHOW TRACE;` statement. The following response is returned:

```
+------+----------+----------------+----------------------------------+---------------+--------------
------------+------+-----------+--------+
| ID   | TYPE     | GROUP_NAME     | DBKEY_NAME                       | TIME_COST(MS) | CONNECTION_TI
ME_COST(MS) | ROWS | STATEMENT | PARAMS |
+------+----------+----------------+----------------------------------+---------------+--------------
------------+------+-----------+--------+
|    0 | Optimize | DRDS           | DRDS                             | 3             | 0.00
|    0 | select 1  | NULL   |
|    1 | Query    | TDDL5_00_GROUP | db218249098_sqa_zmf_tddl5_00_3309 | 7             | 0.15
|    1 | select 1  | NULL   |
+------+----------+----------------+----------------------------------+---------------+--------------
------------+------+-----------+--------+
2 rows in set (0.01 sec)
```

CHECK TABLE tablename

You can execute this SQL statement to check a data table. This SQL statement can be used when you fail to create a table by using a DDL statement.

- If the data table is a partitioned table, this SQL statement allows you to check whether an underlying physical table shard is missing and whether the columns and indexes of the underlying physical table shard are consistent.
- If the data table is a single-database non-partitioned table, this SQL statement allows you to check whether this table exists.

Execute the `CHECK TABLE tddl_mgr_log;` statement. The following response is returned:

```
+------------------------+-------+----------+----------+
| TABLE                  | OP    | MSG_TYPE | MSG_TEXT |
+------------------------+-------+----------+----------+
| TDDL5_APP.tddl_mgr_log | check | status   | OK       |
+------------------------+-------+----------+----------+
1 row in set (0.56 sec)
```

Execute the `CHECK TABLE tddl_mg;` statement. The following response is returned:

```
+-------------------+-------+----------+-----------------------------------------+
| TABLE             | OP    | MSG_TYPE | MSG_TEXT                                |
+-------------------+-------+----------+-----------------------------------------+
| TDDL5_APP.tddl_mg | check | Error    | Table 'tddl5_00.tddl_mg' doesn't exist |
+-------------------+-------+----------+-----------------------------------------+
1 row in set (0.02 sec)
```

SHOW TABLE STATUS [LIKE 'pattern' | WHERE expr]

You can execute this SQL statement to query the information about a table. This statement aggregates the data of all underlying physical table shards.

The following section describes the important columns:

- NAME: indicates the name of the table.
- ENGINE: indicates the storage engine of the table.
- VERSION: indicates the version of the storage engine of the table.
- ROW_FORMAT: indicates the format of the rows in the table. Valid values include Dynamic, Fixed, and Compressed. The value Dynamic indicates that the row length is variable, for example, a VARCHAR or BLOB field. The value Fixed indicates that the row length is constant, for example, a CHAR or INTEGER field.
- ROWS: indicates the number of rows in the table.
- AVG_ROW_LENGTH: indicates the average number of bytes in each row.
- DATA_LENGTH: indicates the data volume of the entire table. Unit: bytes.
- MAX_DATA_LENGTH: indicates the maximum volume of data that can be stored in the table.
- INDEX_LENGTH: indicates the size of the disk space occupied by indexes.
- CREATE_TIME: indicates the time when the table was created.
- UPDATE_TIME: indicates the time when the table was last updated.
- COLLATION: indicates the default character set and character sorting rule of the table.
- CREATE_OPTIONS: indicates all the other options specified when the table was created.

Execute the `SHOW TABLE STATUS LIKE 'multi_db_multi_tbl';` statement. The following response is returned:

```
+--------------------+--------+---------+------------+------+----------------+-------------+----------
-------+--------------+-----------+----------------+----------------------+-------------+------------
+-----------------+----------+----------------+----------+
| NAME               | ENGINE | VERSION | ROW_FORMAT | ROWS | AVG_ROW_LENGTH | DATA_LENGTH | MAX_DATA_
LENGTH | INDEX_LENGTH | DATA_FREE | AUTO_INCREMENT | CREATE_TIME          | UPDATE_TIME | CHECK_TIME
| COLLATION       | CHECKSUM | CREATE_OPTIONS | COMMENT |
+--------------------+--------+---------+------------+------+----------------+-------------+----------
-------+--------------+-----------+----------------+----------------------+-------------+------------
+-----------------+----------+----------------+----------+
| multi_db_multi_tbl | InnoDB |      10 | Compact    |    2 |          16384 |       16384 |
0 |        16384 |         0 |         100000 | 2017-03-27 17:43:57.0 | NULL        | NULL        | utf
8_general_ci | NULL     |                |         |
+--------------------+--------+---------+------------+------+----------------+-------------+----------
-------+--------------+-----------+----------------+----------------------+-------------+------------
+-----------------+----------+----------------+----------+
1 row in set (0.03 sec)
```

You can use the SHOW TABLE STATUS statement and the SCAN hint to view the data volume of each physical table shard.

Execute the `/! TDDL:SCAN='multi_db_multi_tbl'*/SHOW TABLE STATUS LIKE 'multi_db_multi_tbl';` statement. The following response is returned:

```
+---------------------+--------+---------+------------+------+----------------+-------------+--------
--------+--------------+-----------+----------------+----------------------+-------------+------------
+-----------------+----------+----------------+---------+--------------+
| Name                | Engine | Version | Row_format | Rows | Avg_row_length | Data_length | Max_dat
a_length | Index_length | Data_free | Auto_increment | Create_time          | Update_time | Check_time
| Collation       | Checksum | Create_options | Comment | Block_format |
+---------------------+--------+---------+------------+------+----------------+-------------+--------
--------+--------------+-----------+----------------+----------------------+-------------+------------
+-----------------+----------+----------------+---------+--------------+
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    1 |          16384 |       16384 |
0 |        16384 |         0 |              2 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |              0 |       16384 |
0 |        16384 |         0 |              1 | 2017-03-27 17:43:57 | NULL        | NULL        | utf8_
general_ci |     NULL |                |         | Original     |
```

```
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    0 |               0 |        16384 |
0 |         16384 |          0 |              1 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |               0 |        16384 |
0 |         16384 |          0 |              1 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    0 |               0 |        16384 |
0 |         16384 |          0 |              1 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |               0 |        16384 |
0 |         16384 |          0 |              1 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    0 |               0 |        16384 |
0 |         16384 |          0 |              1 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
| multi_db_multi_tbl_1 | InnoDB |      10 | Compact    |    0 |               0 |        16384 |
0 |         16384 |          0 |              1 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
| multi_db_multi_tbl_0 | InnoDB |      10 | Compact    |    1 |           16384 |        16384 |
0 |         16384 |          0 |              3 | 2017-03-27 17:43:57 | NULL          | NULL         | utf8_
general_ci |      NULL |                  |          | Original      |
+----------------------+--------+---------+------------+------+-----------------+-------------+--------
---------+--------------+-----------+----------------+---------------------+-------------+------------
+-----------------+----------+----------------+---------+--------------+
16 rows in set (0.04 sec)
```

12.1.8.5. Statements for querying statistics

This topic describes the statements that you can execute to query statistics.

SHOW [FULL] STATS

You can execute this SQL statement to query the overall statistics. The statistics are instantaneous values. Take note of this point: The returned results of `SHOW FULL STATUS` vary based on the instance version.

The following list describes the important columns:

- QPS: the queries per second (QPS) sent from an application to a instance. In most cases, the QPS is named logical QPS.
- RDS_QPS: the QPS sent from a instance to an ApsaraDB RDS for MySQL instance. In most cases, the QPS is named physical QPS.
- ERROR_PER_SECOND: the total number of errors that occur per second. These errors include SQL syntax errors, primary key conflicts, system errors, and connectivity errors.
- VIOLATION_PER_SECOND: the number of primary key conflicts or unique key conflicts per second.
- MERGE_QUERY_PER_SECCOND: the number of queries on tables per second. Sharding is enabled for a database instance.
- ACTIVE_CONNECTIONS: the number of active connections.
- CONNECTION_CREATE_PER_SECCOND: the number of connections that are created per second.
- RT(MS): the response time (RT) for an SQL query that is sent from an application to a instance. In most cases, the RT is named logical RT.
- RDS_RT(MS): the RT for an SQL query that is sent from a instance to an ApsaraDB RDS for MySQL instance. In most cases, the RT is named physical RT.
- NET_IN(KB/S): the inbound traffic of a instance per second.
- NET_OUT(KB/S): the outbound traffic of a instance per second.
- THREAD_RUNNING: the number of the running threads.

- HINT_USED_PER_SECOND: the number of SQL queries that contain hints per second.
- HINT_USED_COUNT: the total number of SQL queries that contain hints since a database instance is started.
- AGGREGATE_QUERY_PER_SECCOND: the number of aggregate queries per second.
- AGGREGATE_QUERY_COUNT: the total number of aggregate queries. This column indicates the accumulative historical data.
- TEMP_TABLE_CREATE_PER_SECCOND: the number of temporary tables that are created per second.
- TEMP_TABLE_CREATE_COUNT: the total number of temporary tables that are created since a database instance is started.
- MULTI_DB_JOIN_PER_SECCOND: the number of cross-database JOIN queries that are processed by a database instance per second.
- MULTI_DB_JOIN_COUNT: the total number of cross-database JOIN queries that are processed since a database instance is started.

Execute the `SHOW STATS;` statement. The following response is returned:

```
+------+---------+----------+------------------+-----------------+------------------------+---------
-----------+--------+------------+--------------+---------------+----------------+
| QPS  | RDS_QPS | SLOW_QPS | PHYSICAL_SLOW_QPS | ERROR_PER_SECOND | MERGE_QUERY_PER_SECOND | ACTIVE_C
ONNECTIONS | RT(MS) | RDS_RT(MS) | NET_IN(KB/S) | NET_OUT(KB/S) | THREAD_RUNNING |
+------+---------+----------+------------------+-----------------+------------------------+---------
-----------+--------+------------+--------------+---------------+----------------+
| 1.77 |    1.68 |     0.03 |              0.03 |             0.02 |                   0.00 |
7 | 157.13 |      51.14 |       134.49 |          1.48 |              1 |
+------+---------+----------+------------------+-----------------+------------------------+---------
-----------+--------+------------+--------------+---------------+----------------+
1 row in set (0.01 sec)
```

Execute the `SHOW [FULL] STATS;` statement. The following response is returned:

```
+------+---------+----------+------------------+-----------------+----------------------+-----------
-------------+--------------------+-------------------------------+--------+------------+--------------
+---------------+----------------+----------------------+-----------------+--------------------------
-+-----------------------+-------------------------------+-------------------------+------------------
------+----------------------+-------+---------+-------------+------------+
| QPS  | RDS_QPS | SLOW_QPS | PHYSICAL_SLOW_QPS | ERROR_PER_SECOND | VIOLATION_PER_SECOND | MERGE_QUER
Y_PER_SECOND | ACTIVE_CONNECTIONS | CONNECTION_CREATE_PER_SECOND | RT(MS) | RDS_RT(MS) | NET_IN(KB/S)
| NET_OUT(KB/S) | THREAD_RUNNING | HINT_USED_PER_SECOND | HINT_USED_COUNT | AGGREGATE_QUERY_PER_SECOND
| AGGREGATE_QUERY_COUNT | TEMP_TABLE_CREATE_PER_SECOND | TEMP_TABLE_CREATE_COUNT | MULTI_DB_JOIN_PER_S
ECOND | MULTI_DB_JOIN_COUNT | CPU   | FREEMEM | FULLGCCOUNT | FULLGCTIME |
+------+---------+----------+------------------+-----------------+----------------------+-----------
-------------+--------------------+-------------------------------+--------+------------+--------------
+---------------+----------------+----------------------+-----------------+--------------------------
-+-----------------------+-------------------------------+-------------------------+------------------
------+----------------------+-------+---------+-------------+------------+
| 1.63 |    1.68 |     0.03 |              0.03 |             0.02 |                 0.00 |
0.00 |                   6 |                          0.01 | 157.13 |      51.14 |       134.33 |
1.21 |              1 |                 0.00 |              54 |                         0.00 |
663 |                          0.00 |                     512 |                      0.00 |
516 | 0.09% |   6.96% |       76446 |   21326906 |
+------+---------+----------+------------------+-----------------+----------------------+-----------
-------------+--------------------+-------------------------------+--------+------------+--------------
+---------------+----------------+----------------------+-----------------+--------------------------
-+-----------------------+-------------------------------+-------------------------+------------------
------+----------------------+-------+---------+-------------+------------+
1 row in set (0.01 sec)
```

SHOW DB STATUS

You can execute this SQL statement to view the storage and performance information about a physical database in real time. The storage information is obtained from an ApsaraDB RDS for MySQL system table. Therefore, the returned information may be different from the actual storage information.

The following list describes the important columns:

- NAME: the internal tag of a database. Each tag identifies a database. The tag is different from the name of the database.
- CONNECTION_STRING: the information about a connection from a database instance to a database shard.
- PHYSICAL_DB: the name of a database shard. The `TOTAL` row indicates the total storage of all the database shards in a database.
- SIZE_IN_MB: the used storage in a database shard. Unit: MB.
- RATIO: the ratio of the data volume of a database shard to the total data volume of the database.
- THREAD_RUNNING: the number of threads that are running on a physical database. The meaning of this parameter is the same as that of the THREAD_RUNNING parameter in the returned results of the `SHOW GLOBAL STATUS` statement in MySQL. For more information, see MySQL documentation.

Execute the `SHOW DB STATUS;` statement. The following response is returned:

```
+------+---------------------------+-------------------+-------------------+------------+--------+---
-------------+
| ID   | NAME                      | CONNECTION_STRING  | PHYSICAL_DB       | SIZE_IN_MB | RATIO  | TH
READ_RUNNING |
+------+---------------------------+-------------------+-------------------+------------+--------+---
-------------+
|    1 | drds_db_1516187088365daui | 100.100.64.1:59077 | TOTAL             |  13.109375 | 100%   | 3
|
|    2 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0000 |   1.578125 | 12.04% |
|
|    3 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0001 |     1.4375 | 10.97% |
|
|    4 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0002 |     1.4375 | 10.97% |
|
|    5 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0003 |     1.4375 | 10.97% |
|
|    6 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0004 |   1.734375 | 13.23% |
|
|    7 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0005 |   1.734375 | 13.23% |
|
|    8 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0006 |   2.015625 | 15.38% |
|
|    9 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0007 |   1.734375 | 13.23% |
|
+------+---------------------------+-------------------+-------------------+------------+--------+---
-------------+
```

SHOW FULL DB STATUS [LIKE {tablename}]

You can execute this SQL statement to view the storage and performance information about a table in a physical database in real time. The storage information is obtained from an ApsaraDB RDS for MySQL system table. Therefore, the returned information may be different from the actual storage information.

The following list describes the important columns:

- NAME: the internal tag of a database. Each tag identifies a database. The tag is different from the name of the database.
- CONNECTION_STRING: the information about a connection from a database instance to a database shard.

- PHYSICAL_DB: the name of a database shard. If the LIKE keyword is used in a statement, the TOTAL row indicates the total storage of the database shard. If the LIKE keyword is not used in a statement, the TOTAL row indicates the total storage of all the database shards.
- PHYSICAL_TABLE: the name of a table shard in a database shard. If the LIKE keyword is used in a statement, the TOTAL row indicates the total storage of the table shard. If the LIKE keyword is not used in a statement, the TOTAL row indicates the total storage of all the table shards.
- SIZE_IN_MB: the used storage in a database shard. Unit: MB.
- RATIO: the ratio of the data volume of a table shard to the total data volume of all the returned table shards.
- THREAD_RUNNING: the number of threads that are running on a physical database. The meaning of this parameter is the same as that of the THREAD_RUNNING parameter in the returned results of the `SHOW GLOBAL STATUS` statement in MySQL. For more information, see MySQL documentation.

Execute the `SHOW FULL DB STATUS LIKE hash_tb;` statement. The following response is returned:

```
+------+---------------------------+--------------------+------------------+----------------+--------
----+--------+----------------+
| ID   | NAME                      | CONNECTION_STRING  | PHYSICAL_DB      | PHYSICAL_TABLE | SIZE_IN
_MB | RATIO  | THREAD_RUNNING |
+------+---------------------------+--------------------+------------------+----------------+--------
----+--------+----------------+
|    1 | drds_db_1516187088365daui | 100.100.64.1:59077 | TOTAL            |                |     19.
875 | 100%   | 3              |
|    2 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0000 | TOTAL          |    3.03
125 | 15.25% |                |
|    3 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0000 | hash_tb_00     |   1.515
625 | 7.63%  |                |
|    4 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0000 | hash_tb_01     |   1.515
625 | 7.63%  |                |
|    5 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0001 | TOTAL          |
2.0 | 10.06% |                |
|    6 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0001 | hash_tb_02     |   1.515
625 | 7.63%  |                |
|    7 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0001 | hash_tb_03     |   0.484
375 | 2.44%  |                |
|    8 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0002 | TOTAL          |    3.03
125 | 15.25% |                |
|    9 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0002 | hash_tb_04     |   1.515
625 | 7.63%  |                |
|   10 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0002 | hash_tb_05     |   1.515
625 | 7.63%  |                |
|   11 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0003 | TOTAL          |   1.953
125 | 9.83%  |                |
|   12 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0003 | hash_tb_06     |   1.515
625 | 7.63%  |                |
|   13 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0003 | hash_tb_07     |     0.4
375 | 2.2%   |                |
|   14 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0004 | TOTAL          |    3.03
125 | 15.25% |                |
|   15 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0004 | hash_tb_08     |   1.515
625 | 7.63%  |                |
|   16 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0004 | hash_tb_09     |   1.515
625 | 7.63%  |                |
|   17 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0005 | TOTAL          |   1.921
875 | 9.67%  |                |
|   18 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0005 | hash_tb_11     |   1.515
625 | 7.63%  |                |
|   19 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0005 | hash_tb_10     |    0.40
625 | 2.04%  |                |
|   20 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0006 | TOTAL          |    3.03
125 | 15.25% |                |
|   21 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0006 | hash_tb_12     |   1.515
625 | 7.63%  |                |
|   22 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0006 | hash_tb_13     |   1.515
625 | 7.63%  |                |
|   23 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0007 | TOTAL          |      1.
875 | 9.43%  |                |
|   24 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0007 | hash_tb_14     |   1.515
625 | 7.63%  |                |
|   25 | drds_db_1516187088365daui | 100.100.64.1:59077 | drds_db_xzip_0007 | hash_tb_15     |   0.359
375 | 1.81%  |                |
+------+---------------------------+--------------------+------------------+----------------+--------
----+--------+----------------+
```

12.1.8.6. SHOW PROCESSLIST and KILL

This topic describes the SHOW PROCESSLIST and KILL statements.

> **Note**
> - If the version of is 5.1.28-1408022 or later, supports the SHOW PROCESSLIST and KILL statements for logical and physical connections. For more information, see this topic.
> - If the version of is earlier than 5.1.28-1408022, supports the SHOW PROCESSLIST and KILL statements for only physical connections. For more information, see SHOW PROCESSLIST and KILL commands in earlier versions.

SHOW PROCESSLIST

In a instance, you can execute the `SHOW PROCESSLIST` statement to view information including connections to the instance and SQL statements that are being executed in the instance.

Syntax

```
SHOW [FULL] PROCESSLIST
```

Examples

Execute the `SHOW PROCESSLIST;` statement. The following response is returned:

```
     ID: 1971050
   USER: admin
   HOST: 111.111.111.111:4303
     DB: drds_test
COMMAND: Query
   TIME: 0
  STATE:
   INFO: show processlist
1 row in set (0.01 sec)
```

The meanings of the following fields in the result set are described in detail:

- ID: the ID of the connection. The value is a number of the Long type.
- USER: the name of the user who establishes the connection.
- HOST: the IP address and port number of the host that establishes the connection.
- DB: the name of the database to which the connection is established.
- COMMAND: the usage state of the connection. This field can be set to the following values:
  - Query: indicates that the current connection is executing an SQL statement.
  - Sleep: indicates that the current connection is idle.
- TIME: the duration when the connection is in the current state:
  - When the value of COMMAND is Query, this field indicates how long the SQL statement has been being executed over the connection.
  - When the value of COMMAND is Sleep, this field indicates how long the connection has been in the idle state.
- STATE: Currently, this field is meaningless and is constantly empty.
- INFO:

- When the value of COMMAND is Query, this field indicates the content of the SQL statement that is being executed over the connection. If the FULL parameter is not specified, a maximum of the first 30 characters of the SQL statement are returned. If the FULL parameter is specified, a maximum of the first 1,000 characters of the SQL statement are returned.
- When the value of COMMAND is other values, this field is meaningless and left empty.

SHOW PHYSICAL_PROCESSLIST

In a instance, you can execute the SHOW PHYSICAL_PROCESSLIST statement to view information about all the SQL statements that are being executed on underlying ApsaraDB RDS for MySQL instances.

Syntax

```
SHOW [FULL] PHYSICAL_PROCESSLIST
```

When an SQL statement is excessively long, the responses of the `SHOW PHYSICAL_PROCESSLIST` statement may be truncated. In this case, you can execute the `SHOW FULL PHYSICAL_PROCESSLIST` statement to obtain the complete SQL statement.

The meaning of each column in the responses is equivalent to that in the responses of the `SHOW PROCESSLIST` statement. For more information, see SHOW PROCESSLIST Syntax.

> **Note** Different from MySQL, the instance returns a string instead of a number in the ID column of a physical connection.

Execute the `SHOW PHYSICAL_PROCESSLIST;` statement. The following response is returned:

```
*************************** 1. row ***************************
           ID: 0-0-521414
         USER: tddl5
           DB: tddl5_00
      COMMAND: Query
         TIME: 0
        STATE: init
         INFO: show processlist
*************************** 2. row ***************************
           ID: 0-0-521570
         USER: tddl5
           DB: tddl5_00
      COMMAND: Query
         TIME: 0
        STATE: User sleep
         INFO: /*DRDS /88.88.88.88/b67a0e4d8800000/ */ select sleep(1000)
2 rows in set (0.01 sec)
```

KILL

You can execute the KILL statement to terminate an SQL statement that is being executed.

The instance connects to an ApsaraDB RDS for MySQL instance by using the username created by the instance on the ApsaraDB RDS for MySQL instance. Therefore, if you directly connect to the ApsaraDB RDS for MySQL instance, you are not authorized to execute the KILL statement to terminate a request initiated by the instance.

To terminate an SQL statement that is being executed on the instance, you must use tools to connect to the instance. You can use tools such as the MySQL command line and . Then, execute the KILL statement on the instance.

Syntax

```
KILL PROCESS_ID | 'PHYSICAL_PROCESS_ID' | 'ALL'
```

The KILL statement can be used in the following three ways:

- Execute the `KILL PROCESS_ID` statement to terminate a specified logical SQL statement.

  The PROCESS_ID parameter is obtained from the ID column in the responses of the `SHOW [FULL] PROCESSLIST` statement.

  If you execute the `KILL PROCESS_ID` statement in the instance, it will terminate both logical and physical SQL statements that are being executed over this connection. In addition, this connection will be disconnected.

  The instance does not support the `KILL QUERY` statement.

- Execute the `KILL 'PHYSICAL_PROCESS_ID'` statement to terminate a specified physical SQL statement.

  The PHYSICAL_PROCESS_ID parameter is obtained from the ID column in the responses of the `SHOW PHYSICAL_PROCESS_ID` statement.

  > **Note** The PHYSICAL_PROCESS_ID column is a string instead of a number. Therefore, the PHYSICAL_PROCESS_ID parameter must be enclosed in single quotation marks (') in the KILL statement.

  Examples

  Execute the `KILL '0-0-521570';` statement. The following response is returned:

  ```
  Query OK, 0 rows affected (0.01 sec)
  ```

- Execute the `KILL 'ALL'` statement to terminate all the physical SQL statements that are executed by the instance in the current logical database.

  When the underlying ApsaraDB RDS for MySQL instance is overloaded due to several SQL statements, you can execute the `KILL 'ALL'` statement to terminate all the physical SQL statements that are being executed in the current logical database.

  All the physical SQL statements indicated by PROCESS that meet the following conditions can be terminated by the `KILL 'ALL'` statement:

  - The value of the User parameter for the physical SQL statement indicated by PROCESS is a username created by the instance in the ApsaraDB RDS for MySQL instance.
  - The physical SQL statement indicated by PROCESS is executing a query, which means that the value of COMMAND is Query.

12.1.8.7. SHOW PROCESSLIST and KILL statements in earlier versions

This topic describes the SHOW PROCESSLIST and KILL statements in earlier versions.

> **Note**
>
> - If the version of is 5.1.28-1408022 or later, supports the SHOW PROCESSLIST and KILL statements for logical and physical connections. For more information, see SHOW PROCESSLIST and KILL commands.
> - If the version of is earlier than 5.1.28-1408022, supports the SHOW PROCESSLIST and KILL statements for only physical connections. For more information, see this topic.

SHOW PROCESSLIST

In a instance, you can execute the `SHOW PROCESSLIST` statement to view information about all the SQL statements that are being executed on the underlying ApsaraDB RDS for MySQL instances.

Syntax

```
SHOW [FULL] PROCESSLIST
```

When an SQL statement is excessively long, the responses of the `SHOW PROCESSLIST` statement may be truncated. In this case, you can execute the `SHOW FULL PROCESSLIST` statement to obtain the complete SQL statement.

The meaning of each column in the responses is equivalent to that in the responses of the `SHOW PROCESSLIST` statement. For more information, see SHOW PROCESSLIST Syntax.

```
*************************** 1. row ***************************
           ID: 0-0-521414
         USER: tddl5
           DB: tddl5_00
      COMMAND: Query
         TIME: 0
        STATE: init
         INFO: show processlist
    ROWS_SENT: NULL
ROWS_EXAMINED: NULL
    ROWS_READ: NULL
*************************** 2. row ***************************
           ID: 0-0-521570
         USER: tddl5
           DB: tddl5_00
      COMMAND: Query
         TIME: 0
        STATE: User sleep
         INFO: /*DRDS /88.88.88.88/b67a0e4d8800000/ */ select sleep(1000)
    ROWS_SENT: NULL
ROWS_EXAMINED: NULL
    ROWS_READ: NULL
2 rows in set (0.01 sec)
```

KILL

You can execute the KILL statement to terminate an SQL statement that is being executed.

The instance connects to an ApsaraDB RDS for MySQL instance by using the username created by the instance on the ApsaraDB RDS for MySQL instance. Therefore, if you directly connect to the ApsaraDB RDS for MySQL instance, you are not authorized to execute the KILL statement to terminate a request initiated by the instance.

To terminate an SQL statement that is being executed on the instance, you must use tools to connect to the instance. You can use tools such as the MySQL command line and . Then, execute the KILL statement on the instance.

Syntax

```
KILL 'PROCESS_ID' | 'ALL'
```

The KILL statement can be used in the following two ways:

- Execute the `KILL 'PROCESS_ID'` statement to terminate a specified SQL statement.

  The PROCESS_ID parameter is obtained from the ID column in the responses of the `SHOW PROCESSLIST` statement.

> **Note** Different from MySQL, the instance returns a string instead of a number in the ID column. Therefore, the PROCESS_ID parameter must be enclosed in single quotation marks (') in the KILL statement.

Examples

Execute the `KILL '0-0-521570';` statement. The following response is returned:

```
Query OK, 0 rows affected (0.01 sec)
```

- Execute the `KILL 'ALL'` statement to terminate all the SQL statements executed by the instance in the current logical database.

  When the underlying ApsaraDB RDS for MySQL instance is overloaded due to several SQL statements, you can execute the `KILL 'ALL'` statement to terminate all the SQL statements that are being executed in the current logical database.

  All SQL statements indicated by PROCESS that meet the following conditions can be terminated by the `KILL 'ALL'` statement:

  - The value of the User parameter for the physical SQL statement indicated by PROCESS is a username created by the instance in the ApsaraDB RDS for MySQL instance.
  - The physical SQL statement indicated by PROCESS is executing a query, which means that the value of COMMAND is Query.

  instances in earlier versions do not support the `KILL 'ALL'` statement. An error will be reported if this statement is being executed in these instances. To resolve this problem, you can upgrade the version of the instance.

12.1.9. Custom hints

> **Note** This topic is applicable to 5.3 and later. For earlier versions, see PolarDB-X 5.2 hints.

12.1.9.1. Introduction to hints

As a supplement to the SQL syntax, hints play a critical role in relational databases. They allow you to modify execution plans of SQL statements by using the relevant syntax. This way, you can optimize the SQL statements. also provides special hint syntax.

For example, assume that you know the data is stored in table shards in specific database shards. If you need to route an SQL statement directly to the database shards for execution, you can use custom hints provided by .

```
SELECT /*+TDDL:node('node_name')*/ * FROM table_name;
```

In the preceding SQL statement, the part between `/*` and `*/` is a hint. This means that `+TDDL:node('node_name')` is a hint. The hint specifies the ApsaraDB RDS for MySQL database shard where the SQL statement will be executed.

> **Note**
>
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/` .
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/` . In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Syntax of PolarDB-X hints

Basic syntax

```
/*+TDDL: hint_command [hint_command ...] */
/! +TDDL: hint_command [hint_command ...] */
```

hints are based on the MySQL comment syntax. A hint is located between `/*` and `*/` or between `/!` and `*/`, and must begin with `+TDDL:`. The `hint_command` parameter indicates a hint command related to a specific operation. Multiple values of `hint_command` are separated with spaces.

Examples

```
# Query the names of physical table shards in each database shard.
/*+TDDL:scan()*/SHOW TABLES;
# Route the query to database shard 0000 of a read-only ApsaraDB RDS for MySQL instance.
/*+TDDL:node(0) slave()*/SELECT * FROM t1;
```

In the example, `/*+TDDL:scan()*/` and `/*+TDDL:node(0) slave()*/` are hints that begin with `+TDDL:`. The `scan()`, `node(0)`, and `slave()` parameters are hint commands. Hint commands are separated with spaces.

- Use a hint in an SQL statement:

  allows you to use hints in DML, DDL, and data access language (DAL) statements. The following section describes the syntax.

  - For all statements that support hints, you can specify a hint at the beginning of the statements, as shown in the following example:

    ```
    /*+TDDL: ... */ SELECT ...
    /*+TDDL: ... */ INSERT ...
    /*+TDDL: ... */ REPLACE ...
    /*+TDDL: ... */ UPDATE ...
    /*+TDDL: ... */ DELETE ...
    /*+TDDL: ... */ CREATE TABLE ...
    /*+TDDL: ... */ ALTER TABLE ...
    /*+TDDL: ... */ DROP TABLE ...
    /*+TDDL: ... */ SHOW ...
    ...
    ```

  - For DML statements, you can specify a hint behind the first keyword of the statements, as shown in the following example:

    ```
    SELECT /*+TDDL: ... */  ...
    INSERT /*+TDDL: ... */  ...
    REPLACE /*+TDDL: ... */  ...
    UPDATE /*+TDDL: ... */  ...
    DELETE /*+TDDL: ... */  ...
    ...
    ```

    > **Note** Different hints support different statements. For more information, see the hints in the following topics.

- Use multiple hint commands in an SQL statement:

  allows you to use multiple hint commands in a hint in an SQL statement.

  ```
  SELECT /*+TDDL:node(0) slave()*/ ... ;
  ```

  has the following limits on the use of multiple hint commands in a hint in an SQL statement:

```
# A single SQL statement cannot contain multiple hints.
SELECT /*+TDDL:node(0)*/ /*+TDDL:slave()*/ ... ;
# A hint cannot contain duplicate hint commands.
SELECT /*+TDDL:node(0) node(1)*/ ... ;
```

Classification of PolarDB-X hints

hints are classified into the following categories based on operation types:

- Read/write splitting
- Specify a timeout period for an SQL statement
- Specify a database shard to run an SQL statement
- Scan all or some of database shards and table shards

12.1.9.2. Read/write splitting

This topic describes the read/write splitting feature provided by PolarDB-X.

provides transparent read/write splitting at the application layer. Data synchronization between primary and read-only ApsaraDB RDS for MySQL instances has a latency of several milliseconds. If you need to read the changed data immediately after data in the primary ApsaraDB RDS for MySQL instance is changed, you must ensure that the SQL statement for reading data is routed to the primary ApsaraDB RDS for MySQL instance. To meet this demand, provides custom hints for read/write splitting. These custom hints allow you to route SQL statements to a specified primary or read-only ApsaraDB RDS for MySQL instance.

> **Note** This topic is applicable to 5.3 and later. For more information about custom hints in earlier versions, see Read/write splitting.

Syntax

```
/*+TDDL:
    master()
    | slave()
*/
```

The custom hints allow you to specify whether to execute an SQL statement on a primary or read-only ApsaraDB RDS for MySQL instance. When you use the custom hint `/*+TDDL:slave()*/`, if a primary ApsaraDB RDS for MySQL instance has multiple read-only ApsaraDB RDS for MySQL instances, the instance randomly selects a read-only ApsaraDB RDS for MySQL instance based on its weight, to execute the SQL statement.

> **Note**
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

- Execute an SQL statement on a specified primary ApsaraDB RDS for MySQL instance:

```
SELECT /*+TDDL:master()*/ * FROM table_name;
```

After the custom hint `/*+TDDL:master()*/` is added behind the first keyword in the SQL statement, this SQL statement is routed to the primary ApsaraDB RDS for MySQL instance.

- Execute an SQL statement on a specified read-only ApsaraDB RDS for MySQL instance:

```
SELECT /*+TDDL:slave()*/ * FROM table_name;
```

After the custom hint `/*+TDDL:slave()*/` is added behind the first keyword in the SQL statement, this SQL statement is randomly routed to a read-only ApsaraDB RDS for MySQL instance based on its allocated weight.

Considerations

- The custom hints for read/write splitting are only applicable to read SQL statements for non-transactional data. SQL statements for transactional data and write SQL statements are still routed to the primary ApsaraDB RDS for MySQL instance.
- When you use the `/*+TDDL:slave()*/` hint, the instance routes the SQL statement randomly to a read-only ApsaraDB RDS for MySQL instance based on the allocated weight. If no read-only ApsaraDB RDS for MySQL instance is available, no error is reported. Instead, the primary ApsaraDB RDS for MySQL instance is selected to execute the SQL statement.

12.1.9.3. Specify a timeout period for an SQL statement

In , if the amount of time consumed to execute an SQL statement on a node and an ApsaraDB RDS for MySQL instance exceeds the default duration of 900 seconds, the execution times out. You can adjust the timeout period. The duration during which a slow SQL statement is executed may exceed 900 seconds. For this type of slow SQL statement, provides a custom hint to help you adjust the timeout period. You can use this custom hint to adjust the duration in which a SQL statement is executed.

> **Note** This topic is applicable to 5.3 and later. For other versions, see Specify a timeout period for an SQL statement.

Syntax

You can use the following syntax of a custom hint to specify a timeout period for an SQL statement:

```
/*+TDDL:SOCKET_TIMEOUT(time)*/
```

The value specified by `SOCKET_TIMEOUT` is measured in milliseconds. You can use this custom hint to change the timeout period for SQL statements based on your business requirements.

> **Note**
> - You can specify custom hints in the `/*+TDDL:hint_command*/` format or in the `/!+TDDL:hint_command*/` format.
> - In the official MySQL CLI, if you execute SQL statements that contain customer hints in the `/*+TDDL:hint_command*/` format, add the `-c` parameter in the `mysql -u*** -p*** -h***` command that is run to log on to the client. Otherwise, the client deletes custom hints before the client sends the SQL statements to the server for execution. This is because the hints are in the format of a MySQL comment. As a result, the customer hints do not take effect. For more information, see MySQL client options.

Examples

Set the timeout period for a SQL statement to 40 seconds.

```
/*+TDDL:SOCKET_TIMEOUT(40000)*/SELECT * FROM t_item;
```

> **Note** A longer timeout period causes database resources to be consumed for a longer period of time. If a large number of SQL statements are executed over a long period of time within the same interval, a large number of database resources may be consumed. As a result, DRDS database services cannot be provided as expected. To resolve the issue, we recommend that you optimize SQL statements that take a long time to execute if possible.

12.1.9.4. Execute an SQL statement on a specified database shard

When you execute SQL statements on a instance, you may find that some SQL statements are not supported by the instance. You can use the `NODE` hint provided by to route the SQL statements to one or more database shards. If you need to query the data in a specified database shard or the data in a specified table shard of a known database shard, you can use the `NODE` hint to directly route the SQL statement to the database shard.

> **Note** This topic is applicable to 5.3 and later. For more information about custom hints in earlier versions, see Specify a database shard to run an SQL statement.

Syntax

The `NODE` hint allows you to specify a database shard to execute an SQL statement by using its shard name. A shard name uniquely identifies a database shard in a instance. You can execute the SHOW NODE statement to obtain the shard name.

Execute an SQL statement on a database shard by specifying the shard name

This custom hint allows you to specify one or more database shards to execute an SQL statement.

> **Note** Assume that you execute the INSERT statement in a table that uses a sequence. The sequence will not take effect if you use the NODE hint in the INSERT statement. For more information, see Limits and precautions for sequences.

- Execute an SQL statement on a specified database shard:

  ```
  /*+TDDL:node('node_name')*/
  ```

  `node_name` indicates the shard name. This custom hint provided by allows you to route the SQL statement to the database shard specified by `node_name`.

- Execute an SQL statement on multiple database shards:

  ```
  /*+TDDL:node('node_name'[,'node_name1','node_name2'])*/
  ```

  You can specify multiple shard names in the parameters and route the SQL statement to multiple database shards. Separate multiple shard names with commas (,).

Note

- When this custom hint is used, the instance directly routes the SQL statement to the specified database shards. Therefore, the specified shard names in the SQL statement must correspond to existing database shards.
- The `NODE` hint can be used in DML, DDL, and DAL statements.
- hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/` .
- In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/` . In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

The following example shows the responses for executing the `SHOW NODE` statement on a logical database that is named `drds_test` .

```
*************************** 1. row ******************
                ID: 0
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS
  MASTER_READ_COUNT: 212
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 2. row ******************
                ID: 1
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0001_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 3. row ******************
                ID: 2
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0002_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 4. row ******************
                ID: 3
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0003_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 5. row ******************
                ID: 4
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0004_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 6. row ******************
                ID: 5
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0005_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 7. row ******************
                ID: 6
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0006_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 8. row ******************
                ID: 7
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0007_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
8 rows in set (0.02 sec)
```

Each database shard has the `NAME` attribute, which indicates the shard name corresponding to the database shard. Each shard name uniquely corresponds to one database shard name. For example, the shard name `DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0003_RDS` corresponds to the database shard name `drds_test_vtla_0003`. Therefore, after you obtain the shard name, you can use the custom hint provided by to specify the database shard on which you want to execute an SQL statement.

- Execute an SQL statement on database shard 0:

```
SELECT /*TDDL:node('DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS')*/ * FROM table_name;
```

- Execute an SQL statement on multiple database shards:

```
SELECT /*TDDL:node('DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS','DRDS_TEST_1473471355140LRP
RDRDS_TEST_VTLA_0006_RDS')*/ * FROM table_name;
```

  This SQL statement will be executed on the database shards specified by the shard names `DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS` and `DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0006_RDS`.

- View the execution plan of an SQL statement on database shard 0:

```
/*TDDL:node('DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS')*/ EXPLAIN SELECT * FROM table_nam
e;
```

12.1.9.5. Scan some or all of the database shards and table shards

You can route an SQL statement to one or more database shards. You can also route an SQL statement to some or all of the database shards and table shards by using the `SCAN` hint provided by . You can use the `SCAN` hint to route an SQL statement to all database shards at a time. For example, you can view all the table shards in a specified database shard or view the data volume of each physical table shard that corresponds to a specified logical table.

> **Note** This topic is applicable to 5.3 and later. For more information about custom hints in earlier versions, see Scan all database shards and table shards.

The SCAN hint allows you to execute an SQL statement by using the following methods:

- Execute an SQL statement on all table shards in all database shards.
- Execute an SQL statement on all table shards in specified database shards.
- Execute an SQL statement on specified table shards in specified database shards. The names of the physical database shards and table shards are calculated based on given conditions.
- Execute an SQL statement on table shards in database shards by explicitly specifying the names of the physical table shards.

The `SCAN` hint can be used in data manipulation language (DML) statements, DDL statements, and some data access language (DAL) statements.

> **Note**
>
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Syntax

```
SCAN HINT
# Route an SQL statement to all table shards in all database shards.
SCAN()
# Route an SQL statement to all table shards in specified database shards.
SCAN(NODE="node_list")               # Specify the database shards.
# Route an SQL statement to specified table shards in specified database shards. The names of the phys
ical database shards and table shards are calculated based on given conditions.
SCAN(
  [TABLE=]"table_name_list"          # Specify the name of the logical table.
  , CONDITION="condition_string"     # Calculate the names of the physical database shards and table s
hards based on the values of the TABLE and CONDITION parameters.
  [, NODE="node_list"] )             # Filter the results obtained based on the value of the CONDITION
parameter to retain only the results about the specified physical database shards.
# Route an SQL statement to table shards in database shards by explicitly specifying the names of the
physical table shards.
SCAN(
  [TABLE=]"table_name_list"          # Specify the name of the logical table.
  , REAL_TABLE=("table_name_list")   # Specify the name of the physical table shards. The same name of
physical table shards is applied to all physical database shards.
  [, NODE="node_list"] )             # Filter the results obtained based on the value of the CONDITION
parameter to retain only the results about the specified physical database shards.
# Specify the names of physical table shards or logical tables.
table_name_list:
    table_name [, table_name]...
# Specify physical database shards by using GROUP_KEY and GROUP_INDEX. You can obtain their values by
executing the `SHOW NODE` statement.
node_list:
    {group_key | group_index} [, {group_key | group_index}]...
# Execute an SQL WHERE statement. You must specify conditions for each table, such as t1.id = 2 and t2
.id = 2.
condition_string:
    where_condition
```

Examples

- Execute the following SQL statement on all table shards in all database shards:

  ```
  SELECT /*+TDDL:scan()*/ COUNT(1) FROM t1
  ```

  After this statement is executed, the SQL statement is routed to all the physical table shards corresponding to logical table `t1`, and the result sets are merged and returned.

- Execute the following SQL statement on all table shards in specified database shards:

  ```
  SELECT /*+TDDL:scan(node='0,1,2')*/ COUNT(1) FROM t1
  ```

  After this statement is executed, all the physical table shards corresponding to logical table `t1` in database shards 0000, 0001, and 0002 are calculated. Then, the SQL statement is routed to the physical table shards, and the result sets are merged and returned.

- Execute the following SQL statement on specified table shards based on conditions:

  ```
  SELECT /*+TDDL:scan('t1', condition='t1.id = 2')*/ COUNT(1) FROM t1
  ```

  After this statement is executed, all the physical table shards that correspond to logical table `t1` and meet the conditions are calculated. Then, the SQL statement is routed to the specified physical table shards, and the result sets are merged and returned.

- Execute the following JOIN statement on the specified table shards based on conditions:

```
SELECT /*+TDDL:scan('t1, t2', condition='t1.id = 2 and t2.id = 2')*/ * FROM t1 a JOIN t2 b ON a.id
= b.id WHERE b.name = "test"
```

After this statement is executed, all the physical table shards that correspond to logical tables `t1` and `t2` and meet the conditions are calculated. Then, the SQL statement is routed to the specified physical table shards, and the result sets are merged and returned.

> **Notice** Before you use this custom hint, you must ensure that data from logical tables t1 and t2 is partitioned into the same number of table shards in the same number of database shards. Otherwise, the database shards that are calculated by the instance based on the conditions are different, and an error is returned.

- Execute the following SQL statement on table shards in database shards by explicitly specifying the names of the physical table shards:

```
SELECT /*+TDDL:scan('t1', real_table=("t1_00", "t1_01"))*/ COUNT(1) FROM t1
```

After this statement is executed, the SQL statement is routed to table shards `t1_00``t1_01` in all database shards, and the result sets are merged and returned.

- Execute the following JOIN statement on table shards in database shards by explicitly specifying the names of the physical table shards:

```
SELECT /*+TDDL:scan('t1, t2', real_table=("t1_00,t2_00", "t1_01,t2_01"))*/ * FROM t1 a JOIN t2 b ON
a.id = b.id WHERE b.name = "test";
```

After this statement is executed, the SQL statement is routed to table shards `t1_00`, `t2_00`, `t1_01`, and `t2_01` in all database shards, and the result sets are merged and returned.

12.1.9.6. INDEX HINT

This topic describes the syntax of the INDEX hint and provides examples.

- supports global secondary indexes (GSIs). The INDEX hint allows you to obtain query results from a specified GSI.
- The INDEX hint takes effect only for SELECT statements.

> **Note** The version of the MySQL engine in your ApsaraDB RDS for MySQL instance must be 5.7 or later, and the version of your instance must be 5.4.1 or later.

Syntax

```
FORCE INDEX
tbl_name [[AS] alias] [index_hint]
index_hint:
    FORCE INDEX({index_name})INDEX()
/*+TDDL:
    INDEX({table_name | table_alias}, {index_name})
*/
```

The INDEX hint has the following two variants:

- `FORCE INDEX()`: This syntax is the same as that of MySQL FORCE INDEX.
- `INDEX()`: In this syntax, a GSI is specified with a table name (or alias) and an index name. This hint does not take effect in the following cases:
  - The query does not contain the specified table name or alias.
  - The specified GSI is not in the specified table.

> **Note**
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

```
CREATE TABLE t_order (
 `id` bigint(11) NOT NULL AUTO_INCREMENT,
 `order_id` varchar(20) DEFAULT NULL,
 `buyer_id` varchar(20) DEFAULT NULL,
 `seller_id` varchar(20) DEFAULT NULL,
 `order_snapshot` longtext DEFAULT NULL,
 `order_detail` longtext DEFAULT NULL,
 PRIMARY KEY (`id`),
 GLOBAL INDEX `g_i_seller`(`seller_id`) dbpartition by hash(`seller_id`),
 UNIQUE GLOBAL INDEX `g_i_buyer` (`buyer_id`) COVERING(`seller_id`, `order_snapshot`)
  dbpartition by hash(`buyer_id`) tbpartition by hash(`buyer_id`) tbpartitions 3
) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by hash(`order_id`);
```

Specify the `g_i_seller` GSI by using the FORCE INDEX hint in the FROM clause:

```
 SELECT a.*, b.order_id
   FROM t_seller a
     JOIN t_order b FORCE INDEX(g_i_seller) ON a.seller_id = b.seller_id
   WHERE a.seller_nick="abc";
```

Specify the `g_i_buyer` GSI by using the INDEX hint and a table alias:

```
/*+TDDL:index(a, g_i_buyer)*/ SELECT * FROM t_order a WHERE a.buyer_id = 123
```

12.1.10. PolarDB-X 5.2 hints

12.1.10.1. Introduction to hints

As a supplement to the SQL syntax, hints play a critical role in relational databases. They allow you to modify execution plans of SQL statements by using the relevant syntax. This way, you can optimize the SQL statements.

Classification of PolarDB-X hints

provides special hint syntax.

For example, assume that you know the data is stored in table shards in specific database shards. If you need to route an SQL statement directly to the database shards for execution, you can use custom hints provided by .

```
/! TDDL:NODE IN('node_name', ...) */SELECT * FROM table_name;
```

In the preceding SQL statement, the part between `/!` and `*/` is a hint. This means that `TDDL:node in('node_name', ...)` is a hint. The hint specifies the ApsaraDB RDS for MySQL database shard where the SQL statement will be executed.

> **Note**
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Syntax of PolarDB-X hints

Basic syntax

```
/! TDDL:hint command*/
```

hints are based on the MySQL comment syntax. Therefore, a hint is located between `/!` and `*/`, and must begin with `TDDL:`. The `hint command` parameter indicates a hint command related to a specific operation. For example, a hint is added to the following SQL statement to show the name of each database shard.

```
/! TDDL:SCAN*/SHOW TABLES;
```

In this SQL statement, `/! TDDL:SCAN*/` is the hint that begins with `TDDL:`, and `SCAN` is a hint command.

12.1.10.2. Read/write splitting

This topic describes the read/write splitting feature provided by PolarDB-X.

provides transparent read/write splitting at the application layer. Data synchronization between primary and read-only ApsaraDB RDS for MySQL instances has a latency of several milliseconds. If you need to read the changed data immediately after data in the primary ApsaraDB RDS for MySQL instance is changed, you must ensure that the SQL statement for reading data is routed to the primary ApsaraDB RDS for MySQL instance. To meet this demand, provides custom hints for read/write splitting. These custom hints allow you to route SQL statements to a specified primary or read-only ApsaraDB RDS for MySQL instance.

Syntax

```
/! TDDL:MASTER|SLAVE*/
```

The custom hints allow you to specify whether to execute an SQL statement on a primary or read-only ApsaraDB RDS for MySQL instance. When you use the custom hint `/!TDDL:SLAVE*/`, if a primary ApsaraDB RDS for MySQL instance has multiple read-only ApsaraDB RDS for MySQL instances, the instance randomly selects a read-only ApsaraDB RDS for MySQL instance based on its weight, to execute the SQL statement.

> **Note**
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

- Execute an SQL statement on a specified primary ApsaraDB RDS for MySQL instance:

  ```
  /! TDDL:MASTER*/SELECT * FROM table_name;
  ```

  After the custom hint `/! TDDL:MASTER*/` is added behind the keyword of the SQL statement, this SQL statement is routed to the primary ApsaraDB RDS for MySQL instance.

- Execute an SQL statement on a specified read-only ApsaraDB RDS for MySQL instance:

  ```
  /! TDDL:SLAVE*/SELECT * FROM table_name;
  ```

  After the custom hint `/! TDDL:SLAVE*/` is added behind the keyword of the SQL statement, this SQL statement is randomly routed to a read-only ApsaraDB RDS for MySQL instance based on the allocated weight.

Considerations

- The custom hints for read/write splitting are only applicable to read SQL statements for non-transactional data. SQL statements for transactional data and write SQL statements are still routed to the primary ApsaraDB RDS for MySQL instance.
- When you use the `/*+TDDL:slave()*/` hint, the instance routes the SQL statement randomly to a read-only ApsaraDB RDS for MySQL instance based on the allocated weight. If no read-only ApsaraDB RDS for MySQL instance is available, no error is reported. Instead, the primary ApsaraDB RDS for MySQL instance is selected to execute the SQL statement.

12.1.10.3. Perform a switchover for a delayed read-only instance

Assume that you have configured a read-only instance for the primary ApsaraDB RDS for MySQL instance of a logical database in a instance and set read traffic for the primary instance and read-only instance. In this case, routes SQL statements to the primary instance or read-only instance based on the read/write ratio. Assume that PolarDB-X routes the SQL statements to the read-only instance. However, if asynchronous data replication between the primary instance and read-only instance has a high delay, an error is reported or an error result is returned.

PolarDB-X performs a switchover for a read-only instance based on the maximum delay of primary/secondary replication. PolarDB-X determines whether to route the SQL statements to the primary instance or the read-only instance.

Syntax

```
/! TDDL:SQL_DELAY_CUTOFF=time*/
```

You can specify `SQL_DELAY_CUTOFF` in the custom hint. If the SQL_DELAY value for the read-only ApsaraDB RDS for MySQL instance reaches or exceeds the `time` value, SQL statements are routed to the primary ApsaraDB RDS for MySQL instance. SQL_DELAY indicates the primary/secondary replication delay between the instances. time is measured in seconds.

> **Note**
>
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

- Set the primary/secondary replication delay to 5 seconds:

```
/! TDDL:SQL_DELAY_CUTOFF=5*/SELECT * FROM table_name;
```

  In this SQL statement, the value of `SQL_DELAY_CUTOFF` is set to 5. Therefore, if the value of `SQL_DELAY` for the read-only instance reaches or exceeds 5 seconds, SQL statements are routed to the primary instance.

- Use the custom hint in conjunction with other custom hints:

```
/! TDDL:SLAVE AND SQL_DELAY_CUTOFF=5*/SELECT * FROM table_name;
```

  The custom hint that is used to perform a switchover for a delayed read-only ApsaraDB RDS for MySQL instance can be used in conjunction with other hints. By default, an SQL query request is routed to a read-only ApsaraDB RDS for MySQL instance. However, if the primary/secondary replication delay reaches or exceeds 5 seconds, the SQL query request is routed to the primary ApsaraDB RDS for MySQL instance.

12.1.10.4. Specify a timeout period for an SQL statement

In , if the amount of time consumed to execute an SQL statement on a node and an ApsaraDB RDS for MySQL instance exceeds the default duration of 900 seconds, the execution times out. You can adjust the timeout period. The duration during which a slow SQL statement is executed may exceed 900 seconds. For this type of slow SQL statement, provides a custom hint to help you adjust the timeout period. You can use this custom hint to adjust the duration in which a SQL statement is executed.

Syntax

The following syntax of the hint can be used to specify a timeout period for an SQL statement:

```
/!TDDL:SOCKET_TIMEOUT=time*/
```

The value specified by `SOCKET_TIMEOUT` is measured in milliseconds. You can use this custom hint to change the timeout period for SQL statements based on your business requirements.

> **Note**
> - You can specify custom hints in the `/*+TDDL:hint_command*/` format or in the `/!+TDDL:hint_command */` format.
> - In the official MySQL CLI, if you execute SQL statements that contain customer hints in the `/*+TDDL:hint_command*/` format, add the `-c` parameter in the `mysql -u*** -p*** -h***` command that is run to log on to the client. Otherwise, the client deletes custom hints before the client sends the SQL statements to the server for execution. This is because the hints are in the format of a MySQL comment. As a result, the customer hints do not take effect. For more information, see MySQL client options.

Examples

Set the timeout period for a SQL statement to 40 seconds.

```
/!TDDL:SOCKET_TIMEOUT=40000*/SELECT * FROM t_item;
```

> **Note** A longer timeout period causes database resources to be consumed for a longer period of time. If a large number of SQL statements are executed over a long period of time within the same interval, a large number of database resources may be consumed. As a result, DRDS database services cannot be provided as expected. To resolve the issue, we recommend that you optimize SQL statements that take a long time to execute if possible.

12.1.10.5. Specify a database shard to execute an SQL statement

When you execute Structured Query Language (SQL) statements in , some SQL statements may not be supported by . In this case, you can use a custom hint provided by to route the SQL statements to one or more database shards for execution. If you want to query the data in a database shard or table shard, you can use a custom hint to route the SQL statement to the database shard for execution.

Syntax

You can use two methods to specify a database shard to execute an SQL statement that has a custom hint: Specify a database shard name or a database shard key value. A shard name is a unique identifier in a database shard. You can run the `SHOW NODE` control command to obtain the shard name.

> **Note** If the hint for specifying a database shard is used in an INSERT statement that contains a sequence for the destination table, the sequence does not take effect. For more information, see Limits and precautions for sequences.

- Execute an SQL statement on a database shard by specifying a shard name:

  If you specify shard names, two methods can be used to execute SQL statements: Execute SQL statements on one or more database shards.

  - Specify one database shard to execute an SQL statement:

    ```
    /! TDDL:NODE='node_name'*/
    ```

    `node_name` indicates the shard name. The custom hint routes the SQL statement to the database shard specified by `node_name`.

  - Specify multiple database shards to execute an SQL statement:

    ```
    /! TDDL:NODE IN ('node_name'[,'node_name1','node_name2'])*/
    ```

    Use the `in` keyword to specify multiple shard names. This way, the SQL statement can be routed to multiple database shards. Separate multiple shard names with commas (,).

    > **Note** If the custom hint is used, routes the SQL statement to the database shards for execution. Therefore, you must specify the shard names that exist in the database for the SQL statement.

- Execute an SQL statement on a database shard by specifying a shard key value:

  ```
  /! TDDL:table_name.partition_key=value [and table_name1.partition_key=value1]*/
  ```

  In the custom hint, `table_name` indicates the name of a logical table, and the table is a sharded table. `partition_key` indicates a shard key, and `value` indicates the shard key value. In the custom hint, you can use the `and` keyword to specify the shard keys of multiple sharded tables. When the custom hint is used, calculates the database shards and table shards where the SQL statement is to be executed. Then, PolarDB-X routes the SQL statement to the destination database shards.

> **Note**
>
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

The following example shows the result of the `SHOW NODE;` statement executed in the database named `drds_test`:

```
*************************** 1. row ******************
                ID: 0
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS
  MASTER_READ_COUNT: 212
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 2. row ******************
                ID: 1
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0001_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 3. row ******************
                ID: 2
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0002_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 4. row ******************
                ID: 3
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0003_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 5. row ******************
                ID: 4
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0004_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 6. row ******************
                ID: 5
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0005_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 7. row ******************
                ID: 6
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0006_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
*************************** 8. row ******************
                ID: 7
              NAME: DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0007_RDS
  MASTER_READ_COUNT: 29
   SLAVE_READ_COUNT: 0
MASTER_READ_PERCENT: 100%
 SLAVE_READ_PERCENT: 0%
8 rows in set (0.02 sec)
```

Each database shard has the `NAME` attribute that indicates the shard name of the database shard. Each shard name corresponds to one unique database shard name. For example, the shard name `DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0003_RDS` corresponds to the database shard name `drds_test_vtla_0003`. After you have obtained the shard name, you can use the custom hint to specify a database shard to execute an SQL statement.

- Execute an SQL statement on database shard 0:

```
/! TDDL:NODE='DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS'*/SELECT * FROM table_name;
```

- Execute an SQL statement on multiple database shards:

```
/! TDDL:NODE  IN('DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS','DRDS_TEST_1473471355140LRPRD
RDS_TEST_VTLA_0006_RDS')*/SELECT * FROM table_name;
```

  The SQL statement is executed on the shards `DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS` and `D RDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0006_RDS`.

- View the execution plan on a database shard:

```
/! TDDL:NODE='DRDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS'*/EXPLAIN SELECT * FROM table_name
;
```

  After the SQL statement is executed, the execution plan of the `SELECT` statement on the database shard `D RDS_TEST_1473471355140LRPRDRDS_TEST_VTLA_0000_RDS` appears.

- Execute an SQL statement on a database shard by specifying the shard key value:

  does not support subqueries in the `SET` clause of an `UPDATE` statement, because a shard key must be specified for the `UPDATE` statement in . You can use the custom hint to route the statement to a database shard for execution.

  For example, the following statements are used to create two logical tables t1 and t2. The two tables are table shards in database shards.

```
CREATE TABLE `t1` (
    `id` bigint(20) NOT NULL,
    `name` varchar(20) NOT NULL,
    `val` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`id`)
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by hash(`id`) tbpartition by hash(`name`) tbpart
itions 3;
  CREATE TABLE `t2` (
    `id` bigint(20) NOT NULL,
    `name` varchar(20) NOT NULL,
    `val` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`id`)
  ) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by hash(`id`) tbpartition by hash(`name`) tbpart
itions 3;
```

  Execute the following statement:

```
UPDATE t1 SET val=(SELECT val FROM t2 WHERE id=1) WHERE id=1;
```

  If the statement is directly executed in , an error that indicates the statement is not supported is generated. In this case, you can add the custom hint to the SQL statement before you submit it to for execution. Execute the following statement:

```
/! TDDL:t1.id=1 and t2.id=1*/UPDATE t1 SET val=(SELECT val FROM t2 WHERE id=1) WHERE id=1;
```

The statement is routed to the database shard whose `id` is 1 on `t1`. You can run the `explain /! TDDL:t1.id=1 and t2.id=1*/UPDATE t1 SET val=(SELECT val FROM t2 WHERE id=1) WHERE id=1;` command to view the execution plan of the SQL statement:

```
   *************************** 1. row ***************************
   GROUP_NAME: TEST_DRDS_1485327111630IXLWTEST_DRDS_IGHF_0001_RDS
          SQL: UPDATE `t1_2` AS `t1` SET `val` = (SELECT val FROM `t2_2` AS `t2` WHERE `id` = 1) WHE
RE `id` = 1
       PARAMS: {}
   *************************** 2. row ***************************
   GROUP_NAME: TEST_DRDS_1485327111630IXLWTEST_DRDS_IGHF_0001_RDS
          SQL: UPDATE `t1_1` AS `t1` SET `val` = (SELECT val FROM `t2_1` AS `t2` WHERE `id` = 1) WHE
RE `id` = 1
       PARAMS: {}
   *************************** 3. row ***************************
   GROUP_NAME: TEST_DRDS_1485327111630IXLWTEST_DRDS_IGHF_0001_RDS
          SQL: UPDATE `t1_0` AS `t1` SET `val` = (SELECT val FROM `t2_0` AS `t2` WHERE `id` = 1) WHE
RE `id` = 1
       PARAMS: {}
   3 rows in set (0.00 sec)
```

Based on the result set of the `EXPLAIN` command, the SQL statement is rewritten into three statements and routed to the database shards for execution. You can also specify the value of a table shard key to narrow the execution scope of the SQL statement to one table shard. For example, if you run the `explain /! TDDL:t1.id=1 and t2.id=1 and t1.name='1'*/UPDATE t1 SET val=(SELECT val FROM t2 WHERE id=1) WHERE id=1;` command, the following result is returned:

```
   *************************** 1. row ***************************
   GROUP_NAME: TEST_DRDS_1485327111630IXLWTEST_DRDS_IGHF_0001_RDS
          SQL: UPDATE `t1_1` AS `t1` SET `val` = (SELECT val FROM `t2_1` AS `t2` WHERE `id` = 1) WHE
RE `id` = 1
       PARAMS: {}
   1 row in set (0.00 sec)
```

Note Before you use this custom hint, ensure that data from logical tables t1 and t2 is partitioned into the same number of table shards in the same number of database shards. Otherwise, the database shards that are calculated by the instance based on the conditions are different, and an error is returned.

12.1.10.6. Scan all database shards and table shards

You can route an SQL statement to one or more database shards. You can also route an SQL statement to all database shards and table shards by using the SCAN hint provided by . You can use this custom hint to route an SQL statement to all database shards at a time. For example, you can use this custom hint to view all the table shards in a specified database shard. You can also use this custom hint to view the data volume of table shards in each database shard that corresponds to a specified logical table.

Syntax

This hint allows you to route an SQL statement to all database shards for execution and route an SQL statement to all database shards to perform an operation on a specified logical table.

- Route an SQL statement to all database shards for execution:

```
/! TDDL:SCAN*/
```

- Perform an operation on a specified logical table:

```
/! TDDL:SCAN='table_name'*/
```

The `table_name` parameter indicates the name of a logical table in a logical database. This custom hint is provided for table shards in database shards. Make sure that the value of `table_name` is the name of a table shard in database shards.

> **Note**
>
> - hints can be in the formats of `/*+TDDL:hint_command*/` and `/! +TDDL:hint_command*/`.
> - In the MySQL command-line client, you may need to execute an SQL statement that contains a hint in the format of `/*+TDDL:hint_command*/`. In this case, add the `-c` parameter to the logon command. Otherwise, the client deletes the hint before it sends the SQL statement to the server for execution because the hint is in the format of a MySQL comment. In this case, the hint fails to take effect. For more information, see MySQL Client Options.

Examples

- View the data volume of a specified broadcast table in each database shard:

```
/! TDDL:SCAN*/SELECT COUNT(1) FROM table_name;
```

In this SQL statement, `table_name` indicates a broadcast table. This hint allows you to route the SQL statement to each database shard for execution. Therefore, the result sets include the total data volume of the broadcast table `table_name` in all database shards. This statement allows you to conveniently check whether the data of a broadcast table is normal.

- Scan a logical table whose data is partitioned into database shards but not further into table shards:

```
/! TDDL:SCAN*/SELECT COUNT(1) FROM table_name;
```

This hint allows you to route the `SELECT COUNT (1) FROM table_name` statement to each database shard for execution. The `table_name` parameter indicates a logical table in a logical database. Before you use this hint, make sure that each database shard contains the table shard `table_name`. This means that `table_name` is a logical table whose data is partitioned only into database shards but not further into table shards. Otherwise, an error will be returned, indicating that the table is not found.

- Scan a logical table whose data is partitioned into table shards in database shards:

```
/! TDDL:SCAN='table_name'*/SELECT COUNT(1) FROM table_name;
```

When you execute this statement, the instance first calculates all the database shards and table shards corresponding to the logical table `table_name`. Then, the instance generates a COUNT clause for each table shard in each database shard.

- View the execution plans on all database shards:

```
/! TDDL:SCAN='table_name'*/EXPLAIN SELECT * FROM table_name;
```

12.1.11. Distributed transactions

12.1.11.1. Distributed transactions based on MySQL 5.7

This topic describes distributed transactions based on MySQL 5.7.

> **Note**
> - If the version of the MySQL engine in your ApsaraDB RDS for MySQL instance is 5.7 or later and the version of your instance is 5.3.4 or later, XA distributed transactions are automatically enabled. The user experience of XA distributed transactions on PolarDB-X is the same as that on standalone databases that run the MySQL engine. You do not need to run special commands to enable XA distributed transactions.
> - For more information about MySQL and PolarDB-X in other versions, see Distributed transactions based on MySQL 5.6.

How it works

If the version of the MySQL engine in your ApsaraDB RDS for MySQL instance is 5.7 or later, processes distributed transactions based on the XA protocol by default.

Method

The user experience of distributed transactions on is the same as that on standalone databases that run the MySQL engine. For example, you can execute the following statements:

- `SET AUTOCOMMIT=0` : starts a transaction.
- `COMMIT` : commits a transaction.
- `ROLLBACK` : rolls back a transaction.

If SQL statements in a transaction involve only a single shard, treats the transaction as a non-distributed transaction. In this case, the system directly routes the transaction to an ApsaraDB RDS for MySQL instance. If SQL statements in a transaction are used to modify data in multiple shards, automatically upgrades the transaction to a distributed transaction.

12.1.11.2. Distributed transactions based on MySQL 5.6

This topic describes distributed transactions based on MySQL 5.6.

How it works

The XA protocol for MySQL 5.6 is immature. Therefore, independently implements the two-phase commit protocol (2PC) for distributed transactions. If the version of the MySQL engine in your ApsaraDB RDS for MySQL instance is 5.7 or later, we recommend that you use the XA protocol for transactions.

> **Note** The distributed transactions that are described in this topic are intended for users who use MySQL 5.6 or earlier than 5.3.4. For more information about MySQL 5.7 or later and 5.3.4 or later, see Distributed transactions based on MySQL 5.7.

Method

If a transaction involves multiple database shards, declare the transaction as a distributed transaction. If a transaction involves only a single database shard, you do not need to enable distributed transactions. The system treats such a transaction as a non-distributed transaction and directly routes the transaction to an ApsaraDB RDS for MySQL instance.

If you want to enable distributed transactions, perform the following operations:

After you enable distributed transactions, execute the `SET drds_transaction_policy = '...'` statement.

To enable 2PC transactions, execute the following statements in a MySQL command-line client:

```
SET AUTOCOMMIT=0;
SET drds_transaction_policy = '2PC'; -- We recommend that you execute this statement if you use MySQL
5.6.
.... -- Execute SQL statements.
COMMIT; -- You can also execute the ROLLBACK statement.
```

If you use Java Database Connectivity (JDBC), execute the following sample statements to enable 2PC transactions:

```
conn.setAutoCommit(false);
try (Statement stmt = conn.createStatement()) {
    stmt.execute("SET drds_transaction_policy = '2PC'");
}
// ... Execute SQL statements ...
conn.commit(); // You can also execute the ROLLBACK statement.
```

FAQ

How can I use Spring Framework to enable distributed transactions in ?

If you use the annotation `@Transactional` in Spring Framework, you can extend the transaction manager to enable distributed transactions in .

Sample code:

```
import org.springframework.jdbc.datasource.DataSourceTransactionManager;
import org.springframework.transaction.TransactionDefinition;
import javax.sql.DataSource;
import java.sql.Connection;
import java.sql.SQLException;
import java.sql.Statement;
public class DrdsTransactionManager extends DataSourceTransactionManager {
    public DrdsTransactionManager(DataSource dataSource) {
        super(dataSource);
    }
    @Override
    protected void prepareTransactionalConnection(Connection con, TransactionDefinition definition) th
rows SQLException {
        try (Statement stmt = con.createStatement()) {
            stmt.executeUpdate("SET drds_transaction_policy = '2PC'"); // A 2PC transaction is provide
d as an example.
        }
    }
}
```

You can use the following sample code to instantiate the preceding class in Spring Framework:

```
<bean id="drdsTransactionManager" class="my.app.DrdsTransactionManager">
     <property name="dataSource" ref="yourDataSource" />
</bean>
```

If you need to enable distributed transactions in for a class, add the annotation `@Transactional("drdsTransactionManager")` .

12.1.12. DDL operations

12.1.12.1. DDL statements

The data definition language (DDL) statement CREATE TABLE in a instance is similar to that in a MySQL database, and is extended based on the syntax in a MySQL database. To create a table shard in a instance, you must specify the table sharding manner and the database sharding manner in the `drds_partition_options` parameter. The valid values include `DBPARTITION BY` , `TBPARTITION BY` , `TBPARTITIONS` , and `BROADCAST` .

Currently, you can run a DDL statement in the following ways:

- Run the DDL statement through the MySQL command-line client, for example, by using MySQL command lines, Navicat, or MySQL Workbench.
- Connect to the specified instance by using program code and then call the DDL statement for execution.

For the syntax of the CREATE TABLE statement in a MySQL database, see MySQL CREATE TABLE Statement.

12.1.12.2. CREATE TABLE statement

12.1.12.2.1. Overview

This topic describes the syntax, clauses, parameters, and basic methods for creating a table by using a data definition language (DDL) statement.

> **Note** instances do not allow you to directly create a database by using a DDL statement. To create a database, you can Log on to the PolarDB-X console. For the information about how to create a database, see Create a database.

Syntax

```
CREATE [TEMPORARY] TABLE [IF NOT EXISTS] tbl_name
    (create_definition,...)
    [table_options]
    [drds_partition_options]
    [partition_options]
drds_partition_options:
    DBPARTITION BY
        HASH([column])
        [TBPARTITION BY
            { HASH(column)
            | {MM|DD|WEEK|MMDD}(column)}
          [TBPARTITIONS num]
        ]
```

Clauses and parameters for database and table sharding

- `DBPARTITION BY hash(partition_key)` : This parameter specifies the shard key and the sharding algorithm for database sharding. Database sharding by time is not supported.
- `TBPARTITION BY { HASH(column) | {MM|DD|WEEK|MMDD}(column)` : (Optional) This parameter specifies the method of mapping data to a physical table. The value is the same as that of DBPARTITION BY by default.
- `TBPARTITIONS num` : (Optional) This parameter specifies the number of physical tables to be created in each database shard. The default value is 1. If no table sharding is required, you do not need to specify this parameter.

12.1.12.2.2. Create a single-database non-partitioned table

This topic describes how to create a single-database non-partitioned table.

Create a single-database non-partitioned table

```
CREATE TABLE single_tbl(
  id int,
  name varchar(30),
  primary key(id)
);
```

Execute the `SHOW TOPOLOGY FROM single_tbl;` statement to view the node topology of the logical table. Based on the output, a single-database non-partitioned logical table is created only on logical database 0.

```
+------+-------------------------------------------------------------------+------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME |
+------+-------------------------------------------------------------------+------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | single_tbl |
+------+-------------------------------------------------------------------+------------+
1 row in set (0.01 sec)
```

Set parameters

You can also set the `select_statement` parameter when you create a single-database non-partitioned table. If you need to create a partitioned table, you cannot set this parameter.

```
CREATE [TEMPORARY] TABLE [IF NOT EXISTS] tbl_name
    [(create_definition,...)]
    [table_options]
    [partition_options]
    select_statement
```

For example, you can execute the following statement to create a single-database non-partitioned table named single_tbl2 to store the data from the single_tbl table:

```
CREATE TABLE single_tbl2(
  id int,
  name varchar(30),
  primary key(id)
) select * from single_tbl;
```

12.1.12.2.3. Create a logical table partitioned into database shards

This topic describes how to create a logical table whose data is partitioned into database shards but not further into table shards.

Assume that eight database shards are created. You can execute the following statement to create a logical table whose data is hash-partitioned into the eight database shards based on the ID column.

```
CREATE TABLE multi_db_single_tbl(
  id int,
  name varchar(30),
  primary key(id)
) dbpartition by hash(id);
```

Execute the `show topology from multi_db_single_tbl;` statement to view the node topology of the logical table. Based on the output, one table shard is created on each database shard. This means that only database sharding is performed.

```
+------+-------------------------------------------------------------------+---------------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME          |
+------+-------------------------------------------------------------------+---------------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | multi_db_single_tbl |
|    1 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | multi_db_single_tbl |
|    2 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0002_RDS | multi_db_single_tbl |
|    3 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0003_RDS | multi_db_single_tbl |
|    4 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0004_RDS | multi_db_single_tbl |
|    5 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS | multi_db_single_tbl |
|    6 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0006_RDS | multi_db_single_tbl |
|    7 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | multi_db_single_tbl |
+------+-------------------------------------------------------------------+---------------------+
8 rows in set (0.01 sec)
```

12.1.12.2.4. Sharding

This topic describes how to use different methods to implement sharding.

You can use the following methods for sharding:

- Use hash functions for sharding.
- Use double hashing for sharding.
- Use date functions for sharding.

The following examples are based on eight database shards.

Use hash functions for sharding

Create a table for which database sharding and table sharding are implemented. Each database shard contains three physical tables. For database sharding, hashing is implemented based on the id column. For table sharding, hashing is implemented based on the bid column. You can generate hash values for the id column values and distribute the table data to database shards. Then, generate hash values for the bid column values and distribute the data in each database shard to three physical tables.

```
CREATE TABLE multi_db_multi_tbl(
  id int auto_increment,
  bid int,
  name varchar(30),
  primary key(id)
) dbpartition by hash(id) tbpartition by hash(bid) tbpartitions 3;
```

To view the node topology of the logical table, execute the `show topology from multi_db_multi_tbl;` statement. The node topology shows that three table shards are created in each database shard.

```
+------+-------------------------------------------------------------------+-----------------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME            |
+------+-------------------------------------------------------------------+-----------------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | multi_db_multi_tbl_00 |
|    1 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | multi_db_multi_tbl_01 |
|    2 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | multi_db_multi_tbl_02 |
|    3 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | multi_db_multi_tbl_03 |
|    4 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | multi_db_multi_tbl_04 |
|    5 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | multi_db_multi_tbl_05 |
|    6 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0002_RDS | multi_db_multi_tbl_06 |
|    7 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0002_RDS | multi_db_multi_tbl_07 |
|    8 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0002_RDS | multi_db_multi_tbl_08 |
|    9 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0003_RDS | multi_db_multi_tbl_09 |
|   10 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0003_RDS | multi_db_multi_tbl_10 |
|   11 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0003_RDS | multi_db_multi_tbl_11 |
|   12 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0004_RDS | multi_db_multi_tbl_12 |
|   13 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0004_RDS | multi_db_multi_tbl_13 |
|   14 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0004_RDS | multi_db_multi_tbl_14 |
|   15 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS | multi_db_multi_tbl_15 |
|   16 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS | multi_db_multi_tbl_16 |
|   17 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS | multi_db_multi_tbl_17 |
|   18 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0006_RDS | multi_db_multi_tbl_18 |
|   19 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0006_RDS | multi_db_multi_tbl_19 |
|   20 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0006_RDS | multi_db_multi_tbl_20 |
|   21 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | multi_db_multi_tbl_21 |
|   22 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | multi_db_multi_tbl_22 |
|   23 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | multi_db_multi_tbl_23 |
+------+-------------------------------------------------------------------+-----------------------+
24 rows in set (0.01 sec)
```

To view the sharding rule of the logical table, execute the `show rule from multi_db_multi_tbl;` statement. The returned result shows that hashing is implemented for sharding. The shard key for database sharding is id, and the shard key for table sharding is bid.

```
+------+--------------------+-----------+------------------+---------------------+--------------------
+------------------+---------------------+--------------------+
| ID   | TABLE_NAME         | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT
| TB_PARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+--------------------+-----------+------------------+---------------------+--------------------
+------------------+---------------------+--------------------+
|    0 | multi_db_multi_tbl |         0 | id               | hash                | 8
| bid              | hash                | 3                  |
+------+--------------------+-----------+------------------+---------------------+--------------------
+------------------+---------------------+--------------------+
1 row in set (0.01 sec)
```

Use double hashing for sharding

- Usage notes

  A shard key must be a character or a number.

- Routing method

  Calculate a hash value based on the last N characters of a shard key. Then, use RANGE_HASH to calculate a route. N is the third parameter in the function. For example, if you use the RANGE_HASH(COL1, COL2, N) function, the function first selects COL1 and truncates COL1 to obtain the last N characters of the value. If COL1 does not exist, COL2 is selected.

- Scenarios

RANGE_HASH is suitable for scenarios in which two shard keys are used for sharding but only one shard key value is used for queries. In the following example, a database in is partitioned into eight physical databases. The customer has the following requirements:

i. Partition the order table of each service into database shards by buyer ID and order ID.

ii. Use the buyer ID or order ID as the condition to perform a query.

Execute the following data definition language (DDL) statement to create a table:

```
create table test_order_tb (
           id int,
           seller_id varchar(30) DEFAULT NULL,
           order_id varchar(30) DEFAULT NULL,
           buyer_id varchar(30) DEFAULT NULL,
           create_time datetime DEFAULT NULL,
           primary key(id)
   ) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by RANGE_HASH(buyer_id, order_id, 10) tbpartiti
on by RANGE_HASH(buyer_id, order_id, 10) tbpartitions 3;
```

Note

- You cannot modify the two shard keys.
- Data fails to be inserted if the two shard keys point to different database shards or table shards.

Use date functions for sharding

You can use a hash function for sharding. You can also use the date function MM, DD, WEEK, or MMDD for table sharding. The following examples are provided to show the procedure:

- Create a table for which database sharding and table sharding are implemented. For database sharding, hashing is implemented based on the userId column. For table sharding, the actionDate column is used, and the assumption that each week has seven days applies. The WEEK(actionDate) function calculates DAY_OF_WEEK.

  For example, if a value in the actionDate column is 2017-02-27, which is on Monday, the value that the WEEK(actionDate) function returns is 2. In this case, the record is stored in table shard 2 based on the equation 2 % 7 = 2. This table shard is located in a database shard named user_log_2. For another example, if a value in the actionDate column is 2017-02-26, which is on Sunday, the value that the WEEK(actionDate) function returns is 1. In this case, the record is stored in table shard 1 based on the equation 1 % 7 = 1. This table shard is located in a database shard named user_log_1.

  ```
  CREATE TABLE user_log(
  userId int,
  name varchar(30),
  operation varchar(30),
  actionDate DATE
  ) dbpartition by hash(userId) tbpartition by WEEK(actionDate) tbpartitions 7;
  ```

  To view the node topology of the logical table, execute the `show topology from user_log;` statement. Each week has seven days. The node topology shows that seven table shards are created in each database shard. In the following example, an ellipsis (...) is used toomit some data because the returned result is long.

```
+------+-------------------------------------------------------------------+------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME |
+------+-------------------------------------------------------------------+------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_0 |
|    1 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_1 |
|    2 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_2 |
|    3 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_3 |
|    4 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_4 |
|    5 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_5 |
|    6 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log_6 |
|    7 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_0 |
|    8 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_1 |
|    9 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_2 |
|   10 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_3 |
|   11 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_4 |
|   12 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_5 |
|   13 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log_6 |
...
|   49 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_0 |
|   50 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_1 |
|   51 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_2 |
|   52 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_3 |
|   53 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_4 |
|   54 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_5 |
|   55 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log_6 |
+------+-------------------------------------------------------------------+------------+
56 rows in set (0.01 sec)
```

To view the sharding rule of the logical table, execute the `show rule from user_log;` statement. The returned result shows that hashing is implemented for database sharding, and the shard key for database sharding is userId. For table sharding, the date function WEEK is used, and the shard key for table sharding is actionDate.

```
+------+------------+-----------+------------------+---------------------+---------------------+----
--------------+----------------------+--------------------+
| ID   | TABLE_NAME | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT | TB_
PARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+------------+-----------+------------------+---------------------+---------------------+----
--------------+----------------------+--------------------+
|    0 | user_log   |         0 | userId           | hash                | 8                   | act
ionDate       | week                 | 7                  |
+------+------------+-----------+------------------+---------------------+---------------------+----
--------------+----------------------+--------------------+
1 row in set (0.00 sec)
```

View the physical database shard and its physical table to which the SQL statement is routed if the parameters of the database shard key and table shard key are specified.

View the routing of SQL statements

```
mysql> explain select name from user_log where userId = 1 and actionDate = '2017-02-27'\G
*************************** 1. row ***************************
GROUP_NAME: SANGUAN_1490167540907XWDVSANGUAN_BSQT_0001_RDS
       SQL: select `user_log`.`name` from `user_log_2` `user_log` where ((`user_log`.`userId` = 1) AND (`user_log`.`actionDate` = '2017-02-27'))
    PARAMS: {}
1 row in set (0.01 sec)
```

- Create a table for which database sharding and table sharding are implemented. For database sharding, hashing is implemented based on the userId column. For table sharding, the actionDate column is used, and the assumption that each year has 12 months applies. The MM(actionDate) function calculates MONTH_OF_YEAR.

For example, if a value in the actionDate column is 2017-02-27, the value that the MM(actionDate) function returns is 02. In this case, the record is stored in table shard 02 based on the equation 02 % 12 = 02. This table shard is located in a database shard named user_log_02. For another example, if a value in the actionDate column is 2016-12-27, the value that the MM(actionDate) function returns is 12. In this case, the record is stored in table shard 00 based on the equation 12 % 12 = 00. This table shard is located in a database shard named user_log_00.

```
CREATE TABLE user_log2(
userId int,
name varchar(30),
operation varchar(30),
actionDate DATE
) dbpartition by hash(userId) tbpartition by MM(actionDate) tbpartitions 12;
```

To view the node topology of the logical table, execute the `show topology from user_log2;` statement. Each year has 12 months. The node topology shows that 12 table shards are created in each database shard. An ellipsis (...) is used toomit some data because the returned result is long.

```
+------+-------------------------------------------------------------------+--------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME   |
+------+-------------------------------------------------------------------+--------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_00 |
|    1 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_01 |
|    2 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_02 |
|    3 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_03 |
|    4 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_04 |
|    5 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_05 |
|    6 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_06 |
|    7 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_07 |
|    8 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_08 |
|    9 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_09 |
|   10 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_10 |
|   11 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log2_11 |
|   12 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_00 |
|   13 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_01 |
|   14 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_02 |
|   15 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_03 |
|   16 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_04 |
|   17 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_05 |
|   18 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_06 |
|   19 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_07 |
|   20 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_08 |
|   21 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_09 |
|   22 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_10 |
|   23 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | user_log2_11 |
...
|   84 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_00 |
|   85 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_01 |
|   86 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_02 |
|   87 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_03 |
|   88 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_04 |
|   89 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_05 |
|   90 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_06 |
|   91 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_07 |
|   92 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_08 |
|   93 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_09 |
|   94 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_10 |
|   95 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log2_11 |
+------+-------------------------------------------------------------------+--------------+
96 rows in set (0.02 sec)
```

To view the sharding rule of the logical table, execute the `show rule from user_log2;` statement. The returned result shows that hashing is implemented for database sharding, and the shard key for database sharding is userId. For table sharding, the date function MM is used, and the shard key for table sharding is actionDate.

```
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+--------------------+-------------------+
| ID   | TABLE_NAME | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT | TB_
PARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+--------------------+-------------------+
|    0 | user_log2  |         0 | userId           | hash                | 8                  | act
ionDate       | mm                 | 12                |
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+--------------------+-------------------+
1 row in set (0.00 sec)
```

- Create a table for which database sharding and table sharding are implemented. For database sharding, hashing is implemented based on the userId column. For table sharding, the assumption that each month has 31 days applies. The DD(actionDate) function calculates DAY_OF_MONTH.

  For example, if a value in the actionDate column is 2017-02-27, the value that the DD(actionDate) function returns is 27. In this case, the record is stored in table shard 27 based on the equation 27 % 31 = 27. This table shard is located in a database shard named user_log_27.

  ```
  CREATE TABLE user_log3(
  userId int,
  name varchar(30),
  operation varchar(30),
  actionDate DATE
  ) dbpartition by hash(userId) tbpartition by DD(actionDate) tbpartitions 31;
  ```

  To view the node topology of the logical table, execute the `show topology from user_log3;` statement. Each month has 31 days. The node topology shows that 31 table shards are created in each database shard. An ellipsis (...) is used toomit some data because the returned result is long.

```
+------+-------------------------------------------------------------------+--------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME   |
+------+-------------------------------------------------------------------+--------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_00 |
|    1 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_01 |
|    2 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_02 |
|    3 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_03 |
|    4 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_04 |
|    5 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_05 |
|    6 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_06 |
|    7 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_07 |
|    8 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_08 |
|    9 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_09 |
|   10 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_10 |
|   11 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_11 |
|   12 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_12 |
|   13 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_13 |
|   14 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_14 |
|   15 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_15 |
|   16 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_16 |
|   17 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_17 |
|   18 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_18 |
|   19 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_19 |
|   20 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_20 |
|   21 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_21 |
|   22 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_22 |
|   23 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_23 |
|   24 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_24 |
|   25 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_25 |
|   26 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_26 |
|   27 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_27 |
|   28 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_28 |
|   29 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_29 |
|   30 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log3_30 |
...
|  237 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_20 |
|  238 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_21 |
|  239 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_22 |
|  240 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_23 |
|  241 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_24 |
|  242 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_25 |
|  243 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_26 |
|  244 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_27 |
|  245 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_28 |
|  246 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_29 |
|  247 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log3_30 |
+------+-------------------------------------------------------------------+--------------+
248 rows in set (0.01 sec)
```

To view the sharding rule of the logical table, execute the `show rule from user_log3;` statement. The returned result shows that hashing is implemented for database sharding, and the shard key for database sharding is userId. For table sharding, the date function DD is used, and the shard key for table sharding is actionDate.

```
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+--------------------+--------------------+
| ID   | TABLE_NAME | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT | TB_
PARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+--------------------+--------------------+
|    0 | user_log3  |         0 | userId           | hash                | 8                  | act
ionDate       | dd                 | 31                 |
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+--------------------+--------------------+
1 row in set (0.01 sec)
```

- Create a table for which database sharding and table sharding are implemented. For database sharding, hashing is implemented based on the userId column. For table sharding, the assumption that each year has 365 days applies, and the table data is routed to 365 physical tables. The MMDD(actionDate) tbpartitions 365 function calculates DAY_OF_YEAR % 365.

  For example, if a value in the actionDate column is 2017-02-27, the value that the MMDD function returns is 58. In this case, the record is stored in table shard 58. This table shard is located in a database shard named user_log_58.

  ```
  CREATE TABLE user_log4(
  userId int,
  name varchar(30),
  operation varchar(30),
  actionDate DATE
  ) dbpartition by hash(userId) tbpartition by MMDD(actionDate) tbpartitions 365;
  ```

  To view the node topology of the logical table, execute the `show topology from user_log4;` statement. Each year has 365 days. The node topology shows that 365 table shards are created in each database shard. An ellipsis (...) is used toomit some data because the returned result is long.

```
+------+-------------------------------------------------------------------+---------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME    |
+------+-------------------------------------------------------------------+---------------+
...
| 2896 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_341 |
| 2897 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_342 |
| 2898 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_343 |
| 2899 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_344 |
| 2900 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_345 |
| 2901 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_346 |
| 2902 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_347 |
| 2903 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_348 |
| 2904 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_349 |
| 2905 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_350 |
| 2906 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_351 |
| 2907 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_352 |
| 2908 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_353 |
| 2909 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_354 |
| 2910 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_355 |
| 2911 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_356 |
| 2912 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_357 |
| 2913 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_358 |
| 2914 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_359 |
| 2915 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_360 |
| 2916 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_361 |
| 2917 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_362 |
| 2918 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_363 |
| 2919 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log4_364 |
+------+-------------------------------------------------------------------+---------------+
2920 rows in set (0.07 sec)
```

To view the sharding rule of the logical table, execute the `show rule from user_log4;` statement. The returned result shows that hashing is implemented for database sharding, and the shard key for database sharding is userId. For table sharding, the date function MMDD is used, and the shard key for table sharding is actionDate.

```
+------+------------+-----------+-----------------+---------------------+--------------------+----
--------------+---------------------+--------------------+
| ID   | TABLE_NAME | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT | TB_
PARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+------------+-----------+-----------------+---------------------+--------------------+----
--------------+---------------------+--------------------+
|    0 | user_log4  |         0 | userId           | hash                | 8                  | act
ionDate       | mmdd                | 365                |
+------+------------+-----------+-----------------+---------------------+--------------------+----
--------------+---------------------+--------------------+
1 row in set (0.02 sec)
```

- Create a table for which database sharding and table sharding are implemented. For database sharding, hashing is implemented based on the userId column. For table sharding, the assumption that each year has 365 days applies, and the table data is routed to 10 physical tables. The MMDD(actionDate) tbpartitions 10 function calculates DAY_OF_YEAR % 10.

```
CREATE TABLE user_log5(
userId int,
name varchar(30),
operation varchar(30),
actionDate DATE
) dbpartition by hash(userId) tbpartition by MMDD(actionDate) tbpartitions 10;
```

To view the node topology of the logical table, execute the `show topology from user_log5;` statement. Each year has 365 days, and the table data is routed to 10 physical tables. The node topology shows that 10 table shards are created in each database shard. An ellipsis (...) is used toomit some data because the returned result is long.

```
+------+-------------------------------------------------------------------+--------------+
| ID   | GROUP_NAME                                                        | TABLE_NAME   |
+------+-------------------------------------------------------------------+--------------+
|    0 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_00 |
|    1 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_01 |
|    2 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_02 |
|    3 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_03 |
|    4 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_04 |
|    5 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_05 |
|    6 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_06 |
|    7 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_07 |
|    8 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_08 |
|    9 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | user_log5_09 |
...
|   70 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_00 |
|   71 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_01 |
|   72 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_02 |
|   73 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_03 |
|   74 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_04 |
|   75 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_05 |
|   76 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_06 |
|   77 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_07 |
|   78 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_08 |
|   79 | SANGUAN_TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | user_log5_09 |
+------+-------------------------------------------------------------------+--------------+
80 rows in set (0.02 sec)
```

To view the sharding rule of the logical table, execute the `show rule from user_log5;` statement. The returned result shows that hashing is implemented for database sharding, and the shard key for database sharding is userId. For table sharding, the date function MMDD is used, the table data is routed to 10 physical tables, and the shard key for table sharding is actionDate.

```
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+---------------------+--------------------+
| ID   | TABLE_NAME | BROADCAST | DB_PARTITION_KEY | DB_PARTITION_POLICY | DB_PARTITION_COUNT | TB_
PARTITION_KEY | TB_PARTITION_POLICY | TB_PARTITION_COUNT |
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+---------------------+--------------------+
|    0 | user_log5  |         0 | userId          | hash                | 8                  | act
ionDate       | mmdd                | 10                 |
+------+------------+-----------+-----------------+---------------------+--------------------+----
-------------+---------------------+--------------------+
1 row in set (0.01 sec)
```

12.1.12.2.5. Use the primary key as the shard key

When no shard key is specified in the sharding algorithm, the system uses the primary key as the shard key by default. The following examples illustrate how to use the primary key as the database shard key and the table shard key.

Use the primary key as the database shard key

```
CREATE TABLE prmkey_tbl(
id int,
name varchar(30),
primary key(id)
) dbpartition by hash();
```

Use the primary key as the database shard key and the table shard key

```
CREATE TABLE prmkey_multi_tbl(
id int,
name varchar(30),
primary key(id)
) dbpartition by hash() tbpartition by hash() tbpartitions 3;
```

12.1.12.2.6. Create a broadcast table

This topic describes the statement for creating a broadcast table and provides examples.

The BROADCAST clause is used to specify a broadcast table to be created. A broadcast table is a table that is replicated to each database shard, and a synchronization mechanism is used to ensure data consistency between the database shards. Data synchronization between the database shards has a latency of several seconds. This feature allows you to route a join from a PolarDB-X instance to an underlying ApsaraDB RDS for MySQL instance to avoid joins across multiple databases. For more information about how to optimize SQL statements by using broadcast tables, see Overview.

The following statement is an example of creating a broadcast table:

```
CREATE TABLE brd_tbl(
  id int,
  name varchar(30),
  primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 BROADCAST;
```

12.1.12.2.7. Other attributes of the MySQL CREATE TABLE statement

This topic describes other attributes of the MySQL CREATE TABLE statement.

When you create a logical table whose data is partitioned into table shards in database shards, you can also specify other attributes in the MySQL CREATE TABLE statement. For example, you can specify the following attributes:

```
CREATE TABLE multi_db_multi_tbl(
  id int,
  name varchar(30),
  primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by hash(id) tbpartition by hash(id) tbpartitions 3;
```

12.1.12.3. Modify a table

This topic describes the syntax of the DDL statement ALTER TABLE and provides examples.

The following syntax of the ALTER TABLE statement is used to modify a table:

```
ALTER [ONLINE|OFFLINE] [IGNORE] TABLE tbl_name
    [alter_specification [, alter_specification] ...]
    [partition_options]
```

On a instance, you can use this DDL statement to perform routine DDL operations, such as adding a column, adding an index, and modifying a data definition. For more information about the syntax, see MySQL ALTER TABLE Statement.

> **Note** If you need to modify a partitioned table, you are not allowed to modify the shard key.

- Add a column:

```
ALTER TABLE user_log
  ADD COLUMN idcard varchar(30);
```

- Add an index:

```
ALTER TABLE user_log
  ADD INDEX idcard_idx (idcard);
```

- Delete an index:

```
ALTER TABLE user_log
  DROP INDEX idcard_idx;
```

- Modify a field:

```
ALTER TABLE user_log
  MODIFY COLUMN idcard varchar(40);
```

12.1.12.4. Delete a table

This topic describes the syntax of the DROP TABLE statement and provides examples.

The following syntax of the DROP TABLE statement is used to delete a table:

```
DROP [TEMPORARY] TABLE [IF EXISTS]
    tbl_name [, tbl_name] ...
    [RESTRICT | CASCADE]
```

The DROP TABLE statement executed on a logical database is the same as that executed on a MySQL database in terms of syntax. After the statement is executed, the system automatically deletes the corresponding physical table shard. For more information, see MySQL DROP TABLE Statement.

For example, you can execute the following statement to delete the user_log table:

```
DROP TABLE user_log;
```

12.1.12.5. FAQ about DDL statements

This topic provides answers to some frequently asked questions about DDL statements.

What can I do if an error occurs when I create a table?

processes DDL statements in a distributed manner. If an error occurs, the schemas of all table shards are inconsistent with one another. Therefore, you need to perform manual cleanup.

Perform the following operations:

1. Check the basic error descriptions provided by , such as syntax errors. If the error message is too long, the system will prompt you to execute the SHOW WARNINGS statement to view the failure cause of each database shard.
2. Execute the SHOW TOPOLOGY statement to view the topology of physical table shards. For example, after you execute the `SHOW TOPOLOGY FROM multi_db_multi_tbl;` statement, the following output is displayed:

```
+------+-----------------+-----------------------+
| ID   | GROUP_NAME      | TABLE_NAME            |
+------+-----------------+-----------------------+
|    0 | corona_qatest_0 | multi_db_multi_tbl_00 |
|    1 | corona_qatest_0 | multi_db_multi_tbl_01 |
|    2 | corona_qatest_0 | multi_db_multi_tbl_02 |
|    3 | corona_qatest_1 | multi_db_multi_tbl_03 |
|    4 | corona_qatest_1 | multi_db_multi_tbl_04 |
|    5 | corona_qatest_1 | multi_db_multi_tbl_05 |
|    6 | corona_qatest_2 | multi_db_multi_tbl_06 |
|    7 | corona_qatest_2 | multi_db_multi_tbl_07 |
|    8 | corona_qatest_2 | multi_db_multi_tbl_08 |
|    9 | corona_qatest_3 | multi_db_multi_tbl_09 |
|   10 | corona_qatest_3 | multi_db_multi_tbl_10 |
|   11 | corona_qatest_3 | multi_db_multi_tbl_11 |
+------+-----------------+-----------------------+
12 rows in set (0.21 sec)
```

3. Run the `check table multi_db_multi_tbl;` command to check whether the logical table is created. For example, the following output indicates that a physical table shard corresponding to the logical table multi_db_multi_tbl failed to be created.

```
+-------------------------------------------------+-------+----------+-----------------------------
-----------------------------------------------+
| TABLE                                           | OP    | MSG_TYPE | MSG_TEXT
|
+-------------------------------------------------+-------+----------+-----------------------------
-----------------------------------------------+
| andor_mysql_qatest. multi_db_multi_tbl | check | Error    | Table 'corona_qatest_0. multi_db_mul
ti_tbl_02' doesn't exist |
+-------------------------------------------------+-------+----------+-----------------------------
-----------------------------------------------+
1 row in set (0.16 sec)
```

4. Continue to create or delete the logical table in idempotent mode. This way, the remaining physical table shards will be created or deleted.

```
CREATE TABLE IF NOT EXISTS table1
(id int, name varchar(30), primary key(id))
dbpartition by hash(id);
DROP TABLE IF EXISTS table1;
```

What can I do if I failed to create an index or add a column?

The method for handling the failure in creating an index or adding a column is similar to that for the failure in creating a table. For more information, see Handle DDL exceptions.

12.1.12.6. DDL functions for sharding

12.1.12.6.1. Overview

is a database service that supports both database sharding and table sharding.

Support for database sharding and table sharding

The following table lists the support for database sharding and table sharding by DDL sharding functions.

Sharding function	Description	Support for database sharding	Support for table sharding
HASH	Performs a simple modulo operation.	Yes	Yes
UNI_HASH	Performs a simple modulo operation.	Yes	Yes
RIGHT_SHIFT	Performs a signed right shift on the value of the database shard key.	Yes	Yes
RANGE_HASH	Performs hashing when two sharding keys are required.	Yes	Yes
MM	Performs hashing by month.	No	Yes
DD	Performs hashing by date.	No	Yes
WEEK	Performs hashing by week.	No	Yes
MMDD	Performs hashing by month and date.	No	Yes
YYYYMM	Performs hashing by year and month.	Yes	Yes
YYYYWEEK	Performs hashing by year and week.	Yes	Yes
YYYYDD	Performs hashing by year and date.	Yes	Yes
YYYYMM_OPT	Performs optimized hashing by year and month.	Yes	Yes
YYYYWEEK_OPT	Performs optimized hashing by year and week.	Yes	Yes
YYYYDD_OPT	Performs optimized hashing by year and date.	Yes	Yes

Note When you use database sharding and table sharding in , take note of the following points:

- In a instance, the sharding method of a logical table is jointly defined by a sharding function and a shard key. The sharding function specifies the number of shards to be created and the routing algorithm.
- In a instance, database sharding and table sharding use the same sharding method only when they use the same sharding function and the same shard key. This allows the instance to uniquely locate a physical table in a physical database based on the value of the shard key.
- Assume that the database sharding method and the table sharding method of a logical table are different. If the database shard key and table shard key are not specified in an SQL query, the instance scans all database shards or all table shards when it processes the SQL query.

Support for data types in PolarDB-X DDL sharding functions

Different DDL sharding functions support different data types. The following table lists the support for data types in sharding functions. The check mark (√) indicates supported whereas the cross mark (×) indicates not supported.

Support for data types in DDL sharding functions

Sharding function	BIGINT	INT	MEDIUMINT	SMALLINT	TINYINT	VARCHAR	CHAR	DATE	DATETIME	TIMESTAMP	Other types
HASH	√	√	√	√	√	√	√	×	×	×	×
UNI_HASH	√	√	√	√	√	√	√	×	×	×	×
RANGE_HASH	√	√	√	√	√	√	√	×	×	×	×
RIGHT_SHIFT	√	√	√	√	√	×	×	×	×	×	×
MM	×	×	×	×	×	×	×	√	√	√	×
DD	×	×	×	×	×	×	×	√	√	√	×
WEEK	×	×	×	×	×	×	×	√	√	√	×
MMDD	×	×	×	×	×	×	×	√	√	√	×
YYYYMM	×	×	×	×	×	×	×	√	√	√	×
YYYYWEEK	×	×	×	×	×	×	×	√	√	√	×
YYYYDD	×	×	×	×	×	×	×	√	√	√	×
YYYYMM_OPT	×	×	×	×	×	×	×	√	√	√	×
YYYYWEEK_OPT	×	×	×	×	×	×	×	√	√	√	×
YYYYDD_OPT	×	×	×	×	×	×	×	√	√	√	×

Syntax description for PolarDB-X DDL sharding functions

is compatible with the DDL statements in MySQL, and provides the `drds_partition_options` keyword to support database sharding and table sharding, as shown in the following example:

```
CREATE [TEMPORARY] TABLE [IF NOT EXISTS] tbl_name
    (create_definition,...)
    [table_options]
    [drds_partition_options]
    [partition_options]
CREATE [TEMPORARY] TABLE [IF NOT EXISTS] tbl_name
    [(create_definition,...)]
    [table_options]
    [drds_partition_options]
    [partition_options]
    select_statement
drds_partition_options:
    DBPARTITION BY
        { {HASH|YYYYMM|YYYYWEEK|YYYYDD|YYYYMM_OPT|YYYYWEEK_OPT|YYYYDD_OPT}([column])}
        [TBPARTITION BY
            { {HASH|MM|DD|WEEK|MMDD|YYYYMM|YYYYWEEK|YYYYDD|YYYYMM_OPT|YYYYWEEK_OPT|YYYYDD_OPT}(column)
}
          [TBPARTITIONS num]
        ]
```

12.1.12.6.2. HASH

This topic describes how to use the HASH function.

Usage notes

- The data type of shard keys must be integer or string.

- This sharding function has no limits on the version of your instance. By default, it supports all PolarDB-X instance versions.

Routing method

If different shard keys are used to execute the HASH function for database sharding and table sharding, divide a value of the database shard key by the number of database shards and find the remainder. If the key value is a string, the string is first converted into a hash value and then used for route calculation. For example, HASH('8') is equivalent to 8 % D, where D indicates the number of database shards.

If the same shard key is used to execute the HASH function for database sharding and table sharding, divide a value of the shard key by the total number of table shards and find the remainder. Assume that two database shards are created, each database shard has four table shards, table shards 0 to 3 are stored in database shard 0, and table shards 4 to 7 are stored in database shard 1. Based on this routing method, shard key value 15 is distributed to table shard 7 in database shard 1. The equation is 15 % (2 × 4) = 7.

Scenarios

- The function can be used to partition data from a logical table into database shards by user ID or order ID.
- Values of shard keys are strings.

Examples

Assume that you need to partition data from a logical table into database shards by executing the HASH function based on the ID column. Execute the following DDL statement to create a table:

```
create table test_hash_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by HASH(ID);
```

Considerations

The HASH function is a simple modulus operation. The output of the HASH function can be evenly distributed only when the values in the shard key columns are evenly distributed.

12.1.12.6.3. UNI_HASH

This topic describes how to use the UNI_HASH function.

Usage notes

- The data type of shard keys must be integer or string.
- The version of your instance must be 5.1.28-1508068 or later. For more information about the release notes of , see View the instance version.

Routing method

If you execute the UNI_HASH function for database sharding, divide a value of the database shard key by the total number of database shards and find the remainder. If the key value is a string, the string is first converted into a hash value and then used for route calculation. For example, HASH('8') is equivalent to 8 % D, where D indicates the number of database shards.

If the same shard key is used to execute the UNI_HASH function for database sharding and table sharding, divide a value of the database shard key by the number of database shards first. (This step is different from that in the HASH function.) Then, evenly distribute data to the table shards in the database shards.

Scenarios

- The function can be used to partition data from a logical table into database shards by user ID or order ID.
- The data type of shard keys is integer or string.
- You can use this function when you need to partition data from two logical tables into database shards based on the same shard key. In each database shard, the number of table shards for one logical table is different from that for the other logical table. You need to frequently perform joins on the two logical tables based on the shard key.

Comparison with the HASH function

If you execute the UNI_HASH function to partition data from a logical table into database shards but not further into table shards, the routing method is the same as that used in the HASH function. The route is calculated by dividing the values of database shard keys by the number of database shards.

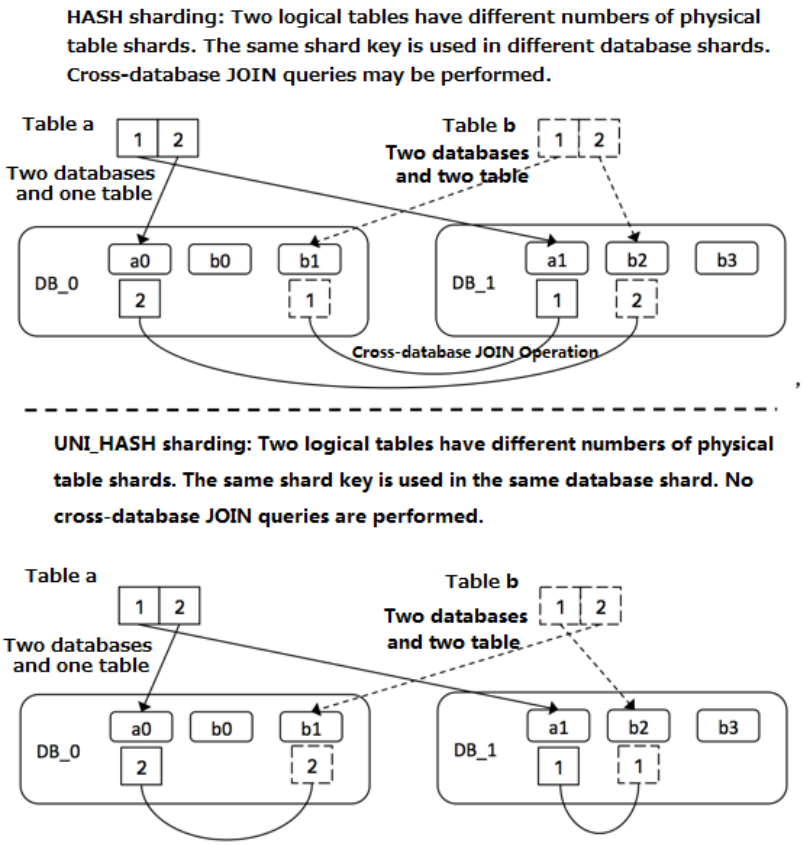
If you use the same shard key to execute the HASH function for database sharding and table sharding, the same key value may be routed to different database shards as the number of table shards changes.

If you use the same shard key to execute the UNI_HASH function for database sharding and table sharding, the same key value is always routed to the same database shard regardless of the number of table shards.

Assume that you execute the HASH or UNI_HASH function to partition data from two logical tables into table shards in database shards based on the same shard key and the total number of table shards for the two logical tables are different. If you perform a join on the two logical tables based on the shard key, the join is performed across database shards when the HASH function is used but in the same database shard when the UNI_HASH function is used.

Assume that you have two logical tables a and b. Each logical table has two database shards. Each database shard corresponding to logical table a stores one table shard, and each database shard corresponding to logical table b stores two table shards. The following figures separately show the results of performing a join on logical tables a and b after the HASH and UNI_HASH functions are used to partition data in the two logical tables.

Comparison between HASH and UNI_HASH



Examples

Assume that you need to partition data from a logical table into four table shards in each database shard by executing the UNI_HASH function based on the ID column. Execute the following CREATE TABLE statement:

```
create table test_hash_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8
dbpartition by UNI_HASH(ID)
tbpartition by UNI_HASH(ID) tbpartitions 4;
```

Considerations

The UNI_HASH function is a simple modulus operation. The output of the HASH function can be evenly distributed only when the values in the shard key columns are evenly distributed.

12.1.12.6.4. RIGHT_SHIFT

This topic describes how to use the RIGHT_SHIFT function for sharding.

Usage notes

- A shard key must be an integer.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

The RIGHT_SHIFT function shifts the value of a database shard key to the right by a specified number of binary digits. The value of a database shard key must be an integer. Then, the function divides the obtained integer by the number of database shards or table shards and returns the remainder. You can specify the number of shifted digits in a data definition language (DDL) statement.

Scenarios

The RIGHT_SHIFT function can implement uniform hashing when the lower-part digits of most shard key values are similar but the higher-part digits significantly vary.

In this example, four shard key values are available: 12340000, 12350000, 12460000, and 12330000. The four lower-part digits of the four values are all 0000. If you enable hashing for the original shard key values, hashing is ineffective. However, if you use the RIGHT_SHIFT(shardKey, 4) function to shift the values of the shard keys to the right by four digits, the values change to 1234, 1235, 1246, and 1233. This improves hashing performance.

Use cases

You can execute the following statement to create a table. In the statement, the id column is used as a shard key and the id column values are shifted to the right by four binary digits. Hashing is implemented based on the obtained values.

```
create table test_hash_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8
dbpartition by RIGHT_SHIFT(id, 4)
tbpartition by RIGHT_SHIFT(id, 4) tbpartitions 2;
```

Notes

The number of digits to shift cannot exceed the number of digits that the integer contains.

12.1.12.6.5. RANGE_HASH

This topic describes how to use the RANGE_HASH function.

Usage notes

- The data type of shard keys must be character or number.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Calculate the hash value based on the last N digits of the value of a shard key. Then, divide the hash value by the number of database shards and find the remainder. The number N is the third parameter in the function.

For example, during the calculation of the RANGE_HASH(COL1, COL2, N) function, COL1 is preferentially selected and then truncated to obtain the last N digits for calculation. If COL1 does not exist, COL2 is selected and truncated for calculation.

Scenarios

The RANGE_HASH function can be used in scenarios where a table needs to be partitioned by two shard keys but queries are performed based on the values of only one shard key.

Examples

Assume that data in a logical database is partitioned into eight database shards and that you have the following requirements:

Partition data into database shards by buyer ID and order ID. However, values of only one ID are available for queries.

You can execute the following DDL statement to create an order table:

```
create table test_order_tb (
    id int,
    buyer_id varchar(30) DEFAULT NULL,
    order_id varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
    ) ENGINE=InnoDB DEFAULT CHARSET=utf8
    dbpartition by RANGE_HASH(buyer_id,order_id, 10)
    tbpartition by RANGE_HASH (buyer_id,order_id, 10) tbpartitions 3;
```

Considerations

- The two shard keys cannot be modified.
- Data insertion fails if data is routed to different database shards or table shards based on the two shard keys.

12.1.12.6.6. MM

This topic describes how to use the MM function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The MM function can be used only for table sharding, but cannot be used for database sharding.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Divide the month in a time value of the database shard key by the total number of table shards, and find the remainder. The remainder is the table shard subscript.

Scenarios

The MM function can be used to partition data into table shards by month. The table shard name indicates a specific month.

Examples

Assume that you need to partition data first into database shards based on the ID column and then into table shards based on the create_time column by month. Meanwhile, you need to map each month to a physical table shard. Execute the following DDL statement to create a table:

```
create table test_mm_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8
dbpartition by HASH(id)
tbpartition by MM(create_time) tbpartitions 12;
```

Considerations

When you partition data into table shards by executing the MM function, make sure that each database shard has no more than 12 table shards because a year has 12 months.

12.1.12.6.7. DD

This topic describes how to use the DD function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The DD function can be used only for table sharding, but cannot be used for database sharding.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Divide the day of the month in a time value of the database shard key by the total number of table shards, and find the remainder. The remainder is the table shard subscript.

Scenarios

The DD function can be used to partition data into table shards based on date, which is the day of a month. The subscript of the table shard name indicates the day of a month. A month has 31 days at most.

Examples

Assume that you need to partition data first into database shards based on the ID column and then into table shards based on the create_time column by day of a month. Meanwhile, you need to map each day of the month to a physical table shard. Execute the following DDL statement to create a table:

```
create table test_dd_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8
dbpartition by HASH(id)
tbpartition by DD(create_time) tbpartitions 31;
```

Considerations

When you partition data into table shards by executing the DD function, make sure that each database shard has no more than 31 table shards because a month has 31 days at most.

12.1.12.6.8. WEEK

This document describes how to use the WEEK function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The WEEK function can be used only for table sharding, but cannot be used for database sharding.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Divide the day of the week in a time value of the database shard key by the total number of table shards, and find the remainder. The remainder is the table shard subscript.

Scenarios

The WEEK function can be used to partition data into table shards by day of a week. The subscript of the table shard name corresponds to each day of a week, from Monday to Sunday.

Examples

Assume that you need to partition data first into database shards based on the ID column and then into table shards based on the create_time column by day of a week. Meanwhile, you need to map each day of a week (from Monday to Sunday) to a physical table shard. Execute the following DDL statement to create a table:

```
create table test_week_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8
dbpartition by HASH(name)
tbpartition by WEEK(create_time) tbpartitions 7;
```

Considerations

When you partition data into table shards by executing the WEEK function, make sure that each database shard has no more than seven table shards because a week has seven days.

12.1.12.6.9. MMDD

This topic describes how to use the MMDD function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The MMDD function can be used only for table sharding, but cannot be used for database sharding.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Divide the day of the year in a time value of the database shard key by the total number of table shards, and find the remainder. The remainder is the table shard subscript.

Scenarios

The MMDD function can be used to partition data into table shards by day of a year. The subscript of the table shard name indicates the day of the year. A year has 366 days at most.

Examples

Assume that you need to partition data first into database shards based on the ID column and then into table shards based on the create_time column by day of a year. Meanwhile, you need to map each day of a year to a physical table shard. Execute the following data definition language (DDL) statement to create a table:

```
create table test_mmdd_tb (
    id int,
    name varchar(30) DEFAULT NULL,
    create_time datetime DEFAULT NULL,
    primary key(id)
) ENGINE=InnoDB DEFAULT CHARSET=utf8
dbpartition by HASH(name)
tbpartition by MMDD(create_time) tbpartitions 365;
```

Considerations

When you partition data into table shards by executing the MMDD function, make sure that each database shard has no more than 366 table shards because a year has 366 days at most.

12.1.12.6.10. YYYYMM

This topic describes how to use the YYYYMM function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Calculate the hash value based on the year and month in a time value of the database shard key. Then, divide the hash value by the number of database shards and find the remainder.

For example, the YYYYMM('2012-12-31 12:12:12') function is equivalent to (2012 × 12 + 12) % D, where D indicates the number of database shards.

Scenarios

The YYYYMM function can be used to partition data into database shards by year and month. We recommend that you use the YYYYMM function with the tbpartition YYYYMM(ShardKey) function.

Assume that data in a logical database is partitioned into eight physical database shards and that you have the following requirements:

- Partition data into the database shards by year and month.
- Distribute data from the same month to the same table shard and ensure that each month within two years corresponds to a separate table shard.
- Directly query data from a physical table shard in a physical database shard when database shard keys and table shard keys are specified in the query.

Then, you can use the YYYYMM function. In the case that each month within two years corresponds to a physical table shard, a total of 24 physical table shards must be created, because a year has 12 months. Data in the logical database is partitioned into eight database shards. Therefore, three physical table shards must be created in each database shard. Execute the following DDL statement to create a table:

```
create table test_yyyymm_tb (
        id int,
        name varchar(30) DEFAULT NULL,
        create_time datetime DEFAULT NULL,
        primary key(id)
    ) ENGINE=InnoDB DEFAULT CHARSET=utf8
    dbpartition by YYYYMM(create_time)
    tbpartition by YYYYMM(create_time) tbpartitions 3;
```

Considerations

- You cannot use the YYYYMM function to distribute data from every month in every year to a separate table shard. Instead, the number of table shards must be fixed.
- After a cycle over months, data from the same month may be routed to the same table shard in the same database shard based on the number of table shards. For example, a cycle is from 2012-03 to 2013-03.

12.1.12.6.11. YYYYWEEK

This topic describes how to use the YYYYWEEK function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Calculate the hash value based on the year and week of the year in a time value of the database shard key. Then, divide the hash value by the number of database shards and find the remainder.

For example, the YYYYWEEK('2012-12-31 12:12:12') function is equivalent to (2013 × 52 + 1) % D, where D indicates the number of database shards. Therefore, the date 2012-12-31 falls on the first week of 2013.

Scenarios

The YYYYWEEK function can be used to partition data into database shards by year and week of the year. We recommend that you use the YYYYWEEK function with the tbpartition YYYYWEEK(ShardKey) function.

Assume that data in a logical database is partitioned into eight physical database shards and that you have the following requirements:

- Partition data into the database shards by year and week of the year.
- Distribute data from the same week to the same table shard and ensure that each week within two years corresponds to a separate table shard.
- Directly query data from a physical table shard in a physical database shard when database shard keys and table shard keys are specified in the query.

Then, you can use the YYYYWEEK function. In the case that each week within two years corresponds to a physical table shard, at least 106 physical table shards must be created, because a year has up to 53 weeks. Data in the logical database is partitioned into eight database shards, and therefore 14 physical table shards must be created in each database shard. The total number of physical database shards is calculated in the following formula: 14 × 8 = 112, where 112 is greater than 106. We recommend that the number of physical table shards be an integer multiple of the number of database shards. Execute the following DDL statement to create a table:

```
create table test_yyyymm_tb (
        id int,
        name varchar(30) DEFAULT NULL,
        create_time datetime DEFAULT NULL,
        primary key(id)
    ) ENGINE=InnoDB DEFAULT CHARSET=utf8
    dbpartition by YYYYWEEK(create_time)
    tbpartition by YYYYWEEK(create_time) tbpartitions 14;
```

Considerations

- You cannot use the YYYYWEEK function to distribute data from every week in every year to a separate table shard. Instead, the number of table shards must be fixed.
- After a cycle over weeks, data from the same week may be routed to the same table shard in the same

database shard based on the number of table shards. For example, a cycle is from the first week in 2012 to the first week in 2013.

12.1.12.6.12. YYYYDD

This topic describes how to use the YYYYDD function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Routing method

Calculate the hash value based on the year and day of the year in a time value of the database shard key. Then, divide the hash value by the number of database shards and find the remainder.

For example, the YYYYDD('2012-12-31 12:12:12') function is equivalent to (2012 × 366 + 365) % D, where D indicates the number of database shards. Therefore, the value 2012-12-31 is the 365th day of 2012.

Scenarios

The YYYYDD function can be used to partition data into database shards by year and day of the year. We recommend that you use the YYYYDD function with the tbpartition YYYYDD(ShardKey) function.

Assume that data in a logical database is partitioned into eight physical database shards and that you have the following requirements:

- Partition data into the database shards by year and day of the year.
- Distribute data from the same week to the same table shard and ensure that each day within two years corresponds to a separate table shard.
- Directly query data from a physical table shard in a physical database shard when database shard keys and table shard keys are specified in the query.

Then, you can use the YYYYDD function. In the case that each day within two years corresponds to a physical table shard, a total of 732 physical table shards must be created, because a year has up to 366 days. Data in the logical database is partitioned into eight database shards, and therefore 92 physical table shards must be created in each database shard. This number is calculated in the following formula: 732/8 = 91.5, which is rounded to 92. We recommend that the number of table shards be an integer multiple of the number of database shards. Execute the following DDL statement to create a table:

```
create table test_yyyydd_tb (
        id int,
        name varchar(30) DEFAULT NULL,
        create_time datetime DEFAULT NULL,
        primary key(id)
    ) ENGINE=InnoDB DEFAULT CHARSET=utf8
    dbpartition by YYYYDD(create_time)
    tbpartition by YYYYDD(create_time) tbpartitions 92;
```

Considerations

- You cannot use the YYYYDD function to distribute data from every day in every year to a separate table shard. Instead, the number of table shards must be fixed.
- After a cycle over dates, data from the same date may be routed to the same table shard in the same database shard based on the number of table shards. For example, a cycle is from 2012-03-01 to 2013-03-01.

12.1.12.6.13. YYYYMM_OPT

This topic describes how to use the YYYYMM_OPT function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The year and month of user data does not randomly increase but naturally increase over time.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

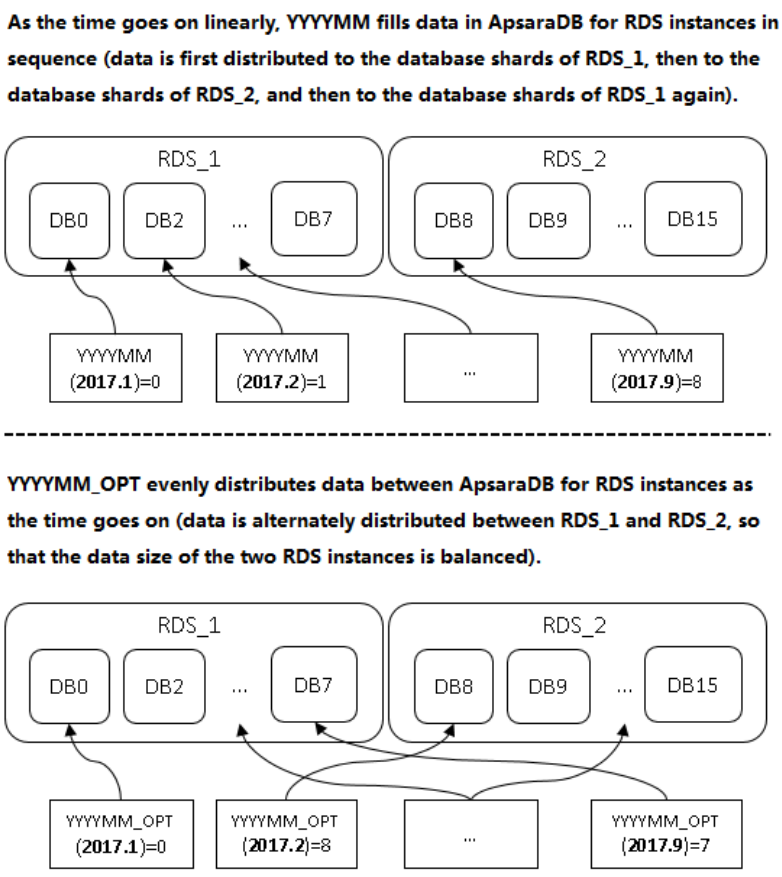
Optimizations

Compared with the YYYYMM function, the YYYYMM_OPT function can ensure that data is evenly distributed across ApsaraDB RDS for MySQL instances as the timeline increases.

Assume that two ApsaraDB RDS for MySQL instances are attached to your instance and that data in your logical table is partitioned into 16 database shards. DB0 to DB7 shards are located on one ApsaraDB RDS for MySQL instance, and DB8 to DB15 shards are located on the other ApsaraDB RDS for MySQL instance.

The following figures compare the mappings when the YYYYMM and YYYYMM_OPT functions are used for database sharding.

Comparison between the YYYYMM and YYYYMM_OPT functions



- The YYYYMM_OPT function evenly distributes data across ApsaraDB RDS for MySQL instances. This maximizes the performance of each ApsaraDB RDS for MySQL instance.
- Choose between the YYYYMM and YYYYMM_OPT functions:
  - If the time of business data sequentially increases and the data volume does not differ much between time points, you can use the YYYYMM_OPT function to evenly distribute data across ApsaraDB RDS for MySQL instances.

- You can use the YYYYMM function if the time of business data randomly increases.

Routing method

- Calculate the hash value based on the year and month of the year in a time value of the database shard key. Then, divide the hash value by the number of database shards and find the remainder.
- The hash calculation based on the database and table shard key considers the evenness of data distribution among the ApsaraDB RDS for MySQL instances attached to your instance.

Scenarios

- Data needs to be partitioned into table shards in database shards by year and month of the year.
- Data must be evenly distributed across the ApsaraDB RDS for MySQL instances attached to your instance.
- The time of the shard key sequentially increases, and data is evenly distributed across all ApsaraDB RDS for MySQL instances from month to month. For example, the number of monthly journal logs increases every month, but the logs are evenly distributed across all ApsaraDB RDS for MySQL instances.

Considerations

- The YYYYMM_OPT function does not allow you to distribute data from every month in every year to a separate table shard. Instead, the number of table shards must be fixed.
- After a cycle over months, data from the same month may be routed to the same table shard in the same database shard based on the number of table shards. For example, a cycle is from 2012-03 to 2013-03.

12.1.12.6.14. YYYYWEEK_OPT

This topic describes how to use the YYYYWEEK_OPT function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Optimizations

- Compared with the YYYYWEEK function, the YYYYWEEK_OPT function can ensure that data is evenly distributed across ApsaraDB RDS for MySQL instances as the timeline increases. The YYYYWEEK_OPT function has the similar effect to the YYYYMM_OPT function.
- The YYYYWEEK_OPT function evenly distributes data across ApsaraDB RDS for MySQL instances. This maximizes the performance of each ApsaraDB RDS for MySQL instance.
- Choose between the YYYYWEEK and YYYYWEEK_OPT functions:
  - If the time of business data sequentially increases and the data volume does not differ much between time points, you can use the YYYYWEEK_OPT function to evenly distribute data across ApsaraDB RDS for MySQL instances.
  - You can use the YYYYWEEK function if the time of business data randomly increases.

Routing method

- Calculate the hash value based on the year and week of the year in a time value of the database shard key. Then, divide the hash value by the number of database shards and find the remainder.
- The hash calculation based on the database and table shard key considers the evenness of data distribution among the ApsaraDB RDS for MySQL instances attached to your instance.

Scenarios

- Data needs to be partitioned into table shards in database shards by year and week of the year.

- Data must be evenly distributed across the ApsaraDB RDS for MySQL instances attached to your instance.
- The time of the shard key sequentially increases, and data is evenly distributed across all ApsaraDB RDS for MySQL instances from week to week. For example, the number of weekly journal logs increases every week, but the logs are evenly distributed across all ApsaraDB RDS for MySQL instances.

Considerations

- The YYYYWEEK_OPT function does not allow you to distribute data from every week in every year to a separate table shard. Instead, the number of table shards must be fixed.
- After a cycle over weeks, data from the same week may be routed to the same table shard in the same database shard based on the number of table shards. For example, a cycle is from the first week in 2012 to the first week in 2013.

12.1.12.6.15. YYYYDD_OPT

This topic describes how to use the YYYYDD_OPT function.

Usage notes

- The data type of shard keys must be DATE, DATETIME, or TIMESTAMP.
- The version of your instance must be 5.1.28-1320920 or later. For more information, see View the instance version.

Optimizations

- Compared with YYYYDD function, the YYYYDD_OPT function can ensure that data is evenly distributed across ApsaraDB RDS for MySQL instances as the timeline increases. The YYYYDD_OPT function has the similar effect to the YYYYMM_OPT function.
- The YYYYDD_OPT function evenly distributes data across ApsaraDB RDS for MySQL instances. This maximizes the performance of each ApsaraDB RDS for MySQL instance.
- Choose between the YYYYDD and YYYYDD_OPT functions:
  - If the time of business data sequentially increases and the data volume does not differ much between time points, you can use the YYYYDD_OPT function to evenly distribute data across ApsaraDB RDS for MySQL instances.
  - You can use the YYYYDD function if the time of business data randomly increases.

Routing method

- Calculate the hash value based on the year and day of the year in a time value of the database shard key. Then, divide the hash value by the number of database shards and find the remainder.
- The hash calculation based on the database and table shard key considers the evenness of data distribution among the ApsaraDB RDS for MySQL instances attached to your instance.

Scenarios

- Data needs to be partitioned into table shards in database shards by year and day of the year.
- Data must be evenly distributed across the ApsaraDB RDS for MySQL instances attached to your instance.
- The time of the shard key sequentially increases, and data is evenly distributed across all ApsaraDB RDS for MySQL instances from day to day. For example, the number of daily journal logs increases every day, but the logs are evenly distributed across all ApsaraDB RDS for MySQL instances.

Considerations

- The YYYYDD_OPT function does not allow you to distribute data from every month in every year to a separate table shard. Instead, the number of table shards must be fixed.
- After a cycle over dates, data from the same date may be routed to the same table shard in the same database shard based on the number of table shards. For example, a cycle is from 2012-03-01 to 2013-03-01.

12.1.13. Automatic protection of high-risk SQL statements

In , the data manipulation language (DML) statements are the same as MySQL statements.

We recommend that you include the shard key in the **SELECT** and **UPDATE** statements of . The **INSERT** statement of must include the shard key and a non-null key value.

By default, disables full-table deletion and updating to prevent misoperation.

The following statements are prohibited by default:

- A DELETE statement without the WHERE or LIMIT condition
- An UPDATE statement without the WHERE or LIMIT condition

If you need to perform full-table deletion or update, you can temporarily skip this limit by using the following hint:

```
HINT: /! TDDL:FORBID_EXECUTE_DML_ALL=false*/
```

Examples

- Full-table deletion is intercepted by default.

  ```
  mysql> delete from tt;
  ERR-CODE: [TDDL-4620][ERR_FORBID_EXECUTE_DML_ALL] Forbid execute DELETE ALL or UPDATE ALL sql. More
  : [http://middleware.alibaba-inc.com/faq/faqByFaqCode.html?faqCode=TDDL-4620]
  ```

  The operation is successful if the following HINT is added:

  ```
  mysql> /! TDDL:FORBID_EXECUTE_DML_ALL=false*/delete from tt;
  Query OK, 10 row affected (0.21 sec)
  ```

- Full-table update is intercepted by default.

  ```
  mysql> update tt set id = 1;
  ERR-CODE: [TDDL-4620][ERR_FORBID_EXECUTE_DML_ALL] Forbid execute DELETE ALL or UPDATE ALL sql. More
  : [http://middleware.alibaba-inc.com/faq/faqByFaqCode.html?faqCode=TDDL-4620]
  ```

  The operation is successful if the following HINT is added:

  ```
  mysql> /! TDDL:FORBID_EXECUTE_DML_ALL=false*/update tt set id = 1;
  Query OK, 10 row affected (0.21 sec)
  ```

- This limit does not apply to DELETE or UPDATE statements that contain the WHERE or LIMIT condition.

  ```
  mysql> delete from tt where id = 1;
  Query OK, 1 row affected (0.21 sec)
  ```

12.1.14. PolarDB-X sequence

12.1.14.1. Overview

A sequence is a 64-digit numeric sequence of the signed BIGINT type in MySQL and is short for sequence in the following description. It aims to create a globally unique and numeric sequence that is sequentially incremental. A sequence can be used to generate values, such as the values of primary key columns and unique index columns.

sequences are used in the following two methods:

- The explicit sequence, which is created and maintained by using the sequence-specific data definition language

(DDL) syntax and is available for independent use. The sequence value can be obtained by using `select seq.nextval;`. `seq` indicates the name of a sequence.

- The implicit sequence, which automatically fills primary keys after AUTO_INCREMENT is defined for the primary keys. The sequence of this type is automatically maintained by .

> **Notice** creates implicit sequences after AUTO_INCREMENT is specified only for sharded tables and broadcast tables. PolarDB-X does not create implicit sequences for tables that are not sharded. The AUTO_INCREMENT value of a table that is not sharded is automatically generated by ApsaraDB RDS for MySQL at the underlying layer.

Types and features of PolarDB-X sequences

PolarDB-X supports the following three types of sequences:

Type (abbreviation)	Globally unique	Consecutive	Monotonically increasing	Monotonically increasing in the same connection	No single point of failure	Data type	Readability
Group Sequence (GROUP)	Yes	No	No	Yes	Yes	All integer types	High
Time-based Sequence (TIME)	Yes	No	Monotonically increasing at the macro level and non-monotonically increasing at the micro level.	Yes	Yes	Only BIGINT is supported.	Low
Simple Sequence (SIMPLE)	Yes	Yes	Yes	Yes	No	All integer types	High

Concepts:

- Consecutive: If the current value is n, the next value must be n + 1. If the next value is not n + 1, it is nonconsecutive.
- Monotonically increasing: If the current value is n, the next value must be a number greater than n.
- Single point: The risk of a single point of failure exists.
- Monotonically increasing at the macro level and non-monotonically increasing at the micro level: An example of this is the `1, 3, 2, 4, 5, 7, 6, 8, ...` sequence. The sequence is monotonically increasing at the macro level and is not monotonically increasing at the micro level.

GROUP (default)

Features

A group sequence is a globally unique sequence that has natural numeric values. A group sequence does not need to be consecutive or monotonically increasing. If the sequence type is not specified, uses group sequences by default.

- Advantages: A group sequence is globally unique and prevents single points of failure. A group sequence

provides excellent performance.

- Disadvantages: A group sequence may contain nonconsecutive values and does not need to start from the initial value. The values of a group sequence cannot be cyclical.

Implementation mechanism

The values of a group sequence are created by multiple nodes to ensure high availability. The values in a segment are nonconsecutive if the values are not all used. For example, this situation may occur due to disconnection.

Time-based Sequence (TIME)

Features

A time-based sequence consists of a timestamp, node ID, and serial number. Such a sequence is globally unique and auto-increment at the macro level. The update of sequence values does not rely on databases. You only need to store sequence names and types in databases instead of storing sequence values in a persistent way. In this case, a time-based sequence delivers excellent performance. You can create sequence values such as `776668092129345536, 776668098018148352, 776668111578333184, and 776668114812141568` .

> **Notice** Sequence values must be of the BIGINT type when they are used in the auto-increment columns of tables.

- Advantages: A time-based sequence is globally unique and delivers excellent performance.
- Disadvantages: The values of a time-based sequence are nonconsecutive. The START WITH, INCREMENT BY, MAXVALUE, and CYCLE or NOCYCLE parameters are invalid for time-based sequences.

Simple Sequence (SIMPLE)

Features

Only simple sequences support the INCREMENT BY, MAXVALUE, and CYCLE or NOCYCLE parameters.

- Advantages: A simple sequence is globally unique and monotonically increasing. Such a sequence has consecutive values.
- Disadvantages: A simple sequence is prone to single points of failure, low performance, and bottlenecks. Use a simple sequence with caution.

Implementation mechanism

Each time a value is generated, the value is stored in a persistent way.

Scenarios

The three types of sequences are globally unique, and can be used in primary key columns and unique index columns.

- In most scenarios, we recommend that you use group sequences.
- Use only simple sequences for services that have high requirements for consecutive sequence values. Take note of the simple sequence performance.
- Assume that you have high requirements for sequence performance, the amount of data inserted to tables is small, and the sequence values are large. In this case, we recommend that you use a time-based sequence. A time-based sequence is compute-bound and does not reply on databases. Such a sequence does not require a lock in computing or persistent storage.

The following example is used to describe how to create a sequence that starts from 100000 and has a step size of 1.

- A simple sequence creates globally unique, consecutive, and monotonically increasing values, such as `100000, 100001, 100002, 100003, 100004, ..., 200000, 200001, 200002, 200003...` Simple sequences are stored in a persistent way. Even after services are restarted upon a single point of failure, values are still generated in a consecutive way from the breakpoint. However, simple sequences have low performance because each value is stored in a persistent way once it is created.

- A group sequence may create values such as `200001, 200002, 200003, 200004, 100001, 100002, 100003...`

  > **Notice**
  > - A group sequence does not need to start from the specified value. However, a group sequence must start from a value that is greater than or equal to the specified value. In this example, the specified value is 100000. In this example, the initial value of the group sequence is 200001.
  > - A group sequence is globally unique but may contain nonconsecutive values, for example, when a node fails or the connection that only uses partial values is closed. In this case, nonconsecutive values are generated. In this example, nonconsecutive values are generated because the values between 200004 and 100001 are missing.

- A time-based sequence may create values such as `776668092129345536, 776668098018148352, 776668111578333184, 776668114812141568...`

12.1.14.2. Use explicit sequences

This topic describes how to use data definition language (DDL) statements to create, modify, delete, and query sequences and how to obtain explicit sequence values.

Create a sequence

Syntax:

```
CREATE [ GROUP | SIMPLE | TIME ] SEQUENCE <name>
[ START WITH <numeric value> ] [ INCREMENT BY <numeric value> ]
[ MAXVALUE <numeric value> ] [ CYCLE | NOCYCLE ]
```

Parameter description:

Parameter	Description	Applicable scope
START WTIH	The initial value of a sequence. If you leave this parameter empty, the default value is 1.	Simple sequences and group sequences
INCREMENT BY	The increment between two adjacent sequence values, also called the interval value or step size. If you leave this parameter empty, the default value is 1.	Simple Sequence
MAXVALUE	The maximum value the sequence can generate. If you leave this parameter empty, the default value is the maximum value of the signed BIGINT type.	Simple Sequence
CYCLE or NOCYCLE	Specifies whether to repeat the sequence that starts from the value specified by START WITH after the sequence reaches the maximum value. Set the value to CYCLE or NOCYCLE. If you leave this parameter empty, the default value is NOCYCLE.	Simple Sequence

> **Note**
> - If the sequence type is not specified, a group sequence is used by default.
> - INCREMENT BY, MAXVALUE, CYCLE, and NOCYCLE are invalid for group sequences.
> - START WITH, INCREMENT BY, MAXVALUE, CYCLE, and NOCYCLE are invalid for time-based sequences.
> - The values of a group sequence are nonconsecutive. The START WITH parameter only provides reference for group sequences. A group sequence may not start from the value specified by START WITH. However, a group sequence must start from a value that is greater than the specified value.

Example 1: Use the following two methods to create a group sequence.

- Method 1: Execute the `CREATE SEQUENCE seq1;` statement. The following result is returned:

```
Query OK, 1 row affected (0.27 sec)
```

- Method 2: Execute the `CREATE GROUP SEQUENCE seq1;` statement. The following result is returned:

```
Query OK, 1 row affected (0.27 sec)
```

Example 2: Execute the `CREATE TIME SEQUENCE seq1;` statement to create a time-based sequence. The following result is returned:

```
Query OK, 1 row affected (0.27 sec)
```

Example 3: Execute the `CREATE SIMPLE SEQUENCE seq2 START WITH 1000 INCREMENT BY 2 MAXVALUE 99999999999 NOCYCLE;` statement to create a simple sequence. The sequence starts from 1000, and has a step size of 2 and a maximum value of 99999999999. The sequence does not continue to generate repeated values after it reaches the maximum value. The following result is returned:

```
Query OK, 1 row affected (0.03 sec)
```

Modify a sequence

allows you to modify sequences in the following ways:

- The START WITH, INCREMENT BY, MAXVALUE, and CYCLE or NOCYCLE parameters of a simple sequence.
- The START WITH parameter of a group sequence.
- The conversion between different sequence types.

Syntax:

```
ALTER SEQUENCE <name> [ CHANGE TO GROUP | SIMPLE | TIME ]
START WITH <numeric value> [ INCREMENT BY <numeric value> ]
[ MAXVALUE <numeric value> ] [ CYCLE | NOCYCLE ]
```

Parameter description:

Parameter	Description	Applicable scope
START WITH	The initial value of a sequence. If you leave this parameter empty, the default value is 1.	Simple sequences and group sequences
INCREMENT BY	The increment between two adjacent sequence values, also called the interval value or step size. If you leave this parameter empty, the default value is 1.	Simple Sequence

Parameter	Description	Applicable scope
MAXVALUE	The maximum value the sequence can generate. If you leave this parameter empty, the default value is the maximum value of the signed BIGINT type.	Simple Sequence
CYCLE or NOCYCLE	Specifies whether to repeat the sequence that starts from the value specified by START WITH after the sequence reaches the maximum value. Set the value to CYCLE or NOCYCLE. If you leave this parameter empty, the default value is NOCYCLE.	Simple Sequence

> **Note**
>
> - The values of a group sequence are nonconsecutive. The START WITH parameter only provides reference for group sequences. A group sequence may not start from the value specified by START WITH. However, a group sequence must start from a value that is greater than or equal to the specified value.
> - Assume that you have specified a START WITH value when you modify a sequence. In this case, the START WITH value takes effect after it is specified. The following sequence value starts from the new START WITH value. For example, if you change the START WITH value to 200 when the sequence value increases to 100, the following sequence value starts from 200.
> - Before you change the START WITH value, analyze the existing sequence values and the speed of generating sequence values to avoid conflicts. Exercise caution when you change the START WITH value.

Example: Execute the `ALTER SEQUENCE seq2 START WITH 3000 INCREMENT BY 5 MAXVALUE 1000000 CYCLE;` statement to modify the simple sequence seq2. Change the initial value to 3000, the step size to 5, and the maximum value to 1000000. The sequence continues to generate repeated values after it reaches the maximum value. The following result is returned:

```
Query OK, 1 row affected (0.01 sec)
```

Conversion between different sequence types

- Use the `CHANGE TO <sequence_type>` clause in the `ALTER SEQUENCE` statement to convert a sequence into another type.
- If you include the `CHANGE TO` clause in `ALTER SEQUENCE`, you must specify the `START WITH` parameter to avoid generating duplicate values. `CHANGE TO` is optional. If you omit this clause, you can leave START WITH empty.

Example: Execute the `ALTER SEQUENCE seq1 CHANGE TO SIMPLE START WITH 1000000;` statement to convert a group sequence into a simple sequence. The following result is returned:

```
Query OK, 1 row affected (0.02 sec)
```

Delete a sequence

Syntax:

```
DROP SEQUENCE <name>
```

Example: Execute the `DROP SEQUENCE seq3;` statement. The following result is returned:

```
Query OK, 1 row affected (0.02 sec)
```

Query a sequence

Syntax:

```
SHOW SEQUENCES
```

Example: Execute the `SHOW SEQUENCES;` statement. In the following returned result, the TYPE column displays the abbreviation of each sequence type:

```
mysql>
+------------+---------------------+-------------+------------+---------------------+-------+--------
+
| NAME       | VALUE               | INCREMENT_BY | START_WITH | MAX_VALUE           | CYCLE | TYPE
|
+------------+---------------------+-------------+------------+---------------------+-------+--------
+
| AUTO_SEQ_1 | 91820513            | 1            | 91820200   | 9223372036854775807 | N     | SIMPLE
|
| AUTO_SEQ_4 | 91820200            | 2            | 1000       | 9223372036854775807 | Y     | SIMPLE
|
| seq_test   | N/A                 | N/A          | N/A        | N/A                 | N/A   | TIME
|
| AUTO_SEQ_2 | 100000              | N/A          | N/A        | N/A                 | N/A   | GROUP
|
| AUTO_SEQ_3 | 200000              | N/A          | N/A        | N/A                 | N/A   | GROUP
|
+------------+---------------------+-------------+------------+---------------------+-------+--------
+
5 rows in set (0.01 sec)
```

Obtain a sequence value

Syntax:

```
< sequence name >.NEXTVAL
```

Example: Execute the `SELECT sample_seq.nextVal FROM dual;` statement. The following result is returned:

```
+--------------------+
| SAMPLE_SEQ.NEXTVAL |
+--------------------+
|             101001 |
+--------------------+
1 row in set (0.04 sec)
```

You can also execute the following SQL statement: `INSERT INTO some_users (name,address,gmt_create,gmt_modified,intro) VALUES ('sun',SAMPLE_SEQ.nextVal,now(),now(),'aa');`. In the statement, SAMPLE_SEQ.nextVal returns a sequence value. The following result is returned:

```
Query OK, 1 row affected (0.01 sec)
```

> **Note** If the AUTO_INCREMENT parameter is specified when you create a table, you do not need to specify an auto-increment column when you execute the `INSERT` statement. automatically maintains the auto-increment column.

Obtain multiple sequence values at a time

Syntax:

```
SELECT < sequence name >.NEXTVAL FROM DUAL WHERE COUNT = < numeric value >
```

Example: Execute the `SELECT sample_seq.nextVal FROM dual WHERE count = 10;` statement. The following result is returned:

```
+--------------------+
| SAMPLE_SEQ.NEXTVAL |
+--------------------+
|             101002 |
|             101003 |
|             101004 |
|             101005 |
|             101006 |
|             101007 |
|             101008 |
|             101009 |
|             101010 |
|             101011 |
+--------------------+
10 row in set (0.04 sec)
```

12.1.14.3. Implicit sequence usage

After AUTO_INCREMENT is configured for a primary key, a sequence automatically maintained by can be used to automatically generate the values of the primary key.

CREATE TABLE

The standard CREATE TABLE syntax is extended so that you can add the sequence type for an auto-increment column. If you do not specify a type, the default type GROUP is used. If a sequence is automatically created by and is associated with a table, the sequence name consists of the AUTO_SEQ_ prefix and the table name.

```
CREATE TABLE <name> (
   <column> ... AUTO_INCREMENT [ BY GROUP | SIMPLE | TIME ],
   <column definition>,
   ...
) ... AUTO_INCREMENT=<start value>
```

SHOW CREATE TABLE

The SHOW CREATE TABLE statement returns the type of the sequence that is used to generate values for an auto-increment column in a table shard or a broadcast table.

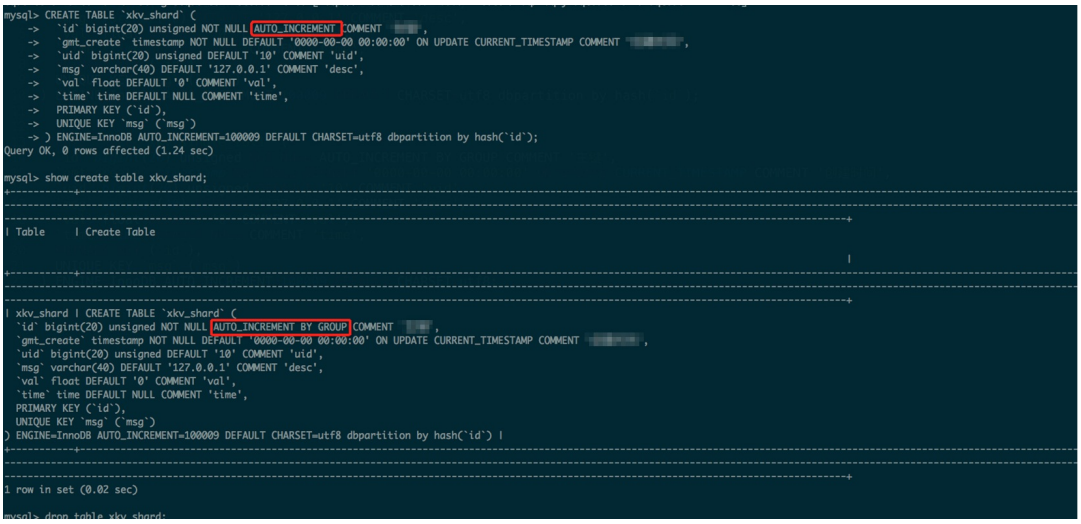
```
SHOW CREATE TABLE <name>
```

Examples

- When you create a table, you specify the AUTO_INCREMENT parameter and you do not specify the sequence

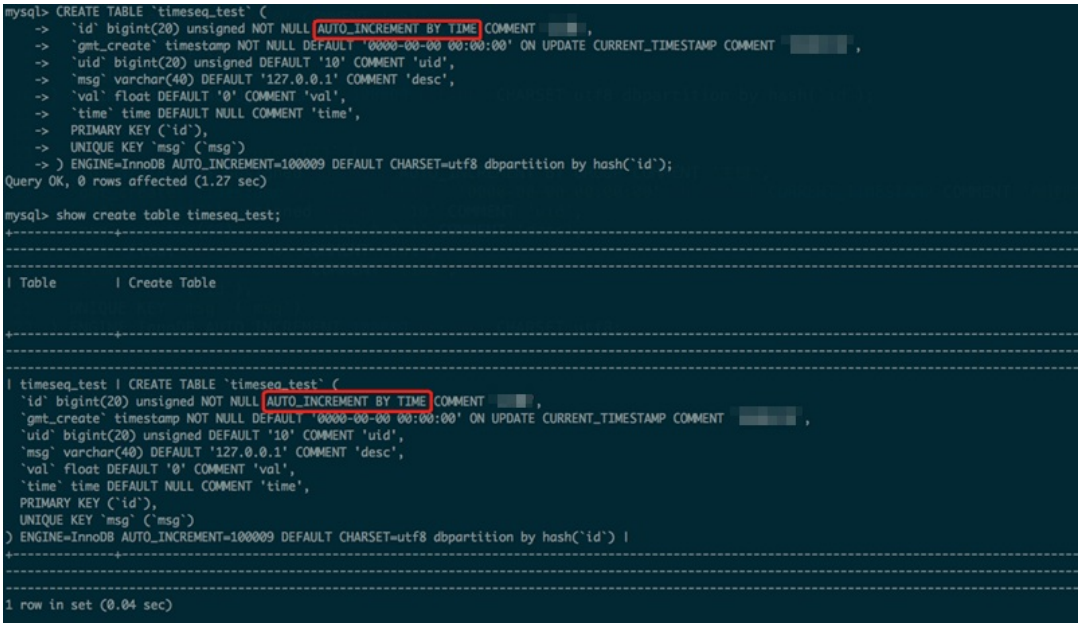
type. In this case, a group sequence is used by default.

Example 1



- When you create a table, specify the AUTO_INCREMENT parameter and specify a time-based sequence as the values of the primary key.

  Example 2



ALTER TABLE

The **ALTER TABLE** statement cannot be used to change the sequence type but can be used to change the start value of the sequence. To change the sequence type of a table, execute the **SHOW SEQUENCES** statement to query the sequence name and the sequence type. Then, execute the **ALTER SEQUENCE** statement to change the sequence type.

Use the following syntax to query the sequence name and the sequence type:

```
SHOW SEQUENCES
```

Use the following syntax to change the sequence type:

```
ALTER SEQUENCE <name> [ CHANGE TO GROUP | SIMPLE | TIME ]
START WITH <numeric value> [ INCREMENT BY <numeric value> ]
[ MAXVALUE <numeric value> ] [ CYCLE | NOCYCLE ]
```

Notice After a sequence is used, we recommend that you do not change the start value specified by the AUTO_INCREMENT parameter. If you want to change the start value, you must analyze the existing sequence values and the rate at which new sequence values are generated. This prevents duplicate sequence values from being generated.

12.1.14.4. Limits and precautions

This topic describes the limits and precautions for sequences.

Limits and precautions

- When a time-based sequence is used in an auto-increment column of a table, the data type of the column must be BIGINT.
- The `START WITH` parameter must be set when the sequence is changed to another type.
- Assume that you need to execute the `INSERT` statement on a non-partitioned database to which only one ApsaraDB RDS for MySQL database is attached. The database automatically optimizes and routes the statement to the ApsaraDB RDS for MySQL database and bypasses the optimizer that allocates sequence values. The INSERT statement is processed the same way if you need to execute this statement on a partitioned database to which single-database table shards are attached and these table shards are not broadcast table shards. The `INSERT INTO ... VALUES (seq.nextval, ...)` syntax is not supported. We recommend that you use the auto-increment column feature of ApsaraDB RDS for MySQL instead.
- Assume that you execute an `INSERT` statement on a PolarDB-X logical table that uses a sequence. For example, you need to execute the `INSERT INTO … VALUES …` or `INSERT INTO … SELECT…` statement. bypasses the optimizer and directly routes the statement to the underlying ApsaraDB RDS for MySQL tables so that the sequence does not take effect. The PolarDB-X logical table creates IDs by using the auto-increment column feature of the ApsaraDB RDS for MySQL tables.
- To allocate IDs for the same table, you can use the sequence feature of or the auto-increment feature of ApsaraDB RDS for MySQL. If you use the two allocation methods together for the same table, duplicate IDs may be created. This makes troubleshooting difficult.

Troubleshoot primary key conflicts

Assume that data is directly written to the underlying ApsaraDB RDS for MySQL database and that the related primary key values are not the sequence values created by the database. The primary key values automatically created by the database may conflict with the directly written data. To troubleshoot this issue, perform the following operations:

1. View the existing sequences by executing the SHOWSEQUENCES SQL statement of . The sequence prefixed with `AUTO_SEQ_` is an implicit sequence. This sequence is generated if you set the `AUTO_INCREMENT` parameter to create the table. Execute the `SHOWSEQUENCES;` statement. The following response is returned:

```
+-------------------------+-------+--------------+------------+-----------+-------+-------+
| NAME                    | VALUE | INCREMENT_BY | START_WITH | MAX_VALUE | CYCLE | TYPE  |
+-------------------------+-------+--------------+------------+-----------+-------+-------+
| AUTO_SEQ_timeseq_test   | N/A   | N/A          | N/A        | N/A       | N/A   | TIME  |
| AUTO_SEQ_xkv_shard_tbl1 | 0     | N/A          | N/A        | N/A       | N/A   | GROUP |
| AUTO_SEQ_xkv_shard      | 0     | N/A          | N/A        | N/A       | N/A   | GROUP |
+-------------------------+-------+--------------+------------+-----------+-------+-------+
3 rows in set (0.04 sec)
```

2. For example, if the t_item table has a conflict and the primary key of this table is ID, execute the `SELECT MAX(`

`id) FROM t_item;` statement to query the largest primary key value of the table from . The following response is returned:

```
+-----------+
| max(id)   |
+-----------+
| 8231 |
+-----------+
1 row in set (0.01 sec)
```

3. Update the related value in the sequence table to a value greater than 8231, such as 9000. Then, no error is returned for the primary key values that are automatically created by the sequence you set in the statement.

   Execute the `ALTER SEQUENCE AUTO_SEQ_USERS START WITH 9000;` statement. The following response is returned:

   ```
   Query OK, 1 row affected (0.01 sec)
   ```

12.1.15. Best practices

12.1.15.1. Determine shard keys

A shard key is a field for database sharding or table sharding. The shard key is used to create sharding rules. A instance horizontally partitions data from a logical table into physical database shards on each ApsaraDB RDS for MySQL instance based on the shard key.

When you perform table sharding, you must comply with the following primary principle: Determine the appropriate entities to which data belongs based on your business logic. You must make sure that most or core SQL operations are performed based on the entities. After you determine the entities, you can use the fields that identify the entities as your shard keys.

In most cases, entities vary based on application scenarios. In the following typical application scenarios, entities are clear and the fields that identify the entities can be used as shard keys:

- User-oriented Internet applications are designed to meet user requirements. Users are the entities and the user ID field can be used as the shard key.
- Seller-oriented e-commerce applications are designed to meet seller requirements. Sellers are the entities and the seller ID field can be used as the shard key.
- Gaming applications are designed to meet gamer requirements. Gamers are the entities and the gamer ID field can be used as the shard key.
- Internet of Vehicles (IoV) applications are designed based on vehicles. Vehicles are the entities and the vehicle ID field can be used as the shard key.
- Tax applications are designed based on taxpayers. Taxpayers are the entities and the taxpayer ID field can be used as the shard key.

In other scenarios, you can also use the appropriate field that identifies the entities as the shard key.

Assume that you need to horizontally partition data in a single table of a seller-oriented e-commerce application.

```
CREATE TABLE sample_order (
  id INT(11) NOT NULL,
  sellerId INT(11) NOT NULL,
  trade_id INT(11) NOT NULL,
  buyer_id INT(11) NOT NULL,
  buyer_nick VARCHAR(64) DEFAULT NULL,
  PRIMARY KEY (id)
);
```

Sellers are the entities. Therefore, you can use the sellerId field as the shard key to perform only database sharding. You can execute the following data definition language (DDL) statement to create a table:

```
CREATE TABLE sample_order (
  id INT(11) NOT NULL,
  sellerId INT(11) NOT NULL,
  trade_id INT(11) NOT NULL,
  buyer_id INT(11) NOT NULL,
  buyer_nick VARCHAR(64) DEFAULT NULL,
  PRIMARY KEY (id)
) DBPARTITION BY HASH(sellerId);
```

If no entity can be used as the shard key, use the following methods to determine the shard key:

- Determine your shard key based on data distribution and data access requests. Make sure that your data is evenly distributed across table shards in database shards if possible. This method can be used if a large number of analytical queries need to be performed and query concurrency stays at 1 in most cases.
- Combine the fields of the string, date, and time data types and use the combined result as your shard key for database sharding or table sharding. This method is applicable to log retrieval.

Assume that a log system records all user operations and that you need to horizontally partition the following single log table:

```
CREATE TABLE user_log (
  userId INT(11) NOT NULL,
  name VARCHAR(64) NOT NULL,
  operation VARCHAR(128) DEFAULT NULL,
  actionDate DATE DEFAULT NULL
);
```

You can combine the user ID and time fields as the shard key to partition the table by week. You can execute the following DDL statement to create a table:

```
CREATE TABLE user_log (
  userId INT(11) NOT NULL,
  name VARCHAR(64) NOT NULL,
  operation VARCHAR(128) DEFAULT NULL,
  actionDate DATE DEFAULT NULL
) DBPARTITION BY HASH(userId) TBPARTITION BY WEEK(actionDate) TBPARTITIONS 7;
```

For more information about how to determine shard keys and how to perform table sharding, see DDL statements.

Notice Whatever shard key and sharding policy are used, avoid querying data only from one shard if possible.

12.1.15.2. Select the number of shards

This topic describes how to select the number of shards.

The number of shards is also referred to as the number of table shards. supports two methods for horizontal splitting: database sharding and table sharding. By default, eight physical database shards are created on each RDS instance. One or more physical table shards can be created on each physical database shard. In most cases, we recommend that each physical table shard contains no more than 5 million rows of data. In general scenarios, you can estimate the data growth in one or two years. Then, divide the estimated total data size by the total number of physical database shards, and divide the result by the recommended maximum data size of 5 million. In this case, you can obtain the number of physical table shards to be created on each physical database shard.

```
Number of physical table shards on each physical database shard = CEILING(Estimated total data size/(Number of RDS instances x 8)/5,000,000)
```

Therefore, if the calculated number of physical table shards is equal to 1, only database sharding is required. This indicates that a physical table shard is created on each physical database shard. You do not need to perform table sharding. If the calculation result is greater than 1, we recommend that you perform database sharding and table sharding. This indicates that multiple physical table shards are created on each physical database shard.

Assume that the estimated total size of a table is about 0.1 billion rows two years later and you have four RDS instances. In this case, calculate the number of shards by using the preceding formula:

```
Number of physical table shards on each physical database shard = CEILING(100,000,000/(4 x 8)/5,000,000) = CEILING(0.625) = 1
```

The result is 1. This indicates that only database sharding is required, and one physical table shard is created on each physical database shard.

Assume that only one RDS instance is available in the preceding example. In this case, calculate the number of shards by using the preceding formula:

```
Number of physical table shards on each physical database shard = CEILING(100,000,000/(1 x 8)/5,000,000) = CEILING(2.5) = 3
```

The result is 3. Therefore, we recommend that you perform database sharding and table sharding, and create three physical table shards on each physical database shard.

12.1.15.3. Basic concepts of SQL optimization

is an efficient and stable distributed relational database service that processes distributed relational operations. optimizes SQL statements differently from standalone relational databases, such as MySQL databases. PolarDB-X focuses on the network I/O overheads in a distributed environment and routes SQL operations to the underlying database shards, such as ApsaraDB RDS for MySQL database shards. This reduces the network I/O overheads and improves the SQL execution efficiency.

provides statements for you to obtain SQL execution information and optimize SQL execution. For example, you can execute the EXPLAIN statements to query SQL execution plans and execute the TRACE statements to query SQL execution processes and overheads. This topic introduces the basic concepts and common statements related to SQL optimization in .

Execution plans

An SQL execution plan (or execution plan) is a set of ordered steps generated to access data. In , execution plans are divided into the execution plans at the layer and the execution plans at the ApsaraDB RDS for MySQL layer. Execution plan analysis is an effective way to optimize SQL statements. It allows you to know whether or ApsaraDB RDS for MySQL has generated optimal execution plans for SQL statements and whether further optimization can be made.

When an SQL statement is executed, the optimizer analyzes the basic information about the SQL statement and the related tables. Then, the optimizer determines the database shards on which the SQL statement is executed. The optimizer also determines the specific SQL statement format, execution policy, and data merging and computing policy for each database shard. This process optimizes SQL statement execution and generates an execution plan at the layer. The execution plan at the ApsaraDB RDS for MySQL layer is the native MySQL execution plan.

provides a set of EXPLAIN statements to display the execution plans that are implemented at different layers or show different levels of details.

The following table briefly describes the EXPLAIN statements available for use in .

EXPLAIN statements

Statement	Description	Example
EXPLAIN { SQL }	Displays the information about the summary execution plan of an SQL statement at the layer. The information includes the database shards on which the SQL statement is executed, physical statements, and general parameters.	EXPLAIN SELECT * FROM test
EXPLAIN DETAIL { SQL }	Displays the information about the detailed execution plan of an SQL statement at the layer. The information includes the statement type, the maximum number of threads that can be used to run queries in parallel, the returned field information, table shards, and database shards.	EXPLAIN DETAIL SELECT * FROM test
EXPLAIN EXECUTE { SQL }	Displays the information about the execution plan at the underlying ApsaraDB RDS for MySQL layer. This statement is equivalent to the EXPLAIN statement that is used by open source MySQL.	EXPLAIN EXECUTE SELECT * FROM test

Execution plans at the layer

The following table describes the fields in the response returned for an execution plan at the layer.

Fields in an execution plan at the layer

Field	Description
GROUP_NAME	The name of a database shard for a logical table. The suffix identifies the specific database shard. This value is consistent with the result of the SHOW NODE statement.
SQL	The SQL statement executed on the database shard.
PARAMS	The SQL statement parameters used when communicates with ApsaraDB RDS for MySQL over the Prepare protocol.

The value of the SQL field can use the following forms:

1. If an SQL statement does not contain the following parts, the execution plan is displayed in the form of SQL statements:
   - An aggregate function that involves multiple database shards
   - A distributed JOIN clause that involves multiple table shards
   - A complex subquery

   For example, you execute the `EXPLAIN SELECT * FROM test;` statement. The following response is returned:

```
+-------------------------------------------+-------------------------------------------+-----
---+
| GROUP_NAME                                | SQL                                       | PARA
MS |
+-------------------------------------------+-------------------------------------------+-----
---+
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0000_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0001_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0002_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0003_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0004_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0005_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0006_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
| TESTDB_1478746391548CDTCTESTDB_OXGJ_0007_RDS | select `test`.`c1`,`test`.`c2` from `test` | {}
|
+-------------------------------------------+-------------------------------------------+-----
---+
8 rows in set (0.04 sec)
```

You can query the group names in the GROUP_NAME column by executing the `SHOW NODE;` statement.

```
+----+-------------------------------------------+------------------+------------------+------
--------------+--------------------+
| ID | NAME                                      | MASTER_READ_COUNT | SLAVE_READ_COUNT | MASTE
R_READ_PERCENT | SLAVE_READ_PERCENT |
+----+-------------------------------------------+------------------+------------------+------
--------------+--------------------+
| 0  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0000_RDS |                69 |                0 | 100%
| 0%                 |
| 1  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0001_RDS |                45 |                0 | 100%
| 0%                 |
| 2  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0002_RDS |                30 |                0 | 100%
| 0%                 |
| 3  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0003_RDS |                29 |                0 | 100%
| 0%                 |
| 4  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0004_RDS |                11 |                0 | 100%
| 0%                 |
| 5  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0005_RDS |                 1 |                0 | 100%
| 0%                 |
| 6  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0006_RDS |                 8 |                0 | 100%
| 0%                 |
| 7  | TESTDB_1478746391548CDTCTESTDB_OXGJ_0007_RDS |                50 |                0 | 100%
| 0%                 |
+----+-------------------------------------------+------------------+------------------+------
--------------+--------------------+
8 rows in set (0.10 sec)
```

2. Execution plans that cannot be expressed by SQL statements can be expressed in custom formats of .

   For example, you execute the `EXPLAIN DETAIL SELECT COUNT(*) FROM test;` statement. The following response is returned:

```
+-----------------------------------------------+----------------------------------+--------+
| GROUP_NAME                                    | SQL                              | PARAMS |
+-----------------------------------------------+----------------------------------+--------+
| TEST_DB_1478746391548CDTCTEST_DB_OXGJ_0000_RDS | Merge as test
    queryConcurrency:GROUP_CONCURRENT
    columns:[count(*)]
    executeOn: TEST_DB_1478746391548CDTCTEST_DB_OXGJ_0000_RDS
        Query from test as test
            queryConcurrency:SEQUENTIAL
            columns:[count(*)]
            tableName:test
            executeOn: TEST_DB_1478746391548CDTCTEST_DB_OXGJ_0000_RDS
        Query from test as test
            queryConcurrency:SEQUENTIAL
            columns:[count(*)]
            tableName:test
            executeOn: TEST_DB_1478746391548CDTCTEST_DB_OXGJ_0001_RDS
        ... ...
        Query from test as test
            queryConcurrency:SEQUENTIAL
            columns:[count(*)]
            tableName:test
            executeOn: TEST_DB_1478746391548CDTCTEST_DB_OXGJ_0007_RDS
| NULL   |
+-----------------------------------------------+----------------------------------+--------+
1 row in set (0.00 sec)
```

The executeOn field indicates the database shard on which the SQL statement is executed. finally merges the results returned by the related database shards.

Execution plans at the ApsaraDB RDS for MySQL layer

Execution plans at the ApsaraDB RDS for MySQL layer are the same as native MySQL execution plans. For more information, see MySQL documentation.

A logical table may consist of multiple table shards that are distributed in different database shards. Therefore, you can view the execution plans at the ApsaraDB RDS for MySQL layer in multiple ways.

1. View the execution plan of an SQL statement on an ApsaraDB RDS for MySQL table shard.

   If the query condition contains a shard key, execute the EXPLAIN EXECUTE statement to query the execution plan on the corresponding table shard. For example, you execute the `EXPLAIN EXECUTE SELECT * FROM test WHERE c1 = 1;` statement. The following response is returned:

```
+----+-------------+-------+-------+---------------+---------+---------+-------+------+-------+
| id | select_type | table | type  | possible_keys | key     | key_len | ref   | rows | Extra |
+----+-------------+-------+-------+---------------+---------+---------+-------+------+-------+
|  1 | SIMPLE      | test  | const | PRIMARY       | PRIMARY | 4       | const |    1 | NULL  |
+----+-------------+-------+-------+---------------+---------+---------+-------+------+-------+
1 row in set (0.04 sec)
```

> **Notice** In some cases, such as if an SQL statement does not contain a shard key, the SQL statement is executed on multiple table shards. In such a case, the EXPLAIN EXECUTE statement returns an execution plan that is on a random ApsaraDB RDS for MySQL table shard.

   To view the execution plan of an SQL statement on a specified table shard, you can add a NODE hint. For example, you execute the `/!TDDL:node='TESTDB_1478746391548CDTCTESTDB_OXGJ_0000_RDS'*/EXPLAIN SELECT * FROM test;` statement. The following response is returned:

```
+----+-------------+-------+------+---------------+-------+---------+-------+------+-------+
| id | select_type | table | type | possible_keys |  key  | key_len |  ref  | rows | Extra |
+----+-------------+-------+------+---------------+-------+---------+-------+------+-------+
|  1 | SIMPLE      | test  | ALL  | NULL          | NULL  | NULL    | NULL  |    2 | NULL  |
+----+-------------+-------+------+---------------+-------+---------+-------+------+-------+
1 row in set (0.04 sec)
```

2. View the execution plans on all ApsaraDB RDS for MySQL table shards.

   To view the execution plans of the SQL statement on all table shards, you can add the SCAN hint in the SQL statement. For example, you execute the `/!TDDL:scan='test'*/EXPLAIN SELECT * FROM test;` statement. The following response is returned:

```
+----+-------------+-------+------+---------------+-------+---------+-------+------+-------+
| id | select_type | table | type | possible_keys |  key  | key_len |  ref  | rows | Extra |
+----+-------------+-------+------+---------------+-------+---------+-------+------+-------+
|  1 | SIMPLE      | test  | ALL  | NULL          | NULL  | NULL    | NULL  |    2 | NULL  |
|  1 | SIMPLE      | test  | ALL  | NULL          | NULL  | NULL    | NULL  |    3 | NULL  |
+----+-------------+-------+------+---------------+-------+---------+-------+------+-------+
2 rows in set (0.08 sec)
```

> **Notice**
>
> i. After you add the hint, only replaces the table names to implement sharding. Then, it directly routes the logical SQL statement to the underlying ApsaraDB RDS for MySQL table shards. It does not process the SQL statement or the result that is returned for the SQL statement.
>
> ii. Execution plans obtained by executing EXPLAIN statements are generated based on static analysis. The corresponding SQL statements are not executed on databases.

TRACE statements

The TRACE statements in can track the SQL execution process and the overheads at each stage. It can be used together with execution plans to optimize SQL statements.

PolarDB-X provides the TRACE and SHOW TRACE statements. These two statements must be used together.

12.1.15.4. SQL optimization methods

12.1.15.4.1. Overview

This topic describes the principles of SQL optimization and methods for optimizing different types of SQL statements in .

Basic principles of SQL optimization

In , SQL computing that can be performed by the underlying ApsaraDB RDS for MySQL instances is called push-down computing. Push-down computing reduces data transfers, decreases overheads at the network layer and layer, and improves the execution efficiency of SQL statements.

Therefore, the basic principle of SQL optimization in is to push down as many computations as possible to the underlying ApsaraDB RDS for MySQL instances.

The following list shows the push-down computations:

- JOIN connections
- Filter conditions, such as the `WHERE` or `HAVING` conditions
- Aggregate computations, such as the `COUNT` and `GROUP BY`
- Sorting, such as `ORDER BY`

- Deduplication, such as `DISTINCT`
- Function computations, such as the `NOW()` function
- Subqueries

> **Notice** The preceding list only describes possible forms of push-down computations. This does not mean that all clauses, conditions, or combinations of clauses or conditions can be pushed down for computing.

SQL statements that have different types and conditions can be optimized in different ways. The following list shows some specific scenarios:

- Single-table SQL optimization
  - Filter condition optimization
  - Optimization of the number of returned rows for a query
  - Grouping and sorting optimization
- Join optimization
  - Optimization of joins that can be pushed down
  - Optimization of distributed joins
- Subquery optimization

12.1.15.4.2. Single-table SQL optimization

This topic describes how to optimize SQL statements for a single table.

Single-table SQL optimization must follow the following rules:

- Make sure that the SQL statement contains the shard key.
- Use an equivalence condition for the shard key whenever possible.
- If the shard key is an IN condition, the number of values after IN must be as small as possible. This means that this number must be far smaller than the number of table shards and remain unchanged as your business grows.
- If the SQL statement does not contain a shard key, use only one of the DISTINCT, GROUP BY, and ORDER BY clauses in the same SQL statement.

Filter condition optimization

horizontally partitions data by the shard key. This requires the filter condition to contain the shard key so that can push queries down to specific database shards based on the shard key values. This way, does not need to scan all table shards.

For example, the shard key of the test table is c1. If the filter condition does not contain this shard key, full table scan is performed.

Execute the `SELECT * FROM test WHERE c2 = 2` statement. The following response is returned:

```
+----+----+
| c1 | c2 |
+----+----+
|  2 |  2 |
+----+----+
1 row in set (0.05 sec)
```

To view the corresponding execution plan, execute the `EXPLAIN SELECT * FROM test WHERE c2 = 2;` statement. The following response is returned:

```
+-----------------------------------------------+-----------------------------------------------
---------------+--------+
| GROUP_NAME                                    | SQL
| PARAMS |
+-----------------------------------------------+-----------------------------------------------
---------------+--------+
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0004_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0007_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0005_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0002_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0003_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0006_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0000_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0001_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c2` = 2) | {}     |
+-----------------------------------------------+-----------------------------------------------
---------------+--------+
8 rows in set (0.00 sec)
```

The smaller the value range of the filter condition that contains the shard key, the faster the query speed of .

Assume that you execute the `SELECT * FROM test WHERE c1 > 1 AND c1 < 4;` statement to query data that contains c1 and meets the condition from the test table. The following response is returned:

```
+----+----+
| c1 | c2 |
+----+----+
|  2 |  2 |
|  3 |  3 |
+----+----+
2 rows in set (0.04 sec)
```

To view the corresponding execution plan, execute the `EXPLAIN SELECT * FROM test WHERE c1 > 1 AND c1 < 4;` statement. The following response is returned:

```
+-----------------------------------------------+-----------------------------------------------
----------------------------------------+--------+
| GROUP_NAME                                    | SQL
| PARAMS |
+-----------------------------------------------+-----------------------------------------------
----------------------------------------+--------+
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0002_RDS | select `test`.`c1`,`test`.`c2` from `test` where ((
`test`.`c1` > 1) AND (`test`.`c1` < 4)) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0003_RDS | select `test`.`c1`,`test`.`c2` from `test` where ((
`test`.`c1` > 1) AND (`test`.`c1` < 4)) | {}     |
+-----------------------------------------------+-----------------------------------------------
----------------------------------------+--------+
2 rows in set (0.00 sec)
```

The equivalence condition is executed faster than the range condition. Assume that you execute the `SELECT * FROM test WHERE c1 = 2;` statement. The following response is returned:

```
+----+----+
| c1 | c2 |
+----+----+
|  2 |  2 |
+----+----+
1 row in set (0.03 sec)
```

To view the corresponding execution plan, execute the `EXPLAIN SELECT * FROM test WHERE c1 = 2;` statement. The following response is returned:

```
+-----------------------------------------------+-------------------------------------------------
---------------+--------+
| GROUP_NAME                                    | SQL
| PARAMS |
+-----------------------------------------------+-------------------------------------------------
---------------+--------+
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0002_RDS | select `test`.`c1`,`test`.`c2` from `test` where (`
test`.`c1` = 2) | {}     |
+-----------------------------------------------+-------------------------------------------------
---------------+--------+
1 row in set (0.00 sec)
```

If you need to insert data into a partitioned table, the inserted field must contain a shard key.

Assume that you specify shard key c1 when you insert data to the test table. Execute the `INSERT INTO test(c1,c2) VALUES(8,8);` statement. The following response is returned:

```
Query OK, 1 row affected (0.07 sec)
```

Optimization of the number of returned rows for a query

When performs a query that contains **LIMIT** [*offset*,] *row_count*, PolarDB-X actually reads records before *offset* in order and directly discards them. When the value of *offset* is large, the query is slow even if the value of *row_count* is small. For example, execute the following SQL statement:

```
SELECT *
FROM sample_order
ORDER BY sample_order.id
LIMIT 10000, 2;
```

Although only the 10,000th and 10,001st records are returned, it takes about 12 seconds to execute the SQL statement because actually reads 10,002 records.

Execute the `SELECT * FROM sample_order ORDER BY sample_order.id LIMIT 10000,2;` statement. The following response is returned.

```
+--------------+------------+--------------+--------------+------------+
| id           | sellerId   | trade_id     | buyer_id     | buyer_nick |
+--------------+------------+--------------+--------------+------------+
| 242012755468 | 1711939506 | 242012755467 | 244148116334 | zhangsan   |
| 242012759093 | 1711939506 | 242012759092 | 244148138304 | wangwu     |
+--------------+------------+--------------+--------------+------------+
2 rows in set (11.93 sec)
```

To view the corresponding execution plan, execute the `EXPLAIN SELECT * FROM sample_order ORDER BY sample_order.id LIMIT 10000,2;` statement. The following response is returned.

```
+------------------+-------------+------------+
| GROUP_NAME       | SQL         | PARAMS     |
+------------------+-------------+------------+
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0004_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0007_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0005_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0002_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0003_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0006_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0000_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0001_RDS | select `sample_order`.`id`,`sample_order`.`sellerId
`,`sample_order`.`trade_id`,`sample_order`.`buyer_id`,`sample_order`.`buyer_nick` from `sample_order`
order by `sample_order`.`id` asc  limit 0,10002 | {}     |
+------------------+---------------------------+--------+
8 rows in set (0.01 sec)
```

To optimize the preceding SQL statement, find the ID set, and use the IN query to match the actual records. The following example shows the SQL query after modification:

```
SELECT *
FROM sample_order o
WHERE o.id IN (
      SELECT id
      FROM sample_order
      ORDER BY id
      LIMIT 10000, 2 );
```

This is to cache IDs in the memory first on the premise that the number of IDs is small. If the shard key of the sample_order table is ID, can also push down such an IN query to different database shards by performing rule-based calculation. This avoids full table scan and unnecessary network I/O operations. View the effect of the SQL query after modification:

Execute the `SELECT * FROM sample_order o WHERE o.id IN ( SELECT id FROM sample_order ORDER BY id LIMIT 10000,2 );` statement. The following response is returned:

```
+--------------+------------+--------------+--------------+------------+
| id           | sellerId   | trade_id     | buyer_id     | buyer_nick |
+--------------+------------+--------------+--------------+------------+
| 242012755468 | 1711939506 | 242012755467 | 244148116334 | zhangsan   |
| 242012759093 | 1711939506 | 242012759092 | 244148138304 | wangwu     |
+--------------+------------+--------------+--------------+------------+
2 rows in set (1.08 sec)
```

The execution time is significantly reduced from 12 seconds to 1.08 seconds.

To view the corresponding execution plan, execute the `EXPLAIN SELECT * FROM sample_order o WHERE o.id IN ( SELECT id FROM sample_order ORDER BY id LIMIT 10000,2 );` statement. The following response is returned:

```
+-----------------------------------------------+-----------------------------------------------------
-------------------------------------------------------------------------------+--------+
| GROUP_NAME                                    | SQL
| PARAMS |
+-----------------------------------------------+-----------------------------------------------------
-------------------------------------------------------------------------------+--------+
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0002_RDS | select `o`.`id`,`o`.`sellerId`,`o`.`trade_id`,`o`.`
buyer_id`,`o`.`buyer_nick` from `sample_order` `o` where (`o`.`id` IN (10002)) | {}     |
| SEQPERF_1478746391548CDTCSEQPERF_OXGJ_0001_RDS | select `o`.`id`,`o`.`sellerId`,`o`.`trade_id`,`o`.`
buyer_id`,`o`.`buyer_nick` from `sample_order` `o` where (`o`.`id` IN (10001)) | {}     |
+-----------------------------------------------+-----------------------------------------------------
-------------------------------------------------------------------------------+--------+
2 rows in set (0.03 sec)
```

Grouping and sorting optimization

In , if an SQL query must use all of the DISTINCT, GROUP BY, and ORDER BY clauses, make sure that the fields after these clauses are the same and that the fields are shard keys. This way, only a small amount of data is returned for the SQL query. This minimizes the network bandwidth consumed by distributed queries and removes the need to retrieve a large volume of data and sort the data in a temporary table. This way, the system performance is maximized.

12.1.15.4.3. Join optimization

Joins in are classified into joins that can be pushed down to ApsaraDB RDS for MySQL and those that cannot be pushed down to ApsaraDB RDS for MySQL. Joins that cannot be pushed down to ApsaraDB RDS for MySQL are distributed joins. The optimization policies for these two types of joins are different.

Optimize joins that can be pushed down

Joins that can be pushed down are classified into the following types:

- Joins between single tables, which are non-partitioned tables.
- The tables involved in the join contain the shard key in the filter condition and use the same sharding algorithm. This means that the data calculated by the sharding algorithm is distributed to the same table shard.
- The tables involved in the join use the shard key as the join condition and use the same sharding algorithm.
- Joins between broadcast tables and partitioned tables. Broadcast tables are also called small table broadcast.

In , optimize joins to those that can be pushed down to and executed on database shards.

Take a join between a broadcast table and a partitioned table as an example. The broadcast table is used as the driving table for the join. The left table for the join is called the driving table. The broadcast table of is replicated to all database shards. When the broadcast table is used as the driving table for the join, the join between this broadcast table and table shards can be pushed down to each database shard. The results are combined for computing. This way, the query performance is improved.

For example, a join is performed on the following three tables, among which the sample_area table is the broadcast table whereas the sample_item and sample_buyer tables are partitioned tables. The query execution time is about 15 seconds.

Execute the `SELECT sample_area.name FROM sample_item i JOIN sample_buyer b ON i.sellerId = b.sellerId JOIN sample_area a ON b.province = a.id WHERE a.id < 110107 LIMIT 0, 10;` statement. The following response is returned:

```
+------+
| name |
+------+
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
+------+
10 rows in set (14.88 sec)
```

If you adjust the join order and move the broadcast table to the leftmost as the driving table for the join, the join is pushed down to each database shard of the instance.

Execute the `SELECT sample_area.name FROM sample_area a JOIN sample_buyer b ON b.province = a.id JOIN sample_item i ON i.sellerId = b.sellerId WHERE a.id < 110107 LIMIT 0, 10;` statement. The following response is returned:

```
+------+
| name |
+------+
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
|  BJ  |
+------+
10 rows in set (0.04 sec)
```

The query execution time decreases from 15 seconds to 0.04 seconds. The query performance is significantly improved.

Notice The broadcast table achieves data consistency among database shards by using a synchronization mechanism. Data synchronization has a latency of several seconds.

Optimization of distributed joins

If a join cannot be pushed down, must complete part of the computing in the query. A join that cannot be pushed down means that the join condition and filter condition do not contain a shard key. Such a join is a distributed join.

Tables in a distributed join are classified into two types based on the data size:

- Small table: A table that contains a small amount of data (less than 100 data records or less data than other tables) that is involved in join computation after filtering.
- Large table: A table that contains a large amount of data (more than 100 data records or more data than other tables) that is involved in join computation after filtering.

In most cases, Nested Loop and its derived algorithms are used in join computation at the layer. If sorting is required for joins, the Sort Merge algorithm is used. When the Nested Loop algorithm is used, a smaller data size in the left table for a join indicates a smaller number of queries performed by on the right table. If the right table has indexes or contains a small volume of data, the join is even faster. In , the left table for a distributed join is called the driving table. To optimize a distributed join, use a small table as the driving table and set as many filter conditions as possible for the driving table.

Take the following distributed join as an example. The query takes about 24 seconds:

Execute the `SELECT t.title, t.price FROM sample_order o, ( SELECT * FROM sample_item i WHERE i.id = 242002396687 ) t WHERE t.source_id = o.source_item_id AND o.sellerId < 1733635660;` statement. The following response is returned:

```
+----------------------------------+--------+
| title                            | price  |
+----------------------------------|--------+
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
+----------------------------------+--------+
10 rows in set (23.79 sec)
```

The preceding join is an INNER JOIN query, and the actual size of the intermediate data involved in the join computation is unknown. Perform COUNT() operations on the o table and the t table to obtain the actual data size.

The o.sellerId < 1733635660 filter in the WHERE condition is only related to the o table. Extract this filter and add it to the COUNT() condition of the o table. Execute the `SELECT COUNT(*) FROM sample_order o WHERE o.sellerId < 1733635660;` statement. The following response is returned:

```
+----------+
| count(*) |
+----------+
|   504018 |
+----------+
1 row in set (0.10 sec)
```

The intermediate result of the o table contains about 500,000 records. Similarly, the t table is a subquery. Directly extract the subquery and add it to the COUNT() query. Then, execute the `SELECT COUNT(*) FROM sample_item i WHERE i.id = 242002396687;` statement. The following response is returned:

```
+----------+
| count(*) |
+----------+
|        1 |
+----------+
1 row in set (0.01 sec)
```

The intermediate result of the t table contains only one record. Therefore, the o table is the large table and the t table is the small table. Use the small table as the driving table for the distributed join. Then, execute the `SELECT t.title, t.price FROM ( SELECT * FROM sample_item i WHERE i.id = 242002396687 ) t, sample_order o WHERE t.source_id = o.source_item_id AND o.sellerId < 1733635660;` statement. The following response is returned:

```
+----------------------------------+--------+
| title                            | price  |
+----------------------------------|--------+
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
| Sample Item for Distributed JOIN | 239.00 |
+----------------------------------+--------+
10 rows in set (0.15 sec)
```

The query time was reduced from about 24 seconds to 0.15 seconds. The query performance is significantly improved.

12.1.15.4.4. Subquery optimization

When you optimize an SQL statement that contains subqueries, push the subqueries down to database shards as many as possible to reduce the computing workload at the layer.

For this purpose, you can try the following two optimization methods:

- Rewrite subqueries into multi-table joins, and optimize the joins.
- Use the shard key in the join condition or filter condition so that can push the queries down to specific database shards to avoid full table scan.

Take the following subquery as an example:

```
SELECT o. *
FROM sample_order o
WHERE NOT EXISTS
     (SELECT sellerId FROM sample_seller s WHERE o.sellerId = s.id);
```

Rewrite the subquery into a join:

```
SELECT o. *
FROM sample_order o LEFT JOIN sample_seller s ON o.sellerId = s.id
WHERE s.id IS NULL;
```

12.1.15.5. Choose a connection pool for an application

A database connection pool is used to manage database connections in a centralized manner. This improves application performance and reduce database loads.

Connection pools offer the following benefits:

- Resource reuse: Connections can be reused to avoid high performance overheads caused by frequent connection establishment and release. Resource reuse can also improve the system stability.
- Improvement of system response efficiency: After the connection initialization is complete, all requests can directly use the existing connections. This avoids the overheads of connection initialization and release and improves the system response efficiency.
- Connection leakage prevention: The connection pool forcibly revokes connections based on the preset revocation policy to avoid connection resource leakage.

We recommend that you use a connection pool to connect applications and databases for business operations. For Java programs, we recommend that you use Druid connection pools.

The following code shows the standard Spring configurations of Druid connection pools:

```
    <bean id="dataSource" class="com.alibaba.druid.pool.DruidDataSource" init-method="init" destroy-me
thod="close">
        <property name="driverClassName" value="com.mysql.jdbc.Driver" />
        <! -- Set the basic properties, including the URL, username, and password. -->
        <property name="url" value="jdbc:mysql://ip:port/db? autoReconnect=true&amp;rewriteBatchedStat
ements=true&amp;socketTimeout=30000&amp;connectTimeout=3000" />
        <property name="username" value="root" />
        <property name="password" value="123456" />
        <! -- Configure the initial connection pool size, maximum number of active connections, and mi
nimum number of idle connections. -->
        <property name="maxActive" value="20" />
        <property name="initialSize" value="3" />
        <property name="minIdle" value="3" />
        <! -- maxWait indicates the timeout period for obtaining the connection. -->
        <property name="maxWait" value="60000" />
        <! -- timeBetweenEvictionRunsMillis indicates the interval for detecting idle connections to b
e closed, in milliseconds. -->
        <property name="timeBetweenEvictionRunsMillis" value="60000" />
        <! -- minEvictableIdleTimeMillis indicates the minimum idle time of a connection in the connec
tion pool, in milliseconds.-->
        <property name="minEvictableIdleTimeMillis" value="300000" />
        <! -- Set the SQL statement used to check whether connections are available. -->
        <property name="validationQuery" value="SELECT 'z'" />
        <! -- Set whether to enable idle connection checks. -->
        <property name="testWhileIdle" value="true" />
        <! -- Set whether to check the connection status before the system obtains a connection. -->
        <property name="testOnBorrow" value="false" />
        <! -- Set whether to check the connection status before the system releases a connection. -->
        <property name="testOnReturn" value="false" />
    </bean>
```

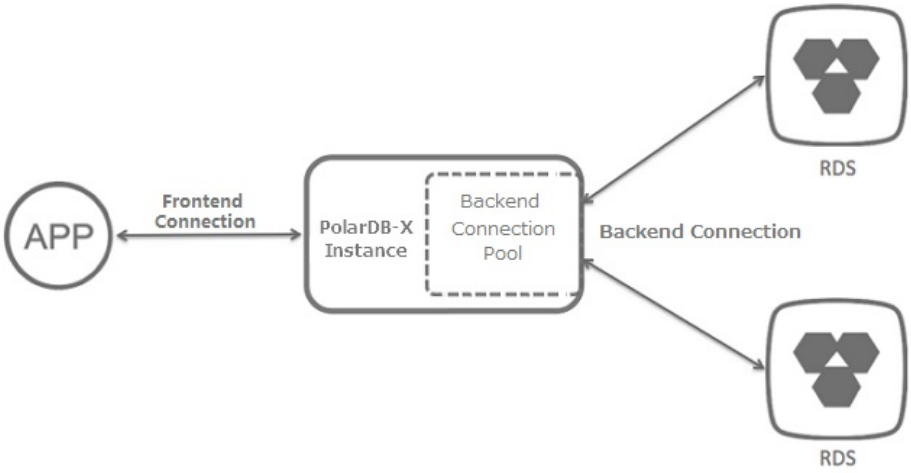
12.1.15.6. Connections to PolarDB-X instances

This topic describes the connections to PolarDB-X instances.

When an application connects to a instance for operation, two types of connections are available from the perspective of the instance:

- Frontend connection: a connection established by an application to a logical database in a instance.
- Backend connection: a connection established by a node in a instance to a physical database in a backend RDS instance.

instance connection diagram



Frontend connection

In theoretical scenarios, the number of frontend connections is limited only by the available memory size and the number of network connections to the nodes of the instance. In actual scenarios, when an application connects to a instance, the nodes of the PolarDB-X instance manage a limited number of connections to perform requested operations. The nodes do not maintain a large number of concurrent persistent connections, such as tens of thousands of concurrent persistent connections. Therefore, the number of frontend connections that a instance can accept is considered unlimited.

The number of frontend connections is unlimited and a large number of idle connections are allowed. Therefore, this method applies to scenarios where a large number of servers on which applications are deployed exist. You must maintain their connections to the instance.

> **Note** Although the number of frontend connections is considered unlimited, operation requests obtained from frontend connections are executed by internal threads of the instance over backend connections. Due to the limited number of internal threads and backend connections, the total number of concurrent requests that the instance can process is limited.

Backend connection

Each node of a instance creates a backend connection pool to automatically manage and maintain the backend connections to the physical databases in the RDS instance. Therefore, the maximum number of connections in the backend connection pool of a instance is related to the maximum number of connections that the RDS instance supports. You can use the following formula to calculate the maximum number of connections in the backend connection pool of a instance:

```
Maximum number of connections in a backend connection pool of a  instance = FLOOR(Maximum number of connections in an RDS instance/Number of physical database shards on the RDS instance/Number of nodes on the  instance)
```

Assume that you have an RDS instance and a instance of the following specifications:

- The RDS instance has eight physical database shards, four cores, and 16 GB memory, and supports a maximum number of 4,000 connections.
- The dedicated instance has 32 cores and 32 GB memory, and each node on the instance has two cores and 2 GB memory. This indicates that the instance has 16 nodes.

You can use the following formula to calculate the maximum number of connections in the backend connection pool of the instance:

```
Maximum number of connections in the backend connection pool of the  instance = FLOOR(4000/8/16) = FLOOR(31.25) = 31
```

> **Note**
>
> - The calculation result by using the preceding formula is the maximum number of connections in the backend connection pool of the instance. In actual scenarios, the instance sets the maximum number of connections in the backend connection pool to a value that is smaller than the calculated maximum value. This reduces the connection pressure.
> - We recommend that you create databases in a instance on a dedicated RDS instance. This indicates that you do not create databases for other applications or instances on the dedicated RDS instance.

Relationship between frontend connections and backend connections

After an application establishes frontend connections to a instance and sends requests for SQL statement execution, nodes of the instance process the requests in an asynchronous way. The nodes obtain backend connections from the internal backend connection pool, and then execute optimized SQL statements on one or more physical database shards.

Nodes of the instance process requests in an asynchronous way, and the frontend connections are not bound to backend connections. Therefore, a small number of backend connections can process a large number of requests for short transactions and simple queries from many concurrent frontend connections. This is why you must focus on the queries per second (QPS) in instead of the number of concurrent connections.

Although the number of frontend connections is considered unlimited, the maximum number of connections maintained in the backend connection pool of a instance is limited. For more information, see Backend connection. Take note of the following points in actual scenarios:

- Avoid long or large transactions on applications. These transactions occupy many or even all backend connections when they are not committed or rolled back for a long time. This reduces the overall concurrent processing capability, and increases the connection pool pressure and response time (RT).
- Monitor and optimize or remove slow queries in the instance to prevent them from occupying an excessive number of backend connections. Otherwise, connections in the backend connection pool are insufficient or the maximum number of backend connections is reached. In this case, the instance or the RDS instance is under greater processing pressure. This may lead to reduced concurrent processing capability, increased RT, or a higher SQL execution failure rate due to execution timeout. For more information about how to troubleshoot and optimize slow queries, see Details about a low SQL statement and Overview.
- Assume that connections are used and queries are executed in a proper way. In this case, if the maximum number of connections in the backend connection pool of the instance is reached, contact Customer Services.

12.1.15.7. Upgrade instance specifications

Database performance can be measured by the response time (RT) and queries per second (QPS). RT reflects the performance of a single SQL statement. You can optimize SQL statements to solve this type of performance problem. When you upgrade the specifications of your instance, the capacity is expanded to improve performance. You can upgrade the specifications for database access business with low latency and high QPS.

The performance of a instance depends on the performance of and ApsaraDB RDS for MySQL. Insufficient performance of a or ApsaraDB RDS for MySQL node can create a bottleneck in the overall performance. This topic describes how to monitor the performance metrics of a instance and upgrade specifications of the PolarDB-X instance to solve performance bottlenecks. For more information about how to determine the performance of an ApsaraDB RDS for MySQL instance and upgrade the ApsaraDB RDS for MySQL instance, see the ApsaraDB RDS for MySQL documentation.

Determine a performance bottleneck of a instance

The QPS and CPU performance of a instance are in positive correlation. If a instance encounters a performance bottleneck, the CPU utilization of the PolarDB-X instance remains high.

Monitor the CPU utilization

1. Open the **Basic Information** page of your instance. In the left-side navigation pane, choose **Monitoring and Alerts > Instance Monitoring**.
2. On the Instance Monitoring page, select a monitoring dimension and the corresponding metrics to view details.

If the CPU utilization **exceeds 90%** or **remains higher than 80%**, the PolarDB-X instance faces a performance bottleneck. If the ApsaraDB RDS for MySQL instance does not encounter any bottlenecks, the current instance specifications cannot meet the QPS performance requirements of the business. Then, you can upgrade the specifications of the PolarDB-X instance.

For more information about performance-related service monitoring scenarios and methods of configuring a CPU utilization alert for a instance.

Upgrade the specifications of your instance

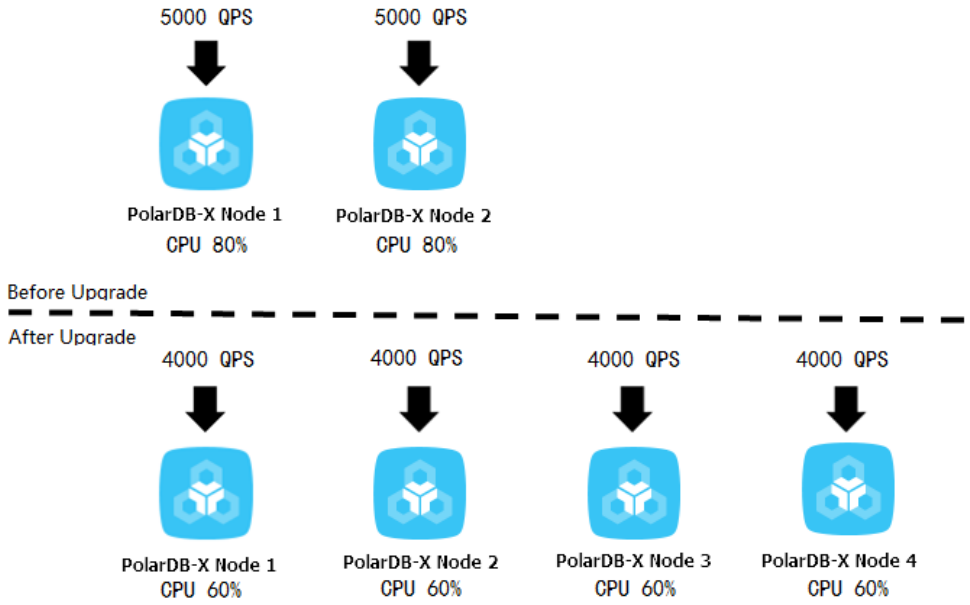
QPS is an important metric to determine whether the instance specifications can meet the business requirements. Each type of instance specification corresponds to a reference QPS value.

Some special SQL statements require more computing, such as temporary table sorting and aggregate computing, in . Therefore, the QPS supported by each instance is lower than the standard value specified in the specifications.

Specifications upgrade of a instance improves the processing performance of the instance by adding nodes to share the QPS. This upgrade method linearly improves the performance of the instance, because nodes are stateless.

For example, Business A requires about 15,000 QPS. The current instance has 4 vCPUs, 4 GB of memory, and 2 nodes, which support only 10,000 QPS. In addition, the CPU utilization of the instance remains high. You can upgrade the instance to 8 vCPUs and 8 GB of memory. After the upgrade, each node handles about 4,000 QPS. Then, the performance meets business needs, and the CPU utilization drops to a reasonable level. The following figure shows the procedure.

specifications upgrade



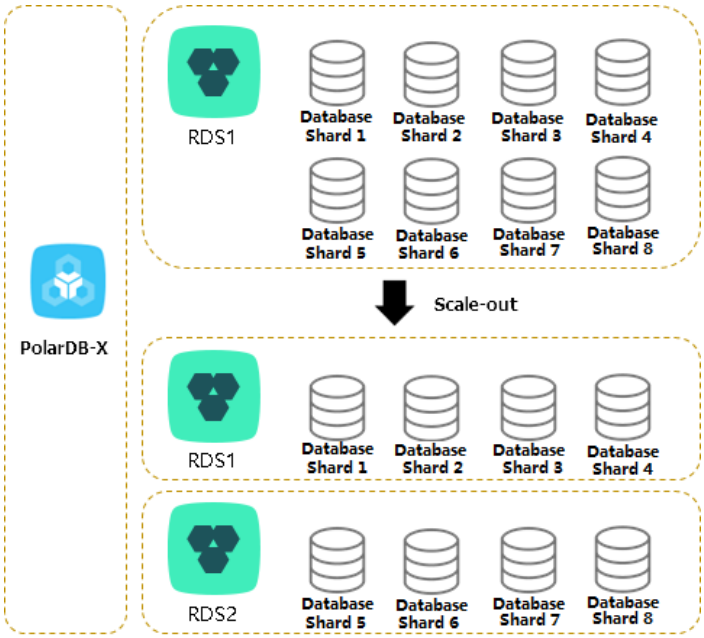
For more information about how to upgrade the specifications of a PolarDB-X instance, see Change specifications.

12.1.15.8. Perform scale-out

In , smooth scale-out improves the overall performance by increasing the number of ApsaraDB RDS for MySQL instances. You can increase the number of ApsaraDB RDS for MySQL instances by performing smooth scale-out to increase the database capacity when the following conditions are met: 1. The IOPS, CPU utilization, disk space, and other metrics of the ApsaraDB RDS for MySQL instance reach their bottlenecks. 2. You are unable to remove the bottlenecks by optimizing SQL statements or upgrading ApsaraDB RDS for MySQL specifications. For example, the disk has been upgraded to the top configuration.

smooth scale-out reduces the pressure on the original ApsaraDB RDS for MySQL instance by migrating database shards to the new ApsaraDB RDS for MySQL instance. Assume that all the eight databases are deployed in one ApsaraDB RDS for MySQL instance before scale-out. After scale-out, the eight databases are deployed in two ApsaraDB RDS for MySQL instances. Therefore, the pressure on a single ApsaraDB RDS for MySQL instance is significantly reduced. The following figure shows the procedure.

scale-out



After multiple scale-out operations, the number of ApsaraDB RDS for MySQL instances may be equal to the number of database shards. In this case, you need to create another instance and ApsaraDB RDS for MySQL databases with the expected capacity, and migrate data to further increase the data capacity. This procedure is complex. We recommend that you consider the data growth expected in the next two to three years and plan the number of ApsaraDB RDS for MySQL instances when you create a database.

Determine whether scale-out is required

You can determine whether smooth scale-out is required based on three ApsaraDB RDS for MySQL metrics: IOPS, CPU utilization, and disk space. You can view these metrics in the ApsaraDB RDS for MySQL console. For more information, see **User Guide > Monitoring and alerting > View monitoring data of system resources and engines** in the *ApsaraDB RDS documentation*.

IOPS and CPU utilization

If you find that the IOPS or CPU utilization remains higher than 80% for a long time or you frequently receive alerts, perform the following steps:

1. Optimize SQL statements. In most cases, you can solve the high CPU utilization problem by using this method. For more information, see Overview.
2. If the problem persists, upgrade the ApsaraDB RDS for MySQL instance. For more information, see **User Guide > Instance management > Change the configurations** in the *ApsaraDB RDS documentation*.
3. When the CPU utilization or IOPS exceeds the threshold, you can set read-only databases to share the load on the primary database. However, read/write splitting affects read consistency. For more information, see the

Read/write splitting documentation.

4. If the problem persists, scale out the instance.

Disk space

ApsaraDB RDS for MySQL has the following types of disk space:

1. Data space: the space occupied by data. The disk usage continues to increase as more data is inserted. We recommend that you keep the remaining disk space higher than 30%.
2. System file space: the space occupied by shared tables and error log files.
3. Binary log file space: the space occupied by binary logs generated during database operation. The more the update transactions, the larger the occupied space.

Whether scale-out is required depends on the data space. When the data space is about to or expected to exceed the disk capacity, you can distribute the data to the database shards on multiple ApsaraDB RDS for MySQL instances by performing scale-out.

Scale-out risks and precautions

scale-out consists of four steps: **configuration > migration > switchover > cleanup**. For more information, see the Perform smooth scale-out documentation.

Take note of the following points before scale-out:

- To reduce the pressure of read operations on the source ApsaraDB RDS for MySQL instance, perform scale-out when the load on the source ApsaraDB RDS for MySQL instance is low.
- During scale-out, do not submit any data definition language (DDL) tasks in the console or connect to the instance to directly execute DDL SQL statements. Otherwise, the scale-out task may fail.
- Scale-out requires that the source table have a primary key. If the source table does not have a primary key, add one first.
- During scale-out, the read and write traffic is switched to the new ApsaraDB RDS for MySQL instance. The switchover process takes 3 to 5 minutes. We recommend that you perform a switchover during off-peak hours.
- Scale-out does not affect the instance before the switchover. Therefore, you can cancel the scale-out by performing a rollback before the switchover.
- The scale-out operation creates pressure on databases. We recommend that you perform this operation during off-peak hours.

12.1.15.9. Troubleshoot slow SQL statements in PolarDB-X

12.1.15.9.1. Details about a low SQL statement

defines an SQL statement that takes more than 1 second to execute as a slow SQL statement. Slow SQL statements in are classified into logical slow SQL statements and physical slow SQL statements. In , an SQL statement is executed step by step on and ApsaraDB RDS for MySQL nodes. Large execution loss on a node will result in slow SQL statements.

- Logical slow SQL statements are slow SQL statements sent from an application to your instance.
- Physical slow SQL statements are slow SQL statements sent from your instance to ApsaraDB RDS for MySQL instance.

Syntax

```
SHOW FULL {SLOW | PHYSICAL_SLOW} [WHERE where_condition]
                                 [ORDER BY col_name [ASC | DESC], ...]
                                 [LIMIT {[offset,] row_count | row_count OFFSET offset}]
```

Description

The `SHOW FULL SLOW` statement shows logical slow SQL statements, which are SQL statements sent from an application to your instance.

The result set of the `SHOW FULL SLOW` statement contains the following columns:

- TRACE_ID: the unique identifier of the SQL statement. A logical SQL statement and the physical SQL statements generated by this logical SQL statement have the same TRACE_ID value. The TRACE_ID value is also sent as a comment to your ApsaraDB RDS for MySQL instance.
- HOST: the IP address of the client that sends the SQL statement.

  > **Notice** The client IP address may not be obtained when you use a virtual private cloud (VPC).

- START_TIME: the time when the instance starts to execute the SQL statement.
- EXECUTE_TIME: the time consumed by the instance to execute the SQL statement.
- AFFECT_ROW: the number of records returned or the number of rows affected by the SQL statement.
- SQL: the statement that is executed.

The `SHOW FULL PHYSICAL_SLOW` statement shows the physical slow SQL statements, which are SQL statements sent from your instance to ApsaraDB RDS for MySQL instance.

The result set of the `SHOW FULL PHYSICAL_SLOW` statement contains the following columns:

- TRACE_ID: the unique identifier of the SQL statement. A logical SQL statement and the physical SQL statements generated by this logical SQL statement have the same TRACE_ID value. The TRACE_ID value is also sent as a comment to your ApsaraDB RDS for MySQL instance.
- GROUP_NAME: the name of a database group. Grouping aims to manage multiple groups of databases that have identical data. For example, the databases can be the primary and secondary databases after data replication that is implemented by ApsaraDB RDS for MySQL. It is used for read/write splitting and primary/secondary switchovers.
- DBKEY_NAME: the name of the database shard on which the SQL statement is executed.
- START_TIME: the time when the instance starts to execute the SQL statement.
- EXECUTE_TIME: the time consumed by the instance to execute the SQL statement.
- SQL_EXECUTE_TIME: the time consumed by the instance to call the ApsaraDB RDS for MySQL instance to execute the SQL statement on the database shard.
- GETLOCK_CONNECTION_TIME: the time that the instance takes to obtain a connection from the connection pool. If the value is large, the ApsaraDB RDS for MySQL connections are exhausted. This is typically due to a large number of slow SQL statements. You can log on to the corresponding ApsaraDB RDS for MySQL instance and execute the SHOW PROCESSLIST statement to locate slow SQL statements.
- CREATE_CONNECTION_TIME: the time that the instance takes to establish a connection to the ApsaraDB RDS for MySQL instance. If the value is large, it is largely because the ApsaraDB RDS for MySQL instance is overloaded or faulty.
- AFFECT_ROW: the number of records returned or the number of rows affected by the SQL statement.
- SQL: the statement that is executed.

Example 1

The following example describes how to locate the execution information of a slow SQL statement on a instance and between and ApsaraDB RDS for MySQL instances.

1. You can query the specified slow SQL statement by setting conditions such as execution time and SQL string match: Assume that you execute the `SHOW FULL SLOW where `SQL` like '%select sleep(50)%';` statement. The following response is returned:

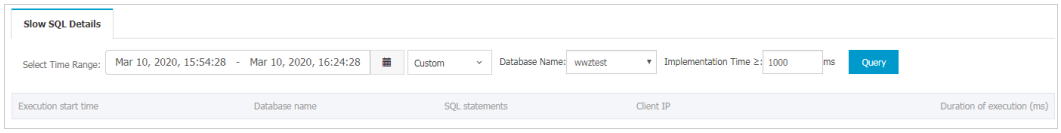
```
+-----------------+-----------+-------------------------+--------------+------------+-------------
-----+
| TRACE_ID        | HOST      | START_TIME              | EXECUTE_TIME | AFFECT_ROW | SQL
|
+-----------------+-----------+-------------------------+--------------+------------+-------------
-----+
| ae0e565b8c00000 | 127.0.0.1 | 2017-03-29 19:28:43.028 |        50009 |          1 | select sleep
(50) |
+-----------------+-----------+-------------------------+--------------+------------+-------------
-----+
1 row in set (0.02 sec)
```

2. Based on the value of TRACE_ID that you obtain from the slow SQL statement, execute the `SHOW FULL PHYSICAL_SLOW where trace_id = 'ae0e565b8c00000';` to query the physical execution information of this SQL statement.

```
+-----------------+-----------------------------------------------------+-------------------------
---------------------+-------------------------+--------------+------------------+-----------------
--------+-------------------------+------------+------------------+
| TRACE_ID        | GROUP_NAME                                          | DBKEY_NAME
| START_TIME              | EXECUTE_TIME | SQL_EXECUTE_TIME | GETLOCK_CONNECTION_TIME | CREATE_CONN
ECTION_TIME | AFFECT_ROW | SQL              |
+-----------------+-----------------------------------------------------+-------------------------
---------------------+-------------------------+--------------+------------------+-----------------
--------+-------------------------+------------+------------------+
| ae0e565b8c00000 | PRIV_TEST_1489167306631PJAFPRIV_TEST_APKK_0000_RDS | rdso6g5b6206sdq832ow_priv
_test_apkk_0000_nfup | 2017-03-29 19:27:53.02 |        50001 |            50001 |
0 |                       0 |          1 | select sleep(50) |
+-----------------+-----------------------------------------------------+-------------------------
---------------------+-------------------------+--------------+------------------+-----------------
--------+-------------------------+------------+------------------+
1 row in set (0.01 sec)
```

3. In the SQL statement details and slow SQL statement records of the ApsaraDB RDS for MySQL instance, you can query the execution information of this SQL statement on the ApsaraDB RDS for MySQL instance based on the value of the TRACE_ID parameter.

   Slow query logs



Example 2

This example describes how to locate the original SQL statement in a instance based on the slow SQL statement that you locate in the ApsaraDB RDS for MySQL instance.

1. The slow query log generated by the ApsaraDB RDS for MySQL instance shows the TRACE_ID value of the slow SQL query. Assume that the value is ae0e55660c00000.
2. Based on the TRACE_ID value that you obtain in Step 1, execute the `SHOW FULL PHYSICAL_SLOW` statement to obtain the physical execution information of this SQL statement. Execute the `SHOW FULL PHYSICAL_SLOW where trace_id = 'ae0e55660c00000';` statement. The following response is returned:

```
+----------------+-------------------------------------------------------+-------------------------
--------------------+-------------------------+-------------+------------------+----------------
---------+-------------------------+------------+------------------+
| TRACE_ID        | GROUP_NAME                                            | DBKEY_NAME
| START_TIME              | EXECUTE_TIME | SQL_EXECUTE_TIME | GETLOCK_CONNECTION_TIME | CREATE_CON
NECTION_TIME | AFFECT_ROW | SQL              |
+----------------+-------------------------------------------------------+-------------------------
--------------------+-------------------------+-------------+------------------+----------------
---------+-------------------------+------------+------------------+
| ae0e55660c00000 | PRIV_TEST_1489167306631PJAFPRIV_TEST_APKK_0000_RDS | rdso6g5b6206sdq832ow_priv
_test_apkk_0000_nfup | 2017-03-29 19:27:37.308 |        10003 |            10001 |
0 |                      0 |          1 | select sleep(10) |
+----------------+-------------------------------------------------------+-------------------------
--------------------+-------------------------+-------------+------------------+----------------
---------+-------------------------+------------+------------------+
1 row in set (0.02 sec)
```

12.1.15.9.2. Locate slow SQL statements

You can locate a slow SQL statement in two ways. You can obtain historical information about a slow SQL statement from slow SQL statement records. Alternatively, you can execute the `SHOW PROCESSLIST` statement to obtain the real-time execution information about a slow SQL statement.

Perform the following steps to troubleshoot slow SQL statements:

1. Locate slow SQL statements.
2. Locate nodes with performance loss.
3. Troubleshoot the performance loss.

> **Note** During troubleshooting, we recommend that you execute the `mysql -hIP -PPORT -uUSER -pPASSWORD -c` statement in the MySQL command-line client to establish the connection. Be sure to add `-c` to prevent the MySQL command-line client from filtering out the comments (default operation) and therefore affecting the execution of hints.

- View slow SQL statement records

  Execute the following statement to query top 10 slow SQL statements. This statement can query logical SQL statements in a instance. One logical SQL statement corresponds to SQL statements that are executed on one or more databases or tables of the ApsaraDB RDS for MySQL instance. For more information, see Details about a low SQL statement.

  Execute the `SHOW SLOW limit 10;` statement. The following response is returned:

```
+----------------+-------------+-------------------------+-------------+------------+-----------
-----------------------------------------------------+
| TRACE_ID        | HOST         | START_TIME              | EXECUTE_TIME | AFFECT_ROW | SQL
|
+----------------+-------------+-------------------------+-------------+------------+-----------
-----------------------------------------------------+
| ac3133132801001 | xx.xxx.xx.97 | 2017-03-06 15:48:32.330 |       900392 |         -1 | select det
ail_url, sum(price) from t_item group by detail_url;  |
......
+----------------+-------------+-------------------------+-------------+------------+-----------
-----------------------------------------------------+
10 rows in set (0.01 sec)
```

- View real-time SQL execution information

If the execution of an SQL statement is slow in the current server, execute the **SHOW PROCESSLIST** statement to view the real-time SQL execution information in the current database. The value in the TIME column indicates how long the current SQL statement has been executed.

Execute the `SHOW PROCESSLIST WHERE COMMAND ! ='Sleep';` statement. The following response is returned:

```
+---------------+-----------+--------------------+-------------+---------+-------------------------
-----------------------------------------------+-----------------------------------------------------------
-------------------------------------------------+-----------+---------------+-----------+
| ID            | USER      | DB                 | COMMAND     | TIME    | STATE
| INFO
| ROWS_SENT | ROWS_EXAMINED | ROWS_READ |
+---------------+-----------+--------------------+-------------+---------+-------------------------
-----------------------------------------------+-----------------------------------------------------------
-------------------------------------------------+-----------+---------------+-----------+
| 0-0-352724126 | ifisibhk0 | test_123_wvvp_0000 | Query       |     13  | Sending data
| /*DRDS /42.120.74.88/ac47e5a72801000/ */select `t_item`.`detail_url`,SUM(`t_item`.`price`) from `
t_i |      NULL |          NULL |      NULL |
| 0-0-352864311 | cowxhthg0 | NULL               | Binlog Dump |     17  | Master has sent all binl
og to slave; waiting for binlog to be updated | NULL
|      NULL |          NULL |      NULL |
| 0-0-402714795 | ifisibhk0 | test_123_wvvp_0005 | Alter       |     114 | Sending data
| /*DRDS /42.120.74.88/ac47e5a72801000/ */ALTER TABLE `Persons` ADD `Birthday` date
|      NULL |          NULL |      NULL |
......
+---------------+-----------+--------------------+-------------+---------+-------------------------
-----------------------------------------------+-----------------------------------------------------------
-------------------------------------------------+-----------+---------------+-----------+
12 rows in set (0.03 sec)
```

The following list describes each column:

- ID: the ID of the connection.
- USER: the username of the database shard on which this SQL statement is executed.
- DB: the specified database. If no database is specified, the value is NULL.
- COMMAND: the type of the statement that is being executed. SLEEP: indicates an idle connection. For more information about other details, see MySQL thread information documentation.
- TIME: the elapsed execution time of the SQL statement, in seconds.
- STATE: the current execution status. For more information, see MySQL thread status documentation.
- INFO: the information used to derive the complete SQL statement. The SQL statement that is being executed may be so long that it cannot be completely displayed. You can derive the complete SQL statement based on information such as business parameters.

In this example, the following slow SQL statement is located:

```
ALTER TABLE `Persons` ADD `Birthday` date
```

12.1.15.9.3. Locate nodes with performance loss

When you locate a slow SQL statement in slow SQL statement records or real-time SQL execution information, you can run the TRACE command to trace the running time of the SQL statement in and ApsaraDB RDS for MySQL to locate the bottleneck.

The TRACE command actually runs the SQL statement, records the time consumed on all nodes, and returns the execution result. For more information about TRACE and other control commands, see Help statements.

> **Note** The TRACE command needs to maintain the context information of the connection. Some GUI clients may use connection pools, which results in command exceptions. Therefore, we recommend that you use the MySQL command line to run the TRACE command.

Run the following command for the identified slow SQL statement:

```
mysql> trace select detail_url, sum(distinct price) from t_item group by detail_url;
    +---------------+---------------+
    | detail_url    | sum(price)    |
    +---------------+---------------+
    | www.xxx.com   | 1084326800.00 |
    | www.xx1.com   | 1084326800.00 |
    | www.xx2.com   | 1084326800.00 |
    | www.xx3.com   | 1084326800.00 |
    | www.xx4.com   | 1084326800.00 |
    | www.xx5.com   | 1084326800.00 |
    ......
    +---------------+---------------+
    1 row in set (7 min 2.72 sec)
```

After the TRACE command is run, run SHOW TRACE to view the result. You can identify the bottleneck of the slow SQL statement based on the time consumption of each component.

```
 mysql> SHOW TRACE;
    +------+------------+--------------+-----------------------------------------------------------------------+---
----------------------------------------+---------------+--------------------------+------+--------
----------------------------------------------------------------------------------------------------------------
---------------------------------+--------+
    | ID   | TIMESTAMP  | TYPE         | GROUP_NAME                                                            | DB
KEY_NAME                                | TIME_COST(MS) | CONNECTION_TIME_COST(MS) | ROWS | STATEME
NT
| PARAMS |
    +------+------------+--------------+-----------------------------------------------------------------------+---
----------------------------------------+---------------+--------------------------+------+--------
----------------------------------------------------------------------------------------------------------------
---------------------------------+--------+
    |    0 |      0.000 | Optimize     | DRDS                                                                  | DR
DS                                      | 2             | 0.00                     |    0 | select
detail_url, sum(price) from t_item group by detail_url
| NULL   |
    |    1 | 423507.342 | Merge Sorted | DRDS                                                                  | DR
DS                                      | 411307        | 0.00                     |    8 | Using M
erge Sorted, Order By (`t_item`.`detail_url` asc )
| NULL   |
    |    2 |      2.378 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0003_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0003_hbpz | 15            | 1.59                     |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    3 |      2.731 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0000_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0000_hbpz | 11            | 1.78                     |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    4 |      2.933 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0004_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0004_hbpz | 15            | 1.48                     |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    5 |      3.111 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0001_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0001_hbpz | 15            | 1.56                     |    1 | select
```

```
 t_item` . detail_url ,SUM(distinct `t_item` . price ) from `t_item` group by `t_item` . detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    6 |      3.323 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0007_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0007_hbpz | 15            | 1.54                      |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    7 |      3.496 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0006_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0006_hbpz | 18            | 1.30                      |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    8 |      3.505 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0005_hbpz | 423507        | 1.97                      |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |    9 |      3.686 | Query        | TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0002_RDS | rd
so6g5b6206sdq832ow_test_123_wvvp_0002_hbpz | 14            | 1.47                      |    1 | select
`t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` orde
r by `t_item`.`detail_url` asc  | NULL   |
    |   10 | 423807.906 | Aggregate    | DRDS                                                                | DR
DS                                          | 1413          | 0.00                      |    1 | Aggrega
te Function (SUM(`t_item`.`price`)), Group By (`t_item`.`detail_url` asc )
| NULL   |
    +------+------------+--------------+---------------------------------------------------------------------+---
-----------------------------------------+---------------+---------------------------+------+--------
-----------------------------------------------------------------------------------------------------------
---------------------------------+--------+
    11 rows in set (0.01 sec)
```

In the returned results of SHOW TRACE, you can determine which node has a long execution time based on the values (in milliseconds) in the TIME_COST column. You can also see the corresponding GROUP_NAME (that is, the or ApsaraDB RDS for MySQL node) and the STATEMENT column information (that is, the SQL statement being executed). By checking whether the value of GROUP_NAME is , you can determine whether the slow node exists in or ApsaraDB RDS for MySQL.

According to the preceding results, the Merge Sorted action on the node and the TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS node of ApsaraDB RDS for MySQL take a lot of time.

12.1.15.9.4. Troubleshoot the performance loss

Slow nodes may exist on the or ApsaraDB RDS for MySQL instance. Troubleshoot the fault accordingly after the cause is determined.

Solution for slow nodes

When the `GROUP_NAME` of a slow node is in the instance, check whether time-consuming computing operations such as Merge Sorted, Temp Table Merge, and Aggregate exist during SQL statement execution. If so, rectify it. For more information, see Overview.

Solution for slow ApsaraDB RDS for MySQL nodes

When the slow node is on the ApsaraDB RDS for MySQL instance, check the execution plan of this SQL statement on the ApsaraDB RDS for MySQL instance.

In , you can run `/! TDDL:node={GROUP_NAME}*/ EXPLAIN` to check the SQL execution plan of an ApsaraDB RDS for MySQL instance. The execution plan displays the SQL execution process information, including inter-table association and index information.

The detailed process is as follows:

1. Based on `GROUP_NAME` , assemble the HINT: `/! TDDL:node='TEST_123_1488766060743ACTJSANGUAN_TEST_123_W`

`VVP_0005_RDS'*/` .

2. Combine the assembled HINT and the statement prefixed by EXPLAIN to form a new SQL statement and run it. The EXPLAIN command does not actually run. It only displays the execution plan of the SQL statement.

The following example describes how to query the execution plan of the identified slow node.

```
    mysql> /! TDDL:node='TEST_123_1488766060743ACTJSANGUAN_TEST_123_WVVP_0005_RDS'*/ EXPLAIN  select `
t_item`.`detail_url`,SUM(distinct `t_item`.`price`) from `t_item` group by `t_item`.`detail_url` order
by `t_item`.`detail_url` asc;
    +----+-------------+--------+------+---------------+-----+---------+-----+---------+--------------
------------------+
    | id | select_type | table  | type | possible_keys | key | key_len | ref | rows    | Extra
|
    +----+-------------+--------+------+---------------+-----+---------+-----+---------+--------------
------------------+
    |  1 | SIMPLE      | t_item | ALL  | NULL          |NULL | NULL    | NULL| 1322263 | Using tempora
ry; Using filesort |
    +----+-------------+--------+------+---------------+-----+---------+-----+---------+--------------
------------------+
    1 row in set (0.01 sec)
```

When the preceding SQL statement is run in ApsaraDB RDS for MySQL, the message `Using temporary; Using filesort` is returned. It indicates that low SQL statement execution is caused by improper use of the index. In this case, you can correct the index and run the SQL statement again.

12.1.15.10. Handle DDL exceptions

When you run any data definition language (DDL) commands of , PolarDB-X performs the corresponding DDL operation on all table shards.

Failures can be divided into two types:

1. A DDL statement fails to be executed in a database shard. DDL execution failure in any database shard may result in inconsistent table shard structures.
2. The system does not respond for a long time after a DDL statement is executed. When you perform a DDL statement on a large table, the system may make no response for a long time due to the long execution time of the DDL statement in a database shard.

Execution failures in database shards may occur for various reasons. For example, the table you want to create already exists, the column you want to add already exists, or the disk space is insufficient.

No response for a long time is generally caused by the long execution time of a DDL statement in a database shard. Taking ApsaraDB RDS for MySQL as an example, the DDL execution time depends mostly on whether the operation is an in-place (directly modifying the source table) or copy (copying data in the table) operation. An in-place operation only requires modification of metadata, while a copy operation reconstructs the whole table and also involves log and buffer operations.

To determine whether a DDL operation is an in-place or copy operation, you can view the returned value of "rows affected" after the operation is completed.

Example:

- Change the default value of a column (this operation is very fast and does not affect the table data at all):

```
Query OK, 0 rows affected (0.07 sec)
```

- Add an index (this operation takes some time, but "0 rows affected" indicates that the table data is not replicated):

```
Query OK, 0 rows affected (21.42 sec)
```

- Change the data type of column (this operation takes a long time and reconstructs all data rows in the table):

```
Query OK, 1671168 rows affected (1 min 35.54 sec)
```

Therefore, before executing a DDL operation on a large table, perform the following steps to determine whether the operation is a fast or slow operation:

1. Copy the table structure to generate a cloned table.
2. Insert some data.
3. Perform the DDL operation on the cloned table.
4. Check whether the value of "rows affected" is 0 after the operation is completed. A non-zero value means that this operation reconstructs the entire table. In this case, you need to perform this operation in off-peak hours.

Solution for failures

DDL operations distribute all SQL statements to all database shards for parallel execution. Execution failure on any database shard does not affect the execution on other database shards. In addition, provides the CHECK TABLE command to check the structure consistency of the table shards. Therefore, failed DDL operations can be performed again, and errors reported on database shards on which the operations have been executed do not affect the execution on other database shards. Make sure that all table shards ultimately have the same structure.

Procedure for handling DDL operation failures

1. Run the **CHECK TABLE** command to check the table structure. If the returned result contains only one row and the status is normal, **the table statuses are consistent**. In this case, go to Step 2. Otherwise, go to Step 3.
2. Run the **SHOW CREATE TABLE** command to check the table structure. If the displayed table structure is the same as the expected structure after the DDL statement is run, the DDL statement is run. Otherwise, go to Step 3.
3. Run the **SHOW PROCESSLIST** command to check the statuses of all SQL statements being executed. If any ongoing DDL operations are detected, wait until these operations are completed, and then perform Steps 1 and 2 to check the table structure. Otherwise, go to Step 4.
4. Perform the DDL operation again on . If the **Lock conflict** error is reported, go to Step 5. Otherwise, go to Step 3.
5. Run the **RELEASE DBLOCK** command to release the DDL operation lock, and then go to Step 4.

The procedure is as follows:

1. Check the table structure consistency

   Run the CHECK TABLE command to check the table structure. When the returned result contains only one row and the displayed status is OK, **the table structures are consistent**.

   > **Notice** If no result is returned after you run CHECK TABLE, retry by using the CLI.

```
mysql> check table `xxxx`;
+----------------------------+-------+----------+----------+
| TABLE                      | OP    | MSG_TYPE | MSG_TEXT |
+----------------------------+-------+----------+----------+
| TDDL5_APP.xxxx             | check | status   | OK       |
+----------------------------+-------+----------+----------+
1 row in set (0.05 sec)
```

2. Check the table structure

   Run the SHOW CREATE TABLE command to check the table structure. If **table structures are consistent** and **correct**, the DDL statement has been run.

```
mysql> show create table `xxxx`;
+---------+-----------------------------------------------------------------------------------------------
--------------------------+
| Table   | Create Table
|
+---------+-----------------------------------------------------------------------------------------------
--------------------------+
|  xxxx   | CREATE TABLE `xxxx` (
`id` int(11) NOT NULL DEFAULT '0',
`NAME` varchar(1024) NOT NULL DEFAULT '',
PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 dbpartition by hash(`id`) tbpartition by hash(`id`) tbpartiti
ons 3                     |
+---------+-----------------------------------------------------------------------------------------------
--------------------------+
1 row in set (0.05 sec)
```

3. Check the SQL statements being executed.

   If some DDL statement executions are slow and no response is received for a long time, you can run the SHOW PROCESSLIST command to check the status of all SQL statements being executed.

```
mysql> SHOW PROCESSLIST WHERE COMMAND ! = 'Sleep';
+---------------+-----------+--------------------+-------------+---------+-------------------------
-----------------------------------------------+-----------------------------------------------------
-----------------------------------------------+-----------+---------------+-----------+
| ID            | USER      | DB                 | COMMAND     | TIME    | STATE
| INFO
| ROWS_SENT | ROWS_EXAMINED | ROWS_READ |
+---------------+-----------+--------------------+-------------+---------+-------------------------
-----------------------------------------------+-----------------------------------------------------
-----------------------------------------------+-----------+---------------+-----------+
| 0-0-352724126 | ifisibhk0 | test_123_wvvp_0000 | Query       |      15 | Sending data
| /*DRDS /xx.xxx.xx.88/ac47e5a72801000/ */select `t_item`.`detail_url`,SUM(`t_item`.`price`) from
`t_i |      NULL |          NULL |      NULL |
| 0-0-352864311 | cowxhthg0 | NULL               | Binlog Dump |      13 | Master has sent all bin
log to slave; waiting for binlog to be updated | NULL
|      NULL |          NULL |      NULL |
| 0-0-402714566 | ifisibhk0 | test_123_wvvp_0005 | Query       |      14 | Sending data
| /*DRDS /xx.xxx.xx.88/ac47e5a72801000/ */select `t_item`.`detail_url`,`t_item`.`price` from `t_i
|      NULL |          NULL |      NULL |
| 0-0-402714795 | ifisibhk0 | test_123_wvvp_0005 | Alter       |     114 | Sending data
| /*DRDS /xx.xxx.xx.88/ac47e5a72801000/ */ALTER TABLE `Persons` ADD `Birthday` date
|      NULL |          NULL |      NULL |
......
+---------------+-----------+--------------------+-------------+---------+-------------------------
-----------------------------------------------+-----------------------------------------------------
-----------------------------------------------+-----------+---------------+-----------+
12 rows in set (0.03 sec)
```

   The value in the TIME column indicates the number of seconds that the command has been executed. If a command execution is too slow, as shown in the figure, you can run the **KILL '0-0-402714795'** command to cancel the slow command.

> **Notice** In , one logical SQL statement corresponds to multiple statements on database shards. Therefore, you may need to kill multiple commands to stop a logical DDL statement. You can determine the logical SQL statement to which a command belongs based on the INFO column in the SHOW PROCESSLIST result set.

4. Handle the lock conflict error

   adds a database lock before performing a DDL operation and releases the lock after the operation. The KILL DDL operation may not release the lock. If you perform the DDL operation again, the following error message will be returned:

   ```
   Lock conflict , maybe last DDL is still running
   ```

   In this case, run **RELEASE DBLOCK** to release the lock. After the command is canceled and the lock is released, run the DDL statement again during off-peak hours or when the service is stopped.

Other problems

Clients cannot display the modified table structures.

To enable some clients to obtain table structures from system tables (such as COLUMNS or TABLES), creates a shadow database in database shard 0 on your ApsaraDB RDS for MySQL instance. The shadow database name must be the same as the name of your logical database. It stores all table structures and other information in the user database.

The client obtains the table structure from the system table of the shadow database. During the processing of DDL exceptions, the table structure may be modified normally in the user database but not in the shadow database due to some reasons. In this case, you need to connect to the shadow database and perform the DDL operation on the table again in the database.

Notice The CHECK TABLE command does not check whether the table structure in the shadow database is consistent with that in the user database.

12.1.15.11. Efficiently scan PolarDB-X data

supports efficient data scanning and uses aggregate functions for statistical summary during full table scan.

The following describes common scanning scenarios:

- **Scan of tables without database or table shards**: transmits the original SQL statement to the backend ApsaraDB RDS for MySQL database for execution. In this case, supports any aggregate functions.
- **Non-full table scan**: transmits the original SQL statement to each single ApsaraDB RDS for MySQL database for execution. For example, when the shard key in the WHERE clause is Equal, non-full table scan is performed. In this case, PolarDB-X also supports any aggregate functions.
- **Full table scan**: Currently, the supported aggregate functions are COUNT, MAX, MIN, and SUM. In addition, LIKE, ORDER BY, LIMIT, and GROUP BY are also supported during full table scan.
- **Parallel scan of all table shards**: If you need to export data from all databases, you can run the SHOW command to view the table topology and scan all table shards in parallel. For more information, see the following section.

Traverse tables by using a hint

1. Run the SHOW TOPOLOGY FROM TABLE_NAME command to obtain the table topology.

   ```
   mysql:> SHOW TOPOLOGY FROM DRDS_USERS;
   +------+------------------+--------------+
   | ID   | GROUP_NAME       | TABLE_NAME   |
   +------+------------------+--------------+
   |    0 | DRDS_00_RDS      | drds_users   |
   |    1 | DRDS_01_RDS      | drds_users   |
   +------+------------------+--------------+
   2 rows in set (0.06 sec)
   ```

   By default, the non-partition table is stored in database shard 0.

2. Traverse each table for TOPOLOGY.
   i. Run the current SQL statement in database shard 0.

```
/! TDDL:node='DRDS_00_RDS'*/ SELECT * FROM DRDS_USERS;
```

   ii. Run the current SQL statement in database shard 1.

```
/! TDDL:node='DRDS_01_RDS'*/ SELECT * FROM DRDS_USERS;
```

> **Notice** We recommend that you run `SHOW TOPOLOGY FROM TABLE_NAME` to obtain the latest table topology before each scan.

Scan data in parallel

allows you to run mysqldump to export data. However, if you want to scan data faster, you can establish multiple sessions for each table shard to scan tables in parallel.

```
mysql> SHOW TOPOLOGY FROM LJLTEST;
+------+----------------+------------+
| ID   | GROUP_NAME     | TABLE_NAME |
+------+----------------+------------+
|    0 | TDDL5_00_GROUP | ljltest_00 |
|    1 | TDDL5_00_GROUP | ljltest_01 |
|    2 | TDDL5_00_GROUP | ljltest_02 |
|    3 | TDDL5_01_GROUP | ljltest_03 |
|    4 | TDDL5_01_GROUP | ljltest_04 |
|    5 | TDDL5_01_GROUP | ljltest_05 |
|    6 | TDDL5_02_GROUP | ljltest_06 |
|    7 | TDDL5_02_GROUP | ljltest_07 |
|    8 | TDDL5_02_GROUP | ljltest_08 |
|    9 | TDDL5_03_GROUP | ljltest_09 |
|   10 | TDDL5_03_GROUP | ljltest_10 |
|   11 | TDDL5_03_GROUP | ljltest_11 |
+------+----------------+------------+
12 rows in set (0.06 sec)
```

As shown above, the table has four database shards, and each database shard has three table shards. Run the following SQL statement to operate on the table shards of the TDDL5_00_GROUP database:

```
/! TDDL:node='TDDL5_00_GROUP'*/ select * from ljltest_00;
```

> **Note** TDDL5_00_GROUP in HINT corresponds to the GROUP_NAME column in the execution results of the SHOW TOPOLOGY command. In addition, the table name in the SQL statement is the table shard name.

At this time, you can establish up to 12 sessions (corresponding to 12 table shards respectively) to process data in parallel.

12.1.16. Appendix: PolarDB-X terms

This topic lists common terms of for your reference.

Term	Description	Remarks

Term	Description	Remarks
	is a distributed database service that was independently developed by Alibaba to solve the bottlenecks of single-instance database services. is compatible with MySQL protocols and syntax. It supports automatic horizontal splitting, online smooth scale-out, auto scaling, and transparent read/write splitting, and provides operations and maintenance (O&M) capabilities for distributed databases throughout their entire lifecycle.	-
TDDL	Taobao Distributed Data Layer (TDDL) was developed by Alibaba and has become a preferred component for nearly 1,000 core applications of Alibaba.	-
Console	The console allows database administrators (DBAs) to isolate resources as needed. You can perform operations, such as instance management, database and table management, read/write splitting configuration, smooth scale-out, monitoring data display, and the IP address whitelist.	-
DRDS Manager	DRDS Manager allows global O&M personnel and DBAs to manage all the resources and monitor the system.	-
Server	Server is the service layer of . Multiple server nodes constitute a server cluster to provide distributed database services, including read/write splitting, routed SQL execution, result merging, dynamic database configuration, and globally unique IDs (GUIDs).	-
Load balancer	server nodes are stateless. Therefore, a request can be routed to a server node in a random way. A load balancer is used to complete this task. Server Load Balancer (SLB) is used for integrated output by Apsara Stack. In normal cases, VIPServer is used for Alibaba middleware output.	-
Diamond	Diamond manages the configuration and storage of . It allows you to configure storage, queries, and notifications. In , Diamond stores the source data of databases and configuration data that includes sharding rules and switches.	-
Data Replication System	Data Replication System migrates and synchronizes data for . The core capabilities of Data Replication System include full data migration and incremental data synchronization. Its derived features include smooth data import, smooth scale-out, and global secondary indexes. Data Replication System requires the support of ZooKeeper and Rtools.	-
instance	A instance consists of multiple server nodes. A instance can contain multiple databases.	-
instance ID	A unique instance ID identifies an instance.	-
Number of nodes on a instance	The number of server nodes on a instance.	-
VIP	The virtual IP addresses (VIPs) of load balancers can be divided into the following types: • 1. Public VIP, which is accessible from the Internet and used for testing in normal cases. • 2. Private VIP, which is accessible only from Alibaba Cloud internal networks.	-

Term	Description	Remarks
VPC	In normal cases, virtual private cloud (VPC) is used for Alibaba Cloud.	-
Region	The region information such as China (Hangzhou) and China (Shanghai). This concept is generally used for Alibaba Cloud.	-
Azone	The zone information such Hangzhou Zone A. This concept is generally used for Alibaba Cloud.	-
Logical SQL statement	A logical SQL statement is an SQL statement sent from an application to .	-
Physical SQL statement	A physical SQL statement is an SQL statement that is obtained after parses a logical SQL statement and sends it to ApsaraDB RDS for MySQL for execution.	Logical SQL statements and physical SQL statements may be the same or different. Logical SQL statements and physical SQL statements may be in a one-to-one or one-to-many mapping.
QPS	The queries per second (QPS) is the average number of logical SQL statements executed by per second in a statistical period.	The QPS does not indicate the number of transactions. Most control statements, such as COMMIT and SET statements, are not counted in QPS.
RT	The response time (RT) is the average response time of logical SQL statements executed by in a statistical period. The unit is millisecond. Use the following formula to calculate the RT of an SQL statement: (Time when writes the last packet of the result set) - (Time when receives the SQL statement)	-
Physical QPS	The physical QPS is the average number of physical SQL statements that executes on ApsaraDB RDS for MySQL per second in a statistical period.	-
Physical RT	The physical RT is the average response time of physical SQL statements executed by on ApsaraDB RDS for MySQL in a statistical period. The unit is millisecond. Use the following formula to calculate the RT of a physical SQL statement: (Time when receives the result set returned by ApsaraDB RDS for MySQL) - (Time when starts to obtain the connection to ApsaraDB RDS for MySQL)	This includes the time of establishing a connection to ApsaraDB RDS for MySQL or obtaining a connection from the connection pool, the network transmission time, and the time of executing the SQL statement by ApsaraDB RDS for MySQL.

Term	Description	Remarks
Connections	The number of connections established between the application and .	This does not indicate the number of connections established between and ApsaraDB RDS for MySQL.
Inbound traffic	The network traffic generated when the application sends SQL statements to .	This traffic is irrelevant to the traffic used for interaction between and ApsaraDB RDS for MySQL.
Outbound traffic	The network traffic generated when sends the result set to the application.	This traffic is irrelevant to the traffic used for interaction between and ApsaraDB RDS for MySQL.
ThreadRunning	The number of threads running on a instance. This parameter can be used to indicate the load of the instance.	-
Global	The total monitoring data of all the databases on a instance.	-
Memory usage	The Java Virtual Machine (JVM) memory usage of a server process.	-
Total memory usage	The memory usage of the machine where the server is located.	This metric is available only when servers are deployed on ECS instances. This metric is generally used for Alibaba Cloud.
CPU utilization	The CPU utilization of the machine where the server is located.	This metric is available only when servers are deployed on ECS instances. This metric is generally used for Alibaba Cloud.
System load	The load of the machine where the server is located.	This metric is available only when servers are deployed on ECS instances. This metric is generally used for Alibaba Cloud.

Term	Description	Remarks
Service port	The port used by servers to provide MySQL-based services for external applications.	The port number is generally 3306. When multiple nodes are deployed on one machine, the port number changes. In most cases, the nodes are physical machines.
Management port	The port used by servers to provide management application program interfaces (APIs).	The port number is generally the service port number plus 100.
Start time	The time when the server node starts.	-
Running time	The continuous running time of the server node since the last startup time.	-
Total memory size	The maximum JVM memory size of the server node.	-
Memory usage	The JVM memory that is used by the server node.	-
Number of nodes	Required. It indicates the number of instance machines. In essence, a instance is a cluster, and the number of nodes indicates the number of machines in the cluster.	-
Instance type	Required. It indicates the type of the instance, including dedicated and shared instances. A dedicated instance works in exclusive mode. A shared instance works in multi-tenant mode. In normal cases, shared instances are used for Alibaba Cloud.	-
Machine type	Required. It indicates the type of the machine where an instance is deployed. You can use a physical machine, a virtual machine such as an ECS instance, or a machine that is automatically selected by the system. The inventory is divided into the physical machine inventory and virtual machine inventory based on the type of machines where the PolarDB-X servers are deployed. The two types of machines cannot be deployed in hybrid mode because their deployment and O&M methods are different.	-
AliUid	Required. The unique identifier (UID) of an instance. On Apsara Stack, a UID is provided by the account system in the deployment environment.	-
Backend port	The backend port of the VIP. For a server, this indicates a service port of the machine where the PolarDB-X server is deployed.	-
Frontend port	The frontend port of the VIP for user access. Each VIP has a set of frontend ports and backend ports. The VIP forwards data from frontend ports to backend ports.	-
Private network or Internet	The network type of the VIP. You can use the following network types: • Internet: the public VIP, which is accessible from the Internet. • Private network: the private VIP, which is accessible from private networks. A private VIP can be a VIP in a VPC.	-

Term	Description	Remarks
lbId	The ID of an SLB instance. This ID is the unique identifier for a VIP. You can manage the VIP by using this ID.	-
Forwarding mode	The forwarding mode of a VIP. You can use the following modes: • FNAT: We recommend that you select this mode if the backend machine is a virtual machine or a VPC is used. • NAT: You can select this mode if the backend machine is a physical machine. This mode is used only for Alibaba Cloud. • Open FNAT: This mode is used only for Alibaba Cloud.	-
VPC ID	The ID of the VPC to be accessed.	-
vSwitch ID	The ID of the vSwitch. This determines the Classless Inter-Domain Routing (CIDR) block to which the VPC VIP of the instance belongs.	-
APPName	The application name of the database. Each database has an application name for loading configurations.	-
UserName	The username that is used to log on to the database.	-
DBName	The name of the database that you want to log on to.	-
IP address whitelist	Only the IP addresses specified in the IP address whitelist can connect to the instance.	-
Read-only instance	Physical database instances are divided into the following two types based on whether data can be written into the instances: • Primary instance: Read and write requests are allowed on such an instance. On Apsara Stack, ApsaraDB RDS for MySQL is supported. On Alibaba Cloud, ApsaraDB RDS is supported. • Read-only instance: Only read requests are allowed on such an instance. On Apsara Stack, ApsaraDB RDS for MySQL is supported. On Alibaba Cloud, ApsaraDB RDS is supported.	-
Read SQL statement	A type of SQL statements that are used to read data, such as the SELECT statement. determines whether an SQL statement is a read-only SQL statement when the statement is not in a transaction. If the SQL statement is in a transaction, PolarDB-X processes it as a write SQL statement during read/write splitting.	-
Read/write splitting	If read-only physical database instances exist, you can configure read/write splitting in the console to allocate read SQL statements to the primary instance and read-only instance in a proportional way. automatically identifies the type of SQL statements and allocates them in a proportional way.	-
Smooth scale-out	Data distribution on physical database instances is adjusted for scale-out in a dynamic way by using horizontal splitting. Scale-out is generally complete in an asynchronous way. You do not need to change business code.	-
Broadcast of small tables	You can synchronize data in a single table in a database to all the database shards in advance. In this case, you can convert the cross-database JOIN query into a pushdown JOIN query that can be run on physical databases.	-

Term	Description	Remarks
Horizontal splitting	Horizontal splitting distributes the data rows that are originally stored in one table to multiple tables by using specified rules to achieve horizontal linear scaling.	-
Sharding	This mode allows you to create multiple database shards on an ApsaraDB RDS for MySQL instance. These database shards constitute a database. In this mode, all the features can be used.	-
Non-sharding	In this mode, a database that has been created on an ApsaraDB RDS for MySQL instance is used as a database. In this mode, only read/write splitting of is allowed, and other features of such as sharding are not allowed.	-
Imported database	An existing database on the ApsaraDB RDS for MySQL instance. This concept applies when a database is created.	-
Read policy	The ratio of read SQL statements assigned by to the primary instance and read-only instance.	-
Full table scan	If no shard field is specified in an SQL statement, runs the SQL statement on all the table shards and summarizes the results. You can disable this feature due to its high overheads.	-
Shard key	The column in a logical table. routes data and SQL statements to a physical table based on this column.	-
Data import	The operation of importing data from an existing ApsaraDB RDS for MySQL instance to a database.	-
Full migration	The operation of migrating all the existing records from a database to . An offset is recorded before full migration starts.	-
Offset	In a MySQL binary log file, each row represents a data change operation. The position of a line in the binary log file is called an offset.	-
Incremental migration	The operation of reading all the MySQL binary log records from the recorded offset, converting them into SQL statements, and executing the statements in . Incremental migration continues before the switchover.	-
Switchover	A step of data import and smooth scale-out, which writes all the remaining incremental records from MySQL binary logs to .	-
Cleanup	The last step of smooth scale-out. In this step, redundant data and configurations generated during smooth scale-out are cleaned.	-
Heterogeneous indexing	For table shards of a database, the WHERE condition of an SQL statement for a query must contain the shard key if possible. In this way, routes the query request to a specific database shard. This makes the query efficient. If the WHERE condition of the SQL statement does not contain the shard key, performs a full table scan. provides heterogeneous indexing to solve this problem. The data in a database shard or table shard of a instance is fully or partially synchronized to another table based on different shard keys. The destination table to which the data is synchronized is called a heterogeneous index table.	-

Term	Description	Remarks
Globally unique numeric sequence	A globally unique numeric sequence in is a 64-digit number of the BIGINT data type in MySQL. Such a sequence is used to ensure that data in a defined unique field such as a primary key or a unique key is globally unique and incremental in sequential order.	-
custom hint	To use in an efficient way, defines some hints to specify special operations.	-

13.AnalyticDB for PostgreSQL

13.1. User Guide

13.1.1. What is AnalyticDB for PostgreSQL?

is a distributed cloud database service that uses multiple compute nodes to provide massively parallel processing (MPP) data warehousing.

is developed based on the open source Greenplum database project and enhanced by Alibaba Cloud. This service has the following features:

- Compatible with Greenplum and all tools that support Greenplum.
- Supports Object Storage Service (OSS), JSON, and HyperLogLog, which is a probabilistic algorithm for cardinality estimation.
- Supports SQL:2003-compliant syntax and Online Analytical Processing (OLAP) aggregate functions to provide flexible hybrid analysis.
- Supports both row store and column store to enhance analytics performance.
- Supports data compression to reduce storage costs.
- Provides online scaling and performance monitoring to enable database administrators (DBAs), developers, and data analysts to focus on improving enterprise productivity and creating core business value instead of managing and maintaining large numbers of MPP clusters.

13.1.2. Quick start

13.1.2.1. Overview

This topic describes all operations that you can perform on an AnalyticDB for PostgreSQL instance, from instance creation to database logon. It provides a quick start guide to the operations on the instance.

- Log on to the AnalyticDB for PostgreSQL console

  You can log on to the console.

- Create an instance

  Before you perform other operations, you must first create an instance in the console.

- Configure a whitelist

  Before you use an instance, add IP addresses or CIDR blocks needed to access your database to the whitelist of the instance to improve the security and stability of the database.

- Create an initial account

  After you create an instance, you must create an initial account to log on to the database.

- Connect to a database

  You can use a client that supports PostgreSQL or Greenplum to connect to a database.

13.1.2.2. Log on to the AnalyticDB for PostgreSQL console

This topic describes how to log on to the console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.

- We recommend that you use the Google Chrome browser.

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Database Services > AnalyticDB for PostgreSQL**.

13.1.2.3. Creates an instance

This topic describes how to create an instance in the console. Before you perform other operations, you must first create an AnalyticDB for PostgreSQL instance.

1. Log on to the AnalyticDB for PostgreSQL console.
2. In the upper-right corner of the page, click **Create Instance**.
3. On the **Create AnalyticDB for PostgreSQL Instance** page, set the following parameters.

Section	Parameter	Description
	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.

Region Section	Parameter	Description
	Region	The region in which you want to create the instance. **Note** If you want to access the instance from an Elastic Compute Service (ECS) instance over a virtual private cloud (VPC), you must create the instance in the same region and zone as those of the ECS instance.
	Zone	The zone in which you want to create the instance.
Basic Settings	Engine	Only the integrated computing and storage version is supported.
	Engine Version	The engine version of the instance.
	Node Type	Specifications for each compute node. The storage space and compute capability of a compute node vary based on the specified specifications.
	Nodes	The number of compute nodes. An instance must contain at least two compute nodes. The performance of an instance scales linearly with the number of compute nodes.
Network	Network Type	The network type of the instance. Valid values: ○ *Classic Network*: Cloud services within the classic network are not isolated from each other. Unauthorized access to a cloud service can be blocked only by the security group or whitelist policy of the service. ○ *VPC*: A VPC helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for improved security. You can create a VPC in advance, or change the network type to VPC after the instance is created.
	VPC	The VPC in which you want to create the instance. **Note** VPC: You can use a VPC to build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC.
	vSwitch	The vSwitch to which the instance is attached.
	IP Whitelist	The IP addresses that are allowed to access the instance.

4. After you set the preceding parameters, click **Submit**.

13.1.2.4. Configure a whitelist

To ensure a secure and stable database, you must add IP addresses or CIDR blocks that are allowed to access the database to a whitelist.

1. Log on to the AnalyticDB for PostgreSQL console.
2. Find the instance that you want to manage and click its ID. The **Basic Information** page appears.

3. In the left-side navigation pane, click **Security Controls**. The **Security Controls** page appears.
4. On the **Whitelist Settings** tab, click **Modify** corresponding to the *default* whitelist. The **Modify Group** page appears.

   > **Note** You can also click Clear corresponding to the *default* whitelist to delete the IP addresses in the default whitelist. Then, click Add Group to create another whitelist.

5. Delete 127.0.0.1 from the *default* whitelist and enter your IP addresses in the whitelist. The following table describes the parameters.

Parameter	Description
Group Name	Specify the name of the whitelist. The name must be 2 to 32 characters in length and can contain lowercase letters, digits, and underscores (_). It must start with a lowercase letter and end with a letter or digit. The default whitelist cannot be modified or deleted.
IP Addresses	Enter the CIDR blocks or IP addresses that are allowed to access the database. Use commas (,) to separate multiple CIDR blocks or IP addresses. ○ A whitelist can contain IP addresses such as 10.10.10.1 and CIDR blocks such as 10.10.10.0/24. This CIDR block indicates that all IP addresses in the 10.10.10.X format have access to the database. ○ The percent sign (%) or 0.0.0.0/0 indicates that all IP addresses are allowed to access the database. **Notice** We recommend that you do not use this configuration because it reduces the security of the database. ○ Default whitelists of new instances contain the loopback address 127.0.0.1. This configuration allows no access from external IP addresses. ○ You can add up to 999 IP addresses or CIDR blocks to a whitelist.

6. Click **OK**.

What's next

- We recommend that you maintain the whitelist on a regular basis to ensure secure access for .
- You can click **Modify** or **Delete** to modify or delete custom whitelists.

13.1.2.5. Create an initial account

After you create an instance, you must create an initial account to log on to the database.

1. Log on to the AnalyticDB for PostgreSQL console.
2. Find the instance that you want to manage and click its ID. The **Basic Information** page appears.
3. In the left-side navigation pane, click **Account Management**. The **Account Management** page appears.
4. In the upper-right corner of the page, click **Create Account**. The **Create Account** page appears.
5. Enter the database account and password, and click **OK**.

Parameter	Description
Account	The name of the account must be 2 to 16 characters in length and can contain lowercase letters, digits, and underscores (_). It must start with a letter and end with a letter or digit.

Parameter	Description
New Password	The password must be 8 to 32 characters in length. It must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters.
Password	Enter the password again.

13.1.2.6. Obtain client tools

The interface protocol of is compatible with Greenplum Community Edition and PostgreSQL 8.2. You can use the Greenplum or PostgreSQL client to connect to .

> **Note**
>
> Apsara Stack is an isolated environment. You must deploy software installation packages to the internal environment.

Graphical client tools

users can directly use client tools that support Greenplum, such as SQL Workbench, Navicat Premium, Navicat for PostgreSQL, and pgadmin III v1.6.3.

Command-line client psql (for RHEL 6, RHEL 7, CentOS 6, and CentOS 7)

For Red Hat Enterprise Linux (RHEL) 6, RHEL 7, CentOS 6, and CentOS 7, download the tools from the following links and decompress the packages to use them:

- For RHEL 6 or CentOS 6, click hybriddb_client_package_el6.
- For RHEL 7 or CentOS 7, click hybriddb_client_package_el7.

Command-line client psql (for other Linux systems)

For other Linux systems, perform the following operations to compile the client tools:

1. Obtain the source code by using one of the following methods:
   - Obtain the git directory. You must first install the git tool.

     ```
     git clone https://github.com/greenplum-db/gpdb.git
     cd gpdb
     git checkout 5d870156
     ```

   - Download the code.

     ```
     wget https://github.com/greenplum-db/gpdb/archive/5d87015609abd330c68a5402c1267fc86cbc9e1f.zip
     unzip 5d87015609abd330c68a5402c1267fc86cbc9e1f.zip
     cd gpdb-5d87015609abd330c68a5402c1267fc86cbc9e1f
     ```

2. Use GCC and other compilers.

   ```
   ./configure
   make -j32
   make install
   ```

3. Use psql and pg_dump. The two tools are located in the following paths:

   psql: */usr/local/pgsql/bin/psql*

   pg_dump: */usr/local/pgsql/bin/pg_dump*

Command-line client psql (for Windows and other systems)

For Windows and other systems, go to the Pivotal website to download HybridDB Client.

13.1.2.7. Connect to a database

Greenplum Database and are both developed based on PostgreSQL 8.2 and fully compatible with its message protocol. users can use tools that support the PostgreSQL 8.2 message protocol, such as libpq, Java Database Connectivity (JDBC), Open Database Connectivity (ODBC), psycopg2, and pgAdmin III.

Context

provides psql, a binary program of Red Hat. For more information about the download link, see Obtain the client tool. The Greenplum official website provides an easy-to-install installation package that includes JDBC, ODBC, and libpq. For more information, see Greenplum official documentation.

> **Note**
> - Apsara Stack is an isolated environment. To access Apsara Stack, you must prepare the necessary software installation packages in advance.
> - By default, instances can be accessed only by clients that are deployed on Elastic Compute Service (ECS) instances within the same region and zone.

psql

psql is a common tool used together with Greenplum, and provides a variety of command functions. Its binary files are located in the *bin* directory of Greenplum. To use psql, perform the following steps:

1. Use one of the following methods to connect to the database:
   - Connection string

     ```
     psql "host=yourgpdbaddress.gpdb.rds.aliyuncs.com port=3432 dbname=postgres user=gpdbaccount password=gpdbpassword"
     ```

   - Specified parameters

     ```
     psql -h yourgpdbaddress.gp.aliyun-inc.com -p 3432 -d postgres -U gpdbaccount
     ```

     The following section describes the parameters:

     - -h: the host address.
     - -p: the port used to connect to the database.
     - -d: the name of the database. The default value is postgres.
     - -U: the account used to connect to the database.

     You can run the `psql --help` command to view more options. You can also run the `\?` commandto view the commands supported in psql.
2. Enter the password to go to the psql shell interface.

   ```
   postgres=>
   ```

References

- For more information about the Greenplum psql usage, see psql.
- also supports psql commands of PostgreSQL. Pay attention to the differences between Greenplum psql and PostgreSQL psql. For more information, see PostgreSQL 8.3.23 Documentation - psql.

pgAdmin III

pgAdmin III is a PostgreSQL graphical client that can be directly used to connect to . For more information, click here. For more information about other graphical clients, see Obtain the client tool.

1. Download pgAdmin III 1.6.3 or earlier.

   You can download pgAdmin III 1.6.3 from the PostgreSQL website. pgAdmin III 1.6.3 supports various operating systems, such as Windows, macOS, and Linux.

   > **Note** is compatible with PostgreSQL 8.2. Therefore, you must use pgAdmin III 1.6.3 or earlier to connect to . pgAdmin 4 and later are not supported.

2. Choose **File > Add Server**.
3. In the New Server Registration dialog box, set the parameters.
4. Click **OK** to connect to .

JDBC

JDBC uses the interface provided by PostgreSQL. Use the following method to download the JDBC driver:

Click here to download the official JDBC of PostgreSQL. Then, add it to the environment variables.

Sample code:

```
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
public class gp_conn {
    public static void main(String[] args) {
        try {
            Class.forName("org.postgresql.Driver");
            Connection db = DriverManager.getConnection("jdbc:postgresql://mygpdbpub.gpdb.rds.aliyuncs
.com:3432/postgres",
                    "mygpdb", "mygpdb");
            Statement st = db.createStatement();
            ResultSet rs = st.executeQuery(
                    "select * from gp_segment_configuration;");
            while (rs.next()) {
                System.out.print(rs.getString(1));
                System.out.print(" | ");
                System.out.print(rs.getString(2));
                System.out.print(" | ");
                System.out.print(rs.getString(3));
                System.out.print(" | ");
                System.out.print(rs.getString(4));
                System.out.print(" | ");
                System.out.print(rs.getString(5));
                System.out.print(" | ");
                System.out.print(rs.getString(6));
                System.out.print(" | ");
                System.out.print(rs.getString(7));
                System.out.print(" | ");
                System.out.print(rs.getString(8));
                System.out.print(" | ");
                System.out.print(rs.getString(9));
                System.out.print(" | ");
                System.out.print(rs.getString(10));
                System.out.print(" | ");
                System.out.println(rs.getString(11));
            }
            rs.close();
            st.close();
        } catch (ClassNotFoundException e) {
            e.printStackTrace();
        } catch (SQLException e) {
            e.printStackTrace();
        }
    }
}
```

Python

Python uses psycopg2 to connect to Greenplum and PostgreSQL. Perform the following operations:

1. Install psycopg2. Use one of the following methods to install psycopg2 in CentOS:
   - Method 1: Run the `yum -y install python-psycopg2` command.
   - Method 2: Run the `pip install psycopg2` command.
   - Method 3: Run the following source code:

```
yum install -y postgresql-devel*
wget http://initd.org/psycopg/tarballs/PSYCOPG-2-6/psycopg2-2.6.tar.gz
tar xf psycopg2-2.6.tar.gz
cd psycopg2-2.6
python setup.py build
sudo python setup.py install
```

2. Run the following commands to set PYTHONPATH and reference it:

```
import psycopg2
 sql = 'select * from gp_segment_configuration;'
 conn = psycopg2.connect(database='gpdb', user='mygpdb', password='mygpdb', host='mygpdbpub.gpdb.rds.aliyuncs.com', port=3432)
 conn.autocommit = True
 cursor = conn.cursor()
 cursor.execute(sql)
 rows = cursor.fetchall()
 for row in rows:
     print row
 conn.commit()
 conn.close()
```

An output similar to the following one is displayed:

```
(1, -1, 'p', 'p', 's', 'u', 3022, '192.168.2.158', '192.168.2.158', None, None)(6, -1, 'm', 'm', 's', 'u', 3019, '192.168.2.47', '192.168.2.47', None, None)(2, 0, 'p', 'p', 's', 'u', 3025, '192.168.2.148', '192.168.2.148', 3525, None)(4, 0, 'm', 'm', 's', 'u', 3024, '192.168.2.158', '192.168.2.158', 3524, None)(3, 1, 'p', 'p', 's', 'u', 3023, '192.168.2.158', '192.168.2.158', 3523, None)(5, 1, 'm', 'm', 's', 'u', 3026, '192.168.2.148', '192.168.2.148', 3526, None)
```

libpq

libpq is the C language interface to AnalyticDB for PostgreSQL. You can use the libpq library to access and manage PostgreSQL databases in a C program. You can find its static and dynamic libraries in the lib directory.

For more information about example programs, see Example Programs.

For more information about libpq, see PostgreSQL 9.4.17 Documentation - Chapter 31. libpq - C Library.

ODBC

PostgreSQL ODBC is an open source version based on the GNU Lesser General Public License (LGPL) protocol. You can download it from the PostgreSQL website.

1. Install the driver.

```
yum install -y unixODBC.x86_64
yum install -y postgresql-odbc.x86_64
```

2. View the driver configurations.

```
cat /etc/odbcinst.ini
# Example driver definitions
# Driver from the postgresql-odbc package
# Setup from the unixODBC package
[PostgreSQL]
Description = ODBC for PostgreSQL
Driver = /usr/lib/psqlodbcw.so
Setup = /usr/lib/libodbcpsqlS.so
Driver64 = /usr/lib64/psqlodbcw.so
Setup64 = /usr/lib64/libodbcpsqlS.so
FileUsage = 1
# Driver from the mysql-connector-odbc package
# Setup from the unixODBC package
[MySQL]
Description = ODBC for MySQL
Driver = /usr/lib/libmyodbc5.so
Setup = /usr/lib/libodbcmyS.so
Driver64 = /usr/lib64/libmyodbc5.so
Setup64 = /usr/lib64/libodbcmyS.so
FileUsage = 1
```

3. Configure the data source name (DSN). Replace `****` in the following code with the corresponding connection information.

```
[mygpdb]
Description = Test to gp
Driver = PostgreSQL
Database = ****
Servername = ****.gpdb.rds.aliyuncs.com
UserName = ****
Password = ****
Port = ****
ReadOnly = 0
```

4. Test the connectivity.

```
echo "select count(*) from pg_class" | isql mygpdb
+---------------------------------------+
| Connected!                            |
|                                       |
| sql-statement                         |
| help [tablename]                      |
| quit                                  |
|                                       |
+---------------------------------------+
SQL> select count(*) from pg_class
+----------------------+
| count                |
+----------------------+
| 388                  |
+----------------------+
SQLRowCount returns 1
1 rows fetched
```

5. After ODBC is connected to the instance, connect the application to ODBC. For more information, see psqlODBC - PostgreSQL ODBC driver and psqlODBC HOWTO - C#.

References

- Pivotal Greenplum official documentation

- PostgreSQL psqlODBC
- Compiling psqlODBC on Unix
- Download ODBC connectors
- Download JDBC connectors
- The PostgreSQL JDBC Interface

13.1.3. Instances

13.1.3.1. Reset the password

When you use , you can reset the password of your database account in the console if you forget your password.

> **Note** To ensure data security, we recommend that you change your password on a regular basis.

1. Log on to the AnalyticDB for PostgreSQL console.
2. Find the target instance and click its ID. The **Basic Information** page appears.
3.
4. Click **Reset Password** in the Actions column corresponding to an account. The **Reset Account Password** page appears.
5. After you enter and confirm the new password, click **OK**.

> **Note** The password must be 8 to 32 characters in length. It must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. We recommend that you do not use a previously used password.

13.1.3.2. View monitoring information

You can go to the monitoring information page in the console to view the operation status of an instance.

1. Log on to the AnalyticDB for PostgreSQL console.
2. Find the instance that you want to manage and click its ID. The **Basic Information** page appears.
3. In the left-side navigation pane, click **Monitoring and Alarms**. The **Monitoring and Alarms** page appears.

   Specify a duration of time n up to seven days in length to view the metrics for that last n period.

13.1.3.3. Switch the network type of an instance

The default network type of an instance is Virtual Private Cloud (VPC). After an instance is created, you can switch its network type between classic network and VPC as needed.

Context

AnalyticDB for PostgreSQL supports two network types: classic network and VPC. Both network types use BGP connections, and are independent of the public network of your service provider. These network types only differ in functions, and you can choose a network type based on your requirements. The two network types are applicable to different scenarios:

- Classic network: IP addresses are allocated by Alibaba Cloud. Classic networks are easy to configure and use. This network type is suitable for users who do not need to perform complex operations, or who only require short deployment cycles.
- VPC: a logically isolated private network. You can customize the network topology and IP addresses and connect through a leased line. This network type is suitable for advanced users.

Warning Switching the network type will cause the database service to stop. Proceed with caution.

1. Log on to the AnalyticDB for PostgreSQL console.
2. Find the target instance and click its ID. The **Basic Information** page appears.
3. In the left-side navigation pane, click **Database Connection**. The **Database Connection** page appears.
4. In the upper-right corner of the page, click **Switch to Classic Network** or **Switch to VPC**.
5. If you click **Switch to VPC**, you must select the destination **VPC** and **VSwitch**. Click **OK**.

   Note To switch the network type to VPC, a VPC and a VSwitch must exist or be created in the zone where the instance is located.

6. If you click **Switch to Classic Network**, click **OK** in the displayed message.

Note After you switch the network type, it takes 3 to 30 minutes for the instance to enter the running state.

13.1.3.4. Restart an instance

To better meet your requirements, updates the minor kernel version on a regular basis. When you create an instance, the latest database kernel is used by default. After a new version is released, you can restart your instance to update the database kernel and use its extended features. This topic describes how to restart an instance.

Warning The database service may be interrupted when you restart an instance.

1. Log on to the AnalyticDB for PostgreSQL console.
2. Find the instance that you want to manage and click its ID. The **Basic Information** page appears.
3. In the upper-right corner of the page, click **Restart Instance**.

Note The restart process takes about 3 to 30 minutes. During the restart period, the instance cannot provide external services. We recommend that you take precautionary measures before you restart the instance. After the instance is restarted and enters the running state, you can access the database.

13.1.3.5. Import data

13.1.3.5.1. Import data from or export data to OSS in parallel

allows you to import data from or export data to Object Storage Service (OSS) tables in parallel by using the OSS external table feature, gpossext. AnalyticDB for PostgreSQL also supports GZIP compression for OSS external tables to reduce file size and storage costs. gpossext can read from and write to TEXT and CSV files, even when they are compressed in GZIP packages.

- Create an OSS external table extension (oss_ext)

  To use an OSS external table, you must first create an OSS external table extension in . You must create an extension for each database that you need to access.

  - To create the extension, execute the `CREATE EXTENSION IF NOT EXISTS oss_ext;` statement.
  - To delete the inextension, execute the `DROP EXTENSION IF EXISTS oss_ext;` statement.

- Import data in parallel

i. Distribute data evenly into multiple files in OSS. We recommend that you set the number of OSS files to an integer that is the multiple of the number of compute nodes in .
ii. Create a READABLE external table in .
iii. Execute the following statement to import data in parallel:

```
INSERT INTO <Destination table> SELECT * FROM <External table>
```

> **Note**
> - The data import performance depends on the OSS performance and resources of instances, such as CPU, I/O, memory, and network resources. To ensure the best import performance, we recommend that you use column store and compression when you create a table. For example, you can specify the following clause: `WITH (APPENDONLY=true, ORIENTATION=column, COMPRESSTYPE=zlib, COMPRESSLEVEL=5, BLOCKSIZE=1048576)` . For more information, see Greenplum Database official documentation on database table creation syntax.
> - To ensure the best import performance, we recommend that you configure OSS and instances within the same region.

- Export data in parallel
  i. Create a WRITABLE external table in .
  ii. Execute the following statement to export data to OSS in parallel:

```
INSERT INTO <External table> SELECT * FROM <Source table>
```

- Create OSS external tables

> **Note** The syntax to create and use external tables is the same as that of Greenplum Database, except for the syntax of location-related parameters.

```
CREATE [READABLE] EXTERNAL TABLE tablename
( columnname datatype [, ...] | LIKE othertable )
LOCATION ('ossprotocol')
FORMAT 'TEXT'
            [( [HEADER]
               [DELIMITER [AS] 'delimiter' | 'OFF']
               [NULL [AS] 'null string']
               [ESCAPE [AS] 'escape' | 'OFF']
               [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
               [FILL MISSING FIELDS] )]
          | 'CSV'
            [( [HEADER]
               [QUOTE [AS] 'quote']
               [DELIMITER [AS] 'delimiter']
               [NULL [AS] 'null string']
               [FORCE NOT NULL column [, ...]]
               [ESCAPE [AS] 'escape']
               [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
               [FILL MISSING FIELDS] )]
[ ENCODING 'encoding' ]
[ [LOG ERRORS [INTO error_table]] SEGMENT REJECT LIMIT count
       [ROWS | PERCENT] ]
CREATE WRITABLE EXTERNAL TABLE table_name
( column_name data_type [, ...] | LIKE other_table )
LOCATION ('ossprotocol')
FORMAT 'TEXT'
               [( [DELIMITER [AS] 'delimiter']
               [NULL [AS] 'null string']
               [ESCAPE [AS] 'escape' | 'OFF'] )]
          | 'CSV'
               [([QUOTE [AS] 'quote']
               [DELIMITER [AS] 'delimiter']
               [NULL [AS] 'null string']
               [FORCE QUOTE column [, ...]] ]
               [ESCAPE [AS] 'escape'] )]
[ ENCODING 'encoding' ]
[ DISTRIBUTED BY (column, [ ... ] ) | DISTRIBUTED RANDOMLY ]
ossprotocol:
   oss://oss_endpoint prefix=prefix_name
    id=userossid key=userosskey bucket=ossbucket compressiontype=[none|gzip] async=[true|false]
ossprotocol:
   oss://oss_endpoint dir=[folder/[folder/]...]/file_name
    id=userossid key=userosskey bucket=ossbucket compressiontype=[none|gzip] async=[true|false]
ossprotocol:
   oss://oss_endpoint filepath=[folder/[folder/]...]/file_name
    id=userossid key=userosskey bucket=ossbucket compressiontype=[none|gzip] async=[true|false]
```

Parameters

Common parameters

Parameter	Description

Parameter	Description
Protocol and endpoint	This parameter is in the `protocol name://oss_endpoint` format. The protocol name is oss. oss_endpoint is the domain name that is used to access OSS in a region. **Note** You can access the database from a virtual private cloud (VPC) host by using an internal endpoint that contains "internal" in the name to avoid generating public traffic.
id	The AccessKey ID of the OSS account.
key	The AccessKey secret of the OSS account.
bucket	The bucket where the data file is located. You must use OSS to create the bucket before you import data.

prefix	The prefix of the path name corresponding to the data file. Prefixes are directly matched and cannot be controlled by regular expressions. The prefix, filepath, and dir parameters are mutually exclusive. Only one of the parameters can be specified at a time.
	• If you create a READABLE external table for data import, all OSS files that contain the specified prefix are imported.
	  ◦ If you set prefix to test/filename, the following files are imported:
	  ▪ test/filename
	  ▪ test/filenamexxx
	  ▪ test/filename/aa
	  ▪ test/filenameyyy/aa
	  ▪ test/filenameyyy/bb/aa
	  ◦ If you set prefix to test/filename/, only the following file out of the preceding files is imported:
	  test/filename/aa
	• If you create a WRITABLE external table for data export, each exported file has a unique name based on this parameter.
	Note One or more files can be exported for each compute node. The names of exported files are in the `prefix_tablename_uuid.x` format. uuid indicates a timestamp in microseconds as an int64 value. x indicates the node ID. You can use an external table for multiple export operations. Each export operation is assigned a uuid value. The files exported during each operation share a uuid value.

Parameter	Description
dir	The virtual folder path in OSS. The prefix, filepath, and dir parameters are mutually exclusive. Only one of the parameters can be specified at a time. • A folder path must end with a forward slash (/). Example: `test/mydir/`. • If you use this parameter when you create an external table for data import, all files in the specified virtual directory (except for its subdirectories and contained files) are imported. Unlike filepath, dir does not require you to specify the names of files in the directory. • If you use this parameter when you create an external table for data export, all data is exported as multiple files within the specified directory. The names of exported files are in the `filename.x` format, where x is a number. The values of x may not be consecutive.

filepath	The file name that contains a path in OSS. The prefix, filepath, and dir parameters are mutually exclusive. Only one of the parameters can be specified at a time. You can specify only the filepath parameter when you create a READABLE external table for data import.
	• The file name includes the file path, but not the bucket name.
	• The file name that is specified for data import must be in the `filename` or `filename.x` format. The values of x must be consecutive numbers that start from 1.
	For example, if filepath is set to filename and OSS contains the following files, the imported files include filename, filename.1, and filename.2, but filename.4 is not imported because filename.3 does not exist.
	`filename`
	`filename.1`
	`filename.2`
	`filename.4`

Import mode parameters

Parameter	Description
async	Specifies whether to enable asynchronous data import. • By default, asynchronous data import is enabled. You can set async to false or f to disable asynchronous data import. • You can enable the worker thread to load data from OSS to accelerate the import performance. By default, asynchronous data import is used. • Asynchronous data import consumes more hardware resources than normal data import.
compressiontype	The compression format of the imported files. Valid values: • none: The import files are not compressed. This is the default value. • gzip: The imported files are compressed in the GZIP format. Only the GZIP format is supported.
compressionlevel	The compression level of the files that are written to OSS. Valid values: 1 to 9. Default value: 6.

Export mode parameters

Parameter	Description
oss_flush_block_size	The size of each data block that is written to OSS. Valid values: 1 to 128. Default value: 32. Unit: MB.
oss_file_max_size	The maximum size for each file that is written to OSS. If the limit is exceeded, subsequent data is written to another file. Valid values: 8 to 4000. Default value: 1024. Unit: MB.
num_parallel_worker	The number of parallel compression threads for data that is written to OSS. Valid values: 1 to 8. Default value: 3.

For data export, take note of the following items:

- WRITABLE is the keyword of the external table for data export. You must specify this keyword when you create an external table.
- Only the prefix and dir parameters are supported for data export. The filepath parameter is not supported.
- You can use the DISTRIBUTED BY clause to write data from compute nodes to OSS based on the specified distribution keys.

Other common parameters

The following table describes the fault-tolerance parameters that can be used for data import and export.

Fault-tolerance parameters

Parameter	Description
oss_connect_timeout	The connection timeout period. Default value: 10. Unit: seconds.
oss_dns_cache_timeout	The Alibaba Cloud DNS (DNS) timeout period. Default value: 60. Unit: seconds.
oss_speed_limit	The minimum rate tolerated. Default value: 1024 bit/s (1 Kbit/s).
oss_speed_time	The maximum amount of time tolerated. Default value: 15. Unit: seconds.

When the default values are used for the preceding parameters, a timeout occurs if the transmission rate is lower than 1 Kbit/s for 15 consecutive seconds. For more information, see **Troubleshooting** in *OSS SDK reference*.

The other parameters are compatible with the external table syntax of Greenplum Database. For more information about the syntax, see Greenplum Database official documentation on external table syntax. The following parameters are included:

- FORMAT: the supported file format, such as TEXT and CSV.
- ENCODING: the data encoding format of a file, such as UTF-8.
- LOG ERRORS refers to improperly imported data that can be ignored and is instead written to to error_table. You can also use the count parameter to specify the error reporting threshold.

Examples

```
# Create a READABLE external table of OSS.
create readable external table ossexample
        (date text, time text, open float, high float,
        low float, volume int)
        location('oss://oss-cn-hangzhou.aliyuncs.com
        prefix=osstest/example id=XXX
        key=XXX bucket=testbucket compressiontype=gzip')
        FORMAT 'csv' (QUOTE '''' DELIMITER E'\t')
        ENCODING 'utf8'
```

```
        LOG ERRORS INTO my_error_rows SEGMENT REJECT LIMIT 5;
create readable external table ossexample
        (date text, time text, open float, high float,
        low float, volume int)
        location('oss://oss-cn-hangzhou.aliyuncs.com
        dir=osstest/ id=XXX
        key=XXX bucket=testbucket')
        FORMAT 'csv'
        LOG ERRORS SEGMENT REJECT LIMIT 5;
create readable external table ossexample
        (date text, time text, open float, high float,
        low float, volume int)
        location('oss://oss-cn-hangzhou.aliyuncs.com
        filepath=osstest/example.csv id=XXX
        key=XXX bucket=testbucket')
        FORMAT 'csv'
        LOG ERRORS SEGMENT REJECT LIMIT 5;
# Create a WRITABLE external table of OSS.
create WRITABLE external table ossexample_exp
        (date text, time text, open float, high float,
        low float, volume int)
        location('oss://oss-cn-hangzhou.aliyuncs.com
        prefix=osstest/exp/outfromhdb id=XXX
        key=XXX bucket=testbucket') FORMAT 'csv'
        DISTRIBUTED BY (date);
create WRITABLE external table ossexample_exp
        (date text, time text, open float, high float,
        low float, volume int)
        location('oss://oss-cn-hangzhou.aliyuncs.com
        dir=osstest/exp/ id=XXX
        key=XXX bucket=testbucket') FORMAT 'csv'
        DISTRIBUTED BY (date);
# Create a heap table named example to which you want to import data.
create table example
        (date text, time text, open float,
         high float, low float, volume int)
         DISTRIBUTED BY (date);
# Import data to the example heap table from the ossexample table in parallel.
insert into example select * from ossexample;
# Export data from the example heap table to OSS in parallel.
insert into ossexample_exp select * from example;
# The following execution plan shows that all compute nodes are involved in the task.
# All compute nodes read data from OSS in parallel. AnalyticDB for PostgreSQL performs a redistributio
n motion operation to compute the data by using a hash algorithm, and then distributes the data to its
compute nodes after computing. After a compute node receives data, it performs an insert operation to
add the data to AnalyticDB for PostgreSQL.
explain insert into example select * from ossexample;
                                        QUERY PLAN
-------------------------------------------------------------------------------------------------
 Insert (slice0; segments: 4)  (rows=250000 width=92)
   ->  Redistribute Motion 4:4  (slice1; segments: 4)  (cost=0.00..11000.00 rows=250000 width=92)
         Hash Key: ossexample.date
         ->  External Scan on ossexample  (cost=0.00..11000.00 rows=250000 width=92)
(4 rows)
# The following query plan shows that each compute node directly exports data to OSS without redistrib
uting the data.
explain insert into ossexample_exp select * from example;
                          QUERY PLAN
---------------------------------------------------------------
 Insert (slice0; segments: 3)  (rows=1 width=92)
```

```
   ->  Seq Scan on example  (cost=0.00..0.00 rows=1 width=92)
(2 rows)
```

TEXT and CSV format description

The following parameters specify the formats of files read from and written to OSS. You can specify the parameters in the external DDL parameters.

- \n: the line feed for TEXT and CSV files.
- DELIMITER: the delimiter of columns.
  - If the DELIMITER parameter is specified, the QUOTE parameter must also be specified.
  - Recommended column delimiters include commas (,), vertical bars (|), and special characters such as \t.
- QUOTE: encloses user data that contains special characters by column.
  - Strings that contain special characters must be enclosed by QUOTE to differentiate user data from control characters.
  - To optimize the efficiency, it is unnecessary to enclose data such as integers in QUOTE characters.
  - QUOTE cannot be the same string as specified in DELIMITER. The default value of QUOTE is double quotation marks (").
  - User data that contains QUOTE characters must also contain ESCAPE characters to differentiate user data from machine code.
- ESCAPE: the escape character.
  - Place an escape character before a special character that needs to be escaped to indicate that it is not a special character.
  - If ESCAPE is not specified, the default value is the same as QUOTE.
  - You can also use other characters such as backslashes (\) as ESCAPE characters, which is used by MySQL.

Default control characters for TEXT and CSV files

Default control characters for TEXT and CSV files

Control character	TEXT	CSV
DELIMITER	Tab (\t)	Comma (,)
QUOTE	double quotation mark (")	double quotation mark (")
ESCAPE	N/A	Same as QUOTE
NULL	Backslash plus N (\N)	Empty string without quotation marks

> **Note** All control characters must be single-byte characters.

SDK troubleshooting

The following Error log information table lists the error logs generated when an error occurs during the import or export process.

Error log information

Keyword	Description
code	The HTTP status code of the error request.

Keyword	Description
error_code	The error code returned by OSS.
error_msg	The error message returned by OSS.
req_id	The UUID used to identify the request. If you require assistance from OSS developers, you can submit a ticket that contains the req_id parameter of the failed request.

You can handle limitout-related errors by using parameters related to oss_ext.

References

- Greenplum Database official documentation on external table syntax
- Greenplum Database official documentation on table creation syntax

13.1.3.5.2. Import data from MySQL

You can use the mysql2pgsql tool to migrate tables from MySQL to , Greenplum Database, PostgreSQL, or Postgres Plus Advanced Server (PPAS).

Background information

mysql2pgsql connects to both the source MySQL database and the destination database. The tool retrieves the data that you want to export from the source MySQL database, and uses a COPY statement to import the data to the destination database. This tool supports simultaneous data import over multiple threads. Each worker thread imports data from some database tables.

To download the binary installation package of mysql2pgsql, click here.

To view instructions on source code compilation of mysql2pgsql, click here.

Procedure

1. Modify the my.cfg configuration file to configure the connection information of source and destination databases.
   i. Modify the connection information of the source MySQL database.

   > **Note** You must have the read permissions on all user tables.

   ```
   [src.mysql]
   host = "192.168.1.1"
   port = "3306"
   user = "test"
   password = "test"
   db = "test"
   encodingdir = "share"
   encoding = "utf8"
   ```

   ii. Modify the connection information of the destination PostgreSQL, PPAS, or database.

   > **Note** You must have the write permissions on the destination table.

   ```
   [desc.pgsql]
   connect_string = "host=192.168.1.2 dbname=test port=3432  user=test password=pgsql"
   ```

2. Use mysql2pgsql to import data.

```
./mysql2pgsql -l <tables_list_file> -d -n -j <number of threads> -s <schema of target able>
```

Parameters

Parameter	Description

-l	Optional. This parameter is used to specify a text file that contains tables to be synchronized. If you do not specify this parameter, all the tables in the database that is specified in the configuration file are synchronized. `<tables_list_file>` specifies the name of a file that contains a collection of tables to be synchronized and conditions for table queries. You can specify the file content in the following format:
	`table1 : select * from table_big where column1 < '2016-08-05'`
	`table2 :`
	`table3`
	`table4: select column1, column2 from tableX where column1 != 10`
	`table5: select * from table_big where column1 >= '2016-08-05'`

-d	Optional. This parameter indicates the table creation DDL statement that creates the destination table but does not synchronize data.
-n	Optional. This parameter must be used along with -d to specify that the table partition definition is not included in the DDL statement.
-j	Optional. This parameter is used to specify the number of threads that are used for data synchronization. If you do not specify this parameter, five concurrent threads are used.
-s	Optional. This parameter is used to specify the schema of the destination table. Only one schema at a time can be specified by the command. If you do not specify this parameter, the data is imported into the table under the public schema.

Typical usage

Full database migration

1. Run the following command to obtain the DDL statements of the corresponding destination table:

   ```
   ./mysql2pgsql -d
   ```

2. Create a table in the destination database based on these DDL statements with the distribution key information added.
3. Run the following command to synchronize all tables:

   ```
   ./mysql2pgsql
   ```

   This command migrates the data from all MySQL tables in the database that is specified in the configuration file to the destination database. By default, five concurrent threads are used to read and import data from involved tables.

Partial table migration

1. Create a file named tab_list.txt and enter the following content:

   ```
   t1
   t2 : select * from t2 where c1 > 138888
   ```

2. Run the following command to synchronize the specified t1 and t2 tables. For the t2 table, only data that meets the c1 > 138888 condition is migrated.

   ```
   ./mysql2pgsql -l tab_list.txt
   ```

13.1.3.5.3. Import data from PostgreSQL

You can use the pgsql2pgsql tool to migrate tables across , Greenplum Database, PostgreSQL, and Postgres Plus Advanced Server (PPAS).

Context

pgsql2pgsql supports the following features:

- Full migration across PostgreSQL, PPAS, Greenplum Database, and .
- Full migration and incremental migration from PostgreSQL or PPAS (version 9.4 or later) to AnalyticDB for PostgreSQL or ApsaraDB RDS for PPAS.

You can download the software packages from the dbsync project library.

- To download the binary installation package of pgsql2pgsql, click here.
- To view instructions on source code compilation of pgsql2pgsql, click here.

Procedure

1. Modify the my.cfg configuration file to configure the connection information of source and destination databases.
   i. Modify the connection information of the source PostgreSQL database.

   > **Note** In the connection information of the source PostgreSQL database, we recommend that you set the user to the owner of the source database.

   ```
   [src.pgsql]
   connect_string = "host=192.168.0.1 dbname=test port=3432  user=test password=pgsql"
   ```

   ii. Modify the connection information of the on-premises temporary PostgreSQL database.

   ```
   [local.pgsql]
   connect_string = "host=192.168.0.2 dbname=test port=3432  user=test2 password=pgsql"
   ```

   iii. Modify the connection information of the destination PostgreSQL database.

   > **Note** You must have the write permissions on the destination table.

   ```
   [desc.pgsql]
   connect_string = "host=192.168.0.2 dbname=test port=3432  user=test3 password=pgsql"
   ```

   > **Note**
   > - If you want to synchronize incremental data, you must have the permissions to create replication slots in the source database.
   > - PostgreSQL 9.4 and later support logic flow replication. Therefore, source databases of these versions support incremental data migration. A kernel supports logic flow replication only after you configure the following parameters:
   >
   >   wal_level = logical
   >
   >   max_wal_senders = 6
   >
   >   max_replication_slots = 6

2. Use pgsql2pgsql to perform full database migration.

   ```
   ./pgsql2pgsql
   ```

By default, the migration program migrates the table data of all users from the source PostgreSQL database to the destination PostgreSQL database.

3. View the status information.

   You can view the status information in a single migration process by connecting to the on-premises temporary database. The status information is stored in the db_sync_status table and includes the start and end time of full data migration, the start time of incremental data migration, and the status of incremental data synchronization.

13.1.3.5.4. Use the \COPY statement to import data

You can use the `\COPY` statement to import data from local text files to instances. Local text must be properly formatted and include necessary delimiters such as commas (,), semicolons (;), or special characters.

Context

- Parallel writing of large amounts of data is unavailable because the `\COPY` statement writes data in series by using the coordinator node. If you need to import a large amount of data in parallel, you can use the data import method based on Object Storage Service (OSS).
- The `\COPY` statement is a psql instruction. If you use the database statement `COPY` instead of the `\COPY` statement, you must note that only stdin is supported. This COPY statement does not support files because the root user does not have the superuser permissions to perform operations on files.
- also allows you to use JDBC to execute the COPY statement. The CopyIn method is encapsulated within JDBC. For more information, see Interface CopyIn.
- For more information about how to use the COPY statement, see COPY.

Procedure

1. Import data by using the following sample code:

```
\COPY table [(column [, ...])] FROM {'file' | STDIN}
     [ [WITH]
       [OIDS]
       [HEADER]
       [DELIMITER [ AS ] 'delimiter']
       [NULL [ AS ] 'null string']
       [ESCAPE [ AS ] 'escape' | 'OFF']
       [NEWLINE [ AS ] 'LF' | 'CR' | 'CRLF']
       [CSV [QUOTE [ AS ] 'quote']
            [FORCE NOT NULL column [, ...]]
       [FILL MISSING FIELDS]
       [[LOG ERRORS [INTO error_table] [KEEP]
       SEGMENT REJECT LIMIT count [ROWS | PERCENT] ]
\COPY {table [(column [, ...])] | (query)} TO {'file' | STDOUT}
      [ [WITH]
        [OIDS]
        [HEADER]
        [DELIMITER [ AS ] 'delimiter']
        [NULL [ AS ] 'null string']
        [ESCAPE [ AS ] 'escape' | 'OFF']
        [CSV [QUOTE [ AS ] 'quote']
             [FORCE QUOTE column [, ...]] ]
      [IGNORE EXTERNAL PARTITIONS ]
```

13.1.4. Databases

13.1.4.1. Overview

The operations based on Greenplum Database in are the same as those in Greenplum Database, including their schemas, supported data types, and user permissions. Except for specific operations that are exclusive to Greenplum Database such as those on distribution keys and append-optimized (AO) tables, you can refer to PostgreSQL for other operations.

References

- Pivotal Greenplum Official Documentation
- GP 4.3 Best Practice

13.1.4.2. Create a database

After you log on to the AnalyticDB for PostgreSQL instance, you can execute SQL statements to create databases.

As in PostgreSQL, you can execute SQL statements to create databases in . For example, after psql is connected to Greenplum, execute the following statements:

```
=> create database mygpdb;
CREATE DATABASE
=> \c mygpdb
psql (9.4.4, server 8.3devel)
You are now connected to database "mygpdb" as user "mygpdb".
```

13.1.4.3. Create a distribution key

is a distributed database where data is distributed across all the data nodes. You must create distribution keys to distribute the data in . Distribution keys are vital to query performance. Distribution keys are used to ensure **even data distribution**. Appropriate selection of keys can significantly improve query performance.

Specify a distribution key

In , tables can be distributed across all compute nodes in hash or random mode. You must specify a distribution key when you create a table. Imported data is distributed to the specific compute node based on the hash value calculated by the distribution key.

```
=> create table vtbl(id serial, key integer, value text, shape cuboid, location geometry, comment text
) distributed by (key);
CREATE TABLE
```

If you do not specify the distribution key, randomly allocates the id field by using the round-robin algorithm. You can specify the distribution key by adding `distributed by (key)` to the table creation statement.

Rules for selecting the distribution key

- Select evenly distributed columns or multiple columns to prevent data skew.
- Select fields that are commonly used for connection operations, especially for highly concurrent statements.
- Select condition columns that feature high concurrency queries and high filterability.
- Do not use random distribution.

13.1.4.4. Construct data

In some test scenarios, you must construct data to fill the database.

1. Create a function that generates random strings.

```
CREATE OR REPLACE FUNCTION random_string(integer) RETURNS text AS $body$
SELECT array_to_string(array
                      (SELECT substring('0123456789ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstu
vwxyz'
                                        FROM (ceil(random()*62))::int
                                        FOR 1)
                       FROM generate_series(1, $1)), '');
$body$
LANGUAGE SQL VOLATILE;
```

2. Create a partition key.

```
CREATE TABLE tbl(id serial, KEY integer, locate geometry, COMMENT text) distributed by (key);
```

3. Construct data.

```
INSERT INTO tbl(KEY, COMMENT, locate)
 SELECT
     KEY,
     COMMENT,
     ST_GeomFromText(locate) AS locate
 FROM
     (SELECT
         (a + 1) AS KEY,
         random_string(ceil(random() * 24)::integer) AS COMMENT,
         'POINT(' || ceil(random() * 36 + 99) || ' ' || ceil(random() * 24 + 50) || ')' AS locate
     FROM
         generate_series(0, 99999) AS a)
     AS t;
```

13.1.4.5. Query data

This topic describes the query statements and how to view the query plans.

Query statement sample

```
=> select * from tbl where key = 751;
| id  | key | value | shape        | locate                                          | comment        |
+-----+-----+-------+--------------+-------------------------------------------------+----------------+
| 751 | 751 | red   | 0101000000000000000C05B400000000000004A40 | B9hPhjeNWPqV |
(1 row)
Time: 513.101 ms
```

View a query plan

```
=> explain select * from tbl where key = 751;
Gather Motion 1:1  (slice1; segments: 1)  (cost=0.00..1519.28 rows=1 width=53)
  ->  Seq Scan on tbl  (cost=0.00..1519.28 rows=1 width=53)
        Filter: key = 751
Settings:  effective_cache_size=8GB; gp_statistics_use_fkeys=on
Optimizer status: legacy query optimizer
```

13.1.4.6. Manage extensions

You can use extensions to expand database features. enables you to manage extensions.

Supported extensions

supports the following extensions:

- PostGIS: processes geographic data.
- MADlib: provides a machine learning function library.
- fuzzystrmatch: implements fuzzy match of strings.
- orafunc: provides compatibility with some Oracle functions.
- oss_ext: reads data from Object Storage Service (OSS).
- HyperLogLog: collects statistics.
- PL/Java: compiles user-defined functions (UDFs) in PL/Java.
- pgcrypto: provides cryptographic functions.
- IntArray: provides integer array-related functions, operators, and indexes.

Create an extension

Execute the following statements to create an extension:

```
CREATE EXTENSION <extension name>;
CREATE SCHEMA <schema name>;
CREATE EXTENSION IF NOT EXISTS <extension name> WITH SCHEMA <schema name>;
```

> **Note**
>
> Before you create a MADlib extension, you must create a plpythonu extension.
>
> ```
> CREATE EXTENSION plpythonu;
> CREATE EXTENSION madlib;
> ```

Delete an extension

Execute the following statements to delete an extension.

```
DROP EXTENSION <extension name>;
DROP EXTENSION IF EXISTS <extension name> CASCADE;
```

> **Note** If an object depends on an extension that you want to delete, you must add the CASCADE keyword to delete the object.

13.1.4.7. Manage users and permissions

This topic describes how to manage users and permissions in .

Manage users

The system prompts you to specify an initial username and password when you create an instance. This initial user is the root user. After the instance is created, you can use the root user account to connect to the database. The system also creates superusers such as aurora and replicator for internal management.

You can run the `\du+` command to view the information of all the users after you connect to the database by using the client tool of PostgreSQL or Greenplum. Example:

```
postgres=> \du+
                                   List of roles
  Role name   |            Attributes             | Member of |  Description
--------------+-----------------------------------+-----------+---------------
 root_user    |                                   |      | rds_superuser
 ...
```

does not provide superuser permissions, but offers a similar role, RDS_SUPERUSER, which is consistent with the permission system of ApsaraDB RDS for PostgreSQL. The root user, such as root_user in the preceding example, has the permissions of the RDS_SUPERUSER role. This permission attribute can only be identified by viewing the user description.

The root user has the following permissions:

- Create databases and users and perform actions such as LOGIN, excluding the SUPERUSER permissions.
- View and modify the data tables of other users and perform actions such as SELECT, UPDATE, DELETE, and changing owners.
- View the connection information of other users, cancel their SQL statements, and kill their connections.
- Create and delete extensions.
- Create other users with RDS_SUPERUSER permissions. Example:

```
CRATE ROLE root_user2 RDS_SUPERUSER LOGIN PASSWORD 'xyz' ;
```

Manage permissions

You can manage permissions at the database, schema, and table levels. For example, if you want to grant read permissions on a table to a user and revoke their write permissions, you can execute the following statements:

```
GRANT SELECT ON TABLE t1 TO normal_user1;
REVOKE UPDATE ON TABLE t1 FROM normal_user1;
REVOKE DELETE ON TABLE t1 FROM normal_user1;
```

13.1.4.8. Manage JSON data

JSON has become a basic data type in the Internet and Internet of things (IoT) fields. For more information about JSON, visit JSON official website. PostgreSQL provides full support for JSON. is optimized by Alibaba Cloud to support the JSON type based on the PostgreSQL syntax.

Check whether the current version supports JSON

Execute the following statement to check whether the current version supports JSON:

```
=> SELECT '""'::json;
```

If the following output is displayed, the JSON type is supported and the instance is ready for use. If the operation fails, restart the instance.

```
 json
------
 ""
(1 row)
```

If the following output is displayed, the JSON type is not supported.

```
ERROR:  type "json" does not exist
LINE 1: SELECT '"'::json;
               ^
```

The preceding command converts data from the string type to the JSON type. PostgreSQL supports operations on JSON data based on this conversion.

JSON conversion in the database

Database operations include reading and writing. The written data is typically converted from the string type to the JSON type. The contents of a string must meet the JSON standard, such as strings, digits, arrays, and objects. Example:

String

```
=> SELECT '"hijson"'::json;
 json
-------
 "hijson"
(1 row)
```

`::` is used for explicit type conversion in PostgreSQL, Greenplum, and . The database calls the input function of the JSON type during the conversion. Therefore, JSON format check is performed. Example:

```
=> SELECT '{hijson:1024}'::json;
ERROR:  invalid input syntax for type json
LINE 1: SELECT '{hijson:1024}'::json;
               ^
DETAIL:  Token "hijson" is invalid.
CONTEXT:  JSON data, line 1: {hijson...
=>
```

In the preceding example, `hijson` must be enclosed in double quotation marks (`"`) because JSON requires the KEY value to be a string. A syntax error is returned when `{hijson:1024}` is entered.

Apart from explicit type conversion, database records can also be converted to JSON.

Typically, JSON is not used for a string or a digit, but an object that contains one or more key-value pairs. can support most JSON scenarios after data is converted from the string type to the object type. Example:

```
=> select row_to_json(row('{"a":"a"}', 'b'));
           row_to_json
---------------------------------
 {"f1":"{\"a\":\"a\"}","f2":"b"}
(1 row)
=> select row_to_json(row('{"a":"a"}'::json, 'b'));
        row_to_json
---------------------------
 {"f1":{"a":"a"},"f2":"b"}
(1 row)
```

You can see the differences between the string and JSON. The entire record is converted into the JSON type.

JSON data types

- Object

  The object is the most frequently used data type in JSON. Example:

```
=> select '{"key":"value"}'::json;
       json
 -----------------
  {"key":"value"}
 (1 row)
```

- Integer and floating point number

  JSON supports only three data types for numeric values: integer, floating-point number, and constant expression. supports the three types.

```
 => SELECT '1024'::json;
  json
 ------
  1024
 (1 row)
 => SELECT '0.1'::json;
  json
 ------
  0.1
 (1 row)
```

  The following information is required in some special situations:

```
 => SELECT '1e100'::json;
  json
 -------
  1e100
 (1 row)
 => SELECT '{"f":1e100}'::json;
     json
 -------------
  {"f":1e100}
 (1 row)
```

  Extra-long numbers are also supported. Example:

```
 => SELECT '9223372036854775808'::json;
         json
 ---------------------
  9223372036854775808
 (1 row)
```

- Array

```
=> SELECT '[[1,2], [3,4,5]]'::json;
        json
 ------------------
  [[1,2], [3,4,5]]
 (1 row)
```

Operators

Operators supported by JSON

```
=> select oprname,oprcode from pg_operator where oprleft = 3114;
 oprname |          oprcode
---------+---------------------------
 ->      | json_object_field
 ->>     | json_object_field_text
 ->      | json_array_element
 ->>     | json_array_element_text
 #>      | json_extract_path_op
 #>>     | json_extract_path_text_op
(6 rows)
```

Basic usage

```
=> SELECT '{"f":"1e100"}'::json -> 'f';
 ?column?
----------
 "1e100"
(1 row)
=> SELECT '{"f":"1e100"}'::json ->> 'f';
 ?column?
----------
 1e100
(1 row)
=> select '{"f2":{"f3":1},"f4":{"f5":99,"f6":"stringy"}}'::json#>array['f4','f6'];
 ?column?
-----------
 "stringy"
(1 row)
=> select '{"f2":{"f3":1},"f4":{"f5":99,"f6":"stringy"}}'::json#>'{f4,f6}';
 ?column?
-----------
 "stringy"
(1 row)
=> select '{"f2":["f3",1],"f4":{"f5":99,"f6":"stringy"}}'::json#>>'{f2,0}';
 ?column?
----------
 f3
(1 row)
```

JSON functions

Supported JSON functions

```
postgres=# \df *json*
                                                  List of functions
   Schema   |           Name            | Result data type |                     Argument data types
|  Type
------------+---------------------------+------------------+-----------------------------------------
-----------------+--------
 pg_catalog | array_to_json             | json             | anyarray
| normal
 pg_catalog | array_to_json             | json             | anyarray, boolean
| normal
 pg_catalog | json_array_element        | json             | from_json json, element_index integer
| normal
 pg_catalog | json_array_element_text   | text             | from_json json, element_index integer
| normal
 pg_catalog | json_array_elements       | SETOF json       | from_json json, OUT value json
| normal
 pg_catalog | json_array_length         | integer          | json
| normal
 pg_catalog | json_each                 | SETOF record     | from_json json, OUT key text, OUT value j
son              | normal
 pg_catalog | json_each_text            | SETOF record     | from_json json, OUT key text, OUT value t
ext              | normal
 pg_catalog | json_extract_path         | json             | from_json json, VARIADIC path_elems text[
]                | normal
 pg_catalog | json_extract_path_op      | json             | from_json json, path_elems text[]
| normal
 pg_catalog | json_extract_path_text    | text             | from_json json, VARIADIC path_elems text[
]                | normal
 pg_catalog | json_extract_path_text_op | text             | from_json json, path_elems text[]
| normal
 pg_catalog | json_in                   | json             | cstring
| normal
 pg_catalog | json_object_field         | json             | from_json json, field_name text
| normal
 pg_catalog | json_object_field_text    | text             | from_json json, field_name text
| normal
 pg_catalog | json_object_keys          | SETOF text       | json
| normal
 pg_catalog | json_out                  | cstring          | json
| normal
 pg_catalog | json_populate_record      | anyelement       | base anyelement, from_json json, use_json
_as_text boolean | normal
 pg_catalog | json_populate_recordset   | SETOF anyelement | base anyelement, from_json json, use_json
_as_text boolean | normal
 pg_catalog | json_recv                 | json             | internal
| normal
 pg_catalog | json_send                 | bytea            | json
| normal
 pg_catalog | row_to_json               | json             | record
| normal
 pg_catalog | row_to_json               | json             | record, boolean
| normal
 pg_catalog | to_json                   | json             | anyelement
| normal
(24 rows)
```

Basic usage

```
=> SELECT array_to_json('{{1,5},{99,100}}'::int[]);
  array_to_json
-----------------
 [[1,5],[99,100]]
(1 row)
=> SELECT row_to_json(row(1,'foo'));
     row_to_json
---------------------
 {"f1":1,"f2":"foo"}
(1 row)
=> SELECT json_array_length('[1,2,3,{"f1":1,"f2":[5,6]},4]');
 json_array_length
-------------------
                 5
(1 row)
=> select * from json_each('{"f1":[1,2,3],"f2":{"f3":1},"f4":null,"f5":99,"f6":"stringy"}') q;
 key |   value
-----+-----------
 f1  | [1,2,3]
 f2  | {"f3":1}
 f4  | null
 f5  | 99
 f6  | "stringy"
(5 rows)
=> select json_each_text('{"f1":[1,2,3],"f2":{"f3":1},"f4":null,"f5":"null"}');
  json_each_text
------------------
 (f1,"[1,2,3]")
 (f2,"{""f3"":1}")
 (f4,)
 (f5,null)
(4 rows)
=> select json_array_elements('[1,true,[1,[2,3]],null,{"f1":1,"f2":[7,8,9]},false]');
  json_array_elements
----------------------
 1
 true
 [1,[2,3]]
 null
 {"f1":1,"f2":[7,8,9]}
 false
(6 rows)
create type jpop as (a text, b int, c timestamp);
=> select * from json_populate_record(null::jpop,'{"a":"blurfl","x":43.2}', false) q;
   a    | b | c
--------+---+---
 blurfl |   |
(1 row)
=>     select * from json_populate_recordset(null::jpop,'[{"a":"blurfl","x":43.2},{"b":3,"c":"2012-01-
20 10:42:53"}]',false) q;
   a    | b |            c
--------+---+--------------------------
 blurfl |   |
        | 3 | Fri Jan 20 10:42:53 2012
(2 rows)
```

Code examples

Create a table

```
create table tj(id serial, ary int[], obj json, num integer);
=> insert into tj(ary, obj, num) values('{1,5}'::int[], '{"obj":1}', 5);
INSERT 0 1
=> select row_to_json(q) from (select id, ary, obj, num from tj) as q;
               row_to_json
-------------------------------------------
 {"f1":1,"f2":[1,5],"f3":{"obj":1},"f4":5}
(1 row)
=> insert into tj(ary, obj, num) values('{2,5}'::int[], '{"obj":2}', 5);
INSERT 0 1
=> select row_to_json(q) from (select id, ary, obj, num from tj) as q;
               row_to_json
-------------------------------------------
 {"f1":1,"f2":[1,5],"f3":{"obj":1},"f4":5}
 {"f1":2,"f2":[2,5],"f3":{"obj":2},"f4":5}
(2 rows)
```

Join multiple tables

```
create table tj2(id serial, ary int[], obj json, num integer);
=> insert into tj2(ary, obj, num) values('{2,5}'::int[], '{"obj":2}', 5);
INSERT 0 1
=> select * from tj, tj2 where tj.obj->>'obj' = tj2.obj->>'obj';
 id |  ary  |    obj    | num | id |  ary  |    obj    | num
----+-------+-----------+-----+----+-------+-----------+-----
  2 | {2,5} | {"obj":2} |   5 |  1 | {2,5} | {"obj":2} |   5
(1 row)
=> select * from tj, tj2 where json_object_field_text(tj.obj, 'obj') = json_object_field_text(tj2.obj,
'obj');
 id |  ary  |    obj    | num | id |  ary  |    obj    | num
----+-------+-----------+-----+----+-------+-----------+-----
  2 | {2,5} | {"obj":2} |   5 |  1 | {2,5} | {"obj":2} |   5
(1 row)
```

Use the JSON function index

```
CREATE TEMP TABLE test_json (
       json_type text,
       obj json
);
=> insert into test_json values('aa', '{"f2":{"f3":1},"f4":{"f5":99,"f6":"foo"}}');
INSERT 0 1
=> insert into test_json values('cc', '{"f7":{"f3":1},"f8":{"f5":99,"f6":"foo"}}');
INSERT 0 1
=> select obj->'f2' from test_json where json_type = 'aa';
 ?column?
----------
 {"f3":1}
(1 row)
=> create index i on test_json (json_extract_path_text(obj, '{f4}'));
CREATE INDEX
=> select * from test_json where json_extract_path_text(obj, '{f4}') = '{"f5":99,"f6":"foo"}';
 json_type |                 obj
-----------+-------------------------------------------
 aa        | {"f2":{"f3":1},"f4":{"f5":99,"f6":"foo"}}
(1 row)
```

> **Note**
>
> JSON data cannot be used as the distribution key and does not support JSON aggregate functions.

The following example describes how to use Python to access the database:

```
#! /bin/env python
import time
import json
import psycopg2
def gpquery(sql):
    conn = None
    try:
        conn = psycopg2.connect("dbname=sanity1x2")
        conn.autocommit = True
        cur = conn.cursor()
        cur.execute(sql)
        return cur.fetchall()
    except Exception as e:
        if conn:
            try:
                conn.close()
            except:
                pass
            time.sleep(10)
        print e
    return None
def main():
    sql = "select obj from tj;"
    #rows = Connection(host, port, user, pwd, dbname).query(sql)
    rows = gpquery(sql)
    for row in rows:
        print json.loads(row[0])
if __name__ == "__main__":
    main()
```

13.1.4.9. Use HyperLogLog

is highly optimized by Alibaba Cloud. It has the features of Greenplum Database and supports HyperLogLog. AnalyticDB for PostgreSQL is suited for industries such as Internet advertising and estimation analysis that require quick estimation of business metrics such as page views (PVs) and unique visitors (UVs).

Create a HyperLogLog extension

Execute the following statement to create a HyperLogLog extension:

```
CREATE EXTENSION hll;
```

Basic types

- Execute the following statement to create a table containing the hll field:

  ```
  create table agg (id int primary key,userids hll);
  ```

- Execute the following statement to convert int to hll_hashval:

  ```
  select 1::hll_hashval;
  ```

Basic operators

- The hll type supports =, !=,<>, ||, and #.

```
select hll_add_agg(1::hll_hashval) = hll_add_agg(2::hll_hashval);
select hll_add_agg(1::hll_hashval) || hll_add_agg(2::hll_hashval);
select #hll_add_agg(1::hll_hashval);
```

- The hll_hashval type supports =, !=,and <>.

```
select 1::hll_hashval = 2::hll_hashval;
select 1::hll_hashval <> 2::hll_hashval;
```

Basic functions

- Hash functions such as hll_hash_boolean, hll_hash_smallint, and hll_hash_bigint.

```
select hll_hash_boolean(true);
select hll_hash_integer(1);
```

- hll_add_agg: converts the int format to the hll format.

```
select hll_add_agg(1::hll_hashval);
```

- hll_union: aggregates the hll fields.

```
select hll_union(hll_add_agg(1::hll_hashval),hll_add_agg(2::hll_hashval));
```

- hll_set_defaults: sets the precision.

```
select hll_set_defaults(15,5,-1,1);
```

- hll_print: displays debugging information.

```
select hll_print(hll_add_agg(1::hll_hashval));
```

Examples

```
create table access_date (acc_date date unique, userids hll);
insert into access_date select current_date, hll_add_agg(hll_hash_integer(user_id)) from generate_seri
es(1,10000) t(user_id);
insert into access_date select current_date-1, hll_add_agg(hll_hash_integer(user_id)) from generate_se
ries(5000,20000) t(user_id);
insert into access_date select current_date-2, hll_add_agg(hll_hash_integer(user_id)) from generate_se
ries(9000,40000) t(user_id);
postgres=# select #userids from access_date where acc_date=current_date;
     ?column?
------------------
 9725.85273370708
(1 row)
postgres=# select #userids from access_date where acc_date=current_date-1;
     ?column?
------------------
 14968.6596883279
(1 row)
postgres=# select #userids from access_date where acc_date=current_date-2;
     ?column?
------------------
 29361.5209149911
(1 row)
```

13.1.4.10. Use the CREATE LIBRARY statement

introduces the CREATE LIBRARY and DROP LIBRARY statements to help you import custom software packages.

Syntax

```
CREATE LIBRARY library_name LANGUAGE [JAVA] FROM oss_location OWNER ownername
CREATE LIBRARY library_name LANGUAGE [JAVA] VALUES file_content_hex OWNER ownername
DROP LIBRARY library_name
```

Parameters

Parameter	Description
library_name	The name of the library that you want to install. If the library that you want to install has the same name as an existing library, you must delete the existing library before you install the new one.
LANGUAGE [JAVA]	The programming language that you want to use. Only PL/Java is supported.
oss_location	The location of the package. You can specify the Object Storage Service (OSS) bucket and object names. Only one object can be specified and the specified object cannot be a compressed file. Sample format: `oss://oss_endpoint filepath=[folder/[folder/]...]/file_name id=userossid key=userosskey bucket=ossbucket`
file_content_hex	The content of the file. The byte stream is in hexadecimal notation. For example, 73656c6563742031 indicates the hexadecimal byte stream of "select 1". You can use this syntax to import packages without the need to use OSS.
ownername	The user.
DROP LIBRARY	Deletes a library.

Examples

- Example 1: Install a JAR package named analytics.jar.

```
create library example language java from 'oss://oss-cn-hangzhou.aliyuncs.com filepath=analytics.jar id=xxx key=yyy bucket=zzz';
```

- Example 2: Import the file content with the byte stream in hexadecimal notation.

```
create library pglib LANGUAGE java VALUES '73656c6563742031' OWNER "myuser";
```

- Example 3: Delete a library.

```
drop library example;
```

- Example 4: View installed libraries.

```
select name, lanname from pg_library;
```

13.1.4.11. Create and use a PL/Java UDF

allows you to compile and upload JAR software packages that are written in PL/Java language, and use these JAR packages to create user-defined functions (UDFs). supports PL/Java 1.5.0 and JVM 1.8. This topic describes how to create a PL/Java UDF. For more information about PL/Java examples, see PL/Java code. For more information about the compiling method, see PL/Java documentation.

Procedure

1. In , execute the following statement to create a PL/Java extension. You need only to execute the statement once for each database.

```
create extension pljava;
```

2. Compile the UDF based on your business requirements. For example, you can use the following code to compile the Test.java file:

```
public class Test
 {
     public static String substring(String text, int beginIndex,
               int endIndex)
             {
                     try {
                         Process process = null;
                         process = Runtime.getRuntime().exec("echo Test running");
                     } catch (Exception e) {
                        return "" + e;
                     }
                     return text.substring(beginIndex, endIndex);
             }
 }
```

3. Compile the manifest.txt file:

```
 Manifest-Version: 1.0
 Main-Class: Test
 Specification-Title: "Test"
 Specification-Version: "1.0"
 Created-By: 1.7.0_99
 Build-Date: 01/20/2016 21:00 AM
```

4. Run the following commands to compile and package the program:

```
javac Test.java
 jar cfm analytics.jar manifest.txt Test.class
```

5. Upload the analytics.jar file generated in Step 4 to Object Storage Service (OSS) by using the following OSS console command:

```
osscmd put analytics.jar oss://zzz
```

6. In , execute the CREATE LIBRARY statement to import the file to .

```
create library example language java from 'oss://oss-cn-hangzhou.aliyuncs.com filepath=analytics.j
ar id=xxx key=yyy bucket=zzz';
```

> **Note** You can only use the filepath variable in the CREATE LIBRARY statement to import files one at a time. The CREATE LIBRARY statement also supports byte streams to import files without the need to use OSS. For more information, see Use the CREATE LIBRARY statement.

7. In , execute the following statements to create and use the UDF:

```
create table temp (a varchar) distributed randomly;
 insert into temp values ('my string');
 create or replace function java_substring(varchar, int, int) returns varchar as 'Test.substring'
language java;
 select java_substring(a, 1, 5) from temp;
```

13.1.5. Table

13.1.5.1. Create a table

You can create tables within your databases.

Syntax

The following statement shows how to create a table. Not all clauses are required. Use the clauses that can fulfill your business needs.

```
CREATE [[GLOBAL | LOCAL] {TEMPORARY | TEMP}] TABLE table_name (
[ { column_namedata_type [ DEFAULT default_expr ]
   [column_constraint [ ... ]
[ ENCODING ( storage_directive [,...] ) ]
]
   | table_constraint
   | LIKE other_table [{INCLUDING | EXCLUDING}
                      {DEFAULTS | CONSTRAINTS}] ...}
   [, ... ] ]
   )
   [ INHERITS ( parent_table [, ... ] ) ]
   [ WITH ( storage_parameter=value [, ... ] )
   [ ON COMMIT {PRESERVE ROWS | DELETE ROWS | DROP} ]
   [ TABLESPACE tablespace ]
   [ DISTRIBUTED BY (column, [ ... ] ) | DISTRIBUTED RANDOMLY ]
   [ PARTITION BY partition_type (column)
       [ SUBPARTITION BY partition_type (column) ]
          [ SUBPARTITION TEMPLATE ( template_spec ) ]
       [...]
    ( partition_spec )
        | [ SUBPARTITION BY partition_type (column) ]
          [...]
    ( partition_spec
      [ ( subpartition_spec
           [(...)]
        ) ]
    )
```

Definition of the column_constraint clause:

```
[CONSTRAINT constraint_name]
   NOT NULL | NULL
   | UNIQUE [USING INDEX TABLESPACE tablespace]
            [WITH ( FILLFACTOR = value )]
   | PRIMARY KEY [USING INDEX TABLESPACE tablespace]
                 [WITH ( FILLFACTOR = value )]
   | CHECK ( expression )
   | REFERENCES table_name [ ( column_name [, ... ] ) ]
            [ key_match_type ]
            [ key_action ]
```

Definition of the storage_directive clause of columns:

```
   COMPRESSTYPE={ZLIB | QUICKLZ | RLE_TYPE | NONE}
    [COMPRESSLEVEL={0-9} ]
    [BLOCKSIZE={8192-2097152} ]
```

Definition of the storage_parameter clause of tables:

```
APPENDONLY={TRUE|FALSE}
BLOCKSIZE={8192-2097152}
ORIENTATION={COLUMN|ROW}
CHECKSUM={TRUE|FALSE}
COMPRESSTYPE={ZLIB|QUICKLZ|RLE_TYPE|NONE}
COMPRESSLEVEL={0-9}
FILLFACTOR={10-100}
OIDS[=TRUE|FALSE]
```

Definition of the table_constraint clause:

```
[CONSTRAINT constraint_name]
UNIQUE ( column_name [, ... ] )
       [USING INDEX TABLESPACE tablespace]
       [WITH ( FILLFACTOR=value )]
| PRIMARY KEY ( column_name [, ... ] )
              [USING INDEX TABLESPACE tablespace]
              [WITH ( FILLFACTOR=value )]
| CHECK ( expression )
| FOREIGN KEY ( column_name [, ... ] )
         REFERENCES table_name [ ( column_name [, ... ] ) ]
         [ key_match_type ]
         [ key_action ]
         [ key_checking_mode ]
```

Valid values of key_match_type:

```
  MATCH FULL
| SIMPLE
```

Valid values of key_action:

```
  ON DELETE
| ON UPDATE
| NO ACTION
| RESTRICT
| CASCADE
| SET NULL
| SET DEFAULT
```

Valid values of key_checking_mode:

```
  DEFERRABLE
| NOT DEFERRABLE
| INITIALLY DEFERRED
| INITIALLY IMMEDIATE
```

Valid values of partition_type:

```
  LIST
| RANGE
```

Definition of the partition_specification clause:

```
partition_element [, ...]
```

Definition of the partition_element clause:

```
   DEFAULT PARTITION name
  | [PARTITION name] VALUES (list_value [,...] )
  | [PARTITION name]
     START ([datatype] 'start_value') [INCLUSIVE | EXCLUSIVE]
     [ END ([datatype] 'end_value') [INCLUSIVE | EXCLUSIVE] ]
     [ EVERY ([datatype] [number | INTERVAL] 'interval_value') ]
  | [PARTITION name]
     END ([datatype] 'end_value') [INCLUSIVE | EXCLUSIVE]
     [ EVERY ([datatype] [number | INTERVAL] 'interval_value') ]
[ WITH ( partition_storage_parameter=value [, ... ] ) ]
[ TABLESPACE tablespace ]
```

Definition of the subpartition_spec or template_spec clause:

```
subpartition_element [, ...]
```

Definition of the subpartition_element clause:

```
   DEFAULT SUBPARTITION name
  | [SUBPARTITION name] VALUES (list_value [,...] )
  | [SUBPARTITION name]
     START ([datatype] 'start_value') [INCLUSIVE | EXCLUSIVE]
     [ END ([datatype] 'end_value') [INCLUSIVE | EXCLUSIVE] ]
     [ EVERY ([datatype] [number | INTERVAL] 'interval_value') ]
  | [SUBPARTITION name]
     END ([datatype] 'end_value') [INCLUSIVE | EXCLUSIVE]
     [ EVERY ([datatype] [number | INTERVAL] 'interval_value') ]
[ WITH ( partition_storage_parameter=value [, ... ] ) ]
[ TABLESPACE tablespace ]
```

Definition of the storage_parameter clause:

```
   APPENDONLY={TRUE|FALSE}
   BLOCKSIZE={8192-2097152}
   ORIENTATION={COLUMN|ROW}
   CHECKSUM={TRUE|FALSE}
   COMPRESSTYPE={ZLIB|QUICKLZ|RLE_TYPE|NONE}
   COMPRESSLEVEL={1-9}
   FILLFACTOR={10-100}
   OIDS[=TRUE|FALSE]
```

Parameters

The Table creation parameters table describes the key parameters for creating a table.

Table creation parameters

Parameter	Description
TABLE_NAME	The name of the table that you want to create.
column_name	The name of a column that you want to create in the table.

Parameter	Description
data_type	The data type of the column. For columns that contain textual data, set the data type to VARCHAR or TEXT. We recommend that you do not use the CHAR type.
DEFAULT default_expr	Specifies a default value for the column. The system assigns this default value to all columns that do not have a value. The default value can be a variable-free expression. Subqueries or cross-references to other columns in the table are not allowed. The data type of the default expression must match the data type of the column. If a column does not have a default value, this parameter is left empty.
ENCODING storage_directive	Specifies the compression type and block size for the column data. This clause is valid only for append-optimized, column-oriented tables. Column compression settings are inherited from the table level to the partition level and then to the sub-partition level. The lowest-level settings have the highest priority.
INHERITS	Configures all columns in the new table to automatically inherit a parent table. You can use INHERITS to create a persistent relationship between the new child table and its parent table. Schema modifications to the parent table are applied to the child table as well. When the parent table is also scanned, the data of the child table is scanned as well.
LIKE other_table	Specifies a table from which the new table automatically copies all column names, data types, NOT NULL constraints, and distribution policies. Storage properties such as append-optimized or partition structure are not copied. Unlike INHERITS, the new table is completely decoupled from the original table after the new table is created.
CONSTRAINT constraint_name	Configures a column or table constraint. If a constraint is violated, the constraint name is displayed in the error message. Constraint names can be used to communicate helpful information to client applications.Constraint names that contain spaces must be enclosed by double quotation marks ('').
WITH (storage_option=value)	Configures storage options for the table or its indexes.
ON COMMIT	The operation that the system performs on the temporary tables at the end of a transaction. Valid values: • **PRESERVE ROWS**: No special action is taken. The data is retained after the transaction is complete. The data is released only when the session is disconnected. • **DELETE ROWS**: All rows in the temporary table are deleted. • **DROP**: The temporary table is deleted.
TABLESPACE tablespace	Specifies the name of the tablespace in which you want to create the new table. If this parameter is not specified, the default tablespace of the database is used.

Parameter	Description
DISTRIBUTED BY	Configures the distribution policy for the database. • DISTRIBUTED BY (column, [...]): specifies the distribution key. The system distributes data based on the distribution key. To evenly distribute data, you must set the distribution key to the primary key of the table or a unique column or a set of columns. • DISTRIBUTED RANDOMLY: randomly distributes data. Note We recommend that you do not use random distribution.
PARTITION BY	Configures a partition key to partition the table. Partitioning large tables improves data access efficiency. To partition a table is to create a top-level (parent) table and multiple lower-level (child) tables. After a partition table is created, its parent table is always empty. Data is stored in the lowest-level child tables. In a multi-level partition table, data is stored only in the lowest-level sub-partitions. Valid values: RANGE, LIST, and a combination of the two.
SUBPARTITION BY	Configures a multi-level partition table.
SUBPARTITION TEMPLATE	Specifies a sub-partition template to create sub-partitions (lower-level child tables). This ensures that all parent partitions have the same sub-partition structure.

Examples

Create a table and configure a distribution key. By default, a primary key is used as a distribution key in .

```
CREATE TABLE films (
code        char(5) CONSTRAINT firstkey PRIMARY KEY,
title       varchar(40) NOT NULL,
did         integer NOT NULL,
date_prod   date,
kind        varchar(10),
len         interval hour to minute
);
CREATE TABLE distributors (
did    integer PRIMARY KEY DEFAULT nextval('serial'),
name   varchar(40) NOT NULL CHECK (name <> '')
);
```

Create a compressed table and configure a distribution key.

```
CREATE TABLE sales (txn_id int, qty int, date date)
WITH (appendonly=true, compresslevel=5)
DISTRIBUTED BY (txn_id);
```

Use sub-partition templates of each level and the default partition to create a three-level partition table.

```
CREATE TABLE sales (id int, year int, month int, day int,
region text)
DISTRIBUTED BY (id)
PARTITION BY RANGE (year)
  SUBPARTITION BY RANGE (month)
    SUBPARTITION TEMPLATE (
       START (1) END (13) EVERY (1),
       DEFAULT SUBPARTITION other_months )
  SUBPARTITION BY LIST (region)
    SUBPARTITION TEMPLATE (
       SUBPARTITION usa VALUES ('usa'),
       SUBPARTITION europe VALUES ('europe'),
       SUBPARTITION asia VALUES ('asia'),
       DEFAULT SUBPARTITION other_regions)
( START (2008) END (2016) EVERY (1),
  DEFAULT PARTITION outlying_years);
```

13.1.5.2. Principles and scenarios of row store, column store, heap tables, and AO tables

supports row store, column store, heap tables, and append-optimized (AO) tables. This topic describes their principles and scenarios.

Row store and column store

Comparison

Dimension	Row store	Column store
Definition	Row store stores data in the form of rows. Each row is a tuple. To read a column, you must deform all of the columns that precede it. Therefore, the costs for accessing the first and the last columns are different.	Column store stores data as columns. Each column corresponds to a file or a batch of files. The cost of reading each column is the same. However, if you want to read multiple columns, you must access multiple files. The more columns you access, the higher the overheads are.
Compression ratio	Low.	High.
Cost of reading a column	The higher the column number, the higher the cost of reading the column.	Same.
Vector computing and JIT architecture	Not suitable. Not suitable for batch computing.	Suitable. More efficient when you access and collect statistics of a batch of data.

Dimension	Row store	Column store
Scenarios	If you need to perform a large number of update and delete operations due to online transaction processing (OLTP) requirements such as when you query table details where multiple columns are returned, you can use row store. You can use partition tables if you have complex requirements. For example, if you want to partition data based on time, you can use row store to query the details of recent data and use column store to obtain more statistics from historical data.	You can use column store if you need data statistics because of online analytical processing (OLAP) requirements. If you want a higher compression ratio, you can use column store.

Heap tables

A heap table is heap storage. All changes to the heap table generate redo logs that can be used to restore data by point in time. However, heap tables cannot implement logical incremental backup because all data blocks in the tables can be changed and it is inconvenient to record the position by using heap storage.

Commit and redo logs are used to ensure reliability when transactions are finished. You can also implement redundancy by using secondary nodes through redo logs.

AO tables

AO tables are used to append data for storage. When you delete the updated data, you can use another bitmap file to mark the row that you want to delete and then use the offset of the bit to determine whether the row was deleted.

When the transaction is finished, you must call the fsync() function to record the offset of the data block that performs the last write operation. Even if the data block contains only a single record, a new data block is appended for the next transaction. The data block is synchronized to the secondary node for data redundancy.

AO tables are not suitable for small transactions because the fsync() function is called at the end of each transaction. This data block is not reused even if space is left.

AO tables are suitable for OLAP scenarios, batch data writing, high compression ratio, and logical backup that supports incremental backup. During backup, you need only to record the offset from the backup and the bitmap deletion mark for each full backup.

Use scenarios of heap tables

- When multiple small transactions are handled, use a heap table.
- When you need to restore data by point in time, use a heap table.

Use scenarios of AO tables

- When you want to use column store, use an AO table.
- When data is written in batches, use an AO table.

13.1.5.3. Enable the column store and compression features

If you want to improve performance, speed up data import, or reduce costs for tables that have infrequent updates and multiple fields, we recommend that you use column store and compression. These features increase the compression ratio by up to threefold to ensure high performance and speed up data import.

To enable the column store and compression features, you must specify the column store and compression options when you create a table. For example, you can add the following clause to the CREATE statement to enable these features. For more information about the table creation syntax, see Create a table.

```
with (APPENDONLY=true, ORIENTATION=column, COMPRESSTYPE=zlib, COMPRESSLEVEL=5, BLOCKSIZE=1048576, OIDS
=false)
```

Note AnalyticDB for PostgreSQL supports only zlib and RLE_TYPE compression algorithms. If you specify the quicklz algorithm, it is automatically converted to zlib.

13.1.5.4. Add a field to a column store table and set the default value

This topic describes how to add a field to a column store table and set the default value for the field, and how to use the ANALYZE statement to view the impact of updated data on the size of the column store table.

Context

In a column store table, each column is stored as a file, and two columns in the same row correspond to each other by using the offset. For example, if you add two fields of the INT8 type, you can quickly locate column B from column A by using the offset.

When you add the field, AO tables are not rewritten. If an AO table contains the records of deleted data, the added field must be filled with the deleted records before using the offset.

Procedure

1. Create three AO column store tables.

```
postgres=# create table tbl1 (id int, info text) with (appendonly=true, blocksize=8192, compressty
pe=none, orientation=column);
NOTICE:  Table doesn't have 'DISTRIBUTED BY' clause -- Using column named 'id' as the Greenplum Da
tabase data distribution key for this table.
HINT:  The 'DISTRIBUTED BY' clause determines the distribution of data. Make sure column(s) chosen
are the optimal data distribution key to minimize skew.
CREATE TABLE
postgres=# create table tbl2 (id int, info text) with (appendonly=true, blocksize=8192, compressty
pe=none, orientation=column);
NOTICE:  Table doesn't have 'DISTRIBUTED BY' clause -- Using column named 'id' as the Greenplum Da
tabase data distribution key for this table.
HINT:  The 'DISTRIBUTED BY' clause determines the distribution of data. Make sure column(s) chosen
are the optimal data distribution key to minimize skew.
CREATE TABLE
postgres=# create table tbl3 (id int, info text) with (appendonly=true, blocksize=8192, compressty
pe=none, orientation=column);
NOTICE:  Table doesn't have 'DISTRIBUTED BY' clause -- Using column named 'id' as the Greenplum Da
tabase data distribution key for this table.
HINT:  The 'DISTRIBUTED BY' clause determines the distribution of data. Make sure column(s) chosen
are the optimal data distribution key to minimize skew.
CREATE TABLE
```

2. Insert 10 million entries to the first two tables and 20 million entries to the third one.

```
postgres=# insert into tbl1 select generate_series(1,10000000),'test';
INSERT 0 10000000
postgres=# insert into tbl2 select generate_series(1,10000000),'test';
INSERT 0 10000000
postgres=# insert into tbl3 select generate_series(1,20000000),'test';
INSERT 0 20000000
```

3. Analyze the tables and display their sizes.

```
postgres=# analyze tbl1;
ANALYZE
postgres=# analyze tbl2;
ANALYZE
postgres=# analyze tbl3;
ANALYZE
postgres=# select pg_size_pretty(pg_relation_size('tbl1'));
 pg_size_pretty
----------------
 88 MB
(1 row)
postgres=# select pg_size_pretty(pg_relation_size('tbl2'));
 pg_size_pretty
----------------
 88 MB
(1 row)
postgres=# select pg_size_pretty(pg_relation_size('tbl3'));
 pg_size_pretty
----------------
 173 MB
(1 row)
```

4. Update all the data in the first table. Display the table size after the update. The size is twice as large as the size before the update.

```
postgres=# update tbl1 set info='test';
UPDATE 10000000
postgres=# analyze tbl1;
ANALYZE
postgres=# select pg_size_pretty(pg_relation_size('tbl1'));
 pg_size_pretty
----------------
 173 MB
(1 row)
```

5. Add fields to the three tables and set the default values.

```
postgres=# alter table tbl1 add column c1 int8 default 1;
ALTER TABLE
postgres=# alter table tbl2 add column c1 int8 default 1;
ALTER TABLE
postgres=# alter table tbl3 add column c1 int8 default 1;
ALTER TABLE
```

6. Analyze the tables and view the table sizes.

```
postgres=# analyze tbl1;
ANALYZE
postgres=# analyze tbl2;
ANALYZE
postgres=# analyze tbl3;
ANALYZE
postgres=# select pg_size_pretty(pg_relation_size('tbl1'));
 pg_size_pretty
----------------
 325 MB
(1 row)
postgres=# select pg_size_pretty(pg_relation_size('tbl2'));
 pg_size_pretty
----------------
 163 MB
(1 row)
postgres=# select pg_size_pretty(pg_relation_size('tbl3'));
 pg_size_pretty
----------------
 325 MB
(1 row)
```

When you add fields to the AO tables, the number of entries in the existing files will prevail. Even if all the entries are deleted, you must initialize the original data in the newly added fields.

13.1.5.5. Configure table partitions

For fact tables and large-sized tables in a database, we recommend that you configure table partitions.

Configure table partitions

You can use the table partitioning feature to add and remove data based on table partitions on a regular basis. You can use the `ALTER TABLE DROP PARTITION` statement to remove all the data in a partition, and use the `ALTER TABLE EXCHANGE PARTITION` statement to import data to a new data partition.

supports range partitioning, list partitioning, and composite partitioning. Range partitioning supports only the numeric or datetime data types of fields.

The following example shows how to use range partitioning in a table:

```
CREATE TABLE LINEITEM (
L_ORDERKEY              BIGINT NOT NULL,
L_PARTKEY               BIGINT NOT NULL,
L_SUPPKEY               BIGINT NOT NULL,
L_LINENUMBER    INTEGER,
L_QUANTITY              FLOAT8,
L_EXTENDEDPRICE FLOAT8,
L_DISCOUNT              FLOAT8,
L_TAX                   FLOAT8,
L_RETURNFLAG    CHAR(1),
L_LINESTATUS    CHAR(1),
L_SHIPDATE              DATE,
L_COMMITDATE    DATE,
L_RECEIPTDATE   DATE,
L_SHIPINSTRUCT  CHAR(25),
L_SHIPMODE              CHAR(10),
L_COMMENT               VARCHAR(44)
)  WITH  (APPENDONLY=true, ORIENTATION=column, COMPRESSTYPE=zlib, COMPRESSLEVEL=5, BLOCKSIZE=1048576,
OIDS=false) DISTRIBUTED BY (l_orderkey)
PARTITION BY RANGE (L_SHIPDATE) (START (date '1992-01-01') INCLUSIVE END (date '2000-01-01') EXCLUSIVE
EVERY (INTERVAL '1 month' ));
```

Principles of table partitioning

The purpose of partitioning is to minimize the amount of data to be scanned during a query. Therefore, partitions must be associated with the query conditions.

- Principle 1: Select the fields that are related to the query conditions to configure partitions and reduce the amount of data to be scanned.
- Principle 2: If multiple query conditions exist, configure sub-partitions to further reduce the amount of data to be scanned.

13.1.5.6. Configure the sort key

A sort key is an attribute of a table. Data on disks is stored in the order of the sort key.

Context

Sort keys have two major advantages:

- Speed up and optimize column-store operations. The min and max meta information the system collects seldom overlaps with each other, which features good filterability.
- Eliminate the need to perform ORDER BY and GROUP BY operations. The data directly read from the disk is ordered as required by the sorting conditions.

Create a table

```
Command:     CREATE TABLE
Description: define a new table
Syntax:
CREATE [[GLOBAL | LOCAL] {TEMPORARY | TEMP}] TABLE table_name (
[ { column_name data_type [ DEFAULT default_expr ]     [column_constraint [ ... ]
[ ENCODING ( storage_directive [,...] ) ]
]
   | table_constraint
   | LIKE other_table [{INCLUDING | EXCLUDING}
                      {DEFAULTS | CONSTRAINTS}] ...}
   [, ... ] ]
   [column_reference_storage_directive [, ] ]
   )
   [ INHERITS ( parent_table [, ... ] ) ]
   [ WITH ( storage_parameter=value [, ... ] )
   [ ON COMMIT {PRESERVE ROWS | DELETE ROWS | DROP} ]
   [ TABLESPACE tablespace ]
   [ DISTRIBUTED BY (column, [ ... ] ) | DISTRIBUTED RANDOMLY ]
   [ SORTKEY (column, [ ... ] )]
   [ PARTITION BY partition_type (column)
       [ SUBPARTITION BY partition_type (column) ]
          [ SUBPARTITION TEMPLATE ( template_spec ) ]
       [...]
    ( partition_spec )
        | [ SUBPARTITION BY partition_type (column) ]
          [...]
    ( partition_spec
      [ ( subpartition_spec
           [(...)]
         ) ]
)
```

Examples:

```
create table test(date text, time text, open float, high float, low float, volume int) with(APPENDONLY
=true,ORIENTATION=column) sortkey (volume);
```

Sort the table

```
VACUUM SORT ONLY [tablename]
```

Modify the sort key

This statement only modifies the catalog and does not sort data. You must execute the `VACUUM SORT ONLY` statement to sort the table.

```
ALTER [[GLOBAL | LOCAL] {TEMPORARY | TEMP}] TABLE table_name SET SORTKEY (column, [ ... ] )
```

Examples:

```
alter table test set sortkey (high,low);
```

13.1.6. Best practices

13.1.6.1. Configure memory and load parameters

To improve database stability, you must configure memory and load parameters.

Background information

is a massively parallel processing (MPP) database service with high computational and resource requirements. AnalyticDB for PostgreSQL consumes all of its resources. This allows AnalyticDB for PostgreSQL to reach high processing speeds but makes it very easy to reach its limits.

When CPU, network, or hard disk resources are exhausted, the worst that can occur is a hardware bottleneck. However, if all memory resources are completely consumed, the database may not respond.

How to avoid OOM errors

Out of memory (OOM) indicates that the system is unable to provide sufficient memory requested by a process. The following prompt appears when OOM errors occur:

```
Out of memory (seg27 host.example.com pid=47093) VM Protect failed to allocate 4096 bytes, 0 MB availa
ble
```

Causes

OOM errors occur due to the following causes:

- Database nodes do not have sufficient memory.
- Kernel parameters related to the memory of the operating system are incorrectly configured.
- Data skew has occurred. This causes a compute node to request a large amount of memory.
- Query skew has occurred. For example, if the grouping fields of some aggregate or window functions are not distribution keys, the data must be redistributed. After the data is redistributed, data is skewed on a specific computer node and results in the node requesting a large amount of memory.

Solutions

1. Modify the queries to request less memory.
2. Use a resource queue of to limit the number of concurrent queries. Reduce the number of queries that are executed within the cluster at the same time to reduce the overall memory requested by the system.
3. Reduce the number of compute nodes deployed on a host. For example, you can deploy 8 compute nodes instead of 16 on a host that has 128 GB of memory to allow each compute node to utilize twice as much memory.
4. Increase the memory of a host.
5. Set the gp_vmem_protect_limit parameter to limit the maximum VMEM that can be used by a single compute node. The memory size of a single host and the number of compute nodes deployed on the host determine the maximum memory size that a single compute node can use on average.
6. For SQL statements that have unpredictable memory usage, you can set the statement_mem parameter in the session to limit the memory usage of a single SQL statement. This prevents a single SQL statement from consuming all available memory.
7. Set the statement_mem parameter at the database level to apply to all the sessions in the database.
8. Use the resource queue of to limit the maximum memory usage of the resource group. Add database users to the resource group to limit the total memory used by these users.

Configure memory-related parameters

You can properly configure the operating system, database parameters, and resource queue to effectively reduce the probability of OOM.

When you calculate the average memory usage of a single compute node on a single host, you must consider both the primary and secondary compute nodes. If the cluster encounters a host failure, the system switches the service from primary compute nodes to the corresponding secondary compute nodes. At this time, the number of compute nodes on the host is greater than usual. Therefore, you must consider the amount of resources occupied by the secondary compute nodes during failover.

The following tables describe how to configure parameters for the operating system kernel and database to avoid OOM.

The following Operating system kernel parameters table describes the parameter configuration of the operating system kernel.

Operating system kernel parameters

Parameter	Description
huge page	Do not configure the huge page parameter of the system. does not support the latest version of PostgreSQL and therefore does not support the huge page feature. The huge page parameter locks a part of the allocated memory. Database nodes are not able to use this part of the memory.
vm.overcommit_memory	If you use the swap space, set this parameter to 2. If you do not use the swap space, set this parameter to 0. Valid values: • 0: The requested memory space cannot exceed the difference between the total memory and the resident set size (RSS). An error is returned only when the memory has been exceeded. • 1: Most processes use the malloc function to request memory, but do not use all of the requested memory. When this parameter is set to 1, the memory requested by the malloc function is allocated in all circumstances unless the memory is not sufficient. • 2: The swap space is also considered when the system calculates the memory space that can be requested. You can request a large amount of memory even if the swap space is triggered.
overcommit_ratio	The larger the value is, the more memory the process can request. However, less space is reserved for the operating system. For more information about the formula that is used to calculate the memory parameters, see Examples to calculate the memory parameters. When this parameter is set to 2, the memory that can be requested cannot exceed swap + memory × overcommit_ratio.

The following Database parameters table describes the parameter configuration of the database.

Database parameters

Parameter	Description
gp_vmem_protect_limit	The maximum percentage of memory that all processes can request on each node. If the value is too large, it may result in a system OOM error or even more serious problems. If the value is too small, SQL statements may not be executed even when the system has enough memory.

Parameter	Description
runaway_detector_activation_percent	Default value: 90. Unit: %. When the amount of memory used by a compute node exceeds runaway_detector_activation_percent × gp_vmem_protect_limit/100, the query is terminated to prevent OOM. The node remains stopped until the memory usage reaches a value lower than runaway_detector_activation_percent × gp_vmem_protect_limit/100. You can use the gp_toolkit.session_level_memory_consumption view to observe the memory usage of each session and runaway information.
statement_mem	The maximum amount of memory that a single SQL statement can request. When the maximum memory is exceeded, spill files are created. Default value: 125. Unit: MB. We recommend that you set this parameter based on the following formula: `(gp_vmem_protect_limit * 0.9) / max_expected_concurrent_queries` **Note** • You can specify the statement_mem parameter in a session. If few concurrent queries exist and if the session needs to execute a query that requires a large amount of memory, you must specify this parameter in the session. • The statement_mem parameter is suitable for limiting memory usage in low concurrency scenarios. If you use statement_mem to limit the memory for high concurrency scenarios, each query is allocated with a very small amount of memory. As a result, the performance of a small number of queries with high memory requirements in high concurrency scenarios is affected. We recommend that you use the resource queue to limit the maximum memory usage in high concurrency scenarios.
gp_workfile_limit_files_per_query	The maximum number of spill files that can be created by each query. When the amount of memory requested by the query exceeds the statement_mem limit, spill files (also known as work files) are created. Spill files are similar to the swap space of the operating system. When the number of spill files used exceeds the limit, the query is terminated. The default value is 0, which indicates that an unlimited number of spill files can be created.
gp_workfile_compress_algorithm	The compression algorithm for spill files. Valid values: NONE and ZLIB. This parameter can optimize storage space or I/O by sacrificing CPU performance. You can set this parameter when the disk space is insufficient or when the spill files encounter a write bottleneck.

Examples to calculate the memory parameters

The following environment configurations are used in the examples:

- Host configuration:

```
Total RAM = 256GB
SWAP = 64GB
```

- Four hosts, each deployed with eight primary compute nodes and eight secondary compute nodes.

  When a host fails, the eight primary compute nodes are distributed to the remaining three hosts. A single host can be deployed with at most three extra primary compute nodes from the failed host. A single host can be deployed with at most 11 primary compute nodes.

  1. Calculate the total memory allocated to by the operating system.

     Reserve 7.5 GB and 5% of memory for the operating system and calculate the available memory for all applications. Then, divide the available memory by an empirical coefficient of 1.7.

     ```
     gp_vmem = ((SWAP + RAM) - (7.5 GB + 0.05 × RAM))/1.7
             = ((64 + 256) - (7.5 + 0.05 * 256)) / 1.7
             = 176
     ```

  2. Use an empirical coefficient of 0.026 to calculate overcommit_ratio.

     ```
     vm.overcommit_ratio = (RAM - (0.026 * gp_vmem)) / RAM
                         = (256 - (0.026 * 176)) / 256
                         = .982
     Set vm.overcommit_ratio to 98.
     ```

  3. Calculate the gp_vmem_protect_limit parameter by dividing gp_vmem by maximum_acting_primary_segments. The maximum_acting_primary_segments parameter indicates the number of primary compute nodes to be run on each other host after one host fails.

     ```
     gp_vmem_protect_limit calculation
     gp_vmem_protect_limit = gp_vmem / maximum_acting_primary_segments
                           = 176 / 11
                           = 16GB
                           = 16384MB
     ```

Configure resource queues

You can use resource queues of to limit the number of concurrent queries and the total memory usage. When a query is being executed, it is added to the corresponding queue and the resources used are recorded in the queue. The resource limit of the queue is applied to all sessions in the queue.

The resource queue in is similar to cgroup in Linux.

The following syntax demonstrates how to create a resource queue:

```
Command:     CREATE RESOURCE QUEUE
Description: create a new resource queue for workload management
Syntax:
CREATE RESOURCE QUEUE name WITH (queue_attribute=value [, ... ])
where queue_attribute is:
   ACTIVE_STATEMENTS=integer
        [ MAX_COST=float [COST_OVERCOMMIT={TRUE|FALSE}] ]
        [ MIN_COST=float ]
        [ PRIORITY={MIN|LOW|MEDIUM|HIGH|MAX} ]
        [ MEMORY_LIMIT='memory_units' ]
|  MAX_COST=float [ COST_OVERCOMMIT={TRUE|FALSE} ]
        [ ACTIVE_STATEMENTS=integer ]
        [ MIN_COST=float ]
        [ PRIORITY={MIN|LOW|MEDIUM|HIGH|MAX} ]
        [ MEMORY_LIMIT='memory_units' ]
```

The Resource queue creation parameters table describes the parameters for creating the resource queue.

Resource queue creation parameters

Parameter	Description
ACTIVE_STATEMENTS	The number of SQL statements that can be concurrently executed (in the active state). A value of -1 indicates that an unlimited number of SQL statements can be concurrently executed.

MEMORY_LIMIT 'memory_units kB, MB or GB'	The maximum memory usage allowed by all SQL statements in the resource queue. A value of -1 indicates unlimited memory usage. However, it is easy to trigger OOM errors because it is limited by the database or system parameters mentioned in the preceding sections.
	The memory usage of SQL statements is limited by resource queues and parameters.
	• When the gp_resqueue_memory_policy parameter is set to none, the limit is the same as that in Greenplum Database earlier than version 4.1.
	• When the gp_resqueue_memory_policy parameter is set to auto and you have specified the statement_mem parameter for a session or at the database level, the allowed memory of a single query exceeds the MEMORY_LIMIT value of the resource queue.
	Example:
	`=> SET statement_mem='2GB';`
	`=> SELECT * FROM my_big_table WHERE column='value' ORDER BY id;`
	`=> RESET statement_mem;`
	• The system parameter max_statement_mem can limit the maximum memory usage at the compute node level. The memory requested by a single query cannot exceed max_statement_mem.
	You can modify the statement_mem parameter at the session level, but do not modify the max_statement_mem parameter. We recommend that you specify max_statement_mem in the following formula:
	`(seghost_physical_memory) / (average_number_concurrent_queries)`
	• When the gp_resqueue_memory_policy parameter is set to eager_free, the database allocates the memory by stages. For example, if a query requests 1 GB of memory in total but needs only 100 MB during each stage, the database allocates 100 MB of memory to the query. You can use eager_free to reduce the possibility of insufficient memory for the query.

MAX_COST float	The maximum cost of the queries that can be concurrently executed by the resource group. The cost is the estimated total cost in the SQL execution plan. The value of the parameter can be specified as a floating-point number such as 100.0 or an exponent such as 1e+2. A value of -1 indicates that the cost is unlimited.
COST_OVERCOMMIT boolean	Specifies whether the limit of max_cost can be exceeded when the system is idle. A value of TRUE indicates that the limit can be exceeded.
MIN_COST float	When the resources requested exceed the limit, the queries are queued. However, if the cost of a query is lower than the min_cost value, the query can be executed without queuing.

Parameter	Description
PRIORITY={MIN|LOW|MEDIUM|HIGH|MAX}	The priority of the current resource queue. When resources are insufficient, CPU resources are allocated to the resource queue that has a higher priority. The SQL statements in the resource queue that has a higher priority can obtain CPU resources first. We recommend that you allocate users that initiate queries that have high real-time requirements to resource queues that have a higher priority. This parameter is similar to the time slice policy that is used in CPU resource groups in a Linux cgroup to schedule real-time and common tasks.

Example of modifying resource queue limits:

```
ALTER RESOURCE QUEUE myqueue WITH (MAX_COST=-1.0, MIN_COST= -1.0);
```

Example of putting a user in a resource queue:

```
ALTER ROLE sammy RESOURCE QUEUE poweruser;
```

The following table describes the parameters of resource queues.

Resource queue parameters

Parameter	Description
gp_resqueue_memory_policy	The memory management policy of the resource queue.
gp_resqueue_priority	Specifies whether to enable query prioritization. Valid values: • On • Off If this parameter is disabled, the existing priority settings are not evaluated.
gp_resqueue_priority_cpucores_per_segment	The number of CPU cores allocated to each compute node. For example, if an 8-core host is configured with two primary compute nodes, you can set the parameter to 4. If no other nodes exist on the primary node, set the parameter to 8. When the CPU is preempted, the SQL statements that are executed in a higher-priority resource group are allocated with CPU resources first.
gp_resqueue_priority_sweeper_interval	The interval at which CPU utilization is recalculated for all active statements. The share value is calculated when the SQL statement is executed. You can calculate the share value based on the priority and gp_resqueue_priority_cpucores_per_segment. The smaller the value and the more frequent the calculation, the better the result brought by the priority settings and the larger the overhead.

Tips for configuring resource queues:

- We recommend that you create a resource queue for each user.

  The default resource queue of is pg_default. If no queue is created, all users are assigned to pg_default. This operation is not recommended. We recommend that you create a resource queue for each user. Typically, a database user corresponds to a business. Different database users may correspond to different businesses or users, such as business users, analysts, developers, and DBAs.

- We recommend that you do not use superusers to execute queries.

  Queries initiated by superusers are limited only by the preceding parameters and not by the resource queue. We recommend that you do not use superusers to execute queries if you want to use resource queues to limit the use of resources.

- ACTIVE_STATEMENTS indicates the SQL statements that can be concurrently executed within a resource queue. When the cost of a query is lower than the minimum cost specified by min_cost, the query can be executed without being queued.
- You can specify the MEMORY_LIMIT parameter to set the allowed maximum memory usage of all the SQL statements in a resource queue. The statement_mem parameter has a higher priority that can break through the limit of resource queues.

  > **Note** The memory of all resource queues cannot exceed gp_vmem_protect_limit.

- You can distinguish businesses by configuring the priorities of resource queues.

  For example, the report-related business has a top priority, while common businesses and analysts have lower priorities. In this case, you can create three resource queues that have the max, high, and medium priorities, respectively.

- If the number of resources requested at different periods of time varies, you can run the crontab command to periodically adjust the limits of resource queues based on usage patterns.

  For example, the queue of analysts has a top priority during the day, while the queue of the report-related business has a lower priority at night. does not allow you to configure resource limits by period of time. Therefore, you can only externally deploy tasks by using the ALTER RESOURCE QUEUE statement.

- You can use the view provided by gp_toolkit to view the resource usage of the resource queues.

```
gp_toolkit.gp_resq_activity
gp_toolkit.gp_resq_activity_by_queue
gp_toolkit.gp_resq_priority_backend
gp_toolkit.gp_resq_priority_statement
gp_toolkit.gp_resq_role
gp_toolkit.gp_resqueue_status
```

14.KVStore for Redis

14.1. User Guide

14.1.1. What is KVStore for Redis?

KVStore for Redis is a database service that is compatible with open source Redis protocols. KVStore for Redis is based on a highly available hot standby architecture and can scale to meet the requirements of high-performance and low-latency read/write operations.

Features

- KVStore for Redis supports various data types, such as strings, lists, sets, sorted sets, hash tables, and streams. This service also supports advanced features, such as transactions, message subscription, and message publishing.
- KVStore for Redis Enhanced Edition (Tair), which is a key-value pair cloud caching service, is an advanced version of KVStore for Redis Community Edition.

Instance editions

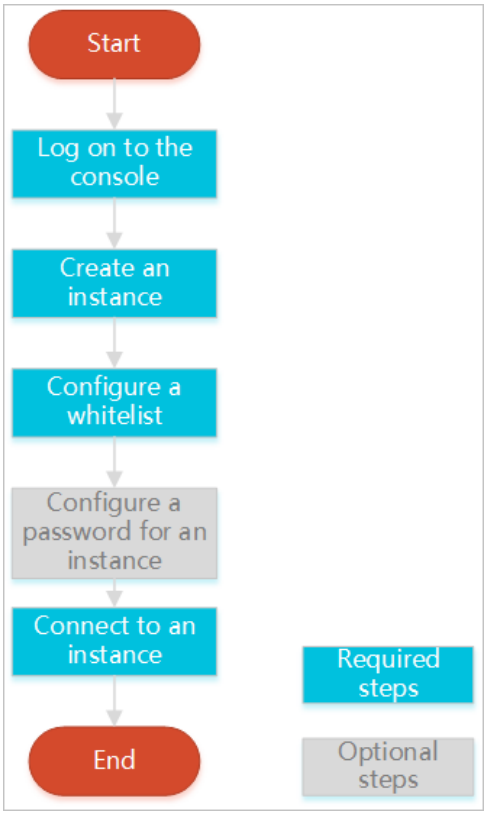
Edition	Overview
Community Edition instances	KVStore for Redis Community Edition is compatible with the data cache service of open source Redis engines. It supports master-replica instances, cluster instances, and read/write splitting instances.
Performance-enhanced instances of KVStore for Redis Enhanced Edition	KVStore for Redis Enhanced Edition provides a multi-threading model and integrates some features of Alibaba Tair. KVStore for Redis Enhanced Edition (Tair) supports multiple data structures of Tair and is suitable for diverse scenarios.

14.1.2. Quick Start

14.1.2.1. Get started with KVStore for Redis

This topic describes all operations that you can perform on an instance from instance creation to database logon. This topic helps you understand how to create and manage an instance.

The following figure shows how to manage a KVStore for Redis instance.



- Log on to the KVStore for Redis console

  This topic describes how to log on to the KVStore for Redis console.

- Create an instance

  KVStore for Redis supports classic networks and virtual private clouds (VPCs). You can create KVStore for Redis instances in these networks.

- Configure a whitelist

  Before you use a KVStore for Redis instance, add IP addresses or CIDR blocks that are used to access the database to the whitelist of the instance to improve the security and stability of the database.

- If you do not specify a password when you create the instance, specify a password on the **Instance Information** page.

- Connect to a KVStore for Redis instance

  To connect to the KVStore for Redis instance, you can use a client that supports Redis protocols or use the Redis command-line interface (redis-cli) tool.

14.1.2.2. Log on to the Apsara Uni-manager Management Console

This topic describes how to log on to the Apsara Uni-manager Management Console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

14.1.2.3. Create a KVStore for Redis instance

This topic describes how to create a KVStore for Redis instance in the KVStore for Redis console.

Procedure

1. Log on to the KVStore for Redis console.
2. Click **Create Instance** in the upper-right corner of the page.
3. Set the following parameters.

   Configure the following parameters.

Section	Parameter	Description
Basic Settings	**Organization**	Select the organization to which the KVStore for Redis instance that you want to create belongs.
	Resource Set	Select the resource set to which the instance belongs. **Notice** After you select a resource set, the instance is accessible only to the members of the specified resource set.
	Region	The region where the KVStore for Redis instance is deployed.

Region Section	Parameter	Description
	Zone	The zone where the KVStore for Redis instance is deployed.
Specifications	**Edition**	◦ **community**: This edition is compatible with the open source Redis protocol and provides high performance. ◦ **enterprise**: This edition is developed based on the community edition. The read /write performance of the community edition can be three times that of the community edition. It is integrated with multiple self-developed Redis modules to improve the applicability. For more information, see *Performance-enhanced instances of KVStore for Redis Enhanced Edition (Tair)* in *Product Introduction*.
	Chip Architecture	The chip architecture of the machine to which the KVStore for Redis instance belongs. **Note** If a message prompts that you are not authorized when you specify this parameter, contact the operations administrator to authorize your account.
	Engine Version	The version of the Redis engine. Valid values: 4.0 and 5.0.
	Architecture Type	◦ **Standard**: runs in a master-replica architecture, provides high-performance caching services, and ensures high data reliability. ◦ **Cluster**: eliminates the performance bottleneck that is caused by a single-threading model. You can use the high-performance cluster instance to process large-capacity workloads. ◦ **Read/Write Splitting**: ensures high availability and high performance, and supports multiple specifications. The read/write splitting architecture allows a large number of concurrent requests to read hot data from read replicas. This reduces the loads on the master node and minimizes the O&M cost.
	Node Type	**Master-replica** is automatically selected. The node type has one master node and one replica node.
	Instance Type	The specifications of the instance. The maximum number of connections and maximum internal bandwidth vary based on the instance specification. For more information, see *Instance specifications* in *Product Introduction*.

Section	Parameter	Description
Network	Network Type	◦ Classic network: Cloud services in a classic network are not isolated. Unauthorized access to a cloud service is blocked by security groups or the service whitelist. ◦ VPC: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize the route table, CIDR block, and gateway of a VPC. You can also migrate applications to the cloud without service interruption. You can use an Express Connect circuit or VPN to connect data centers to cloud resources in a VPC. **Note** Before you can select VPC, you must create a VPC. For more information, see *Create a VPC* and *Create a vSwitch* in *VPC User Guide*.
Password	Instance Name	Enter a name that can help you identify the instance. ◦ The name must be 2 to 128 characters in length and ◦ The name can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter.
	Password Setting	You can select **Set Now** or **Set after Purchase**.
	Logon Password	Enter a password to access the instance. The password must meet the following requirements: ◦ The password must be 8 to 30 characters in length. ◦ The password must contain uppercase letters, lowercase letters, and digits. The password must not contain special characters.
	Confirm Password	Enter the specified password again.

4. After you configure the parameters, click **Submit**.

14.1.2.4. Configure a whitelist

Before you use a KVStore for Redis instance, add IP addresses or CIDR blocks that are used to access the database to the whitelist of the instance to improve the security and stability of the database.

Context

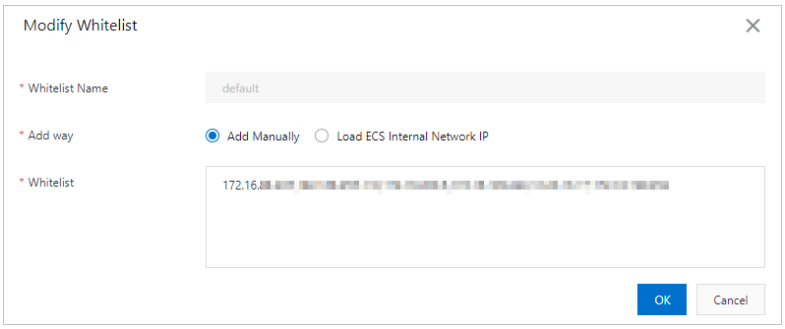
Note A properly configured whitelist can ensure a higher level of security protection for your KVStore for Redis instance. We recommend that you maintain the whitelist on a regular basis.

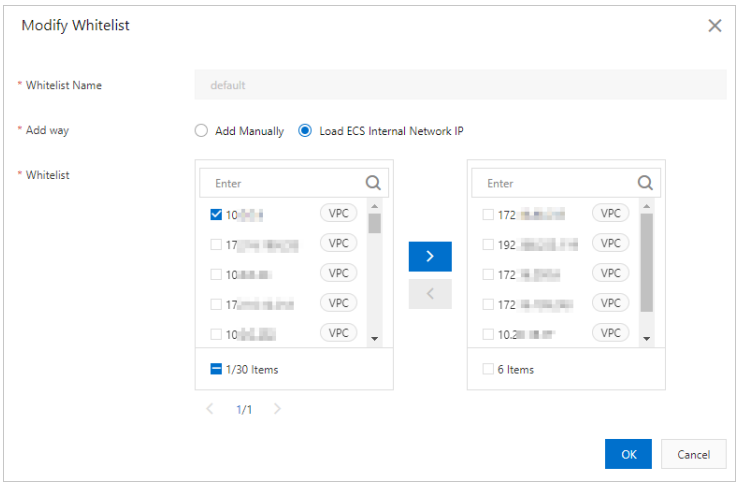
Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Whitelist Settings**.
4. Find the IP address whitelist that you want to manage and click **Modify**.

> **Note** You can also click **Add Whitelist** to create an IP address whitelist. The name of the IP address whitelist must be 2 to 32 characters in length and can contain lowercase letters, digits, and underscores (_). The name of the whitelist must start with a lowercase letter and end with a lowercase letter or digit.

5. In the dialog box that appears, perform one of the following operations:

Action	Procedure
Manually modify the IP address whitelist	Enter IP addresses or CIDR blocks. Manually modify the IP address whitelist  **Note** ◦ Separate multiple IP addresses with commas (,). You can add up to 1,000 unique IP addresses. Supported formats are specific IP addresses such as 10.23.12.24 and CIDR blocks such as 10.23.12.24/24. /24 indicates the length of the IP address prefix. An IP address prefix can be 1 to 32 bits in length. ◦ If you set the prefix length to 0, for example, 0.0.0.0/0 or 127.0.0.1/0, all IP addresses are allowed to access the instance. In this case, the security risk of your instance is high. Proceed with caution.

Action	Procedure
Add private IP addresses of ECS instances to an IP address whitelist	i. Click **Load ECS Internal Network IP**. ii. Select IP addresses based on your business requirements. Add private IP addresses of ECS instances  **Note** To find the ECS instance to which a specific IP address is assigned, you can move the pointer over the IP address. Then, the system displays the ID and name of the ECS instance to which the IP address is assigned.
Remove IP addresses from the IP address whitelist	To remove all IP addresses from the IP address whitelist and retain the IP address whitelist, click **Delete**.

6. Then, click **OK**.

14.1.2.5. Connect to an instance

14.1.2.5.1. Use a Redis client

KVStore for Redis is compatible with open source Redis. You can connect to KVStore for Redis and open source Redis in a similar manner. Therefore, you can use a client that is compatible with the Redis protocols to connect to KVStore for Redis. You can connect to a KVStore for Redis instance by using clients of different programming languages.

Prerequisites

The private IP address of an Elastic Compute Service (ECS) instance or the public IP address of an on-premises machine is added to a whitelist of the instance. For more information, see Configure a whitelist.

Obtain connection information

When you use a client to connect to a KVStore for Redis instance, you must obtain the following information and specify the information in the code:

Information	Description
Endpoint	You can find the endpoint in the **Connection Information** section on the **Instance Information** page. **Note** KVStore for Redis instances support multiple types of endpoints. We recommend that you use endpoints in a VPC for higher security and lower network latency.
Port number	The default port number is 6379.
The account of the instance. This parameter is not required by some clients.	By default, a KVStore for Redis instance provides a database account that is named after the instance ID, such as, r-bp10noxlhcoim2****.
The password of the account.	If you forget your password, you can reset the password. For more information, see Change the password.

Commonly used clients

For the list of clients supported by Redis, see Redis clients.

- Jedis client
- PhpRedis client
- Redis-py client
- C or C++ client
- .NET client
- Node-redis client
- C# client StackExchange.Redis

Jedis client

1. Download and install the Jedis client. For more information, see Jedis.
2. Select a connection method to meet your business requirements.
   - JedisPool-based connection: This connection method is recommended.
     a. Launch the Eclipse client, create a project, and then configure the following pom file:

```
<dependency>
<groupId>redis.clients</groupId>
<artifactId>jedis</artifactId>
<version>2.7.2</version>
<type>jar</type>
<scope>compile</scope>
</dependency>
```

b. Enter the following code in the project to add the relevant applications:

```
import org.apache.commons.pool2.PooledObject;
import org.apache.commons.pool2.PooledObjectFactory;
import org.apache.commons.pool2.impl.DefaultPooledObject;
import org.apache.commons.pool2.impl.GenericObjectPoolConfig;
import redis.clients.jedis.HostAndPort;
import redis.clients.jedis.Jedis;
import redis.clients.jedis.JedisPool;
import redis.clients.jedis.JedisPoolConfig;
```

c. Enter the following code in the project based on the Jedis client version and modify the code based on the comments.

> **Note** For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see Obtain connection information.

- Jedis 2.7.2

```
JedisPoolConfig config = new JedisPoolConfig();
//Maximum number of idle connections. You can configure this parameter. Make sure that th
e specified value does not exceed the maximum number of connections that the KVStore for
Redis instance supports.
config.setMaxIdle(200);
//Maximum number of connections. You can configure this parameter. Make sure that the spe
cified value does not exceed the maximum number of connections that the KVStore for Redis
instance supports.
config.setMaxTotal(300);
config.setTestOnBorrow(false);
config.setTestOnReturn(false);
String host = "*.aliyuncs.com";
String password = "Password";
JedisPool pool = new JedisPool(config, host, 6379, 3000, password);
Jedis jedis = null;
try {
jedis = pool.getResource();
/// ... do stuff here ... for example
jedis.set("foo", "bar");
String foobar = jedis.get("foo");
jedis.zadd("sose", 0, "car");
jedis.zadd("sose", 0, "bike");
Set<String> sose = jedis.zrange("sose", 0, -1);
} finally {
if (jedis != null) {
jedis.close();
}
}
/// ... when closing your application:
pool.destroy();
```

- Jedis 2.6 or Jedis 2.5

```
JedisPoolConfig config = new JedisPoolConfig();
//Maximum number of idle connections. You can configure this parameter. Make sure that th
e specified value does not exceed the maximum number of connections that the KVStore for
Redis instance supports.
config.setMaxIdle(200);
//Maximum number of connections. You can configure this parameter. Make sure that the spe
cified value does not exceed the maximum number of connections that the KVStore for Redis
instance supports.
config.setMaxTotal(300);
config.setTestOnBorrow(false);
config.setTestOnReturn(false);
String host = "*.aliyuncs.com";
String password = "Password";
JedisPool pool = new JedisPool(config, host, 6379, 3000, password);
Jedis jedis = null;
boolean broken = false;
try {
     jedis = pool.getResource();
     /// ... do stuff here ... for example
     jedis.set("foo", "bar");
     String foobar = jedis.get("foo");
     jedis.zadd("sose", 0, "car");
     jedis.zadd("sose", 0, "bike");
     Set<String> sose = jedis.zrange("sose", 0, -1);
}
catch(Exception e)
{
     broken = true;
} finally {
if (broken) {
     pool.returnBrokenResource(jedis);
} else if (jedis != null) {
     pool.returnResource(jedis);
 }
}
```

- Single Jedis connection: This connection method does not allow a client to automatically reconnect to the KVStore for Redis instance after a connection times out. Therefore, this connection is not recommended.

  Launch the Eclipse client, create a project, enter the following code, and then modify the code based on the comments.

  > **Note** For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see Obtain connection information.

```
import redis.clients.jedis.Jedis;
public class jedistest {
public static void main(String[] args) {
try {
     String host = "xx.kvstore.aliyuncs.com";//You can find the connection address in the conso
le.
     int port = 6379;
     Jedis jedis = new Jedis(host, port);
     //Authentication information.
     jedis.auth("password");//password
     String key = "redis";
     String value = "aliyun-redis";
     //Select a database. Default value: 0.
     jedis.select(1);
     //Specify a key.
     jedis.set(key, value);
     System.out.println("Set Key " + key + " Value: " + value);
     //Obtain the configured key and value.
     String getvalue = jedis.get(key);
     System.out.println("Get Key " + key + " ReturnValue: " + getvalue);
     jedis.quit();
     jedis.close();
}
catch (Exception e) {
 e.printStackTrace();
 }
}
}
```

3. Run the project. If Eclipse returns the following result, it indicates that the client is connected to the KVStore for Redis instance.

```
Set Key redis Value aliyun-redis
Get Key redis ReturnValue aliyun-redis
```

PhpRedis client

1. Download and install the PhpRedis client. For more information, see PhpRedis.
2. Enter the following code in a PHP editor and modify the code based on the comments.

> **Note** For more information about how to obtain the connection address, account, and password of the KVStore for Redis instance, see Obtain connection information.

```
<?php
 /* Replace the parameter values with the endpoint and port number of the instance. */
 $host = "r-bp10noxlhcoim2****.redis.rds.aliyuncs.com";
 $port = 6379;
 /* Replace the following parameter values with the ID and password of the instance. */
 $user = "test_username";
 $pwd = "test_password";
 $redis = new Redis();
 if ($redis->connect($host, $port) == false) {
        die($redis->getLastError());
   }
 if ($redis->auth($pwd) == false) {
        die($redis->getLastError());
  }
  /* You can manage the database after you pass the authentication. For more information, visit ht
tps://github.com/phpRedis/phpredis. */
 if ($redis->set("foo", "bar") == false) {
        die($redis->getLastError());
 }
 $value = $redis->get("foo");
 echo $value;
 ?>
```

3. Run the preceding code to connect to the instance.

   For more information, see PhpRedis.

Redis-py client

1. Download and install the redis-py client. For more information, see redis-py.
2. Enter the following code in a Python editor and modify the code based on the comments.

   > **Note** For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see Obtain connection information.

```
#!/usr/bin/env python
#-*- coding: utf-8 -*-
import redis
#Replace the value of the host parameter with the endpoint of the instance and replace the value o
f the port parameter with the port number.
host = 'localhost'
port = 6379
#Replace the following parameter value with the password of the instance.
pwd = 'test_password'
r = redis.StrictRedis(host=host, port=port, password=pwd)
#You can perform database operations after you establish a connection. For more information, visit
https://github.com/andymccurdy/redis-py.
r.set('foo', 'bar');
print r.get('foo')
```

3. Run the preceding code to connect to the instance.

C or C++ client

1. Run the following commands to download, compile, and install the C client:

```
git clone https://github.com/redis/hiredis.git
cd hiredis
make
sudo make install
```

2. Enter the following code in a C or C++ editor and modify the code based on the comments.

> **Note** For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see Obtain connection information.

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <hiredis.h>
int main(int argc, char **argv) {
unsigned int j;
redisContext *c;
redisReply *reply;
if (argc < 4) {
        printf("Usage: example xxx.kvstore.aliyuncs.com 6379 instance_id password\n");
        exit(0);
}
const char *hostname = argv[1];
const int port = atoi(argv[2]);
const char *instance_id = argv[3];
const char *password = argv[4];
struct timeval timeout = { 1, 500000 }; // 1.5 seconds
c = redisConnectWithTimeout(hostname, port, timeout);
if (c == NULL || c->err) {
if (c) {
        printf("Connection error: %s\n", c->errstr);
        redisFree(c);
} else {
        printf("Connection error: can't allocate redis context\n");
}
exit(1);
}
/* AUTH */
reply = redisCommand(c, "AUTH %s", password);
printf("AUTH: %s\n", reply->str);
freeReplyObject(reply);
/* PING server */
reply = redisCommand(c,"PING");
printf("PING: %s\n", reply->str);
freeReplyObject(reply);
/* Set a key */
reply = redisCommand(c,"SET %s %s", "foo", "hello world");
printf("SET: %s\n", reply->str);
freeReplyObject(reply);
/* Set a key using binary safe API */
reply = redisCommand(c,"SET %b %b", "bar", (size_t) 3, "hello", (size_t) 5);
printf("SET (binary API): %s\n", reply->str);
freeReplyObject(reply);
/* Try a GET and two INCR */
reply = redisCommand(c,"GET foo");
printf("GET foo: %s\n", reply->str);
freeReplyObject(reply);
reply = redisCommand(c,"INCR counter");
printf("INCR counter: %lld\n", reply->integer);
```

```
freeReplyObject(reply);
/* again ... */
reply = redisCommand(c,"INCR counter");
printf("INCR counter: %lld\n", reply->integer);
freeReplyObject(reply);
/* Create a list of numbers, from 0 to 9 */
reply = redisCommand(c,"DEL mylist");
freeReplyObject(reply);
for (j = 0; j < 10; j++) {
        char buf[64];
        snprintf(buf,64,"%d",j);
        reply = redisCommand(c,"LPUSH mylist element-%s", buf);
        freeReplyObject(reply);
    }
/* Let's check what we have inside the list */
reply = redisCommand(c,"LRANGE mylist 0 -1");
if (reply->type == REDIS_REPLY_ARRAY) {
        for (j = 0; j < reply->elements; j++) {
        printf("%u) %s\n", j, reply->element[j]->str);
}
}
freeReplyObject(reply);
/* Disconnects and frees the context */
redisFree(c);
return 0;
}
```

3. Compile the code.

```
gcc -o example -g example.c -I /usr/local/include/hiredis -lhiredis
```

4. Perform a test run and connect to the instance.

```
 example xxx.kvstore.aliyuncs.com 6379 instance_id password
```

.NET client

Warning If you need to switch or select a database from multiple databases in a cluster instance, you must set the cluster_compat_enable parameter to *0* and restart the client application. This disables the support for the cluster syntax of open source Redis. Otherwise, the system sends the following error message: `Multiple databases are not supported on this server; cannot switch to database` . For more information, see Parameter configuration.

1. Run the following command to download the .NET client.

```
 git clone https://github.com/ServiceStack/ServiceStack.Redis
```

2. Create a .NET project on the .NET client.
3. Add a reference. The reference file is stored in the library file directory ServiceStack.Redis/lib/tests.
4. Enter the following code in the .NET project and modify the code based on the comments. For more information, see ServiceStack.Redis.

Note For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see Obtain connection information.

```
using System;
 using System.Collections.Generic;
 using System.Linq;
```

```
using System.Text;
using System.Threading.Tasks;
using ServiceStack.Redis;
namespace ServiceStack.Redis.Tests
{
        class Program
{
public static void RedisClientTest()
{
        string host = "127.0.0.1";/*The endpoint of the host.*/
        string password = "password";/*The password*/
        RedisClient redisClient = new RedisClient(host, 6379, password);
        string key = "test-aliyun";
        string value = "test-aliyun-value";
        redisClient.Set(key, value);
        string listKey = "test-aliyun-list";
        System.Console.WriteLine("set key " + key + " value " + value);
        string getValue = System.Text.Encoding.Default.GetString(redisClient.Get(key));
        System.Console.WriteLine("get key " + getValue);
        System.Console.Read();
}
public static void RedisPoolClientTest()
{
        string[] testReadWriteHosts = new[] {
        "redis://password@127.0.0.1:6379"/* redis://password@endpoint:port */
};
RedisConfig.VerifyMasterConnections = false;//Required.
PooledRedisClientManager redisPoolManager = new PooledRedisClientManager(10/*Number of connection
pools*/, 10/*Connection pool timeout value*/, testReadWriteHosts);
for (int i = 0; i < 100; i++){
        IRedisClient redisClient = redisPoolManager.GetClient();//Obtain the connection.
        RedisNativeClient redisNativeClient = (RedisNativeClient)redisClient;
        redisNativeClient.Client = null;//KVStore for Redis does not support the CLIENT SETNAME c
ommand. Set Client to null.
try
{
        string key = "test-aliyun1111";
        string value = "test-aliyun-value1111";
        redisClient.Set(key, value);
        string listKey = "test-aliyun-list";
        redisClient.AddItemToList(listKey, value);
        System.Console.WriteLine("set key " + key + " value " + value);
        string getValue = redisClient.GetValue(key);
        System.Console.WriteLine("get key " + getValue);
        redisClient.Dispose();//
}catch (Exception e)
{
        System.Console.WriteLine(e.Message);
}
}
        System.Console.Read();
}
static void Main(string[] args)
{
        //Single-connection mode.
        RedisClientTest();
        //Connection-pool mode.
        RedisPoolClientTest();
}
}
```

```
}
```

Node-redis client

1. Download and install the node-redis client.

```
npm install hiredis redis
```

2. Enter the following code in the node-redis client and modify the code based on the comments.

> **Note** For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see Obtain connection information.

```
var redis = require("redis"),
client = redis.createClient(<port>, <"host">, {detect_buffers: true});
client.auth("password", redis.print)
```

Parameters:

- <port>: the service port number of the KVStore for Redis database. The default port number is 6379.
- <"host">: the endpoint of the KVStore for Redis instance.

Configuration examples:

```
var redis = require("redis"),
client = redis.createClient(6379, "r-abcdefg.redis.rds.aliyuncs.com", {detect_buffers: true});
client.auth("password", redis.print)
```

3. Run the preceding code to connect to the KVStore for Redis instance.
4. Use KVStore for Redis.

```
//Write data to the instance.
 client.set("key", "OK");
 //Query data on the instance. The returned data is of the STRING type.
 client.get("key", function (err, reply) {
 console.log(reply.toString()); // print `OK`
 });
 //If the input parameter is a buffer, the returned value is also a buffer.
 client.get(new Buffer("key"), function (err, reply) {
 console.log(reply.toString()); // print `<Buffer 4f 4b>`
 });
 client.quit();
```

C# client StackExchange.Redis

> **Warning** If you need to switch or select a database from multiple databases in a cluster instance, you must set the cluster_compat_enable parameter to *0* and restart the client application. This disables the support for the cluster syntax of open source Redis. Otherwise, the system sends the following error message: `RedisCommandException: Multiple databases are not supported on this server; cannot switch to database: 1`. For more information, see Parameter configuration.

1. Download and install the StackExchange.Redis client.
2. Add a reference.

```
using StackExchange.Redis;
```

3. Initialize ConnectionMultiplexer.

ConnectionMultiplexer is the core of StackExchange.Redis and is shared and reused in the entire application. You must use ConnectionMultiplexer as a singleton. ConnectionMultiplexer is initialized in the following method:

> **Note**
> - For more information about how to obtain the connection endpoint and password of the KVStore for Redis instance, see .
> - ConfigurationOptions contains multiple options, such as keepAlive, connectRetry, and name. For more information, see ConfigurationOptions.

```
// redis config
private static ConfigurationOptions configurationOptions = ConfigurationOptions.Parse("127.0.0.1:
6379,password=xxx,connectTimeout=2000");
 //the lock for singleton
private static readonly object Locker = new object();
 //singleton
private static ConnectionMultiplexer redisConn;
//singleton
public static ConnectionMultiplexer getRedisConn()
{
    if (redisConn == null)
    {
        lock (Locker)
        {
            if (redisConn == null || !redisConn.IsConnected)
            {
                redisConn = ConnectionMultiplexer.Connect(configurationOptions);
            }
        }
    }
    return redisConn;
}
```

4. GetDatabase() returns a lightweight object. You can obtain this object from the object of ConnectionMultiplexer.

```
redisConn = getRedisConn();
var db = redisConn.GetDatabase();
```

5. You can use the client to perform database operations.

> **Note** The following examples describe the commands for common data types. These commands are slightly different from the Redis-native commands.

   - String

```
//set get
string strKey = "hello";
string strValue = "world";
bool setResult = db.StringSet(strKey, strValue);
Console.WriteLine("set " + strKey + " " + strValue + ", result is " + setResult);
//incr
string counterKey = "counter";
long counterValue = db.StringIncrement(counterKey);
Console.WriteLine("incr " + counterKey + ", result is " + counterValue);
//expire
db.KeyExpire(strKey, new TimeSpan(0, 0, 5));
Thread.Sleep(5 * 1000);
Console.WriteLine("expire " + strKey + ", after 5 seconds, value is " + db.StringGet(strKey));
//mset mget
KeyValuePair<RedisKey, RedisValue> kv1 = new KeyValuePair<RedisKey, RedisValue>("key1", "value1
");
KeyValuePair<RedisKey, RedisValue> kv2 = new KeyValuePair<RedisKey, RedisValue>("key2", "value2
");
db.StringSet(new KeyValuePair<RedisKey, RedisValue>[] {kv1,kv2});
RedisValue[] values = db.StringGet(new RedisKey[] {kv1.Key, kv2.Key});
Console.WriteLine("mget " + kv1.Key.ToString() + " " + kv2.Key.ToString() + ", result is " + va
lues[0] + "&&" + values[1]);
```

- Hash

```
string hashKey = "myhash";
//hset
db.HashSet(hashKey,"f1","v1");
db.HashSet(hashKey,"f2", "v2");
HashEntry[] values = db.HashGetAll(hashKey);
//hgetall
Console.Write("hgetall " + hashKey + ", result is");
for (int i = 0; i < values.Length;i++)
{
  HashEntry hashEntry = values[i];
  Console.Write(" " + hashEntry.Name.ToString() + " " + hashEntry.Value.ToString());
}
Console.WriteLine();
```

- List

```
//list key
string listKey = "myList";
//rpush
db.ListRightPush(listKey, "a");
db.ListRightPush(listKey, "b");
db.ListRightPush(listKey, "c");
//lrange
RedisValue[] values = db.ListRange(listKey, 0, -1);
Console.Write("lrange " + listKey + " 0 -1, result is ");
for (int i = 0; i < values.Length; i++)
{
 Console.Write(values[i] + " ");
}
Console.WriteLine();
```

- Set

```
//set key
string setKey = "mySet";
//sadd
db.SetAdd(setKey, "a");
db.SetAdd(setKey, "b");
db.SetAdd(setKey, "c");
//sismember
bool isContains = db.SetContains(setKey, "a");
Console.WriteLine("set " + setKey + " contains a is " + isContains );
```

- Sorted Set

```
string sortedSetKey = "myZset";
//sadd
db.SortedSetAdd(sortedSetKey, "xiaoming", 85);
db.SortedSetAdd(sortedSetKey, "xiaohong", 100);
db.SortedSetAdd(sortedSetKey, "xiaofei", 62);
db.SortedSetAdd(sortedSetKey, "xiaotang", 73);
//zrevrangebyscore
RedisValue[] names = db.SortedSetRangeByRank(sortedSetKey, 0, 2, Order.Ascending);
Console.Write("zrevrangebyscore " + sortedSetKey + " 0 2, result is ");
for (int i = 0; i < names.Length; i++)
{
  Console.Write(names[i] + " ");
}
Console.WriteLine();
```

14.1.2.5.2. Use redis-cli

The redis-cli tool is a command-line interface (CLI) of Redis. You can use redis-cli to connect to a KVStore for Redis instance from an Elastic Compute Service (ECS) instance or on-premises machine and manage data.

Prerequisites

- The ECS instance and KVStore for Redis instance are deployed in the same classic network or virtual private cloud (VPC).
- The private IP address of the ECS instance is added to the whitelist of the KVStore for Redis instance. For more information, see Configure a whitelist.
- The ECS instance runs a Linux operating system and open source Redis is installed. For more information, see Redis official website.

Procedure

1. Log on to the CLI of the ECS instance and run the following command to connect to the Redis instance:

   ```
   src/redis-cli -h <hostname> -p <port>
   ```

   Parameters

Parameter	Description
<hostname>	The internal endpoint of the KVStore for Redis instance. You can find the endpoint in the **Connection Information** section on the **Instance Information** page.
<port>	The service port number of the KVStore for Redis instance. Default value: 6379.

   Example

   ```
   src/redis-cli -h r-bp1zxszhcgatnx****.redis.rds.aliyuncs.com -p 6379
   ```

2. Run the following command to verify the password:

```
AUTH <password>
```

<password>: the password that you specify when you create the instance. If you forget your password, you can reset the password. For more information, see Change the password.

Example:

```
AUTH testaccount:Rp829dlwa
```

If the password verification is successful, the following result is returned:

```
OK
```

14.1.3. Instance management

14.1.3.1. Change a password

Procedure

1.
2. On the **Instance List** page, click the ID of the instance.
3.
4. In the dialog box that appears, enter the current password and a new password.

> **Note**
> - If you forget your password, you can click **Forgot password** in the Change Password dialog box and enter a new password.
> - The password must be 8 to 32 characters in length.
> - The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include `! @ # $ % ^ & * ( ) _ + - =` .

5. Click **OK**.

14.1.3.2. Configure a whitelist

Before you use a KVStore for Redis instance, add IP addresses or CIDR blocks that are used to access the database to the whitelist of the instance to improve the security and stability of the database.

Context

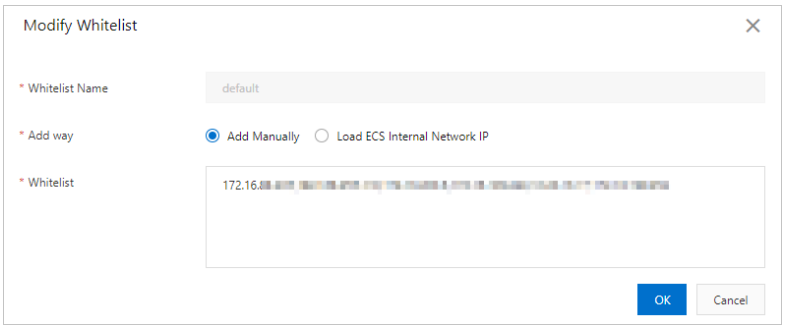
> **Note** A properly configured whitelist can ensure a higher level of security protection for your KVStore for Redis instance. We recommend that you maintain the whitelist on a regular basis.

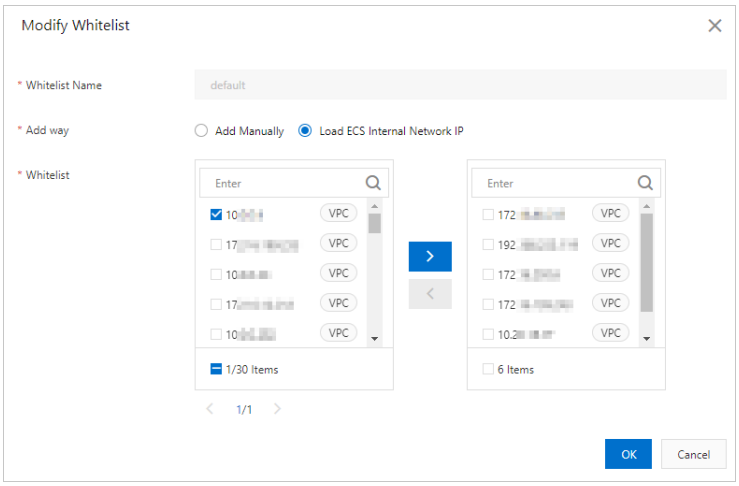
Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Whitelist Settings**.
4. Find the IP address whitelist that you want to manage and click **Modify**.

> **Note** You can also click **Add Whitelist** to create an IP address whitelist. The name of the IP address whitelist must be 2 to 32 characters in length and can contain lowercase letters, digits, and underscores (_). The name of the whitelist must start with a lowercase letter and end with a lowercase letter or digit.

5. In the dialog box that appears, perform one of the following operations:

Action	Procedure
Manually modify the IP address whitelist	Enter IP addresses or CIDR blocks. Manually modify the IP address whitelist  **Note** ○ Separate multiple IP addresses with commas (,). You can add up to 1,000 unique IP addresses. Supported formats are specific IP addresses such as 10.23.12.24 and CIDR blocks such as 10.23.12.24/24. /24 indicates the length of the IP address prefix. An IP address prefix can be 1 to 32 bits in length. ○ If you set the prefix length to 0, for example, 0.0.0.0/0 or 127.0.0.1/0, all IP addresses are allowed to access the instance. In this case, the security risk of your instance is high. Proceed with caution.

Action	Procedure
Add private IP addresses of ECS instances to an IP address whitelist	i. Click **Load ECS Internal Network IP**. ii. Select IP addresses based on your business requirements. Add private IP addresses of ECS instances  **Note** To find the ECS instance to which a specific IP address is assigned, you can move the pointer over the IP address. Then, the system displays the ID and name of the ECS instance to which the IP address is assigned.
Remove IP addresses from the IP address whitelist	To remove all IP addresses from the IP address whitelist and retain the IP address whitelist, click **Delete**.

6. Then, click **OK**.

14.1.3.3. Change configurations

Precautions

After configuration changes, the system migrates data and switches the instance. The instance is disconnected for a few seconds during this process. We recommend that you upgrade or downgrade the instance during off-peak hours.

Procedure

1.
2. On the **Instance List** page, click the ID of the instance.
3. On the page that appears, configure the required parameters.

Parameter	Description

Parameter	Description
Architecture Type	The architecture type of the instance. ○ **Standard**: runs in a master-replica architecture, provides high-performance caching services, and ensures high data reliability. ○ **Cluster**: eliminates the performance bottleneck that is caused by a single-threading model. You can use the high-performance cluster instance to process large-capacity workloads. ○ **Read/Write Splitting**: ensures high availability and high performance, and supports multiple specifications. The read/write splitting architecture allows a large number of concurrent requests to read hot data from read replicas. This reduces the loads on the master node and minimizes the O&M cost.
Instance Type	The specifications of the instance. The maximum connections and maximum internal network bandwidth vary based on the instance type.

4. Click **Submit**.

14.1.3.4. Specify a maintenance window

Context

Procedure

1.
2. On the **Instance List** page, click the ID of the instance.
3. In the **Basic Information** section, click **Settings** on the right of **Maintenance Window**.
4. Select a time period for maintenance and click **Save**.

> **Note** The time periods are in UTC+8.

14.1.3.5. Upgrade the minor version

Context

Procedure

1.
2. On the **Instance List** page, click the ID of the instance.
3.
4.

14.1.3.6. Configure SSL encryption

This topic describes how to enhance link security by enabling Secure Sockets Layer (SSL) encryption and installing SSL certification authority (CA) certificates on your application services. The SSL encryption feature encrypts network connections at the transport layer to improve data security and ensure data integrity during communication.

Prerequisites

- The major version of your ApsaraDB for Redis instance is Redis 2.8. The instance can be a standard instance or a cluster instance.
- The major version of your instance is Redis 4.0 or Redis 5.0. The instance must be a cluster instance.

Context

Procedure

1.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **SSL Settings**.
4.
5. In the dialog box that appears, turn on **Enable SSL Certificate**.

   > **Note** If this option is not supported in the current version of the instance, upgrade the minor version. For more information, see Upgrade the minor version.

6. Click **OK**.

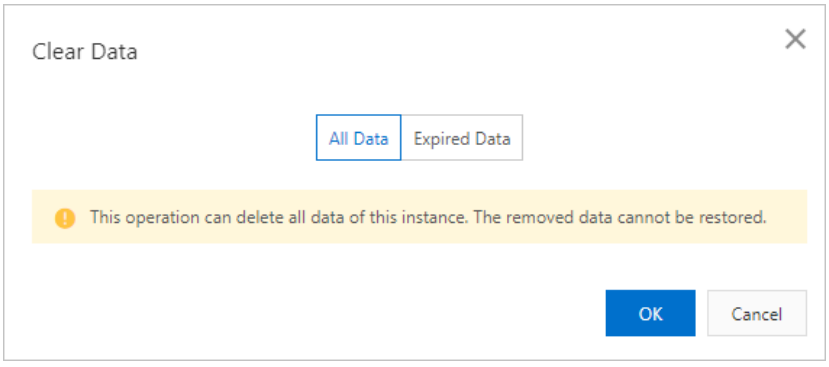
   > **Note**
   > - The result of the operation is displayed after a short period of time.
   > - In the upper-right corner of the SSL Settings page, you can also click **Update Validity** and **Download CA Certificate** to perform relevant operations.

14.1.3.7. Delete data

You can delete all data or expired data of an ApsaraDB for Redis instance in the ApsaraDB for Redis console.

Procedure

1.
2. On the **Instance List** page, click the ID of the instance.
3.
4. In the dialog box that appears, select the data that you want to delete.



   - **All Data**: runs the **FLUSHALL** command to delete all data of the instance. Deleted data cannot be recovered.
   - **Expired Data**: runs the **SCAN** command to delete all expired data of the instance. Deleted data cannot be recovered. You can select **Update Now** or Update During Maintenance. For more information, see Set a maintenance window.

> **Warning** Data deletion immediately takes effect and deleted data cannot be recovered. This may affect your business. Proceed with caution. We recommend that you back up the data of an ApsaraDB for Redis instance before you delete data. For more information, see Back up data manually.

5. In the message that appears, click **OK**.

> **Note** If you select **All Data**, you can select whether to back up the data after you click **OK**.

14.1.3.8. Release an instance

Procedure

1.
2. On the **Instances** page, click the instance ID or choose > **Release** in the **Actions** column.

> **Warning** After an instance is released, the instance cannot be restored. Proceed with caution. We recommend that you back up your data before you release an instance.

3.

14.1.3.9. Manage database accounts

KVStore for Redis allows you to create up to 20 database accounts for an instance. You can grant permissions to these accounts and manage your instance to prevent user errors.

Prerequisites

The engine version of the instance is Redis 4.0 or later.

> **Note** If the engine version of the instance is not Redis 4.0, only the default account that is created when you create the instance is available. For more information about how to change the password of the default account, see Change the password.

Context

Create an account

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. On the **Instances** page, find the instance that you want to manage and click the instance ID. In the left-side navigation pane, click **Account Management**.

> **Note** If Account Management is unavailable for an instance of Redis 4.0 or later, you must upgrade the minor version of the instance. For more information, see Upgrade the minor version.

4.
5. In the dialog box that appears, configure the required parameters and click **OK**. The following table describes the parameters.

Parameter	Description

Parameter	Description
Account	The account name must be 1 to 16 characters in length. The name can contain lowercase letters, digits, and underscores (_) and must start with a letter.
Privilege	Specify the permissions that are granted to the account. Valid values: Read-only, Read/Write, and Replicate. If you select Replicate, you can run the SYNC and PSYNC commands after you connect to an instance by using your account. **Note** You can create accounts that have the replicate permissions only for standard instances of Redis 4.0 or later.
Password Settings	Specify a password for the account. The password must be 8 to 32 characters in length. The password must contain at least three of the following types of characters: uppercase letters, lowercase letters, digits, and special characters. The following special characters are supported: `!@ # $ % ^ & * ( ) + - = _` .
Confirm Password	Enter the password again.
Description	The description of the account.

14.1.3.10. Restart an instance

Procedure

1. Log on to the KVStore for Redis console.
2. In the **Actions** column, choose > **Restart**.

Warning During the restart, the instance may be disconnected for a few seconds. We recommend that you restart instances during off-peak hours. You must also make sure that your application supports automatic reconnection.

3. In the dialog box that appears, specify the upgrade time and click **OK**.
   - Restart Immediately: restarts the instance immediately.
   - Restart Within Maintenance Window: restarts the instance within the preset maintenance window.

14.1.3.11. Export the list of instances

Procedure

1. Log on to the KVStore for Redis console.
2.
3.

14.1.3.12. Use a Lua script

Support for Lua commands

> **Note** If the EVAL command cannot be executed, for example, the "`ERR command eval not support for normal user`" message appears, you can Upgrade the minor version. During the upgrade, the instance may be disconnected and become read-only for a few seconds. We recommend that you upgrade instance versions during off-peak hours.

Limits on Lua scripts

A Lua script, which is supported by the cluster instance of KVStore for Redis, has the following limits to ensure that all operations in the script are performed within the same hash slot:

- The Lua script uses the **redis.call/redis.pcall** function to run Redis commands. . For Redis commands, the keys must be passed by using the KEYS array, which cannot be replaced by Lua variables. If the KEYS array is not used, the following error message is returned:

  ```
  -ERR bad lua script for redis cluster, all the keys that the script uses should be passed using the KEYS array\r\n
  ```

- All keys that are used by the script must be allocated to the same hash slot. If the keys are allocated to different hash slots, the following error message is returned:

  ```
  -ERR eval/evalsha command keys must be in same slot\r\n
  ```

- Keys must be included in all commands that you want to run. If the keys are not included in all commands, the following error message is returned:

  ```
  -ERR for redis cluster, eval/evalsha number of keys can't be negative or zero\r\n
  ```

- The following Pub/Sub commands are not supported: **PSUBSCRIBE**, **PUBSUB**, **PUBLISH**, **PUNSUBSCRIBE**, **SUBSCRIBE**, and **UNSUBSCRIBE**.
- The UNPACK function is not supported.

> **Note** If you do not want to impose the preceding limits but can make sure that all operations are performed in the same hash slot in the code, you can set the script_check_enable parameter to 0 in the console to disable the backend script check.

14.1.4. Connection management

14.1.4.1. View endpoints

Context

> **Note**
> - The virtual IP address of a KVStore for Redis instance may change when you maintain or modify the instance. To ensure that the connection is available, we recommend that you use an endpoint to access the KVStore for Redis instance.
> - For more information about how to apply for a public endpoint, see Applies for a public connection string.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the **Connection Information** section, you can view the private and public endpoints of the instance.

> **Note** By default, only a private endpoint is provided by a KVStore for Redis instance. If you want to connect to a KVStore for Redis instance over the Internet, you must apply for a public endpoint. For more information, see Apply for a public endpoint.

14.1.4.2. Apply for a public endpoint

This topic describes how to apply for a public endpoint for a KVStore for Redis instance.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the **Connection Information** section, click **Apply for Public Endpoint**.
4. In the dialog box that appears, enter an endpoint and a port number and click **OK**.

   > **Note**
   > - The custom endpoint must be 8 to 64 characters in length and can contain lowercase letters and digits. The endpoint must start with a lowercase letter.
   > - The custom port number ranges from 1024 to 65535. The default value is 6379.
   > - After you apply for a public endpoint, you must add the public endpoint to the IP address whitelist of the instance. This way, you can connect to the instance over the Internet. For more information, see Configure a whitelist.

5. On the **Instance Information** page, view the **Public Endpoint** in the **Connection Information** section.

   > **Note** If you no longer use the public endpoint, click **Release Public Endpoint** next to **Public Endpoint** to release the endpoint.

14.1.4.3. Modify the endpoint of an KVStore for Redis instance

Prerequisites

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3.
4. In the **Modify Public Endpoint** dialog box, set the following parameters:

Parameter	Description
Connection type	Select **Internal Endpoint** or **Public Endpoint**.
Endpoint	Set the prefix of the endpoint. ○ The endpoint can contain lowercase letters and digits. ○ It must start with a lowercase letter. ○ The endpoint must be 8 to 64 characters in length.

Parameter	Description
Port	Specify a port number. Valid value: 1024 to 65535. Note It takes about 10 minutes for the modified port number of the public endpoint to take effect. You can refresh the page to view the latest port number information.

5. In the **Modify Public Endpoint** dialog box, modify **Connection Type**, **Endpoint**, and **Port**, and then click **OK**.

Note
- The custom endpoint prefix must be 8 to 64 characters in length and can contain lowercase letters and digits. It must start with a lowercase letter.
- The custom port number ranges from 1024 to 65535. The default value is 6379.

14.1.5. Performance monitoring

14.1.5.1. Query monitoring data

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Performance Monitor**.
4. Select the start and end time and click **OK**.

Note For more information about the metrics, see Understand metrics.

14.1.5.2. Select metrics

Context

KVStore for Redis supports more than 10 monitoring groups. By default, the Performance Monitor page displays the metrics of the basic monitoring group. You can click **Customize Metrics** to switch to the metrics of other monitoring groups. The following table describes the monitoring groups.

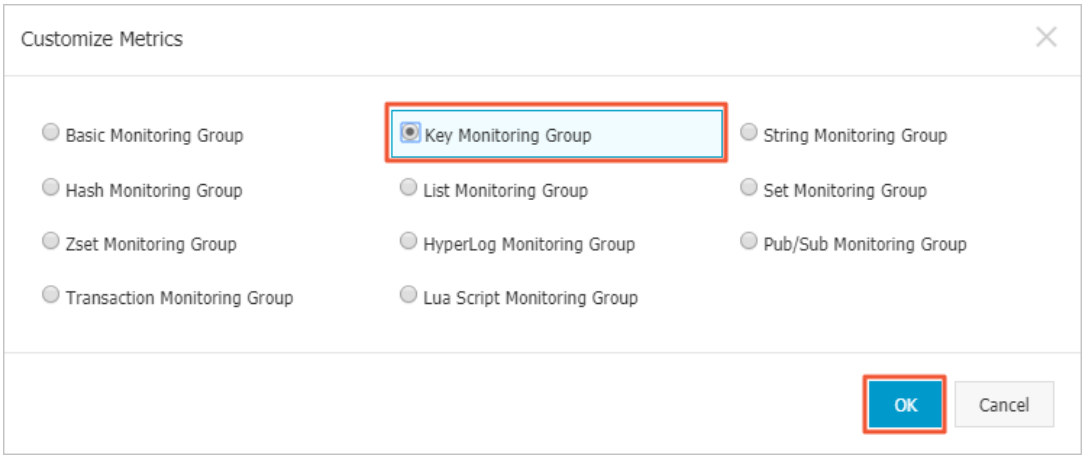
Monitoring group	Description
Basic monitoring group	The basic monitoring information about an instance, such as the queries per second (QPS), bandwidth, and memory usage of the instance.
Key monitoring group	The monitoring information about the use of key-value related commands, such as the number of times that the DEL and EXITS commands are executed.
String monitoring group	The monitoring information about the use of string commands, such as the number of times that the APPEND and MGET commands are executed.
Hash monitoring group	The monitoring information about the use of hash commands, such as the number of times that the HGET and HDEL commands are executed.

Monitoring group	Description
List monitoring group	The monitoring information about the use of list commands, such as the number of times that the BLPOP and BRPOP commands are executed.
Set monitoring group	The monitoring information about the use of set commands, such as the number of times that the SADD and SCARD commands are executed.
Zset monitoring group	The monitoring information about the use of zset commands, such as the number of times that the ZADD and ZCARD commands are executed.
HyperLog monitoring group	The monitoring information about the use of HyperLogLog commands, such as the number of times that the PFADD and PFCOUNT commands are executed.
Pub/Sub monitoring group	The monitoring information about the use of publication and subscription commands, such as the number of times that the PUBLISH and SUBSCRIBE commands are executed.
Transaction monitoring group	The monitoring information about the use of transaction commands, such as the number of times that the WATCH, MULTI, and EXEC commands are executed.
Lua script monitoring group	The monitoring information about the use of Lua script commands, such as the number of times that the EVAL and SCRIPT commands are executed.

For more information about the definitions of the metrics in each monitoring group, see Understand metrics.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Performance Monitor**.
4.
5. In the dialog box that appears, specify a monitoring group and click **OK**.



On the Historical Monitoring Data page, the metrics in the selected monitoring group appear.

14.1.5.3. Modify the data collection interval

Context

Procedure

1. Log on to the KVStore for Redis console.

2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Performance Monitor**.
4.
5.

14.1.5.4. Understand metrics

KVStore for Redis updates more than 10 monitoring groups of metrics in real time. This allows you to monitor the status of KVStore for Redis instances. This topic describes the metrics of the monitoring groups.

Metrics of basic monitoring groups

Metric	Unit	Description	Statistical method
CpuUsage	%	The CPU utilization	Check the CPU utilization when the monitoring data is collected.
UsedMemory	Bytes	The amount of the used memory.	Check the memory usage when the monitoring data is collected.
TotalQps	Counts/s	The number of requests that are received by the instance per second.	Divide the number of requests in a monitoring cycle by the number of seconds in the monitoring cycle.
ConnCount	Counts	The number of connections.	Check the number of connections when collecting monitoring data.
InFlow	KBps	The amount of data received by the instance per second.	Divide the amount of data that is received in a monitoring cycle by the number of seconds in the monitoring cycle.
OutFlow	KBps	The amount of data sent by the instance per second.	Divide the amount of data that is sent in a monitoring cycle by the number of seconds in the monitoring cycle.
FailedCount	Counts/s	The average number of abnormal requests per second.	Divide the total number of abnormal requests in a monitoring cycle by the number of seconds in the monitoring cycle.
AvgRt	us	The average response time of all requests. **Note** For more information, see Response time (RT) metrics.	Divide the processing time of all requests in a monitoring cycle by the number of requests in the monitoring cycle.
MaxRt	us	The maximum response time of requests. **Note** For more information, see Response time (RT) metrics.	The maximum time that is required to process a request in a monitoring cycle.

Metric	Unit	Description	Statistical method
Keys	Counts	The total number of keys.	The number of keys when the monitoring data is collected.
Expires	Counts	The total number of keys for which an expiration time is configured.	The total number of keys for which an expiration time is set when the monitoring data is collected.
ExpiredKeys	Counts	The total number of expired keys.	The cumulative sum when the monitoring data is collected. After the instance is restarted, the cumulative sum is calculated again.
EvictedKeys	Counts	The total number of keys that are evicted because the memory is exhausted.	The cumulative sum when the monitoring data is collected. After the instance is restarted, the cumulative sum is calculated again.
request	Bytes	The total amount of request data received by KVStore for Redis nodes in a monitoring cycle.	See the description of this metric.
response	Bytes	The total amount of response data sent by KVStore for Redis nodes in a monitoring cycle.	See the description of this metric.
request_max	Bytes	The maximum amount of data in a request in a monitoring cycle.	See the description of this metric.
response_max	Bytes	The maximum amount of data in a response in a monitoring cycle.	See the description of this metric.
traffic_control_input	Counts	The number of times that downstream throttling is triggered.	Monitor the cumulative sum in a monitoring cycle.
traffic_control_output	Counts	The number of times that uplink throttling is triggered.	Monitor the cumulative sum in a monitoring cycle.
traffic_control_input_status	Counts	Indicates whether downstream throttling was triggered in a monitoring cycle. A value of 0 indicates that throttling was not triggered. A value of 1 indicates that throttling was triggered.	See the description of this metric.
traffic_control_output_status	Counts	Indicates whether upstream throttling was triggered in a monitoring cycle. A value of 0 indicates that throttling was not triggered. A value of 1 indicates that throttling was triggered.	See the description of this metric.
hit_rate	%	The request hit rate, which is the probability that data exists in a KVStore for Redis instance for a data access request.	Calculate the percentage of the hit requests to the total number of requests in a monitoring cycle.
hit	Counts	The number of hit requests.	Check the number of hit requests in a monitoring cycle.

Metric	Unit	Description	Statistical method
miss	Counts	The number of missed requests.	Check the number of missed requests in a monitoring cycle.
evicted_keys_per_sec	Counts/s	The number of keys that are evicted per second.	Divide the total number of keys that are evicted in a monitoring cycle by the number of seconds in the monitoring cycle.

Metrics in other monitoring groups

The system also uses other metrics to monitor specific types of data or specific features. The metrics are classified into:

- Metrics that indicate the number of times that commands are used. For example, the del, dump, and exists metrics that are used to monitor keys indicate the number of times that the DEL, DUMP, and EXISTS commands are executed.
- Response time (RT) metrics of commands. For example, the metrics that end with avg_rt, such as del_avg_rt dump_avg_rt, and exists_avg_rt, in the key monitoring group are used to monitor the average response time of the DEL, DUMP, and EXISTS commands in a monitoring cycle.

Response time (RT) metrics

All monitoring groups have RT metrics. RT metrics end with Rt or rt. For example, the AvgRt and MaxRt metrics are used in the basic monitoring group and the del_avg_rt and exists_avg_rt metrics are used to monitor keys.

The AvgRt and MaxRt metrics in the basic monitoring group are the most frequently used RT metrics. These metrics have different meanings for proxy nodes and data nodes.

- For a cluster instance or a read/write splitting instance, the AvgRt metric of a proxy node indicates the average time consumed by the proxy node to process all commands. The following process shows how a proxy node processes a command:
  i. The proxy node receives a command and forwards the command to a data node.
  ii. The data node processes the command and responds to the proxy node.
  iii. The proxy node returns the command processing result.

  The AvgRt metric of the proxy node includes the amount of time consumed by the data node to process a command and the time that is required to wait for the command to be processed. This metric also includes the amount of time consumed for network communication between the proxy node and the data node.
- For data nodes of a cluster instance or a read/write splitting instance or for a standard instance, the AvgRt metric indicates the average time consumed by a data node to process all commands. This metric records the period of time from the time when the data node receives the command to the time when the data node returns the result. This metric does not include the time consumed by the proxy node to process a command and the time that is required for network communication.
- The MaxRt metric indicates the maximum response time of requests. The statistical method of this metric is similar to the statistical method of the AvgRt metric for all KVStore for Redis instances.

14.1.6. Parameter configuration

Context

Configure parameters in the KVStore for Redis console

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3.

4.
5.

14.1.7. Backup and recovery

14.1.7.1. Automatically back up data

An increasing number of applications use Redis for persistent storage. In this case, an automatic backup mechanism is required to back up data on a regular basis so that you can restore data if user errors occur. KVStore for Redis uses Redis database backup (RDB) snapshots to back up data on replica nodes. The backup process does not have negative impacts on the performance of your instance. You can configure a custom backup policy in the console.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Backup and Recovery**.
4. Click the **Backup Settings** tab.
5. Click **Edit** and specify Backup Cycle and Backup Time.
   - **Retention Days**: The number of days for which backups are retained. This parameter is set to seven days and cannot be changed.
   - **Backup Cycle**: You can select one or more days in a week. By default, one backup is created per day.
   - **Backup Time**: You can specify a period of time in hours within a day. We recommend that you back up data during off-peak hours.
6.

14.1.7.2. Back up an instance

You can initiate a manual backup task in the console at any time.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Backup and Recovery**.
4. In the upper-right corner of the page, click **Create Backup**.
5. Click **OK**.

> **Note** On the **Data Backup** tab, you can select a time range to query existing backups. Backups are retained for seven days.

14.1.7.3. Download backup files

To archive backup files for a long period, you can copy the URLs in the console and download the database backup files to an on-premises machine.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Backup and Recovery**.

4. Find the backup file that you want to download and click **Download** in the **Actions** column.

> **Note** For a cluster instance, you must download the backup file for each data shard at the same point in time to ensure data consistency.

14.1.7.4. Restore data

KVStore for Redis allows you to restore data from a specified backup set to the current KVStore for Redis instance.

Prerequisites

The instance must be a master-replica or cluster instance.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Backup and Recovery**.
4. Perform one of the following operations based on the architecture of your KVStore for Redis instance:
   - Master-replica instance: Find the backup set that you want to restore and click **Restore Data** in the **Actions** column.
   - Cluster instance: Select the backup sets of all data shards that were generated at the same point in time and click **Restore Data** in the upper-right corner.

> **Warning** Risks may occur when you restore data. Proceed with caution. Verify the data that you want to restore before you restore the data.

5. In the message that appears, read the content and click **Continue**.

   You can also restore backup data by cloning an instance. For more information, see Clone an instance.

14.1.7.5. Clone an instance

KVStore for Redis allows you to create an instance from a specified backup set. The data in the new instance is the same as the data in the backup set. This feature can be applied in scenarios such as data recovery, quick workload deployment, and data verification.

Prerequisites

The instance must be a master-replica or cluster instance.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, click **Backup and Recovery**.
4. Find the backup set that you want to restore and click **Clone Instance** in the **Actions** column.

> **Note** For a cluster instance, you must select the backup file for each data shard at the same point in time.

5. In the message that appears, click **OK**.
6. On the Restore Instance page, configure the parameters and click **Submit**.

> **Note** For more information about the configurations of the new instance, see Create an instance.

14.1.8. CloudDBA

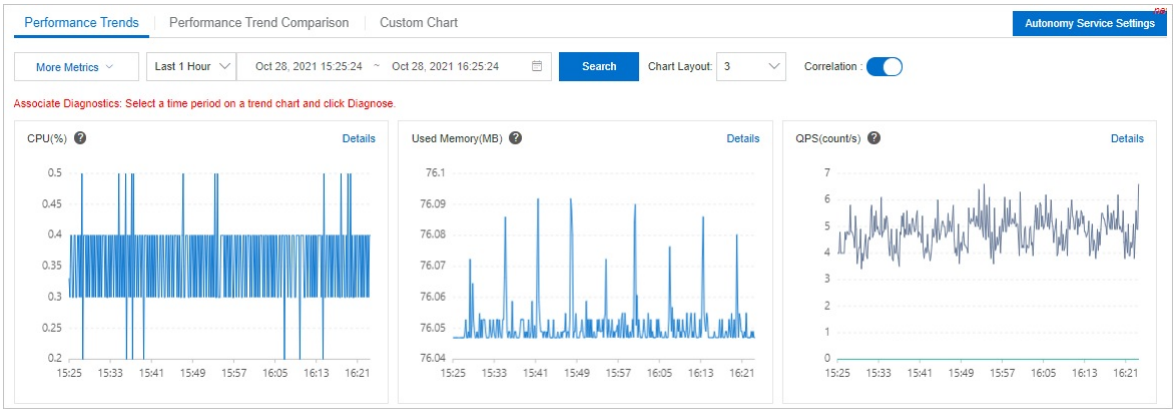
14.1.8.1. Performance trends

CloudDBA provides the performance trends feature that allows you to monitor the basic performance of a KVStore for Redis instance and the operational trends within a specified period of time. The performance trends include the CPU utilization, memory usage, queries per second (QPS), total connections, response time, data transfer, and key hit ratio.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, choose **CloudDBA > Performance Trends**.
4. You can use the following methods to view performance trends:

> **Note** If the KVStore for Redis instance uses a cluster architecture, the Performance Trends page displays the information about the nodes. The performance data during the last 1 hour is displayed. If you click the node ID, you can view the details of a specified node.



- **Performance trends**

  On the **Performance Trends** tab, specify a time range and click **Search**.

  > **Note**
  > - By default, **Correlation** is enabled. If you move the pointer over the CPU chart to view the CPU metric of the KVStore for Redis instance at 09:00, other charts also display other metrics of the instance at 09:00.
  > - To view the definition of the performance metric and the performance trend, click **Definition?** and **Details** in the upper-right corner of the chart.

- **Performance trend comparison**

  To compare the performance trends within two periods of time, click the **Performance Trend Comparison** tab, specify two periods of time, select more metrics, and then click **Search**.

- **Custom chart**

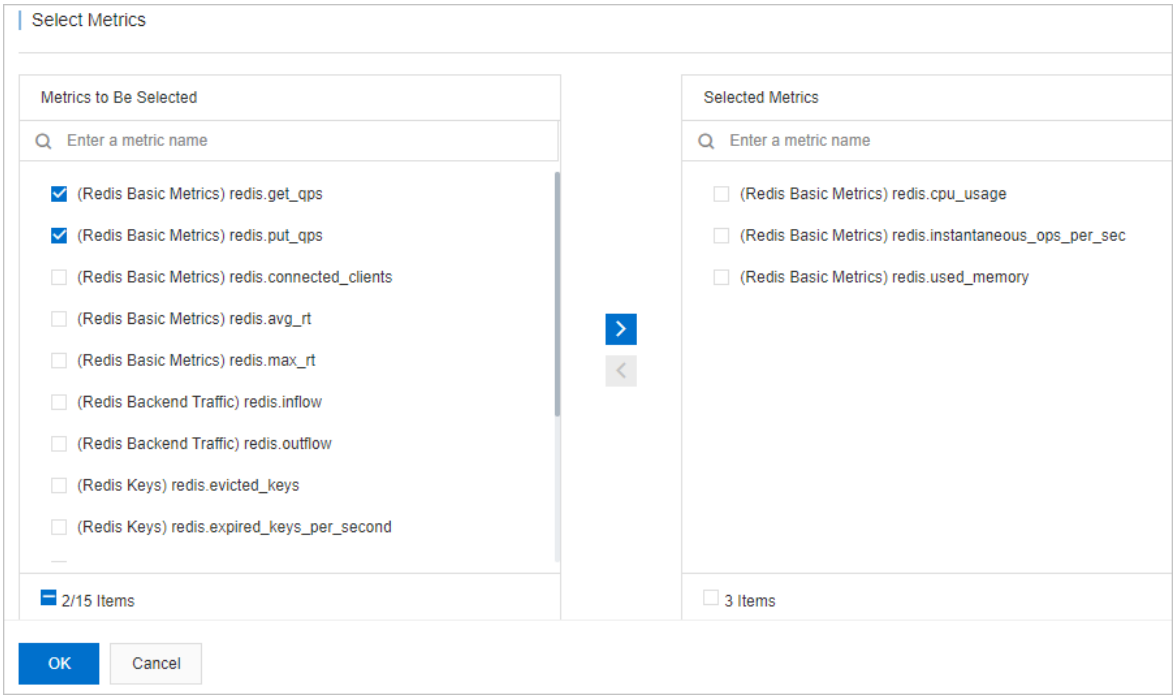
The preceding two methods display the basic metrics of a KVStore for Redis instance. If you want to display only basic metrics, you can configure custom performance trend charts. For more information, see Add a performance trend chart.

14.1.8.2. Add a performance trend chart

The default performance trends tab displays the basic performance metrics of a KVStore for Redis instance. You can add a chart that contains only specified performance metrics to analyze the performance trends of your instances. This topic describes how to add a performance trend chart to a dashboard for KVStore for Redis instances.

Procedure

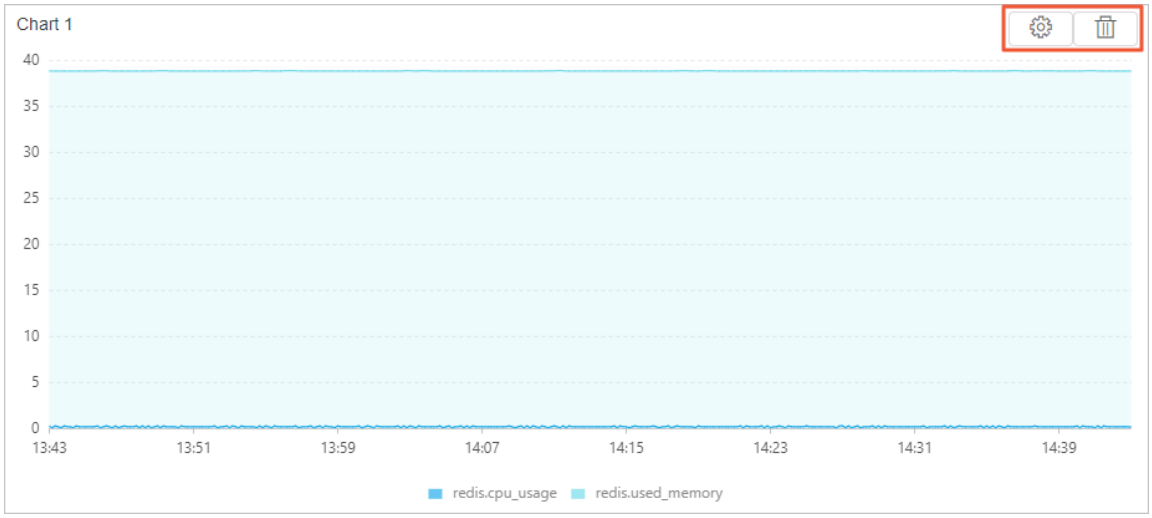
1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, choose **CloudDBA > Performance Trends**.
4. Click the **Custom Chart** tab.
5. Click **Add Monitoring Dashboard**. In the Create Monitoring Dashboard dialog box, enter a dashboard name and click **OK**.
6. Click **+ Add Chart** or **Add Monitoring Chart**. Select the metrics that you want to add and click **OK**.



7. (Optional)You can view, modify, and delete monitoring dashboards.
   - View a monitoring dashboard

     Select the monitoring dashboard, specify a time range, and then click **Search**.
   - Modify a monitoring dashboard

     Click the following icons to modify or delete a chart in the monitoring dashboard.



- Delete a monitoring dashboard

  Choose **Operate Dashboard > Delete Monitoring Dashboard**.

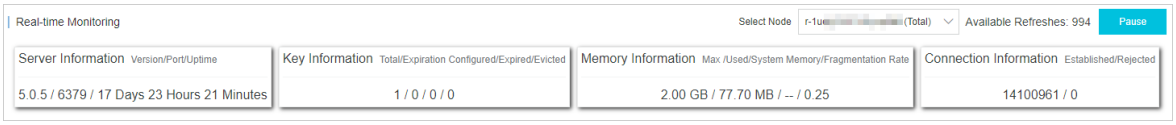
14.1.8.3. View performance metrics in real time

CloudDBA allows you to view the performance metrics of KVStore for Redis instances in real time. The performance metrics include information about CPU utilization, memory usage, queries per second (QPS), network traffic, servers, keys, clients, and connections.

Procedure

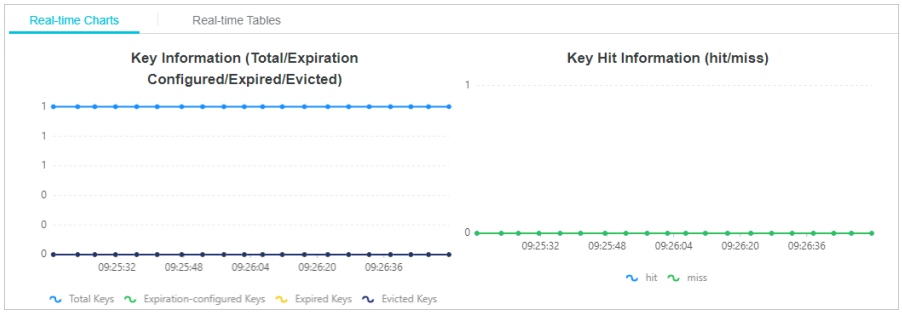
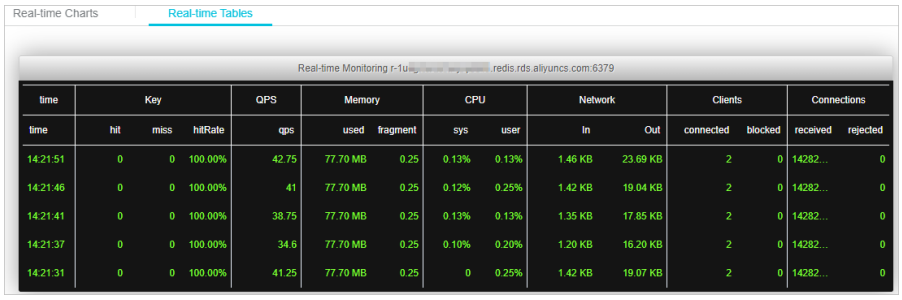
1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, choose **CloudDBA > Real-time Performance**.
4. Select a view based on your business requirements.

   In the upper part of the page, performance metrics are displayed in real time. The metrics include information about the server, keys, memory, clients, and connections. The details about the performance metrics are displayed in **Real-time Charts** and **Real-time Tables**.



> **Note** The metrics are automatically updated every 5 minutes. Therefore, you can view real-time changes in performance. The remaining updating times are displayed in the upper-right corner. To stop updating the performance metrics, you can click **Pause**.

View	Description

View	Description
Real-time Charts	Displays the real-time performance metrics of an instance in curve charts, such as key information, key hit information, key hit ratio, CPU utilization, memory information, QPS, and network traffic. 
Real-time Tables	Displays the information about keys, QPS, memory usage, CPU utilization, network traffic, clients, and connections. The table displays up to 999 records. A new record is added to the table every 5 seconds. 

14.1.8.4. Instance sessions

Instance sessions allow you to view the information about sessions between a KVStore for Redis instance and a client in real time, which includes the client information, commands that are run, and connection duration. You can also terminate abnormal sessions based on your business requirements.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, choose **CloudDBA > Instance Sessions**.
4. Perform the operations that are described in the following table based on your business requirements.
   - View sessions: By default, the details of all sessions are displayed. You can move the pointer over a specific parameter name to view information.

     > **Note**
     > - You can enter keywords in the **search box** to filter session information.
     > - To refresh instance session information, click **Refresh** in the upper-left corner or enable **Auto Refresh** to automatically refresh the page every 30 seconds.

   - Terminate sessions: Press the Shift key and select the specified session. To terminate the selected session, click **Kill Selected** in the upper-right corner. To terminate all sessions, click **Kill All**.

> **Warning** To avoid unexpected results, we recommend that you do not terminate system-level sessions.

- View session statistics: Session statistics record the total number of clients, active clients, and source IP addresses of instance sessions.

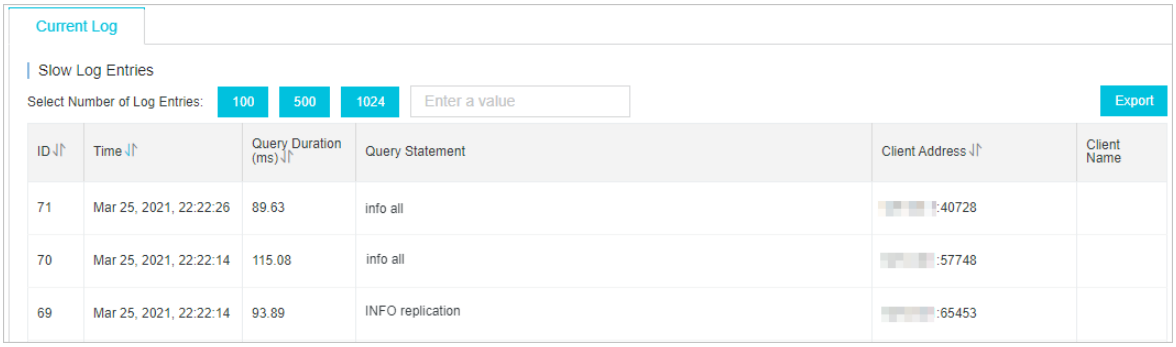
> **Note** In the **Statistics by Source** table, click the icon on the right of the source IP address to modify the source alias. In the Total Sessions column, click a value to view the details about a source IP address.

14.1.8.5. Slow queries

Slow queries reduce the stability of KVStore for Redis instances. To monitor and analyze slow queries, you can view the details about slow query logs in CloudDBA.

Procedure

1. Log on to the KVStore for Redis console.
2. On the **Instance List** page, click the ID of the instance.
3. In the left-side navigation pane, choose **CloudDBA > Slow Queries**.
4. Query the details about the slow query logs.



> **Note** You can select the number of log entries to be displayed or enter keywords in the search box to filter log entries.

15.ApsaraDB for MongoDB

15.1. User Guide

15.1.1. Usage notes

You must get familiar with the precautions and limits of ApsaraDB for MongoDB before you start.

To ensure the stability and security of ApsaraDB for MongoDB instances, take note of the limits, see Instance types in *ApsaraDB for MongoDB ApsaraDB for MongoDB limits*.

ApsaraDB for MongoDB limits

Item	Limit
Scale out nodes	You cannot scale out secondary nodes.
	• When a replica set instance is created, three nodes are added. • ApsaraDB for MongoDB provides a primary node, a secondary node, and a hidden node for each replica set instance. The hidden node is invisible to you. • You cannot scale out secondary nodes.
Restart an instance	You must restart an ApsaraDB for MongoDB instance in the ApsaraDB for MongoDB console or by calling the API operation.

15.1.2. Log on to the ApsaraDB for MongoDB console

This topic describes how to log on to the ApsaraDB for MongoDB console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

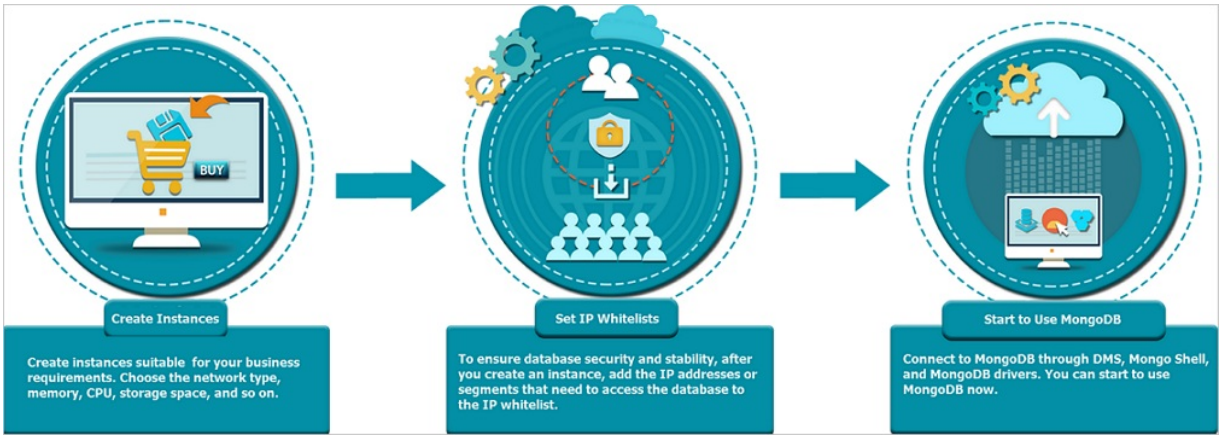
     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products** > **Database Services** > **ApsaraDB for MongoDB**.

15.1.3. Quick start

15.1.3.1. Use ApsaraDB for MongoDB

This topic is a quick start guide to basic usage operations for ApsaraDB for MongoDB, such as creating an instance, configuring a whitelist, and connecting to an instance. Flowcharts are used to describe the basic procedures in ApsaraDB for MongoDB, and guide you to create an ApsaraDB for MongoDB instance.



- Create an ApsaraDB for MongoDB instance

  An instance is a virtual database server on which you can create and manage multiple databases.
- Configure a whitelist for an ApsaraDB for MongoDB instance

  After you create an ApsaraDB for MongoDB instance, you need to configure a whitelist for the instance to allow external devices to access the instance.

  A whitelist can enhance access security for ApsaraDB for MongoDB instances. We recommend that you update the whitelist on a regular basis. The normal services of the instance are not affected if you configure a whitelist.
- Connect to a replica set instance by using the mongo shell

  After you create an instance and configure a whitelist, you can use the mongo shell to connect to the instance.

15.1.3.2. Create an ApsaraDB for MongoDB instance

This topic describes how to create an instance in the ApsaraDB for MongoDB console.

Create a replica set instance

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click **Create Instance** in the upper-left corner.
4. On the **Create ApsaraDB for MongoDB Instance** page, configure the parameters described in the following table.

Section	Parameter	Description
Basic Settings	**Organization**	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	**Region**	The region where the instance is deployed.
	Zone	The zone where the instance is deployed. **Note** If you select dual zones, the instance supports zone-disaster recovery across two data centers.
	Chip Architecture	The chip architecture of the host where the instance is deployed. **Note** If you do not have permissions to select an option, contact the operations administrator to grant such permissions to your account.
	Database Engine	The value is fixed at **MongoDB**.
	Engine Version	The database engine version of the instance. Valid values: ○ 3.0 ○ 3.4 ○ 4.0 ○ 4.2

Section	Parameter	Description
Specifications	**Node Type**	The following node types are available in ApsaraDB for MongoDB: ○ Three-node Replica Set: uses dedicated memory and I/O resources but shares CPU and storage resources with other general-purpose instances on the same server. ○ Dedicated Instance: uses dedicated CPU, memory, storage, and I/O resources to ensure long-term stable performance. In this case, an instance is not affected by other instances on the same server. ○ Dedicated Host: exclusively uses all resources of a server. This is the top configuration of exclusive specifications.
	Node Specifications	The node specifications of the instance. For more information, see descriptions in the ApsaraDB for MongoDB console.
	Storage Capacity (GB)	The storage capacity of the instance, which includes the storage capacity for data, system files, log files, and transaction files. For more information, see *Instance types* in *ApsaraDB for MongoDB Product Introduction*.
Network	**Network Type**	The following network types are available in ApsaraDB for MongoDB: ○ **Classic Network**: Cloud services on the classic network are not isolated. Unauthorized access can be blocked only by security groups or whitelists. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for enhanced security. **Note** If you select the **VPC** network type, you must configure the **VPC** and **vSwitch** parameters.

Section	Parameter	Description
	VPC	The VPC in which the instance resides. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch of the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
Password Settings	Instance Name	The name of the instance. The name must meet the following requirements: ○ The name must start with a letter. ○ The name can contain digits, letters, underscores (_), and hyphens (-). ○ The name must be 2 to 256 characters in length.
	Password Setting	Specifies when to set a password. The following options are available: ○ **Set Now**: immediately sets the logon password. ○ **Set after Purchase**: sets the logon password after you create the instance. For more information, see Reset the password for an ApsaraDB for MongoDB instance.
	Logon Password	The password used to log on to the database. The password must meet the following requirements: ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include !#$%^&*()_+= ○ The password must be 8 to 32 characters in length.

Section	Parameter	Description
	Confirm Password	Enter the password again. The password you enter here must be the same as that you entered in the Logon Password field.

5. Click **Submit**.

Create a sharded cluster instance

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click **Create Instance**.
4. On the **Create ApsaraDB for MongoDB Sharded Cluster Instance** page, configure the parameters described in the following table.

Section	Parameter	Description
Basic Settings	**Organization**	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	**Region**	The region where the instance is deployed.
	Zone	The zone where the instance is deployed. **Note** If you select dual zones, the instance supports zone-disaster recovery across two data centers.
Specifications	**Chip Architecture**	The chip architecture of the host where the instance is deployed. **Note** If you do not have permissions to select an option, contact the operations administrator to grant such permissions to your account.
	Database Engine	The value is fixed at **MongoDB**.
	Engine Version	The database engine version of the instance. Valid values: ○ 3.4 ○ 4.0 ○ 4.2

Section	Parameter	Description
Network	Network Type	The following network types are available in ApsaraDB for MongoDB: ○ **Classic Network**: Cloud services on the classic network are not isolated. Unauthorized access can be blocked only by security groups or whitelists. ○ **VPC**: A VPC helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for enhanced security. **Note** If you select the **VPC** network type, you must configure the **VPC** and **vSwitch** parameters.
	VPC	The VPC in which the instance resides. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch of the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
Mongos Specifications	Mongos Specifications	The specifications of the mongos node. For more information, see descriptions in the ApsaraDB for MongoDB console.
	Quantity	The number of mongos nodes. You can select 2 to 32 mongos nodes.
Shard Specifications	Shard Specifications	The specifications of the shard node. For more information, see descriptions in the ApsaraDB for MongoDB console.
	Storage Capacity (GB)	The storage capacity of the shard node, which includes the storage capacity for data, system files, log files, and transaction files. For more information, see *Instance types* in *ApsaraDB for MongoDB Product Introduction*.

Section	Parameter	Description
	Quantity	The number of shard nodes. You can select 2 to 32 shard nodes.
Config Server Specifications	**Config Server Type**	The specifications of the Configserver node. The value is fixed at **1 core, 2 GB**.
	Disk (GB)	The storage capacity of the Configserver node. The value is fixed at 20 GB.
Password Settings	**Instance Name**	The name of the instance. The name must meet the following requirements: ○ The name must start with a letter. ○ The name can contain digits, letters, underscores (_), and hyphens (-). ○ The name must be 2 to 256 characters in length.
	Password Setting	Specifies when to set a password. The following options are available: ○ **Set Now**: immediately sets the logon password. ○ **Set after Purchase**: sets the logon password after you create the instance. For more information, see Reset the password for an ApsaraDB for MongoDB instance.
	Logon Password	The password used to log on to the database. The password must meet the following requirements: ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include !#$%^&*()_+= ○ The password must be 8 to 32 characters in length.
	Confirm Password	Enter the password again. The password you enter here must be the same as that you entered in the Logon Password field.

5. Click **Submit**.

15.1.3.3. Reset the password for an ApsaraDB for MongoDB instance

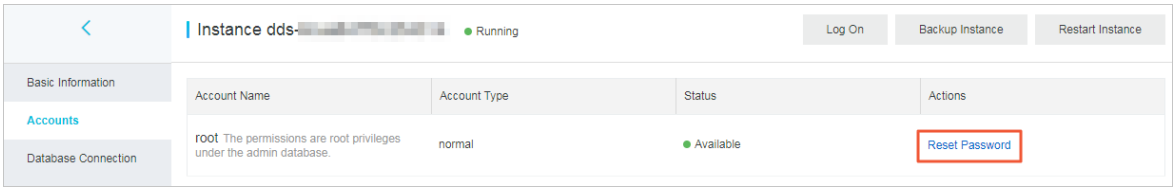
This topic describes how to reset your password in the ApsaraDB for MongoDB console.

Context

Notice We recommend that you change your password on a regular basis to ensure data security.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. In the left-side navigation pane, click **Accounts**.
5. Click **Reset Password** in the Actions column and configure the parameters in the Reset Password panel.



Parameters for resetting a password describes the parameters.

Parameters for resetting a password

Parameter	Description
New Password	Specify the new password of the account based on the following rules: ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include !#$%^&*()_+= ○ The password must be 8 to 32 characters in length.
Confirm New Password	Enter the password again. The password you enter here must be the same as that in New Password.

6. Click **OK**.

15.1.3.4. Configure a whitelist for an ApsaraDB for MongoDB instance

This topic describes how to configure a whitelist for an ApsaraDB for MongoDB instance. Before you use an ApsaraDB for MongoDB instance, you must add the IP addresses or CIDR blocks that you use for database access to a whitelist of this instance. This improves database security and stability. Proper configuration of whitelists can enhance access security of ApsaraDB for MongoDB. We recommend that you maintain the whitelists on a regular basis.

Context

- The system creates a default whitelist for each instance. This whitelist can be modified or cleared but cannot be deleted.
- After an ApsaraDB for MongoDB instance is created, the system automatically adds the IP address 127.0.0.1 to the **default** whitelist of this instance. The IP address 127.0.0.1 indicates that no IP addresses are allowed to access this instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > Whitelist Settings**.
5. On the page that appears, use one of the following methods to add IP addresses to a whitelist:
   - Manually modify a whitelist
     a. Click ⋮ in the **Actions** column corresponding to a whitelist and select **Manually Modify**.
     b. In the Manually Modify panel, enter IP addresses or CIDR blocks in the **IP White List** field.

     > **Note**
     > - Separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.
     >
     >   Supported formats are 0.0.0.0/0, IP addresses such as 10.23.12.24, or CIDR blocks such as 10.23.12.24/24. /24 indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits.
     > - If the IP address whitelist is empty or contains only `0.0.0.0/0`, all devices are granted access. This poses risks to your ApsaraDB for MongoDB instance. We recommend that you add only the IP addresses or CIDR blocks of your own web servers to the whitelist.

     c. Click **OK**.
   - Load internal IP addresses of ECS instances
     a. Click ⋮ in the **Actions** column corresponding to a whitelist and select **Import ECS Intranet IP**.
     b. In the **Import ECS Intranet IP** panel, select the IP addresses that you want to add to the IP address whitelist and click > to add these IP addresses to the whitelist.
     c. Click **OK**.

15.1.3.5. Connect to an instance

15.1.3.5.1. Use DMS to log on to an ApsaraDB for MongoDB instance

You can use Data Management (DMS) to log on to an ApsaraDB for MongoDB instance.

Prerequisites

An IP address whitelist is configured. For more information about how to configure an IP address whitelist, see Configure a whitelist for an ApsaraDB for MongoDB instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click in the **Actions** column corresponding to the instance and select **Manage**.
4. In the upper-right corner of the page, click **Log On**.

   > **Note** For a sharded cluster instance, you must also select a mongos node.

   Log on to the **DMS** console.
5. In the **Login instance** dialog box, configure the parameters described in the following table.

Parameter	Description
Database Type	The database engine of the instance. By default, this parameter is set to the database engine of the instance that you want to access.
Instance Region	The region where the instance is deployed. By default, this parameter is set to the region where the current instance is deployed.
Connection string address	The connection string of the instance. By default, this parameter is set to the connection string of the current instance.
Database Name	The name of the database. By default, this parameter is set to admin.
Database Account	The account used to connect to the database. By default, this parameter is set to root.
Database Password	The password of the account used to connect to the database.

6. Click **Login**.

   > **Note** You can select **Remember password** to eliminate the need to manually enter the password again the next time you log on to the database.

15.1.3.5.2. Use the mongo shell to connect to an ApsaraDB for MongoDB instance

This topic describes how to use the mongo shell to connect to an ApsaraDB for MongoDB instance. The mongo shell is a database management tool provided by ApsaraDB for MongoDB. You can install it on your client or in an Elastic Compute Service (ECS) instance.

Prerequisites

- The version of the mongo shell is the same as that of your ApsaraDB for MongoDB instance. This ensures successful authentication. For more information about the installation procedure, see Install MongoDB.

  > **Note** You can select a MongoDB version corresponding to your client version in the upper-left corner of the page.

- The IP address of your client is added to a whitelist of the ApsaraDB for MongoDB instance. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections** to view connection strings.

   > **Note**
   > - Replica set instances: View the connection string or connection string URI of a node.
   > - Sharded cluster instances: View the connection string or connection string URI of a mongos node.
   >
   > For more information about connection strings, see Overview of replica set instance connections or Overview of sharded cluster instance connections.

5. Connect to a database of the instance from your client or ECS instance where the mongo shell is installed.
   - Replica set instances
     - High-availability connection (recommended)

       You can use a connection string URI to connect to both the primary and secondary nodes of the instance. This ensures that your application is always connected to the primary node and the read and write operations of your application are not affected even if the roles of the primary and secondary nodes are switched.

       Command syntax:

       ```
       mongo "<ConnectionStringURI>"
       ```

       > **Note**
       > - The connection string URI must be enclosed in a pair of double quotation marks ("").
       > - <ConnectionStringURI>: the **connection string URI** of the instance.
       >
       >   You must replace `****` in the **connection string URI** with the database password. For more information about how to set a database password, see Reset the password for an ApsaraDB for MongoDB instance.

       Example:

       ```
       mongo "mongodb://root:****@dds-**********.mongodb.rds.intra.env17e.shuguang.com:3717,dds-****
       ******.mongodb.rds.intra.env17e.shuguang.com:3717/admin?replicaSet=mgset-*****"
       ```

- Single-node connection

  In most cases, you can directly connect to the primary, secondary, or read-only node. When the primary node fails, the system automatically switches to the secondary node and the secondary node becomes the primary node. This affects the read and write operations of your application.

  Command syntax:

  ```
  mongo --host <host> -u <username> -p --authenticationDatabase <database>
  ```

  > **Note**
  > - <host>: the endpoint used to log on to the primary or secondary node.
  > - <username>: the username used to log on to a database of the instance. The initial username is **root**.
  > - <database>: the name of the database corresponding to the username if authentication is enabled. If the username is root, enter admin as the database name.

  Example:

  ```
  mongo --host dds-bp**********.mongodb.rds.aliyuncs.com:3717 -u root -p --authenticationDatabase admin
  ```

  When `Enter password:` is displayed, enter the password of the username and press the Enter key. If you forget the password of the root username, you can reset the password. For more information, see Reset the password for an ApsaraDB for MongoDB instance.

  > **Note** The password characters are not displayed when you enter the password.

- Sharded cluster instances
  - High-availability connection (recommended)

    You can use a connection string URI to connect to a database of the instance. If one mongos node fails, another mongos node takes over business to ensure the high availability of the connection.

    Command syntax:

    ```
    mongo "<ConnectionStringURI>"
    ```

    > **Note**
    > - The connection string URI must be enclosed in a pair of double quotation marks ("").
    > - <ConnectionStringURI>: the **connection string URI** of the instance.
    >
    >   You must replace `****` in the **connection string URI** with the database password. For more information about how to set a database password, see Reset the password for an ApsaraDB for MongoDB instance.

    Example:

    ```
    mongo "mongodb://root:****@s-**********.mongodb.rds.intra.env17e.shuguang.com:3717,s-**********.mongodb.rds.intra.env17e.shuguang.com:3717/admin"
    ```

- Mongos node connection

  Command syntax:

  ```
  mongo --host <mongos_host> -u <username> -p --authenticationDatabase <database>
  ```

  > **Note**
  > - <mongos_host>: the endpoint used to log on to a mongos node.
  > - <username>: the username used to log on to a database of the instance. The initial username is **root**.
  > - <database>: the name of the database corresponding to the username if authentication is enabled. If the username is root, enter admin as the database name.

  Example:

  ```
  mongo --host s-bp**********.mongodb.rds.aliyuncs.com:3717 -u root -p --authenticationDatabase admin
  ```

15.1.3.5.3. Introduction to connection strings and URIs

15.1.3.5.3.1. Overview of replica set instance connections

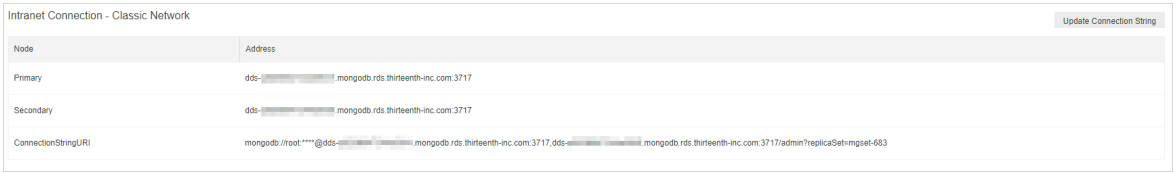
ApsaraDB for MongoDB supports both connection strings and connection string URIs. You can use a connection string to connect to the primary or secondary node, and use a connection string URI to connect to both of them. For high availability, we recommend that you use connection string URIs to connect your application to both primary and secondary nodes. This topic provides an overview of replica set instance connections.

Prerequisites

A whitelist is configured for the replica set instance. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

View connection strings

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections** to view connection strings.



Description of connection strings

Item	Description

Item	Description
Connection address type	• Classic network endpoint: Classic network endpoints are used for communication over the classic network. In the classic network, Apsara Stack services are not isolated. To block unauthorized traffic, you must configure security groups or IP address whitelists. • VPC endpoint: Virtual private cloud (VPC) endpoints are used for communication over VPCs. A VPC is an isolated network that provides higher security and higher performance than the classic network. By default, ApsaraDB for MongoDB provides VPC endpoints for ApsaraDB for MongoDB instances to ensure high security and high performance.
Role	• Primary: the primary node in the replica set instance. If you connect to this node, you can perform read and write operations on the databases of the replica set instance. • Secondary: the secondary node in the replica set instance. If you connect to this node, you can perform only read operations on the databases of the replica set instance. • Connection String URI: ApsaraDB for MongoDB allows you to use a connection string URI to connect to a replica set instance to achieve load balancing and high availability.
Connection string	The connection string of a primary or secondary node is in the following format: `<host>:<port>` • <host>: the endpoint used to connect to the replica set instance. • <port>: the port number used to connect to the replica set instance.
Connection string URI	A connection string URI is in the following format: `mongodb://[username:password@]host1[:port1][,host2[:port2],...[,hostN[:portN]]][/[database][?options]]` • mongodb://: the prefix of the connection string URI. • username:password@: the username and password used to log on to the replica set instance. You must separate them with a colon (:). • hostX:portX: the endpoint and port number used to connect to the replica set instance. • /database: the name of the database corresponding to the username if authentication is enabled. • ?options: the additional options that are used to connect to the replica set instance. **Note** If your application is in a production environment, we recommend that you use a connection string URI to connect to the instance. This way, when a node fails, the read and write operations of your application are not affected as a result of the failover.

Related information

- Connect to a replica set instance by using the mongo shell

15.1.3.5.3.2. Overview of sharded cluster instance connections

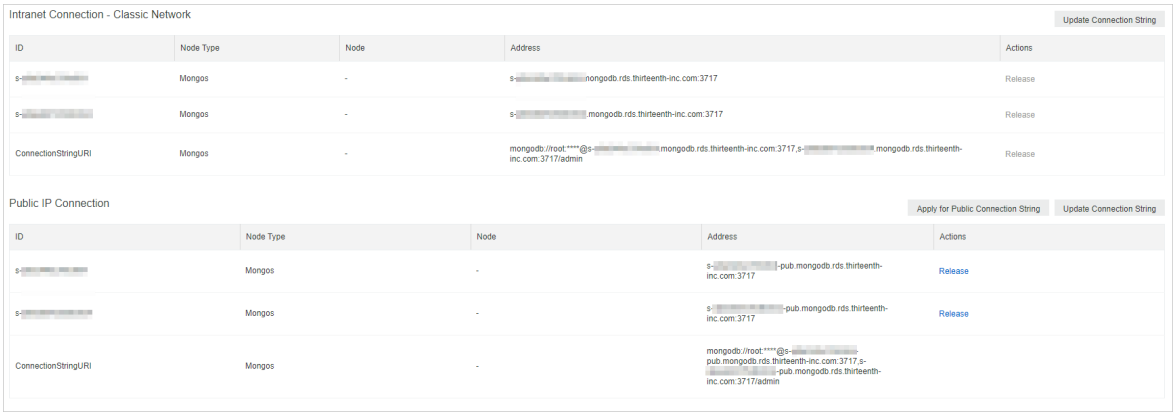
ApsaraDB for MongoDB supports both connection strings and connection string URIs. You can use a connection string to connect to a single mongos node, and use a connection string URI to connect to multiple mongos nodes. For high availability, we recommend that you use connection string URIs to connect your application to multiple mongos nodes. This topic provides an overview of sharded cluster instance connections.

View connection strings

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see

Log on to the ApsaraDB for MongoDB console.

2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections** to view connection strings.



Description of connection strings

Item	Description
Connection address type	• Classic network endpoint: Classic network endpoints are used for communication over the classic network. In the classic network, Apsara Stack services are not isolated. To block unauthorized traffic, you must configure security groups or IP address whitelists. • VPC endpoint: Virtual private cloud (VPC) endpoints are used for communication over VPCs. A VPC is an isolated network that provides higher security and higher performance than the classic network. By default, ApsaraDB for MongoDB provides VPC endpoints for ApsaraDB for MongoDB instances to ensure high security and high performance. • Public endpoint: Public endpoints are used for communication over the Internet. If you connect to an ApsaraDB for MongoDB instance over the Internet, the instance may be exposed to security risks. By default, ApsaraDB for MongoDB does not provide public endpoints for ApsaraDB for MongoDB instances. If you want to connect to an ApsaraDB for MongoDB instance from a device outside of Apsara Stack (such as an on-premises device), you must apply for a public endpoint. For more information, see Apply for a public endpoint for an ApsaraDB for MongoDB instance.
Mongos node ID	The connection string of a mongos node is in the following format: `<host>:<port>` • <host>: the endpoint used to connect to the sharded cluster instance. • <port>: the port number used to connect to the sharded cluster instance. **Note** During routine tests, you can use a connection string to directly connect to a mongos node.

Item	Description
Connection string URI	A connection string URI is in the following format: `mongodb://[username:password@]host1[:port1][,host2[:port2],...[,hostN[:portN]]][/[database][?options]]` • mongodb://: the prefix of the connection string URI. • username:password@: the username and password used to log on to the sharded cluster instance. You must separate them with a colon (:). • hostX:portX: the endpoint and port number used to connect to the sharded cluster instance. • /database: the name of the database corresponding to the username if authentication is enabled. • ?options: the additional options that are used to connect to the sharded cluster instance. Note If your application is in a production environment, we recommend that you use a connection string URI to connect to the sharded cluster instance. Then, your client can automatically distribute your requests to multiple mongos nodes to balance loads. If a mongos node fails, your client automatically redirects requests to other mongos nodes in the normal state.

15.1.4. Instances

15.1.4.1. Create an ApsaraDB for MongoDB instance

This topic describes how to create an instance in the ApsaraDB for MongoDB console.

Create a replica set instance

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click **Create Instance** in the upper-left corner.
4. On the **Create ApsaraDB for MongoDB Instance** page, configure the parameters described in the following table.

Section	Parameter	Description
Basic Settings	**Organization**	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
	Region	The region where the instance is deployed.

Region Section	Parameter	Description
	Zone	The zone where the instance is deployed. **Note** If you select dual zones, the instance supports zone-disaster recovery across two data centers.
Specifications	Chip Architecture	The chip architecture of the host where the instance is deployed. **Note** If you do not have permissions to select an option, contact the operations administrator to grant such permissions to your account.
	Database Engine	The value is fixed at **MongoDB**.
	Engine Version	The database engine version of the instance. Valid values: ○ 3.0 ○ 3.4 ○ 4.0 ○ 4.2
	Node Type	The following node types are available in ApsaraDB for MongoDB: ○ Three-node Replica Set: uses dedicated memory and I/O resources but shares CPU and storage resources with other general-purpose instances on the same server. ○ Dedicated Instance: uses dedicated CPU, memory, storage, and I/O resources to ensure long-term stable performance. In this case, an instance is not affected by other instances on the same server. ○ Dedicated Host: exclusively uses all resources of a server. This is the top configuration of exclusive specifications.

Section	Parameter	Description
	Node Specifications	The node specifications of the instance. For more information, see descriptions in the ApsaraDB for MongoDB console.
	Storage Capacity (GB)	The storage capacity of the instance, which includes the storage capacity for data, system files, log files, and transaction files. For more information, see *Instance types* in *ApsaraDB for MongoDB Product Introduction*.
Network	**Network Type**	The following network types are available in ApsaraDB for MongoDB: ○ **Classic Network**: Cloud services on the classic network are not isolated. Unauthorized access can be blocked only by security groups or whitelists. ○ **VPC**: A virtual private cloud (VPC) helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for enhanced security. **Note** If you select the **VPC** network type, you must configure the **VPC** and **vSwitch** parameters.
	VPC	The VPC in which the instance resides. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch of the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.

Section	Parameter	Description
Password Settings	Instance Name	The name of the instance. The name must meet the following requirements: ○ The name must start with a letter. ○ The name can contain digits, letters, underscores (_), and hyphens (-). ○ The name must be 2 to 256 characters in length.
	Password Setting	Specifies when to set a password. The following options are available: ○ **Set Now**: immediately sets the logon password. ○ **Set after Purchase**: sets the logon password after you create the instance. For more information, see Reset the password for an ApsaraDB for MongoDB instance.
	Logon Password	The password used to log on to the database. The password must meet the following requirements: ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include !#$%^&*()_+= ○ The password must be 8 to 32 characters in length.
	Confirm Password	Enter the password again. The password you enter here must be the same as that you entered in the Logon Password field.

5. Click **Submit**.

Create a sharded cluster instance

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click **Create Instance**.
4. On the **Create ApsaraDB for MongoDB Sharded Cluster Instance** page, configure the parameters described in the following table.

Section	Parameter	Description
Basic Settings	Organization	The organization to which the instance belongs.
	Resource Set	The resource set to which the instance belongs.
Region	Region	The region where the instance is deployed.
	Zone	The zone where the instance is deployed. Note If you select dual zones, the instance supports zone-disaster recovery across two data centers.
Specifications	Chip Architecture	The chip architecture of the host where the instance is deployed. Note If you do not have permissions to select an option, contact the operations administrator to grant such permissions to your account.
	Database Engine	The value is fixed at MongoDB.
	Engine Version	The database engine version of the instance. Valid values: ○ 3.4 ○ 4.0 ○ 4.2

Section	Parameter	Description
Network	Network Type	The following network types are available in ApsaraDB for MongoDB: ○ **Classic Network**: Cloud services on the classic network are not isolated. Unauthorized access can be blocked only by security groups or whitelists. ○ **VPC**: A VPC helps you build an isolated network environment on Apsara Stack. You can customize route tables, CIDR blocks, and gateways in a VPC. We recommend that you select VPC for enhanced security. **Note** If you select the **VPC** network type, you must configure the **VPC** and **vSwitch** parameters.
	VPC	The VPC in which the instance resides. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
	vSwitch	The vSwitch of the instance. **Note** When **Network Type** is set to **VPC**, you must specify this parameter.
Mongos Specifications	Mongos Specifications	The specifications of the mongos node. For more information, see descriptions in the ApsaraDB for MongoDB console.
	Quantity	The number of mongos nodes. You can select 2 to 32 mongos nodes.
Shard Specifications	Shard Specifications	The specifications of the shard node. For more information, see descriptions in the ApsaraDB for MongoDB console.
	Storage Capacity (GB)	The storage capacity of the shard node, which includes the storage capacity for data, system files, log files, and transaction files. For more information, see *Instance types* in *ApsaraDB for MongoDB Product Introduction*.

Section	Parameter	Description
	Quantity	The number of shard nodes. You can select 2 to 32 shard nodes.
Config Server Specifications	**Config Server Type**	The specifications of the Configserver node. The value is fixed at **1 core, 2 GB**.
	Disk (GB)	The storage capacity of the Configserver node. The value is fixed at 20 GB.
Password Settings	**Instance Name**	The name of the instance. The name must meet the following requirements: ○ The name must start with a letter. ○ The name can contain digits, letters, underscores (_), and hyphens (-). ○ The name must be 2 to 256 characters in length.
	Password Setting	Specifies when to set a password. The following options are available: ○ **Set Now**: immediately sets the logon password. ○ **Set after Purchase**: sets the logon password after you create the instance. For more information, see Reset the password for an ApsaraDB for MongoDB instance.
	Logon Password	The password used to log on to the database. The password must meet the following requirements: ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include !#$%^&*()_+= ○ The password must be 8 to 32 characters in length.
	Confirm Password	Enter the password again. The password you enter here must be the same as that you entered in the Logon Password field.

5. Click **Submit**.

15.1.4.2. View the details of an ApsaraDB for MongoDB instance

This topic describes how to view the details of an ApsaraDB for MongoDB instance, such as the basic information, internal network connection information, status, and configurations.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. Go to the basic information page by using one of the following methods:
   - Find the instance that you want to view and click its ID to go to the **Basic Information** page. Then, you can view the details of the instance.
   - Click ⋮ in the **Actions** column corresponding to the instance that you want to view and select **Manage** to go to the **Basic Information** page. Then, you can view the details of the instance.

15.1.4.3. Restart an ApsaraDB for MongoDB instance

You can manually restart an ApsaraDB for MongoDB instance when the number of connections exceeds the threshold or performance issues occur on the instance. This topic describes how to restart an ApsaraDB for MongoDB instance.

Prerequisites

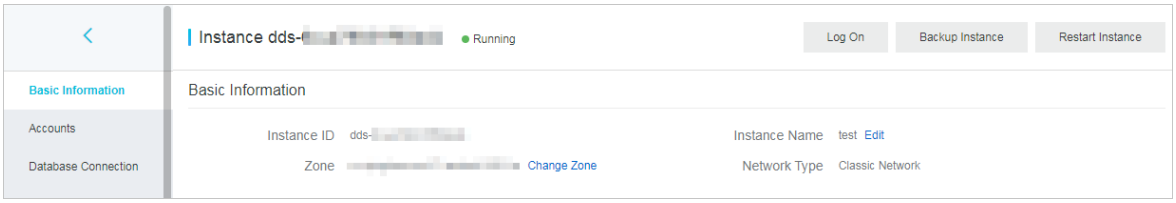
The instance is in the **Running** state.

Context

> **Note** When an ApsaraDB for MongoDB instance is restarted, all its connections are closed. Plan your operations in advance before you restart an ApsaraDB for MongoDB instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. Click **Restart Instance** in the upper-right corner.



> **Note** You can also click ⋮ in the **Actions** column corresponding to the instance and select **Restart**.

5. In the **Restart Instance** message, click **OK**.

15.1.4.4. Change the configurations of an ApsaraDB for MongoDB instance

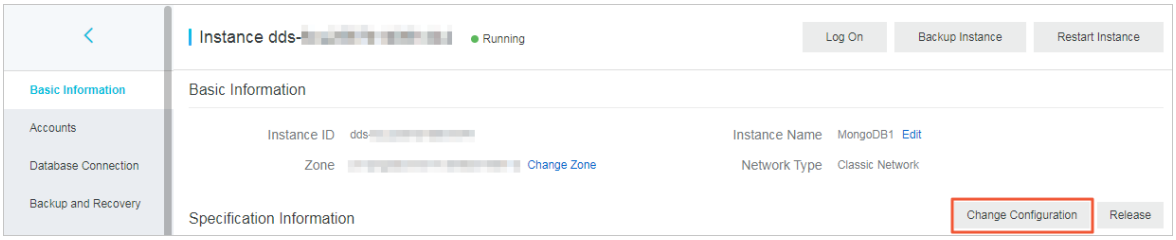
This topic describes how to change the configurations of an ApsaraDB for MongoDB instance. You can upgrade or downgrade an ApsaraDB for MongoDB instance to meet your business needs.

Prerequisites

The instance is an ApsaraDB for MongoDB replica set instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. Click **Change Configuration** in the upper-right corner of the Specification Information section to go to the **Change Specifications** page.



> **Note** To go to the **Change Specifications** page, you can also choose ⋮ > **Change Configuration** in the **Actions** column corresponding to the instance on the **Replica Set Instances** page.

5. On the **Change Specifications** page, change the instance configurations.

   You can change values of the following parameters:

   - **Node Type**
   - **Node Specifications**
   - **Storage Capacity (GB)**
6. After you change the instance configurations, click **Submit**.

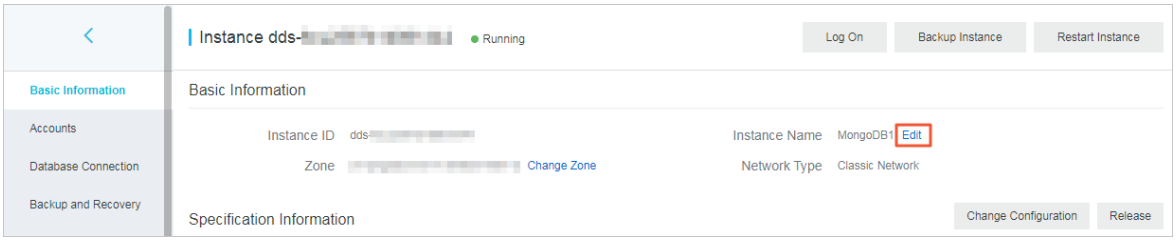
15.1.4.5. Change the name of an ApsaraDB for MongoDB instance

This topic describes how to change the name of an ApsaraDB for MongoDB instance to facilitate management.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.

2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. Click **Edit** next to **Instance Name**.



> **Note**
> - The instance name must start with a letter. It cannot start with http:// or https://.
> - The instance name can contain letters, underscores (_), hyphens (-), and digits.
> - The instance name must be 2 to 128 characters in length.

5. Click **OK**.

15.1.4.6. Reset the password for an ApsaraDB for MongoDB instance

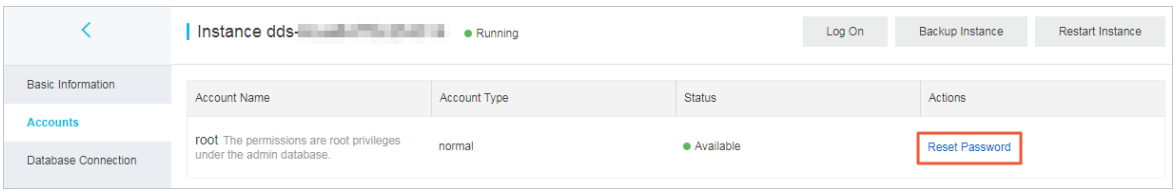
This topic describes how to reset your password in the ApsaraDB for MongoDB console.

Context

> **Notice** We recommend that you change your password on a regular basis to ensure data security.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. In the left-side navigation pane, click **Accounts**.
5. Click **Reset Password** in the Actions column and configure the parameters in the Reset Password panel.



Parameters for resetting a password describes the parameters.

Parameters for resetting a password

Parameter	Description
New Password	Specify the new password of the account based on the following rules: ○ The password must contain at least three of the following character types: uppercase letters, lowercase letters, digits, and special characters. Special characters include !#$%^&*()_+= ○ The password must be 8 to 32 characters in length.
Confirm New Password	Enter the password again. The password you enter here must be the same as that in New Password.

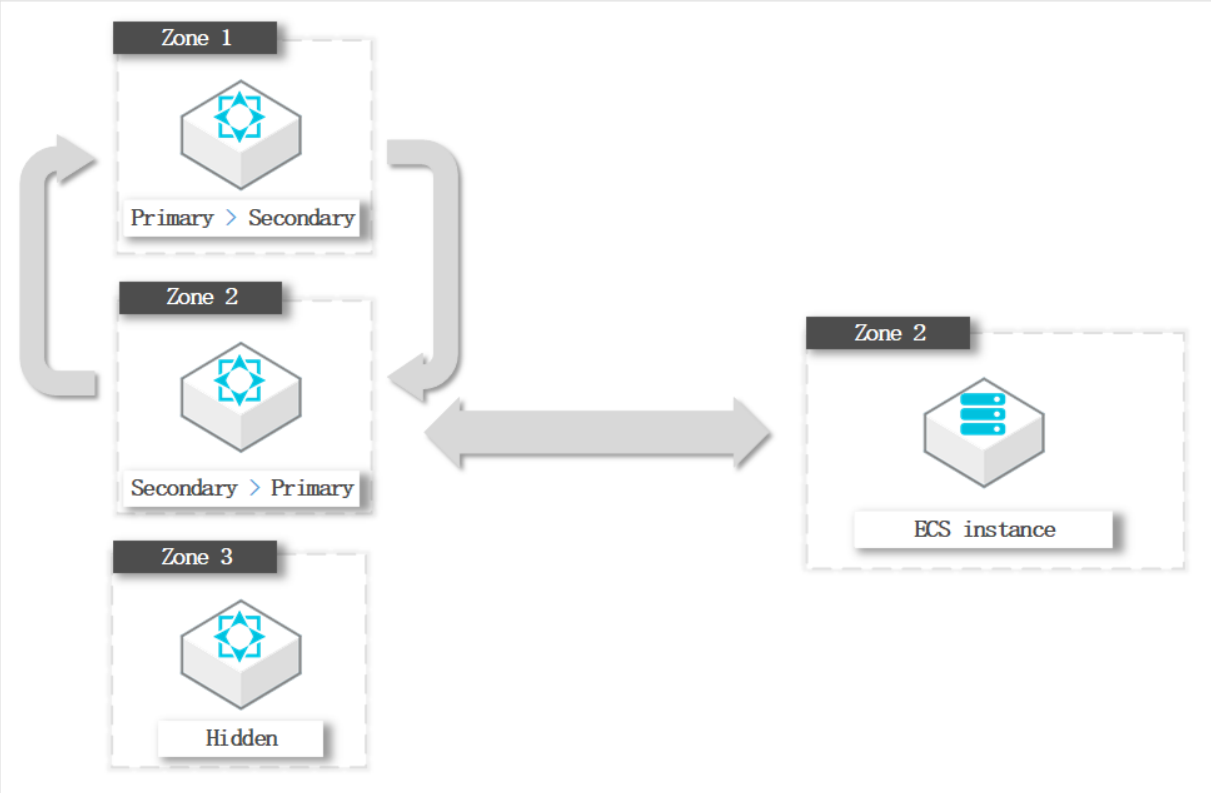
6. Click **OK**.

15.1.4.7. Switch node roles

You can switch the node roles of an ApsaraDB for MongoDB instance in the ApsaraDB for MongoDB console based on your business deployment.

Typical scenario

When an Elastic Compute Service (ECS) instance and an ApsaraDB for MongoDB instance are connected in the same zone over an internal network, the latency is minimal. If they are connected across different zones, the latency increases and the performance of the ApsaraDB for MongoDB instance and your business is affected.



In this example, the ECS instance to which the application belongs is in Zone 2. If the primary node of the ApsaraDB for MongoDB instance is in Zone 1, the ECS instance needs to connect to the primary node across zones.

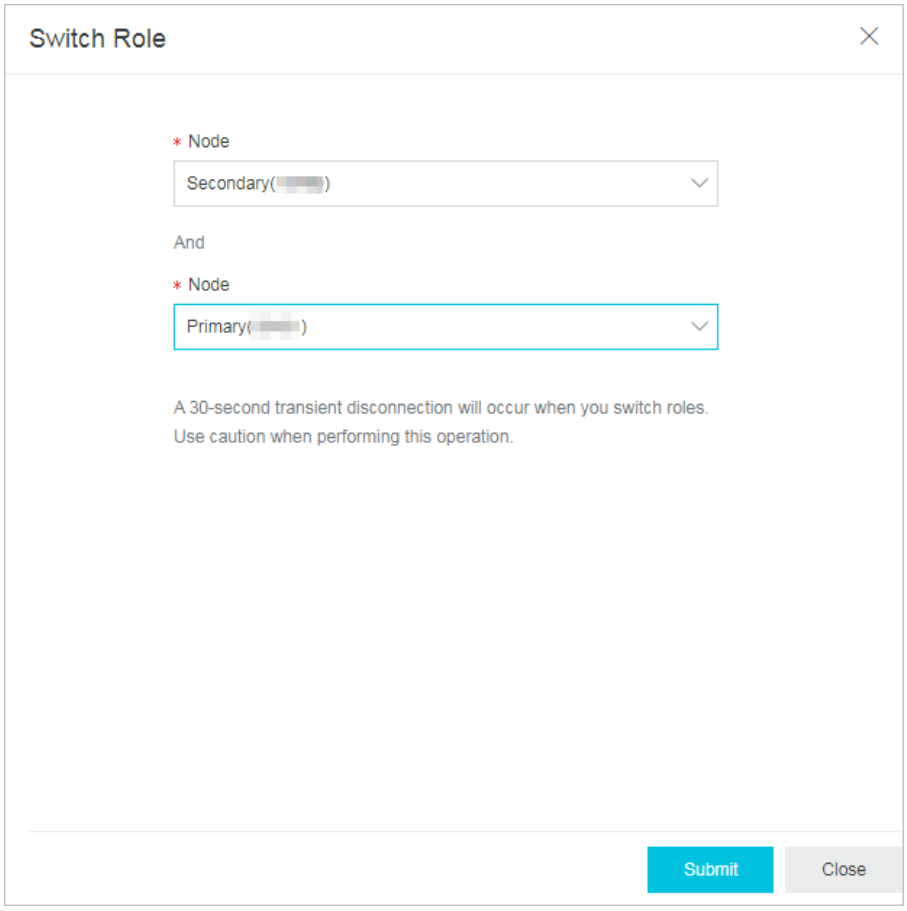
To optimize the business deployment architecture, you can switch roles between the primary and secondary nodes. In this example, you can change the role of the node in Zone 2 to primary and the role of the node in Zone 1 to secondary. The ECS and ApsaraDB for MongoDB instances can be connected in the same zone.

Precautions

- Each time you switch node roles, the instance may experience a transient connection of up to 30 seconds. Perform this operation during off-peak hours or make sure that your application can automatically re-establish a connection.
- Each time you switch node roles, only the node roles are changed and the zones and role IDs of nodes remain unchanged.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Service Availability**.
5. On the **Service Availability** page, perform the subsequent operations based on the corresponding instance architecture.
   - Replica set instances
     a. Click **Switch Role** in the upper-right corner of the page.
     b. In the **Switch Role** panel, select the node for which you want to switch the role.

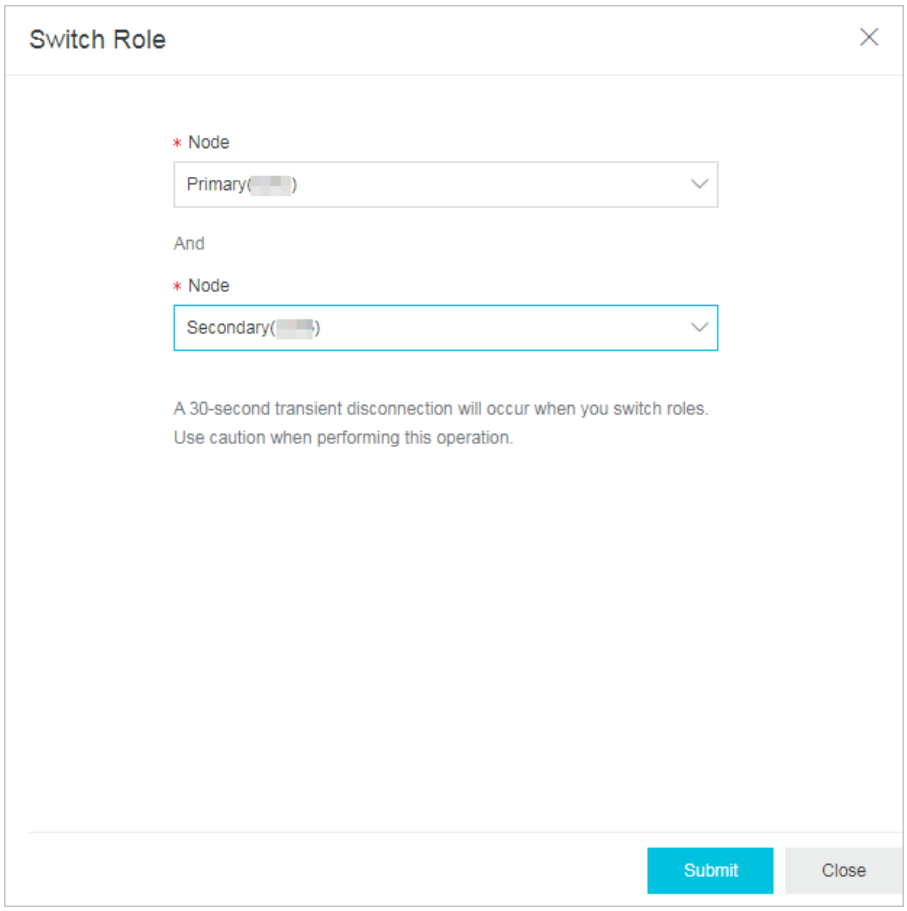


     c. Click **OK**.
   - Sharded cluster instances

> **Note** For sharded cluster instances, you can manage only the zone distribution of shard and Configserver nodes.

a. In the upper-right corner of the **Zone Distribution for Shards** or **Zone Distribution for Configservers** section, click **Switch Role**.

b. In the **Switch Role** panel, select the node for which you want to switch the role.



c. Click **OK**.

15.1.4.8. Migrate an ApsaraDB for MongoDB instance across zones in the same region

This topic describes how to migrate an ApsaraDB for MongoDB instance across zones in the same region. After instances are migrated to other zones, the attributes, specifications, and connection strings of the instances remain unchanged.

Prerequisites

- The ApsaraDB for MongoDB instance is a replica set instance or sharded cluster instance that runs MongoDB 4.2 or later.
- Transparent data encryption (TDE) is not enabled for the ApsaraDB for MongoDB instance.
- The destination zone must be in the same region as the current zone of the instance.
- If the instance is in a VPC, make sure that a vSwitch is created in the destination zone before you start migration. For more information, see *Create a VPC* and *Create a vSwitch* in *VPC User Guide*.
- The instance does not have a public endpoint. If you have applied for a public endpoint, you must release it

before migration. For more information, see Release a public connection string.

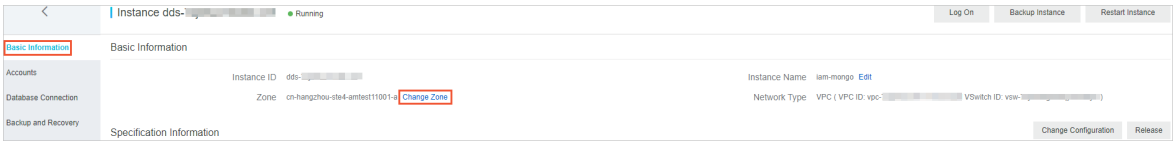
Precautions

- If the instance is in a VPC, you cannot change the VPC when the instance is in the migration process.
- The time required varies based on factors such as the network conditions, task queue status, and data volume. We recommend that you migrate the instance across zones during off-peak hours.
- During the migration, a transient connection of 30 seconds occurs. Make sure that your application is configured to reconnect to the instance after it is disconnected.
- The virtual IP addresses (VIPs) of the instance, such as 172.16.88.60, are changed when the instance is migrated across zones. If your application uses the original VIP, the application cannot connect to the instance after the migration.

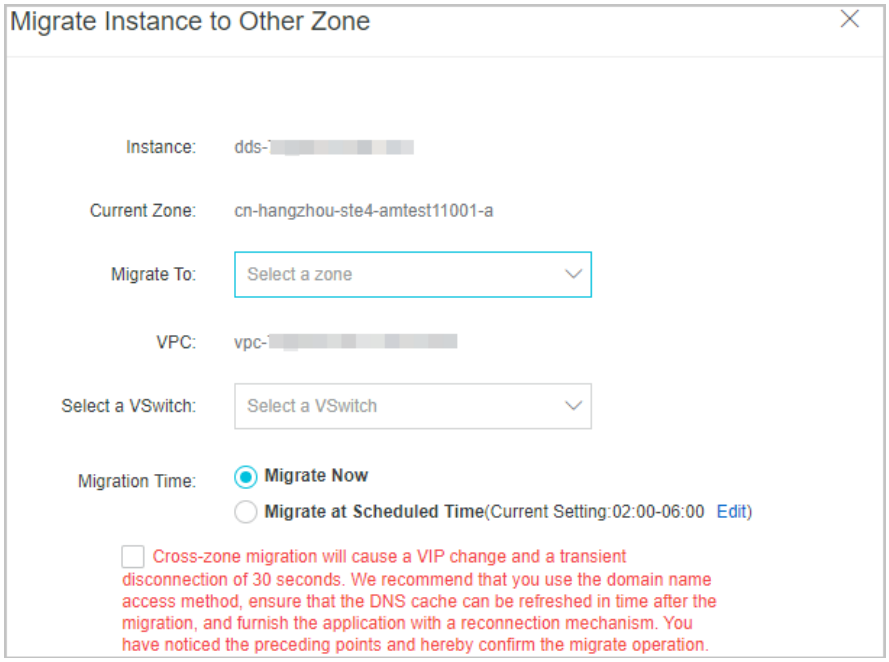
> **Note** We recommend that you use a connection string URI to connect to the instance, which ensures high availability. For more information, see Overview of replica set instance connections or Overview of sharded cluster instance connections.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click in the **Actions** column corresponding to the instance and select **Manage**.
4. In the **Basic Information** section, click **Change Zone**.

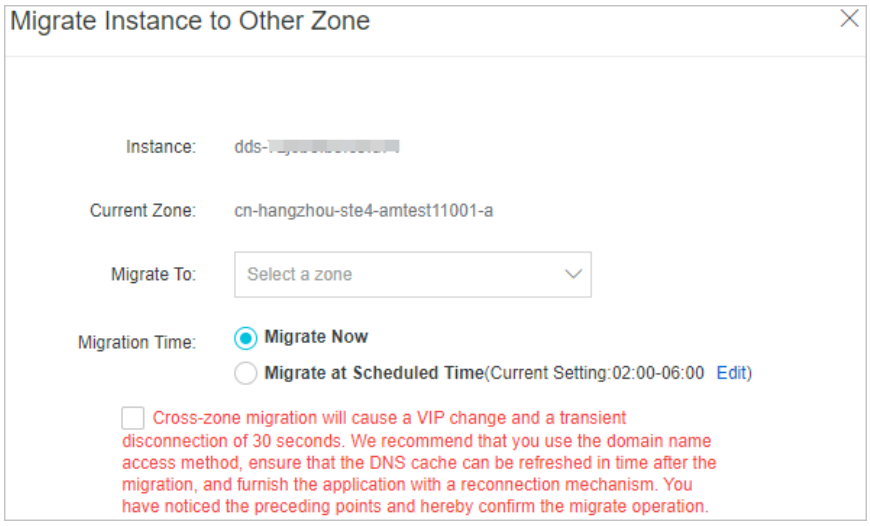


5. In the **Migrate Instance to Other Zone** panel, configure parameters based on the network type of the instance.
   - VPC



Parameter	Description
Migrate To	Select the destination zone.
Select a VSwitch	Select the destination vSwitch.
Migration Time	Select the time when you want to start the migration. ■ **Switch Immediately after Migration**: The migration immediately starts. When the instance status changes to Running, the migration is complete. ■ **Switch within Maintenance Window**: The migration starts during the specified period. You can click **Edit** next to **Switch within Maintenance Window** to change the period.

- Classic network



Parameter	Description
Migrate To	Select the destination zone.

Parameter	Description
Migration Time	Select the time when you want to start the migration. ■ **Switch Immediately after Migration**: The migration immediately starts. When the instance status changes to Running, the migration is complete. ■ **Switch within Maintenance Window**: The migration starts during the specified period. You can click **Edit** next to **Switch within Maintenance Window** to change the period.

6. Read the message that is displayed and select the check box next to the message.
7. Click **OK**.

15.1.4.9. Release an ApsaraDB for MongoDB instance

This topic describes how to manually release an ApsaraDB for MongoDB instance to meet your business needs.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. On the page that appears, click **Release** in the lower-right corner of the **Basic Information** section.

   > **Note** You can also click ⋮ in the **Actions** column corresponding to the instance and select **Release**.

5. In the **Release Instance** message, click **OK**.

   > **Warning** After you release an ApsaraDB for MongoDB instance, data in the instance can no longer be recovered. Proceed with caution.

15.1.4.10. Primary/secondary failover

15.1.4.10.1. Configure a primary/secondary failover for a replica set instance

By default, an ApsaraDB for MongoDB replica set instance consists of three nodes. ApsaraDB for MongoDB provides connection strings for you to connect to the primary node and a secondary node. The other secondary node is hidden as a backup to ensure high availability. If a node fails, the high availability system of ApsaraDB for MongoDB automatically triggers a primary/secondary failover to ensure the availability of the instance. You can also manually trigger a primary/secondary failover for an ApsaraDB for MongoDB instance in scenarios such as routine disaster recovery drills.

Prerequisites

The instance is in the **Running** state.

Context

After a primary/secondary failover is triggered for a replica set instance, the system switches roles between the primary and secondary nodes in the instance.

Precautions

Each time you trigger a primary/secondary failover for an instance, the instance may experience a transient connection of up to 30 seconds. Make sure that your applications can automatically re-establish a connection.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. On the **Replica Set Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
3. On the page that appears, click **Failover** in the **Node List** section.

4. In the **Failover** message, click **OK**.
5. The instance state changes to **Switching role**. When the instance state changes to **Running**, the failover is successful.

> **Note**
> - The failover operation is complete in about one minute.
> - If you have connected to the instance by using the connection string of the primary node, you are connecting to a secondary node after a failover and you have no write permissions on the instance. For more information about how to use the connection string of the new primary node to connect to the instance, see Overview of replica set instance connections.

15.1.4.10.2. Configure a primary/secondary failover for a sharded cluster instance

By default, each shard or Configserver node of a sharded cluster instance consists of three nodes. If a node fails, the high availability system of ApsaraDB for MongoDB automatically triggers a primary/secondary failover to ensure the availability of the shard or Configserver node. You can also manually trigger a primary/secondary failover for an ApsaraDB for MongoDB instance in scenarios such as routine disaster recovery drills.

Prerequisites

The instance is in the **Running** state.

Context

After a primary/secondary failover is triggered, the system switches roles between the primary and secondary nodes in a shard node.

Precautions

Each time you trigger a primary/secondary failover for an instance, the instance may experience a transient connection of up to 30 seconds. We recommend that you perform this operation during off-peak hours and ensure that your applications can automatically re-establish a connection.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. In the **Shard List** or **ConfigServer List** section, click ⋮ in the **Actions** column corresponding to the node for which you want to perform failover and select **Failover**.

   > **Note** You can separately trigger a primary/secondary failover for each shard node. The failover operation takes effect only for the current shard node and does not affect other shard nodes of the same sharded cluster instance.

5. In the **Failover** message, click **OK**.

   > **Note** The failover operation is complete in about one minute.

15.1.4.11. View monitoring data

This topic describes the performance metrics provided by ApsaraDB for MongoDB to check the status of ApsaraDB for MongoDB instances. You can view instance monitoring data in the ApsaraDB for MongoDB console.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Monitoring Data**.
5. On the page that appears, specify a time range to view the monitoring data.

   Instance monitoring data is collected every 300 seconds.

Metric	Description
CPU utilization	cpu_usage: the CPU utilization of the instance.
Memory usage	mem_usage: the memory usage of the instance.
IOPS usage	The IOPS of the instance. The following items are included: ◦ data_iops: the IOPS of the data disk. ◦ log_iops: the IOPS of the disk that stores logs.
IOPS usage percentage	iops_usage: the ratio of the IOPS used by the instance to the maximum IOPS allowed.

Metric	Description
Disk usage	The total disk space used by the instance. The following items are included: ○ ins_size: the total space used. ○ data_size: the disk space used by data files. ○ log_size: the disk space used by log files.
Disk usage percentage	disk_usage: the ratio of the total disk space used by the instance to the maximum disk space that can be used.
QPS	The queries per second (QPS) of the instance. The following items are included: ○ The number of insert operations. ○ The number of query operations. ○ The number of delete operations. ○ The number of update operations. ○ The number of getMore operations. ○ The number of command operations.
Connections	current_conn: the number of current connections to the instance.
Cursors	The number of cursors used by the instance. The following items are included: ○ total_open: the number of cursors that are opened. ○ timed_out: the number of cursors that timed out.
Network traffic	The network traffic of the instance. The following items are included: ○ bytes_in: the inbound network traffic. ○ bytes_out: the outbound network traffic. ○ num_requests: the number of requests that are processed.
Read/write queues	The length of the queues that are waiting for global locks for the instance. The following items are included: ○ gl_cq_total: the length of the queue that is waiting for both global read and write locks. ○ gl_cq_readers: the length of the queue that is waiting for global read locks. ○ gl_cq_writers: the length of the queue that is waiting for global write locks.
WiredTiger	The cache metrics of the WiredTiger engine used by the instance. The following items are included: ○ bytes_read_into_cache: the amount of data that is read into the cache. ○ bytes_written_from_cache: the amount of data that is written from the cache to the disk. ○ maximum_bytes_configured: the size of the maximum available disk space that is configured.
Primary/secondary replication latency	repl_lag: the latency in data synchronization between the primary and secondary nodes of the instance.

15.1.5. Backup and restoration

15.1.5.1. Configure automatic backup for an ApsaraDB for MongoDB instance

This topic describes how to configure automatic backup for an ApsaraDB for MongoDB instance. ApsaraDB for MongoDB can automatically back up data based on the default backup policy or the backup policy you specify.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Backup and Recovery**.
5. In the upper-left corner of the page, click **Backup Settings**.
6. In the **Backup Settings** panel, configure the parameters in the following table.

Parameter	Description
Retention Days	The backup retention period is fixed to seven days.
Backup Time	The period of time during which you want to back up data. We recommend that you specify a time period that is during off-peak hours.
Day of Week	The days of a week on which you want to back up data. You can select multiple days.

7. Click **OK**.

15.1.5.2. Manually back up an ApsaraDB for MongoDB instance

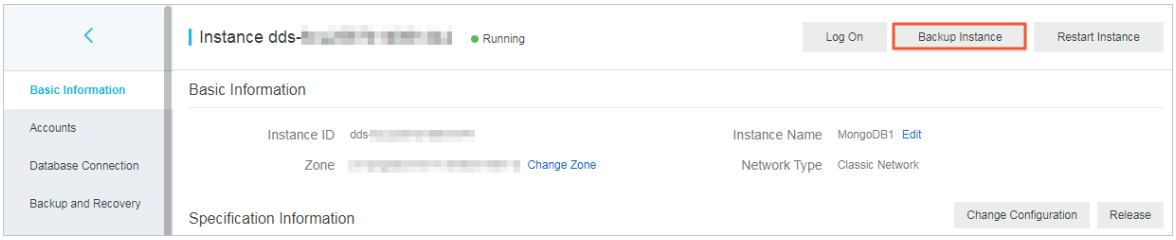
This topic describes how to manually back up an ApsaraDB for MongoDB instance. ApsaraDB for MongoDB supports both automatic backup and manual backup. You can configure a backup policy for the system to automatically back up your ApsaraDB for MongoDB instance based on the backup cycle you specify.

Backup methods

- Physical backup: This method backs up physical database files of an ApsaraDB for MongoDB instance. Compared with logical backup, physical backup provides faster data backup and restoration.
- Logical backup: The mongodump tool is used to store operation records of databases in a logical backup file. Logical backup restores data in the form of playback commands during restoration.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
4. In the upper-right corner of the page, click **Back up Instance**.



5. In the **Back up Instance** panel, select a backup method from the **Backup Method** drop-down list and then click **OK**.

15.1.5.3. Restore data to the current ApsaraDB for MongoDB instance

This topic describes how to restore data to the current ApsaraDB for MongoDB instance. This helps minimize the data loss caused by incorrect operations.

Prerequisites

The instance is a replica set instance with three nodes.

Background information

- The time required to restore data to your current instance varies depending on factors such as the data volume, task queue status, and network conditions. When the status of the instance changes to **Running**, the restoration is complete.
- If you restore data to your current instance, all existing data is overwritten and cannot be restored.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. On the **Replica Set Instances** page, click the ID of an instance, or click ⋮ in the **Actions** column corresponding to the instance and select **Manage**.
3. In the left-side navigation pane, click **Backup and Recovery**.
4. On the **Backup and Recovery** page that appears, find the backup set and choose ⋮ > **Data Recovery** in the Actions column.

   > **Note** If you have upgraded the database version, you cannot use the backup files of the earlier database version to restore data.

5. In the **Roll Back Instance** message, click **OK**.

   > **Note** The instance status becomes **Restoring from Backup** after you click **OK**. You can click **Refresh** in the upper-right corner of the **Backup and Recovery** page to update the instance status. The restoration is complete when the instance status changes to **Running**.

15.1.5.4. Download a backup file

ApsaraDB for MongoDB allows you to download backup files that can be used to restore databases. This topic describes how to download a backup file in the ApsaraDB for MongoDB console.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Backup and Recovery**.
5. On the page that appears, click ⋮ in the **Actions** column corresponding to a backup file and select **Download**.
6. In the **Download Backup** dialog box, click **Copy** next to the download URL and then click **OK**.
7. Select an appropriate method to download the backup file based on your business environment. You can run the wget command, or paste the copied download URL into the address bar of your browser and press the Enter key to download the backup file.

15.1.6. Database connections

15.1.6.1. Modify a public or internal endpoint of an ApsaraDB for MongoDB instance

This topic describes how to modify a public or internal endpoint of an ApsaraDB for MongoDB instance in the ApsaraDB for MongoDB console.

Limits

Instance architecture	Limit
Replica set instance	Public and internal endpoints can be modified for both primary and secondary nodes.
Sharded cluster instance	Only the public and internal endpoints of mongos nodes can be modified.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections**.
5. In the **Internal Connections** or **Public Connections** section, click **Update Connection String**.
6. In the **Update Connection String** panel, specify a new endpoint.

   > **Note** Only the prefix of the endpoint can be modified. You must abide by the following rules when you modify the prefix:
   >
   > - The prefix must start with a lowercase letter.
   > - The prefix must end with a lowercase letter or digit.
   > - The prefix must be 8 to 64 characters in length and can contain lowercase letters, digits, and hyphens (-).

7. Click **Submit**.

What's next

After you modify the public or internal endpoint, you must connect a client or an application to your ApsaraDB for MongoDB instance by using the new endpoint.

15.1.6.2. Use DMS to log on to an ApsaraDB for MongoDB instance

You can use Data Management (DMS) to log on to an ApsaraDB for MongoDB instance.

Prerequisites

An IP address whitelist is configured. For more information about how to configure an IP address whitelist, see Configure a whitelist for an ApsaraDB for MongoDB instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance, or click in the **Actions** column corresponding to the instance and select **Manage**.
4. In the upper-right corner of the page, click **Log On**.

   > **Note** For a sharded cluster instance, you must also select a mongos node.

   Log on to the **DMS** console.
5. In the **Login instance** dialog box, configure the parameters described in the following table.

Parameter	Description
Database Type	The database engine of the instance. By default, this parameter is set to the database engine of the instance that you want to access.
Instance Region	The region where the instance is deployed. By default, this parameter is set to the region where the current instance is deployed.
Connection string address	The connection string of the instance. By default, this parameter is set to the connection string of the current instance.
Database Name	The name of the database. By default, this parameter is set to admin.
Database Account	The account used to connect to the database. By default, this parameter is set to root.
Database Password	The password of the account used to connect to the database.

6. Click **Login**.

   > **Note** You can select **Remember password** to eliminate the need to manually enter the password again the next time you log on to the database.

15.1.6.3. Use the mongo shell to connect to an ApsaraDB for MongoDB instance

This topic describes how to use the mongo shell to connect to an ApsaraDB for MongoDB instance. The mongo shell is a database management tool provided by ApsaraDB for MongoDB. You can install it on your client or in an Elastic Compute Service (ECS) instance.

Prerequisites

- The version of the mongo shell is the same as that of your ApsaraDB for MongoDB instance. This ensures successful authentication. For more information about the installation procedure, see Install MongoDB.

  > **Note** You can select a MongoDB version corresponding to your client version in the upper-left corner of the page.

- The IP address of your client is added to a whitelist of the ApsaraDB for MongoDB instance. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections** to view connection strings.

   > **Note**
   > - Replica set instances: View the connection string or connection string URI of a node.
   > - Sharded cluster instances: View the connection string or connection string URI of a mongos node.
   >
   > For more information about connection strings, see Overview of replica set instance connections or Overview of sharded cluster instance connections.

5. Connect to a database of the instance from your client or ECS instance where the mongo shell is installed.
   - Replica set instances
     - High-availability connection (recommended)

       You can use a connection string URI to connect to both the primary and secondary nodes of the instance. This ensures that your application is always connected to the primary node and the read and write operations of your application are not affected even if the roles of the primary and secondary nodes are switched.

       Command syntax:

       ```
       mongo "<ConnectionStringURI>"
       ```

       > **Note**
       > - The connection string URI must be enclosed in a pair of double quotation marks ("").
       > - <ConnectionStringURI>: the **connection string URI** of the instance.
       >
       >   You must replace `****` in the **connection string URI** with the database password. For more information about how to set a database password, see Reset the password for an ApsaraDB for MongoDB instance.

       Example:

       ```
       mongo "mongodb://root:****@dds-**********.mongodb.rds.intra.env17e.shuguang.com:3717,dds-****
       ******.mongodb.rds.intra.env17e.shuguang.com:3717/admin?replicaSet=mgset-*****"
       ```

- Single-node connection

  In most cases, you can directly connect to the primary, secondary, or read-only node. When the primary node fails, the system automatically switches to the secondary node and the secondary node becomes the primary node. This affects the read and write operations of your application.

  Command syntax:

  ```
  mongo --host <host> -u <username> -p --authenticationDatabase <database>
  ```

  > **Note**
  > - <host>: the endpoint used to log on to the primary or secondary node.
  > - <username>: the username used to log on to a database of the instance. The initial username is **root**.
  > - <database>: the name of the database corresponding to the username if authentication is enabled. If the username is root, enter admin as the database name.

  Example:

  ```
  mongo --host dds-bp**********.mongodb.rds.aliyuncs.com:3717 -u root -p --authenticationDatabase admin
  ```

  When `Enter password:` is displayed, enter the password of the username and press the Enter key. If you forget the password of the root username, you can reset the password. For more information, see Reset the password for an ApsaraDB for MongoDB instance.

  > **Note** The password characters are not displayed when you enter the password.

- Sharded cluster instances
  - High-availability connection (recommended)

    You can use a connection string URI to connect to a database of the instance. If one mongos node fails, another mongos node takes over business to ensure the high availability of the connection.

    Command syntax:

    ```
    mongo "<ConnectionStringURI>"
    ```

    > **Note**
    > - The connection string URI must be enclosed in a pair of double quotation marks ("").
    > - <ConnectionStringURI>: the **connection string URI** of the instance.
    >
    >   You must replace `****` in the **connection string URI** with the database password. For more information about how to set a database password, see Reset the password for an ApsaraDB for MongoDB instance.

    Example:

    ```
    mongo "mongodb://root:****@s-**********.mongodb.rds.intra.env17e.shuguang.com:3717,s-**********.mongodb.rds.intra.env17e.shuguang.com:3717/admin"
    ```

- Mongos node connection

  Command syntax:

```
mongo --host <mongos_host> -u <username> -p --authenticationDatabase <database>
```

> **Note**
>
> - <mongos_host>: the endpoint used to log on to a mongos node.
> - <username>: the username used to log on to a database of the instance. The initial username is **root**.
> - <database>: the name of the database corresponding to the username if authentication is enabled. If the username is root, enter admin as the database name.

  Example:

```
mongo --host s-bp**********.mongodb.rds.aliyuncs.com:3717 -u root -p --authenticationDatabase admin
```

15.1.6.4. Apply for a public endpoint for a sharded cluster instance

This topic describes how to apply for a public endpoint for an ApsaraDB for MongoDB instance when you want to connect to this instance over the Internet.

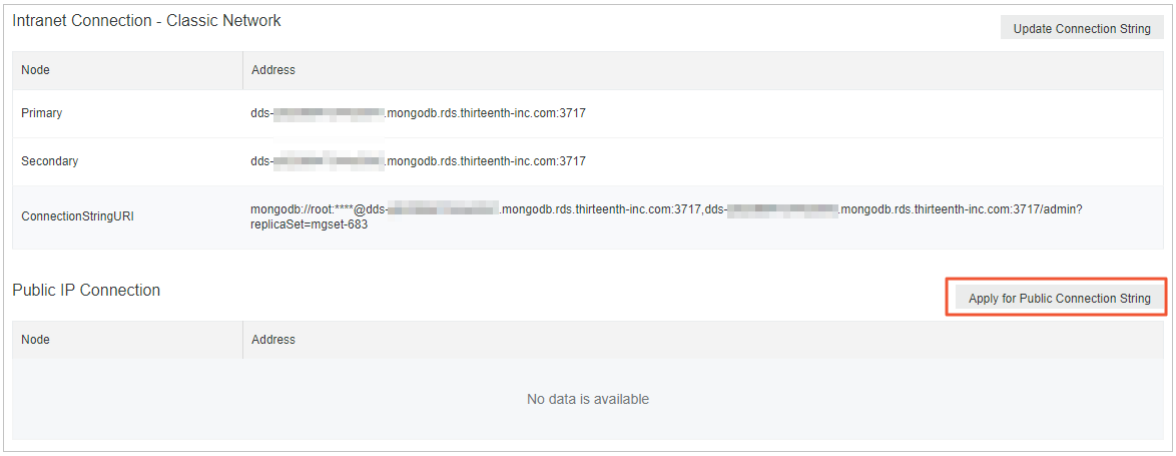
Context

The following table describes the Virtual Private Cloud (VPC) and classic network endpoints supported by ApsaraDB for MongoDB.

Connection address type	Description
VPC endpoint	• A VPC is an isolated network that provides higher security and performance than the classic network. • By default, ApsaraDB for MongoDB provides VPC endpoints for an ApsaraDB for MongoDB instance to ensure high security and high performance.
Classic network endpoint	Cloud services on the classic network are not isolated. Unauthorized access can be blocked only by using security groups or whitelists.
Public endpoint	• Your ApsaraDB for MongoDB instance is at risk when you connect to it over the Internet. For this reason, ApsaraDB for MongoDB does not provide public endpoints by default. • If you want to connect to an ApsaraDB for MongoDB instance from a device outside of Apsara Stack (such as an on-premises device), you must apply for a public endpoint.

Apply for a public endpoint for a replica set instance

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. On the **Replica Set Instances** page, click the ID of an instance.
3. In the left-side navigation pane, click **Database Connections**.
4. In the **Public Connections** section, click **Apply for Public Connection String**.



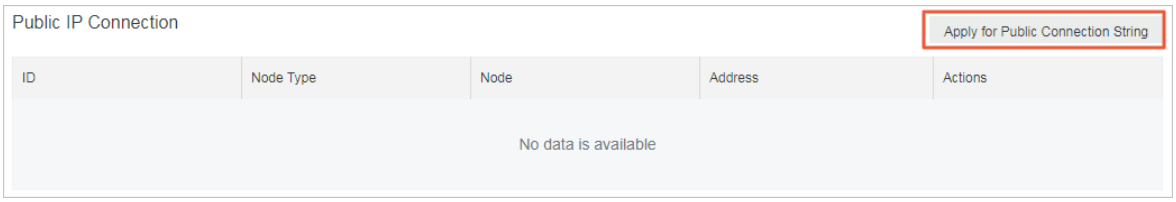
5. In the **Apply for Public Connection String** message, click **OK**.

> **Note** If you want to connect to a replica set instance by using a public endpoint, you must add the public IP address of your client to a whitelist of this instance. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

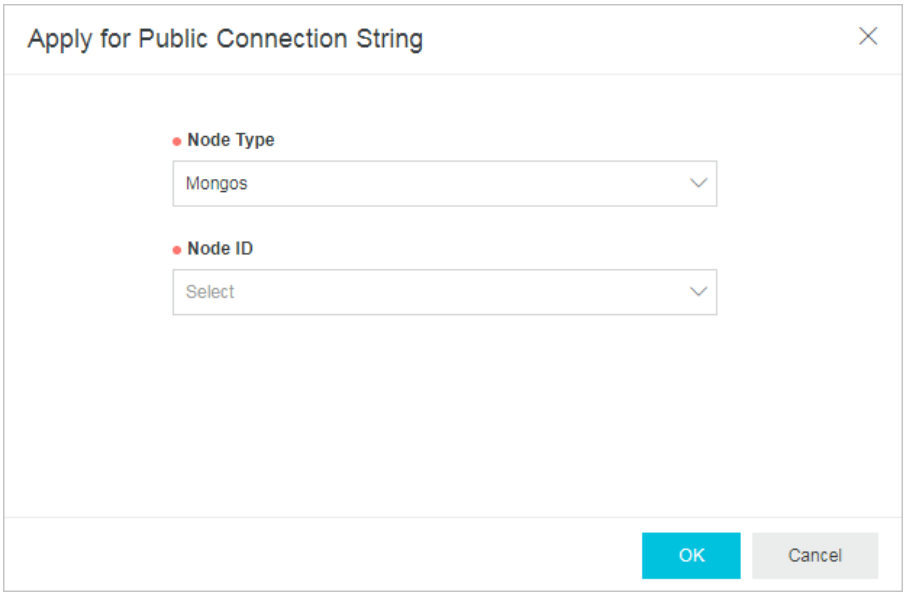
After the application is complete, the replica set instance generates new endpoints for both the primary and secondary nodes and the corresponding connection string URI. For more information, see Overview of replica set instance connections.

Apply for a public endpoint for a sharded cluster instance

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections**.
5. In the **Public Connections** section, click **Apply for Public Connection String**.



6. In the **Apply for Public Connection String** panel, select a node ID from the **Node ID** drop-down list and click **OK**.



7. (Optional)If you want to apply for public endpoints for multiple nodes in a sharded cluster instance, repeat the preceding steps.

> **Note** To apply for a public endpoint for another node in the instance, you must wait until the state of the instance becomes **Running**.

References

- To ensure data security, we recommend that you release a public endpoint if you no longer need it. For more information, see Release a public connection string.
- Before you connect to a database over the Internet, we recommend that you enable SSL encryption. For more information, see Use the mongo shell to connect to an ApsaraDB for MongoDB database in SSL encryption mode.

15.1.6.5. Release a public endpoint

To ensure data security, you can release public endpoints that are no longer needed in the ApsaraDB for MongoDB console.

Precautions

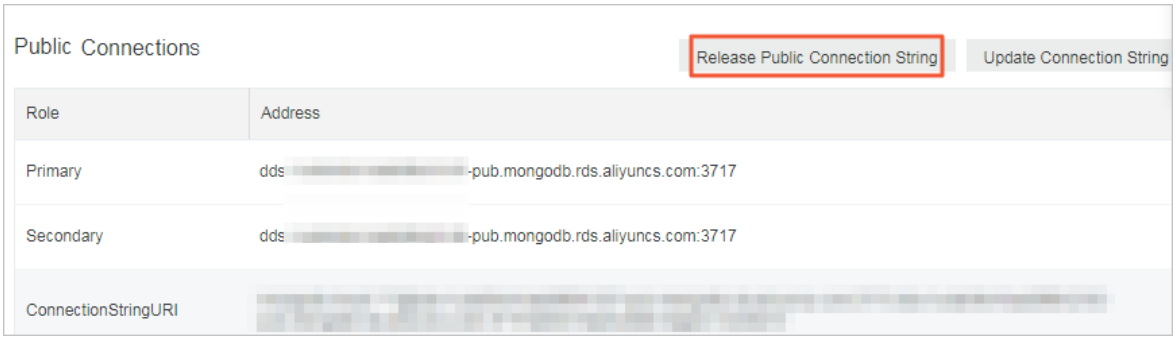
- You can release one or more public endpoints of the mongos nodes for a sharded cluster instance.
- After a public endpoint is released for an instance or node, you cannot connect to the instance or node by using the public endpoint.
- After the public endpoint is released, we recommend that you delete the corresponding public IP address from the whitelist to ensure data security. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

Release a public endpoint for a replica set instance

> **Note** After the public endpoint of a replica set instance is released, the public endpoints of the primary and secondary nodes are released.

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. On the **Replica Set Instances** page, click the ID of an instance.
3. In the left-side navigation pane, click **Database Connections**.

4. In the **Public Connections** section, click **Release Public Connection String**.



5. In the **Release Public Connection String** message, click **OK**.

Release a public endpoint for a sharded cluster instance

> **Note**
> - You can release one or more public endpoints of the mongos, shard, and Configserver nodes for a sharded cluster instance.
> - After the public endpoint of a shard or Configserver node is released, the public endpoints of the primary and secondary nodes are released.

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections**.
5. In the **Public Connections** section, find the mongos node for which you want to release the public endpoint.
6. Click **Release** in the **Actions** column.

> **Note** You can repeat this step to release the public endpoints of other nodes as needed. To release the public endpoint of another node in the instance, you must wait until the public endpoint of the current node is released or the state of the current instance becomes **Running**.

7. In the **Release Public Connection String** message, click **OK**.
8. (Optional)You can repeat this step to release the public endpoints of multiple nodes in a sharded cluster instance.

> **Note** To release the public endpoint of another node in the instance, you must wait until the state of the instance becomes **Running**.

15.1.6.6. Overview of replica set instance connections

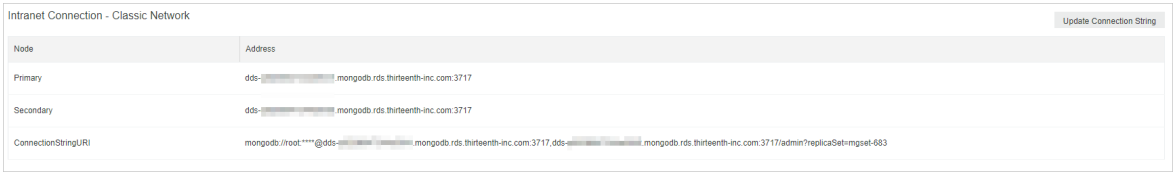
ApsaraDB for MongoDB supports both connection strings and connection string URIs. You can use a connection string to connect to the primary or secondary node, and use a connection string URI to connect to both of them. For high availability, we recommend that you use connection string URIs to connect your application to both primary and secondary nodes. This topic provides an overview of replica set instance connections.

Prerequisites

A whitelist is configured for the replica set instance. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

View connection strings

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections** to view connection strings.



Description of connection strings

Item	Description
Connection address type	• Classic network endpoint: Classic network endpoints are used for communication over the classic network. In the classic network, Apsara Stack services are not isolated. To block unauthorized traffic, you must configure security groups or IP address whitelists. • VPC endpoint: Virtual private cloud (VPC) endpoints are used for communication over VPCs. A VPC is an isolated network that provides higher security and higher performance than the classic network. By default, ApsaraDB for MongoDB provides VPC endpoints for ApsaraDB for MongoDB instances to ensure high security and high performance.
Role	• Primary: the primary node in the replica set instance. If you connect to this node, you can perform read and write operations on the databases of the replica set instance. • Secondary: the secondary node in the replica set instance. If you connect to this node, you can perform only read operations on the databases of the replica set instance. • Connection String URI: ApsaraDB for MongoDB allows you to use a connection string URI to connect to a replica set instance to achieve load balancing and high availability.
Connection string	The connection string of a primary or secondary node is in the following format: `<host>:<port>` • <host>: the endpoint used to connect to the replica set instance. • <port>: the port number used to connect to the replica set instance.

Item	Description
Connection string URI	A connection string URI is in the following format: `mongodb://[username:password@]host1[:port1][,host2[:port2],...[,hostN[:portN]]][/[database][?options]]` • mongodb://: the prefix of the connection string URI. • username:password@: the username and password used to log on to the replica set instance. You must separate them with a colon (:). • hostX:portX: the endpoint and port number used to connect to the replica set instance. • /database: the name of the database corresponding to the username if authentication is enabled. • ?options: the additional options that are used to connect to the replica set instance. **Note** If your application is in a production environment, we recommend that you use a connection string URI to connect to the instance. This way, when a node fails, the read and write operations of your application are not affected as a result of the failover.

Related information

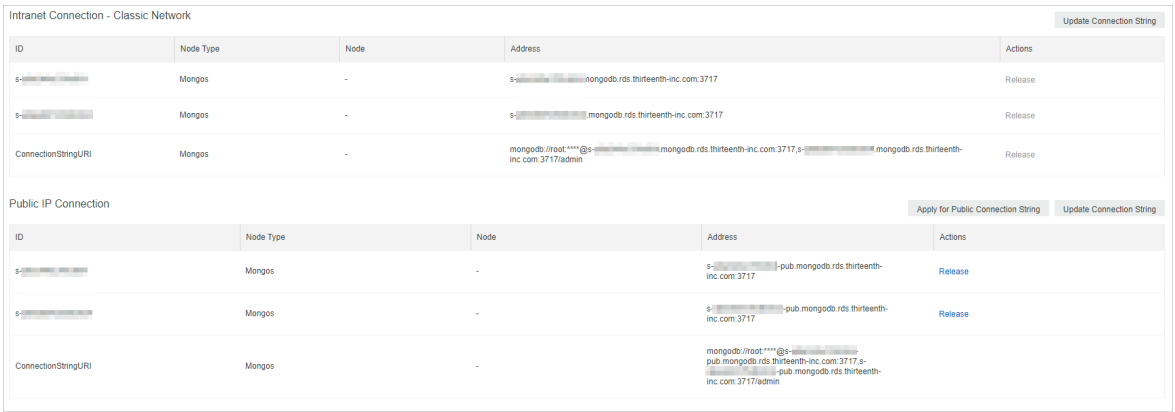
- Connect to a replica set instance by using the mongo shell

15.1.6.7. Overview of sharded cluster instance connections

ApsaraDB for MongoDB supports both connection strings and connection string URIs. You can use a connection string to connect to a single mongos node, and use a connection string URI to connect to multiple mongos nodes. For high availability, we recommend that you use connection string URIs to connect your application to multiple mongos nodes. This topic provides an overview of sharded cluster instance connections.

View connection strings

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, click **Database Connections** to view connection strings.



Description of connection strings

Item	Description
Connection address type	• Classic network endpoint: Classic network endpoints are used for communication over the classic network. In the classic network, Apsara Stack services are not isolated. To block unauthorized traffic, you must configure security groups or IP address whitelists. • VPC endpoint: Virtual private cloud (VPC) endpoints are used for communication over VPCs. A VPC is an isolated network that provides higher security and higher performance than the classic network. By default, ApsaraDB for MongoDB provides VPC endpoints for ApsaraDB for MongoDB instances to ensure high security and high performance. • Public endpoint: Public endpoints are used for communication over the Internet. If you connect to an ApsaraDB for MongoDB instance over the Internet, the instance may be exposed to security risks. By default, ApsaraDB for MongoDB does not provide public endpoints for ApsaraDB for MongoDB instances. If you want to connect to an ApsaraDB for MongoDB instance from a device outside of Apsara Stack (such as an on-premises device), you must apply for a public endpoint. For more information, see Apply for a public endpoint for an ApsaraDB for MongoDB instance.
Mongos node ID	The connection string of a mongos node is in the following format: `<host>:<port>` • <host>: the endpoint used to connect to the sharded cluster instance. • <port>: the port number used to connect to the sharded cluster instance. **Note** During routine tests, you can use a connection string to directly connect to a mongos node.
Connection string URI	A connection string URI is in the following format: `mongodb://[username:password@]host1[:port1][,host2[:port2],...[,hostN[:portN]]][/[database][?options]]` • mongodb://: the prefix of the connection string URI. • username:password@: the username and password used to log on to the sharded cluster instance. You must separate them with a colon (:). • hostX:portX: the endpoint and port number used to connect to the sharded cluster instance. • /database: the name of the database corresponding to the username if authentication is enabled. • ?options: the additional options that are used to connect to the sharded cluster instance. **Note** If your application is in a production environment, we recommend that you use a connection string URI to connect to the sharded cluster instance. Then, your client can automatically distribute your requests to multiple mongos nodes to balance loads. If a mongos node fails, your client automatically redirects requests to other mongos nodes in the normal state.

15.1.7. Data security

15.1.7.1. Configure a whitelist for an ApsaraDB for MongoDB instance

This topic describes how to configure a whitelist for an ApsaraDB for MongoDB instance. Before you use an ApsaraDB for MongoDB instance, you must add the IP addresses or CIDR blocks that you use for database access to a whitelist of this instance. This improves database security and stability. Proper configuration of whitelists can enhance access security of ApsaraDB for MongoDB. We recommend that you maintain the whitelists on a regular basis.

Context

- The system creates a default whitelist for each instance. This whitelist can be modified or cleared but cannot be deleted.
- After an ApsaraDB for MongoDB instance is created, the system automatically adds the IP address 127.0.0.1 to the **default** whitelist of this instance. The IP address 127.0.0.1 indicates that no IP addresses are allowed to access this instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > Whitelist Settings**.
5. On the page that appears, use one of the following methods to add IP addresses to a whitelist:
   - Manually modify a whitelist
     a. Click ⋮ in the **Actions** column corresponding to a whitelist and select **Manually Modify**.
     b. In the Manually Modify panel, enter IP addresses or CIDR blocks in the **IP White List** field.

     > **Note**
     > - Separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added.
     >
     >   Supported formats are 0.0.0.0/0, IP addresses such as 10.23.12.24, or CIDR blocks such as 10.23.12.24/24. /24 indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits.
     > - If the IP address whitelist is empty or contains only `0.0.0.0/0`, all devices are granted access. This poses risks to your ApsaraDB for MongoDB instance. We recommend that you add only the IP addresses or CIDR blocks of your own web servers to the whitelist.

     c. Click **OK**.
   - Load internal IP addresses of ECS instances
     a. Click ⋮ in the **Actions** column corresponding to a whitelist and select **Import ECS Intranet IP**.
     b. In the **Import ECS Intranet IP** panel, select the IP addresses that you want to add to the IP address whitelist and click > to add these IP addresses to the whitelist.
     c. Click **OK**.

15.1.7.2. Create or delete a whitelist

This topic describes how to create or delete whitelists. Whitelists consist of the IP addresses allowed to access specific databases.

Context

If your business involves multiple applications and you need to add a whitelist for each of them, you can sort the IP addresses of the applications into different whitelists.

Create a whitelist

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > Whitelist Settings**.
5. Click **Create Whitelist**.
6. In the **Create Whitelist** panel, set **Group Name** and **IP White List** and click **OK**.

Parameter	Description
Group Name	The name of the whitelist. The name must comply with the following rules: ○ The name must start with a lowercase letter. ○ The name must end with a lowercase letter or digit. ○ The name must be 2 to 32 characters in length.
IP White List	The IP addresses or CIDR blocks that you want to add to the whitelist. **Note** ○ Separate multiple IP addresses with commas (,). A maximum of 1,000 different IP addresses can be added to a whitelist. Supported formats are 0.0.0.0/0, IP addresses such as 10.23.12.24, or CIDR blocks such as 10.23.12.24/24. In the preceding example, /24 indicates the length of the IP address prefix in the CIDR block. An IP address prefix can contain 1 to 32 bits. ○ If the IP address whitelist is empty or contains only `0.0.0.0/0`, all devices are granted access. This poses risks to your ApsaraDB for MongoDB instance. We recommend that you add only the IP addresses or CIDR blocks of your own web servers to the whitelist.

Delete a whitelist

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > Whitelist Settings**.
5. Click ⋮ in the **Actions** column corresponding to the whitelist that you want to delete, and then select **Delete Whitelist Group**.

Note You cannot delete the **default** whitelist.

6. In the **Delete Whitelist Group** message, click **OK**.

15.1.7.3. Audit logs

This topic describes audit logs provided in the ApsaraDB for MongoDB console. You can query the statement execution logs, operation logs, and error logs of an ApsaraDB for MongoDB instance to identify and analyze faults.

Context

The audit log feature records all operations that a client performs on a connected database. This feature provides references for you to perform fault analysis, behavior analysis, and security auditing because you can obtain the operation execution details from the audit logs. Audit logs are essential in the regulatory operations of Finance Cloud and other core business scenarios.

> **Note** Audit logs are stored for seven days, after which they are deleted.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > Audit Logs**.
5. In the upper-left corner of the page, click **Enable Audit Log**.
6. In the **Enable Audit** message, click **OK**.

Result

After the audit log feature is enabled, specify the time range, database name, database user, and keyword to query audit logs. You can also use the following options:

- **Export File**: exports an audit log file.
- **File List**: displays a list of audit log files.
- **Disable Audit Log**: stops the collection of information on database operations and deletes the saved audit logs.

15.1.7.4. Configure SSL encryption for an ApsaraDB for MongoDB instance

To enhance link security, you can enable SSL encryption and install SSL certification authority (CA) certificates on your application services. SSL encryption can encrypt network connections at the transport layer to improve data security and ensure data integrity. This topic describes operations related to SSL encryption.

Prerequisites

- The instance is a replica set instance.
- The MongoDB version of the instance is 3.4, 4.0, or 4.2.

Notes

When you enable or disable SSL encryption or update SSL CA certificates for an instance, the instance is restarted. Plan your operations in advance and make sure that your applications can automatically re-establish a connection.

> **Note** When an instance is restarted, all its nodes are restarted in turn and each node goes through a transient connection of about 30 seconds. If the instance contains more than 10,000 collections, the transient connections last longer.

Precautions

- You can download SSL CA certificate files only from the ApsaraDB for MongoDB console.
- After you enable SSL encryption for an instance, the CPU utilization of the instance is significantly increased. We recommend that you enable SSL encryption only when encryption needs arise. For example, you can enable SSL encryption when you connect to an ApsaraDB for MongoDB instance over the Internet.

  > **Note** Internal network connections are more secure than Internet connections and do not need SSL encryption.

- After you enable SSL encryption for an instance, both SSL and non-SSL connections are supported.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > SSL**.
5. Perform the corresponding operations based on your needs.

   > **Note** When you enable or disable SSL encryption or update SSL CA certificates for an instance, the instance is restarted. Plan your operations in advance and make sure that your applications can automatically re-establish a connection.

Operation	Prerequisite	Procedure
Enable SSL encryption	The SSL encryption state is **Disabled**.	i. Turn on **SSL Status**. ii. In the **Restart Instance** message, click **OK**.
Update an SSL CA certificate	The SSL encryption state is **Enabled**.	i. Click **Update Certificate**. ii. In the **Restart Instance** message, click **OK**.
Download an SSL CA certificate file	The SSL encryption state is **Enabled**.	Click **Download Certificate** to download an SSL CA certificate file to your computer.
Disable SSL encryption	The SSL encryption state is **Enabled**.	i. Turn off **SSL Status**. ii. In the **Restart Instance** message, click **OK**.

15.1.7.5. Configure TDE for an ApsaraDB for MongoDB instance

This topic describes how to configure Transparent Data Encryption (TDE) for an ApsaraDB for MongoDB instance. Before data files are written to disks, TDE encrypts the data files. When data files are loaded from disks to the memory, TDE decrypts the data files. TDE does not increase the sizes of data files. When you use TDE, you do not need to modify your application that uses the ApsaraDB for MongoDB instance. To enhance data security, you can enable the TDE feature for an instance in the ApsaraDB for MongoDB console.

Prerequisites

The MongoDB version of the instance is 4.0 or 4.2.

> **Note** Before you enable TDE, you can create a MongoDB 4.0 or 4.2 instance to test the compatibility between your application and the database version. You can release the instance after the test is complete.

Notes

- When you enable TDE, your instance is restarted, and your application is disconnected from the instance. We recommend that you enable TDE during off-peak hours and make sure that your application can reconnect to the instance after it is disconnected.
- TDE increases the CPU utilization of your instance.

Precautions

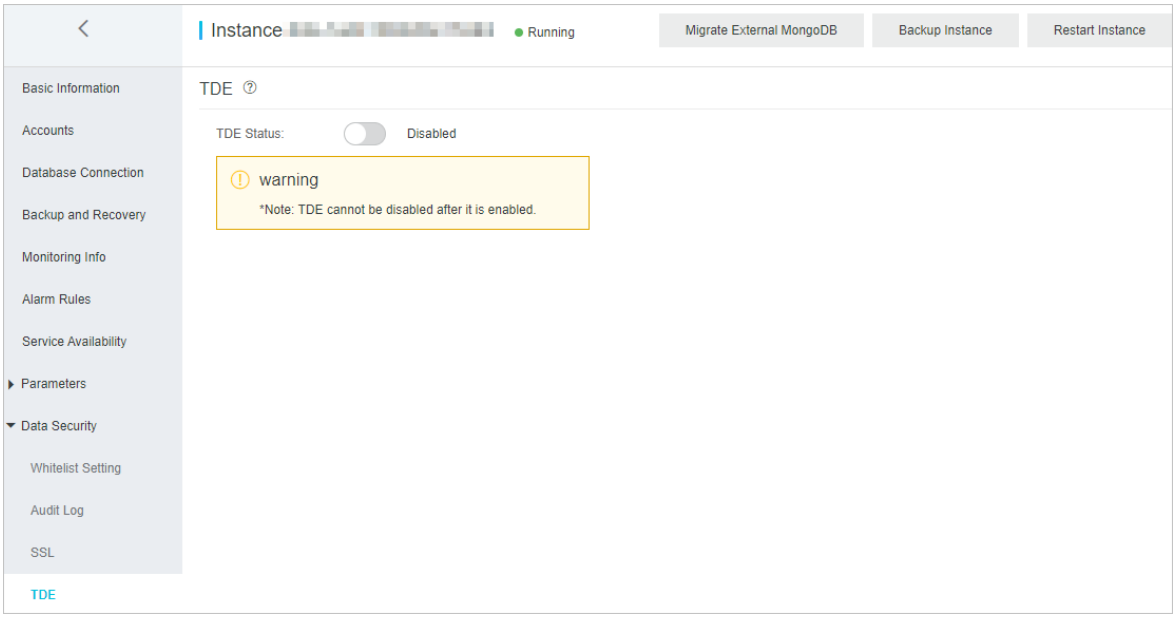
- You cannot disable TDE after it is enabled.
- You can enable TDE for an instance and disable encryption for a collection.

  > **Note** In special business scenarios, you can choose not to encrypt a collection when you create it. For more information, see Disable encryption for a specified collection.

- After you enable TDE, only new collections are encrypted. Existing collections are not encrypted.
- Key Management Service (KMS) generates and manages the keys used by TDE. ApsaraDB for MongoDB does not provide keys or certificates required for encryption.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **Data Security > TDE**.
5. On the **TDE** page, turn on **TDE Status**.



6. In the **Restart Instance** message, click **OK**.

The instance state changes to **Modifying TDE**. After the state changes to **Running**, TDE is enabled.

Disable encryption for a specified collection

After you enable TDE, all new collections are encrypted. When you create a collection, you can perform the following steps to disable encryption for the collection:

1. Connect to a replica set instance by using the mongo shell. For more information, see Connect to a replica set instance by using the mongo shell.
2. Run the following command to create a collection with encryption disabled:

```
db.createCollection("<collection_name>",{ storageEngine: { wiredTiger: { configString: "encryption=(name=none)" } } })
```

> **Note** <collection_name>: the name of the collection.

Example:

```
db.createCollection("customer",{ storageEngine: { wiredTiger: { configString: "encryption=(name=none)" } } })
```

15.1.7.6. Use the mongo shell to connect to an ApsaraDB for MongoDB database in SSL encryption mode

This topic describes how to use the mongo shell to connect to an ApsaraDB for MongoDB database in SSL encryption mode. SSL encryption can encrypt network connections at the transport layer to improve data security and ensure data integrity.

Prerequisites

- The instance is a replica set instance, and the database version of the instance is 3.4, 4.0, or 4.2.
- The mongo shell of the required version is installed on the on-premises server or Elastic Compute Service (ECS) instance from which you want to connect to the database. For more information about the installation procedure, see Install MongoDB.
- SSL encryption is enabled for the instance. For more information, see Configure SSL encryption.
- The IP address of the on-premises server or ECS instance is added to a whitelist of the ApsaraDB for MongoDB instance. For more information, see Configure a whitelist for an ApsaraDB for MongoDB instance.

Precautions

After you enable SSL encryption for an instance, the CPU utilization of the instance is significantly increased. We recommend that you enable SSL encryption only when encryption needs arise.

Procedure

An on-premises server with a Linux operating system is used in the following example.

1. Download an SSL CA certificate package. For more information, see Configure SSL encryption.
2. Decompress the package and upload the certificate files to the on-premises server or ECS instance where the mongo shell is installed.

> **Note** In this example, the *.pem* file is uploaded to the */root/sslcafile/* directory of the on-premises server.

3. On the on-premises server or ECS instance that has the mongo shell installed, run the following command to connect to a database of the ApsaraDB for MongoDB instance:

```
mongo --host <host> -u <username> -p --authenticationDatabase <database> --ssl --sslCAFile <sslCAFile_path> --sslAllowInvalidHostnames
```

> **Note**
> - If you want to connect to a database of the ApsaraDB for MongoDB instance over an internal network, make sure that the ApsaraDB for MongoDB instance has the same network type as the ECS instance. If the network type is VPC, make sure that the two instances reside within the same virtual private cloud (VPC).
> - <host>: the endpoint of the primary or secondary node for a replica set instance or of the mongos node for a sharded cluster instance. For more information, see Overview of replica set instance connections or Overview of sharded cluster instance connections.
> - <username>: the username you use to log on to a database of the ApsaraDB for MongoDB instance. The initial username is root.
> - <database>: the name of the database corresponding to the username if authentication is enabled. If the username is root, enter admin as the database name.
> - <sslCAFile_path>: the path of the SSL CA certificate files.

Example:

```
mongo --host dds-bpxxxxxxxx-pub.mongodb.rds.aliyuncs.com:3717 -u root -p --authenticationDatabase admin --ssl --sslCAFile /root/sslcafile/ApsaraDB-CA-Chain.pem  --sslAllowInvalidHostnames
```

4. When `Enter password:` is displayed, enter the password of the database user and press the Enter key.

> **Note**
> - The password characters are not displayed when you enter the password.
> - If you forget the password of the root user, you can reset the password. For more information, see Reset the password for an ApsaraDB for MongoDB instance.

15.1.8. Zone-disaster recovery

15.1.8.1. Create a dual-zone replica set instance

This topic describes how to create a dual-zone replica set instance. ApsaraDB for MongoDB provides a zone-disaster recovery solution to ensure the reliability and availability of your replica set instance. This solution deploys the three nodes of a replica set instance across two different zones within one region. The components in these zones exchange data over an internal network. When one of the two zones becomes unavailable due to unexpected events such as a power or network failure, the high-availability (HA) system switches services over to another zone.

Deployment policies

The primary, secondary, and hidden nodes of a replica set instance are deployed in two different zones within one region.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances**.
3. On the **Replica Set Instances** page, click **Create Instance**.

4. On the **Create ApsaraDB for MongoDB Instance** page, configure parameters.

> Note
> - For the **Zone** parameter, you must select dual zones, such as the amtest17001-a and amtest17001-b zones of the cn-qingdao-env11e-MAZ1 region.
> - For more information, see Create a replica set instance.

5. Click **Submit**.

15.1.8.2. Create a dual-zone sharded cluster instance

This topic describes how to create a dual-zone sharded cluster instance. ApsaraDB for MongoDB provides a zone-disaster recovery solution to ensure the reliability and availability of your sharded cluster instance. This solution deploys the components of a sharded cluster instance across two different zones within one region. The components in these zones exchange data over an internal network. When one zone becomes unavailable due to unexpected events such as a power or network failure, the high availability (HA) system automatically switches business over to the other zone.

Deployment policies

The components of a sharded cluster instance are deployed across two different zones within one region.

- Mongos nodes are evenly deployed across all data centers. At least two mongos nodes are deployed at a time, with each in one zone. Each new mongos node added later is deployed to one of the zones in turn.
- The primary, secondary, and hidden nodes in each shard node are not deployed to the two zones in sequence. The deployment of these nodes may change when manual switchover or HA failover between primary and secondary nodes is triggered.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Sharded Cluster Instances**.
3. On the **Sharded Cluster Instances** page, click **Create Instance**.
4. On the **Create ApsaraDB for MongoDB Sharded Cluster Instance** page, configure parameters.

> Note
> - For the **Zone** parameter, you must select dual zones, such as the amtest17001-a and amtest17001-b zones of the cn-qingdao-env11e-MAZ1 region.
> - For more information, see Create a sharded cluster instance.

5. Click **Submit**.

15.1.9. CloudDBA

15.1.9.1. Performance trends

This topic describes how to view performance trends in specific ranges, compare performance trends, and customize charts to view performance trends on an ApsaraDB for MongoDB instance.

Go to the Performance page

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.

2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **CloudDBA > Performance Trends**.

> **Note** For more information about performance trends, see *Performance trends* in *Database Autonomy Service User Guide*.

15.1.9.2. Real-time performance

This topic describes how to view real-time monitoring statistics of an ApsaraDB for MongoDB instance, such as read/write latency, queries per second (QPS), operations, connections, and network traffic.

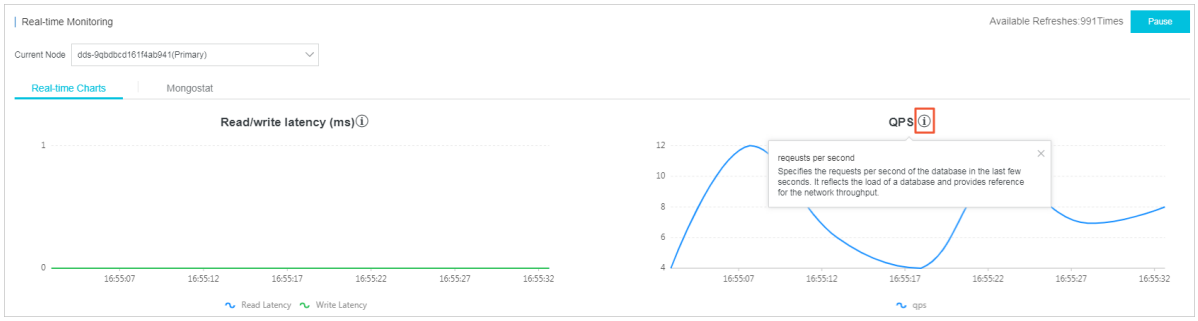
Go to the Real-time Performance page

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **CloudDBA > Real-time Performance**.

Overview of the Real-time Monitoring page

On the Real-time Monitoring page, you can click the Real-time Charts or Mongostat tab to view monitoring statistics. When you refresh the **Real-time Monitoring** page, the information on the Real-time Charts and Mongostat tabs is refreshed, and **Available Refreshes** is reset in the upper-right corner.

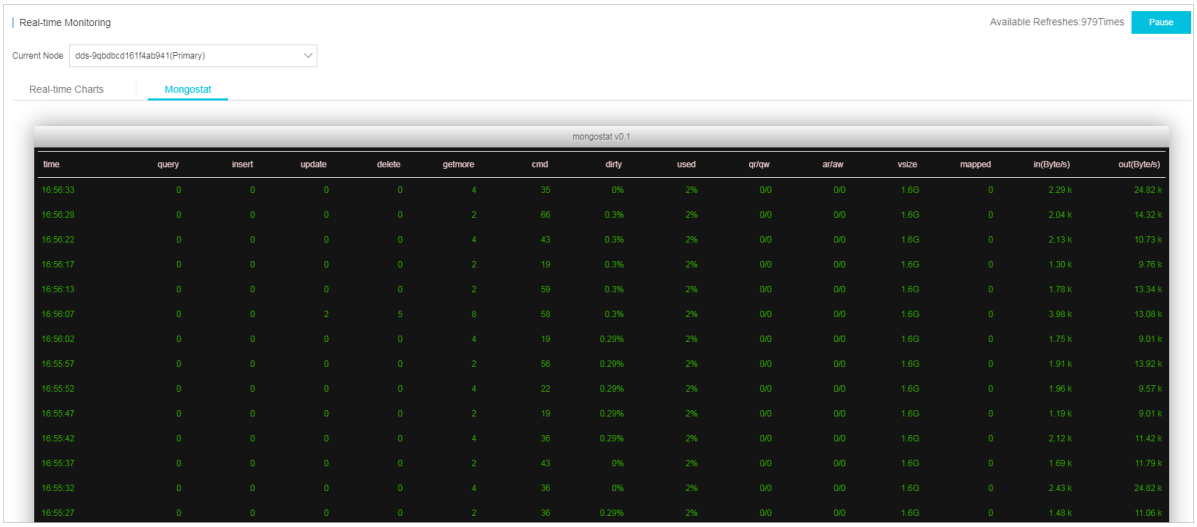
- Real-time Charts



By default, content on the **Real-time Charts** tab is displayed when you go to the Real-time Monitoring page. Line charts on the tab are refreshed every 5 seconds to provide up-to-date performance trends.

> **Note** You can move the pointer over ⓘ to view the detailed information of performance parameters.

- mongostat



Click the **Mongostat** tab. On the tab, you can view Mongostat command outputs. A new line of monitoring data is added every 5 seconds. The tab can contain up to 999 lines of information.

> **Note** For more information about Mongostat command outputs, see MongoDB official documentation.

15.1.9.3. Instance sessions

This topic describes how to view real-time monitoring statistics of an ApsaraDB for MongoDB instance, such as read/write latency, queries per second (QPS), operations, connections, and network traffic.

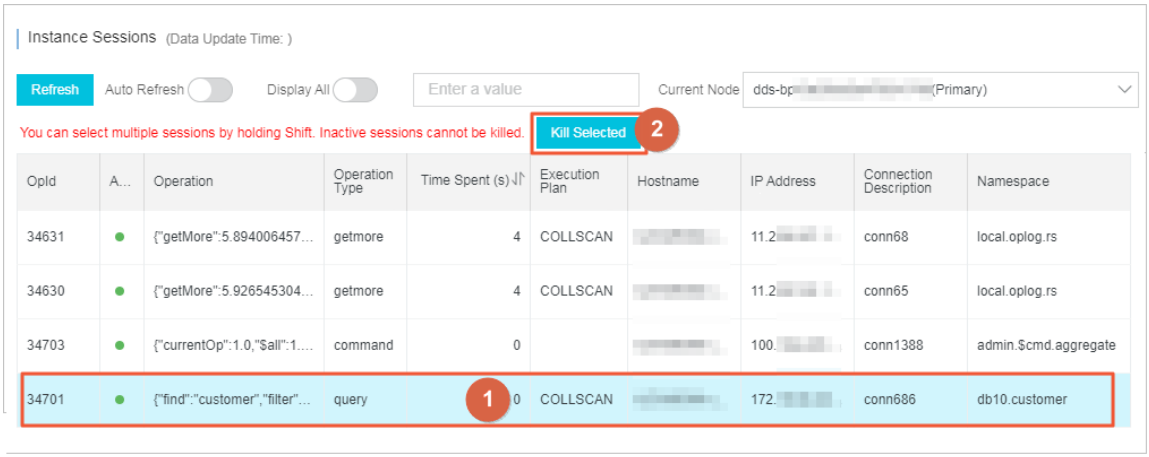
View instance sessions

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **CloudDBA > Sessions**.
   - If you turn on **Auto Refresh**, the system updates session data on the page every 5 seconds.
   - By default, the system displays only active sessions. You can turn on **Display All** to view both active and inactive sessions.
   - In the **Session Statistics** section, you can view information about sessions in the **Overview**, **Statistics by Client**, and **Statistics by Namespace** charts.

Terminate instance sessions

> **Warning** To avoid unexpected results, we recommend that you do not terminate system-level sessions.

1. In the **Instance Sessions** section, select the session that you want to terminate.



2. Click **Kill Selected**.
3. In the message that appears, click **OK**.

15.1.9.4. Storage analysis

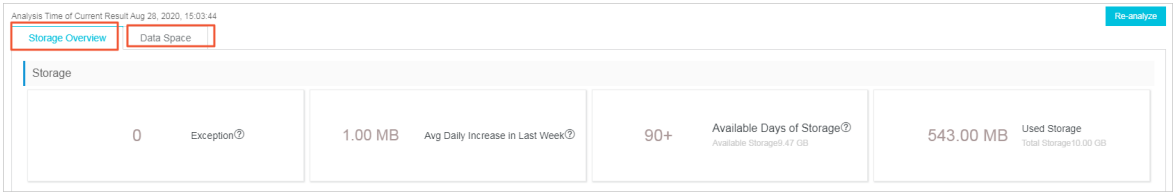
This topic describes how to view information about the storage analysis feature, including **Storage Overview**, **Exceptions**, **Storage Trend**, **Tablespaces**, and **Data Space**. The information helps you identify and resolve exceptions in the database storage to ensure database stability.

Prerequisites

The database version of the ApsaraDB for MongoDB instance is 4.0 or 4.2.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **CloudDBA > Storage Analysis**.
5. In the upper-right corner, click **Re-analyze**. Then, wait until the analysis is complete.
6. Click the **Storage Overview** or **Data Space** tab to view analysis results.

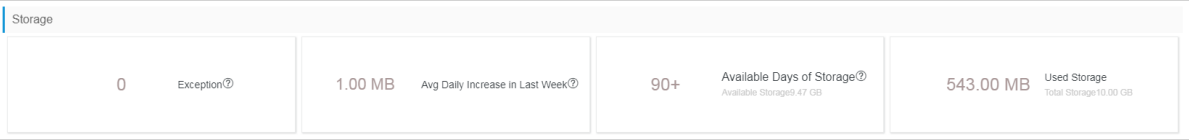


- For more information about Storage Overview, see Storage Overview.
- For more information about Data Space, see Data Space.

Storage Overview

This tab contains four sections: **Storage**, **Exceptions**, **Storage Trend**, and **Tablespaces**.

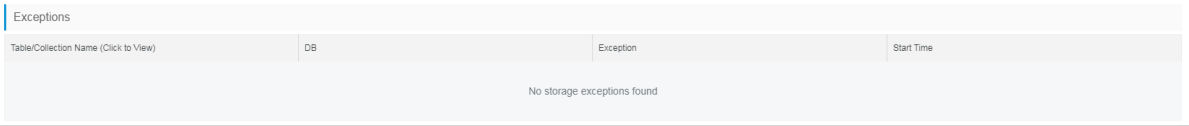
- **Storage** section



Parameter	Description
Exception	The number of detected storage exceptions in the instance. ApsaraDB for MongoDB can detect the following types of exceptions: ◦ More than 90% of the storage capacity is used. ◦ The available physical storage capacity will be exhausted within seven days. ◦ A single collection contains more than 10 indexes.
Avg Daily Increase in Last Week	The average daily increase of storage usage in the instance over the last seven days. Formula: (Size of available storage capacity at the time of collection - Size of available storage capacity seven days ago)/7. **Note** ◦ The increase is the average daily increase over the last seven days at the time of collection. ◦ This parameter is suitable for scenarios in which the traffic volume remains stable. The value of this parameter is inaccurate in the event of abrupt storage changes that are caused by batch import, deletion of historical data, instance migration, or instance rebuilding.
Available Days of Storage	The number of days during which storage capacity is available in the instance. Formula: Size of available storage capacity/Average daily increase over the last week. **Note** ◦ 90+ indicates that the storage capacity is sufficient for more than 90 days of usage. ◦ This parameter is suitable for scenarios in which the traffic volume remains stable. The value of this parameter is inaccurate in the event of abrupt changes in storage caused by batch import, deletion of historical data, instance migration, or instance rebuilding.
Used Storage	The size of used storage capacity in contrast to the total size of storage capacity.

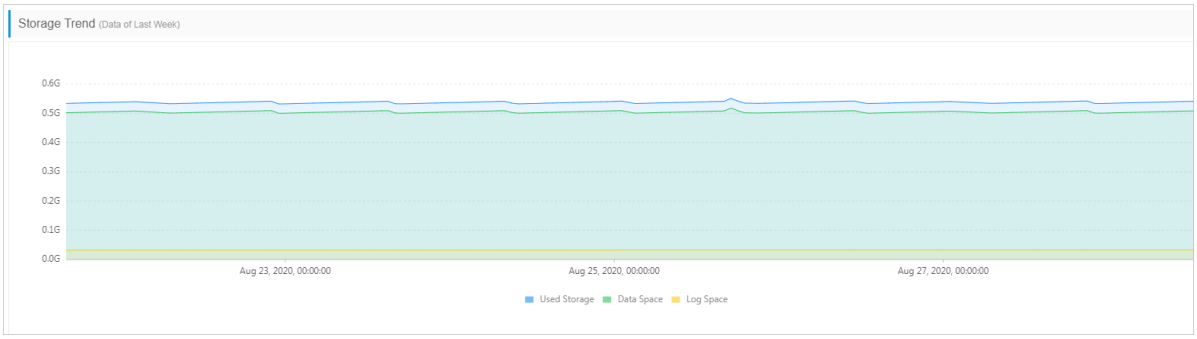
- **Exceptions** section

  Information about detected storage exceptions. You can resolve the exceptions based on the information in this section.



- **Storage Trend** section

Changes in storage usage over the last week, such as changes in the **used storage**, **data space**, and **log space**.



- **Tablespaces** section

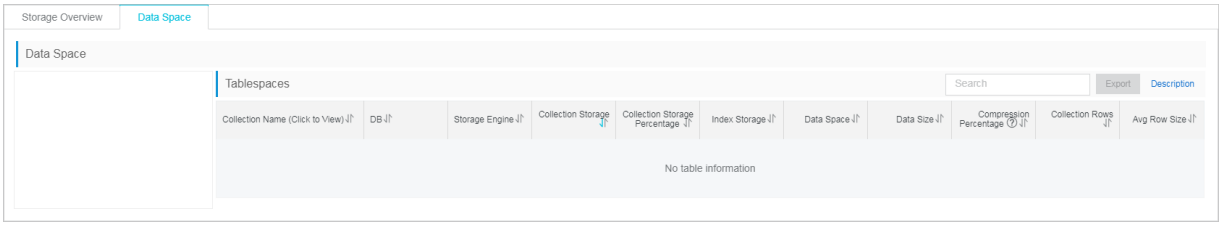
  Information about all tables, such as the database name, storage engine, and collection storage.

> **Note** You can click the name of a collection to view its indexes.

Data Space

The Data Space tab shows the total storage capacity and tablespace information of each database.

> **Note**
> - You can click the name of a data space to view its tablespace information.
> - You can click the name of a collection to view its indexes.



15.1.9.5. Slow query logs

This topic describes how to view slow query logs of an ApsaraDB for MongoDB instance. You can identify, analyze, diagnose, and track slow query logs to create indexes, which improves the utilization of resources in the instance.

Procedure

1. Log on to the ApsaraDB for MongoDB console. For more information about how to log on to the console, see Log on to the ApsaraDB for MongoDB console.
2. In the left-side navigation pane, click **Replica Set Instances** or **Sharded Cluster Instances**.
3. On the **Replica Set Instances** or **Sharded Cluster Instances** page, click the ID of an instance.
4. In the left-side navigation pane, choose **CloudDBA > Slow Query Logs**.

> **Note** By default, slow query logs generated in the past 15 minutes are displayed in the trend chart. You can specify a time range and click **Search** to view slow query logs. The maximum time range is one day.

5. View details of slow query logs by using one of the following methods:

   Method 1

   i. Click the **Slow Log Details** tab in the lower part of the page.

   ii. On the **Slow Log Details** tab, select the database that you want to query.

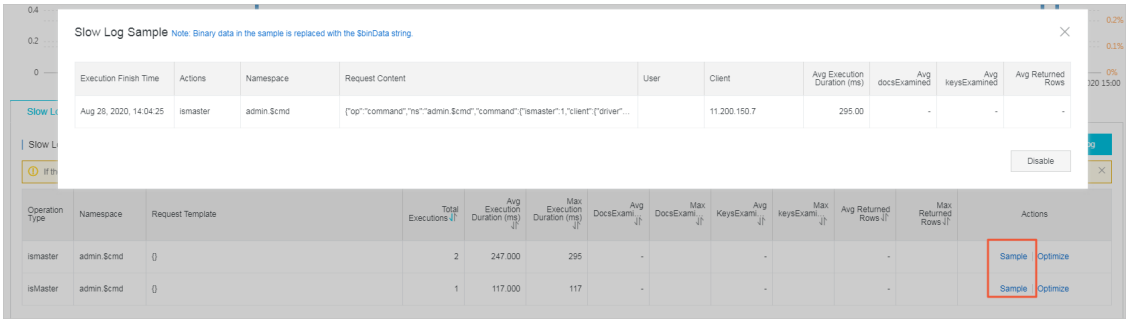
   > **Note** If the request content of the database is hidden, you can move the pointer over the corresponding content in the **Request Content** column and view the complete content.

   Method 2

   i. In the Slow Log Trend chart, click a point in time. Then, you can view the statistics of the slow query logs generated at the point in time on the **Slow Log Statistics** tab.

   ii. On the **Slow Log Statistics** tab, click **Sample** in the **Actions** column.
   In the **Slow Log Sample** dialog box, you can view details of the slow query log.

   > **Note** If the request content of the database is hidden, you can move the pointer over the corresponding content in the **Request Content** column and view the complete content.

Export slow query logs

You can click **Export Slow Log** on the **Slow Log Statistics** tab to save the slow query log information to your computer.

16.Data Management (DMS)

16.1. User Guide

16.1.1. What is DMS?

Data Management (DMS) is a fully managed service that is provided by Apsara Stack. You can use this service to manage data, table schemas, R&D processes, R&D specifications, users, permissions, and access security.

Supported databases

- Relational databases:
  - MySQL: ApsaraDB RDS for MySQL, PolarDB-X, MySQL databases from other cloud service providers, and self-managed MySQL databases
  - SQL Server: ApsaraDB RDS for SQL Server, SQL Server databases from other cloud service providers, and self-managed SQL Server databases
  - PostgreSQL: ApsaraDB RDS for PostgreSQL, PostgreSQL databases from other cloud service providers, and self-managed PostgreSQL databases
  - Self-managed Dameng (DM) databases
  - Self-managed Oracle databases
  - ApsaraDB for OceanBase and self-managed OceanBase databases
- NoSQL databases:
  - Redis: ApsaraDB for Redis, Redis databases from other cloud service providers, and self-managed Redis databases
  - MongoDB: ApsaraDB for MongoDB, MongoDB databases from other cloud service providers, and self-managed MongoDB databases
  - Graph Database (GDB)
- Online analytical processing (OLAP) databases:
  - AnalyticDB for MySQL
  - AnalyticDB for PostgreSQL

> **Note** Self-managed databases are databases that are installed on Apsara Stack Elastic Compute Service (ECS) instances, instances from other cloud service providers, or servers in data centers.

Features

- DMS provides support for the entire database development process. You can design table schemas in an on-premises environment based on the design specifications. Before you publish SQL statements to an online environment, DMS can review the add, remove, modify, and query operations in the statements. Then, you can publish schemas to the specified environment as needed.
- DMS provides fine-grained access control at the database, table, or field level. You can perform all operations on databases in the DMS console. The operations can be traced and audited.
- DMS allows you to configure operation specifications and approval processes for multiple modules based on your business requirements. These modules include the schema design, data change, data export, and permission application.
- DMS integrates database development with database interaction. You can manage databases without the need to switch between database endpoints at a high frequency by using database accounts and passwords.
- DMS provides the task orchestration feature that allows you to orchestrate and schedule SQL tasks for databases. You can use this feature to perform a variety of operations with ease. For example, you can use this

feature to dump historical data or generate periodical reports.

16.1.2. Quick start

16.1.2.1. Log on to the DMS console

This topic uses Google Chrome to describe how to log on to the Data Management (DMS) console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Database Services > Data Management**.
6. Set the **Organization** and **Region** parameters and click **DMS**.

> **Note**
> - If you log on to the DMS console as a DMS administrator and your account is added to multiple tenants, you can move the pointer over the  icon in the upper-right corner and select **Switch tenant** to switch to another tenant.
> - By default, the previous version of the DMS console appears after the logon. For more information about how to use the new version of the DMS console, see Experience the new version of the DMS console. For more information about how to return to the previous version of the DMS console, see Return to the previous version of the DMS console.

16.1.2.2. Customize the top navigation bar

Data Management (DMS) allows you to customize the top navigation bar so that you can access the features that are commonly used in DMS with ease.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon.
3. Turn on the 

switch in the upper-right corner so that you can customize the top navigation bar.
   - on: You can customize the top navigation bar.
   - off: By default, all features are displayed. You cannot customize the top navigation bar.
4. Find a feature that you want to add to favorites and click the  icon. The feature is displayed in the top navigation bar.

> **Note** If a great many features are added to favorites, move the pointer over **More** in the top navigation bar to view the favorite features that are hided.

16.1.2.3. Register database instances with DMS

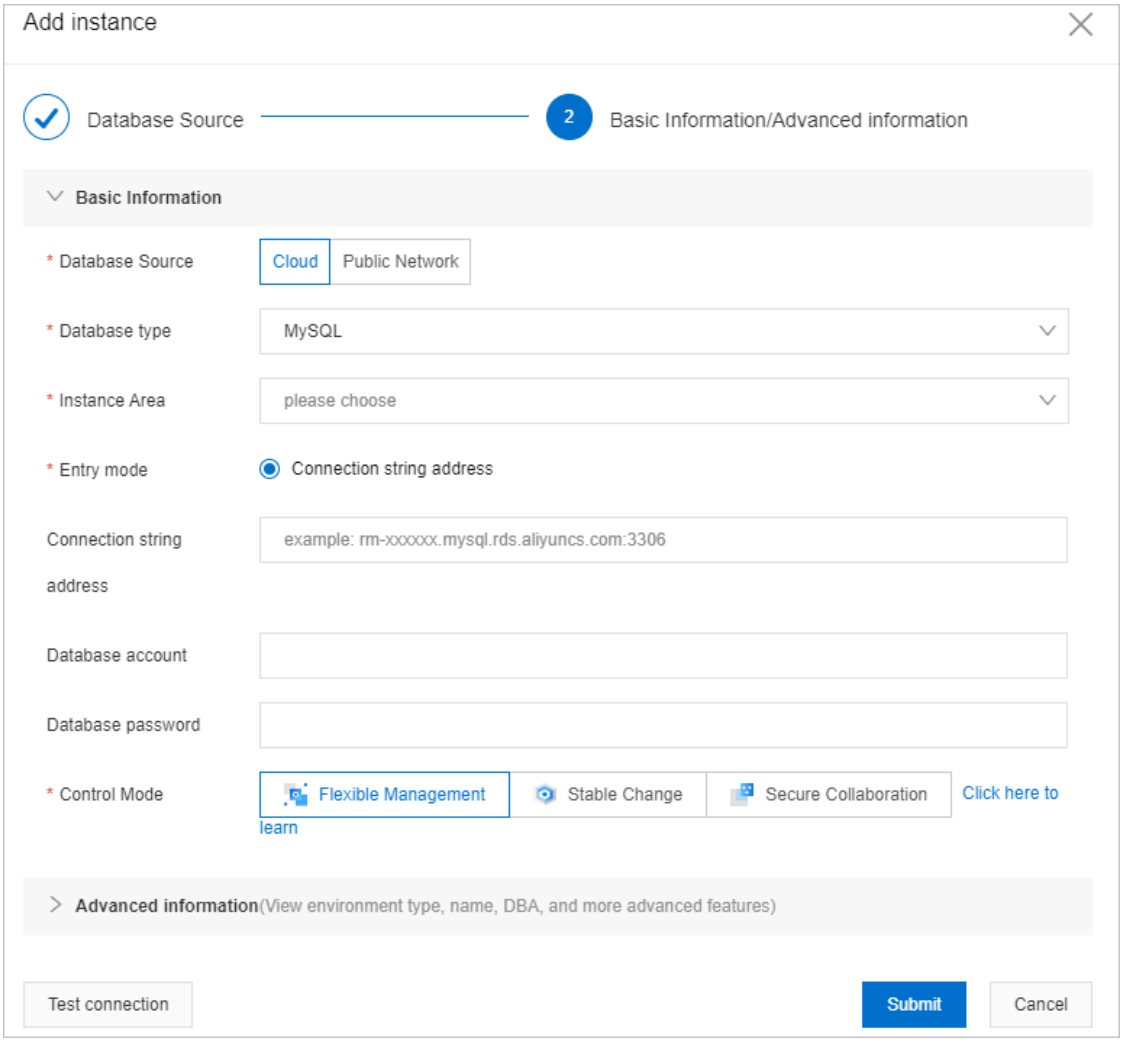
To manage database instances in Data Management (DMS), you must register the database instances with DMS. DMS allows you to register ApsaraDB instances and self-managed database instances that are hosted over the Internet. This topic describes how to register an ApsaraDB RDS for MySQL instance with DMS.

Procedure

1. Log on to the DMS console.
2. In the left-side navigation pane, click **Add Instance**.

> **Note** You can also click the icon in the upper-left corner to add an instance.

3. In the Add Instance dialog box, click the **Cloud** tab.
4. On the Cloud tab, select a database type.
5. In the Add instance dialog box, set the parameters as required. This example shows you how to register an ApsaraDB RDS for MySQL instance with DMS.



Section	Parameter	Description
Basic Information	**Data Source**	The source of the database instance. In this example, select **Cloud**.
	Database Type	The type of the database instance. In this example, select **MySQL**.
	Instance Region	The region where the database instance resides. Select a region from the drop-down list.
	Entry mode	The method that you use to log on to the database instance. Default value: **Connection string address**. This value cannot be changed.
	Connection string address	The endpoint of the database instance. The endpoint contains a port number.
	Database Account	The username that you use to log on to the database instance.
	Database password	The password that you use to log on to the database instance.

Section	Parameter	Description
	Control Mode	The control mode that is used to manage the database instance. For more information, see Control modes. **Note** If you set this parameter to **Security Collaboration**, you must set the **Security Rules** parameter.
Advanced Information	**Environment type**	The environment of the database instance.
	Instance Name	The name of the database instance.
	Enable DSQL	Specifies whether to enable the cross-database query feature. To enable the cross-database query feature, you must specify a database link name. For more information, see Cross-database query.
	Lock-free Schema Change	Specifies whether to enable the lock-free schema change feature. **Note** This parameter is displayed only for MySQL databases.
	Enable SSL	Specifies whether to allow DMS to connect to the database instance by using SSL connections. After this feature is enabled, DMS can connect to the database instance by using SSL connections. SSL encrypts network connections at the transport layer to improve the security and integrity of data in transit. However, SSL increases the response time of network connections. Before you use SSL connections, make sure that the SSL encryption feature is enabled for the database instance. Valid values: ◦ **Default (DMS automatically checks whether self-negotiation is enabled for the database instance.)**: DMS automatically checks whether the SSL encryption feature is enabled for the database instance. If the SSL encryption feature is enabled, DMS connects to the database instance by using SSL connections. Otherwise, DMS connects to the database instance without encryption. ◦ **Open**: DMS connects to the database instance by using SSL connections. This value is invalid if you disable the SSL encryption feature for the database instance. ◦ **Close**: DMS does not connect to the database instance by using SSL connections. **Note** This parameter is displayed only for MySQL databases.
	DBA	The database administrator (DBA) of the database instance. The DBA can grant permissions to users.
	query timeout(s)	The timeout period for the execution of an SQL query statement. If the execution of an SQL query statement lasts longer than the specified timeout period, the execution of the statement is terminated to protect the database.

Section	Parameter	Description
	export timeout(s)	The timeout period for the execution of an SQL export statement. If the execution of an SQL export statement lasts longer than the specified timeout period, the execution of the statement is terminated to protect the database.

6. After the parameters are set, click **Basic Information** and then **Test connection** in the lower-left corner to verify the settings.

   > **Note** If the connectivity test fails, check the specified parameter values based on the error message.

7. Click **Submit**.

Result

After the preceding steps are performed, the ApsaraDB RDS for MySQL instance is registered with DMS. You can view and manage your database instance in the left-side navigation pane of the DMS console.

16.1.2.4. Add a user

Data Management (DMS) allows you to manage users. You can add users and assign the required roles to each user based on your business requirements.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **System > User**.

   > **Note** On the User tab, you can perform the required operations on the existing users. For example, you can edit, disable, enable, or delete a user.

3. Click **New**.
4. In the Add User dialog box, set the required parameters. The following table describes the parameters.

Parameter	Description
Alibaba Cloud Account	The ID of an Apsara Stack tenant account. You can enter one of the following IDs: ○ The ID of an Apsara Stack tenant account. You can obtain the ID from the account owner. ○ The ID of a Resource Access Management (RAM) user. You can obtain the required ID from the Service-linked Roles page of the Apsara Uni-manager Management Console.

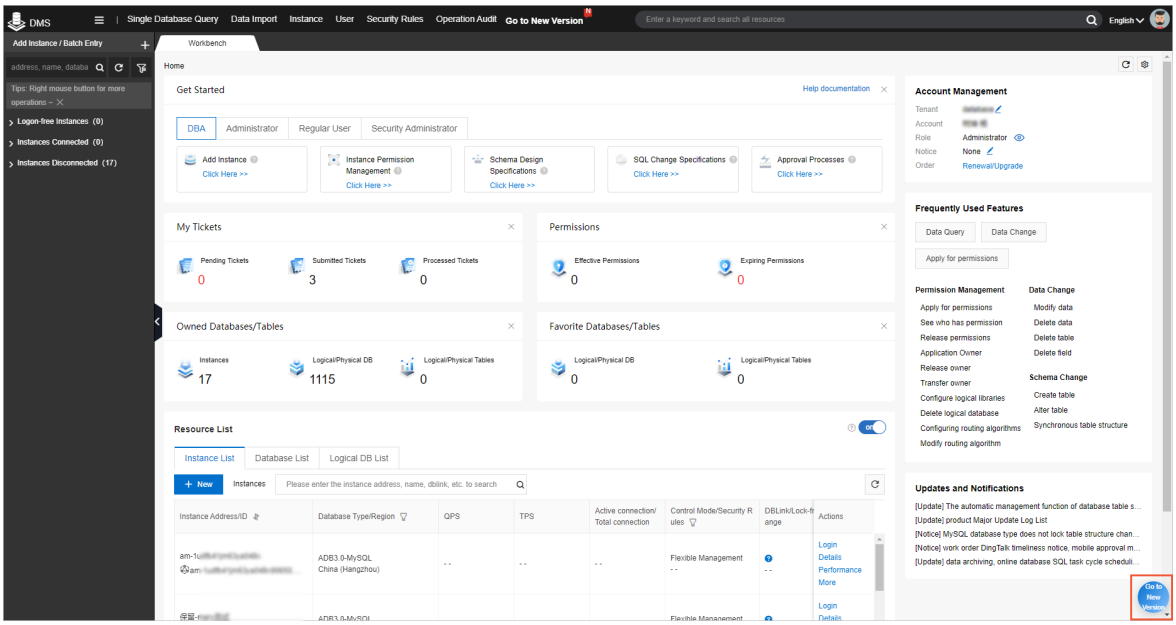
Parameter	Description
Role	The role that you want to assign to the user based on your business requirements. Valid values: ○ Regular User ○ DBA ○ Administrator ○ Security Administrator ○ Technical Support **Note** For more information about the features that are supported by each role, see Features that are supported by each role.

5. Click **OK**.

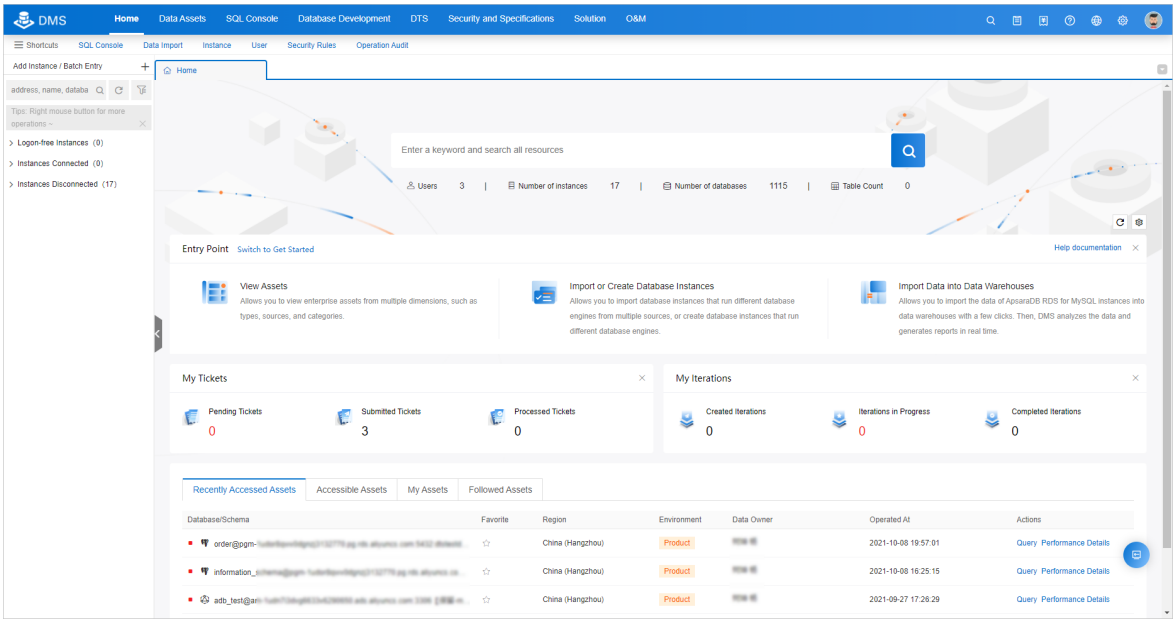
16.1.2.5. Experience the new version of the DMS console

The new version of the DMS console classifies functional modules based on scenarios, and integrates the features of Data Transmission Service (DTS). This topic describes how to go to the new version of the DMS console and return to the previous version of the DMS console.

1. Log on to the DMS console.
2. On the **Workbench** tab, click the **Go to New Version** icon in the lower-right corner.



3. In the message that appears, click **Leave** to go to the new version of the DMS console, as shown in the following figure.



> **Note** For more information about how to return to the previous version of the DMS console, see Return to the previous version of the DMS console.

Return to the previous version of the DMS console

1. Log on to the DMS console.
2. In the top navigation bar, click **Data Assets**.
3. In the left-side navigation pane, click **Home**.
4. Click the icon in the lower-right corner of the page to return to the previous version of the DMS console.

16.1.3. Control modes

DMS provides three control modes for you to manage instances: Flexible Management, Stable Change, and Security Collaboration. You can specify a control mode for each instance.

Control mode	Description	Scenario	Logon method
Flexible management	This control mode allows you to manage the visualized data and schemas of multiple types of databases. It also provides a variety of data management solutions. This simplifies the use of databases and facilitates management.	• Database instances do not require strict control. • Database instances are used by a single user.	A database account and the related password.

Control mode	Description	Scenario	Logon method
Stable change	• This control mode provides multiple solutions to ensure database reliability. These solutions allow you to change data without the need to lock the related table or schema. • All features that are included in the flexible management control mode are available.	• Database instances require a high level of availability. This ensures that these database instances function as expected for an extended period of time. • Database instances are used by a small-sized group that includes multiple users.	A database account and the related password.
Security collaboration	• This control mode provides multiple solutions to ensure data security. These solutions include fine-grained access control at the database, table, or field level and sensitive data management. • This control mode allows you to produce enterprise-specific database DevOps solutions through custom design specifications and approval processes. • All features that are included in the flexible management and stable change control modes are available.	• Ensure the data security of database instances. • Implement strict access control over development or change workflows. • Manage compliance for enterprises.	Logon-free through authorization.

Note The instances that are managed in Stable Change mode consume the billing quota of the instances that are managed in Security Collaboration mode.

16.1.4. Features that are supported by each role

DMS provides the following roles: regular user, DBA, security administrator, and DMS administrator. This topic describes the features that are supported by each role.

Category	Feature	Regular user	DBA	Security administrator	DMS administrator	Description
Permission	Permission management	√	√	√	√	You can use this feature to apply for permissions on instances, databases, tables, and sensitive fields. You can also view permissions that you have.
	Data Changes	√	√	√	√	You can use this feature to initialize data for a newly published project, clean up historical data, fix bugs, or run a test.
	Data Import	√	√	√	√	You can use this feature to import a large amount of data to your databases at a time.

Category	Feature	Regular user	DBA	Security administrator	DMS administrator	Description
Data Plans	Data Export	√	√	√	√	You can use this feature to export a large amount of data for analysis or export the required data.
	Data Tracking	√	√	√	√	If specific data fails to meet your requirements due to reasons such as misoperation, you can use this feature to restore data to the normal state.
	Test Data Generate	√	√	√	√	Some business scenarios may require frequent data preparation. In this case, you can use this feature to generate test data to ensure data security and discreteness and improve production efficiency.
	Data Warehouse Development	√	√	√	√	DMS uses a database as a computing engine and integrates various tools and services, such as Data Transmission Service (DTS) and Data Lake Analytics (DLA), in the database ecosystem for data warehouse development. You can use this feature to develop and manage data warehouses in DMS with ease.
	Data Service	√	√	√	√	You can use this feature to export data at the field or row level, display data in a visualized manner, and publish API operations to the Alibaba Cloud Marketplace for sale.
	Database Clone	√	√	√	√	You can use this feature to clone MySQL databases.
Schemas	Schema Design	√	√	√	√	When you develop or optimize projects or process new business requirements, you can use this feature to change schemas. For example, you can use this feature to create a table or modify an existing table.
	Table Sync	√	√	√	√	You can use this feature to compare and synchronize the schemas of tables in different environments, such as online and offline environments. This feature helps ensure the consistency of schemas.
Optimization	SQL Review	√	√	√	√	You can use this feature to prevent SQL statements that do not use indexes or do not conform to database development standards. This feature helps protect against SQL injection attacks.

Category	Feature	Regular user	DBA	Security administrator	DMS administrator	Description
SQLConsole	Single Database query	√	√	√	√	You can write SQL statements to query data in a single database. This feature can be used to verify business code, analyze product effects, and identify issues in an online environment.
	Cross-database Query	√	√	√	√	You can use this feature to perform join queries across online heterogeneous databases that are deployed in different environments.
System Management	Instance management	×	√	×	√	You can use this feature to manage instances. For example, you can register, view, or edit instances.
	User management	×	×	×	√	You can use this feature to manage users. For example, you can add, view, or edit users as needed.
	Task management	×	√	×	√	You can use this feature to manage tasks. For example, you can create, start, or stop tasks.
	Configuration management	×	×	×	√	You can use this feature to view and modify system configurations, or view the historical modifications of the configurations.
Security management	Security Rules	×	√	×	√	You can use this feature to configure security rules. Only SQL statements that conform to the security rules can be executed.
	Approval Processes	×	√	×	√	Approval processes are associated with security rules. You can configure different approval processes for different types of tickets.
	Operation Logs	×	√	√	√	Operations logs record data changes. Each record contains information such as the user who performed the operation, operation details, and time at which the operation was performed. You can use this feature to track historical user operations at any time.
	Access IP Whitelists	×	×	×	√	After you configure an access IP whitelist, only the IP addresses or Classless Inter-Domain Routing (CIDR) blocks in the whitelist can access the resources within your DMS tenant. This effectively enhances data security.

Category	Feature	Regular user	DBA	Security administrator	DMS administrator	Description
	Sensitive Data	×	√	√	√	You can use this feature to manage sensitive data. For example, you can use algorithms to de-identify sensitive data or adjust the security levels of sensitive data.
Tickets	Ticket management	√	√	√	√	You can use this feature to configure notification methods. DMS can notify you of the approval or execution status of tickets by using DingTalk notifications or emails.

16.1.5. Apply for permissions

You can apply for the query, change, and export permissions on a database, table, or column. After the database owner approves your application, you can query, change, and export data.

Permissions

- Query permissions: the permissions to execute SQL statements in the SQLConsole to query the data of the object on which you want to apply for the permission.
- Change permissions: the permissions to submit tickets to change data or synchronize data in a database or table. You cannot change data without approval.
- Export permissions: the permissions to submit tickets to export data from the object on which you want to apply for the permission. You cannot export data without approval.

Permission categories that are supported by each control mode

Permission category	Permissions	Control mode		
		Flexible management	Stable change	Security collaboration
Instance logon	After you obtain the instance logon permission on an instance, you can use the preset database account and password to log on to the instance.	√	√	×
Database	Database permissions are classified into query, export, and change permissions. After you obtain the database permissions on a database, you can access the following resources of the database: 1. All the insensitive fields. 2. All the tables to which row-level control settings are not applied. 3. All new tables. • Query permission: You can execute SQL statements in the SQLConsole to query data. • Change permission: You can submit data change and data import tickets. • Export permission: You can submit data export tickets.	×	×	√

Permission category	Permissions	Control mode		
		Flexible management	Stable change	Security collaboration
Table	Table permissions are classified query, export, and change permissions. After you obtain the table permissions on a table, you have full access to all data in the table except sensitive fields. • Query permission: You can execute SQL statements in the SQLConsole to query data. • Change permission: You can submit data change and data import tickets. • Export permission: You can submit data export tickets.	×	×	√
Sensitive field	Sensitive field permissions are classified into query, export, and change permissions. After you obtain the sensitive field permissions on sensitive fields in a table, you have full access to all sensitive fields in the table. Before you apply for the sensitive field permissions, you must obtain the database and table permissions to which the sensitive fields belong. • Query permission: You can execute SQL statements in the SQLConsole to query data. • Change permission: You can submit data change and data import tickets. • Export permission: You can submit data export tickets.	×	×	√
Database owner	• The owner of a database can manage the permissions on the database. For example, the owner of a database can grant or revoke permissions on the database and tables in the database. • The owner of a database can query all data in the database except sensitive or confidential fields. The owner can also submit tickets to perform operations on the data and schemas in the database without the need to apply for permissions. • DMS automatically identifies and assigns database owners to owner nodes in approval processes.	√	√	√
Table owner	• The owner of a table can manage the permissions on the table. For example, the owner can grant or revoke permissions on the table. • The owner of a table can query all data in the table except sensitive or confidential fields.	√	√	√

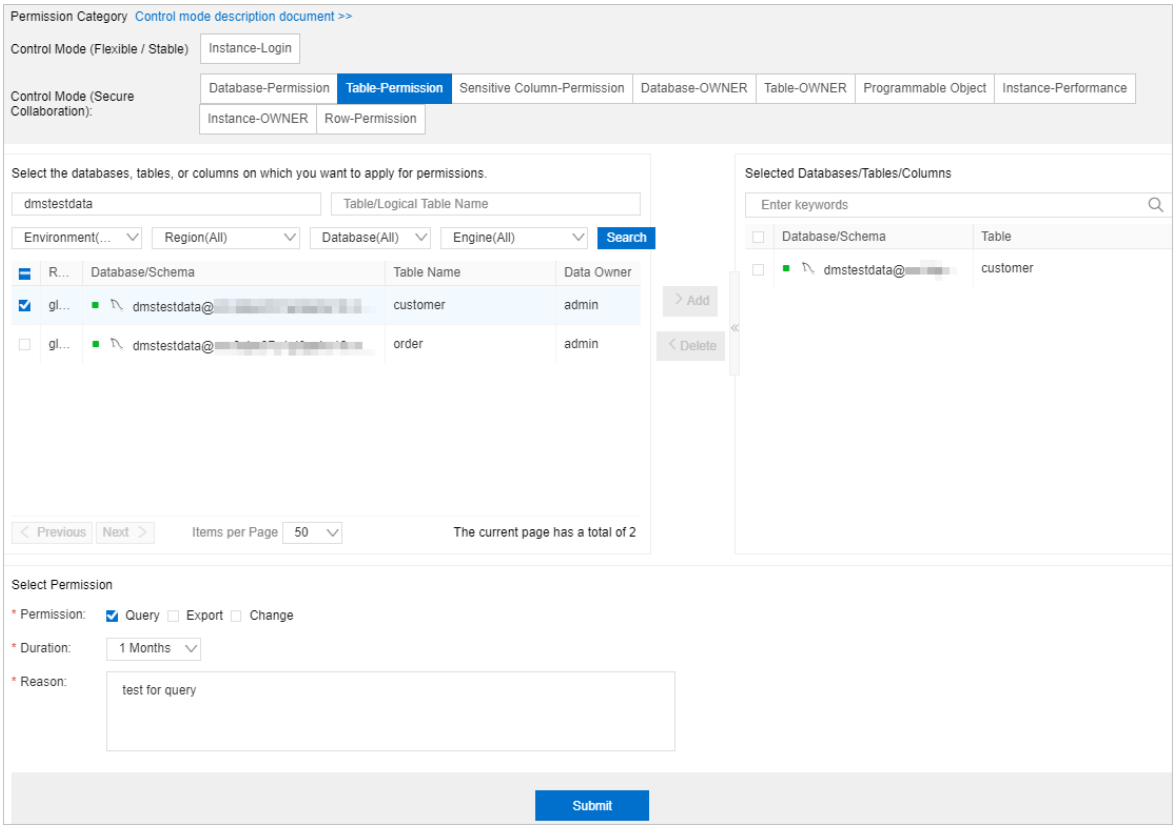
Permission category	Permissions	Control mode		
		Flexible management	Stable change	Security collaboration
Programmable object	Programmable object permissions are classified into query, export, and change permissions. • Query permission: You can execute SQL statements in the SQLConsole to query data. • Change permission: You can submit data change and data import tickets. • Export permission: You can submit data export tickets.	×	×	×
Instance performance	You can apply for permissions to view the performance of instances that are managed in security collaboration mode.	×	×	√
Instance owner	• The owner of an instance can manage the permissions on the instance. For example, the owner of an instance can grant or revoke permissions on the instance. • The owner of an instance can query all data in the databases of the instance except sensitive or confidential fields. The owner can also submit tickets to perform operations on the data and schemas of the instance without the need to apply for permissions.	√	√	√
Row	Row permissions are classified into query, export, and change permissions. You can apply for permissions on specific values of a managed field in a table. You can also apply for permissions on all values of a managed field in a table. • Query permission: You can execute SQL statements in the SQLConsole to query data. • Change permission: You can submit data change and data import tickets. • Export permission: You can submit data export tickets.	×	×	×

Apply for permissions

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Permissions** and select a permission category. For more information about permission categories, see Permission categories that are supported by each control mode.

   > **Note** In the top search box, you can search for databases or tables by name. In the search results, find the required database or table and click **Access apply** in the **Actions** column.

3. Set the required parameters of the permission for which you want to apply.



i. Select the category of the permission for which you want to apply.

ii. Select the databases, tables, or columns on which you want to apply for permissions.

> **Note** Enter keywords, specify filter conditions, and then click Search to search for databases or tables. The keywords that you enter can contain percent signs (%) as wildcards. In the search results, select the databases or tables on which you want to apply for permissions. Then, click Add.

iii. Select the type of permissions for which you want to apply and specify the duration for which you want to have the permission. Then, enter the reason for your application.

4. Click **Submit** and wait for approval.

> **Note** You can view the status of application ticket in the My Tickets section of the **Workbench** tab.

Manage permissions

Management type	Action	Description
Active management	Release permissions	On the Workbench tab, click **Effective Permissions** in the Permissions section. Select the object on which you want to release permissions and click **Release Permission**.
	Renew permissions	On the Workbench tab, click **Expiring Permissions** in the Permissions section to view and check the permissions that are about to expire. If you want to renew a permission, submit a ticket to apply for the renewal.
Passive management	N/A	The owner of a database can view and assess the rationality of permission assignments at any time and manage the permissions that are granted.

> **Note** All the operations that you perform to apply for, release, revoke, and grant permissions are recorded in operation logs. To view the operation logs, choose **System Management > Security > Operation Logs** in the top navigation bar.

16.1.6. SQLConsole

16.1.6.1. Cmd Tab

Data Management (DMS) provides the Cmd Tab feature as a CLI for you to write and execute SQL statements. The executed SQL statements and execution results are displayed in the upper part of the Cmd Tab tab. This topic describes the GUI of the Cmd Tab tab and how to use the Cmd Tab tab.

Prerequisites

- The database to be queried is a relational database, such as a MySQL, an Oracle, or an SQL Server database.
- You are granted the query permissions on the database or table that you want to query.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **SQLConsole > Cmd Tab**.
3. Select the database that you want to query from the drop-down list or enter a keyword in the field to search for the database. After you select the database, click **Confirm**.
4. Enter the SQL statement to be executed in the lower part of the Cmd Tab tab and press `Ctrl+Enter` or click **Execute** to execute the SQL statement.

16.1.6.2. Single database query

The single database query feature allows you to execute various SQL statements in the SQLConsole of the console with ease. You can use this feature to visualize the add, delete, modify, and query operations on the data in a database. This feature applies to scenarios, such as data queries and data development.

Prerequisites

You are granted the query permission on the database or table that you want to query.

Precautions

- A table may contain sensitive or confidential fields. You do not have permissions to access these fields. Therefore, the values of these fields are displayed as `******` in the query results. For more information, see Manage sensitive data.
- By default, a maximum of 200 data rows can be returned for each query. If you are an administrator, you can change this value based on your business requirements. To change this value, perform the following steps: 1. Log on to the Data Management (DMS) console. 2. In the top navigation bar, choose **System > Security > Security Rules**.
- A full scan can be performed on a table that does not exceed 10 GB in size. If you are an administrator, you can change this value based on your business requirements. To change this value, perform the following steps: 1. Log on to the DMS console. 2. In the top navigation bar, choose **System > Security > Security Rules**.
- By default, the timeout period to execute a single SQL statement is 60 seconds. If you are an administrator, you can change this value in the **Advanced information** section of the **Edit instance** dialog box.

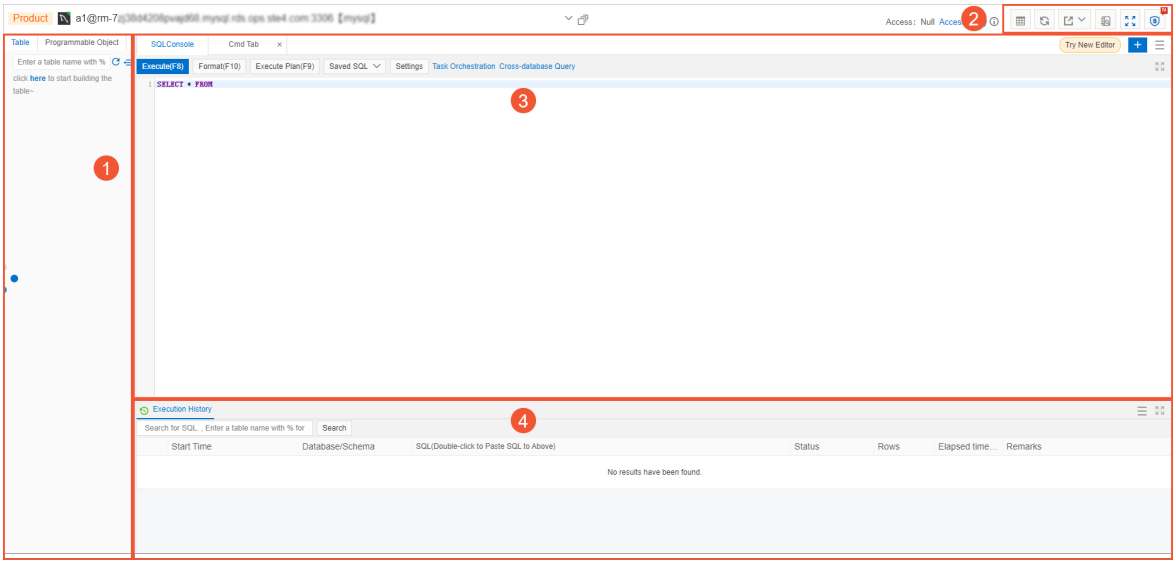
Procedure

1. Log on to the DMS console.

2. In the top navigation bar, choose **More > SQLConsole > Single Database Query**.

   > **Note** To go to the **SQLConsole** tab, you can also double-click the required database in the left-side instance list of the DMS console.

3. Select the database that you want to query from the drop-down list. You can also search for databases by keyword. After you find and select the required database, click **Confirm**.
4. Enter the SQL statement to be executed on the SQLConsole tab and click **Execute**.



GUI of the SQLConsole

No.	Section	Description
①	Visual operation section	In this section, you can visually manage your database. ○ **Tables** You can view all tables, fields, and indexes of the current database. You can also right-click a table in the database to modify the table schema, import data to the table, or export data from the table. ○ **Programmable objects** You can create, view, execute, and manage programmable objects, such as views, stored procedures, functions, triggers, and events. **Note** A maximum of 1,000 entries can be displayed. ○ Key-value pair information **Note** The key-value pair information can be displayed only for a NoSQL database.

No.	Section	Description
②	Extended feature section	In this section, shortcuts to extended features are provided. You can click the icons of the features to use the features. The following table describes the icons. ○ : the **Tables** icon. You can click the icon to view the details about the table. Then, click the icon to return to the SQLConsole tab. ○ : the **Sync Metadata** icon. After you click this icon, DMS collects most recent metadata information about the database, such as tables, fields, indexes, and programmable objects. This way, you can manage permissions on tables, fields, and programmable objects based on the security level. ○ : the Export icon. You can click this icon to export the data of the database, table schemas of the database, or table creation statements. ○ : the **Operation audit** icon. You can click this icon to view the information about all data query and data change records. For example, you can query the information about an operation, the operator, and the time when the operation is performed. For more information, see View operations logs.
③	Command running section	In this section, you can write and execute SQL statements to manage the current database. You can also format SQL statements, create execution plans, save commonly used SQL statements, and configure display settings. **Note** You can click the icon to add multiple query tabs.
④	Execution result section	In this section, you can view the execution results after SQL statements are executed. You can also view the details about a single row and add, delete, or modify data. **Note** You can click the **Execution History** tab to view the historical execution records. For example, you can view the time at which the execution of an SQL statement started, the affected database, and the details about the SQL statement. You can also export the execution results as required.

16.1.6.3. Cross-database query

Data Management (DMS) provides the cross-database query feature that allows you to perform cross-database queries on online heterogeneous data sources that are deployed in different environments. The cross-database query feature allows you to perform cross-database queries on databases and tables in database instances that are added to DMS.

Prerequisites

- The type of the database instance that you want to query is MySQL, SQL Server, PostgreSQL, PolarDB-X, or Redis.
- The cross-database query feature is enabled for each database instance.

Note If the cross-database query feature is not enabled for a database instance, you can enable the feature in the Edit instance dialog box in the DMS console. Then, enter a database link name in the Advanced information section. The name of each database link must be unique.

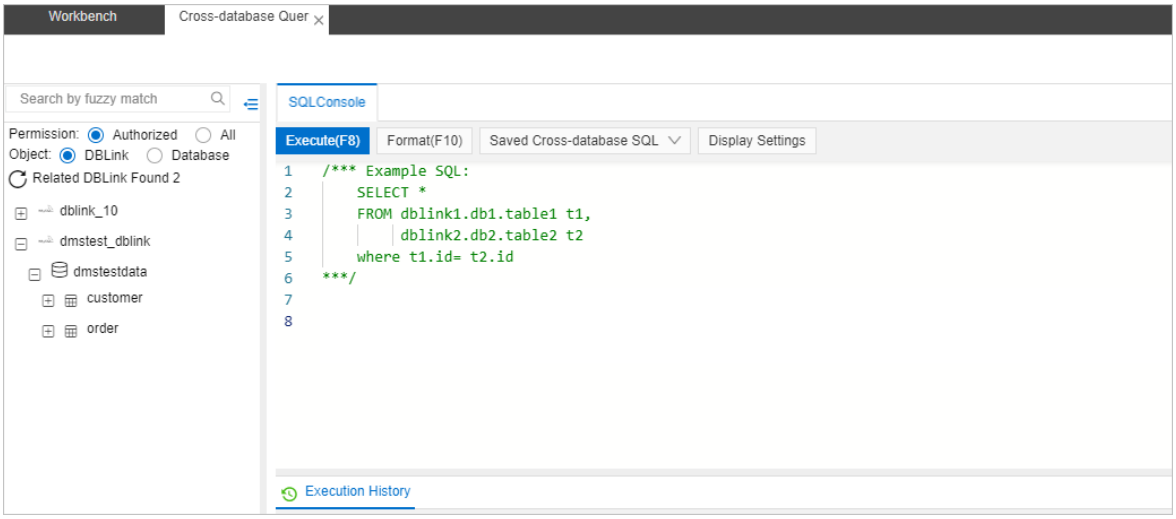
- You are granted the query permission on each database or table that you want to query.

Limits

You can perform cross-database queries only on physical databases. This feature is unavailable for logical databases.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > SQLConsole > Cross-database Query**.
3. In the left-side pane of the **Cross-database Query** tab, view the databases on which you have the query permission or all the databases.
4. On the **SQLConsole** tab, enter an SQL statement. You can perform cross-database queries on databases and tables on which you have the query permission in database instances.



> **Note**
> - You must specify the table that you want to query in the format of `<DBLinkName>.<databaseName>.<tableName>`, for example, `dmstest_dblink.dmstestdata.customer`.
> - In the left-side list, you can double-click a table on which you have the query permission or drag the table to the SQLConsole tab. An SQL statement that is used to query data in the table is automatically generated.

5. Click **Execute(F8)**. You can view the execution results and execution history in the Execution History section.

> **Note** DMS also provides the format, saved cross-database SQL, and display settings features. You can use these features based on your business requirements.

16.1.6.4. Manage schema versions

After you change the schema of a table in a database in , DMS adds the latest schema to the schema version list of the database. You can download and compare schema versions, and restore an earlier schema version in the schema version list.

Prerequisites

You have the permissions to query the data of the table or the database to which the table belongs. For more information, see Apply for permissions.

Overview

Schema versions are defined based on a database and store the schema information of all the tables in the database. If the schema of a table in the database is changed, a new schema version is saved. For more information, see Save new schema versions.

If a database instance that has five databases is managed in Security Collaboration mode, each database can contain 50 schema versions. In other words, for a database instance that is managed in Security Collaboration mode, a maximum of 50 schema versions can be retained for each database in the instance.

Limits

- The following database engines are supported:
  - MySQL: ApsaraDB RDS for MySQL databases, PolarDB-X databases, AnalyticDB for MySQL databases, MySQL databases in third-party clouds, and self-managed MySQL databases.
  - SQL Server: ApsaraDB RDS for SQL Server databases, SQL Server databases in third-party clouds, and self-managed SQL Server databases.
  - PostgreSQL: ApsaraDB RDS for PostgreSQL databases, AnalyticDB for PostgreSQL databases, PostgreSQL databases in third-party clouds, and self-managed PostgreSQL databases.
  - OceanBase
  - DamengDB
- The following content shows the maximum number of schema versions that can be retained for each database in database instances that are managed in different control modes:
  - Flexible Management: 3
  - Stable Change: 20
  - Security Collaboration: 50
- You cannot manage schema versions for the following databases:
  - Databases that contain more than 1,024 tables
  - System databases such as the information_schema and system databases in a MySQL database instance

Procedure

1. Log on to the DMS console.
2. , click the database instance where the database that you want to manage resides, right-click the database, and then select **Version management**.

   > **Note** On the SQL Console tab of the database that you want to manage, move the pointer over the icon and select **Version management**.

3. On the **Database version list** tab, find the version number of the schema that you want to manage. The following table describes the operations that you can perform.

Operation	Description
View	View the detailed information of the schema version.
Preview script	Preview the SQL script that is used to generate the schema version.
Table structure comparison	Synchronize the schema version to another database, or compare the schema version with a schema version in another database. For more information, see Schema synchronization.

Operation	Description
Structural recovery	Synchronize the schema version that you want to restore to an empty database. For more information, see Initialize empty databases.

Save new schema versions

New schema versions are saved when the following operations are performed in DMS:

- SQL statements are executed on the SQL Console tab to change schemas.
- SQL statements are executed to change schemas when you submit tickets for normal data change, lock-free data change, schema design, or schema synchronization.
- SQL statements are executed to change schemas by a DMS administrator.

> **Note**
>
> If the schema of a table in a database is changed in environments other than DMS, you can perform the following operations in DMS to synchronize the latest schema of the database:
>
> - If the instance where the database resides is managed in Security Collaboration mode, click the icon in the upper-right corner of the SQL Console tab to synchronize the latest schema.
> - If the instance where the database resides is managed in Flexible Management or Stable Change mode, click the icon in the upper-left corner of the SQL Console tab to synchronize the latest schema.

16.1.6.5. Generate a risk audit report

allows you to generate risk audit reports for database instances. Risk audit reports collect and assess various risks that are involved in the O&M of database instances. Risk audit reports also provide optimization suggestions for you to improve the security and stability of your instances.

Overview

A risk audit report is generated based on a database instance in DMS. The report diagnoses and analyzes the risks that are involved in the O&M of the instance or a specific database in the instance.

The following table describes the risk audit items that are contained in risk audit reports.

Risk audit item	Description	Supported database engines

Risk audit item	Description	Supported database engines
SQL audit	For this item, DMS checks whether the SQL statements that are executed in the DMS console to manage the current database instance conform to the R&D specifications. By default, DMS checks the SQL statements that are executed in the previous week. The statements include those that are executed on the SQL Console tab and those that are executed after tickets are submitted, such as Normal Data Modify and Lockless change tickets. For example, DMS may find the following accidental operation: A whole table was accidentally updated because the `WHERE` clause was missing in an `UPDATE` statement. **Note** This audit item is checked based on optimization suggestions for SQL review.	MySQL Self-managed MySQL databases, ApsaraDB RDS for MySQL databases, PolarDB for MySQL databases, PolarDB-X databases, and AnalyticDB for MySQL databases
Metadata	For this item, DMS assesses the risks of all the schemas in the current database instance. For example, DMS may identify the following risk: An auto-increment primary key of the INT type runs out of valid values. **Note** This audit item is checked based on optimization suggestions for SQL review.	MySQL Self-managed MySQL databases, ApsaraDB RDS for MySQL databases, PolarDB for MySQL databases, PolarDB-X databases, and AnalyticDB for MySQL databases
Sensitive Data	For this item, DMS checks whether the current database instance contains sensitive fields. For example, if the instance contains sensitive fields, such as mobile numbers, ID card numbers, or passwords, DMS checks whether these fields are prone to sensitive data breaches.	• MySQL Self-managed MySQL databases, ApsaraDB RDS for MySQL databases, PolarDB for MySQL databases, PolarDB-X databases, and AnalyticDB for MySQL databases • SQL Server Self-managed SQL Server databases and ApsaraDB RDS for SQL Server databases • PostgreSQL Self-managed PostgreSQL databases and PolarDB for PostgreSQL databases • MaxCompute

Limits

- Only DMS administrators, security administrators, database administrators (DBAs), instance owners, and database owners can generate risk audit reports.
- You can keep only a limited number of risk audit reports for an instance. The number varies based on the control mode of the instance.
  - For an instance that is managed in Flexible Management mode, you can keep up to three reports. You cannot view the details of the reports.
  - For an instance that is managed in Stable Change mode, you can keep up to 20 reports.

- For an instance that is managed in Security Collaboration mode, you can keep up to 50 reports.

Procedure

1. Log on to the DMS console.
2. , right-click the instance for which you want to generate a risk audit report and choose **Audit > Risk Audit**.

   > **Note** On the SQL Console tab of the database, move the pointer over the icon and select **Risk Audit**.

3. Click **Real-time Diagnostics**.

   > **Note** By default, DMS does not automatically diagnose an instance. If this is the first time for the instance to be diagnosed, you can click **Diagnose**.

4. In the dialog box that appears, select the risk audit items and click **Diagnose**.

   Wait until the status of the report that is being generated changes to **Completed**.

5. After the report is generated, you can view the diagnosis details of each database in the instance.

16.1.6.6. Super SQL mode

Data Management (DMS) provides the super SQL mode feature. After you enable this feature as a DMS administrator or a database administrator (DBA), all SQL statements that you execute on the SQLConsole tab are executed without being affected by security rules.

Prerequisites

- You are a DMS administrator or a DBA.
- An instance is managed in Security Collaboration mode.

Context

To enhance the stability and security of databases, DMS administrators and DBAs may configure security rules for the databases. For example, a security rule is configured to prevent unauthorized users from executing DML statements in a production database on the SQLConsole tab. They can execute those statements only by submitting a ticket. However, these security rules may cause inconvenience to privileged users, such as DMS administrators and DBAs.

In view of this, DMS provides the **super SQL mode** feature. If you enable this feature as a DMS administrator or a DBA, all SQL statements that you execute on the SQLConsole tab are executed without being affected by security rules.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **SQLConsole > Single Database query**.

   > **Note** To go to the **SQLConsole** tab, you can also double-click the database that you want to query in the left-side navigation pane of the DMS console.

3. Select the database that you want to query from the drop-down list or enter a keyword in the field to search for the database. After you select the database, click **Confirm**.
4. On the SQLConsole tab, click the icon in the upper-right corner. In the message that appears, click **OK**.

Then, the outside borders of the SQLConsole tab turn orange. This indicates that the **super SQL mode** feature is enabled. The SQL statements that you enter on the SQLConsole tab are directly executed.

> **Notice** After you enable this feature as a DMS administrator or a DBA, all SQL statements that you execute on the SQLConsole tab are executed without being affected by security rules.

To disable the **super SQL mode** feature, click the icon in the upper-right corner.

16.1.7. Data plans

16.1.7.1. Change data

DMS provides the data change feature that allows you to change data. This topic describes how to use the data change feature to change data.

Context

DMS allows you to submit data change tickets to initialize data for a newly published project, clear historical data, fix bugs, or run a test. The operations that you can perform to change data include insert, update, delete, or truncate operations.

Data change types

Type	Description

Normal data modify	You can use the normal data modify feature to perform the following data changes:
	• Perform normal data changes.
	• Perform lock-free schema changes. You can perform this type of operations to change character sets and collations for tables, adjust time zones, and change column data types. Compared with normal data change operations, lock-free schema changes can be performed to achieve the following benefits:
	  ◦ Ensures business continuity regardless of whether tables are locked due to database schema changes.
	  ◦ Ensures consistent synchronization between the primary and secondary databases regardless of whether the latency occurs due to native online data definition language (DDL) operations that are performed on the databases.
	  ◦ Reclaims tablespaces and reduces fragmentation rates without the need to lock tables. You no longer need to use the OPTIMIZE TABLE statement that results in table locks.
	Note You can use this feature only for MySQL databases. Before you use this feature, you must set the OnlineDDL parameter to Open(DMS OnlineDDL first) in the **Advanced information** section when you add or edit an instance. For more information, see Register database instances with DMS.

Lock-free data modify	You can use this feature to change a large amount of data. For example, you can use this feature to delete historical data and update all fields in a table. Multiple SQL statements for data changes are divided and executed at the same time based on the primary key or unique key. This limits the consumption of database performance and space. **Note** You can use this feature only for MySQL databases.

Type	Description
History data clean	You can use this feature to regularly clean historical data. This way, the stability of the production environment is not affected when you obtain historical data. **Note** You can use this feature only for MySQL databases.
Programmable object	Databases provide programmable objects such as stored functions and stored procedures. This feature allows you to use the programmable objects to standardize management processes and provide audit records.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Data Plans** and select a data change type.

   Note This topic describes how to submit a ticket that is used to **modify normal data**.

3. Specify the required parameters for the data change ticket.

Parameter	Description
Reason Category	The reason for the data change. This helps you identify the ticket in the future.
Business Background	The purpose or objective of the data import. This reduces communication costs.
Change Stakeholder	The stakeholders of the data change. All specified stakeholders can view the details about the ticket and participate in the approval process. Unauthorized users, except for administrators and DBAs, cannot view the details about the ticket.
Execution Method	The method that is used to execute the ticket based on your business requirements.
Database	The database on which you have the change permission. If you have only query permissions on the database or the change permission on the tables of the database, you cannot submit a data change ticket.
Affected Rows	The estimated number of data rows that are affected by the data change. To obtain the actual number of affected rows, you can write an SQL statement that includes the COUNT function on the SQLConsole tab.
SQL Statements for Change	The executable SQL statements that are used to export data. You can upload a file to provide the SQL statements. DMS verifies the syntax of the SQL statements when you submit the ticket. If the syntax is invalid, DMS rejctes the ticket.
SQL Statements for Rollback	The executable SQL statements that are used to roll back the data import operation. You can write the SQL statements in the SQL Text field or upload an SQL that includes the required SQL statements.
Attachments	The images or files that are uploaded to add more information about the data change.

4. After you configure the settings, click **Submit**.
5. After your ticket passes the precheck, click **Submit for Approval**. In the message that appears, click **OK**.
6. After the ticket is approved, click **Execute Change**.
7. Set the Execute Immediately parameter and click **Confirm Execution**.

> **Note** By default, the Execute Immediately switch is turned on. You can turn off the **Execute Immediately** switch and specify a point in time to run the ticket. The system automatically runs the ticket at the specified point in time.

8. Wait until the execution is completed.

16.1.7.2. Import data

Data Management (DMS) provides the data import feature that allows you to import large amounts of data to a database with ease. This saves manpower and resources.

Supported databases

- Self-managed MySQL databases and ApsaraDB RDS for MySQL databases
- PolarDB-X databases

Supported file formats for data import

- TXT format.
- SQL script. By default, you can use only the INSERT and REPLACE statements to import data to database instances that are managed in Security Collaboration mode. If you want to use other SQL statements to import data, modify the security rules for data import as a database administrator (DBA) or DMS administrator. To modify the security rules, click the **SQL Correct** tab on the **Security Rules** tab and set the Checkpoints parameter to **Batch Data import rules**.
- CSV format. Values in a CSV file must be separated by commas (,). The first row must be field names.

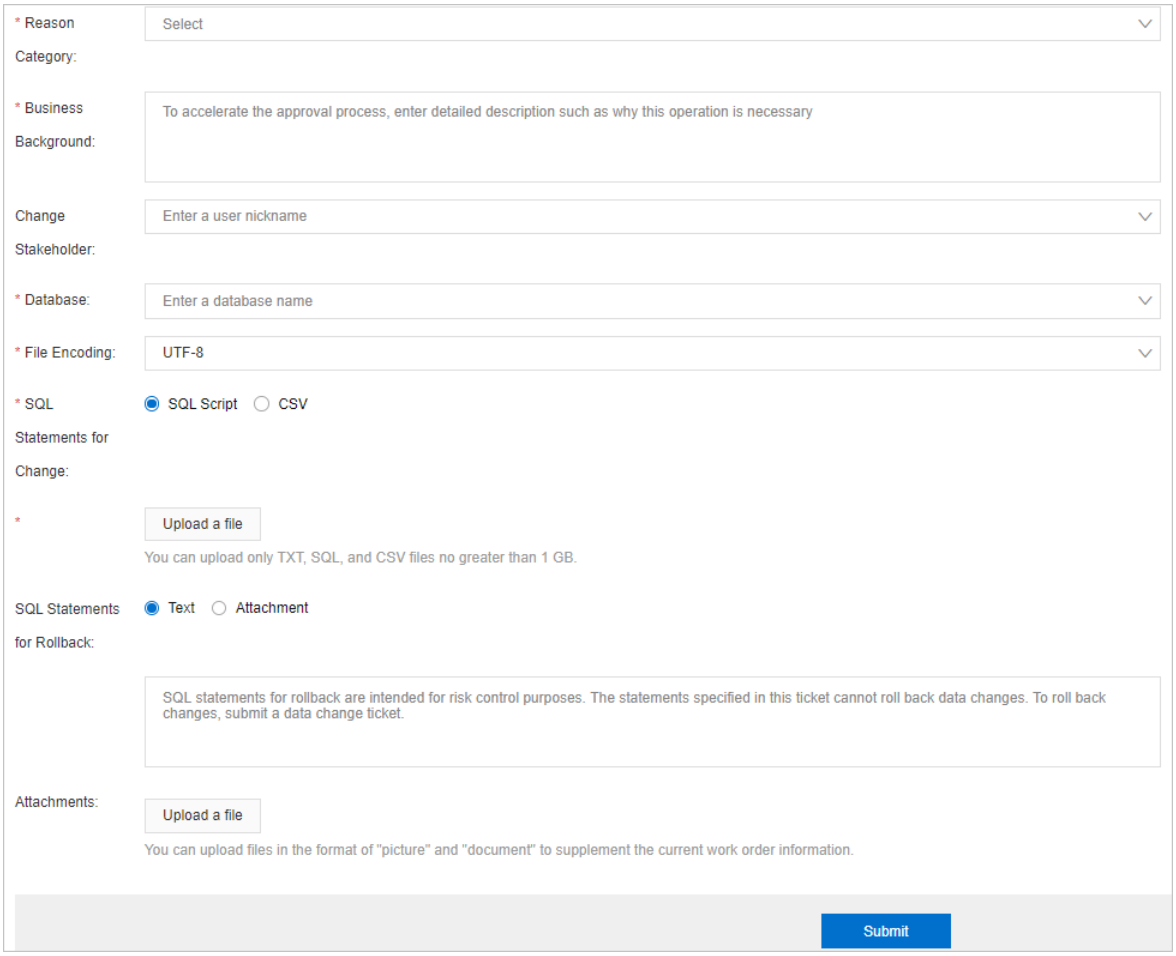
> **Note** The file size cannot exceed 5 GB.

Usage notes

- If you need to change only a small amount of data, we recommend that you submit a Normal Data Modify or Lockless change ticket to ensure stable data change. For more information, see Change data.
- If you submit a Large Data Import ticket to import a large amount of data to a table, the table will be locked even if you set the OnlineDDL parameter to Open (DMS OnlineDDL first) for the database instance.
- We recommend that you use SQL statements with better performance to import a large amount of data, such as the INSERT, UPDATE, and DELETE statements. Indexes of primary keys are used in the UPDATE and DELETE statements.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Plans > Data Import**.
3. On the Large Data Import tab, set the parameters that are described in the following table.



Parameter	Description
Reason Category	The reason for the data import. This helps you identify the ticket in the future.
Business Background	The purpose or objective of the data import. This reduces unnecessary communication.
Change Stakeholder	The stakeholders involved in the data import. All the specified stakeholders can view the ticket details and assist in the approval process. Irrelevant users other than DMS administrators and DBAs are not allowed to view the ticket details.
Database	The database on which you have the change permissions. You cannot submit a data import ticket if you have only the permissions to query data in the database or change data in tables.
File Encoding	The encoding algorithm to be used by the database.
SQL Statements for Rollback	The executable SQL statements for rolling back the data import. You can write the SQL statements in the SQL Text field or upload an SQL script that includes the required SQL statements.
Attachments	The images or files that are uploaded to add more information about the data import.

4. After you configure the settings, click **Submit**.
5. After your ticket passes the precheck, click **Submit for Approval**. In the message that appears, click **OK**.

6. After the ticket is approved, click **Execute Change**.
7. Set the Execute Immediately parameter and click **Confirm Execution**.

> **Note** By default, the Execute Immediately switch is turned on. You can turn off the **Execute Immediately** switch and specify a point in time to run the ticket. The system automatically runs the ticket at the specified point in time.

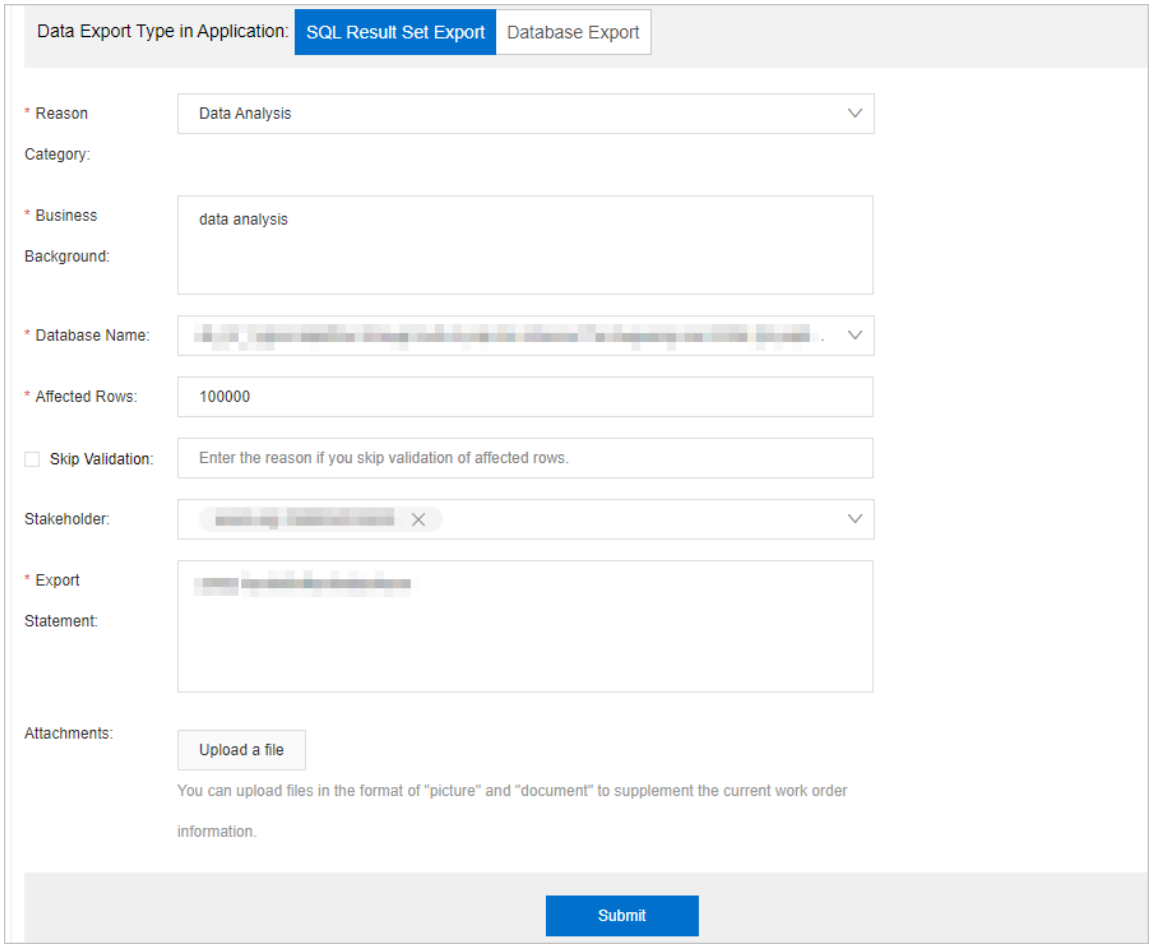
8. Wait until the execution is completed.

16.1.7.3. Data export

DMS provides the data export feature. You can use this feature to export a database or SQL result sets. Then, you can extract the required data for data analysis.

Procedure

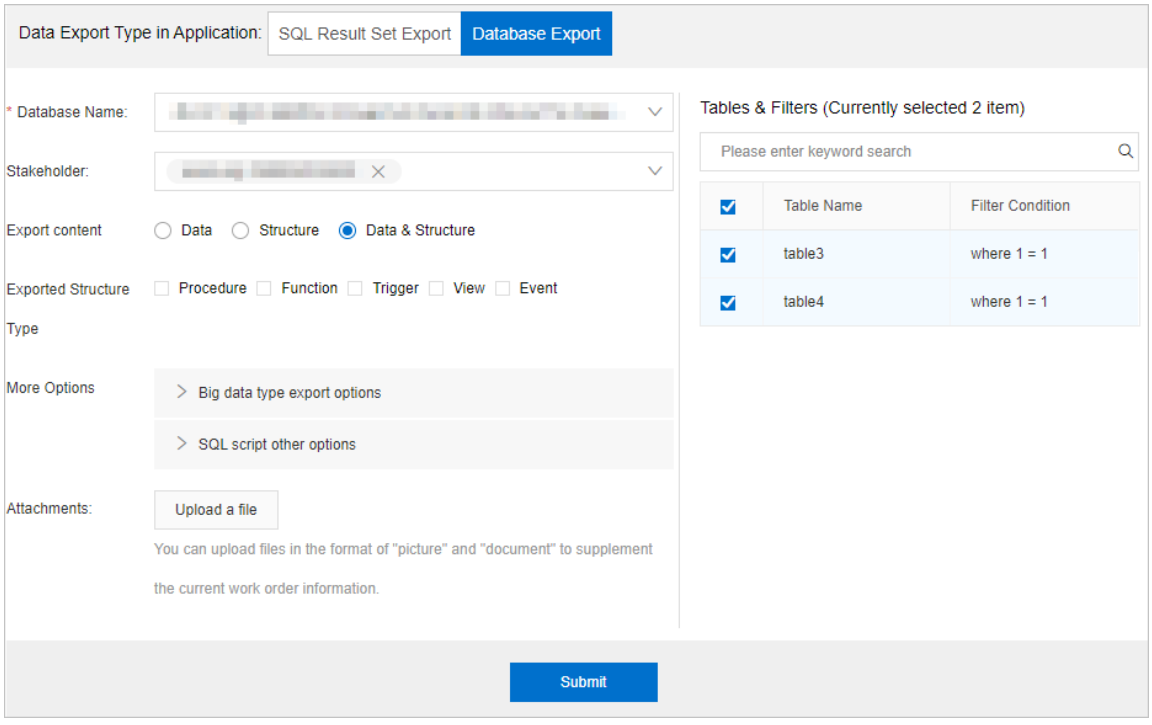
1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Data Plans** and select **SQL Result Set Export** or **Database Export**.
3. On the Data Export Ticket Application tab, set the required parameters.
   - Set the required parameters on the SQL Result Set Export tab



Parameter	Description

Parameter	Description
Reason Category	The reason for the data export. This helps you find the ticket in a timely manner in the future.
Business Background	The purpose or objective of the data export. This parameter reduces communication costs.
Database Name	The database on which you have the export permission.
Affected Rows	The estimated number of data rows that are affected by the data export. To obtain the actual number of affected rows, use the `COUNT` function in SQL statements on the SQL Editor tab.
Skip Validation	Specifies whether to skip validation. If you select **Skip Validation**, you must enter a reason in the field next to the check box. **Warning** If you select **Skip Validation**, DMS does not check the number of rows that are affected by the data export. If a large amount of data is exported, potential risks to your business increase. Proceed with caution.
Stakeholder	The stakeholders of the data export. All specified stakeholders can view the details about the ticket and are included in the approval process. Unauthorized users, except for administrators and DBAs, cannot view the details about the ticket.
Export Statement	The executable SQL statement that is used to export data. Example: `select * from testtable`. DMS verifies the syntax of the SQL statement when you submit the ticket. If the syntax is invalid, you cannot submit the ticket.
Attachments	The images or files that are uploaded to add more information about the data export.

- Configure a database export ticket



Parameter	Description
Database Name	The database on which you have the export permission. After you select the database, you must select the table to which you want to export data and configure filter conditions in the Tables & Filters section.
Reason Category	The reason for the data export. This helps you find the ticket in a timely manner in the future.
Business Background	The purpose or objective of the data export. This parameter reduces communication costs.
Stakeholder	The stakeholders of the data export. All specified stakeholders can view the details about the ticket and are included in the approval process. Unauthorized users, except for DMS administrators and DBAs, cannot view the details about the ticket.
Export content	The type of data that you want to export. Valid values: **Data**, **Structure**, and **Data & Structure**.
Exported Structure Type	The type of schema that you want to export.
More Options	The other objects that you want to export. Click **Big data type export options** or **SQL script other options** and select the required options.
Attachments	The images or files that are uploaded to add more information about the data export.

4. After you complete the configurations, click **Submit** and wait for approval.

   > **Note** When you export an SQL result set, DMS prechecks the SQL statements. After the SQL statements pass the precheck, click **Submit for Approval**. In the message that appears, click **OK**.

5. After your ticket is approved, go to the **Workbench** tab and click **Submitted Tickets** in the My Tickets section.
6. Find the data export ticket that is submitted and click the ticket number.
7. In the **Execute/Automatic Execution** section, click **Download Exported File**.

16.1.7.4. Generate test data

DMS provides the test data generation feature that allows you to generate data in a quick manner. You can generate test data for functional or performance tests.

Prerequisites

- A relational database is created to store test data that you generate. The relational database can be a self-managed MySQL, ApsaraDB RDS for MySQL, AnalyticDB for MySQL, or PolarDB-X database.
- A table is created. You can use the schema design feature to create a table. For more information, see Design a schema.

Context

Functional tests or performance tests often require test data. You can use one of the following methods to generate test data:

- Write test data. This method has low efficiency and is not applicable to scenarios in which a large amount of test data is required.
- Use scripts. This method requires high costs and the data that is generated by using this method cannot meet

discreteness requirements.

- Export data from a production environment as test data. This method is not secure and may cause data leak.

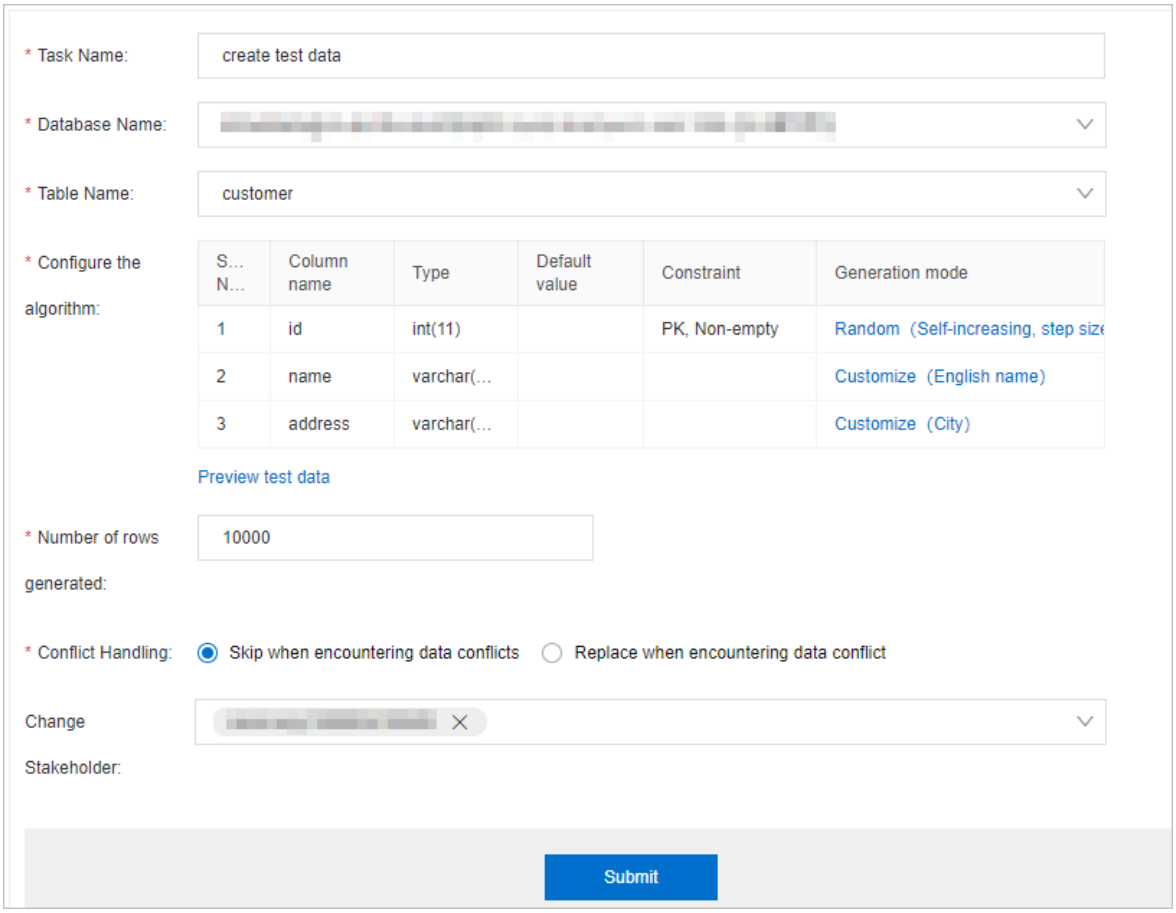
DMS provides the test data generation feature that allows you to generate test data in a quick, efficient, and secure manner. You can use this feature to control the discreteness of the data that is generated.

Precautions

- You can use this feature to generate test data in only one table at a time. To generate test data in multiple tables, submit a ticket.
- To prevent database overload due to the instantaneous generation of excessive data, DMS allows you to perform traffic throttling. Check the following examples for your reference.
  - About 1 minute is required to generate one million rows of data in a table that has four fields.
  - About 2 to 3 minutes are required to generate one million rows of data in a table that has 40 fields.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Data Plans > Test Data Generation**.
3. In the upper-right corner, click **Test Data Generation**.
4. In the Test data build ticket application dialog box, set the required parameters.



Parameter	Description
Task Name	The name of the task. The name can help you identify and manage the task in the future.

Parameter	Description
Database Name	The name of the database that includes the required table.
Table Name	The name of the table in which you want to generate test data. You can search for tables by keyword and then select a table from the matched results. After you select a table, the **Configure the algorithm** parameter that contains the field information of the table is displayed. **Note** You can use this feature to generate test data only in one table at a time. To generate test data in multiple tables, submit a ticket.
Configure the algorithm	The algorithms that are used to generate test data. Find the required field and click the value of the Generation mode parameter next to the field. Then, set the required parameters in Generation mode dialog box based on your business requirements. **Note** For example, you can use the random, customize, or enumeration algorithm to generate test data of the STRING type. The customize algorithm can be used to generate multiple industry-standard types of data.
Number of rows generated	The number of data rows that you want to generate.
Conflict Handling	The method that is used to handle conflicts based on your business requirements.
Change Stakeholder	The stakeholders of the ticket. All specified stakeholders can view the details about the ticket and are included in the approval process. Unauthorized users, except for DMS administrators and DBAs, cannot view the details about the ticket.

5. After you configure the settings, click **Submit**.
6. After your ticket is approved, DMS generates test data.

16.1.7.5. Clone databases

The database clone feature allows you to replicate data at the database level. This topic describes how to use the database clone feature.

Prerequisites

- A MySQL database is used.
- A database instance is managed in flexible management mode. You have logged on to the database instance in the Data Management (DMS) console.

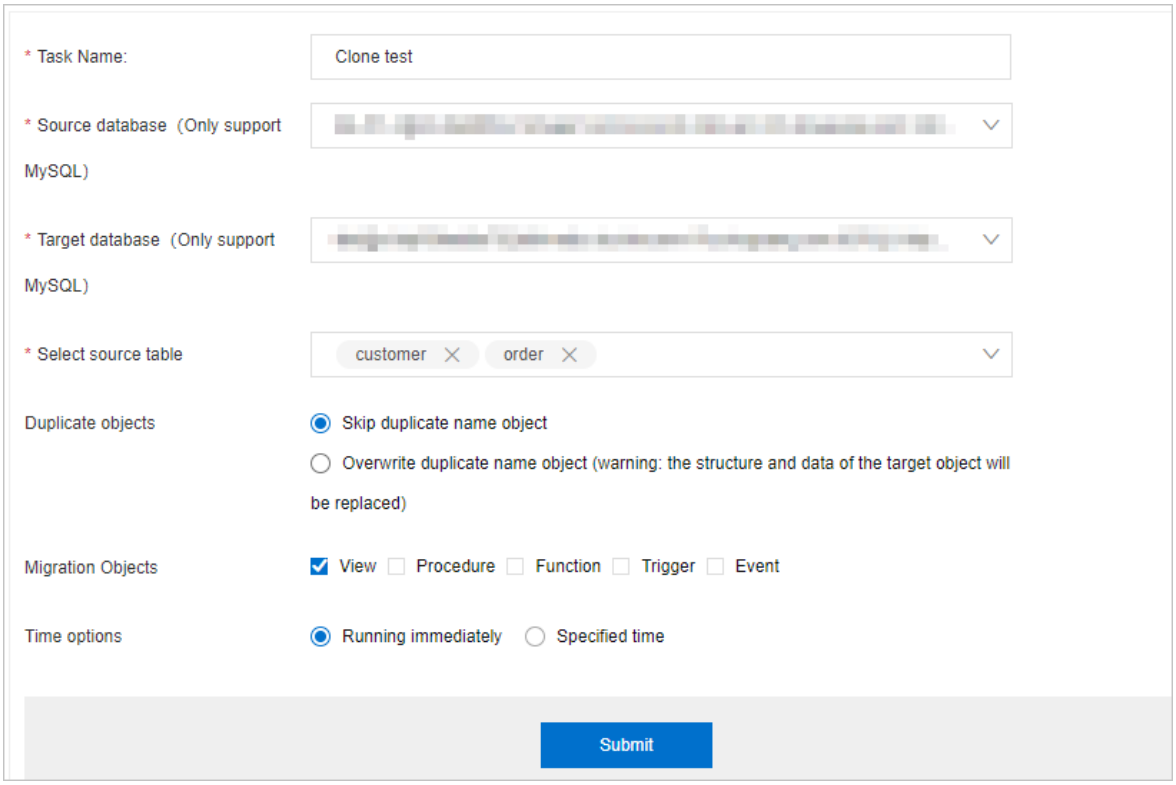
Scenarios

- Create a full database backup.
- Initialize databases that are deployed in different environments, such as development and test environments.
- Copy data from a database in an online environment to a database in an offline environment for data processing and analysis.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Data Plans > Database Clone**.
3. In the upper-right corner, click **Database Clone**.

4. In the Apply step, set the required parameters.



Parameter	Description
Task Name	The name of the task. The name is used to identify and manage the task.
Source database (Only support MySQL)	The source database that you want to clone. You can search for databases by keyword and then select a database from the matched results.
Target database (Only support MySQL)	The destination database to which you want to write the data that is cloned from the source database. You can search for databases by keyword and then select a database from the matched results. **Note** The destination database must be different from the source database.
Select source table	The tables that you want to clone from the source database. You can search for tables by keyword and then select a table from the matched results. **Note** To clone all tables, set this parameter to All Tables.
Duplicate objects	The method that is used to handle object conflicts based on your business requirements. ◦ **Skip duplicate name object**: The system skips objects that have a duplicate name. ◦ **Overwrite duplicate name object (warning: the structure and data of the target object will be replaced)**: If the two objects have the same name, the schema and data of the object in the destination database are overwritten by the schema and data of the object in the source database.

Parameter	Description
Migration Objects	The objects that you want to clone. In addition to tables, you can simultaneously clone other objects from the source database to the destination database. These objects include views, stored procedures, functions, triggers, and events.
Time options	Valid values: **Running immediately** and **Specified time**. If you set the Time options parameter to **Specified time**, you must specify a date and time to run the task. ○ **Running immediately**: The task is run immediately after the ticket is approved. ○ **Specified time**: DMS automatically runs the task to clone data at a specified point in time.

5. After you configure the settings, click **Submit**.
6. After the ticket is approved, the task is automatically run at a specified point in time.

16.1.8. Data factory

16.1.8.1. Task orchestration (new)

16.1.8.1.1. Orchestrate tasks

DMS provides the task orchestration feature that you can use to orchestrate various types of tasks and then schedule and run the tasks. You can create a task flow that consists of one or more task nodes. This allows you to schedule tasks in complex scenarios and improves efficiency of data development.

Prerequisites

The task orchestration feature supports the following relational databases: MySQL, SQL Server, PostgreSQL, PolarDB-X, PolarDB-O, self-managed Dameng (DM), self-managed Oracle, and self-managed OceanBase.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Factory > Task Orchestration (New)**.
3. In the left-side navigation pane, click the  icon.
4. Create a task flow.
   i. Click **Create Task Flow** in the lower part of the tab.
   ii. In the **Create Task Flow** dialog box, enter a name and the description for the task flow.
   iii. Click **OK**.
5. Configure a task node.

   Note You can repeat this step to configure multiple task nodes for the task flow.

   i. From the **Task Type** pane, drag a task node to the blank area of the canvas based on your business requirements.
   ii. Click the task node on the canvas.

iii. Configure the task node. In the following example, a Single Instance SQL node is configured.

Parameter	Description
Database	a. Click the **Node Information** tab. b. Select the database that you want to manage from the drop-down list below **Node Information**. **Note** You can view the schemas of tables in the database on the **Metadata** tab. c. Enter the SQL statements to be executed in the SQL editor. d. Click **Save**.
Variable Setting	■ Click the **Variable Setting** tab, click **Node Variable**, and then configure node variables. For more information, see Configure a time variable. ■ Click **Task Flow Variable** and configure task flow variables. For more information, see Configure a time variable.
Task rerun config	Click the **Advanced Settings** tab. You can turn off or turn on **Enable re-run of failed task**.

6. Connect the node to its upstream and downstream nodes.

   Move the pointer over the upstream node, click the hollow dot on the left side of the upstream node, and then drag the connection line to the node that you are configuring.

   In the following example, a **Table status check** node is connected to the **Single Instance SQL** node. Move the pointer over the **Table status check** node, click the hollow dot on the left side of the **Table status check** node, and then drag the connection line to the **Single Instance SQL** node.

   Note In the task flow shown in the preceding figure, the Table status check node is executed before the Single Instance SQL node.

7. Click the blank area of the canvas to configure the task flow.

   i. Configure basic properties. Click the **Task Flow Information** tab. In the **Properties** section, modify the task flow name, owner, and stakeholders, turn on or turn off **Enable message notification** switch, and then select the error handling policy and concurrency control policy.

   ii. Specify the scheduling information.

Parameter	Description
Enable Scheduling	Turn on **Enable Scheduling** to enable scheduling for the task flow.
Scheduling Type	Set this parameter based on your business requirements. ■ If you set this parameter to **Cyclic scheduling**, you must set the **Effective Time**, **Scheduling Cycle**, and **Specific Point in Time** parameters. ■ If you set this parameter to **Schedule once**, you need to set only the **Specific Point in Time** parameter.

8. Publish the task flow. For more information, see Publish task flows.

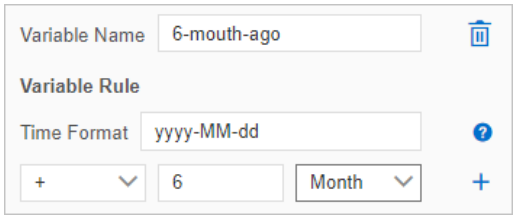
16.1.8.1.2. Configure variables

This topic describes the two types of variables that are used in the task orchestration feature. This topic also describes how to configure a time variable.

Variable types

- Node variables: A node variable is a time variable that is used in a node. Node variables are not displayed in the input variables of downstream nodes. Node variables are invisible to downstream nodes and cannot be referenced in downstream nodes.
- Task flow variable: A task flow variable is a time variable that can be used in all nodes of a task flow. Task flow variables are displayed in the input variables of downstream nodes. Task flow variables are visible to downstream nodes and can be referenced in downstream nodes.

Configure a time variable

Parameter	Description
Variable Name	The name of the custom variable. **Note** To delete a configured variable, click the icon.
Variable Rule	The time format of the time variable. • **Time Format**: Specify the time format that you require. For more information, see Time formats. • Offset: A time variable is defined based on the value of the bizdate variable that indicates one day before the current date. You can configure an offset to the value of the bizdate variable. For example, you can set the variable name to `6_month_ago`, the time format to yyyy-MM-dd, and the offset to - 6 Month. If the current date is 2021-8-12, the value of the variable `${6_month_ago}` is 2021-02-11.  **Note** After you configure a time variable, you can reference the time variable in the format of `${Variable name}` in SQL statements in the SQL editor on the right. You can click **SQLPreview** to view the value of the time variable.

Time formats

The following table describes the time formats that variables support.

Element	Description	Sample format	Sample value
Anno Domini (AD)	G indicates AD.	Gyyyy	AD 2021

Element	Description	Sample format	Sample value
Year	• y or yyyy: the year of the current day. • yy: the last two digits of the year. • Y: the year of the last day in the current week. The last day of the week is Sunday.	yyyy	2021
Month	M: the month of the current year. Valid values of M: [1,12]. Valid values of MM: [01,12].，MMM将返回一月至十二月	MM	08
Week	• w: the week of the current year. Valid values of w: [1,52]. Valid values of ww: [01,52]. • W: the week of the current month. Valid values: [1,5].	ww	13
Day	• D: the day of the current year. Valid values of D: [1,365]. Valid values of DD: [01,365]. Valid values of DDD: [001,365]. • d: the day of the current month. Valid values of d: [1,31]. Valid values of dd: [01,31].	D	360
Day of the week	• E: the day of the week. Valid values: Monday to Sunday. • e: the number that indicates the day of the week. Valid values: [1,7]. A value of 1 indicates Monday. Note To count Sunday as the first day of the week, you can set the offset to + 1 Day.	e	1
Ante meridiem (AM) or post meridiem (PM)	a: indicates whether the point in time is before or after the midday. If the time in point is from 00:00 to 11:59, the return value is AM. If the time in point is from 12:00 to 23:59, the return value is PM.	a	AM
Hour	• H: the hour of the current day. A value of 0 indicates the first hour of the day. Valid values of H: [0,23]. Valid values of HH: [00,23]. • h: the hour of the half day. A value of 1 indicates the first hour of the half day. Valid values of h: [1,12]. Valid values of hh: [01,12]. • K: the hour of the half day. A value of 0 indicates the first hour of the half day. Valid values of K: [0,11]. Valid values of KK: [00,11]. • k: the hour of the current day. A value of 1 indicates the first hour of the day. Valid values of k: [1,24]. Valid values of kk: [01,24].	HH	10
Minute	m: the minute of the hour. Valid values of m: [0,59]. Valid values of mm: [00,59]	m	27
Second	• s: the second of the minute. • S: the millisecond of the minute.	ss	08
Time zone	z: the time zone.	z	UTC+08:00

Sample time formats

Sample format	Sample value
`yyyy-MM-dd`	2021-08-12
`yyyyMM01`	20210801
`HH:mm:ss`	11:05:21
`yyyyMMdd HH:mm:ss`	20210812 11:05:21

16.1.8.1.3. Publish task flows

After you configure or modify a task flow, you must publish the latest task flow. This effectively prevents the modified task flow from being published before the modifications to the task flow are confirmed.

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Factory > Task Orchestration (New)**.
3. Click the name of the task flow that you want to manage to go to the task flow details tab.
4. (Optional)Try run the task flow.
   i. In the upper-left corner of the canvas, click **Try Run**.

   You can also click the icon on the right side of Try Run and select a try run method.

   - **Run at a Specific Point in Time**: In the **Run at a Specific Point in Time** dialog box, select a date and time, and click **OK**.
   - **Run at a Specific Time Range**: In the **Run at a Specific Time Range** dialog box, set the start time and end time, set the scheduling cycle, and then click **OK**.

   ii. In the **Alert** message, click **OK**.

   iii. Click the blank area of the canvas. In the low part of the task flow details tab, click the **Execution Logs** tab and check whether the try run of the task is successful.
   - If the last line of the execution logs contains `status SUCCEEDED`, the try run of the task flow trial is successful.
   - If the last line of the execution logs contains `status FAILED`, the try run of the task flow fails.

   > **Note** If the try run fails, you can find the failed node and the reason in the execution logs. Modify the node configurations and try again.

5. Publish the task flow.
   i. In the upper-left corner of the canvas, click **Publish**.
   ii. In the **Publish** dialog box, enter remarks information and click **OK**.

   > **Note** If you no longer need to use the task flow, you can unpublish the task flow by performing the following steps:
   >
   > i. In the upper-left corner of the canvas, click **Unpublish**.
   >
   > ii. In the message that appears, click **OK** to unpublish the task flow.

6. Check whether the task is published.

i. In the upper part of the canvas, click **Go to O&M**.

ii. On the right side of the page, view the **Released** parameter to check whether the task is published.

- **Published**: The task flow is in published.
- **Not published**: The task flow is not published.

16.1.8.2. Data warehouse development

16.1.8.2.1. Overview

Data Management (DMS) provides the data warehouse development feature. This feature uses databases as the computing engine and integrates a variety of tools and services in the database ecosystem. This allows you to develop and manage data warehouses with ease. This feature is designed to provide you with a one-stop development platform for data integration, processing, visualization, and value mining.

Benefits

- A variety of data warehouse engine types

  You can choose a data warehouse engine type based on your enterprise scale, data volume, and requirement for real-time performance. For example, you can choose AnalyticDB for MySQL or ApsaraDB RDS for MySQL as the data warehouse engine type.

- Two development modes

  The data warehouse development feature of DMS provides two development modes: task orchestration and professional development. The two modes meet different business requirements.

  - Task orchestration: This development mode allows you to develop a data warehouse by creating task flows and writing SQL scripts for task nodes. You do not need expertise in data warehouse development. You need only to focus on your business logic.
  - Professional development: This development mode meets the requirements of professional warehouse developers. It provides capabilities such as theme management, hierarchical management, production, release, multi-person collaboration, and data quality control. These capabilities empower professional warehouse development solutions for your enterprise.

    > **Note** Some of the capabilities are planned to be supported soon.

- Support for offline and real-time data warehouses

  The data warehouse development feature supports offline data synchronization and task scheduling. This allows you to develop offline data warehouses with ease in DMS. In addition, DMS is integrated with Data Transmission Service (DTS) and cloud-native data warehouses. This allows you to build a real-time data warehouse system based on the real-time synchronization feature of DTS and cloud-native data warehouse engines. Then, you can develop data and consume data in real time in DMS.

- Unified management of online and offline data

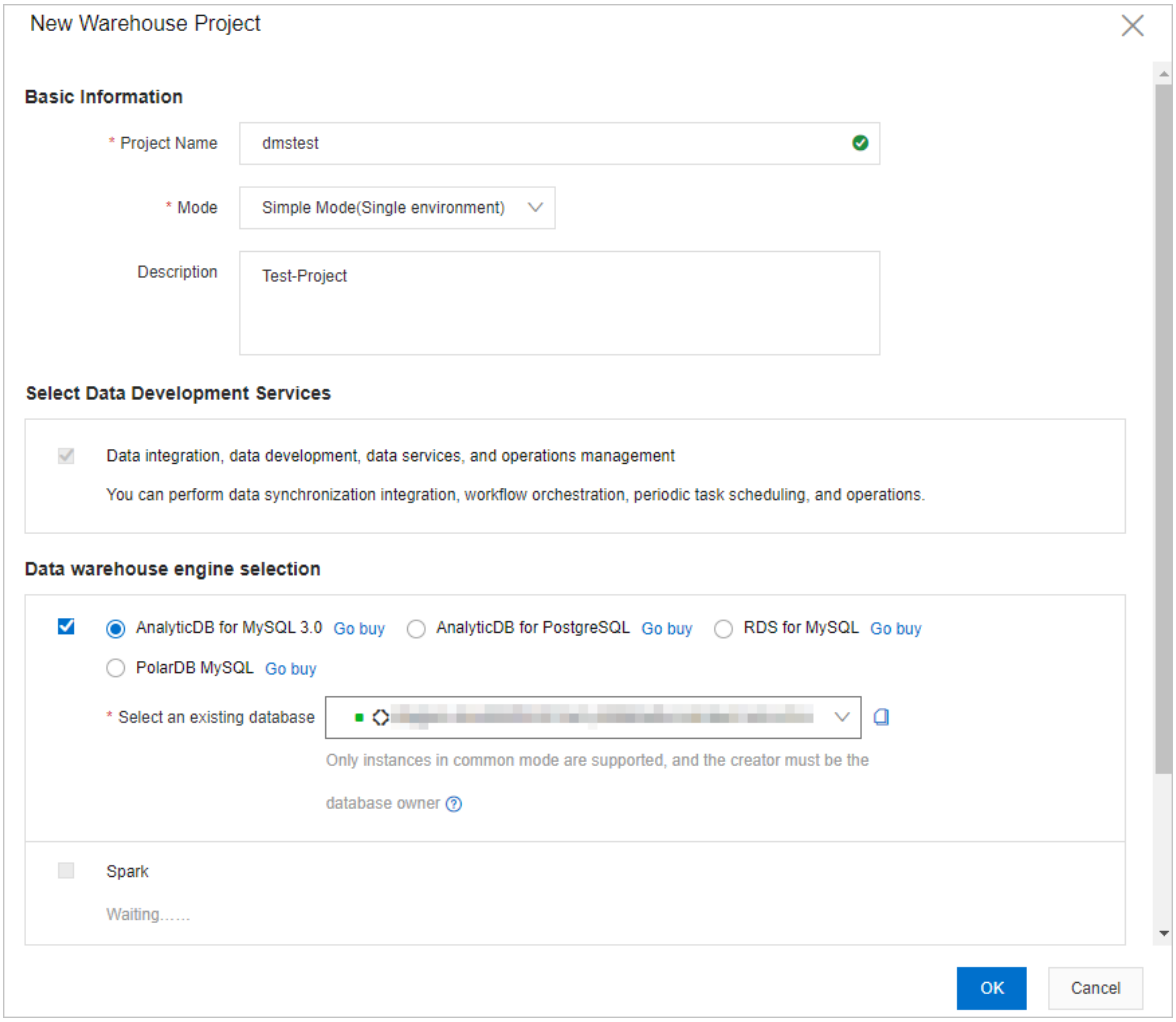
  DMS supports unified database management and permission management. You can manage your online transaction processing (OLTP) databases and online analytical processing (OLAP) databases in a centralized manner in DMS. This avoids security issues that are caused by the isolation between offline and online systems.

16.1.8.2.2. Create a data warehouse project

Before you can use the data warehouse development feature of Data Management (DMS), you must create a data warehouse project and select a data source for data warehouse development. This topic describes how to create a data warehouse project.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Factory > Data Warehouse Development**.
3. On the left-side navigation submenu, click the icon.
4. Click the icon to the right of **Data warehouse**.
5. In the New Warehouse Project dialog box, set the parameters as required.



Section	Parameter	Description
Basic Information	**Project Name**	The name of the project. Specify a descriptive name for easy identification.
	Mode	The mode of the project. Set this parameter to **Simple Mode(Single environment)**. This way, you can use the same database in a development environment and a production environment.
	Description	The description of the project.
Select Data Development Services	N/A	DMS automatically completes the configuration in this section.

Section	Parameter	Description
Data warehouse engine selection	Select a type of data warehouse engine.	After you select a data warehouse engine, such as **AnalyticDB for MySQL 3.0**, select a database from the **Select an existing database** drop-down list. **Note** Only the databases of the instances that are managed in Security Collaboration mode are available in the drop-down list. You must be the owner of the selected database.

6. Click **OK**.

What's next

Create or import an internal table

16.1.8.2.3. Create or import an internal table

After you create a data warehouse project in Data Management (DMS), you must create or import an internal table for the project. An internal table refers to a table that exists in the data warehouse engine. This topic describes how to create or import an internal table.

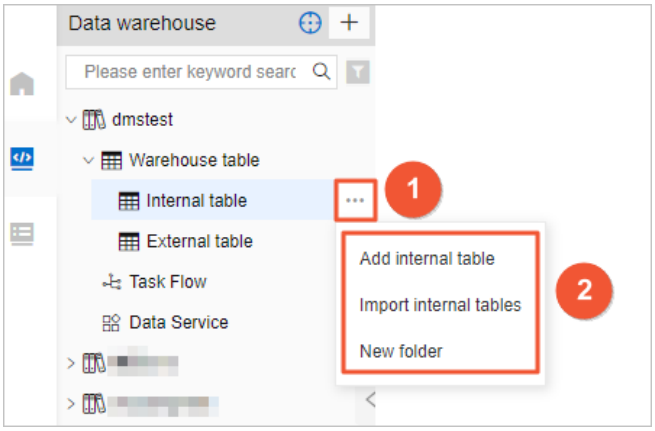
Prerequisites

You have the change permissions on the database for which you want to create or import an internal table. For information about how to apply for permissions, see Apply for permissions.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **Data Factory > Data Warehouse Development**.
3. On the left-side navigation submenu, click the icon.
4. In the left-side navigation pane, expand a data warehouse project and then expand Warehouse table. Move the pointer over **Internal table** and then the More icon that appears. Then, select an option from the menu to perform one of the following operations.

   Note You cannot configure external tables in DMS.

   - Create an internal table:
     a. Select **Add internal table**.

b. On the tab that appears, enter an SQL statement to create a table.

c. Click **Execute**.

- Import an internal table:

> **Note** The data warehouse development feature does not support real-time synchronization of tables that are created by using other means such as a command-line tool. You can import such tables to data warehouse projects in DMS.

a. Select **Import internal tables**.

b. In the Import internal tables dialog box, select the table that you want to import from the **Choose table** drop-down list and enter a description in the Remarks field.

c. Click **OK**.

- Create a folder:

> **Note** If you have a large number of tables, you can use folders to organize and classify the tables.

a. Select **New folder**.

b. In the New folder dialog box, enter a folder name.

c. Click **OK**.

What's next

Manage task flows

16.1.8.2.4. Manage task flows

Data Management (DMS) supports task flows and timed scheduling. You can configure a variety of task nodes in task flows. This can meet your requirements for data archiving, integration, and processing.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **Data Factory > Data Warehouse Development**.
3. On the left-side navigation submenu, click the  icon.
4. In the left-side navigation pane, expand a data warehouse project. Move the pointer over **Task Flow** and then the

   icon that appears.
5. In the New Task Flow dialog box, enter a name and the description for the task flow.
6. Click **OK**.
7. On the tab that appears, configure task nodes for the task flow.

> **Note** The configurations of a task flow in professional development mode are basically the same as the configurations of a task flow in task orchestration mode. For more information, see Step 5 in the *Task orchestration* topic.

16.1.8.2.5. Use the data service feature

In Data Management (DMS), the data warehouse development feature is integrated with the data service feature. The data service feature allows you to export the data that is managed by DMS. This feature is applicable to scenarios in which you need to export data at the column or row level, visualize data, or perform complex analysis.

Limits

When you use the data service feature to create an API for a data warehouse, the data source of the API must be a table in the data warehouse project.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Data Factory > Data Warehouse Development**.

   > **Note** You can also choose **More > Data Factory > Data Service**.

3. On the left-side navigation submenu, click the </> icon.
4. In the left-side navigation pane, expand the required data warehouse project and double-click **Data Service**.

Configure an API

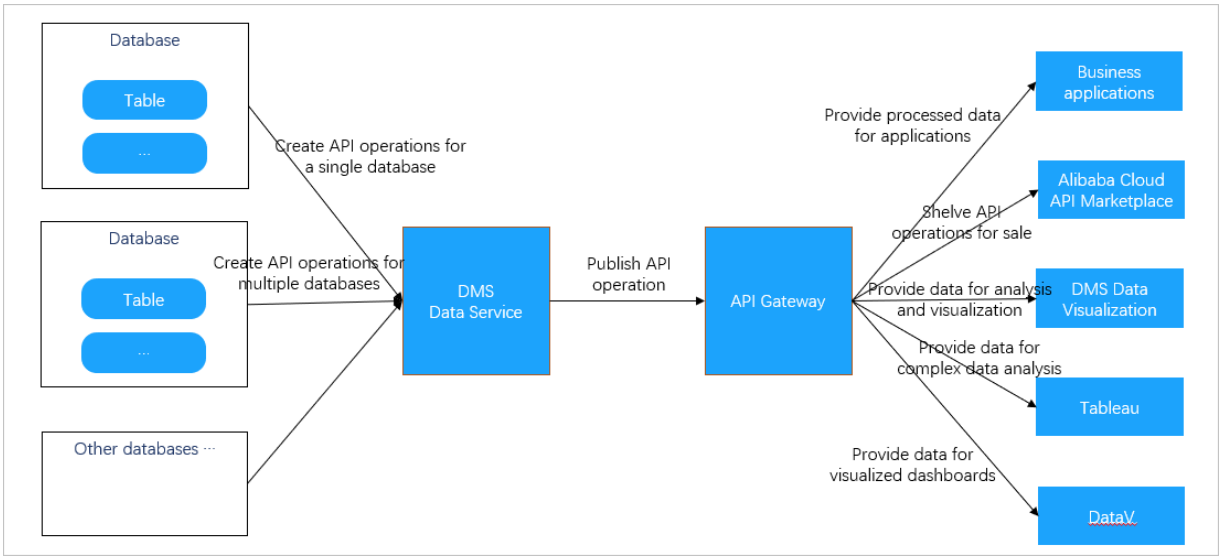
For information about how to configure an API, see Develop an API, Unpublish or test an API, Test an API, and Call an API.

16.1.8.3. Data service

16.1.8.3.1. Overview

Data Management (DMS) provides the data service feature, which allows you to export the data that is managed by DMS. This feature is applicable to scenarios where you need to export data at the column or row level, display data in a visualized manner, or perform complex analysis.

Features



- You can use the data service feature to create APIs that can be called to access the data that is managed by DMS. When you create the APIs, you can apply the security control features that are used for SQL execution in the SQLConsole, such as permission control and data de-identification.
- The data service feature works based on a serverless architecture. This feature frees you from the concern about

the infrastructure of the runtime environment, such as servers and networks. You need to focus only on how to create APIs and design data query logic. This avoids operations and maintenance (O&M) overheads that are generated by using traditional architectures.

- The data service feature is fully integrated with API Gateway. You can use this feature to publish APIs to API Gateway. This way, you can use all the features that are provided by API Gateway, such as API permission control, IP address-based access control, throttling, metering and billing, and SDKs.

Scenarios

Scenario	Description
Minimize data exposure	Assume that you need to export the data that is managed by DMS to an external environment. In this case, APIs can be called to export the data of specific rows or columns to the external environment. To export the data of specific rows, specify a filter condition in the SQL statement. To export the data of specific columns, specify the columns in the SQL statement. Compared with data export of a whole table, this minimizes data exposure and ensures data security.
Connect visualization tools to databases	Most visualization tools can connect to databases by calling APIs. You can connect a visualization tool to your database by calling an API, instead of by using a username and a password. This method is easy to implement and avoids account exposure.
Sell APIs in the Alibaba Cloud Marketplace	If you want to provide paid or free data for other users, publish an API to the Alibaba Cloud Marketplace.
Provide processed data for applications	After data is processed and summarized by using the data warehouse development feature of DMS, APIs can be created and provided for applications to read the processed data from DMS to meet business needs. To modify the logic of data reading, you need only to modify the query logic of the required API without the need to republish the application.

16.1.8.3.2. Develop an API

The data service feature of Data Management (DMS) allows you to develop APIs with ease. This topic describes how to create and manage APIs.

Prerequisites

API Gateway is activated. For more information, see the documentation of *API Gateway*.

Context

The data service feature allows you to export the data that is managed by DMS. This feature is applicable to various scenarios. These scenarios include data export at the column or row level, data visualization, or complex data analysis. For more information, see Overview.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Data Factory > Data Service**.
3. In the left-side navigation pane, click the **API Development** tab.
4. On the APIManagement tab, click **New API** in the upper-right corner.
5. On the tab that appears, set the required parameters.

i. Set the required parameters on the AttributeConfiguration tab.

Parameter	Description
APIName	The name of the API. The name must be 4 to 100 characters in length, and can contain letters, digits, and underscores (_). The name must start with a letter.
Description	Optional. The description of the API. Enter an informative description, for example, a description of the data that you want the API to return or the scenarios in which the API can be called.
Path	The path of the API. The path must start with a forward slash (/) and can contain letters, digits, underscores (-), and hyphens (-). The specified path forms a part of the URL that is used to call the API. A URL that is used to call an API must be in the `https://{Domain name}{Path}` format. For example, if the domain name is `xxxx-cn-hangzhou.alicloudapi.com` and the path is `/item/monthly_data`, the URL that is used to call the API is `https://xxxx-cn-hangzhou.alicloudapi.com/item/monthly_data`.
ReturnFormat	The format in which you want the API to return data. Valid value: **JSON**.
RequestMode	The request method. Valid values: **POST** and **GET**.
TimeOut (MS)	The maximum period of time that the system can wait until an API request expires. Unit: milliseconds. If the execution time of an API exceeds the specified timeout period, the system returns a timeout error. Maximum value: 30000.
Returns the maximum number of records	The maximum number of entries that can be returned for an API request. This parameter limits the number of entries that can be returned for each API query. **Note** If the database instance is managed in security collaboration mode, the value of this parameter must be less than the maximum number of entries that is specified in the security rules.

ii. Click the **ExecuteConfiguration** tab and set the required parameters.

Parameter	Description
Instance query type	■ **Single InstanceQuery**: You can call the API to read data from only one database instance. ■ **Cross-instanceQuery**: You can write dynamic SQL statements for the API to query data across multiple database instances. **Note** If you set this parameter to **Cross-instanceQuery**, you need only to enter dynamic SQL statements in the **QuerySQL** field.
Data source	The database that is queried by the API. You can search for databases on which you have query permissions by keyword and then select a database.

Parameter	Description
ConfigurationMode	■ **Table boot mode**: You can configure data query by selecting a table and fields. ■ **Script mode**: You must configure a data query by specifying variables and writing SQL statements. **Note** After you set this parameter to **Script mode**, you need only to enter SQL statements in the **QuerySQL** field.
SelectTable	The table to be queried. You can search for tables by keyword.
FieldList	The fields in the selected table. You can specify the required fields as request parameters or response parameters.
Script mode	The mode in which an SQL script is written to define the data query logic. **Note** You can set the **ConfigurationMode** parameter to **Table boot mode** or Script mode.
QuerySQL	The SQL statement that is used to query the data in the table. After you enter an SQL statement, click **ParsingScript** to verify the syntax and to parse the request parameters and response parameters. **Note** ■ Custom variables are supported. Custom variables can be mapped as request parameters in API requests. The variables of an SQL statement must be specified in the `${Variable name}` format. For example, the `select item_id, item_name from ex_item where category=${category}` SQL statement include a variable named `category`. ■ If you set the **Instance query type** parameter to **Cross-instanceQuery**, you must use the syntax of cross-database query SQL statements. For more information, see Cross-database query.

iii. Click the **Request Parameters** tab and set the required parameters.

Parameter	Description
ParametersName	The name of the request parameter. ■ The name can contain letters, digits, hyphens (-), and underscores (_). ■ The name must start with a letter or an underscore (_). ■ The name must be 1 to 50 characters in length.
FieldName	The name of the field that is specified by the request parameter. The field name is specified on the **ExecuteConfiguration** tab and cannot be changed.
Cannot be empty	Specifies whether the request parameter is required.
Description	The description of the request parameter.
Data type	The data type of the request parameter. The data type is used to check whether the value of the request parameter in an API request is valid. Valid values: String, Integer, and Floating point. Default value: String. **Note** This parameter affects the SQL statement that is executed when the API is called.
Example value	The sample value of the request parameter. You can use the sample values that are provided in SDKs and documentation as references when you call API operations.
Default value	The default value of the request parameter. If the request parameter is optional and not specified in the API request, the default value is used.

iv. Click the **Return parameter** tab and set the required parameters.

Parameter	Description
ParametersName	The name of the response parameter. ■ The name can contain letters, digits, hyphens (-), and underscores (_). ■ The name must start with a letter or an underscore (_). ■ The name must be 1 to 50 characters in length.
FieldName	The name of the field that is returned. The name cannot be changed.
Description	The description of the response parameter.
Data type	The data type of the response parameter. Valid values: String, Integer, and Floating point. Default value: String. This parameter is used by DMS to convert the type of the data in API responses. This parameter affects the JSON data that is returned.
Example value	The sample value of the response parameter. You can use the sample values that are provided in SDKs and documentation as references to help you understand API responses.

6. Click **Save**.
7. In the left-side navigation pane, click the **API Development** tab.
8. Perform the following operations to manage the API based on your business requirements:

- Publish the API

  On the APIManagement tab, find the required API and click **Publish** in the **Operation** column of the API. In the message that appears, click **OK**.

- Modify the API

  On the APIManagement tab, find the required API and click **Modify** in the **Operation** column of the API. Modify the configurations of the API based on the descriptions in Step 5 and click **Save**.

- Delete the API:

  On the APIManagement tab, find the required API and click **Delete** in the **Operation** column of the API. In the message that appears, click **OK**.

16.1.8.3.3. Unpublish or test an API

This topic describes how to unpublish or test an API that has been published.

Prerequisites

An API is created. For more information, see Develop an API.

Procedure

1. Log on to the DMS console.
2.
3. Click the **API Publish** tab on the left side.
   The APIPublishList tab displays all the published APIs.
4. Find the API that you want to manage and perform the following operations based on your business requirements:
   - Unpublish the API:

     Click **Offline** in the **Operation** column. In the message that appears, click **OK**.

   - Test the API:

     Click **Test** in the **Operation** column. For more information, see Test an API.

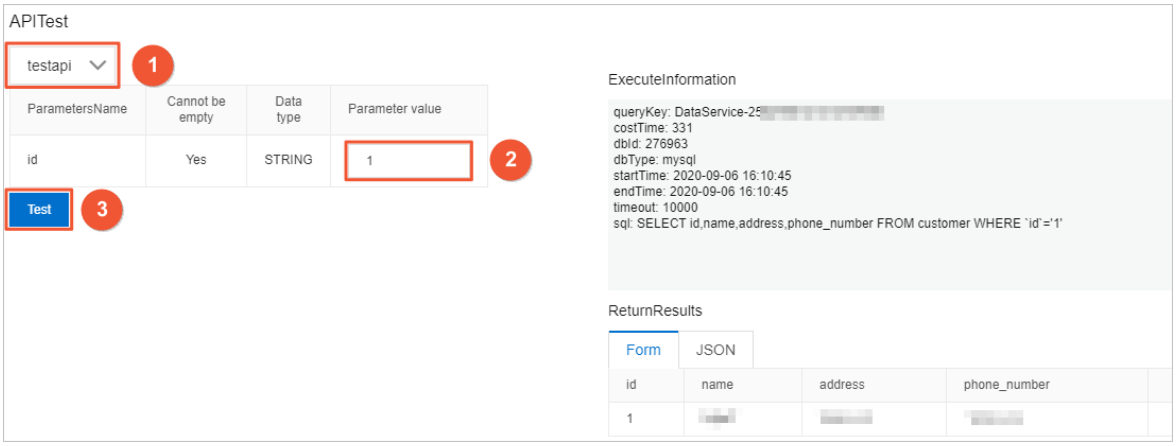
16.1.8.3.4. Test an API

After you create an API, you can test the API to verify whether the API meets your business requirements.

Prerequisites

Procedure

1. Log on to the DMS console.
2.
3. Click the **API Test** tab on the left side.
4. On the APITest tab, test an API.



i. Select the API that you want to test from the drop-down list.

ii. Enter values in the Parameter value column.

iii. Click **Test**.
After the test is complete, the execution information and return results appear on the right side. You can evaluate whether the API meets your business requirements based on the information.

> **Note** You can click the **JSON** tab in the **ReturnResults** section so that the return results are displayed in the JSON format.

16.1.8.3.5. Call an API

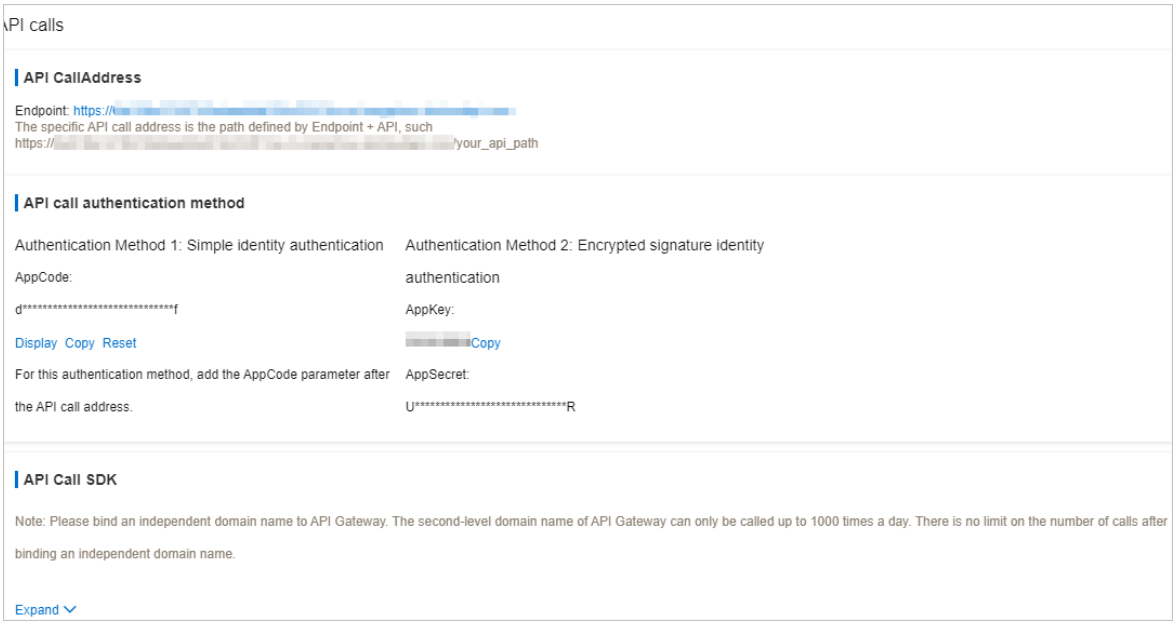
After you create, publish, and test an API, you can call the API in an application by using an SDK.

Prerequisites

- An API is created and published. For more information, see Develop an API.
- API Gateway is activated. For more information, see the documentation of *API Gateway*.

Procedure

1. Log on to the DMS console.
2.
3. Click the **API Call** tab on the left side.
4. View the API call address and the authentication information.



- **Simple identity authentication**: requires only an AppCode. This authentication method is suitable for calling APIs by using URLs. This authentication method has a low security level and is generally used in scenarios in which data visualization is involved, such as calling APIs in DataV.
- **Encrypted signature identity authentication**: requires an AppKey and an AppSecret, which are used to dynamically generate an encrypted signature for calling an API. This authentication method has a high security level.

5. Call the API in an application by using an SDK.

> **Note** For more information about how to call an API in an application by using an SDK, see the documentation of *API Gateway*.

16.1.8.4. Data visualization

16.1.8.4.1. Overview

This topic introduces the basic concepts, design philosophy, and scenarios of the data visualization feature of Data Management (DMS).

Background information

DMS allows you to manage databases and query data in the SQLConsole where results are returned in the form of a table. However, if you want to analyze business characteristics in scenarios such as trend analysis and growth comparison, tables cannot meet the requirements, and data visualization is required. To resolve this issue, DMS provides the data visualization feature. You can use this feature to gain insights into your business and make better business decisions.

Basic concepts

The data visualization feature provides a three-layer model for you to visualize data in various forms, including datasets, charts, and dashboards or big screens. You can execute SQL statements in the SQLConsole to obtain datasets and convert the datasets to common charts such as line charts, pie charts, column charts, circular charts, table charts, dual Y-axis charts, and funnel charts. Then, on a dashboard or big screen, you can freely combine and lay out these charts based on your analysis logic or methodology to visually present your business data.

> **Note** For example, you can use indicator cards to display the overall metrics of your business, such as a transaction volume and unique visitors (UVs). Then, you can use a line chart to present the growth trend of the transaction volume and a column chart to compare transactions in all regions. Finally, you can use a table chart with a filter to query region-specific data.

Design philosophy

Two core concepts of data visualization are datasets and charts. Datasets are also called data views, and charts are also called visualization components.

- Datasets represent the structured form of data. Data logic, permissions, and services are all based on this form.
- Charts represent the visual form of data. Data presentation, interaction, and guidance are all based on this form.

  > **Note** Datasets and charts complement each other to provide the same data in two different forms and help you better understand data.

- Dashboards or big screens are used for quick data analysis and custom data visualization. You can combine charts on dashboards or big screens as needed. This can satisfy the data visualization needs of most users.

Scenarios

- Analyze data in a secure and custom manner

  The data visualization feature is based on the security control feature in DMS. This ensures that data is authorized before it is visualized.

  - You can set the configurations only once to implement the advanced filtering, advanced control, interaction, drilling, download, and sharing of visual components. This facilitates data analysis and decision-making. For example, you can use this feature to compare data and analyze the geographic information of data, data distribution, data trends, and data clusters with ease.
  - Dashboards use automatic layouts. They can be used for most visual reports that require simple configuration and need to be viewed and shared with ease.
  - Big screens use custom layouts. They can be used for specific visual reports that require additional modifier elements and need to be retained for a long period of time. Time and efforts are required to configure a big screen in these scenarios, such as a big screen for massive online promotions.

- Monitor operations in real time

  In the data factory of DMS, you can synchronize your business data in real time to AnalyticDB or ApsaraDB RDS where data can be analyzed. Then, you can visualize data that is analyzed in AnalyticDB or ApsaraDB RDS. This way, you can monitor database performance in real time and make sure that data flows are seamlessly connected. In addition, you can compare data to detect anomalies and handle key-link issues. Pivot-driven mode and chart-driven mode are provided for chart configuration. You can apply these modes to different scenarios based on your business requirements.

16.1.8.4.2. Terms

This topic describes the terms related to the data visualization feature of DMS.

Term	Description
dimension	A dimension represents an attribute of business data, such as time, region, gender, and category. A dimension contains a collection of discrete values, based on which a measure is obtained.
measure	A measure is a statistical value that is obtained from an aggregation operation. For example, unique visitor (UV) and transaction volume are both measures.

Term	Description
dataset	A dataset is a collection of data in the form of a two-dimensional table. The data is generated after an SQL statement is executed to query a database. Therefore, you must prepare a database and an SQL statement to obtain a dataset.
chart	A chart is a graph that visualizes data to present a data feature. For example, a line chart shows a data trend, a table chart displays detailed data, a bar chart compares data, and a pie chart highlights percentages. Different charts may need different numbers of dimensions and measures. For example, a line chart requires one dimension and one or more measures.
dashboard	A dashboard is a visualization tool where multiple charts are combined to present business data in a comprehensive way. Charts can be laid out more flexibly on dashboards than in traditional visual reports. You can divide a dashboard into sections and adjust the size and position of the chart in each section. This can optimize the dashboard layout and provide user-friendly interactions. You can also configure a global filter that allows you to filter data across charts on a dashboard and display data that is queried.
dashboard collection	A dashboard collection is used to manage a group of dashboards that are related to each other. For example, you can create a dashboard collection that is specific to products and put the Products Sold dashboard, the Products Added to Shopping Cart dashboard, and the Products Returned dashboard to this dashboard collection. This provides convenient and unified management. In a dashboard collection, you can create recursive directories to classify the dashboards.
big screen	On a big screen, you can combine multiple charts as you can on a dashboard. You can also use auxiliary graphics, such as images and rectangles, on a big screen. This allows you to create layouts in a more flexible way. Different from dashboards that you can divide into sections, big screens adopt an absolute positioning layout. You can freely drag charts and auxiliary graphics on a big screen. This meets the requirement for more flexible data visualization.

16.1.8.4.3. Go to the Data Visualization tab

This topic describes the methods that you can use to go to the Data Visualization tab.

Go to the Data Visualization tab in the DMS console

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Factory > Data Visualization**.

Go to the Data Visualization tab from the SQL Console tab

1. Go to the **SQL Console** tab. For more information, see SQLConsole.
2. At the top of the **SQL Console** tab, click **Data Visualization**.

16.1.8.4.4. Manage datasets

A dataset is a collection of data in the form of a two-dimensional table. The data is generated after an SQL statement is executed to query a database. This topic describes how to manage datasets.

Create a dataset

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. On the **Dataset management** tab, click the + icon.

4. In the **Writing SQL** step, set the parameters and click **Execute**.

Parameter	Description
Name	The name of the dataset. The dataset name must be unique for each Data Management (DMS) user.
Description	The description of the dataset.
Select an existing database	The database to be queried. You must have permissions to query the database.
Search table/field name	The name of the table or field that you want to search for and select.

5. After the SQL statement is executed, click **Next Step**.
6. In the **Edit dataset model** step, set the **Data type** and **Visualization type** parameters for each field based on your requirements.

   In this step, you must specify each queried field as a dimension or measure and specify a visualization type for each field.

Parameter	Description
Data type	Valid values: ○ **Dimension**: the scope, aspect, or angle of measures. ○ **Measure**: the statistical value that is obtained after an aggregation operation. To show how transaction volume changes over time, you can set the Data type parameter for the time field to **Dimension** and that for the transaction volume field to **Measure**. For more information, see Terms.
Visualization type	Valid values: ○ **Digital** ○ **String** ○ **Date** ○ **Geography: Country** ○ **Geography: Provinces** ○ **Geography: City** To show how transaction volume changes over time, you can set the Visualization type parameter for the time field to **Date** and that for the transaction volume field to **Digital**.

7. Click **Save**. The dataset is created, and you are navigated to the Dataset management tab.

Modify a dataset

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. On the **Dataset management** tab, find the dataset that you want to modify and click the icon in the **Operation** column.
4. In the **Writing SQL** step, configure the SQL query statement and click **Execute**.

   For more information about how to set the other parameters in the **Writing SQL** step, see Create a dataset.
5. After the SQL statement is executed, click **Next Step**.

6. In the **Edit dataset model** step, set the **Data type** and **Visualization type** parameters for each field based on your requirements.

   For more information about the **Data type** and **Visualization type** parameters, see Create a dataset.

7. Click **Save**. The dataset is modified, and you are navigated to the Dataset management tab.

Delete a dataset

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. On the **Dataset management** tab, find the dataset that you want to delete and click the icon in the **Operation** column.
4. In the message that appears, click **OK**.

> **Note** Before a dataset is deleted, DMS checks whether the dataset is referenced by a chart. If the dataset is referenced, you cannot delete the dataset. To delete the dataset, you must go to the chart editing page to cancel the reference or delete the chart that references the dataset.

Grant permissions on a dataset

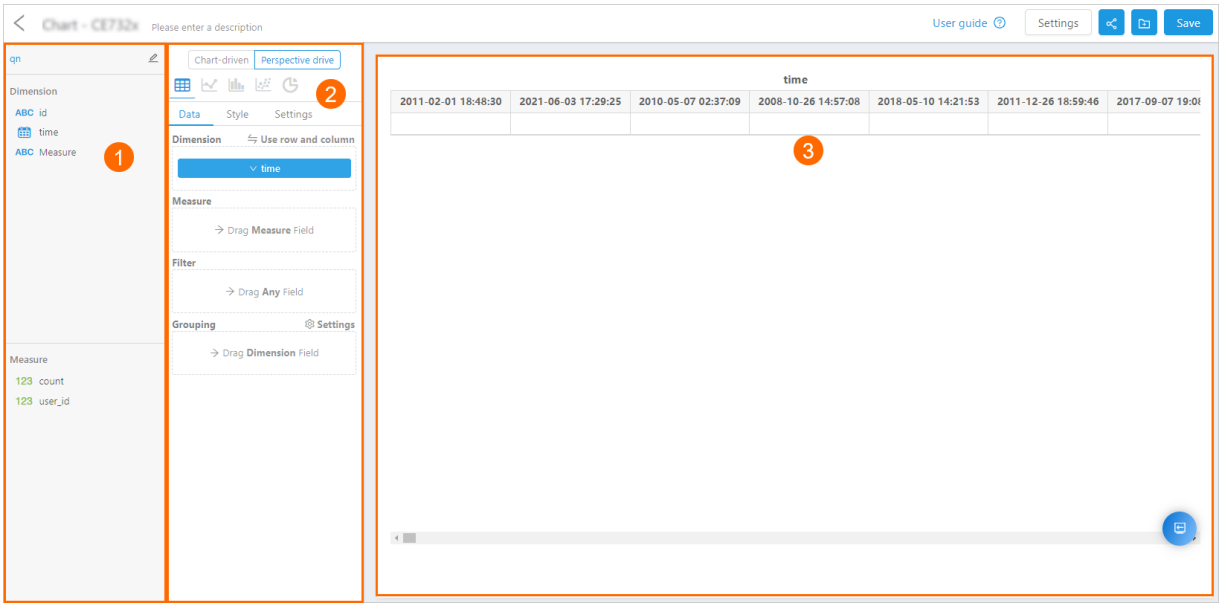
1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. On the **Dataset management** tab, find the dataset on which you want to grant permissions and click the icon in the **Operation** column.
4. In the **Share** dialog box, specify the permissions to be granted on the dataset and the users to whom you want to grant the permissions. Then, click **Authorize**.

16.1.8.4.5. Manage charts

A chart is the smallest unit for data visualization. A chart is generated by aggregating and grouping SQL query results based on the dataset model provided by a dataset, and then visualizing the processed data by using code.

Each chart must be associated with a dataset. In the chart editor, the original SQL statement and the dataset model in a dataset are used to generate a new SQL statement. Then, the new SQL statement is executed to query data that is to be visually presented in a chart. This topic describes features of charts and how to use charts.

Chart editor



The following table describes the three sections of the chart editor.

No.	Section	Description
①	Dataset model display section	After you select a dataset in the upper-left corner of this section, its dataset model will be automatically displayed below, including dimensions and measures. You can drag fields in this section to corresponding areas on the **Data** tab in the chart configuration section.
②	Chart configuration section	At the top of this section, you can select a configuration mode and a chart type. Move the pointer over a chart type icon to view the numbers of dimensions and measures required for this chart type. Below the chart type icons are three tabs. • On the Data tab, you can specify which fields in the dataset model are needed for the chart. You can drag fields from the dataset model display section to corresponding areas on this tab. • On the Style tab, you can set the display style of the chart. • On the Settings tab, you can configure functional modules for the chart, such as filters and cache.
③	Chart display section	This section displays the chart based on the dataset model and your configurations in the chart configuration section.

Configuration modes

The chart configuration section provides you with the following two configuration modes that are based on different visual presentation logic: including the **pivot-driven mode** and **chart-driven mode**. This meets different requirements in various scenarios.

Configuration mode	Description	Scenario
Chart-driven mode	The chart-driven mode represents the general visual presentation logic that is based on chart classification. Various charts can be configured by using the chart-driven mode. In this mode, dimensions and measures can be regarded as fixed configuration items, along with style settings, for configuring a chart.	The chart-driven mode is applicable to most data visualization scenarios.

Configuration mode	Description	Scenario
Pivot-driven mode	The pivot-driven mode represents the visual presentation logic that is based on pivot tables. A chart configured by using the pivot-driven mode can be regarded as the visualization of a pivot table through coding. Dimensions and measures in the pivot table are converted to axes in the chart for graphic display. In this mode, you can configure different graphic coding for each measure. The lowest-level dimension can be used as a common dimension axis.	The pivot-driven mode is applicable to scenarios in which a small amount of data needs to be freely analyzed on clients.

Data configuration

To complete data configuration, you must drag fields from the dataset model display section to corresponding areas on the **Data** tab in the chart configuration section. When you are dragging a field, the areas where the field can be placed are highlighted on the Data tab.

Dimension

Only categorical fields can be placed in the Dimension area. Values of each field in this area will be grouped in the new SQL statement to be generated.

Measure

- Only fields of the NUMERIC type can be placed in the Measure area. Values of each field in this area will be aggregated in the new SQL statement to be generated. To specify how a field will be aggregated, you can click the field name and select an aggregate function. The following six aggregate functions are supported:
  - sum
  - avg
  - count
  - count_distinct
  - max
  - min
- You can specify the following data formats for fields in the Measure area:
  - Default: the default format.
  - Numeric value: You can specify a unit, the number of decimal places to which numbers are rounded, and whether to use the thousands separator.
  - Currency: You can specify a unit, the number of decimal places to which numbers are rounded, whether to use the thousands separator. You can also prefix and suffix texts.
  - Percentage: You can specify the number of decimal places to which numbers are rounded.
  - Scientific type: You can specify the number of decimal places to which numbers are rounded.

Filter

Both dimension and measure fields can be placed in the Filter area. Values of fields in this area will be used to specify filter conditions in the SQL statement to be generated. Data can be filtered by values, conditions, and dates.

> **Note** Filter methods applicable to each field in this area correspond to the visualization type that you specified for the field when you configure the dataset model.

Filter method	Description	Visualization type

Filter method	Description	Visualization type
By conditions	Filtering by conditions is the most flexible filter method. In the Filter configuration dialog box, you can specify filter conditions for the selected field as needed. You can also use the logical AND and logical OR operators to specify multiple conditions.	• Numeric value • String • Geographical value
By values	In the Filter configuration dialog box, a set of distinct values of the selected field are listed in the left-side section. You can select values from this section and add them to the right-side section to filter data.	• String • Geographical value
By dates	You can query the selected field in a specified period of time. The period of time can be fixed or dynamic.	Date

Grouping

You can place only one dimension field in the Grouping area. Values of the field in this area will be grouped in the new SQL statement to be generated. In the chart to be generated, values of the field will be grouped. Each group has a distinct color. The legend of the chart will help you differentiate groups with different colors. If you want measures of a dimension to be displayed as several groups in a chart, you can use color settings to differentiate the groups.

For example, a dataset model contains the trade_date, zone, and price fields, respectively representing the transaction date, transaction region, and transaction volume. You want a bar chart to show aggregated transaction volumes of each region by day. In this case, you can place the trade_date field in the Dimension area, the zone field in the Grouping area, and the price field in the Measure area.

> **Note** When you generate a pie chart, you require a dimension field whose values are to be grouped. In this case, you must place the dimension field in the Grouping area instead of the Dimension area.

General settings for fields

- Field alias

  Click a field name and select **Field settings**. In the **Field settings** dialog box, specify a field alias. You can specify the following two types of field aliases:

  - Permanent alias.
  - Dynamic alias: Dynamic aliases are generated by writing code in JavaScript and can be used with variables. DMS provides the Moment.js library that can be used with variables to dynamically generate aliases for fields of the DATE type. Dynamic aliases can only be generated for fields in table charts.

- Field description

  Click a field name and select **Field settings**. In the **Field settings** dialog box, enter a description. You can add field descriptions only for charts that are configured by using the **chart-driven mode**.

- Field sorting

  Click a field name, select **Sort Type** or Sort, and then select a sorting method. The following table describes the available sorting methods.

Sorting method	Description
Default	Do not sort the field values.
Ascending order	Sort the field values in ascending order in the SQL statement to be generated.

Sorting method	Description
Descending order	Sort the field values in descending order in the SQL statement to be generated.
Customize	You can drag field values to sort them. In the chart to be generated, the field values will be displayed in the specified order. Only categorical fields support this method.

Other configurations

- Size: If you are configuring a scatter chart, you must place a field of the NUMERIC type in the Size area. The field is used to code points in terms of size. Values of the field will be aggregated in the SQL statement to be generated.
- Prompt information: You can set prompts only for certain Cartesian charts. Only fields of the NUMERIC type can be placed in the Prompt information area. Values of each field in this area will be aggregated in the SQL statement to be generated.
- When you configure a chart by using the pivot-driven mode, you can place fields of any type in the Label area. Values of each categorical field in this area will be grouped and values of each field of the NUMERIC type will be aggregated in the SQL statement to be generated.
- When you configure a scatter chart by using the pivot-driven mode, you must place a field of the NUMERIC type in the x data axis area. The field is specified as the measure of the X-axis. Values of the field will be aggregated in the SQL statement to be generated.
- When you configure a dual Y-axis chart by using the chart-driven mode, you can specify a measure for the left Y-axis and a measure for the right Y-axis.

Chart configuration

After you configure fields, you can select a chart type by clicking the corresponding icon at the top of the chart configuration section.

> **Note** You can move the pointer over an icon to check the prerequisites of the chart type. If your data configuration does not meet all the prerequisites, the icon is dimmed. Only after all the prerequisites are met does the icon become highlighted. Click the icon and a chart appears in the chart display section.

On the **Style** tab of the chart configuration section, you can customize the chart display style.

Functionality settings

On the **Settings** tab of the chart configuration section, you can configure a filter and cache and specify whether to automatically load data.

- Filter

  You can create a filter or edit an existing filter by clicking **Settings** next to Filter.

  > **Note** To configure a filter, you must have defined **variables** when you configure the dataset.

  After you add a chart with a filter to a dashboard, you can click the button in the upper-left corner of the dashboard card. In the filter panel that appears, specify filter conditions and then click **Query** in the lower-right corner of the panel. Conditions that you specified in the chart filter and conditions that you specified in the global filter of the dashboard take effect at the same time.

- Cache

  You can enable or disable caching for a chart. You can also specify a validity period for the cached data.

> **Note** If you enable caching for a chart, the result of the first query for the chart on a dashboard or big screen will be stored in the cache. The SQL query statement will be used as a key. Within the validity period, DMS returns the cached result for the same SQL query statement with no need to access the data source.

- Automatically load data

  In scenarios where data query is frequent, you may not want data to be loaded immediately when you open a dashboard. In this case, you can set **Automatically load data** to **No**. This parameter is set to **Yes** by default.

Create a chart

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Chart**.
4. On the Chart management page, click the icon.
5. Configure the dataset and the chart.
6. In the upper-right corner, click **Save**.

Modify a chart

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Chart**.
4. On the **Chart management** page, find the chart that you want to modify and click the icon in the **Operation** column.
5. After the configurations of the dataset and the chart are modified, click **Save**.

Delete a chart

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Chart**.
4. On the **Chart management** page, find the chart that you want to delete and click the icon in the **Operation** column.
5. In the message that appears, click **OK**.

Duplicate a chart

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Chart**.
4. On the **Chart management** page, find the chart that you want to duplicate and click the icon in the **Operation** column.
5. In the **Copy chart** dialog box, set the **Chart name** and **Description** parameters.
6. Click **OK**.

Grant permissions on a chart

1. Log on to the DMS console.

2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Chart**.
4. On the **Chart management** page, find the chart on which the permissions you want to grant and click the icon in the **Operations** column.
5. In the dialog box that appears, select the user to which you want to grant permissions and the permissions to be granted.
6. Click the button in the dialog box.

16.1.8.4.6. Manage dashboards

Dashboards are the other type of visualization applications that are provided by DMS. Dashboards support automatic layout and provide interactive capabilities to help you create visual reports. This topic describes features of dashboards and how to use dashboards.

Create a dashboard collection

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Data Display**.
4. In the **Dashboard** section of the **Resource management** page, click **New Dashboard collection**.
5. In the **New Dashboard collection** dialog box, set the **Name** and **Description** parameters.
6. Click **Save**.

Add a chart to a dashboard

1. In a dashboard, click the icon in the upper-right corner. The New Chart dialog box appears, which contains the list of available charts.
2. Select the chart that you want to add to the dashboard and click **Next Step**.
3. Select an option from the **Data Refresh mode** drop-down list. The following modes are supported:
   - Manual refresh: You must click the **Synchronize data** icon in the upper-right corner of the chart to refresh data.
   - Scheduled refresh: The system automatically refreshes data in the chart at the specified interval in the unit of seconds.
4. Click **Save**. The chart is added to the dashboard.

Configure a chart

- To refresh data in a chart, click the icon in the upper-right corner of the chart to trigger a query and synchronize data.

  > **Note** If cache is enabled for the chart, the cached content is also refreshed when you click this icon.

- To edit a chart, click the icon in the upper-right corner of the chart.
- To view a chart in full screen, click the icon in the upper-right corner of the chart.
- To change the data refresh mode of a chart or delete a chart, click the icon in the upper-right corner of the chart and select Basic information or Delete as required.

Use the auto layout feature

- You can drag the lower-right corner of a chart to adjust its size.
- You can drag the top of a chart to adjust its position.
- Dashboards adopt a fluid layout. If the width of the display window is equal to or greater than 768 pixels, charts on a dashboard are displayed based on specified percentages. If the width of the display window is less than 768 pixels, charts are displayed in the mode for mobile devices.
- When you adjust the size or position of a chart, other charts in the same dashboard automatically adapt to the change, as designed in a fluid layout.

Configure filter interactions

You can configure filter interactions between charts in the same dashboard. Click the icon in the upper-right corner of a dashboard. The Linkage relationship settings dialog box appears.

- You can configure multiple filter interactions for a dashboard. For each filter interaction, you need to specify a trigger, an associated field, and a mapping relationship between the trigger and the associated field.
  - A trigger is a field that triggers a filter interaction. This field can be a dimension or an aggregated measure. Only fields that are used in the selected chart, instead of all fields in the dataset model that corresponds to the selected chart, can be specified as triggers.
  - An associated field can be any field or variable in the dataset model that corresponds to the selected chart. The data type of a trigger must be the same as the data type of its associated field.
- You can configure multiple filter interactions that have the same trigger. This way, one trigger is associated with multiple charts. A relationship diagram is displayed on the right of the Linkage relationship settings dialog box, showing the filter interactions between charts.

  > **Note** You can also configure multiple filter interactions whose associated fields belong to the same chart. All filter conditions, which are the mapping relationships defined in these filter interactions, take effect at the same time.

- After you configure filter interactions, each chart to which a trigger belongs has an icon in the upper-left corner. Move the pointer over the icon and an action prompt appears.

Configure global filters

You can configure global filters for a dashboard. Global filters allow you to filter data, within one or across multiple charts, by defining filter conditions or replacing variables. Click the icon in the upper-right corner of a dashboard. The Global filter settings dialog box appears.

- **Basic settings**

  The Global filter settings dialog box is divided into three sections: **Filter list**, **Associated chart** and **Category**, and **Filter configuration**.

- **Filter list**

  In the **Global filter settings** dialog box, click the plus sign (+) to the right of Filter list. A global filter is created with the default name: New Filter. To rename or delete the filter, move the pointer over the filter name and click the corresponding icon that appears to the right of the filter name.

- **Associated chart** and **Category**

  In the **Associated chart** section, select the charts that you want to associate with the global filter. Then, in the **Category** section, select at least one associated **field** or **variable** that you want to associate with the global filter. The input of the global filter is used as the value of the corresponding field to form a filter condition or used to replace the corresponding variable in the SQL query statement.

- **Filter configuration**

You can configure the following types of filters:

- **Drop-down list**

  Drop-down lists can be set only for global filters whose associated fields are dimensions. The options of a drop-down list are a set of distinct values of the associated field.

  - If the associated charts of a global filter are based on different datasets, the options of the drop-down list include values of all the associated fields in all the datasets.
  - If you set the Type parameter to Drop-down menu, the system automatically executes an SQL statement to query associated fields in the datasets and sets the values of these fields as options of the drop-down list. Therefore, you can enable cache and set a validity period.
  - If you do not want to use the values of associated fields as options of the drop-down list, select Custom options and click the plus sign (+) that appears. In the Edit custom options dialog box, enter one pair of option text and option value per line. Separate each option text and option value with a space. If the option text and option value are the same, you need only to enter one of them.

  If you select associated variables in the Category section, you still can use field values as options of the drop-down list.

  In the Filter configuration section, you can also enable the Multiple choice feature.

  > **Note** Assume that you select associated variables in the Category section and select Multiple choice for the drop-down list. If you select multiple options from the drop-down list, the selected options are converted to `'Option 1','Option 2','Option 3'` to replace the corresponding variable. Therefore, you need to include `in ()` in the SQL statement to ensure correct execution.

- **Date selection**

  Date selection can be set only for global filters whose associated fields are dimensions. The selected date can be converted to one of the following formats:

  - Date. Example: 2019-01-01.
  - Date and time, accurate to seconds. Example: 2019-01-01 12:00:00.
  - Date and time, accurate to minutes. Example: 2019-01-01 12:00.
  - Month. Example: 2019-01.
  - Week. Example: 2019-5th week.
  - Year. Example: 2019.

  You can set a default value for a date filter. The default value can be a specified date, whether the date is fixed on the timeline or moves on the timeline.

  > **Note** You can set default values only for date filters.

  If you select Multiple choice for a date filter, you can set the Date format parameter only to Date, Month, or Year. To specify multiple months or years as default values, you must select dates in the corresponding months or years.

  > **Note** Similar to drop-down list filters, if you select associated variables in the Category section and select Multiple choice for a date filter, you also need to modify the SQL statement to ensure correct execution.

- **Date range**

  Date ranges can be set only for global filters whose associated fields are dimensions. The formats to which a selected date range can be converted are the same as those for date filters. When you associate a date range filter with variables, you must select two variables, one as the start time and the other as the end time.

- **Text input box**

  Text input boxes can be set only for global filters whose associated fields are dimensions. If you set the query mode of a text input box filter to Auto query, after you enter a value in the input box, you still need to press the Enter key to trigger a query.

- **Number range input box**

  Number range input boxes can be set only for global filters whose associated fields are metrics. If you set the query mode of a number range input box filter to Auto query, after you enter a value in the input box, you still need to press the Enter key to trigger a query. When you associate a number range input box filter with variables, you must select two variables, one as the start value and the other as the end value.

- **Filter hierarchy**

  In the tree menu section, you can drag and drop filters to set a filter hierarchy. A parent filter is used as a drop-down list to filter options of its child filter.

- **Query mode**

  In the lower-right corner of the Global filter settings dialog box, you can set the query mode to Auto query or Manual query.

  - **Auto query**: Changes to the value of the filter immediately trigger a query. For text input box filters and number range input box filters, even if you set the query mode to **Auto query**, you still need to press the Enter key to trigger a query after you enter a value.
  - **Manual query**: If you set the query mode to Manual query, a **Query** button and a **Reset** button appear on the right of the global filter bar. To trigger a query, select or enter a value on the left of the global filter bar and click **Query**.

Perform drilling in a chart

To perform drilling in a chart, select the chart elements that you want to drill on, right-click, and then select the dimension that you want to drill to.

> **Note** If a chart contains a trigger configured for filter interactions, you cannot perform drilling in the chart.

- Each chart that allows drilling has an icon in the upper-left corner. If you move the pointer over the icon, an action prompt appears.
- If a chart allows drilling, its drilling path is displayed in the lower-left corner. You can drill to a level by clicking the level name in the drilling path.

  > **Note** For example, you can select the Shanghai, Shenzhen and Guangzhou chart elements to drill to the `education` field. This way, the chart displays the education level and salary distribution in Shanghai, Shenzhen, and Guangzhou.

- The drilling feature is implemented differently in charts configured by using the pivot-driven mode and in charts configured by using the chart-driven mode.
  - Pivot-driven mode

    In charts configured by using the pivot-driven mode, drilling is implemented by adding dimensions to or removing dimensions from filter conditions.

    > **Note** Pivot tables support roll-up or drill-down operations. A roll-up operation is to remove a dimension from a filter condition. A drill-down operation is to add a dimension to a filter condition. Pivot tables also allow you to drill down to a row or a column.

- Chart-driven mode
  - In table charts configured by using the chart-driven mode, drilling is implemented in a similar way it is implemented in pivot tables.
  - In other charts configured by using the chart-driven mode, a drilling operation replaces an existing dimension in a filter condition with the dimension you want to drill to. This way, data changes are presented from another perspective.

  You can perform drilling in the following types of charts configured by using the chart-driven mode:

  - Table chart
  - Column chart
  - Line chart
  - Scatter chart
  - Pie chart
  - Funnel chart
  - Secondary Y-axis chart

Edit a dashboard collection

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Data Display**.
4. In the **Dashboard** section of the **Resource management** page, click the Edit icon in the upper-right corner of the dashboard collection that you want to edit.

Delete a dashboard collection

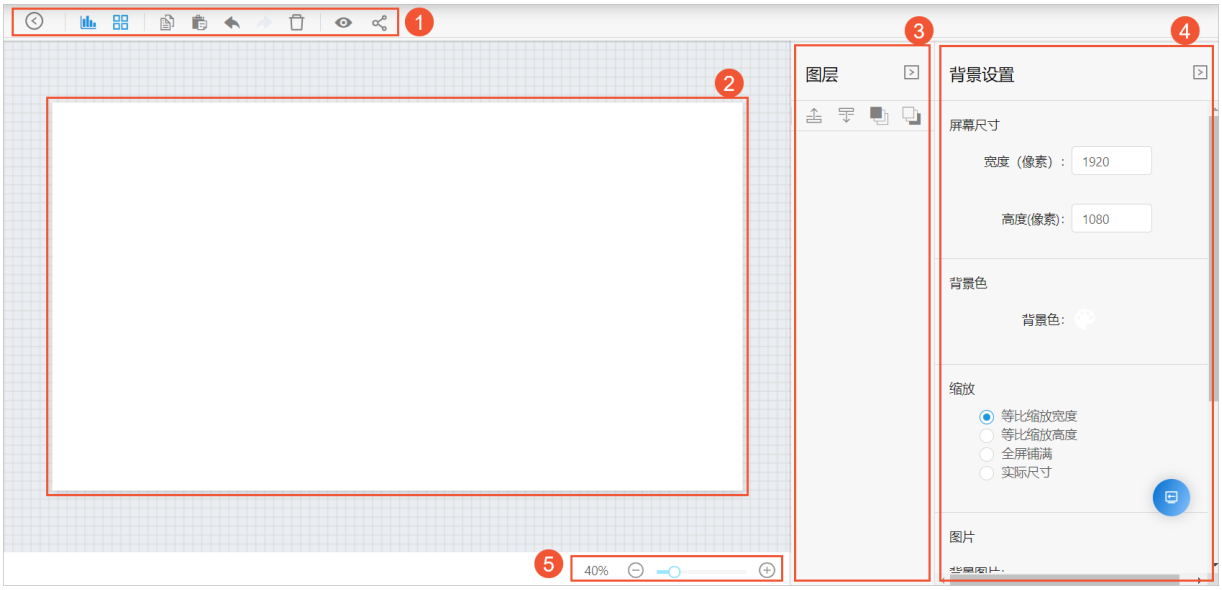
1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Data Display**.
4. In the **Dashboard** section of the **Resource management** page, find the dashboard collection that you want to delete.
5. Click the icon.
6. In the message that appears, click **OK**.

16.1.8.4.7. Manage big screens

In addition to dashboards, Data Management (DMS) provides another type of visualization applications: big screens. Big screens provide customizable layouts and styles. You can combine visualization components with bountiful built-in auxiliary graphics to create visualized big screens of various visual styles.

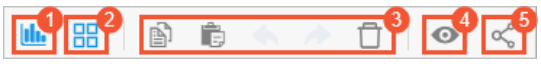
Dashboards and big screens are designed for different purposes. Dashboards are used to generate visual reports. Big screens are widely used for data viewing and presentation in a static or scrolling way. Therefore, it usually takes more time to create a big screen than to create a dashboard. This topic describes how to stylize a big screen.

Big screen editor



No.	GUI element
①	Toolbar
②	Canvas
③	Layer panel
④	Parameter settings panel
⑤	Zoom tool

Toolbar



No.	GUI element	Description
①	**Chart**	You can click this icon to add charts to a big screen.
②	**Auxiliary graphics**	You can click this icon to add the following four types of auxiliary graphics: rectangle, label, video, and timer.
③	**Operation icons**	You can click the operation icons to manage layers on the canvas. The following operation icons are provided: **Copy**, **Paste**, **Undo**, **Redo**, and **Delete**.
④	**Preview**	You can click this icon to open a new web page to view the display effect of the big screen that you are editing.
⑤	Share	You can click this icon to share visualized data.

- **Chart**

  You can click the Chart icon to add charts. To add a chart, perform the following steps:

  i. Click the **Chart** icon.

  ii. In the **Add Chart** dialog box, select a chart and click **Next Step**.

  iii. Select a mode from the **Data Refresh mode** drop-down list and click **Save**.

DMS provides the following two methods to update data:

- **Manual refresh**: If you select this mode, you must click **Synchronize Data** in the upper-right corner of the chart on the big screen each time you want to update data.
- **Scheduled refresh**: The system automatically updates data in the chart at the specified interval in the unit of seconds.

iv. Click **Save**. The chart is added to the canvas.

You can drag a chart to adjust its position. You can also drag the chart in the lower-right corner to resize the chart.

- **Auxiliary graphics**

DMS provides four types of auxiliary graphics. You can click the **Auxiliary graphics** icon and select an auxiliary graphic from the list.

Auxiliary graphic	Description
Rectangle	This auxiliary graphic is generally used to decorate the background of charts. You can set the background color, background picture, and borders for a rectangle.
Label	This auxiliary graphic is generally used to display texts. You can set the text font, margin, background color, and borders for a label.
Video	This auxiliary graphic is used to play online videos.
Time	This auxiliary graphic is used as a ticking clock.

- **Operation icons**

Operation icon	Description	Shortcut key in Mac OS	Shortcut key in Windows
Copy	You can click this icon to copy a chart or an auxiliary graphic.	Cmd+C	Ctrl+C
Paste	You can click this icon to paste a chart or an auxiliary graphic.	Cmd+V	Ctrl+V
Undo	When you edit the canvas, you can click the Undo icon to roll back an operation or click the Redo icon to restore an operation.	N/A	N/A
Redo			
Delete	You can click this icon to delete a chart or an auxiliary graphic.	Delete	Backspace or Delete

> **Note**
> - You can click the **Copy**, **Paste**, or **Delete** icon to copy, paste, or delete multiple layers at a time.
> - To select multiple layers, press and hold the Cmd or Alt key on the keyboard and click the layers.

- **Preview**

After you create a big screen, you can click the **Preview** icon in the toolbar to preview the big screen on a new web page.

Canvas

Each chart or auxiliary graphic that is added to the canvas can be viewed as a **layer**. On the canvas, you can perform the following operations:

- You can drag a layer to adjust its position. You can also select a layer and use the arrow keys on the keyboard to fine-tune the position of the layer.

  > **Note** When you are dragging a layer, a prompt appears to the right of the layer to show the current position of the layer in the form of coordinates. Guides are also displayed on the canvas to help you align the layer.

- When you edit a part of the canvas, you can adjust the slider of the zoom tool to zoom in or zoom out on the canvas.
- You can press and hold the Cmd or Alt key on the keyboard, click multiple layers to select them, and then copy, delete, or align the layers at a time.
- To deselect one or more selected layers, click the blank area on the canvas.
- To edit a chart, click the **Edit** icon in the upper-right corner of the chart.

Layer panel

In the Layer panel, you can perform the following operations:

- Select a layer: You can select a layer on the canvas by selecting the corresponding item in the layer list.
- Select multiple layers: You can select multiple layers on the canvas by pressing and holding the Cmd or Alt key on the keyboard and selecting the corresponding items in the layer list.
- Adjust the position of a layer on the Z axis: After you select a layer, you can click the icons above the layer list to adjust the position of the layer on the Z axis. The following icons are provided: **Move up**, **Move down**, **Top**, and **Bottom**.

Parameter settings panel

The content that is displayed in the parameter settings panel varies with the layer that you select.

- By default, the **Background settings** panel is displayed as the parameter settings panel. The following table describes the parameters in the Background settings panel.

Parameter	Description
Screen size	You can adjust the size of the canvas based on the display terminal of the big screen.
Background color	You can specify a background color for the big screen.
Zoom	◦ **Proportional scaling width**: The width of the canvas is the same as the width of the display terminal. The height of the canvas is proportionally scaled. ◦ **Proportional scaling height**: The height of the canvas is the same as the height of the display terminal. The width of the canvas is proportionally scaled. ◦ **Full screen**: Both the height and width of the canvas are the same as the height and width of the display terminal. In this mode, the canvas may be deformed on the display terminal. ◦ **Actual Size**
Picture	You can upload a background picture for the big screen.

- If a single layer is selected, the **Chart** panel is displayed as the parameter settings panel. You can set the following parameters in the Chart panel:
  - **Chart size**
  - **Chart location**
  - **Background**
  - **Border**
  - **Data Refresh mode**

- If multiple layers are selected, the **Layer alignment** panel is displayed as the parameter settings panel. You can click the following icons to operate the layers: **Top alignment**, **Left alignment**, **Center horizontally**, **Vertically centered**, **Right alignment**, and **Bottom alignment**.

Create a big screen

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Data Display**.
4. In the **Big Screen** section of the **Resource management** page, click **New big screen**.
5. In the **New Big Screen** dialog box, set the **Name** and **Description** parameters.
6. Click **Save**. You can view the new big screen in the **Big Screen** section.

Edit a big screen

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Data Display**.
4. In the **Big Screen** section of the **Resource management** page, find the big screen that you want to edit.
5. On the page that appears, edit the big screen based on your requirements. For more information about how to edit a big screen, see Big screen editor.

Delete a big screen

1. Log on to the DMS console.
2. Go to the Data Visualization tab.
3. In the left-side navigation pane, click **Data Display**.
4. In the **Big Screen** section of the **Resource management** page, find the big screen that you want to delete.
5. Click the icon.
6. In the message that appears, click **OK**.

16.1.8.5. Use the category feature

As the business develops and the number of tables increases, provides the category feature to help you classify tables. This way, administrators, developers, and O&M engineers can manage or use the tables more conveniently.

Prerequisites

- A relational database or data warehouse is used. For more information, see Supported databases.
- You are a DMS administrator or DBA. For information about user roles, see Features that are supported by each role.

Manage categories

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Factory > Category**.
3. On the tab that appears, click **Create Category**.
4. In the **Category Name** field of the dialog box that appears, enter the name of a category and click **OK**. The created category is displayed in the left-side category tree.

> **Note** By default, DMS creates the **Uncategorized** category. All the tables that are not added to categories belong to this category.

5. On the right side of a category, move the pointer over the ··· icon and perform an operation to manage the category. You can perform the following operations:
   - **Modify the name of a category**

     To modify the name of a category, move the pointer over the ··· icon and select **Change**.
   - **Create a subcategory**

     To create a subcategory, move the pointer over the ··· icon and select **Create Subcategory**.

     You can create up to four category levels. If a table is added to a category, you cannot create a subcategory for this category.
   - **Delete a category**

     To delete a category, move the pointer over the ··· icon and select **Delete**.

     > **Note** If a subcategory is created in this category or tables are added to this category, this category cannot be deleted.

Add a table to a category

Each table can be added to only one category. If you add a table to another category, the table is removed from the existing category.

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Data Factory > Category**.
3. Click the category to which you want to add a table.

   > **Note** To view subcategories, you can click the › icon on the left side of the category.

4. On the page of the category, click **Quick Add** in the upper-right corner of the page.

   > **Note** You can also use the following method to add a table to a category.
   > - On the page of the **Uncategorized** category, find the table that you want to manage and click **Associate Category** in the Actions column.

5. In the dialog box that appears, search for and select the table that you want to add, and then click **OK**.

   > **Note** You can add multiple tables to a category at a time.

Remove a table from a category

To remove a table from a category, find the table on the page of the category and click **Remove from Category** in the Actions column.

16.1.9. Schemas

16.1.9.1. Schema design

Data Management (DMS) provides the schema design feature. This feature allows you to change schemas with ease. This topic describes how to change schemas.

Prerequisites

The destination database is a MySQL, a PolarDB-X, or an ApsaraDB for OceanBase database.

Context

When you create projects, process new business requirements, or optimize business operations, you may need to change schemas. These schema operations include creating and editing tables. For example, you may need to add or delete fields or indexes, adjust field attributes, or adjust the index composition. In these scenarios, you can use the schema design feature of DMS.

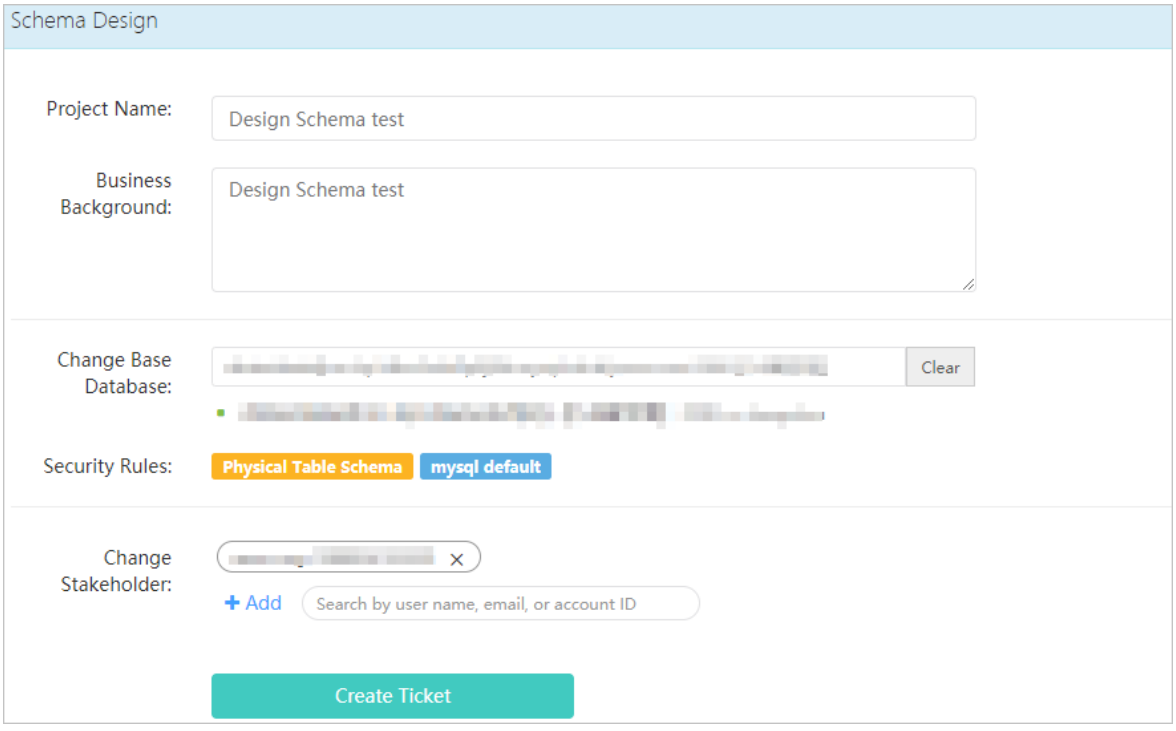
- This feature allows multiple users to simultaneously change a schema in the DMS console at the same time.
- This feature allows you to send verified scripts to other environments. This ensures consistency between schemas in different environments.

Precautions

When you submit a schema design ticket to delete a table, make sure that the table is created by using a schema design ticket.

Procedure

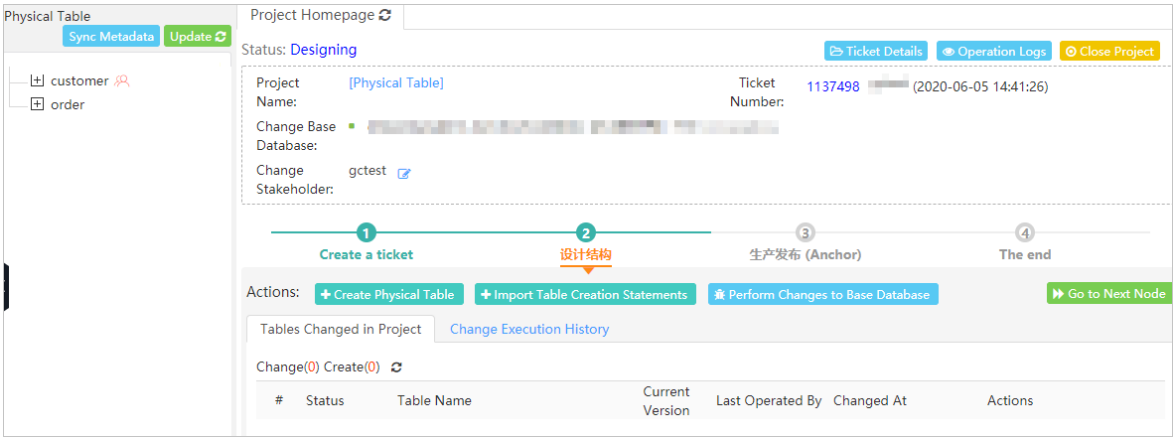
1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Schemas > Schema Design**.
3. In the upper-right corner of the Schema Design tab, click **Schema Design**.
4. On the Schema Design tab, specify the required parameters for a schema design ticket.



Parameter	Description
Project Name	The name of the project. Specify a name that can help you identify the project.

Parameter	Description
Project description	The background information about the project, such as the purpose or objective of the project. The description is used to reduce communication costs.
Change Base Database	The database whose schema you want to change. You can search for databases by keyword. Prefix match is applied. Only databases on which you have permissions in test or development environments are displayed. **Note** You must have at least the query, export, or change permissions on the database that you select.
Security Rules	No configurations are required. The default setting is specified.
Change Stakeholder	The stakeholders of the changes. The specified stakeholders can view the ticket details and are included in the approval process. Unauthorized users, except for administrators and database administrators (DBAs), cannot view the ticket details.

5. Click **Create Ticket**.
6. Change a schema based on your business requirements.



- Create a table:
  a. Click **Create Physical Table**.

     Note If the destination database is a logical database, click **Create Logical Table**.

  b. On the Create Physical Table tab, set the required parameters. The parameters include the table name, character set, fields, and indexes.
  c. Click **Save**.

     Note After you click Save, DMS verifies the specified information based on design specifications. If the information does not comply with the design specifications, a message appears.

  d. After the information passes the precheck, click **Confirm Changes and Submit to Save**.
- Change the schema of a table:
  a. In the left-side table list, right-click the name of the required table.
  b. On the menu that appears, select **Design Table**.

c. Change the schema as required and click **Save**.

> **Note** After you click Save, DMS verifies the specified information based on design specifications. If the information does not comply with the design specifications, a message appears.

d. After the specified information passes the verification, click **Confirm Changes and Submit to Save**.

7. After the schema is changed, click **Perform Changes to Base Database**.
8. In the Perform Changes to Base Database dialog box, set the Execution Strategy parameter to **Execute Now** or **Schedule**.
9. Click **Submit for Execution** and wait until the ticket is approved.
10. After the ticket is approved, click **Go to Next Node**.

> **Note**
> - After the ticket is approved, DMS applies the changes at the specified point in time. If you do not specify the execution time, the changes are automatically applied after the ticket is approved at the last approval node. You can view the execution status and operation logs. After all changes are applied, you can repeat the preceding procedure to change the schema again. If no additional changes are required for the schema, click Go to Next Node.
> - After the ticket is submitted to the next node, whether you can go back to the previous node is subject to the predefined design specifications.

11. In the Go to Next Node message, click **Go to Next Node**.
12. On the Project Homepage tab, click **Perform Changes to Target Database**.
13. In the Perform Changes to Target Database dialog box, set the Target Database and Execution Strategy parameters and click **Submit for Execution**.

> **Note** The required database must reside in a production environment.

14. Wait until the ticket is approved and the changes are applied.
15. Click **Go to Next Node**.
    The schema design process ends and the ticket is closed.

16.1.9.2. Schema synchronization

Data Management (DMS) provides the schema synchronization feature. You can use this feature to compare the schemas of two databases, generate a script to synchronize schemas, and then run the script on the destination database. This topic describes the schema synchronization feature and how to synchronize schemas.

Prerequisites

The source databases and destination databases are ApsaraDB for OceanBase or MySQL databases.

Precautions

- You cannot synchronize schemas to a destination database that resides in a production environment.
- The empty database initialization feature allows you to synchronize some or all tables from a physical or logical database.

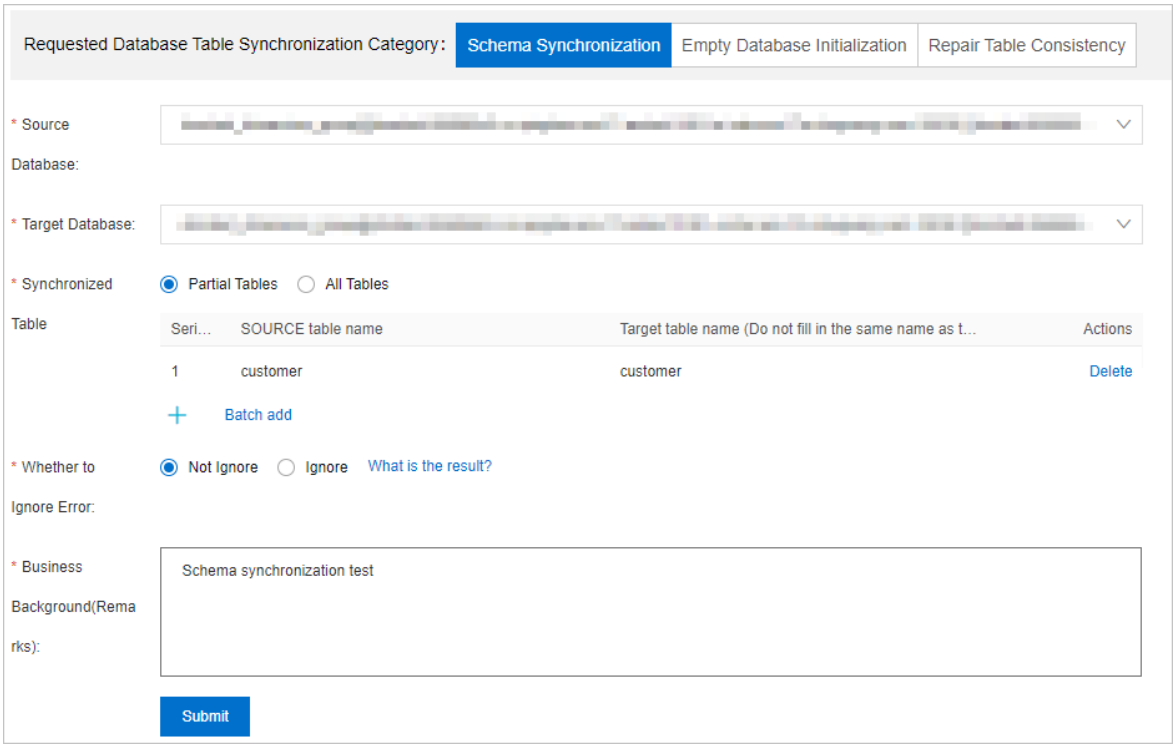
Scenarios

You can use the schema synchronization feature to synchronize schemas and ensure schema consistency in the following scenarios:

- Synchronize data between a database in a production environment and a database in a test environment.
- Synchronize data between different databases that are deployed in a test environment.
- Synchronize data between different databases that are deployed in a production environment.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Schemas > Schema Synchronization**.
3. Specify the required parameters for a schema synchronization ticket.



Parameter	Description
Source Database	The name of the source database from which you want to synchronize schemas. You must have the read permissions on the source database.
Target Database	The name of the destination database to which you want to synchronize schemas. You must have the change permissions on the destination database.
Synchronized Table	The tables that you want to synchronize. Valid values: ◦ **Partial Tables**: Synchronize one or more tables in the source database. You can click **Batch Add** to add multiple tables. **Note** If you do not set this parameter, the names of the destination tables are the same as the names of the source tables. ◦ **All Tables**: Synchronize all tables in the source database.

Parameter	Description
Whether to Ignore Error	○ **Not Ignore**: If an error occurs when SQL scripts are executed in serial mode, the system immediately stops executing the current and remaining SQL scripts. ○ **Ignore**: If an error occurs when SQL scripts are executed, DMS stops executing the current SQL script and continues to execute the next statement until all remaining SQL scripts are executed.

4. Click **Submit**. DMS starts to analyze the schemas.
5. Check the comparison results.

> **Note** If the schemas are changed when the system analyzes the schemas, click Re-analyze in the Schema Analysis step.

6. Verify the script that is used to synchronize schemas and click **Submit and Synchronize to Target Database**.

> **Note** If the schemas of the source database and destination database are the same, you do not need to submit the script, and the schema synchronization ticket is closed.

16.1.9.3. Synchronize shadow tables

Data Management (DMS) provides the shadow table synchronization feature to automatically create a shadow table based on the schema of a source table. DMS generates the name of the shadow table by attaching a prefix or suffix to the name of the source table. You can use this feature for end-to-end stress testing.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **Schemas > Shadow Table Synchronize**.
3. On the **Table/Database Synchronization Application** page, set the parameters that are described in the following table.

Parameter	Description
Source Database	The database whose data is to be synchronized.
Prefix / Suffix	The prefix or suffix that is used to create a shadow table name. The name can be in the **Prefix** + Source table name format or Source table name + **Suffix** format. You can use a custom prefix or suffix as needed. By default, the Prefix + Source table name format is used. Default shadow table name: `__test_Source table name`.
Synchronized Table	The tables whose schemas you want to synchronize. Valid values: ○ **Partial Tables** ○ **All Tables**

Parameter	Description
Synchronization Policy	The policy that is used for shadow table synchronization. Valid values: ○ **Synchronize Now**: DMS immediately synchronizes the tables after you submit the ticket. In this case, the tables are synchronized only once. ○ **Scheduled Synchronization**: DMS synchronizes the tables at the specified time on a regular basis. You can use a crontab expression to schedule synchronization based on your requirements. The minimum interval for synchronization is 1 hour. By default, the shadow tables start to be synchronized at 02:00 every day. For more information, see the Crontab expressions section of this topic.
Whether to Ignore Error	Specifies whether to skip errors that occur when SQL statements are being executed. Valid values: ○ **Not Ignore**: If an error occurs when SQL statements are being executed, DMS stops executing the current and subsequent SQL statements. ○ **Ignore**: If an error occurs when SQL statements are being executed, DMS skips the current SQL statement and continues to execute subsequent SQL statements until all remaining statements are executed.
Business Background(Remarks)	The business background of the project, such as the purposes and objectives of the project.

4.
5.
6.

Crontab expressions

If you need to schedule the synchronization task to be run in a more precise manner, you can use a crontab expression. The interval for running the task can be specified by using a combination of minutes, hours, days, weeks, or months.

A crontab expression consists of five fields of the NUMERIC type. Valid values of each field:

- **Minutes**: `0 to 59`.
- **Hours**: `0 to 23`. A value of 0 indicates the midnight.
- **Days**: `1 to 31`. A value of this field indicates a specific day of a month.
- **Months**: `1 to 12`. A value of 1 indicates January, and a value of 2 indicates February. Similarly, the specific month that is indicated by a specific value can be obtained.
- **Weeks**: `1 to 7`. A value of 1 indicates Sunday, and a value of 2 indicates Monday. In other words, the seven week days from Sunday to Saturday are indicated by values 1 to 7.

Usage notes

- Specify the time for running a stress testing task by the day or week. You cannot specify the day and week at the same time. After you specify one of the preceding two values, you must set the other value to `?`. A value of ? indicates an unspecified value. For example, if you schedule the task to be run on the first and second days of each month, the **Weeks** field must be set to `?`.
- Limit the characters in a crontab expression to English special characters. The special characters can be wildcards such as asterisks (*) and question marks (?).
- Separate multiple values with commas (,).
- Use a hyphen (-) to indicate a value range. For example, if you set the **Days** field to `1-5`, the task is scheduled to be run on the first to fifth days of a month.
- Use a forward slash (/) to indicate an interval for running the task. For example, if you set the Days field to `*/2`

, the task is scheduled to be run every two days.

Crontab expression examples

- To schedule the task to be run at 23:00 every Saturday and Sunday, use the following crontab expression: 0 23 ? * 7,1.
- To schedule the task to be run at 09:30 on the fifth, fifteenth, and twenty-fifth days of each month, use the following crontab expression: 30 9 5,15,25 * ?.
- To schedule the task to be run at 00:00 every two days, use the following crontab expression: 0 0 */2 * ?.

16.1.9.4. Initialize empty databases

DMS provides the empty database initialization feature. This feature allows you to compare the schemas of two databases, generate a script that is used to synchronize data from the source database to the destination database, and run the script on the destination database. To use this feature, the destination databases must be empty. This topic describes how to initialize empty databases.

Prerequisites

- The source databases and destination databases are MySQL or ApsaraDB for OceanBase databases.
- The destination databases are empty databases that do not contain tables.

Features

The empty database initialization feature allows you to synchronize some or all tables from a physical or logical database.

Scenario

Synchronize data between databases that are deployed in different regions or units. For example, this feature is applicable to the following scenarios:

- Synchronize data between a database that is deployed in a production environment and a database that is deployed in a test environment.
- Synchronize data between different databases that are deployed in the test environment.
- Synchronize data between different databases that are deployed in the production environment.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Schemas > Empty Database Initialization**.
3. On the Table/Database Synchronization Application tab, set the required parameters to create an empty database initialization ticket.

Parameter	Description
Source Database	The name of the source database from which data is synchronized. You must have the read permissions on the source database.
Target Database	The name of the destination database to which data is synchronized. You must have the write permissions on the destination database. **Note** The type of the destination database must be the same as the type of the source database.

Parameter	Description
Initialized Table	The tables that you want to synchronize. Valid values: ○ **Partial Tables**: Synchronize one or more tables in the source database. To add a table, click the **+** icon and specify a name for the source table. Note ■ You can also click **Batch Add**. In the Batch Add dialog box, select the required tables and click **Batch Add**. ■ If you do not set this parameter, the names of the destination tables are the same as the names of the source tables. ○ **All Tables**: Synchronize all tables in the source database.
Whether to Ignore Error	○ **Not Ignore**: If an error occurs when an SQL script is being executed in serial mode, the system immediately stops executing the current and remaining SQL scripts. ○ **Ignore**: If an error occurs when an SQL script is being executed, DMS stops executing the current SQL script and continue to execute the next statement until all remaining SQL scripts are executed.

4.
5.
6.

16.1.9.5. Repair table consistency

DMS provides the table consistency repairing feature. This feature is used to compare schemas between tables in databases that are deployed in different environments, provides an efficient way to identify schema differences, and execute SQL statements that are specific to the required environment. This ensures schema consistency between different environments.

Prerequisites

The source databases and destination databases are MySQL or ApsaraDB for OceanBase databases.

Scenarios

- Ensure the schema consistency between physical tables that are deployed in the test environment and the production environment.
- Ensure the schema consistency between physical tables and logical tables in a physical database or a logical database.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > Schemas > Table Consistency Repairing**.
3. On the Table/Database Synchronization Application tab, set the required parameters to create a Repair Table Consistency ticket.

Parameter	Description
Base Database(Physical Database)	The source database based on which schema consistency is to be repaired. You must have the query permissions on the source database.

Parameter	Description
Target Database	The destination database whose data is to be modified. You must have the change permissions on the destination database.
Repaired Table	The tables between which schema consistency is to be repaired. To add tables, click the **+** icon and specify the required table names. **Note** If you do not specify the destination table name, the system names the destination table after the name of the specified source table.
Whether to Ignore Error	◦ **Not Ignore**: If an error occurs when SQL statements are being executed in serial mode, the system immediately stops executing the current and remaining SQL statements. ◦ **Ignore**: If an error occurs when DMS is executing an SQL statement, DMS stops executing the current SQL statement and continues to execute the remaining SQL statement.
Business Background	The business background of the ticket. This parameter reduces communication costs.

4.
5.
6.

16.1.10. SQL review

DMS provides the SQL review feature. You can use this feature to remove SQL statements that do not use indexes or do not conform to database development standards. This way, the risk of SQL injection attacks is reduced.

Prerequisites

SQL reviews are performed before the related code is published to an online environment. Therefore, you must set the environment type of the required database instance to **Test** in the DMS console.

Context

When you develop a project, you must execute SQL statements on databases to add, delete, modify, and query data so that you can implement business logic and visualize data. Before the project is published, you must review all SQL statements that you want to execute. Make sure that all SQL statements conform to database development standards to ensure business continuity.

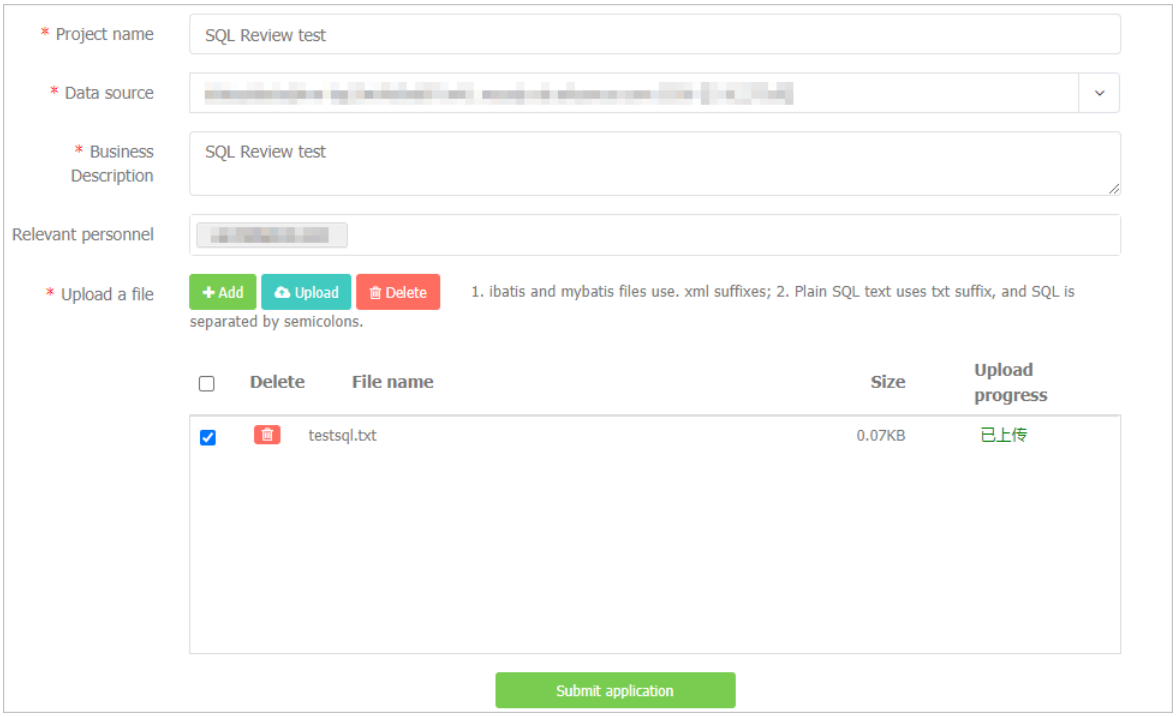
If DBAs manually review all SQL statements in sequence, this process requires excessive human resources and reduces the efficiency of research and development (R&D). However, the SQL review feature provides a quick method to review SQL statements and optimization suggestions.

Limits

- Only XML or TXT files can be uploaded.
- Tables that are specified in SQL statements must exist in the specified database. Otherwise, the system cannot review these SQL statements.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **Optimization > SQL Review**.
3. Set the required parameters of the SQL review ticket.



Parameter	Description
Project Name	Enter a project name based on your business requirements so that the ticket can be distinguished from other projects in subsequent processes.
Database	Select a test database for the project as the destination database. You must have the change permission on the database.
Business Background	Enter the information about the business scope of the project to help relevant users obtain the details about the project.
Change Stakeholder	Enter an at sign (@) and select the required user. **Note** You can repeat this operation to select multiple users.
Upload a file	Click **Add**, select the required files, and then click **Upload**. **Note** - The iBATIS and MyBatis files must be in the XML format. - SQL statements are saved as TXT files. Separate multiple SQL statements with semicolons (;). - To remove an added file, select the file and click **Delete** next to the file.

4. Click **Submit**.
5. View the result of the SQL review.

> **Note**
> - If the SQL statements conform to database development standards and use indexes, the result indicates that the SQL review succeeds and no recommended indexes are provided.
> - If the SQL statements conform to database development standards and do not use indexes, the result indicates that the SQL review succeeds and recommended indexes are provided.
> - If the SQL statements do not conform to database development standards, the result indicates that the SQL review fails.

6. Find a failed SQL review and click **View reason** to check the reason. You can also click **Details**, **Adjust SQL**, or **More** in the **Operation** column of the failed SQL review to perform the required operations.

   > **Note** After you modify the SQL statements and click **Confirm** to allow the system to review the SQL statements again. For dynamic SQL statements in XML files, you must enumerate each combination of SQL statements.

7. When all SQL statements pass the SQL review, click **Inspection results**.
8. In the dialog box that appears, click **Submit for approval** and wait for approval.

   > **Note** The approval process of the ticket varies based on the security rules that are configured for the current database instance.

16.1.11. System management

16.1.11.1. Manage instances

Data Management (DMS) allows you to manage database instances. For example, you can export the information about instance configurations.

Prerequisites

You are a database administrator (DBA) or a DMS administrator.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > Instance**.
3. On the **Instance List** tab, select one or more database instances that you want to manage. Then, you can perform the following operations based on your business requirements:

   > **Note** You can click **Expand filter** to show more filter conditions.

   - Add an instance

     Click **New** and register a database instance with DMS. For more information, see Register database instances with DMS.

   - Modify the information for multiple instances at a time

     Click **Batch edit**. In the dialog box that appears, modify the instance information and click **OK**.

     > **Note** The database instances that you select must be of the same database type, such as MySQL.

   - Synchronize the data dictionary

Click **Sync dictionary**. In the message that appears, click **OK**.

> **Note**
> - If you change schemas for a database instance by using DMS, DMS automatically synchronizes the data dictionary of the instance.
> - If you change schemas for a database instance by using a service other than DMS, you must manually synchronize the data dictionary of the instance.

- Disable or enable one or more instances

  Click **Forbidden instance** or **Enable instance**. In the message that appears, click **OK**.

  > **Note**
  > - After you disable a database instance, the instance is removed from the left-side instance list. DMS users can no longer find databases or tables in this instance in the DMS console.
  > - After you enable a database instance, the instance appears in the left-side instance list. Databases in this instance become available. Relevant permissions that have been granted to DMS users on this instance also become valid.

- Remove one or more instances

  Click **Delete Instance**. In the message that appears, click **OK**. After you remove a database instance, the instance is removed from the left-side instance list. DMS users can no longer use databases in this instance in the DMS console. Relevant permissions that have been granted to DMS users on this instance also become invalid and are revoked.

  > **Note** On the **Instance List** tab, you can find database instances in the Delete state and enable these instances to recover them.

- Export configuration information

  Click **Export config**. The browser automatically downloads a CSV file named *instances*. You can use Excel or a text editor to view this file.

- Configure access control

  Click **Access control**. In the dialog box that appears, turn on or off the switch for access control and click **OK**. The IP addresses of DMS servers are automatically added to the whitelists of the specified database instances.

  > **Note** The destination database instances must be ApsaraDB instances.

- More operations

  You can find a database instance and click Details in the Actions column to view the details about databases and tables in this instance. You can also move the pointer over More and perform other operations. For example, you can log on to the instance or modify the information about the instance.

16.1.11.2. Database management

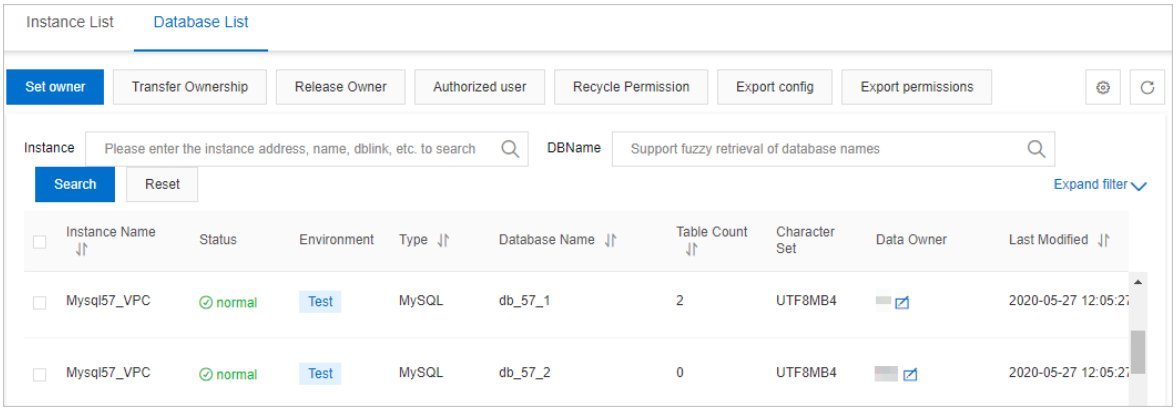
On the Database List tab, you can manage databases. For example, you can specify the database owner, transfer the ownership, revoke the owner permission, grant and revoke user permissions, and export the information about database configurations or permissions.

Procedure

1. Log on to the DMS console.

2. In the top navigation bar, move the pointer over the **More** icon and choose **System > Instance**.
3. In the top navigation bar, choose **System > Instance**.
4. Click the **Database List** tab.
5. Set filter conditions and select one or more databases that you want to manage. Then, you can perform the following operations based on your business requirements:

> **Note** You can click **Expand filter** to show more filter conditions.



- Specify an owner

  Specify an owner for the selected databases. You can specify multiple owners for multiple databases at a time.

- Transfer the ownership

  Transfer the ownership of the selected databases to a user. If you transfer the ownership of multiple databases at a time, you can select only a user that assumes the ownership of all of the databases as the original owner.

- Revoke owner permissions

  Revoke the owner permission from the owners of the selected databases.

- Grant permissions

  Grant the query, export, or change permission on the selected databases to one or more users. You can also specify an expiration time for the permission.

- Revoke permissions

  Revoke the query, export, or change permission on the selected databases from one or more users. If a user does not have the related permissions, the following message appears: `No corresponding permissions. You do not need to recycle or release permissions`.

- Export configurations

  Export the configurations of the selected databases to an Excel file. The configurations include the instance status, environment, DBA, and owner.

- Export permission information

  Export the permission information about the selected databases to an Excel file. The permission information includes the database information, users, permissions, and users who grant the permissions.

- Other operations

  You can click Tables in the Actions column of a database to view the details about tables in the database. You can also move the pointer over More and select the required operation that you want to perform. For example, you can query data in the database, manage permissions, view the details about the instance to which the database belongs, and find the instance on the Instance List tab.

16.1.11.3. Manage users

Data Management (DMS) allows you to manage users. For example, you can manage user permissions and roles.

Prerequisites

You are a DMS administrator.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > User**.
3. Select the user that you want to manage. Then, you can perform the following operations based on your business requirements:
   - Add a user

     Click **New** and set parameters to add a user. For more information, see Add a user.
   - Edit a user

     Click **Change** in the **Actions** column. In the dialog box that appears, modify the user configurations and click **Confirm Change**.
   - Disable or enable a user

     Move the pointer over **More** in the **Actions** column and select **Disable** or **Enable**.
   - Remove a user

     Move the pointer over **More** in the **Actions** column and select **Delete**. In the message that appears, click **OK**.
   - Grant permissions

     Move the pointer over **Authorize** in the **Actions** column and select the object on which you want to grant permissions to the user. For example, you can select **Authorize instance**, **Authorize database**, **Authorize table**, **Authorize Line**, or **Authorize sensitive column**. In the dialog box that appears, enter a keyword to filter objects, set the Permission and Expire Date parameters, and then click **OK**.
   - Release permissions

     Move the pointer over **More** in the **Actions** column and select **Permission Details**. You can set conditions to filter the permissions that are granted to the user, select the permission type that you want to manage, and then click **Release Permission** to release the permissions.

16.1.11.4. Enable metadata access control

Data Management (DMS) provides the metadata access control feature. You can use this feature to allow users to view the information about and access a database or database instance on which they have permissions. Before this feature is enabled, regular users can query all databases and database instances within the current tenant account. After this feature is enabled, you can allow specific users to view the information about and access the databases or database instances on which they have permissions. This further enhances the data security of your enterprise.

Background information

As a centralized data management service, DMS provides different roles that are assigned different permissions. This helps you manage data in your enterprise in a secure manner. After you enable metadata access control for a database instance or database, only users who have permissions on the instance or database can view the information about and access the instance or database. This way, users can view the information about and access only databases on which they have permissions. This further enhances data security.

> **Note** In DMS, permissions on a database include the query, export, and change permissions. If you have one of these permissions on a database, you can view the following information about the database:
>
> - Information about the database. You can search for the database in the search box in the upper part of the left-side navigation pane or in the top navigation bar of the DMS console. Alternatively, you can search for the database in the search box of the Select the databases, tables, or columns on which you want to apply for permissions field on the Permission Application Ticket page. Whether you can query the data in the database depends on whether you have the query permissions on the database.
> - Information about the instance to which the database belongs. Whether you can view the information about other databases in this instance depends on whether you have permissions on the other databases.

You can enable metadata control access for the following objects:

- A user: The user can view the information about and access only databases on which the user has permissions.
- A database: Only users who have permissions on the database can view the information about and access the database.
- A database instance: Only users who have permissions on the database instance can view the information about and access the database instance. If a user has permissions on a database in this database instance, the user can view the information about and access this database.

Enable metadata access control for a user

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > User**.

   > **Note** You are a DMS administrator.

3. Find the user for whom you want to enable metadata access control, move the pointer over **More** in the **Actions** column, and then select **Access control**.
4. In the User access control dialog box, turn on **Metadata access control** and click **OK**.

Enable metadata access control for a database instance

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > Instance**.

   > **Note** You are a database administrator (DBA) or a DMS administrator.

3. On the **Instance List** tab, find the instance for which you want to enable metadata access control. Then, select the instance and click **Access control** in the upper part of this tab.

   > **Note** You can also enable metadata access control for multiple instances at a time. Select multiple instances and click **Access control** in the upper part of this tab.

4. In the dialog box that appears, turn on **Metadata access control** and click **OK**.

Enable metadata access control for a database

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > Instance**.

   > **Note** You are a DBA or a DMS administrator.

3. On the **Database List** tab, find the database for which you want to enable metadata access control. Move

the pointer over **More** in the **Actions** column and select **Access control**.

> **Note** You can also enable metadata access control for multiple databases at a time. Select multiple databases and click **Access control** in the upper part of this tab.

4. In the dialog box that appears, turn on **Metadata access control** and click **OK**.

16.1.11.5. Manage tasks

The task management feature allows you to manage various tasks that are created by using tickets. You can also use this feature to directly create or manage tasks.

Prerequisites

You are a database administrator (DBA) or an administrator.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > System > Task**.
3. On the **Task** tab, view and manage tasks that are created by using tickets.
4. Find a task and perform one of the following operations based on your business requirements: For example, you can **pause**, **retry**, or**delete** a task.
   - Pause a task

     Click **Pause** to pause a task.
   - Retry a task

     If a task is in the **Failure** state, click **Retry** to run the task again.
   - Delete a task

     Click **Delete** to delete a task. After a task is deleted, the status of the task changes to **Delete** and the task cannot be run.
   - Create a task

     Click **Add SQL task**. In the Add SQL task dialog box, enter the task description, the database that you want to manage, and the SQL statements that you want to execute. Then, click **Submit Task**.

16.1.11.6. Configuration management

DMS allows you to manage system configurations. To implement flexible management, you must log on to the DMS console as an administrator to modify the required system configurations.

Prerequisites

An administrator account is required to perform the operation.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > System > Configuration**.
3. Find the required parameter and click **Change** in the **Actions** column of the parameter.

> **Note** You can click **Change History** to view the change history of the parameter.

4. In the Change Parameter Configuration dialog box, enter the required value.

5. Click **Confirm Change**.

Types of data changes

key	value	Description
config_correct	Modify Config	Modifies configurations.
project_init_data	Init Project Data	Initializes the data for a project.
program_bug	Program Bug	Fixes a bug.
require_deal_without_backend_function	Requirements Without Backend Function	Manages the data of an application that does not support backend management.
history_data_clear	History Data Clean	Clears historical data.
test	Test	Runs a test.
mis_operation	Mis Operation	Restores data after a misoperation.
others	Others	Changes data for other reasons.

16.1.11.7. Database grouping

This topic describes how to create a database group in Data Management (DMS). You can use this feature to apply a data change or a schema change to all of the databases in a database group with ease.

Prerequisites

The databases that you want to add to a database group meet the following conditions:

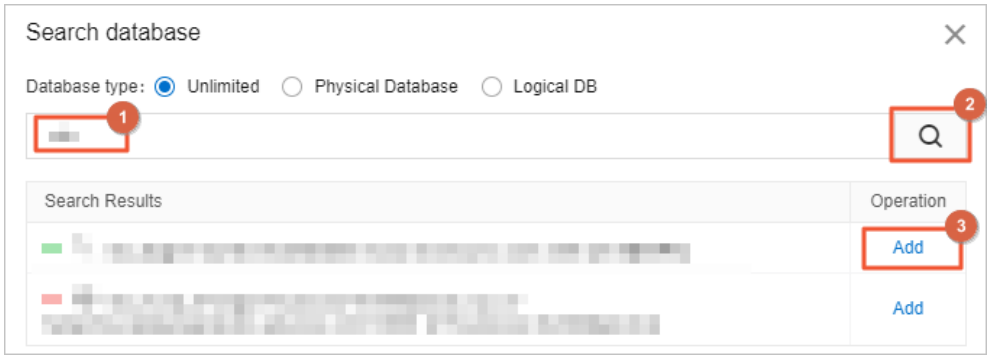
- All of the instances to which the databases belong are managed in Security Collaboration mode.
- All of the databases are physical databases or logical databases.
- All of the databases are deployed in the same environment, such as the development environment.
- The engines of the databases are of the same type. For example, all of the databases are MySQL databases.

Create a database group

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > Database grouping**.
3. Click **New Group**.
4. In the **NewGrouping** dialog box, perform the following steps:
   i. Enter a group name in the **Group name** field.
   ii. Set the **Grouping type** parameter to **General grouping**.

> **Note** You cannot set this parameter to **Remote live**. This feature will be available soon.

iii. Click **Add database**. In the **Search database** dialog box, enter a prefix in the search box to search for databases. Select one or more databases to be grouped from the matched results and click **Add** in the Operation column.



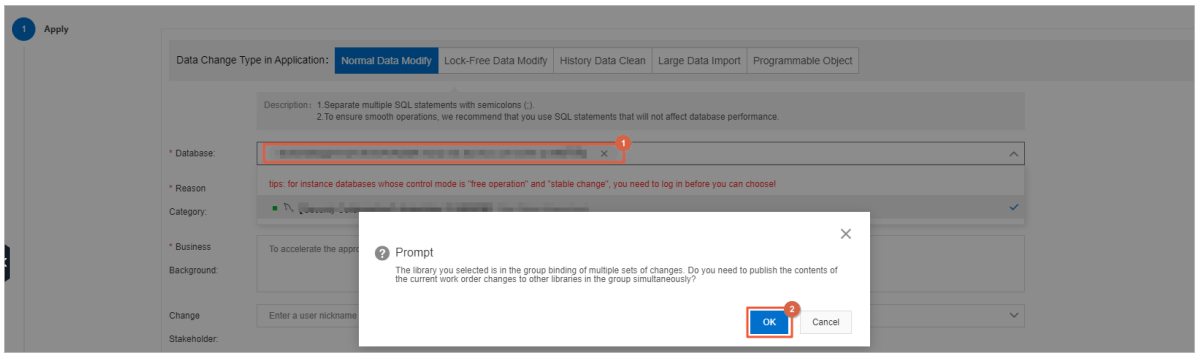
iv. After you add all the databases to be grouped, click the icon in the upper-right corner to close the Search database dialog box.

5. After you complete the configurations, click **Save**.

Scenarios

- Data change

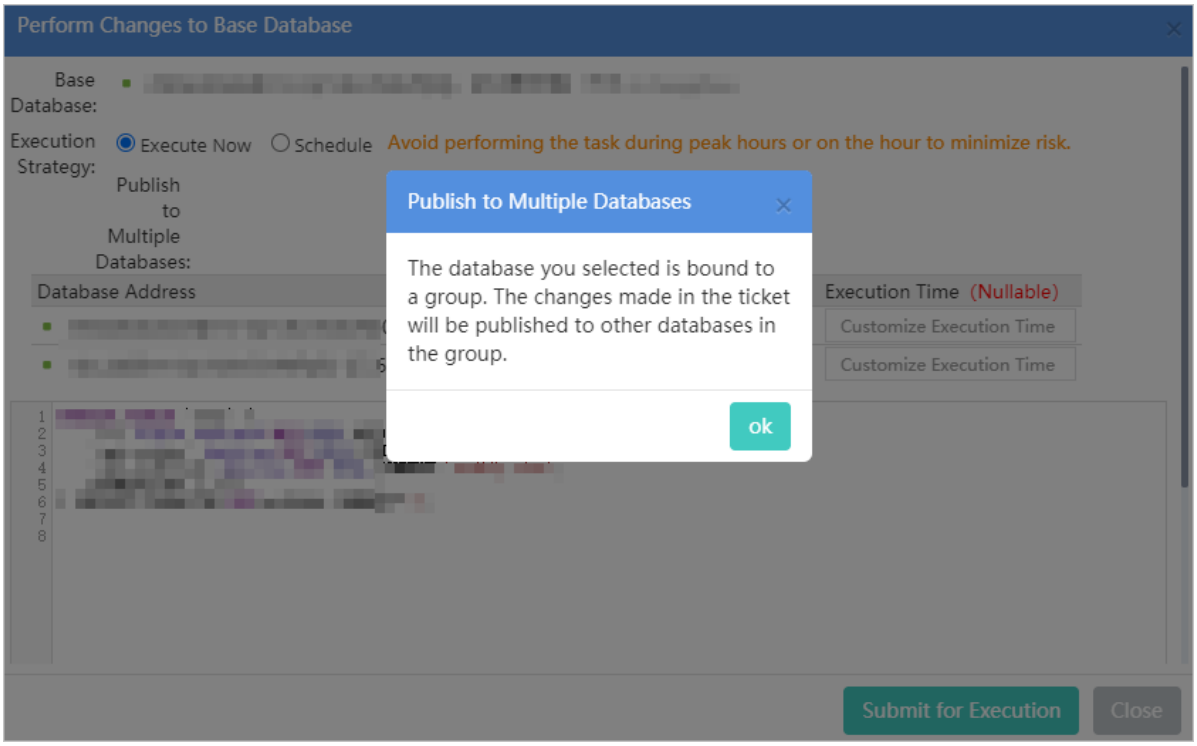
  For example, you want to create a ticket to perform a data change on a database, and the database belongs to a database group. After you select the database, DMS displays a message to remind you that the selected database belongs to a database group. If you click **OK**, DMS adds all the other databases in the group as the databases on which the data change will be performed. This saves your effort in selecting databases one by one. If you click **Cancel**, the other databases in the group will not be selected. The following figure shows the message.



  This feature applies to the data change and data import tickets that are supported by DMS. For more information about how to create a ticket, see Change data and Import data.

- Schema design

For example, you want to create a schema design ticket and select a database that belongs to a database group as a base database. After you click **Perform Changes to Base Database**, DMS displays a message. This message is used to remind you that the base database belongs to a database group and the current operation will apply to all the other databases in the group. The following figure shows the message.



For more information about how to use the schema design feature, see Design a schema.

16.1.11.8. Security management

16.1.11.8.1. Manage security rules

Security rules are implemented by using a collection of domain-specific languages (DSLs) to control user access to databases based on several factors. These factors include the type of databases, the syntax of database operations, and the number of affected rows. You can use security rules to standardize database operations, development processes, and approval processes as required. This topic describes how to manage security rules.

Prerequisites

You are a database administrator (DBA) or an administrator.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > System > Security Rules**.
3. Perform one of the following operations based on your business requirements:
   - Create a rule set

     Click **Create Rule Set**. In the Create Rule Set dialog box, set the Engine Type, Rule Set Name, and Remarks parameters, and click **Submit**.
   - Edit a rule set
     a. Find the required rule set and click **Edit** in the **Actions** column of the rule set.

b. In the left-side pane, click the required rule subset, for example, **SQLConsole**. In the right-side pane, select a checkpoint.
c. Find the required rule and click **Edit** next to the rule. For more information about the related syntax, see DSL syntax for security rules.

> **Note** You can disable or delete a rule.

- Create a similar rule set
  a. Find the required rule set and click **Create As** in the **Actions** column of the rule set.
  b. In the dialog box that appears, enter a name and a description for the new rule set.
  c. Click **Submit**. The system copies the configurations of the original rule set to the new rule set.
- Delete a rule set

  Find the required rule set and click **Delete** in the **Actions** column of the rule set. In the message that appears, click **OK**.

  > **Note**
  > - A deleted rule set cannot be recovered. Proceed with caution.
  > - You can delete only custom rule sets. You cannot delete built-in rule sets.
- Set a rule set as the default rule set

  Find the required rule set and click **Set as Default** next to the rule set. In the message that appears, click **OK**. The rule set is used as the default rule set for the related database engine.

16.1.11.8.2. DSL syntax for security rules

DMS provides a domain-specific language (DSL) to describe security rules. You can use the DSL syntax to define security rules. This allows you to define database development standards based on your business requirements.

Overview

The DSL syntax can include one or more conditions and related actions that are specified by an IF-ELSE statement.

> **Note** The if clause is required. Zero or more elseif clauses can be specified. Zero or one else clause can be specified.

Example 1: If Condition 1 is met, DMS performs Action 1.

```
if
    Condition 1
then
    Action 1
end
```

Example 2: If Condition 1 is met, DMS performs Action 1. If Condition 2 is met, DMS performs Action 2. If Condition 1 and Condition 2 are not met, DMS performs Action 3.

> **Note** If the `else Action 3` clause is removed and Condition 1 and Condition 2 are not met, DMS performs no action.

```
if
    Condition 1
then
    Action 1
elseif
    Condition 2
then
    Action 2
[else Action 3]
end
```

DSL syntax

- Conditional clauses

  DMS uses conditional clauses to evaluate whether to perform actions. The result of a conditional clause is true or false. A conditional clause consists of one or more connectors, operators, and factors. Connectors include AND and OR. Factors are predefined system variables. The following examples are valid conditional clauses:

```
1. true                    // This is the simplest conditional clause. The result is true.
2. 1 > 0
3. 1 > 0 and 2 > 1
4. 1 <= 0 or 1 == 1
```

- Connectors

  Connectors include AND and OR. The AND connector has higher priority than the OR connector. The two connectors have lower priority than operators. For example, a conditional clause is `1 <= 0 or 1 == 1`. DMS evaluates the result of the `1 <= 0` expression and the result of the `1 == 1` expression. Then, DMS evaluates the result of the OR expression based on the preceding results.

- Operators

  Operators are used to connect factors and constants to perform logical operations. The following table describes the operators that are supported by DMS.

Operator	Description	Examples
==	Evaluates whether a value is equal to another value.	1 == 1
!=	Evaluates whether a value is not equal to another value.	1 != 2
>	Evaluates whether a value is greater than another value.	1 > 2
>=	Evaluates whether a value is greater than or equal to another value.	1 >= 2
<	Evaluates whether a value is less than another value.	1 < 2
<=	Evaluates whether a value is less than or equal to another value.	1 <= 2
in	Evaluates whether a value belongs to an array of values.	'a' in ['a', 'b', 'c']
not in	Evaluates whether a value does not belong to an array of values.	'a' not in ['a', 'b', 'c']

Operator	Description	Examples
matchs	Evaluates whether a string matches a regular expression.	'idxaa' matchs 'idx\w+'
not matchs	Evaluates whether a string does not match a regular expression.	'idxaa' not matchs 'idx\w+'
isBlank	Evaluates whether a value is empty.	'' isBlank
isNotBlank	Evaluates whether a value is not empty.	'' isNotBlank

> **Note**
> - If you need to use a backslash (\) in a regular expression, you must add another backslash (\) as an escape character before the backslash that you want to use. For example, if you want to write the `idx_\w+` expression, you must enter `idx_\\w+`.
> - If a conditional clause includes nested expressions, we recommend that you enclose the required expressions in parentheses (). For example, a conditional clause is `1 <= 2 == true`. To specify the priority, you can change the clause to `(1 <= 2) == true`. DMS first evaluates the result of the `1 <= 2` expression in the parentheses.

- Factors

  A factor is a predefined variable in DMS. You can use factors to obtain the context to be validated by security rules. The context includes command categories and the number of affected rows. A factor name is prefixed by `@fac.`.Each tab of the Security Rules tab includes different factors for different checkpoints. The following table describes the factors that are supported by DMS.

Factor	Description
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.
@fac.sql_type	The type of the SQL statement, for example, UPDATE or INSERT. For more information, see the SQL subcategories that are described in the "SQLConsole for relational databases" topic.
@fac.detail_type	The type of the data change. Valid values: ○ COMMON: a normal data modify ticket ○ CHUNK_DML: a lock-free data modify ticket ○ PROCEDURE: a programmable object ticket ○ CRON_CLEAR_DATA: a history data clean ticket ○ BIG_FILE: a large data import ticket
@fac.is_logic	A Boolean value that indicates whether the affected database is a logical database.
@fac.extra_info	Other information about the ticket. This factor is not in use.
@fac.is_ignore_affect_rows	A Boolean value that indicates whether to skip the validation.

Factor	Description
@fac.insert_rows	The number of data rows to be inserted.
@fac.update_delete_rows	The number of data rows to be updated.
@fac.max_alter_table_size	The size of the largest tablespace in which the table to be modified is stored.
@fac.is_has_security_column	A Boolean value that indicates whether sensitive fields are specified in the SQL statement to be executed.
@fac.security_column_list	The sensitive fields that are specified in the SQL statement to be executed.
@fac.risk_level	The risk level that is identified.
@fac.risk_reason	The reason based on which the operation is identified as this risk level.

Note You can use factors in conditional clauses. For example, you can write `@fac.sql_type == 'DML'` to evaluate whether an SQL statement is a DML statement.

- Action clauses

  An action indicates an operation that is performed when the if clause evaluates to true. For example, DMS can disable the submission of a ticket, select an approval process, approve a ticket, or reject a ticket. An action indicates the usage of a security rule. An action name is prefixed by `@act.` .Each tab of the Security Rules tab includes different actions for different checkpoints. The following table describes the actions that are supported by DMS.

Action	Description
@act.allow_submit	Requires the submission of SQL statements to be executed in a ticket.
@act.allow_execute_direct	Allows the execution of SQL statements in the SQLConsole.
@act.forbid_execute	Disables the execution of SQL statements.
@act.mark_risk	Marks the risk level of a data change. Example: `@act.mark_risk 'medium-level risk: online environment'` .
@act.do_not_approve	Specifies the ID of an approval template.
@act.choose_approve_template	
@act.choose_approve_template_with_reason	

- Predefined functions

  DMS provides predefined functions that can be used in conditional clauses and action clauses. A function name is prefixed by `@fun.` .

Function	Description	Format
@fun.concat	Connects strings to form a single string. Output: a string. Input: multiple strings.	@fun.concat('d', 'm', 's') // The output is the string 'dms'. @fun.concat('[Development standards] The [', @fac.column_name, '] You must enter remarks.') // The output is a prompt that reminds the user who submits the ticket to enter a value in the field.
@fun.char_length	Calculates the length of a string. Output: an integer. Input: a string.	@fun.char_length('dms') // The output is 3. @fun.char_length(@fac.table_name) // The output is the length of the table name.
@fun.is_char_lower	Evaluates whether all the letters in a string are lowercase letters. Output: true or false. Input: a string.	@fun.is_char_lower('dms') // The output is true. @fun.is_char_lower(@fac.table_name) // If the output is true, it indicates that all the letters in the table name are lowercase.
@fun.is_char_upper	Evaluates whether all the letters in a string are uppercase letters. Output: true or false. Input: a string.	@fun.is_char_upper('dms') // The output is false. @fun.is_char_upper(@fac.table_name) // If all the letters in the table name are uppercase letters, the output is true.
@fun.array_size	Counts the number of values in an array. Output: an integer. Input: an array of values.	@fun.array_size([1, 2, 3]) // The output is 3. @fun.array_size(@fac.table_index_array) // The output is the number of indexes of the table.
@fun.add	Adds multiple numeric values. Output: a numeric value. Input: multiple numeric values.	@fun.add(1, 2, 3) // 6
@fun.sub	Deducts a numeric value from another numeric value. Output: a numeric value. Input: two numeric values.	@fun.sub(6, 1) // 5
@fun.between	Evaluates whether a value belongs to a specific closed range. The supported data types are NUMERIC, DATE, and TIME. Output: true or false. Input: three values. The first value is the value to be evaluated. The second value indicates the lower limit. The third value indicates the upper limit.	@fun.between(1, 1, 3) // The output is true because the value 1 belongs to [1, 3]. @fun.between(2, 1, 3) // The output is true because the value 2 belongs to [1, 3]. @fun.between(7, 1, 3) // The output is false because the value 7 does not belong to [1, 3]. @fun.between(@fac.export_rows, 2001, 100000) // If the number of exported rows belongs to [2001, 100000], the output is true. @fun.between(@fun.current_datetime(), '2019-10-31 00:00:00', '2019-11-04 00:00:00') // If the current date and time belong to [2019-10-31 00:00:00, 2019-11-04 00:00:00], the output is true. @fun.between(@fun.current_date(), '2019-10-31', '2019-11-04') // If the current date belongs to [2019-10-31, 2019-11-04], the output is true. @fun.between(@fun.current_time(), '13:30:00', '23:59:59') // If the current time belongs to [13:30:00, 23:59:59], the output is true.
@fun.current_datetime	Returns the current date and time, in the format of yyyy-MM-dd HH:mm:ss. Output: a string. Input: none.	@fun.current_datetime() // For example, the output is 2019-10-31 00:00:00.
@fun.current_date	Returns the current date, in the format of yyyy-MM-dd. Output: a string. Input: none.	@fun.current_date() // For example, the output is 2020-01-13.

Function	Description	Format
@fun.current_time	Returns the current time, in the format of HH:mm:ss. Output: a string. Input: none.	@fun.current_time() // For example, the output is 19:43:20.

DSL configuration examples

Limit the number of SQL statements in a ticket: If the number of SQL statements in a ticket exceeds 1,000, DMS rejects the ticket and returns the related message.

```
if
    @fac.sql_count > 1000
then
    @act.reject_execute 'The number of SQL statements in a ticket cannot exceed 1,000.'
else
    @act.allow_execute
end
```

Allows the submission of only data manipulation language (DML) statements: If the SQL statements in a ticket are DML statements such as the UPDATE, DELETE, and INSERT statements, DMS allows the execution of the statements.

```
if
    @fac.sql_type in [ 'UPDATE','DELETE','INSERT','INSERT_SELECT']
then
    @act.allow_submit
end
```

16.1.11.8.3. Configure security rules for a database instance

This topic describes how to configure security rules for a database instance.

Prerequisites

- You are a database administrator (DBA) or an administrator.
- The control mode of the database instance is **security collaboration**.

Procedure

1. Log on to the DMS console.
2. In the left-side navigation pane, right-click the required database instance.
3. In the shortcut menu that appears, choose **Control Mode > Security Collaboration** and select the required security rule set.

> **Note** You can also change the security rule set of the database instance on the Instance tab. For more information, see Instance management.

16.1.11.8.4. Customize approval processes

Data Management (DMS) allows you to configure instance-level security rules so that you can customize different approval processes for different database instances or database operations. However, instance-level security rules have some limits in the production environment. This topic describes how to customize an approval process.

Prerequisites

You are a database administrator (DBA) or an administrator.

Context

- Each database instance has only one DBA. However, multiple DBAs are included in an approval process to ensure business continuity regardless of whether one of the DBAs is unavailable.
- If multiple business units share the same database in a database instance, each business unit must approve the tickets for their respective business operations in an approval process.

To resolve the issues in the preceding scenarios, you can customize approval processes.

Precautions

- Do not assign only one approver to an approval node. We recommend that you assign at least two approvers to each approval node and at least two data owners to a database.
- You can assign a maximum of three data owners to a database. If multiple business units share the same database, you can specify these business units in an approval process by performing the following steps: Create an approval node and add the data owners of the business units as approvers. Then, add the new node instead of the system node Owner to an approval template.

Procedure

This topic describes how to customize an approval process and specify multiple DBAs in the approval process. You can perform similar steps to customize an approval process in other scenarios.

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > System > Approval Processes**.
3. Create an approval node.
   i. Click the **Approval Node** tab. Then, click **Create Approval Node**.
   ii. The following table describes the parameters that you can specify for the approval node.

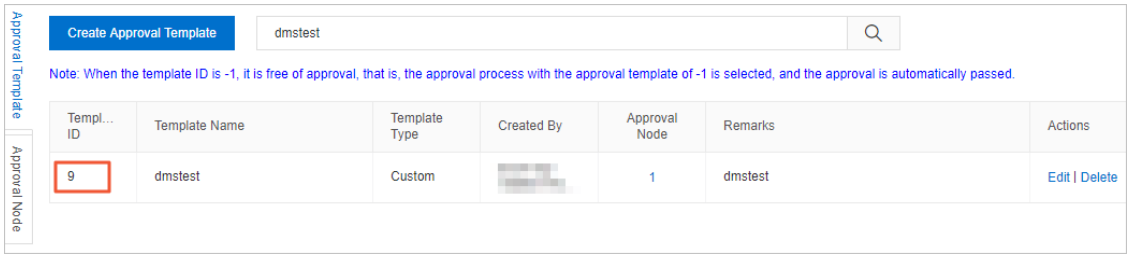
Parameter	Description
Node Name	The name of the approval node. The name must be globally unique.
Remarks	The description of the approval node. This parameter distinguishes the approval node from other approval nodes.
Approver	The Apsara Stack tenant accounts of the approvers for the approval node. You can search for approvers by keyword. Prefix match is used. **Note** In this example, three approvers are selected.

   iii. Click **Submit**.
4. Create an approval template.
   i. Click the **Approval Template** tab. Then, click **Create Approval Template**.

ii. The following table describes the parameters that you can specify for the approval template.

Parameter	Description
Template Name	The name of the approval template. The name must be globally unique.
Remarks	The description of the approval template. This parameter distinguishes the approval template from other approval templates.
Approval Node	Click Add Node and select the required approval nodes. In this example, the system node Owner and the approval node that is created in Step 3 are selected to allow multiple DBAs to participate in the approval process. **Note** The approval process is implemented based on the values of the Approval Order parameter in ascending order.

iii. Click **Submit**.
After the approval template is created, you can view the template ID. In this example, the template ID is 9.



5. Apply a new approval process.

This example shows how to edit a rule that is applied to medium-level risk approval processes under the **Risk Approval Rules** checkpoint. You can perform similar steps to apply a rule to other scenarios.

i. In the top navigation bar, choose **System > Security > Security Rules**.
ii. Find the required rule set that you want to edit and click **Edit** in the **Actions** column of the rule set.
iii. In the left-side navigation pane, click the **SQL Correct** tab.
iv. Select **Risk Approval Rules** as the checkpoint.
v. Find the rule that is related to the medium-level risk approval process and click **Edit**.
vi. In the **Rule DSL** field, change the template ID.

Note In this example, change `3` to `9`, as shown in the preceding figure. The ID 9 is the ID of the approval template that is created in Step 4.

vii. Click **Submit**.

Result

If the data change tickets that you submit match the rule, all specified DBAs receive ticket approval notifications and can participate in the approval process.

16.1.11.8.5. Operation audit

Data Management (DMS) provides the operation audit feature in addition to the basic features of operation log management. You can use this feature to troubleshoot database issues with ease and audit the operations that are performed on databases. You can also use this feature to view and manage the SQL statements that are used in the SQLConsole, tickets, logon information, and operation logs.

Features

The following table describes the two modules of the operation audit feature in DMS: Operation Logs and Operation audit.

Module	Description	Item
Operation Logs	Displays the logs of all the operations that are performed in DMS.	Includes the logs of management and configuration operations, SQL statements that are used in the SQLConsole, tickets, and logon information.
Operation audit	Displays all the operations that are performed on databases in DMS. **Note** This module provides a user interface (UI) for you to audit operations in a centralized manner. This also helps you troubleshoot database issues with ease.	Includes SQL statements that are used in the SQLConsole, tickets, and logon information. **Note** Only a DMS administrator, a database administrator (DBA), a ticket submitter, and stakeholders involved in the ticket approval process are allowed to view the ticket details.

Log data is permanently retained in DMS. You can access and view the log data of the instances that are managed in **Stable Change** or **Security Collaboration** mode at any time.

Note You can view the log data of the instances that are managed in **Flexible Management** mode only for the last seven days. To view all log data, change the control mode of the instances.

Links and supported roles

The following table describes the roles that you can assume to use the operation audit feature. It also shows you how to go to the Operation audit tab in the DMS console.

Auditing dimension	Limit	Link to operation audit	Supported role

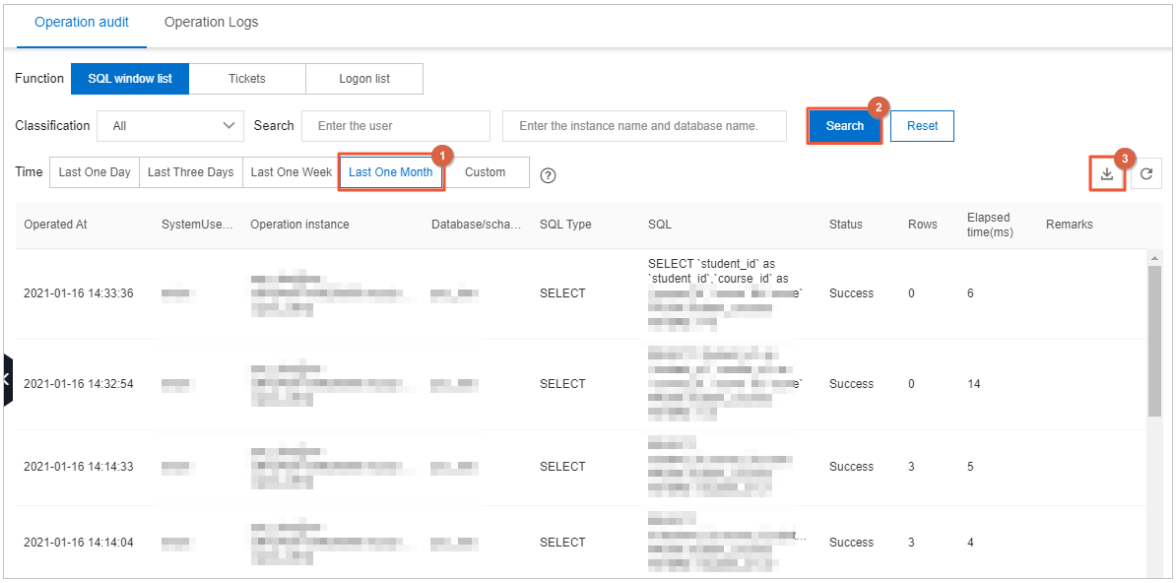
Auditing dimension	Limit	Link to operation audit	Supported role
Database	You can view and audit only the operations that are performed on the current database.	• On the SQLConsole tab of the database that you want to audit, click the icon in the upper-right corner. • In the left-side navigation pane of the DMS console, click the instance in which the database you want to audit resides, right-click the database, and then select **Operation audit**.	You can be a DMS administrator, a security administrator, a DBA, an instance owner, or a regular user. **Note** If you are a regular user, you can view and audit only the operations that you performed on the current database.
Instance	You can view and audit only the operations that are performed on the current instance.	In the left-side navigation pane of the DMS console, right-click the instance that you want to audit and select **Operation audit**.	You can be a DMS administrator, a security administrator, a DBA, an instance owner, or a regular user. **Note** If you are a regular user, you can view and audit only the operations that you performed on the current instance.
Global	You can view and audit all the operations that are performed in DMS.	In the top navigation bar, move the pointer over the **More** icon and choose **System > Operation audit**.	You can be a DMS administrator, a security administrator, or a DBA.

View and download operation records

This example shows you how to view and download all the SQL statements that are used in the SQLConsole in the last month.

1. Log on to the DMS console.
2. In the top navigation bar, move the pointer over the **More** icon and choose **System > Operation audit**.
   By default, a list of SQL statements appears.
3. Set the Time parameter to **Last One Month** and click **Search**.
   Then, DMS returns the search results.
4. Click the icon to download the results.

   Then, DMS exports an XLSX file that contains the search results on the current page.



> **Note** To preview and export more results, you can set the **Items per page** parameter to **100**.

16.1.11.8.6. Configure IP whitelists

DMS allows you to configure IP whitelists to control the service scope of DMS. You can allow user access to DMS only from specific trusted network environments.

Prerequisites

You are an administrator.

Procedure

1. Log on to the DMS console.
2. In the top navigation, choose **More > System > Access IP Whitelists**.
3. Perform the following operations based on your business requirements:
   - Enable or disable the whitelist control feature

     Click **Click to Open** or **Click to Close** to enable or disable the whitelist control feature.
   - Create a whitelist
     a. Click **Create Whitelist**.
     b. In the dialog box that appears, enter the IP addresses and description.

     > **Note**
     > - Separate IP addresses with semicolons (;). Make sure that each IP address in a whitelist is unique.
     > - You can specify IP addresses such as 10.23.12.24 or CIDR blocks such as 10.23.12.24/24, where /24 indicates the length of the IP address prefix in the CIDR block. The IP address prefix can be 1 to 32 bits in size.
     > - The 0.0.0.0/0 value indicates that all IP addresses are allowed.

     c. Click **Submit**.
   - Edit a whitelist
     a. Find the required IP address whitelist and click **Edit** in the **Actions** column of the IP address.

b. In the dialog box that appears, modify the IP address information.

c. Click **Submit**.

- Delete a whitelist

a. Find the required IP address whitelist and click **Delete** in the **Actions** column of the IP address.

b. In the message that appears, click **OK**.

> **Note** You cannot delete all IP address whitelists. At least one IP address whitelist must be retained.

16.1.11.8.7. Row-level control

In some cases, different users may access different rows in the same table, which can be achieved by using views. Data Management (DMS) provides an alternative solution that is called the row-level control feature to control access at the row level.

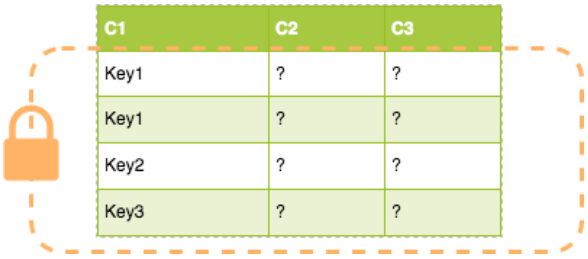
Prerequisites

You are a security administrator, a database administrator (DBA), or an administrator.

Context

Row-level control is used to provide horizontal data protection for tables. All the rows in a table are distinguished by one or more specified values. These values are called control values. To access a row that corresponds to a control value in the DMS console, you must have permissions on the row.

> **Note** A control value may correspond to multiple rows. If a user has permissions on a control value that corresponds to multiple rows, the user has permissions on all the rows that correspond to the control value.



Limits

- The sensitive data management feature applies only to relational databases, such as MySQL. However, this feature is unavailable for NoSQL databases.
- You can use the row-level control feature only on database instances managed in security collaboration mode.
- This feature applies only to physical databases. However, this feature is unavailable for logical databases.
- When you execute SQL statements to query, modify, or delete the data of a row-level control table, the following limits are set on filter conditions.

i. The control field must be specified in SQL statements to filter data.

ii. The system controls access to all the rows of a row-level control table. Users who do not have permissions on all rows can use only the `=` and `IN` operators to specify a control field. The control value that is specified in an SQL statement must be one of the control values for the table.

iii. Users who do not have permissions on all rows cannot use some operators, such as OR, XOR, and logical NOT.

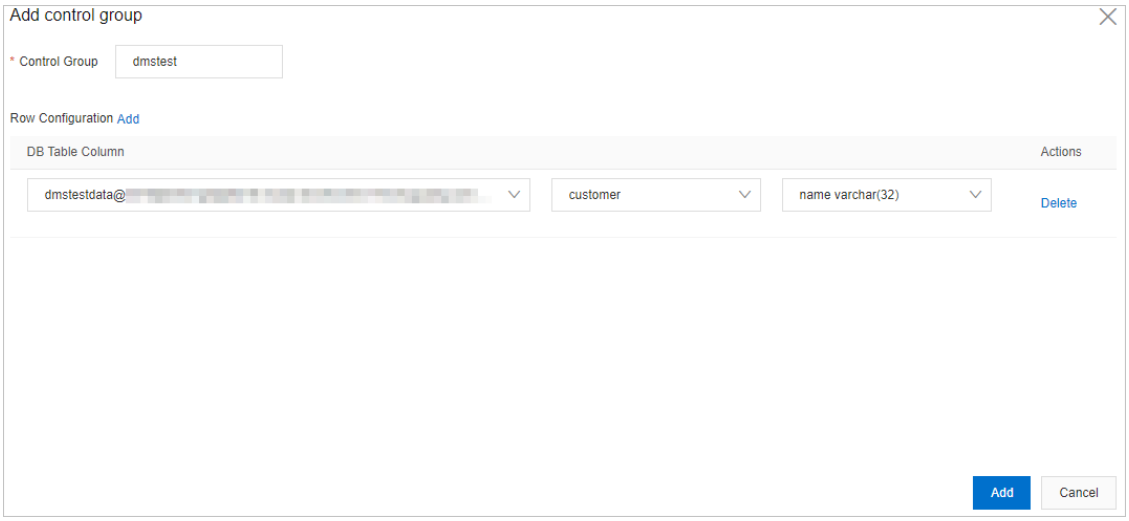
Terms

Term	Description
row permission	You can apply for permissions on a control value to access rows that correspond to the control value. Permissions on the rows of a table are defined as row permissions and are incorporated into the existing permissions of DMS. Permissions that can be controlled in security collaboration mode include permissions on databases, tables, columns (fields), and rows.
single control value	When a user applies for permissions on the rows of a row-level control table, the user can select **Single** to apply for permissions on a single control value. **Note** A control value may correspond to multiple rows. If a user has permissions on a control value that corresponds to multiple rows, the user has permissions on all the rows that correspond to this control value.
all control values	When a user applies for permissions on the rows of a row-level control table, the user can select **ALL** to apply for permissions on all control values. After the application is approved, the user has permissions on all the rows of the table. In this case, the user can access the entire row-level control table without limits. Even if the control values are changed or more control values are added, the user still has permissions on all the rows of the table.
row-level control table	A table that requires row-level control is called a row-level control table.
control field	A control field is a field to which the control values of a row-level control table are added.
control group	A control group is a group of row-level control tables that have the same control values. For example, if Table A and Table B have the same control values, you can add the two tables to a control group. This way, you can manage the two tables at the same time by using one set of control values.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > System > Sensitive Data**.
3. Click the **Row Level Security** tab.
4. Add a control group.
   i. Click **Add control group**.

ii. In the Add control group dialog box, set the required parameters.



Parameter	Description
Control Group	Enter a name for the control group.
Row Configuration	Click **Add** to add a row configuration in which you can specify a database, table, and field. **Note** You can repeat this step to add multiple row configurations.
DB Table Column	Search for databases by keyword and select a database. Then, select a table and a field from the drop-down lists. **Note** The selected field is the control field.

iii. Click **Add**.

5. Add control values.
   i. Find the new control group and click **Details** in the **Actions** column of the control group.
   ii. Click **Add Row Value**.
   iii. In the Import Row Value dialog box, specify whether to append row values and enter the required row values.

   Note Separate multiple row values with commas (,).

   iv. Click **Import**.

What to do next

After you configure row-level control settings for a table, a user may still have no permission on a control value that corresponds to one or more rows in the table. In this case, an error appears when the user queries row data. The error indicates that the user does not have permissions to access the row. The user can apply for permissions on the control value to access the rows. For more information, see Apply for permissions.

16.1.11.8.8. Manage sensitive data

Data Management (DMS) allows you to manage all classified sensitive and confidential fields in a unified manner. You can configure encryption algorithms for sensitive and confidential fields. This improves the control over the data masking feature.

Prerequisites

You are a security administrator, a database administrator (DBA), or an administrator.

Context

When you query a table that contains sensitive or confidential fields on which and you do not have permissions on the fields, the values of the fields are displayed as `******` in the query results. In this case, sensitive data is fully masked. In some scenarios, developers or test engineers may need to view a part of sensitive data for troubleshooting. To meet this requirement, you can configure masking algorithms to show some sensitive data.

Limits

- The sensitive data management feature applies only to relational databases such as MySQL. However, this feature is unavailable for NoSQL databases.
- To use this feature, the required database instance must be managed in security collaboration mode.

Procedure

1. Log on to the DMS console.
2. Specify security levels for fields in the required table.

   > **Note** If security levels are specified for the fields, skip this step.

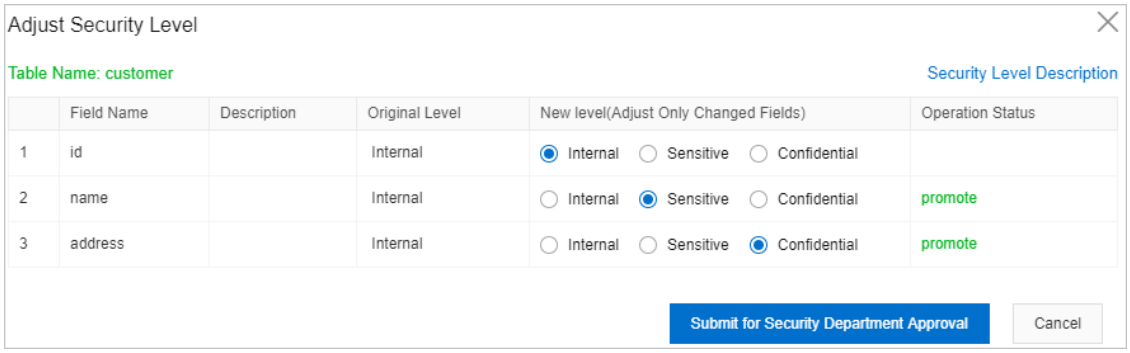
   i. In the left-side database instance list, click the icon next to the required database instance to show the databases in the instance.

   ii. Find the required database, right-click the database, and then select **Tables**.

   iii. Click the + icon next to the table name to show the table details.

   iv. Click **Adjust**.

   v. In the Adjust Security Level dialog box, change the security levels of fields.

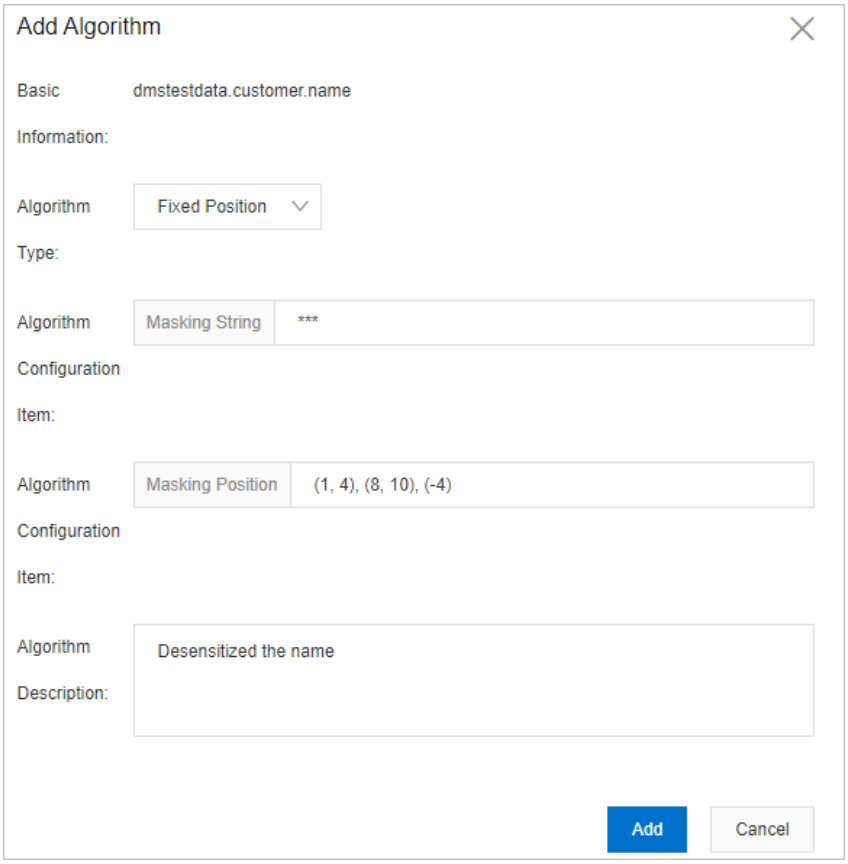
   vi. Click **Submit for Security Department Approval**.

   > **Note** The application to increase the security level of a field is automatically approved. The application to decrease the security level of a field is approved based on the approval process specified by an administrator or DBA.

   vii. In the message that appears, click **OK**.

3. In the top navigation bar, choose **More > System > Sensitive Data**.

4. Find the required field and click **Add Algorithm** in the **Actions** column of the field.
5. In the dialog box that appears, configure a masking algorithm.



Parameter	Description
Algorithm Type	The type of the algorithm. You can select an algorithm type based on your business requirements.

Parameter	Description
Algorithm Configuration Item	The algorithm configuration items vary based on the specified algorithm type. ○ **Fixed Position** algorithm type You must set the Masking String and Masking Position parameters. For example, you can set the Masking String parameter to ***. The Masking Position parameter specifies the positions of the characters to be masked in the field values. The positions are in the format of coordinates. Examples: ■ (1, 4): masks the first four characters. You can also enter (4) to simplify the format. ■ (-4): masks the last four characters. **Note** You can specify a maximum of three positions. For example, `(1, 4), (8, 10), (-4)` indicates to mask the first four characters, the eighth to tenth characters, and the last four characters. ○ **Fixed Character** algorithm type You must set the Masking String and Character to Be Replaced parameters. **Note** The Character to Be Replaced parameter specifies the characters that you want to mask in the format of a string. You can specify a maximum of three strings. ○ **Full Masking** algorithm type You need to set only the Masking String parameter.
Algorithm Description	Enter a description that can help you identify the algorithm.

6. Click **Add**.

16.1.11.8.9. Data protection

Data Management (DMS) provides the data protection feature. This topic describes how to use the data protection feature.

Context

Data protection is one of the greatest challenges in the process of data application. Data is the core of an enterprise. Most critical and sensitive information is stored as structured data. The information includes ID numbers, bank account numbers, phone numbers, customer records, medical records, transaction records, and salary records. Security events may cause enormous economic loss and damage the reputation of enterprises. These security events include data tampering, data theft, and data misuse.

To ensure data security, you need to understand how to protect your data from these security risks. The data protection feature helps the security management team achieve the following goals:

- Intelligently recognizes and classifies sensitive data. Then, you can group the data based on the security level of fields for DMS.
- Audits databases and prevents against data loss.
- Provides an efficient method to identify data application mode. Then, you can use the user and entity behavior analytics identification model and big data security expert rules of Ant Financial to identify and manage risks.
- Provides unified masking SDKs to protect sensitive information. These SDKs are used to intelligently identify sensitive data that exists in the system content that is displayed and mask the data. These operations are performed based on the definition of sensitive information and masking policies that are specified in the data

protection feature. These SDKs are also used to provide a unified method to manage masking rules within an enterprise, and increase the efficiency of security management.

Control access

Data protection is a tool that is provided by DMS to data security administrators. To enable and use the data protection feature, you must log on to the DMS console with security administrator permissions.

1. Log on to the DMS console.

   > **Note** You must use an account with administrator permissions to log on to the DMS console.

2. In the top navigation bar, choose **More > System > User**.
3. Click **Change**.
4. Select **Security Administrator** and click **Confirm Change**.
5. Use the account to re-log on to the Log on to the DMS console again.
6. In the top navigation bar, choose **More > Security > Data Protection**.

   > **Note** To enable a user to use the data protection feature, you must log on to the DMS console with security administrator permission to enable the feature and authorize the user to use the feature.

Data classification

The data protection feature classifies fields based on the custom automatic classification setting that is specified in the metadata of these fields. The classification results are used to update the levels of these DMS fields. This feature facilitates the access control over DMS fields.

1. Log on to the DMS console.

   > **Note** You must use an account with security administrator permissions to log on to the DMS console.

2. In the top navigation bar, choose **More > Security > Data Protection**.
3. In the left-side navigation pane, choose **Rule Configuration > Data Identification Rules**.
4. In the upper-right corner, click **Create Rule**. Then, set the **Data Type**, **Data Name**, **Owner** and **Remarks** parameters.
5. Click **Next**.
6. Set the Level parameter. Valid values: Internal, Sensitive, and Confidential. Select Field Scanning in the Data Recognition Rules field.
7. Click **Next**. After the rule is configured, click **Save and Enable**.

   > **Note**
   > - You can view all rules on the **Data Identification Rules** page. You can also modify, disable, or enable a rule on this page.
   > - After a rule is applied, the system identifies and classifies data based on metadata every hour on the hour. The security level of a field in DMS Enterprise is updated based on the classification result. This enables field-level access control over DMS operations.

Manually correct data

The Manual Data Correction page shows all identified fields. This way, you can verify these fields. If some fields are incorrectly identified, you can remove these fields or change the related types. After data is corrected, the results are immediately synchronized to the DMS Enterprise console.

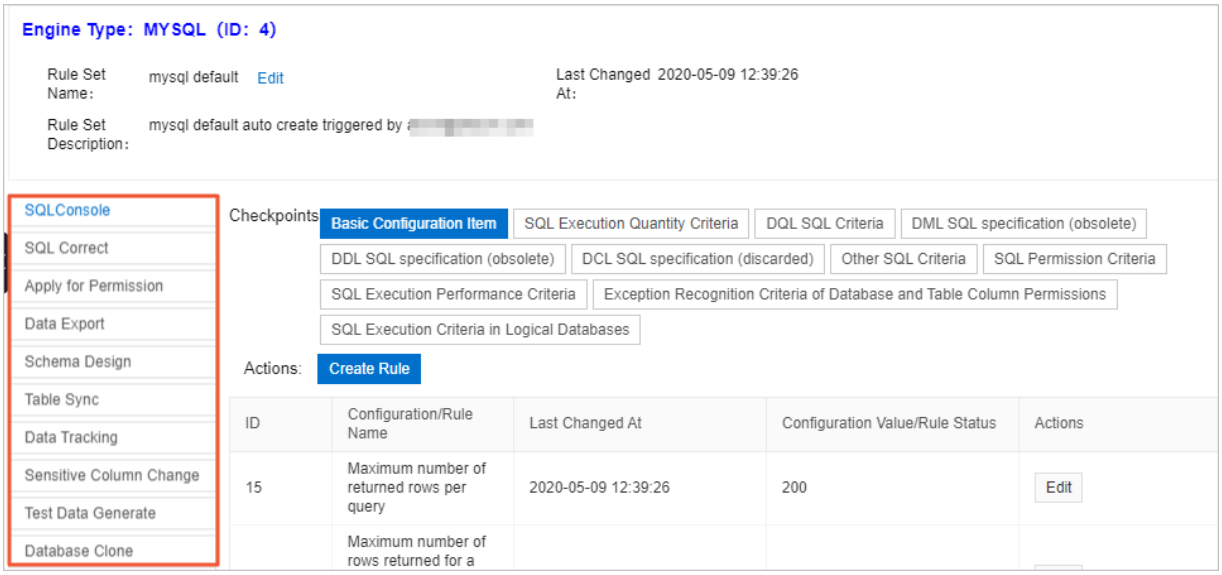
Data detection

The data detection feature is used to collect statistics based on the results of data identification from multiple dimensions and show the details of fields that are identified in the field details list. These dimensions include the security level and related instance.

16.1.11.9. Security rules

16.1.11.9.1. Overview of security rule sets

Security rule sets are implemented by using a collection of domain-specific languages (DSLs) to control user access to databases based on several factors. These factors include the type of databases, the syntax of database operations, and the number of affected rows. You can use security rule sets to standardize database operations, development processes, and approval processes as required.



This topic describes the features that are supported by security rule sets. You can click the link of a feature to view the information about the feature. The information includes the basic configuration items, checkpoints, factors, actions, and supported statements or commands.

- SQLConsole for relational databases
- SQLConsole for MongoDB
- SQLConsole for Redis
- Data change
- Permission application
- Data export
- Schema design
- Database and table synchronization
- Sensitive field change
- Test data generation
- Database cloning

16.1.11.9.2. Manage security rules under checkpoints

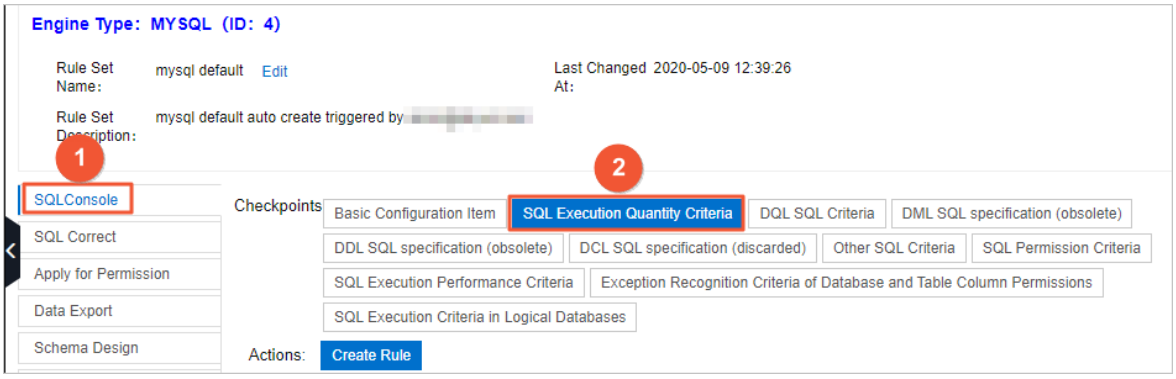
This topic describes how to configure a security rule under the SQL Execution Quantity Criteria checkpoint on the SQL Editor tab. You can configure security rules under other checkpoints by using a method similar to that described in this topic.

Procedure

1. Log on to the DMS console.
2. In the top navigation bar, choose **More > System > Security Rules**.
3. Find the required security rule set and click **Edit** in the **Actions** column of the security rule.

   > **Note** In this example, the security rule that is configured for MySQL databases is used.

4. Click a tab and then click a checkpoint based on your business requirements. In this example, click the **SQL Editor** tab and then click the **SQL Execution Quantity Criteria** checkpoint.



   > **Note**
   > - For more information about the tabs and checkpoints, see Overview of security rule sets.
   > - You can click **Create Rule** to create a security rule. For more information about the syntax, see DSL syntax for security rules.

5. Find the required security rule and click **Edit** in the **Actions** column of the security rule.

   > **Note** You can click **Disable** to disable a security rule or click **Delete** to delete a security rule.

6. In the Change Rule - SQL Editor dialog box, modify the DSL statements of the security rule based on your business requirements. For more information about the syntax, see DSL syntax for security rules. In this example, change the maximum number of SQL statements that can be executed at a time from 1,000 to 500.

   > **Note**
   > - A large number of security rule templates are provided for each checkpoint. You can click **Load from Template Database** to use a template.
   > - For more information about factors and actions, see Overview of security rule sets.

7. Click **Submit**.

16.1.11.9.3. SQLConsole for relational databases

DMS allows you to manage relational and non-relational databases on the SQLConsole tab. The definition and classification of security rules are different for relational and non-relational databases. This topic describes the security rules for relational databases on the SQLConsole tab, such as MySQL databases.

Default security rules

- Constraints on SQL statement categories: No constraints are imposed on data query language (DQL) statements. By default, DML statements, DDL statements, data control language (DCL) statements, and SQL statements that

cannot be identified by DMS are all blocked. To execute DML, DDL, or DCL statements on the SQLConsole tab, you must configure and enable corresponding security rules.

- Constraints on permissions on databases, tables, and fields: By default, users can perform operations on databases, tables, and fields without permission validation. To enable permission validation, you must configure and enable security rules under the **SQL Permission Criteria** checkpoint. For more information, see Supported checkpoints.

Basic configuration items

Configuration item	Description
Maximum number of returned rows per query	The maximum number of rows that can be returned for a query.
Maximum number of rows returned for a single query with sensitive column conditions	The maximum number of rows that can be returned for a query that contains query conditions for sensitive fields.
Limit the maximum allowed SQL full table scan (MB)	The maximum size of data that can be scanned. Before an SQL statement is executed, DMS checks the execution plan. If the size of the data to be scanned exceeds the specified threshold, the SQL statement fails to be executed. **Note** This item can be configured only for MySQL and Oracle databases.
Turn off the execution of change SQL validation affects the number of rows and prompts	Specifies whether to check the number of rows to be affected and display a prompt before DMS executes an SQL statement to change data. By default, this item is disabled.
How many rows does result set page support	The maximum number of rows that can be returned in the query result set on the SQLConsole tab.
Does the result set support paging	Specifies whether the query result set can be displayed on multiple pages on the SQLConsole tab.
Does the result set support editing	Specifies whether the query result set can be edited on the SQLConsole tab.

Supported checkpoints

Note Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description
SQL Execution Quantity Criteria	Allows you to limit the number of SQL statements that can be submitted at a time.
DQL SQL Criteria	Allows you to set constraints on DQL statements.

Checkpoint	Description
Other SQL Criteria	Allows you to set constraints on multiple categories of SQL statements. Different enterprises may define different high-risk SQL statements, which may include specific subcategories of DML, DCL, and DDL statements. **Note** You can also set constraints on SQL statements that cannot be identified by DMS.
SQL Permission Criteria	Allows you to set constraints on the execution of SQL statements from the aspect of permissions. For example, DMS checks whether a user has the required permissions on the corresponding databases, tables, and fields.
SQL Execution Performance Criteria	Allows you to set constraints on the execution of SQL statements from the aspect of performance. For example, you can specify that a DML statement is not executed if the number of rows to be affected by the statement exceeds the specified threshold, or that a DDL statement is not executed if the size of the table involved exceeds the specified threshold.
Exception Recognition Criteria of Database and Table Column Permissions	After a user submits SQL statements on the SQLConsole tab, DMS parses the SQL statements and checks whether the user has the required permissions on the corresponding databases, tables, and fields. You can configure security rules under this checkpoint to ensure that if exceptions occur when DMS parses complex SQL statements, these statements can be executed. **Note** If you configure and enable security rules under the **Exception Recognition Criteria of Database and Table Column Permissions** checkpoint, security rules under the SQL Permission Criteria, DQL SQL Criteria, Other SQL Criteria, and SQL Execution Performance Criteria checkpoints are automatically disabled.
SQL Execution Criteria in Logical Databases	This checkpoint is reserved for logical databases and not suitable for physical databases.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`.The following table describes the supported factors for relational databases on the SQLConsole tab.

Factor	Description
@fac.sql_count	The number of SQL statements that are submitted at a time.
@fac.select_sql_count	The number of DQL statements among the SQL statements that are submitted at a time.
@fac.dml_sql_count	The number of DML statements among the SQL statements that are submitted at a time.
@fac.sql_type	The category and subcategory of the SQL statement. For more information, see Supported SQL statements.
@fac.sql_sub_type	
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.

Factor	Description
@fac.fulltable_delete	A Boolean value that indicates whether the current SQL statement deletes a full table. Valid values: *true* and *false*.
@fac.fulltable_update	A Boolean value that indicates whether the current SQL statement updates a full table. Valid values: *true* and *false*.
@fac.current_sql	The current SQL statement.
@fac.user_is_admin	A Boolean value that indicates whether the current user is a DMS administrator. Valid values: *true* and *false*.
@fac.user_is_dba	A Boolean value that indicates whether the current user is a DBA. Valid values: *true* and *false*.
@fac.user_is_inst_dba	A Boolean value that indicates whether the current user is the DBA of the current instance. Valid values: *true* and *false*.
@fac.user_is_sec_admin	A Boolean value that indicates whether the current user is a security administrator. Valid values: *true* and *false*.
@fac.sql_affected_rows	The number of rows to be affected by the current SQL statement. **Warning** This factor triggers COUNT operations, which may affect the database performance. Use this factor with caution.
@fac.sql_relate_table_store_size	The estimated total size of the table to be accessed by the current SQL statement. Unit: MB. **Note** This value is estimated based on the metadata that is obtained by DMS. It is not an actual value.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions for relational databases on the SQLConsole tab.

Action	Description
@act.reject_execute	Rejects the request to execute the current SQL statement.
@act.allow_execute	Allows the current SQL statement to be executed.
@act.reject_sql_type_execute	Rejects the request to execute a specific subcategory of SQL statements. You must specify an SQL statement subcategory after the action name. Example: `@act.reject_sql_type_execute 'UPDATE'`.
@act.allow_sql_type_execute	Allows a specific subcategory of SQL statements to be executed. You must specify an SQL statement subcategory after the action name. Example: `@act.allow_sql_type_execute 'UPDATE'`.

Action	Description
@act.check_dml_sec_column_permission	Checks whether a user has the required permissions on sensitive fields. If the user does not have the permissions, the DML statement for data change is not executed.
@act.uncheck_dml_sec_column_permission	Does not check whether a user has the required permissions on sensitive fields.
@act.check_sql_access_permission	Checks whether a user has the required permissions, such as query and change permissions, on the databases, tables, and fields that are involved in the SQL statements to be executed.
@act.uncheck_sql_access_permission	Does not check whether a user has the required permissions on the objects that are involved in the SQL statements to be executed.
@act.enable_sec_column_mask	De-identifies sensitive fields in query result sets that are returned for SQL statements that are submitted by users who do not have permissions on the sensitive fields.
@act.disable_sec_column_mask	Does not de-identify sensitive fields in query result sets that are returned for SQL statements that are submitted by users who do not have permissions on the sensitive fields.

Supported SQL statements

Category	Subcategory
DQL	• SELECT • DESC • EXPLAIN • SHOW
DML	• INSERT • INSERT_SELECT • REPLACE • REPLACE_INTO • UPDATE • DELETE • MERGE

Category	Subcategory
DDL	• DATABASE_OP • CREATE • CREATE_INDEX • CREATE_VIEW • CREATE_SEQUENCE • CREATE_TABLE • CREATE_SELECT • TRUNCATE • DROP_INDEX • DROP_VIEW • DROP_TABLE • RENAME • ALTER • ALTER_INDEX • ALTER_VIEW • ALTER_TABLE • ALTER_SEQUENCE • CREATE_FUNCTION • CREATE_PROCEDURE • ALTER_FUNCTION • ALTER_PROCEDURE • DROP_FUNCTION • DROP_PROCEDURE
DCL	• GRANT • DECLARE • SET • ANALYZE • FLUSH • OPTIMIZE • KILL

16.1.11.9.4. SQLConsole for MongoDB

DMS allows you to manage relational and non-relational databases on the SQLConsole tab. The definition and classification of security rules are different for relational and non-relational databases. This topic describes the security rules for MongoDB databases on the SQLConsole tab.

Basic configuration items

Maximum number of returned rows per query: the maximum number of rows that can be returned for a query.

Supported checkpoints

Note Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description
User Permission Validation	Allows you to specify whether to check the permissions of specific users when they submit commands.
Collection Statement Criteria	Allows you to specify whether to allow DMS to run a specific category of commands.
DB Statement Criteria	
Cache Query Statement Criteria	
User Management Statement Criteria	
Role Management Statement Criteria	
Replication Set Statement Criteria	
Sharding Statement Criteria	

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as command categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`.The following table describes the supported factors for MongoDB databases on the SQLConsole tab.

Factor	Description
@fac.sql_sub_type	The subcategory of the current command. For more information about the supported commands, see Supported MongoDB commands.
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as `DEV` or `PRODUCT`.
@fac.current_sql	The current command.
@fac.user_is_admin	A Boolean value that indicates whether the current user is a DMS administrator. Valid values: *true* and *false*.
@fac.user_is_dba	A Boolean value that indicates whether the current user is a DBA. Valid values: *true* and *false*.
@fac.user_is_inst_dba	A Boolean value that indicates whether the current user is the DBA of the current instance. Valid values: *true* and *false*.
@fac.user_is_sec_admin	A Boolean value that indicates whether the current user is a security administrator. Valid values: *true* and *false*.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions for MongoDB databases on the SQLConsole tab.

Action	Description
@act.reject_execute	Rejects the request to run the current command.
@act.allow_execute	Allows the current command to be run.
@act.reject_sql_type_execute	Rejects the request to run a specific subcategory of commands. You must specify a subcategory after the action name. Example: `@act.reject_sql_type_execute 'UPDATE'` .
@act.allow_sql_type_execute	Allows a specific subcategory of commands to be run. You must specify a subcategory after the action name.

Supported MongoDB commands

Category	Subcategory	Command
Collection commands	Query commands	• aggregate • find • findOne • count • distinct • getIndexes • getShardDistribution • isCapped • stats • dataSize • storageSize • totalIndexSize • totalSize
	Data update commands	• insert • save • findAndModify • remove • update
	Collection modification commands	• drop • renameCollection
	Index modification commands	• createIndex • createIndexes • dropIndexes • reIndex
	Other commands	validate

Category	Subcategory	Command
Database commands	Database query commands	• commandHelp • currentOp • getCollectionInfos • getCollectionNames • getLastError • getLastErrorObj • getLogComponents • getPrevError • getProfilingStatus • getReplicationInfo • getSiblingDB • help • isMaster • listCommands • printCollectionStats • printReplicationInfo • version • serverBuildInfo • serverStatus,stats
	Collection creation commands	createCollection
	High-risk commands	• dropDatabase • fsyncLock • fsyncUnlock • killOp • repairDatabase • resetError • runCommand
Commands related to the query plan cache	Read commands	• getPlanCache • getPlansByQuery • listQueryShapes
	Write commands	clearPlansByQuery
User management commands	User query commands	• getUser • getUsers
	User modification commands	• createUser • changeUserPassword • dropUser • dropAllUsers • grantRolesToUser • revokeRolesFromUser • updateUser

Category	Subcategory	Command
Role management commands	Role query commands	• getRole • getRoles
	Role modification commands	• createRole • dropRole • dropAllRoles • grantPrivilegesToRole • revokePrivilegesFromRole • revokeRolesFromRole • updateRole
Replica set commands	N/A	• help • printReplicationInfo • status • conf
Sharding commands	N/A	• getBalancerState • isBalancerRunning

16.1.11.9.5. SQLConsole for Redis

DMS allows you to manage relational and non-relational databases on the SQLConsole tab. The definition and classification of security rules are different for relational and non-relational databases. This topic describes the security rules for Redis databases on the SQLConsole tab.

Supported checkpoints

Note Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description
Permission Execution Statement Criteria	Allows you to set constraints on the permissions for command execution.
Statement Criteria: Keys	Allows you to specify whether to check the permissions of specific users when they submit commands.
Statement Criteria: String	Allows you to specify whether to allow the execution of various Redis commands.
Statement Criteria: List	
Statement Criteria: SET	
Statement Criteria: SortedSet	
Statement Criteria: Hash	

Checkpoint	Description
Statement Criteria: Other	

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`. The following table describes the supported factors for Redis databases on the SQLConsole tab.

Factor	Description
@fac.cmd_type	The type of the Redis command. For more information about valid values, see Supported Redis commands.
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as `DEV` or `PRODUCT`.
@fac.is_read	A Boolean value that indicates whether the current command is a read command. Valid values: *true* and *false*.
@fac.is_write	A Boolean value that indicates whether the current command is a write command. Valid values: *true* and *false*.
@fac.current_sql	The current command.
@fac.user_is_admin	A Boolean value that indicates whether the current user is a DMS administrator. Valid values: *true* and *false*.
@fac.user_is_dba	A Boolean value that indicates whether the current user is a DBA. Valid values: *true* and *false*.
@fac.user_is_inst_dba	A Boolean value that indicates whether the current user is the DBA of the current instance. Valid values: *true* and *false*.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`. The following table describes the supported actions for Redis databases on the SQLConsole tab.

Action	Description
@act.reject_execute	Rejects the request to run the current command.
@act.allow_execute	Allows the current command to be run.

Supported Redis commands

Category	Subcategory	Command

Category	Subcategory	Command
Key-related commands	Key-related read commands	• EXISTS • TTL • PTTL • RANDOMKEY • TYPE • SCAN • OBJECTS
	Key-related write commands	• DEL • DUMP • EXPIRE • EXPIREART • MOVE • PERSIST • PEXPIRE • PEXPIREAT • RENAME • RENAMENX • RESTORE • SORT • TOUCH • UNLIMK • WAIT • MIGRATE
String-related commands	String-related read commands	• GET • GETRANGE • BITCOUNT • GETBIT • MGET • STRLEN • BITOPS

Category	Subcategory	Command
	String-related write commands	• APPEND • BITFIELD • BITOP • DECR • DECRBY • GETSET • INCR • INCRBY • INCRBYFLOAT • MSET • MSETNX • PSETEX • SET • SETNX
List-related commands	List-related read commands	• LINDEX • LLEN • LRANGE
	List-related write commands	• BLPOP • BRPOP • BRPOPLPUSH • LINSERT • LPOP • LPUSH • LPUSHX • LREM • LSET • LTRIM • RTOP • RPOPLPUSH • RPUSH • RPUSHX
Set-related commands	Set-related read commands	• SCARD • SISMEMBER • SRANDMEMBER • SSCAN
	Set-related write commands	• SADD • SMOVE • SPOP • SREM

Category	Subcategory	Command
Sorted set-related commands	Sorted set-related read commands	• ZCARD • ZCOUNT • ZLEXCOUNT • ZRANGE • ZRANGEBYLEX • ZRANGEBYSCORE • ZRANK • ZREVRNGE • ZREVRANGEBYLEX • ZREVRANGEBYSCORE • ZREVRANK • ZSCAN • ZSCORE
	Sorted set-related write commands	• ZADD • ZINCRBY • ZINTERSTORE • ZPOPMAX • ZPOPMIN • ZREM • ZUNIONSTORE • BZPOPMIN • BZPOPMAX
Hash-related commands	Hash-related read commands	• HEXISTS • HGET • HLEN • HMGET • HSCAN • HSTRLEN
	Hash-related write commands	• HDEL • HINCRBY • HINCRBYFLOAT • HMESET • HSET • HSETNX

16.1.11.9.6. Data change

In DMS, you can execute SQL statements for data changes. However, the execution requires a high level of security. DMS allows you to configure security rules on the SQL Correct tab to validate the submission and approval of tickets for data changes. Only the SQL statements that are validated by the security rules can be executed.

Background information

Based on a DSL, new security rules are flexible to use. You can apply new security rules to define risk levels for tickets so that a ticket can be submitted to the approval process that is designed for the specified risk level. For more information, see DSL syntax for security rules.

Basic configuration items

Configuration item	Description
Data change default approval Template	By default, this approval template takes effect if you do not configure different approval rules for data changes at different risk levels under the Risk Approval Rules checkpoint. In the Switch Approval Template dialog box, you can change the approval process of the default approval template. For more information, see Customize approval processes.
Data Change risk level list	This risk level list defines risk levels that are used in the **Risk Identification Rules** and **Risk Approval Rules** checkpoints to identify and classify risks in data changes. You can set risk levels based on the type and scenario of data changes. Data changes at different risk levels are submitted to different approval processes. DMS allows you to set the following four risk levels: • *low*: a low risk level. • *middle*: a medium risk level. • *high*: a high risk level. • *highest*: a critical risk level.

Supported checkpoints

Note Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description

SQL execution rules	SQL execution rules are used to limit the SQL statements that can be submitted for execution. If you do not enable SQL execution rules, all SQL statements that are used for data changes cannot be executed. Assume that you want to use DML statements to change the data of a database in an online environment. You can create the following SQL execution rule:
	Example:
	`if`
	`  @fac.env_type not in ['product']`
	`  and`
	`  @fac.sql_type in [ 'UPDATE','DELETE','INSERT']`
	`then`
	`  @act.allow_submit`
	`end`
	Note:
	The preceding rule specifies that you can only submit data change tickets to execute UPDATE, DELETE, and INSERT statements on a database that is deployed in an online environment.

Checkpoint	Description
Risk Identification Rules	If a ticket conforms to the preset SQL execution rules, DMS continues to validate the ticket based on the risk identification rules. Risk identification rules are used to identify and classify risks in data changes. You can create risk identification rules based on your database environment, the number of rows in which data is affected, and the categories and subcategories of SQL statements. **Note** Different risk identification rules apply to different check items. DMS automatically identifies the highest risk level for a data change. For example, if the risk level of a data change is identified as high, medium, and low by one, three, and five risk identification rules, DMS assumes that the data change is at high risk. Example: `if @fac.env_type not in ['product','pre'] then @act.mark_risk 'low 'Low risk level: offline environment' end` Note: The preceding rule specifies that if the destination database is deployed in an offline environment, data changes are at low risk.
Risk Approval Rules	After the risk level of a data change is identified by the risk identification rules, DMS processes the ticket based on the risk approval rules. You can customize risk approval rules under the **Risk Approval Rules** checkpoint. **Note** • If a data change does not hit risk approval rules, DMS uses the default approval template that is specified under the Basic Configuration Item checkpoint to process the ticket. • By default, an offline environment is identified as a factor at low risk and requires no approval.
Batch Data import rules	These rules apply only to the validation of data import tickets. You can use the default rules that are provided in templates, or configure rules based on your actual needs.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`. The following table describes the supported factors on the SQL Correct tab.

Factor	Description
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.
@fac.sql_type	The type of the SQL statement. The value is the subcategory of the SQL statement, such as UPDATE or INSERT. For more information, see Supported SQL statements.

Factor	Description
@fac.detail_type	The type of the data change. Valid values: • *COMMON*: a Normal Data Modify ticket. • *CHUNK_DML*: a Lock-Free Data Modify ticket. • *PROCEDURE*: a Programmable Object ticket. • *CRON_CLEAR_DATA*: a History Data Clean ticket. • *BIG_FILE*: a Large Data Import ticket.
@fac.is_logic	A Boolean value that indicates whether the database to be affected is a logical database.
@fac.extra_info	The additional information about the data change. This factor is not in use.
@fac.is_ignore_affect_rows	A Boolean value that indicates whether to skip the validation.
@fac.insert_rows	The number of rows of data to be inserted.
@fac.update_delete_rows	The number of rows of data to be updated.
@fac.max_alter_table_size	The size of the largest tablespace where the table to be modified is stored.
@fac.is_has_security_column	A Boolean value that indicates whether the SQL statement to be executed involves sensitive fields.
@fac.security_column_list	A list of sensitive fields that the SQL statement to be executed involves.
@fac.risk_level	The risk level of the operation that is to be performed by the SQL statement.
@fac.risk_reason	The reason for identifying the operation to be performed as at the risk level.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.` .The following table describes the supported action on the SQL Correct tab.

Action	Description
@act.allow_submit	Requires the submission of SQL statements to be executed in a ticket.
@act.allow_execute_direct	Allows the execution of SQL statements in the SQLConsole.
@act.forbid_execute	Forbids the execution of SQL statements.
@act.mark_risk	Marks the risk level of a data change. Example: `@act.mark_risk 'Medium risk level: online environment'` .
@act.do_not_approve	Specifies the ID of an approval template.
@act.choose_approve_template	
@act.choose_approve_template_with_reason	

16.1.11.9.7. Permission application

DMS allows you to configure security rules on the Access apply tab to validate applications for permissions, including permissions on instances, databases, and tables.

Background information

In DMS, security rules are flexible to use. You can apply security rules to define risk levels for tickets so that a ticket can be submitted to the approval process that is designed for the specified risk level. For more information about the DSL syntax, see DSL syntax for security rules.

Basic configuration items

The following table describes the basic configuration items that are supported on the Access apply tab.

Configuration item	Description
[Instance-permission application] default approval Template	By default, this approval template takes effect if you do not set different approval processes for instance permission applications at different risk levels under the **Validation for Instance Permission Application** checkpoint. **Note** In the Switch Approval Template dialog box, you can change the approval process of the default approval template.
[DB-permission application] default approval Template	By default, this approval template takes effect if you do not set different approval processes for database permission applications at different risk levels under the **Database Permission Application Validation** checkpoint.
Table-permission request default approval Template	By default, this approval template takes effect if you do not set different approval processes for table permission applications at different risk levels under the **Table Permission Application Validation** checkpoint.
[Programmable object-permission application] default approval Template	By default, this approval template takes effect if you do not set different approval processes for programmable object permission applications at different risk levels under the **Programmable object verification** checkpoint.
[Field-permission application] default approval Template	By default, this approval template takes effect if you do not set different approval processes for sensitive field permission applications at different risk levels under the **Sensitive Field Application Validation** checkpoint.
Line-permission application default approval Template	By default, this approval template takes effect if you do not set different approval processes for row permission applications at different risk levels under the **Line permission application verification** checkpoint.
[Owner-application] default approval template (when the resource has no Owner)	By default, this approval template takes effect if you do not set different approval processes for data ownership applications at different risk levels under the **Owner Application Validation** checkpoint and the data that is involved in the application has no owner.
[Owner-application] default approval template (when the resource has an Owner)	By default, this approval template takes effect if you do not set different approval processes for data ownership applications at different risk levels under the **Owner Application Validation** checkpoint and the data that is involved in the application has one or more owners.

Supported checkpoints

When a user submits a ticket to apply for permissions, DMS checks whether the ticket conforms to rules that are specified under checkpoints. The ticket can be submitted only after DMS determines that the ticket conforms to all rules that are specified under checkpoints.

> **Note** Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description
Owner Application Validation	Allows you to set approval processes or constraints for **Instance-OWNER**, **Table-OWNER**, and **Database-OWNER** tickets.
Validation for Instance Permission Application	Allows you to set approval processes or constraints for **Instance-Performance** and **Instance-Login** tickets.
Database Permission Application Validation	Allows you to set approval processes or constraints for **Database-Permission** tickets.
Table Permission Application Validation	Allows you to set approval processes or constraints for **Table-Permission** tickets.
Programmable object verification	Allows you to set approval processes or constraints for **Programmable Object** tickets.
Sensitive Field Application Validation	Allows you to set approval processes or constraints for **Sensitive Column-Permission** tickets.
Line permission application verification	Allows you to set approval processes or constraints for **Row-Permission** tickets.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and database names. A factor name starts with the prefix `@fac.`.The following table describes the supported factors on the Access apply tab.

Factor	Description
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.
@fac.schema_name	The name of the database.
@fac.perm_apply_duration	The period of time during which the applicant needs the permission. Unit: hours.
@fac.column_security_level	The security level of the field. Valid values: • *sensitive* • *confidential* • *inner*

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported action on the Access apply tab.

Action	Description
@act.forbid_submit_order	Forbids a ticket from being submitted.
@act.do_not_approve	Specifies the ID of an approval template.
@act.choose_approve_template	
@act.choose_approve_template_with_reason	

16.1.11.9.8. Data export

DMS allows you to manage security rules on the Data Export tab to validate the permissions of applicants on involved databases, tables, sensitive fields, and rows during the submission and approval of data export tickets. This helps ensure data security.

Basic configuration items

Data export default approval Template: the default approval template that takes effect if you do not set different approval processes for data export tickets at different risk levels under the Approval Rule Validation checkpoint. In the Switch Approval Template dialog box, you can change the approval process of the default approval template. For more information, see Customize approval processes.

Supported checkpoints

Note Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description
Pre-check Validation	Allows you to specify whether to validate the permissions of applicants on involved databases, tables, sensitive fields, and rows by configuring security rules.
Approval Rule Validation	Allows you to submit data export tickets to different approval processes by configuring security rules. For example, you can submit tickets for exporting more than a specific number of rows to an approval process and other tickets to another approval process.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`.The following table describes the supported factors on the Data Export tab.

Factor	Description
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.
@fac.is_ignore_export_rows_check	A Boolean value that indicates whether to skip the check on the number of rows to be affected.
@fac.export_rows	The number of rows to be exported.
@fac.include_sec_columns	A Boolean value that indicates whether the data to be exported contains sensitive fields.

Factor	Description
@fac.sec_columns_list	The sensitive fields that require or do not require approval before data is exported. The sensitive fields are displayed in the format of `Table name.Field name,[Table name.Field name, ...]`.
@fac.user_is_admin	A Boolean value that indicates whether the applicant is a DMS administrator.
@fac.user_is_dba	A Boolean value that indicates whether the applicant is a DBA.
@fac.user_is_inst_dba	A Boolean value that indicates whether the applicant is the DBA of the current instance.
@fac.user_is_sec_admin	A Boolean value that indicates whether the applicant is a security administrator.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions on the Data Export tab.

Action	Description
@act.do_not_approve	Allows a ticket to be processed without approval.
@act.choose_approve_template	Specifies an approval template.
@act.choose_approve_template_with_reason	Specifies an approval template with a reason provided.
@act.forbid_submit_order	Forbids a ticket from being submitted.
@act.enable_check_permission	Validates the permissions of an applicant on involved databases and tables.
@act.disable_check_permission	Does not validate the permissions of an applicant on involved databases and tables.
@act.enable_check_sec_column	Validates the permissions of an applicant on involved sensitive fields.
@act.disable_check_sec_column	Does not validate the permissions of an applicant on involved sensitive fields.

16.1.11.9.9. Schema design

DMS allows you to configure security rules on the Schema Design tab to check the design rules and risk identification rules that apply to schema design tickets. This helps ensure data security.

Basic configuration items

Configuration item	Description
Enable non-peer Publishing	Specifies whether to enable non-peer publishing. By default, data changes to a table can be published only to a table with the same name in another database. After you enable non-peer publishing, you can perform data changes on all tables. **Warning** This feature may bring high risks. We recommend that you proceed with caution and enable this feature only for special requirements.
R & D process	The whole process of a schema design ticket. It is the most important configuration item on the Schema Design tab. For more information about the parameters of the configuration item, see Parameters involved in the R&D process.
Field type configuration	The supported data types of fields to be added.
Index type configuration	The supported data types of indexes to be added.
It is forbidden to modify the original field data type	Specifies whether to prohibit the data types of the original fields from being modified when the original table is to be modified.
Prohibit deleting original fields	Specifies whether to prohibit the existing fields from being deleted when the original table is to be modified. **Note** We recommend that you enable this feature because deleting existing fields may bring high risks.
Prohibit renaming original fields	Specifies whether to prohibit the existing fields from being renamed when the original table is to be modified. **Note** We recommend that you enable this feature because reaming existing fields may bring high risks.
Table character set license configuration	The range of character sets that are allowed to be used when you create a table. For example, you can specify utf8 and utf8mb4.
Default approval template for Structural design	The default approval template that is used for a schema design ticket if you do not configure the Approval Rule Validation checkpoint. In the Switch Approval Template dialog box, you can change the approval process of the default approval template.
When published, the ticket will automatically advance to the end state	The point that is used to stop the schema change process. If you enable this feature, after the node that is set as the anchor in the R&D process is run, DMS automatically turns the ticket to the Finished state. **Note** To use this feature, you must set the last node in the R&D process as the anchor.

Supported checkpoints

> **Note** Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

The Schema Design tab contains the following two processes:

- Process of saving changes: DMS provides the following three checkpoints for this process. The checkpoints validate the table headings, fields, and indexes.
  - **Save Changes and Validate Header**
  - **Save Changes and Validate Field**
  - **Save Changes and Validate Index**
- Process of applying changes: DMS provides the following five checkpoints for this process. The first four checkpoints identify the risks that arise from changing schemas without locking tables, and the last checkpoint assigns an approval process to each type of risk.
  - **Table Creation Risk Control**
  - **Field Change Risk Control**
  - **Index Change Risk Control**
  - **SQL Execution Risk Control**
  - **Approval Rule Validation**

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`. The following table describes the supported factors on the Schema Design tab.

Factor	Description
@fac.table_kind	The type of the table whose schema is to be changed. Valid values: • *new*: a newly created table. • *old*: an existing table.
@fac.column_kind	The type of the field to be changed. Valid values: • *new*: a newly created field. • *old*: an existing field.
@fac.xxxx_old	The value of an existing field or index that is used for comparison.
@fac.column_is_primary	A Boolean value that indicates whether the current field serves as a primary key. Valid values: *true* and *false*.
@fac.column_type_support_default	A Boolean value that indicates whether the data type of the current field supports a default value. Valid values: *true* and *false*. **Note** For example, a field of the CHAR type supports a default value, whereas a field of the TEXT type does not.
@fac.index_kind	The type of the index to be changed. Valid values: • *new*: a newly created index. • *old*: an existing index.
@fac.index_column_count	The number of fields in the index.

Factor	Description
@fac.change_type	The type of the schema change to be performed by DDL statements. Valid values: • *add*: adds one or more fields or indexes. • *modify*: modifies one or more fields or indexes. • *delete*: deletes one or more fields or indexes.
@fac.altered_table_size	The size of the table whose schema is to be changed. Unit: MB.
@fac.online_execute	A Boolean value that indicates whether the schema change can be performed in an online environment. Valid values: *true* and *false*.
@fac.change_risk_level	The risk level of the schema change. Valid values: • *high*: a high risk level. • *middle*: a medium risk level. • *low*: a low risk level.
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions on the Schema Design tab.

Action	Description	Format
@act.block_submit	Blocks the submission of the schema change and displays the error message. This action can be used in the process of saving changes.	@act.block_submit 'Reason for blocking the submission'
@act.show_warning	Displays the error message without blocking the submission of the schema change. This action can be used in the process of saving changes.	@act.show_warning 'Error message'
@act.mark_middle_risk	Specifies that the schema change is at medium risk. This action can be used in the process of identifying the risk level.	@act.mark_middle_risk 'Reason for the identification'
@act.mark_high_risk	Specifies that the schema change is at high risk. This action can be used in the process of identifying the risk level.	@act.mark_high_risk 'Reason for the identification'
@act.forbid_submit_publish	Rejects the ticket. This action can be used in the process of setting the approval process.	@act.forbid_submit_publish 'Reason for the rejection'
@act.do_not_approve		
@act.choose_approve_template	Specifies the ID of an approval	

Action	Description	N/A Format
@act.choose_approve_template_with_reason	Specifies the ID of an approval template.	

Parameters involved in the R&D process

Parameter	Description
Step	• The type of the node. Valid values: • **Design**: The design node in the R&D process is generated by default and cannot be removed. It determines the environment where the schema change is designed. • **Publish**: A publish node in the R&D process is used to publish the schema change after the change is designed. You can set multiple publish nodes.
Node Name	The name of the node. The node name can be up to 10 characters in length.
Database Environment	The environment where the node is run.
Execution Strategy	• The way in which the node is run. Valid values: • **Immediately**: The node is run immediately after it is approved. • **Periodically**: The node is run at the time that you specify. **Note** If a node is approved before the specified point in time, it is run as scheduled. Otherwise, the node is interrupted and not run.
Can Go Back	Specifies whether a publish node can be rolled back to the design node.
Can Skip	Specifies whether the current node can be skipped.
Anchor	The point that is used to stop the schema change process. If you set a node as the anchor, after the node is published, the nodes that follow the anchor cannot be run and the schema change process ends. At this time, the ticket enters the Published state.
Actions	The operation that you can perform on a publish node. You can remove a publish node as required.

16.1.11.9.10. Database and table synchronization

DMS allows you to configure security rules on the Table Sync tab to validate operations that are related to schema synchronization, empty database initialization, and table consistency repair.

Basic configuration items

Configuration item	Description
Enable execution capability	Specifies whether to enable SQL-based synchronization. If this configuration item is set to OFF, applicants can compare table schemas but cannot execute SQL statements to synchronize databases and tables. Other configuration items and security rules you set under checkpoints on the Table Sync tab also become invalid.
Database table synchronization default approval Template	The default approval template for database and table synchronization applications. You can use the default approval template or click Switch Approval Template and select another template. For more information, see Customize approval processes.

Configuration item	Description
Analysis phase script Expiration Time (unit: hours)	The timeout period of the analysis phase. You can set an appropriate timeout period in which synchronization can be canceled if schemas are changed in the destination database.

Supported checkpoints

The Table Sync tab contains three checkpoints that are corresponding to the three features that are supported by the tab. The three checkpoints are unrelated to each other. For example, when you submit a Schema Synchronization ticket, only the basic configuration items and the security rules that are specified under the Schema Synchronization Validation checkpoint are used to validate the ticket.

> **Note** Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Checkpoint	Description
Schema Synchronization Validation	Allows you to set approval processes or constraints for Schema Synchronization tickets.
Empty Database Initialization Validation	Allows you to set approval processes or constraints for Empty Database Initialization tickets.
Table Consistency Repair Validation	Allows you to set approval processes or constraints for Repair Table Consistency tickets.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`. The following table describes the supported factors on the Table Sync tab.

Factor	Description
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.
@fac.schema_name	The name of the schema.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`. The following table describes the supported actions on the Table Sync tab.

Action	Description
@act.forbid_submit_order	Forbids a ticket from being submitted. The statement is in the following format: `@act.forbid_submit_order 'Reason for forbidding the submission of the ticket'`.
@act.do_not_approve	
@act.choose_approve_template	Specifies the ID of an approval template.

Action	Description
@act.choose_approve_template_with_reason	

16.1.11.9.11. Sensitive field change

The topic describes the security rules on the Sensitive Column Change tab.

Basic configuration items

Sensitive column default approval Template: the default approval template that takes effect if you do not set approval processes for tickets that apply to change the security levels of sensitive fields under the Approval Rule Validation checkpoint. In the Switch Approval Template dialog box, you can change the approval process of the default approval template.

Supported checkpoints

Approval Rule Validation: When a user submits a ticket to change the security level of a sensitive field, DMS checks whether the ticket conforms to the rules that are specified under the Approval Rule Validation checkpoint.

> **Note** Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`.The following table describes the supported factors on the Sensitive Column Change tab.

Factor	Description

@fac.column_level_change_type	The type of security level change that the applicant wants to perform on a sensitive field. Valid values:
	• *upper*: raises the current security level, including the following three cases:
	  ◦ From inner to sensitive
	  ◦ From inner to confidential
	  ◦ From sensitive to confidential
	• *sensitive_to_inner*: lowers the security level from sensitive to inner.
	• *confidential_to_sensitive*: lowers the security level from confidential to sensitive.
	• *confidential_to_inner*: lowers the security level from confidential to inner.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions on the Sensitive Column Change tab.

Action	Description

Action	Description
@act.forbid_submit_order	Forbids a ticket from being submitted. The statement is in the following format: `@act.forbid_submit_order 'Reason for forbidding the submission of the ticket'`.
@act.do_not_approve	Specifies the ID of an approval template.
@act.choose_approve_template	
@act.choose_approve_template_with_reason	

16.1.11.9.12. Test data generation

This topic describes the security rules on the Test Data Generate tab.

Supported checkpoints

Approval rule validation: When a user submits a ticket to generate test data, DMS checks whether the ticket conforms to the rules that are specified under the Approval rule validation checkpoint.

> **Note** Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Supported factors

A factor is a built-in variable in DMS. You can use factors to obtain the context to be validated by security rules, such as SQL statement categories and the number of rows to be affected. A factor name starts with the prefix `@fac.`.The following table describes the supported factors on the Test Data Generate tab.

Factor	Description
@fac.env_type	The type of the environment. The value is the display name of the environment type, such as DEV or PRODUCT.
@fac.schema_name	The name of the schema.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions on the Test Data Generate tab.

Action	Description
@act.forbid_submit_order	Forbids a ticket from being submitted. The statement is in the following format: `@act.forbid_submit_order 'Reason for forbidding the submission of the ticket'`.
@act.do_not_approve	Specifies the ID of an approval template.
@act.choose_approve_template	
@act.choose_approve_template_with_reason	

16.1.11.9.13. Database cloning

This topic describes the security rules on the Database Clone tab.

Basic configuration items

Database clone default approval Template: the default approval template that takes effect if you do not set approval processes for database clone tickets under the Approval rule validation checkpoint. In the Switch Approval Template dialog box, you can change the approval process of the default approval template.

Supported checkpoints

Approval rule validation: When a user submits a database clone ticket, DMS checks whether the ticket conforms to the rules that are specified under the Approval rule validation checkpoint.

> **Note** Various security rule templates are built in each checkpoint. You can configure security rules based on these templates or create custom security rules as needed.

Supported actions

An action in a security rule is an operation that DMS performs when the if condition in the rule is met. For example, DMS can forbid the submission of a ticket, approve a ticket, or reject a ticket. An action in a security rule denotes the purpose of the security rule. An action name starts with the prefix `@act.`.The following table describes the supported actions on the Database Clone tab.

Action	Description
@act.forbid_submit_order	Forbids a ticket from being submitted. The statement is in the following format: `@act.forbid_submit_order 'Reason for forbidding the submission of the ticket'`.
@act.do_not_approve	Specifies the ID of an approval template.
@act.choose_approve_template	
@act.choose_approve_template_with_reason	

17.Server Load Balancer (SLB)

17.1. User Guide

17.1.1. What is SLB?

This topic provides an overview of Server Load Balancer (SLB). SLB distributes inbound network traffic across multiple Elastic Compute Service (ECS) instances that act as backend servers based on forwarding rules. You can use SLB to improve the responsiveness and availability of your applications.

Overview

After you add ECS instances that are deployed in the same region to a SLB instance, SLB uses virtual IP addresses (VIPs) to virtualize these ECS instances into backend servers in a high-performance server pool that ensures high availability. Client requests are distributed to the ECS instances based on forwarding rules.

SLB checks the health status of the ECS instances and automatically removes unhealthy ones from the server pool to eliminate single points of failure (SPOFs). This enhances the resilience of your applications.

Components

SLB consists of three components:

- SLB instances

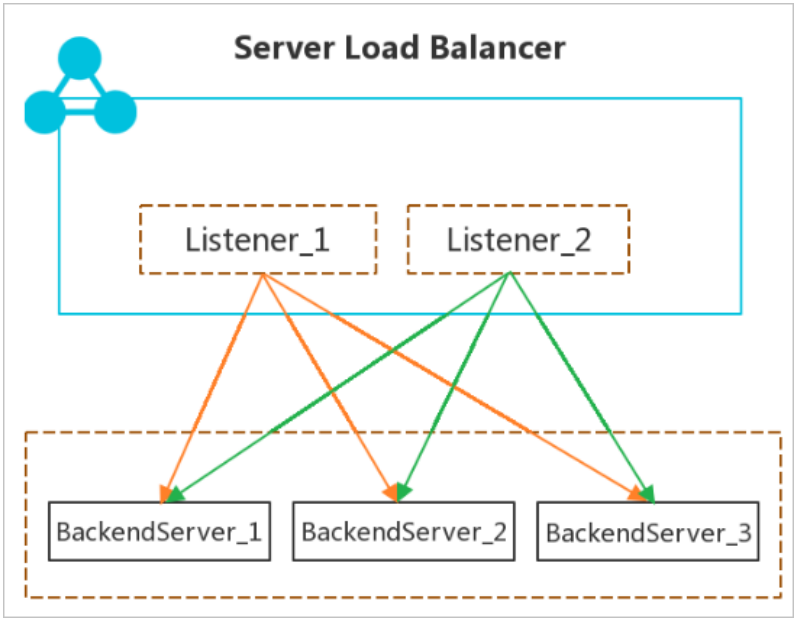
  A SLB instance is a running SLB service entity that receives traffic and distributes traffic to backend servers. To get started with SLB, you must create a SLB instance and add at least one listener and two ECS instances to the SLB instance.

- Listeners

  A listener checks client requests and forwards them to backend servers. It also performs health checks on backend servers.

- Backend servers

  ECS instances are used as backend servers to receive distributed requests. You can separately add ECS instances to the server pool, or use vServer groups or primary/secondary server groups to add and manage ECS instances in batches.



Benefits

- High availability

  SLB is designed with full redundancy that avoids SPOFs and supports zone-disaster recovery.

  SLB can be scaled based on application loads and can provide continuous service during traffic fluctuations.

- High scalability

  You can increase or decrease the number of backend servers to adjust the load balancing capability of your applications.

- Cost-effectiveness

  SLB can save 60% of load balancing costs compared with using traditional hardware solutions.

- Security

  You can use SLB with Apsara Stack Security to defend your applications against up to 5 Gbit/s DDoS attacks.

- High concurrency

  A SLB cluster supports hundreds of millions of concurrent connections and a single SLB instance supports tens of millions of concurrent connections.

17.1.2. Log on to the SLB console

This topic describes how to go to the Server Load Balancer (SLB) console after you log on to the Apsara Uni-manager Management Console by using the Chrome browser.

Prerequisites

- The domain name of the Apsara Uni-manager Management Console is obtained from the engineer that deploys the service before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.

- You have enabled MFA and bound an MFA device.

  Enter a six-digit MFA authentication code and click **Authenticate**.

  > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Networking > Server Load Balancer**.

17.1.3. Quick start

17.1.3.1. Overview

This topic describes how to create an Internet-facing Classic Load Balancer (CLB) instance.

> **Note** Before you create a CLB instance, you must plan the configuration, which includes the region, type, and billing mode of the CLB instance.

This topic includes the following operations:

1. Create a CLB instance

   A CLB instance is a running entity that provides the CLB service.

2. Configure a CLB instance

   After you create a CLB instance, you must add listeners and backend servers.

3. Release an SLB instance

   If a CLB instance is no longer needed, delete the instance to avoid unexpected fees.

17.1.3.2. Make preparations

This topic describes the considerations for configuring a Classic Load Balancer (CLB) instance. Before you create a CLB instance, you must determine the listener type and network type of the CLB instance.

Select the region of the CLB instance

When you select a region, take note of the following issues:

- To reduce the network latency and increase the downloading speed, we recommend that you select the region that is closest to your users.
- To provide more stable and reliable load balancing services, CLB supports primary/secondary zone deployment in most regions. This implements disaster recovery across data centers in the same region. We recommend that you select a region that supports primary/secondary zone deployment.
- CLB does not support cross-region deployment. Therefore, you must deploy the CLB instance and backend Elastic Compute Service (ECS) instances in the same region.

Select the network type of the CLB instance

CLB provides load balancing services to the Internet and private networks:

- If you want to use CLB to distribute requests from the Internet, create an Internet-facing CLB instance.

  An Internet-facing CLB instance is assigned a public IP address to receive requests from the Internet.

- If you want to use CLB to distribute requests from private networks, create an internal-facing CLB instance.

  An internal-facing CLB instance is only assigned a private IP address of Alibaba Cloud. You must access the CLB instance through networks of Alibaba Cloud.

Select a listener protocol

CLB supports Layer 4 (TCP and UDP) and Layer 7 (HTTP and HTTPS) load balancing:

- A Layer 4 listener directly distributes requests to backend servers without modifying packet headers. After a client request reaches a Layer 4 listener of CLB, CLB uses the backend port that is configured for the listener to establish a TCP connection with a backend ECS instance.
- A Layer 7 listener functions as a reverse proxy. After a client request reaches a Layer 7 listener of CLB, CLB establishes a new TCP connection over HTTP with a backend server, instead of directly forwarding the request to the backend server.

  Compared with Layer 4 listeners, Layer 7 listeners require an additional step of Tengine processing. In addition, the performance of Layer 7 listeners can also be affected by factors such as insufficient client ports or excessive backend server connections. Therefore, we recommend that you use Layer 4 listeners for high-performance load-balancing services.

Create backend servers

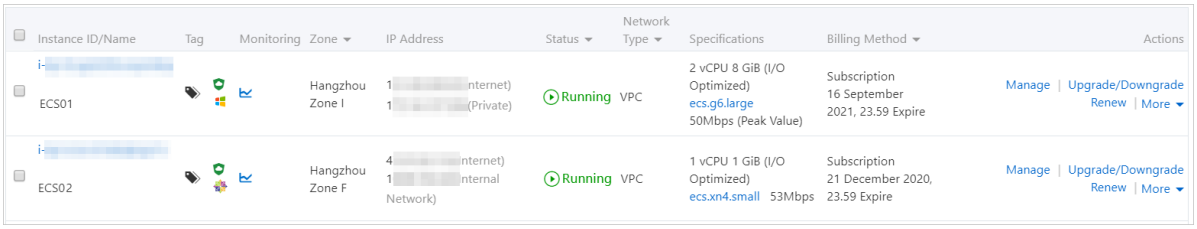
Before you use the CLB service, you must create ECS instances, deploy applications on the ECS instances, and then add the ECS instances to your CLB instance to process client requests.

When you create and configure ECS instances, take note of the following issues:

- Region and zones of the ECS instances

  Make sure that the ECS instances and the CLB instance are deployed in the same region. In addition, make sure that the CLB instance and the ECS instances belong to the same VPC. We recommend that you deploy the ECS instances in different zones to improve the availability.

  In this example, two ECS instances named ECS01 and ECS02 are created in the **China (Hangzhou)** region. The following figure shows their basic configurations.



- Configurations

  In this example, two static web pages are hosted on ECS01 and ECS02 by using Apache.

  - Enter the elastic IP address (EIP) that is associated with ECS01 in the address bar of your browser.
  - Enter the EIP that is associated with ECS02 in the address bar of your browser.

  No additional configurations are required after you deploy applications on the ECS instances. However, if you want to use a Layer 4 (TCP or UDP) listener and the ECS instances run Linux, make sure that the following parameters in the *net.ipv4.conf* file under */etc/sysctl.conf* are set to 0:

```
net.ipv4.conf.default.rp_filter = 0
net.ipv4.conf.all.rp_filter = 0
net.ipv4.conf.eth0.rp_filter = 0
```

17.1.3.3. Create an SLB instance

This topic describes how to create a Server Load Balancer (SLB) instance. An SLB instance is a running entity of the SLB service. You can add multiple listeners and backend servers to an SLB instance.

Prerequisites

- Elastic Compute Service (ECS) instances are created and applications are deployed on the ECS instances.

- The ECS instances and the SLB instance belong to the same organization. In addition, the security group rules of the ECS instances allow access from port 80 (HTTP) and port 443 (HTTPS).

Procedure

1. Log on to the SLB console.
2. In the left-side navigation pane, choose **Instances > Instances**.
3. On the **Instances** page, click **Create Instance**.
   - **Organization**: Select an organization for the SLB instance from the drop-down list.

     > **Note** Make sure that the organization of the SLB instance is the same as the organization of its backend servers.

   - **Resource Set**: Select a resource set for the SLB instance from the drop-down list.
   - **Region**: Select the region where you want to deploy the SLB instance.
   - **Zone**: Select a zone for the SLB instance from the drop-down list.
   - **Instance Name**: Enter a name for the SLB instance in the Instance Name field.

     The name must be 2 to 128 characters in length, and can contain letters, digits, full-width characters, hyphens (-), colons (:), periods (.), and underscores (_). Line breaks and spaces are supported. It must start with a letter and cannot start with `http://` or `https://`.
   - **Instance Edition**: Select one of the following options: shared-performance and guaranteed-performance. Shared-performance SLB instances share resources with each other. The performance of shared-performance SLB instances is not guaranteed. The performance of guaranteed-performance SLB instances varies by type.
   - **Instance Type**: Select the type of network traffic that you want to distribute. Valid values: Internal Network and Internet.
   - **Network Type**: Select the network type of the SLB instance. Valid values: Classic Network and VPC.
   - **IP Version**: Select an IP version.
   - **IP Address**: Enter a service IP address for the SLB instance. Make sure that the service IP address is valid. Otherwise, the SLB instance cannot be created. If you do not set this parameter, the system automatically allocates an IP address to the SLB instance.

     > **Note** The private IP address that you specify must belong to the destination CIDR block of the vSwitch.

4. Click **Submit**.

What's next

Configure a CLB instance

17.1.3.4. Configure a CLB instance

This topic describes how to configure a instance. Before you can use a CLB instance to forward traffic, you must add at least one listener and one group of backend servers to the CLB instance. The following example describes how to add a TCP listener and two Elastic Compute Service (ECS) instances to a CLB instance. The ECS instances are named ECS 01 and ECS 02. The ECS instances function as backend servers and host static web pages.

Procedure

1.
2. On the **Instances** page, find the CLB instance and click **Configure Listener** in the **Actions** column.
3. In the **Protocol and Listener** wizard, set the following parameters and click **Next**.

Parameter	Description
Select Listener Protocol	Select the protocol of the listener. In this example, **TCP** is selected.
Listening Port	Specify the port that the CLB instance uses to receive requests and forward the requests to backend servers. In this example, the port number is set to **80**. **Note** The ports on which a CLB instance listens must be unique.
Listener Name	Enter a name for the listener.
Advanced	Click **Modify** to configure advanced settings. In this topic, the default values are used. For more information, see Add a TCP listener.

4. In the **Backend Servers** wizard, select a backend server type.

   In this topic, **Default Server Group** is selected. Click **Add More**.

   i. In the **My Servers** panel, select ECS 01 and ECS 02 that you created, and click **Next**.

   ii. Set weights for the servers and click **Add**.

   A backend server that has a higher weight receives more requests. In this topic, the default values are used.

   iii. On the **Default Server Group** tab, specify backend ports that are available to receive requests, and click **Next**.

   You can specify the same port for multiple backend servers of a CLB instance. In this example, the port number is set to 80.

5. In the **Health Check** wizard, configure health checks and click **Next**. The default values are used in this topic.

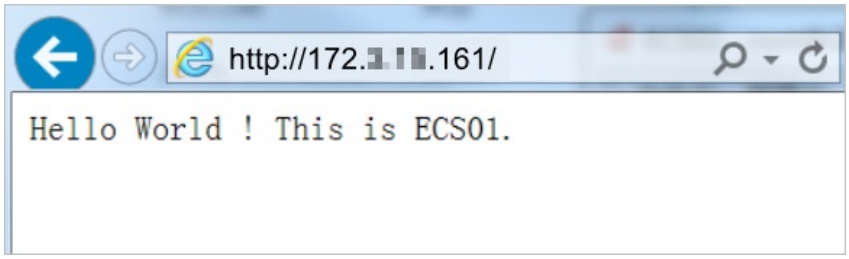
   After you enable health checks for the CLB instance, the CLB instance periodically checks whether the backend ECS instances are healthy. When the CLB instance detects an unhealthy ECS instance, the CLB instance distributes the requests to other healthy ECS instances. When the unhealthy ECS instance recovers, the CLB instance starts to distribute requests to the ECS instance again.

6. In the **Configuration Review** wizard, check the configuration and click **Submit**.

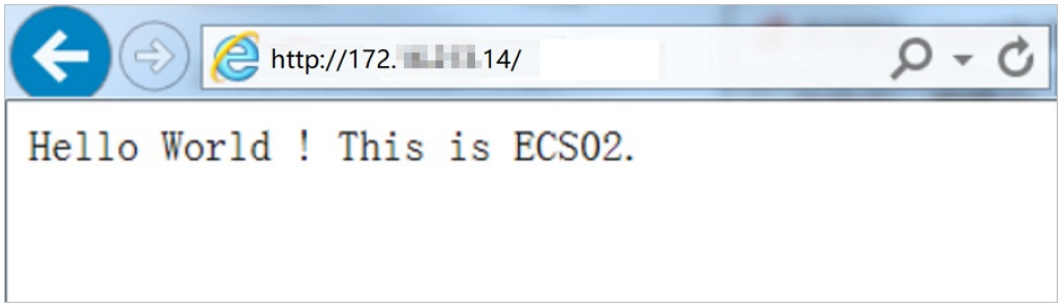
7. Click **OK** to go back to the **Instances** page. Then, click the ↻ icon to refresh the page.

8. Enter the service address of the CLB instance into the address bar of the browser to check the connectivity. We recommend that you run the test multiple times.

   ECS01

   ECS02



17.1.3.5. Release an SLB instance

If an SLB instance is no longer needed, you can release the instance to save costs. The backend ECS instances will not be deleted or affected after you delete an SLB instance.

Procedure

1. Log on to the SLB console.
2. On the **Instances** page, find the instance and click > **Release** in the Actions column, or select the instance and click **Release** at the lower part of the page.
3. In the **Release** dialog box, select **Release Now**.

   > **Note** The system performs release operations at 30-minute and hour marks. However, billing for the SLB instance is stopped at the specified release time.

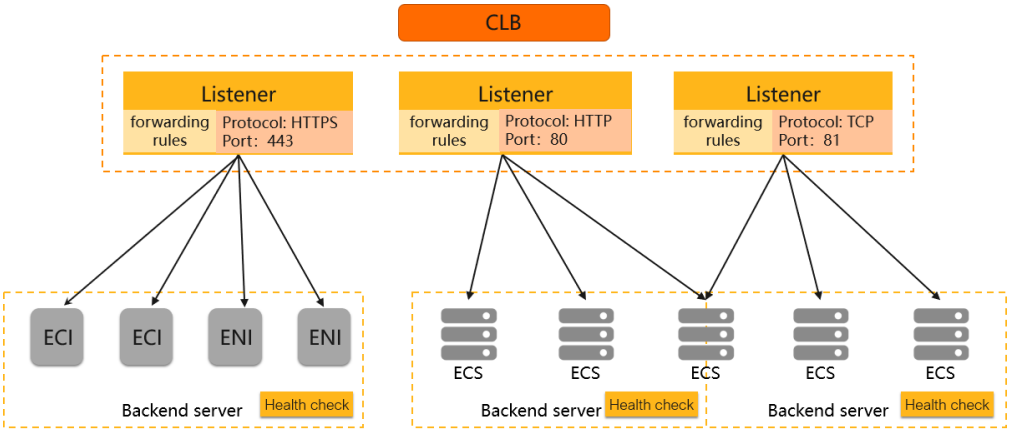
4. Click **Next**.
5. Click **OK** to release the SLB instance.

   > **Note** Pay-as-you-go SLB instances cannot be restored once deleted. We recommend that you exercise caution when you release SLB instances.

17.1.4. SLB instances

17.1.4.1. Overview

Classic Load Balancer (CLB) instances receive requests from clients and forward requests to backend servers. To use the load balancing service, you must create a CLB instance and add listeners and backend servers to the instance.



Instance types

Alibaba Cloud provides Internet-facing and internal-facing CLB instances.

Internet-facing CLB instances

When you create an Internet-facing CLB instance, the system automatically assigns a public address to the CLB instance. You can associate your domain name with the public IP address. CLB instances receive requests from clients over the Internet and forward requests to backend servers based on the forwarding rules that you specify for listeners.

An Internet-facing CLB instance has the following features:

- The system assigns a public IP address to the CLB instance. You cannot disassociate the public IP address from the CLB instance.
- When the subscription billing method is used, only the pay-by-bandwidth metering method is supported. When the pay-as-you-go billing method is used, the pay-by-bandwidth and pay-by-data-transfer metering methods are supported.

Internal-facing CLB instances

Internal-facing CLB instances provide services over private networks. Internal-facing CLB instances receive requests from internal networks and forward requests to backend servers based on listener rules.

Internal-facing CLB instances that are associated with elastic IP addresses (EIPs) can process requests from the Internet. An internal-facing CLB instance has the following features when it provides services over the Internet:

- An internal-facing CLB instance associated with an EIP can provide services over the Internet. You can disassociate the EIP from the CLB instance as needed.
- An EIP that is associated with an EIP bandwidth plan supports the pay-by-95th-percentile-bandwidth billing method in addition to the subscription and pay-as-you-go billing methods.

The network types supported by internal-facing CLB instances vary based on billing methods.

- Subscription internal-facing CLB instances support the following network types: classic network and virtual private cloud (VPC).
  - VPC

    If an internal-facing CLB instance is deployed in a VPC, the IP address of the CLB instance is allocated from the CIDR block of the vSwitch that is attached to the VPC. The internal-facing CLB instance can be accessed by only Elastic Compute Service (ECS) instances in the same VPC.
  - Classic network

    If the internal-facing CLB instance is deployed in a classic network, the private IP address of the CLB instance is allocated and managed by Alibaba Cloud. The internal-facing CLB instance can be accessed by only ECS instances in classic networks.

    > **Notice** Internal-facing CLB instances of the classic network type are no longer available.

- Pay-as-you-go internal-facing CLB instances support only the VPC network type.

  Internal-facing CLB instances support PrivateLink. Internal-facing CLB instances can receive requests from other VPCs through PrivateLink connections and distribute the requests to backend servers based on listener rules.

Instance types and specifications

> **Note** If you require a higher QPS, you can purchase Application Load Balancer (ALB) instances.

Resources are shared among all shared-resource CLB instances. Therefore, the performance of shared-resource CLB instances is not guaranteed.

Notice Shared-resource CLB instances are no longer available for purchase.

Feature	High-performance CLB instance	Shared-resource CLB instance
Resource allocation	Exclusive resources	Shared resources
Service uptime guaranteed by the service-level agreement (SLA)	99.95%	Not supported
IPv6	√	-
Server Name Indication (SNI)	√	-
Blacklists and whitelists	√	-
Elastic network interface (ENI) mounting	√	-
Assigning secondary IP addresses to ENIs that are associated with ECS instances	√	-
HTTP-to-HTTPS redirection	√	-
Consistent hashing	√	-
TLS security policies	√	-
HTTP2	√	-
WebSocket or WebSocket Secure	√	-

Note In the preceding table, "√" indicates that a feature is supported, and "-" indicates that a feature is not supported.

17.1.4.2. Create an SLB instance

This topic describes how to create a Server Load Balancer (SLB) instance. An SLB instance is a running entity of the SLB service. You can add multiple listeners and backend servers to an SLB instance.

Prerequisites

- Elastic Compute Service (ECS) instances are created and applications are deployed on the ECS instances.
- The ECS instances and the SLB instance belong to the same organization. In addition, the security group rules of the ECS instances allow access from port 80 (HTTP) and port 443 (HTTPS).

Procedure

1. Log on to the SLB console.
2. In the left-side navigation pane, choose **Instances > Instances**.
3. On the **Instances** page, click **Create Instance**.
   - **Organization**: Select an organization for the SLB instance from the drop-down list.

     Note Make sure that the organization of the SLB instance is the same as the organization of its backend servers.

- **Resource Set**: Select a resource set for the SLB instance from the drop-down list.
- **Region**: Select the region where you want to deploy the SLB instance.
- **Zone**: Select a zone for the SLB instance from the drop-down list.
- **Instance Name**: Enter a name for the SLB instance in the Instance Name field.

  The name must be 2 to 128 characters in length, and can contain letters, digits, full-width characters, hyphens (-), colons (:), periods (.), and underscores (_). Line breaks and spaces are supported. It must start with a letter and cannot start with `http://` or `https://`.
- **Instance Edition**: Select one of the following options: shared-performance and guaranteed-performance. Shared-performance SLB instances share resources with each other. The performance of shared-performance SLB instances is not guaranteed. The performance of guaranteed-performance SLB instances varies by type.
- **Instance Type**: Select the type of network traffic that you want to distribute. Valid values: Internal Network and Internet.
- **Network Type**: Select the network type of the SLB instance. Valid values: Classic Network and VPC.
- **IP Version**: Select an IP version.
- **IP Address**: Enter a service IP address for the SLB instance. Make sure that the service IP address is valid. Otherwise, the SLB instance cannot be created. If you do not set this parameter, the system automatically allocates an IP address to the SLB instance.

  > **Note** The private IP address that you specify must belong to the destination CIDR block of the vSwitch.

4. Click **Submit**.

What's next

Configure a CLB instance

17.1.4.3. Start or stop an instance

This topic describes how to start or stop a Classic Load Balancer (CLB) instance. A stopped CLB instance does not receive or forward client requests.

Procedure

1. Log on to the SLB console.
2. In the top navigation bar, select the region where the CLB instance is deployed.
3. On the **Instances** page, find the CLB instance and choose ⋮ > **Start** or ⋮ > **Stop** in the **Actions** column.
4. To start or stop multiple CLB instances at a time, select the CLB instances and click **Start** or **Stop** at the lower part of the page.

Related information

References

- SetLoadBalancerStatus: sets the status of a CLB instance.

17.1.4.4. Tags

17.1.4.4.1. Tag overview

This topic provides an overview of tags in SLB. SLB provides the tag management feature that allows you to classify SLB instances by using tags.

Each tag consists of a key and a value. Before you use tags, note the following limits:

- Tags must be added to SLB instances.
- Each SLB instance can have a maximum of ten tags. You can add or remove a maximum of 5 tags at a time.
- The key of each tag added to an SLB instance must be unique. If a tag with the same key already exists, the tag is overwritten with the new value.

17.1.4.4.2. Add tags

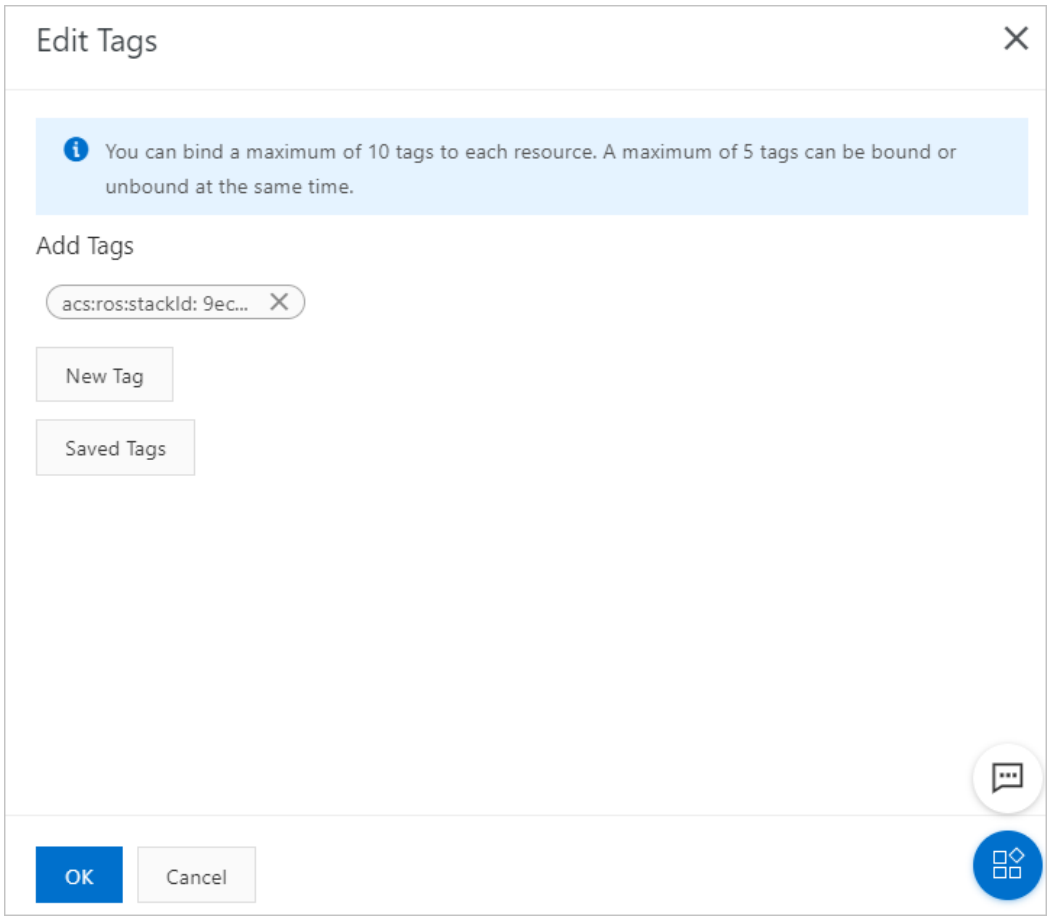
This topic describes how to add tags to a CLB instance.

Procedure

1. Log on to the SLB console
2. In the left-side navigation pane, choose **Instances > Server Load Balancers**.
3. In the **Actions** column, choose ⋮ **> Edit Tags**.
4. Edit tags in the **Edit Tags** dialog box.

   To add a tag, perform the following operations:

   - To add an existing tag, click **Saved Tags** and then select a tag.
   - To create and add a new tag, click **New Tag** in the **Edit Tags** dialog box, enter the key and value of the new tag, and then click **OK** next to the value.



5. Click **OK**.

17.1.4.4.3. Query CLB instances by tag

This topic describes how to use tags to query CLB instances.

Procedure

1. Log on to the SLB console.
2. In the left-side navigation pane, choose **Instances > Server Load Balancers**.
3. Click **Select a tag** and select a tag to filter instances.

> **Note** To clear the search condition, move the pointer over the selected tag and click the displayed deletion icon next to it.

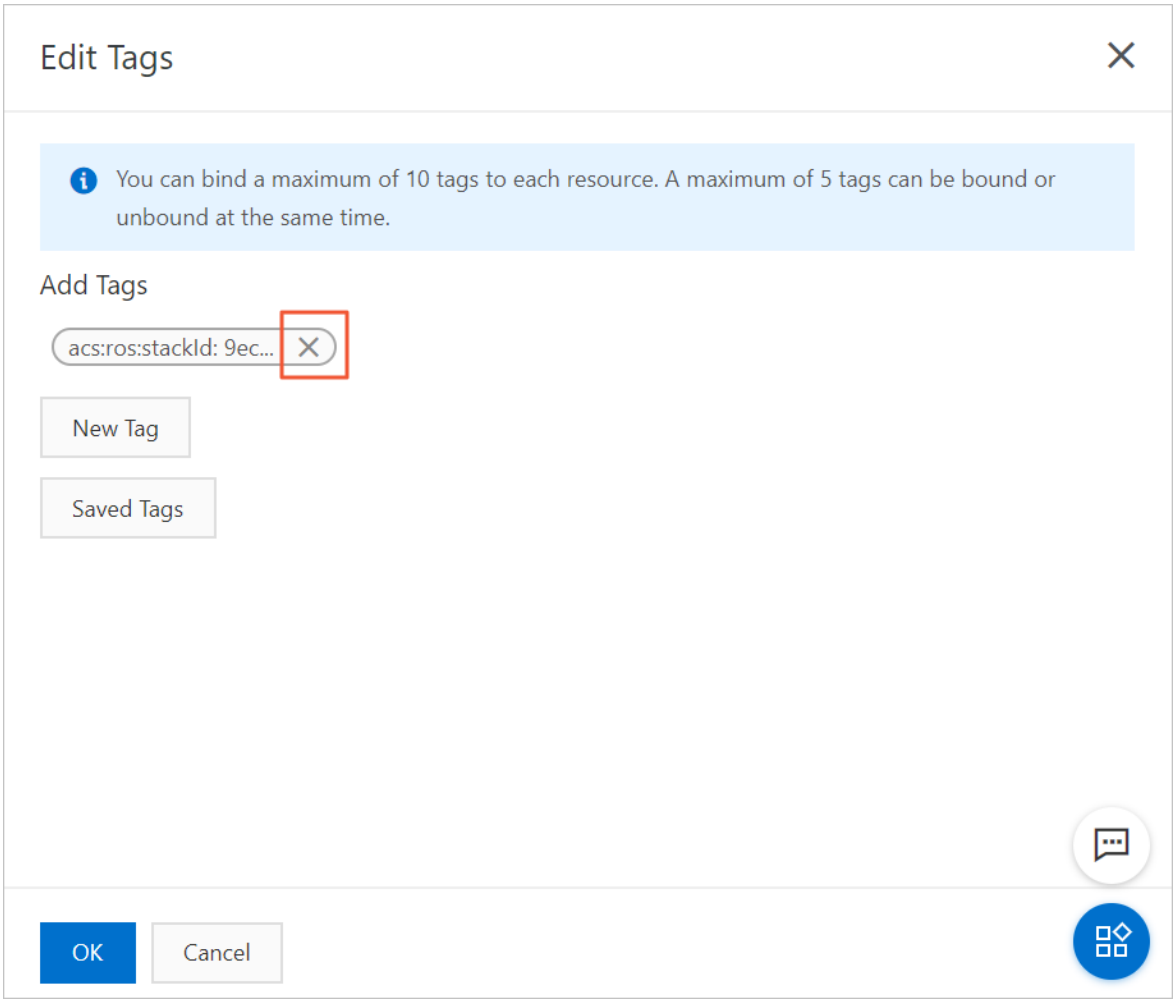
17.1.4.4.4. Remove a tag

This topic describes how to remove tags from a CLB instance. You can only remove tags for one CLB instance at a time.

Procedure

1. Log on to the SLB console.
2. In the left-side navigation pane, choose **Instances > Server Load Balancers**.
3. In the **Actions** column, choose ⋮ **> Edit Tags**.
4. In the **Edit Tags** dialog box, click the deletion icon next to the tags to be removed, and then click **OK**.

> **Note** If a tag is removed from a CLB instance and is not added to other instances, the tag is deleted from the system.



17.1.4.5. Release an SLB instance

This topic describes how to release a Server Load Balancer (SLB) instance. You can immediately release SLB instances .

Procedure

1. Log on to the SLB console.
2. In the left-side navigation pane, choose **Instances > Instances**.
3. Find the SLB instance that you want to release and choose ⋮ **> Release** in the Actions column.
4. In the **Release** panel, **click Release Now**.
5. Click **Next**.
6. Confirm the displayed information and click **OK** to release the instance.

17.1.5. Listeners

17.1.5.1. Listener overview

This topic provides an overview of listeners. After you create a Classic Load Balancer (CLB) instance, you must configure one or more listeners for it. A listener checks for connection requests and then distributes the requests to backend servers based on the forwarding rules that are defined by a specified scheduling algorithm.

CLB provides Layer 4 (TCP or UDP) and Layer 7 (HTTP or HTTPS) listeners. The following table lists the features and use cases of these listeners.

Protocol	Description	Scenario
TCP	• A connection-oriented protocol. A reliable connection must be established before data can be sent and received. • Session persistence is based on source IP addresses. • Source IP addresses are visible at the network layer. • Data is transmitted at a fast rate.	• Applicable to scenarios that require high reliability and data accuracy but can tolerate low speeds, such as file transmission, sending or receiving emails, and remote logons. • Web applications that do not have custom requirements. For more information, see Add a TCP listener.
UDP	• A connectionless protocol. UDP transmits data packets directly instead of making a three-way handshake with the other party before UDP sends data. UDP does not provide error recovery or data re-transmission. • Fast data transmission but relatively low reliability.	Applicable to scenarios where real-time transmission is more important than reliability, such as video chats and real-time financial market pushes. For more information, see Add a UDP listener.
HTTP	• An application-layer protocol that is used to package data. • Cookie-based session persistence. • Use the X-Forward-For header to obtain the real IP addresses of clients.	Applicable to scenarios that require data content identification, such as web applications and small mobile games. For more information, see Add an HTTP listener.
HTTPS	• Encrypted data transmission that prevents unauthorized access. • Centralized certificate management service. You can upload certificates to CLB. The decryption operations are completed directly on CLB.	Applicable to scenarios that require encrypted transmission. For more information, see Add an HTTPS listener.

17.1.5.2. Add a TCP listener

This topic describes how to add a TCP listener to a Classic Load Balancer (CLB) instance. TCP allows you to transmit data in a reliable and accurate manner but at relatively low speeds. Therefore, you can use TCP to transfer files, send or receive emails, and perform remote logons. You can add a TCP listener to forward TCP requests.

Step 1: Configure a TCP listener

1.
2. Use one of the following methods to open the listener configuration wizard:
   - On the **Instances** page, find the CLB instance and click **Configure Listener** in the **Actions** column.
   - On the **Instances** page, find the CLB instance that you want to manage and click the ID of the instance. On the **Listener** tab, click **Add Listener**.
3. Set the following parameters and click **Next**.

Parameter	Description
Select Listener Protocol	Select **TCP**.
Listening Port	Set the listening port that is used to receive requests and forward them to backend servers. Valid values: 1 to 65535. You can set a TCP or UDP listener to listen on all ports within a specified port range.
Listener Name	Specify a name for the listener.
Advanced	Click **Modify** to configure advanced settings.
Scheduling Algorithm	Select a scheduling algorithm. ◦ **Weighted Round-Robin (WRR)**: Backend servers that have higher weights receive more requests than backend servers that have lower weights. ◦ **Round-Robin (RR)**: Requests are distributed to backend servers in sequence. ◦ **Consistent Hash (CH)**: ▪ **Tuple**: specifies consistent hashing that is based on four factors: source IP address, destination IP address, source port, and destination port. Requests that contain the same information based on the four factors are distributed to the same backend server. ▪ **Source IP**: specifies consistent hashing that is based on source IP addresses. Requests from the same source IP address are distributed to the same backend server. **Note** Only high-performance CLB instances support the CH algorithm.
Enable Session Persistence	Specify whether to enable session persistence. After session persistence is enabled, CLB forwards all requests from a client to the same backend server. For TCP listeners, session persistence is implemented based on IP addresses. Requests from the same IP address are forwarded to the same backend server.
Enable Peak Bandwidth Limit	Specify whether to set the bandwidth limit of the listener. If a CLB instance is billed based on bandwidth usage, you can set different maximum bandwidth values for different listeners. This limits the amount of traffic that flows through each listener. The sum of the maximum bandwidth values of all listeners that are added to a CLB instance cannot exceed the maximum bandwidth value of the CLB instance. By default, this feature is disabled and all listeners share the bandwidth of the CLB instance.
Idle Timeout	Specify the timeout period of idle TCP connections. Unit: seconds. Valid values: 10 to 900.
Proxy Protocol	Use the proxy protocol to pass client IP addresses to backend servers. **Note** You cannot enable this feature in scenarios where PrivateLink is used.
Obtain Client Source IP Address	Specify whether to retrieve the real IP addresses of clients. Only Layer 4 listeners support this feature. By default, this feature is enabled.

Parameter	Description
Automatically Enable Listener After Creation	Specify whether to immediately enable the listener after it is created. By default, listeners are enabled after they are created.

Step 2: Add backend servers

After you configure the listener, you must add backend servers to process client requests. You can use the default server group that is configured for the CLB instance. You can also configure a vServer group or a primary/secondary server group, or enable the primary/secondary mode for the listener. For more information, see Backend server overview.

1. On the **Backend Servers** wizard page, select the type of the server group to which requests are forwarded. The default server group is used in this example.

   Select **Default Server Group** and click **Add More**.

2. In the **My Servers** panel, select the ECS instances that you want to add as backend servers and click **Next**.
3. On the **Configure Ports and Weights** wizard page, specify the weights of the backend servers that you added. A backend server with a higher weight receives more requests.

   > **Note** If the weight of a backend server is set to 0, no request is distributed to the backend server.

4. Click **Add**. On the Default Server Group tab, specify the ports that you want to open on the backend servers to receive requests. The backend servers are the ECS instances that you selected. Valid values: 1 to 65535.

   You can specify the same port on different backend servers that are added to a CLB instance.

5. Click **Next**.

Step 3: Configure health checks

CLB performs health checks to check the availability of the ECS instances that serve as backend servers. The health check feature improves overall service availability and reduces the impact of backend server failures.

On the **Health Check** wizard page, click **Modify** to modify the health check configurations. For more information, see Configure health checks.

Step 4: Submit the configurations

1. On the **Confirm** wizard page, check the configurations. You can click **Modify** to modify the configurations.
2. After you confirm the configurations, click **Submit**.
3. When Configuration Successful appears, click **OK**.

   After you configure the listener, you can view the listener on the Listener tab.

17.1.5.3. Add a UDP listener

This topic describes how to add a UDP listener to a Classic Load Balancer (CLB) instance. UDP applies to services that prioritize real-time content delivery over reliability, such as video conferencing and real-time quote services. You can add a UDP listener to forward UDP packets.

Context

Before you configure a UDP listener, take note of the following items:

- You cannot specify ports 250, 4789, or 4790 for UDP listeners. They are system reserved ports.
- Fragmentation is not supported.
- You cannot view source IP addresses by using the UDP listeners of a CLB instance in the classic network.
- The following operations take effect 5 minutes after they are performed on a UDP listener:

- Remove backend servers.
- Set the weight of a backend server to 0 after it is detected unhealthy.

Prerequisites

A CLB instance is created. For more information, see Create an SLB instance.

Step 1: Configure a UDP listener

1.
2. Use one of the following methods to open the listener configuration wizard:
   - On the **Instances** page, find the CLB instance and click **Configure Listener** in the **Actions** column.
   - On the **Instances** page, find the CLB instance that you want to manage and click the ID of the instance. On the **Listeners** tab, click **Add Listener**.
3. Set the following parameters and click **Next**.

Parameter	Description
Select Listener Protocol	Select **UDP**.
Listening Port	Set the listening port that is used to receive requests and forward them to backend servers. Valid values: 1 to 65535. You can set a TCP or UDP listener to listen on all ports within a specified port range.
Listener Name	Specify a name for the listener.
Advanced	Click **Modify** to configure advanced settings.

Scheduling Algorithm	Select a scheduling algorithm.
	- **Weighted Round-Robin (WRR)**: Backend servers that have higher weights receive more requests than backend servers that have lower weights.
	- **Round-Robin (RR)**: Requests are distributed to backend servers in sequence.
	- **Consistent Hash (CH)**:
	  - **Source IP**: specifies consistent hashing that is based on source IP addresses. Requests from the same source IP address are distributed to the same backend server.
	  - **Tuple**: specifies consistent hashing that is based on four factors: source IP address, destination IP address, source port, and destination port. Requests that contain the same information based on the four factors are distributed to the same backend server.
	  - **QUIC ID**: specifies consistent hashing that is based on Quick UDP Internet Connections (QUIC) IDs. Requests that contain the same QUIC ID are distributed to the same backend server.
	Notice QUIC is implemented based on draft-ietf-quic-transport-10 and is rapidly evolving. Therefore, compatibility is not guaranteed for all QUIC versions. We recommend that you perform tests before you apply the protocol to a production environment.

Parameter	Description
Enable Session Persistence	Specify whether to enable session persistence. After session persistence is enabled, the listener forwards all requests from the same client to the same backend server.
Enable Connection Draining	After connection draining is enabled, connections to backend servers can function as expected during the specified timeout period after the backend servers are removed or fail health checks. **Note** This feature is available for only users in a whitelist. To use this feature, .
Enable Peak Bandwidth Limit	Specify whether to set a maximum bandwidth value for the listener. If a CLB instance is billed based on bandwidth usage, you can set different maximum bandwidth values for different listeners. This limits the amount of traffic that flows through each listener. The sum of the maximum bandwidth values of all listeners that are added to a CLB instance cannot exceed the maximum bandwidth value of the CLB instance. By default, this feature is disabled and all listeners share the bandwidth of the CLB instance.
Proxy Protocol	Use the proxy protocol to pass client IP addresses to backend servers. **Note** You cannot enable this feature in scenarios where PrivateLink is used.
Obtain Client Source IP Address	Specify whether to reserve the real IP addresses of clients. Only Layer 4 listeners support this feature. By default, this feature is enabled. **Note** You cannot view source IP addresses by using the UDP listeners of a CLB instance in the classic network. To obtain source IP addresses, enable **Proxy Protocol**.
Automatically Enable Listener After Creation	Specify whether to immediately enable the listener after it is created. By default, this feature is enabled.

Step 2: Add backend servers

After you configure the listener, you must add backend servers to process client requests. You can use the default server group that is configured for the CLB instance. You can also configure a vServer group or a primary/secondary server group, or enable the primary/secondary mode for the listener. For more information, see Backend server overview.

1. On the **Backend Servers** wizard page, select the type of the server group to which requests are forwarded. The default server group is used in this example.

   Select **Default Server Group** and click **Add More**.

2. In the **My Servers** panel, select the ECS instances that you want to add as backend servers and click **Next**.
3. On the **Configure Ports and Weights** wizard page, specify the weights of the backend servers that you added. A backend server with a higher weight receives more requests.

   Note If the weight of a backend server is set to 0, no request is distributed to the backend server.

4. Click **Add**. On the Default Server Group tab, specify the ports that you want to open on the backend servers to receive requests. The backend servers are the ECS instances that you selected. Valid values: 1 to 65535.

   You can specify the same port on different backend servers that are added to a CLB instance.

5. Click **Next**.

Step 3: Configure health checks

CLB performs health checks to check the availability of the ECS instances that serve as backend servers. The health check feature improves overall service availability and reduces the impact of backend server failures.

On the **Health Check** wizard page, click **Modify** to modify the health check configurations. For more information, see Configure health checks.

Step 4: Submit the configurations

1. On the **Confirm** wizard page, check the configurations. You can click **Modify** to modify the configurations.
2. After you confirm the configurations, click **Submit**.
3. When Configuration Successful appears, click **OK**.

   After you configure the listener, you can view the listener on the Listener tab.

17.1.5.4. Add an HTTP listener

This topic describes how to add an HTTP listener to a Classic Load Balancer (CLB) instance. HTTP is applicable to applications that must identify data from different sources, such as web applications and mobile games. You can add HTTP listeners to forward HTTP requests.

Step 1: Configure an HTTP listener

1. Log on to the SLB console.
2. Use one of the following methods to open the listener configuration wizard:
   - On the **Instances** page, find the CLB instance that you want to manage and click **Configure Listener** in the **Actions** column.
   - On the **Instances** page, click the ID of the CLB instance that you want to manage. On the **Listener** tab, click **Add Listener**.
3. Set the following parameters and click **Next**.

Parameter	Description
Select Listener Protocol	Select the protocol of the listener. In this example, **HTTP** is selected.
Listening Port	Set the listening port that is used to receive requests and forward them to backend servers. Valid values: 1 to 65535.
Listener Name	Enter a name for the listener.
Advanced	Click **Modify** to configure advanced settings.
Scheduling Algorithm	Select a scheduling algorithm. ○ **Weighted Round-Robin (WRR)**: Backend servers that have higher weights receive more requests than backend servers that have lower weights. ○ **Round-Robin (RR)**: Requests are evenly and sequentially distributed to backend servers.

Parameter	Description
Redirection	Specify whether to redirect traffic from the HTTP listener to an HTTPS listener. **Note** Before you enable redirection, make sure that an HTTPS listener is created.
Enable Session Persistence	Specify whether to enable session persistence. After session persistence is enabled, CLB forwards all requests from the same client to the same backend server. HTTP session persistence is implemented through cookies. CLB allows you to use the following methods to process cookies: ◦ **Insert cookie**: If you select this option, you only need to specify the timeout period of the cookie. CLB inserts a cookie (SERVERID) into the first HTTP or HTTPS response that is sent to a client. The next request from the client will contain this cookie, and the listener will forward this request to the recorded backend server. ◦ **Rewrite cookie**: If you select this option, you can specify the cookie that you want to insert into an HTTP or HTTPS response. You must specify the timeout period and the lifetime of a cookie on a backend server. After you specify a cookie, CLB overwrites the original cookie with the specified cookie. The next time CLB receives a client request that carries the specified cookie, the listener distributes the request to the recorded backend server.
Enable Peak Bandwidth Limit	Specify whether to set a bandwidth limit for the listener. Unit: Mbit/s. Valid values: 0 to 5120. If a CLB instance is billed based on bandwidth usage, you can set different maximum bandwidth values for different listeners. This limits the amount of traffic that flows through each listener. The sum of the maximum bandwidth values of all listeners that are added to a CLB instance cannot exceed the maximum bandwidth value of the CLB instance. By default, this feature is disabled and all listeners share the bandwidth of the CLB instance.
Idle Timeout	Specify the timeout period of idle connections. Unit: seconds. Valid values: 1 to 60. If no request is received within the specified timeout period, CLB closes the current connection. CLB creates a new connection when a new connection request is received.
Request Timeout	Specify the request timeout period. Unit: seconds. Valid values: 1 to 180. If no response is received from the backend server within the request timeout period, CLB returns an HTTP 504 error to the client.
Enable Gzip Compression	Specify whether to enable Gzip compression to compress specific types of files. Gzip supports the following file types: text/xml, text/plain, text/css, application/javascript, application/x-javascript, application/rss+xml, application/atom+xml, and application/xml.

Parameter	Description
Add HTTP Header Fields	Select the custom HTTP header field that you want to add.
Obtain Client Source IP Address	Specify whether to retrieve the real IP address of the client. By default, this feature is enabled.
Automatically Enable Listener After Creation	Specify whether to immediately enable the listener after it is created. By default, this feature is enabled.

Step 2: Add backend servers

After you configure the listener, you must add backend servers to process client requests. You can use the default server group that is configured for the CLB instance. You can also create a vServer group or a primary/secondary server group. For more information, see Backend server overview.

The default server group is selected in this example.

1. On the **Backend Servers** wizard page, select **Default Server Group**. Then, click **Add More**.
2. In the **My Servers** panel, select the Elastic Compute Service (ECS) instances that you want to add as backend servers and click **Next**.
3. On the **Configure Ports and Weights** wizard page, specify the weights of the backend servers that you want to add. A backend server that has a higher weight receives more requests than a backend server that has a lower weight.

   > **Note** If the weight of a backend server is set to 0, the backend server no longer accepts requests.

4. Click **Add**. On the Default Server Group tab, specify the ports that you want to open on the backend servers to receive requests. Valid values: 1 to 65535. Click **Next**.

   You can specify the same port on different backend servers that are added to a CLB instance.

Step 3: Configure health checks

CLB performs health checks to check the availability of the ECS instances that serve as backend servers. The health check feature improves overall service availability and reduces the impact of backend server failures.

On the **Health Check** wizard page, click **Modify** to modify the health check configurations. For more information, see Configure health checks.

Step 4: Submit the configurations

1. On the **Confirm** wizard page, check the configurations. You can click **Modify** to modify the configurations.
2. After you confirm the configurations, click **Submit**.
3. When Configuration Successful appears, click **OK**.

   After you configure the listener, you can view the listener on the Listener tab.

17.1.5.5. Add an HTTPS listener

This topic describes how to add an HTTPS listener to a Classic Load Balancer (CLB) instance. HTTPS is intended for applications that require encrypted data transmission. You can add an HTTPS listener to forward HTTPS requests.

Step 1: Configure a UDP listener

1.
2. Use one of the following methods to open the listener configuration wizard:
   - On the **Instances** page, find the CLB instance that you want to manage and click **Configure Listener** in the **Actions** column.

- On the **Instances** page, click the ID of the CLB instance that you want to manage. On the **Listener** tab, click **Add Listener**.

3. Set the following parameters and click **Next**.

Parameter	Description
Select Listener Protocol	Select the protocol type of the listener. In this example, **HTTPS** is selected.
Listening Port	Set the listening port that is used to receive requests and forward them to backend servers. Valid values: 1 to 65535.
Listener Name	Enter a name for the listener.
Advanced	Click **Modify** to configure advanced settings.
Scheduling Algorithm	Select a scheduling algorithm. - **Weighted Round-Robin (WRR)**: Backend servers that have higher weights receive more requests than backend servers that have lower weights. - **Round-Robin (RR)**: Requests are distributed to backend servers in sequence.
Enable Session Persistence	Specify whether to enable session persistence. After session persistence is enabled, CLB forwards all requests from the same client to the same backend server. HTTP session persistence is implemented through cookies. CLB allows you to use the following methods to process cookies: - **Insert cookie**: If you select this option, you only need to specify the timeout period of the cookie. CLB inserts a cookie (SERVERID) into the first HTTP or HTTPS response that is sent to a client. The next request from the client will contain this cookie, and the listener will forward this request to the recorded backend server. - **Rewrite cookie**: If you select this option, you can specify the cookie that you want to insert into an HTTP or HTTPS response. You must specify the timeout period and the lifetime of a cookie on a backend server. After you specify a cookie, CLB overwrites the original cookie with the specified cookie. The next time CLB receives a client request that carries the specified cookie, the listener distributes the request to the recorded backend server.
Enable Peak Bandwidth Limit	Specify whether to set a maximum bandwidth value for the listener. If a CLB instance is billed based on bandwidth usage, you can set different maximum bandwidth values for different listeners. This limits the amount of traffic that flows through each listener. The sum of the maximum bandwidth values of all listeners that are added to a CLB instance cannot exceed the maximum bandwidth value of the CLB instance. By default, this feature is disabled and all listeners share the bandwidth of the CLB instance.
Idle Timeout	Specify the timeout period of idle connections. Unit: seconds. Valid values: 1 to 60. If no request is received within the specified timeout period, CLB closes the current connection. CLB creates a new connection when a new connection request is received.

Parameter	Description
Request Timeout	Specify the request timeout period. Unit: seconds. Valid values: 1 to 180. If no response is received from the backend server within the request timeout period, CLB returns an HTTP 504 error to the client.
Enable Gzip Compression	Specify whether to enable Gzip compression to compress specific types of files. Gzip supports the following file types: text/xml, text/plain, text/css, application/javascript, application/x-javascript, application/rss+xml, application/atom+xml, and application/xml.
Add HTTP Header Fields	Select the custom HTTP header field that you want to add.
Obtain Client Source IP Address	Specify whether to retrieve the real IP addresses of clients. By default, this feature is enabled.
Automatically Enable Listener After Creation	Specify whether to immediately enable the listener after it is created. By default, this feature is enabled.

Step 2: Configure an SSL certificate

When you add an HTTPS listener, you must upload a server certificate or certification authority (CA) certificate , as shown in the following table.

Certificate	Description	Required for one-way authentication	Required for mutual authentication
Server certificate	The certificate that is used to identify the server. Your browser uses the server certificate to check whether the certificate sent by the server is signed and issued by a trusted CA.	Yes You must upload the server certificate to the certificate management system of CLB.	Yes You must upload the server certificate to the certificate management system of CLB.
Client certificate	The certificate that is used to identify the client. The server identifies the client by checking the certificate sent by the client. You can sign a client certificate with a self-signed CA certificate.	No	Yes You must install the client certificate on the client.
CA certificate	The server uses a CA certificate to verify the signature on the client certificate. If the signature is invalid, the connection request is denied.	No	Yes You must upload the CA certificate to the certificate management system of CLB.

Before you upload a certificate, take note of the following rules:

- CLB supports the following public key algorithms: RSA 1024, RSA 2048, RSA 4096, ECDSA P-256, ECDSA P-384, and ECDSA P-521.
- The certificate that you want to upload must be in the PEM format.
- After you upload a certificate to CLB, CLB can manage the certificate. You do not need to bind the certificate to backend servers.
- It may take a few minutes to upload, load, and verify the certificate. Therefore, an HTTPS listener is not enabled

immediately after it is created. It takes about 1 to 3 minutes to enable an HTTPS listener.

- The ECDHE cipher suite used by HTTPS listeners supports forward secrecy. It does not support the security enhancement parameters that are required by the DHE cipher suite. Therefore, you cannot upload certificates (PEM files) that contain the `BEGIN DH PARAMETERS` field. For more information, see Certificate requirements.
- HTTPS listeners do not support Server Name Indication (SNI). You can choose TCP listeners and configure SNI on backend servers.
- By default, the timeout period of session tickets for HTTPS listeners is 300 seconds.
- The actual amount of data transfer on an HTTPS listener is larger than the billed amount because a portion of data is used for handshaking.
- If a large number of connections are established, a large amount of data is used for handshaking.

1. On the **SSL Certificates** wizard page, select the server certificate that you uploaded. You can also click **Create Server Certificate** to upload a server certificate.
2. To enable mutual authentication or configure a TLS security policy, click **Modify** next to **Advanced**.
3. Enable mutual authentication, and select an uploaded CA certificate. You can also create a CA certificate.

Step 3: Add backend servers

After configuring the listener, you must add backend servers to process client requests. You can use the default server group that is configured for the CLB instance. You can also create a vServer group or a primary/secondary server group. For more information, see Backend server overview.

The default server group is selected in this example.

1. On the **Backend Servers** wizard page, select **Default Server Group**. Then, click **Add More**.
2. In the **My Servers** panel, select the Elastic Compute Service (ECS) instances that you want to add as backend servers and click **Next**.
3. Set weights for the selected ECS instances in the **Weight** column.

   > **Note**
   > - An ECS instance with a higher weight receives more requests. The default weight is 100. You can click **Reset** to set the **weight** to the default value.
   > - If you set the weight of a server to 0, the server does not receive requests.

4. Click **Add**. On the Default Server Group tab, specify the ports that you want to open on the backend servers to receive requests. Valid values: 1 to 65535. Click **Next**.

   You can specify the same port on different backend servers that are added to a CLB instance.

Step 4: Configure health checks

CLB performs health checks to check the availability of the ECS instances that serve as backend servers. The health check feature improves overall service availability and reduces the impact of backend server failures.

Step 5: Submit the configurations

1. On the **Confirm** wizard page, check the configurations. You can click **Modify** to modify the configurations.
2. After you confirm the configurations, click **Submit**.
3. When Configuration Successful appears, click **OK**.

   After you configure the listener, you can view the listener on the Listener tab.

17.1.5.6. Manage TLS security policies

When you create or configure an HTTPS listener for a high-performance Classic Load Balancer (CLB) instance, you can select a Transport Layer Security (TLS) security policy.

Select a TLS security policy

When you create or configure an HTTPS listener, you can modify the advanced settings on the **SSL Certificates** page and select a TLS security policy. For more information, see Add an HTTPS listener.

TLS security policies

A TLS security policy contains TLS protocol versions and cipher suites that are available for HTTPS. A later version of TLS ensures higher security of HTTPS communication than an earlier version. However, a later version is less compatible with browsers than an earlier version.

TLS security policy	Supported TLS version	Supported cipher suite
tls_cipher_policy_1_0	TLS 1.0, TLS 1.1, and TLS 1.2	ECDHE-ECDSA-AES128-GCM-SHA256, ECDHE-ECDSA-AES256-GCM-SHA384, ECDHE-ECDSA-AES128-SHA256, ECDHE-ECDSA-AES256-SHA384, ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES256-SHA384, AES128-GCM-SHA256, AES256-GCM-SHA384, AES128-SHA256, AES256-SHA256, ECDHE-ECDSA-AES128-SHA, ECDHE-ECDSA-AES256-SHA, ECDHE-RSA-AES128-SHA, ECDHE-RSA-AES256-SHA, AES128-SHA, AES256-SHA, and DES-CBC3-SHA
tls_cipher_policy_1_1	TLS 1.1 and TLS 1.2	ECDHE-ECDSA-AES128-GCM-SHA256, ECDHE-ECDSA-AES256-GCM-SHA384, ECDHE-ECDSA-AES128-SHA256, ECDHE-ECDSA-AES256-SHA384, ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES256-SHA384, AES128-GCM-SHA256, AES256-GCM-SHA384, AES128-SHA256, AES256-SHA256, ECDHE-ECDSA-AES128-SHA, ECDHE-ECDSA-AES256-SHA, ECDHE-RSA-AES128-SHA, ECDHE-RSA-AES256-SHA, AES128-SHA, AES256-SHA, and DES-CBC3-SHA
tls_cipher_policy_1_2	TLSv1.2	ECDHE-ECDSA-AES128-GCM-SHA256, ECDHE-ECDSA-AES256-GCM-SHA384, ECDHE-ECDSA-AES128-SHA256, ECDHE-ECDSA-AES256-SHA384, ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES256-SHA384, AES128-GCM-SHA256, AES256-GCM-SHA384, AES128-SHA256,AES256-SHA256, ECDHE-ECDSA-AES128-SHA, ECDHE-ECDSA-AES256-SHA, ECDHE-RSA-AES128-SHA, ECDHE-RSA-AES256-SHA, AES128-SHA, AES256-SHA, and DES-CBC3-SHA
tls_cipher_policy_1_2_strict	TLSv1.2	ECDHE-ECDSA-AES128-GCM-SHA256, ECDHE-ECDSA-AES256-GCM-SHA384, ECDHE-ECDSA-AES128-SHA256, ECDHE-ECDSA-AES256-SHA384, ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES256-SHA384, ECDHE-ECDSA-AES128-SHA, ECDHE-ECDSA-AES256-SHA, ECDHE-RSA-AES128-SHA, and ECDHE-RSA-AES256-SHA
tls_cipher_policy_1_2_without_sha1	TLSv1.2	ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES256-SHA384, AES128-GCM-SHA256, AES256-GCM-SHA384, and AES256-SHA256

TLS security policy	Supported TLS version	Supported cipher suite
tls_cipher_policy_1_2_strict_with_1_3	TLS 1.2 and TLS 1.3	TLS_AES_128_GCM_SHA256, TLS_AES_256_GCM_SHA384, TLS_CHACHA20_POLY1305_SHA256, TLS_AES_128_CCM_SHA256, TLS_AES_128_CCM_8_SHA256, ECDHE-ECDSA-AES128-GCM-SHA256, ECDHE-ECDSA-AES256-GCM-SHA384, ECDHE-ECDSA-AES128-SHA256, ECDHE-ECDSA-AES256-SHA384, ECDHE-RSA-AES128-GCM-SHA256, ECDHE-RSA-AES256-GCM-SHA384, ECDHE-RSA-AES128-SHA256, ECDHE-RSA-AES256-SHA384, ECDHE-ECDSA-AES128-SHA, ECDHE-ECDSA-AES256-SHA, ECDHE-RSA-AES128-SHA, and ECDHE-RSA-AES256-SHA

About the preceding security policies

Cipher suites supported by TLS security policies

TLS security policy		tls_cipher_policy_1_0	tls_cipher_policy_1_1	tls_cipher_policy_1_2	tls_cipher_policy_1_2_strict	tls_cipher_policy_1_2_strict_with_1_3
TLS		1.2, 1.1, and 1.0	1.1 and 1.2	1.2	1.2	1.2 and 1.3
CIPHER	ECDHE-RSA-AES128-GCM-SHA256	√	√	√	√	√
	ECDHE-RSA-AES256-GCM-SHA384	√	√	√	√	√
	ECDHE-RSA-AES128-SHA256	√	√	√	√	√
	ECDHE-RSA-AES256-SHA384	√	√	√	√	√
	AES128-GCM-SHA256	√	√	√	-	-
	AES256-GCM-SHA384	√	√	√	-	-
	AES128-SHA256	√	√	√	-	-
	AES256-SHA256	√	√	√	-	-
	ECDHE-RSA-AES128-SHA	√	√	√	√	√
	ECDHE-RSA-AES256-SHA	√	√	√	√	√
	AES128-SHA	√	√	√	-	-
	AES256-SHA	√	√	√	-	-
	DES-CBC3-SHA	√	√	√	-	-
	TLS_AES_128_GCM_SHA256	-	-	-	-	√

TLS security policy		tls_cipher_policy_1_0	tls_cipher_policy_1_1	tls_cipher_policy_1_2	tls_cipher_policy_1_2_strict	tls_cipher_policy_1_2_strict_with_1_3
	TLS_AES_256_GCM_SHA384	-	-	-	-	√
	TLS_CHACHA20_POLY1305_SHA256	-	-	-	-	√
	TLS_AES_128_CCM_SHA256	-	-	-	-	√
	TLS_AES_128_CCM_8_SHA256	-	-	-	-	√
	ECDHE-ECDSA-AES128-GCM-SHA256	-	-	-	-	√
	ECDHE-ECDSA-AES256-GCM-SHA384	-	-	-	-	√
	ECDHE-ECDSA-AES128-SHA256	-	-	-	-	√
	ECDHE-ECDSA-AES256-SHA384	-	-	-	-	√
	ECDHE-ECDSA-AES128-SHA	-	-	-	-	√
	ECDHE-ECDSA-AES256-SHA	-	-	-	-	√

Note The √ sign in the preceding table indicates that a cipher suite is supported, while the - sign indicates that a cipher suite is not supported.

17.1.5.7. Configure forwarding rules

This topic describes how to configure forwarding rules for a Server Load Balancer (SLB) instance. You can configure domain name-based or URL-based forwarding rules for an SLB instance that uses Layer 7 listeners. Layer 7 listeners distribute requests destined for different domain names or URLs to different Elastic Compute Service (ECS) instances.

Context

You can add multiple forwarding rules to a listener. Each forwarding rule is associated with a unique server group. Each server group contains one or more ECS instances. For example, you can configure a listener to forward read requests to one server group and write requests to another server group. This allows you to optimize load balancing among your server resources.

SLB forwards requests based on the following rules:

- If a request matches a domain name-based or URL-based forwarding rule of a listener, the request is forwarded to the corresponding server group based on the forwarding rule.
- If a request does not match the domain name-based or URL-based forwarding rules of a listener but the listener is associated with a server group, the request is forwarded to the server group.
- If none of the preceding conditions are met, requests are forwarded to the ECS instances in the default server

group of the SLB instance.

Procedure

1.
2. Click the ID of the SLB instance that you want to manage. On the **Listener** tab, find the listener that you want to manage.

   You can configure domain name-based or URL-based forwarding rules only for HTTP and HTTPS listeners.
3. Click **Configure Forwarding Rule** in the **Actions** column.
4. Configure forwarding rules based on the following information:
   - Configure a domain name-based forwarding rule
     - When you configure a domain name-based forwarding rule, leave the URL field empty. You do not need to enter a forward slash (/) in this field. The domain name can contain only letters, digits, hyphens (-), and periods (.).
     - Domain-based forwarding rules support both exact matching and wildcard matching. For example, www.aliyun.com is an exact domain name, whereas *.aliyun.com and *.market.aliyun.com are wildcard domain names. When a request matches multiple domain name-based forwarding rules, an exact match prevails over wildcard matches, as described in the following table. Domain name matching rule

Type	Request URL	Domain name matching rule (√ indicates that the domain name is matched whereas x indicates that the domain name is not matched.)		
		www.aliyun.com	*.aliyun.com	*.market.aliyun.com
Exact match	www.aliyun.com	√	x	x
Wildcard match	market.aliyun.com	x	x	x
	info.market.aliyun.com	x	x	√

   - Configure a URL-based forwarding rule
     - When you configure a URL-based forwarding rule, leave the Domain Name field empty.
     - The URL can contain only letters, digits, and hyphens `(-)./%?#&`
     - The URL must start with a forward slash (/).

       > **Note** If you enter only a forward slash (/) in the URL field, the URL-based forwarding rule is invalid.

     - URL-based forwarding rules support string matching and adopt sequential matching. Examples: `/admin`, `/bbs_`, and `/ino_test`.
   - Configure both domain name-based and URL-based forwarding rules

     You can configure both domain name-based and URL-based forwarding rules to forward traffic destined for different URLs of the same domain name. We recommend that you configure a default forwarding rule with the URL field left empty in case errors are returned when the URLs of requests are not matched.

     For example, the domain name of a website is `www.aaa.com`. You are required to forward requests destined for `www.aaa.com/index.html` to Server Group 1 and forward requests destined for other URLs of the domain name to Server Group 2. To meet the preceding requirements, you must configure two forwarding rules, as shown in the following figure. Otherwise, a 404 error code is returned when a request destined for the `www.aaa.com` domain name does not match all forwarding rules.
5. Click **Save**.

17.1.5.8. Enable access control

This topic describes how to enable access control for a listener. You can enable access control for each listener of a Classic Load Balancer (CLB) instance. You can set whitelists for different listeners to regulate network access control.

Procedure

1. Log on to the SLB console.
2. Click the ID of the CLB instance.
3. Click the **Listener** tab, find the listener that you want to manage, and then choose ⋮ > **Set Access Control** in the **Actions** column.
4. Set the following parameters and click **OK**.

Parameter	Configuration method
Enable Access Control	Enable access control.
Access Control Method	**Whitelist**: After you set a whitelist for a listener, the listener forwards only requests from IP addresses or CIDR blocks that are added to the whitelist. Risks may arise if the whitelist is improperly set. After a whitelist is configured, only IP addresses in the whitelist can access the CLB listener. If you enable a whitelist but the whitelist does not contain an IP address, the listener forwards all requests.
Access Control List	Select a network ACL. IPv6 instances can be associated only with IPv6 network ACLs, and IPv4 instances can be associated only with IPv4 network ACLs. **Note** Separate multiple IP entries with commas (,). You can add up to 300 IP entries to each network ACL. IP entries must be unique within each network ACL.

17.1.5.9. Disable access control

This topic describes how to disable access control for a listener.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. On the page that appears, click the **Listener** tab.
4. Find the target listener, and choose  > **Set Access Control** in the Actions column.
5. In the **Access Control Settings** dialog box, disable access control and then click **OK**.

17.1.6. Backend servers

17.1.6.1. Backend server overview

Before you use a Classic Load Balancer (CLB) instance, you must specify Elastic Compute Service (ECS) instances as the backend servers of the CLB instance to receive client requests.

Introduction

CLB creates a server group for multiple ECS instances in the same region and sets a virtual IP address for the server group to ensure high performance and high availability. You can also use vServer groups to manage backend servers. You can associate listeners with different server groups. This way, CLB can distribute requests to backend servers that use different ports.

> **Note** If you associate a listener with a vServer group, the listener distributes requests to the ECS instances in the vServer group instead of the ECS instances in the default server group.

Limits

You can add ECS instances to or remove ECS instances from a CLB instance anytime. CLB distributes network traffic across groups of ECS instances. Before you use CLB, make sure that health checks are enabled and at least one ECS instance works as expected to ensure service stability.

When you add backend servers, take note of the following items:

- You can add two ECS instances that run different operating systems to a CLB instance. However, the applications deployed on the ECS instances must be the same and have consistent data. To simplify management and maintenance, we recommend that you add ECS instances that use the same operating system to a CLB instance.
- You must configure a listener for each application deployed on an ECS instance. Each CLB instance supports up to 50 listeners. CLB uses listening ports to receive client requests and forward the requests to backend ports that are used by the applications on ECS instances.
- You can specify a weight for each ECS instance in a server group. An ECS instance with a higher weight receives more requests.
- If session persistence is enabled, requests may not be evenly distributed to each backend server. We recommend that you disable session persistence and check whether the issue persists.

  If requests are not evenly distributed, troubleshoot the issue by performing the following steps:

  i. Count the numbers of access log entries generated on different ECS instances within a specified time period.
  ii. Check whether the numbers of access log entries generated on different ECS instances have deviations based on the CLB configurations. If session persistence is enabled, you must exclude access log entries that contain the same IP address. If the ECS instances have different weights, you must check whether the numbers of access log entries are also weighted.
- When an ECS instance performs hot migration, persistent connections to CLB may be closed. Make sure that your applications are configured with the automatic reconnection mechanism.

Primary/secondary mode

In primary/secondary mode, a listener is associated with a primary server group and a secondary server group. If the number of unhealthy ECS instances in the primary server group reaches the failover threshold, requests are distributed to the secondary server group. Each server group contains one or more ECS instances. A listener associated with a primary server group and a secondary server group provides higher reliability than a listener associated with a primary server and a secondary server.

Before you enable the primary/secondary mode for a listener, make sure that the listener is associated with at least two vServer groups. For more information about how to create a vServer group, see Create a vServer group. You can enable the primary/secondary mode for a listener when you add backend servers to the listener.

> **Note** Only TCP and UDP listeners support the primary/secondary mode.

Default server group

You can add ECS instances to the default server group of a listener to receive requests. If a listener is not associated with a vServer group or a primary/secondary server group, requests are distributed to the ECS instances in the default server group.

Before a CLB instance can process requests, you must add at least one backend server to the default server group to receive requests. For more information, see Add a default backend server.

vServer groups

You can create vServer groups for CLB to distribute different requests to different backend servers. To allow CLB to distribute requests based on domain names and URLs, you can specify vServer groups in domain name-based forwarding rules and URL-based forwarding rules. For more information, see Create a vServer group.

Primary/secondary server groups

A primary/secondary server group contains only two ECS instances. One ECS instance serves as the primary server and the other ECS instance serves as the secondary server. CLB does not perform health checks on the secondary server in a primary/secondary server group. If the primary server is down, network traffic is automatically distributed to the secondary server. When the primary server recovers, traffic is switched back to the primary server. For more information, see Create a primary/secondary server group.

> **Note** You can add primary/secondary server groups only to TCP and UDP listeners.

Related information

- Add a default backend server
- Create a vServer group
- Create a primary/secondary server group

17.1.6.2. Default server groups

17.1.6.2.1. Add a default backend server

This topic describes how to add a default backend server. Before you use the Classic Load Balancer (CLB) service, you must add at least one default backend server to receive client requests.

Prerequisites

Before you add an Elastic Compute Service (ECS) instance to the default server group, make sure that the following requirements are met:

- A CLB instance is created. For more information, see Create an SLB instance.
- ECS instances are created and applications are deployed on the ECS instances to receive requests.

Procedure

1. Log on to the SLB console.
2. Find the CLB instance that you want to manage and click its ID.
3. Click the **Default Server Group** tab.
4. Click **Add**.
5. In the **My Servers** panel, for **Select Servers**, select one or more ECS instances that you want to add to the default server group.
6. Click **Next**.
7. For **Configure Ports and Weights**, specify the weight of each ECS instance.

   An ECS instance that has a higher weight receives more requests.

You can change the weight of a server and then move the pointer over to change the weights of other servers:

- If you click **Replicate to Below**, the weights of all servers below the current server are set to the weight of the current server.
- If you click **Replicate to Above**, the weights of all servers above the current server are set to the weight of the current server.
- If you click **Replicate to All**, the weights of all servers in the default server group are set to the weight of the current server.
- If you click **Reset**, when you clear the weight of the current server, the weights of all servers in the default server group are cleared.

> **Notice**
> - Valid values of weights: 0 to 100. If you set the weight of a server to 0, the server does not receive requests.
> - If two servers have the same weight, only one server receives requests.

8. Click **Add**.
9. Click **OK**.

17.1.6.2.2. Add IDC servers to the default server group

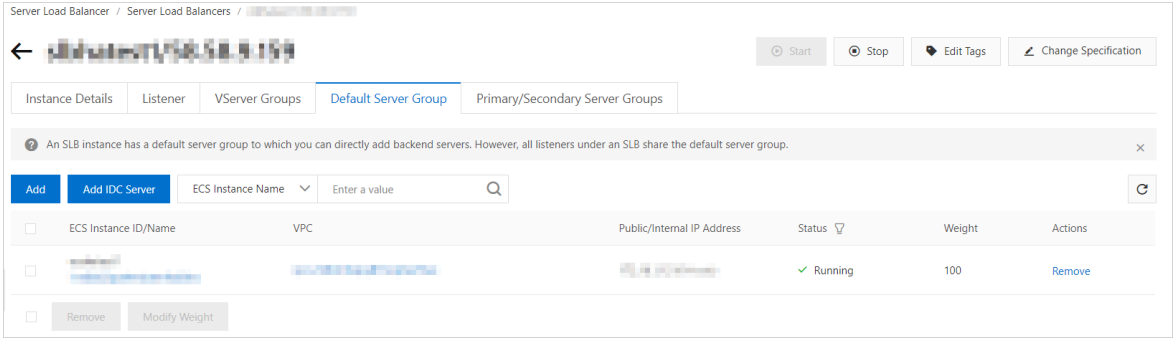
This topic describes how to add servers in on-premises Internet Data Centers (IDCs) as default backend servers to the default server group of an SLB instance. Before you use the SLB service, you must add at least one default backend server to receive client requests forwarded by SLB.

Prerequisites

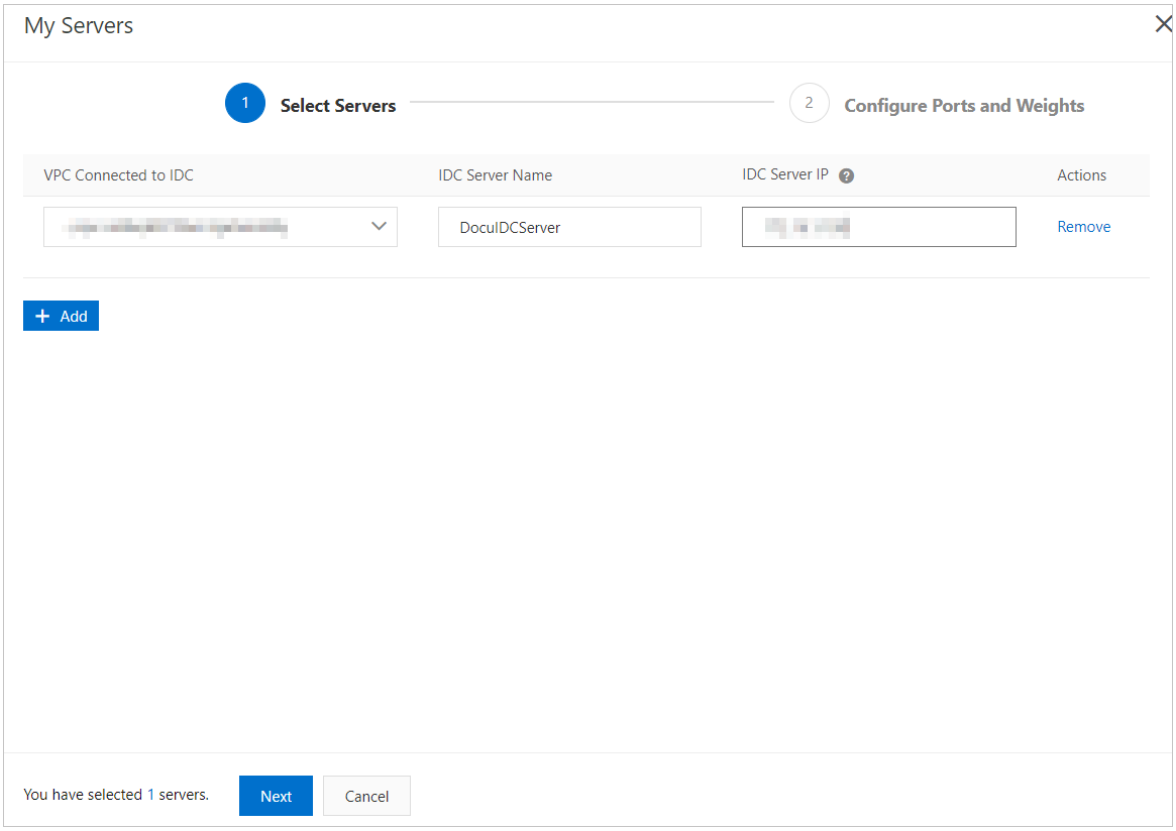
Applications are deployed on the IDC servers, and the IDC servers are ready to receive distributed requests.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **Default Server Group** tab.
4. Click **Add IDC Server**.



5. In the **My Servers** dialog box, click **Add**.
6. Select a VPC from the VPC Connected to IDC drop-down list, enter a name for the IDC server, and specify the IP address of the IDC server.



7. Click **Next**.
8. In the **Configure Ports and Weights** step, specify the weight of each added IDC server.

   An IDC server with a higher weight receives more requests.

   You can change the weight of a server and then move the pointer over to change the weights of multiple servers:

   - Click **Replicate to Below**: The weights of all servers below the current server are set to the weight of the current server.
   - Click **Replicate to Above**: The weights of all servers above the current server are set to the weight of the current server.
   - Click **Replicate to All**: The weights of all servers in the default server group are set to the weight of the current server.
   - Click **Reset**: The weight fields of all servers in the default server group are cleared.

   > **Notice** If the weight of a backend server is set to 0, this backend server no longer receives new requests.

9. Click **Add**.
10. Click **OK**.

17.1.6.2.3. Change the weight of a backend server

This topic describes how to change the weight of a backend server to adjust the proportion of requests sent to the backend server.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **Default Server Group** tab.
4. Move the pointer over the weight value of the target backend server and click the  icon.
5. Change the weight and then click **OK**.

   A backend server (ECS instance or IDC server) with a higher weight receives more requests.

   > **Notice** The weight value ranges from 0 to 100. If the weight of a backend server is set to 0, no requests are sent to the backend server.

17.1.6.2.4. Remove a backend server

This topic describes how to remove a backend server that is no longer needed.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **Default Server Group** tab.
4. Find the target backend server and click **Remove** in the **Actions** column.
5. In the dialog box that appears, click **OK**.

17.1.6.3. VServer groups

17.1.6.3.1. Create a vServer group

This topic describes how to create a vServer group for a Server Load Balancer (SLB) instance. A vServer group contains Elastic Compute Service (ECS) instances that function as backend servers. If you associate a vServer group with a listener, the listener distributes requests only to backend servers in the vServer group.

Prerequisites

Before you create a vServer group, make sure that the following conditions are met:

- An SLB instance is created. For more information, see Create an SLB instance.
- ECS instances are created and applications are deployed on the ECS instances to process requests.

Context

Take note of the following items before you create a vServer group for an SLB instance:

- ECS instances are added to a vServer group and the corresponding SLB instance must be deployed in the same region.
- An ECS instance can be added to multiple vServer groups.
- A vServer group can be associated with multiple listeners of an SLB instance.
- A vServer group consists of ECS instances and application ports.

Procedure

1. Log on to the SLB console.
2. Find the SLB instance and click its instance ID.

3. Click the **VServer Groups** tab.
4. On the **VServer Groups** tab, click **Create VServer Group**.
5. On the **Create VServer Group** page, set the parameters.
   i. In the **VServer Group Name** field, enter a name for the vServer group.
   ii. Click **Add**. On the **My Servers** wizard page, select the ECS instances that you want to add.
   iii. Click **Next**.
   iv. Set the Port and Weight parameters for each ECS instance, and then click **Add**.

   Set the Port and Weight parameters based on the following information:

   - **Port**: The backend port opened on an ECS instance to receive requests.

     You can set the same port number for multiple backend servers of an SLB instance.
   - **Weight**: An ECS instance with a higher weight receives more requests.

     > **Notice** If the weight of an ECS instance is set to 0, the ECS instance no longer receives new requests.

   You can click  to specify the ports and weights of the added ECS instances in batches.

   - **Replicate to Below**: The ports or weights of all servers below the current server are set to the port or weight of the current server.
   - **Replicate to Above**: The ports or weights of all servers above the current server are set to the port or weight of the current server.
   - **Replicate to All**: The ports or weights of all servers in the vServer group are set to the port or weight of the current server.
   - **Reset**: If the port or weight of the current server is cleared, the ports or weights of all servers in the vServer group are also cleared.
6. Click **Create**.

17.1.6.3.2. Add IDC servers to a VServer group

This topic describes how to create a VServer group and then add IDC servers to the VServer group. You can add ECS instances and IDC servers as backend servers to a VServer group. If you associate a VServer group with a listener, the listener distributes requests only to the backend servers in the VServer group instead of other backend servers.

Prerequisites

Before you create a VServer group, make sure that applications are deployed on the IDC servers and the IDC servers are ready to receive distributed requests.

Context

Note the following items before you create a VServer group:

- An IDC server can be added to multiple VServer groups.
- A VServer group can be associated with multiple listeners of an SLB instance.
- The settings of the VServer group include the settings of IDC servers and application ports.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **VServer Groups** tab.

4. On the **VServer Groups** tab, click **Create VServer Group**.
5. On the **Create VServer Group** page, configure the VServer group.
   i. In the **VServer Group Name** field, enter a name for the VServer group.
   ii. Click **Add IDC Server**.
   iii. In the **My Servers** dialog box, click **Add**.
   iv. Select a VPC from the VPC Connected to IDC drop-down list, enter a name for the IDC server, and specify the IP address of the IDC server.

   The IP address of the IDC server must be accessible to the VPC.
   v. Click **Next**.
   vi. Specify a port and weight for each IDC server, and then click **Add**.

   Set the ports and weights based on the following information:

   - **Port**: The backend port opened on an IDC server to receive requests. Multiple ports can be added to an IDC server.

     You can set the same port number for multiple backend servers of the same SLB instance.
   - **Weight**: An IDC server with a higher weight receives more requests.

     > **Notice** If the weight of an IDC server is set to 0, the IDC server no longer receives new requests.

   You can change the weight of a server and then move the pointer over  to change the weights of multiple servers:

   - Click **Replicate to Below**: The weights of all servers below the current server are set to the weight of the current server.
   - Click **Replicate to Above**: The weights of all servers above the current server are set to the weight of the current server.
   - Click **Replicate to All**: The weights of all servers in the VServer group are set to the weight of the current server.
   - Click **Reset**: The weight fields of all servers in the VServer group are cleared.

   > **Notice** If the weight of a backend server is set to 0, this backend server no longer receives new requests.

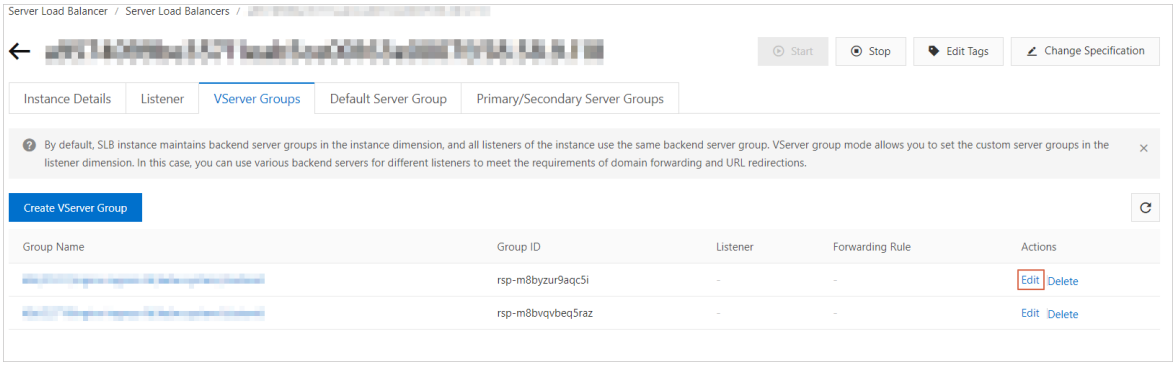
   vii. Click **Add**.
6. Click **Create**.

17.1.6.3.3. Modify a VServer group

This topic describes how to modify the settings of ECS instances or IDC servers in a VServer group.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **VServer Groups** tab.
4. Find the target VServer group and then click **Edit** in the Actions column.



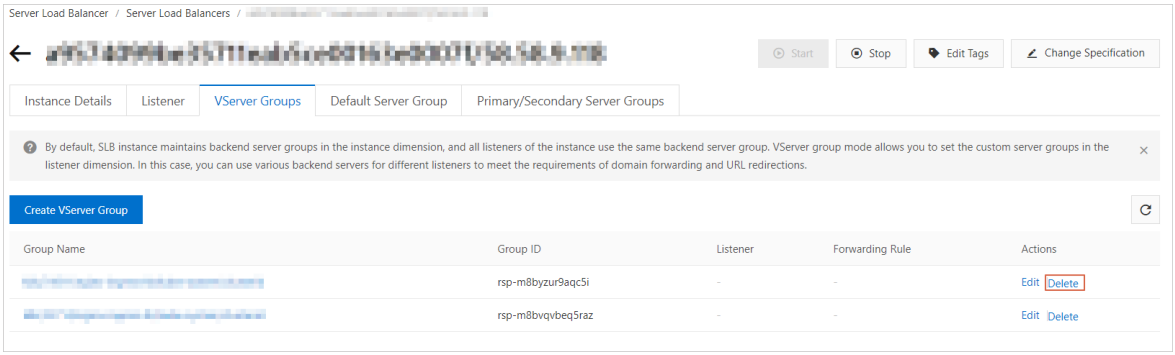
5. Modify the ports and weights of ECS instances or IDC servers, and then click **Save**.

17.1.6.3.4. Delete a VServer group

This topic describes how to delete a VServer group that is no longer needed for traffic distribution.

Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **VServer Groups** tab.
4. Find the target VServer group, and then click **Delete** in the Actions column.



5. In the dialog box that appears, click **OK**.

17.1.6.4. Active/standby server groups

17.1.6.4.1. Create a primary/secondary server group

This topic describes how to create a primary/secondary server group and then add Elastic Compute Service (ECS) instances to the primary/secondary server group. A primary/secondary server group contains a primary server and a secondary server that can fail over to prevent service interruption. By default, the primary server handles all requests that are distributed by the SLB instance. When the primary server fails, requests are redirected to the secondary server.

Prerequisites

Before you create a primary/secondary server group, make sure that the following requirements are met:

- A Server Load Balancer (SLB) instance is created. For more information, see Create an SLB instance.
- ECS instances are created and applications are deployed on the ECS instances to process requests.

> **Note** Only TCP and UDP listeners support primary/secondary server groups.

Procedure

1. Log on to the SLB console.
2. Find the SLB instance and click its instance ID.
3. Click the **Primary/Secondary Server Groups** tab.
4. On the **Primary/Secondary Server Groups** tab, click **Create Primary/Secondary Server Group**.
5. On the **Create Primary/Secondary Server Group** page, configure the primary/secondary server group.
   i. In the **Primary/Secondary Server Group Name** field, enter a name for the primary/secondary server group.
   ii. Click **Add**. In the **My Servers** panel, select the ECS instances that you want to add.

   You can add only two ECS instances to a primary/secondary server group.
   iii. Click **Next**.
   iv. On the **Configure Ports and Weights** wizard page, specify the backend ports that you want to open on the ECS instances to receive requests. If you want to open more than one port on an ECS instance, click **Add Port** in the **Actions** column.

   You can open the same port on different backend servers that are connected to the same SLB instance.
   v. Click **Add**.
6. On the **Create Primary/Secondary Server Group** page, select an ECS instance in the **Type** column as the primary server.
7. Click **Create**.

17.1.6.4.2. Add IDC servers to a primary/secondary server group

This topic describes how to create a primary/secondary server group and then add IDC servers to the primary/secondary server group. You can use a primary/secondary server group to implement failover between a primary server and a secondary server. By default, the primary server handles all distributed requests. When the primary server fails, traffic is redirected to the secondary server.

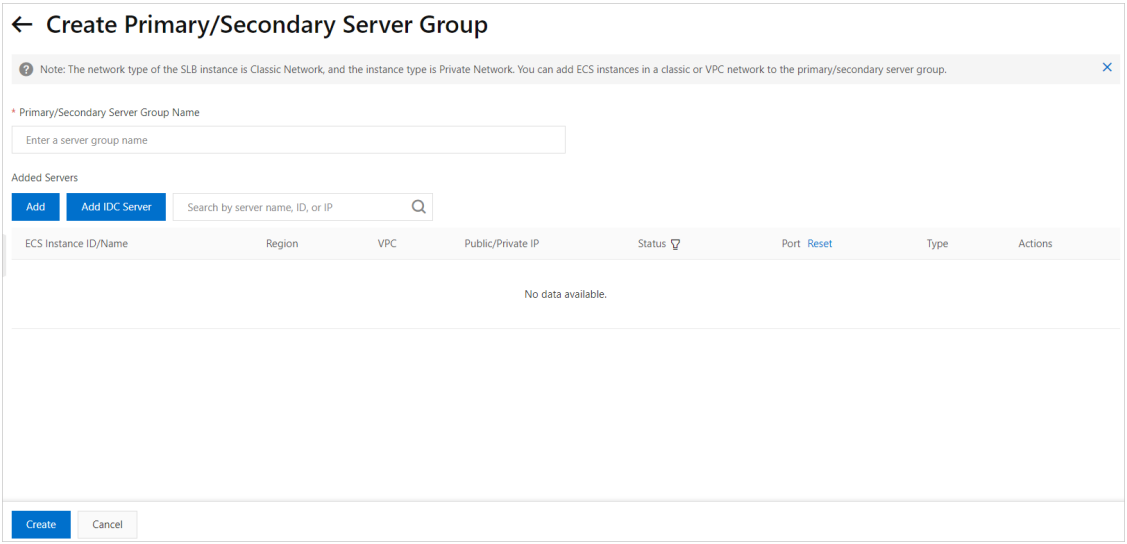
Prerequisites

The IDC servers are created, configured to deploy applications, and ready to receive distributed requests.

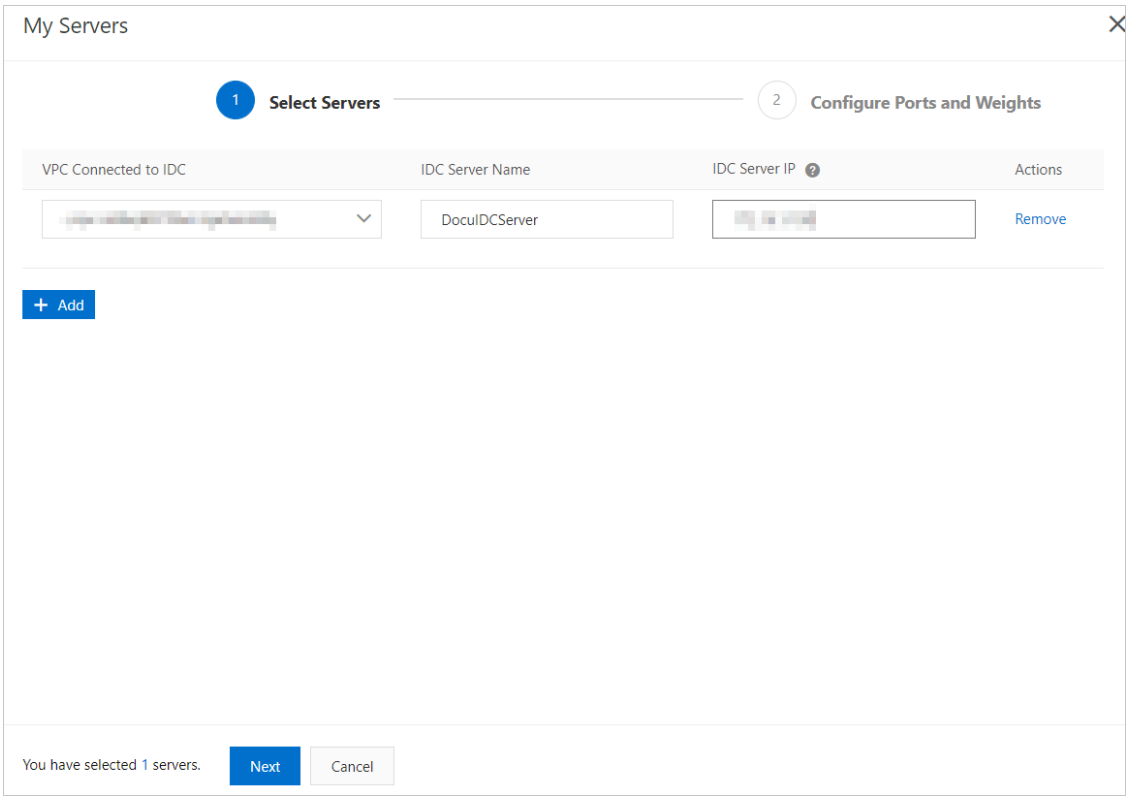
Procedure

1. Log on to the SLB console.
2. Find the target SLB instance and click its instance ID.
3. Click the **Primary/Secondary Server Groups** tab.
4. On the **Primary/Secondary Server Groups** tab, click **Create Primary/Secondary Server Group**.
5. On the **Create Primary/Secondary Server Group** page, configure the primary/secondary server group.

i. In the **Primary/Secondary Server Group Name** field, enter a name for the primary/secondary server group, and then click **Add IDC Server**.



ii. In the **My Servers** dialog box, click **Add**.

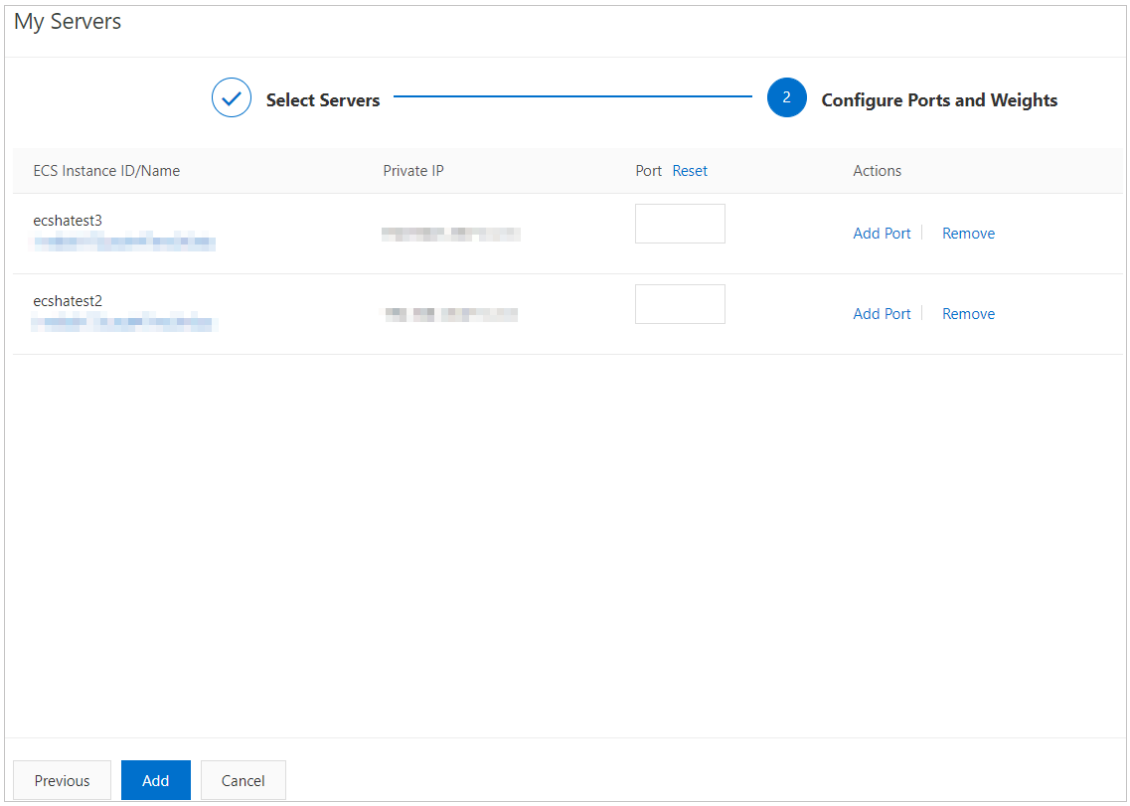


iii. Select a VPC from the VPC Connected to IDC drop-down list, enter a name for the IDC server, and specify the IP address of the IDC server.

The IP address of the IDC server must be accessible to the VPC.

iv. Click **Next**.

v. Configure the backend ports opened on ECS instances to receive requests, and then click **Add**.



You can set multiple ports for an IDC server.

vi. Set a backend server as the primary server.

vii. Click **Create**.

17.1.6.4.3. Delete a primary/secondary server group

This topic describes how to delete a primary/secondary server group of a Server Load Balancer (SLB) instance. If a primary/secondary server group is no longer needed to forward traffic, you can delete the primary/secondary server group.

Procedure

1. Log on to the SLB console.
2. Find the SLB instance and click its instance ID.
3. Click the **Primary/Secondary Server Groups** tab.
4. On the **Primary/Secondary Server Groups** tab, find the primary/secondary server group that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

17.1.7. Health check

17.1.7.1. Health check overview

This topic describes the health check feature of Server Load Balancer (SLB). SLB checks the availability of Elastic Compute Service (ECS) instances that act as backend servers by performing health checks. The health check feature improves the overall availability of your frontend business and mitigates the impacts of exceptions that occur on backend ECS instances.

After you enable the health check feature, SLB stops distributing requests to ECS instances that are declared unhealthy and distributes new requests to healthy ECS instances. When the unhealthy ECS instances have recovered, SLB starts forwarding requests to these ECS instances again.

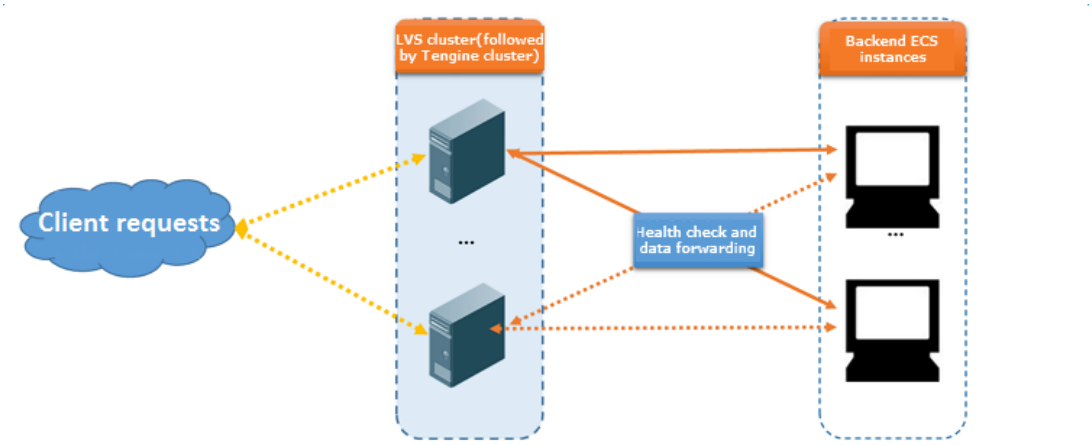
If your business is highly sensitive to traffic loads, frequent health checks may impact the availability of normal business. To reduce the impacts of health checks on your business, you can reduce the health check frequency, increase the health check interval, or change Layer 7 health checks to Layer 4 health checks. We recommend that you do not disable the health check feature to ensure business continuity.

Health check process

SLB is deployed in clusters. Node servers in the LVS or Tengine cluster forward data and perform health checks.

The node servers in the LVS cluster forward data and perform health checks independently and in parallel based on configured load balancing policies. If an LVS node server detects that a backend ECS instance is unhealthy, this node server no longer sends new client requests to this ECS instance. This operation is synchronized among all node servers in the LVS cluster.

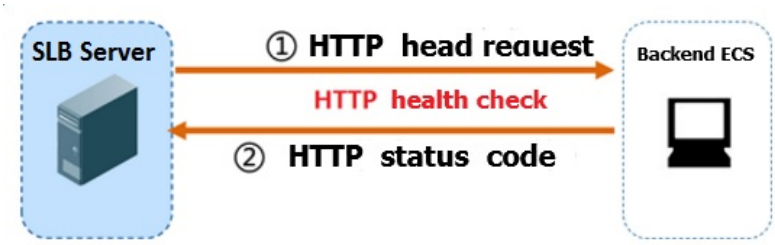
SLB uses the CIDR block of 100.64.0.0/10 for health checks. Make sure that backend ECS instances do not block this CIDR block. You do not need to configure a security group rule to allow access from this CIDR block. However, if you have configured security rules such as iptables, you must allow access from this CIDR block. 100.64.0.0/10 is reserved by Alibaba Cloud. Other users cannot use any IP addresses within this CIDR block, and therefore no relevant security risks exist.



Health checks of HTTP or HTTPS listeners

For Layer 7 (HTTP or HTTPS) listeners, SLB checks the status of backend ECS instances by sending HTTP HEAD requests. The following figure shows the process.

For HTTPS listeners, certificates are managed in SLB. To improve system performance, HTTPS is not used for data exchange (including health check data and business interaction data) between SLB and backend ECS instances.



The following section describes the health check process of a Layer 7 listener:

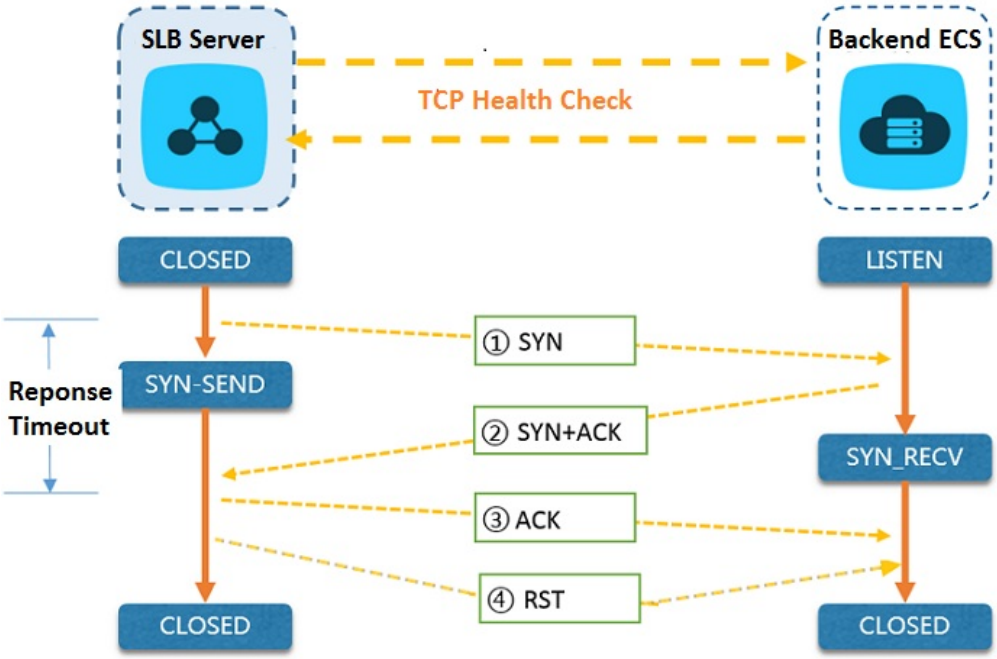
1. A Tengine node server sends an HTTP HEAD request that contains the configured domain name to the internal IP address, health check port, and health check path of a backend ECS instance based on health check settings.
2. After the backend ECS instance receives the request, the ECS instance returns an HTTP status code based on

the running status.

3. If the Tengine node server does not receive a response from the backend ECS instance within the specified response timeout period, the backend server is declared unhealthy.
4. If the Tengine node server receives a response from the backend ECS instance within the specified response timeout period, the node server compares the response with the configured status code. If the response contains the status code that indicates a healthy server, the backend server is declared healthy. Otherwise, the backend server is declared unhealthy.

Health checks of TCP listeners

For TCP listeners, SLB checks the status of backend servers by establishing TCP connections to improve health check efficiency. The following figure shows the process.



The following section describes the health check process of a TCP listener:

1. An LVS node server sends a TCP SYN packet to the internal IP address and health check port of a backend ECS instance.
2. After the backend ECS instance receives the request, the ECS instance returns an SYN-ACK packet if the corresponding port is listening normally.
3. If the LVS node server does not receive a packet from the backend ECS instance within the specified response timeout period, the backend ECS instance is declared unhealthy. Then, the node server sends an RST packet to the backend ECS instance to terminate the TCP connection.
4. If the LVS node server receives a packet from the backend ECS instance within the specified response timeout period, the node server determines that the service runs properly and the health check succeeds. Then, the node server sends an RST packet to the backend ECS instance to terminate the TCP connection.

Note A TCP three-way handshake is conducted to establish a TCP connection. After the LVS node server receives the SYN+ACK packet from the backend ECS instance, the node server sends an ACK packet, and then immediately sends an RST packet to terminate the TCP connection.

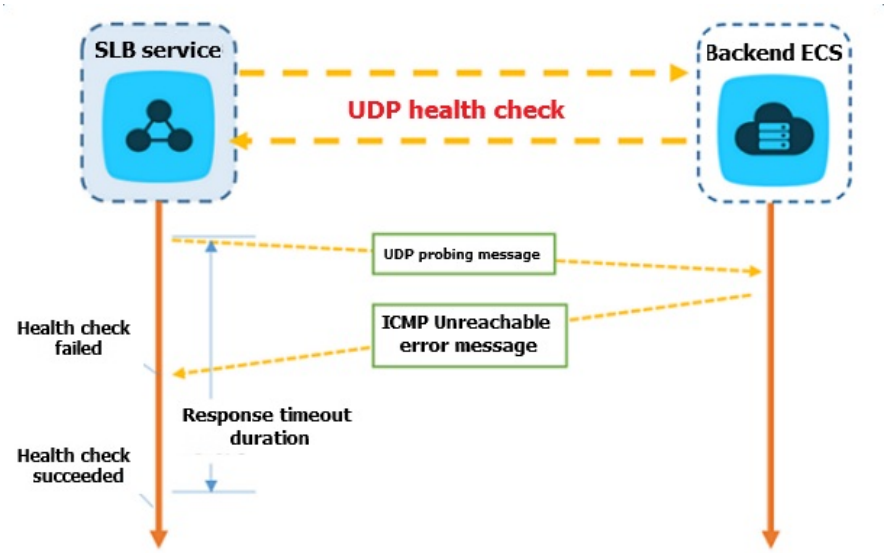
This process may cause backend ECS instances to think that an error such as an abnormal exit occurred in the TCP connection. Then, these instances may report a corresponding error message, such as `Connection reset by peer`, in logs such as Java connection pool logs.

Solution:

- You can implement HTTP health checks.
- If you have enabled the feature of obtaining actual client IP addresses on backend ECS instances, you can ignore connection errors caused by the access of the SLB CIDR block.

Health checks of UDP listeners

For UDP listeners, SLB checks the status of backend ECS instances by sending UDP packets. The following figure shows the process.



The following section describes the health check process of a UDP listener:

1. An LVS node server sends a UDP packet to the internal IP address and health check port of an ECS instance based on health check configurations.
2. If the corresponding port of the ECS instance is not listening normally, the system returns an ICMP error message, such as `port XX unreachable`. Otherwise, no message is returned.
3. If the LVS node server receives the ICMP error message within the response timeout period, the backend ECS instance is declared unhealthy.
4. If the LVS node server does not receive any messages from the backend ECS instance within the response timeout period, the ECS instance is declared healthy.

Note For UDP health checks, the health check result may not reflect the real status of a backend ECS instance in the following situation:

If the backend ECS instance uses a Linux operating system, the speed at which ICMP messages in high concurrency scenarios are sent is limited due to the ICMP attack prevention feature of Linux. In this case, even if a service exception occurs, SLB may declare the backend ECS instance healthy because the error message `port XX unreachable` is not returned. Consequently, the health check result deviates from the actual service status.

Solution:

You can specify a request and a response for UDP health checks. The ECS instance is considered healthy only when the specified response is returned. However, the client must be configured accordingly to return responses.

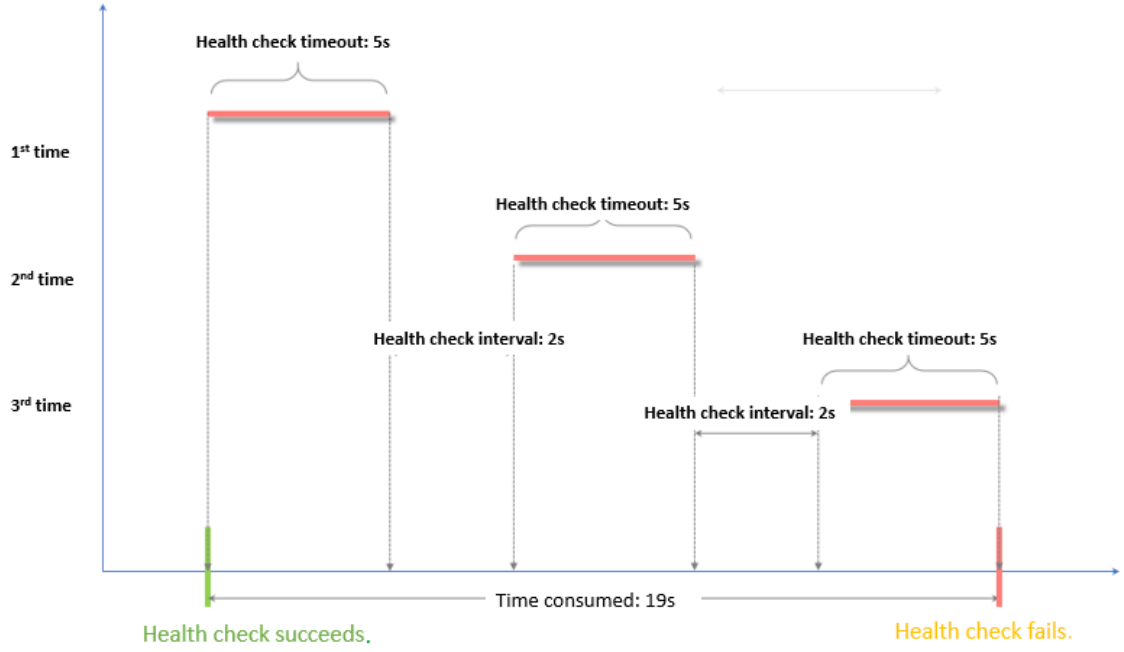
Health check time window

The health check feature effectively improves the availability of your services. However, to avoid impacts on system availability caused by frequent switching after failed health checks, the health check status switches only when health checks successively succeed or fail for a specified number of times within a certain time window. The health check time window is determined by the following factors:

- Health check interval: how often health checks are performed
- Response timeout: the length of time to wait for a response
- Health check threshold: the number of consecutive successes or failures of health checks

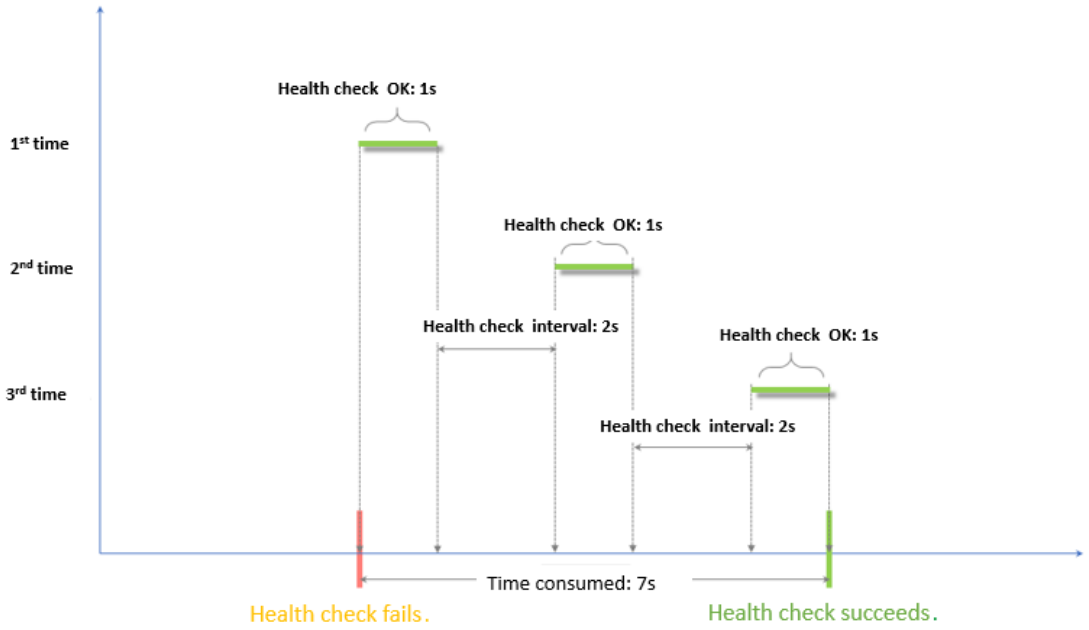
The health check time window is calculated based on the following formula:

- Time window for health check failures = Response timeout × Unhealthy threshold + Health check interval × (Unhealthy threshold - 1)



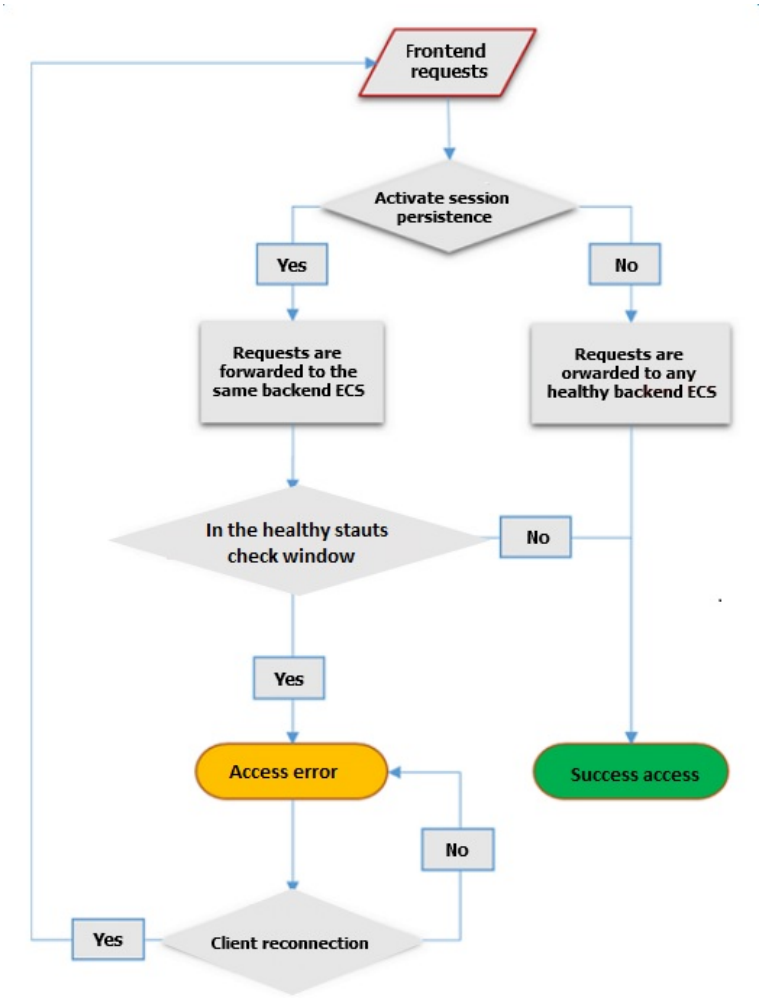
- Time window for health check successes = Response time of a successful health check × Healthy threshold + Health check interval × (Healthy threshold - 1)

> **Note** The response time of a successful health check is the duration from the time when the health check request is sent to the time when the response is received. When TCP health checks are used, the response time is short and almost negligible because only whether the specific port is alive is checked. For HTTP health checks, the response time depends on the performance and load of the application server and is typically within a few seconds.



The health check result has the following impacts on request forwarding:

- If the health check of the backend ECS instance fails, new requests are distributed to other backend ECS instances. This does not affect client access.
- If the health check of the backend ECS instance succeeds, new requests are distributed to this instance. The client access is normal.
- If an exception occurs on the backend ECS instance and a request arrives during a time window for health check failures, the request is still sent to the backend ECS instance. This is because the number of failed health checks has not reached the unhealthy threshold (3 times by default). In this case, the client access fails.



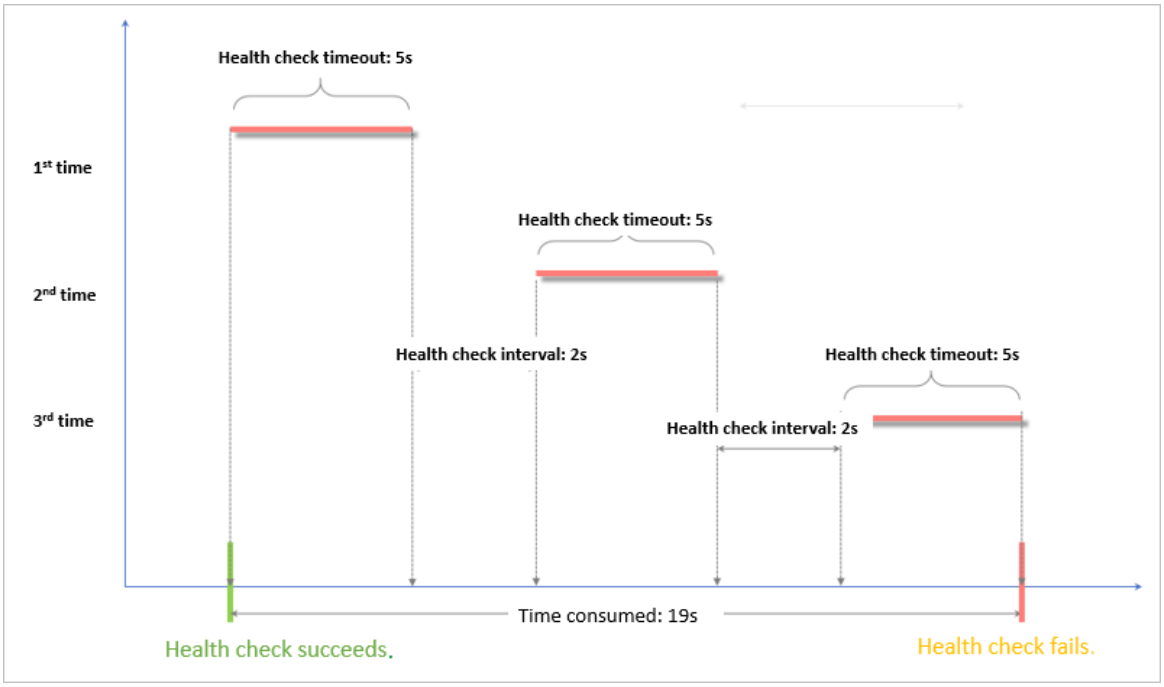
Examples of health check response timeout and health check interval

The following health check settings are used in these examples:

- **Response Timeout Period: 5 Seconds**
- **Health Check Interval: 2 Seconds**
- **Healthy Threshold: 3 Times**
- **Unhealthy Threshold: 3 Times**

Time window for health check failures = Response timeout × Unhealthy threshold + Health check interval × (Unhealthy threshold - 1). That is, 5 × 3 + 2 × (3 - 1) = 19 seconds. If the response time of a health check exceeds 19 seconds, the health check fails.

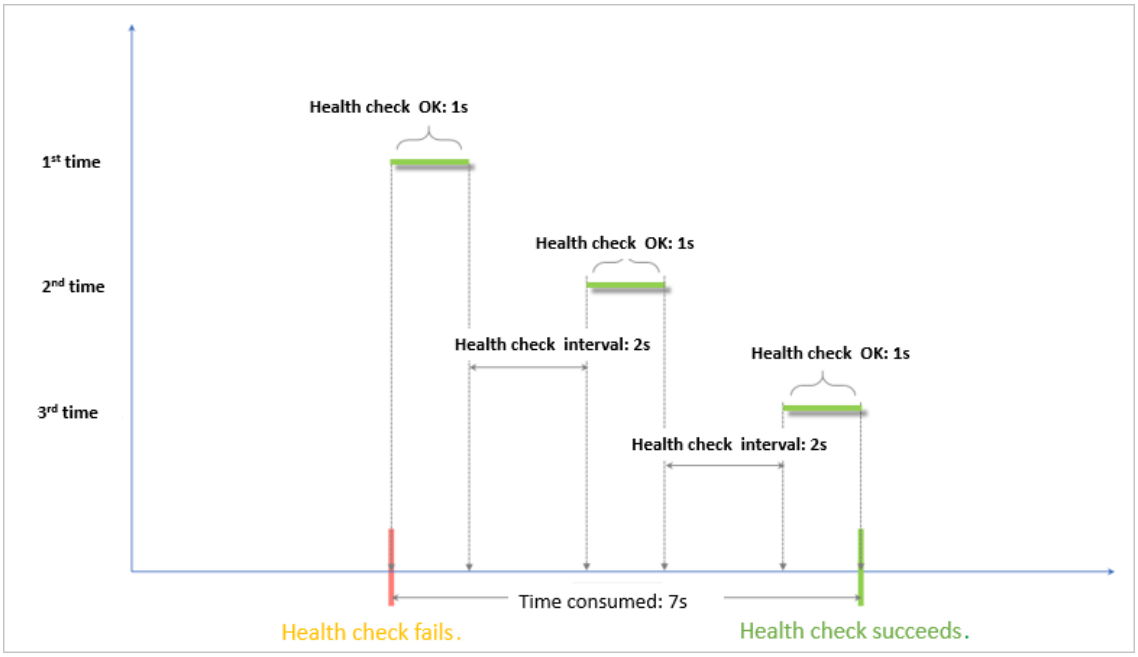
The following figure shows the time window from a healthy status to an unhealthy status.



Time window for health check successes = Response time of a successful health check × Healthy threshold + Health check interval × (Healthy threshold - 1). That is, (1 × 3) + 2 × (3 - 1) = 7 seconds. If the response time of a successful health check is less than seven seconds, the health check succeeds.

> **Note** The response time of a successful health check is the duration from the time when the health check request is sent to the time when the response is received. When TCP health checks are used, the response time is short and almost negligible because only whether the specific port is alive is checked. For HTTP health checks, the response time depends on the performance and load of the application server and is typically within a few seconds.

The following figure shows the time window from an unhealthy status to a healthy status (assume that it takes 1 second for the server to respond to a health check request).



Domain name setting in HTTP health checks

When HTTP health checks are used, you can set a domain name for health checks. The setting is optional. Some application servers verify the host field in requests. In this case, the request header must contain the host field. If a domain name is configured in health check setting, SLB adds the domain name to the host field when SLB forwards a request to an application server. If no domain name is configured, the health check request is denied by the application server because it does not contain a host field and the health check may fail. If your application server verifies the host field in requests, you must configure a domain name to make sure that the health check feature works.

17.1.7.2. Configure health checks

This topic describes how to configure health checks. You can configure health checks when you create a listener or for an existing listener. The default health check settings can meet your requirements in most cases.

Procedure

1. Log on to the SLB console.
2. Find an SLB instance and click the instance ID.
3. On the page that appears, click the **Listener** tab.
4. Click **Add Listener**, or find an existing listener and click **Modify Listener** in the **Actions** column.
5. Click **Next** to go to the **Health Check** step and configure the health check.

   We recommend that you use the default settings when you configure health checks.

   Health check parameters

Parameter	Description
Health Check Protocol	Select the protocol that the SLB instance uses when it performs health checks. For TCP listeners, both TCP health checks and HTTP health checks are supported. ◦ A TCP health check implements detection at the network layer by sending SYN packets to check whether a port is open. ◦ An HTTP health check verifies the health of a backend server by sending HEAD or GET requests to simulate browser access.
Health Check Method (for the HTTP and HTTPS health checks only)	Health checks of Layer 7 (HTTP or HTTPS) listeners support both the HEAD and GET methods. The HEAD method is used by default. If your backend application server does not support the HEAD method or if the HEAD method is disabled, the health check may fail. To solve this issue, you can use the GET method instead. If the GET method is used and the response size exceeds 8 KB, the response is truncated. However, the health check result is not affected.

Parameter	Description
Health Check Path and **Health Check Domain Name (Optional)** (for the HTTP health checks only)	By default, SLB sends HTTP HEAD requests to the default homepage configured on the application server through the internal IP address of the backend ECS instance to perform health checks. If you do not use the default homepage of the application server for health checks, you must specify the path for health checks. Some application servers verify the host field in requests. In this case, the request header must contain the host field. If a domain name is configured in health check settings, SLB adds this domain name to the host field when SLB forwards a health check request to one of the preceding application servers. If no domain name is configured, SLB does not include the host field in requests and the requests are rejected by the application server, which may cause health checks to fail. If your application server verifies the host field in requests, you must configure a domain name in health check settings to ensure that the health check feature functions properly.
Normal Status Code (for the HTTP health checks only)	Select the HTTP status code that indicates successful health checks. Default values: http_2xx and http_3xx.
Health Check Port	The detection port used by the health check feature to access backend servers. By default, the backend port configured for the listener is used. **Note** If a VServer group or a primary/secondary server group is configured for the listener, and the ECS instances in the group use different ports, leave this parameter empty. SLB uses the backend port of each ECS instance to perform health checks.
Response Timeout	The length of time to wait for a health check response. If the backend ECS instance does not send an expected response within the specified period of time, the health check fails. Valid values: 1 to 300. Unit: seconds. Default value for UDP listeners: 10. Default value for HTTP, HTTPS, and TCP listeners: 5.
Health Check Interval	The interval between two consecutive health checks. All nodes in the LVS cluster perform health checks independently and in parallel on backend ECS instances at the specified interval. The health check statistics of a single ECS instance cannot reflect the health check interval because the nodes perform health checks at different times. Valid values: 1 to 50. Unit: seconds. Default value for UDP listeners: 5. Default value for HTTP, HTTPS, and TCP listeners: 2.

Parameter	Description
Unhealthy Threshold	The number of consecutive failed health checks that must occur on a backend ECS instance before this ECS instance is declared unhealthy. Valid values: 2 to 10. Default value: 3.

6. Click **Next**.

17.1.7.3. Disable the health check feature

This topic describes how to disable the health check feature for a Classic Load Balancer (CLB) instance. If you disable the health check feature, requests may be distributed to unhealthy backend Elastic Compute Service (ECS) instances by . This causes service disruptions. We recommend that you enable the health check feature.

Procedure

1. Log on to the SLB console.
2. On the **Instances** page, click the ID of the instance that you want to manage.
3. On the **Listener** tab, find the listener for which you want to disable the health check feature and click **Modify Listener** in the **Actions** column.
4. On the **Configure Listener** page, click **Next** to proceed to the **Health Check** wizard page.
5. Turn off the Enable Health Check switch and click **Next**.
6. Click **Submit** and click **OK**.

17.1.8. Certificate management

17.1.8.1. Certificate overview

This topic provides an overview of the certificates that can be deployed on SLB instances. To use an HTTPS listener, you must upload the required third-party server certificate and digital identification issued by a certificate authority (CA) to SLB. You do not need to configure certificates on backend servers after uploading the certificates to SLB.

To upload a third-party certificate, you must have the files that contain the public key and private key of the certificate.

HTTPS server certificates and client CA certificates are supported.

You can create a maximum of 100 certificates per account.

17.1.8.2. Certificate requirements

Server Load Balancer (SLB) supports only certificates in the PEM format. Before you upload a certificate, make sure that the certificate content, certificate chain, and private key meet the corresponding format requirements.

Certificates issued by a root CA

If the certificate was issued by a root certification authority (CA), the received certificate is the only one that needs to be uploaded to SLB. In this case, the website that is configured with this certificate is regarded as a trusted website and does not require additional certificates.

The certificate must meet the following format requirements:

- The certificate must start with `-----BEGIN CERTIFICATE-----` and end with -----END CERTIFICATE-----.
- Each line (except the last line) must contain 64 characters. The last line can contain 64 or fewer characters.
- The certificate content cannot contain spaces.

Certificates issued by an intermediate CA

If the certificate was issued by an intermediate CA, the received certificate file contains multiple certificates. You must upload both the server certificate and the required intermediate certificates to SLB.

The format of the certificate chain must meet the following requirements:

- The server certificate must be put first and the content of the one or more required intermediate certificates must be put underneath without blank lines between the certificates.
- The certificate content cannot contain spaces.
- Blank lines are not allowed between the certificates. Each line must contain 64 characters. For more information, see RFC1421.
- Certificates must meet the corresponding format requirements. In most cases, the intermediate CA provides instructions about the certificate format when certificates are issued. The certificates must meet the format requirements.

The following section provides a sample certificate chain:

```
-----BEGIN CERTIFICATE-----
-----END CERTIFICATE-----
-----BEGIN CERTIFICATE-----
-----END CERTIFICATE-----
-----BEGIN CERTIFICATE-----
-----END CERTIFICATE-----
```

Public keys of certificates

SLB supports the following public key algorithms:

- RSA 1024
- RSA 2048
- RSA 4096
- ECDSA P-256
- ECDSA P-384
- ECDSA P-521

RSA private keys

When you upload a server certificate, you must upload the private key of the certificate.

An RSA private key must meet the following format requirements:

- The private key must start with `-----BEGIN RSA PRIVATE KEY-----` and end with -----END RSA PRIVATE KEY-----, and these parts must also be uploaded.
- Blank lines are not allowed in the content. Each line (except the last line) must contain 64 characters. The last line can contain 64 or fewer characters. For more information, see RFC1421.

You may use an encrypted private key. For example, the private key starts with `----BEGIN PRIVATE KEY-----` and ends with -----END PRIVATE KEY-----, or starts with `-----BEGIN ENCRYPTED PRIVATE KEY-----` and ends with -----END ENCRYPTED PRIVATE KEY-----. The private key may also contain `Proc-Type: 4,ENCRYPTED`. In this case, you must first run the following command to convert the private key:

```
openssl rsa -in old_server_key.pem -out new_server_key.pem
```

17.1.8.3. Upload certificates

This topic describes how to create and upload certificates to Server Load Balancer (SLB). Before you create an HTTPS listener, you must upload the required server certificate and CA certificate to SLB. You do not need to configure certificates on backend servers after you upload the certificates to SLB.

Prerequisites

- A server certificate is purchased.
- A CA certificate and a client certificate are generated.

Context

Note that you can create up to 100 certificates with each account.

Procedure

1. In the left-side navigation pane, click **Certificates**.
2. On the Certificates page, click **Create Certificate**.
3. In the **Create Certificate** panel, set the required parameters and click **Create**.

Parameter	Description
Certificate Name	Enter a name for the certificate. The name must be 1 to 80 characters in length, and can contain only letters, digits, hyphens (-), forward slashes (/), periods (.), underscores (_), and asterisks (*).
Organization	The organization to which the certificate belongs.
Resource Group	The resource set to which the certificate belongs.
Certificate Standard	Select the type of certificate. You can select **International Standard Certificate** or **National Standard Certificate**.
Public Key Certificate	The content of the server certificate. Paste the content into the editor. Click **Example** to view the valid certificate formats. For more information, see Certificate requirements.
Private Key	The private key of the server certificate. Paste the private key into the editor. Click **Example** to view the valid certificate formats. For more information, see Certificate requirements. **Notice** A private key is required only when you upload a server certificate.
Region	The region where you want to deploy the certificate.

4. Click **Create**.

17.1.8.4. Generate a CA certificate

When you configure an HTTPS listener, you can use a self-signed CA certificate. This topic describes how to generate a CA certificate and use the CA certificate to sign a client certificate.

Generate a CA certificate by using Open SSL

1. Run the following commands to create a *ca* folder in the */root* directory and then create four subfolders under the *ca* folder.

```
sudo mkdir ca
cd ca
sudo mkdir newcerts private conf server
```

- *newcerts* is used to store the digital certificate signed by the CA certificate.
- *private* is used to store the private Key of the CA certificate.
- *conf* is used to store the configuration files used for simplifying parameters.
- *server* is used to store the server certificate.

2. Create an *openssl.conf* file that contains the following information in the *conf* directory.

```
[ ca ]
default_ca = foo
[ foo ]
dir = /root/ca
database = /root/ca/index.txt
new_certs_dir = /root/ca/newcerts
certificate = /root/ca/private/ca.crt
serial = /root/ca/serial
private_key = /root/ca/private/ca.key
RANDFILE = /root/ca/private/.rand
default_days = 365
default_crl_days= 30
default_md = md5
unique_subject = no
policy = policy_any
[ policy_any ]
countryName = match
stateOrProvinceName = match
organizationName = match
organizationalUnitName = match
localityName = optional
commonName      = supplied
emailAddress    = optional
```

3. Run the following command to generate a private Key.

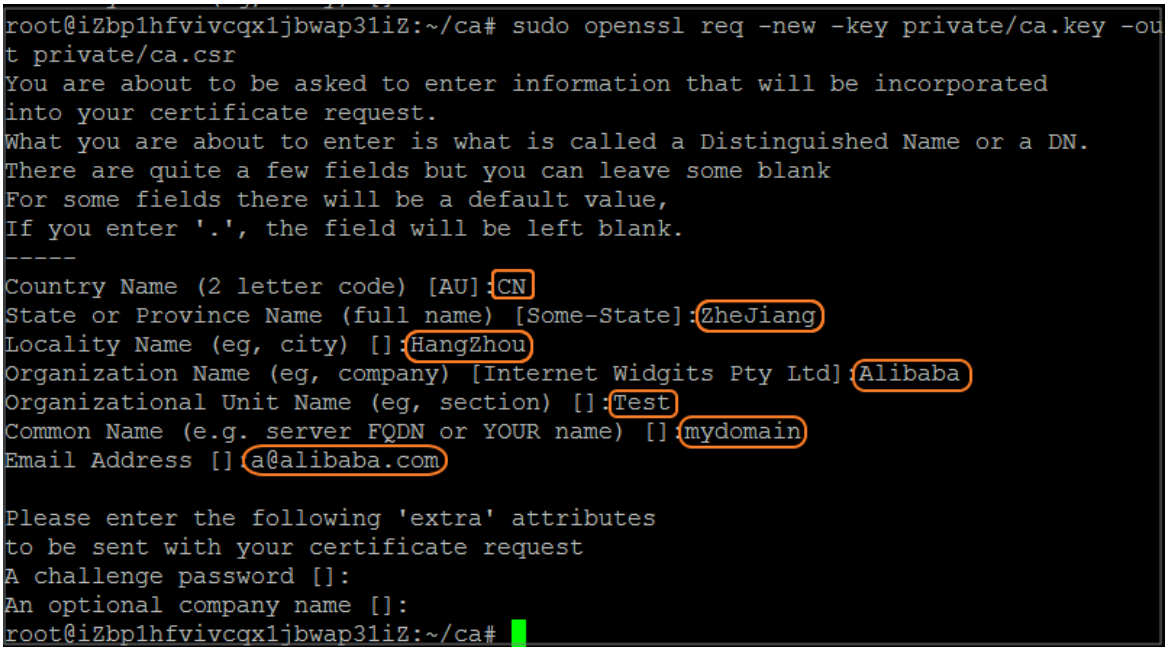
```
cd /root/ca
sudo openssl genrsa -out private/ca.key
```

The following figure is an example of the key generation.

```
root@iZbp1hfvivcqx1jbwap31iZ:~/ca/conf# cd /root/ca
root@iZbp1hfvivcqx1jbwap31iZ:~/ca# sudo openssl genrsa -out private/ca.key
Generating RSA private key, 2048 bit long modulus
.....................................................................
.....+++
..+++
e is 65537 (0x10001)
```

4. Run the following command and input the required information according to the prompts. Press Enter to generate a *csr* file.

```
sudo openssl req -new -key private/ca.key -out private/ca.csr
```



> **Note**
>
> Common Name is the domain name of the SLB instance.

5. Run the following command to generate a *crt* file:

```
sudo openssl x509 -req -days 365 -in private/ca.csr -signkey private/ca.key -out private/ca.crt
```

6. Run the following command to set the start sequence number for the private Key, which can be any four characters.

```
sudo echo FACE > serial
```

7. Run the following command to create a CA Key library:

```
sudo touch index.txt
```

8. Run the following command to create a certificate revocation list for removing the client certificate:

```
sudo openssl ca -gencrl -out /root/ca/private/ca.crl -crldays 7 -config "/root/ca/conf/openssl.con
f"
```

The output is:

```
Using configuration from /root/ca/conf/openssl.conf
```

Sign the client certificate

1. Run the following command to generate a *users* folder under the *ca* directory to store the client Key.

```
sudo mkdir users
```

2. Run the following command to create a Key for the client certificate:

```
sudo openssl genrsa -des3 -out /root/ca/users/client.key 1024
```

> **Note**
>
> Enter a pass phrase when creating the Key. It is the password to protect the private Key from unauthorized access. Enter the same password twice.

3. Run the following command to create a *csr* file for the client Key.

```
sudo openssl req -new -key /root/ca/users/client.key -out /root/ca/users/client.csr
```

Enter the pass phrase set in the previous step and other required information when prompted.

> **Note**
>
> A challenge password is the password of the client certificate. Note that it is not the password of the client Key.

4. Run the following command to sign the client Key.

```
sudo openssl ca -in /root/ca/users/client.csr -cert /root/ca/private/ca.crt -keyfile /root/ca/private/ca.key -out /root/ca/users/client.crt -config "/root/ca/conf/openssl.conf"
```

Enter *y* twice when prompted to confirm the operation.

```
root@iZbp1hfvivcqx1jbwap31iZ:~/ca# sudo openssl ca -in /root/ca/users/client.csr
 -cert /root/ca/private/ca.crt -keyfile /root/ca/private/ca.key -out /root/ca/us
ers/client.crt -config "/root/ca/conf/openssl.conf"
Using configuration from /root/ca/conf/openssl.conf
Check that the request matches the signature
Signature ok
The Subject's Distinguished Name is as follows
countryName           :PRINTABLE:'CN'
stateOrProvinceName   :ASN.1 12:'ZheJiang'
localityName          :ASN.1 12:'HangZhou'
organizationName      :ASN.1 12:'Alibaba'
organizationalUnitName:ASN.1 12:'Test'
commonName            :ASN.1 12:'mydomain'
emailAddress          :IA5STRING:'a@alibaba.com'
Certificate is to be certified until Jun  4 15:28:55 2018 GMT (365 days)
Sign the certificate? [y/n]:y


1 out of 1 certificate requests certified, commit? [y/n]y
Write out database with 1 new entries
Data Base Updated
root@iZbp1hfvivcqx1jbwap31iZ:~/ca#
```

5. Run the following command to convert the certificate to a *PKCS12* file.

```
sudo openssl pkcs12 -export -clcerts -in /root/ca/users/client.crt -inkey /root/ca/users/client.key -out /root/ca/users/client.p12
```

Follow the prompts to enter the pass phrase of client Key. Then enter the password used for exporting the client certificate. This password is used to protect the client certificate, which is required when you install the client certificate.

6. Run the following commands to view the generated client certificate:

```
cd users
ls
```

17.1.8.5. Convert the certificate format

Server Load Balancer (SLB) supports PEM certificates only. Certificates in other formats must be converted to the PEM format before they can be uploaded to SLB. We recommend that you use Open SSL for conversion.

Convert DER to PEM

DER: This format is usually used on a Java platform. The certificate file suffix is generally *. der*, *. cer*, or *. crt*.

- Run the following command to convert the certificate format:

```
openssl x509 -inform der -in certificate.cer -out certificate.pem
```

- Run the following command to convert the private key:

```
openssl rsa -inform DER -outform PEM -in privatekey.der -out privatekey.pem
```

Convert P7B to PEM

P7B: This format is usually used in a Windows server and Tomcat.

Run the following command to convert the certificate format:

```
openssl pkcs7 -print_certs -in incertificate.p7b -out outcertificate.cer
```

Convert PFX to PEM

PFX: This format is usually used in a Windows server.

- Run the following command to extract the certificate:

```
openssl pkcs12 -in certname.pfx -nokeys -out cert.pem
```

- Run the following command to extract the private key:

```
openssl pkcs12 -in certname.pfx -nocerts -out key.pem -nodes
```

17.1.8.6. Replace a certificate

This topic describes how to replace a certificate with a new certificate. We recommend that you replace certificates before they expire to avoid service interruption.

Procedure

1. Log on to the SLB console.
2. Click the ID of the Server Load Balancer (SLB) instance for which you want to replace the certificate and select the **Listener** tab.
3. Find the HTTPS listener for which you want to replace the certificate and click **Manage Certificate** in the **Actions** column.
4. On the **Manage Certificate** page, select Add Server Certificate.

   You can also select **Create Server Certificate** or **Purchase Certificate**. For more information about how to create a server certificate, see Overview.
5. On the **Advanced Settings** tab, click **Modify** and select whether to enable mutual authentication and TLS security policy. For more information, see Add an HTTPS listener.
6. Click **OK**.
7. On the **Certificates** page, find the expired certificate and click **Delete**.
8. In the message that appears, click **OK**.

> **Note** If the certificate is associated with another listener, it cannot be deleted.

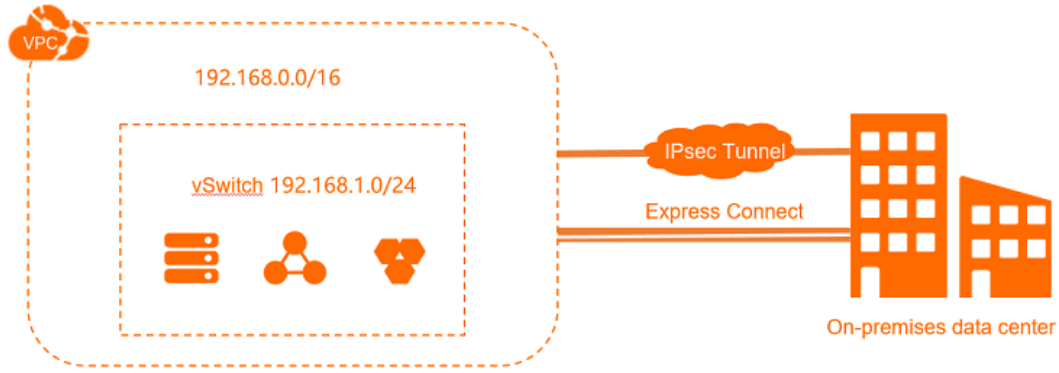
18.Virtual Private Cloud (VPC)

18.1. User Guide

18.1.1. What is a VPC?

A virtual private cloud (VPC) is a private network dedicated for your use. You have full control over your VPC. For example, you can specify the CIDR block and configure route tables and gateways. In a VPC, you can deploy Apsara Stack resources, such as Elastic Compute Service (ECS) instances, ApsaraDB RDS instances, and Server Load Balancer (SLB) instances.

Furthermore, you can connect your VPC to other VPCs or on-premises networks to create a custom network environment. This way, you can migrate applications to the cloud and extend data centers.



Components

Each Virtual Private Cloud consists of at least one private network segment, one router, and at least one switch.

- Private CIDR blocks

  When you create a VPC and a vSwitch, you must specify the private IP address range for the VPC in CIDR notation.

  You can use the standard private CIDR blocks listed in the following table and their subsets as CIDR blocks for your VPCs. For more information, see The network planning section in *User Guide*.

CIDR blocks	Number of available private IP addresses (system reserved ones excluded)
192.168.0.0/16	65,532
172.16.0.0/12	1,048,572
10.0.0.0/8	16,777,212

- VRouter

  A VRouter is a hub that connects all VSwitches in a VPC and serves as a gateway between the VPC and other networks. After a VPC is created, the system creates a VRouter for the VPC. Each vRouter is associated with a route table.

  For more information, see the Route table overview topic in *User Guide*.

- vSwitches

A vSwitch is a basic network component that connects different cloud resources in a VPC. After you create a VPC, you can create a vSwitch to divide your VPC into multiple subnets. vSwitches deployed in a VPC can communicate with each other over the private network. You can deploy your applications in vSwitches that belong to different zones to improve service availability.

For more information, see the Create a vSwitch topic in *User Guide*.

18.1.2. Log on to the VPC console

This topic describes how to log on to the Virtual Private Cloud (VPC) console of Apsara Uni-manager by using the Google Chrome browser.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Networking > Virtual Private Cloud**.

18.1.3. Quick start

18.1.3.1. Design networks

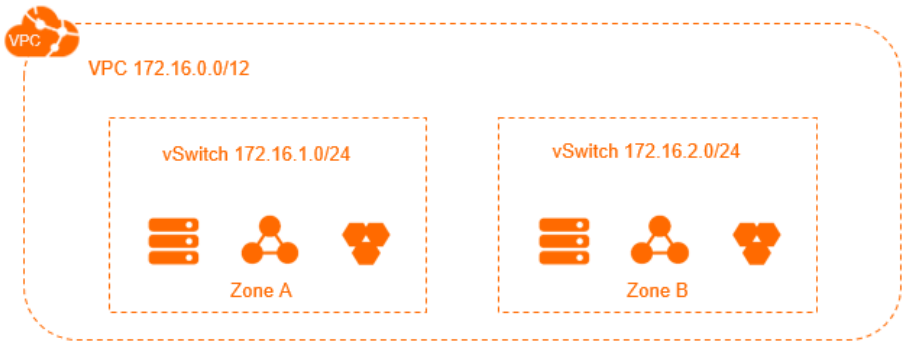
Before you create virtual private clouds (VPCs) and vSwitches, you must decide the quantities and CIDR blocks of VPCs and vSwitches.

- How many VPCs do I use?
- How many vSwitches do I use?
- How do I specify CIDR blocks?
- How do I specify CIDR blocks if I want to connect a VPC to another VPC or a data center?

How many VPCs do I use?

- One VPC

  If you do not want to deploy your business systems across regions or isolate different business systems, we recommend that you use one VPC.

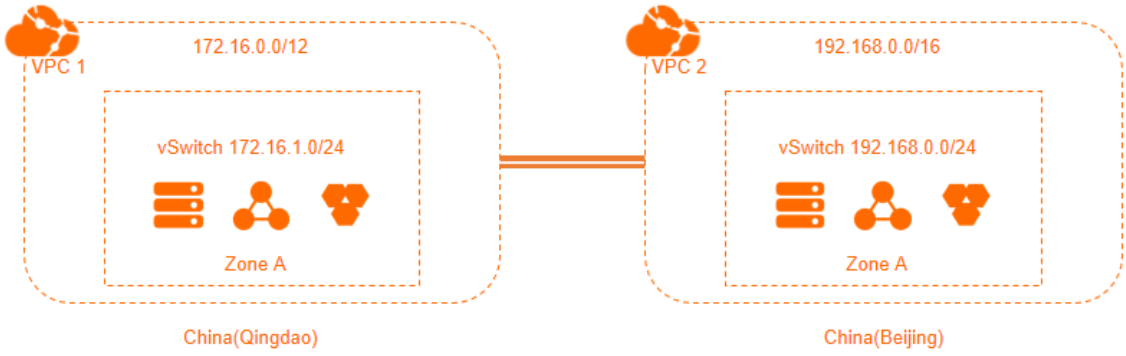


- Multiple VPCs

  In the following scenarios, we recommend that you use multiple VPCs:

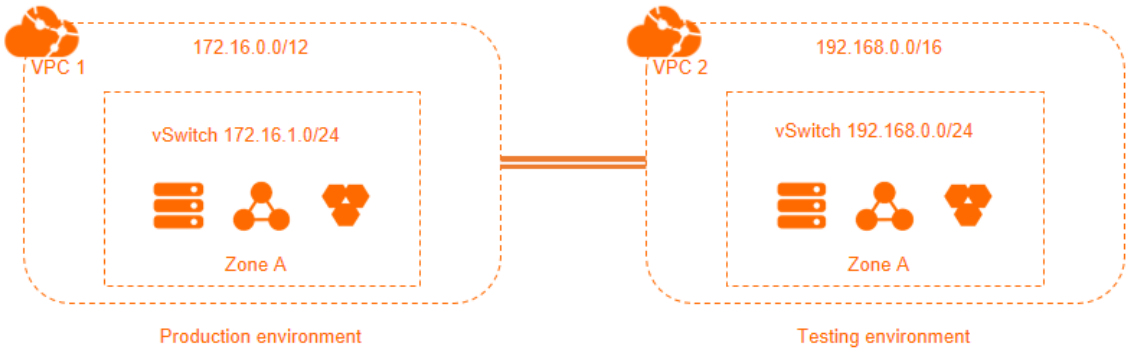
  - Cross-region deployment

    A VPC cannot be deployed across regions. If you want to deploy your business system across regions, you must create VPCs in these regions. Then, you can use services such as Express Connect and VPN Gateway to connect the VPCs.



- Business system isolation

  If you want to isolate business systems deployed in the same region, you must create multiple VPCs. The following figure shows how the test environment is isolated from the production environment by using VPCs.



How many vSwitches do I use?

We recommend that you create at least two vSwitches in different zones for a VPC. This implements cross-zone disaster recovery.

The network latency between different zones in a region is low. However. you still need to verify the network latency in your actual business system. This is because the network latency may be higher than expected due to complex data processing or cross-zone data transmission. We recommend that you optimize and adjust your business system to strike a balance between availability and latency.

In addition, the size and design of your business system must also be taken into consideration when you create vSwitches. If your frontend systems are publicly accessible and your backend systems require Internet access, you can deploy the frontend systems in different vSwitches for redundancy. Then, deploy your backend systems in the remaining vSwitches.

How do I specify CIDR blocks?

When you create VPCs and vSwitches, you must specify their private IP addresses in CIDR block notation.

- Specify VPC CIDR blocks

  You can specify 192.168.0.0/16, 172.16.0.0/12, 10.0.0.0/8, or their subsets as the CIDR blocks of your VPCs. When you specify VPC CIDR blocks, take note of the following rules:

  - If you have only one VPC and the VPC does not need to communicate with a data center, you can specify one of the preceding CIDR blocks or their subsets as the VPC CIDR block.
  - If you have multiple VPCs, or you want to set up a hybrid cloud environment between a VPC and your data center, we recommend that you specify the subsets of the preceding CIDR blocks for your VPCs. In this case, we recommend that you set the subnet mask length to 16 bits or less.

- Specify vSwitch CIDR blocks

  The CIDR block of a vSwitch must be a subset of the CIDR block of the VPC to which the vSwitch belongs. For example, if the CIDR block of a VPC is 192.168.0.0/16, the CDIR block of a vSwitch that belongs to the VPC can range from 192.168.0.0/17 to 192.168.0.0/29.

  When you specify vSwitch CIDR blocks, take note of the following rules:

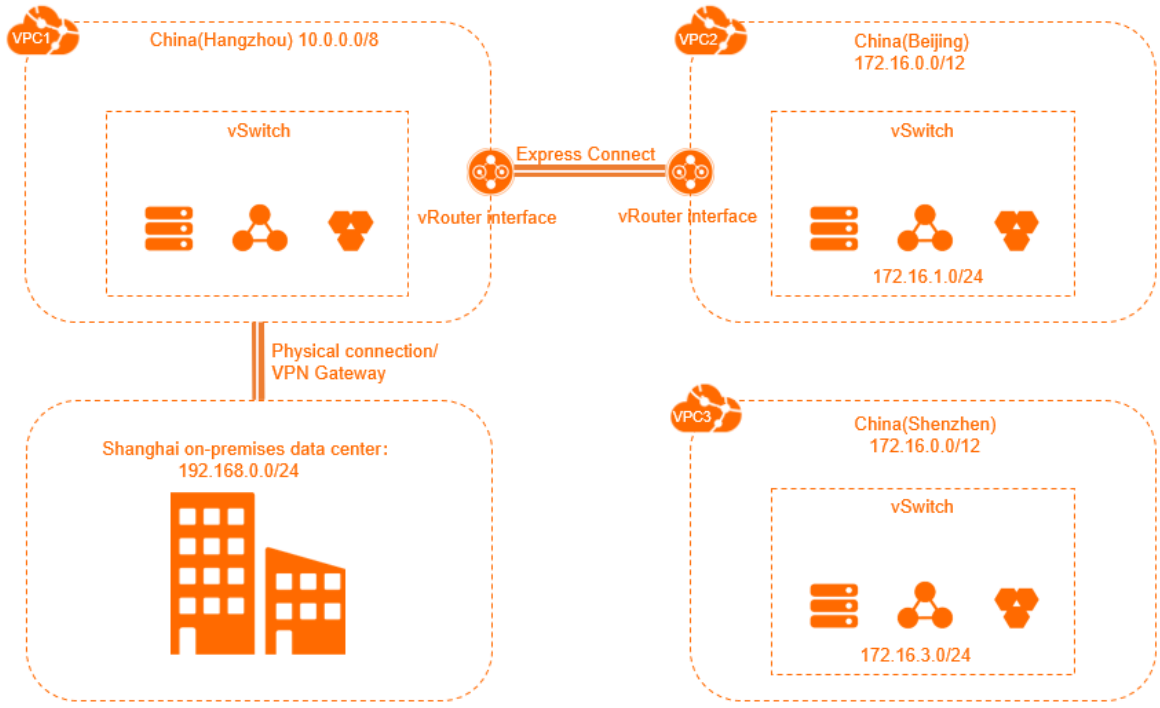
  - The subnet mask of a vSwitch CIDR block must be 16 to 29 bits in length. This allows the vSwitch to provide 8 to 65,536 IP addresses. A 16-bit subnet mask provides IP addresses for 65,532 Elastic Compute Service (ECS) instances, which can meet your needs in most cases. A subnet mask of more than 29 bits in length provides few IP addresses.
  - The first IP address and the last three IP addresses of each vSwitch CIDR block are reserved for the system. For example, if the CIDR block of a vSwitch is 192.168.1.0/24, the IP addresses 192.168.1.0, 192.168.1.253, 192.168.1.254, and 192.168.1.255 are reserved.

- When you specify the CIDR block of a vSwitch, you must also take into consideration the number of ECS instances that you want to attach to the vSwitch.

How do I specify CIDR blocks if I want to connect a VPC to another VPC or a data center?

Before you connect your VPC to another VPC or a data center, you must make sure that the VPC CIDR block does not conflict with that of the peer network.

For example, you have three VPCs: VPC1 in the China (Hangzhou) region, VPC2 in the China (Beijing) region, and VPC3 in the China (Shenzhen) region, as shown in the following figure. An Express Connect circuit is used to connect VPC1 and VPC2. VPC3 does not communicate with other VPCs, but may need to communicate with VPC2 in the future. Additionally, you have a data center located in Shanghai, and you need to connect it to VPC1 by using an Express Connect circuit.



In this example, the CIDR block of VPC2 is different from the CIDR block of VPC1, but is the same with the CIDR block of VPC3. Take into consideration that VPC2 and VPC3 need to communicate with each other through private connections, the vSwitches in these VPCs are assigned different CIDR blocks. The vSwitch CIDR blocks of VPCs that need to communicate with each other must be different. However, these VPCs can have the same CIDR block.

When you specify CIDR blocks for multiple VPCs that need to communicate with each other, take note of the following rules:

- The preferred practice is to specify different CIDR blocks for different VPCs. You can use the subsets of standard CIDR blocks if the given CIDR blocks are insufficient.
- If you cannot assign different CIDR blocks to the VPCs, try to specify different CIDR blocks for vSwitches in these VPCs.
- If you cannot assign different CIDR blocks to all of the vSwitches in these VPCs, make sure that different CIDR blocks are specified for the vSwitches that need to communicate with each other.

18.1.3.2. Create an IPv4 VPC

This topic describes how to create a virtual private cloud (VPC) with an IPv4 CIDR block and how to create an Elastic Compute Service (ECS) instance in the VPC.

Prerequisites

To deploy cloud resources in a VPC, you must first plan the network. For more information, see Design networks.

Step 1: Create a VPC

1. Log on to the VPC console.
2. On the **VPCs** page, click **Create VPC**.
3. On the **Create VPC** page, set the parameters as described in the following table and click **Submit**.

Parameter	Description
Organization	Select the organization to which the VPC belongs.
Resource Set	Select the resource set to which the VPC belongs.
Region	Select the region where you want to deploy the VPC.
Sharing Scope	Select the sharing scope of the VPC. ○ **Current Resource Set**: Only the administrator of the current resource set can create resources for the shared VPC. ○ **Current Organization and Subordinate Organizations**: Only the administrators of the current organization and its subordinate organizations can use the VPC to create resources. ○ **Current Organization**: Only the administrator of the current organization can use the VPC to create resources. **Current Resource Set** is selected in this example.
VPC Name	Enter a name for the VPC. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`. **VPCtest** is used in this example.
IPv4 CIDR Block	Specify an IPv4 CIDR block for the VPC. The following settings are supported: ○ **Recommended CIDR Block**: You can select the following standard IPv4 CIDR blocks: 192.168.0.0/16 and 172.16.0.0/16. ○ **Custom CIDR Block**: You can use 192.168.0.0/16, 172.16.0.0/16, or a subset of the preceding CIDR blocks. The subnet mask must be 8 to 28 bits in length. Example: 192.168.0.0/16. In this example, 192.168.0.0/16 is used as the IPv4 CIDR block of the VPC. **Note** After a VPC is created, you cannot change its IPv4 CIDR block.
IPv6 CIDR Block	Specify whether to assign an IPv6 CIDR block. ○ **Do Not Assign**: The system does not assign an IPv6 CIDR block to the VPC. ○ **Assign**: The system automatically assigns an IPv6 CIDR block to the VPC. **Do Not Assign** is selected in this example.

Parameter	Description
Description	Enter a description for the VPC. The description must be 2 to 256 characters in length, and can contain letters, digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.

4. Click **Back to Console**. On the **VPCs** page, you can view the VPCs that are created.

Step 2: Create a vSwitch

1. In the left-side navigation pane, click **vSwitch**.
2. On the **vSwitch** page, click **Create vSwitch**.
3. On the **vSwitch** page, set the parameters as described in the following table and click **Submit**.

Parameter	Description
Organization	Select the organization to which the vSwitch belongs.
Resource Set	The resource set to which the vSwitch belongs.
Region	Select the region where you want to deploy the vSwitch.
Zone	Select the zone where you want to deploy the vSwitch. In a VPC, a vSwitch can be deployed only in one zone. You cannot deploy a vSwitch across zones. However, you can deploy cloud resources in vSwitches that belong to different zones to achieve cross-zone disaster recovery. **Note** An instance can be attached only to one vSwitch.
Sharing Scope	Select the sharing scope of the vSwitch. ○ **Current Resource Set**: Only the administrator of the current resource set can create resources in the shared vSwitch. ○ **Current Organization and Subordinate Organizations**: Only the administrators of the current organization and its subordinate organizations can create resources in the shared vSwitch. ○ **Current Organization**: Only the administrator of the current organization can use the vSwitch to create resources. **Current Resource Set** is selected in this example.
VPC	Select the VPC in which you want to create the vSwitch. VPCtest is selected in this example.
Dedicated for Out-of-cloud Physical Machines	Specify whether the vSwitch is dedicated for bare-metal servers. For more information about bare-metal servers, see the **Bare-metal servers in VPCs** topic in *Bare-metal Server Management Service User Guide*. **No** is selected in this example.

Parameter	Description
vSwitch Name	Enter a name for the vSwitch. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`.
IPv4 CIDR Block	Specify an IPv4 CIDR block for the vSwitch. The default IPv4 CIDR block is used in this example.
IPv6 CIDR Block	Specify an IPv6 CIDR block for the vSwitch. **Do Not Assign** is selected in this example.
Description	Enter a description for the vSwitch. The description must be 2 to 256 characters in length, and can contain letters, digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.

Step 3: Create a security group

1. In the top navigation bar, choose **Products > Elastic Computing > Elastic Compute Service**.
2. Choose **Networks and Security > Security Groups**.
3. On the **Security Groups** page, click **Create Security Group**.
4. On the **Create Security Group** page, set the parameters as described in the following table and click **Submit**.

Parameter	Description
Organization	Select the organization to which the security group belongs.
Resource Set	Select the resource set to which the security group belongs.
Region	Select the region where you want to deploy the security group. Make sure that the security group and the VPC belong to the same region.
Zone	Select the zone where you want to deploy the security group.
Sharing Scope	Select the sharing scope of the security group. ○ **Current Resource Set**: Only the administrator of the current resource set can create resources for the security group. ○ **Current Organization and Subordinate Organizations**: Only the administrators of the current organization and its subordinate organizations can create resources for the security group. ○ **Current Organization**: Only the administrator of the current organization can create resources for the security group. **Current Resource Set** is selected in this example.
VPC	Select the VPC to which the security group belongs.

Parameter	Description
Security Group Name	Enter a name for the security group. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`.
Description	Enter a description for the security group. The description must be 2 to 256 characters in length, and can contain letters, digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.

Step 4: Create an ECS instance

1. In the top navigation bar, choose **Products > Networking > Virtual Private Cloud**.
2. In the left-side navigation pane, click **vSwitch**.
3. In the top navigation bar, select the region where the vSwitch is deployed.
4. On the **vSwitch** page, find the vSwitch that you have created and choose **Create > ECS Instance** in the **Actions** column.
5. On the **Create ECS Instance** page, set the parameters and click **Submit**.

   For more information about how to configure ECS instances, see **Create an instance** in the **Quick start** of the *ECS user guide*.

18.1.3.3. Create an IPv6 VPC

This topic describes how to create a virtual private cloud (VPC) that supports IPv6 CIDR blocks and then create an Elastic Compute Service (ECS) instance that is assigned an IPv6 address in the VPC to access IPv6 services.

Step 1: Create a VPC and a vSwitch

Before you deploy cloud resources in a VPC, you must create a VPC and a vSwitch.

Perform the following steps to create a VPC and a vSwitch:

1. Log on to the VPC console.
2. On the **VPCs** page, click **Create VPC**.
3. On the **Create VPC** page, set the following parameters to configure the VPC and click **OK**.

Parameter	Description
Organization	Select the organization to which the VPC belongs.
Resource Set	Select the resource set to which the VPC belongs.
Region	Select the region where you want to deploy the VPC.

Parameter	Description
Sharing Scope	Specify the scope of entities that are allowed to use the VPC. ○ **Current Resource Set**: If you select this option, the administrator of the current resource set can create resources in the VPC. ○ **Current Organization and Subordinate Organizations**: If you select this option, administrators that belong to the current organization and subordinate organizations can create resources in the VPC. ○ **Current Organization**: If you select this option, administrators that belong to the current organization can create resources in the VPC. In this example, **Current Resource Set** is selected.
VPC Name	Enter a name for the VPC. The name must be 2 to 128 characters in length, and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter or Chinese character and cannot start with `http://` or `https://`. In this example, **VPCtest** is entered.
IPv4 CIDR Block	Specify the IPv4 CIDR block of the VPC. You can specify an IPv4 CIDR block in one of the following ways: ○ **Recommended CIDR Block**: You can use one of the following standard IPv4 CIDR blocks: 192.168.0.0/16, 172.16.0.0/12, and 10.0.0.0/8. ○ **Custom CIDR Block**: Enter 192.168.0.0/16, 172.16.0.0/12, 10.0.0.0/8, or a subset of these CIDR blocks. The subnet mask must be 8 to 28 bits in length. For example, you can enter 192.168.0.0/16. In this example, Recommended CIDR Block is selected and 192.168.0.0/16 is selected as the IPv4 CIDR block of the VPC. **Note** After you create a VPC, you cannot change its IPv4 CIDR block.
IPv6 CIDR Block	Specify whether to assign an IPv6 CIDR block. ○ **Do Not Assign**: If you select this option, the system does not assign an IPv6 CIDR block to the VPC. ○ **Assign**: If you select this option, the system automatically assigns an IPv6 CIDR block to the VPC. In this example, **Assign** is selected.
Description	Enter a description for the VPC. The description must be 2 to 256 characters in length and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter or Chinese character. It cannot start with `http://` or `https://`.

4. Click **Back to Console**. In the left-side navigation pane, click **VSwitches**.
5. On the **VSwitches** page, click **Create VSwitch**.
6. On the **vSwitch** page, set the following parameter to configure the vSwitch and click **Submit**.

Parameter	Description
Organization	Select the organization to which the vSwitch belongs.

Parameter	Description
Resource Set	Select the resource set to which the vSwitch belongs.
Region	Select the region where you want to deploy the vSwitch.
Zone	Select the zone where you want to deploy the vSwitch. In a VPC, each vSwitch can be deployed in only one zone. You cannot deploy a vSwitch across zones. However, you can deploy cloud resources in vSwitches that belong to different zones to achieve zone-disaster recovery. **Note** Each cloud resource can be added to only one vSwitch.
Sharing Scope	Specify the scope of entities that are allowed to use the vSwitch. ○ **Current Resource Set**: If you select this option, the administrator of the current resource set can create resources in the vSwitch. ○ **Current Organization and Subordinate Organizations**: If you select this option, administrators that belong to the current organization and subordinate organizations can create resources in the vSwitch. ○ **Current Organization**: If you select this option, administrators that belong to the current organization can can create resources in the vSwitch. In this example, **Current Resource Set** is selected.
VPC	Select the VPC where you want to deploy the vSwitch. In this example, VPCtest is selected.
Dedicated for Out-of-cloud Physical Machines	Specify whether the vSwitch to be created is dedicated to bare-metal servers. For more information about bare-metal servers, see the **Bare-metal servers in VPCs** topic in *Bare-metal Server Management Service User Guide*. In this example, **No** is selected.
vSwitch Name	Enter a name for the vSwitch. The name must be 2 to 128 characters in length, and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter or Chinese character. It cannot start with `http://` or `https://`.
IPv4 CIDR Block	Enter an IPv4 CIDR block for the vSwitch. In this example, the default IPv4 CIDR block is used.
IPv6 CIDR Block	Enter an IPv6 CIDR block for the vSwitch. In this example, the default IPv6 CIDR block is used.
Description	Enter a description for the vSwitch. The description must be 2 to 256 characters in length, and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter or Chinese character. It cannot start with `http://` or `https://`.

Step 2: Create a security group

Perform the following steps to create a security group:

1. In the top navigation bar, choose **Products > Elastic Computing > Elastic Compute Service**.
2. Choose **Networks and Security > Security Groups**.
3. On the **Security Groups** page, click **Create Security Group**.
4. On the **Create Security Group** page, set the following parameters to configure the security group and click **Submit**.

Parameter	Description
Organization	Select the organization to which the security group belongs.
Resource Set	Select the resource set to which the security group belongs.
Region	Select the region to which the security group belongs. Make sure that the security group and the VPC belong to the same region.
Zone	Select the zone to which the security group belongs.
VPC	Select the VPC to which the security group belongs.
Security Group Name	Enter a name for the security group. The name must be 2 to 128 characters in length, and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter or Chinese character. It cannot start with `http://` or `https://`.
Description	Enter a description for the security group. The description must be 2 to 256 characters in length, and can contain digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter or Chinese character. It cannot start with `http://` or `https://`.

Step 3: Create and configure an ECS instance

After you create a VPC and a vSwitch, you must create an ECS instance and assign an IPv6 address to the ECS instance. You must associate this IPv6 address with the network interface controller (NIC) of the ECS instance.

Perform the following steps to create and configure an ECS instance:

1. In the top navigation bar, choose **Products > Networking > Virtual Private Cloud**.
2. In the left-side navigation pane, click **VSwitches**.
3. Select the region where the vSwitch is created.
4. On the **VSwitches** page, find the vSwitch that you want to manage and choose **Create > ECS Instance** in the **Actions** column.
5. On the **Create ECS Instance** page, configure the ECS instance and click **Submit**.

   In this example, **Assign** is selected. Therefore, an IPv6 IP address is assigned to the ECS instance. For more information about other parameters that you are required to specify when you create an ECS instance, see **Create an ECS instance** in **Quick Start** of *Elastic Compute Service User Guide*.
6. Return to the **Instances** page and click the instance ID to view the IPv6 address that is assigned to the ECS instance.
7. Configure a static IPv6 address.
   - If the image of your ECS instance supports DHCPv6, you do not need to manually configure a static IPv6 address. DHCPv6 enables automatic configuration of IPv6 addresses. Therefore, if your ECS instance image supports DHCPv6, the ECS instance can use the assigned IPv6 address to communicate within the private

network.

The following images support DHCPv6:

- Linux images:
  - CentOS 7.6 IPV6 64Bit
  - CentOS 6.10 64Bit
  - SUSE Linux Enterprise Server 12 SP4 64Bit
- Windows Server images

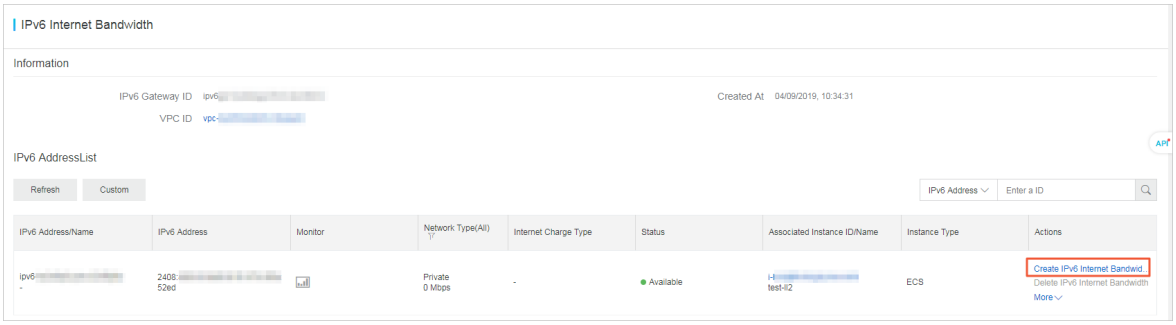
- If the image of your ECS instance does not support DHCPv6, you must manually configure an IPv6 address for the ECS instance. We recommend that you refer to the related documentation for each image for configuration guidance.

Step 4: Purchase an IPv6 Internet bandwidth plan

By default, IPv6 addresses are only used for communication within private networks. If you want to allow an instance that is assigned an IPv6 address to access the Internet or receive requests from IPv6 clients over the Internet, you must purchase an Internet bandwidth plan for the IPv6 address.

Perform the following steps to purchase an Internet bandwidth plan for the IPv6 address:

1. In the top navigation bar, choose **Products > Networking > IPv6 Gateway**.
2. Select the region where the IPv6 gateway is created.
3. On the **IPv6 Gateway** page, find the IPv6 gateway that you want to manage and click **Manage** in the **Actions** column.
4. In the left-side navigation pane, click **IPv6 Internet Bandwidth**.
5. On the **IPv6 Internet Bandwidth** page, find the IPv6 address that you want to manage and click **Enable IPv6 Internet Bandwidth** in the **Actions** column.



6. Select a bandwidth plan and click **OK**.

   The maximum IPv6 Internet bandwidth for an IPv6 gateway of the Free, Enterprise, or Enhanced Edition is 2 Gbit/s.

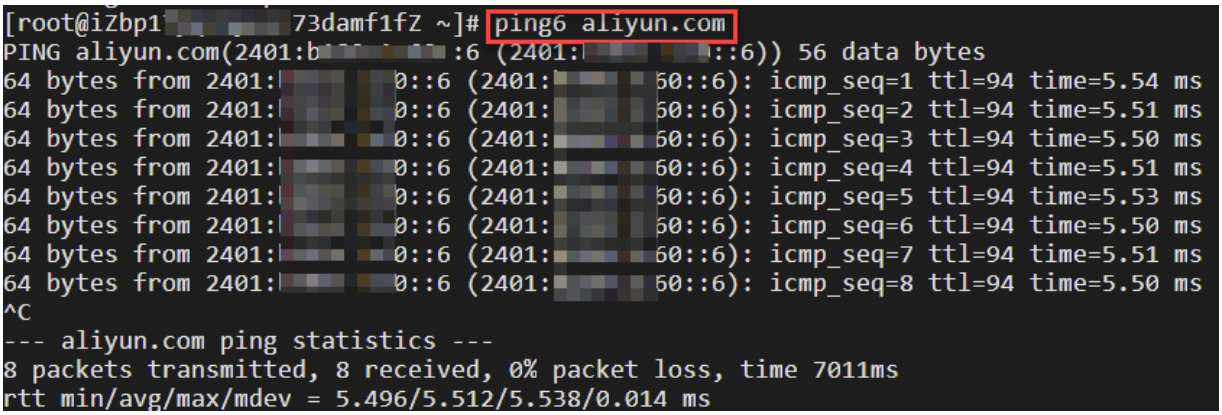
Step 5: Configure security group rules

IPv4 and IPv6 addresses are independent of each other. If the current security group rules do not apply to your IPv6 services, you must configure security group rules for the ECS instances to regulate communication with IPv6 addresses.

For more information about how to configure security rules, see the **Add security group rules** chapter in **Security Groups** of *Elastic Compute Service User Guide*.

Step 6: Test the network connectivity

Log on to an ECS instance and **ping** an IPv6 service to test the network connectivity.



18.1.4. VPCs and VSwitches

18.1.4.1. Overview

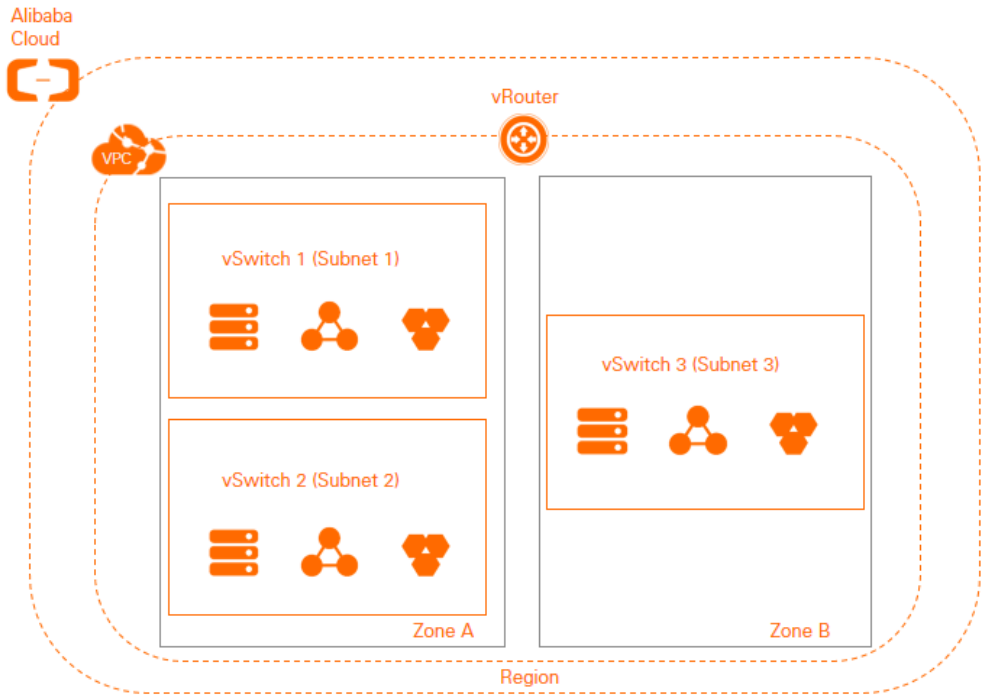
Before you can use cloud resources in a virtual private cloud (VPC), you must create a VPC and a vSwitch. You can create multiple vSwitches to divide a VPC into multiple subnets. By default, the subnets in a VPC can communicate with each other.

VPCs and vSwitches

VPCs are private networks dedicated to tenants.

> **Note** You cannot directly deploy cloud resources in a VPC. You must deploy cloud resources in the vSwitches of a VPC.

A vSwitch is a basic network device in a VPC and is used to connect cloud resources. You can deploy a VPC only in one region and cannot deploy a VPC across regions. However, a VPC covers all zones of the region to which the VPC belongs. You can create one or more vSwitches in a zone to divide a VPC into subnets.



CIDR blocks and IP addresses

VPCs support both IPv4 and IPv6. By default, VPCs use IPv4. You can enable IPv6 based on your business requirements.

VPCs support the dual-stack mode. In dual-stack mode, resources in a VPC can communicate through both IPv4 and IPv6 addresses. IPv4 and IPv6 addresses are independent of each other. Therefore, you must configure routes and security groups for both IPv4 and IPv6 addresses.

The following table lists the differences between IPv4 and IPv6 addresses.

IPv4 VPC	IPv6 VPC
32 binary digits. An IPv4 address appears as four decimal numbers separated by periods (.).Example: 192.168.1.11.	128 binary digits. An IPv6 address appears as eight hexadecimal numbers separated by colons (:). Example: 2001:db8:ffff:ffff:ffff:ffff:ffff:ffff.
By default, IPv4 is enabled.	You can enable IPv6 as needed.
The subnet mask of a VPC CIDR block can range from /8 to /24.	The subnet mask of a VPC CIDR block is /61.
The subnet mask of a vSwitch CIDR block can range from /16 to /29.	The subnet mask of a vSwitch CIDR block is /64.
You can specify an IPv4 CIDR block.	You cannot specify an IPv6 CIDR block. The system automatically assigns an IPv6 CIDR block to your VPC from the IPv6 address pool.

IPv4 VPC	IPv6 VPC
Supported by all instance types.	Not supported by some instance types. For more information, see **Instance types** under **What is ECS?** in the *Elastic Compute Service User Guide*.
IPv4 elastic IP addresses (EIPs) are supported.	IPv6 EIPs are not supported.
VPN gateways and NAT gateways are supported.	VPN gateways and NAT gateways are not supported.

By default, IPv4 and IPv6 addresses provided for VPCs support only communication over private networks. Cloud resources deployed in different vSwitches that belong to the same VPC can communicate with each other over private networks. To connect a VPC to another VPC or a data center, you can use Express Connect or VPN Gateway.

To enable cloud resources in a VPC to communicate with the Internet, perform the following operations:

- IPv4 addresses

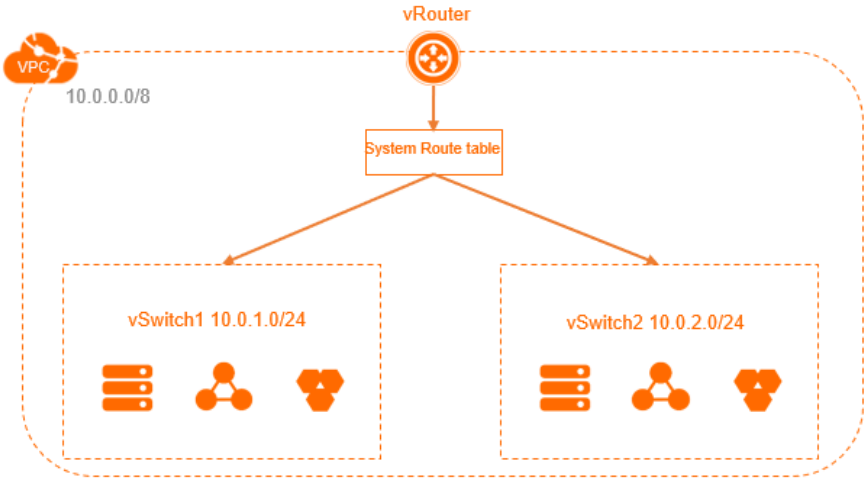
  You can configure NAT gateways or associate EIPs with Elastic Compute Service (ECS) instances in a VPC. This way, the ECS instances can communicate with the Internet through IPv4 addresses.

- IPv6 addresses

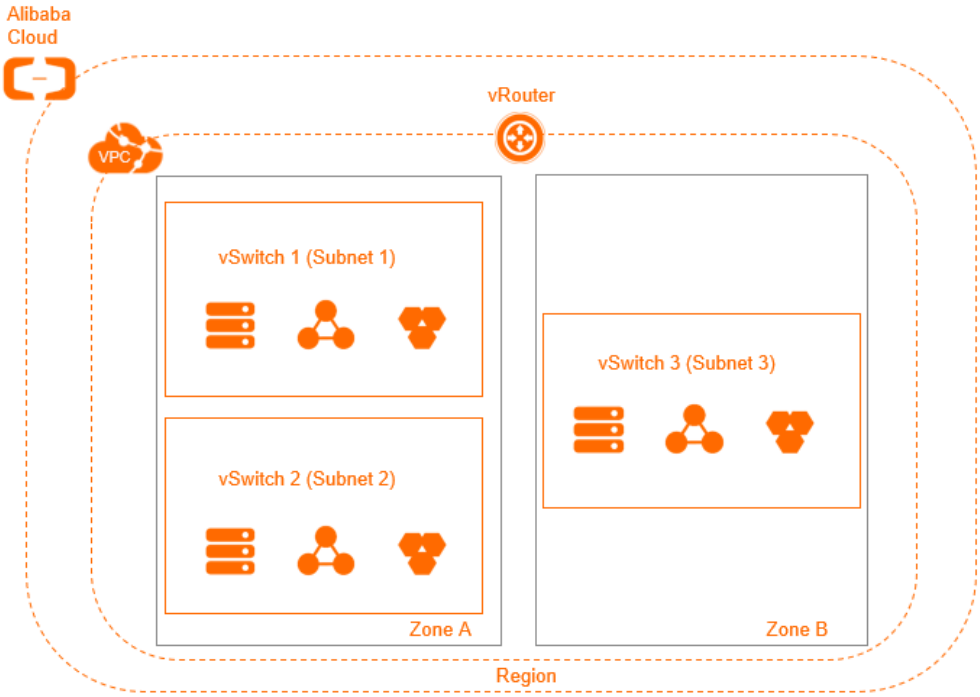
  To enable cloud resources in a VPC to communicate with the Internet through IPv6 addresses, you must purchase an IPv6 public bandwidth plan. You can configure egress-only rules for IPv6 addresses. This allows cloud resources in the VPC to access the Internet through IPv6 addresses. However, IPv6 clients cannot access the cloud resources over the Internet.

Routes

After you create a VPC, the system creates a system route table for the VPC and adds system routes to the route table. The system routes are used to route traffic within the VPC. A VPC has only one system route table. You cannot create or delete a system route table.



You can create and associate custom route tables with vSwitches to facilitate network management. A vSwitch can be associated only with one route table. For more information, see Create a custom route table.



If multiple route entries match the destination IP address, the route entry with the longest subnet mask is preferably used to determine the next hop. This ensures that the traffic is routed to the most precise destination. You can also add a custom route entry to route traffic to a specified destination. For more information, see Add a custom route entry.

18.1.4.2. VPC management

18.1.4.2.1. Create a VPC

Virtual private clouds (VPCs) are private networks dedicated to tenants. You have full control over your VPC. For example, you can specify CIDR blocks, and configure route tables and gateways for your VPC. This topic describes how to create a VPC.

Prerequisites

Before you create a VPC, you must plan your networks. For more information, see Design networks.

Procedure

1. Log on to the VPC console.
2. In the top navigation bar, select the region where you want to deploy the VPC.

   > **Note** You must deploy the VPC in the same region as the cloud resources that you want to deploy in this VPC.

3. On the **VPC** page, click **Create VPC**.
4. On the **Create VPC** page, set the parameters as described in the following table and click **Submit**.

Parameter	Description
Organization	Select the organization to which the VPC belongs.
Resource Set	Select the resource set to which the VPC belongs.
Region	Select the region where you want to deploy the VPC.
Sharing Scope	Select the sharing scope of the VPC. ○ **Current Resource Set**: Only the administrator of the current resource set can create resources for the shared VPC. ○ **Current Organization and Subordinate Organizations**: Only the administrators of the current organization and its subordinate organizations can use the VPC to create resources. ○ **Current Organization**: Only the administrator of the current organization can use the VPC to create resources.
VPC Name	Enter a name for the VPC. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`.
IPv4 CIDR Block	Specify an IPv4 CIDR block for the VPC. The following settings are supported: ○ **Recommended CIDR Block**: You can select the following standard IPv4 CIDR blocks: 192.168.0.0/16 and 172.16.0.0/16. ○ **Custom CIDR Block**: You can use 192.168.0.0/16, 172.16.0.0/16, or a subset of the preceding CIDR blocks. The subnet mask must be 8 to 28 bits in length. Example: 192.168.0.0/24. **Note** After a VPC is created, you cannot change its IPv4 CIDR block.
IPv6 CIDR Block	Specify whether to assign an IPv6 CIDR block. ○ **Do Not Assign**: The system does not assign an IPv6 CIDR block to the VPC. ○ **Assign**: The system automatically assigns an IPv6 CIDR block to the VPC. **Note** If you select **Do Not Assign**, you can click **Enable IPv6 CIDR Block** in the **IPv6 CIDR Block** column on the **VPC** page after the VPC is created. In the Enable IPv6 CIDR Block dialog box, you can select **Enable IPv6 CIDR Block of all VSwitches in VPC** to enable IPv6 for all vSwitches in the VPC. The system automatically creates a free IPv6 gateway for this VPC, and assigns an IPv6 CIDR block with the subnet mask /61, such as 2XX1:db8::/61. By default, IPv6 addresses can be used for communication only within private networks. If you want to enable mutual access between an IPv6 address and the Internet, you must purchase an Internet bandwidth plan for the IPv6 address. For more information, see the **Activate IPv6 Internet bandwidth** section of the **Manage IPv6 Internet bandwidth** topic of the *IPv6 gateway user guide*.
Description	Enter a description for the VPC. The description must be 2 to 256 characters in length, and can contain letters, digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.

18.1.4.2.2. Add a secondary IPv4 CIDR block

This topic describes how to expand a virtual private cloud (VPC) by adding a secondary IPv4 CIDR block to the VPC.

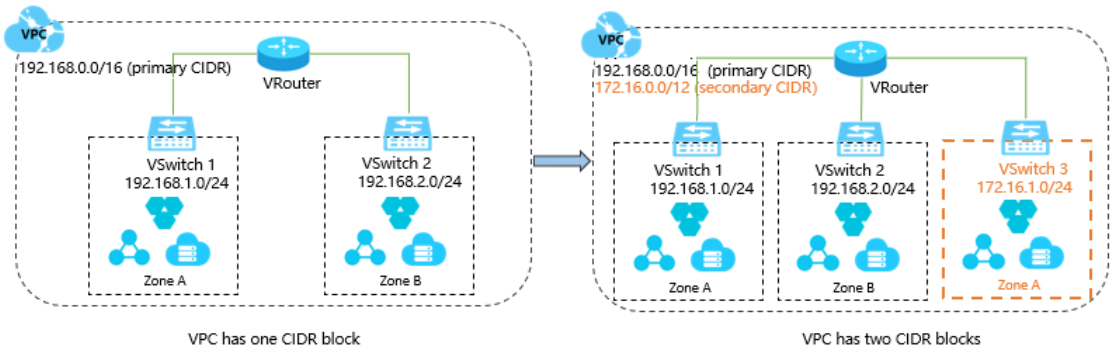
Prerequisites

A VPC is created. For more information, see 创建和管理专有网络.

Context

When you create a VPC, the IPv4 CIDR block that you specified is the primary CIDR block. After the VPC is created, the primary IPv4 CIDR block of the VPC cannot be modified. However, you can add a secondary IPv4 CIDR block to expand the VPC. After you add the secondary IPv4 CIDR block, both the primary and secondary IPv4 CIDR blocks take effect. You can create a vSwitch with the primary or secondary IPv4 CIDR block. However, each vSwitch belongs to only one VPC CIDR block.

The system automatically adds a vSwitch route to the VPC route table when you create a vSwitch with the primary or secondary IPv4 CIDR block. The destination CIDR block of a vSwitch route is the CIDR block with which the vSwitch is created. The CIDR block range cannot be the same as or larger than those of other routes in the route table of the VPC.



Note You can add only one secondary IPv4 CIDR block to a VPC and cannot increase the quota.

Procedure

1. Log on to the VPC console.
2. In the top navigation bar, select the region where the VPC is deployed.
3. On the **VPC** page, find the VPC and click **Manage** in the **Actions** column.
4. On the **CIDRs** tab, click **Add IPv4 CIDR**.
5. In the **Add Secondary CIDR** panel, set the following parameters and click **OK**.

Parameter	Description
VPC	The VPC to which you want to add the secondary IPv4 CIDR block.

Parameter	Description
Secondary CIDR	Select a method to configure the secondary IPv4 CIDR block: ○ **Default CIDR Block**: You can specify one of the following standard IPv4 CIDR blocks as the secondary IPv4 CIDR block: 192.168.0.0/16 and 172.16.0.0/12. ○ **Custom CIDR Block**: You can specify one of the following standard IPv4 CIDR blocks and their subnets as the secondary IPv4 CIDR block: 192.168.0.0/16 and 172.16.0.0/12. When you add a secondary IPv4 CIDR block, take note of the following limits: ○ The CIDR block cannot start with 0. The subnet mask must be 8 to 24 bits in length. ○ The secondary IPv4 CIDR block cannot overlap with the primary IPv4 CIDR block or an existing secondary IPv4 CIDR block. For example, if the primary IPv4 CIDR block of a VPC is 192.168.0.0/16, you cannot specify one of the following CIDR blocks as the secondary IPv4 CIDR block: ■ A CIDR block that is larger than 192.168.0.0/16, such as 192.168.0.0/8. ■ 192.168.0.0/16. ■ A CIDR block that is smaller than 192.168.0.0/16, such as 192.168.0.0/24.

What's next

Work with vSwitches

18.1.4.2.3. Delete a secondary IPv4 CIDR block

This topic describes how to delete a secondary IPv4 CIDR block of a virtual private cloud (VPC). However, you cannot delete the primary IPv4 CIDR block of a VPC.

Prerequisites

Before you delete a secondary IPv4 CIDR block, make sure that you have deleted the vSwitch that is created within the secondary IPv4 CIDR block. For more information, see Delete a vSwitch.

Procedure

1. Log on to the VPC console.
2. In the top navigation bar, select the region where the VPC is deployed.
3. On the **VPCs** page, find the VPC that you want to manage and click **Manage** in the **Actions** column.
4. On the **CIDRs** tab, find the secondary IPv4 CIDR block that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

18.1.4.2.4. Modify the name and description of a VPC

This topic describes how to modify the name and description of a virtual private cloud (VPC).

Procedure

1. Log on to the VPC console.
2. In the top navigation bar, select the region where your VPC is deployed.
3. On the **VPCs** page, find the target VPC network and click **Manage** in the **Actions** column.
4. In the **VPC Details** section, click **Edit** next to **Name**. In the dialog box that appears, enter a new name for the VPC and click **OK**.

The name must be 2 to 128 characters in length and can contain letters, Chinese characters, digits, underscores (_), and hyphens (-). It must start with a letter or a Chinese character.

5. Click **Edit** next to **Description**. In the dialog box that appears, enter a new description, and click **OK**.

   The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

18.1.4.2.5. Delete a VPC

This topic describes how to delete a virtual private cloud (VPC). After you delete a VPC, the vRouters and route tables associated with the VPC are also deleted.

Prerequisites

Before you delete a VPC, make sure that the following requirements are met:

- No vSwitch exists in the VPC. If a vSwitch exists in the VPC, delete the vSwitch first. For more information, see Delete a vSwitch.
- No IPv6 gateway exists in the VPC. If an IPv6 gateway exists in the VPC, delete the IPv6 gateway first.

Procedure

1. Log on to the VPC console.
2. In the top navigation bar, select the region where the VPC is deployed.
3. On the **VPCs** page, find the VPC that you want to delete, and click **Delete** in the **Actions** column.
4. In the **Delete VPC** message, click **OK**.

18.1.4.2.6. Manage tags

You can use tags to label virtual private clouds (VPCs), and then search and filter VPCs by tag.

Procedure

1. Log on to the VPC console.
2. In the top navigation bar, select the region where the VPC that you want to manage is deployed.
3. On the **VPCs** page, find the VPC that you want to manage, move the pointer over in the **Tags** column, and then click **Add** in the pop-up message.
4. In the **Configure Tags** dialog box, specify the key and value as described in the following table and click **OK**.

Parameter	Description
Tag Key	The key of a tag. You can select or enter a key. The key must be 1 to 64 characters in length, and cannot start with `aliyun` or `acs:`. It cannot contain `http://` or `https://`.
Tag Value	The value of a tag. You can select or enter a value. The value cannot exceed 128 characters in length, and cannot start with `aliyun` or `acs:`. It cannot contain `http://` or `https://`.

5. Return to the **VPCs** page and click **Filter by Tag**. In the **Filter by Tag** dialog box, specify tag keys and values to filter VPCs.

18.1.4.3. VSwitch management

18.1.4.3.1. Create a vSwitch

A vSwitch is a basic network component that connects different cloud resources in a virtual private cloud (VPC).

Context

After you create a VPC, you can create vSwitches to divide the VPC into one or more subnets. vSwitches within the same VPC can communicate with each other. Cloud resources must be deployed in vSwitches. You can deploy applications in vSwitches that belong to different zones to improve service availability.

> **Note** vSwitches do not support multicasting or broadcasting.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **vSwitch**.
3. In the top navigation bar, select the region where the VPC that you want to manage is deployed.
4. On the **vSwitch** page, click **Create vSwitch**.
5. On the **Create vSwitch** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the vSwitch belongs.
Resource Set	Select the resource set to which the vSwitch belongs.
Region	Select the region where you want to deploy the vSwitch.
Zone	Select the zone where you want to deploy the vSwitch. In a VPC, a vSwitch can be deployed only in one zone. You cannot deploy a vSwitch across zones. However, you can deploy cloud resources in vSwitches that belong to different zones to achieve cross-zone disaster recovery. **Note** An instance can be attached only to one vSwitch.
Sharing Scope	Select the sharing scope of the vSwitch. ○ **Current Resource Set**: Only the administrator of the current resource set can create resources in the shared vSwitch. ○ **Current Organization and Subordinate Organizations**: Only the administrators of the current organization and its subordinate organizations can create resources in the shared vSwitch. ○ **Current Organization**: Only the administrator of the current organization can use the vSwitch to create resources.
VPC	Select the VPC in which you want to create the vSwitch.
Dedicated for Out-of-cloud Physical Machines	Specify whether the vSwitch is dedicated for bare-metal servers. For more information about bare-metal servers, see the **Bare-metal servers in VPCs** topic in *Bare-metal Server Management Service User Guide*.

Parameter	Description
vSwitch Name	Enter a name for the vSwitch. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`.
IPv4 CIDR Block	Specify an IPv4 CIDR block for the vSwitch. ○ You must specify the IP address range of the vSwitch in CIDR notation. The subnet mask must be 16 to 29 bits in length. It means that 8 to 65,536 IP addresses can be provided. ○ The CIDR block of a vSwitch must be a subset of the CIDR block of the VPC to which the vSwitch belongs. ○ The first IP address and the last three IP addresses of each vSwitch CIDR block are reserved for the system. For example, if the CIDR block of a vSwitch is 192.168.1.0/24, the IP addresses 192.168.1.0, 192.168.1.253, 192.168.1.254, and 192.168.1.255 are reserved. ○ The CIDR block of a vSwitch cannot be the same as the destination CIDR block in a route entry of the VPC to which the vSwitch belongs. However, the CIDR block of the vSwitch can be a subset of the destination CIDR block of the route entry. ○ After a vSwitch is created, you cannot modify its CIDR block.
IPv6 CIDR Block	Specify an IPv6 CIDR block for the vSwitch. ○ You must check whether IPv6 is enabled for the specified VPC. If IPv6 is not enabled, you cannot assign an IPv6 CIDR block to the vSwitch. ○ If IPv6 is enabled for the VPC, you can enter a decimal number that ranges from 0 to 255 to define the last 8 bits of the IPv6 CIDR block of the vSwitch. For example, if the IPv6 CIDR block of the VPC is 2XX1:db8::/64, you can enter 255 to define the last 8 bits of the IPv6 CIDR block. In this case, the IPv6 CIDR block of the vSwitch is 2XX1:db8:ff::/64. ff is the hexadecimal value of 255. **Note** If IPv6 is not enabled when you create a vSwitch, you can enable IPv6 after the vSwitch is created. To do this, go to the **vSwitch** page, click **Enable IPv6 CIDR** in the **IPv6 CIDR Block** column.
Description	Enter a description for the vSwitch. The description must be 2 to 256 characters in length, and can contain letters, digits, underscores (_), hyphens (-), periods (.), colons (:), and commas (,). It must start with a letter and cannot start with `http://` or `https://`.

18.1.4.3.2. Create cloud resources in a vSwitch

You cannot directly deploy cloud resources in a virtual private cloud (VPC). You can deploy cloud resources only in a vSwitch that belongs to a VPC. This topic describes how to create cloud resources in a vSwitch.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **vSwitches**.
3. Select the region of the VPC to which the vSwitch belongs.
4. On the **vSwitches** page, find the vSwitch, click **Create** in the **Actions** column, and select the cloud resource that you want to create.

   You can create Elastic Compute Service (ECS) instances, ApsaraDB RDS instances, and Server Load Balancer

(SLB) instances in a vSwitch.

5. On the page that appears, set the parameters.

18.1.4.3.3. Modify a vSwitch

This topic describes how to modify the name and description of a vSwitch.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **vSwitches**.
3. Select the region of the VPC to which the vSwitch belongs.
4. On the **vSwitches** page, find the vSwitch that you want to manage and click **Manage** in the **Actions** column.
5. In the **vSwitch Basic Information** section, click **Edit** next to **Name** to modify the name of the vSwitch.

   The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter.
6. Click **Edit** next to **Description** to modify the description of the vSwitch.

   The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

18.1.4.3.4. Delete a vSwitch

This topic describes how to delete a vSwitch. After you delete a vSwitch, you cannot deploy cloud resources in it.

Prerequisites

Before you delete a vSwitch, make sure that the following requirements are met:

- All instances deployed in the vSwitch are deleted, such as Elastic Compute Service (ECS) instances, Server Load Balancer (SLB) instances, and ApsaraDB RDS instances.
- All resources associated with the vSwitch are deleted, such as high-availability virtual IP addresses (HAVIPs) and Source Network Address Translation (SNAT) entries.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **vSwitches**.
3. Select the region of the VPC to which the vSwitch belongs.
4. On the **vSwitch** page, find the vSwitch that you want to delete and click **Delete** in the **Actions** column.
5. In the **Delete vSwitch** message, click **OK**.

18.1.5. Route tables

18.1.5.1. Overview

After you create a virtual private cloud (VPC), the system automatically creates a system route table for the VPC and adds system routes to the route table. The system routes are used to route traffic within the VPC. You cannot create or delete system routes. However, you can create custom routes to route traffic from a specified CIDR block to a specified destination.

Route tables

After you create a VPC, the system creates a system route table to manage routes of the VPC. By default, vSwitches in the VPC use this route table. You cannot create or delete the system route table of a VPC. However, you can create a custom route table and associate it with a vSwitch to manage your network in a more flexible way. When you associate a vSwitch with a custom route table, the vSwitch is automatically disassociated from the system route table. For more information, see Add subnet routes to a route table.

Each entry in a route table is a *route entry*. A route entry specifies the destination to which network traffic is routed and consists of the destination CIDR block, next hop type, and next hop. Route entries include system route entries and custom route entries.

System routes

After you create a VPC, the system automatically adds the following system routes to the route table:

- A route entry with a destination CIDR block of 100.64.0.0/10. This route is used for communication among cloud resources within the VPC.
- Route entries whose destination CIDR blocks are the same as the CIDR blocks of the vSwitches in the VPC. These routes are used for communication among cloud resources within the vSwitches.

For example, if you create a VPC whose CIDR block is 192.168.0.0/16 and two vSwitches whose CIDR blocks are 192.168.1.0/24 and 192.168.0.0/24, three system routes are automatically added to the route table of the VPC. The following table describes the system routes.

Destination CIDR block	Next hop	Type
100.64.0.0/10	-	System route
192.168.1.0/24	-	System route
192.168.0.0/24	-	System route

Custom routes

You can add custom routes to replace system routes or route traffic to a specified destination. You can specify the following types of next hops when you create a custom route:

- Elastic Compute Service (ECS) instance: Traffic that is destined for the destination CIDR block is routed to a specified ECS instance in the VPC.

  You can select this type if the application deployed on the ECS instance needs to access the Internet or other applications.
- VPN gateway: Traffic destined for the destination CIDR block is routed to a specified VPN gateway.

  You can select this type if you want to connect a VPC to another VPC or an on-premises network through the VPN gateway.
- NAT gateway: Traffic destined for the destination CIDR block is routed to a specified NAT gateway.

  You can select this type if you want to connect a VPC to the Internet through the NAT gateway.
- Router interface (to VPC): Traffic that is destined for the destination CIDR block is routed to a specified VPC.

  You can select this type if you want to connect two VPCs through Express Connect circuits.
- Router interface (to VBR): Traffic that is destined for the destination CIDR block is routed to a specified virtual border router (VBR).

  You can select this type if you want to connect a VPC to an on-premises network through Express Connect circuits.
- Secondary ENI: Traffic that is destined for the destination CIDR block is routed to a specified secondary elastic network interface (ENI).
- IPv6 gateway: Traffic that is destined for the destination CIDR block is routed to a specified IPv6 gateway.

You can select this type if you want to implement IPv6 communication through an IPv6 gateway.

IPv6 routes

If IPv6 is enabled for your VPC, the following route entries are automatically added to the system route table of the VPC:

- A custom route entry whose destination CIDR block is ::/0 and whose next hop is the IPv6 gateway. Cloud resources deployed in the VPC use this route to access the Internet through IPv6 addresses.
- A system route entry of which the destination CIDR block is the IPv6 CIDR block of a vSwitch. This route is used for communication within the vSwitch.

> **Note** If you create a custom route table and associate the custom route table with a vSwitch for which IPv6 is enabled, you must add a custom route entry whose destination CIDR block is ::/0 and next hop is the IPv6 gateway instance. For more information, see Add a custom route entry.

Routing rules

If multiple route entries match the destination IP address, the route entry with the longest subnet mask is preferably used to determine the next hop. This ensures that the traffic is routed to the most precise destination.

For example, the following table describes the route table of a VPC.

Destination CIDR block	Next hop type	Next hop	Route entry type
100.64.0.0/10	-	-	System route
192.168.0.0/24	-	-	System route
0.0.0.0/0	Instance	i-12345678	Custom route
10.0.0.0/24	Instance	i-87654321	Custom route

The route entries that are destined for `100.64.0.0/10` and `192.168.0.0/24` are system route entries. The route entries that are destined for `0.0.0.0/0` and `10.0.0.0/24` are custom route entries. Traffic that is destined for `0.0.0.0/0` is routed to the ECS instance `i-12345678`, whereas traffic destined for `10.0.0.0/24` is routed to the ECS instance `i-87654321`. Based on the preceding rule, traffic destined for `10.0.0.1` is routed to the ECS instance `i-87654321`, whereas traffic destined for `10.0.1.1` is routed to the ECS instance `i-12345678`.

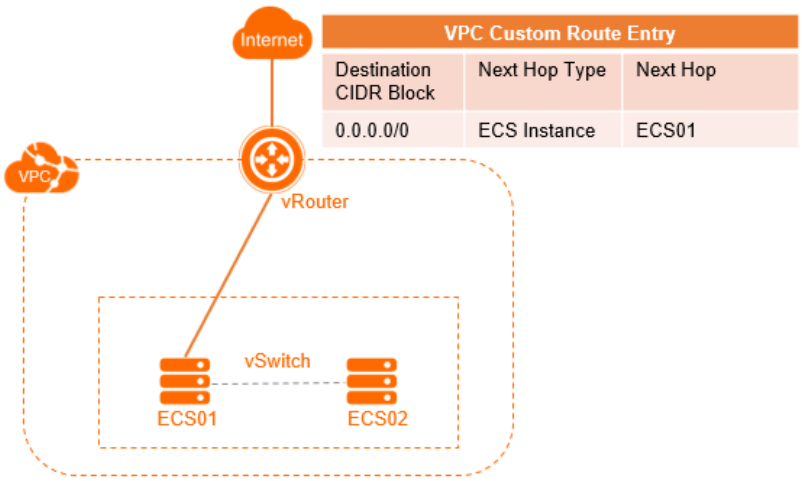
Examples

You can add custom route entries to a route table to control inbound and outbound traffic that is transmitted over the VPC.

- Routing within a VPC

  The following figure shows a NAT gateway that is deployed on an ECS instance (ECS01) in a VPC. To enable the cloud resources in the VPC to access the Internet through the ECS instance, you must add the following route entry to the route table.

Destination CIDR block	Next hop type	Next hop
0.0.0.0/0	ECS instance	ECS01



- Connect two VPCs by using Express Connect

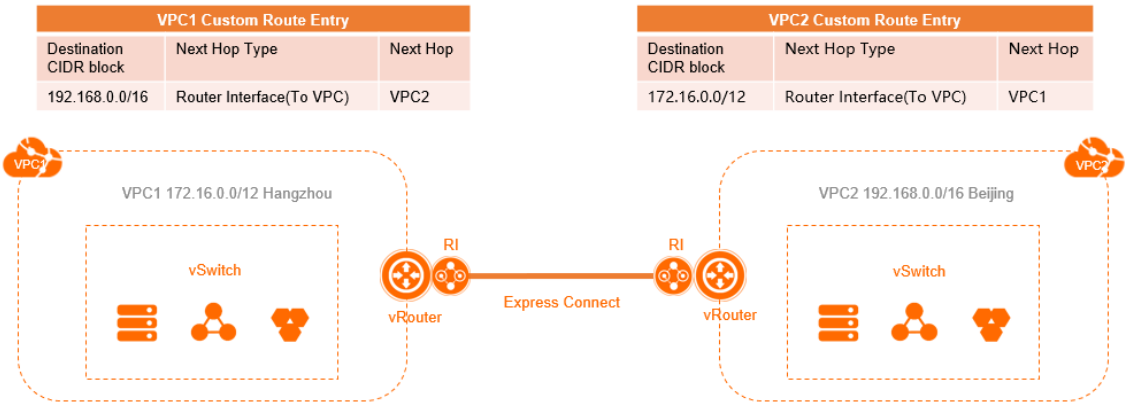
  The following figure shows that VPC1 (172.16.0.0/12) is connected to VPC2 (192.168.0.0/16) through Express Connect. After you create router interfaces, you must add the following route entries to the VPCs:

  - Add the following route entry to VPC1

Destination CIDR block	Next hop type	Next hop
192.168.0.0/16	Router interface (to VPC)	VPC2

  - Add the following route entry to VPC2

Destination CIDR block	Next hop type	Next hop
172.16.0.0/12	Router interface (to VPC)	VPC1



- Connect two VPCs through a VPN gateway

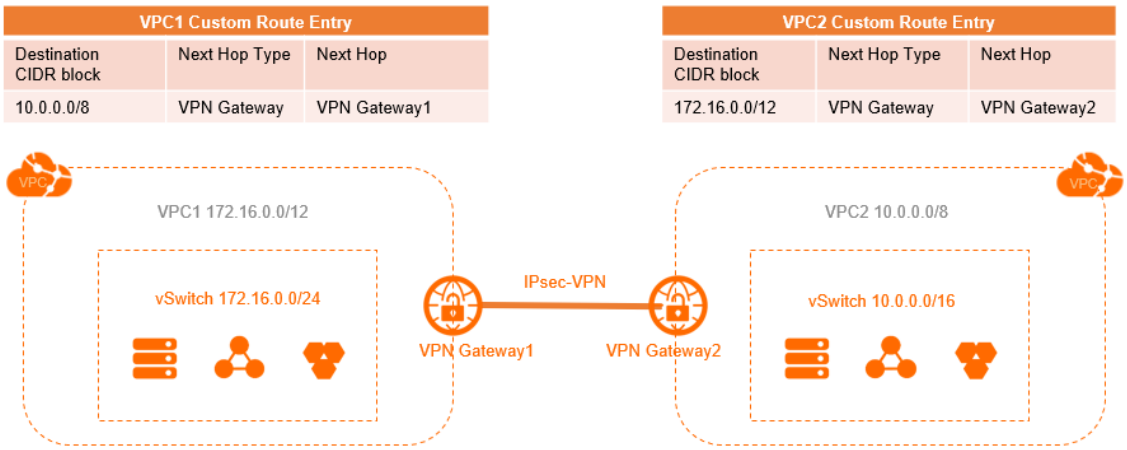
  The following figure shows that VPC1 (172.16.0.0/12) is connected to VPC2 (10.0.0.0/8) through a VPN gateway. After you configure the VPN gateway, you must add the following route entries to the VPCs.

- Add the following route entry to VPC1

Destination CIDR block	Next hop type	Next hop
10.0.0.0/8	VPN gateway	VPN gateway1

- Add the following route entry to VPC2

Destination CIDR block	Next hop type	Next hop
172.16.0.0/12	VPN gateway	VPN gateway2



- Connect a VPC to a data center through Express Connect

  The following figure shows that a VPC is connected to an on-premises network through Express Connect. After you configure a connection over an Express Connect circuit and a VBR, you must add the following route entries:

  - Add the following route entry to the VPC

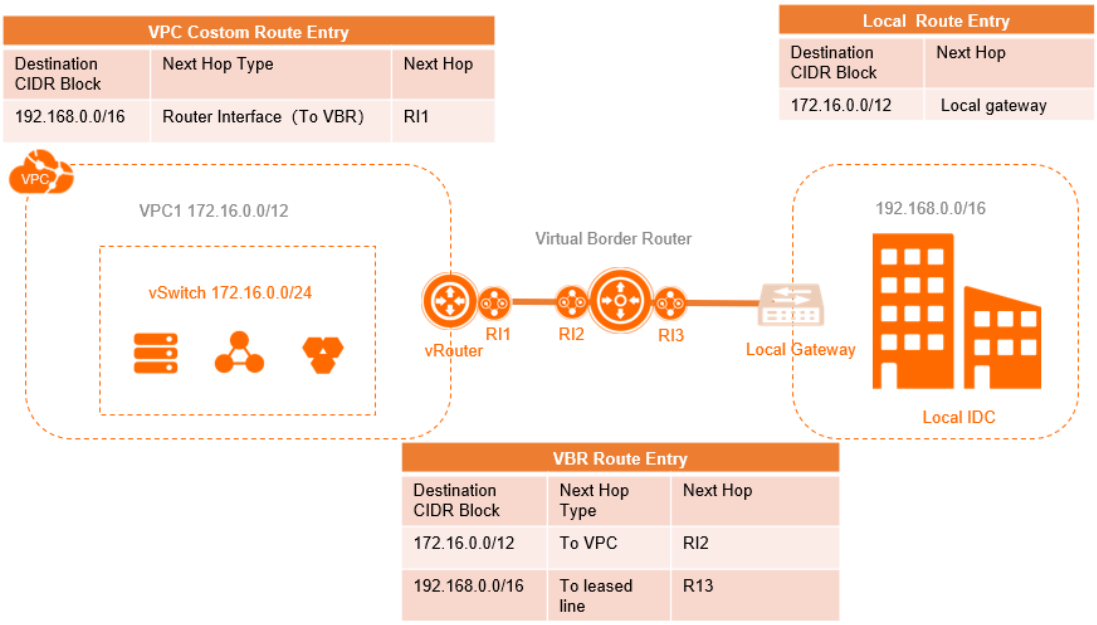
Destination CIDR block	Next hop type	Next hop
192.168.0.0/16	Router interface (general routing)	Router interface RI1

  - Add the following route entries to the VBR

Destination CIDR block	Next hop type	Next hop
192.168.0.0/16	Express Connect circuit	Router interface RI3
172.16.0.0/12	VPC	Router interface RI2

- Add the following route entry to the on-premises network

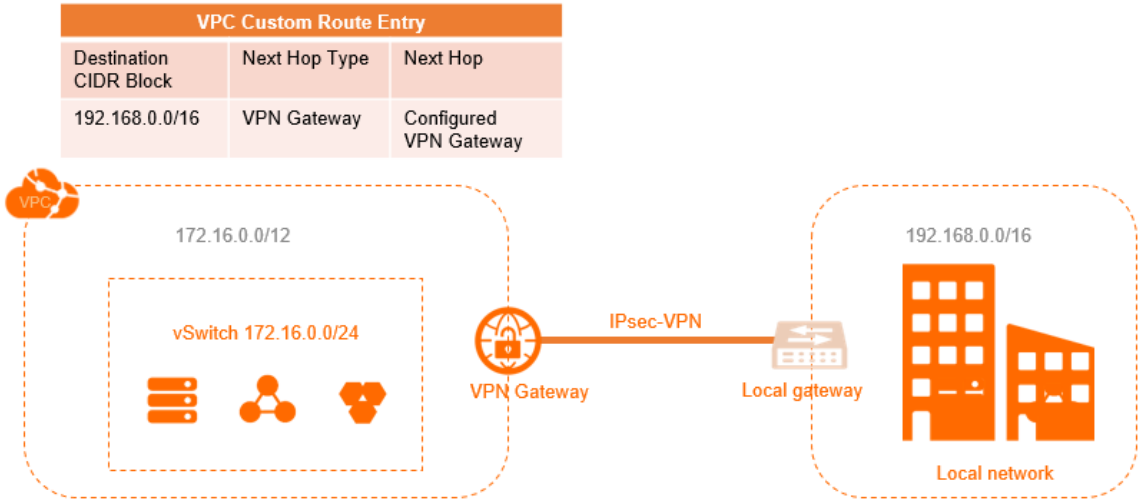
Destination CIDR block	Next hop type	Next hop
172.16.0.0/12	-	On-premises gateway device



- Connect a VPC to a data center through a VPN gateway

  The following figure shows that a VPC (172.16.0.0/12) is connected to a data center (192.168.0.0/16) through a VPN gateway. After you configure the VPN gateway, you must add the following route entry to the VPC.

Destination CIDR block	Next hop type	Next hop
192.168.0.0/16	VPN gateway	The VPN gateway that you created



18.1.5.2. Create a custom route table

This topic describes how to work with custom route tables. A route table consists of route entries. Each route entry specifies the destination to which network traffic is routed. You can create a custom route table to manage the inbound and outbound network traffic of subnets in a virtual private cloud (VPC).

Create a custom route table

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region where you want to create the route table.
4. On the **Route Tables** page, click **Create Route Table**.
5. In the **Create Route Table** panel, configure the route table as described in the following table and click **OK**.

Parameter	Description
VPC	Select the VPC to which the route table belongs.
Name	Enter a name for the route table. The name must be 2 to 128 characters in length and can contain digits, underscores (_), and hyphens (-). The name must start with a letter.
Description	Enter a description for the route table. The description must be 2 to 256 characters in length, and cannot start with `http://` or `https://`.

After you create a custom route table, you can go to the **Route Tables** page to view the route table. The type of route table is displayed as **Custom** in the **Route Table Type** column.

Disassociate a route table from a vSwitch

You can disassociate a custom route table from a vSwitch. After a custom route table is disassociated from a vSwitch, the vSwitch is automatically associated with the system route table.

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. On the **Route Tables** page, find the route table that you want to manage and click its ID.
5. In the **Route Table Details** section, click the **Associated vSwitch** tab. Find the vSwitch that you want to manage and click **Unbind** in the **Actions** column.
6. In the **Unbind Route Table** message, click **OK**.

Delete a custom route table

You can delete custom route tables. However, you cannot delete system route tables. If the custom route table that you want to delete is associated with a vSwitch, you must first disassociate the custom route table from the vSwitch.

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. On the **Route Tables** page, find the route table that you want to delete and click **Delete** in the **Actions** column.
5. In the **Delete Route Table** message, click **OK**.

18.1.5.3. Add a custom route entry

This topic describes how to add a custom route entry. After you create a virtual private cloud (VPC), the system creates a system route table and adds system route entries to the route table. The system route entries are used to route traffic within the VPC. You cannot create or delete system route entries. However, you can create custom route entries to route traffic from source CIDR blocks to specific destinations.

Context

Each item in the route table is a route entry. A route entry specifies the destination to which network traffic is routed and consists of the destination CIDR block, next hop type, and next hop. Route entries include system route entries and custom route entries. You can add custom route entries to both the system route table and custom route tables.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. On the **Route Tables** page, find the route table and click **Manage** in the **Actions** column.
5. In the **Route Table Details** section, click **Add Route Entry**.
6. In the **Add Route Entry** panel, set the following parameters and click **OK**.

Parameter	Description
Destination CIDR Block	Enter a destination CIDR block for the route entry. ○ **IPv4 CIDR Block**: The destination CIDR block is an IPv4 CIDR block. ○ **IPv6 CIDR Block**: The destination CIDR block is an IPv6 CIDR block.

Parameter	Description
Next Hop Type	Select the next hop type. Valid values: ○ **ECS Instance**: Traffic destined for the specified CIDR block is routed to the specified Elastic Compute Service (ECS) instance. Select this type if you want to route traffic to an ECS instance for centralized traffic forwarding and management. For example, you can configure an ECS instance as the Internet-facing gateway to route traffic from other ECS instances to the Internet. ○ **HaVip**: Traffic destined for the specified CIDR block is routed to the high-availability virtual IP address (HAVIP) that you select. ○ **VPN Gateway**: Traffic destined for the specified CIDR block is routed to the specified VPN gateway. ○ **NAT Gateway**: Traffic destined for the specified CIDR block is routed to the specified NAT gateway. ○ **Secondary ENI**: Traffic destined for the specified CIDR block is routed to the secondary elastic network interface (ENI) that you select. ○ **Router Interface (To VPC)**: Traffic destined for the specified CIDR block is routed to the VPC that you select. Select this type if you want to connect VPCs through Express Connect circuits. ○ **Router Interface (To VBR)**: Traffic destined for the specified CIDR block is routed to the router interface that is associated with a virtual border router (VBR). Select this type if you want to connect the VPC to a data center through Express Connect circuits. If you select Router Interface (To VBR), you must also select a routing mode: ■ **General Routing**: Select an associated router interface. ■ **Active/Standby Routing**: Select two instances as the next hop. The active route has a weight of 100 and the standby route has a weight of 0. The standby route takes over when the active route fails to pass the health check. ■ **Load Balancing Routing**: Select two to four router interfaces as the next hop. The peer router of each router interface must be a VBR. You can set the weight of each instance to an integer from 1 to 255. The default value is 100. The weights of the instances must be the same. This way, traffic can be evenly distributed to the next-hop instances. ○ **IPv6 Gateway**: Traffic destined for the specified CIDR block is routed to the specified IPv6 gateway. **Note** If you set **Destination CIDR Block** to **IPv6 CIDR Block**, you can select **IPv6 Gateway** and **Router Interface (To VBR)** for **Next Hop Type**.
ECS Instance, VPN Gateway, NAT Gateway, Secondary ENI, HaVip, Router Interface (To VPC), Router Interface (To VBR), IPv6 Gateway	Select an instance as the next hop.

18.1.5.4. Export route entries

This topic describes how to export route entries from a route table for backup.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. On the **Route Tables** page, find the route table and click **Manage** in the **Actions** column.
5. In the **Route Table Details** section, click the **Route Entry List** tab, and then click **Export**.

   The route entries are exported to a *.csv* file in your local computer.

18.1.5.5. Modify a route table

This topic describes how to modify the name and description of a route table.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. On the **Route Tables** page, find the route table and click its ID.
5. In the **Route Table Details** section, click **Edit** next to **Name**. In the dialog box that appears, enter a new name for the route table and click **OK**.

   The name must be 2 to 128 characters in length and can contain letters, Chinese characters, digits, underscores (_), and hyphens (-). It must start with a letter or a Chinese character.
6. Click **Edit** next to **Description**. In the dialog box that appears, enter a new description of the route table, and click **OK**.

   The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

18.1.5.6. Delete a custom route entry

This topic describes how to delete a custom route entry. A route table consists of one or more route entries. Each route entry specifies the destination network to which traffic is routed. You can delete custom route entries. However, you cannot delete system route entries.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. In the **Route Table Details** section, click **Add Route Entry**.
5. On the **Route Entry List** tab, find the custom route that you want to delete and click **Delete** in the **Actions** column.
6. In the **Delete Route Entry** message, click **OK**.

18.1.5.7. Add subnet routes to a route table

You can create a custom route table for a virtual private cloud (VPC) and add subnet routes to the custom route table. Then, you can associate the custom route table with a vSwitch to manage the traffic of the vSwitch. This facilitates network management.

Prerequisites

A VPC is created. For more information, see Create a VPC.

Step 1: Create a custom route table

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region where you want to create the route table.
4. On the **Route Tables** page, click **Create Route Table**.
5. In the **Create Route Table** panel, configure the route table as described in the following table and click **OK**.

Parameter	Description
VPC	Select the VPC to which the route table belongs.
Name	Enter a name for the route table. The name must be 2 to 128 characters in length and can contain digits, underscores (_), and hyphens (-). The name must start with a letter.
Description	Enter a description for the route table. The description must be 2 to 256 characters in length, and cannot start with `http://` or `https://`.

After you create a custom route table, you can go to the **Route Tables** page to view the route table. The type of route table is displayed as **Custom** in the **Route Table Type** column.

Step 2: Add subnet routes

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table belongs.
4. On the **Route Tables** page, find the route table and click **Manage** in the **Actions** column.
5. In the **Route Table Details** section, click **Add Route Entry**.
6. In the **Add Route Entry** panel, set the following parameters and click **OK**.

Parameter	Description
Destination CIDR Block	Enter a destination CIDR block for the route entry. ○ **IPv4 CIDR Block**: The destination CIDR block is an IPv4 CIDR block. ○ **IPv6 CIDR Block**: The destination CIDR block is an IPv6 CIDR block.

Parameter	Description
Next Hop Type	Select the next hop type. Valid values: ○ **ECS Instance**: Traffic destined for the specified CIDR block is routed to the specified Elastic Compute Service (ECS) instance. Select this type if you want to route traffic to an ECS instance for centralized traffic forwarding and management. For example, you can configure an ECS instance as the Internet-facing gateway to route traffic from other ECS instances to the Internet. ○ **HaVip**: Traffic destined for the specified CIDR block is routed to the high-availability virtual IP address (HAVIP) that you select. ○ **VPN Gateway**: Traffic destined for the specified CIDR block is routed to the specified VPN gateway. ○ **NAT Gateway**: Traffic destined for the specified CIDR block is routed to the specified NAT gateway. ○ **Secondary ENI**: Traffic destined for the specified CIDR block is routed to the secondary elastic network interface (ENI) that you select. ○ **Router Interface (To VPC)**: Traffic destined for the specified CIDR block is routed to the VPC that you select. Select this type if you want to connect VPCs through Express Connect circuits. ○ **Router Interface (To VBR)**: Traffic destined for the specified CIDR block is routed to the router interface that is associated with a virtual border router (VBR). Select this type if you want to connect the VPC to a data center through Express Connect circuits. If you select Router Interface (To VBR), you must also select a routing mode: ■ **General Routing**: Select an associated router interface. ■ **Active/Standby Routing**: Select two instances as the next hop. The active route has a weight of 100 and the standby route has a weight of 0. The standby route takes over when the active route fails to pass the health check. ■ **Load Balancing Routing**: Select two to four router interfaces as the next hop. The peer router of each router interface must be a VBR. You can set the weight of each instance to an integer from 1 to 255. The default value is 100. The weights of the instances must be the same. This way, traffic can be evenly distributed to the next-hop instances. ○ **IPv6 Gateway**: Traffic destined for the specified CIDR block is routed to the specified IPv6 gateway. **Note** If you set **Destination CIDR Block** to **IPv6 CIDR Block**, you can select **IPv6 Gateway** and **Router Interface (To VBR)** for **Next Hop Type**.
ECS Instance, VPN Gateway, NAT Gateway, Secondary ENI, HaVip, Router Interface (To VPC), Router Interface (To VBR), IPv6 Gateway	Select an instance as the next hop.

Step 3: Associate the custom route table with a vSwitch

You can associate the custom route table with a vSwitch to manage the routes of the vSwitch. A vSwitch can be associated only with one system route table or one custom route table. Perform the following operations to associate the custom route table with a vSwitch:

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Route Tables**.
3. In the top navigation bar, select the region to which the route table that you want to manage belongs.
4. On the **Route Tables** page, find the custom route table that you want to manage and click its ID.
5. In the **Route Table Details** section, click the **Associated vSwitch** tab and click **Associate vSwitch**.
6. In the **Associate vSwitch** panel, select the vSwitch that you want to associate and click **OK**.

18.1.6. HAVIPs

18.1.6.1. Overview

High-availability virtual IP addresses (HAVIPs) are private IP addresses that can be created and released as independent resources. You can use HAVIPs with high-availability (HA) software such as Keepalived to provide active/standby services. This improves the availability of your services.

Introduction

Each Elastic Compute Service (ECS) instance is assigned a private IP address as the primary IP address. You can associate HAVIPs with an ECS instance to increase the number of private IP addresses available for the ECS instance. Both the primary IP address and HAVIPs of an ECS instance can be used to access networks. In addition, you can use HAVIPs with HA software such as Keepalived to provide active/standby services. This improves the availability of your services. You can associate an HAVIP with ECS instances by using the following methods:

- Directly associate an HAVIP with ECS instances.

  Each HAVIP can be associated with two ECS instances. After an HAVIP is associated with ECS instances, the ECS instances can send Address Resolution Protocol (ARP) messages to advertise the HAVIP. After the ECS instances advertise the HAVIP, one of the ECS instances serves as the primary instance, and the other ECS instance serves as the secondary instance. If the primary ECS instance is down, the secondary ECS instance takes over to provide services.

- Attach secondary elastic network interfaces (ENIs) to ECS instances. Then, associate the HAVIP with the secondary ENIs.

  Each HAVIP can be associated with ENIs of two ECS instances. After the HAVIP is associated with the ENIs, the ECS instances can send ARP messages to advertise the HAVIP. After the ECS instances advertise the HAVIP, one of the ECS instances serves as the primary instance, and the other ECS instance serves as the secondary instance. If the primary ECS instance is down, the secondary ECS instance takes over to provide services.

  > **Note** Before you associate an HAVIP with secondary ENIs, make sure that the secondary ENIs are attached to two ECS instances.

HAVIPs have the following features:

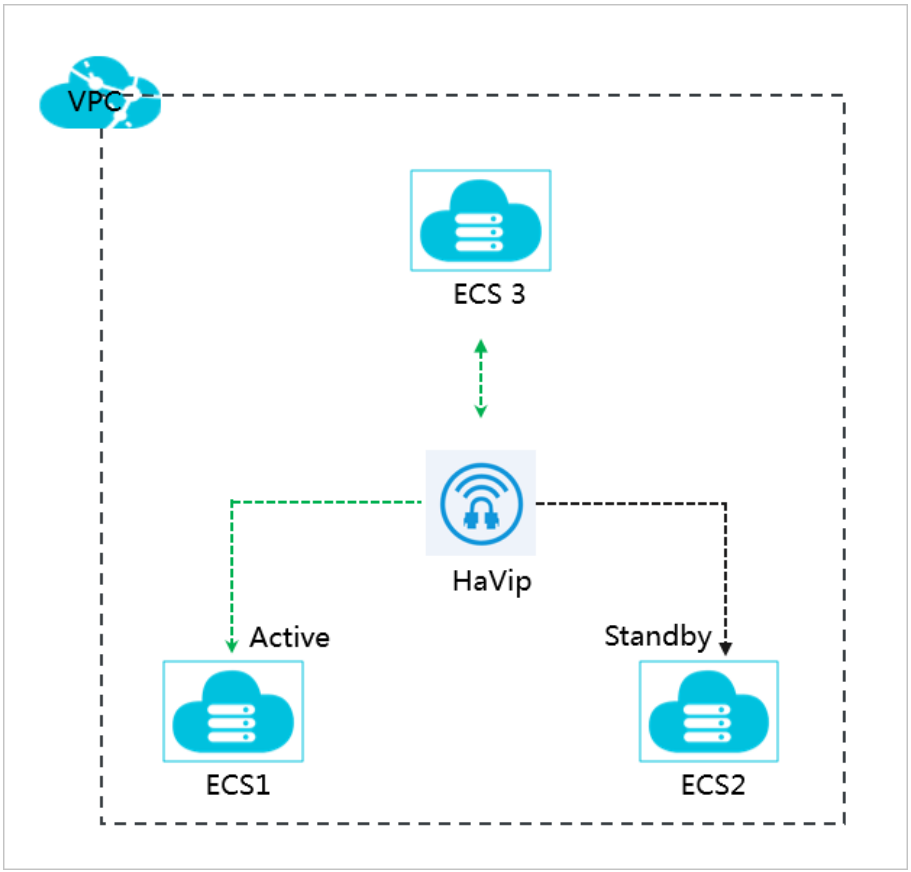
- HAVIPs are floating private IP addresses and are not statically assigned to specified ECS instances. HAVIPs can be associated with or disassociated from ECS instances through ARP announcements.
- An HAVIP can be associated only with ECS instances or ENIs that belong to the same vSwitch.
- You can associate each HAVIP with two ECS instances or two secondary ENIs. However, you cannot associate an HAVIP with an ECS instance and a secondary ENI at the same time.

Scenarios

HAVIPs support flexible configurations in the following scenarios:

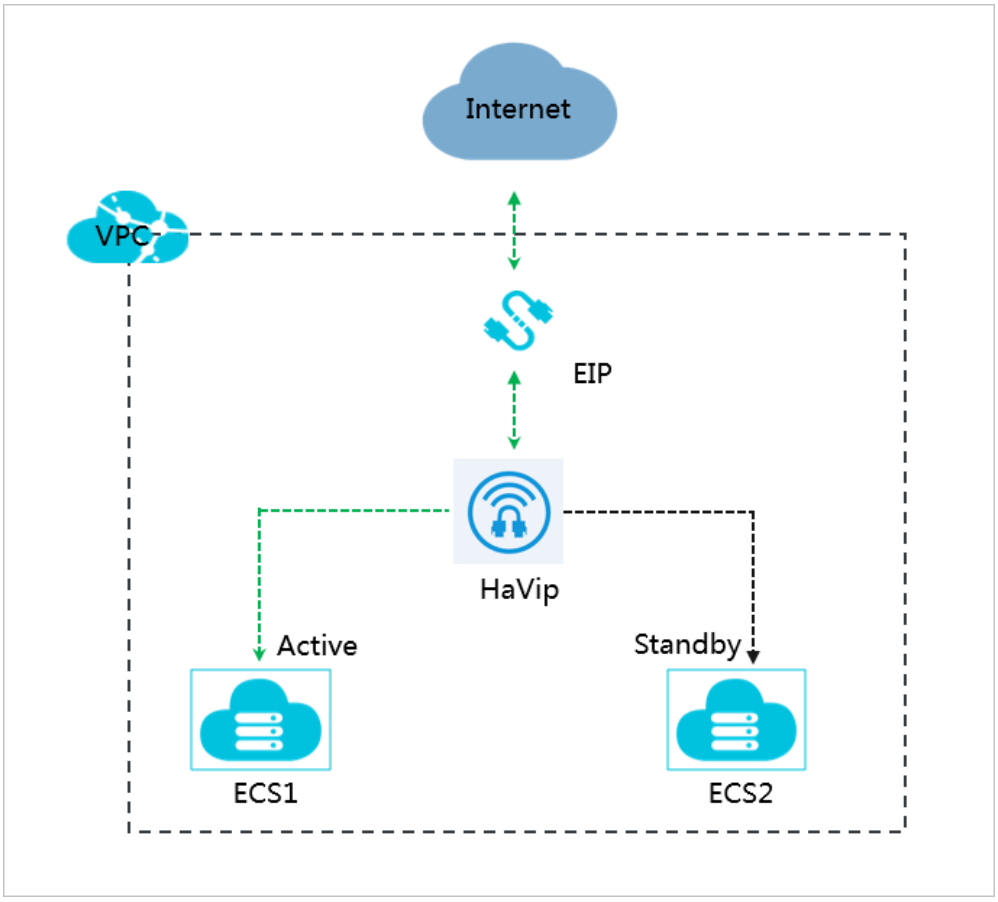
- Scenario 1: Internal-facing HA services

  In the following figure, two ECS instances are assigned the same HAVIP. Keepalived is configured for the ECS instances to provide an internal-facing HA service. Other instances in the same virtual private cloud (VPC) can access this service. The HAVIP serves as the service address. If the primary ECS instance is down, the secondary ECS instance takes over. This improves the availability of your services.



- Scenario 2: Internet-facing HA services

  In the following figure, two ECS instances are assigned the same HAVIP. Keepalived is configured and the HAVIP is associated with an elastic IP address (EIP) for the ECS instances to provide an Internet-facing HA service. The EIP that is associated with the HAVIP serves as the service address. If the primary ECS instance is down, the secondary ECS instance takes over. This improves the availability of your services.



Limits

Before you use HAVIPs, take note of the following limits.

Item	Limit
Number of HAVIPs that can be created in each VPC	5
Number of HAVIPs that can be associated with each ECS instance	5
Number of HAVIPs that can be associated with each ENI	5
Number of ECS instances that can be associated with each HAVIP	2
Number of ENIs that can be associated with each HAVIP	2
Number of route entries that point an HAVIP in each VPC	5

Item	Limit
Whether HAVIPs support broadcasting or multicasting	N/A **Note** HAVIPs support only unicasting. To implement high availability through third-party software such as keepalived, you must modify the configuration file to change the communication method to unicasting.

18.1.6.2. Create an HAVIP

A high-availability virtual IP address (HAVIP) is a private IP address that can be created and released as an independent resource. This topic describes how to create an HAVIP in the console.

Prerequisites

A virtual private cloud (VPC) and vSwitches are created. For more information, see Create a VPC and Create a vSwitch.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top navigation bar, select the region where you want to create the HAVIP.
4. On the **HaVip** page, click **Create HaVip**.
5. On the **Create HAVIP** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the HAVIP belongs.
Resource Set	Select the resource set to which the HAVIP belongs.
Region	Select the region where you want to create the HAVIP.
VPC	Select the VPC to which the HAVIP belongs.
vSwitch	Select the vSwitch to which the HAVIP that you want to create belongs.
Private IP Address	Specify a private IP address for the HAVIP. **Note** You must specify an idle private IP address that falls within the CIDR block of the vSwitch.

18.1.6.3. Associate HAVIPs with backend cloud resources

18.1.6.3.1. Associate an HAVIP with an ECS instance

This topic describes how to associate a high-availability virtual IP address (HAVIP) with an Elastic Compute Service (ECS) instance that is deployed in a virtual private clouds (VPC). After you associate an HAVIP with an ECS instance, the ECS instance can send Address Resolution Protocol (ARP) messages to advertise the HAVIP. This way, the ECS instance can use more than one private IP address. Each HAVIP can be associated with at most two ECS instances.

Prerequisites

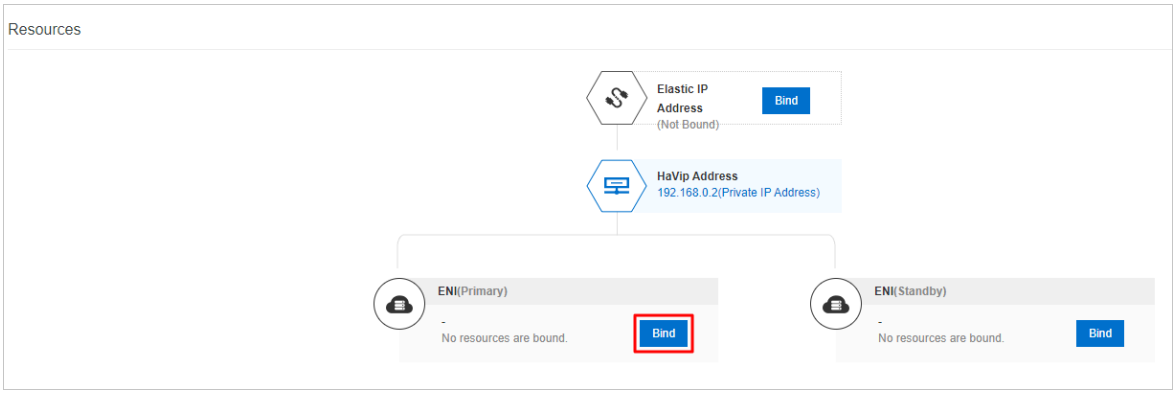
An ECS instance is created. For more information, see in the **Create an ECS instance** topic in **Quick Start** of *Elastic Compute Service User Guide*.

Context

You can associate an HAVIP with two ECS instances or two secondary elastic network interfaces (ENIs). However, you cannot associate an HAVIP with an ECS instance and a secondary ENI at the same time.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top navigation bar, select the region where the HAVIP is created.
4. On the **HaVip Addresses** page, find the HAVIP that you want to manage and click **Manage** in the **Actions** column.
5. In the **Resources** section, find **ENI (Primary)** or **ENI (Standby)** and click **Bind**.



6. In the dialog box that appears, set the following parameters to associate the HAVIP with an ECS instance.

Parameter	Description
Resource Type	Select the type of resource with which you want to associate the HAVIP. Valid values: ○ **ECS Instances** ○ **ENI** In this example, **ECS Instances** is selected.
Bind Resource	Select the ECS instance with which you want to associate the HAVIP. The ECS instance that you select must meet the following requirements: ○ The ECS instance is deployed in a VPC. ○ The ECS instance and the HAVIP belong to the same vSwitch.

7. Click **OK**.

18.1.6.3.2. Associate an HAVIP with an ENI

This topic describes how to associate a high-availability virtual IP address (HAVIP) with ENIs that are attached to Elastic Compute Service (ECS) instances. Then, the ECS instances can send Address Resolution Protocol (ARP) messages to advertise the HAVIP. This way, the ECS instances can use more than one private IP address.

Prerequisites

An ENI is created. For more information, see the **Create an Elastic Network Interface** topic in the **Elastic Network Interface** chapter of *Elastic Compute Service User Guide*.

Context

You can associate each HAVIP with two ECS instances or two ENIs. However, you cannot associate an HAVIP with an ECS instance and an ENI at the same time.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top navigation bar, select the region where the HAVIP is created.
4. On the **HaVip Addresses** page, find the HAVIP that you want to associate and click its ID.
5. In the **Resources** section, find **ENI (Primary)** or **ENI (Standby)** and click **Bind**.
6. In the dialog box that appears, set the following parameters and click **OK**.

Parameter	Description
Resource Type	Select the type of resource with which you want to associate the HAVIP. Valid values: ○ **ECS Instance** ○ **ENI** In this example, **ENI** is selected.
Bind Resource	Select the ENI with which you want to associate the HAVIP. The ENI and the HAVIP must belong to the same vSwitch.

18.1.6.4. Associate HAVIPs with EIPs

This topic describes how to associate high-availability virtual IP addresses (HAVIPs) with elastic IP addresses (EIPs). After you associate an HAVIP with an EIP, the HAVIP can use the EIP to provide services over the Internet.

Prerequisites

An EIP is created. For more information, see **Create an EIP** in **Quick Start** of the *Elastic IP Address User Guide*.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top menu bar, select the region where the HAVIP is created.
4. On the **HaVip Addresses** page, find the HAVIP that you want to manage, and choose **More > Bind EIP Address** in the **Actions** column.
5. In the dialog box that appears, select the EIP with which you want to associate the HAVIP and click **OK**.

   The EIP with which you want to associate the HAVIP must meet the following requirements:

   ○ The EIP and HAVIP are created in the same region.

- The EIP must be in the Available state.

18.1.6.5. Disassociate HAVIPs from backend cloud resources

18.1.6.5.1. Disassociate an HAVIP from an ECS instance

This topic describes how to disassociate a high-availability virtual private IP address (HAVIP) from an Elastic Compute Service (ECS) instance. After you disassociate an HAVIP from an ECS instance, the ECS instance cannot send Address Resolution Protocol (ARP) messages to advertise the HAVIP.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top navigation bar, select the region where the HAVIP is created.
4. On the **HaVip** page, find the HAVIP that you want to manage and click **Manage** in the **Actions** column.
5. In the **Resources** section, find the ECS instance that you want to manage and click **Unbind**.
6. In the message that appears, click **OK**.

18.1.6.5.2. Disassociate an HAVIP from an ENI

This topic describes how to disassociate a high-availability virtual IP address (HAVIP) from an elastic network interface (ENI). After you disassociate an HAVIP from an ENI, the Elastic Compute Service (ECS) instance with which the ENI is associated cannot send Address Resolution Protocol (ARP) messages to advertise the HAVIP.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top navigation bar, select the region where the HAVIP is created.
4. On the **HaVip** page, find the HAVIP that you want to disassociate and click **Manage** in the **Actions** column.
5. In the **Resources** section, find the ENI that you want to manage and click **Unbind**.
6. In the message that appears, click **OK**.

18.1.6.6. Disassociate an HAVIP from an EIP

This topic describes how to disassociate a high-availability virtual IP address (HAVIP) from an elastic IP address (EIP).

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip Addresses**.
3. Select the region to which the HAVIP belongs.
4. On the **HaVip Addresses** page, find the HAVIP that you want to manage and choose **More > Unbind with EIP** in the **Actions** column.
5. In the message that appears, click **OK**.

18.1.6.7. Delete an HAVIP

This topic describes how to delete a high-availability virtual IP address (HAVIP).

Prerequisites

- The HAVIP that you want to delete is not associated with an elastic IP address (EIP). If the HAVIP is associated with an EIP, disassociate the HAVIP from the EIP first. For more information, see Disassociate HAVIPs from EIPs.
- The HAVIP is not associated with an Elastic Compute Service (ECS) instance. If the HAVIP is associated with an ECS instance, disassociate the HAVIP from the ECS instance. For more information, see Disassociate an HAVIP from an ECS instance.
- The HAVIP is not associated with a secondary elastic network interface (ENI). If the HAVIP is associated with a secondary ENI, disassociate the HAVIP from the secondary ENI first. For more information, see Disassociate HAVIPs from secondary ENIs.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **HaVip**.
3. In the top navigation bar, select the region where the HAVIP is created.
4. On the **HaVip** page, find the HAVIP that you want to manage and choose **More > Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

18.1.7. Network ACLs

18.1.7.1. Overview

Network access control lists (ACLs) allow you to regulate access control for a virtual private cloud (VPC). You can create network ACL rules and associate a network ACL with a vSwitch. This allows you to control inbound and outbound traffic of Elastic Compute Service (ECS) instances that are associated with the vSwitch.

Features

Network ACLs have the following features:

- A network ACL is used to filter inbound and outbound network traffic of ECS instances that are associated with a vSwitch in a VPC. The network traffic forwarded to ECS instances by Server Load Balancer (SLB) instances is also filtered.
- Network ACLs are stateless. You must set both inbound and outbound rules. Otherwise, the system may fail to respond to requests.
- If you create a network ACL that does not contain a rule, all inbound and outbound access is denied.
- If a network ACL is associated with a vSwitch, the network ACL does not filter the traffic forwarded between ECS instances that are associated with the vSwitch.

Network ACL rules

You can add rules to or delete rules from a network ACL. Changes to the rules are automatically synchronized to the associated vSwitch. By default, an inbound rule and an outbound rule are automatically added to a newly created network ACL. These rules allow all inbound and outbound network traffic transmitted through the associated vSwitch. You can delete the default rules. The following table lists the default inbound and outbound rules.

- Default inbound rule

Priority	Protocol type	Source CIDR block	Source port range	Action	Type
1	all	0.0.0.0/0	-1/-1	Allow	Custom

- Default outbound rule

Priority	Protocol type	Destination CIDR block	Destination port range	Action	Type
1	all	0.0.0.0/0	-1/-1	Allow	Custom

A network ACL contains the following parameters:

- Priority: A smaller value indicates a higher priority. The system attempts to match requests against rules in descending order of priority. Rule 1 has the highest priority. If a request matches a rule, the system applies the rule to the request and ignores the other rules.

  For example, the following rules are added and requests destined for IP address 172.16.0.1 are sent from an ECS instance. In the following table, the requests match Rules 2 and 3. Rule 2 has a higher priority than Rule 3. Therefore, the system applies Rule 2. Based on the action of Rule 2, the requests are denied.

Priority	Protocol type	Destination CIDR block	Destination port range	Action	Type
1	all	10.0.0.0/8	-1/-1	Allow	Custom
2	all	172.16.0.0/12	-1/-1	Deny	Custom
3	all	172.16.0.0/12	-1/-1	Allow	Custom

- Policy: specifies whether to allow or deny specific traffic.
- Protocol: the protocol type. Available options include All, ICMP, GRE, TCP, and UDP.
- Source CIDR block: the source CIDR block from which inbound traffic is transmitted.
- Destination IP address: the destination IP addresses to which outbound traffic is transmitted.
- Destination port range: the range of destination ports to which the inbound rule applies.
- Destination port range: the range of destination ports to which the outbound rule applies.

Comparison between network ACLs and security groups

Network ACLs control data transmitted through associated vSwitches while security groups control data transmitted through associated ECS instances. The following table describes the differences between network ACLs and security groups.

Network ACL	Security group
Applies to vSwitches.	Applied to instances.
Stateless: Returned traffic must be allowed by inbound rules.	Stateful: Returned traffic is automatically allowed and not affected by rules.
The system attempts to match requests against rules in descending order of priority. Not all rules are matched.	The system matches a request against all rules before a rule is applied.
Each vSwitch can be associated only with one network ACL.	Each ECS instance can be added to more than one security group.

The following figure shows how network ACLs and security groups are applied to ensure network security.

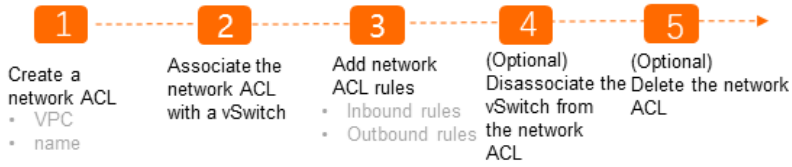
Limits

The following table describes the limits on network ACLs.

Item	Limit
Maximum number of network ACLs that can be created in each VPC	200
Maximum number of network ACLs that can be associated with a vSwitch	1
Maximum number of rules that can be added to a network ACL	• Inbound rules: 20 • Outbound rules: 20

Procedure

The following flowchart shows how to use a network ACL.



18.1.7.2. Scenarios

If you are familiar with the ports that are commonly used by ECS instances, you can specify them in access control list (ACL) rules to facilitate precise network traffic filtering. This topic describes the ports that are commonly used by ECS instances and the application scenarios of these ports.

Ports

The following table lists the ports and the services that use these ports.

Port	Service	Description
21	FTP	The FTP port. It is used to upload and download files.
22	SSH	The SSH port. It is used to log on to Linux instances in the command line method by using username and password pairs.
23	Telnet	The Telnet port. It is used to remotely log on to ECS instances.
25	SMTP	The SMTP port. It is used to send emails.
80	HTTP	The HTTP port. It is used to access services such as IIS, Apache, and NGINX.
110	POP3	The POP3 port. It is used to send and receive emails.
143	IMAP	The Internet Message Access Protocol (IMAP) port. It is used to receive emails.
443	HTTPS	The HTTPS port. It is used to access services. The HTTPS protocol can implement encrypted and secure data transmission.
1433	SQL Server	The TCP port of SQL Server. It is used for SQL Server to provide external services.

Port	Service	Description
1434	SQL Server	The UDP port of SQL Server. It is used to return the TCP/IP port occupied by SQL Server.
1521	Oracle	The Oracle communication port. ECS instances that run Oracle SQL must have this port open.
3306	MySQL	The MySQL port. It is used for MySQL databases to provide external services.
3389	Windows Server Remote Desktop Services	The Windows Server Remote Desktop Services port. It is used to log on to a Windows instance.
8080	Proxy port	An alternative to port 80. It is commonly used for WWW proxy services.

Custom network ACLs

Inbound rules and Outbound rules describe a network ACL example for VPCs that support IPv4 addresses only.

- The inbound rules in effective order 1, 2, 3, and 4 respectively allow HTTP, HTTPS, SSH, and RDP traffic to the vSwitch. Outbound response rules are those in effective order 3.
- The outbound rules in effective order 1 and 2 respectively allow HTTP and HTTPS traffic from the vSwitch. Outbound response rules are those in effective order 5.
- The inbound rule in effective order 6 denies all inbound IPv4 traffic. This rule ensures that packets that do not match any other rules are denied.
- The outbound rule in effective order 4 denies all outbound IPv4 traffic. This rule ensures that packets that do not match any other rules are denied.

> **Note** An inbound or outbound rule must correspond to an inbound or outbound rule that allows response traffic.

Inbound rules

Effective order	Protocol	Source IP addresses	Destination port range	Action	Description
1	TCP	0.0.0.0/0	80/80	Accept	Allows inbound HTTP traffic from any IPv4 addresses.
2	TCP	0.0.0.0/0	443/443	Accept	Allows inbound HTTPS traffic from any IPv4 addresses.
3	TCP	0.0.0.0/0	22/22	Accept	Allows inbound SSH traffic from any IPv4 addresses.
4	TCP	0.0.0.0/0	3389/3389	Accept	Allows inbound RDP traffic from any IPv4 addresses.
5	TCP	0.0.0.0/0	32768/65535	Accept	Allows inbound IPv4 traffic from the Internet. This port range is for reference only. For more information on how to select appropriate ephemeral ports, see Ephemeral ports.
6	All	0.0.0.0/0	-1/-1	Drop	Denies all inbound IPv4 traffic.

Outbound rules

Effective order	Protocol	Destination IP addresses	Destination port range	Action	Description
1	TCP	0.0.0.0/0	80/80	Accept	Allows outbound IPv4 HTTP traffic from the vSwitch to the Internet.
2	TCP	0.0.0.0/0	443/443	Accept	Allows outbound IPv4 HTTPS traffic from the vSwitch to the Internet.
3	TCP	0.0.0.0/0	32768/65535	Accept	Allows outbound IPv4 traffic from the vSwitch to the Internet. This port range is for reference only. For more information on how to select appropriate ephemeral ports, see Ephemeral ports.
4	All	0.0.0.0/0	-1/-1	Drop	Denies all outbound IPv4 traffic.

Network ACLs for SLB

If the ECS instance in the vSwitch acts as the backend server of an SLB instance, you must add the following network ACL rules.

- Inbound rules

Effective order	Protocol	Source IP addresses	Destination port range	Action	Description
1	SLB listener protocol	Client IP addresses allowed to access the SLB instance	SLB listener port	Accept	Allows inbound traffic from specified client IP addresses.
2	Health check protocol	100.64.0.0/10	Health check port	Accept	Allows inbound traffic from health check IP addresses.

- Outbound rules

Effective order	Protocol	Destination IP addresses	Destination port range	Action	Description
1	All	Client IP addresses allowed to access the SLB instance	-1/-1	Accept	Allows all outbound traffic to specified client IP addresses.

Effective order	Protocol	Destination IP addresses	Destination port range	Action	Description
2	All	100.64.0.0/10	-1/-1	Accept	Allows outbound traffic to health check IP addresses.

Ephemeral ports

Clients use different ports to initiate requests. You can select different port ranges for network ACL rules based on the client type. The following table lists ephemeral port ranges for common clients.

Client	Port range
Linux	32768/61000
Windows Server 2003	1025/5000
Windows Server 2008 and later	49152/65535
NAT gateway	1024/65535

18.1.7.3. Create a network ACL

A network access control list (ACL) allows you to manage network access in a virtual private cloud (VPC). You can create network ACLs in the VPC console.

Prerequisites

A VPC is created. For more information, see Create a VPC.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where you want to create the network ACL.
4. On the **Network ACL** page, click **Create Network ACL**.
5. In the **Create Network ACL** dialog box, set the following parameters and click **OK**.

Parameter	Description
Organization	Select the organization to which the network ACL belongs.
Resource Set	Select the resource set to which the network ACL belongs.
Region	Select the region where you want to deploy the network ACL.
Name	The name of the network ACL. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter.

Parameter	Description
Description	Enter a description for the network ACL. The description must be 2 to 256 characters in length, and can contain digits, underscores (_), colons (:), and hyphens (-). The name must start with a letter and cannot start with `http://` or `https://`.
VPC	Select the VPC for which you want to create the network ACL. **Note** The VPC and network ACL must be deployed in the same region.

What's next

- Associate a network ACL with a VSwitch
- Add an inbound rule
- Add an outbound rule

18.1.7.4. Associate a network ACL with a vSwitch

After you create a network access control list (ACL), you can associate it with a vSwitch. This way, you can use the network ACL to control the traffic that flows through the Elastic Compute Service (ECS) instances that are connected to the vSwitch.

Prerequisites

Before you associate a network ACL with a vSwitch, make sure that the following requirements are met:

- A network ACL is created. For more information, see Create a network ACL.
- A vSwitch is created. The vSwitch and network ACL must belong to the same VPC. For more information, see Create a vSwitch.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region to which the network ACK that you want to manage belongs.
4. On the **Network ACL** page, find the network ACL that you want to manage and click **Manage** in the **Actions** column.
5. On the **Resources** tab, click **Bind Resource**.
6. In the **Bind Resource** panel, select the vSwitch that you want to associate and click **OK**.

Note The network ACL and vSwitch must belong to the same VPC. A vSwitch can be associated only with one network ACL.

18.1.7.5. Add network ACL rules

18.1.7.5.1. Add an inbound rule

This topic describes how to add an inbound rule to a network access control list (ACL). You can use inbound rules to control Internet or internal network traffic destined for Elastic Compute Service (ECS) instances connected to a vSwitch.

Prerequisites

A network ACL is created. For more information, see Work with network ACLs.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where the network ACL is created.
4. On the **Network ACL** page, find the network ACL that you want to manage and click **Inbound Rule** in the **Actions** column.
5. On the **Inbound Rule** tab, click **Create Inbound Rule**.
6. In the **Create Inbound Rule** panel, set the following parameters and click **OK**.

Parameter	Description
Name	Enter a name for the inbound rule. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter. It cannot start with `http://` or `https://` .
Effective order	The order in which the inbound rule takes effect. Valid values: 1 to 20. A smaller number indicates a higher priority.
Action	Select an action for the inbound rule. Valid values: ○ **Accept**: accepts network traffic that is destined for the ECS instances connected to the vSwitch. ○ **Drop**: denies network traffic that is destined for the ECS instances connected to the vSwitch.
Protocol Type	Select a transport layer protocol. Valid values: ○ **ALL** ○ **ICMP** ○ **GRE** ○ **TCP** ○ **UDP**
Source IP Addresses	The source CIDR block from which data is transmitted. Default value: 0.0.0.0/32.
Destination Port Range	Enter the destination port range. Valid values: 1 to 65535. Use a forward slash (/) to separate the highest and lowest values in a port range, for example, 1/200 or 80/80. -1/-1 indicates that all ports are available. You cannot set the value to -1/-1.

18.1.7.5.2. Add an outbound rule

This topic describes how to add an outbound rule to a network access control list (ACL). You can use outbound rules to control how Elastic Compute Service (ECS) instances connected to a vSwitch access the Internet or other internal networks.

Prerequisites

A network ACL is created. For more information, see Work with network ACLs.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where the network ACL is created.
4. On the **Network ACL** page, find the network ACL and click **Outbound Rule** in the **Actions** column.
5. On the **Outbound Rule** tab, click **Create Outbound Rule**.
6. In the **Create Outbound Rule** panel, set the following parameters and click **OK**.

Parameter	Description
Name	Enter a name for the outbound rule. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), and hyphens (-). The name must start with a letter. It cannot start with `http://` or `https://`.
Effective order	The order in which the outbound rule takes effect. Valid values: 1 to 20. A smaller number indicates a higher priority.
Policy	Select an action for the outbound rule. Valid values: ○ **Accept**: allows ECS instances connected to the vSwitch to access the Internet or other internal networks. ○ **Drop**: forbids ECS instances connected to the vSwitch to access the Internet or other internal networks.
Protocol Type	Select a transport layer protocol. Valid values: ○ **ALL** ○ **ICMP** ○ **GRE** ○ **TCP** ○ **UDP**
Destination IP Addresses	The destination CIDR block to which data is transmitted. Default value: 0.0.0.0/32.
Destination Port Range	Enter the destination port range. Valid values: 1 to 65535. Use a forward slash (/) to separate the highest and lowest values in a port range, for example, 1/200 or 80/80. -1/-1 indicates that all ports are available. You cannot set the value to -1/-1.

18.1.7.5.3. Change the priority of a network ACL rule

Rules added to network access control lists (ACLs) take effect in descending order of priority. A smaller value indicates a higher priority. You can change the priority of a network ACL rule to meet your business requirements.

Change the priority of an inbound rule

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where the network ACL is created.
4. On the **Network ACL** page, find the network ACL that you want to manage and click **Manage** in the **Actions** column.
5. Click the **Inbound Rule** tab and click **Sort**.
6. In the **Sort** panel, change the priority of the rule by dragging and dropping the rule. Then, click **OK**.

> **Note** Rules are listed in descending order of priority.

Change the priority of an outbound rule

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where the network ACL is created.
4. On the **Network ACL** page, find the network ACL and click **Manage** in the **Actions** column.
5. Click the **Outbound Rule** tab and click **Sort**.
6. In the **Sort** panel, change the priority of the rule by dragging and dropping the rule. Then, click **OK**.

> **Note** Rules are listed in descending order of priority.

18.1.7.6. Disassociate a network ACL from a vSwitch

This topic describes how to disassociate a network access control list (ACL) from a vSwitch. After you disassociate a network ACL from a vSwitch, the network ACL no longer controls the traffic of Elastic Compute Service (ECS) instances that belong to the vSwitch.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where the network ACL is created.
4. On the **Network ACL** page, find the network ACL and click **Manage** in the **Actions** column.
5. On the **Resources** tab, find the vSwitch and click **Unbind** in the **Actions** column.
6. In the **Unbind Network ACL** message, click **OK**.

18.1.7.7. Delete a network ACL

This topic describes how to delete a network access control list (ACL).

Prerequisites

Make sure that the network ACL is not associated with a vSwitch. If the network ACL is associated with a vSwitch, disassociate the network ACL from the vSwitch first. For more information, see Disassociate a VSwitch from a network ACL.

Procedure

1. Log on to the VPC console.
2. In the left-side navigation pane, click **Network ACL**.
3. In the top navigation bar, select the region where the network ACL is created.

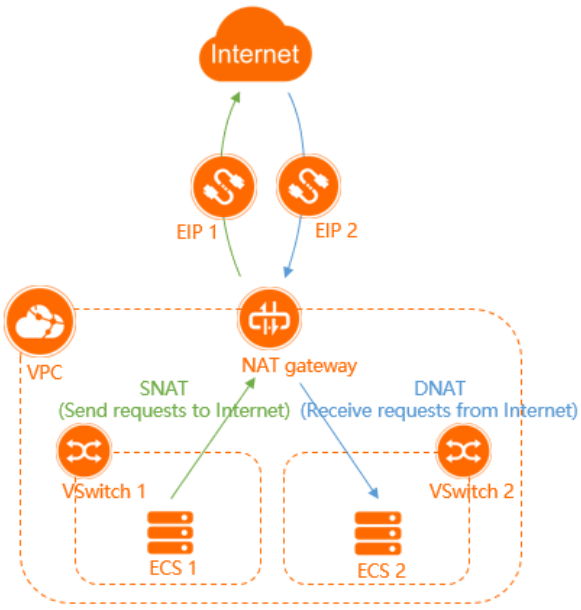
4. On the **Network ACL** page, find the network ACL that you want to delete and click **Delete** in the **Actions** column.
5. In the **Delete Network ACL** message, click **OK**.

19.NAT Gateway

19.1. User Guide

19.1.1. What is NAT Gateway?

NAT gateways are enterprise-class gateways that provide the Source Network Address Translation (SNAT) and Destination Network Address Translation (DNAT) features. Each NAT gateway provides a throughput capacity of up to 100 Gbit/s. NAT gateways also support cross-zone disaster recovery.



Features

You must associate public IP addresses with NAT gateways so that the NAT gateways can function as expected. After you create a NAT gateway, you can associate elastic IP addresses (EIPs) with the NAT gateway.

NAT gateways provide the SNAT and DNAT features.

Feature	Description
SNAT	SNAT allows ECS instances that are deployed in a VPC to access the Internet when no public IP address is associated with these ECS instances.
DNAT	DNAT maps the EIPs of a NAT gateway to ECS instances. This way, the ECS instances can receive requests from the Internet.

19.1.2. Log on to the NAT Gateway console

This topic describes how to log on to the Apsara Uni-manager Management Console to manage your NAT gateways. The Google Chrome browser is used as an example.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.
5. In the top navigation bar, choose **Products > Networking > Virtual Private Cloud**.
6. In the left-side navigation pane, choose **Internet Access > NAT Gateway**.

19.1.3. Quick Start

19.1.3.1. Overview

This topic describes how to configure SNAT and DNAT. You can configure SNAT and DNAT to enable Elastic Compute Service (ECS) instances in a virtual private cloud (VPC) to communicate with the Internet through a NAT gateway.

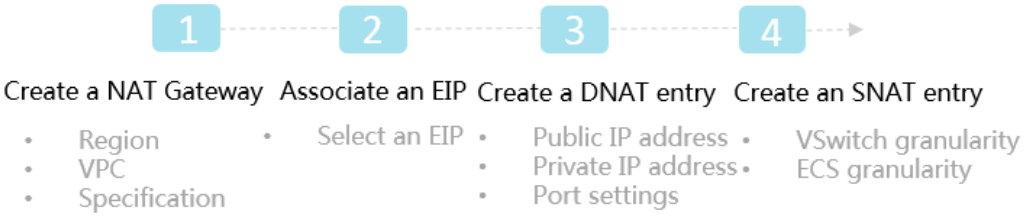
Prerequisites

Before you start, make sure that the following conditions are met:

- A VPC is created. For more information, see the **Create a VPC** topic in the **Quick Start** chapter of *VPC User Guide*.
- An ECS instance is created in the VPC. For more information, see the **Create an ECS instance** topic in the **Quick Start** chapter of *ECS User Guide*.
- An elastic IP address (EIP) is created. For more information, see the **Create an EIP** topic in the **Quick Start** chapter of *EIP User Guide*.

Procedure

The ECS instance used as an example in this topic is deployed in a VPC but is not assigned a public IP address. The following flowchart shows the configuration process.



1. Create a NAT gateway

   NAT gateways are enterprise-class gateways that provide network address translation services for accessing the Internet and VPCs. Before you configure SNAT and DNAT rules, you must create a NAT gateway.

   For more information, see Create a NAT gateway.

2. Associate an EIP with the NAT gateway

   A NAT gateway works as expected only after it is associated with an EIP. After a NAT gateway is created, you must associate an EIP with the NAT gateway.

   For more information, see Associate an EIP with a NAT gateway.

3. Create a DNAT entry

   NAT gateways support the DNAT feature. This feature allows you to map the public IP addresses of NAT gateways to ECS instances. This way, the ECS instances can provide Internet-facing services. DNAT supports port mapping and IP mapping.

   For more information, see Create a DNAT entry.

4. Create an SNAT entry

   NAT gateways support the SNAT feature. This feature allows ECS instances that are not assigned public IP addresses in a VPC to access the Internet through a NAT gateway.

   For more information, see Create an SNAT entry.

19.1.3.2. Create a NAT gateway

NAT gateways are enterprise-class gateways that provide network address translation services for accessing the Internet and virtual private clouds (VPCs). Before you configure SNAT and DNAT rules, you must create a NAT gateway.

Prerequisites

A VPC is created. For more information, see the **Create a VPC** topic in the **Quick Start** chapter of *VPC User Guide*.

Procedure

1. Log on to the NAT Gateway console.
2. On the **NAT Gateway** page, click **Create NAT Gateway**.
3. Set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the NAT gateway belongs.
Resource Set	Select the resource group to which the NAT gateway belongs.
Region	Select the region where you want to create the NAT gateway.

Parameter	Description
VPC	Select the VPC to which the NAT gateway belongs. If you cannot find the VPC in the list, perform the following operations: ○ Check whether the VPC is already associated with a NAT gateway. Each VPC can be associated only with one NAT gateway. ○ Check whether the VPC contains a custom route entry whose destination CIDR block is 0.0.0.0/0. If the custom route entry exists, delete it. ○ If your account is a Resource Access Management (RAM) user, check whether the RAM user is authorized to access the VPC. If the RAM user is unauthorized, contact the owner of the Apsara Stack tenant account that created the RAM user to grant permissions.
Sharing Scope	Select the sharing scope of the VPC. ○ **Current Resource Set**: Only the administrator of the current resource set can use the VPC to create resources. ○ **Current Organization and Subordinate Organization**: Only the administrators of the current organization and its subordinate organization can create resources in the shared VPC. ○ **Current Organization**: Only the administrator of the current organization can use create resources in the shared VPC.
Specification	Select the size of the NAT gateway. Valid values: ○ **Small**: supports at most 10,000 SNAT connections. ○ **Medium**: supports at most 50,000 SNAT connections. ○ **Large**: supports at most 200,000 SNAT connections. ○ **Super Large-1**: supports at most 1,000,000 SNAT connections. **Note** The maximum number of SNAT connections is limited by the size of a NAT gateway. However, the gateway size does not affect the maximum number of DNAT connections.
Name	Enter a name for the NAT gateway. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), hyphens (-), and periods (.). The name must start with a letter and cannot start with `http://` or `https://`.

19.1.3.3. Associate an EIP with a NAT gateway

This topic describes how to associate an elastic IP address (EIP) with a NAT gateway. To enable Internet access for a NAT gateway, you must associate an EIP with the NAT gateway. After you create a NAT gateway, you can associate EIPs with the NAT gateway.

Prerequisites

Before you associate an EIP with a NAT gateway, make sure that the following requirements are met:

- A NAT gateway is created. For more information, see Create a NAT gateway.
- An EIP is created. For more information, see the **Apply for EIPs** topic in the **Quick Start** chapter of *EIP User Guide*.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway with which you want to associate an EIP and choose ⋮ > **Bind Elastic IP Address** in the **Actions** column.
4. In the **Bind Elastic IP Address** panel, set the following parameters and click **OK**.

Parameter	Description
Usable EIP list	Select the EIP that is used to access the Internet.
vSwitch	Select the vSwitch for which you want to add SNAT entries. After you select a vSwitch, the system automatically adds SNAT entries for the vSwitch. Then, cloud services attached to the vSwitch can access the Internet. You can skip this step and manually add SNAT entries after you associate an EIP with the NAT gateway. For more information, see Create an SNAT entry.

19.1.3.4. Create a DNAT entry

This topic describes how to create a DNAT entry. DNAT can map public IP addresses of NAT gateways to Elastic Compute Service (ECS) instances. This way, the ECS instances can provide Internet-facing services. DNAT supports port mapping and IP mapping.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure DNAT** in the **Actions** column.
4. On the **DNAT Table** page, click **Create DNAT Entry**.
5. In the **Create DNAT Entry** panel, set the following parameters and click **OK**.

Parameter	Description
Public IP Address	Select an available public IP address. **Note** If a public IP address is already used in an SNAT entry, the public IP address cannot be used in a DNAT entry.
Private IP Address	Specify the ECS instance that uses the DNAT entry to access the Internet. You can specify the private IP address of the ECS instance in the following ways: ○ **Auto Fill**: Select the ECS instance from the drop-down list. ○ **Manually Input**: Enter the private IP address of the ECS instance. **Note** This private IP address must fall within the CIDR block of the virtual private cloud (VPC). You can also enter the private IP address of an existing ECS instance.

Parameter	Description
Port Settings	Select a DNAT mapping method: ○ **All**: This method uses IP mapping. All requests destined for the elastic IP address (EIP) are forwarded to the ECS instance. ○ **Specific Port**: This method uses port mapping. The NAT gateway forwards requests that use the specified protocol and port to the specified port of the ECS instance. After you select Specific Port, specify the **Public Port** (the external port), **Private Port** (the internal port), and **IP Protocol** (the protocol over which data is transferred).
Entry Name	Enter a name for the DNAT entry. The name must be 2 to 128 characters in length, and can contain digits, underscores (_), and hyphens (-). It must start with a letter.

19.1.3.5. Create an SNAT entry

This topic describes how to create an SNAT entry. SNAT can provide proxy services for Elastic Compute Service (ECS) instances. ECS instances that do not have public IP address assigned in virtual private clouds (VPCs) can access the Internet by using SNAT.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure SNAT** in the **Actions** column.
4. On the **SNAT Table** page, click **Create SNAT Entry**.
5. In the **Create SNAT Entry** panel, set the following parameters and click **OK**.

Parameter	Description
VSwitch Granularity	
VSwitch	Select a vSwitch in the VPC. All ECS instances in the vSwitch can access the Internet by using SNAT. **Note** SNAT entries do not take effect on ECS instances that are assigned public IP addresses. For example, an ECS instance may be assigned a static public IP address, associated with an elastic IP address (EIP), or configured with DNAT IP mapping. In this case, the ECS instance uses the public IP address instead of the SNAT entry to access the Internet.
VSwitch CIDR Block	The CIDR block of the selected vSwitch.

Parameter	Description
Public IP Address	Select an EIP that is used to access the Internet. You can select one or more EIPs to create an SNAT IP address pool. **Note** An EIP that is already used in a DNAT entry cannot be used in an SNAT entry.
ECS Granularity	
Available ECS Instances	Select an ECS instance in the VPC. The ECS instance can access the Internet by using the specified EIP. Make sure that the following requirements are met: ○ The ECS instance is in the Running state. ○ The ECS instance is not assigned an EIP or a static public IP address.
ECS CIDR Block	The CIDR block of the ECS instance.
Public IP Address	Select an EIP that is used to access the Internet. **Note** An EIP that is already used in a DNAT entry cannot be used in an SNAT entry.

19.1.4. Manage a NAT gateway

19.1.4.1. Overview

NAT gateways are enterprise-class gateways that support SNAT and DNAT features. Each NAT gateway provides a forwarding capacity of 10 Gbit/s. NAT gateways support cross-zone disaster recovery.

Sizes of NAT gateways

NAT gateways are available in multiple sizes, including small, middle, large, and super large-1. The size of a NAT gateway determines the SNAT performance, which includes the maximum number of connections and the number of new connections per second. However, the size of a NAT gateway does not affect the DNAT performance. The following table describes available sizes of NAT gateways.

Size	Maximum number of SNAT connections	Number of new SNAT connections per second
Small	10,000	1,000
Medium	50,000	5,000
Large	200,000	10,000
Super Large-1	1,000,000	50,000

When you specify the size of a NAT gateway, take note of the following limits:

- CloudMonitor monitors only the maximum number of SNAT connections for NAT gateways. CloudMonitor does not monitor the number of new SNAT connections per second.

- The timeout period of SNAT connections on NAT gateways is 900 seconds.
- To avoid timeouts of SNAT connections caused by network congestion or Internet jitter, make sure that your applications support automatic reconnection.
- NAT gateways do not support packet fragmentation.
- If you use the same destination public IP address and port, the maximum number of concurrent connections is based on the number of elastic IP addresses (EIPs) that are associated with the NAT gateway. Each EIP that is associated with the NAT gateway supports up to 55,000 concurrent connections. If N EIPs are associated with the NAT gateway, the maximum number of concurrent connections that the NAT gateway supports is N × 55,000.

19.1.4.2. Create a NAT gateway

NAT gateways are enterprise-class gateways that provide network address translation services for accessing the Internet and virtual private clouds (VPCs). Before you configure SNAT and DNAT rules, you must create a NAT gateway.

Prerequisites

A VPC is created. For more information, see the **Create a VPC** topic in the **Quick Start** chapter of *VPC User Guide*.

Procedure

1. Log on to the NAT Gateway console.
2. On the **NAT Gateway** page, click **Create NAT Gateway**.
3. Set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the NAT gateway belongs.
Resource Set	Select the resource group to which the NAT gateway belongs.
Region	Select the region where you want to create the NAT gateway.
VPC	Select the VPC to which the NAT gateway belongs. If you cannot find the VPC in the list, perform the following operations: ○ Check whether the VPC is already associated with a NAT gateway. Each VPC can be associated only with one NAT gateway. ○ Check whether the VPC contains a custom route entry whose destination CIDR block is 0.0.0.0/0. If the custom route entry exists, delete it. ○ If your account is a Resource Access Management (RAM) user, check whether the RAM user is authorized to access the VPC. If the RAM user is unauthorized, contact the owner of the Apsara Stack tenant account that created the RAM user to grant permissions.
Sharing Scope	Select the sharing scope of the VPC. ○ **Current Resource Set**: Only the administrator of the current resource set can use the VPC to create resources. ○ **Current Organization and Subordinate Organization**: Only the administrators of the current organization and its subordinate organization can create resources in the shared VPC. ○ **Current Organization**: Only the administrator of the current organization can use create resources in the shared VPC.

Parameter	Description
Specification	Select the size of the NAT gateway. Valid values: ○ **Small**: supports at most 10,000 SNAT connections. ○ **Medium**: supports at most 50,000 SNAT connections. ○ **Large**: supports at most 200,000 SNAT connections. ○ **Super Large-1**: supports at most 1,000,000 SNAT connections. **Note** The maximum number of SNAT connections is limited by the size of a NAT gateway. However, the gateway size does not affect the maximum number of DNAT connections.
Name	Enter a name for the NAT gateway. The name must be 2 to 128 characters in length, and can contain letters, digits, underscores (_), hyphens (-), and periods (.). The name must start with a letter and cannot start with `http://` or `https://`.

19.1.4.3. Modify a NAT gateway

You can modify the name and description of a NAT gateway.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to modify and click **Manage** in the **Actions** column.
4. On the **Basic Information** page, click **Edit** next to **Instance Name**. In the dialog box that appears, enter a name for the NAT gateway and click **OK**.

   The name must be 2 to 128 characters in length and can contain digits, underscores (_), and hyphens (-). The name must start with a letter.
5. Click **Edit** next to **Description**. In the dialog box that appears, enter a description for the NAT gateway and click **OK**.

   The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

19.1.4.4. Delete a NAT gateway

You can delete a NAT gateway that is no longer in use.

Prerequisites

Before you delete a NAT gateway, make sure that the following requirements are met:

- No elastic IP address (EIP) is associated with the NAT gateway. If an EIP is associated with the NAT gateway, disassociate the EIP from the NAT gateway first. For more information, see Disassociate an EIP from a NAT gateway.
- The DNAT table does not contain DNAT entries. If the DNAT table contains DNAT entries, delete them. For more information, see Delete a DNAT entry.
- The SNAT table does not contain SNAT entries. If the SNAT table contains SNAT entries, delete them. For more information, see Delete an SNAT entry.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to delete and choose ⋮ > **Delete** in the **Actions** column.
4. In the dialog box that appears, click **OK**.

> **Note** If you select **Delete (Delete NAT gateway and resources)**, the DNAT and SNAT entries of the NAT gateway are automatically deleted. The EIPs associated with the NAT gateway are also disassociated.

19.1.4.5. Manage tags

You can mark and classify NAT gateways by adding tags to NAT gateways. This allows you to search and filter NAT gateways in a more efficient way.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway to which you want to add tags, move the pointer over the icon in the **Tag** column, and then click **Add**.
4. In the **Configure Tags** dialog box, set the following parameters and click **OK**.

Parameter	Description
Tag Key	The key of the tag. You can select or enter a key. The key must be 1 to 64 characters in length, and cannot start with `aliyun` or `acs:`. It cannot contain `http://` or `https://`.
Tag Value	The value of the tag. You can select or enter a value. The value cannot exceed 128 characters in length, and cannot start with `aliyun` or `acs:`. It cannot contain `http://` or `https://`.

5. Return to the **NAT Gateway** page and click **Filter by Tag**. In the **Filter by Tag** dialog box, you can specify tag keys and values to filter NAT gateways.

19.1.5. Manage EIPs

19.1.5.1. Associate an EIP with a NAT gateway

This topic describes how to associate an elastic IP address (EIP) with a NAT gateway. To enable Internet access for a NAT gateway, you must associate an EIP with the NAT gateway. After you create a NAT gateway, you can associate EIPs with the NAT gateway.

Prerequisites

Before you associate an EIP with a NAT gateway, make sure that the following requirements are met:

- A NAT gateway is created. For more information, see Create a NAT gateway.
- An EIP is created. For more information, see the **Apply for EIPs** topic in the **Quick Start** chapter of *EIP User Guide*.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway with which you want to associate an EIP and choose ⋮ > **Bind Elastic IP Address** in the **Actions** column.
4. In the **Bind Elastic IP Address** panel, set the following parameters and click **OK**.

Parameter	Description
Usable EIP list	Select the EIP that is used to access the Internet.
vSwitch	Select the vSwitch for which you want to add SNAT entries. After you select a vSwitch, the system automatically adds SNAT entries for the vSwitch. Then, cloud services attached to the vSwitch can access the Internet. You can skip this step and manually add SNAT entries after you associate an EIP with the NAT gateway. For more information, see Create an SNAT entry.

19.1.5.2. Disassociate an EIP from a NAT gateway

This topic describes how to disassociate an elastic IP address (EIP) from a NAT gateway. After an EIP is disassociated from a NAT gateway, the NAT gateway can no longer access the Internet by using the EIP.

Prerequisites

Make sure that the EIP to be disassociated is not used in an SNAT entry or a DNAT entry. If the EIP is used in an SNAT or a DNAT entry, delete the SNAT or DNAT entry first. For more information, see Delete an SNAT entry and Delete a DNAT entry.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and choose ⋮ > **Unbind Elastic IP Address** in the **Actions** column.
4. In the **Unbind Elastic IP Address** panel, select the EIP that you want to disassociate, and click **OK**.

19.1.6. Manage a DNAT table

19.1.6.1. DNAT overview

NAT Gateway supports the DNAT feature. You can create DNAT entries to map public IP addresses to Elastic Compute Service (ECS) instances in a virtual private cloud (VPC). This way, the ECS instances can provide Internet-facing services.

DNAT entries

You can configure port mapping when you create a DNAT entry. After the DNAT entry is created, requests destined for the specified public IP address are forwarded to the ECS instances within a VPC based on the port mapping rule.

Each DNAT entry consists of the following items:

- **Public IP address**: the elastic IP address (EIP) associated with the NAT gateway.

- **Private IP address**: the private IP address assigned to the ECS instance in the VPC.
- **Public Port**: the external port on which requests from the Internet are received.
- **Private Port**: the internal port to which the requests received on the external port are forwarded.
- **IP Protocol**: the protocol used by the ports.

Port mapping and IP mapping

The DNAT feature supports port mapping and IP mapping:

- Port mapping

  After you configure port mapping for a NAT gateway, the NAT gateway forwards the requests that are destined for the specified public IP address to the specified ECS instance. The requests are forwarded based on the specified source and destination port and the specified protocol. The DNAT entries in the following table are used as an example:

  - Entry 1: The NAT gateway forwards requests that are destined for 1.1.XX.XX and TCP port 80 to 192.168.1.1 and TCP port 80.
  - Entry 2:The NAT gateway forwards requests that are destined for 2.2.XX.XX and UDP port 8080 to 192.168.1.2 and TCP port 8000.

DNAT entry	Public IP address	Public port	Private IP address	Private port	Protocol
Entry 1	1.1.XX.XX	80	192.168.1.1	80	TCP
Entry 2	2.2.XX.XX	8080	192.168.1.2	8000	UDP

- IP mapping

  After you configure IP mapping for a NAT gateway, the NAT gateway forwards all the requests that are destined for the specified public IP address to the specified ECS instance. The DNAT entry in the following table is used as an example.The NAT gateway forwards all the requests that are destined for 3.3.XX.XX to the ECS instance whose IP address is 192.168.1.3.

DNAT entry	Public IP address	Public port	Private IP address	Private port	Protocol
Entry 3	3.3.XX.XX	All	192.168.1.3	All	All

19.1.6.2. Create a DNAT entry

This topic describes how to create a DNAT entry. DNAT can map public IP addresses of NAT gateways to Elastic Compute Service (ECS) instances. This way, the ECS instances can provide Internet-facing services. DNAT supports port mapping and IP mapping.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure DNAT** in the **Actions** column.
4. On the **DNAT Table** page, click **Create DNAT Entry**.
5. In the **Create DNAT Entry** panel, set the following parameters and click **OK**.

Parameter	Description
Public IP Address	Select an available public IP address. **Note** If a public IP address is already used in an SNAT entry, the public IP address cannot be used in a DNAT entry.
Private IP Address	Specify the ECS instance that uses the DNAT entry to access the Internet. You can specify the private IP address of the ECS instance in the following ways: ○ **Auto Fill**: Select the ECS instance from the drop-down list. ○ **Manually Input**: Enter the private IP address of the ECS instance. **Note** This private IP address must fall within the CIDR block of the virtual private cloud (VPC). You can also enter the private IP address of an existing ECS instance.
Port Settings	Select a DNAT mapping method: ○ **All**: This method uses IP mapping. All requests destined for the elastic IP address (EIP) are forwarded to the ECS instance. ○ **Specific Port**: This method uses port mapping. The NAT gateway forwards requests that use the specified protocol and port to the specified port of the ECS instance. After you select Specific Port, specify the **Public Port** (the external port), **Private Port** (the internal port), and **IP Protocol** (the protocol over which data is transferred).
Entry Name	Enter a name for the DNAT entry. The name must be 2 to 128 characters in length, and can contain digits, underscores (_), and hyphens (-). It must start with a letter.

19.1.6.3. Modify a DNAT entry

This topic describes how to modify a DNAT entry. You can modify the public IP address, private IP address, and port in a DNAT entry.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure DNAT** in the **Actions** column.
4. On the **DNAT Table** page, find the DNAT entry that you want to modify and click **Edit** in the **Actions** column.
5. In the **Edit DNAT Entry** panel, modify the public IP address, private IP address, or port in the DNAT entry, and then click **OK**.

19.1.6.4. Delete a DNAT entry

This topic describes how to delete a DNAT entry. If you no longer need an Elastic Compute Service (ECS) instance to serve Internet-facing applications, you can delete the DNAT entry created for the ECS instance.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure DNAT** in the **Actions** column.
4. On the **DNAT Table** page, find the DNAT entry that you want to delete and click **Delete** in the **Actions** column.
5. In the message that appears, click **OK**.

19.1.7. Manage an SNAT table

19.1.7.1. SNAT table overview

This topic describes how to configure SNAT. SNAT allows Elastic Compute Service (ECS) instances that are not assigned public IP addresses in a virtual private cloud (VPC) to access the Internet through a NAT gateway.

SNAT entries

You can add SNAT entries to an SNAT table to allow ECS instances to access the Internet.

Each SNAT entry consists of the following items:

- **vSwitches or ECS instances**: the vSwitch or ECS instances that require Internet access by using SNAT.
- **Public IP address**: the public IP address used to access the Internet.

vSwitch granularity and ECS granularity

SNAT entries can be created based on the following granularity to enable ECS instances in a VPC to access the Internet.

- vSwitch granularity

  If you create an SNAT entry for a vSwitch, the ECS instances attached to the vSwitch access the Internet by using the public IP address specified in the SNAT entry and through the NAT gateway on which the SNAT entry is created. By default, all ECS instances attached to the vSwitch can use the specified public IP address to access the Internet.

  > **Note** If an ECS instance has a public IP address, for example, the ECS instance is assigned a static public IP address, associated with an elastic IP address (EIP), or has DNAT IP mapping configured, the ECS instance uses the public IP address to access the Internet instead of the SNAT feature.

- ECS granularity

  If you create an SNAT entry for an ECS instance, the ECS instance can access the Internet by using the specified public IP address. The ECS instance accesses the Internet by using the public IP address specified in the SNAT entry and through the NAT gateway on which the SNAT entry is created.

19.1.7.2. Create an SNAT entry

This topic describes how to create an SNAT entry. SNAT can provide proxy services for Elastic Compute Service (ECS) instances. ECS instances that do not have public IP address assigned in virtual private clouds (VPCs) can access the Internet by using SNAT.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.

3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure SNAT** in the **Actions** column.
4. On the **SNAT Table** page, click **Create SNAT Entry**.
5. In the **Create SNAT Entry** panel, set the following parameters and click **OK**.

Parameter	Description
VSwitch Granularity	
VSwitch	Select a vSwitch in the VPC. All ECS instances in the vSwitch can access the Internet by using SNAT. **Note** SNAT entries do not take effect on ECS instances that are assigned public IP addresses. For example, an ECS instance may be assigned a static public IP address, associated with an elastic IP address (EIP), or configured with DNAT IP mapping. In this case, the ECS instance uses the public IP address instead of the SNAT entry to access the Internet.
VSwitch CIDR Block	The CIDR block of the selected vSwitch.
Public IP Address	Select an EIP that is used to access the Internet. You can select one or more EIPs to create an SNAT IP address pool. **Note** An EIP that is already used in a DNAT entry cannot be used in an SNAT entry.
ECS Granularity	
Available ECS Instances	Select an ECS instance in the VPC. The ECS instance can access the Internet by using the specified EIP. Make sure that the following requirements are met: ○ The ECS instance is in the Running state. ○ The ECS instance is not assigned an EIP or a static public IP address.
ECS CIDR Block	The CIDR block of the ECS instance.
Public IP Address	Select an EIP that is used to access the Internet. **Note** An EIP that is already used in a DNAT entry cannot be used in an SNAT entry.

19.1.7.3. Modify an SNAT entry

This topic describes how to modify the public IP address in an SNAT entry.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure SNAT** in the **Actions** column.

4. On the **SNAT Table** page, find the SNAT entry that you want to modify and click **Edit** in the **Actions** column.
5. In the **Edit SNAT Entry** panel, change the public IP address and click **OK**.

19.1.7.4. Delete an SNAT entry

This topic describes how to delete an SNAT entry. You can delete an SNAT entry if the Elastic Compute Service (ECS) instances that do not have public IP addresses in a virtual private cloud (VPC) no longer need the SNAT service to access the Internet.

Procedure

1. Log on to the NAT Gateway console.
2. In the top navigation bar, select the region where the NAT gateway is deployed.
3. On the **NAT Gateway** page, find the NAT gateway that you want to manage and click **Configure SNAT** in the **Actions** column.
4. On the **SNAT Table** page, find the SNAT entry that you want to delete and click **Remove** in the **Actions** column.
5. In the message that appears, click **OK**.

19.1.8. NAT service plan

19.1.8.1. Create a NAT service plan

You can associate an elastic IP address (EIP) or a NAT service plan to a NAT gateway. However, you can choose only one of them for the NAT gateway. If you want to associate a NAT service plan with the NAT gateway, you must create a NAT service plan first. Then, you can configure SNAT or DNAT for the NAT gateway. A NAT service plan consists of public IP addresses and Internet bandwidth.

Procedure

1. Log on to the NAT Gateway console.
2. On the **NAT Gateways** page, find the target NAT gateway and choose **Purchase NAT Bandwidth Package** in the **Internet Shared Bandwidth** column.
3. On the **Bandwidth Package Details** page, click **Purchase**.
4. On the **NAT Bandwidth Package** page, set the following parameters, and click **Submit**.

Parameter	Description
Region	Indicates the region for which the NAT service plan is purchased.
Billing methods	Select the billing method of the NAT service plan. Only **By Bandwidth** is supported.
Bandwidth (Mbit/s)	Enter a bandwidth value for the NAT service plan that you want to purchase. The maximum value is 5000 Mbit/s.
Name	Enter a name for the NAT service plan. The name must be 2 to 128 characters in length and can contain letters, digits, underscores (_), and hyphens (-). It must start with a letter or Chinese character.

Parameter	Description
Description	Enter a description for the NAT service plan. The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.
Quantity	Enter the number of NAT bandwidth plans that you want to purchase.

19.1.8.2. Modify the bandwidth of a NAT service plan

This topic describes how to modify the bandwidth of a NAT bandwidth plan. The modification takes effect immediately.

Procedure

1. Log on to the NAT Gateway console.
2. On the **NAT Gateways** page, find the target NAT gateway and click the ID of the NAT service plan in the **Internet Shared Bandwidth** column.
3. On the **Bandwidth Package Details** page, click the target NAT service plan, and then choose **Modify Bandwidth**.
4. On the **Modify Bandwidth** page, modify the bandwidth, and then click **Submit**.

   Each NAT bandwidth plan supports a maximum of 5,000 Mbit/s in bandwidth.

19.1.8.3. Add an IP address

This topic describes how to add IP addresses to a NAT service plan. The added IP addresses can be used to create Source Network Address Translation (SNAT) and Destination Network Address Translation (DNAT) rules.

Procedure

1. Log on to the NAT Gateway console.
2. On the **NAT Gateways** page, find the target NAT gateway and click the ID of the NAT service plan in the **Internet Shared Bandwidth** column.
3. On the **Bandwidth Package Details** page, click the target NAT service plan, and then choose **Add IP Address**.
4. On the **Modify IP Addresses** page, enter the number of IP addresses to be added, and then click **Submit**.

19.1.8.4. Release an IP address

This topic describes how to release IP addresses in a NAT service plan. The NAT service plan must contain at least one IP address.

Prerequisites

Before you release an IP address in the NAT service plan, make sure that the IP address is not used in Source Network Address Translation (SNAT) and Destination Network Address Translation (DNAT) entries. If the IP address is used in an SNAT or DNAT entry, delete the SNAT or DNAT entry first. For more information, see Delete a DNAT entry and Delete an SNAT entry.

Procedure

1. Log on to the NAT Gateway console.

2. On the **NAT Gateways** page, find the target NAT gateway and click the ID of the NAT service plan in the **Internet Shared Bandwidth** column.
3. On the **Bandwidth Package Details** page, click the target NAT service plan.
4. In the **Public IP List** section, find the target IP address, and click **Release** in the **Actions** column.
5. In the **Release IP** dialog box, click **OK**.

19.1.8.5. Delete a NAT service plan

This topic describes how to delete a service plan.

Prerequisites

Before you start, make sure that the following requirements are met:

- Delete the IP addresses that are used in Destination Network Address Translation (DNAT) entries. For more information, see Delete a DNAT entry.
- Delete the IP addresses that are used for Source Network Address Translation (SNAT) entries. For more information, see Delete an SNAT entry.

Procedure

1. Log on to the NAT Gateway console.
2. On the **NAT Gateways** page, find the target NAT gateway and click the ID of the NAT service plan in the **Internet Shared Bandwidth** column.
3. On the **Bandwidth Package Details** page, find the target NAT service plan and click **Delete**.
4. In the **Delete Shared Internet Shared Bandwidth** dialog box, click **OK**.

19.1.9. Anti-DDoS Origin Basic

A distributed denial-of-service (DDoS) attack is a malicious network attack against one or more systems, which can crash the targeted network. Alibaba Cloud provides up to 5 Gbit/s of basic anti-DDoS protection for a NAT gateway free of charge. Anti-DDoS Origin Basic can effectively prevent DDoS attacks.

How Anti-DDoS Origin Basic works

After you enable Anti-DDoS Origin Basic, traffic from the Internet must pass through Alibaba Cloud Security before the traffic arrives at the NAT gateway. Anti-DDoS Origin Basic scrubs and filters common DDoS attacks at Alibaba Cloud Security. Anti-DDoS Origin Basic protects your services against attacks such as SYN floods, UDP floods, ACK floods, ICMP floods, and DNS Query floods.

Anti-DDoS Origin Basic specifies the traffic scrubbing and blackhole triggering thresholds based on the bandwidth limit of the elastic IP address (EIP) that is associated with the NAT gateway. When the inbound traffic reaches the threshold, traffic scrubbing or blackhole is triggered:

- Traffic scrubbing: When the attack traffic from the Internet exceeds the scrubbing threshold or matches the attack traffic pattern, Alibaba Cloud Security starts to scrub the attack traffic. Traffic scrubbing includes packet filtering, bandwidth capping, and traffic throttling.
- Blackhole: When the attack traffic from the Internet exceeds the blackhole triggering threshold, blackhole is triggered and all inbound traffic is dropped.

Traffic scrubbing and blackhole triggering thresholds

The following table describes the methods that are used to calculate the traffic scrubbing and blackhole triggering thresholds on NAT gateways.

Bandwidth limit of the EIP	Traffic scrubbing threshold (bit/s)	Traffic scrubbing threshold (pps)	Default blackhole triggering threshold
Lower than or equal to 800 Mbit/s	800 Mbit/s	120,000	1.5 Gbit/s
Higher than 800 Mbit/s	Predefined bandwidth	Predefined bandwidth × 150	Predefined bandwidth × 2

If the bandwidth limit of the EIP is 1,000 Mbit/s, the traffic scrubbing threshold (bit/s) is 1,000 Mbit/s, the traffic scrubbing threshold (pps) is 150,000, and the default blackhole triggering threshold is 2 Gbit/s.

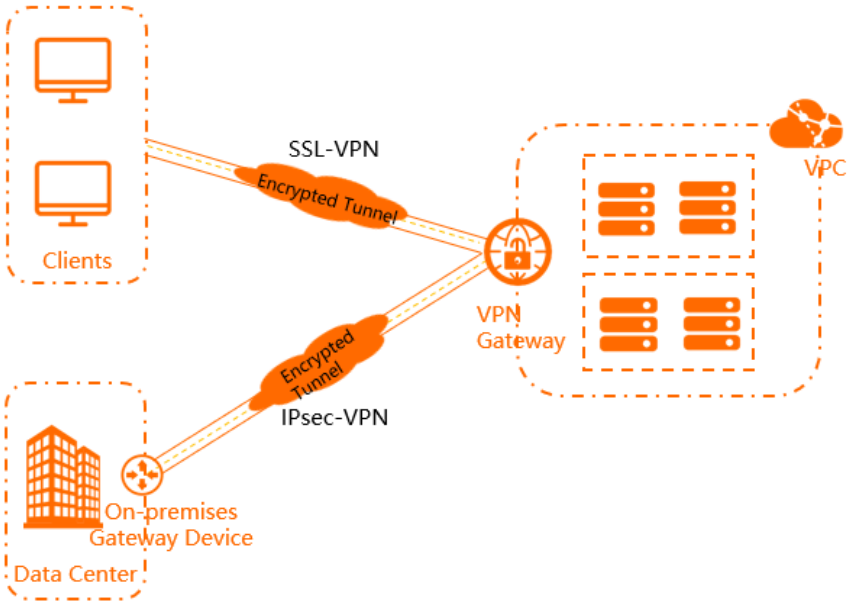
20.VPN Gateway

20.1. User Guide

20.1.1. What is VPN Gateway?

VPN Gateway is an Internet-based service that securely and reliably connects enterprise data centers, office networks, and Internet terminals to virtual private clouds (VPCs) of Alibaba Cloud through encrypted channels.

> **Note** To comply with the relevant national regulations and policies, Alibaba Cloud VPN Gateway does not provide Internet access services.



Features

VPN Gateway supports both IPsec-VPN connections and SSL-VPN connections.

- IPsec-VPN

  IPsec-VPN connects networks based on routes. It facilitates the configuration and maintenance of VPN policies, and provides flexible traffic routing methods.

  You can use IPsec-VPN to connect a data center to a VPC or connect two VPCs. IPsec-VPN supports the IKEv1 and IKEv2 protocols. All on-premises gateway devices that support these two protocols can connect to VPN gateways on Alibaba Cloud.

- SSL-VPN

  SSL-VPN connects networks based on the OpenVPN architecture. After you deploy the required resources, you can load the SSL client certificate to your client and initiate an SSL-VPN connection between the client and a VPC. This way, your client can access applications and services in the VPC.

Benefits

- High security: You can use the IKE and IPsec protocols to encrypt data for secure and reliable data transmission.
- High availability: VPN Gateway adopts the hot-standby architecture to achieve failover within a few seconds,

session persistence, and zero service downtime.

- Cost-effectiveness: The encrypted Internet-based connections provided by VPN Gateway are more cost-effective than Express Connect circuits.
- Ease of use: VPN Gateway is a ready-to-use service. VPN gateways start to work immediately after they are deployed.

20.1.2. Log on to the VPN Gateway console

This topic describes how to log on to the Apsara Uni-manager Management Console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Networking > Virtual Private Cloud**.

20.1.3. Get started with IPsec-VPN

20.1.3.1. IPsec-VPN overview

You can connect a data center to a virtual private cloud (VPC) by establishing an IPsec-VPN connection. This topic describes how to configure IPsec-VPN.

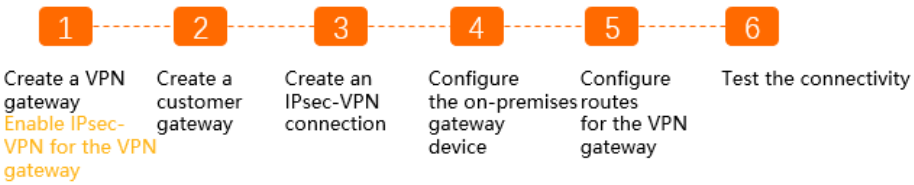
Prerequisites

Before you use IPsec-VPN to connect a data center to a VPC, make sure that the following requirements are met:

- The gateway device in the data center supports the IKEv1 and IKEv2 protocols.

  IPsec-VPN supports the IKEv1 and IKEv2 protocols. All gateway devices that support the two protocols can connect to VPN gateways on Alibaba Cloud.
- A static public IP address is assigned to the gateway device in the data center.
- The CIDR block of the data center does not overlap with the CIDR block of the VPC.
- You must make sure that the security group rules applied to the Elastic Compute Service (ECS) instances in the VPC allow gateway devices in the data center to access cloud resources.

Procedure



1. Create a VPN gateway

   You must enable the IPsec-VPN feature after you create the VPN gateway. You can establish more than one IPsec-VPN connection to each VPN gateway.
2. Create a customer gateway

   You must load the configuration of the gateway device in the data center to a customer gateway on Alibaba Cloud.
3. Create an IPsec-VPN connection

   An IPsec-VPN connection is a VPN tunnel between the VPN gateway and the gateway device in the data center. The data center can exchange encrypted data with Alibaba Cloud only after an IPsec-VPN connection is established.
4. Configure the gateway device in the data center

   You must load the configuration of the VPN gateway on Alibaba Cloud to the gateway device in the data center.
5. Add routes to the VPN gateway

   You must add routes to the VPN gateway and advertise these routes to the VPC route table. Then, the VPC and the data center can communicate with each other. For more information, see VPN Gateway route overview.
6. Verify the connectivity

   Log on to an ECS instance that is not assigned a public IP address in the VPC. Then, run the **ping** command to **ping** the private IP address of a server that resides in the data center.

Common scenarios

Connect a data center to a VPC

20.1.3.2. Connect a data center to a VPC

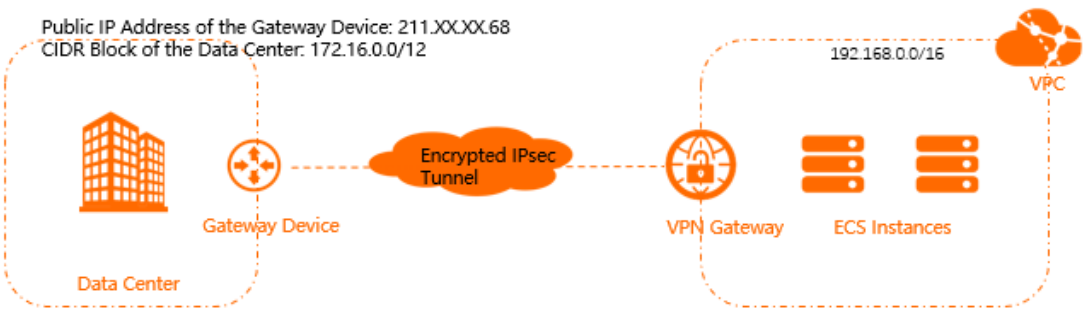
This topic describes how to use IPsec-VPN to connect a data center to a virtual private cloud (VPC). After you establish an IPsec-VPN connection, the data center and the VPC can communicate with each other.

Prerequisites

- The gateway device in the data center supports the IKEv1 and IKEv2 protocols. All gateway devices that support these protocols can connect to the VPN gateway.
- A static public IP address is assigned to the gateway device in the data center.
- The CIDR block of the data center does not overlap with the CIDR block of the VPC.
-

Context

The following scenario is used as an example in this topic. An enterprise has created a VPC on Alibaba Cloud. The CIDR block of the VPC is 192.168.0.0/16. The CIDR block of the data center is 172.16.0.0/16. The static public IP address of the gateway device in the data center is 211.XX.XX.68. To meet business requirements, the enterprise needs to connect the data center to the VPC. You can establish an IPsec-VPN connection between the data center and the VPC, as shown in the following figure. This way, the data center and VPC can share resources with each other.



Step 1: Create a VPN gateway

1.
2.
3.
4. On the **Create VPN Gateway** page, set the following parameters and click **Submit**:
   - **Organization**: Select the organization to which the VPN gateway belongs.
   - **Resource Set**: Select the resource set to which the VPN gateway belongs.
   - **Region**: Select the region where you want to deploy the VPN gateway.

     > **Note** Make sure that the VPC and the VPN gateway are deployed in the same region.

   - **Name**: Enter a name for the VPN gateway.

     The name must be 2 to 128 characters in length, and can contain letters, digits, periods (.), colons (:), underscores (_), and hyphens (-). It must start with a letter, and cannot start with `http://` or `https://`.
   - **VPC**: Select the VPC to be associated with the VPN gateway.
   - **vSwitch**: Select the vSwitch to which you want to attach the VPN gateway.
   - **Bandwidth**: Specify the bandwidth limit of the VPN gateway. Unit: Mbit/s. The bandwidth is used for data transfer over the Internet.

- **IPsec-VPN**: Specify whether to enable IPsec-VPN for the VPN gateway. In this example, select **Enable**.

  After IPsec-VPN is enabled, you can create IPsec-VPN connections between a data center and a VPC, or between two VPCs.

- **SSL-VPN**: Specify whether to enable SSL-VPN. In this example, **Disable** is selected.

  SSL-VPN connections are point-to-site connections. SSL-VPN allows you to connect a client to Alibaba Cloud without the need to configure a customer gateway.

5. Return to the VPN Gateways page to view the created VPN gateway.

   The created VPN gateway is in the **Preparing** state. The VPN gateway changes to the **Normal** state after about 1 to 5 minutes. After the VPN gateway changes to the **Normal** state, the VPN gateway is ready for use.

Step 2: Create a customer gateway

1.
2. In the top navigation bar, select the region where you want to create the customer gateway.

   > **Note** Make sure that the customer gateway and the VPN gateway to be connected are deployed in the same region.

3. On the **User Gateway** page, click **Create Customer Gateway**.
4. On the **Create Customer Gateway** page, set the following parameters and click **Submit**:
   - **Organization**: Select the organization to which the customer gateway belongs.
   - **Resource Set**: Select the resource set to which the customer gateway belongs.
   - **Region**: Select the region where you want to deploy the customer gateway.

     > **Note** Make sure that the customer gateway and the VPN gateway to be connected are deployed in the same region.

   - **Zone**: Select the zone where you want to deploy the customer gateway.
   - **Name**: Enter a name for the customer gateway.

     The name must be 2 to 128 characters in length and can contain digits, hyphens (-), and underscores (_). The name must start with a letter and cannot start with `http://` or `https://`.

   - **IP Address**: Enter the public IP address of the gateway device in the data center that you want to connect to the VPC. In this example, 211.XX.XX.68 is used.
   - **Description**: Enter a description for the customer gateway.

     The description must be 2 to 256 characters in length and can contain digits, hyphens (-), underscores (_), periods (.), commas (,), and colons (:). The description must start with a letter, and cannot start with `http://` or `https://`.

Step 3: Create an IPsec-VPN connection

1.
2. In the top navigation bar, select the region where you want to create the IPsec-VPN connection.

   > **Note** Make sure that the IPsec-VPN connection and the VPN gateway to be connected are deployed in the same region.

3. On the **IPsec Connections** page, click **Create IPsec Connection**.
4. On the **Create IPsec Connection** page, configure the IPsec-VPN connection based on the following information and click **Submit**:

- **Organization**: Select the organization to which the IPsec-VPN connection belongs.
- **Resource Set**: Select the resource set to which the IPsec-VPN connection belongs.
- **Region**: Select the region to which the IPsec-VPN connection belongs.
- **Zone**: Select the zone to which the IPsec-VPN connection belongs.
- **Name**: Enter a name for the IPsec-VPN connection.
- **VPN Gateway**: Select the created VPN gateway.
- **Customer Gateway**: Select the customer gateway to be connected through the IPsec-VPN connection.
- **Source CIDR Block**: Enter the CIDR block of the VPC where the VPN gateway is deployed. 192.168.0.0/16 is used in this example.
- **Destination CIDR Block**: Enter the CIDR block of the data center. In this example, 172.16.0.0/16 is used.
- **Effective Immediately**: Select whether to immediately start connection negotiations.
  - **Yes**: starts negotiations immediately after you complete the configuration.
  - **No**: starts negotiations when data transfer is detected.
- **Advanced Settings**: Select **Default**.

  By default, a pre-shared key is automatically generated.

Step 4: Load the configuration of the IPsec-VPN connection to the gateway device in the data center

1.
2. On the **IPsec Connections** page, find the IPsec-VPN connection that you want to manage, and choose **More > Download Configuration** in the **Actions** column.
3. Load the configuration of the IPsec-VPN connection to the gateway device in the data center. For detailed information about the configurations, consult the manufacturer of your gateway device.

Step 5: Configure routes for the VPN gateway

1.
2. On the **VPN Gateway** page, find the VPN gateway that you want to manage and click its ID.
3. On the **Destination-based Routing** tab, click **Add Route Entry**.
4. In the **Add Route Entry** panel, set the following parameters and click **OK**:
   - **Destination CIDR Block**: Enter the CIDR block of the data center. In this example, 172.16.0.0/16 is used.
   - **Next Hop Type**: Select **IPsec Connection**.
   - **Next Hop**: Select the IPsec-VPN connection that you created.
   - **Publish to VPC**: Specify whether to automatically advertise new routes to the VPC route table. In this example, **Yes** is selected.
   - **Weight**: Select a weight for the route. In this example, **100** is selected.
     - **100**: specifies a high priority for the route.
     - **0**: specifies a low priority for the route.

     > **Note** If two destination-based routes are configured with the same destination CIDR block, you cannot set the weights of both routes to 100.

Step 6: Verify the connectivity

1. Log on to an Elastic Compute Service (ECS) instance that is not assigned a public address in the VPC.
2. Run the **ping** command to access a server in the data center and verify the connectivity.

```
[root@iZm5[redacted]slbZ ~]# ping 172.16.1.188
PING 172.16.1.188 (172.16.1.188) 56(84) bytes of data.
64 bytes from 172.16.1.188: icmp_seq=1 ttl=62 time=23.8 ms
64 bytes from 172.16.1.188: icmp_seq=2 ttl=62 time=23.7 ms
64 bytes from 172.16.1.188: icmp_seq=3 ttl=62 time=23.7 ms
64 bytes from 172.16.1.188: icmp_seq=4 ttl=62 time=23.7 ms
^Z
[1]+  Stopped                 ping 172.16.1.188
[root@iZm5ea8[redacted]xslbZ ~]#
```

20.1.4. Get started with SSL-VPN

20.1.4.1. SSL-VPN overview

SSL-VPN allows clients to connect to a virtual private cloud (VPC) and access applications and services that are deployed in the VPC in a secure manner. This topic describes how to use SSL-VPN.

Prerequisites

Before you use SSL-VPN to establish a connection between a client and a VPC, make sure that the following requirements are met:

- The private CIDR block of the client does not overlap with the private CIDR block of the VPC. Otherwise, the client and the VPC cannot communicate with each other.
- The client can access the Internet.
- You have read and understand the security group rules that apply to the Elastic Compute Service (ECS) instances in the VPC, and make sure that the security rules allow the client to access the ECS instances.

Procedure

1. Create a VPN gateway.

   Create a VPN gateway and enable the SSL-VPN feature.

2. Create an SSL server.

   On the SSL server, specify the private CIDR block that the client needs to access and the CIDR block that is used by the client.

3. Create an SSL client certificate.

   Create and download a client certificate based on the SSL server configuration.

4. Configure the client.

   Download and install VPN software on the client, load the SSL client certificate, and then initiate an SSL-VPN connection.

5. Verify the connectivity.

   Open the CLI on the client, and run the **ping** command to ping an ECS instance in the VPC.

Basic scenarios

Connect a client to a VPC

20.1.4.2. Connect a client to a VPC

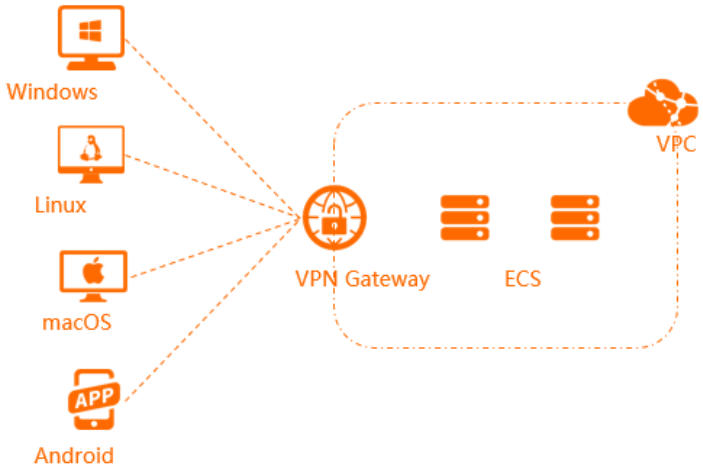
This topic describes how to connect a client to a virtual private cloud (VPC) by using SSL-VPN.

Prerequisites

- The private CIDR block of the client does not overlap with the private CIDR block of the VPC. Otherwise, the client and the VPC cannot communicate with each other.
- The client can access the Internet.
-

Context

The scenario in the following figure is used as an example to describe how to connect Linux, Windows, and Mac clients to a VPC by using SSL-VPN.



Step 1: Create a VPN gateway

1.
2.
3.
4. On the **Create VPN Gateway** page, set the following parameters and click **Submit**:
   - **Organization**: Select the organization to which the VPN gateway belongs.
   - **Resource Set**: Select the resource set to which the VPN gateway belongs.
   - **Region**: Select the region where you want to deploy the VPN gateway.

     > **Note** Make sure that the VPC and the VPN gateway are deployed in the same region.

   - **Name**: Enter a name for the VPN gateway.

     The name must be 2 to 128 characters in length, and can contain letters, digits, periods (.), colons (:), underscores (_), and hyphens (-). It must start with a letter, and cannot start with `http://` or `https://`.
   - **VPC**: Select the VPC to be associated with the VPN gateway.
   - **vSwitch**: Select the vSwitch to which you want to attach the VPN gateway.
   - **Bandwidth**: Specify the bandwidth limit of the VPN gateway. The bandwidth is used for data transfer over the Internet.

- **IPsec-VPN**: Specify whether to enable IPsec-VPN for the VPN gateway. In this example, **Disable** is selected.
- **SSL-VPN**: Specify whether to enable SSL-VPN. In this example, **Enable** is selected.

  SSL-VPN connections are point-to-site connections. SSL-VPN allows you to connect a client to Alibaba Cloud without the need to configure a customer gateway.
- **SSL Connections**: Specify the maximum number of concurrent SSL connections that the VPN gateway supports.

5. Return to the **VPN Gateways** page to view the created VPN gateway.

   The created VPN gateway is in the **Preparing** state. The VPN gateway changes to the **Normal** state after about 1 to 5 minutes. After the VPN gateway changes to the **Normal** state, the VPN gateway is ready for use.

Step 2: Create an SSL server

1.
2. In the top navigation bar, select the region where you want to create the SSL server.

   > **Note** Make sure that the SSL server and the created VPN gateway are deployed in the same region.

3.
4. On the **Create SSL Server** page, set the following parameters and click **Submit**:
   - **Organization**: Select the organization to which the SSL server belongs.
   - **Resource Set**: Select the resource set to which the SSL server belongs.
   - **Region**: Select the region where you want to deploy the SSL server.
   - **Zone**: Select the zone where you want to deploy the SSL server.
   - **Name**: Enter a name for the SSL server.
   - **VPN Gateway**: Select the created VPN gateway.
   - **Source CIDR Block**: Enter the CIDR block of the network to which you want to connect. Click ⊕ to add more CIDR blocks. You can add the CIDR block of a VPC, a vSwitch, or an on-premises network.
   - **Client CIDR Block**: Enter the CIDR block that the client uses to connect to the SSL server. Example: 192.168.10.0/24.
   - **Advanced Settings**: Select **Default**.

Step 3: Create and download an SSL client certificate

1.
2.
3. On the **Create SSL Client Certificate** page, set the following parameters and click **Submit**:
   - **Organization**: Select the organization to which the SSL client certificate belongs.
   - **Resource Set**: Select the resource set to which the SSL client certificate belongs.
   - **Region**: Select the region where you want to create the SSL client certificate.
   - **Zone**: Select the zone where you want to create the SSL client certificate.
   - **Name**: Enter a name for the SSL client certificate.
   - **VPN Gateway**: Select the VPN gateway with which you want to associate the SSL client certificate.
   - **SSL Server**: Select the SSL server with which you want to associate the SSL client certificate.
4. On the **SSL Client** page, find the created SSL client certificate and click **Download** in the **Actions** column.

   The SSL client certificate is downloaded to your client.

Step 4: Configure the client

The following section describes how to configure Linux, Mac, and Windows clients.

- Linux client
  i. Run the following command to install OpenVPN:

  ```
  yum install -y openvpn
  ```

  ii. Decompress the SSL client certificate package that you downloaded and copy the SSL client certificate to the */etc/openvpn/conf/* directory.
  iii. Go to the */etc/openvpn/conf/* directory and run the following command to start OpenVPN:

  ```
  openvpn --config /etc/openvpn/conf/config.ovpn --daemon
  ```

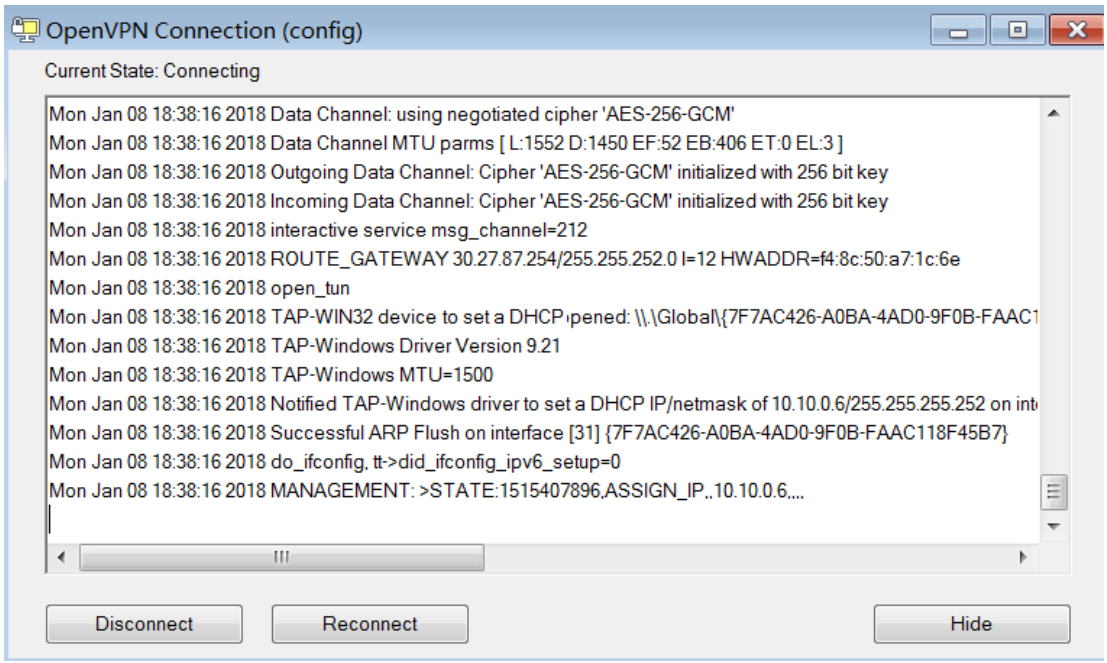
- Windows client
  i. Download and install OpenVPN.

  Download OpenVPN.

  ii. Decompress the downloaded SSL client certificate package and copy the SSL client certificate to the *OpenVPN\config* directory.

  In this example, the certificate is copied to the *C:\Program Files\OpenVPN\config* directory. You must copy the certificate to the directory where OpenVPN is installed.

  iii. Start OpenVPN and click **Connect** to initiate a connection.



- Mac client
  i. Run the following command to install OpenVPN:

  ```
  brew install openvpn
  ```

  > **Note** Make sure that homebrew is installed before you install OpenVPN.

  ii. Copy the SSL client certificate package that you downloaded in Step 3 to the configuration directory of OpenVPN and decompress the package. Then, initiate an SSL-VPN connection.
    a. Back up all configuration files in the */usr/local/etc/openvpn* folder.

b. Run the following command to delete the configuration files of OpenVPN:

```
rm /usr/local/etc/openvpn/*
```

c. Run the following command to copy the downloaded SSL client certificate package to the configuration directory of OpenVPN:

```
cp cert_location /usr/local/etc/openvpn/
```

`cert_location` specifies the path where the SSL client certificate that you downloaded in Step 3 is installed. Example: */Users/example/Downloads/certs6.zip*.

d. Run the following command to decompress the SSL client certificate package:

```
cd  /usr/local/etc/openvpn/
unzip /usr/local/etc/openvpn/certs6.zip
```

e. Run the following command to initiate a connection:

```
sudo /usr/local/opt/openvpn/sbin/openvpn --config /usr/local/etc/openvpn/config.ovpn
```

Step 5: Verify the connectivity

1. Open the CLI on the client.
2. To verify the connectivity, you can run the **ping** command to access an Elastic Compute Service (ECS) instance in the VPC.

20.1.5. Manage a VPN Gateway

20.1.5.1. Create a VPN gateway

This topic describes how to create a VPN gateway. You must create a VPN gateway before you can use the IPsec-VPN and SSL-VPN features. After you create a VPN gateway, a public IP address is assigned to the VPN gateway.

Procedure

1.
2.
3.
4. On the **Create VPN Gateway** page, set the following parameters for the VPN gateway, and then click **Submit**.

Parameter	Description
Organization	Select the organization to which the VPN gateway belongs.
Resource Set	Select the resource set to which the VPN gateway belongs.
Region	Select the region where you want to create the VPN gateway. You can create IPsec-VPN connections on VPN gateways to connect a data center to a VPC or connect two VPCs. Make sure that the VPC and the VPN gateway associated with the VPC are deployed in the same region.
Instance Name	Enter a name for the VPN gateway. The name must be 2 to 128 characters in length, and can contain letters, digits, periods (.), underscores (_), and hyphens (-). It must start with a letter, and cannot start with `http://` or `https://`.

Parameter	Description
VPC	Select the VPC to be associated with the VPN gateway.
vSwitch	Select the vSwitch to which you want to attach the VPN gateway.
Bandwidth	Select a maximum bandwidth value for the VPN gateway. The bandwidth is used for data transfer over the Internet.
IPsec-VPN	Specify whether to enable IPsec-VPN for the VPN gateway. The default value is **Enable IPsec**. After IPsec-VPN is enabled, you can create a secure IPsec tunnel to connect a data center to a VPC, or connect two VPCs.
SSL-VPN	Specify whether to enable SSL-VPN for the VPN gateway. The default value is **Disable**. SSL-VPN connections are point-to-site connections. SSL-VPN allows you to connect a client to Alibaba Cloud without the need to configure a customer gateway.
SSL Connections	Select the maximum number of concurrent SSL connections that the VPN gateway supports. **Note** You can set this parameter only after SSL-VPN is enabled.

20.1.5.2. Modify a VPN gateway

This topic describes how to modify the name and description of a VPN gateway.

Prerequisites

A VPN gateway is created. For more information, see Create and manage a VPN gateway.

Procedure

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway, and click in the **Instance ID/Name** column. In the dialog box that appears, enter a new name and click **OK**.

   The name must be 2 to 100 characters in length, and can contain digits, underscores (_), and hyphens (-). It must start with a letter.
5. Click

   in the **Description** column. In the dialog box that appears, enter a new description and click **OK**.

   The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

20.1.5.3. Configure routes of a VPN Gateway

20.1.5.3.1. Route overview

After you create an IPsec-VPN connection by using a VPN gateway, you must add a route to the VPN gateway.

Route-based IPsec-VPN allows you to route network traffic in multiple ways, and facilitates the configuration and maintenance of VPN policies.

You can add the following two types of route to a VPN gateway:

- Policy-based routes.
- Destination-based routes.

Policy-based routes

Policy-based routes forward traffic based on source and destination IP addresses.

For more information, see Add a policy-based route entry.

> **Note** Policy-based routes take precedence over destination-based routes.

Destination-based routes

Destination-based routes forward traffic to specified destination IP addresses.

For more information, see Create a destination-based route.

20.1.5.3.2. Work with a policy-based route

A policy-based route forwards traffic based on source and destination IP addresses. This topic describes how to create, advertise, modify, and delete a policy-based route.

Prerequisites

An IPsec-VPN connection is created. For more information, see Create an IPsec-VPN connection.

Add a policy-based route

After you create an IPsec-VPN connection, you can create a policy-based route for the IPsec-VPN connection.

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. Click the **Policy-based Routing** tab, and then click **Add Route Entry**.
6. In the **Add Route Entry** panel, set the following parameters and click **OK**.

Parameter	Description
Destination CIDR block	Enter the private CIDR block that you want to access.
Source CIDR Block	Enter the private CIDR block of the VPC.
Next Hop Type	Select IPsec Connection.
Next Hop	Select the IPsec-VPN connection for which you want to create the policy-based route.

Parameter	Description
Publish to VPC	Specify whether to advertise the route to the VPC route table. Valid values: ○ **Yes**: automatically advertises the route to the route table of the VPC. We recommend that you select this value. ○ **No**: does not advertise the route to the VPC route table. **Note** If you select **No**, you must manually advertise the route to the VPC route table.
Weight	Select a weight. Valid values: ○ **100**: specifies a high priority for the policy-based route. ○ **0**: specifies a low priority for the policy-based route. **Note** If two policy-based routes are configured with the same destination CIDR block, you cannot set the weights of the routes to 100.

Advertise a policy-based route

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Policy-based Routing** tab, find the policy-based route that you want to advertise and click **Publish** in the **Actions** column.
6. In the **Publish Route Entry** message, click **OK**.

   If you want to withdraw the policy-based route, click **Unpublish**.

Modify a policy-based route

You can change the weight of a policy-based route.

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Policy-based Routing** tab, find the policy-based route that you want to modify and click **Edit** in the **Actions** column.
6. In the panel that appears, specify a new weight for the route and click **OK**.

Delete a policy-based route

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Policy-based Routing** tab, find the policy-based route that you want to delete and click **Delete** in the **Actions** column.
6. In the **Delete Route Entry** message, click **OK**.

20.1.5.3.3. Manage destination-based routes

Destination-based routing is a technique that routes network traffic to specified destination IP addresses. This topic describes how to create, advertise, modify, and delete a destination-based route.

Prerequisites

An IPsec-VPN connection is created. For more information, see Create an IPsec-VPN connection.

Create a destination-based route

After you create an IPsec-VPN connection, you can create a destination-based route for the IPsec-VPN connection.

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Destination-based routing** tab, click **Add Route Entry**.
6. In the **Add Route Entry** dialog box, set the following parameters and click **OK**.

Parameter	Description
Destination CIDR block	The private CIDR block that you want to access.
Next Hop Type	Select IPsec Connection.
Next Hop	Select the IPsec-VPN connection for which you want to create a destination-based route.
Publish to VPC	Specify whether to advertise the destination-based route to the virtual private cloud (VPC) route table. ○ **Yes**: automatically advertises the route to the route table of the VPC. We recommend that you select this value. ○ **No**: does not advertise the destination-based route to the VPC route table. **Note** If you select **No**, you must manually advertise the destination-based route to the VPC route table.
Weight	Select a weight: ○ **100**: specifies a high priority for the destination-based route. ○ **0**: specifies a low priority for the destination-based route. **Note** If two destination-based routes are configured with the same destination CIDR block, you cannot set the weights of the routes to 100.

Advertise the destination-based route

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Destination-based Routing** tab, find the destination-based route that you want to manage and

click **Publish** in the **Actions** column.

6. In the **Publish Route Entry** message, click **OK**.

   If you want to withdraw the destination-based route, click **Unpublish**.

Modify the destination-based route

You can change the weight of the destination-based route.

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Destination-based Routing** tab, find the destination-based route that you want to manage and click **Edit** in the **Actions** column.
6. In the panel that appears, specify the weight of the destination-based route and click **OK**.

Delete the destination-based route

1. Log on to the VPN Gateway console.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway and click its ID.
5. On the **Destination-based Routing** tab, find the destination-based route that you want to delete and click **Delete** in the **Actions** column.
6. In the **Delete Route Entry** message, click **OK**.

20.1.5.4. Delete a VPN gateway

This topic describes how to delete a VPN gateway that you no longer need. After you delete a VPN gateway, you can no longer establish IPsec-VPN or SSL-VPN connections to the VPN gateway.

Prerequisites

Before you delete a VPN gateway, make sure that the following conditions are met:

- The IPsec-VPN connections established to the VPN gateway are deleted. For more information, see Delete an IPsec-VPN connection.
- The SSL server associated with the VPN gateway is deleted. For more information, see Delete an SSL server.

Procedure

1.
2.
3.
4. On the **VPN Gateways** page, find the VPN gateway that you want to delete and click **Delete** in the **Actions** column.
5. In the **Delete VPN Gateway** message, click **OK**.

20.1.6. Manage a customer gateway

20.1.6.1. Create a customer gateway

This topic describes how to create a customer gateway. You can use a customer gateway to establish an IPsec-VPN connection between a virtual private cloud (VPC) and a data center or between two VPCs. After you create a customer gateway, you can update the information about a gateway device in the data center to Alibaba Cloud. Then, you can connect the customer gateway to a VPN gateway. A customer gateway can connect to multiple VPN gateways.

Procedure

1. Log on to the VPN Gateway console.
2.
3. In the top navigation bar, select the region where you want to create the customer gateway.

   > **Note** Make sure that the customer gateway and the VPN gateway to be connected belong to the same region.

4. On the **Customer Gateways** page, click **Create Customer Gateway**.
5. On the **Create Customer Gateway** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the customer gateway belongs.
Resource Set	Select the resource set to which the customer gateway belongs.
Region	Select the region where you want to deploy the customer gateway. **Note** Make sure that the customer gateway and the VPN gateway to be connected belong to the same region.
Zone	Select the zone where you want to deploy the customer gateway.
Name	Enter a name for the customer gateway. The name must be 2 to 128 characters in length and can contain digits, hyphens (-), and underscores (_). The name must start with a letter and cannot start with `http://` or `https://`.
IP Address	Enter the static public IP address of the gateway device in the data center.
Description	Enter a description for the customer gateway. The description must be 2 to 256 characters in length and can contain digits, hyphens (-), underscores (_), periods (.), commas (,), and colons (:). The description must start with a letter, and cannot start with `http://` or `https://`.

20.1.6.2. Modify a customer gateway

This topic describes how to modify the name and description of a customer gateway.

Prerequisites

A customer gateway is created. For more information, see Create a customer gateway.

Procedure

1. Log on to the VPN Gateway console.

2.
3.
4. On the **Customer Gateways** page, find the customer gateway, click

   in the **Instance ID/Name** column. In the dialog box that appears, enter a name and click **OK**.

   The name must be 2 to 128 characters in length and can contain digits, underscores (_), and hyphens (-). It must start with a letter.
5. Click

   in the **Description** column. In the dialog box that appears, enter a new description and click **OK**.

   The description must be 2 to 256 characters in length and cannot start with `http://` or `https://`.

20.1.6.3. Delete a customer gateway

This topic describes how to delete a customer gateway.

Procedure

1. Log on to the VPN Gateway console.
2.
3.
4. On the **Customer Gateways** page, find the customer gateway that you want to delete, and then click **Delete** in the **Actions** column.
5. In the **Delete Customer Gateway** message, click **OK**.

20.1.7. Configure IPsec-VPN connections

20.1.7.1. Manage an IPsec-VPN connection

20.1.7.1.1. Create an IPsec-VPN connection

This topic describes how to create an IPsec-VPN connection. After you create a VPN gateway and a customer gateway, you can create an IPsec-VPN connection between the two gateways for encrypted data transmission.

Procedure

1.
2.
3. In the top navigation bar, select the region where you want to create the IPsec-VPN connection.
4. On the **IPsec Connections** page, click **Create IPsec Connection**.
5. On the **Create IPsec Connection** page, configure the IPsec-VPN connection based on the following information and click **Submit**.

Parameter	Description
Organization	Select the organization to which the IPsec-VPN connection belongs.
Resource Set	Select the resource set to which the IPsec-VPN connection belongs.
Region:	Select the region where the IPsec-VPN connection is created.

Parameter	Description
Zone	Select the zone where the IPsec-VPN connection is created.
Name	Enter a name for the IPsec-VPN connection. The name must be 2 to 128 characters in length, and can contain digits, hyphens (-), and underscores (_). It must start with a letter and cannot start with `http://` or `https://`.
VPN Gateway	Select the VPN gateway to be connected through the IPsec-VPN connection.
Customer Gateway	Select the customer gateway to be connected through the IPsec-VPN connection.
Local Network	Enter the CIDR block of the VPC to be connected to the data center. The CIDR block is used in Phase 2 negotiations. You can add multiple CIDR blocks of the VPC only if IKEv2 is used.
Remote Network	Enter the CIDR block of the data center to be connected to the VPC. This CIDR block is used in Phase 2 negotiations. You can add multiple CIDR blocks of the data center only if IKEv2 is used.
Effective Immediately	Specify whether to start connection negotiations immediately. ○ **Yes**: starts connection negotiations after the configuration is completed. ○ **No**: starts negotiations when traffic is detected.
Advanced Configurations	Select the type of the advanced configuration. ○ **Default**: Use default settings. ○ **Configure**: Use custom settings.
Advanced configuration: IKE configuration	
Pre-Shared Key	Enter the pre-shared key used for authentication between the VPN gateway and the customer gateway. You can specify a key, or use the default key that is randomly generated by the system.
Version	Select an IKE version. ○ **ikev1** ○ **ikev2** IKEv1 and IKEv2 are supported. Compared with IKEv1, IKEv2 simplifies the Security Association (SA) negotiation process and provides better support for scenarios where multiple CIDR blocks are used. We recommend that you select IKEv2.
Negotiation Mode	Select the negotiation mode of IKEv1. ○ **main**: This mode offers higher security during negotiations. ○ **aggressive**: This mode is faster and has a higher success rate. Connections negotiated in both modes ensure the same security level of data transmission.
Encryption Algorithm	Select the encryption algorithm to be used in Phase 1 negotiations. Supported algorithms are **aes**, **aes192**, **aes256**, **des**, and **3des**.

Parameter	Description
Authentication Algorithm	Select the authentication algorithm to be used in Phase 1 negotiations. Supported algorithms are **sha1** and **md5**.
DH Group	Select the Diffie-Hellman key exchange algorithm to be used in Phase 1 negotiations.
SA Life Cycle (seconds)	Specify the lifecycle of the SA after Phase 1 negotiations succeed. Default value: **86400**. Unit: seconds.
LocalId	Specify the ID of the VPN gateway. The ID is used in Phase 1 negotiations. The default value is the public IP address of the VPN gateway. If you set LocalId to FQDN, we recommend that you set Negotiation Mode to **Aggressive**.
RemoteId	Specify the ID of the customer gateway. The ID is used in Phase 1 negotiations. The default value is the public IP address of the customer gateway. If you set RemoteId to FQDN, we recommend that you select set Negotiation Mode to **Aggressive**.
Advanced configuration: IPSec configuration	
Encryption Algorithm	Select the encryption algorithm to be used in Phase 2 negotiations. Supported algorithms are **aes**, **aes192**, **aes256**, **des**, and **3des**.
Authentication Algorithm	Select the authentication algorithm to be used in Phase 2 negotiations. Supported algorithms are **sha1** and **md5**.
DH Group	Select the Diffie-Hellman key exchange algorithm to be used in Phase 2 negotiations. ◦ If you select a value other than **disabled**, the PFS feature is enabled by default, which necessitates key update for every renegotiation. Therefore, you must also enable PFS for the client. ◦ For clients that do not support PFS, select **disabled**.
SA Life Cycle (seconds)	Specify the lifecycle of the SA after Phase 2 negotiations succeed. Default value: **86400**. Unit: seconds.

20.1.7.1.2. Modify an IPsec-VPN connection

After you create an IPsec-VPN connection, you can modify its configurations.

Procedure

1.
2.
3.
4. On the **IPsec Connections** page, find the IPsec-VPN connection that you want to manage, and click **Edit** in the **Actions** column.
5. On the **Modify IPsec Connections** panel, modify the name, advanced configurations, CIDR block, and then click **OK**.

   For more information about the parameters, see Create an IPsec-VPN connection.

20.1.7.1.3. Download the configuration file of an IPsec-VPN connection

This topic describes how to download the configuration file of an IPsec-VPN connection and load the configuration file to an on-premise gateway device.

Procedure

1.
2.
3.
4. On the **IPsec Connections** page, find the IPsec-VPN connection that you want to manage and choose **More > Download Configuration** in the **Actions** column.

> **Note** The values of RemoteSubnet and LocalSubnet in the downloaded configuration file are opposite to the values that you specified when you create the IPsec-VPN connection. For a VPN gateway, RemoteSubnet refers to the CIDR block of the data center, whereas LocalSubnet refers to the CIDR block of the VPC. For a gateway device, LocalSubnet refers to the CIDR block of the data center, whereas RemoteSubnet refers to the CIDR block of the VPC.

20.1.7.1.4. Configure a security group

This topic describes how to configure a security group to control the inbound and outbound traffic of Elastic Compute Service (ECS) instances in the security group after an IPsec-VPN connection is created.

Procedure

1.
2.
3.
4. On the **IPsec Connections** page, find the IPsec-VPN connection that you want to manage and choose **More > Configure Routing Group** in the **Actions** column.
5. In the **Configure Routing Group** panel, set the following parameters and click **OK**.

Parameter	Description
Security Group	Select the security group to which you want to add the security group rule.
Rule Direction	Select the direction of data transfer that the rule controls. ◦ **Outbound**: controls data transfer from the ECS instances in the security group to the Internet or other ECS instances. ◦ **Inbound**: controls data transfer from the Internet or other ECS instances to the ECS instances in the security group.
Action	Specify the action to be performed on the matching requests. ◦ **Allow**: accepts requests. ◦ **Deny**: drops requests without returning messages. If two security group rules use the same settings except for different actions, the **Deny** rule prevails over the **Allow** rule.
Protocol Type	The protocol of the security group rule.

Parameter	Description
Port Range	Enter a port range for the security group rule. Valid values: -1 and 1 to 65535. You cannot enter only -1. Examples: ○ 1/200 specifies ports 1 to 200. ○ 80/80 specifies port 80. ○ -1/-1 specifies all ports.
Priority	Set the priority of the rule. Valid values: 1 to 100. The default value is 1, which indicates the highest priority.
Authorization Mode	Specify the type of addresses that the security group rule permits or blocks. You can select only **Address**, which indicates CIDR blocks.
ENI Type	Specify the type of data transfer that the security group rule controls. ○ **Internal**: controls data transfer within Alibaba Cloud. ○ **External**: controls data transfer over the Internet.
Authorization IP Addresses	Specify the CIDR blocks that you want the security group rule to accept or block. You can specify at most 10 CIDR blocks.
Enable Automatically Configure Routers	Specify whether to automatically propagate routes. The feature is enabled by default.
Description	Enter a description for the security group rule. The description must be 2 to 256 characters in length, and cannot start with `http://` or `https://`. You can leave this parameter empty.

20.1.7.1.5. View IPsec-VPN connection logs

This topic describes how to view IPsec-VPN connection logs that are generated within the last month to troubleshoot connection errors. The time range of a log is set to 10 minutes.

Procedure

1.
2.
3.
4. On the **IPsec Connections** page, find the IPsec-VPN connection that you want to manage and choose **More > View Logs** in the **Actions** column.
5. In the **IPsec Connection Logs** Panel, set the time range and view the log.

20.1.7.1.6. Delete an IPsec-VPN connection

This topic describes how to delete an IPsec-VPN connection.

Procedure

1.
2.
3.
4. On the **IPsec Connections** page, find the IPsec-VPN connection that you want to delete, and click **Delete** in the **Actions** column.
5. In the **Delete IPsec Connection** message, click **OK**.

20.1.7.2. MTU considerations

The maximum transmission unit (MTU) is the size of the largest packet that can be transmitted over a network layer protocol, such as TCP. Packets are measured in bytes. The MTU takes both the sizes of headers and data into account.

Segments transmitted over an IPsec tunnel are encrypted and then encapsulated into packets for routing purpose. The size of a segment must fit the MTU of the packet that carries the segment. Therefore, the MTU of the segment must be smaller than the MTU of the packet.

Gateway MTU

You must set the MTU of the local VPN gateway to a value no greater than 1,360 bytes. We recommend that you set the MTU to 1,360 bytes.

The TCP protocol negotiates the maximum segment length (MSS) of each packet segment between the sender and the receiver. We recommend that you set the TCP MSS of the on-premises VPN gateway to 1,359 bytes to facilitate the encapsulation and transfer of TCP packets.

20.1.8. Configure SSL-VPN

20.1.8.1. Manage an SSL server

20.1.8.1.1. Create an SSL server

This topic describes how to create an SSL server. Before you can create an SSL-VPN connection, you must create an SSL server.

Prerequisites

A VPN gateway is created and SSL-VPN is enabled for the VPN gateway. For more information, see Create and manage a VPN gateway.

Procedure

1. Log on to the VPN Gateway console.
2.
3. In the top navigation bar, select the region where you want to create the SSL server.
4. On the **SSL Servers** page, click **Create SSL Server**.
5. On the **Create SSL Server** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the SSL server belongs.
Resource Set	Select the resource set to which the SSL server belongs.
Region	Select the region where you want to deploy the SSL server.

Parameter	Description
Zone	Select the zone where you want to deploy the SSL server.
Name	Enter a name for the SSL server. The name must be 2 to 128 characters in length, and can contain digits, hyphens (-), and underscores (_). It must start with a letter and cannot start with `http://` or `https://`.
VPN Gateway	Select the VPN gateway that you want to associate with the SSL server. Make sure that SSL-VPN is enabled for the VPN gateway.
Local Network	Enter the CIDR block that the client needs to access through the SSL-VPN connection. It can be the CIDR block of a VPC, a vSwitch, a data center connected to a VPC through an Express Connect circuit, or a cloud service such as ApsaraDB RDS or OSS. Click **+** to add more CIDR blocks. **Note** The subnet mask of the specified CIDR block must be 16 to 29 bits in length.
Client CIDR Block	Enter the CIDR block from which an IP address is allocated to the virtual NIC of the client. Do not enter the private CIDR block of the client. When a client accesses the server through an SSL-VPN connection, the VPN gateway assigns an IP address from the specified client CIDR block to the client. **Notice** ◦ Make sure that the CIDR block of the destination network and the client CIDR block do not overlap with each other. ◦ Make sure that the number of IP addresses that the client CIDR block provides is at least four times the number of SSL-VPN connections. For example, if you specify 192.168.0.0/24 as the client CIDR block, the system first divides a subnet CIDR block with a subnet mask of 30 from 192.168.0.0/24. 192.168.0.4/30, which provides up to four IP addresses, is used as the subnet CIDR block in this example. Then, the system allocates an IP address from 192.168.0.4/30 to the client and uses the other three IP addresses to ensure network communication. In this case, one client consumes four IP addresses. Therefore, to ensure that an IP address can be allocated to your client, you must make sure that the number of IP addresses that the client CIDR block provides is at least four times the number of SSL-VPN connections.

Parameter	Description
Advanced Configurations	Select the type of advanced configurations. ○ **Default**: Use the default advanced configurations. ○ **Custom**: Use custom configurations. You can set the following parameters: ■ **Protocol**: Select the protocol used by the SSL-VPN connection. Valid values: UDP and TCP. ■ **Port**: Specify the port used by the SSL-VPN connection. Invalid values: 22, 2222, 22222, 9000, 9001, 9002, 7505, 80, 443, 53, 68, 123, 4510, 4560, 500, and 4500. ■ **Encryption Algorithm**: Select the encryption algorithm used by the SSL connection. Valid values: AES-128-CBC, AES-192-CBC, AES-256-CBC, and none. ■ **Enable Compression**: Specify whether to compress the data that is transmitted over the SSL-VPN connection.

20.1.8.1.2. Modify an SSL server

After you create an SSL server, you can modify its configurations.

Procedure

1.
2.
3.
4. On the **SSL Servers** page, find the SSL server that you want to manage and click **Modify** in the **Actions** column.
5. On the **Edit SSL Server** page, modify the name, server CIDR block, client CIDR block, and advanced settings of the SSL server, and then click **OK**.

   For more information about the parameters on the buy page, see Create an SSL server.

20.1.8.1.3. Configure a security group

This topic describes how to configure a security group to control the inbound and outbound traffic of Elastic Compute Service (ECS) instances in the security group after an SSL-VPN connection is created.

Procedure

1.
2.
3.
4. On the **SSL Servers** page, find the SSL server that you want to manage and click **Configure Routing Group** in the **Actions** column.
5. In the **Configure Routing Group** panel, set the following parameters and click **OK**.

Parameter	Description
Security Group	Select the security group to which you want to add the security group rule.

Parameter	Description
Rule Direction	Select the direction of data transfer that the rule controls. ○ **Outbound**: controls data transfer from the ECS instances in the security group to the Internet or other ECS instances. ○ **Inbound**: controls data transfer from the Internet or other ECS instances to the ECS instances in the security group.
Action	Specify the action to be performed on the matching requests. ○ **Allow**: accepts requests. ○ **Deny**: drops requests without returning messages. If two security group rules use the same settings except for different actions, the **Deny** rule prevails over the **Allow** rule.
Protocol Type	The protocol of the security group rule.
Port Range	Enter a port range for the security group rule. Valid values: -1 and 1 to 65535. You cannot enter only -1. Examples: ○ 1/200 specifies ports 1 to 200. ○ 80/80 specifies port 80. ○ -1/-1 specifies all ports.
Priority	Set the priority of the rule. Valid values: 1 to 100. The default value is 1, which indicates the highest priority.
Authorization Mode	Specify the type of addresses that the security group rule permits or blocks. You can select only **Address**, which indicates CIDR blocks.
ENI Type	Specify the type of data transfer that the security group rule controls. ○ **Internal**: controls data transfer within Alibaba Cloud. ○ **External**: controls data transfer over the Internet.
Authorization IP Addresses	Specify the CIDR blocks that you want the security group rule to accept or block. You can specify at most 10 CIDR blocks.
Description	Enter a description for the security group rule. The description must be 2 to 256 characters in length, and cannot start with `http://` or `https://`. You can leave this parameter empty.

20.1.8.1.4. Delete an SSL server

This topic describes how to delete an SSL server.

Procedure

1.
2.
3.

4. On the **SSL Servers** page, find the SSL server that you want to delete and click **Delete** in the **Actions** column.
5. In the **Delete SSL Server** message, click **OK**.

20.1.8.2. Manage an SSL client certificate

20.1.8.2.1. Create an SSL client certificate

After you create an SSL server, you must create an SSL client certificate based on the configuration of the SSL server.

Prerequisites

You have created an SSL server. For more information, see Create an SSL server.

Procedure

1.
2.
3. In the top navigation bar, select the region where you want to create the SSL client.
4. On the **SSL Clients** page, click **Create Client Certificate**.
5. On the **Create SSL Client Certificate** page, configure the client certificate based on the following information, and then click **Submit**.

Parameter	Description
Organization	Select the organization to which the SSL client belongs.
Resource Set	Select the resource set to which the SSL client belongs.
Region	Select the region where the SSL client is deployed.
Zone	Select the zone where the SSL client is deployed.
Name	Enter a name for the SSL client certificate. The name must be 2 to 128 characters in length, and can contain digits, hyphens (-), and underscores (_). It must start with a letter and cannot start with `http://` or `https://`.
VPN Gateway	Select the VPN gateway that you want to associate with the SSL client certificate.
SSL Server	Select the SSL server that you want to associate with the SSL client certificate.

20.1.8.2.2. Download an SSL client certificate

This topic describes how to download an SSL client certificate. You can download an SSL client certificate in the VPN Gateway console.

Prerequisites

An SSL client certificate is created. For more information, see Create an SSL client certificate.

Procedure

1.
2.

3.
4. On the **SSL Clients** page, find the SSL client certificate that you want to download and click **Download** in the **Actions** column.

20.1.8.2.3. Delete an SSL client certificate

This topic describes how to delete an SSL client certificate.

Procedure

1.
2.
3.
4. On the **SSL Clients** page, find the SSL client certificate that you want to delete, and click **Delete** in the **Actions** column.
5. In the **Delete Client Certificate** dialog box, click **OK**.

20.1.8.3. Query SSL-VPN connection logs

This topic describes how to query the logs of an SSL server and an SSL client. To troubleshoot the issues in an SSL-VPN connection, you can query the logs of the SSL server and the SSL client to which the connection is established.

Context

You can query logs that are created within the most recent month. The time range that you can specify for a query cannot exceed 10 minutes.

Query the logs of an SSL server.

1.
2.
3.
4. On the **SSL Servers** page, find the SSL server that you want to manage and click **View Logs** in the **Actions** column.
5. In the **SSL VPN Connection Logs** panel, specify the time range to query logs.

Query the logs of an SSL client

1.
2.
3.
4. On the **SSL Clients** page, find the SSL client certificate that you want to manage and click **View Logs** in the **Actions** column.
5. In the **SSL-VPN Client Logs** panel, specify the time range to query logs.

21.Elastic IP Address

21.1. User Guide

21.1.1. EIP overview

An elastic IP address (EIP) is a public IP address that you can purchase and use as an independent resource. You can associate an EIP with an Elastic Compute Service (ECS) instance, an internal-facing Server Load Balancer (SLB) instance, or a secondary elastic network interface (ENI) deployed in a virtual private cloud (VPC). You can also associate an EIP with a NAT gateway or a high-availability virtual IP address (HAVIP).

An EIP is a NAT IP address provisioned in the Internet-facing gateway of Alibaba Cloud and is mapped to the associated cloud resource by using NAT. After an EIP is associated with a cloud resource, the cloud resource can use the EIP to communicate with the Internet.

Differences between an EIP and the static public IP address of an ECS instance

The following table describes the differences between an EIP and the static public IP address of an ECS instance.

Item	EIP	Static public IP address
Supported network	VPC	VPC
Used as an independent resource	Supported	Not supported
Associated with and disassociated from an ECS instance at any time	Supported	Not supported
Displayed in the ENI information of the associated ECS instance	No	No

Benefits

EIPs have the following benefits:

- Purchase and use as independent resources

  You can purchase and use an EIP as an independent resource. EIPs are not bundled with other computing or storage resources.

- Associate with resources at any time

  You can associate an EIP with a cloud resource as needed. You can also disassociate and release an EIP at any time.

- Modify bandwidth limits on demand

  You can modify the bandwidth limit of an EIP at any time to meet your business requirements. The modification immediately takes effect.

21.1.2. Log on to the EIP console

This topic describes how to log on to the Apsara Uni-manager Management Console to manage your elastic IP addresses (EIPs). The Google Chrome browser is used as an example.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.

- A browser is available. We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 8 to 20 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Login**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Networking > Elastic IP Address**.

21.1.3. Quick start

21.1.3.1. Overview

This topic describes how to create an elastic IP address (EIP) and associate the EIP with an Elastic Compute Service (ECS) instance. This allows the ECS instance to access the Internet.

This topic includes the following operations:

1. Apply for an EIP

   An EIP is a public IP address that you can purchase and use as an independent resource. Before you get started, you must apply for an EIP.

2. Associate an EIP with an ECS instance

   You can associate an EIP with an ECS instance that is deployed in a virtual private cloud (VPC). After the ECS instance is associated with an EIP, the ECS instance can access the Internet.

3. Disassociate an EIP from a cloud resource

   If you do not want the ECS instance to access the Internet, you can disassociate the EIP from the ECS instance.

4. Release an EIP

   You can release an EIP that is no longer in use.

21.1.3.2. Apply for an EIP

This topic describes how to apply for an elastic IP address (EIP). An EIP is a public IP address that you can purchase and use as an independent resource.

Procedure

1. Log on to the EIP console.
2. On the **Elastic IP Addresses** page, click **Create EIP**.
3. On the **Create Elastic IP Address** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the EIP belongs.
Resource Set	Select the resource set to which the EIP belongs.
Region	Select the region where you want to create the EIP. Make sure that the EIP and the cloud resource to be associated with the EIP are deployed in the same region.
Quantity	Enter the number of EIPs that you want to purchase.
EIP Name	Enter the name of the EIP.
Connection Type	Select the connection type of the EIP.
Network Type	Select a network type for the EIP. ◦ **Internet**: The EIP is used for communication over the Internet. ◦ **Hybrid Cloud**: The EIP is used to establish communication within a hybrid cloud. For example, if you want to allow a data center to access the Internet by using SNAT and DNAT, you must select this type.
IP Address	Enter the EIP for which you want to apply. Make sure that you enter an idle IPv4 address. If the IPv4 address is already in use, your application fails. **Note** If you do not specify an IPv4 address, the system randomly allocates one.
Maximum Bandwidth	The bandwidth limit of the EIP. Unit: Mbit/s.

21.1.3.3. Associate an EIP with an ECS instance

This topic describes how to associate an elastic IP address (EIP) with an Elastic Compute Service (ECS) instance that is deployed in a virtual private cloud (VPC). After the ECS instance is associated with an EIP, the ECS instance can access the Internet.

Prerequisites

An ECS instance is created. For more information, see the **Create an instance** topic in the **Quick Start** chapter of

ECS User Guide.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to manage and click **Bind Resource** in the **Actions** column.
4. In the **Bind Elastic IP Address to Resources** dialog box, set the following parameters and click **OK**.

Parameter	Description
Instance Type	Select **ECS Instance**.
Binding mode	Select the mode in which you want to associate the EIP. Only Normal mode is supported. In Normal mode: ○ The EIP is associated with the ECS instance in NAT mode. Both the private IP address and public IP address of the ECS instance are available for use. ○ The EIP is not displayed in the operating system. To query the EIP of the ECS instance, call the DescribeInstances operation. ○ The EIP does not support NAT application layer gateway (ALG) protocols such as H.323, Session Initiation Protocol (SIP), Domain Name System (DNS), Real-Time Streaming Protocol (RTSP), and Trivial File Transfer Protocol (TFTP).
Select an instance to bind	Select the ECS instance with which you want to associate the EIP. Make sure that the following requirements are met: ○ The ECS instance is deployed in a VPC. ○ The ECS instance is in the Running or Stopped state. ○ Each ECS instance can be associated only with one EIP. ○ The ECS instance and the EIP are created in the same region. ○ The ECS instance is not assigned a static public IP address. In addition, the ECS instance is not associated with another EIP. **Note** To quickly find the ECS instance that you want to manage, you can use one of the following methods: ○ ECS attributes: Select **ECS Instance Name** or **ECS Instance ID** from the drop-down list and then enter the name or ID of the ECS instance. ○ Tags: In the **Filter by Tag** drop-down list, select the tag key and value of the ECS instance. If you use this method, make sure that the ECS instance is added with the specified tag.

21.1.3.4. Disassociate an EIP from a cloud resource

This topic describes how to disassociate an elastic IP address (EIP) from a cloud resource. After the EIP is disassociated from the cloud resource, the cloud resource can no longer access the Internet by using the EIP.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to disassociate and click **Unbind** in the **Actions**

column.

4. In the message that appears, click **OK**.

21.1.3.5. Release an EIP

This topic describes how to release an elastic IP address (EIP) that you no longer need.

Prerequisites

The EIP is disassociated with the cloud resource. For more information, see Disassociate an EIP from a cloud resource.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to release and choose > **Release** in the **Actions** column.
4. In the **Release EIP** message, click **OK**.

21.1.4. Manage EIPs

21.1.4.1. Apply for an EIP

This topic describes how to apply for an elastic IP address (EIP). An EIP is a public IP address that you can purchase and use as an independent resource.

Procedure

1. Log on to the EIP console.
2. On the **Elastic IP Addresses** page, click **Create EIP**.
3. On the **Create Elastic IP Address** page, set the following parameters and click **Submit**.

Parameter	Description
Organization	Select the organization to which the EIP belongs.
Resource Set	Select the resource set to which the EIP belongs.
Region	Select the region where you want to create the EIP. Make sure that the EIP and the cloud resource to be associated with the EIP are deployed in the same region.
Quantity	Enter the number of EIPs that you want to purchase.
EIP Name	Enter the name of the EIP.
Connection Type	Select the connection type of the EIP.

Parameter	Description
Network Type	Select a network type for the EIP. ○ **Internet**: The EIP is used for communication over the Internet. ○ **Hybrid Cloud**: The EIP is used to establish communication within a hybrid cloud. For example, if you want to allow a data center to access the Internet by using SNAT and DNAT, you must select this type.
IP Address	Enter the EIP for which you want to apply. Make sure that you enter an idle IPv4 address. If the IPv4 address is already in use, your application fails. **Note** If you do not specify an IPv4 address, the system randomly allocates one.
Maximum Bandwidth	The bandwidth limit of the EIP. Unit: Mbit/s.

21.1.4.2. Bind an EIP to a cloud instance

21.1.4.2.1. Associate an EIP with an ECS instance

This topic describes how to associate an elastic IP address (EIP) with an Elastic Compute Service (ECS) instance that is deployed in a virtual private cloud (VPC). After the ECS instance is associated with an EIP, the ECS instance can access the Internet.

Prerequisites

An ECS instance is created. For more information, see the **Create an instance** topic in the **Quick Start** chapter of *ECS User Guide*.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to manage and click **Bind Resource** in the **Actions** column.
4. In the **Bind Elastic IP Address to Resources** dialog box, set the following parameters and click **OK**.

Parameter	Description
Instance Type	Select **ECS Instance**.
Binding mode	Select the mode in which you want to associate the EIP. Only Normal mode is supported. In Normal mode: ○ The EIP is associated with the ECS instance in NAT mode. Both the private IP address and public IP address of the ECS instance are available for use. ○ The EIP is not displayed in the operating system. To query the EIP of the ECS instance, call the DescribeInstances operation. ○ The EIP does not support NAT application layer gateway (ALG) protocols such as H.323, Session Initiation Protocol (SIP), Domain Name System (DNS), Real-Time Streaming Protocol (RTSP), and Trivial File Transfer Protocol (TFTP).

Parameter	Description
Select an instance to bind	Select the ECS instance with which you want to associate the EIP. Make sure that the following requirements are met: ○ The ECS instance is deployed in a VPC. ○ The ECS instance is in the Running or Stopped state. ○ Each ECS instance can be associated only with one EIP. ○ The ECS instance and the EIP are created in the same region. ○ The ECS instance is not assigned a static public IP address. In addition, the ECS instance is not associated with another EIP. **Note** To quickly find the ECS instance that you want to manage, you can use one of the following methods: ○ ECS attributes: Select **ECS Instance Name** or **ECS Instance ID** from the drop-down list and then enter the name or ID of the ECS instance. ○ Tags: In the **Filter by Tag** drop-down list, select the tag key and value of the ECS instance. If you use this method, make sure that the ECS instance is added with the specified tag.

21.1.4.2.2. Associate an EIP with an SLB instance

This topic describes how to associate an elastic IP address (EIP) with a Server Load Balancer (SLB) instance. After you associate an EIP with an SLB instance, the SLB instance can forward requests from the Internet.

Prerequisites

An SLB instance is created. For more information, see the **Create an SLB instance** topic in the **Quick Start** chapter of *SLB User Guide*.

Procedure

1. Log on to the EIP console.
2.
3.
4. In the **Bind Elastic IP Address to Resources** dialog box, set the following parameters and click **OK**.

Parameter	Description
Instance Type	Select **SLB Instance**.

Parameter	Description
Select an instance to bind	Select the SLB instance with which you want to associate the EIP. Make sure that the following requirements are met: ○ The SLB instance is deployed in a virtual private cloud (VPC). ○ The SLB instance and the EIP are created in the same region. ○ Each SLB instance can be associated only with one EIP. **Note** ○ When you associate an EIP with an internal-facing SLB instance, your backend service associated with the SLB instance will be temporarily interrupted. Therefore, we recommend that you perform this operation during off-peak hours or associate the service with another SLB instance first. ○ To quickly find the SLB instance that you want to manage, you can use one of the following methods: ■ SLB attributes: Select **Name** or **SLB Instance ID** from the drop-down list and then enter the name or ID of the SLB instance. ■ Tags: In the **Filter by Tag** drop-down list, select the tag key and value of the SLB instance. If you use this method, make sure that the SLB instance is added with the specified tag.

21.1.4.2.3. Associate an EIP with a NAT gateway

This topic describes how to associate an elastic IP address (EIP) with a NAT gateway. After you associate an EIP with a NAT gateway, you can specify the EIP in DNAT or SNAT entries.

Prerequisites

A NAT gateway is created. For more information, see the **Create a NAT gateway** topic in the **Quick Start** chapter of *NAT Gateway User Guide*.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to manage and click **Bind Resource** in the **Actions** column.
4. In the **Bind Elastic IP Address to Resources** dialog box, set the following parameters and click **OK**.

Parameter	Description
Instance Type	Select **NAT Gateway**.

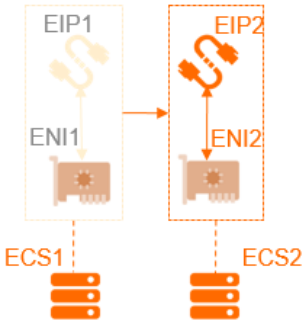
Parameter	Description
Select an instance to bind	Select the NAT gateway with which you want to associate the EIP. The NAT gateway and the EIP must be created in the same region. **Note** To quickly find the NAT gateway that you want to manage, you can use one of the following methods: ○ NAT attributes: Select **NAT Gateway Name** or **NAT Gateway ID** from the drop-down list and then enter the name or ID of the NAT gateway. ○ Tags: Click **Filter by Tag**. In the **Filter by Tag** dialog box, enter the **Tag Key** and **Tag Value** of the NAT gateway and click **Search**. If you use this method, make sure that the NAT gateway is added with the specified tag.

21.1.4.2.4. Associate an EIP with a secondary ENI

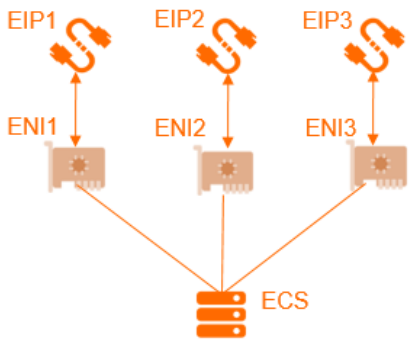
21.1.4.2.4.1. Overview

You can associate elastic IP addresses (EIPs) with elastic network interfaces (ENIs). Then, you can associate the ENIs with an Elastic Compute Service (ECS) instance. This way, the ECS instance can use multiple EIPs. You can use EIPs to improve the service availability, flexibility, and scalability.

Each ENI is assigned a private IP address. After you associate an EIP with an ENI, both the private IP address and the EIP are available for the ENI. You can change the private IP address and public IP address of an ECS instance by replacing the secondary ENI that is associated with the ECS instance. When you replace the secondary ENI of an ECS instance, the reliability and availability of your service are not affected.



You can associate multiple ENIs with an ECS instance. Make sure that an EIP is associated with each ENI. This way, the ECS instance can use multiple EIPs. The ECS instance can use the EIPs to provide Internet-facing services. You can configure security group rules for the ECS instance to control access from the Internet.



Association modes

You can associate an EIP with an ENI in NAT mode.

The following table describes the features in NAT mode.

Item	NAT mode
Whether the EIP is displayed in the ENI information of the operating system	No
Types of ENIs that can be associated with EIPs	Primary and secondary ENIs
Number of EIPs that can be associated with a primary ENI	1
Number of EIPs that can be associated with a secondary ENI	Depends on the number of private IP addresses of the secondary ENI **Note** Each EIP is mapped to a private IP address of a secondary ENI. If a secondary ENI is assigned 10 private IP addresses, at most 10 EIPs can be associated with the secondary ENI.
Whether the private network features of a secondary ENI are available after an EIP is associated with the secondary ENI	Yes
Supported protocols	EIPs deployed as NAT application layer gateways (ALGs) do not support protocols such as H.323, Session Initiation Protocol (SIP), Domain Name System (DNS), and Real-Time Streaming Protocol (RTSP).

21.1.4.2.4.2. Associate an EIP with a secondary ENI in normal mode

This topic describes how to associate an elastic IP address (EIP) with a secondary elastic network interface (ENI) in normal mode. After you associate an EIP with a secondary ENI, both the private IP addresses and public IP addresses of the secondary ENI are available for use. The EIP is not displayed in the secondary ENI information.

Prerequisites

Before you associate an EIP with a secondary ENI in normal mode, make sure that the following requirements are met:

- A secondary ENI is created in a virtual private cloud (VPC). The secondary ENI and the EIP are created in the same region. For more information, see the **Create an ENI** topic in the **Elastic Network Interface** chapter of *ECS User Guide*.
- The secondary ENI is not associated with an Elastic Compute Service (ECS) instance. If the secondary ENI is associated with an ECS instance, disassociate the secondary ENI from the ECS instance, associate the EIP with the secondary ENI in normal mode, and then associate the secondary ENI with the ECS instance. For more information, see the **Disassociate a secondary ENI from an ECS instance** topic in the **Elastic Network Interface** chapter of *ECS User Guide*.

Procedure

1. Log on to the EIP console.
2. In the top navigation bar, select the region where the EIP is deployed.
3. On the **Elastic IP Addresses** page, find the EIP that you want to manage and click **Bind Resource** in the

Actions column.

4. In the **Bind Elastic IP Address to Resources** dialog box, set the following parameters and click **OK**.

Parameter	Description
Instance Type	The type of the instance. In this example, **Secondary ENI** is selected.
Binding mode	Select **Normal**. In Normal mode: ○ The number of EIPs that can be associated with a secondary ENI depends on the number of private IP addresses that are assigned to the secondary ENI. ○ The EIP is associated with the ENI in NAT mode. Both the private IP addresses and public IP addresses of the ENI are available for use. ○ The EIP is not displayed in the operating system. To query the EIP, call the DescribeEipAddresses operation. ○ The EIP does not support NAT application layer gateway (ALG) protocols such as H.323, Session Initiation Protocol (SIP), Domain Name System (DNS), Real-Time Streaming Protocol (RTSP), and Trivial File Transfer Protocol (TFTP).
Select an instance to bind	Select the secondary ENI with which you want to associate the EIP. Make sure that the secondary ENI meets the following requirements: ○ The secondary ENI is deployed in a VPC. ○ The secondary ENI and the EIP are created in the same region.

21.1.4.3. Increase the bandwidth limit of an EIP

You can increase the bandwidth limit of elastic IP addresses (EIPs). After you increase the bandwidth limit of an EIP, the changes immediately takes effect.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to manage and choose > **Modify Configuration** in the **Actions** column.
4. On the **Change Specifications** page, specify a new bandwidth limit and click **Submit**.

21.1.4.4. Disassociate an EIP from a cloud resource

This topic describes how to disassociate an elastic IP address (EIP) from a cloud resource. After the EIP is disassociated from the cloud resource, the cloud resource can no longer access the Internet by using the EIP.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to disassociate and click **Unbind** in the **Actions** column.
4. In the message that appears, click **OK**.

21.1.4.5. Release an EIP

This topic describes how to release an elastic IP address (EIP) that you no longer need.

Prerequisites

The EIP is disassociated with the cloud resource. For more information, see Disassociate an EIP from a cloud resource.

Procedure

1. Log on to the EIP console.
2.
3. On the **Elastic IP Addresses** page, find the EIP that you want to release and choose > **Release** in the **Actions** column.
4. In the **Release EIP** message, click **OK**.

22.Apsara Stack Security

22.1. User Guide

22.1.1. What is Apsara Stack Security?

Apsara Stack Security is a solution that provides a full suite of security features, such as network, server, application, data, and security management to protect Apsara Stack assets.

Background information

Traditional security solutions for IT services use hardware products such as firewalls and intrusion prevention systems (IPSs) to detect attacks on network perimeters and protect networks against attacks.

Cloud computing features low costs, on-demand flexible configuration, and high resource utilization. As cloud computing develops, an increasing number of enterprises and organizations use cloud computing services instead of traditional IT services. Cloud computing environments do not have definite network perimeters. As a result, traditional security solutions cannot effectively safeguard cloud assets.

With the powerful data analysis capabilities and professional security operations team of Alibaba Cloud, Apsara Stack Security provides integrated security protection services for networks, applications, and servers.

Complete security solution

Apsara Stack Security consists of Apsara Stack Security Standard Edition and optional security services and provides a comprehensive security solution.

Security domain	Service name	Description
Security management	Threat Detection Service (TDS)	Monitors traffic and overall security status to audit and manage assets in a centralized manner.
Server security	Server Guard	Protects Elastic Compute Service (ECS) instances against intrusions and malicious code.
	Server Security	Protects physical servers against intrusions.
Application security	Web Application Firewall (WAF)	Protects web applications against attacks and ensures that users of mobile devices and PCs can access web applications over the Internet in a secure manner.
Network Security	Anti-DDoS	Ensures the availability of network links and improves business continuity.
Data security	Sensitive Data Discovery and Protection (SDDP)	Prevents data leaks and helps your business system meet compliance requirements.
O&M audit	Security Audit	Summarizes and analyzes logs. This way, security auditors can detect and eliminate risks at the earliest opportunity.
Security O&M service	On-premises security service	Helps you build and optimize the cloud security system to protect your business system against attacks by using security features of Apsara Stack Security and other Apsara Stack services.

22.1.2. Usage notes

Before you log on to Apsara Stack Security Center, you must verify that your computer meets the configuration requirements.

For more information about the configuration requirements, see Configuration requirements.

Configuration requirements

Item	Requirement
Browser	• Internet Explorer: V11 or later • Google Chrome (recommended): V42.0.0 or later • Mozilla Firefox: V30 or later • Safari: V9.0.2 or later • GmSSL browser that runs the Chrome kernel: V69.0.0 or later
Operating system	• Windows XP • Windows 7 or later • macOS

22.1.3. Quick start

22.1.3.1. User roles and permissions

This topic describes the user roles involved in Apsara Stack Security.

All roles in Apsara Stack Security Center are provided by default. You cannot add custom roles. Before you log on to Apsara Stack Security Center, make sure that your account is assigned the required role. For more information, see Default roles in Apsara Stack Security.

Default roles in Apsara Stack Security

Role	Permission
System administrator of Apsara Stack Security Center	Manages and configures system settings for Apsara Stack Security Center. The system administrator has permissions to manage Apsara Stack accounts, synchronize data, configure alerts, and configure global settings.
Security administrator of Apsara Stack Security Center	Monitors the security status across Apsara Stack and configures security policies for each functional module of Apsara Stack Security. The security administrator has permissions on all features under Threat Detection, Network Security, Application Security, Server Security, Physical Server Security, and Asset Management. **Note** The permissions on Web Application Firewall (WAF) must be separately assigned.
Department security administrator	Monitors the security status of cloud resources in a specific department and configures security policies for each functional module of Apsara Stack Security for this department. The department security administrator has permissions on all features under Threat Detection, Network Security, Application Security, Server Security, Physical Server Security, and Asset Management. In addition, the department security administrator can specify alert notification methods and alert contacts in the department. **Note** The permissions on WAF must be separately assigned.

Role	Permission
Auditor of Apsara Stack Security Center	Conducts security audits across Apsara Stack. The auditor can view audit events and raw logs, configure audit policies, and access all features under Security Audit.

If you do not have an account that assumes the required role, contact the administrator to create an account and assign the role to the account. For more information, see the **Create a user** topic in *Apsara Uni-manager Management Console User Guide*.

22.1.3.2. Log on to Apsara Stack Security Center

This topic describes how to log on to Apsara Stack Security Center.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Security > Alibaba Cloud Security**.
6. On the **Apsara Stack Security Center** page, select a value for **Region**.
7. Click **Access with Authorized Role** to access Apsara Stack Security Center.

22.1.4. Threat Detection Service

22.1.4.1. Overview

This topic introduces the basic concepts related to Threat Detection Service (TDS).

TDS provides comprehensive protection for enterprises. It can monitor vulnerabilities, intrusions, web attacks, DDoS attacks, threat intelligence, and public opinions. TDS uses modeling and analysis to obtain key information based on traffic characteristics, host behavior, and host operation logs. In addition, TDS identifies intrusions that cannot be detected by traffic inspection or file scan. You can use the input of cloud analysis models and intelligence data to discover sources and behavior of attacks and assess threats.

TDS provides the following features:

- Overview: provides a security situation overview and information about security screens.
- Security Alerts: displays security alerts that occur in your business system.
- Attack Analysis: displays application attacks and brute-force attacks that occur in your system.
- Cloud Service Check: checks whether risks exist in the configurations of Apsara Stack services.
- Application Whitelist: allows you to create and apply application whitelist policies to your servers that require special protection. After you create the policies, Apsara Stack Security identifies trusted, suspicious, and malicious programs based on intelligent learning. This prevents unauthorized programs from running.
- Assets: manages servers and cloud services on Apsara Stack.
- Security Reports: allows you to configure security report tasks on Apsara Stack.

22.1.4.2. Security overview

22.1.4.2.1. View security overview information

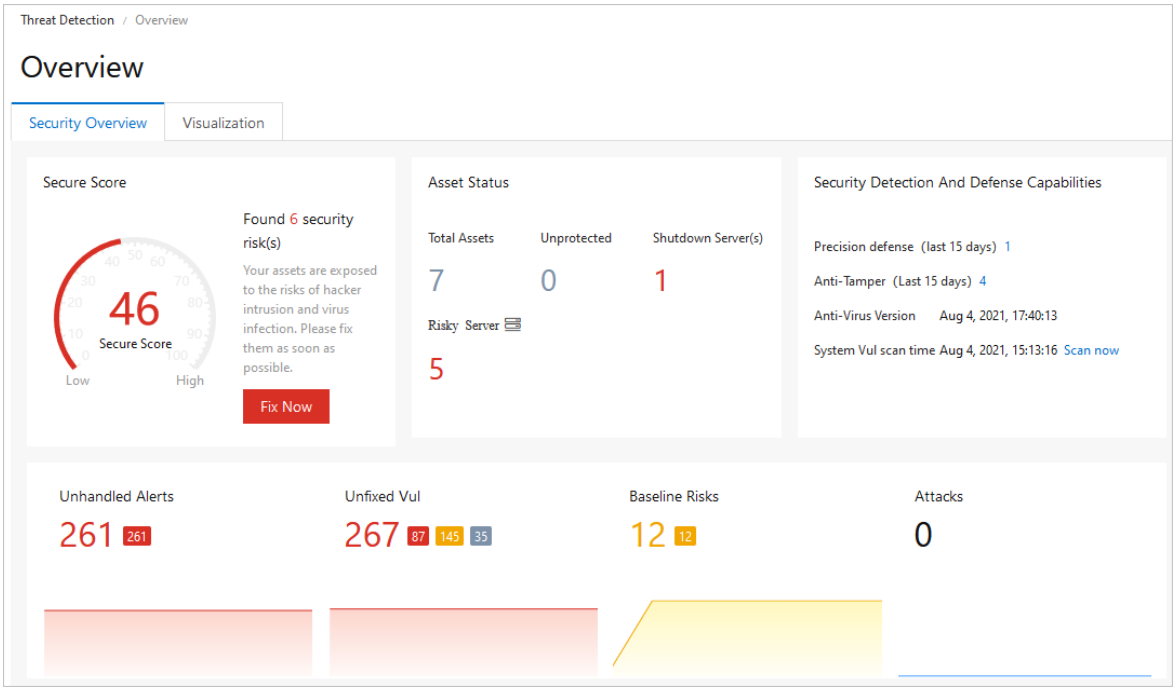
This topic describes how to view security statistics, attack trends, and network traffic information on the Apsara Stack platform.

Context

The **Security Overview** tab provides an overview of detected security events, the latest threats, and inherent vulnerabilities of the system. A security administrator can view information on the **Security Overview** tab to better understand the security posture of the system.

Procedure

1.
2.
3. click **Overview**.
4. On the **Security Overview** tab, view the security posture of the Apsara Stack platform.



Sections on the Security Overview tab

Section	Description
Secure Score	The security score of assets and the number of detected security risks.
Asset Status	The total number of assets and the numbers of servers that are not protected, servers that are stopped, and servers that are at risk.
Security Detection And Defense Capabilities	The numbers of precise defense events and anti-tampering events over the last 15 days, the time when the antivirus database was last updated, and the time when vulnerability scanning was last performed. This allows you to obtain the defense situation and security status of your assets in real time.
Threat statistics	The numbers of alerts that are not handled, vulnerabilities that are not fixed, baseline risks, and attacks.
Configuration Assessment Risks	The risks in the baseline configurations of cloud services.
Issue Resolved	Statistics on alerts, vulnerabilities, and baseline risks that have been processed over the last 15 days. The statistics are displayed in a bar and trend chart.

22.1.4.3. Security alerts

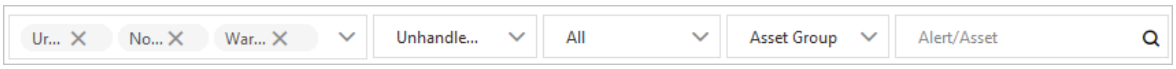
22.1.4.3.1. View security alerts

This topic describes how to view security alerts on the Security Alerts page.

Procedure

1.
2.
3. click **Security Alerts**.
4. (Optional)Specify filter conditions to search for security alerts.

> **Note** If you want to view all alerts, do not specify the conditions.



Filter condition	Description
Alert level	The alert level. You can select one or more levels. Valid values: ○ Urgent ○ Warning ○ Notice
Alert status	The alert status. Valid values: ○ Unhandled Alerts ○ Handled
Alert type	The alert type. Select **All** or a specific type.
Affected asset group	The affected asset group. Select **Asset Group** or a specific group.
Alert name or asset keyword	The alert name or the keywords of affected assets.

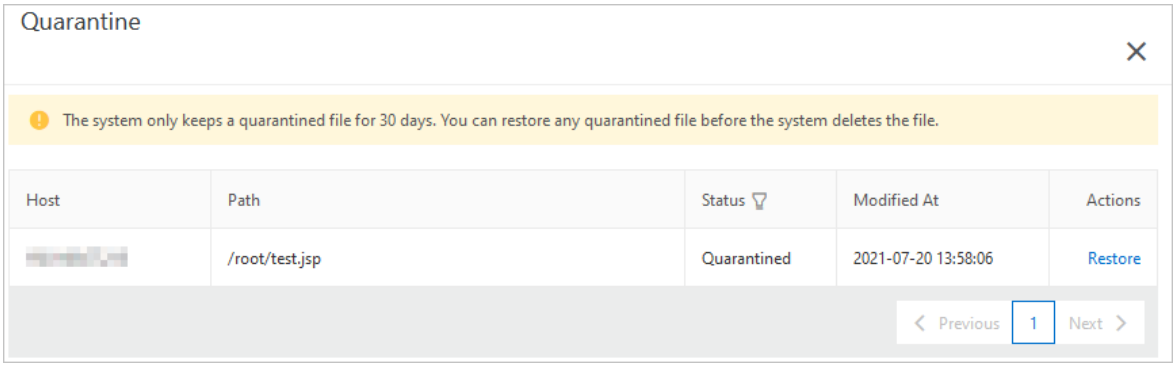
5. View security alerts and their details in the alert list.

22.1.4.3.2. Manage quarantined files

This topic describes how to manage threat files that are quarantined by the system. The system deletes a quarantined file 30 days after the file is quarantined. You can restore the file before it is deleted.

Procedure

1.
2.
3. click **Security Alerts**.
4. In the upper-right corner of the **Alerts** page, click **Quarantine**.
5. In the **Quarantine** panel, view the information about a quarantined file, such as the IP address of the host, path, status, and operation time.



6. (Optional)If a file is incorrectly quarantined, click **Restore** in the **Actions** column to restore the file.

> **Notice** Before you restore a quarantined file, make sure that the file is normal and does not bring risks.

The restored file is removed from the Quarantine panel and is displayed in the security alert list again.

22.1.4.3.3. Configure security alerts

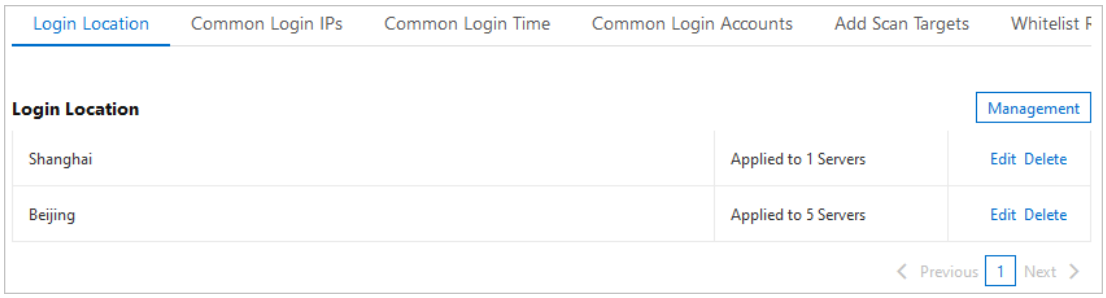
This topic describes how to configure security alerts, which allow you to specify approved logon locations and web directories to scan.

Procedure

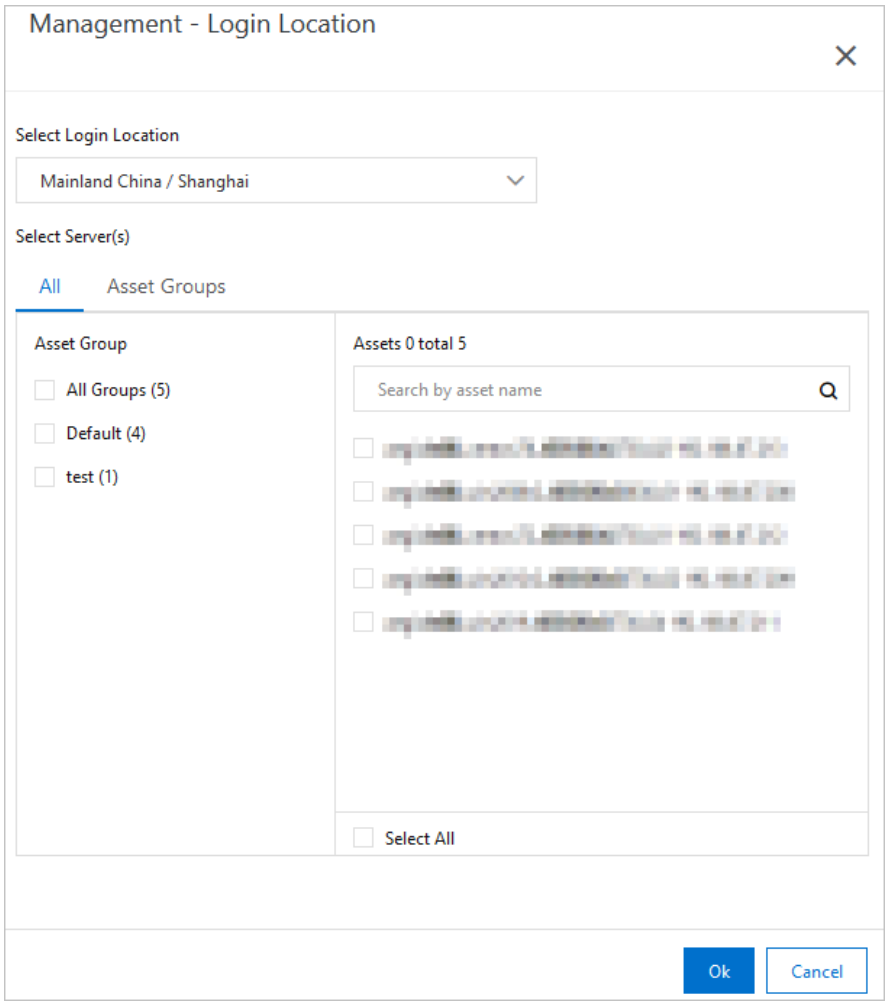
1.
2.
3. click **Security Alerts**.
4. In the upper-right corner of the **Alerts** page, click **Settings**.

   In the **Settings** panel, you can perform the following operations:

   - **Add an approved logon location**
     a. Click **Management** to the right of **Login Location**.



b. In the **Management - Login Location** panel, select the logon location that you want to add and select the servers that allow logons from the added location.



c. Click **Ok**.

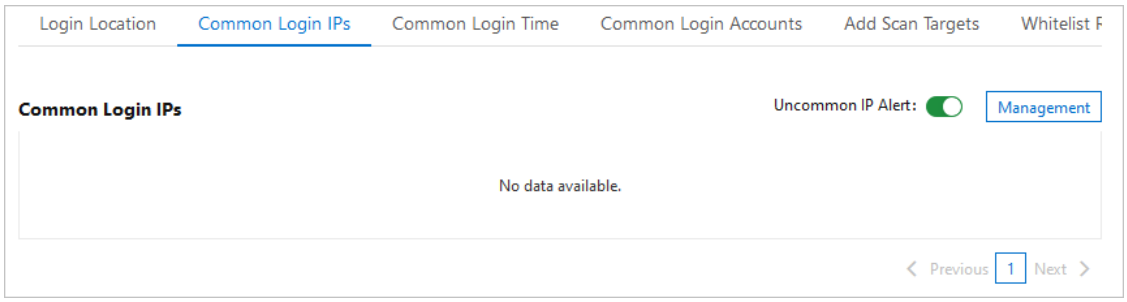
Threat Detection Service (TDS) allows you to **edit** and **delete** added logon locations.

- Find the required logon location and click **Edit** on the right to change the servers that allow logons from this location.
- Find the required logon location and click **Delete** on the right to delete the logon location.

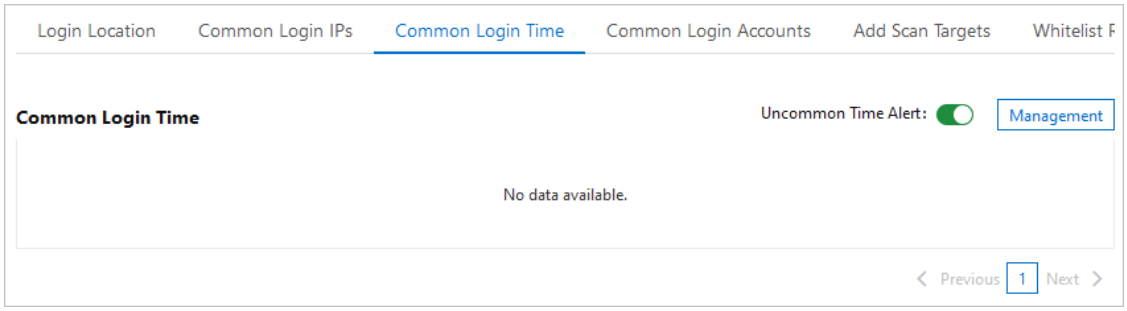
- **Configure advanced logon settings**

> **Note** When you configure advanced logon settings, you can specify the IP addresses, accounts, and time ranges that are allowed for logons to your assets. After you configure these settings, alerts are triggered if your assets receive logon requests that do not meet the requirements. The procedure to configure advanced logon settings is similar to that to configure **approved logon locations**. You can follow the preceding procedure to **add**, **edit**, or **delete** advanced logon settings.

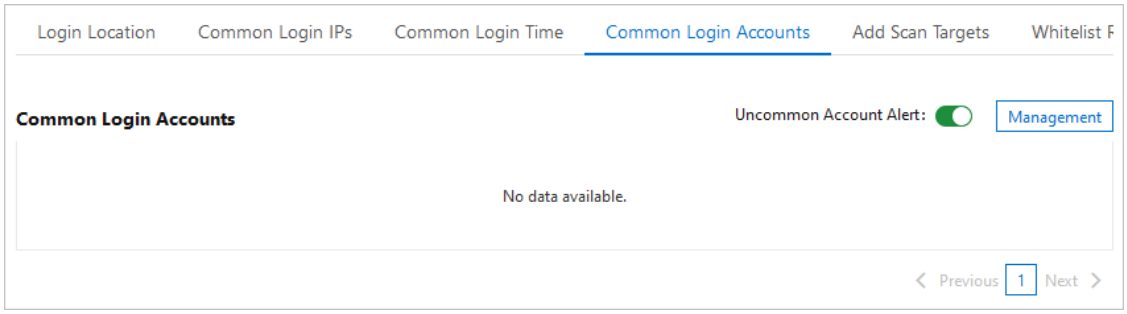
- On the right of **Common Login IPs**, turn on or off Uncommon IP Alert. After the switch is turned on, alerts are triggered if your assets receive logon requests from unapproved IP addresses.



- On the right of **Common Login Time**, turn on or off Uncommon Time Alert. After the switch is turned on, alerts are triggered if your assets receive logon requests at unapproved time.

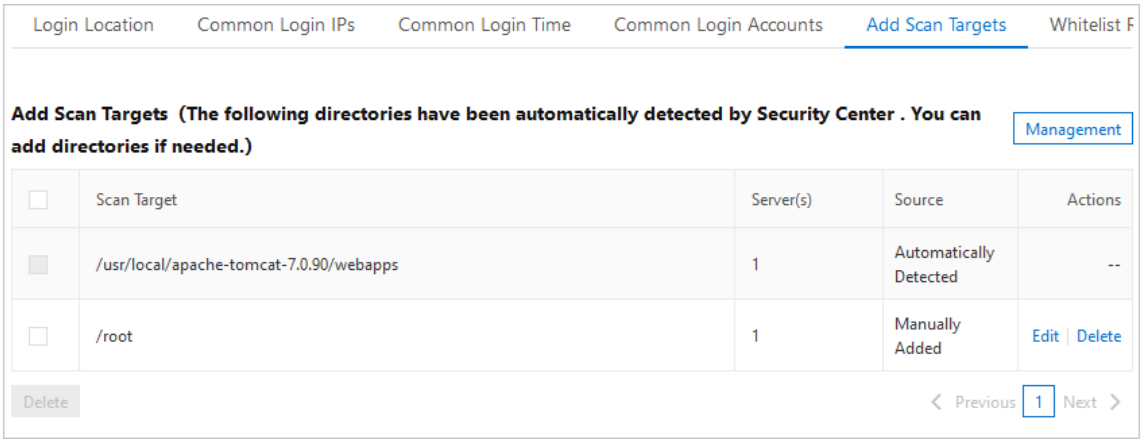


- On the right of **Common Login Accounts**, turn on or off Uncommon Account Alert. After the switch is turned on, alerts are triggered if your assets receive logon requests from unapproved accounts.



- **Add a web directory to scan**

  Apsara Stack Security automatically scans web directories of your servers and runs dynamic and static scan tasks. You can also manually add other web directories of your servers.



  a. On the right of **Add Scan Targets**, click **Management**.

b. Enter a valid web directory and select the servers on which the directory is scanned. The web directory is added to the scan list.

> **Note** Root directories are not allowed. This ensures performance and efficiency.

c. Click **Ok**.

22.1.4.4. Attack analysis

This topic describes the statistics provided by the attack analysis feature. The statistics include the total number of attacks, distribution of attack types, top five attack sources, top five attacked assets, and an attack list.

Background information

The attack analysis feature provides basic attack detection and prevention capabilities in Apsara Stack Security Center. We recommend that you optimize firewalls and enhance business security to develop a more fine-grained and in-depth defense system.

On the **Attack Awareness** page, you can specify a time range to view these attack details. You can view the attack analysis statistics of the current day, last 7 days, or last 15 days. You can also set Time Range to **Custom** to view the statistics of a time range within the last 30 days.

- **Attacks**: the total number of attacks detected in your assets within a specified time range.
- **Attack Type Distribution**: the attack types and the number of attacks for each type.
- **Top 5 Attack Sources**: the top five IP addresses from which the most attacks are launched.
- **Top 5 Attack Assets**: the top five assets that are attacked the most frequently.
- **Attack list**: the details about each attack. The details include the attack time, source IP address, attacked asset, attack type, and total number of attacks.

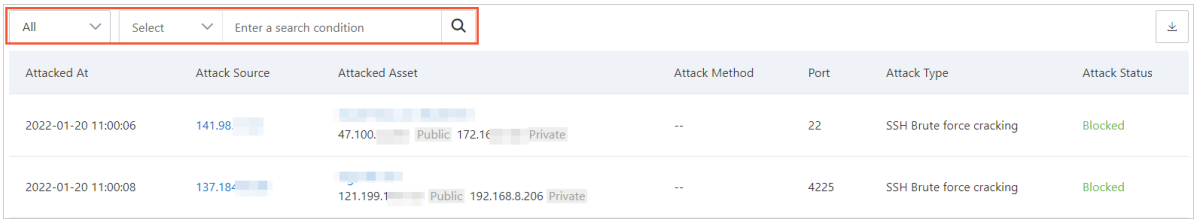
  > **Note** The attack list displays a maximum of 10,000 attacks. You can specify **Time Range** to view details about the attacks that occur over the specified time range.

  Parameters in the attack list

Parameter	Description
Attacked At	The time at which an attack occurs.
Attack Source	The source IP address of an attack.
Attacked Asset	The name, public IP address, and private IP address of an attacked asset.
Attacks	The total number of times that an attack is launched.
Attack Type	The type of an attack. The types of attacks that can be detected include SSH brute-force attacks and remote code execution attacks.

- Search for an attack.

To search for an attack and view the details about the attack, specify search conditions above the attack list. Search conditions include the attack type, attacked asset, source IP address, and port number.



- View the details of an attacked asset.

  To view the details about an attacked asset, move the pointer over the name of the attacked asset.

- Export the attack list.

  To export and save the attack list to your computer, click the icon in the upper-left corner above the attack list. The attack list is exported to an Excel file.

22.1.4.5. Cloud service check

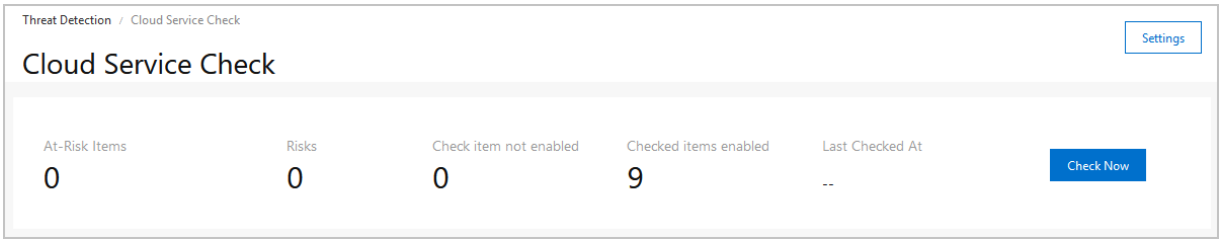
22.1.4.5.1. Overview

Threat Detection Service (TDS) provides the cloud service check feature. This feature allows you to check for the configuration risks of your Apsara Stack services. This topic describes the features and check items that are supported by the cloud service check feature.

Background information

The cloud service check feature allows you to perform network access control and data security checks. The checks help you detect configuration risks of your Apsara Stack services and provide repair solutions.

You can view the number of **Checked items enabled** on the **Cloud Service Check** page.



Cloud service check list

The following table describes the check items.

Type	Supported item	Description
	PolarDB - Backup configurations	Checks whether the automatic backup feature is enabled for PolarDB. Regular backups help you improve database security. You can restore data if an error occurs in your database. PolarDB supports automatic backup. We recommend that you enable automatic backup to create a backup on a daily basis.

Type	Supported item	Description
Data security	Container Registry - Repository permission configurations	Checks whether a Container Registry repository is set to private. Container Registry supports public and private repositories. Public repositories allow users to anonymously download images over the Internet. If images in a repository contain sensitive information, we recommend that you set the repository to private. If images in a repository do not contain sensitive information, ignore related alerts.
	OSS - Server-side encryption	Checks whether the data encryption feature is enabled for Object Storage Service (OSS) buckets. OSS supports server-side encryption to secure data that is persistently stored in OSS. We recommend that you enable server-side encryption to protect sensitive data.
	OSS - Sensitive information leakage scans	Checks whether access permissions on sensitive files in OSS buckets are required.
	ApsaraDB RDS - Cross-region backup configurations	Checks whether the cross-region backup feature is enabled for ApsaraDB RDS instances. ApsaraDB RDS for MySQL provides the cross-region backup feature that automatically synchronizes local backup files to OSS buckets in another region. This implements geo-disaster recovery. We recommend that you enable the cross-region backup feature.
	KVStore for Redis - Backup configurations	Checks whether the data backup feature is enabled for KVStore for Redis instances.
	ApsaraDB for MongoDB - SSL encryption	Checks whether SSL encryption is enabled for ApsaraDB for MongoDB databases. We recommend that you enable the SSL encryption feature to improve the security of data links in ApsaraDB for MongoDB databases.
	ApsaraDB for MongoDB - Backup configurations	Checks whether the automatic backup feature is enabled for ApsaraDB for MongoDB databases. Regular backups help you improve database security. You can restore data if an error occurs in your database. ApsaraDB for MongoDB provides automatic backup policies. We recommend that you enable automatic backup to create a backup on a daily basis.
	ECS - Disk encryption	Checks whether encryption is enabled for disks on Elastic Compute Service (ECS) instances.
	ECS - Automatic snapshot policies	Checks whether the automatic snapshot feature is enabled for the disks on ECS instances. The automatic snapshot feature improves the security of ECS instances and supports disaster recovery.
	OSS - Bucket permissions	Checks whether the OSS bucket ACL is set to *private*.
	OSS - Logging	Checks whether the logging feature is enabled for OSS.
	OSS - Cross-region replication	Checks whether the cross-region replication feature is enabled for OSS.
	ApsaraDB RDS - Database security policies	Checks whether the SQL audit, SSL encrypted transmission, and transparent database encryption features are enabled for ApsaraDB RDS databases.

Type	Supported item	Description
	ApsaraDB RDS - Backup configurations	Checks whether the data backup feature is enabled for ApsaraDB RDS instances.
	SSL Certificates Service - Expiration check	Checks whether your SSL certificate expires. If your SSL certificate expires, you are not allowed to use SSL Certificates Service.
Network access control	ECS - Security group policies	Checks the security group policies of ECS. We recommend that you grant permissions to users based on the principle of least privilege. We also recommend that you specify 0.0.0.0/0 only for the ports that must be open to all services, such as port 80, 443, 22, or 3389.
	OSS - Bucket hotlink protection	Checks whether the hotlink protection feature is enabled for OSS buckets. The OSS hotlink protection feature checks the Referer header to deny access from unauthorized users. We recommend that you enable this feature.
	VPC - DNAT rules	Checks whether a port is open to the Internet. When you create a DNAT rule for a NAT gateway that is deployed in a virtual private cloud (VPC), we recommend that you do not open internal management ports to the Internet. Do not open all ports or an important port, for example, ports 22, 3389, 1433, or 3306, to the Internet.
	Apsara Stack Security - Back-to-origin configuration for Anti-DDoS	Checks whether Anti-DDoS is configured to allow the requests from only Web Application Firewall (WAF) back-to-origin IP addresses. After you set up Anti-DDoS or WAF, you must hide the IP addresses of the backend servers to prevent attacks on the cloud assets.
	Apsara Stack Security - WAF back-to-origin configurations	Checks whether WAF allows requests from only WAF back-to-origin IP addresses. After you set up Anti-DDoS or WAF, you must hide the IP addresses of the backend servers to prevent attacks on the cloud assets.
	SLB - IP address whitelist configurations	Checks the access control configurations of Server Load Balancer (SLB) instances. Checks whether access control is enabled for HTTP and HTTPS services and checks whether 0.0.0.0/0 is added to the IP address whitelist.
	SLB - Open ports	Checks whether SLB opens ports to the Internet for forwarding unnecessary public services.
	ApsaraDB RDS - IP address whitelist configurations	Checks whether a whitelist is configured for ApsaraDB RDS and whether the whitelist contains 0.0.0.0/0. If the whitelist contains 0.0.0.0/0, requests from all IP addresses are allowed. We recommend that you configure the whitelist to allow requests only from specific IP addresses.
	KVStore for Redis - IP address whitelist configurations	Checks whether a whitelist is configured for KVStore for Redis and whether the whitelist contains 0.0.0.0/0. If the whitelist contains 0.0.0.0/0, requests from all IP addresses are allowed. We recommend that you configure the whitelist to allow requests only from specific IP addresses.

Type	Supported item	Description
	AnalyticDB for PostgreSQL - IP address whitelist configurations	Checks whether a whitelist is configured for AnalyticDB for PostgreSQL and whether the whitelist contains 0.0.0.0/0. If the whitelist contains 0.0.0.0/0, requests from all IP addresses are allowed. We recommend that you configure the whitelist to allow requests only from specific IP addresses.
	PolarDB - IP address whitelist configurations	Checks whether a whitelist is configured for PolarDB and whether the whitelist contains `0.0.0.0/0`. If the whitelist contains 0.0.0.0/0, requests from all IP addresses are allowed. We recommend that you configure the whitelist to allow requests only from specific IP addresses.
	ApsaraDB for MongoDB - IP address whitelist configurations	Checks whether a whitelist is configured for ApsaraDB for MongoDB and whether the whitelist contains `0.0.0.0/0`. If the whitelist contains 0.0.0.0/0, requests from all IP addresses are allowed. We recommend that you configure the whitelist to allow requests only from specific IP addresses.

22.1.4.5.2. Run cloud service checks

Threat Detection Service (TDS) provides the cloud service check feature. This feature allows you to check for security risks in the configurations of your cloud services. This topic describes how to manually run cloud service checks on your cloud services. This topic also describes how to specify a detection interval for periodic automatic checks.

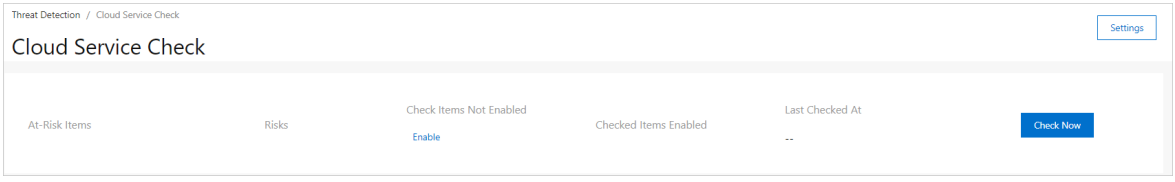
Context

Apsara Stack Security supports manual checks and periodic automatic checks to scan for security risks in the configurations of cloud services.

- Manual checks: On the **Cloud Service Check** page, you can click **Check Now** to check for security risks in the configurations of your cloud services.
- Periodic automatic checks: By default, Apsara Stack Security automatically runs checks during the time range `00:00 - 06:00` every other day. You can also customize a time range for periodic automatic checks. This way, you can detect and handle the security risks in the configurations of your cloud services at the earliest opportunity.

Manual checks

1.
2.
3. click **Cloud Service Check**.
4. On the **Cloud Service Check** page, click **Check Now** to check whether the configurations of all your cloud services contain risks and the number of affected assets.

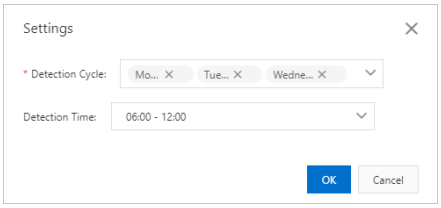


> **Note** Do not perform other operations until the check is complete.

After the check is complete, the detected risks are listed based on risk severities in descending order.

Automatic checks

1.
2.
3. click **Cloud Service Check**.
4.
5. In the **Settings** dialog box, configure the **Detection Cycle** and **Detection Time** parameters.



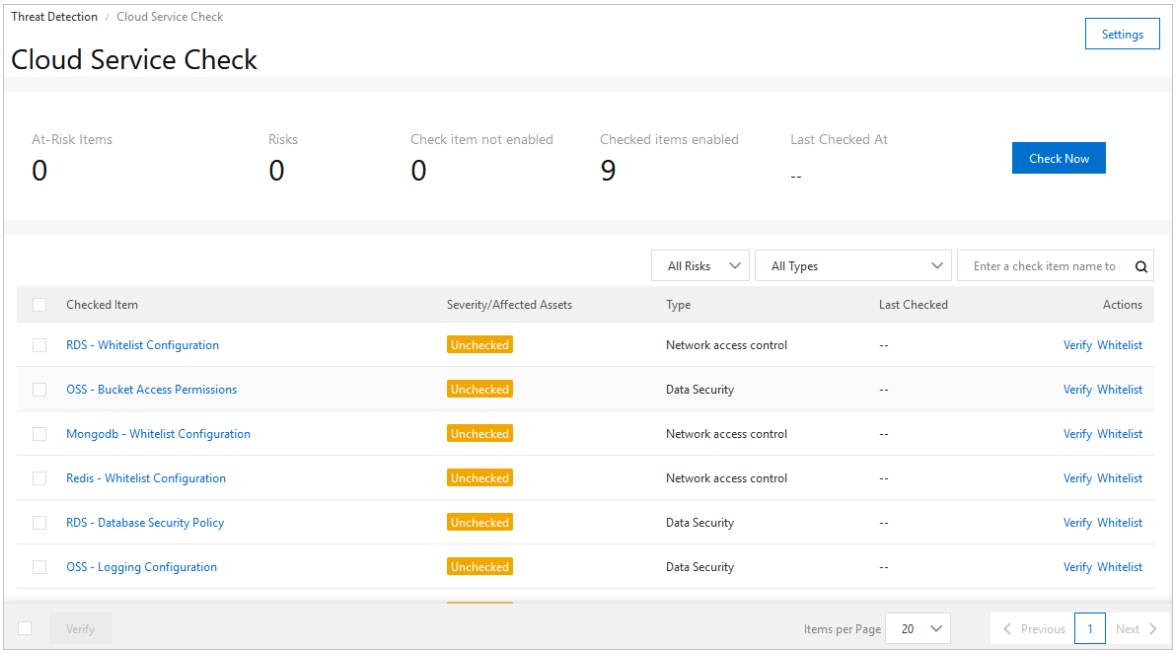
   - **Detection Cycle**: Valid values are Monday to Sunday. You can select multiple values.
   - **Detection Time**: Valid values are `24:00 - 06:00`, `06:00 - 12:00`, `12:00 - 18:00`, and `18:00 - 24:00`. You can select only one time range.
6. Click **Ok**.
   During the selected time range, Apsara Stack Security automatically runs checks based on all check items.

22.1.4.5.3. View the check results of configuration assessment for your cloud services and handle the detected risks

This topic describes how to view the check results of configuration assessment for your cloud services and handle the detected configuration risks in Apsara Stack Security. You can view the check items, details of check items, potential impacts caused by the detected configuration risks, and suggestions on how to handle the detected configuration risks. You can handle the detected configuration risks on the Cloud Service Check page in a centralized manner.

View check results

1.
2.
3. click **Cloud Service Check**.
4. On the **Cloud Service Check** page, view the details of check results.



- **View the statistics of the last check**

  You can view the total number of at-risk items and the numbers of risks at different levels in the **At-Risk Items** section, and the number of assets on which risks are detected in the **Risks** section. You can also view the number of disabled check items in the Check Items Not Enabled section, the number of enabled check items in the Checked Items Enabled section, and the time when the check was last performed in the Last Checked At section.

- **View check items**

  You can view the information about the check items in the check item list. The information includes the risk severities of check items, the number of affected assets, the types of affected assets, the types of check items, and the time when the check was last performed.

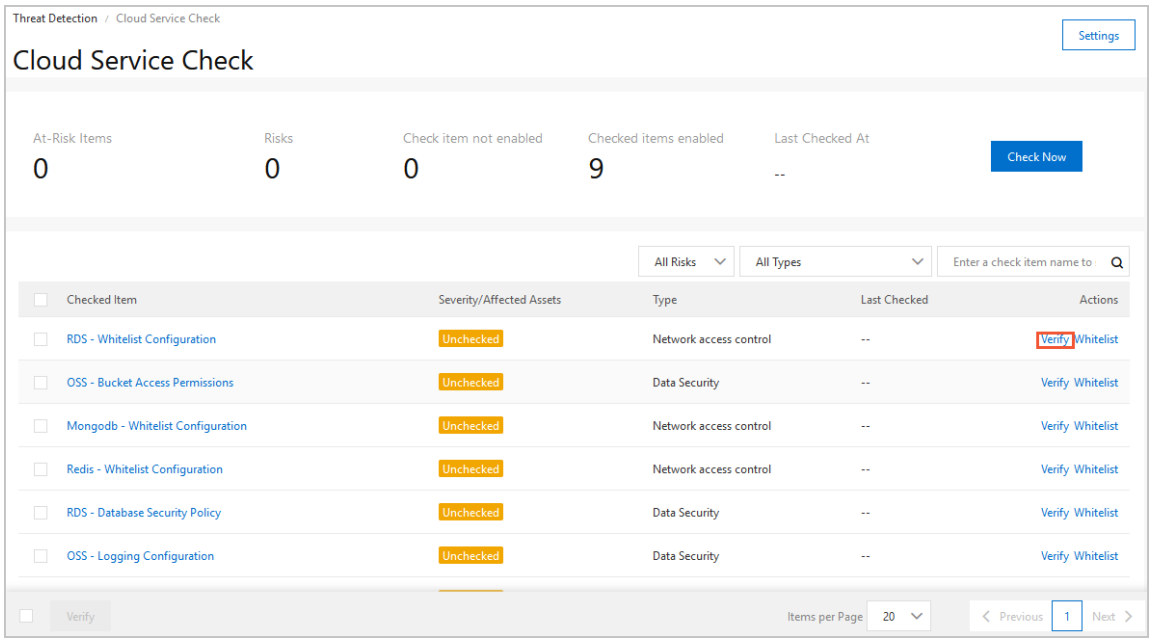
- **View the details of check results**

  You can click the name of a check item in the **Checked Item** column to go to the panel that displays the details of the check item. In the panel, you can view the description of the check item, potential impacts caused by the detected risks, and suggestions on how to handle the risks.

Handle the detected configuration risks of your cloud services

1.
2.
3. click **Cloud Service Check**.
4. On the **Cloud Service Check** page, handle the configuration risks detected on your cloud services.
   - **Verify the configurations after modification**

If you have modified the configurations for which risks are detected, find the check item in the check item list and click **Verify** in the Actions column to check whether the new configurations are at risk.



- **Add check items to a whitelist**

  If you trust a check item for which risks are detected, find the check item in the check item list and click **Whitelist** in the Actions column to add the check item to a whitelist. Then, the state of the check item is displayed as **Ignored** in the Severity/Affected Assets column. **Ignored** check items are not counted in the total number of at-risk items in the **At-Risk Items** section.

  In the check item list, you can click **Remove** to remove the ignored check items from the whitelist.

  > Note After you add a check item to the whitelist, the risk that is detected for the check item is ignored only for this time. If the risk is detected again, Apsara Stack Security still displays the check result of this check item.

22.1.4.6. Assets

22.1.4.6.1. View the security status of a server

The Assets page displays security information about each protected server. The information includes the virtual private could (VPC) where each server resides, server status, and risk status. This topic describes how to search for specific servers and view the security status of these servers. This topic also describes how to specify search conditions and select the items that you want to display on the Assets page.

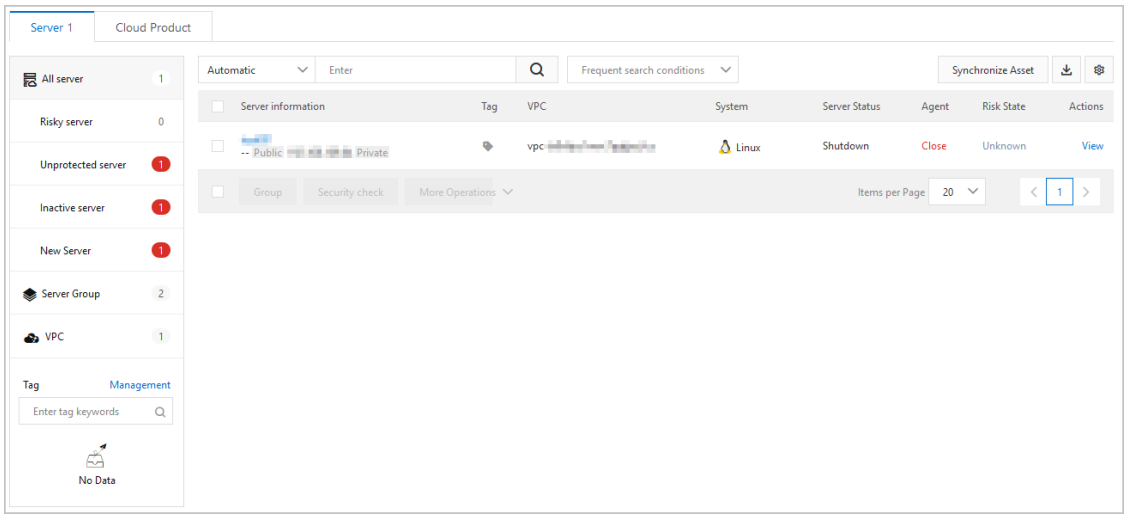
Procedure

1.
2.
3. click **Assets**.
4. On the **Server(s)** tab of the Assets page, view the security status of each server.

   You can perform the following operations:

   - **Filter servers by status**

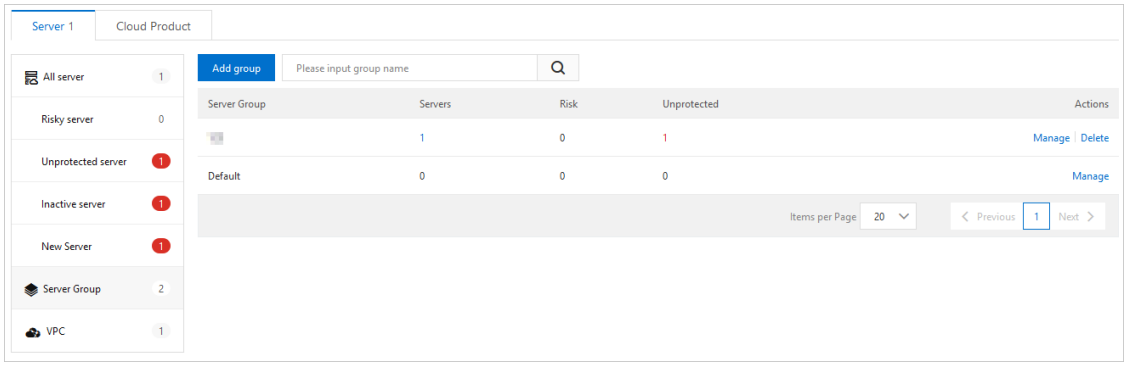
- In **All Servers**, you can view the numbers of all servers, risky servers, unprotected servers, new servers, and servers that are shut down.



  To view the security information about a server, you must click the name of the server or click **Fix** in the **Actions** column. For more information, see View the details of a single asset.

- You can click **Risk Servers**, **Unprotected Servers**, **Shutdown Server(s)**, or **New Server(s)** to view security information about specific servers.

- **Filter servers by group**
  - You can click **Server Group** to view the numbers of all servers, servers that are at risk, and unprotected servers in each server group. You can also view the total number of server groups.

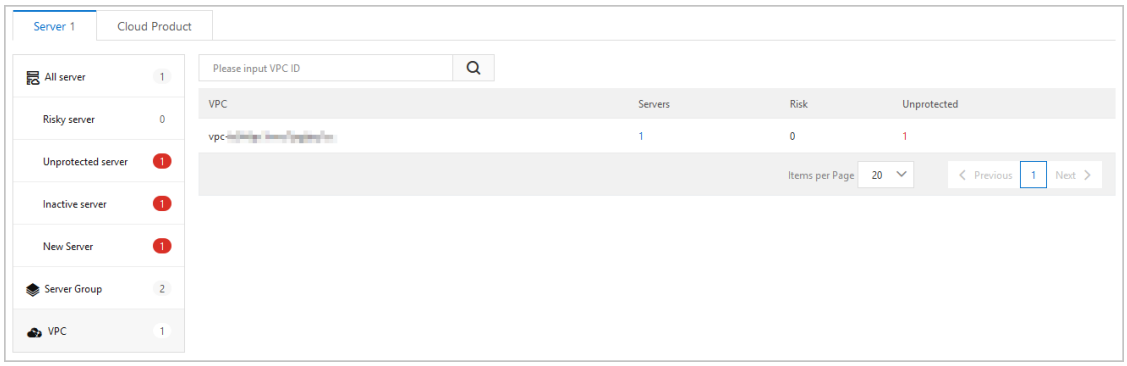


  To manage server groups, you can click **Manage** or **Delete** in the Actions column. For more information, see Manage asset groups.

  - You can find a server group and click the number in the **Servers**, **Risk**, or **Unprotected** column to view the security information about specific servers in this group.

- **Filter servers by VPC ID**

- You can click **VPC** to view the numbers of all servers, servers that are at risk, and unprotected servers in each VPC. You can also view the total number of VPCs.



- You can find a VPC and click the number in the **Servers**, **Risk**, or **Unprotected** column to view the security information about specific servers in this VPC.

- **Filter servers by tag**

  In the navigation tree, you can click a tag to view the security information about servers to which the tag is added.

- **Filter servers by condition**

  If you click **All Servers**, **Server Group**, **VPC**, or a tag in the navigation tree, you can specify filter conditions above the right-side list to search for specific servers.

  - Use one filter condition to search for specific servers:

    You can select a filter condition and select or enter keywords to search for specific servers. The filter conditions include **Internet IP**, **Private IP**, **Instance name**, **System**, **Baseline problems**, **Vul problems**, **Alert problems**, **Risk Status**, **Online or Offline**, **Tag**, **Group name**, **OS**, and **Is there a snapshot risk**.

    > **Note**
    >
    > You can specify multiple filter conditions at a time and specify a Boolean operator for the conditions. The following list describes the Boolean operators:
    >
    > - Boolean operators:
    >   - **AND**: specifies the **AND** logical relation for the conditions.
    >   - **OR**: specifies the **OR** logical relation for the conditions.
    > - If you want to search for servers that meet at least one of the filter conditions, you must set the Boolean operator to **OR**.
    > - If a filter condition requires you to enter a keyword, you must enter the keyword and click the **Search** icon. Results are displayed only after you click the Search icon.

  - Use multiple filter conditions to search for specific servers:

    If you select multiple filter conditions, they are all applied to search for specific servers.

    You can also click **Server Group**, **VPC**, or a tag, and use the search box above the asset list to search for specific servers.

- **Save frequently used filter conditions**

  You can save the filter conditions that are applied as frequently used search conditions. To save the conditions, click **Save** above the right-side list, and enter a name in the **Save condition** dialog box. Then, you can select the saved conditions from the **Frequently used search conditions** drop-down list on the right of the Search icon.

- **Customize displayed items**

On the **Assets** page, you can click the icon in the upper-right corner. Then, you can select the items that you want to display on the Assets page.

22.1.4.6.2. View the security status of cloud services

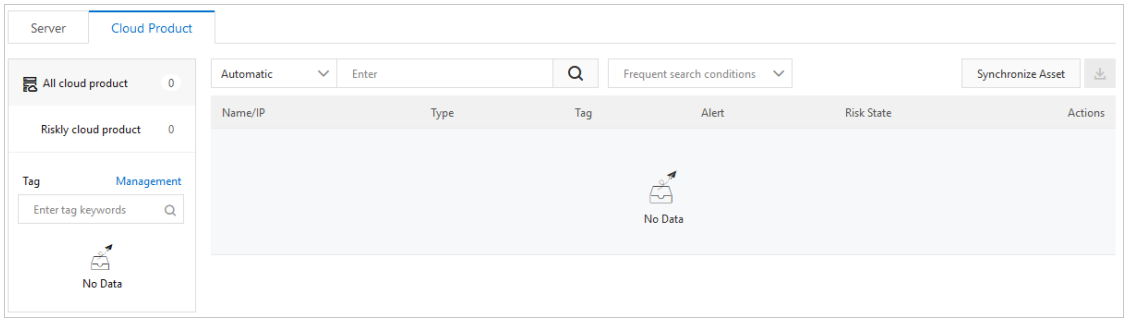
The Assets page displays the security information about each protected cloud service. The information includes the at-risk services and the types of services such as Server Load Balancer (SLB) and NAT Gateway. This topic describes how to configure search conditions to view the security status of cloud services.

Procedure

1.
2.
3. click **Assets**.
4. On the **Cloud Service** tab of the Assets page, view the security status of cloud services.

   You can perform the following operations based on your business requirements:

   - **Search by asset status**
     - In **All Cloud Services**, you can view the numbers of **all cloud services** and **risky cloud services**. You can also view the security status of all cloud services.

     - You can click **Risk Cloud Services** to view the cloud services that are at risk.

     You can click the name of the required cloud service or click **View** in the Actions column that corresponds to a service to view detailed information. For more information, see View the details of a single asset.

   - **Search by asset type**

     Cloud services are classified into two asset types:

     - **SLB**
     - **NAT**

     In the navigation tree of the Cloud Service tab, you can view the number of cloud services of each type. You can click **SLB** or **NAT** to view the security status of the required cloud service.

   - **Search by tag**

     In the **Tag** section of the navigation tree, you can view the number of cloud services to which each tag is added. You can click a tag to view the security status of cloud services to which the tag is added.

   - **Filter by search condition**

     You can click **All Cloud Services**, **SLB**, or **NAT** in the navigation tree and configure search conditions in the search box above the **asset** list to search for specific assets.

     For example, you can click **All Cloud Services** and configure search conditions to search for specific assets.

- Use multiple subconditions to search for specific assets:

  Select a condition from the drop-down list of the search box above the **asset** list, and select a subcondition or enter a keyword into the search box to search for specific assets. Supported search conditions are **Internet IP**, **Instance name**, **Alert problems**, **Risk Status**, **Tag**, and **Group name**.

- Use multiple filter conditions to search for specific assets:

  Apply multiple search conditions.

  - You can click **SLB**, **NAT**, or a tag in the **Tag** section and configure conditions in the search box above the **asset** list to search for specific assets.
  - You can also click **All Cloud Services**, **SLB**, or **NAT**, and select a tag in the **Tag** section to search for specific assets.

- **Save frequently used search conditions**

  You can save the filter conditions that are applied as frequently used search conditions. Click **Save** below the search box and enter a name in the **Save condition** dialog box. Then, you can select the saved search condition from the Frequently used search conditions drop-down list on the right of the search box.

22.1.4.6.3. View the details of a single asset

The Assets page provides details about all assets. These details include basic information, alert management status, baseline check analysis, and asset fingerprints. This topic describes how to view the details of a server or a cloud service.

Context

The **Assets** page provides basic information about all assets. Different types of assets, such as servers and cloud services, are managed in different ways.

The following table lists the features that are supported for servers and cloud services on the **Assets** page. The following list describes the marks that are used to indicate whether a feature is supported for servers or cloud services:

- Cross (×): not supported.
- Tick (√): supported.

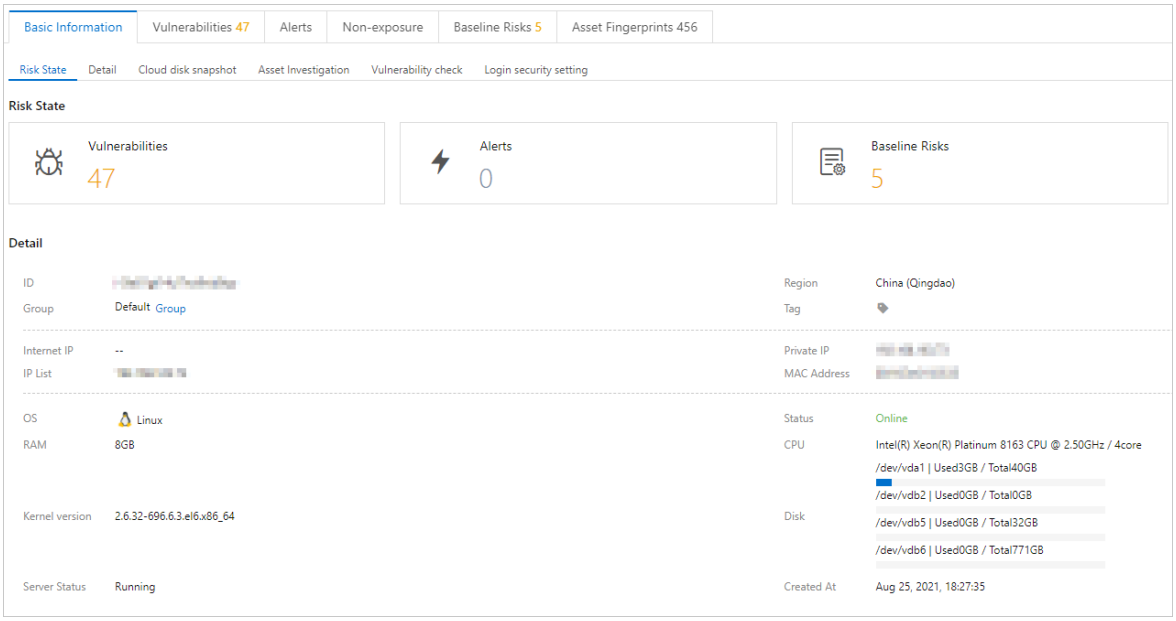
Feature	Description	Server	Cloud service
Basic Information	Risk State: displays the number of risks detected on an asset. The following types of risks can be detected: • **Vulnerabilities** • **Alerts** • **Baseline Risks**	√	√ (Only alerts can be processed.)
	Detail: displays the configuration and protection status of an asset. You can specify a group and a tag for the asset.	√	√ (Asset grouping is not supported.)
	Asset Investigation: displays asset fingerprints, including ports, software, processes, and accounts.	√	X
	Vulnerability check: displays the types of vulnerabilities that can be detected. You can specify the types of vulnerabilities that you want to detect for an asset.	√	X

Feature	Description	Server	Cloud service
	Login security setting: displays the approved logon locations, IP addresses, time ranges, and accounts that are added. You can manage relevant alerts for an asset.	√	X
Vulnerabilities	Displays the results of vulnerability detection on an asset.	√	X
Alerts	Displays the alerts that are generated for an asset.	√	√
Baseline Risks	Displays the results of a baseline check on an asset.	√	X
Asset Fingerprints	Displays the details of asset fingerprints for an asset.	√	X

Procedure

1.
2.
3. click **Assets**.
4. On the **Assets** page, click the **Server(s)** or **Cloud Service** tab.
5. On the **Server(s)** or **Cloud Service** tab, find the required asset and click its name.
6. View the details of the asset.

   On the asset details page, click the **Basic Information**, **Vulnerabilities**, **Alerts**, **Non-exposure**, **Baseline Risks**, or **Asset Fingerprints** tab to view relevant details.

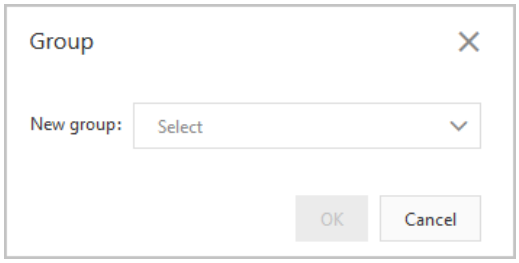


The following list describes the details of the asset:

- **Basic Information**: This tab consists of sections in which you can view asset details and manage the asset.
  - **Risk State**: This section displays the numbers of vulnerabilities, alerts, and baseline risks on the asset. You can click the number under Vulnerabilities, Alerts, or Baseline Risks to view the details.

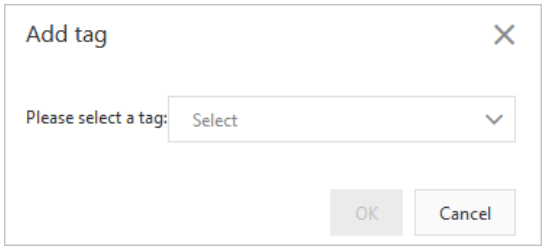
- **Detail**: This section displays information about the asset configuration and security protection settings, and allows you to manage asset tags and groups.
  - **Change asset groups**

    Click **Group**. In the **Group** dialog box, select a new group and click **OK**.

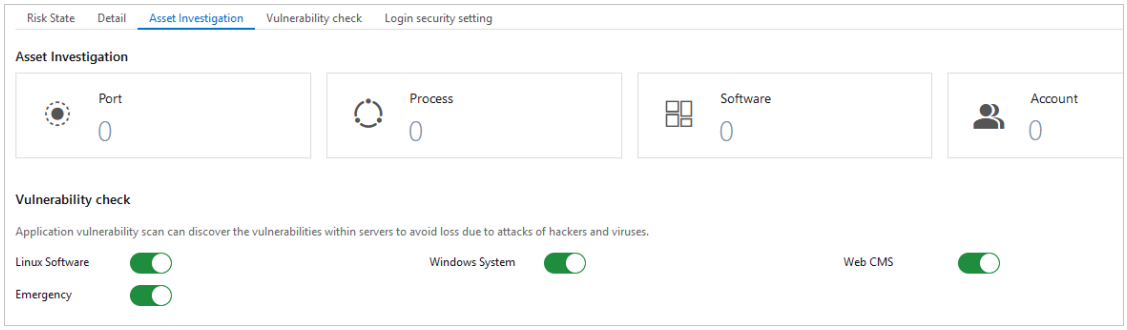
  - **Modify tags**

    Click the icon. In the **Add tag** dialog box, select a tag and click **OK**.

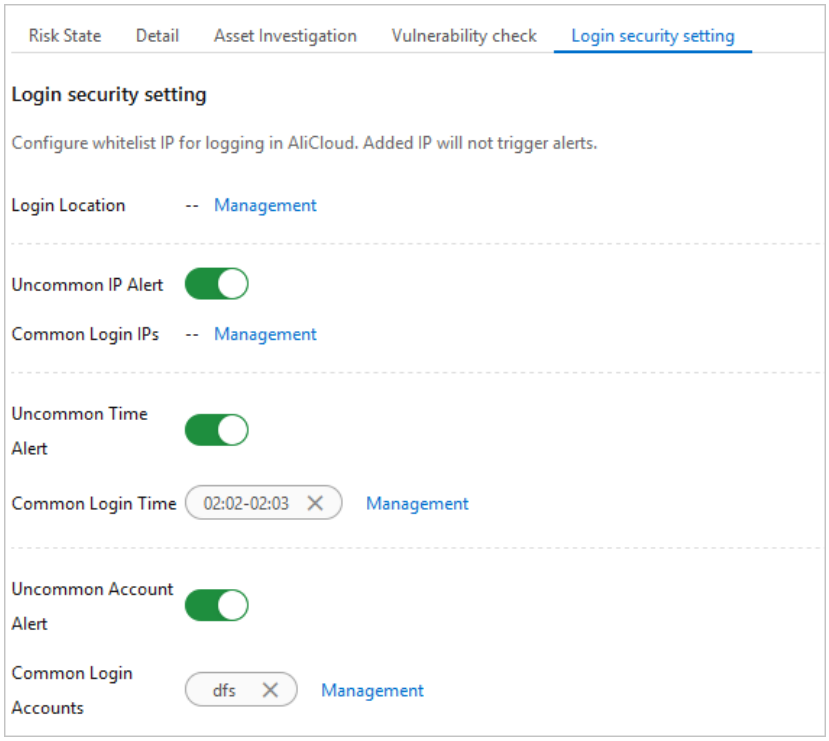
    You can click the icon on the right of a tag to delete the tag.

- **Asset Investigation**: This section displays the fingerprints of an asset. You can click the number under an item to go to the **Asset Fingerprints** tab to view the details.

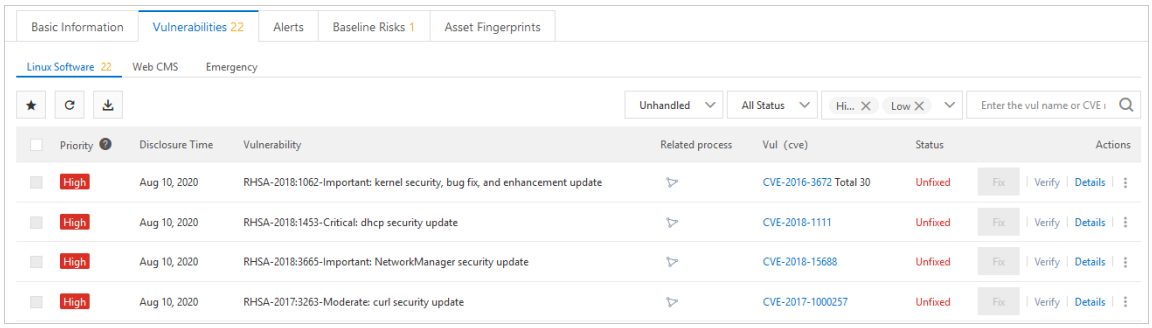
  

- **Vulnerability check**: This section displays vulnerability check items that are enabled or disabled for an asset. You can enable or disable different types of vulnerability checks for the asset. The vulnerabilities include Linux software vulnerabilities, Windows system vulnerabilities, Web CMS vulnerabilities, and urgent vulnerabilities.

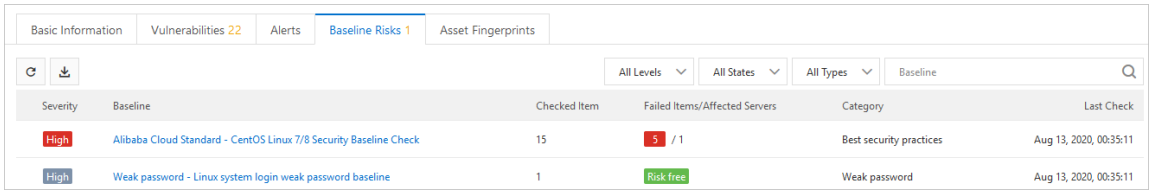
- **Login security setting**: This section allows you to specify approved logon locations, configure advanced logon settings, and turn on or turn off alerting for unapproved IP addresses, time, and accounts. The advanced logon settings include approved IP addresses, time ranges, and accounts. You can also specify approved IP addresses, time ranges, and accounts for a specific asset.



- **Vulnerabilities**: This tab displays vulnerabilities detected on an asset.



- **Alerts**: This tab displays alerts generated for an asset.
- **Baseline Risks**: This tab displays baseline risks of an asset.

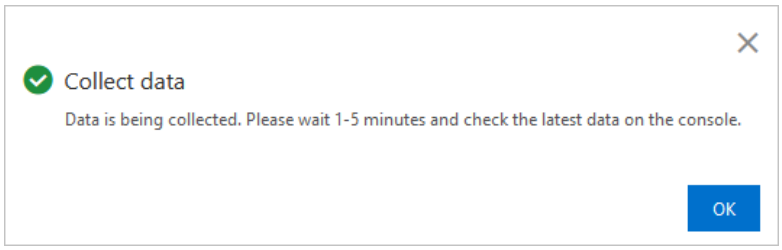


- **Asset Fingerprints**: This tab displays the fingerprints, including ports, processes, software, and accounts of an asset.

  You can manually collect the latest fingerprints of an asset.

  a. You can click the **Port**, **Software**, **Process**, **Account**, or **Scheduled Tasks** tab. In the upper-right corner, click **Collect data now**.

b. In the **Collect data** message, click **OK**.



After the data collection task is submitted, it takes one to five minutes to collect the fingerprints of the required asset. After the data collection task is complete, you can view the latest fingerprints of the asset.

22.1.4.6.4. Enable and disable server protection

This topic describes how to enable and disable server protection.

Procedure

1.
2.
3. click **Assets**.
4. On the **Server(s)** tab of the page that appears, enable or disable server protection for specified servers.
   - **Enable server protection**

     Select one or more servers where the agent is in the **Close** state, and choose **More Operations > Turn on protection**.

     After server protection is enabled, the status in the Agent column changes to **Enable**.
   - **Disable server protection**

     If you confirm that a server does not require protection from Apsara Stack Security, you can disable protection for the server. Select one or more servers where the agent is in the **Enable** state, and choose **More Operations > Suspend Protection**.

     > **Note** After server protection is disabled, Apsara Stack Security stops protecting your servers. For example, Apsara Stack Security no longer detects vulnerabilities or generates alerts for detected risks. We recommend that you proceed with caution.

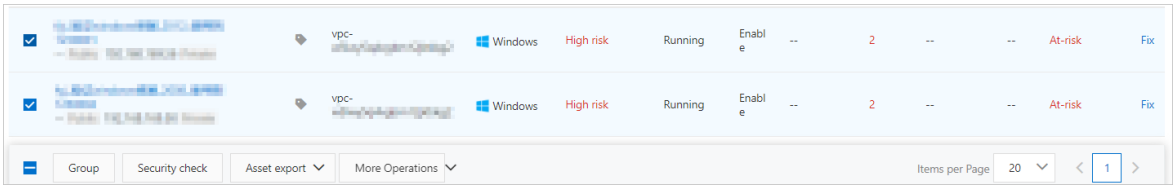
     After server protection is disabled, the status of the agent on your servers changes to **Close**.

22.1.4.6.5. Perform a quick security check

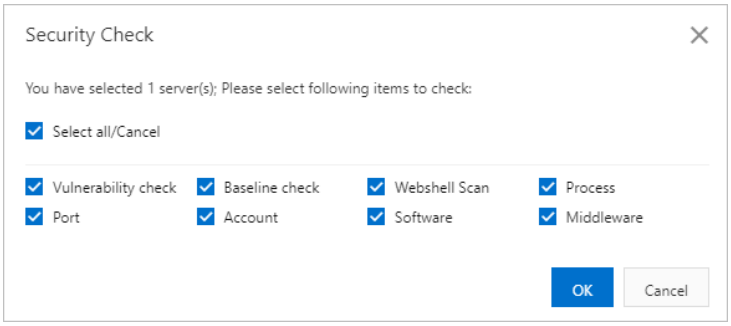
The Server(s) tab of the Assets page allows you to run security checks. You can dispatch security check tasks to scan for vulnerabilities, baseline risks, or webshells, and collect asset fingerprints on a specific server. The asset fingerprints are ports, software, processes, and accounts. This topic describes how to perform a security check on servers.

Procedure

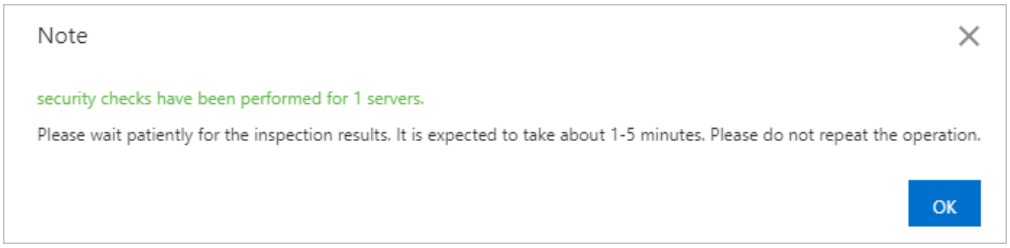
1.
2.
3. click **Assets**.
4. On the **Server(s)** tab, select one or more servers on which you want to perform a security check.
5. In the lower part of the page, click **Security check**.



6. In the **Security Check** dialog box, select check items.



7. Click **OK** to start the check.
8. In the message that appears, click **OK**.



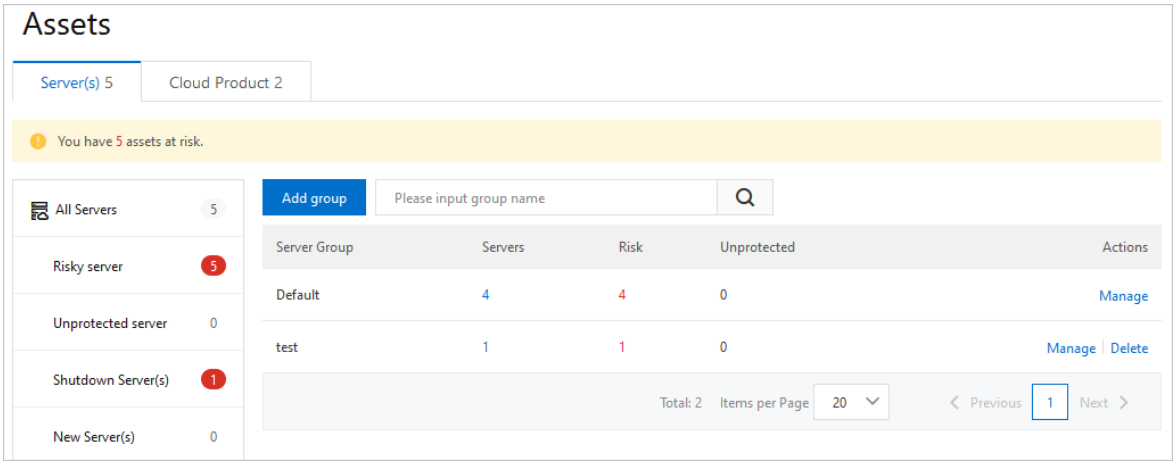
After the security check is complete, the check results are automatically displayed on the details pages of the selected servers.

22.1.4.6.6. Manage server groups

This topic describes how to create, modify, delete, and replace server groups.

Create a server group

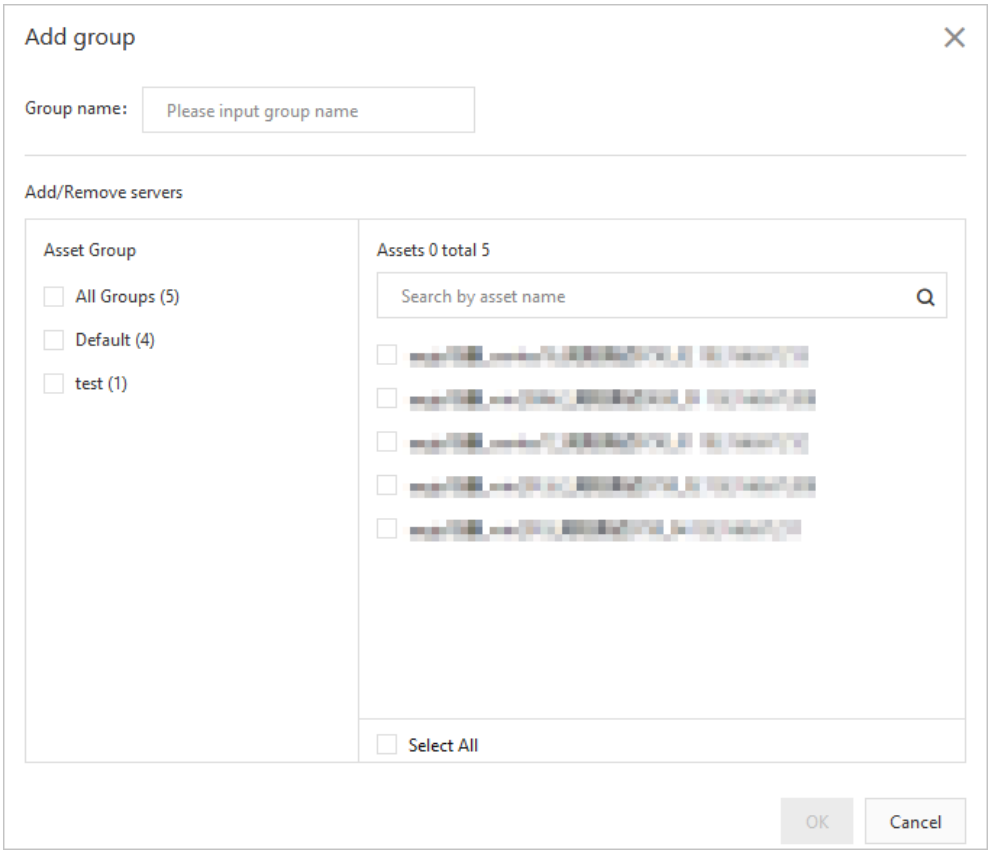
1.
2.
3. click **Assets**.
4. On the **Server(s)** tab of the page that appears, click **Server Group** in the navigation tree.



Note By default, the assets that are not grouped are in the **Default** group.

5. Click **Add Group**.
6. In the **Add Group** dialog box, configure parameters for the new group.

   To configure the parameters, perform the following steps:



   i. Enter a name for the new group in the **Group name** field.

   ii. Add servers to the new group.

   You can add servers in the Default group to the new group. You can also move servers from another group to the new group. To add or move servers, select **Default** or other groups in the **Asset Group** section, and select or clear the check boxes that correspond to the required servers in the asset list in the right area of the section.

7. Click **OK**.

In the server group list, you can view the new group.

Modify or delete a server group

The following procedure describes how to modify or delete a server group. When you modify a server group, you can rename the group or adjust the servers in the group.

1.
2.
3. click **Assets**.
4. On the **Server(s)** tab of the page that appears, click **Server Group** in the navigation tree.
5. Find the server group that you want to modify or delete. In the Actions column, click **Manage** or **Delete**.

   You can perform the following operations based on your business requirements:

   - **Modify the group**
     a. In the Actions column, click **Manage**. The Group dialog box appears.
     b. In the **Group** dialog box, select the group in the **Asset Group** section.
     c. In the right area of the section, clear the check boxes that correspond to the required servers in the asset list.
     d. Click **OK**. The server group is modified.
   - **Delete the group**

     In the Actions column, click **Delete**. In the message that appears, click **OK**.

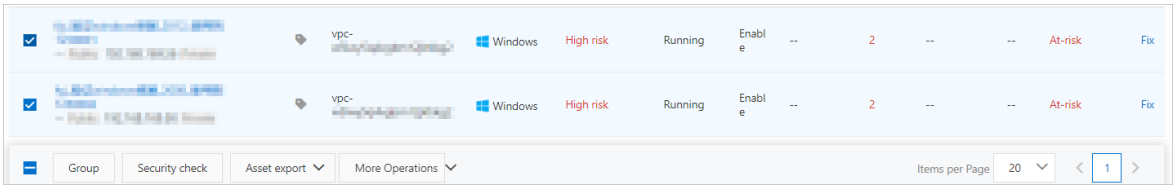
     > **Note** After you delete a group, servers in this group are moved to the **Default** group.

Replace a server group

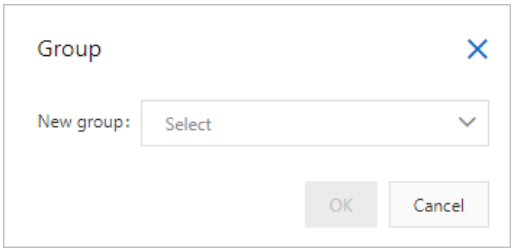
You can add servers to a server group to manage multiple servers at a time. We recommend that you add the same types of servers to a server group. For example, if you configure a baseline check template, you can specify a server group and apply the template to all servers in the group. You can also filter and view servers based on server groups.

To add servers to a specific server group, perform the following steps:

1.
2.
3. click **Assets**.
4. On the Server(s) tab of the page that appears, select one or more servers and click **Group** in the lower part of the page.



5. In the **Group** dialog box, select a new server group.



6. Click **OK**.

22.1.4.6.7. Manage asset tags

This topic introduces asset importance tags and describes how to create, modify, and delete custom tags.

Context

Apsara Stack Security provides the asset importance tags described in the following table to classify assets. You can select appropriate importance tags for your assets.

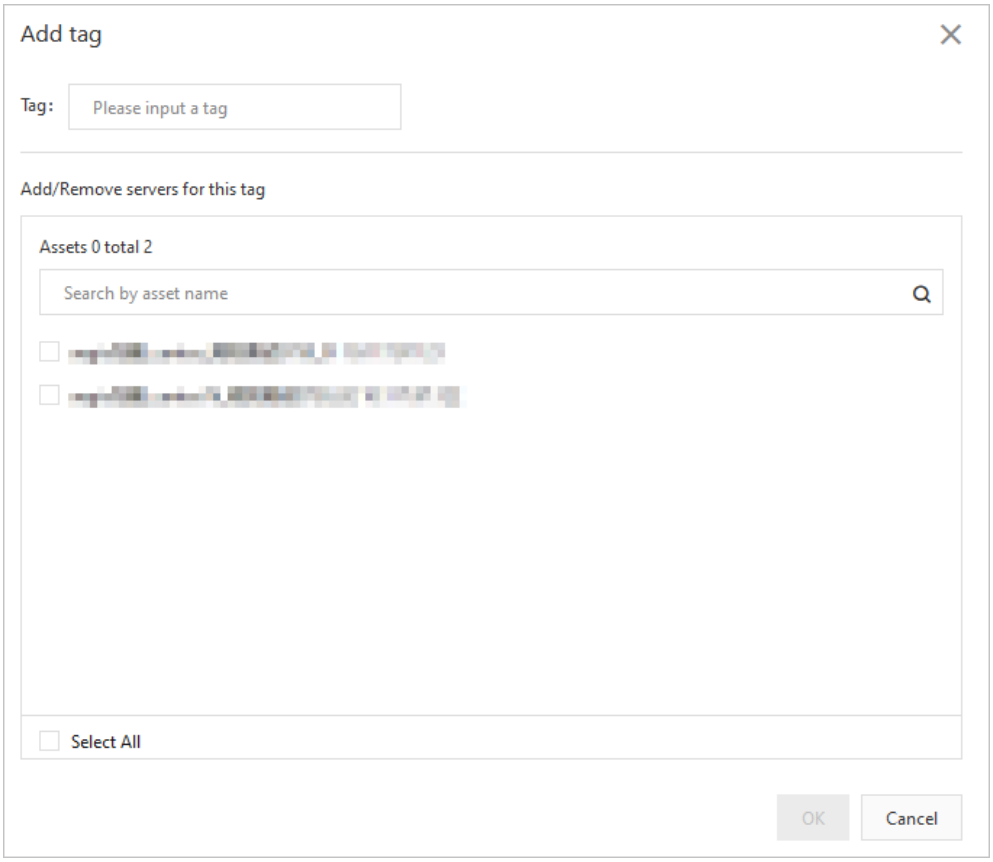
An asset importance tag is transformed to an **asset importance score**. An **asset importance score** is used to calculate a vulnerability priority score. You can determine whether to preferentially fix a vulnerability based on the vulnerability priority score. We recommend that you add importance asset tags to core assets. Apsara Stack Security prompts you to fix vulnerabilities based on the importance of each asset. The following table describes the relationships between asset importance tags and asset importance scores.

Asset importance tag	Asset importance score	Recommendation
Important Assets	1.5	Assets that are related to crucial business or store core business data. Virus intrusion into the assets adversely affects the system and causes major loss.
General Assets	1	Assets that are related to non-crucial business and are highly replaceable. Virus intrusion into the assets causes less impact on the system.
Test Assets	0.5	Assets for functional or performance tests, or assets that can cause less impact on the system.

Note If you do not add asset importance tags, the **General Assets** tag is automatically added to each asset. This tag indicates that the asset importance score is 1.

Create a custom tag

1.
2.
3. click **Assets**.
4. On the **Assets** page, click the **Server(s)** or **Cloud Service** tab.
5. In the navigation tree of the **Server(s)** or **Cloud Service** tab, click **Management** to the right of **Tag**.
6. In the **Add tag** dialog box, enter the tag name in the **Tag** field.



7. In the **Asset Group** section, select a server group. Then, select the required servers to add the new tag to the selected servers in the right area of the section.
8. Click **OK**.

   In the asset list of the Server(s) or Cloud Service tab, you can click the icon in the **Tag** column to add the new tag to an asset.

   > **Note** You can add multiple tags to one asset. All tags of an asset are displayed in the **Tag** column.

Modify or delete a custom tag

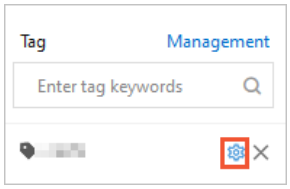
The following procedure describes how to modify or delete a custom tag. When you modify a tag, you can rename the tag or adjust the servers to which the tag is added.

1.
2.
3. click **Assets**.
4. On the **Assets** page, click the **Server(s)** or **Cloud Service** tab.
5. On the **Server(s)** or **Cloud Service** tab, modify or delete a tag.

   Perform the following operations to modify or delete a tag:

   - **Modify a tag**

a. Find the tag that you want to modify and move the pointer over the icon to the right of the tag. Then, click the icon.

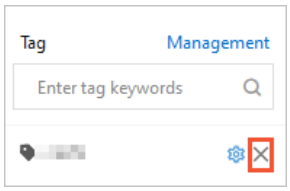


b. In the **Tag** dialog box, enter a new name in the **Tag** field, add the tag to more servers, or remove the tag from specific servers.

c. Click **OK**.

- **Delete a tag**

  Find the tag that you want to delete and move the pointer over the icon to the right of the tag. Then, click the icon. In the message that appears, click **OK**.



22.1.4.7. Application whitelist

The application whitelist feature prevents unauthorized programs from running on your servers and provides a trusted running environment for your servers.

Context

The application whitelist feature allows you to apply application whitelist policies to your servers that require special protection. Apsara Stack Security identifies trusted, suspicious, and malicious programs based on the policies. Then, you can add the identified programs to an application whitelist based on your business requirements. This prevents unauthorized programs from running. This feature protects your servers from untrusted and malicious programs and improves resource usage.

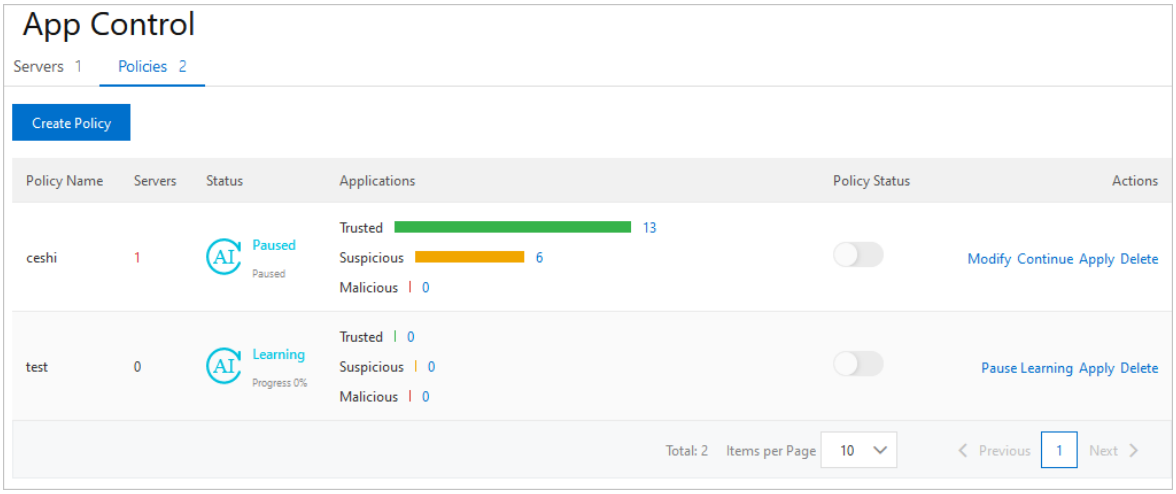
After you create an application whitelist policy, you can apply it to a server that requires special protection. Then, Apsara Stack Security scans for suspicious or malicious programs on the server and generates alerts for the programs that are not in the application whitelist.

> **Note** If a program that is not in the application whitelist starts, an alert is generated. The program may be a normal program that is newly started or a malicious program that is inserted into your compromised server. If the program is a normal program, a frequently used program, or a third-party program installed by you, we recommend that you add the program to the application whitelist. After you add the program to the application whitelist, Apsara Stack Security no longer generates alerts for this program the next time the program starts. If the program is malicious, we recommend that you immediately delete this program and check whether the configuration files such as cron tasks are tampered with.

Step 1: Create an application whitelist policy

1.
2.
3. click **Application Whitelists**.
4.

5. On the **Policies** page, click **Create Policy**.
6. In the Create Policy step of the **Create Whitelist Policy** panel, configure the following parameters:
   - **Policy Name**: the name of the application whitelist policy.
   - **Intelligent Learning Duration**: the duration for Apsara Stack Security to perform intelligent learning. Valid values: 1 Day, 3 Days, 7 Days, and 15 Days. The intelligent learning feature uses machine learning to automatically collect and categorize large amounts of alert data. Apsara Stack Security can identify suspicious or malicious programs based on the collected data.
   - **Servers for Intelligent Learning**: the servers to which you want to apply the application whitelist policy.
7. Click **Next**. The application whitelist policy is created.
   After the application whitelist policy is created, the policy details are displayed in the policy list on the Policies tab.



The following table describes the parameters in the list of application whitelist policies.

Parameter	Description
Policy Name	The name of the application whitelist policy.
Servers	The number of servers to which the application whitelist policy is applied.
Status	The status of the policy. Valid values: ○ **Applied**: Intelligent learning is complete. The policy has been applied to servers. ○ **Pending Confirmation**: Intelligent learning is complete. The policy needs to be confirmed and enabled. After intelligent learning is complete, you must turn on the switch in the **Policy Status** column to enable this policy. The policy takes effect only after it is enabled. Apsara Stack Security automatically identifies the programs on your servers as trusted, suspicious, or malicious programs. ○ **Paused**: Intelligent learning is manually paused. You can click **Continue** in the Actions column to resume intelligent learning. ○ **Learning**: Intelligent learning is in progress. After an application whitelist policy is created, Apsara Stack Security automatically performs intelligent learning based on the policy. The status of a new application whitelist policy is **Learning**.
Applications	The number of programs of each type on all servers to which the policy is applied. The program types include **trusted**, **suspicious**, and **malicious**.

Parameter	Description
Actions	The operations that you can perform on a policy. You can perform the following operations: ○ **Apply**: Add or remove servers to which the policy is applied in the **Apply Whitelist Policy** panel. ○ **Modify**: Modify the policy in the **Modify Whitelist Policy** panel. You can change the values of **Policy Name** and **Intelligent Learning Duration**, and add or remove the servers on which intelligent learning is automatically performed. ○ **Pause Learning**: Pause intelligent learning. ○ **Continue**: Resume intelligent learning. After you click **Continue**, the status of the policy changes to **Learning**. You can view the learning progress of the policy in the **Status** column. ○ **Delete**: Delete the policy. After the policy is deleted, the servers to which the policy is applied are no longer protected by the policy.

Step 2: Apply the created application whitelist policy to servers

1.
2.
3. click **Application Whitelists**.
4. On the **Application Whitelist** page, click **Servers**
5. On the **Servers** tab, click **Add Server**.
6. In the **Add Server** panel, configure the parameters.

   In the **Add Server** panel, configure the following parameters:

   - **Whitelist Policy**: Select the created application whitelist policy from the drop-down list.
   - **Event Handling**: The default value is **Alert**, which indicates that Apsara Stack Security generates an alert when a suspicious program is detected.

     If a program that is not in the application whitelist starts, Apsara Stack Security automatically generates an alert. You can click the number in the **Suspicious Events** column to go to the **Alerts** tab of the server details page and view the alert details.
   - **Servers**: Select the server to which you want to apply the application whitelist policy. You can select multiple servers.

     To search for a server, enter the server name in the **Servers** search box and click the search icon. Fuzzy match is supported.
7. Click **OK**. The application whitelist policy is applied to the selected servers.
   After the application whitelist is created, you can view the protected servers and the name of the application whitelist policy in the server list on the **Servers** tab.

   The **Servers** tab displays the following information of a protected server:

   - **Server Name/IP**: the name and IP address of the server to which the application whitelist policy is applied.
   - **Whitelist Policy**: the name of the application whitelist policy that is applied to the server.
   - **Suspicious Events**: the number of programs that are not in the application whitelist and have started. If a suspicious program starts on the server, Apsara Stack Security detects the program and generates an alert.
   - **Event Handling**: The default value is **Alert**, which indicates that Apsara Stack Security generates an alert when a suspicious program is detected.

If a program that is not in the application whitelist starts, Apsara Stack Security automatically generates an alert. You can click the number in the **Suspicious Events** column to go to the **Alerts** tab of the server details page and view the alert details.

- **Actions**: After you click **Delete** in the **Actions** column, the server is removed from the application whitelist policy.

  After you click Delete in the Actions column, the application whitelist policy becomes invalid for the server. In this case, if a program that is added to the application whitelist starts on this server, Apsara Stack Security generates an alert.

Add a program to or remove a program from an application whitelist

After you configure an application whitelist policy for your server, you can view the detailed information in the server list on the **Servers** tab. The information includes the details of the protected server and the name of the application whitelist policy that is applied to the server. You can click a policy name in the **Whitelist Policy** column to view the programs running on the server. You can also view the numbers of trusted, suspicious, and malicious programs and their detailed information.

The following information about each program on the server is displayed:

- **Type**: the type of the program. Programs are classified into trusted, suspicious, and malicious programs.
- **Process Name**: the name of the program.
- **Hash**: the hash function of the program. A hash function is used to identify whether a program is unique. This helps protect servers against malicious programs.
- **Path**: the file path of the program on the server.
- **Degree of Trustability**: the degree of trustability for the program. The value of this parameter is determined by Apsara Stack Security. Valid values: 0%, 60%, and 100%. The value 0% indicates malicious programs, 60% indicates suspicious programs, and 100% indicates trusted programs.

  > **Note** We recommend that you handle the program whose Degree of Trustability is 0% at the earliest opportunity.

- **Actions**: the operations that can be performed on the program. You can determine whether to add the program to the whitelist based on the services deployed on your server. You can perform the following operations:
  - **Add to Whitelist**: If you trust the program, add it to the whitelist.
  - **Remove from Whitelist**: After you remove the program from the whitelist, Apsara Stack Security identifies the program as untrusted. If this program starts, Apsara Stack Security generates an alert.

22.1.4.8. Vulnerability scan

22.1.4.8.1. Quick start

This topic describes how to get started with the vulnerability scan feature.

The following procedure shows how to use the vulnerability scan feature:

1. Configure the following detection items and the required cycles based on your environment requirements:
   - Overall Monitoring: Configure detection features and the monitoring cycle of each detection feature. For more information, see Configure overall monitoring.
   - Basic Monitoring: Configure Weak Password Vulnerability Monitoring, Operation Security Vulnerability Monitoring, CMS Application Vulnerability Monitoring, and Baseline Monitoring. For more information, see Configure basic monitoring.
   - Web Monitoring: Configure the monitoring cycle and the types of web vulnerabilities that you want to monitor. For more information, see Configure web monitoring.

- Whitelist: Add the assets that do not require detection to the whitelist. For more information, see Configure a whitelist.

2. Import assets that require vulnerability scans.
   - Import internal assets: Configure a scan engine to import your internal assets in virtual private clouds (VPCs). For more information, see Configure a scan engine for internal assets.
   - Import Internet assets: Import your Internet assets. For more information, see Import assets.

   > **Note** The number of imported assets cannot exceed the specified upper limit.

3. View and confirm the results of vulnerability scans.
   - View the overall information to obtain the results of vulnerability scans. For more information, see View the information on the Overview page.
   - View and confirm vulnerability risks. For more information, see Manage security vulnerabilities.
   - View and confirm host compliance risks. For more information, see Manage host compliance risks.
4. Generate vulnerability scan reports.

   Generate reports to audit the vulnerabilities and baseline risks on assets on a regular basis. For more information, see Create a report.

22.1.4.8.2. View the information on the Overview page

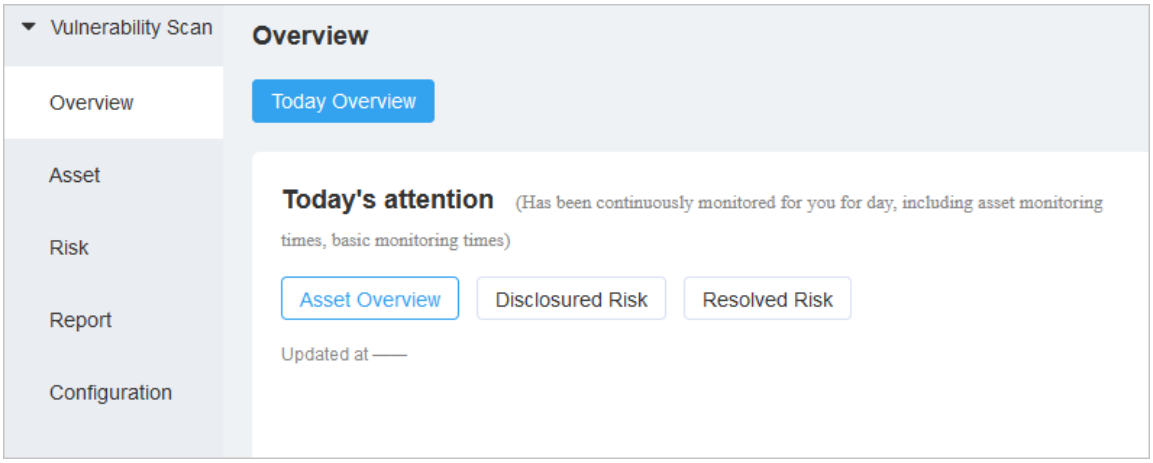
This topic describes the overall results of vulnerability scans. Security administrators can understand the vulnerability situation based on the overall results.

Context

The vulnerability scan feature can identify the following vulnerabilities: web security vulnerabilities, content management system (CMS) application vulnerabilities, weak password vulnerabilities, O&M security vulnerabilities, and baseline security vulnerabilities.

Procedure

1.
2.
3. click **Overview**.
4. View the overall results of vulnerability scans.



Section	Description
Today's attention	View Asset Overview, Disclosured Risk, and Resolved Risk of the current day. ◦ **Asset Overview**: displays the numbers of hosts, websites, and domain names for the current day and provides a security score for the current assets. The radar chart on the right shows the distribution of web security vulnerabilities, CMS application vulnerabilities, weak password vulnerabilities, O&M security vulnerabilities, and baseline security vulnerabilities. ◦ **Disclosured Risk**: displays the numbers of high-risk vulnerabilities, medium-risk vulnerabilities, and low-risk vulnerabilities, and the total number of these vulnerabilities for the current day. These vulnerabilities are not fixed. The **Disclosured Risk Distribution** section on the right displays the distribution of unfixed vulnerabilities. ◦ **Resolved Risk**: displays the numbers of high-risk vulnerabilities, medium-risk vulnerabilities, and low-risk vulnerabilities, and the total number of these vulnerabilities for the current day. These vulnerabilities are fixed. The **Resolved Risk Distribution** section on the right displays the distribution of fixed vulnerabilities.
Asset Risk Top 5	View the top five assets that are at risk on the **Security Vulnerabilities** and **Host Compliance** tabs. These assets are displayed by asset or group.
Risk Monitoring Trend	View the trend charts of vulnerabilities on the **Security Vulnerabilities** and **Host Compliance** tabs. Fixed and unfixed vulnerabilities are identified by lines in different colors. You can move the pointer over a line to view the numbers of unfixed vulnerabilities and fixed vulnerabilities for the specific day.
Asset Monitoring Trend	View the trends in the numbers of protected hosts and websites. Hosts and websites are identified by lines in different colors. You can move the pointer over a line to view the number of protected hosts and websites for the specific day.
Risk Asset Ranking List	View the rankings of assets that are at risk on the **Latest Risk** and **High Risk** tabs.
Port Service Statistics	View the statistics on the **Port** and **Host Service** tabs.

22.1.4.8.3. Asset management

22.1.4.8.3.1. View the results of asset analysis

This topic describes how to view the analysis results of websites and hosts.

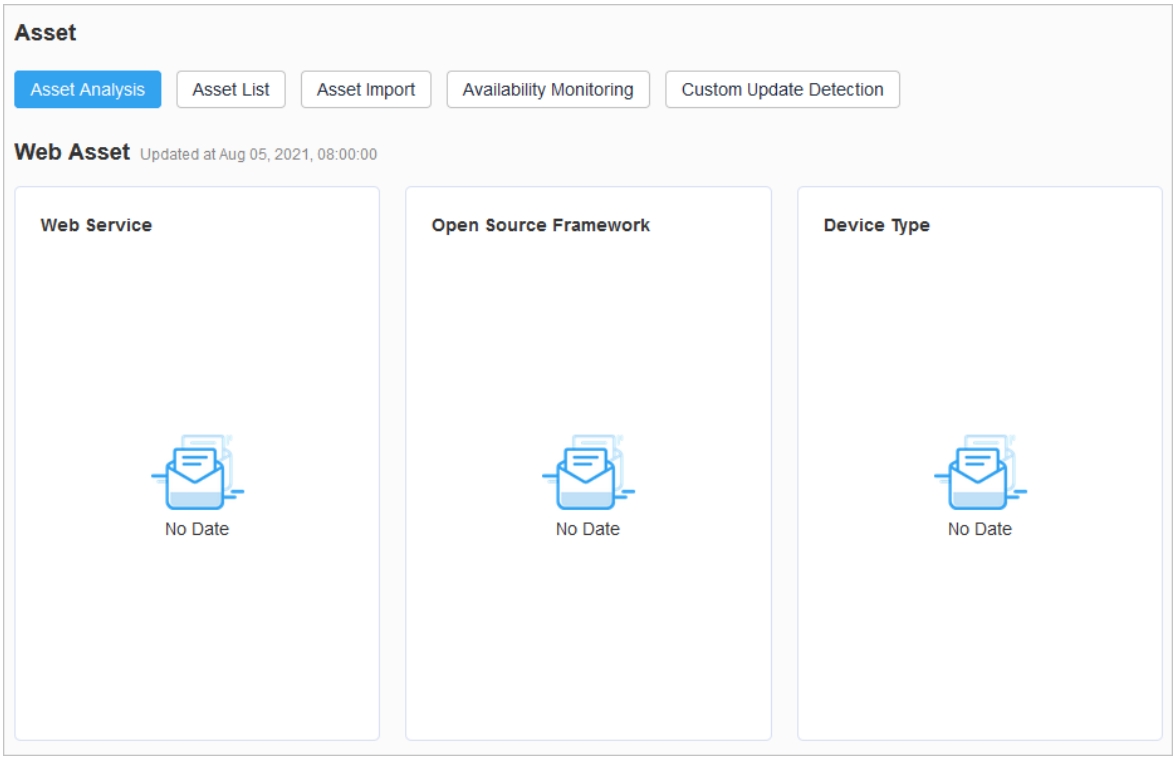
Context

The asset analysis feature allows you to view the analysis results of websites and hosts. For the websites, you can view Web Service, Open Source Framework, and Device Type. For the hosts, you can view Host Port, Host Service, and Operation System.

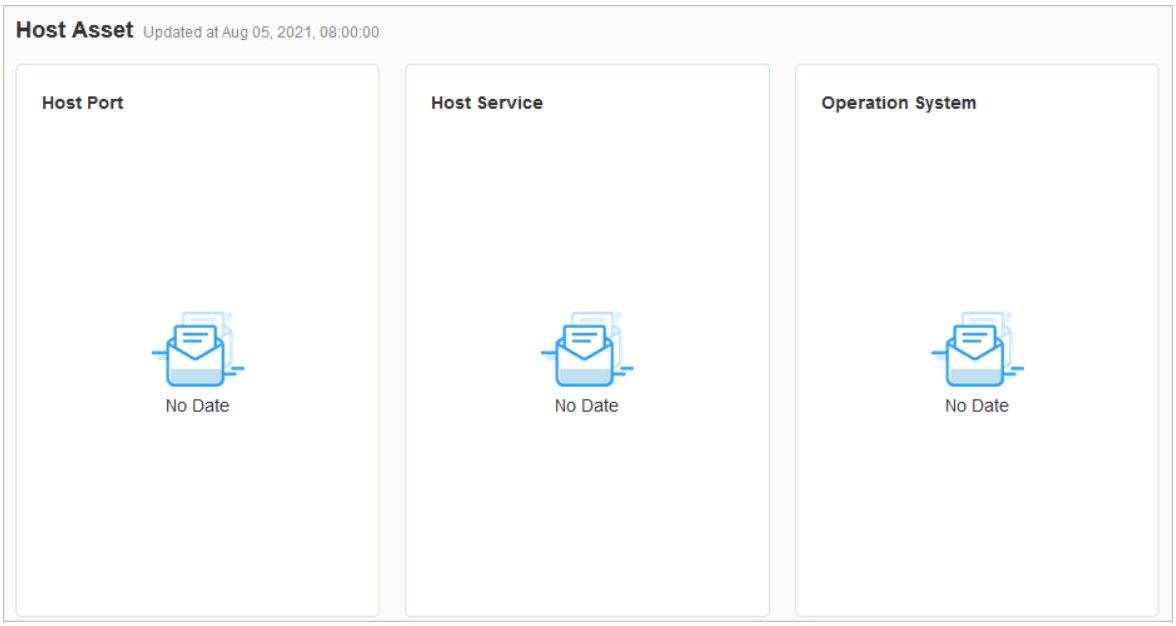
Procedure

1.

2.
3. In the **Vulnerability Scan** pane, click **Asset**.
4. On the **Asset** page, click the **Asset Analysis** tab to view websites.



5. View hosts.



22.1.4.8.3.2. Import assets

This topic describes how to import Internet assets.

Context

The vulnerability scan feature works only on imported assets. If you want to scan the vulnerabilities of your assets,

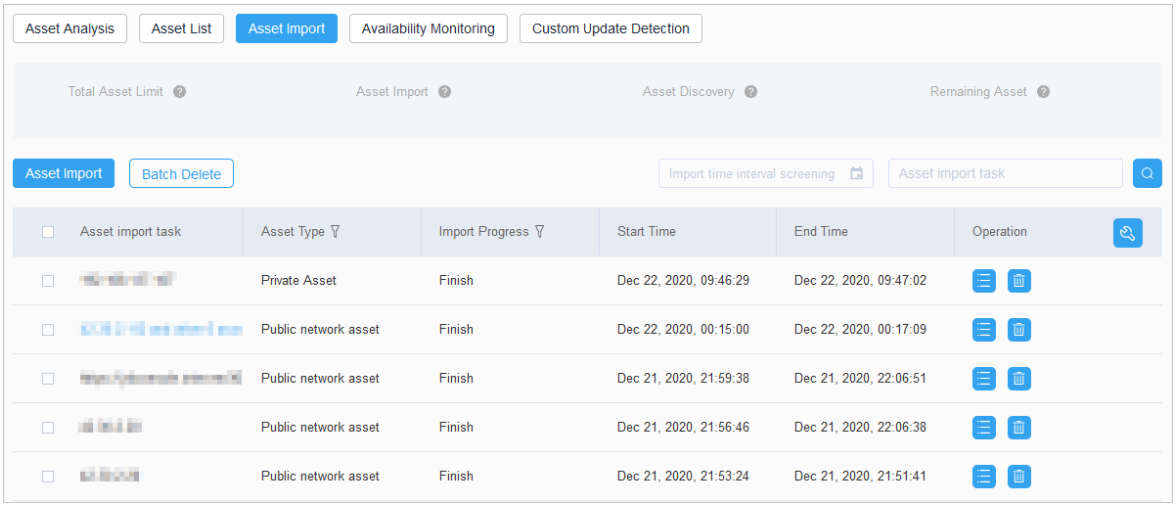
you must import your assets.

The assets that the feature supports include Internet assets and internal assets. The internal assets refer to the assets in a virtual private cloud (VPC).

- To import internal assets, you must add a scan engine. For more information, see Configure a scan engine for internal assets.
- To import Internet assets, perform the operations provided in this topic.

Procedure

1.
2.
3. In the **Vulnerability Scan** pane, click **Asset**.
4. On the **Asset** page, click the **Asset Import** tab to view imported assets.



5. Click **Asset Import**. On the **Public network asset Import** page, create an asset import task.
   i. In the **Import Asset** section, select **Manual Import** and enter the required assets in the field. Then, read and select the disclaimer.
      - You can enter domain names, URLs, IP addresses, and CIDR blocks.
      - You can enter the information about multiple assets at a time. Press Enter after you enter the information about one asset.
      - You cannot enter the information about the assets in VPCs.
      - The number of imported assets must be less than the number of remaining assets supported by the platform.

        > **Note** For example, if the number of remaining assets supported by the platform is 100 and 90 assets are entered, all the assets can be scanned. If 110 assets are entered, only 100 assets can be scanned, and the 10 assets that remain cannot be scanned.

   ii. In the **Asset Info** section, group the imported assets and configure an owner and a tag for the assets.
      - **Asset Group**: Select a group from the drop-down list. You can click the icon to create, edit, or delete a group.
      - **Person in charge**: Select an owner from the drop-down list. You can click the icon to create, edit, or delete an owner.
      - **Asset Tag**: Click **Add Tag** to add a tag to the imported assets.

iii. In the **Import Set** section, select the operations that you want to perform after the assets are imported.

Operation		Description
Asset Discovery	*Auto Import subdomains*	Automatically queries the subdomain assets of the imported domain names.
	Auto import associated IP	Automatically adds IP address assets that are mapped to the domain names.
	Auto synchronize tags and groups	Applies the group and tag of the imported assets to the assets that are discovered by the system.
Web Asset	*Open WEB Monitoring*	Enables the web monitoring feature on the imported website assets. If you want to select the web monitoring rules to use, click the icon. In the dialog box that appears, select the required web monitoring rules. For more information about how to configure web monitoring rules, see Configure web monitoring.

iv. In the **Whitelist** section, add the assets that do not need to be scanned.

You can enter IP addresses and URLs. If you add more than one asset, you must press Enter after you enter the information about an asset.

v. Click **Save**.

6. Manage the created asset import task.

After the asset import task is created, you can view the task in the task list. You can also perform the following operations on the task.

Icon	Description
	View the details, result, and process of the asset import task.
	Delete the asset import task.

22.1.4.8.3.3. Manage assets

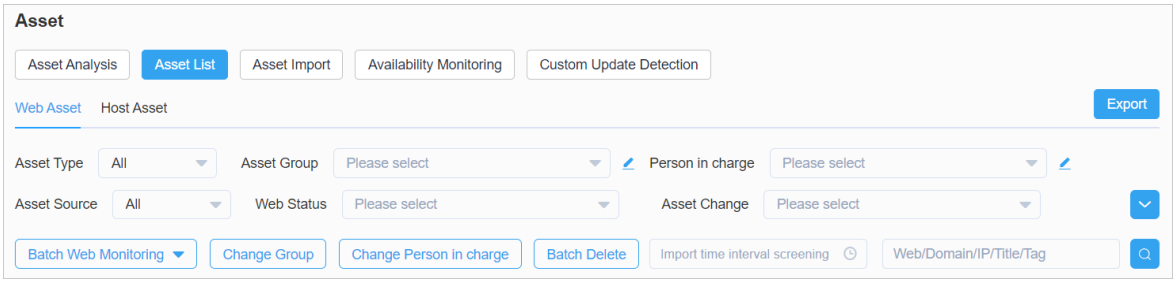
This topic describes how to view and manage assets.

Context

You can view the information about assets in the asset list. If the purpose or owner of an asset changes, security administrators can move the asset to another group or change the owner.

Procedure

1.
2.
3. In the Vulnerability Scan pane, click **Asset**.
4. On the **Asset** page, click the **Asset List** tab to view assets.



5. Click the **Web Asset** tab and manage websites.
   i. Specify filter conditions to search for websites.

      The filter feature allows you to search for websites in a more efficient manner.

Filter condition	Description
Asset Type	The type of the asset that you want to view.
Asset Source	The source from which the asset is imported. Valid values: **Manual** and **System Find**.
Asset Group	The group to which the asset belongs.
Web Status	The status of the website.
Person in charge	The owner of the website.
Asset Change	The change status of the asset. Valid values: **All**, **New**, **Update**, **No Update**, and **Offline**.
Web Monitoring	The monitoring status of the website.
Risk Level	The risk level of the asset.
Web Service	The service type and version of the website.
WAF Recognition	Specifies whether the asset is identified by Web Application Firewall (WAF).
Open Source Framework	The open source framework type of the asset.
Device Type	The type of the device.
Time range	The time range during which the asset is imported.
Key information	The key information of the asset. The key information includes the website, domain name, IP address, title, and tag.

   ii. Click **Export** to export the asset list to an EXCEL file.

iii. Select one or more websites that you want to manage. The following table describes the operations that you can perform on the websites.

Operation	Description
Batch Web Monitoring	Allows you to enable or disable web monitoring for multiple websites. ■ **Batch Open Monitoring**: To enable web monitoring for multiple websites, select Batch Open Monitoring from the drop-down list of Batch Web Monitoring. ■ **Batch Stop Monitoring**: To disable web monitoring for multiple websites, select Batch Stop Monitoring from the drop-down list of Batch Web Monitoring
Change Group	Allows you to change the group of multiple assets at a time.
Change Person in charge	Allows you to change the owner of multiple assets at a time.
Batch Delete	Allows you to delete multiple assets at a time. After the assets are deleted, the assets are not scanned by the vulnerability scan feature.

6. Click the **Host Asset** tab and manage hosts.
   i. Specify filter conditions to search for hosts.

   The filter feature allows you to search for hosts in a more efficient manner.

Filter condition	Description
Asset Type	The asset type, such as Internet assets or a specific VPC.
Asset Source	The source from which the asset is imported. Valid values: **Manual** and **System Find**.
Asset Group	The group to which the asset belongs.
Person in charge	The owner of the asset.
Asset Change	The change status of the asset. Valid values: **All**, **New**, **Update**, **No Update**, and **Offline**.
Risk Level	The risk level of the asset. Valid values: **All**, **High**, **Middle**, **Low**, and **Security**.
SurviveStatus	The liveness status of the asset. Valid values: **Alive** and **Close**.
Operation System	The operating system of the host.
Host Port	The port of the host.
CDN Recognition	Specifies whether Content Delivery Network (CDN) is configured for the asset.
Host Service	The service of the host.
Time range	The time range during which the asset is imported.
Key information	The key information of the asset. The key information includes the IP address, host, tag, and domain name.

   ii. Click **Export** to export the asset list to an EXCEL file.

iii. Select one or more hosts that you want to manage. The following table describes the operations that you can perform on the hosts.

Operation	Description
Change Group	Allows you to change the group of multiple assets at a time.
Change Person in charge	Allows you to change the owner of multiple assets at a time.
Batch Delete	Allows you to delete multiple assets at a time. After the assets are deleted, the assets are not scanned by the vulnerability scan feature.

22.1.4.8.3.4. Manage asset availability

This topic describes how to manage the availability of assets.

Context

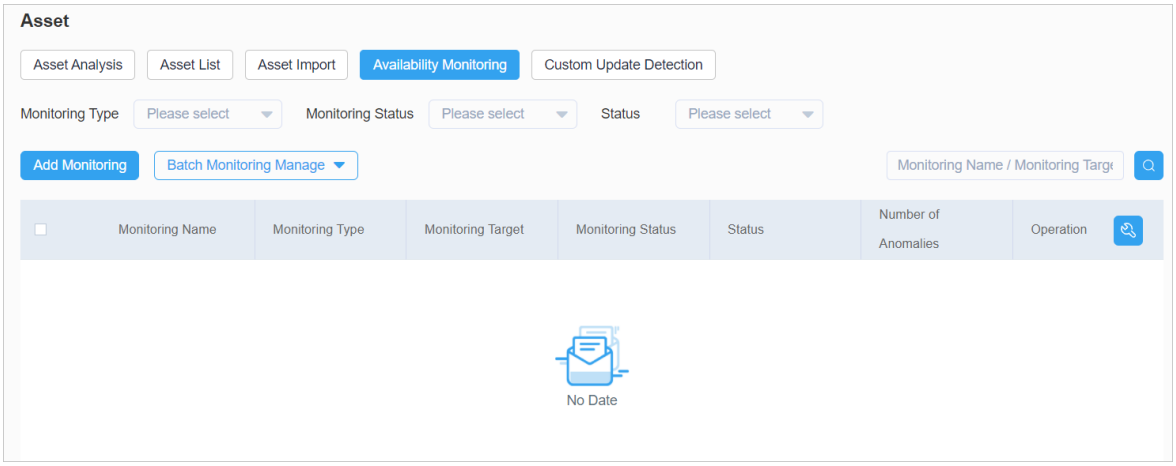
If hosts or websites are used for a long period of time, they may become unavailable due to errors. Availability monitoring allows a security administrator to discover unavailable assets. Then, the security administrator can troubleshoot the issues that cause the assets to become unavailable.

Availability monitoring supports the following methods:

- HTTP monitoring: This method is used to monitor websites.
- PING monitoring: This method is used to monitor hosts.

Procedure

1.
2.
3. In the **Vulnerability Scan** pane, click **Asset**.
4. On the **Asset** page, click the **Availability Monitoring** tab to view availability monitoring tasks.



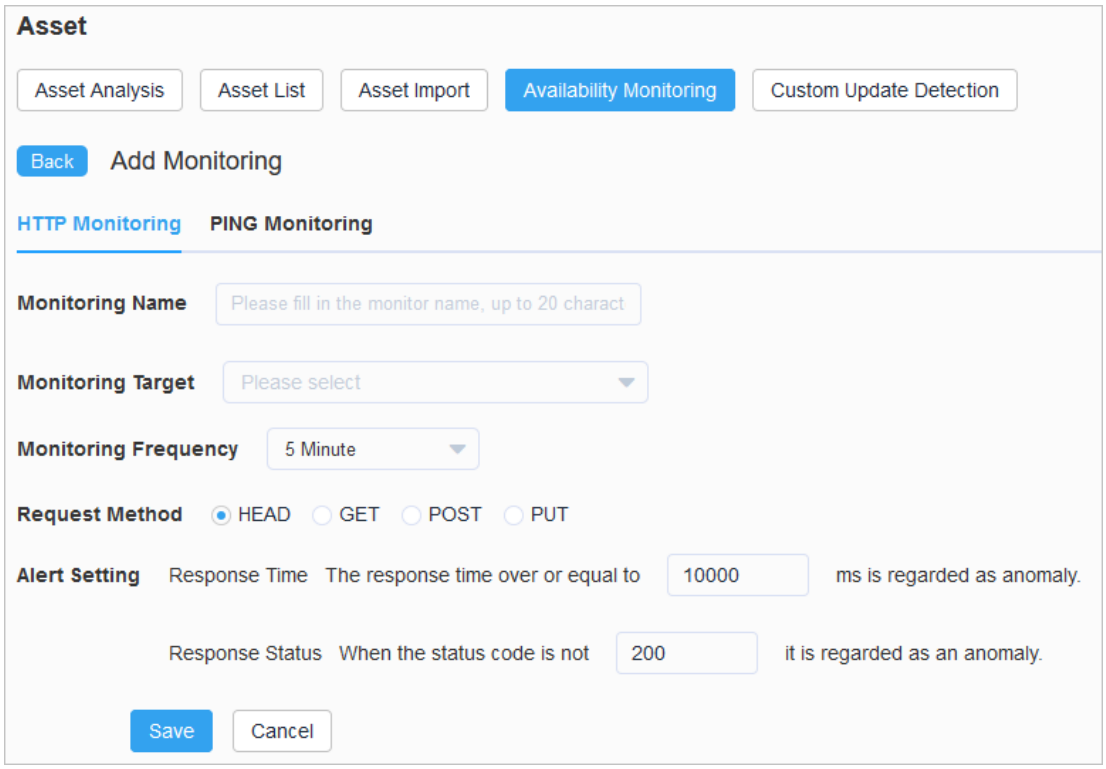
5. Create an availability monitoring task.

   Availability monitoring supports HTTP monitoring and PING monitoring.

   - To create an HTTP monitoring task, perform the following steps:

     a. Click **Add Monitoring**.

b. Click the **HTTP Monitoring** tab.



c. Configure the following parameters.

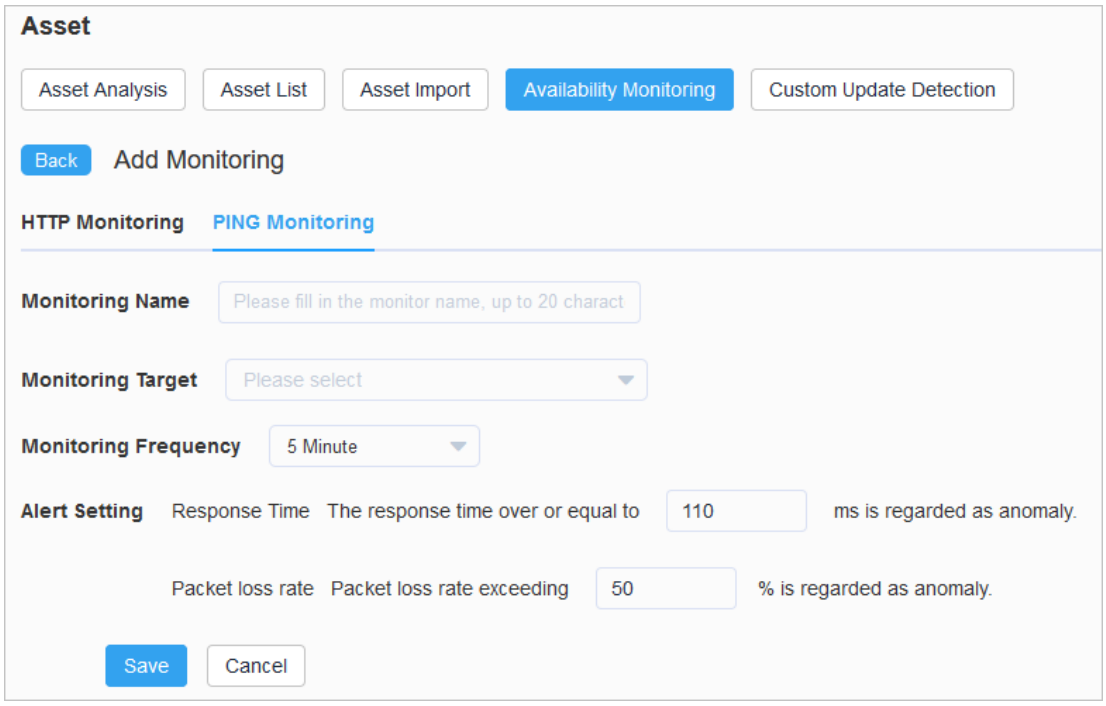
Parameter	Description
Monitoring Name	The name of the availability monitoring task.
Monitoring Target	The website that you want to monitor.
Monitoring Frequency	The interval at which you want to monitor the website. Valid values: **1 Minute**, **5 Minute**, **15 Minute**, and **30 Minute**.
Request Method	The request method that is used to send HTTP request packets. Valid values: **HEAD**, **GET**, **POST**, and **PUT**.
Alert Setting	The policy based on which Apsara Stack Security reports alerts. If one of the following conditions is met, the website is unavailable: ■ Response Time: If the actual response time is greater than the specified value, an exception occurs. ■ Response Status: If an unexpected status code is returned, an exception occurs.

d. Click **Save**.

- To create a PING monitoring task, perform the following steps:

a. Click **Add Monitoring**.

b. Click the **PING Monitoring** tab.



c. Configure the following parameters.

Parameter	Description
Monitoring Name	The name of the availability monitoring task.
Monitoring Target	The host that you want to monitor.
Monitoring Frequency	The interval at which you want to monitor the host. Valid values: **1 Minute**, **5 Minute**, **15 Minute**, and **30 Minute**.
Alert Setting	The policy based on which Apsara Stack Security reports alerts. If one of the following conditions is met, the host is unavailable: ■ Response Time: If the actual response time is greater than the specified value, an exception occurs. ■ Response Status: If an unexpected status code is returned, an exception occurs.

d. Click **Save**.

6. Manage more than one availability monitoring task at a time.

   You can manage more than one availability monitoring task in the monitoring task list at a time.

   - Start more than one availability monitoring task at a time.

     Select more than one availability monitoring task and choose **Batch Monitoring Manage > Batch Open Monitoring**.

   - Stop more than one availability monitoring task at a time.

     Select more than one availability monitoring task and choose **Batch Monitoring Manage > Batch Stop Monitoring**.

   - Delete more than one availability monitoring task at a time.

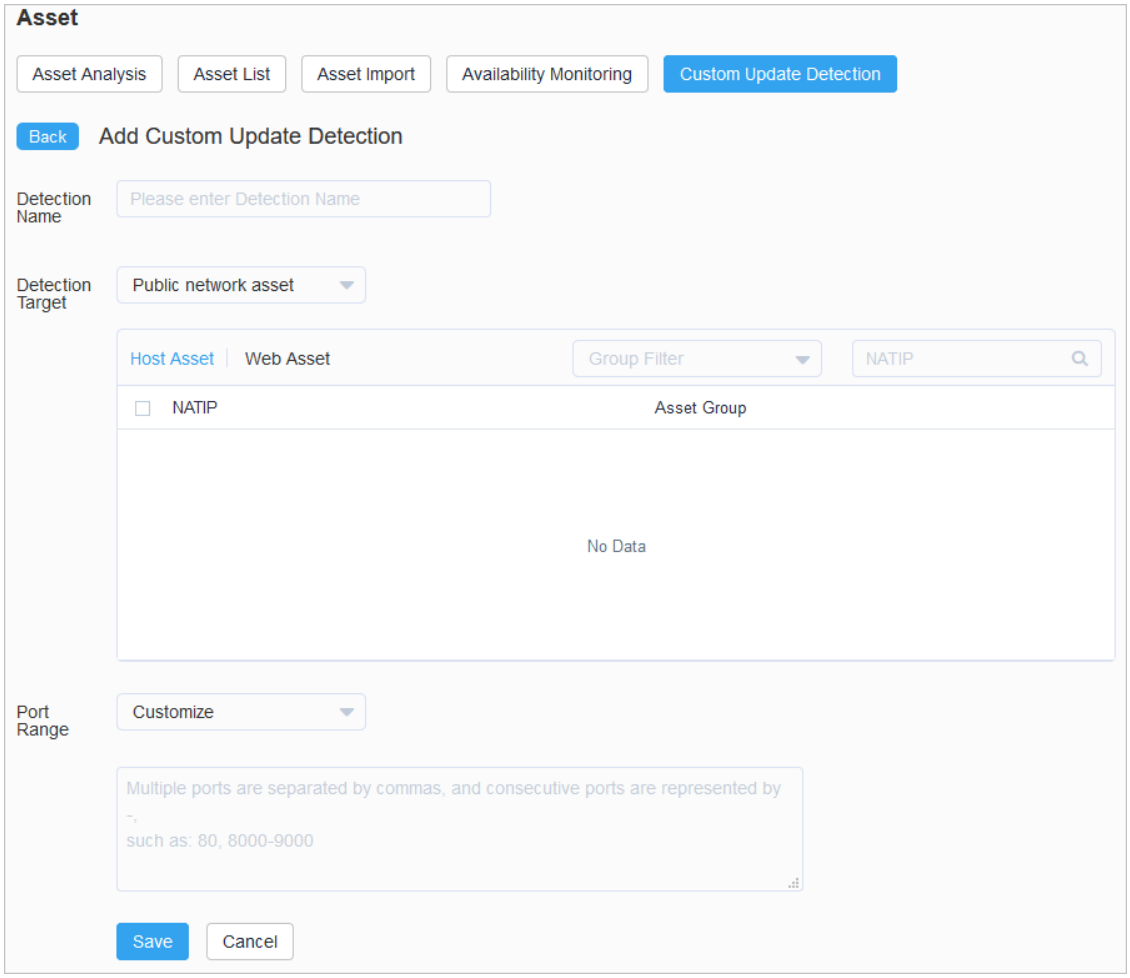
     Select more than one availability monitoring task and choose **Batch Monitoring Manage > Batch Delete Monitoring**.

22.1.4.8.3.5. Manage custom update detection tasks

This topic describes how to manage custom update detection tasks.

Procedure

1.
2.
3. click **Asset**.
4. On the **Asset** page, click the **Custom Update Detection** tab to view custom update detection tasks.
5. Create a custom update detection task.
   i. Click **Add Detection**.
   ii. On the **Add Custom Update Detection** page, configure the parameters.



Parameter	Description
Detection Name	The name of the custom update detection task.

Parameter	Description
Detection Target	The asset that you want to detect. The value is fixed as **Public network asset**. a. Click **Host Asset** or **Web Asset** based on your business requirements. b. Select the assets that you want to detect from the asset list. **Note** You can search for assets by using **Group Filter** or **NAT IP**.
Port Range	The range of ports that you want to detect. Valid values: **Customize**, **Full Port**, **Top100**, and **Top1000**. **Note** The **Port Range** parameter appears only if you click Host Asset when you configure the **Detection Target** parameter. ■ **Customize**: You can specify custom ports to detect. ■ **Full Port**: All ports are detected. ■ **Top 100**: Top 100 ports are detected. ■ **Top 1000**: Top 1,000 ports are detected.

iii. Click **Save**.

22.1.4.8.4. Risk management

22.1.4.8.4.1. Manage vulnerabilities

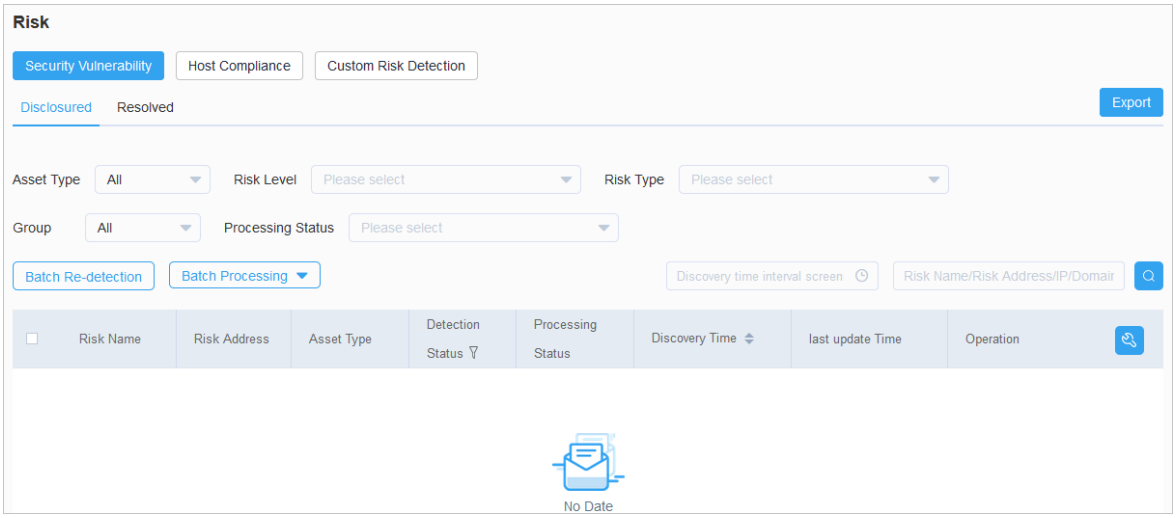
This topic describes how to view and handle the vulnerabilities that are detected by the vulnerability scan feature.

Context

On the **Security Vulnerability** tab, security administrators can view the vulnerabilities that are detected by the vulnerability scan feature.

Procedure

1.
2.
3. click **Risk**.
4. On the **Risk** page, click the **Security Vulnerability** tab to view vulnerabilities.



- Click the **Disclosured** or **Resolved** tab to view unfixed vulnerabilities or fixed vulnerabilities.
- View risk statistics. The statistics include **Disclosured Risk**, **Resolved Risks**, **Unconfirmed Risk**, **Confirmed Risk**, and **Ignored Risk**.

5. Specify search conditions to view specific vulnerabilities. The conditions include **Asset Type** and **Risk Level**.
6. Handle vulnerabilities.

   Security administrators can analyze and confirm whether the vulnerabilities affect the security of assets based on the vulnerability information.

   - Confirm risks

     If a vulnerability affects the security of assets, confirm the risk after the security vulnerability is fixed.

     a. Find the vulnerability and click the icon in the **Operation** column.

     b. In the drop-down list, select **Confirm Risk**.

     c. In the dialog box that appears, click **OK**.

   - Ignore risks

     If a vulnerability is a false positive or does not affect the security of assets, ignore the risk.

     a. Find the vulnerability and click the icon in the **Operation** column.

     b. In the drop-down list, select **Ignore Risk**.

     c. In the dialog box that appears, click **OK**.
7. Click **Export** to export the list of vulnerabilities to your computer.

22.1.4.8.4.2. Manage host compliance risks

This topic describes how to view and confirm host compliance risks.

Context

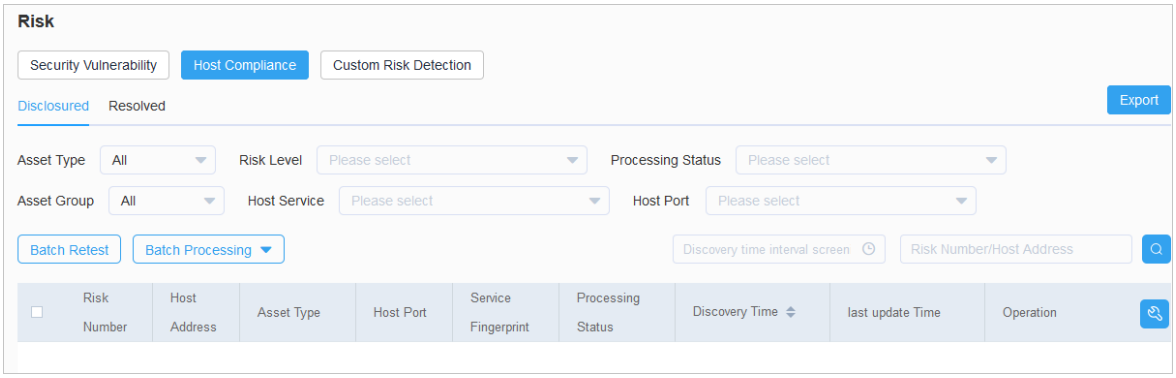
On the **Host Compliance** tab, security administrators can view the host compliance issues that are detected by the vulnerability scan feature.

Procedure

1.
2.

3. click **Risk**.
4. On the **Risk** page, click the **Host Compliance** tab to view host compliance risks.

   You can click the **Disclosured** or **Resolved** tab to view **unfixed vulnerabilities** or **fixed vulnerabilities**.



5. Specify conditions to search for host compliance risks. The conditions include **Asset Type** and **Risk Level**.
6. Handle host compliance risks.

   Security administrators can analyze and confirm whether host compliance risks affect the security of assets based on the risk information.

   - Confirm risks

     If a host compliance risk affects the security of assets, harden the security of hosts and confirm the risk.

     a. Find the risk and click the icon in the **Operation** column.

     b. In the drop-down list, select **Confirm Risk**.

     c. In the message that appears, click **Sure**.

   - Ignore risks

     If a host compliance risk proves to be a false positive or does not affect the security of assets, ignore the risk.

     a. Find the risk and click the icon in the **Operation** column.

     b. In the drop-down list, select **Ignore Risk**.

     c. In the message that appears, click **Sure**.
7. Click **Export** to export the list of host compliance risks to your computer.

22.1.4.8.4.3. Create a custom risk detection task

This topic describes how to create a custom risk detection task.

Procedure

1.
2.
3. choose **Risk > Custom Risk Detection**.
4. On the Custom Risk Detection tab, click **Add Detection**.
5. On the **Add Custom Risk Detection** page, configure the parameters.

Parameter	Description
Detection Name	The name of the custom risk detection task.
Detection Target	The asset on which you want to perform risk detection. The value is fixed as **Public network asset**.
Emergency Detection	The switch that is used to enable or disable the emergency detection feature. If you enable this feature, you can select emergency detection items from the detection item list.
Basic Risk Detection	The switch that is used to enable or disable the basic risk detection feature. For more information about how to configure this feature, see Configure basic monitoring.
WEB Risk Detection	The switch that is used to enable or disable the web risk detection feature. For more information about how to configure this feature, see Configure web monitoring.

6. Click **Save**.

22.1.4.8.5. Report management

22.1.4.8.5.1. Create a report

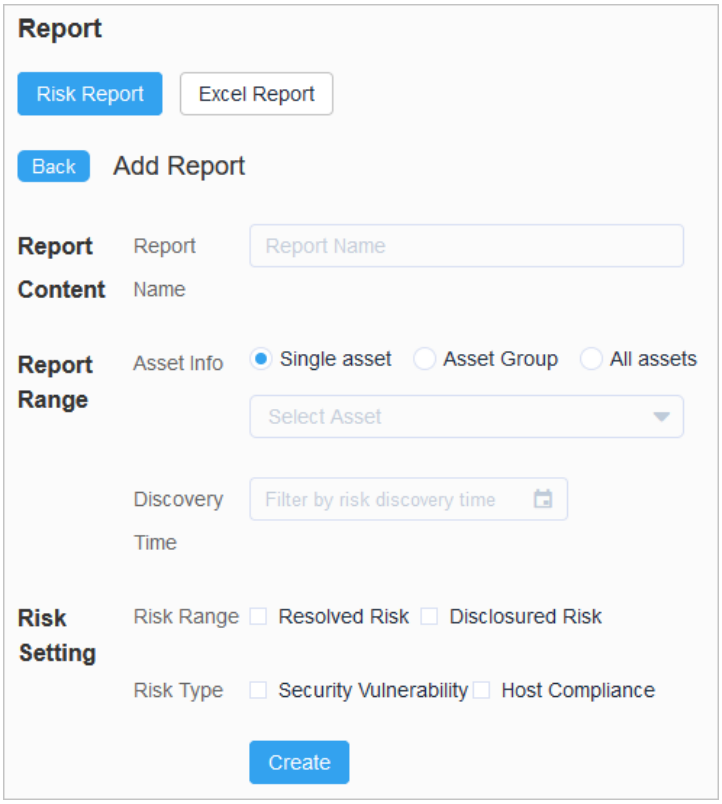
This topic describes how to create a report.

Context

A security administrator can create a report to view the security postures of specific assets during a period of time and implement security measures as required.

Procedure

1.
2.
3. click **Report**.
4. On the **Risk Report** tab, click **Add Report**.
5. Configure the following parameters.



Parameter		Description
Report Content	**Report Name**	The name of the report that you want to create.
Report Range	**Asset Info**	The scope of assets that you want to include in the report. Valid values: **Single asset**, **Asset Group**, and **All assets**. ○ **Single asset**: Select an asset. ○ **Asset Group**: Select an asset group and a tag. **Note** After you select an asset group and a tag, the assets in the group that have the selected tag are included in the report. ○ **All assets**: Select assets by tag.
	Discovery Time	The time range in which you want to perform risk detection.
Risk Setting	**Risk Range**	The scope of risks that you want to include in the report. Valid values: **Resolved Risk** and **Disclosured Risk**.
	Risk Type	The types of risks that you want to include in the report. Valid values: **Security Vulnerability** and **Host Compliance**.

6. Click **Create**.

Result

After the report is created, it appears in the report list on the **Report** page.

22.1.4.8.5.2. Delete multiple reports at a time

This topic describes how to delete multiple reports at a time.

Context

You can delete multiple reports that you no longer need at a time to save storage space.

Procedure

1.
2.
3. click **Report**.
4. Click **Risk Report**.
5. In the report list, select the reports that you want to delete.
6. Click **Batch Delete**.

22.1.4.8.6. Configuration management

22.1.4.8.6.1. Configure overall monitoring

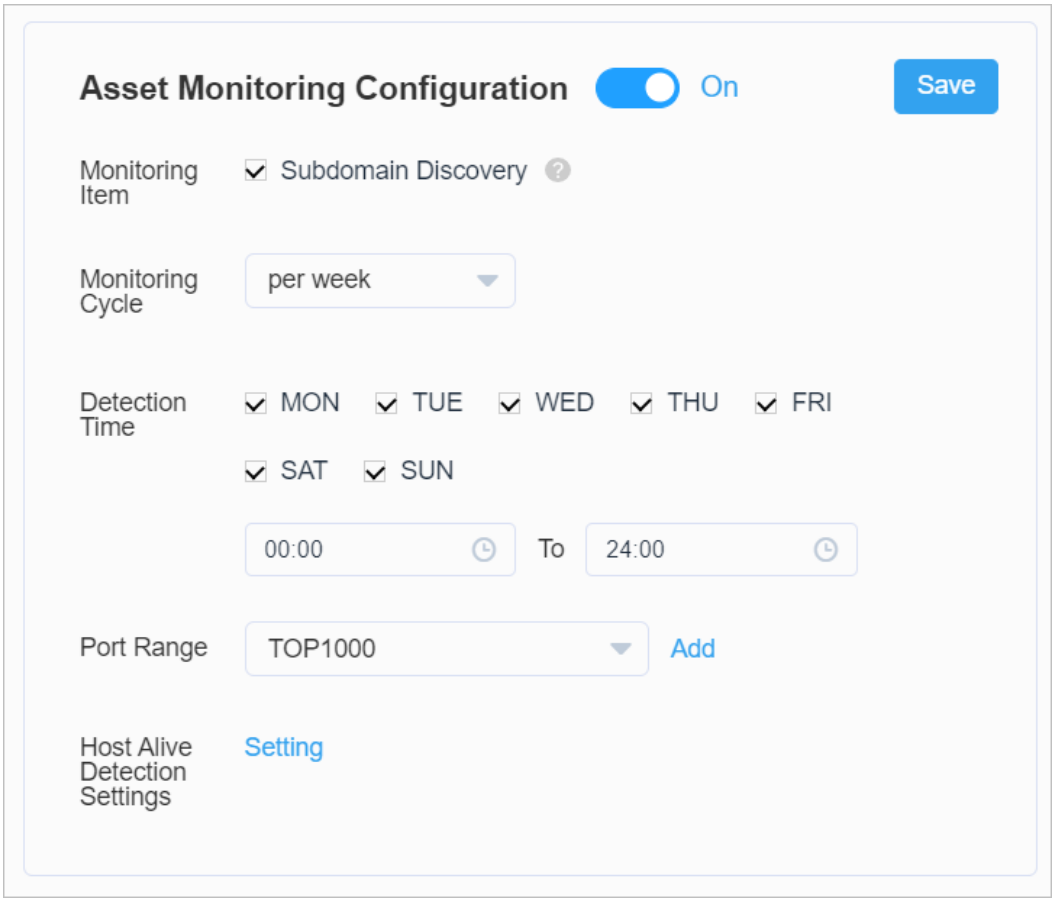
This topic describes how to configure overall monitoring for the vulnerability scan feature. Overall monitoring includes Asset Monitoring Configuration, Base Risk Monitoring Configuration, External Risk Monitoring Configuration, and Scan Configuration.

Context

Overall monitoring allows you to configure detection features and the monitoring cycle for each detection feature.

Procedure

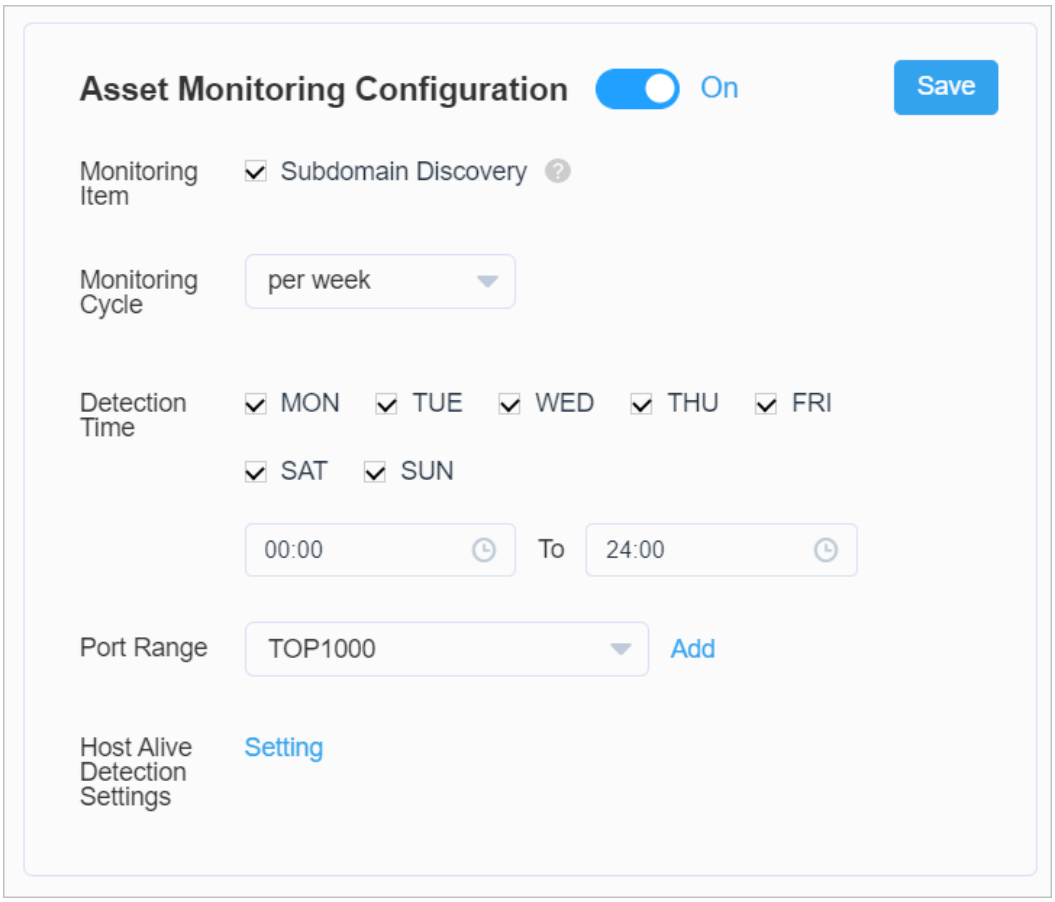
1.
2.
3. click **Configuration**.
4. On the **Monitoring Configuration** tab, click **Overall Monitoring**.
5. In the **Monitoring Status** section, view the status of overall monitoring.



6. Configure detection features.

   Detection features include **Asset Monitoring Configuration**, **Base Risk Monitoring Configuration**, **External Risk Monitoring Configuration**, and **Scan Configuration**.

In this step, the **Asset Monitoring Configuration** detection feature is used as an example.



i. Turn on Asset Monitoring Configuration to enable the asset monitoring feature.

- After the switch is turned on, the switch is in the **On** state. In the On state, the switch is blue. After the switch is turned off, the switch is in the **Off** state. In the Off state, the switch is gray.
- You must turn on **Asset Monitoring Configuration** and **Base Risk Monitoring Configuration** to enable the two features. External Risk Monitoring Configuration and Scan Configuration are automatically enabled.

ii. Configure the following parameters.

Asset Monitoring Configuration

Parameter	Description
Monitoring Item	The item that you want to monitor. Valid value: **Subdomain Discovery**. If you want to import assets, you must set the Import Set parameter to **Auto Import subdomains**. Then, subdomains are automatically imported. If you select **Subdomain Discovery**, Apsara Stack Security regularly discovers subdomains for assets whose Import Set parameter is set to **Auto Import subdomains**.

Parameter	Description
Monitoring Cycle	The cycle based on which you want to perform detection. Valid values: **customization**, **per week**, and **per month**. ■ **customization**: Specify the interval at which you want to perform detection. Unit: days. ■ **per week**: Specify the days of each week on which you want to perform detection. ■ **per month**: Specify the days of each month on which you want to perform detection.
Detection Time	The time when you want to perform detection. The time varies based on the value of the **Monitoring Cycle** parameter. ■ If you set the **Monitoring Cycle** parameter to **customization**, select a time range of the day in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **per week**, select the days of each week and the time range in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **per month**, select the days of each month and the time range in which you want to perform detection. **Note** For example, if you set the **Monitoring Cycle** parameter to **per week** and select **Monday** to **Sunday** and 00:00:00 to 24:00:00 for the **Detection Time** parameter, Apsara Stack Security performs detection 24 hours a day, 7 days a week.
Port Range	The ports on which you want to perform detection. Valid values: **customization**, **Full port**, **TOP100**, and **TOP1000**. ■ **customization**: Specify the ports to scan. ■ **Full port**: Scan all ports. ■ **TOP100**: Scan top 100 ports. You can click **Add** to add more ports. ■ **TOP1000**: Scan top 1,000 ports. You can click **Add** to add more ports.
Host Alive Detection Settings	The option that is used to check whether a host is running. By default, the ping feature is used to check whether a host is running. If the host has the ping feature disabled, the status of the host is checked based on top 20 ports and custom ports. To specify custom ports, click **Settings**. In the Host Alive Detection Settings dialog box, specify the ports in the **Custom Port** field.

Base Risk Monitoring Configuration

Parameter	Description

Parameter	Description
Monitoring Item	The item that you want to monitor. Valid values: **Weak Password**, **Common Vulnerabilities**, **Baseline Monitoring**, and **Host Compliance**. ■ **Weak Password**: Attackers can guess passwords or launch brute-force attacks to crack passwords. Then, the attackers can obtain relevant permissions. If you select this item, weak password vulnerabilities can be identified. ■ **Common Vulnerabilities**: Web vulnerabilities and CMS application vulnerabilities are included. If you select this item, common vulnerabilities can be identified. This way, you can install patches at the earliest opportunity. ■ **Baseline Monitoring**: Risks in host configuration and account configuration are detected. ■ **Host Compliance**: Host compliance risks are detected.
Monitoring Cycle	The cycle based on which you want to perform detection. Valid values: **customization**, **per week**, and **per month**. ■ **customization**: Specify the interval at which you want to perform detection. Unit: days. ■ **per week**: Specify the days of each week on which you want to perform detection. ■ **per month**: Specify the days of each month on which you want to perform detection.
Detection Time	The time when you want to perform detection. The time varies based on the value of the **Monitoring Cycle** parameter. ■ If you set the **Monitoring Cycle** parameter to **customization**, select a time range of the day in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **per week**, select the days of each week and the time range in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **per month**, select the days of each month and the time range in which you want to perform detection. **Note** For example, if you set the **Monitoring Cycle** parameter to **per week** and select **Monday** to **Sunday** and 00:00:00 to 24:00:00 for the **Detection Time** parameter, Apsara Stack Security performs detection 24 hours a day, 7 days a week.

External Risk Monitoring Configuration

Parameter	Description
Monitoring Item	The item that you want to monitor. Valid value: **Code Disclosure**. If you select Code Disclosure, Apsara Stack Security detects leaked source code of your assets.

Parameter	Description
Monitoring Cycle	The cycle based on which you want to perform detection. Valid values: **customization**, **per week**, and **per month**. ■ **customization**: Specify the interval at which you want to perform detection. Unit: days. ■ **per week**: Specify the days of each week on which you want to perform detection. ■ **per month**: Specify the days of each month on which you want to perform detection.
Detection Time	The time when you want to perform detection. The time varies based on the value of the **Monitoring Cycle** parameter. ■ If you set the **Monitoring Cycle** parameter to **customization**, select a time range of the day in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **per week**, select the days of each week and the time range in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **per month**, select the days of each month and the time range in which you want to perform detection. **Note** For example, if you set the **Monitoring Cycle** parameter to **per week** and select **Monday** to **Sunday** and 00:00:00 to 24:00:00 for the **Detection Time** parameter, Apsara Stack Security performs detection 24 hours a day, 7 days a week.

Scan Configuration

Parameter	Description
Risk Re-detection	The time at which you want to perform detection again. If risks are detected in assets, Apsara Stack Security scans the assets again each day at the time you specify.
Asset Scanning Rate	The asset scan rate. Valid values: **Slow Mode**, **General Mode**, **Fast Mode**, and **Turbo Mode**.
Risk Scanning Rate	The risk scan rate. Valid values: **Slow Mode**, **General Mode**, and **Fast Mode**.
UserAgent Setting	The User-Agent property.

iii. Click **Save**.

22.1.4.8.6.2. Configure basic monitoring

This topic describes how to configure basic monitoring.

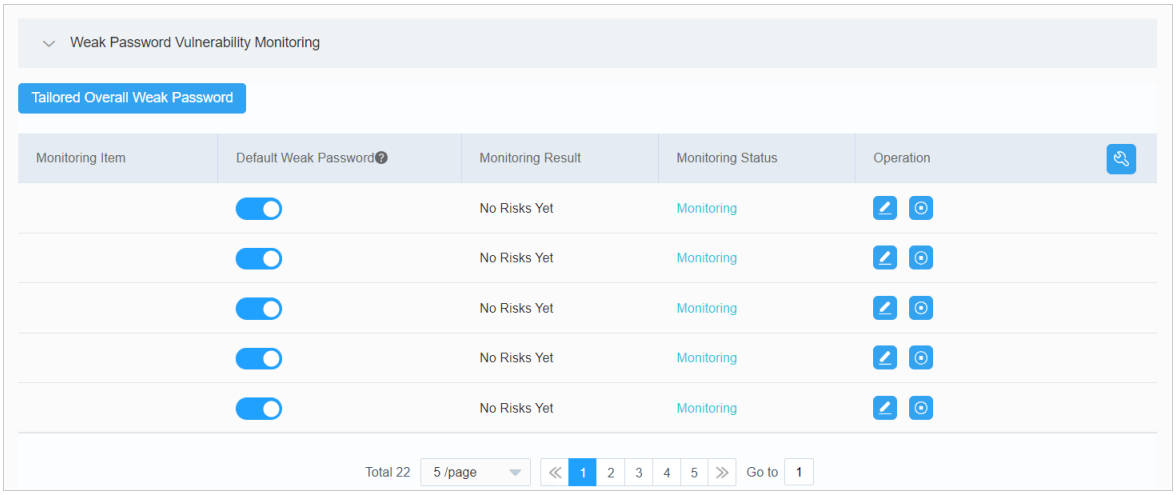
Context

Basic monitoring includes Weak Password Vulnerability Monitoring, Operation Security Vulnerability Monitoring, CMS Application Vulnerability Monitoring, and Baseline Monitoring.

Procedure

1.
2.

3. click **Configuration**.
4. On the **Configuration** page, click the **Monitoring Configuration** tab.
5. Click the **Basic Monitoring** tab. Then, click **Weak Password Vulnerability Monitoring** to configure rules to monitor weak passwords.



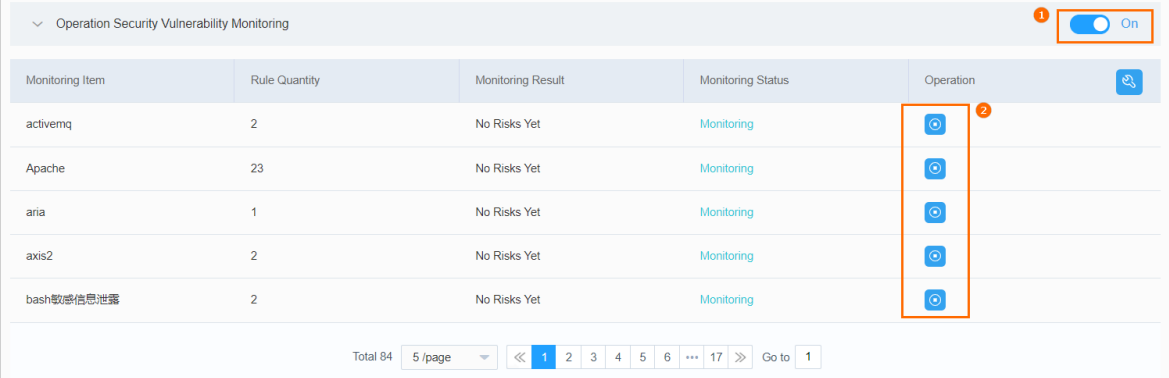
   - By default, all monitoring items on weak passwords use the default weak password library.
   - To disable a monitoring item, perform the following step:

     Find the monitoring item and click the icon in the **Operation** column.
   - To enable a monitoring item, perform the following step:

     Find the monitoring item and click the icon in the **Operation** column.
   - To specify custom weak passwords for a monitoring item, perform the following steps. In this example, **MySQL Weak Password Vulnerability** is used.
     a. In the **Default Weak Password** column, turn off the switch. The switch status changes to .
     b. In the **Operation** column, click the icon.
     c. In the **Customize MySQL Weak Password** dialog box, specify custom weak passwords.
     d. Click **Yes**.
   - To apply the same custom weak passwords to multiple monitoring items, perform the following steps:
     a. In the **Default Weak Password** column, turn off the switches for the monitoring items to which you want to apply the same custom weak passwords. The switch status changes to .

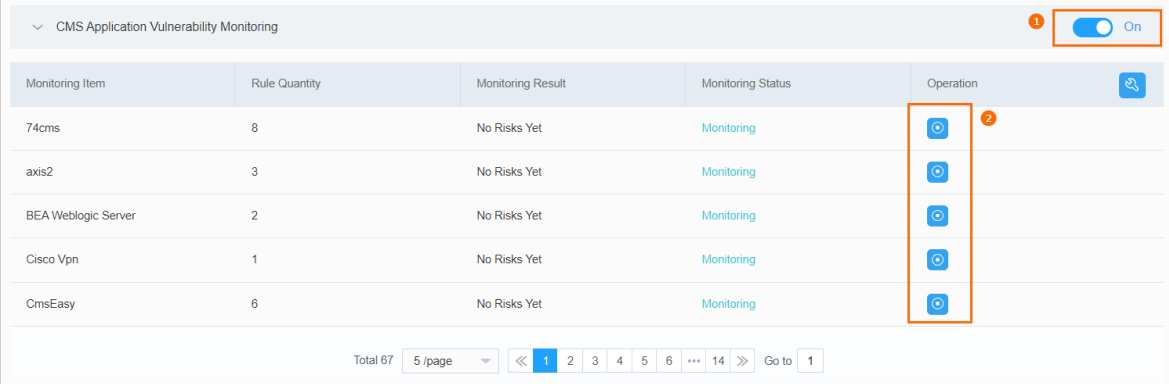
        > **Note** If you want to apply a custom weak password to a monitoring item, you must turn off the switch in the Default Weak Password column of the monitoring item.

     b. Click **Tailored Overall Weak Password**.
     c. In the **Tailored Overall Weak Password** dialog box, specify custom weak passwords.
     d. Click **Yes**.
6. Click **Operation Security Vulnerability Monitoring** and configure O&M vulnerability monitoring.



No.	Description
1	The switch that is used to enable or disable the **Operation Security Vulnerability Monitoring** feature. We recommend that you enable this feature to enhance system security.
2	The switch that is used to enable or disable a monitoring item. You can disable monitoring items based on your business requirements.

7. Click **CMS Application Vulnerability Monitoring** and configure monitoring on content management system (CMS) application vulnerabilities.



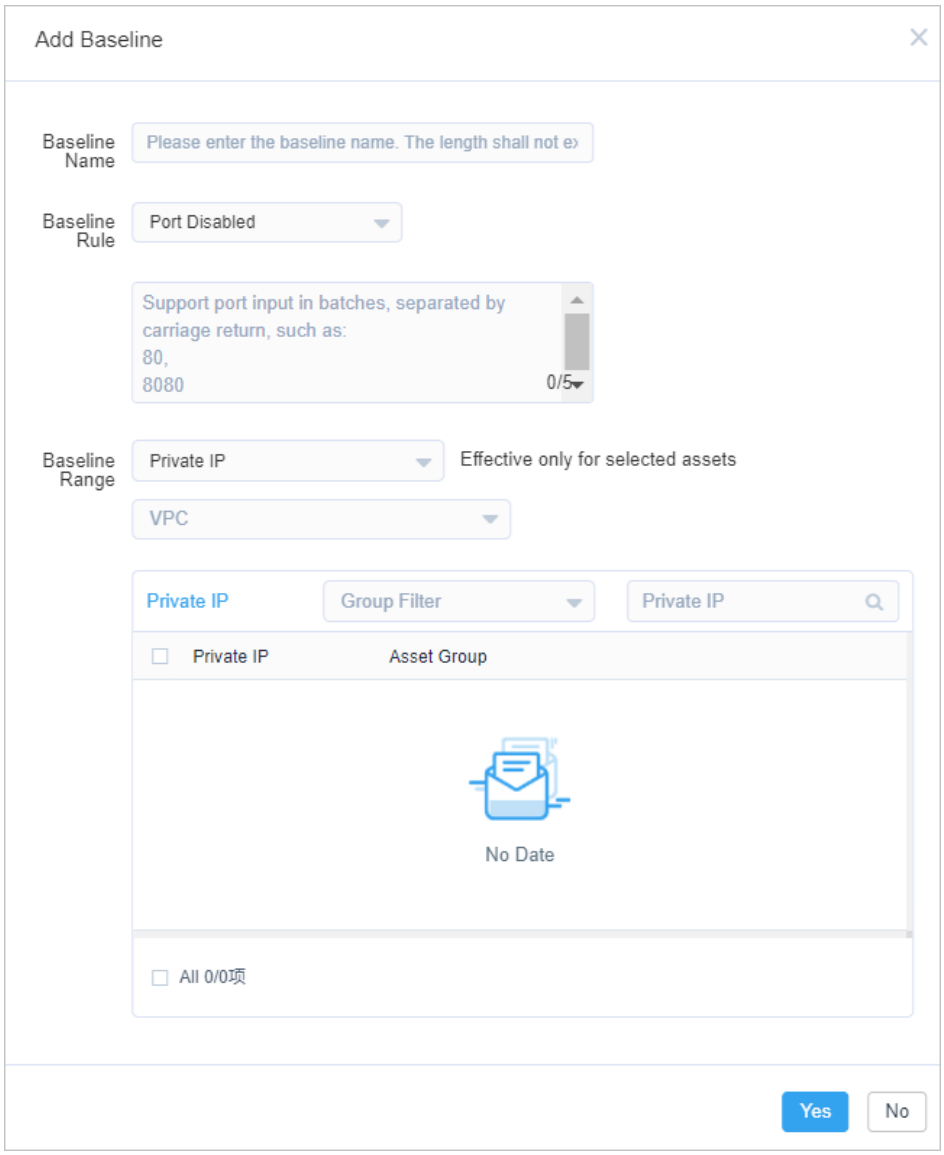
No.	Description
1	The switch that is used to enable or disable the **CMS Application Vulnerability Monitoring** feature. We recommend that you enable this feature to enhance system security.
2	The switch that is used to enable or disable a monitoring item. You can disable monitoring items based on your business requirements.

8. Click **Baseline Monitoring** and configure baseline monitoring.

   To add a baseline monitoring item, perform the following steps:

   i. Click **Add**.

ii. In the **Add Baseline** dialog box, configure the baseline monitoring item.



In this example, a baseline monitoring item is added to block Telnet-based access.

Parameter	Description
Baseline Name	The name of the baseline monitoring item. Example: Block Telnet-based access.
Baseline Rule	The detection rule that is used by the baseline monitoring item. This rule checks whether hosts use disabled ports or run disabled services. Valid values: ■ **Port Disabled**: ports that you want to disable. Example: 23. ■ **Service Disabled**: services that you want to disable. Example: Telnet.
Baseline Range	The scope of assets to which the baseline monitoring item can be applied. Valid values: **Private IP** and **NatIP**. You must specify this parameter and select assets.

iii. Click **Yes**.

22.1.4.8.6.3. Configure web monitoring

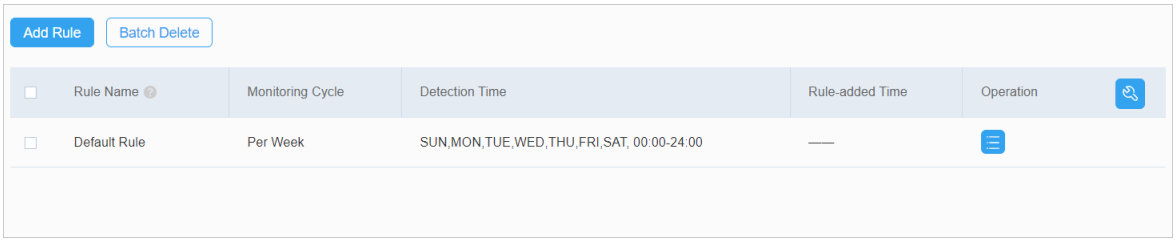
This topic describes how to configure web monitoring.

Context

Web monitoring allows you to configure monitoring items for monitoring web vulnerabilities. You can also configure conditions to block website crawlers.

Procedure

1.
2.
3. click **Configuration**.
4. On the **Monitoring Configuration** tab, click the **Web Monitoring** tab to view existing rules.



> **Note** **Default rule** is created by the system. You can only view details of the default rule, but cannot modify or delete it.

5. Create a web monitoring rule.
   i. Click **Add Rule**.
   ii. On the **Add Web Monitoring Rule** page, configure the following parameters.

Parameter	Description
Rule Name	The name of the web monitoring rule.
Monitoring Cycle	The monitoring cycle. Valid values: **Customization**, **Per Week**, and **Per Month**. ■ **Customization**: Specify the interval at which you want to perform detection. ■ **Per Week**: Specify the days of each week on which you want to perform detection. ■ **Per Month**: Specify the days of each month on which you want to perform detection.

Parameter	Description
Detection Time	The time when you want to perform detection. The time varies based on the value of the **Monitoring Cycle** parameter. ■ If you set the **Monitoring Cycle** parameter to **Customization**, select a time range of the day in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **Per Week**, select the days of each week and the time range in which you want to perform detection. ■ If you set the **Monitoring Cycle** parameter to **Per Month**, select the days of each month and the time range in which you want to perform detection. **Note** For example, if you set the **Monitoring Cycle** parameter to **Per Week** and select **Monday** to **Sunday** and 00:00:00 to 24:00:00 for the **Detection Time** parameter, Apsara Stack Security performs detection 24 hours a day, 7 days a week.
Monitoring Options	The types of the web vulnerabilities that you want to monitor. Supported operations: **Select All**, **Inverse**, and **Clear**.
UserAgent	The User-Agent field of an HTTP request packet. The User-Agent field identifies the application type, operating system, software developer, and version number of the proxy software that initiates the request.
Cookies	The cookie parameters.
Key Page	The web directories or pages that you want to monitor.
Excluded Page	The web directories or pages that you do not want to monitor.
Crawler Depth	The capturing depth of crawlers. Valid values: **10**, **15**, and **30**.
URL Numbers	The number of URLs that are used for crawling. Valid values: **500**, **1000**, and **2000**.
scanning Frequency	The scan frequency of web monitoring. Valid values: **Request 10 Times Per Second** and **Request 15 Times Per Second**.

iii. Click **Yes**.

6. Manage the web monitoring rule.

Icon	Description
	Modify the rule.
	Delete the rule.
Batch Delete	If you want to delete multiple rules, select the rules and click **Batch Delete**.

22.1.4.8.6.4. Configure a whitelist

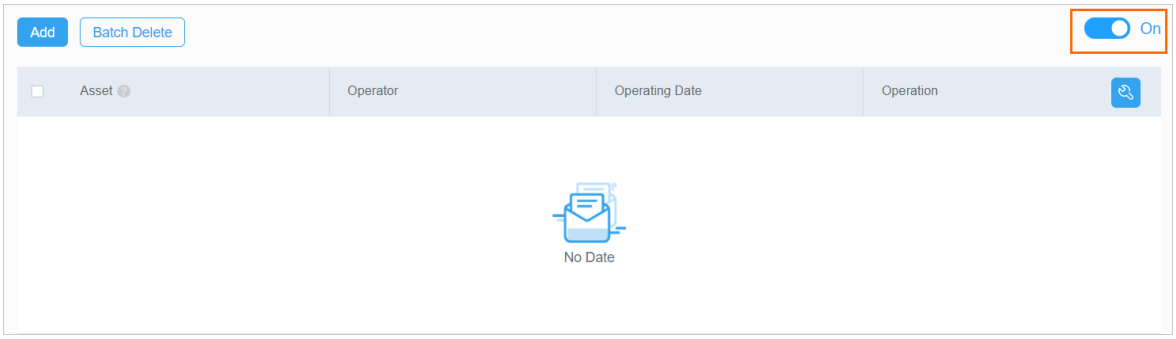
This topic describes how to configure a whitelist.

Context

Apsara Stack Security does not scan the assets that are added to a whitelist. Before you add assets to a whitelist, make sure that the assets are secure.

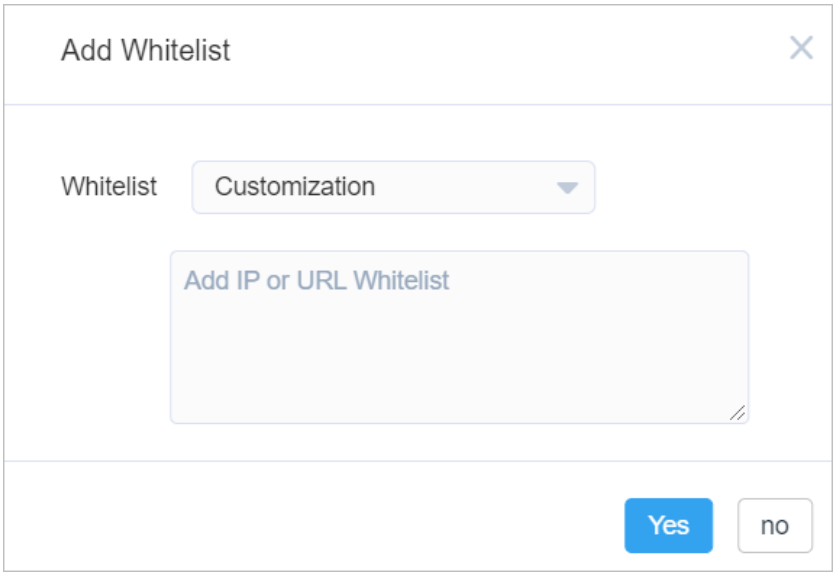
Procedure

1.
2.
3. click **Configuration**.
4. On the Configuration page, click the **Monitoring Configuration** tab. Then, click the **Whitelist** tab to view the assets that are added to a whitelist.



> **Note** By default, the whitelist feature is enabled. If you do not want to use the whitelist feature, turn off the switch in the upper-right corner.

5. Add assets to a whitelist.
   i. Click **Add**.
   ii. In the **Add Whitelist** dialog box, configure the parameters.



   - If you select **Asset Group** for the **Whitelist** parameter, select a group from the second drop-down list. The assets in this group are added to the whitelist.
   - If you select **Customization** for the **Whitelist** parameter, enter the IP addresses or URLs that you want to add to the whitelist in the field that appears.

   iii. Click **Yes**.
6. Manage the assets that are added to a whitelist.

- Remove an asset from a whitelist

  Find the asset and click the icon in the **Operation** column to delete the asset.

- Remove multiple assets from a whitelist at a time

  Select the assets and click **Batch Delete** to remove the assets from a whitelist at a time.

22.1.4.8.6.5. Configure a scan engine for internal assets

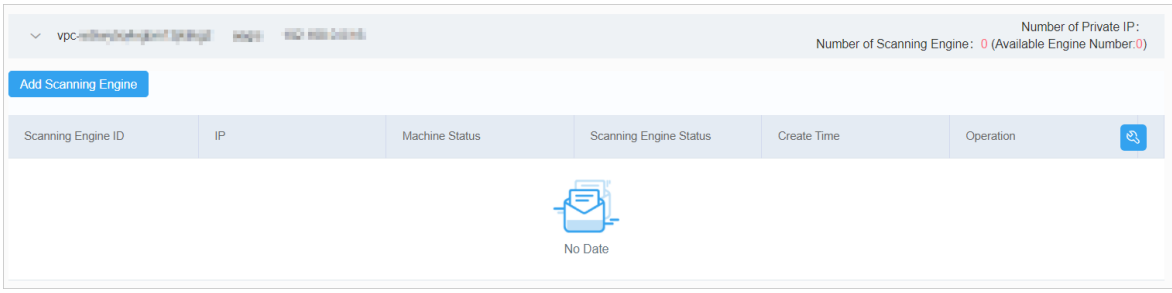
This topic describes how to configure a scan engine for internal assets, such as the assets of a virtual private cloud (VPC).

Context

You must add a scan engine for a VPC before you can scan for vulnerabilities on the assets of the VPC.

Procedure

1.
2.
3. click **Configuration**.
4. On the **Configuration** page, click the **Scan Engine Manage** tab.
5. Click the **Private-sector assets** tab.
6. Click the name of the VPC whose assets you want to scan.



7. Click **Add Scan Engine**.
8. In the **Add Scan Engine** dialog box, select a vSwitch for the VPC from the **vSwitch** drop-down list.
9. Click **OK**.

22.1.4.9. Create a security report

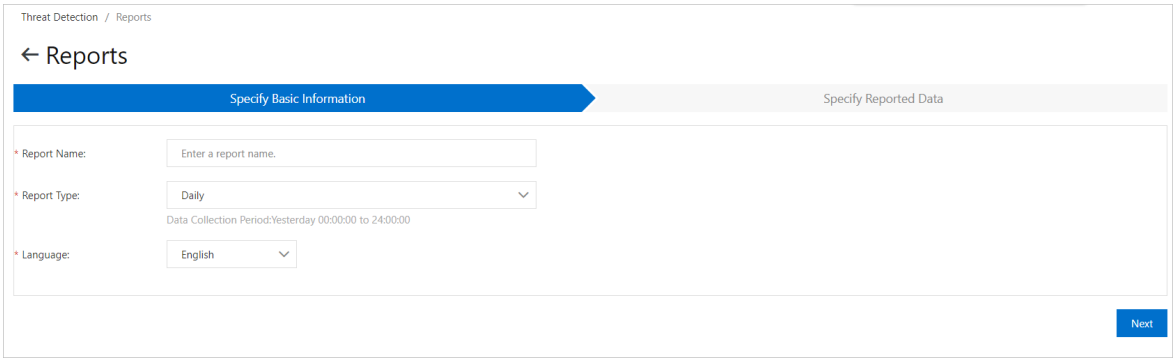
Security reports help monitor the security status of your assets. You can specify the content, types of statistics, and email addresses of recipients to create a security report. This topic describes how to create a security report.

Procedure

1.
2.
3. click **Security Reports**.
4. On the **Reports** page, click **Create Report**.

   Notice In addition to the default security report created by Apsara Stack Security, you can create a maximum of nine security reports.

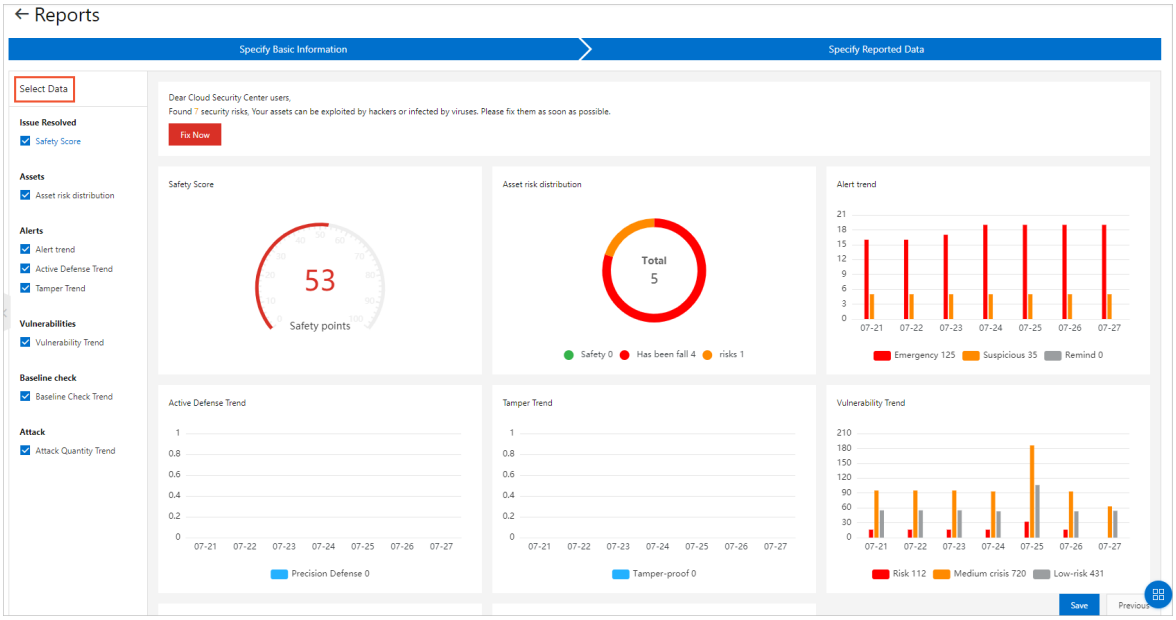
5. In the **Specify Basic Information** step, configure the parameters.



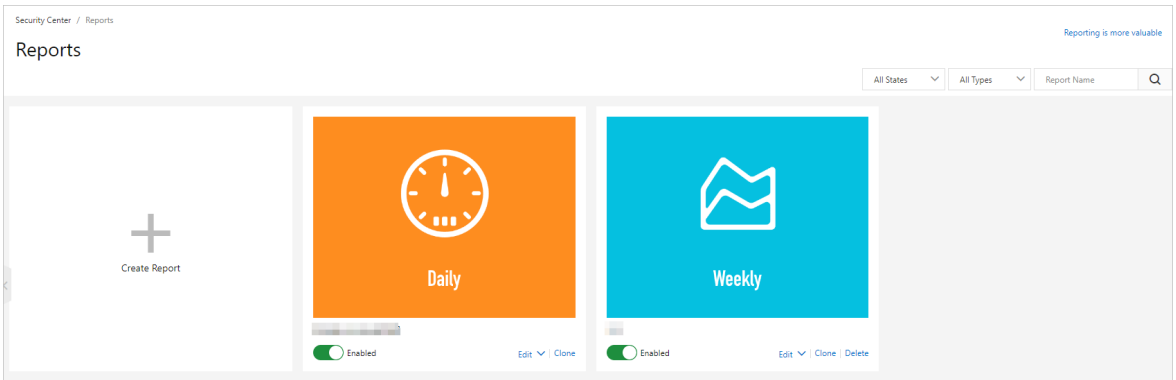
Configure the following parameters:

- **Report Name**: Enter a name for the security report.
- **Report Type**: Select a report type from the drop-down list. Valid values: *Daily*, *Weekly*, and *Monthly*.
- **Language**: Select a natural language for the report. Valid values: **简体中文** and **English**.

6. Click **Next**.
7. In the **Specify Reported Data** step, select the types of data that you want to view in the security report. You can select assets, alerts, vulnerabilities, baselines, attacks, and other data related to security operations.



8. Click **Save report content**. The security report is created.
You can view the newly created security report on the **Reports** page.



22.1.5. Network Traffic Monitoring System

22.1.5.1. View traffic trends

This topic describes how to view the network traffic trends and the statistics about inbound and outbound traffic.

Context

The security administrator can analyze traffic trends and obtain the traffic rate, peaks, and troughs. The security administrator can also view the top five IP addresses that have the largest volume of traffic and identify malicious IP addresses.

Procedure

1. Log on to Apsara Stack Security Center.
2. Choose **Network Security > Network Traffic Monitoring**.
3. In the upper-right corner of the **Traffic Trends** page, select the time range, which can be **Last 1 Hour**, **Last 24 Hours**, or **Last 7 Days**.
4. View network traffic information.
   - Network traffic trends

     View the network traffic trends in the time range that you selected. The network traffic trends include the trend of inbound traffic and the trend of outbound traffic. The inbound and outbound traffic is measured in bit/s.
   - Inbound Traffic

     View the information about Inbound Sessions, Inbound Applications, and Destination IPs with Most Requests.
   - Outbound Traffic

     View the information about Outbound Sessions, Outbound Applications, and Source IPs with Most Requests.
5. (Optional)Click the icon to export traffic trends as a PDF file.

22.1.5.2. View traffic at the Internet border

This topic describes how to view traffic at the Internet border. You can obtain up-to-date information about network security.

Prerequisites

The Network Traffic Monitoring System module is purchased and deployed at the egress (ISW) of Apsara Stack. **This module is used to audit, analyze, and manage both inbound and outbound traffic at Internet borders**.

Context

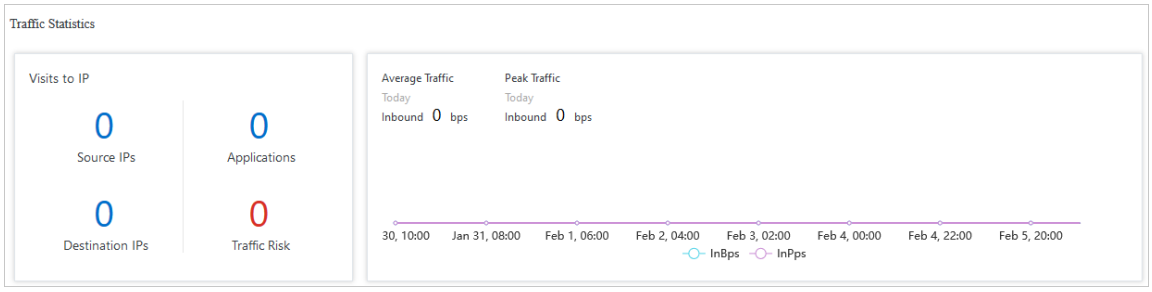
You can use traffic information to identify abnormal Internet traffic and block malicious requests.

Procedure

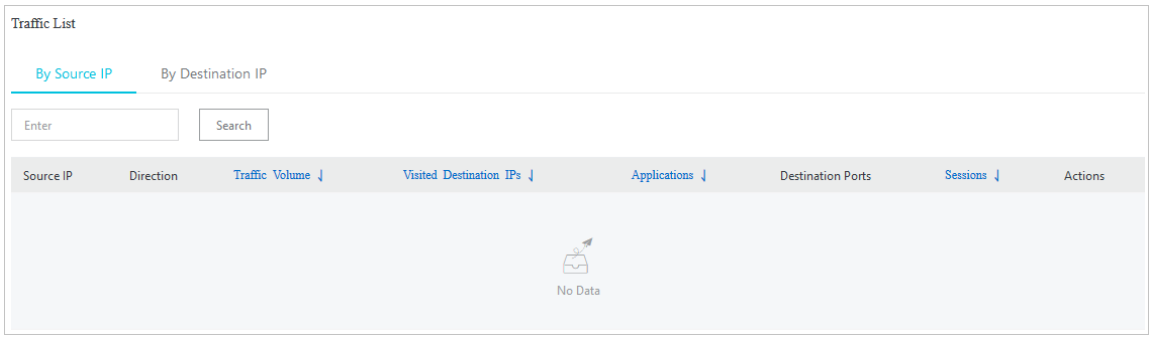
1.
2. Then, click **Network Traffic Monitoring System** in the **Network Security** section.
3. click **Internet Border**.
4. Specify traffic filter conditions.

Item	Description
1	Specify the traffic direction. Valid values: *Inbound* and *Outbound*. ○ *Inbound*: The traffic flows from the Internet to the internal network. ○ *Outbound*: The traffic flows from the internal network to the Internet.
2	Specify whether you want to view the traffic from the IP address or application dimension. Valid values: *By IP* and *By Application*.
3	Specify the time range. Valid values: *Last 1 Hour*, *Last 24 Hours*, and *Last 7 Days*.

5. View details about the traffic at the Internet border.
   - **Traffic Statistics**



     - The **Visits to IP** section includes **Source IPs**, **Destination IPs**, **Applications**, and **Traffic Risk**.
     - In the traffic chart on the right, you can view **Average Traffic**, **Peak Traffic**, and traffic trends.
   - **Traffic List**



     In the **Traffic List** section, you can view traffic details.
6. In the **Traffic List** section, view abnormal traffic of the specified IP address.
   - If *Inbound* is specified, you can view abnormal traffic on the **By Destination IP** tab of the **Traffic List** section.
   - If *Outbound* is specified, you can view abnormal traffic in the **Traffic List** section.

22.1.5.3. View traffic at the internal network border

This topic describes how to view the traffic at the internal network border. You can obtain up-to-date information about network security based on the traffic.

Prerequisites

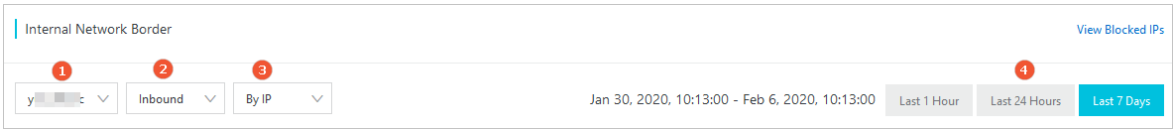
The Network Traffic Monitoring System module is purchased and deployed at the ingress (CSW) of Apsara Stack. **You can use this module to audit, analyze, and manage inbound and outbound traffic that is routed over leased lines between data centers and virtual private clouds (VPCs).**

Context

You can identify abnormal traffic from the internal network based on traffic information and block malicious requests.

Procedure

1. Log on to Apsara Stack Security Center.
2. Choose **Network Security > Network Traffic Monitoring**.
3. click **Internal Network Border**. On the **Internal Network Border** page, specify traffic filter conditions.



Item	Description
1	Select a VPC name from the drop-down list.
2	Specify the traffic direction. Valid values: *Inbound* and *Outbound*. ○ *Inbound*: The traffic flows from the Internet to the internal network. ○ *Outbound*: The traffic flows from the internal network to the Internet.
3	Specify whether you want to view the traffic that flows through the internal network border by IP address or application. Valid values: *By IP* and *By Application*.
4	Specify the time range. Valid values: *Last 1 Hour*, *Last 24 Hours*, and *Last 7 Days*.

4. View details about the traffic at the internal network border.
   - **Traffic Statistics**
     - The **Visits to IP** section includes **Source IPs**, **Destination IPs**, **Applications**, and **Traffic Risk**.
     - In the traffic chart on the right, you can view the average traffic, peak traffic, and traffic trends.
   - **Traffic List**
     - In the **Traffic List** section, you can select **By Source IP** or **By Destination IP** to view traffic details by source or destination IP address.
     - If **By IP** is specified, you can view the abnormal traffic of the specified IP address in the **Traffic List** section.

22.1.5.4. Create packet capture tasks

This topic describes how to create a packet capture task. You can enable the packet capture feature to capture network data packets of IP addresses and ports, and then analyze the packets. This way, you can diagnose faults, analyze attacks, and identify security risks to network communications.

Procedure

1.
2. Then, click **Network Traffic Monitoring System** in the **Network Security** section.
3. click **Packet Capture**.
4. On the **Packet Capture** page, click **Create Packet Capture Task**.
5. In the **Create Packet Capture Task** panel, configure parameters and click **OK**.

Parameter	Description
Task Name	The name of the packet capture task. We recommend that you enter an informative name, such as a name that indicates the purpose of the task.
Maximum Bytes	The maximum number of bytes that can be captured in a packet. If the number of bytes in a packet exceeds this value, the excessive bytes are discarded.
Duration (s)	The maximum duration of the packet capture task. Unit: seconds.
Protocol	The transmission protocol of packets. Valid values: ◦ **All** ◦ **TCP** ◦ **UDP** ◦ **ICMP**
IP Address Type	The IP protocol of packets. Valid values: **IPV4** and **IPV6**.
Direction	The direction of packets. Valid values: **Bidirectional**, **In**, and **Out**.
IP	The **IP address** of packets.
Port	The **port** of packets.

Result

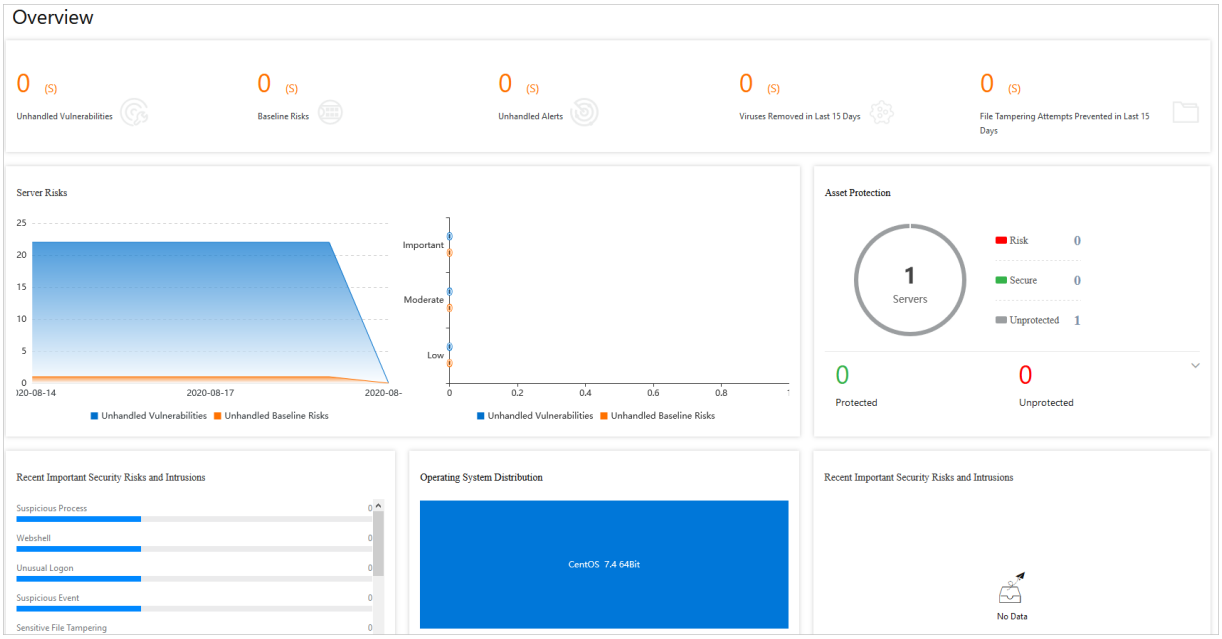
You can go to the **Packet Capture** page to view the packet capture task that you created and the status of the task.

22.1.6. Server security

22.1.6.1. Server security overview

This topic describes how to view the details about the security of servers on the server security overview page of Apsara Stack Security Center. This helps security administrators understand the security status of the servers. The servers refer to servers on the cloud.

To view the details about the security of servers, log on to Apsara Stack Security Center and choose **Server Security > Overview**. On the page that appears, you can view detailed information on the following sections: overall statistics, Server Risks, Asset Protection, Operating System Distribution, and Recent Important Security Risks and Intrusions.



- **Overall statistics**: This section displays the numbers of security vulnerabilities and security events on servers. For security vulnerabilities, you can view **Unhandled Vulnerabilities** and **Baseline Risks**. For security events, you can view **Unhandled Alerts**, **Viruses Removed in Last 15 Days**, and **File Tampering Attempts Prevented in Last 15 Days**.
- **Server Risks**: This section displays the number of unhandled vulnerabilities, the number of baseline risks, and the distribution of risk levels.
- **Asset Protection**: This section displays the number of protected servers and the number of offline servers.
- **Recent Important Security Risks and Intrusions**: This section displays the recent important risks and events on your servers. You can click a risk or an event to view the details.
- **Operating System Distribution**: This section displays your servers by operating system.

22.1.6.2. Server fingerprints

22.1.6.2.1. Manage listening ports

This topic describes how to view information about the listening port of a server. The information helps you identify suspicious listening behavior.

Context

This topic is suitable for the following scenarios:

- Check for servers that listen on a specific port.
- Check for ports that a specific server listens.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. On the **Asset Fingerprints** page, click the **Port** tab to view **listening ports**, **network protocols**, and server

information.

You can search for a port by using the port number, server process name, server name, or server IP address.

In the server information list, you can view the **process**, **IP address**, and **latest scan time** of a server.

22.1.6.2.2. Manage software versions

This topic describes how to regularly view and collect the software version information about a server. This helps you check your software assets.

Context

This topic covers the following scenarios:

- Check for software assets that are installed without authorization.
- Check for outdated software assets.
- Locate affected assets if vulnerabilities are detected.

Procedure

1. Log on to Apsara Stack Security Center.
2. Choose **Server Security > Sever Guard**.
3. In the left-side navigation pane, click **Server Fingerprints**.
4. On the page that appears, click the **Software** tab. On the tab, view all the **software assets** that are in use and the **number of the servers** that use the software assets.

   You can search for specific software by using its name, version, installation directory, server name, or IP address.
5. Click software to view the details, such as the software versions and the servers that use the software.

   You can click the icon in the upper-right corner to download a software version table to your computer for subsequent asset check.

22.1.6.2.3. Manage processes

This topic describes how to regularly collect the process information on a server and record changes. This way, you can view process information and historical process changes.

Context

This task is suitable for the following scenarios:

- Check for servers on which a specific process runs.
- Check for processes that run on a specific server.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. click **Server Fingerprints**.
4. On the page that appears, click the **Process** tab. On the tab, view all running processes and the number of servers that run these processes.

   You can search for a process by using the **process name**, **running user**, **startup parameter**, or **server name or IP address**.
5. Click the name of a process to view the details of the process, such as the servers, paths, and startup parameters.

22.1.6.2.4. Manage account information

This topic describes how to regularly collect the account information on a server and record the changes to the accounts. This way, you can check your accounts and view historical changes to your accounts.

Context

You can use the information collected in this topic for the following scenarios:

- Check for servers on which a specific account is created.
- Check for accounts that are created on a server.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. On the **Asset Fingerprints** page, click the **Account** tab.
5. View all the logged-on accounts and the numbers of servers on which the accounts are created.

   You can search for an account by using the account name, root permissions, server name, or server IP address.
6. Click an account name to view the details, such as the server information, root permissions, and user group.

22.1.6.2.5. Manage scheduled tasks

This topic describes how to view scheduled tasks on servers.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. On the **Asset Fingerprints** page, click the **Scheduled Tasks** tab.
5. View the paths of all tasks and the number of servers that run these tasks.

   You can search for a task by using the task path, server name, or IP address.
6. Click a task path to view the details, such as the servers, executed commands, and task cycles.

22.1.6.2.6. Set the fingerprint collection frequency

You can set the frequency at which the data of running processes, system accounts, listening ports, and software versions is collected.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. In the upper-right corner of the **Asset Fingerprints** page, click **Settings**.
5. Select the collection frequency from each drop-down list.
6. Click **OK** to complete the configuration.

22.1.6.3. Threat protection

22.1.6.3.1. Vulnerability management

22.1.6.3.1.1. Handle Linux software vulnerabilities

This topic describes how to handle Linux software vulnerabilities.

Context

Apsara Stack Security automatically scans the software that is installed on your servers to detect the vulnerabilities provided in the Common Vulnerabilities and Exposures (CVE) list. Apsara Stack Security also sends you alerts about the detected vulnerabilities. In addition, Apsara Stack Security provides commands that you can use to fix vulnerabilities and allows you to verify vulnerability fixes.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. In the left-side navigation pane, click **Vulnerabilities**.
4. On the **Linux Software** tab of the page that appears, view the detected Linux software vulnerabilities.

   > **Note** You can search for a specific vulnerability by using the search and filter features.

5. Find a vulnerability and click its name. In the panel that appears, you can view the details about the vulnerability and the servers that are affected by the vulnerability.

   > **Note** You can find affected servers by using the search and filter features.

   - **Detail**: This tab displays the basic information about the vulnerability, including the name, Common Vulnerability Scoring System (CVSS) score, description, and solution.
   - **Pending vulnerability**: This tab displays the servers that are affected by the vulnerability.
6. Handle the vulnerability based on its impact.

   Actions on vulnerabilities

Action	Description
Generate Fix Command	Select this action to generate the commands that are used to fix the vulnerability. You can then log on to the server to run these commands.
Fix	Select this action to fix the vulnerability.
Restarted and Verified	If a vulnerability fix takes effect only after a server restart, you must restart the server after the status of the vulnerability changes to **Fixed (To Be Restarted)**. After the restart, click **Restarted and Verified**.
Ignore	Select this action to ignore the vulnerability. The system no longer generates alerts for or reports ignored vulnerabilities.
Verify	Click **Verify** to verify the vulnerability fix. If you do not manually verify a fix, the system automatically verifies the fix within 48 hours after the vulnerability is fixed.

   You can handle a vulnerability for one or more affected servers at a time.

   - To handle a vulnerability for one affected server, find the server and select an action in the **Actions** column of the server.

- To handle a vulnerability for multiple affected servers, select the servers and select an action in the lower-left corner.

22.1.6.3.1.2. Handle Windows system vulnerabilities

This topic describes how to handle Windows system vulnerabilities.

Context

Apsara Stack Security automatically checks whether the latest Microsoft updates are installed on your servers, and notifies you of the detected vulnerabilities. Apsara Stack Security also automatically detects and fixes major vulnerabilities on your servers.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. In the left-side navigation pane, click **Vulnerabilities**. On the page that appears, click the **Windows System** tab.
4. View the detected Windows system vulnerabilities.

   > **Note** You can search for a specific vulnerability by using the search and filter features.

5. Find a vulnerability and click its name. In the panel that appears, you can view details about the vulnerability and the servers that are affected by the vulnerability.

   > **Note** You can find affected servers by using the search and filter features.

   - **Detail**: This tab displays the basic information about the vulnerability, including the name, Common Vulnerability Scoring System (CVSS) score, description, and solution.
   - **Pending vulnerability**: This tab displays the servers that are affected by the vulnerability.
6. Handle the vulnerability based on its impact. Actions on vulnerabilities describes the actions.

   Actions on vulnerabilities

Action	Description
Fix	Select this action to fix the vulnerability. The system caches an official Windows patch in the cloud. Your server can automatically download the patch for updates.
Ignore	Select this action to ignore the vulnerability. The system no longer generates alerts for or reports ignored vulnerabilities.
Verify	Click **Verify** to verify the vulnerability fix.
Restarted and Verified	If a vulnerability fix takes effect only after a server restart, you must restart the server after the status of the vulnerability changes to **Fixed (To Be Restarted)**. After the restart, click **Restarted and Verified**.

   You can handle a vulnerability for one or more affected servers at a time.

   - To handle a vulnerability for one affected server, find the server and select an action in the **Actions** column of the server.
   - To handle a vulnerability for multiple affected servers, select the servers and select an action in the lower-left corner.

22.1.6.3.1.3. Handle Web-CMS vulnerabilities

This topic describes how to handle Web-CMS vulnerabilities.

Context

The feature of Web-CMS vulnerability detection obtains information about the latest vulnerabilities and provides patches in the cloud. This helps you detect and fix vulnerabilities.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. In the left-side navigation pane, click **Vulnerabilities**. On the page that appears, click the **Web CMS** tab.
4. View all detected vulnerabilities.

   > **Note** You can search for a specific vulnerability by using the search and filter features.

5. Find a vulnerability and click its name. In the panel that appears, you can view details about the vulnerability and the servers that are affected by the vulnerability.

   > **Note** You can find affected servers by using the search and filter features.

6. Handle the vulnerability based on its impact. Actions on vulnerabilities describes the actions.

   Actions on vulnerabilities

Action	Description
Fix	If you select this action, the system replaces the web files that are affected by the vulnerability on your server to fix the Web-CMS vulnerability. **Note** Before you fix the vulnerability, we recommend that you back up the web files affected by the vulnerability. For more information about the paths of the web files, click Details in the Actions column.
Ignore	Select this action to ignore the vulnerability. The system no longer generates alerts for or reports ignored vulnerabilities.
Verify	After a vulnerability is fixed, you can click **Verify** to verify the vulnerability fix. If you do not manually verify the fix of a vulnerability, the system automatically verifies the fix within 48 hours after the vulnerability is fixed.
Undo Fix	For vulnerabilities that have been fixed, click **Undo Fix** to restore the web files that have been replaced.

   You can handle a vulnerability for one or more affected servers at one time.

   - To handle a vulnerability for one affected server, select an action in the **Actions** column of the server.
   - To handle a vulnerability for multiple affected servers, select the servers and select an action in the lower-left corner.

22.1.6.3.1.4. Handle urgent vulnerabilities

This topic describes how to handle urgent vulnerabilities.

Context

Apsara Stack Security automatically detects vulnerabilities on servers, such as the unauthorized Redis access vulnerability and Struts S2-052 vulnerability, and generates alerts for detected vulnerabilities. After you fix a vulnerability, you can also check whether the fix is successful.

Procedure

1.
2.
3. click **Vulnerabilities**.
4. On the **Vulnerabilities** page, click the **Emergency** tab.
5. View all vulnerabilities.

   You can search for a specific vulnerability by using the search and filter features.
6. Click the name of a vulnerability. In the panel that appears, view the details in the following sections: **Details**, **Suggestions**, and **Affected Assets**.

   You can find affected assets by using the search and filter features.
7. Handle the vulnerability based on its impact. Actions on vulnerabilities describes the actions.

   Follow the instructions to fix the vulnerabilities on the **Emergency** tab.

   Actions on vulnerabilities

Action	Description
Ignore	Select this action to ignore the vulnerability. The system no longer generates alerts for or reports ignored vulnerabilities.
Verify	Click **Verify** to verify the vulnerability fix. If you do not manually verify a fix, the system automatically verifies the fix within 48 hours after the vulnerability is fixed.

   You can handle a vulnerability for one or more affected assets at a time.

   - To handle a vulnerability for one affected asset, find the asset and select an action in the **Actions** column of the asset.
   - To handle a vulnerability for multiple affected assets, select the assets and select an action in the lower-left corner.

22.1.6.3.1.5. Configure vulnerability handling policies

You can enable or disable automatic detection for different types of vulnerabilities and enable vulnerability detection for specific servers. You can also set a duration during which invalid vulnerabilities are retained and configure a vulnerability whitelist.

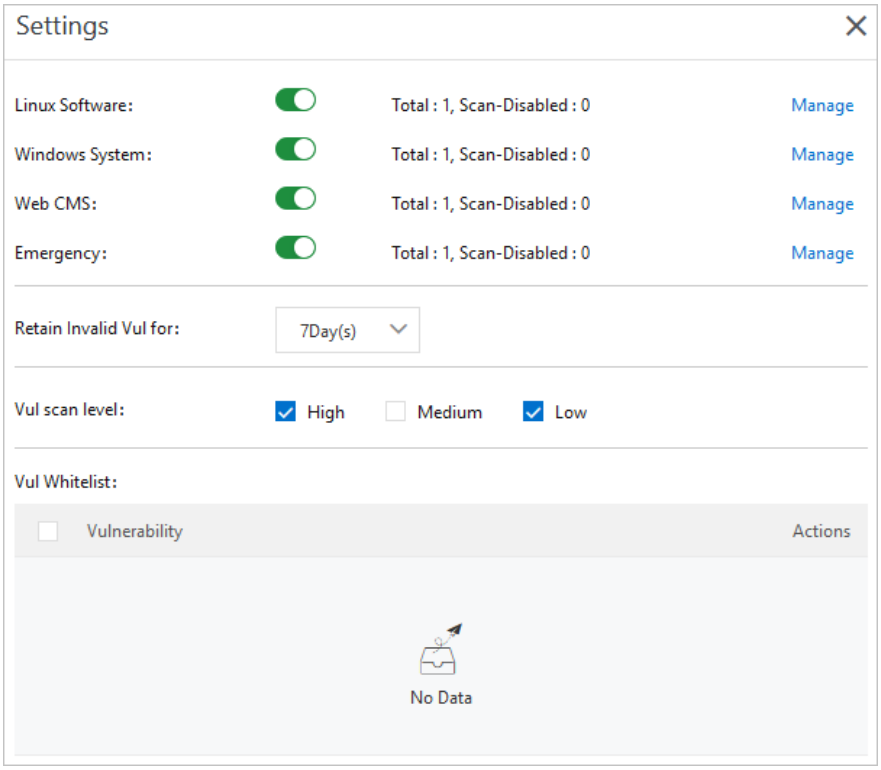
Context

A vulnerability whitelist allows you to ignore specific vulnerabilities. You can add multiple vulnerabilities in the vulnerability list to the whitelist. The system does not detect vulnerabilities that are added to the whitelist. You can configure the vulnerability whitelist on the vulnerability settings page.

Procedure

1.
2.
3. click **Vulnerabilities**.

4. In the upper-right corner of the page that appears, click **Settings** to configure vulnerability handling policies. In the panel that appears, perform the following operations:



   - Find a vulnerability type and enable or disable detection for vulnerabilities of this type.
   - Click **Manage** next to a vulnerability type and specify the servers on which vulnerabilities of this type are detected.
   - Select a time duration during which invalid vulnerabilities are retained for Retain Invalid Vul for. Valid values: 7Day(s), 30Day(s), and 90Day(s).

     > **Note** If you do not take an action on a detected vulnerability, the system determines that the alert, which indicates that a vulnerability is detected, is invalid. The system deletes the vulnerability after the specified duration.

   - Select the vulnerability severities for scanning for Vul scan level. Valid values:
     - **High**: You must fix the vulnerabilities of this severity at the earliest opportunity.
     - **Medium**: You can fix the vulnerabilities of this severity later.
     - **Low**: You can ignore the vulnerabilities of this severity for now.
   - Select vulnerabilities in the Vul Whitelist section and click **Remove** in the Actions column to enable the system to detect these vulnerabilities and generate alerts for these vulnerabilities.

22.1.6.3.2. Baseline check

22.1.6.3.2.1. Baseline check overview

The baseline check feature automatically checks the security configurations on servers and provides detailed check results and suggestions for baseline reinforcement.

Description

After you enable the baseline check feature, Apsara Stack Security automatically checks for risks related to the operating systems, accounts, databases, passwords, and security compliance configurations of your servers, and provides reinforcement suggestions. For more information, see Baselines.

By default, a full baseline check is automatically performed from 00:00 to 06:00 every day. You can create and manage scan policies for baseline checks. When you create or modify a policy, you can specify the baselines, interval, and time period, and select the servers to which you want to apply this policy. For more information, see Add a custom baseline check policy.

Precautions

By default, the following baselines are disabled. To check these baselines, make sure that these baselines do not affect your business and select them when you customize a scan policy.

- Baselines related to weak passwords for specific applications such as MySQL, PostgreSQL, and SQL Server

  > **Note** If these baselines are enabled, the system attempts to log on to servers with weak passwords. The logon attempts consume server resources and generate a large number of logon failure records.

- Baselines related to China classified protection of cybersecurity
- Baselines related to the Center for Internet Security (CIS) standard

Baselines

Category	Baseline
High risk exploit	• High risk exploit - CouchDB unauthorized access high exploit risk • High risk exploit - Docker unauthorized access high vulnerability risk • High risk exploit - Elasticsearch unauthorized access high exploit vulnerability risk • High risk exploit - Memcached unauthorized access high exploit vulnerability risk • High risk exploit - Apache Tomcat AJP File Read/Inclusion Vulnerability • High risk exploit - ZooKeeper unauthorized access high exploit vulnerability risk

Category	Baseline
	Security baseline check against the Alibaba Cloud standard: • Alibaba Cloud Standard-Aliyun Linux 2 Security Baseline Check • Alibaba Cloud Standard - CentOS Linux 6 Security Baseline Check • Alibaba Cloud Standard - CentOS Linux 7 Security Baseline Check • Alibaba Cloud Standard - Debian Linux 8 Security Baseline • Alibaba Cloud Standard - Redhat Linux 6 Security Baseline Check • Alibaba Cloud Standard - Redhat Linux 7 Security Baseline Check • Alibaba Cloud Standard - Ubuntu Security Baseline Check • Alibaba Cloud Standard - Windows Server 2008 R2 Security Baseline Check • Alibaba Cloud Standard - Windows 2012 R2 Security Baseline • Alibaba Cloud Standard - Windows 2016/2019 R2 Security Baseline
	Security baseline check against the CIS standard: • Alibaba Cloud Aliyun Linux 2 CIS Benchmark • CIS CentOS Linux 6 LTS Benchmark • CIS CentOS Linux 7 LTS Benchmark • CIS Debian Linux 8 Benchmark • CIS Ubuntu Linux 14 LTS Benchmark • CIS Ubuntu Linux 16/18 LTS Benchmark • CIS Microsoft Windows Server 2008 R2 Benchmark • CIS Microsoft Windows Server 2012 R2 Benchmark • CIS Microsoft Windows Server 2016/2019 R2 Benchmark

Category	Baseline
CIS and China's Protection of Cybersecurity	Baseline check on compliance of China classified protection of cybersecurity level III: • Aliyun Linux 2 Baseline for China classified protection of cybersecurity-Level III • CentOS Linux 6 Baseline for China classified protection of cybersecurity-Level III • CentOS Linux 7 Baseline for China classified protection of cybersecurity-Level III • Debian Linux 8 Baseline for China classified protection of cybersecurity-Level III • Redhat Linux 6 Baseline for China classified protection of cybersecurity-Level III • Redhat Linux 7 Baseline for China classified protection of cybersecurity-Level III • SUSE Linux 10 Baseline for China classified protection of cybersecurity-Level III • SUSE Linux 11 Baseline for China classified protection of cybersecurity-Level III • SUSE Linux 12 Baseline for China classified protection of cybersecurity-Level III • Ubuntu 14 Baseline for China classified protection of cybersecurity-Level III • Waiting for Level 3-Ubuntu 16/18 compliance regulations inspection • China's Level 3 Protection of Cybersecurity - Windows Server 2008 R2 Compliance Baseline Check • Windows 2012 R2 Baseline for China classified protection of cybersecurity-Level III • Windows 2016/2019 R2 Baseline for China classified protection of cybersecurity-Level III

Category	Baseline
Best security practices	• Alibaba Cloud Standard-Aliyun Linux 2 Security Baseline Check • Alibaba Cloud Standard - Apache Security Baseline Check • Alibaba Cloud Standard - CentOS Linux 6 Security Baseline Check • Alibaba Cloud Standard - CentOS Linux 7/8 Security Baseline Check • Alibaba Cloud Standard - Debian Linux 8 Security Baseline • Alibaba Cloud Standard - IIS 8 Security Baseline Check • Alibaba Cloud Standard - Memcached Security Baseline Check • Alibaba Cloud Standard - MongoDB 3.x Security Baseline Check • Alibaba Cloud Standard - Mysql Security Baseline Check • Alibaba Cloud Standard - Nginx Security Baseline Check • Alibaba Cloud Standard - Redhat Linux 6 Security Baseline Check • Alibaba Cloud Standard - Redhat Linux 7 Security Baseline Check • Alibaba Cloud Standard - Redis Security Baseline Check • Alibaba Cloud Standard - Ubuntu Security Baseline Check • Alibaba Cloud Standard - Windows Server 2008 R2 Security Baseline Check • Alibaba Cloud Standard - Windows 2012 R2 Security Baseline • Alibaba Cloud Standard - Windows 2016/2019 R2 Security Baseline • Alibaba Cloud Standard-Apache Tomcat Security Baseline

Category	Baseline
Weak password	• Weak Password-MongoDB Weak Password baseline(support version 2. X) • Weak password - Ftp login weak password baseline • Weak password - Linux system login weak password baseline • Weak password - MongoDB login weak password baseline • Weak password - SQL Server DB login weak password baseline • Weak password - Mysql DB login weak password baseline • Weak password - Mysql DB login weak password baseline(Windows version) • Weak password - PostgreSQL DB login weak password baseline • Weak password - Redis DB login weak password baseline • Weak password - rsync login weak password baseline • Weak password - svn login weak password baseline

22.1.6.3.2.2. Configure baseline check policies

This topic describes how to create, modify, and delete baseline check policies. This topic also describes how to specify baseline check levels.

Context

By default, the baseline check feature uses the **default policy** to check the baseline risks of assets. You can also customize baseline check policies based on your business requirements. For example, you can customize a baseline check policy to check the compliance with classified protection requirements (MLPS level 2).

Procedure

1.
2.
3. click **Baseline Check**.
4. In the upper-right corner of the page that appears, click **Manage Policies**. In the **Manage Policies** panel, create, modify, or delete a baseline check policy. You can also modify the default policy.
   - In the upper-right corner of the panel, click **+Create Policy** to customize a baseline check policy. Then, click **Ok**.

Parameter	Description
Policy Name	Enter a policy name.
Schedule	Set the time interval for scheduled scan tasks to Every 1 Day, Every 3 Day, Every 7 Day, or Every 30 Day. Then, select one of the following time ranges for scheduled scan tasks: 00:00 to 06:00, 06:00 to 12:00, 12:00 to 18:00, and 18:00 to 24:00.

Parameter	Description
Check Items	Select the baseline items that need to be checked from the following categories: High risk exploit, Container security, CIS and China's Protection of Cybersecurity, Best security practices, and Weak password.
Servers	Select the server groups to which you want to apply the baseline check policy. **Note** By default, newly purchased servers are added to the **Default** group under **Asset Groups**. To apply this policy to the newly purchased servers, select **Default**.

- Click **Edit** or **Delete** next to the created policy to modify or delete it.

  Note You cannot restore a policy after you delete it.

- Find the **Default** policy and click **Edit** in the **Actions** column to modify the server groups to which the default policy is applied.

  Note You cannot delete the default policy or modify the baseline items of the default policy. You can only modify the server groups to which the default policy is applied.

- In the lower part of the **Manage Policies** panel, specify the baseline check levels. Valid values: High, Medium, and Low.

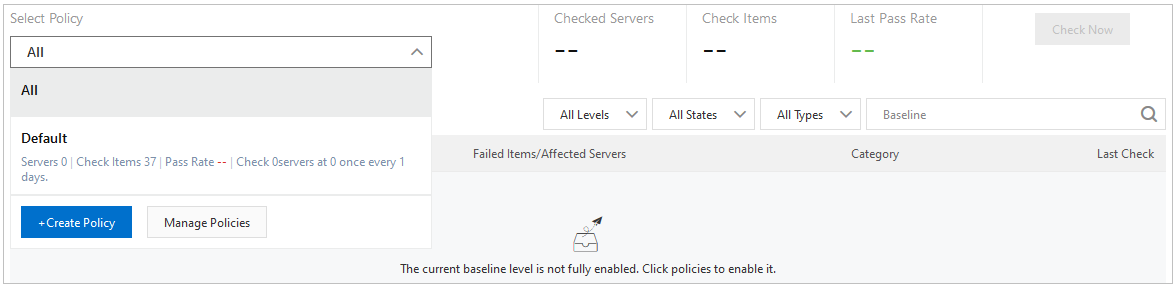
5. Click **Ok**.

22.1.6.3.2.3. View baseline check results and handle baseline risks

Apsara Stack Security Center provides detailed baseline check results and suggestions on how to handle baseline risks. This topic describes how to view baseline check results and handle baseline risks in Apsara Stack Security Center. The check results include information about affected assets, details of check items, and suggestions on how to handle baseline risks.

View the summary of baseline check results

1. Log on to Apsara Stack Security Center.
2.
3. click **Baseline Check**.
4. In the upper part of the **Baseline Check** page, view the summary of baseline check results. You can filter data by policy.



You can select a policy from the **Select Policy** drop-down list to view the following information:

- **Checked Servers**: The number of servers on which the baseline check runs. These servers are specified in the selected baseline check policy.
- **Check Items**: The number of **check items** specified in the selected baseline check policy.
- **Last Pass Rate**: The pass rate of the check items in the last baseline check.

  If the number below **Last Pass Rate** is green, the pass rate is high. If the number is red, a large number of baseline risks have been detected on the checked servers. We recommend that you go to the details page to view and handle the baseline risks.

View all baselines

1. Log on to Apsara Stack Security Center.
2.
3. click **Baseline Check**.
4. Select **All** from the **Select Policy** drop-down list.
   The **Baseline Check** page displays details about all baselines, including **Baseline**, **Checked Item**, **Failed Items/Affected Servers**, **Category**, and **Last Check**.

   > **Note** You can also select a baseline check policy from the **Select Policy** drop-down list to view the baselines specified in this policy.

View details about a baseline

1. Log on to Apsara Stack Security Center.
2.
3. click **Baseline Check**.
4. In the **Baseline** column, click a baseline to view its details.
5. In the details panel, handle the baseline risks.
   - Find the asset that you want to handle and click **View** in the **Actions** column to open the **At-Risk Items** panel.
   -

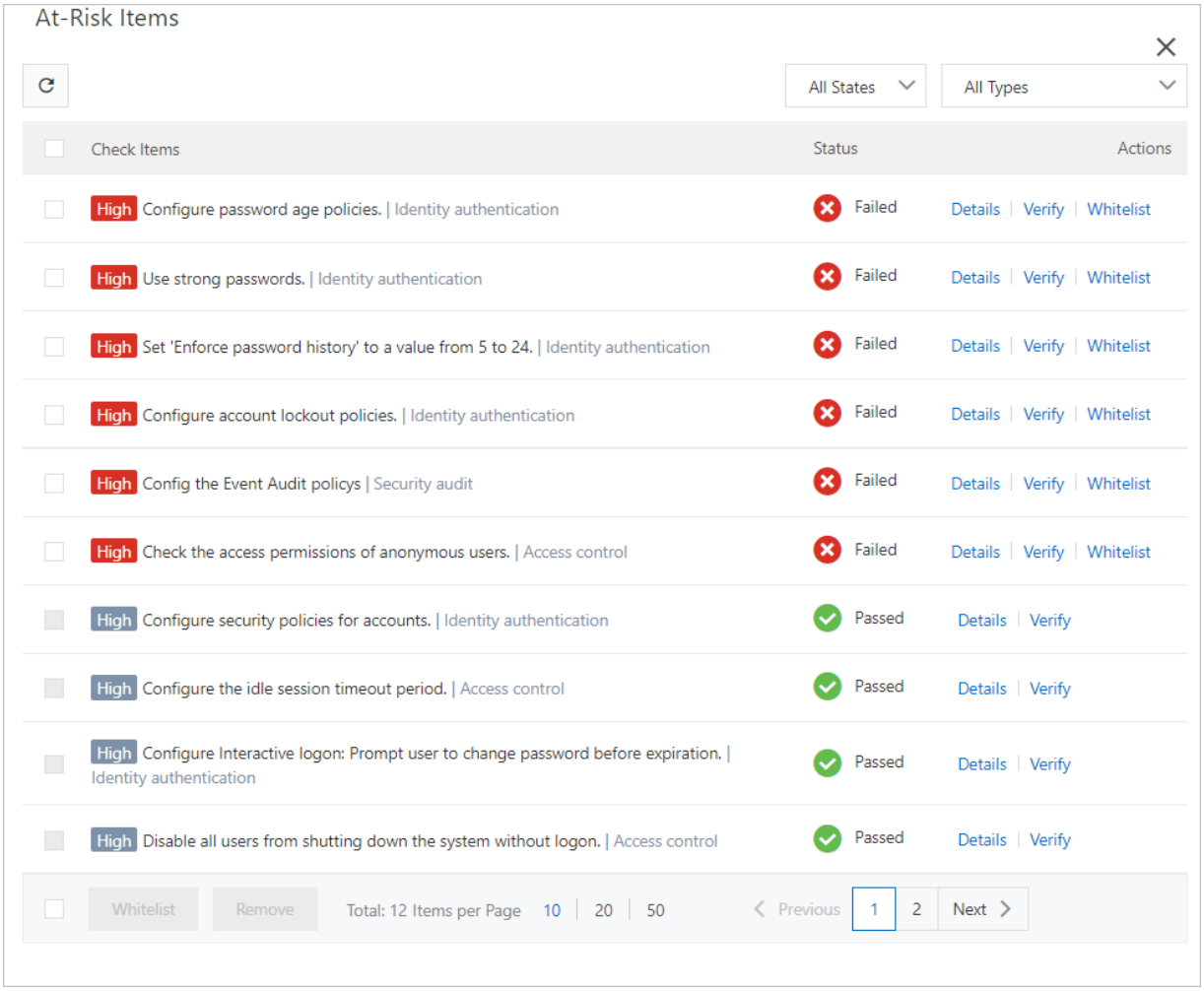
View details about a baseline risk

1. Find the baseline that you want to handle and click it. In the panel that appears, find the asset that you want to handle and click **View** in the **Actions** column to view the details about the risk items.
   You can view the check items of the asset and the statuses of the check items. The status can be **Passed** or **Failed**.
2.

> **Note** We recommend that you follow the suggestions to handle risk items whose status is **Failed** at the earliest opportunity, especially high-risk items.

Handle baseline risks

In the At-Risk Items panel, handle baseline risks.



- **Add check items to the whitelist**

  If you want to disable alerts for a check item, click **Whitelist** to add the check item to the whitelist. Check items in the whitelist do not trigger alerts.

  > **Note** You can also select multiple check items and click **Whitelist** in the lower-left corner to add the check items to the whitelist at a time.

- **Fix risks**

  You can fix only the baseline risks that are detected based on the Alibaba Cloud standard at a time. You can select multiple servers on which the same baseline risk is detected and fix the risk.

  > **Notice** Risk fixing may cause service interruptions. We recommend that you back up your service data before risk fixing.

- **Remove check items from the whitelist**

  If you want to enable alerts for a check item in the whitelist, you can click **Remove** to remove the check item from the whitelist. You can remove one or more check items from the whitelist at a time. After a check item is removed from the whitelist, the check item triggers alerts again.

- **Verify the fix of a baseline risk**

  If you do not manually perform the verification, Apsara Stack Security automatically verifies the fix based on the detection interval specified in the baseline check policy.

22.1.6.4. Intrusion prevention

22.1.6.4.1. Intrusion events

22.1.6.4.1.1. Intrusion event types

If Server Guard detects sensitive file tampering, suspicious processes, webshells, unusual logons, or malicious processes, it generates alerts. Based on these alerts, you can monitor the security status of your assets and handle potential threats at the earliest opportunity.

Apsara Stack Security provides statistics on enabled alerts and defense items. These statistics help you monitor the overall security of your assets. You can view the statistics on the **Intrusions** page.

Alerts

The following table describes the alerts.

Alert	Description
Threat intelligence	Identify potential threats to your assets based on the threat intelligence of Apsara Stack Security. Threat intelligence can correlate threat information to analyze and process the information. If threats are detected, threat intelligence can generate alerts. This helps improve the detection efficiency and response speed. Threat intelligence can detect the following items: • Malicious domain names • Malicious IP addresses • IP addresses of dark web services • IP addresses of command and control (C&C) servers • IP addresses of mining pools • Malicious URLs • Malicious download sources
Unusual Logon	Detect unusual logons to your servers. You can specify approved logon IP addresses, time periods, and accounts. Logons from unapproved IP addresses,time periods, or accounts trigger alerts. You can manually add approved logon locations or configure the system to automatically update approved logon locations. You can also specify assets on which alerts are triggered when unapproved logon locations are detected. Server Guard can detect the following events: • Logons to Elastic Compute Service (ECS) instances from unapproved IP addresses • Logons to ECS instances from unapproved locations • Execution of unusual commands after SSH-based logons to ECS instances • Brute-force attacks on SSH passwords of ECS instances
Webshell	Use engines developed by Alibaba Cloud to scan common webshell files. Server Guard supports scheduled scan tasks, provides real-time protection, and quarantines webshell files. • Server Guard scans the entire web directory early in the morning on a daily basis. A change made to files in the web directory triggers dynamic detection. • You can specify the assets on which Server Guard scans for webshells. • You can quarantine or ignore detected trojan files. You can also restore the quarantined trojan files.

Alert	Description
Precision defense	The **antivirus** feature provides precise protection from common ransomware, DDoS trojans, mining programs, trojans, malicious programs, webshells, and computer worms.
Suspicious Account	Detect logons to your assets from unapproved accounts.
Cloud threat detection	Detect threats in other cloud services.
Persistence	Detect suspicious scheduled tasks on servers and generate alerts when advanced persistent threats (APTs) to the servers are detected.
Unusual Network Connection	Detect disconnections or unusual network connections.
Suspicious Process	Detect whether suspicious processes exist.
Malicious Process	Scan your servers in real time. An agent is used to collect process information, and the information is uploaded to the cloud for detection. If viruses are detected, alerts are generated. You can handle detected viruses in Apsara Stack Security Center. Server Guard can detect the following malicious activities and processes: • Access to malicious IP addresses • Mining programs • Self-mutating trojans • Malicious programs • Trojans
Sensitive File Tampering	Check whether sensitive files on your servers are maliciously modified. The sensitive files include preloaded configuration files in Linux shared libraries.
Other	Detect other types of attacks, such as DDoS attacks.
Web Application Threat Detection	Detect intrusions that use web applications.
Application intrusion event	Detect intrusions that use system application components.

22.1.6.4.1.2. View and handle alert events

This topic describes how to view and handle detected alert events on the Intrusions page.

Background information

After alert events are detected, the alerts events are displayed on the **Intrusions** page in Apsara Stack Security Center. If the detected alert events are not handled, they are displayed in the **Unhandled Alerts** list on the **Intrusions** page. After the alert events are handled, the status of the alert events changes from **Unhandled Alerts** to **Handled**.

> **Note** Apsara Stack Security Center retains the records of **Unhandled Alerts** and **Handled** on the **Intrusions** page. By default, the records of **Unhandled Alerts** are displayed.

View alert events

1.
2.

3. click **Intrusions**.
4. On the page that appears, search for or view all alert events. You can also view the details about the alert events.

Handle alert events

1.
2.
3. click **Intrusions**.
4. On the **Intrusions** page, find the alert event that you want to handle and click **Handle** in the **Actions** column. In the dialog box that appears, configure Process Method and click **Process Now**.

   > **Note** If the alert event is related to multiple exceptions, the panel that shows alert event details appears after you click **Handle**. You can handle the exceptions in the panel.

   - **Ignore**: If you ignore the alert event, the status of the alert event changes to **Handled**. Server Guard no longer generates alerts for the event.
   - **Add To Whitelist**: If the alert event is a false positive, you can add the alert event to the whitelist. Then, the status of the alert event changes to **Handled**. Server Guard no longer generates alerts for the event. In the **Handled** list, you can click **Cancel whitelist** to remove the alert event from the whitelist.

     > **Note** When Server Guard generates a false alert on a normal process, this alert is considered a false positive. A common false positive is a **suspicious process that sends TCP packets**. The false positive notifies you that suspicious scans on other devices are detected on your servers.

   - **Batch unhandled**: This method allows you to batch handle multiple alert events. Before you batch handle multiple alert events, we recommend that you view the details about the alert events.
5. (Optional)If you confirm that one or more alert events are false positives or need to be ignored, go to the **Intrusions** page. Then, select the alert events and click **Ignore Once** or **Whitelist**.

Export alert events

1.
2.
3. click **Intrusions**.
4. In the upper-left corner above the alert event list on the **Intrusions** page, click the  icon to export the list.

   After the list is exported, the **Done** message appears in the upper-right corner of the Intrusions page.
5. In the **Done** notification of the **Alerts** page, click **Download**.
   The alert list is downloaded to your computer.

22.1.6.4.1.3. View exceptions related to an alert

Server Guard supports automatic analysis of exceptions related to an alert. You can click an alert name in the alert list to view and handle all exceptions that are related to the alert. You can also view the results of automatic attack tracing to analyze the exceptions.

Context

- Security Center automatically associates alerts with exceptions in real time to detect potential threats.
- Exceptions related to an alert are listed in chronological order. This allows you to analyze and handle the exceptions to improve the emergency response mechanism of your system.

- An automatically correlated alert is identified by the icon.

Procedure

1.
2.
3. click **Intrusions**.
4. On the Intrusions page, click the **name of the alert** that you want to handle. The alert details panel appears.
5. In the alert details panel, view the details and related exceptions of the alert. Then, handle the exceptions.
   - View alert details

     You can view the assets that are affected by the alert, the first and latest time when the alert was triggered, and the details about the related exceptions.
   - View affected assets

     You can move the pointer over the name of an **affected asset** to view the details about the asset. The details include information about all the alerts, vulnerabilities, baseline risks, and asset fingerprints on the asset.
   - View and handle **related exceptions**

     In the **Related Exceptions** section, you can view the details about all the exceptions that are related to the alert. You can also view suggestions on how to handle the exceptions.
     - Click **Note** to the right of an exception to add a note for the exception.
     - Click the icon to the right of a note to delete the note.

22.1.6.4.1.4. Use the file quarantine feature

Sever Guard can quarantine malicious files. Quarantined files are listed in the Quarantine panel of the Intrusions page. You can restore a quarantined file with a few clicks. However, 30 days after a file is quarantined, the system automatically deletes the file. This topic describes how to view and restore quarantined files.

Procedure

1.
2.
3. click **Intrusions**.
4. In the upper-right corner of the **Intrusions** page, click **Quarantine**.

   In the **Quarantine** panel, you can perform the following operations:
   - View information about quarantined files. The information includes server IP addresses, directories in which the files are stored, file status, and modification time.
   - Click **Restore** in the **Actions** column to restore a quarantined file. The restored file appears in the alert list.

22.1.6.4.1.5. Configure alerts

This topic describes how to configure alerts. You can specify approved logon locations and customize web directories to scan.

Context

Server Guard supports advanced logon settings. You can configure more fine-grained logon detection rules. For example, you can specify approved logon IP addresses, logon time ranges, and logon accounts to block

unauthorized requests that are sent to your assets.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. click **Intrusions**.
4. In the upper-right corner of the page that appears, click **Settings**.

   Configure the parameters on different tabs.

   - **Add an approved logon location**
     a. In the **Login Location** section, click **Management** on the right.
     b. Select the logon location that you want to specify as the approved logon location and select the servers that allow logons from the specified location.
     c. Click **Ok**.

     Server Guard allows you to **edit** or **delete** approved logon locations that you have specified.

     - To change the servers that allow logons from an approved location, find the approved location and click **Edit** on the right.
     - To delete an approved logon location, find the logon location and click **Delete** on the right.
   - **Configure advanced logon settings**

     > **Note** When you configure advanced logon settings, you can specify the IP addresses, accounts, and time ranges that are allowed for logons to your assets. After the advanced logon settings are configured, Server Guard generates alerts if your assets receive unauthorized logon requests. The procedure of configuring advanced logon settings is similar to the procedure of configuring **Login Location**. You can **add**, **edit**, or **delete** advanced logon settings in a similar manner.

     - Turn on or turn off Uncommon IP Alert to the right of **Common Login IPs**. If you turn on Uncommon IP Alert and your assets receive logon requests from unapproved IP addresses, alerts are triggered.
     - Turn on or turn off Uncommon Time Alert to the right of **Common Login Time**. If you turn on Uncommon Time Alert and your assets receive logon requests during unapproved time ranges, alerts are triggered.
     - Turn on or turn off Uncommon Account Alert to the right of **Common Login Accounts**. If you turn on Uncommon Account Alert and your assets receive logon requests from unapproved accounts, alerts are triggered.
   - **Add web directories to scan**

     Server Guard automatically scans web directories of data assets in your servers and runs dynamic and static scan tasks. You can also manually add other web directories.

     a. In the **Add Scan Targets** section, click **Management** on the right.
     b. Specify a valid web directory and select the servers on which the specified web directory is scanned.

     > **Note** To ensure the scan performance and efficiency, we recommend that you do not specify a root directory.

     c. Click **Ok**.

22.1.6.4.1.6. Cloud threat detection

The cloud threat detection feature provided by Server Guard is integrated with widely-used antivirus engines. The feature detects viruses based on large amounts of threat intelligence data provided by Alibaba Cloud and the exception detection model designed by Alibaba Cloud. This model is designed based on machine learning and deep learning. This way, the cloud threat detection feature can provide full-scale and dynamic antivirus protection to safeguard your servers.

The cloud threat detection feature scans hundreds of millions of files on a daily basis and protects millions of servers on the cloud.

Detection capabilities

The cloud threat detection feature uses the Server Guard agent to collect process information and scans the retrieved data for viruses in the cloud. If a malicious process is detected, you can stop the process and quarantine the source files.

The cloud threat detection feature provides the following capabilities:

- **Deep learning engine developed by Alibaba Cloud**: The deep learning engine is built on deep learning technology and a large number of attack samples. The engine detects malicious files on the cloud and automatically identifies potential threats to supplement traditional antivirus engines.
- **Cloud sandbox developed by Alibaba Cloud**: The cloud sandbox feature allows you to simulate cloud environments and monitor attacks launched by malicious samples. The cloud sandbox feature automatically detects threats and offers dynamic analysis and detection capabilities based on big data analytics and machine learning modeling techniques.
- **Integration with major antivirus engines**: The cloud threat detection feature is integrated with major antivirus engines and updates its virus library in real time.
- **Threat intelligence detection**: The cloud threat detection feature works with the exception detection module to detect malicious processes and operations based on threat intelligence data provided by Alibaba Cloud Security.

Detectable virus types

The cloud threat detection feature is developed based on the security technologies and expertise of Alibaba Cloud. The feature provides end-to-end security services, including threat intelligence collection, data masking, threat identification, threat analysis, and malicious file quarantine and restoration. You can quarantine and restore files that contain viruses in the Security Center console.

The cloud threat detection feature can detect the following types of viruses.

Virus	Description
Mining program	A mining program consumes server resources and mines cryptocurrency without authorization.
Computer worm	A computer worm uses computer networks to replicate itself and spread to a large number of computers within a short period of time.
Ransomware	Ransomware, such as WannaCry, uses encryption algorithms to encrypt files and prevent users from accessing the files.
Trojan	A trojan is a program that allows an attacker to access information about servers and users, gain control of the servers, and consume system resources.
DDoS trojan	A DDoS trojan hijacks servers and uses zombie servers to launch DDoS attacks, which interrupts your service.
Backdoor	A backdoor is a malicious program injected by an attacker. Then, the attacker can use the backdoor to control the server or launch attacks.

Virus	Description
Computer virus	A computer virus inserts malicious code into normal programs and replicates the code to infect the whole system.
Malicious program	A malicious program may pose threats to system and data security.

Benefits

- **Self-developed and controllable**: The cloud threat detection feature is based on deep learning, machine learning, and big data analytics with a large number of attack and defense practices. The feature uses multiple detection engines to dynamically protect your assets against viruses.
- **Lightweight**: The cloud threat detection feature consumes only 1% of CPU resources and 50 MB of memory.
- **Dynamic**: The cloud threat detection feature dynamically retrieves startup logs of processes to monitor the startup of viruses.
- **Easy to manage**: You can manage all servers and view their status at any time in the Security Center console.

Threat detection limits

Apsara Stack Security Center allows you to detect and process security alerts, scan for and fix vulnerabilities, analyze attacks, and check security settings. Apsara Stack Security Center can analyze alerts and automatically trace attacks. This allows you to protect your assets. Apsara Stack Security supports a wide range of protection features. We recommend that you install the latest system patches on your assets. We also recommend that you use security services, such as Cloud Firewall and Web Application Firewall (WAF), to better protect your assets against attacks.

> **Note** Attacks and viruses are evolving, and security breaches may occur in various business environments. We recommend that you use the alerting, vulnerability detection, baseline check, and configuration assessment features provided by Apsara Stack Security to better protect your assets against attacks.

22.1.6.4.2. Website tamper-proofing

22.1.6.4.2.1. Overview

Tamper protection monitors website directories in real time, restores modified files or directories, and protects websites from trojans, hidden links, and uploads of violent and illicit content.

Background information

To make illegal profits or conduct business attacks, attackers exploit vulnerabilities in websites to insert illegal hidden links and tamper with the websites. Defaced web pages affect normal user access and may lead to serious economic losses, damaged brand reputation, or political risks.

Tamper protection allows you to add Linux and Windows processes to the whitelist and update protected files in real time.

How tamper protection works

The Security Center agent automatically collects the list of processes that attempt to modify files in the protected directories of the protected servers. It identifies unusual processes and file changes in real time and blocks unusual processes.

The alert list is displayed on the Tamper Protection page. You can view unusual file changes, the corresponding processes, and the number of attempts made by each process in the alert list. If a file is modified by a trusted process, you can add the process to the whitelist. After the process is added to the whitelist, tamper protection no longer blocks the process. In scenarios where the content of websites, such as news and education websites, is frequently modified, the whitelist saves you the effort of frequently enabling and disabling tamper protection.

Versions of operating systems and kernels supported by tamper protection

OS	Supported operating system version	Supported kernel version
Windows	Windows Server 2008 and later	All versions
CentOS	6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 7.0, 7.1, 7.2, 7.3, 7.4, 7.5, and 7.6	• 2.6.32-x • 3.10.0-x
Ubuntu	14, 16, and 18	• 3.13.0-32-generic • 3.13.0-86-generic • 4.4.0-62-generic • 4.4.0-63-generic • 4.4.0-93-generic • 4.4.0-151-generic • 4.4.0-117-generic • 4.15.0-23-generic • 4.15.0-42-generic • 4.15.0-45-generic • 4.15.0-52-generic

Note

- The preceding table lists kernel versions supported by tamper protection. Servers that use an unsupported kernel version cannot use tamper protection. Make sure that your server uses a supported kernel version. If a kernel version is not supported, you must upgrade it to a supported version. Otherwise, you cannot add processes to the whitelist.
- Before you upgrade the server kernel, back up your asset data.

22.1.6.4.2.2. Configure tamper protection

The Server Security feature allows you to configure tamper protection for web pages.

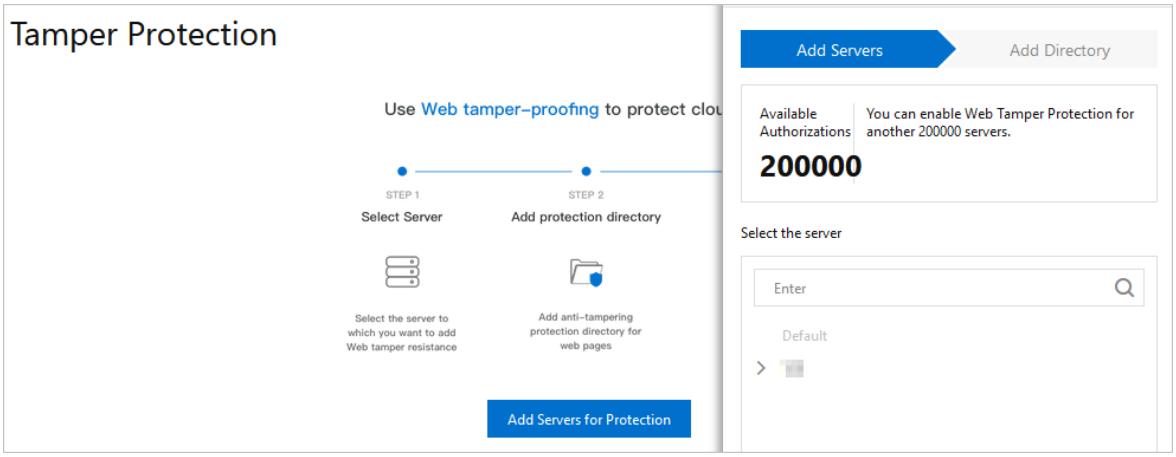
Limits

- For each server, you can add a maximum of 10 directories for protection.
- If you want to add directories that are on a Windows server, the directories must meet the following requirements: The size of each directory does not exceed 20 GB. Each directory contains no more than 2,000 folders. The number of directory levels does not exceed 20. The size of each file does not exceed 3 MB.
- If you want to add directories that are on a Linux server, the directories must meet the following requirements: The size of each directory does not exceed 20 GB. Each directory contains no more than 3,000 folders. The number of directory levels does not exceed 20. The size of each file does not exceed 3 MB.
- Before you add a directory for protection, make sure that the directory meets the preceding requirements.
- We recommend that you exclude file formats that do not require protection, such as *LOG*, *PNG*, *JPG*, *MP4*, *AVI*, and *MP3*. Multiple file formats can be separated by semicolons (;).

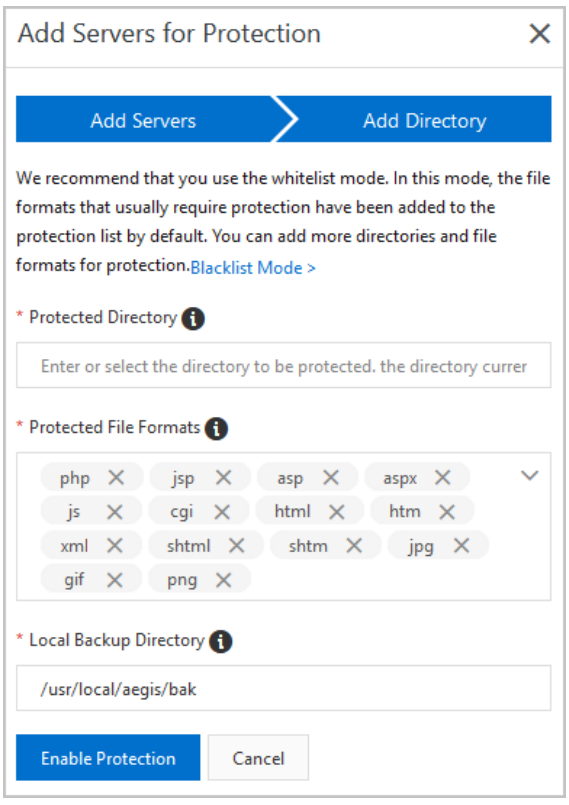
Procedure

1.
2.
3. click **File Tamper Protection**.

4. On the **Tamper Protection** page, click **Add Servers for Protection**.
5. In the **Add Servers for Protection** panel, select a server that you want to protect.



6. Click **Next** to go to the **Add Directory** step.
7. In the **Add Directory** step, configure the parameters.



Select a protection mode. The settings of other parameters vary based on the protection mode. You can select **Whitelist Mode** or **Blacklist Mode**. In whitelist mode, tamper protection is enabled for the specified directories and file formats. In blacklist mode, tamper protection is enabled for the subdirectories, file formats, and files that are not specified. By default, Whitelist Mode is selected.

- The following table describes the parameters that you must configure if you select Whitelist Mode.

Parameter	Description

Parameter	Description
Protected Directory	Enter the path of the directory that you want to protect. **Note** Servers that run Linux or Windows operating systems use different path formats. Enter a valid directory path based on the type of your operating system.
Protected File Formats	Select file formats that you want to protect from the drop-down list, such as *js*, *html*, *xml*, and *jpg*.
Local Backup Directory	The default path in which backup files of the protected directories are stored. By default, Apsara Stack Security assigns */usr/local/aegis/bak* to Linux servers and *C:\Program Files (x86)\Alibaba\Aegis\bak* to Windows servers. You can change the default path based on your business requirements.

- The following table describes the parameters that you must configure if you select Blacklist Mode.

Parameter	Description
Protected Directory	Enter the path of the directory that you want to protect.
Excluded Sub-Directories	Enter the subdirectories that do not require tamper protection. Click **Add Sub-Directory** to add more subdirectories. Apsara Stack Security does not provide tamper protection for files in the excluded subdirectories.
Excluded File Formats	Select file formats that you do not want to protect from the drop-down list. Valid values: **log**, **txt**, and **ldb**. Apsara Stack Security does not provide tamper protection for the files in the excluded formats.
Excluded Files	Enter the path of the file for which you do not want to protect. Click **Add File** to add more files. Apsara Stack Security does not provide tamper protection for the excluded files.
Local Backup Directory	The default path in which backup files of the protected directories are stored. By default, Apsara Stack Security assigns */usr/local/aegis/bak* to Linux servers and *C:\Program Files (x86)\Alibaba\Aegis\bak* to Windows servers. You can change the default path based on your business requirements.

8. Click **Enable Protection**.
   After you enable this feature for a server, the server name is displayed on the Management tab of the **Tamper Protection** page.

   Note By default, tamper protection is in the **Off** state for the server. To enable tamper protection for the server, you must **turn on** the switch in the Protection column on the **Tamper Protection** page.

9. On the **Tamper Protection** page, find the server that you add. Then, click the **Management** tab and turn on

the switch in the **Protection** column to enable tamper protection for the server.

> **Note** By default, tamper protection is in the **Off** state for the server. To enable tamper protection for the server, you must turn on the switch in the Protection column on the **Tamper Protection** page.

After tamper protection is enabled, the status of the server changes to **Running**.

> **Note** If the status of the server is **Exception**, move the pointer over **Exception** in the Status column to view the cause and click **Retry** to enable tamper protection again.

What to do next

After you enable tamper protection for a server, you can go to the **Alerts** page and select Webpage Tampering from the alert type drop-down list to view the alerts generated upon tampering events.

> **Note**
>
> Tamper protection does not take effect immediately after you configure the protected directory, and you can still write files to the directory. In this case, you must go to the **Management** page, disable **Protection** for the server where the directory is located, and then enable **Protection** again.

Handling suggestions for abnormal protection states

State	Description	Suggestion
Initializing	Tamper protection is being initialized.	If this is the first time that you enable tamper protection for a server, the protection status becomes **Initializing**. Wait until tamper protection is enabled.
Running	Tamper protection is enabled and running as expected.	None.
Exception	An error occurred when tamper protection was enabled.	Move the pointer over **Exception** in the Status column to view the exception cause and click Retry.
Not Initialized	Tamper protection is disabled.	Turn on the switch in the **Protection** column to enable tamper protection.

22.1.6.4.2.3. View protection status

This topic describes how to view the status of tamper protection for your assets.

Context

The tamper protection feature monitors changes to the files in website directories in real time and blocks suspicious file changes. To view the status of and details about the tamper protection feature, you must log on to Apsara Stack Security Center and choose **Server Security > Intrusion Prevention > File Tamper Protection**. The following information is displayed:

- Tamper protection overview

  You can view the numbers of files that are changed on the current day and in the last 15 days, the number of protected servers, and the number of protected directories.

- Distribution of protected file types

Protected file types include TXT, PNG, MSI, and ZIP. You can also add more types of files for tamper protection based on your business requirements.

> **Note** All types of files for tamper protection can be added.

- Top five files

  This section shows the names and paths of the top five files that are ranked based on the number of changes to files in descending order in the last 15 days.

- Tamper protection alerts

  This section lists the alerts generated for blocked suspicious changes to files for your assets. You can view details about the alerts, including the severity, alert name, affected assets, paths of files with suspicious changes, and protection status.

  > **Note**
  > - If an alert is reported more than 100 times, we recommend that you handle the alert at your earliest opportunity.
  > - Only alerts at the **Medium** level are displayed in the console.
  > - Only alerts in the **Defended** state are displayed. These alerts are triggered when the tamper protection feature blocks suspicious processes that attempt to modify files without authorization.

22.1.6.4.3. Configure the antivirus feature

Server Guard provides the antivirus feature. This feature allows you to configure settings for virus and webshell detection.

Detect and remove viruses

The antivirus feature can automatically quarantine common Internet viruses, such as common trojans, ransomware, mining programs, and DDoS trojans. Apsara Stack Security experts check and verify all automatically quarantined viruses to avoid false positives.

If the virus blocking feature is disabled, Server Guard generates alerts when viruses are detected. You can handle the detected viruses only in Apsara Stack Security Center. We recommend that you enable the virus blocking feature to improve the security of your servers.

1.
2.
3. click **Virus Defence**.
4. On the **Anti-virus** tab of the page that appears, click **Scan**.
5. In the dialog box that appears, select the servers that you want to scan.
6. Click **Scan**.
7. On the **Anti-virus** page, click the **Real-time protection** tab and turn on Virus Blocking to enable the virus blocking feature.

   After the virus blocking feature is enabled, Server Guard quarantines common viruses that are detected. Quarantined viruses are listed on the Alerts page. To filter these viruses, you can select the **Precision defense** type.

Detect and remove webshells

1.
2.
3. click **Virus Defence**.

4. Specify servers for webshell detection.
   i. In the **Webshell Detection** section, click **Manage**.
   ii. Select the servers for which you want to enable webshell detection.
   iii. Click **OK** to complete the configuration.

22.1.6.5. Log retrieval

22.1.6.5.1. Log retrieval overview

The log retrieval function provided by Server Security allows you to manage logs scattered in various systems of Apsara Stack in a centralized manner, so that you can easily identify the causes of issues that occur on your servers.

The log retrieval function supports storage of logs for 180 days and query of logs generated within 30 days.

Benefits

The log retrieval function provides the following benefits:

- **End-to-end log retrieval platform**: Allows you to retrieve logs of various Apsara Stack services in a centralized manner and trace issues easily.
- **Cloud-based SaaS service**: Allows you to query logs on all servers in Apsara Stack without additional installment and deployment.
- Supports TB-level data retrieval. It also allows you to add a maximum of 50 inference rules (Boolean expressions) in a search condition and obtain full-text search results within several seconds.
- Supports a wide range of log sources.
- Supports log shipping, which allows you to import security logs to Log Service for further analysis.

Scenarios

You can use log retrieval to meet the following requirements:

- **Security event analysis**: When a security event is detected on a server, you can retrieve the logs to identify the cause and assess the damage and affected assets.
- **Operation audit**: You can audit the operation logs on a server to identify high-risk operations and serious issues in a meticulous way.

Supported log types

Log types

Log type	Description
Logon history	Log entries about successful system logons
Brute-force attack	Log entries about system logon failures that are generated during brute-force attacks
Process snapshot	Log entries about processes on a server at a specific time
Listening port snapshot	Log entries about listening ports on a server at a specific time
Account snapshot	Log entries about account logon information on a server at a specific time
Process initiation	Log entries about process initiation on a server
Network connection	Log entries about active connections from a server to external networks

22.1.6.5.2. Query logs

This topic describes how to search for and view server logs.

Procedure

1.
2.
3. click **Log Retrieval**.
4. Specify search conditions.

Search condition	Description
Log source	The log source that you want to query. For more information, see Log sources.
Field	The field that is recorded for the log source. For more information, see Log sources.
Keyword	The keyword of the field.
Logical operator	The equality operator.
+	The inference rules in a search condition for a log source.
Add conditions	The search conditions for different log sources.

5. Click **Search** and view the search result.
   - **Reset**: Click **Reset** to clear the search condition configurations.
   - **Saved Searches**: Click **Saved Searches** to select and use the search condition configurations that you saved.

22.1.6.5.3. Supported log sources and fields

This topic describes the log sources and fields that are supported by the log retrieval feature.

The log retrieval feature allows you to query the following types of log sources. You can click a log source link to view the fields that can be retrieved.

Log source	Description
Logon history	Log entries about successful system logons
Logs of brute-force attacks	Log entries about failed system logons during brute-force attacks
Process snapshot logs	Log entries about processes on a server at a specific point in time
Logs of listening port snapshots	Log entries about listening ports on a server at a specific point in time
Account snapshot logs	Log entries about account-based logons on a server at a specific point in time
Process startup logs	Log entries about process startups on a server
Network connection logs	Log entries about active connections from a server to the Internet.

Logon history

The following table describes the fields that you can use to query the logon history.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
warn_ip	string	The source IP address used for the logon.
warn_port	string	The logon port.
warn_user	string	The username used for the logon.
warn_type	string	The logon type.
warn_count	string	The number of logon attempts.

Logs of brute-force attacks

The following table describes the fields that you can use to query logs of brute-force attacks.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
warn_ip	string	The source IP address of the attack.
warn_port	string	The target port of the attack.
warn_user	string	The target username of the attack.
warn_type	string	The attack type.
warn_count	string	The number of brute-force attack attempts.

Process startup logs

The following table describes the fields that you can use to query process startup logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
pid	string	The ID of the process.
groupname	string	The user group.
ppid	string	The ID of the parent process.
uid	string	The ID of the user.
username	string	The username.

Field	Data type	Description
filename	string	The file name.
pfilename	string	The name of the parent process file.
cmdline	string	The command line.
filepath	string	The path of the process file.
pfilepath	string	The path of the parent process file.

Logs of listening port snapshots

The following table describes the fields that you can use to query logs about listening port snapshots.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
src_port	string	The listening port.
src_ip	string	The listening IP address.
proc_path	string	The path of the process file.
pid	string	The ID of the process.
proc_name	string	The name of the process.
proto	string	The protocol.

Account snapshot logs

The following table describes the fields you can use to query account snapshot logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
perm	string	Indicates whether the user has root permissions.
home_dir	string	The home directory.
warn_time	string	The time when a password expiration notification is sent.
groups	string	The group to which the user belongs.
login_ip	string	The IP address of the last logon.
last_chg	string	The time when the password was last changed.

Field	Data type	Description
shell	string	The Linux shell command.
domain	string	The Windows domain.
tty	string	The logon terminal.
account_expire	string	The time when the account expires.
passwd_expire	string	The time when the password expires.
last_logon	string	The last logon time.
user	string	The username.
status	string	The account status. Valid values: • 0: disabled • 1: normal

Process snapshot logs

The following table describes the fields that you can use to query process snapshot logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
path	string	The path of the process file.
start_time	string	The time when the process was started.
uid	string	The ID of the user.
cmdline	string	The command line.
pname	string	The name of the parent process.
name	string	The name of the process.
pid	string	The ID of the process.
user	string	The username.
md5	string	The MD5 hash value of the process file. If the size of the process file exceeds 1 MB, the system does not calculate the MD5 hash value of the process file.

Network connection logs

The following table describes the fields that you can use to query network connection logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
src_ip	string	The source IP address.
src_port	string	The source port.
proc_path	string	The path of the process file.
dst_port	string	The destination port.
proc_name	string	The name of the process.
dst_ip	string	The destination IP address.
status	string	The status.

22.1.6.5.4. Logical operators

The log retrieval feature supports multiple search conditions. You can add multiple logical operators to one search condition for one log source, or combine multiple search conditions for several log sources by using different logical operators. This topic describes the logical operators that are supported in log retrieval. Examples are provided to help you understand these operators.

The following table describes the logical operators that are supported in log retrieval.

Logical operators

Logical operator	Description
and	Binary operator. This operator is in the format of `query 1 and query 2`, which indicates the intersection of the query results of `query 1` and `query 2`. **Note** If no logical operators are used for multiple keywords, the default operator is AND.
or	Binary operator. This operator is in the format of `query 1 or query 2`, which indicates the union of the query results of `query 1` and `query 2`.
not	Binary operator. This operator is in the format of `query 1 not query 2`, which indicates the results that match `query 1` but do not match `query 2`. This format is equivalent to `query 1 - query 2`. **Note** If you use only `not query 1`, the log data that does not contain the query results of `query 1` is returned.

22.1.6.6. Settings

22.1.6.6.1. Install the Server Guard agent

This topic describes how to install the Server Guard agent.

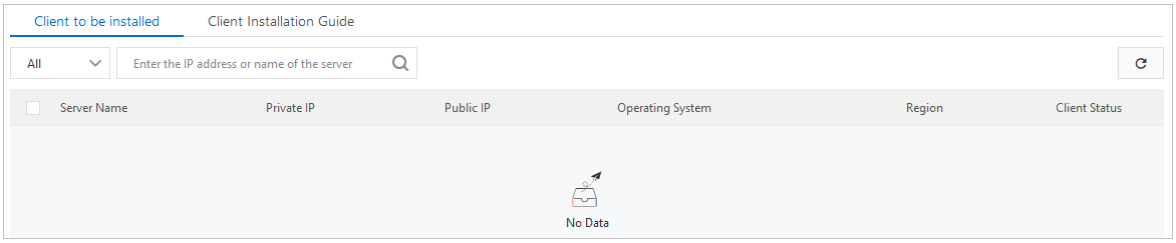
Context

To use the protection features provided by Server Guard, you must install the Server Guard agent on the operating system of your server.

Procedure

1.
2.
3. click **Client Installation**.
4. (Optional)On the page that appears, click the **Client to be installed** tab to view the number of the servers on which the Server Guard agent is not installed. On this tab, you can also view information about these servers.

   You can specify the operating system type, server IP address, or server name to search for a server.



5. Click the **Client Installation Guide** tab.
6. Download and install the Server Guard agent based on the operating system of your server.
   - **Windows**
     a. In the left-side pane of the page, click **Click to download** to download the installation package to your computer.
     b. Upload the installation package to your server. For example, you can use an FTP client to upload the installation package to your server.
     c. Run the installation package on your server as an administrator.

     > **Note** If you install the agent on a server that is not in Alibaba Cloud, you are prompted to enter the installation verification key. You can find the installation verification key on the Client Installation Guide tab.

   - **Linux**
     a. In the right-side pane of the page, select Alibaba Cloud Server or Non-Alibaba Server.
     b. Select the installation command for your 32-bit or 64-bit operating system and click **Copy** to copy the command.
     c. Log on to your Linux server as an administrator.
     d. Run the installation command on your Linux server to download and install the Server Guard agent.

22.1.6.6.2. Manage protection modes

This topic describes how to manage protection modes for a server to improve the performance and security of the server.

Procedure

1.
2.
3. click **Protection Mode**.
4. On the **Protection First Mode** page, click **Manage** next to Protection Mode.

   Configure protection modes for servers.

   - **Business First Mode**: In this mode, the peak CPU utilization is less than 10%, and the peak memory usage is less than 50 MB.
   - **Protection First Mode**: In this mode, the peak CPU utilization is less than 20%, and the peak memory usage is less than 80 MB.
5. Click **OK**.

22.1.7. Physical server security

22.1.7.1. Create and grant permissions to a security administrator account

The physical server security feature is used to ensure the security of physical servers on the platform side. This feature requires you to use a dedicated security administrator account for the platform. This topic describes how to create and grant permissions to a security administrator account.

Procedure

1. Log on to the Apsara Uni-manager Management Console as a system administrator.

   For more information, see the **"Log on to the Apsara Uni-manager Management Console"** topic of *Apsara Uni-manager Management Console User Guide*.
2. Create a dedicated organization that is used to manage the security of physical servers, and obtain the primary key of the organization.

   > **Notice** Make sure that the organization is used only to manage the security of physical servers. Do not add Elastic Compute Service (ECS) instances to the organization.

   i. Create the dedicated organization.

   For more information, see **Enterprise Center > Organization Management > Create Organization** in *Apsara Uni-manager Management Console User Guide*.

   ii. Obtain the **primary key** of the newly created organization.

   For more information, see **Enterprise Center > Organization Management > Obtain the AccessKey pair of an organization** in *Apsara Uni-manager Management Console User Guide*.
3. Create a dedicated account to manage the security of physical servers.

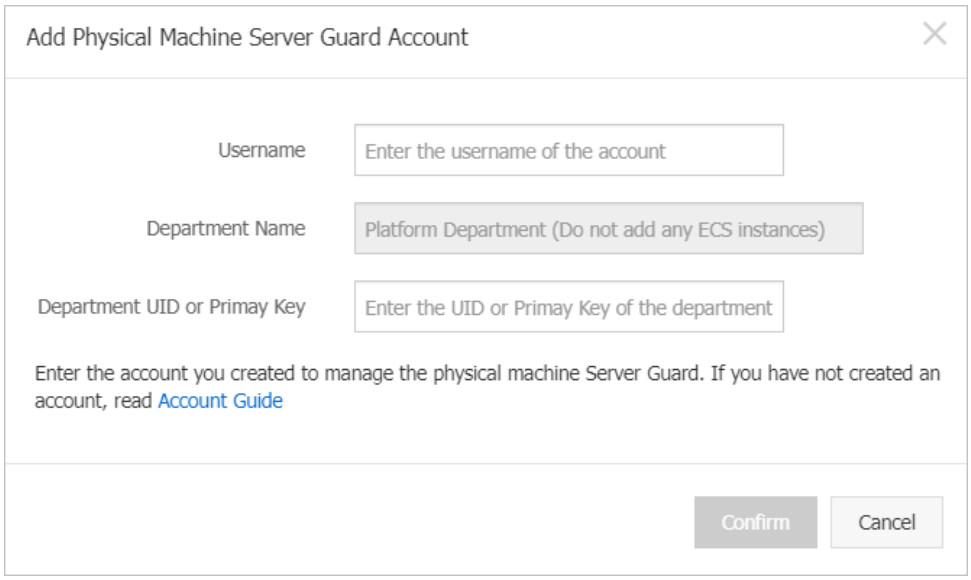
   For more information, see **Enterprise Center > User Management > System User Management > Create User** in *Apsara Uni-manager Management Console User Guide*.

> **Note** When you create the account, take note of the following points for the **organization** and **role**:
>
> - In the **Organization** section, select the organization that is created in the previous step.
> - In the **Role** section, select **Platform Security Configuration Administrator** and **Security System Configuration Administrator**.

4. Log on to Apsara Stack Security Center by using the newly created account.

   For more information, see Log on to Apsara Stack Security Center.

5. Add the **primary key** of the newly created organization to the protection configuration of physical servers.
   i.
   ii. click **Global Settings**
   iii. On the **Global Settings** page, click the **Physical Machine Protection Configurations** tab.
   iv. Click **Add Account**.
   v. In the **Add Physical Machine Server Guard Account** dialog box, configure the **Username** and **Department UID or Primay Key** parameters.



   - **Username**: Enter the account that you created in Step 3.
   - **Primary Key**: Enter the primary key that you obtained in Step 2.

   vi. Click **Confirm**.

Result

After the settings are complete, you can use the dedicated security administrator account that is created in this section to ensure the security of physical servers on the platform side.

22.1.7.2. Physical servers

22.1.7.2.1. Manage physical server groups

This topic describes how to manage physical server groups. To facilitate the security management of physical servers, you can add the physical servers to groups and view their security events by group.

Context

By default, physical servers do not belong to a server group. You must add your physical servers to a server group. If you delete a group, all the physical servers in the group are retained but no longer belong to a server group.

Procedure

1. Log on to Apsara Stack Security Center by using an Apsara Stack tenant account.

   > **Note** For more information about the Apsara Stack tenant account, see Create and grant permissions to a security administrator account.

2.
3. In the left-side navigation pane, choose **Physical Server Security > Servers**.
4. In the left-side group pane, manage sever groups.
   - Create a group.

     Click the Add Subgroup icon next to **All Servers** or a specific group, enter a group name, and click **OK**.

     > **Note** The system supports a maximum of three levels of groups.

   - Modify a group.

     Click the Modify Group Name icon next to the target group, enter a new name, and click **OK**.

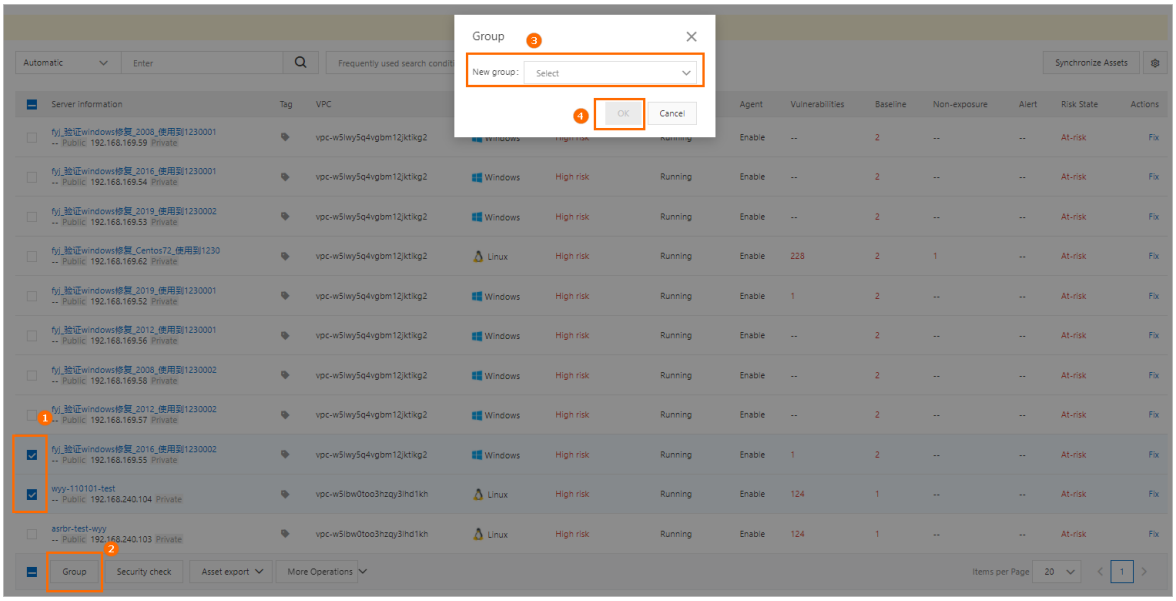
   - Delete a group.

     Click the Delete icon next to the target group. In the message that appears, click **OK**.

     > **Note** After you delete a group, all servers in the group are automatically moved to the **default** group.

   - Sort groups.

     Click **Manage Groups** to sort groups in descending order by priority.
5. Change the server group of specific physical servers.

   i. Select servers from the list on the right.

   ii. Click **Change Group**.

   iii. In the Change Group dialog box that appears, select a group from the drop-down list.

iv. Click **OK**.

22.1.7.2.2. Manage physical servers

This topic describes how to manage servers. On the Servers page, you can view the status of servers protected by Server Guard.

Procedure

1. Log on to Apsara Stack Security Center by using an Apsara Stack tenant account.

   > **Note** For more information about the Apsara Stack tenant account, see Create and grant permissions to a security administrator account.

2.
3. click **Servers**.
4. (Optional)Search for a server.

   To view the agent status of a server, enter the server IP address in the search bar, and click **Search**. Detailed server information, such as security information, is displayed.

5. View the agent status and detailed security information of the server.

   Click

   in the upper-right corner of the page to select the information columns you want to display. The following table lists the information categories.

Category	Information
Basic information	○ Server IP/Name ○ Tag ○ OS ○ Region
Agent status	Agent Status
Threat prevention	○ Vulnerability ○ Baseline Risk
Intrusion detection	○ Unusual Logons ○ Webshells ○ Suspicious Servers
Server fingerprints	○ Processes ○ Ports ○ Root Accounts/Total Accounts

6. Manage servers.

Action	Description

Action	Description
Change Group	Select servers and click **Change Group** to add the selected servers to a new group.
Modify Tag	Select servers and click **Modify Tag** to modify tags for the servers.
Security Inspection	Select servers and click **Security Inspection** to select the items to be checked.
Delete External Servers	Select **external** servers, and choose **More > Delete External Servers**.
Disable Protection	Select the servers whose agent status is **Online**, and choose **More > Disable Protection**. This temporarily disables protection for these servers to reduce server resource consumption.
Enable Protection	Select the servers whose agent status is **Disable Protection**, and choose **More > Enable Protection**. This enables protection for these servers.

22.1.7.3. Intrusion events

22.1.7.3.1. Intrusion event types

If Server Guard detects sensitive file tampering, suspicious processes, webshells, unusual logons, or malicious processes, it generates alerts. Based on these alerts, you can monitor the security status of your assets and handle potential threats at the earliest opportunity.

Apsara Stack Security provides statistics on enabled alerts and defense items. These statistics help you monitor the overall security of your assets. You can view the statistics on the **Intrusions** page.

Alerts

The following table describes the alerts.

Alert	Description
Threat intelligence	Identify potential threats to your assets based on the threat intelligence of Apsara Stack Security. Threat intelligence can correlate threat information to analyze and process the information. If threats are detected, threat intelligence can generate alerts. This helps improve the detection efficiency and response speed. Threat intelligence can detect the following items: • Malicious domain names • Malicious IP addresses • IP addresses of dark web services • IP addresses of command and control (C&C) servers • IP addresses of mining pools • Malicious URLs • Malicious download sources

Alert	Description
Unusual Logon	Detect unusual logons to your servers. You can specify approved logon IP addresses, time periods, and accounts. Logons from unapproved IP addresses,time periods, or accounts trigger alerts. You can manually add approved logon locations or configure the system to automatically update approved logon locations. You can also specify assets on which alerts are triggered when unapproved logon locations are detected. Server Guard can detect the following events: • Logons to Elastic Compute Service (ECS) instances from unapproved IP addresses • Logons to ECS instances from unapproved locations • Execution of unusual commands after SSH-based logons to ECS instances • Brute-force attacks on SSH passwords of ECS instances
Webshell	Use engines developed by Alibaba Cloud to scan common webshell files. Server Guard supports scheduled scan tasks, provides real-time protection, and quarantines webshell files. • Server Guard scans the entire web directory early in the morning on a daily basis. A change made to files in the web directory triggers dynamic detection. • You can specify the assets on which Server Guard scans for webshells. • You can quarantine or ignore detected trojan files. You can also restore the quarantined trojan files.
Precision defense	The **antivirus** feature provides precise protection from common ransomware, DDoS trojans, mining programs, trojans, malicious programs, webshells, and computer worms.
Suspicious Account	Detect logons to your assets from unapproved accounts.
Cloud threat detection	Detect threats in other cloud services.
Persistence	Detect suspicious scheduled tasks on servers and generate alerts when advanced persistent threats (APTs) to the servers are detected.
Unusual Network Connection	Detect disconnections or unusual network connections.
Suspicious Process	Detect whether suspicious processes exist.
Malicious Process	Scan your servers in real time. An agent is used to collect process information, and the information is uploaded to the cloud for detection. If viruses are detected, alerts are generated. You can handle detected viruses in Apsara Stack Security Center. Server Guard can detect the following malicious activities and processes: • Access to malicious IP addresses • Mining programs • Self-mutating trojans • Malicious programs • Trojans
Sensitive File Tampering	Check whether sensitive files on your servers are maliciously modified. The sensitive files include preloaded configuration files in Linux shared libraries.
Other	Detect other types of attacks, such as DDoS attacks.

Alert	Description
Web Application Threat Detection	Detect intrusions that use web applications.
Application intrusion event	Detect intrusions that use system application components.

22.1.7.3.2. View and handle alert events

This topic describes how to view and handle detected alert events on the Intrusions page.

Background information

After alert events are detected, the alerts events are displayed on the **Intrusions** page in Apsara Stack Security Center. If the detected alert events are not handled, they are displayed in the **Unhandled Alerts** list on the **Intrusions** page. After the alert events are handled, the status of the alert events changes from **Unhandled Alerts** to **Handled**.

> **Note** Apsara Stack Security Center retains the records of **Unhandled Alerts** and **Handled** on the **Intrusions** page. By default, the records of **Unhandled Alerts** are displayed.

View alert events

1.
2.
3. click **Intrusions**.
4. On the page that appears, search for or view all alert events. You can also view the details about the alert events.

Handle alert events

1.
2.
3. click **Intrusions**.
4. On the **Intrusions** page, find the alert event that you want to handle and click **Handle** in the **Actions** column. In the dialog box that appears, configure Process Method and click **Process Now**.

   > **Note** If the alert event is related to multiple exceptions, the panel that shows alert event details appears after you click **Handle**. You can handle the exceptions in the panel.

   - **Ignore**: If you ignore the alert event, the status of the alert event changes to **Handled**. Server Guard no longer generates alerts for the event.
   - **Add To Whitelist**: If the alert event is a false positive, you can add the alert event to the whitelist. Then, the status of the alert event changes to **Handled**. Server Guard no longer generates alerts for the event. In the **Handled** list, you can click **Cancel whitelist** to remove the alert event from the whitelist.

     > **Note** When Server Guard generates a false alert on a normal process, this alert is considered a false positive. A common false positive is a **suspicious process that sends TCP packets**. The false positive notifies you that suspicious scans on other devices are detected on your servers.

   - **Batch unhandled**: This method allows you to batch handle multiple alert events. Before you batch handle multiple alert events, we recommend that you view the details about the alert events.
5. (Optional)If you confirm that one or more alert events are false positives or need to be ignored, go to the

Intrusions page. Then, select the alert events and click **Ignore Once** or **Whitelist**.

Export alert events

1.
2.
3. click **Intrusions**.
4. In the upper-left corner above the alert event list on the **Intrusions** page, click the icon to export the list.

   After the list is exported, the **Done** message appears in the upper-right corner of the Intrusions page.
5. In the **Done** notification of the **Alerts** page, click **Download**.
   The alert list is downloaded to your computer.

22.1.7.3.3. View exceptions related to an alert

Server Guard supports automatic analysis of exceptions related to an alert. You can click an alert name in the alert list to view and handle all exceptions that are related to the alert. You can also view the results of automatic attack tracing to analyze the exceptions.

Context

- Security Center automatically associates alerts with exceptions in real time to detect potential threats.
- Exceptions related to an alert are listed in chronological order. This allows you to analyze and handle the exceptions to improve the emergency response mechanism of your system.
- An automatically correlated alert is identified by the icon.

Procedure

1.
2.
3. click **Intrusions**.
4. On the Intrusions page, click the **name of the alert** that you want to handle. The alert details panel appears.
5. In the alert details panel, view the details and related exceptions of the alert. Then, handle the exceptions.
   - View alert details

     You can view the assets that are affected by the alert, the first and latest time when the alert was triggered, and the details about the related exceptions.
   - View affected assets

     You can move the pointer over the name of an **affected asset** to view the details about the asset. The details include information about all the alerts, vulnerabilities, baseline risks, and asset fingerprints on the asset.
   - View and handle **related exceptions**

     In the **Related Exceptions** section, you can view the details about all the exceptions that are related to the alert. You can also view suggestions on how to handle the exceptions.
     - Click **Note** to the right of an exception to add a note for the exception.
     - Click the icon to the right of a note to delete the note.

22.1.7.3.4. Use the file quarantine feature

Sever Guard can quarantine malicious files. Quarantined files are listed in the Quarantine panel of the Intrusions page. You can restore a quarantined file with a few clicks. However, 30 days after a file is quarantined, the system automatically deletes the file. This topic describes how to view and restore quarantined files.

Procedure

1.
2.
3. click **Intrusions**.
4. In the upper-right corner of the **Intrusions** page, click **Quarantine**.

   In the **Quarantine** panel, you can perform the following operations:

   - View information about quarantined files. The information includes server IP addresses, directories in which the files are stored, file status, and modification time.
   - Click **Restore** in the **Actions** column to restore a quarantined file. The restored file appears in the alert list.

22.1.7.3.5. Configure alerts

This topic describes how to configure alerts. You can specify approved logon locations and customize web directories to scan.

Context

Server Guard supports advanced logon settings. You can configure more fine-grained logon detection rules. For example, you can specify approved logon IP addresses, logon time ranges, and logon accounts to block unauthorized requests that are sent to your assets.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. click **Intrusions**.
4. In the upper-right corner of the page that appears, click **Settings**.

   Configure the parameters on different tabs.

   - **Add an approved logon location**
     a. In the **Login Location** section, click **Management** on the right.
     b. Select the logon location that you want to specify as the approved logon location and select the servers that allow logons from the specified location.
     c. Click **Ok**.

     Server Guard allows you to **edit** or **delete** approved logon locations that you have specified.

     - To change the servers that allow logons from an approved location, find the approved location and click **Edit** on the right.
     - To delete an approved logon location, find the logon location and click **Delete** on the right.
   - **Configure advanced logon settings**

> **Note** When you configure advanced logon settings, you can specify the IP addresses, accounts, and time ranges that are allowed for logons to your assets. After the advanced logon settings are configured, Server Guard generates alerts if your assets receive unauthorized logon requests. The procedure of configuring advanced logon settings is similar to the procedure of configuring **Login Location**. You can **add**, **edit**, or **delete** advanced logon settings in a similar manner.

- Turn on or turn off Uncommon IP Alert to the right of **Common Login IPs**. If you turn on Uncommon IP Alert and your assets receive logon requests from unapproved IP addresses, alerts are triggered.
- Turn on or turn off Uncommon Time Alert to the right of **Common Login Time**. If you turn on Uncommon Time Alert and your assets receive logon requests during unapproved time ranges, alerts are triggered.
- Turn on or turn off Uncommon Account Alert to the right of **Common Login Accounts**. If you turn on Uncommon Account Alert and your assets receive logon requests from unapproved accounts, alerts are triggered.

- **Add web directories to scan**

  Server Guard automatically scans web directories of data assets in your servers and runs dynamic and static scan tasks. You can also manually add other web directories.

  a. In the **Add Scan Targets** section, click **Management** on the right.
  b. Specify a valid web directory and select the servers on which the specified web directory is scanned.

  > **Note** To ensure the scan performance and efficiency, we recommend that you do not specify a root directory.

  c. Click **Ok**.

22.1.7.3.6. Cloud threat detection

The cloud threat detection feature provided by Server Guard is integrated with widely-used antivirus engines. The feature detects viruses based on large amounts of threat intelligence data provided by Alibaba Cloud and the exception detection model designed by Alibaba Cloud. This model is designed based on machine learning and deep learning. This way, the cloud threat detection feature can provide full-scale and dynamic antivirus protection to safeguard your servers.

The cloud threat detection feature scans hundreds of millions of files on a daily basis and protects millions of servers on the cloud.

Detection capabilities

The cloud threat detection feature uses the Server Guard agent to collect process information and scans the retrieved data for viruses in the cloud. If a malicious process is detected, you can stop the process and quarantine the source files.

The cloud threat detection feature provides the following capabilities:

- **Deep learning engine developed by Alibaba Cloud**: The deep learning engine is built on deep learning technology and a large number of attack samples. The engine detects malicious files on the cloud and automatically identifies potential threats to supplement traditional antivirus engines.
- **Cloud sandbox developed by Alibaba Cloud**: The cloud sandbox feature allows you to simulate cloud environments and monitor attacks launched by malicious samples. The cloud sandbox feature automatically detects threats and offers dynamic analysis and detection capabilities based on big data analytics and machine learning modeling techniques.
- **Integration with major antivirus engines**: The cloud threat detection feature is integrated with major antivirus engines and updates its virus library in real time.
- **Threat intelligence detection**: The cloud threat detection feature works with the exception detection module to detect malicious processes and operations based on threat intelligence data provided by Alibaba

Cloud Security.

Detectable virus types

The cloud threat detection feature is developed based on the security technologies and expertise of Alibaba Cloud. The feature provides end-to-end security services, including threat intelligence collection, data masking, threat identification, threat analysis, and malicious file quarantine and restoration. You can quarantine and restore files that contain viruses in the Security Center console.

The cloud threat detection feature can detect the following types of viruses.

Virus	Description
Mining program	A mining program consumes server resources and mines cryptocurrency without authorization.
Computer worm	A computer worm uses computer networks to replicate itself and spread to a large number of computers within a short period of time.
Ransomware	Ransomware, such as WannaCry, uses encryption algorithms to encrypt files and prevent users from accessing the files.
Trojan	A trojan is a program that allows an attacker to access information about servers and users, gain control of the servers, and consume system resources.
DDoS trojan	A DDoS trojan hijacks servers and uses zombie servers to launch DDoS attacks, which interrupts your service.
Backdoor	A backdoor is a malicious program injected by an attacker. Then, the attacker can use the backdoor to control the server or launch attacks.
Computer virus	A computer virus inserts malicious code into normal programs and replicates the code to infect the whole system.
Malicious program	A malicious program may pose threats to system and data security.

Benefits

- **Self-developed and controllable**: The cloud threat detection feature is based on deep learning, machine learning, and big data analytics with a large number of attack and defense practices. The feature uses multiple detection engines to dynamically protect your assets against viruses.
- **Lightweight**: The cloud threat detection feature consumes only 1% of CPU resources and 50 MB of memory.
- **Dynamic**: The cloud threat detection feature dynamically retrieves startup logs of processes to monitor the startup of viruses.
- **Easy to manage**: You can manage all servers and view their status at any time in the Security Center console.

Threat detection limits

Apsara Stack Security Center allows you to detect and process security alerts, scan for and fix vulnerabilities, analyze attacks, and check security settings. Apsara Stack Security Center can analyze alerts and automatically trace attacks. This allows you to protect your assets. Apsara Stack Security supports a wide range of protection features. We recommend that you install the latest system patches on your assets. We also recommend that you use security services, such as Cloud Firewall and Web Application Firewall (WAF), to better protect your assets against attacks.

> **Note** Attacks and viruses are evolving, and security breaches may occur in various business environments. We recommend that you use the alerting, vulnerability detection, baseline check, and configuration assessment features provided by Apsara Stack Security to better protect your assets against attacks.

22.1.7.4. Server fingerprints

22.1.7.4.1. Manage listening ports

This topic describes how to view information about the listening port of a server. The information helps you identify suspicious listening behavior.

Context

This topic is suitable for the following scenarios:

- Check for servers that listen on a specific port.
- Check for ports that a specific server listens.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. On the **Asset Fingerprints** page, click the **Port** tab to view **listening ports**, **network protocols**, and server information.

   You can search for a port by using the port number, server process name, server name, or server IP address.

   In the server information list, you can view the **process**, **IP address**, and **latest scan time** of a server.

22.1.7.4.2. Manage software versions

This topic describes how to regularly view and collect the software version information about a server. This helps you check your software assets.

Context

This topic covers the following scenarios:

- Check for software assets that are installed without authorization.
- Check for outdated software assets.
- Locate affected assets if vulnerabilities are detected.

Procedure

1. Log on to Apsara Stack Security Center.
2. Choose **Server Security > Sever Guard**.
3. In the left-side navigation pane, click **Server Fingerprints**.
4. On the page that appears, click the **Software** tab. On the tab, view all the **software assets** that are in use and the **number of the servers** that use the software assets.

   You can search for specific software by using its name, version, installation directory, server name, or IP address.
5. Click software to view the details, such as the software versions and the servers that use the software.

   You can click the icon in the upper-right corner to download a software version table to your computer for subsequent asset check.

22.1.7.4.3. Manage processes

This topic describes how to regularly collect the process information on a server and record changes. This way, you can view process information and historical process changes.

Context

This task is suitable for the following scenarios:

- Check for servers on which a specific process runs.
- Check for processes that run on a specific server.

Procedure

1. Log on to Apsara Stack Security Center.
2.
3. click **Server Fingerprints**.
4. On the page that appears, click the **Process** tab. On the tab, view all running processes and the number of servers that run these processes.

   You can search for a process by using the **process name**, **running user**, **startup parameter**, or **server name or IP address**.
5. Click the name of a process to view the details of the process, such as the servers, paths, and startup parameters.

22.1.7.4.4. Manage account information

This topic describes how to regularly collect the account information on a server and record the changes to the accounts. This way, you can check your accounts and view historical changes to your accounts.

Context

You can use the information collected in this topic for the following scenarios:

- Check for servers on which a specific account is created.
- Check for accounts that are created on a server.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. On the **Asset Fingerprints** page, click the **Account** tab.
5. View all the logged-on accounts and the numbers of servers on which the accounts are created.

   You can search for an account by using the account name, root permissions, server name, or server IP address.
6. Click an account name to view the details, such as the server information, root permissions, and user group.

22.1.7.4.5. Manage scheduled tasks

This topic describes how to view scheduled tasks on servers.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. On the **Asset Fingerprints** page, click the **Scheduled Tasks** tab.
5. View the paths of all tasks and the number of servers that run these tasks.

You can search for a task by using the task path, server name, or IP address.

6. Click a task path to view the details, such as the servers, executed commands, and task cycles.

22.1.7.4.6. Set the fingerprint collection frequency

You can set the frequency at which the data of running processes, system accounts, listening ports, and software versions is collected.

Procedure

1.
2.
3. click **Server Fingerprints**.
4. In the upper-right corner of the **Asset Fingerprints** page, click **Settings**.
5. Select the collection frequency from each drop-down list.
6. Click **OK** to complete the configuration.

22.1.7.5. Log retrieval

22.1.7.5.1. Supported log sources and fields

This topic describes the log sources and fields that are supported by the log retrieval feature.

The log retrieval feature allows you to query the following types of log sources. You can click a log source link to view the fields that can be retrieved.

Log source	Description
Logon history	Log entries about successful system logons
Logs of brute-force attacks	Log entries about failed system logons during brute-force attacks
Process snapshot logs	Log entries about processes on a server at a specific point in time
Logs of listening port snapshots	Log entries about listening ports on a server at a specific point in time
Account snapshot logs	Log entries about account-based logons on a server at a specific point in time
Process startup logs	Log entries about process startups on a server
Network connection logs	Log entries about active connections from a server to the Internet.

Logon history

The following table describes the fields that you can use to query the logon history.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
warn_ip	string	The source IP address used for the logon.

Field	Data type	Description
warn_port	string	The logon port.
warn_user	string	The username used for the logon.
warn_type	string	The logon type.
warn_count	string	The number of logon attempts.

Logs of brute-force attacks

The following table describes the fields that you can use to query logs of brute-force attacks.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
warn_ip	string	The source IP address of the attack.
warn_port	string	The target port of the attack.
warn_user	string	The target username of the attack.
warn_type	string	The attack type.
warn_count	string	The number of brute-force attack attempts.

Process startup logs

The following table describes the fields that you can use to query process startup logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
pid	string	The ID of the process.
groupname	string	The user group.
ppid	string	The ID of the parent process.
uid	string	The ID of the user.
username	string	The username.
filename	string	The file name.
pfilename	string	The name of the parent process file.
cmdline	string	The command line.
filepath	string	The path of the process file.

Field	Data type	Description
pfilepath	string	The path of the parent process file.

Logs of listening port snapshots

The following table describes the fields that you can use to query logs about listening port snapshots.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
src_port	string	The listening port.
src_ip	string	The listening IP address.
proc_path	string	The path of the process file.
pid	string	The ID of the process.
proc_name	string	The name of the process.
proto	string	The protocol.

Account snapshot logs

The following table describes the fields you can use to query account snapshot logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
perm	string	Indicates whether the user has root permissions.
home_dir	string	The home directory.
warn_time	string	The time when a password expiration notification is sent.
groups	string	The group to which the user belongs.
login_ip	string	The IP address of the last logon.
last_chg	string	The time when the password was last changed.
shell	string	The Linux shell command.
domain	string	The Windows domain.
tty	string	The logon terminal.
account_expire	string	The time when the account expires.

Field	Data type	Description
passwd_expire	string	The time when the password expires.
last_logon	string	The last logon time.
user	string	The username.
status	string	The account status. Valid values: • 0: disabled • 1: normal

Process snapshot logs

The following table describes the fields that you can use to query process snapshot logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
path	string	The path of the process file.
start_time	string	The time when the process was started.
uid	string	The ID of the user.
cmdline	string	The command line.
pname	string	The name of the parent process.
name	string	The name of the process.
pid	string	The ID of the process.
user	string	The username.
md5	string	The MD5 hash value of the process file. If the size of the process file exceeds 1 MB, the system does not calculate the MD5 hash value of the process file.

Network connection logs

The following table describes the fields that you can use to query network connection logs.

Field	Data type	Description
uuid	string	The ID of the client.
IP	string	The IP address of the server.
src_ip	string	The source IP address.
src_port	string	The source port.

Field	Data type	Description
proc_path	string	The path of the process file.
dst_port	string	The destination port.
proc_name	string	The name of the process.
dst_ip	string	The destination IP address.
status	string	The status.

22.1.7.5.2. Logical operators

The log retrieval feature supports multiple search conditions. You can add multiple logical operators to one search condition for one log source, or combine multiple search conditions for several log sources by using different logical operators. This topic describes the logical operators that are supported in log retrieval. Examples are provided to help you understand these operators.

The following table describes the logical operators that are supported in log retrieval.

Logical operators

Logical operator	Description
and	Binary operator. This operator is in the format of `query 1 and query 2`, which indicates the intersection of the query results of `query 1` and `query 2`. **Note** If no logical operators are used for multiple keywords, the default operator is AND.
or	Binary operator. This operator is in the format of `query 1 or query 2`, which indicates the union of the query results of `query 1` and `query 2`.
not	Binary operator. This operator is in the format of `query 1 not query 2`, which indicates the results that match `query 1` but do not match `query 2`. This format is equivalent to `query 1 - query 2`. **Note** If you use only `not query 1`, the log data that does not contain the query results of `query 1` is returned.

22.1.7.5.3. Query logs

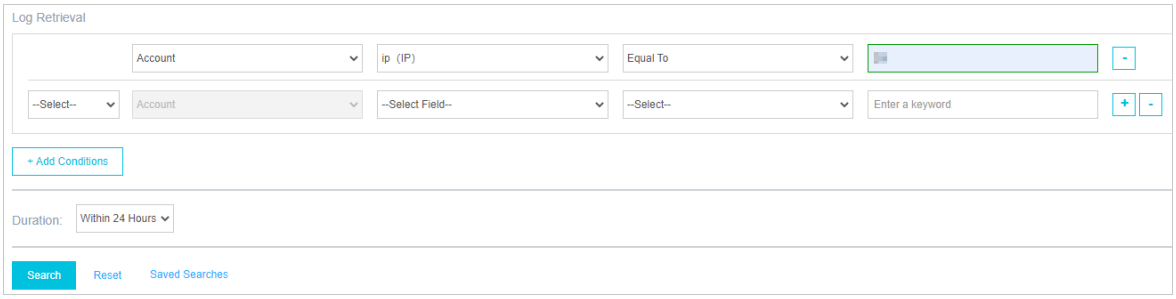
This topic describes how to search for and view physical server logs.

Procedure

1. Log on to Apsara Stack Security Center by using an Apsara Stack tenant account.

> **Note** For more information about the Apsara Stack tenant account, see Create and grant permissions to a security administrator account.

2.
3. click **Log Retrieval**.
4. Specify search conditions.



> **Note** For more information about log sources, log fields, and logical operators, see Supported log sources and fields and Inference rules and logical operators.

5. Click **Search** and view the search result.
   - **Reset**: Click **Reset** to clear the search condition configurations.
   - **Save Search**: Click **Save Search** to save the search condition configurations which you can use to search for logs in the future.
   - **Saved Searches**: Click **Saved Searches** to select and use a search condition that you saved.

22.1.7.6. Configure security settings for physical servers

This topic describes how to configure security settings for physical servers. You can enable or disable periodic trojan scans. You can also specify the working mode of the Server Guard agent.

Procedure

1. Log on to Apsara Stack Security Center by using an Apsara Stack tenant account.

   > **Note** For more information about the Apsara Stack tenant account, see Create and grant permissions to a security administrator account.

2. In the left-side navigation pane, choose **Physical Server Security > Settings**.
3. Enable periodic trojan scans for physical servers.
   i. In the Trojan Scan section, click **Manage**.
   ii. In the All Servers section, select the physical servers on which you want to perform periodic trojan scans. Then, click the rightwards arrow.
   iii. Click **OK**.
4. On the **Protection First Mode** page, click **Manage** next to Protection Mode.

   Configure protection modes for servers.
   - **Business First Mode**: In this mode, the peak CPU utilization is less than 10%, and the peak memory usage is less than 50 MB.
   - **Protection First Mode**: In this mode, the peak CPU utilization is less than 20%, and the peak memory usage is less than 80 MB.

22.1.8. Application security

22.1.8.1. Quick start

This topic helps you get started with the features of Web Application Firewall (WAF).

WAF uses intelligent semantic analysis algorithms to identify web attacks. WAF also uses a learning model to enhance its analysis capabilities and meet your daily security protection requirements without relying on traditional rule libraries.

The following content describes the procedure for using WAF:

1. Customize WAF protection rules.

   WAF provides default protection policies. You can also customize policies that suit your business requirements.

   - For more information about how to configure protection policies, see Configure protection policies.
   - For more information about how to configure custom rules, see Create a custom rule.
   - For more information about how to configure HTTP flood protection rules, see Configure an HTTP flood protection rule.

2. Add websites that you want to protect.

   WAF can protect Internet websites and virtual private cloud (VPC) websites.

   - For more information about how to add an Internet website to WAF for protection, see Add an Internet website for protection.
   - For more information about how to add a VPC website to WAF for protection, see Add a VPC website for protection.

3. Configure Domain Name System (DNS) resolution.

   For more information about how to change the DNS-resolved source IP address for a website to a virtual IP address assigned by WAF, see Modify DNS resolution settings.

4. View WAF protection results.

   - For more information about how to view the protection overview, see View protection overview.
   - For more information about how to view the service access information, see View Web service access information.
   - For more information about how to view the detection logs for web attacks, see View attack detection logs.
   - For more information about how to view the detection logs for HTTP flood attacks, see View HTTP flood protection logs.

22.1.8.2. Detection overview

22.1.8.2.1. View protection overview

This topic describes how to view the Web Application Firewall (WAF) protection overview.

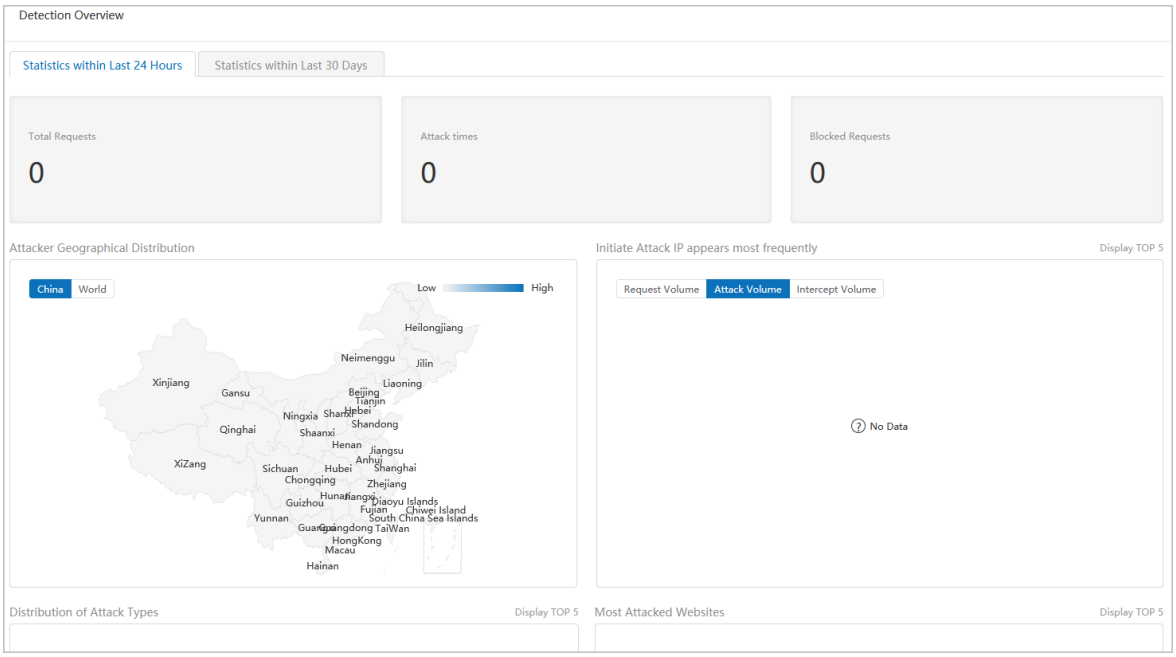
Context

The Detection Overview page displays information such as the statistics of previous attacks, the geographical distribution of attackers, the total number of requests, and the number of blocked requests. You can also view details about the attacks. This way, you can customize rules to protect your web services.

Procedure

1.

2.
3. click **Detection Overview**.
4. On the **Detection Overview** page, view data on the **Statistics within Last 24 Hours** and **Statistics within Last 30 Days** tabs.



- Total Requests

  Displays the total number of requests.

- Attack times

  Displays the total number of attacks.

- Blocked Requests

  Displays the number of blocked requests.

- Attacker Geographical Distribution

  Displays the distribution of attackers on a map. You can select a map of China or a map of the world.

  Displays both the numbers of total requests and blocked requests.

- Initiate Attack IP appears most frequently (Display TOP 5)

  Displays the top five IP addresses from which the most attacks are launched in a bar chart. The x-axis indicates the numbers of requests. The y-axis indicates the IP addresses.

- Distribution of Attack Types (Display TOP 5)

  Displays the distribution of the top five attack types and the number of attacks of each type in a bar chart.

- Most Attacked Websites (Display TOP 5)

  Displays the top five attacked websites and the number of attacks on each website in a bar chart.

22.1.8.2.2. View access information

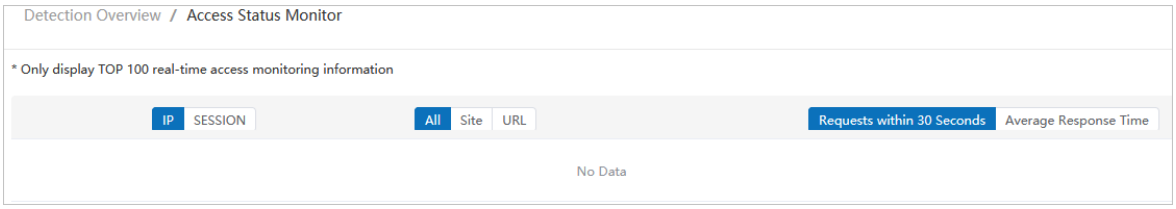
This topic describes how to view access information about web services.

Context

Web Application Firewall (WAF) monitors the access of web services. This way, security administrators can analyze the service access information to detect vulnerabilities and improve security of the services.

Procedure

1.
2.
3. choose **Statistic > Access Status Monitor**.
4. Filter access records to view details.



22.1.8.3. Protection logs

22.1.8.3.1. View attack detection logs

This topic describes how to view attack detection logs.

Context

These logs allow you to analyze attacks on your web services. You can update the protection policies and custom rules, and fix the web service vulnerabilities based on the analysis results.

Procedure

1.
2.
3. choose **Detection Logs > Attack Detection Logs**.
4. Click **Filter**, specify filter conditions, and then click **OK**.

> **Note** If you specify multiple conditions, they are evaluated by using a logical AND. The system returns the required logs only when all the conditions are met.

5. View the attack detection logs.

22.1.8.3.2. View HTTP flood protection logs

This topic describes how to view HTTP flood protection logs.

Context

These logs allow you to analyze HTTP flood attacks on your web services. In addition, you can update the HTTP flood protection rules and HTTP flood whitelist, and fix the web service vulnerabilities based on the analysis results.

Procedure

1.
2.
3. choose **Detection Logs > HTTP Flood Detection Logs**.
4. Click **Filter**, specify filter conditions, and then click **OK**.

> **Note** If you specify multiple conditions, they are evaluated by using a logical AND. The system returns the required logs only when all the conditions are met.

5. View the HTTP flood detection logs.

   The blocked HTTP flood attacks, related rules, and attack time are displayed.

22.1.8.3.3. View system operation logs

This topic describes how to view system operation logs.

Procedure

1.
2.
3. choose **Detection Logs > System operation log**.
4. View the system operation logs.

   The usernames, content, IP addresses, and creation time are displayed.

22.1.8.3.4. View access logs

This topic describes how to view access logs.

Procedure

1.
2.
3. choose **Detection Logs > Access Log**.
4. Click **Filter**, specify filter conditions, and then click **OK**.

> **Note** If you specify multiple conditions, they are evaluated by using a logical AND. The system returns the required logs only when all the conditions are met.

5. View the access logs.

   The requested addresses, destination IP addresses, source IP addresses, methods, response status codes, and time are displayed.

22.1.8.4. Protection configuration

22.1.8.4.1. Configure protection policies

This topic describes how to configure Web Application Firewall (WAF) protection policies.

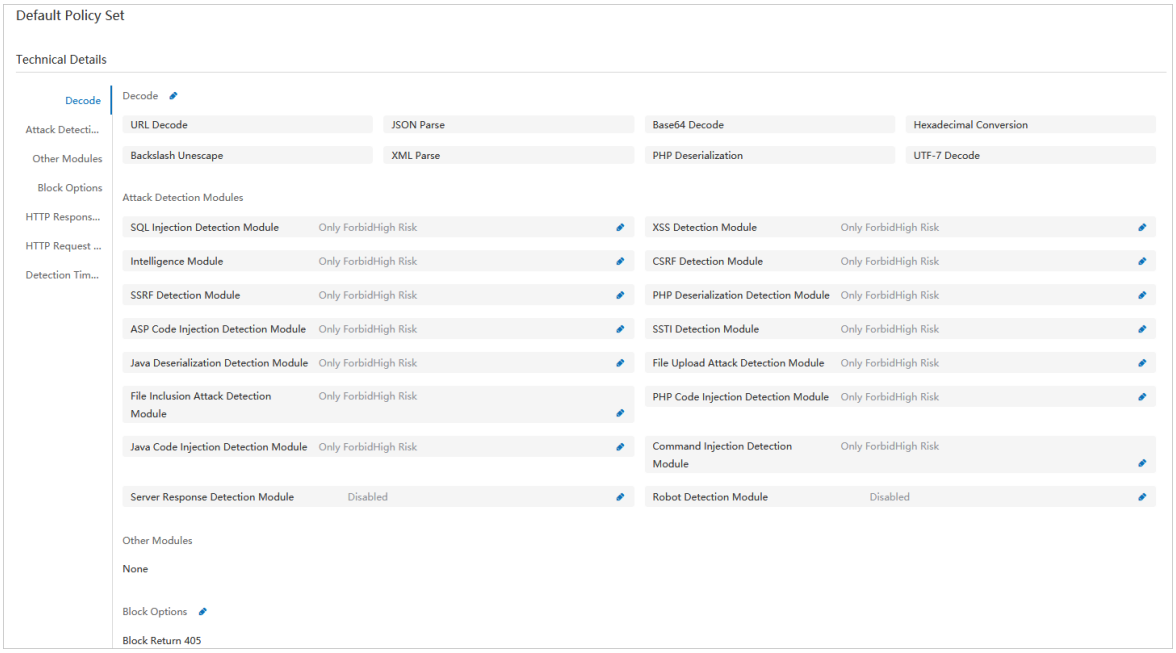
Context

WAF provides a default protection policy. You can also customize protection policies to suit your business requirements.

Procedure

1.
2.
3. choose **Protection Configuration > Website Protection Policies**.
4. Click **Add a protection policy**. In the panel that appears, configure **Policy name** and click **Confirm**.

5. In the **Actions** column of the new protection policy, click the icon to view details.



Parameter	Description
Decode	Select algorithms that you want to use to decode the requests.
Attack Detection Modules	Specify the types of attacks that you want to detect and the risk levels of attacks that you want to block.
Block Options	Specify the HTTP status code and image that you want WAF to return when it blocks an attack.
HTTP Response Detection	Configure **Enable HTTP response processing** and **Response Detection Max Body Size**.
HTTP Request Detection	Configure **Response Detection Max Body Size**.
Detection Timeout	Configure **Enable Detection Timeout** and **Timeout Threshold**.

For example, perform the following steps to configure modules in the **Attack Detection Modules** section:

i. Move the pointer over a specific module in the **Attack Detection Modules** section. In this example, move the pointer over **SQL Injection Detection Module** and click the **modify** icon.

ii. In the **SQL Injection Detection Module** panel, configure the following parameters.

Parameter	Description
Enabled	Specify whether to enable the detection module.
Blocking Threshold	Valid values: **NotForbid**, **Only ForbidHigh Risk**, **ForbidMedium or High Risk**, and **Forbid All**.
Record Threshold	Valid values: **Notrecord**, **Only recordHigh Risk**, **recordMedium or High Risk**, and **record All**.
Detect Non-Injected SQL	Specify whether to enable detection for NoSQL injection vulnerabilities.

iii. Click **OK**.

6. Manage protection policies.

   To delete a protection policy, select the protection policy. Then, in the upper-right corner, choose **More > Delete Selected Protection Policies**. In the message that appears, click **OK**.

   > **Note** You cannot delete the default protection policy.

22.1.8.4.2. Create a custom rule

This topic describes how to create a custom rule for Web Application Firewall (WAF).

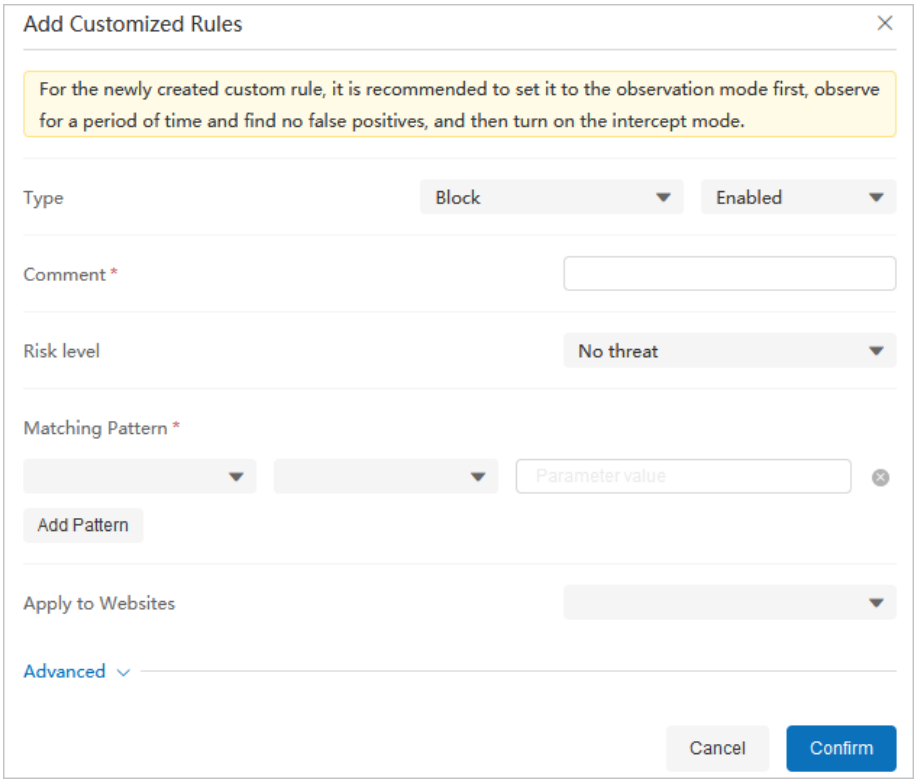
Context

You can create custom rules to meet different requirements for intrusion detection. You can create, edit, or delete custom rules as an administrator. You can use custom rules to filter out requests that meet specific conditions.

Multiple custom rules are evaluated by using a logical **OR**. If two custom rules use the same conditions but trigger different actions such as blocking traffic or allowing traffic, WAF runs the first rule.

Procedure

1.
2.
3. In the left-side navigation tree of the **WAF** page, choose **Protection Configuration > Customized Rules**.
4. In the upper-right corner, click **Add Rule**. In the **Add Customized Rules** panel, configure the parameters.



Parameters used to create a custom rule

Parameter	Description
Type	The operating mode of the rule. Valid values: **Block**, **Allow**, **Monitor**, and **Detection module control**. ○ **Block**: If an HTTP request meets the conditions of the rule, the HTTP request is blocked. ○ **Allow**: If an HTTP request meets the conditions of the rule, the HTTP request is allowed. ○ **Monitor**: If an HTTP request meets the conditions of the rule, the HTTP request is recorded and allowed. ○ **Detection module control**
Comment	The remarks about the rule. We recommend that you enter the purpose of the rule.
Risk level	The risk level. Valid values: **No threat**, **Low Risk**, **Medium Risk**, and **High Risk**.
Matching Pattern	The conditions that trigger the rule. Click **Add Pattern** to specify more than one condition. Multiple conditions are evaluated by using a logical **AND**. The custom rule takes effect only when all conditions are met.
Apply to Websites	The websites that you want the rule to protect.
Log Recording Option	Specifies whether to record a log when the rule is triggered. The default value is Enable Log Recording. After Log Recording Option is set to Enable Log Recording, all interception events are recorded in the intrusion detection logs.
Attack Type	The type of attack that you want the rule to block.

Parameter	Description
Expiration Time	The time at which the rule expires.

5. Click **Confirm**.
6. Manage custom rules.
   - Edit a rule.

     To edit a rule, click the icon in the **Actions** column.
   - Enable a rule.

     To enable a rule that is disabled, select the rule and choose **More** > **Enable Selected Rules**.
   - Disable a rule.

     To disable a rule that is enabled, select the rule and choose **More** > **Disable Selected Rules**.
   - Export a rule.

     To export a rule, select the rule and choose **More** > **Export Selected Rules**.
   - Delete a rule.

     To delete a rule that you no longer need, select the rule and choose **More** > **Delete Selected Rules**.

22.1.8.4.3. Configure an HTTP flood protection rule

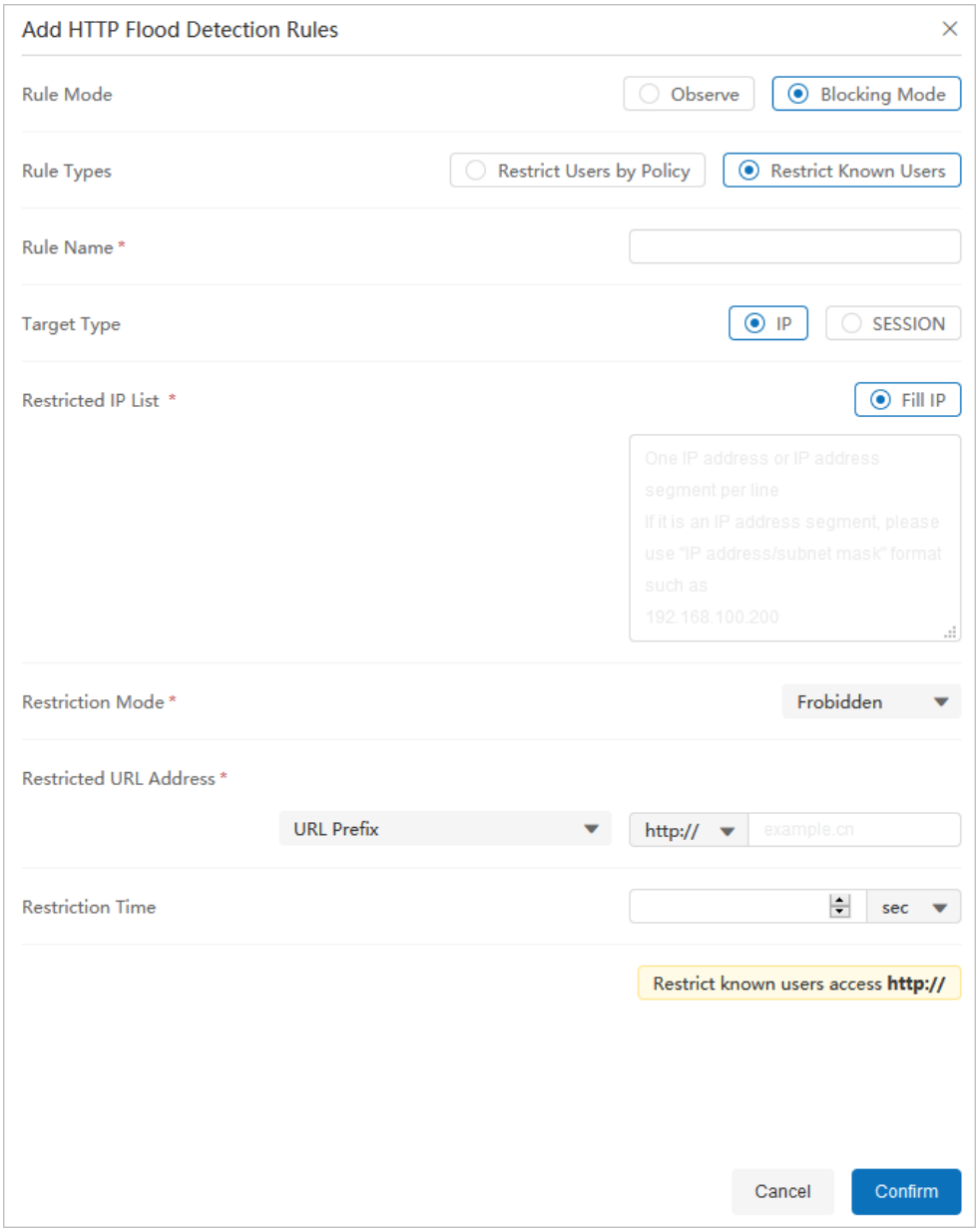
This topic describes how to configure an HTTP flood protection rule.

Context

An HTTP flood attack is a type of DDoS attack that targets the application layer. Attackers use proxy servers or zombies to overwhelm targeted web servers by sending a large number of HTTP requests.

Create an HTTP flood protection rule

1.
2.
3. choose **Protection Configuration** > **HTTP Flood Detection**.
4. Click **Add Rule**. The **Add HTTP Flood Detection Rules** panel appears.
5. Configure parameters and click **Confirm**.



Parameter	Description
Rule Mode	The action on requests after the HTTP flood protection rule is triggered. Valid values: **Blocking Mode** and **Observe**. ○ Blocking Mode: limits the requests that trigger the HTTP flood protection rule. ○ Observe: records the requests that trigger the HTTP flood protection rule, but does not limit the requests.

Parameter	Description
Rule Types	The type of the HTTP flood protection rule. Valid values: **Restrict Users by Policy** and **Restrict Known Users**. The difference between the two types is determined by whether requests of users are initiated from a specific IP address or in a specific session. ○ **Restrict Users by Policy**: limits requests that meet all the configuration items of the HTTP flood protection rule. Configuration items include **Restriction Trigger Threshold**, **Restricted URL Address**, **Restriction Mode**, **Restriction Time**, and **Statistical Range of Visits** in the Advanced section. ○ **Restrict Known Users**: limits requests that are initiated from specific IP addresses or in specific sessions based on the HTTP flood protection rule. To achieve this purpose, you must configure the IP address or session list and the limit mode. After you configure the list, the HTTP flood protection rule limits requests based on the list.
Rule Name	The name of the HTTP flood protection rule.
Target Type	The type of source for requests that are limited. Valid values: **IP** and **SESSION**. **Note** If you set **Target Type** to **SESSION**, you can apply the HTTP protection rule only to a website whose **User Identification** is set to **WAF User System**. For more information, see Add an Internet website for protection.
Restriction Trigger Threshold	If you set **Rule Types** to **Restrict Users by Policy**, you must configure the triggering conditions for the HTTP flood protection rule.
Restricted URL Address	If you set **Rule Types** to **Restrict Users by Policy**, you must specify the URL addresses that are protected based on the HTTP flood protection rule. ○ **URL Prefix** ○ **URL** ○ **Record all IP addresses**
Restricted IP List or **Restricted SESSION List**	If you set **Rule Types** to **Restrict Known Users**, you must enter the IP addresses or sessions from which you want to limit requests based on the setting of **Target Type**. You can enter only one IP address or session in each line.
Restricted URL Address	If you set **Rule Types** to **Restrict Known Users**, you must specify the URL addresses that are protected based on the HTTP flood protection rule. ○ **URL Prefix** ○ **URL** ○ **Restrict user access to all addresses**

Parameter	Description
Restriction Mode	The mode in which the HTTP flood protection rule limits requests. Valid values: ○ **Forbidden**: The rule blocks specific sources from accessing the specified URL address. ○ **Frequency control**: The rule limits the frequency at which specific sources access the specified URL address.
Restriction Time	The time at which the action specified in the HTTP flood protection rule takes effect.
Statistical Range of Visits	If you set **Rule Type** to **Restrict Users by Policy**, you can specify the range of data records to limit requests in the Advanced section. ○ Statistics Full Access Data: If you select this option, the frequency of requests is limited when the requests are forwarded to WAF and meet the HTTP flood protection rule, regardless of which data records the requests access. This decreases system performance. ○ Statistics TOP Access Data: If you select this option, the frequency of requests is limited when the requests meet the preceding conditions and access the top 100 data records. This option helps minimize the decrease in system performance. You can select this option when the number of accessed data records is larger than 100. Note that the top 100 data records are measured based on real-time monitoring.

Manage HTTP flood protection rules

1.
2.
3. choose **Protection Configuration > HTTP Flood Detection**.
4. In the rule list, manage existing HTTP flood protection rules. Modification is allowed.
   - Search for a rule.

     Click **Filter**. In the **Filter Item** panel, specify filter conditions.
   - Enable a rule.

     Select a rule that is disabled and choose **More > Enable Selected Rules**.
   - Disable a rule.

     Select a rule that is enabled and choose **More > Disable Selected Rules**.
   - Delete a rule.

     Select a rule and choose **More > Delete Selected Rules**.

22.1.8.4.4. Configure the HTTP flood whitelist

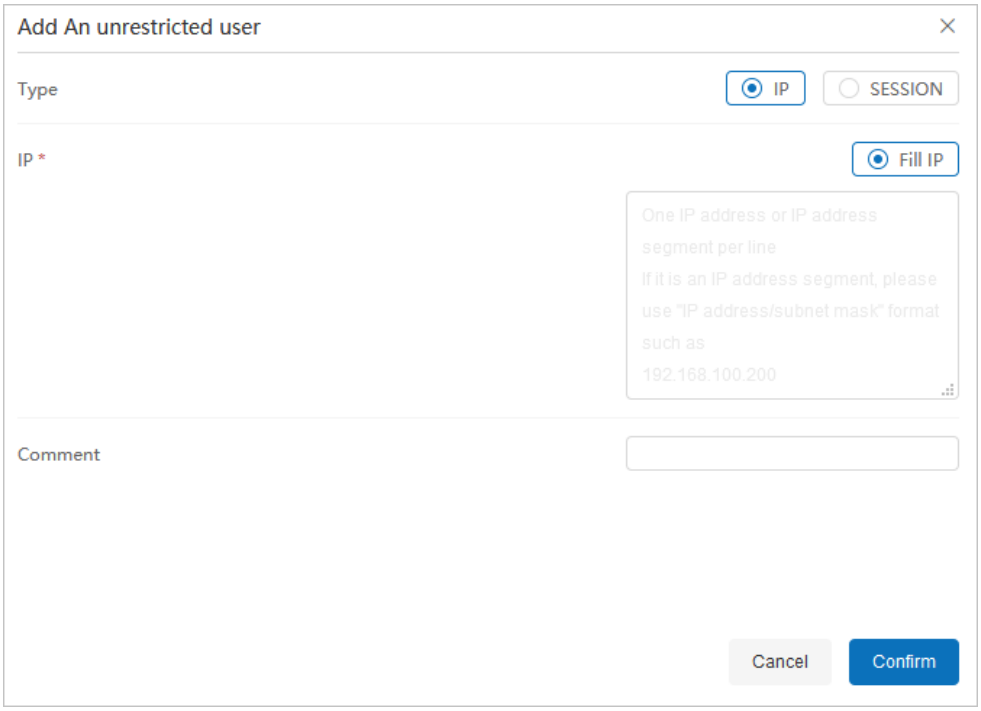
This topic describes how to configure the HTTP flood whitelist.

Context

If a request source is trusted, you can add this request source to the HTTP flood whitelist to allow all requests from this source.

Procedure

1.
2.
3. In the left-side navigation tree of the **WAF** page, choose **Protection Configuration > HTTP Flood Detection**.
4. On the **HTTP Flood Detection Whitelist** tab, click **Add Whitelist Item**, add a trusted request source, and then click **Confirm**.



Parameter	Description
Type	Select the type of the request source. Valid values: **IP** and **SESSION**.
IP or **SESSION**	Specify the IP addresses or sessions based on the setting of **Type**. You can enter only one IP address or session in each line.
Comment	Enter remarks for the request source.

5. Manage request sources in the whitelist.
   - Search for a request source in the whitelist.

     Click **Filter**. In the panel that appears, specify a filter condition or click **Add Filter Item** to specify more filter conditions.
   - Remove a request source from the whitelist.

     Select the request source and choose **More > Delete Selected Items**.

22.1.8.4.5. Manage SSL certificates

This topic describes how to upload or delete SSL certificates.

Context

After you upload an SSL certificate on the **Certificate Management** page, you can select this certificate when you add an HTTPS website for protection on the **Protection Site Management** page.

> **Note** When you add an HTTPS website for protection on the **Protection Site Management** page, you must select the SSL certificate that corresponds to the domain of the HTTPS website.

Procedure

1.
2.
3. choose **Protection Configuration > Certificate Management**.
4. Upload a new certificate.
   i. Click **Upload SSL Certificate**.
   ii. In the **Name** field, enter a **name for the new certificate**.
   We recommend that you enter the domain name for easier management.

   > **Note** If your Certificate Authority (CA) certificate and private key are in the same file, select **Include private key in certificate file**.

   iii. In the **File** section, upload the CA certificate file and private key file.
   iv. Configure **Certificate Password**.
   v. Click **Confirm**.
5. (Optional)Delete the uploaded SSL certificate.
   You can delete expired SSL certificates.
   i. In the SSL certificate list, select the certificate that you want to delete.
   ii. Choose **More > Delete selected SSL certificate**.
   iii. In the message that appears, click **OK**.

22.1.8.4.6. Add Internet websites for protection

This topic describes how to add Internet websites to Web Application Firewall (WAF).

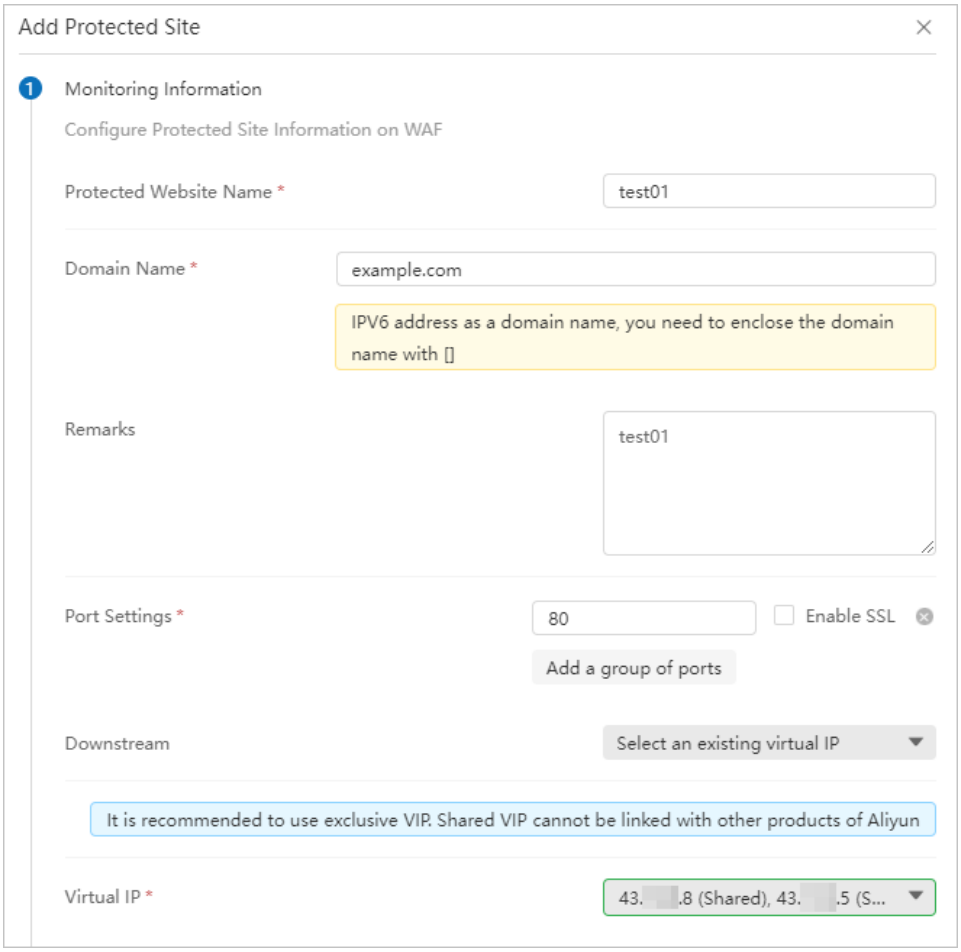
Context

WAF can protect the following types of websites:

- Internet websites.
- Virtual Private Cloud (VPC) websites. For more information about how to add VPC websites to WAF, see Add a VPC website for protection.

Procedure

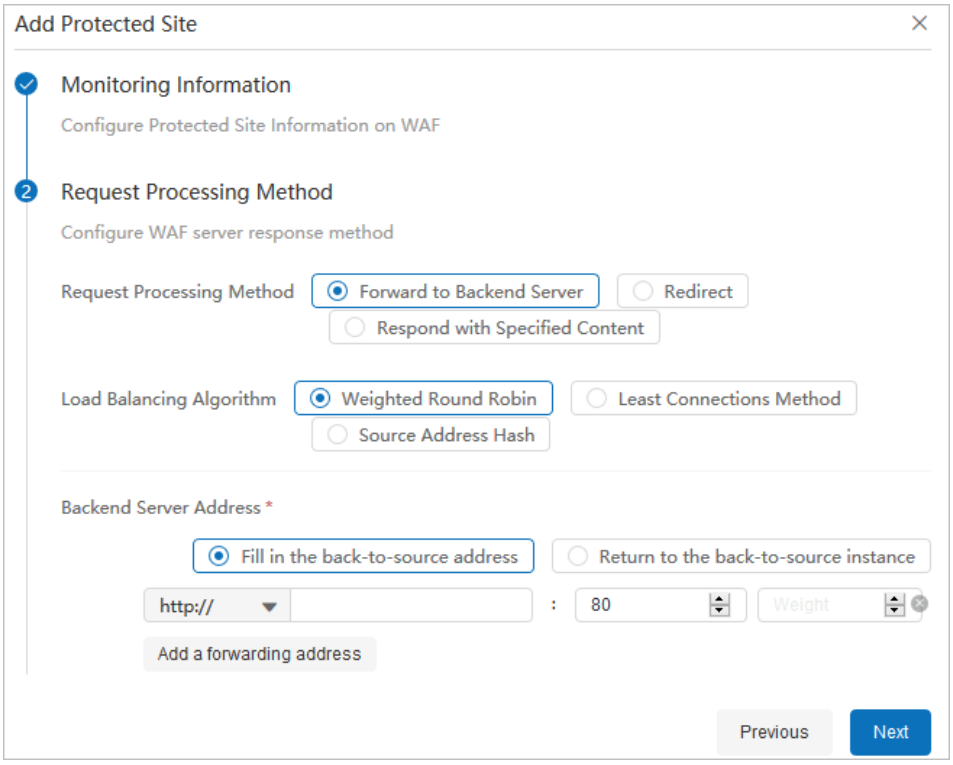
1.
2.
3. , choose **Protection Configuration > Protection Site Management**.
4. On the **Internet Websites** tab, click **Add a site**.
5. In the Monitoring Information step, configure parameters and click **Next**.
   Specify the Internet website that you want WAF to protect. WAF can protect both HTTP and HTTPS websites.



Parameter	Description
Protected Website Name	The name of the website that you want WAF to protect.
Domain Name	The domain name of the website. ◦ You can use an asterisk (*) as a wildcard. ◦ If you specify multiple domain names, separate them with commas (,).
Port Settings	The port that WAF listens on. ◦ If the website supports HTTPS requests, select **Enable SSL** and upload an HTTPS certificate. ◦ If the website can be accessed over multiple ports, click **Add a group of ports** to add the required ports.

Parameter	Description
Cert Setting	The HTTPS certificate of the website. Valid values: *Upload a New Certificate* and *Choose an Existing Certificate*. ○ *Upload a New Certificate*: If the HTTPS certificate of the website has not been uploaded to WAF, select this option. By default, the HTTPS certificate and private key are separately uploaded. If you select **Include private key in certificate file**, upload only one file that contains both the HTTPS certificate and private key. ○ *Choose an Existing Certificate*: If the HTTPS certificate of the website has been uploaded to WAF, select this option. Then, select the required HTTPS certificate from the drop-down list. **Note** This parameter is required only if you select **Enable SSL** next to the **listening port** field.
Name	The name of the HTTPS certificate. **Note** This parameter is required only if you select **Enable SSL** next to the **listening port** field and select Upload a New Certificate.
File	The HTTPS certificate and private key. By default, the HTTPS certificate and private key are separately uploaded. If you select **Include private key in certificate file** next to **Name**, upload only one file that contains both the HTTPS certificate and private key. **Note** This parameter is required only if you select **Enable SSL** next to the **listening port** field and select Upload a New Certificate.
Virtual IP	The IP address type and virtual IP address. **Note** You can select an IPv6 address as the virtual IP address for WAF. By default, WAF provides 10 virtual IP addresses. You can add more virtual IP addresses based on your business requirements. **Note** A virtual IP address is available only for the department to which the creator of the virtual IP address belongs.

6. In the Request Processing Method step, configure parameters and click **Next**.



Response mode	Parameter	Description
Forward to Backend Server	**Load Balancing Algorithm**	The algorithm for load balancing. Valid values: **Weighted Round Robin**, **Source Address Hash**, and **Least Connections Method**.
	Backend Server Address	The address of the origin server to which WAF forwards inbound traffic. Valid values: **Fill in the back-to-source address** and **Return to the back-to-source instance**. ○ **Fill in the back-to-source address**: Enter the address of the origin server. If you enter multiple addresses, load balancing is performed based on the specified **load balancing algorithm**. ○ **Return to the back-to-source instance**: Enter the address of a specific ECS or SLB instance. If you enter multiple addresses, load balancing is performed based on the specified **load balancing algorithm**.
	X-Forwarded-For	The pass-through mode of the actual source IP address. The X-Forwarded-For (XFF) header is used to identify the actual source IP address of an HTTP client. The header is used for traffic forwarding services, such as HTTP proxy and load balancing.

Response mode	Parameter	Description
Redirect	**Response Status Code**	The HTTP status code that WAF returns when it redirects inbound traffic to a specified address. Valid values: **301**, **302**, **307**, and **308**. ○ **301**: The requested page is permanently moved to another URL. ○ **302**: The requested page is temporarily moved to another URL. The requester must continue to use the original URL for future requests. ○ **307**: The requested page is temporarily moved to another URL. The requester must continue to use the original URL for future requests.
	Redirect address	The destination URL for redirection.
Respond with Specified Content	**Response Status Code**	The HTTP status code that WAF returns when it returns specified content. Valid value: **200**, **400**, **401**, **402**, **404**, **405**, **500**, **503**, and **504**.
	Response	The content to return. For example, you can upload an image for the **Response** parameter. If a user visits the website, WAF returns the uploaded image.

7. In the Protection Policy step, configure parameters and click Next. Then, go to the **Finish** step.

> **Note** You can configure a protection policy only if you set **Request Processing Method** to **Forward to Backend Server**.

Parameter	Description
Protection Policy	Select a WAF protection policy. For more information, see Configure protection policies.
User Identification	Specify whether to enable the user identification feature. **Note** If you enabled HTTP flood protection for the protected website and set **Target Type** to **SESSION** when you configured the HTTP flood protection rule, you must set **User Identification** to **WAF User System**.

22.1.8.4.7. Add VPC websites for protection

This topic describes how to add virtual private cloud (VPC) websites to Web Application Firewall (WAF) for protection.

Context

WAF can protect the following types of websites:

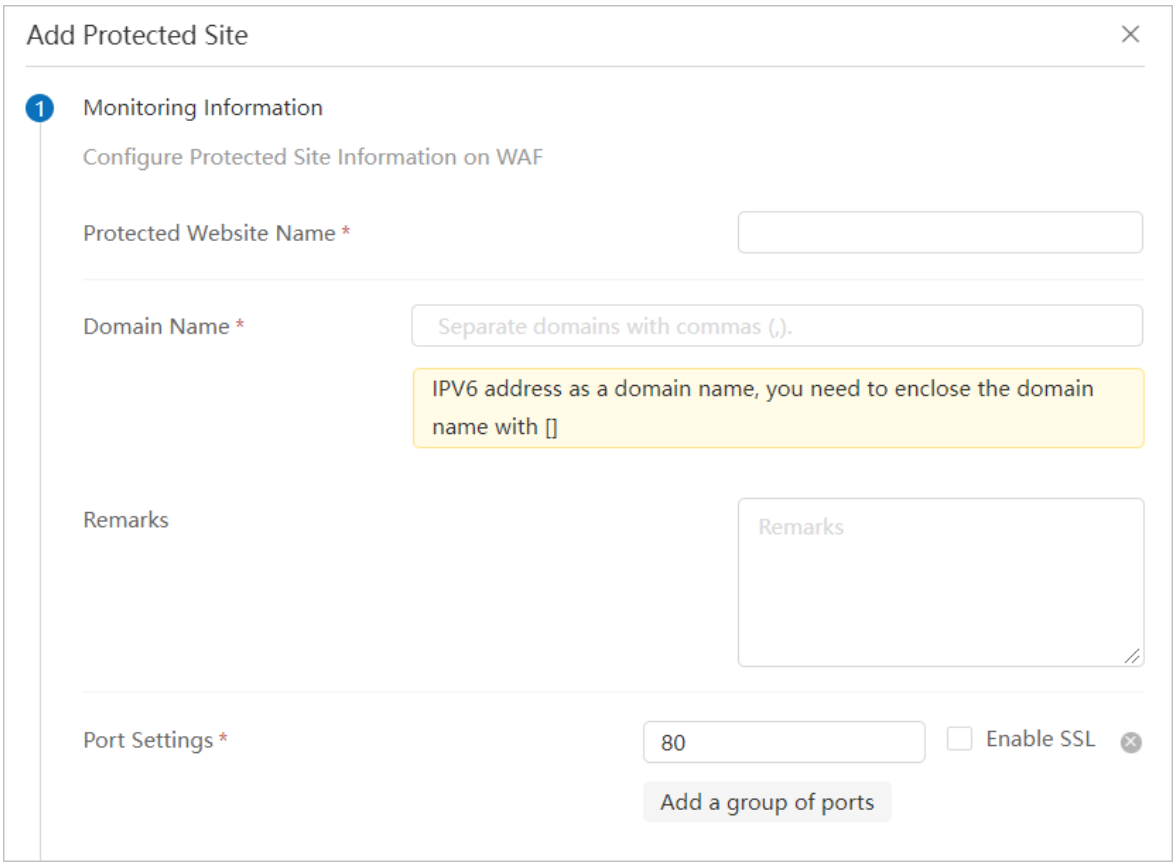
- Internet websites. For more information about how to add an Internet website for protection, see Add an Internet website for protection.

- VPC websites.

Procedure

1. Log on to Apsara Stack Security Center.
2. Choose **Application Security > Web Application Firewall**.
3. In the left-side navigation pane, choose **Protection Configuration > Protection Site Management**.
4. Click the **VPC Websites** tab. Then, click **Add a site**. The **Add Protected Site** panel appears.
5. In the Monitoring Information step, configure parameters and click **Next**.

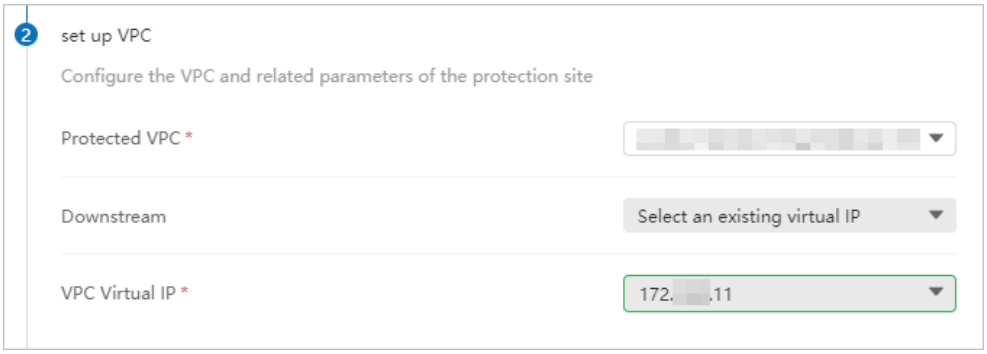
   Specify the VPC website that you want WAF to protect. WAF can protect both HTTP and HTTPS websites.



Parameter	Description
Protected Website Name	The name of the website that you want WAF to protect.
Domain Name	The domain name of the website. ○ You can use an asterisk (*) as a wildcard. ○ If you specify multiple domain names, separate them with commas (,).
Port Settings	The port that WAF listens on. ○ If the website supports HTTPS requests, select **Enable SSL** and upload an HTTPS certificate. ○ If the website can be accessed over multiple ports, click **Add a group of ports** to add the required ports.

Parameter	Description
Cert Setting	The HTTPS certificate of the website. Valid values: **Upload a New Certificate** and **Choose an Existing Certificate**. **Note** Specify this parameter only if you select **Enable SSL** next to **Port Settings**. ○ **Upload a New Certificate**: If the HTTPS certificate used by the website is not uploaded to WAF, select this option. By default, the HTTPS certificate and private key are separately uploaded. If you select **Include private key in certificate file**, upload only a file that contains both the HTTPS certificate and private key. ○ **Choose an Existing Certificate**: If the HTTPS certificate used by the website is uploaded to WAF, select this option. Then, select the HTTPS certificate from the drop-down list.
Name	The name of the HTTPS certificate. **Note** Specify this parameter only if you select **Enable SSL** next to **Port Settings** and set Cert Setting to Upload a New Certificate.
File	The HTTPS certificate and private key to upload. By default, the HTTPS certificate and private key are separately uploaded. If you select **Include private key in certificate file** next to **Name**, upload only a file that contains both the HTTPS certificate and private key. **Note** Specify this parameter only if you select **Enable SSL** next to **Port Settings** and set Cert Setting to Upload a New Certificate.

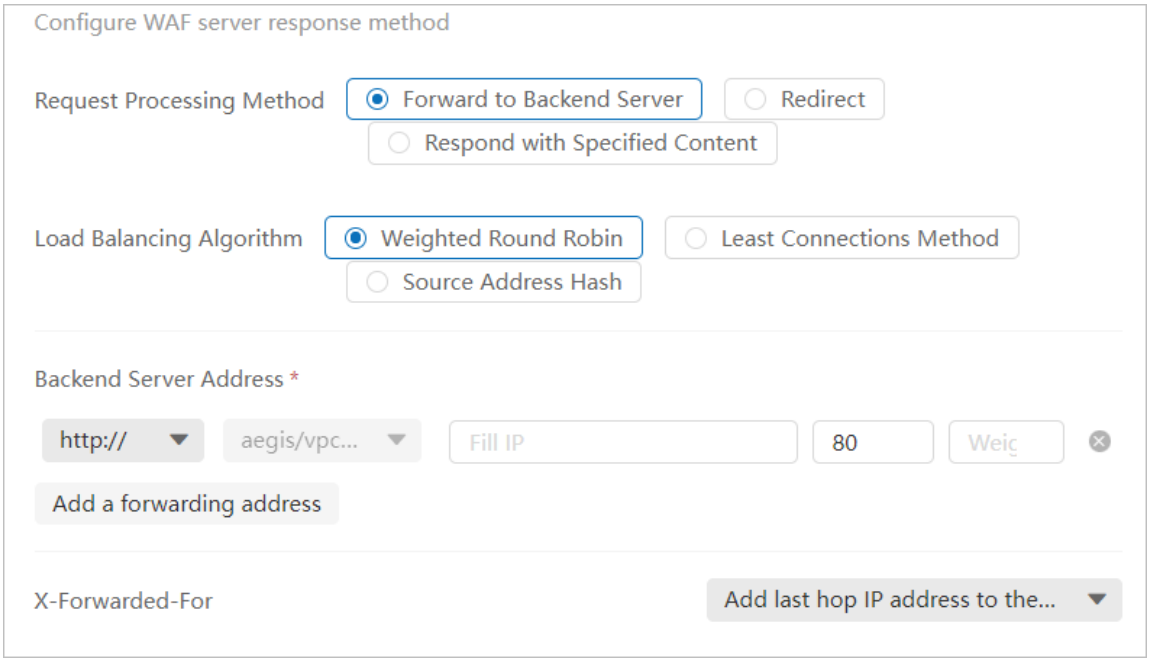
6. In the set up VPC step, configure parameters and click **Next**.



Parameter	Description
Protected VPC	The VPC to which the website belongs.
Virtual Switch	The vSwitch associated with the specified VPC.
Create Virtual IP Method	The method to create a virtual IP address. Valid values: **Select an existing virtual IP** and **Create virtual IP**.

Parameter	Description
VPC Virtual IP	○ If you set **Create Virtual IP Method** to **Select an existing virtual IP**, select an existing virtual IP address from the **VPC Virtual IP** drop-down list. ○ If you set **Create Virtual IP Method** to **Create virtual IP**, click **Click to Create Vip** next to **VPC Virtual IP** to generate a virtual IP address.

7. In the Request Processing Method step, configure parameters and click **Next**.



Request processing method	Parameter	Description
Forward to Backend Server	**Load Balancing Algorithm**	The algorithm for load balancing. Valid values: **Weighted Round Robin**, **Source Address Hash**, and **Least Connections Method**.
	Backend Server Address	The address of the origin server to which WAF forwards inbound traffic. Valid values: **Fill in the back-to-source address** and **Return to the back-to-source instance**. ○ **Fill in the back-to-source address**: Enter the address of the origin server. If you enter multiple addresses, load balancing is performed based on the specified **load balancing algorithm**. ○ **Return to the back-to-source instance**: Enter the address of a specific Elastic Compute Service (ECS) or Server Load Balancer (SLB) instance. If you enter multiple addresses, load balancing is performed based on the specified **load balancing algorithm**.

Request processing method	Parameter	Description
	X-Forwarded-For	The pass-through mode of the actual source IP address. The X-Forwarded-For (XFF) header is used to identify the actual source IP address of an HTTP client. The header is used for traffic forwarding services, such as HTTP proxy and load balancing.
Redirect	**Response Status Code**	The HTTP status code that WAF returns when it redirects inbound traffic to a specified address. Valid values: **301**, **302**, **307**, and **308**. ○ **301**: The requested page is permanently moved to another URL. ○ **302**: The requested page is temporarily moved to another URL. The requester must continue to use the original URL for future requests. ○ **307**: The requested page is temporarily moved to another URL. The requester must continue to use the original URL for future requests.
	Redirect address	The destination URL for redirection.
Respond with Specified Content	**Response Status Code**	The HTTP status code that WAF returns when it returns specified content. Valid value: **200**, **400**, **401**, **402**, **404**, **405**, **500**, **503**, and **504**.
	Response	The content to return. For example, you can upload an image for the **Response with Specified Content** parameter. If a user visits the website, WAF returns the uploaded image.

8. In the Protection Policy step, configure parameters and click Next. Then, go to the **Finish** step.

Parameter	Description
Protection Policy	Select a WAF protection policy. For more information, see Configure protection policies.
User Identification	Specify whether to enable the user identification feature.

22.1.8.4.8. Verify the configurations of a website on your on-premises server

This topic describes how to verify the configurations of a website on your on-premises server.

Context

Before you use Web Application Firewall (WAF) to scrub traffic destined for a website, we recommend that you verify the configurations of the website on your on-premises server. After you add the virtual IP address and the domain of a website to the hosts file on your on-premises server, the request to access the domain from a local browser passes through WAF first.

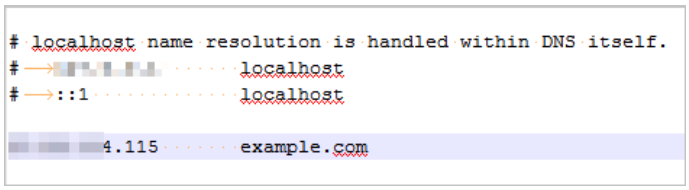
Procedure

1.
2. Add the virtual IP address and domain name to the *hosts* file on your on-premises server.

   If your computer runs Windows 7, the hosts file is stored in the following path: *C:\Windows\System32\drivers\etc\hosts*.

   i. Open the hosts file by using a text editor, such as Notepad.

   ii. Add the following content to the end of the file: `<The virtual IP address that is assigned by WAF><Protected domain name>` .

   > **Note** The IP address preceding the domain name is the virtual IP address that is assigned by WAF.

3. Ping the protected domain name from your on-premises server.

   The returned IP address must be the virtual IP address that is assigned by WAF in the hosts file. If the returned IP address is still the IP address of the origin server, refresh the local Domain Name System (DNS) cache.

4. Enter the domain name in the address bar of your browser and press Enter.

   If the access configurations on WAF are correct, you can visit the website.

5. Verify the protection capability of WAF.

   Simulate a web attack request and check whether WAF blocks the request.

   For example, add `/?alert(xss)` after the URL. If you try to visit `www.example.com/?alert(xss)` , WAF is expected to block the request.

22.1.8.4.9. Modify DNS resolution settings

This topic describes how to modify the Domain Name System (DNS) resolution settings to connect your website to Web Application Firewall (WAF).

Context

Before you can modify the DNS resolution settings, you must verify the settings on your computer and make sure that the settings are correct. Then, the traffic destined for your website can be redirected to WAF after you modify the settings.

The domain name of a protected website may not be resolved by a DNS provider. For example, a website may use a Server Load Balancer (SLB) instance to connect to the Internet. In this case, you can perform the following operation to protect the website by using WAF: Specify the virtual IP address that WAF assigns to your website as the back-to-origin IP address of the SLB instance.

Procedure

1.
2.

3. choose **Protection Configuration > Protection Site Management**.
4. Find the website whose DNS resolution settings you want to modify and click the icon in the **Actions** column.
5. On the **Basic Information** tab, record the virtual IP address assigned to the website.
6. Log on to the console of the DNS provider and find the DNS resolution settings for the domain name of the website. Then, change the IP address in the A record to the virtual IP address assigned to the website.

> **Note** We recommend that you set the TTL to 600 seconds in DNS resolution settings. The larger the TTL is, the longer it takes to synchronize and update DNS records.

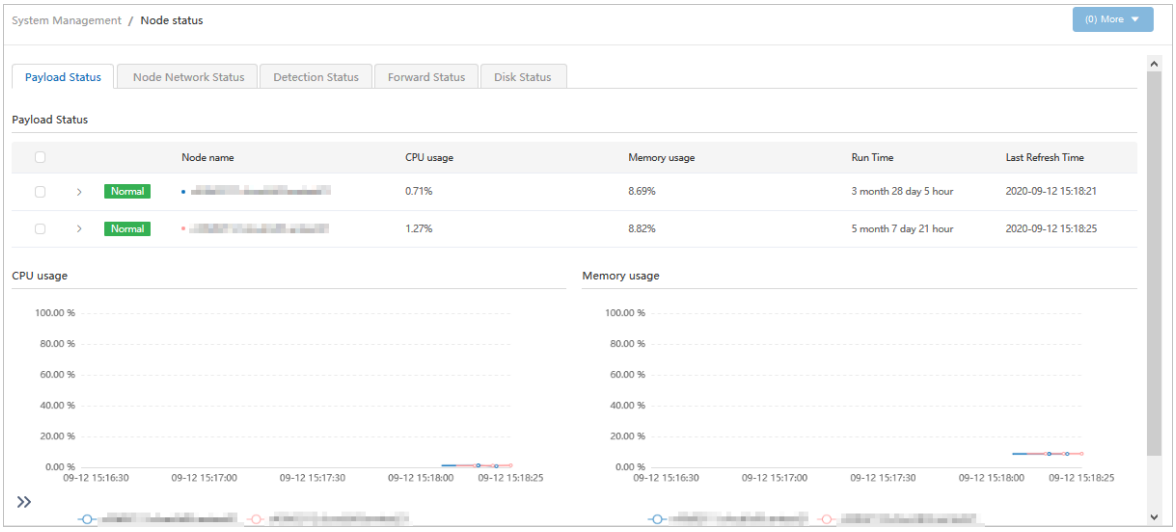
22.1.8.5. System management

22.1.8.5.1. View the load status of nodes

This topic describes how to view the load status of Web Application Firewall (WAF) nodes. The status information includes CPU utilization and memory usage. You can identify faults based on the status and check whether scale-out or scale-up is required.

Procedure

1.
2.
3. In the left-side navigation tree of the **WAF** page, choose **System Info > Node status**.
4. On the **Payload Status** tab, view the load status of WAF nodes.



In the Payload Status section, you can view the CPU utilization and memory usage of WAF nodes. In the CPU usage and Memory usage sections, you can view the changes in CPU utilization and memory usage over a specific period of time.

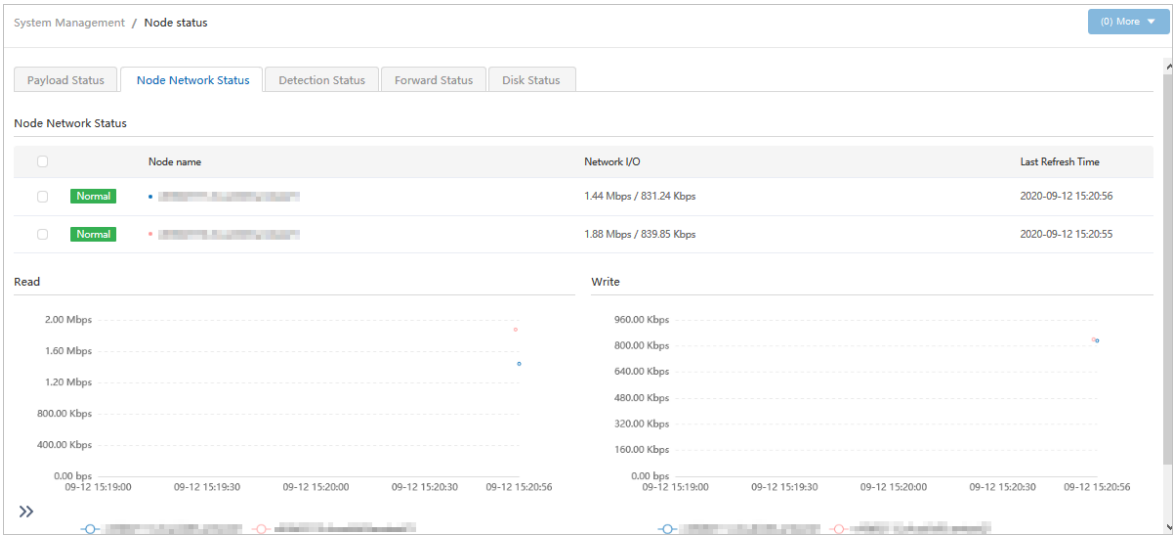
22.1.8.5.2. View the network status of nodes

This topic describes how to view the network status of Web Application Firewall (WAF) nodes. The status information includes network I/O, traffic detection status, and traffic forwarding status.

Node network status

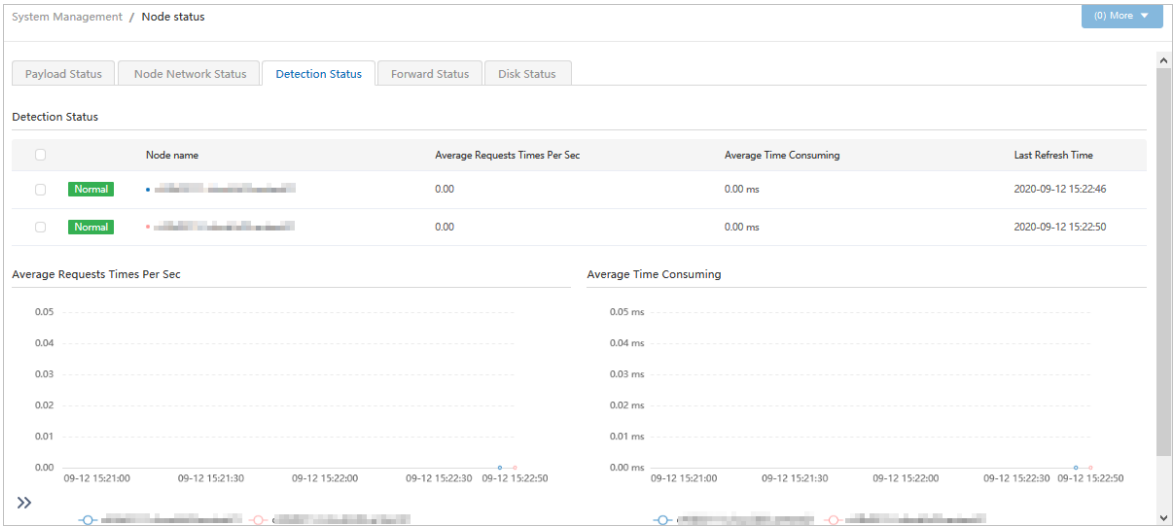
1.

2.
3. choose **System Info > Node status**.
4. On the **Node status** page, click the **Node Network Status** tab.
5. View the network I/O of WAF nodes.



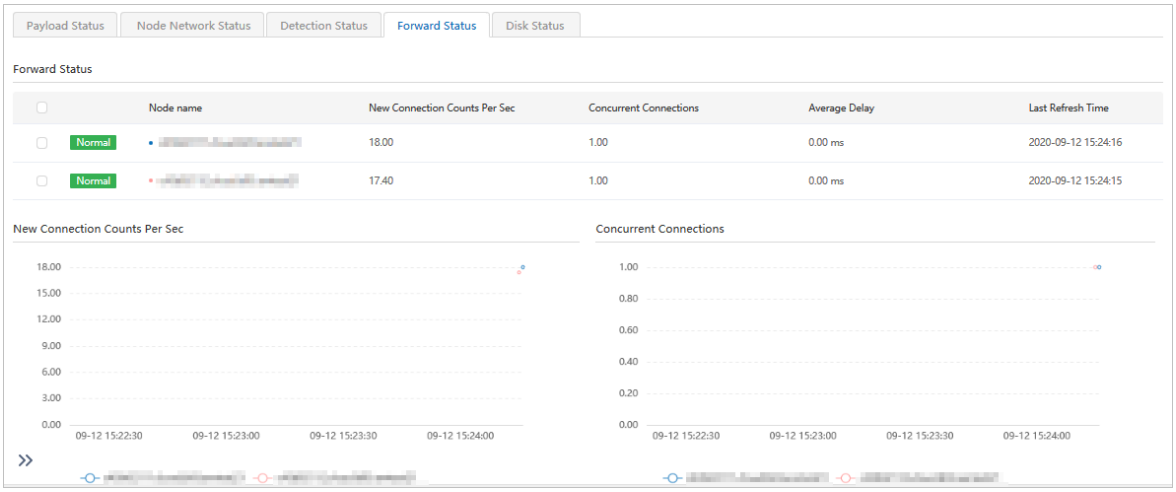
Traffic detection status

1.
2.
3. choose **System Info > Node status**.
4. On the **Node Status** page, click the **Detection Status** tab.
5. View the traffic detection status of WAF nodes.



Traffic forwarding status

1.
2.
3. choose **System Info > Node status**.
4. On the **Node status** page, click the **Forward Status** tab.
5. View the traffic forwarding status of WAF nodes.

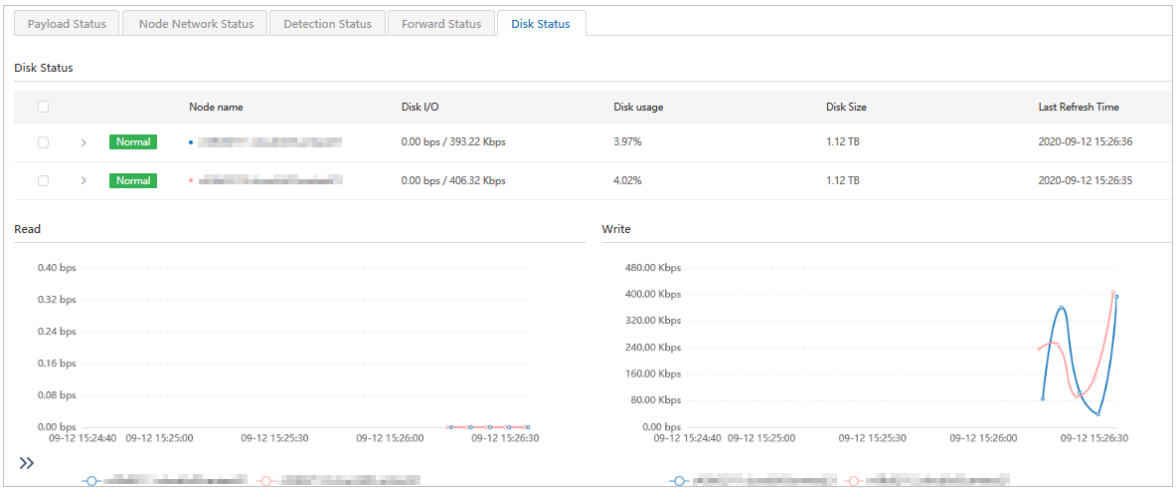


22.1.8.5.3. View the disk status of nodes

This topic describes how to view the disk status of Web Application Firewall (WAF) nodes. You can identify faults based on the status and check whether scale-out or scale-up is required.

Procedure

1.
2.
3. choose **System Info > Node status**.
4. Click the **Disk Status** tab to view the disk status of WAF nodes.



In the Disk Status section, you can view the disk I/O and disk usage of WAF nodes. In the **Read** and **Write** sections, you can view the changes in disk reads and writes over a specific period of time.

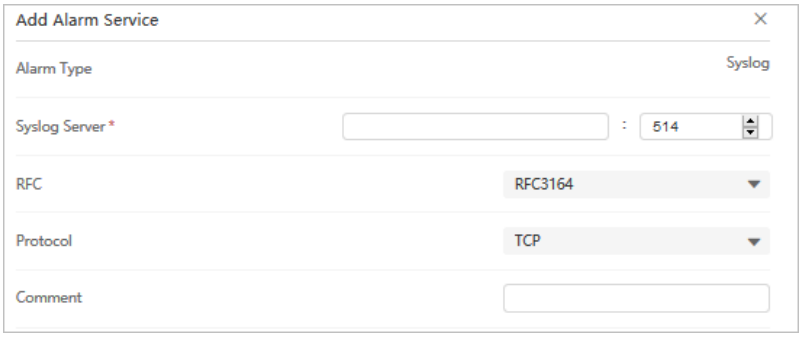
22.1.8.5.4. Configure alerts

This topic describes how to add a syslog server to Web Application Firewall (WAF). After the syslog server is added, WAF alert logs can be pushed to the syslog server over the syslog protocol.

Procedure

1.
2.

3. choose **System Settings > Syslog Configuration** .
4. On the **Alarm Service Configuration** tab, click **Add alarm service**.
5. In the **Add Alarm Service** panel, configure parameters.



Parameter	Description
Syslog Server	The IP address and port number of the syslog server.
RFC	The Request for Comments (RFC) document that defines the syslog protocol. Valid values: **RFC3164** and **RFC5424**.
Protocol	The transmission protocol. Valid values: **TCP** and **UDP**.
Comment	The description of the syslog server. This information facilitates subsequent identification and management.
General	The type of alert. Valid values: System Management and System Monitor and Alarm.
Security	The module whose alert logs are sent to the syslog server.

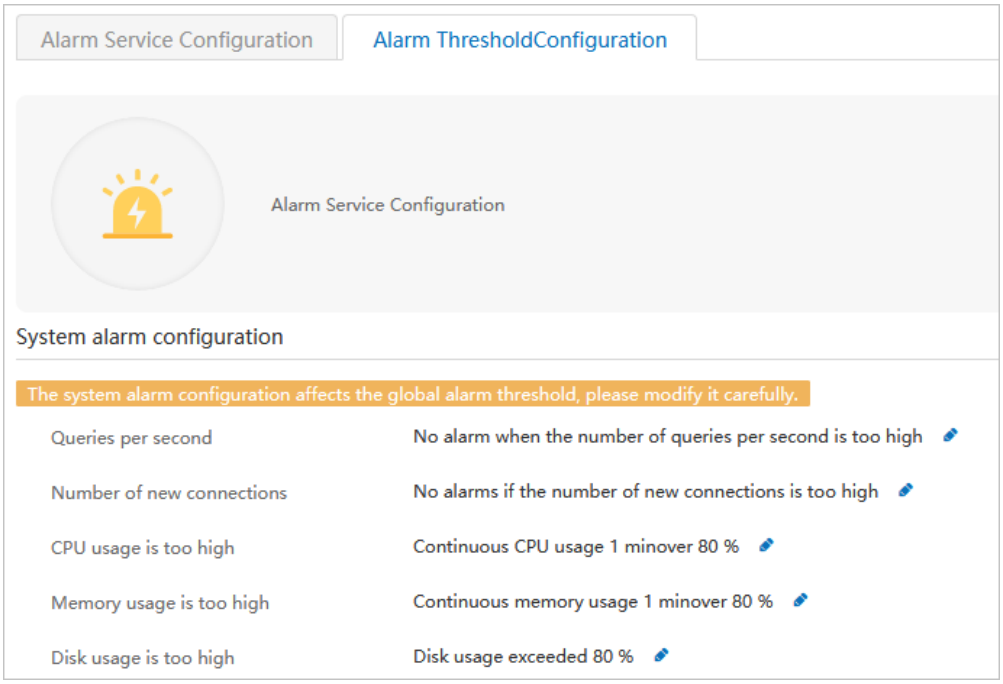
6. Click **Confirm**. The newly added syslog server appears in the list of the Alarm Service Configuration tab.
7. Find the newly added syslog server and click the icon in the **Operation** column to test whether alerts are sent.
   - If a message appears, indicating that the alert test is successful, the syslog server is added.
   - If an error message appears, WAF cannot connect to the syslog server.

22.1.8.5.5. Configure alert thresholds

This topic describes how to configure alert thresholds.

Procedure

1.
2.
3. choose **System Settings > Syslog Configuration** .
4. Click the **Alarm ThresholdConfiguration** tab and click the icon next to the threshold that you want to modify.
5. In the panel that appears, specify the threshold.



Threshold	Description
Queries per second	If queries per second exceed this threshold, alerts are sent. If this threshold is set to 0, no alerts are sent.
Number of new connections	If a large number of new connections exist, no alerts are sent.
CPU usage is too high	If CPU utilization exceeds this threshold in a specific period of time, alerts are sent.
Memory usage is too high	If memory usage exceeds this threshold in specific a period of time, alerts are sent.
Disk usage is too high	If disk usage exceeds this threshold, alerts are sent.

6. Click **OK**.

22.1.9. Security Operations Center (SOC)

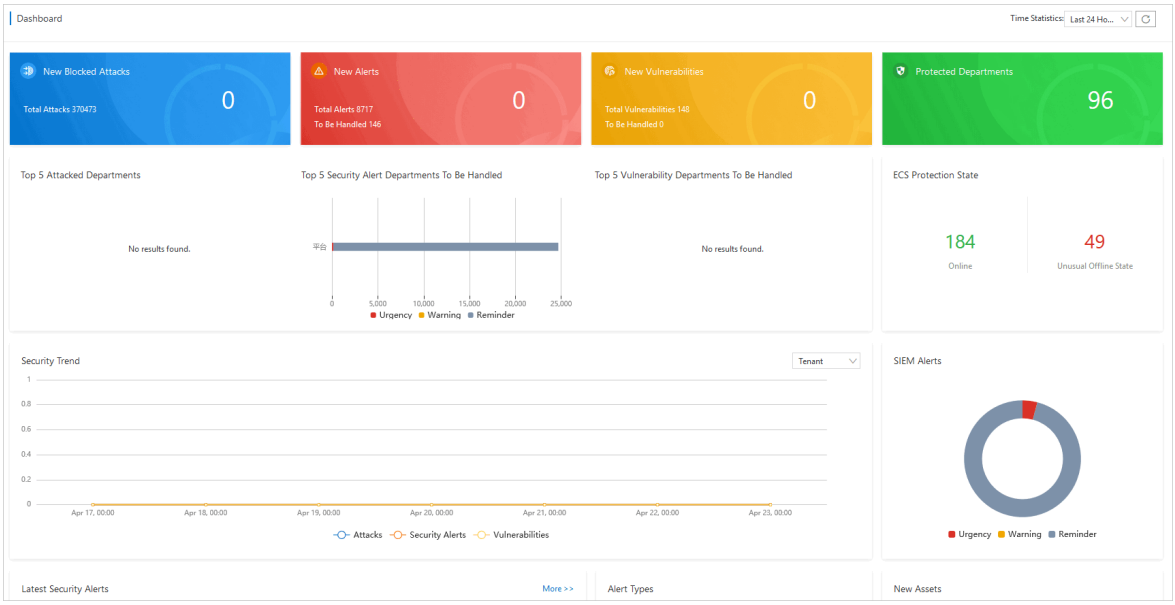
22.1.9.1. View the dashboard

This topic describes how to view the overall security information about the Apsara Stack network environment.

Procedure

1.
2.
3. click **Overview**.
4. In the upper-right corner of the **Dashboard** page, select a time range from the **Time Statistics** drop-down list.

   Valid values: *Last 24 Hours*, *Last 7 Days*, and *Last 30 Days*.
5. View the overall security information.



The Dashboard page displays the following information:

- **New Blocked Attacks**, **New Alerts**, **New Vulnerabilities**, and **Protected Departments**
- **Top 5 Attacked Departments**, **Top 5 Security Alert Departments To Be Handled**, and **Top 5 Vulnerability Departments To Be Handled**
- **Security Trend**, which supports a switchover between Tenant and Platform
- **Latest Security Alerts** and **Alert Types**
- **Latest Attacks** and **Attack Types**.
- **ECS Protection State**, **New Assets**, and **Protected Assets**

22.1.9.2. Security Monitoring

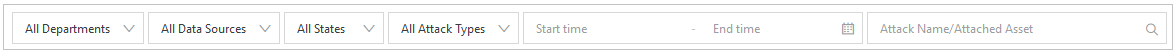
22.1.9.2.1. View security monitoring data of tenants

This topic describes how to view the security monitoring data of tenants on the Attack Protections, Security Alerts, and Vulnerabilities tabs.

Attack Protections

1.
2.
3. click **Tenant Security Monitoring**.
4. On the **Attack Protections** tab, view all attack events.

   You can specify search conditions to search for attack events based on the following table.



Search condition	Description
Department	The department to which the assets affected by the attack belong.
Data source	The data source.
Status	The attack status.

Search condition	Description
Attack type	The attack type.
Start time and end time	The time range to query the attack event.
Attack name or asset keyword	The attack name or the keywords of affected assets.

5. View details in the attack event list.
6. Click the icons in the upper-left corner to refresh or export the list.

   The following list describes how to perform the operations:

   - Click the icon to refresh the attack event list.
   - Click the icon to export the attack event list.
7. Find an attack event. In the Actions column, block requests from a specific IP address, create a tag for the event, or view logs and details of the event.

   The following list describes how to perform the operations:

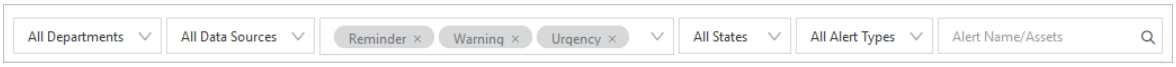
   - Block requests from a specific IP address: Click **Block IP Addresses**. In the **Block IP Addresses** dialog box, configure parameters to block requests from a specific IP address. For more information, see Block IP Addresses.

     In the upper-right corner, click **View Blocked IPs** to view details about the blocked IP addresses.
   - Create a tag: Click **Tag**. In the **Customize Tag** dialog box, create a tag for the attack event and click **OK**.
   - View logs: Click **View Log**. On the **Cloud Tenant Logs** tab of the **Log Audit** page, view the logs of tenants.
   - View details: Click **Details** to view the details of the attack event.

Security Alerts

1.
2.
3. click **Tenant Security Monitoring**.
4. Click the **Security Alerts** tab.
5. (Optional)Specify search conditions.

   Note If you want to view all security alerts, skip this step.



Search condition	Description
Department	The department to which the assets affected by the alert belong.
Data source	The data source.

Search condition	Description
Level	The alert level. You can select one or more alert levels. Valid values: ○ Urgency ○ Warning ○ Reminder
Status	The alert status.
Type	The alert type. You can select **All Alert Types** or a specific alert type.
Start time and end time	The time range to query the alert.
Alert name or asset keyword	The alert name or the keywords of affected assets.

6. View details in the security alert list.
7. Click the icons in the upper-left corner to refresh or export the list.

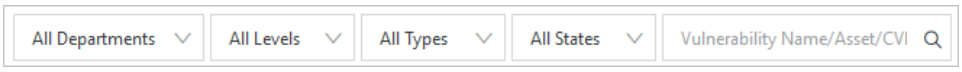
The following list describes how to perform the operations:

- Click the icon to refresh the security alert list.
- Click the icon to export the security alert list.

Vulnerabilities

1.
2.
3. click **Tenant Security Monitoring**.
4. Click the **Vulnerabilities** tab.
5. Click the **Vulnerabilities** or **Server Configurations** tab.
   - **Vulnerabilities**: provides vulnerability information.
   - **Server Configurations**: provides information about server baseline risks.
6. Specify search conditions.

> **Note** If you want to view all vulnerabilities or server baseline risks, skip this step.



Search condition	Description
Department	The department to which the assets affected by the vulnerability or server baseline risk belong.
Level	The level of the vulnerability or server baseline risk.
Type	The type of the vulnerability or server baseline risk.
Status	The status of the vulnerability or server baseline risk.

Search condition	Description
Start time and end time	The time range to query the vulnerability or server baseline risk.
Vulnerability or risk name, asset keyword, or CVE ID keyword	The name of the vulnerability or server baseline risk, or the keywords of affected assets or CVE IDs.

7. Click the icons in the upper-left corner to refresh or export the list of vulnerabilities or server baseline risks.

The following list describes how to perform the operations:

- Click the icon to refresh the list of vulnerabilities or server baseline risks.
- Click the icon to export the list of vulnerabilities or server baseline risks.

22.1.9.2.2. View security monitoring data of the Apsara Stack platform

This topic describes how to view the security monitoring data of the Apsara Stack platform on the Attack Protections, Security Alerts, and Vulnerabilities tabs.

Attack Protections

1.
2. Choose **Global Platform Security > Security Operations Center**.
3. In the left-side navigation pane, click **Platform Security Monitoring**.
4. On the **Attack Protections** tab, view all attack events.

   You can specify search conditions to search for attack events based on the following table.

Search condition	Description
Data source	The data source.
Status	The attack status.
Attack type	The attack type.
Start time and end time	The time range to query the attack event.
Attack name or asset keyword	The attack name or the keywords of affected assets.

5. View details in the attack event list.
6. Click the icons in the upper-left corner to refresh or export the list.

   The following list describes how to perform the operations:

   - Click the icon to refresh the attack event list.
   - Click the icon to export the attack event list.

7. Find an attack event. In the Actions column, block requests from a specific IP address, create a tag for the event, or view logs and details of the event.

   The following list describes how to perform the operations:

   - Block requests from a specific IP address: Click **Block IP Addresses**. In the **Block IP Addresses** dialog box, configure parameters to block requests from a specific IP address. For more information, see Block IP Addresses.

     In the upper-right corner, click **View Blocked IPs** to view details about the blocked IP addresses.
   - Create a tag: Click **Tag**. In the **Customize Tag** dialog box, create a tag for the attack event and click **OK**.
   - View logs: Click **View Logs**. On the **Cloud Platform Logs** tab of the **Log Audit** page, view the logs of the platform.
   - View details: Click **Details** to view the details of the attack event.

Security Alerts

1.
2. Choose **Global Platform Security > Security Operations Center**.
3. In the left-side navigation pane, click **Platform Security Monitoring**.
4. On the **Platform Security Monitoring** page, click the **Security Alerts** tab.
5. (Optional)Specify search conditions.

   > **Note** If you want to view all security alerts, skip this step.

Search condition	Description
Data source	The data source.
Level	The alert level. You can select one or more alert levels. Valid values: ○ Urgency ○ Warning ○ Reminder
Status	The alert status.
Type	The alert type. You can select **All Alert Types** or a specific alert type.
Start time and end time	The time range to query the alert.
Alert name or asset keyword	The alert name or the keywords of affected assets.

6. View details in the security alert list.
7. Click the icons in the upper-left corner to refresh or export the list.

   The following list describes how to perform the operations:

   - Click the icon to refresh the security alert list.
   - Click the icon to export the security alert list.

Vulnerabilities

1.
2. Choose **Global Platform Security > Security Operations Center**.
3. In the left-side navigation pane, click **Platform Security Monitoring**.
4. On the **Platform Security Monitoring** page, click the **Vulnerabilities** tab.
5. On the **Platform Baseline** tab, specify search conditions.

   > **Note** If you want to view all baseline risks of the platform, skip this step.

Search condition	Description
Level	The level of the baseline risk.
Status	The status of the baseline risk.
Start time and end time	The time range to query the baseline risk.
Risk name or asset keyword	The risk name or the keywords of affected assets.

6. Click the icons in the upper-left corner to refresh or export the list of platform baseline risks.

   The following list describes how to perform the operations:

   - Click the  icon to refresh the list of platform baseline risks.
   - Click the  icon to export the list of platform baseline risks.

22.1.9.2.3. View the global traffic

This topic describes how to view global traffic, including the average traffic, peak traffic, overall traffic trends, traffic of tenants, and traffic of platforms.

Procedure

1.
2. Choose **Global Platform Security > Security Operations Center** .
3. In the left-side navigation pane, click **Global Traffic Analysis**.
4. On the **Global Traffic Analysis** page, view the global traffic.

   You can view the following information on this page:

   - View the average traffic and peak traffic
     a. In the upper-right corner of the **Global Traffic Analysis** page, select a time range and traffic direction.

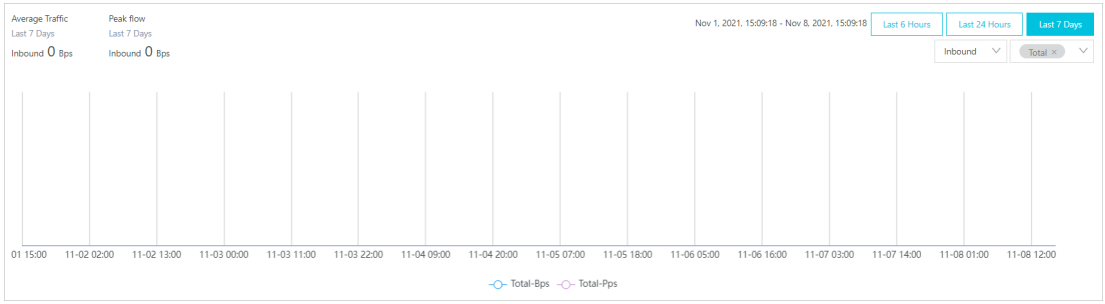
        Valid values of time ranges: **Last 6 Hours**, **Last 24 Hours**, and **Last 7 Days**.

        Valid values of traffic directions: **Inbound** and **Outbound**.

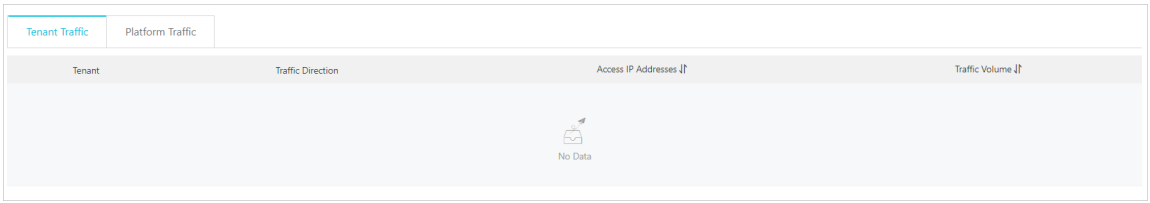
        

b. In the upper-left corner of the Global Traffic Analysis page, view the average and peak traffic of the specified traffic direction within the specified time range.

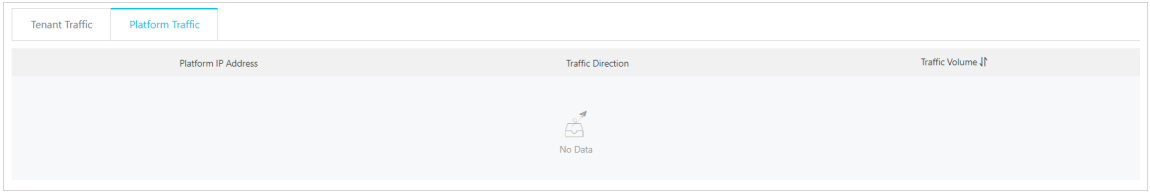
- View traffic trends
  a. In the upper-right corner of the **Global Traffic Analysis** page, select a traffic type.
  b. View the overall traffic trends of each traffic type within the specified time range.



- View the traffic of tenants on the **Tenant Traffic** tab



- View the traffic of platforms on the **Platform Traffic** tab



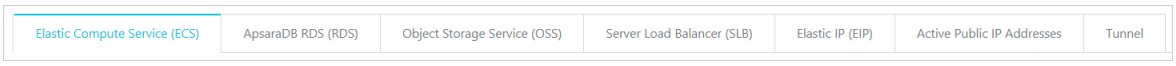
22.1.9.3. Asset Management

22.1.9.3.1. View tenant assets

This topic describes how to view the assets of users. The assets include Elastic Compute Service (ECS) instances, ApsaraDB RDS instances, Object Storage Service (OSS) buckets, Server Load Balancer (SLB) instances, and elastic IP addresses (EIPs).

Procedure

1.
2. Choose **Global Platform Security > Security Operations Center**.
3. In the left-side navigation pane, click **Tenant Assets**.
4. Select the required service. Example: **Elastic Compute Service (ECS)**.



5. Specify search conditions to view an asset.

> **Note** If you want to view all assets, skip this step.

Search condition	Description
Department	The department to which the asset belongs.
VPC	The virtual private cloud (VPC) to which the asset belongs.
Status	The running status of the asset.
New	Specifies whether the asset to query is newly added.
Server name or IP address	The keywords of the asset name.

6. View asset information in the asset list.

22.1.9.3.2. View platform assets

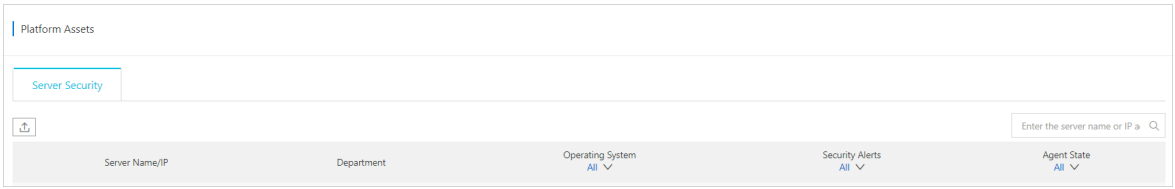
This topic describes how to view the assets of the platform.

Procedure

1.
2.
3. click **Platform Assets**.
4. On the **Platform Assets** page, search for assets.

   > **Note** If you want to view all assets, skip this step.

5. View asset information in the asset list.



22.1.9.4. Log Analysis

22.1.9.4.1. View the Log Overview page

This topic describes how to view the logs that are displayed in the widgets on the Log Overview page.

Procedure

1.
2.
3. click **Log Overview**.
4. On the **Log Overview** page, view the widgets of the logs.

   You can perform the following operations to **modify** or **delete** the widgets on the **Log Overview** page:

   - To modify a widget, click **Modify** in the upper-right corner of the widget.

a. In the **Modify** dialog box, reconfigure the **Chart Type**, **Category**, and **Value** parameters. When you configure the **Category** and **Value** parameters, take note of the following points:
   - **Category**: If you set **Chart Type** to **Bar Chart**, **Line Chart**, **Pie Chart**, or **Sheet**, you must specify this parameter.
   - **Value**: If you set **Chart Type** to **Pie Chart** or **Individual Value Plot**, you must specify this parameter.

b. Click **Refresh** to preview the widget in the right side of the Modify dialog box.

c. Above the widget, enter a new name to rename the widget.

d. Click **OK**. The widget is updated on the **Log Overview** page.

- To delete a widget, click **Delete** in the upper-right corner of the widget.

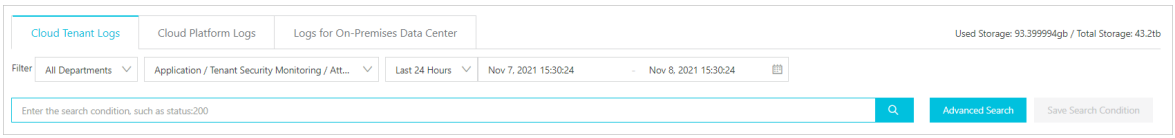
The widgets on the **Log Overview** page are created on the **Log Audit** page. To create a widget, click **Please go to the log audit page to add a chart** in the **Add custom visualization chart** section in the lower part of the **Log Overview** page. On the **Log Audit** page, create a custom widget. For more information, see View global logs.

22.1.9.4.2. View global logs

This topic describes how to view global logs. Global logs are classified into logs of tenants, logs of the platform, and logs of data centers based the department to which the logs belong.

View a log widget

1.
2.
3. click **Log Audit**.
4. Click the tab on which you want to view the logs of a department. For example, if you want to view the logs of tenants, click the **Cloud Tenant Logs** tab.



5. Specify search conditions and click the icon. Then, you can view the logs that meet the search conditions in the sections of log distribution chart and log list.

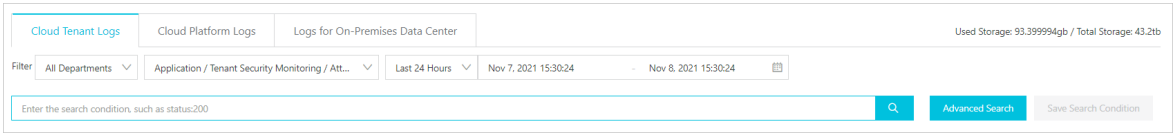
Search condition	Description
Department	If you want to view the logs of a tenant, you can specify this search condition. Select the department to which the tenant belongs.
Log Source	Select the type of the system from which you want to collect the logs, the name of the system from which you want to collect the logs, and the log type from the drop-down list.
Duration	Select the time range within which you want to collect the logs. Valid values: **Last 15 Minutes**, **Last 30 Minutes**, **Last 24 Hours**, **Last 7 Days**, **Last 30 Days**, and **Custom**.
Start time and end time	Specify the start and end time within which you want to collect the logs.

Search condition	Description
Log content	Enter the log content in the search box. If you want to save the search conditions as frequently used search conditions, click **Save Search Condition**. If you want to use these frequently used search conditions, click **Historical Records**. Then, you can view the logs that meet the search conditions.

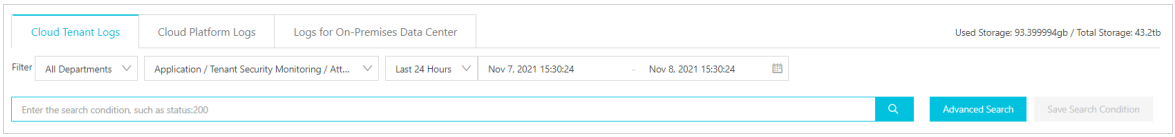
If you want to search for logs by using JSON domain-specific language (DSL) statements, click **Advanced Search**. In the **Advanced Search** dialog box, enter JSON DSL statements and click **Submit**.

Create a log widget

1.
2.
3. , click **Log Audit**.
4. Click the tab on which you want to view the logs of a department. For example, if you want to view the logs of tenants, click the **Cloud Tenant Logs** tab.



5. In the lower part of the Log Audit page, click the **Visualization** tab.
6. On the **Visualization** tab, configure parameters in the **Chart** section.



Parameter	Description
Chart Type	The type of widget that you want to display on the **Overview** page. Valid values: **Bar Chart**, **Line Chart**, **Pie Chart**, **Individual Value Plot**, and **Sheet**.
Category	This parameter is required only if you select **Bar Chart**, **Line Chart**, **Pie Chart**, or **Sheet** for **Chart Type**. The type of item that you want to display in the horizontal axis or the column header of the widget.
Value Category	The type of item that you want to display in the vertical axis or the row header of the widget.
Value Type	The type of value that you want to display in the widget. Valid values: **count**, **max**, **min**, **avg**, **sum**, **unique_count**, and **median**. If you want to display multiple value types in the widget, click **Add Value Field**. Then, you can configure the **Value Category** and **Value Type** parameters.

7. In the upper-left corner of the preview section on the Visualization tab, enter a name for the widget.
8. Click **Update**.

   After the widget is created, you can view the widget on the **Log Overview** page.

To configure the content to be displayed in the log list, you can perform the following steps: In the lower part of the Log Audit page, click the **Logs** tab. Then, click the icon. To export the log list, click the icon.

22.1.9.4.3. Log configurations

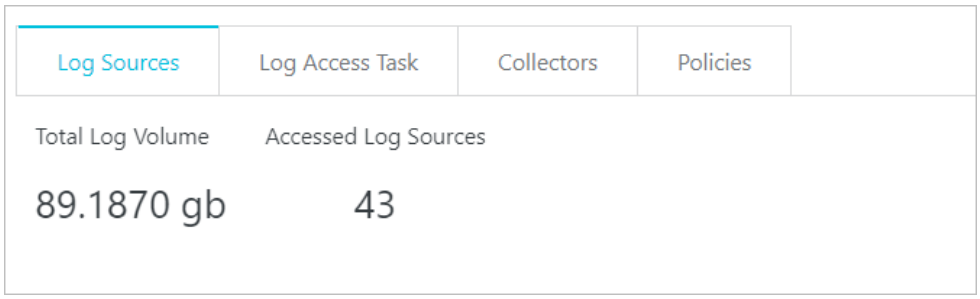
22.1.9.4.3.1. Manage log sources

This topic describes how to view and manage the log sources that are connected to Security Operations Center (SOC).

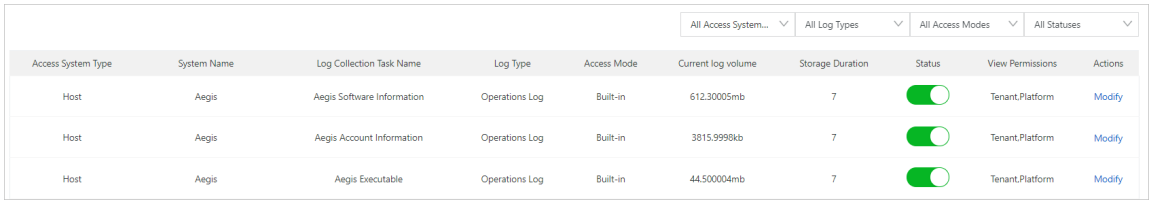
1.
2.
3. click **Log Configurations**.
4. On the **Log Configurations** page, click the **Log Sources** tab.
5. On the **Log Sources** tab, view the log overview and log list.

   This tab provides the following information:

   - In the upper section of the **Log Sources** tab, you can view the total volume of log data and the total number of connected log sources.



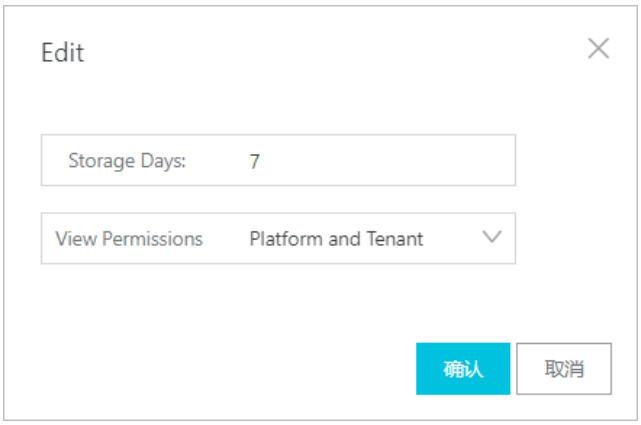
   - In the upper-right corner above the log source list, you can specify conditions to search for log sources.



Search condition	Description
Access system type	The type of the source system in the log source that is connected to SOC. Valid values: **Host**, **Storage**, **Application**, **Networking**, **Data**, **Security**, and **Other**.
Log type	The type of log that is collected by SOC. Valid values: **Operations Log**, **Operational Log**, **Alert Log**, and **Others**.
Access mode	The mode that is used to collect logs. Valid values: **Custom** and **Built-in**.
Status	The status of the log source that is connected to SOC. Valid values: **On** and **Off**. In the log source list, find a log source and click the icon in the **Status** column to set the log source to on or off. In the Tips message, click **OK**.

6. Find a log source and click **Modify** in the Actions column.

7. In the **Edit** dialog box, configure the Storage Days and View Permissions parameters. Then, click **OK**.



22.1.9.4.3.2. Create a log collection task

This topic describes how to create a log collection task.

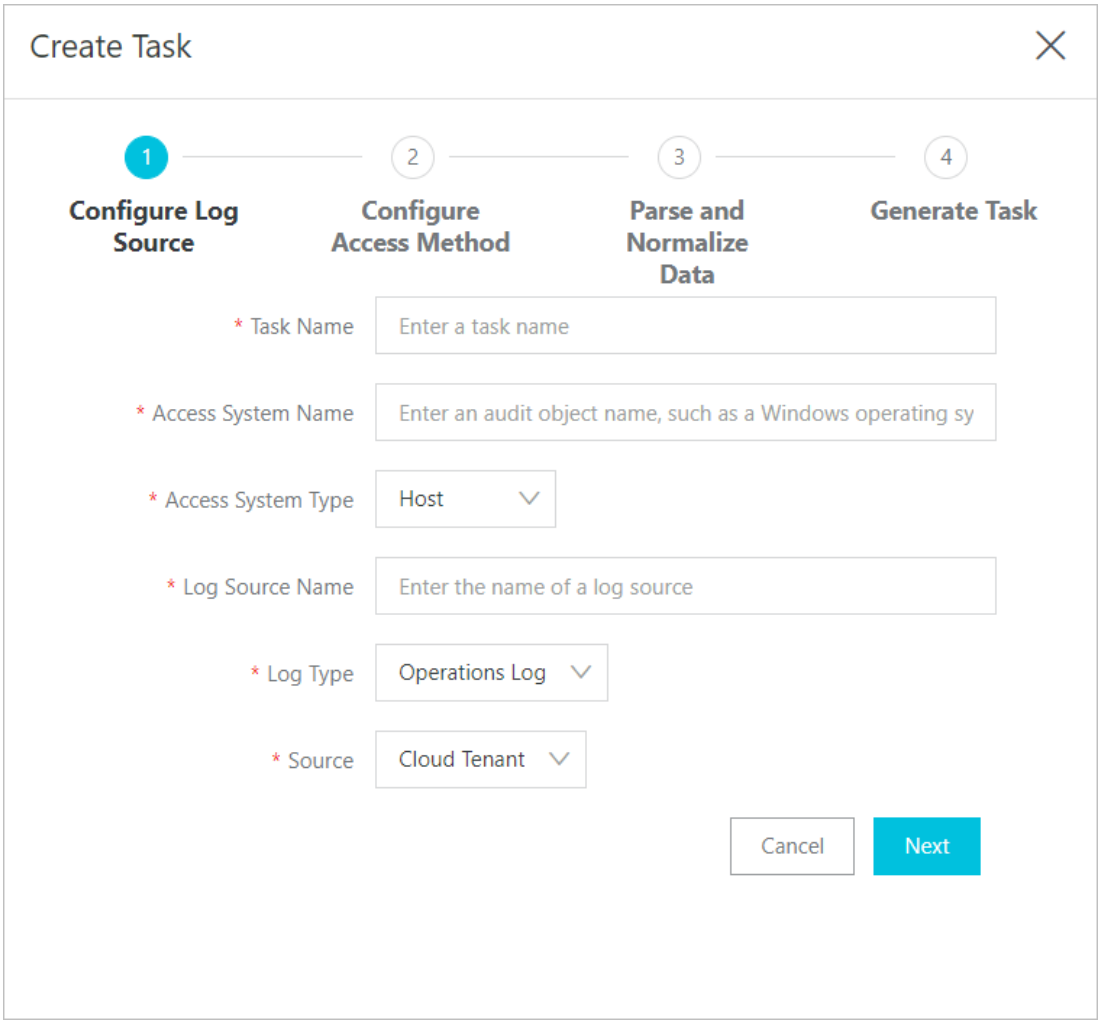
Prerequisites

If you use Logtail to collect logs, make sure that the following conditions are met:

- The Logtail agent is installed.
- A server group for which you want to configure Logtail is created.

For more information, see Manage log collectors.

1.
2.
3. click **Log Configurations**.
4. On the **Log Configurations** page, click the **Log Access Task** tab.
5. On the **Log Access Task** tab, click **Add**.
6. In the **Configure Log Source** step of the **Create Task** wizard, configure parameters and click **Next**.

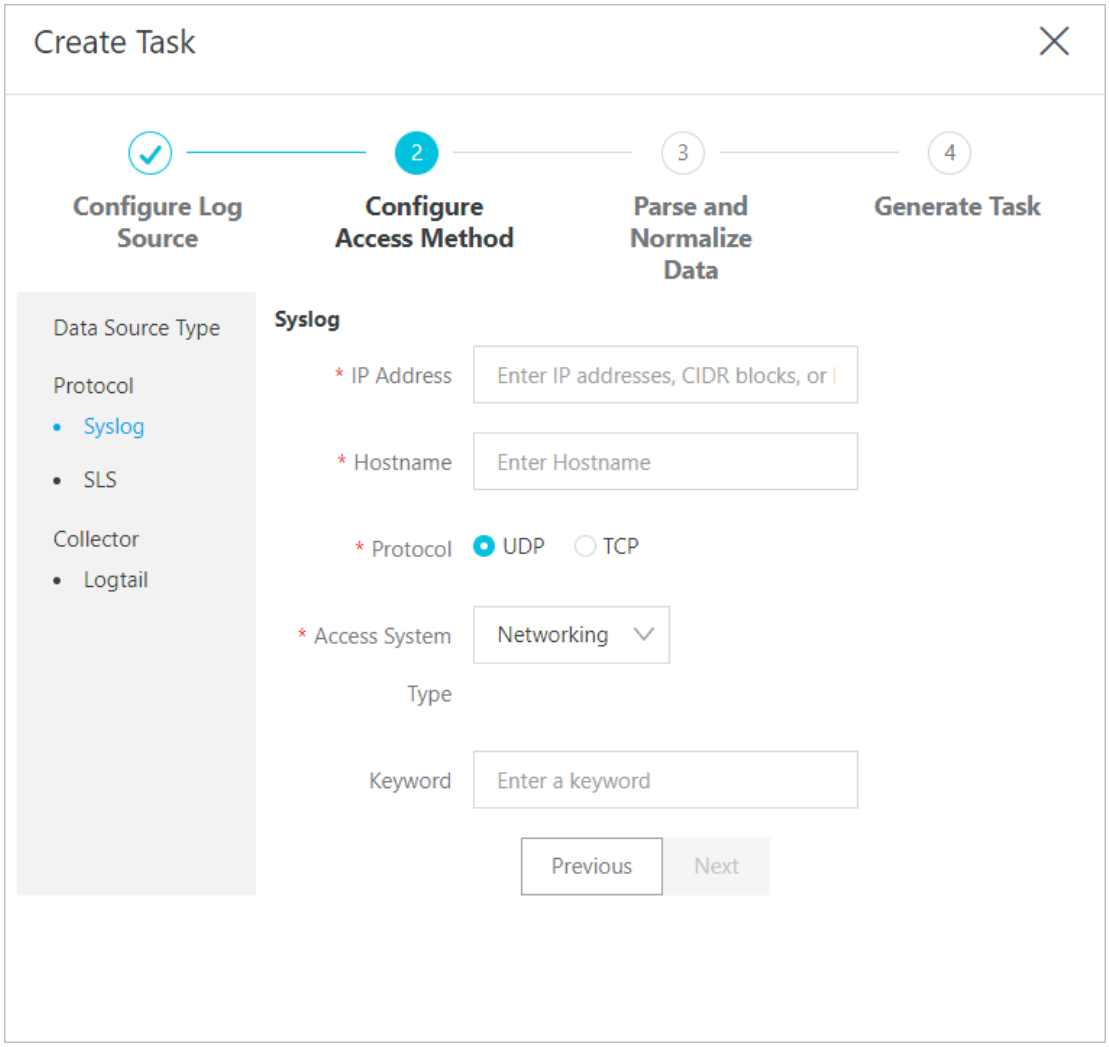


Parameter	Description
Task Name	The name of the log collection task.
Access System Name	The name of the source system from which you want to collect the logs. Example: Windows operating system.
Access System Type	The type of the source system from which you want to collect the logs. Valid values: **Host**, **Storage**, **Application**, **Networking**, **Data**, **Security**, and **Other**.
Log Source Name	The subtype of the log type.
Log Type	The type of log that you want to collect. Valid values: **Operations Log**, **Operational Log**, **Alert Log**, and **Others**.
Source	The department to which the logs belong. Valid values: **Cloud Tenant**, **Cloud Platform**, and **On-Premises Data Center**.

7. In the left-side navigation tree of the **Configure Access Method** step, select a data source type, configure parameters, and then click **Next**.

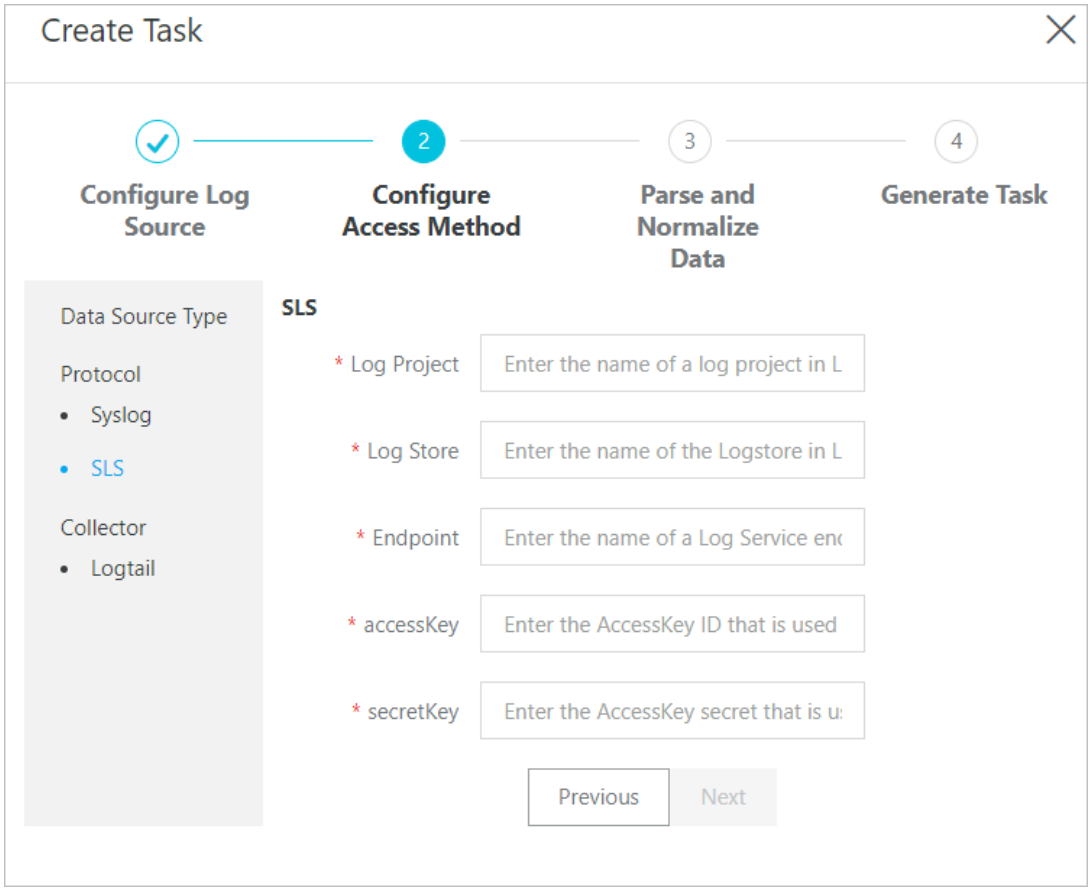
   Valid values of Data Source Type: Syslog, SLS, and Logtail.

   - If you select Syslog, you must configure the following parameters.



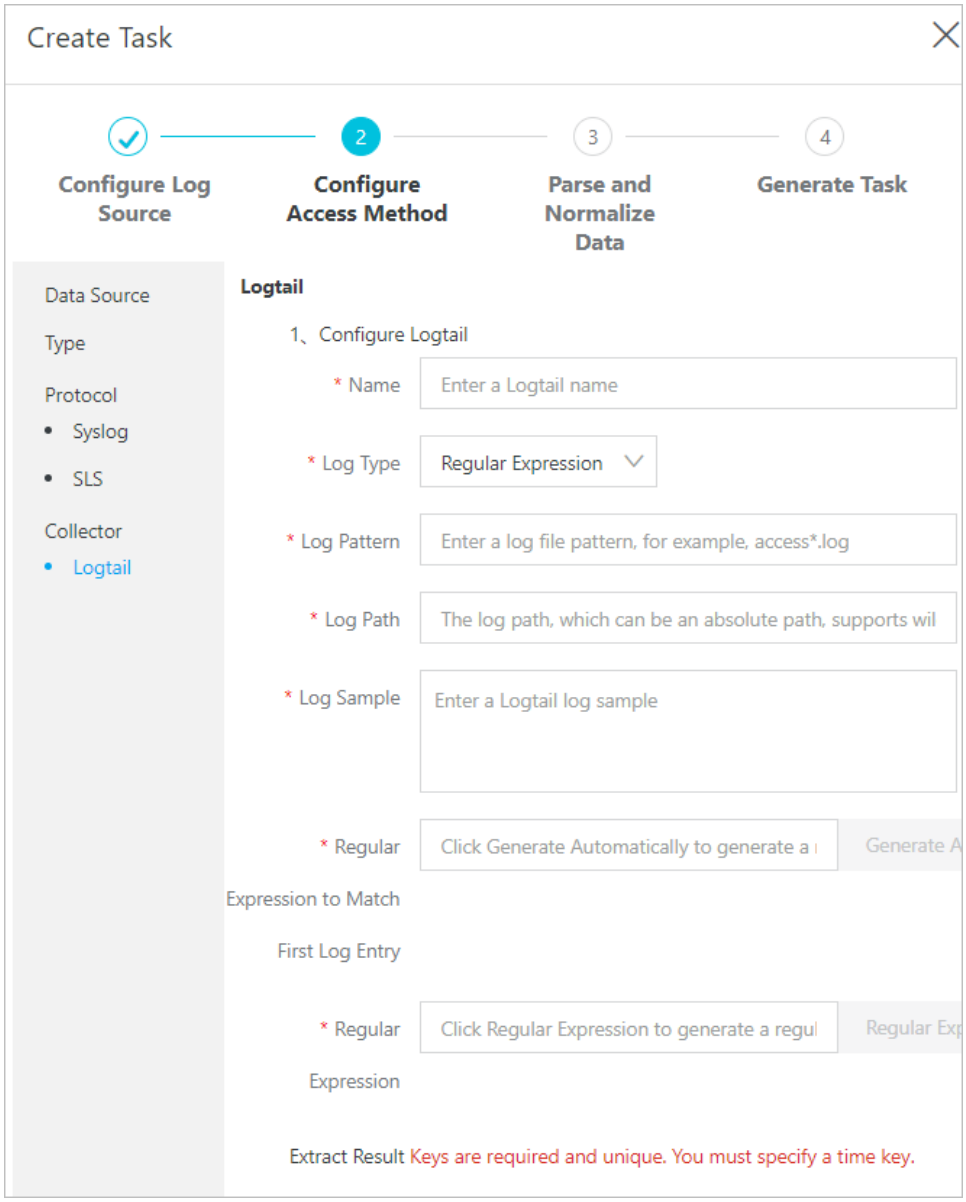
Parameter	Description
IP Address	The IP address or CIDR block used to report syslog logs.
Protocol	The network protocol used when the logs are collected. Valid values: **UDP** and **TCP**.
Access System Type	The type of source system from which you want to collect the logs. Valid values: **Host**, **Application**, **Networking**, and **Other**.
Keyword	The keyword of log data.

- If you select SLS, you must configure the following parameters.



Parameter	Description
Log Project	The name of the project in Log Service.
Log Store	The name of the Logstore in Log Service.
Endpoint	The endpoint used to connect to the project in a specific region.
accessKey	The AccessKey ID of your account used to access Log Service.
secretKey	The AccessKey secret of your account used to access Log Service.

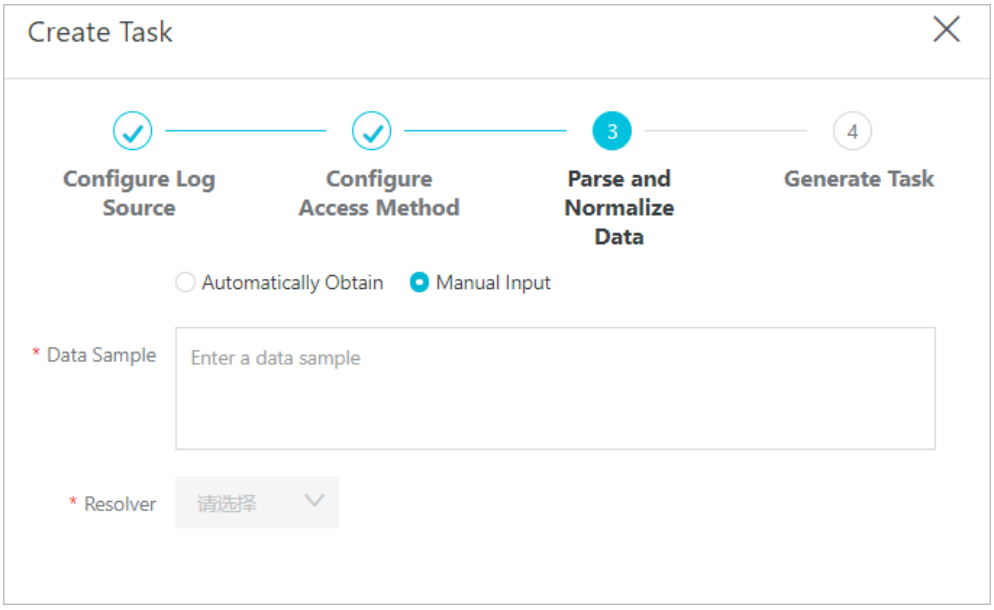
- If you select Logtail, you must configure the following parameters.



Parameter	Description
Name	The name of the log collector.
Log Type	The type of log that you want to collect. Valid values: **JSON**, **Apsara**, **Separator**, and **Regular Expression**.
Log Pattern	The format of log file names. Example: *access*.log*.
Log Path	The path used to store the logs that you want to collect. Absolute paths and relative paths are supported. You can use wildcards in relative paths.
Log Sample	A sample log entry from the logs that you want to collect.
Regular Expression to Match First Log Entry	After you enter the **sample log entry**, click **Generate Automatically**. A regular expression is generated to match the first line of the log entry.

Parameter	Description
Regular Expression	Click **Regular Expression**. In the **Generate Regular Expression** dialog box, select the fields that you want to extract and click **Generate Regular Expression**. After the regular expression is generated, click **OK**.
Date Format	The date format that is automatically generated based on the extracted time fields.
Apply to Server Group	The server group for which you want to configure Logtail.

8. In the **Parse and Normalize Data** step, configure parameters and click **Next**.



Parameter	Description
Data acquisition method	The method used to obtain the sample log entry. Valid values: **Automatically Obtain** and **Manual Input**.
Data Sample	If you select **Manual Input**, you must enter the sample log entry.
Resolver	The parser, which can be selected based on the sample log entry. Valid values: **JSON** and **jsonArray**.
Extract Values	The content that can be extracted from the logs. Click **Add Field**. In the **Add Field** dialog box, configure the **Original Extract Key**, **Original Extract Value**, **Enrich Data**, and **TargetType** parameters. Then, click **OK**. Valid values of **Enrich Data**: **Retain Original Field**, **Tag Field**, **Delete Field**, **Rename Field**, **Automatically Fill System Time**, **Assign Value to Constant**, and **Assign Value to Variable**.

9. In the **Generate Task** step, configure the **Log Source Name**, **Storage Duration**, and **View Permissions** parameters.
10. Click **Save**.

    After the log collection task is created, the task appears in the list of log collection tasks. You can publish, modify, or delete the task in the Actions column of the task.

    - **Publish**: After the log collection task is created, logs are not automatically collected. You must click

Publish. After the task is published, the on icon appears in the **Access Status** column of the task.

You can click the switch in the **Access Status** column to change the status of the log collection task.

- **Edit**: You can click **Edit** to change the configurations in the **Parse and Normalize Data** and **Generate Task** steps. You cannot change the data source type or log collector.
- **Delete**: If you no longer need the log collection task, you can click **Delete** in the Actions column of the task.

22.1.9.4.3.3. Manage log collectors

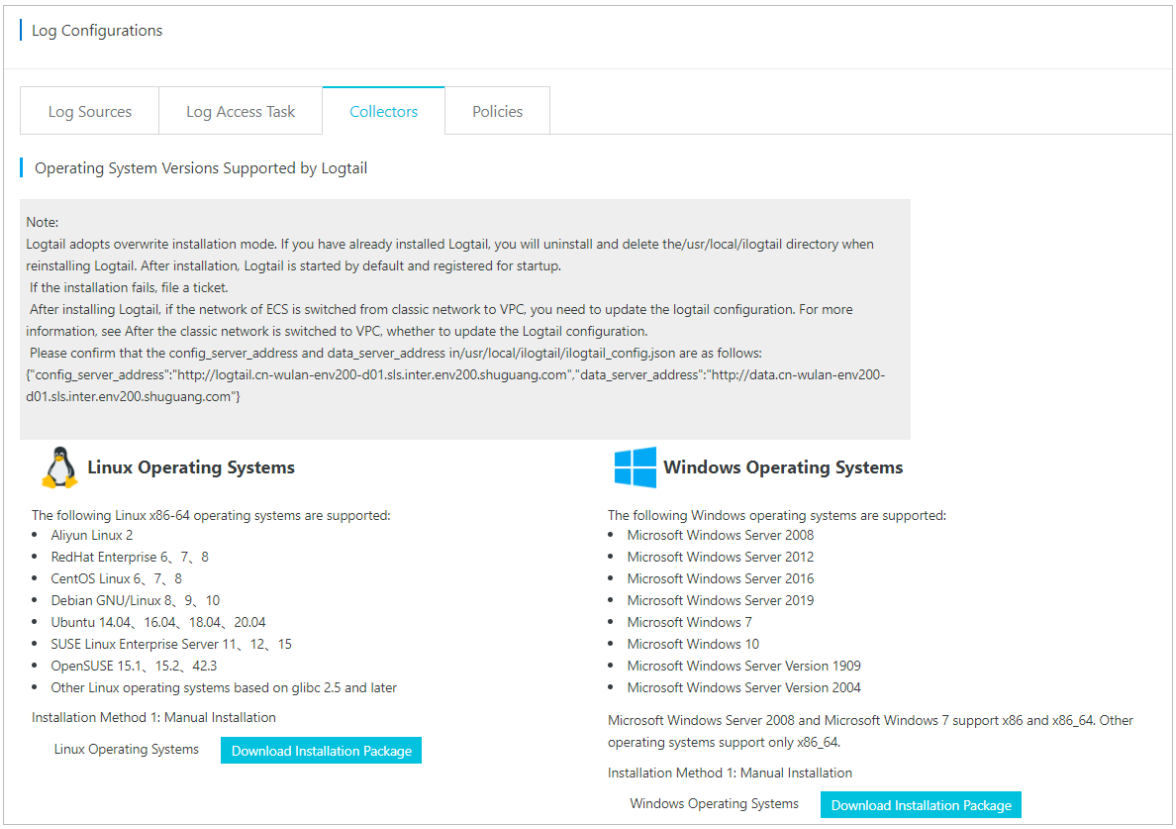
Before you can use Logtail to collect logs, you must install the Logtail agent, add log collectors, and create server groups. This topic describes how to install the Logtail agent, add log collectors, and create server groups.

Install the Logtail agent

If the Logtail agent has been installed, the system automatically uninstalls the existing version of the Logtail agent, deletes the */usr/local/ilogtail* directory, and then reinstalls the Logtail agent. After the new version of the Logtail agent is installed, it automatically runs, and a startup application is added to the registry.

1.
2.
3. click **Log Configurations**.
4. On the **Log Configurations** page, click the **Collectors** tab.
5. Install the Logtail agent.

   On the **Collectors** tab, install the Logtail agent based on the operating system of the server and the installation method that is supported by the Logtail agent.



- If the server runs a Linux operating system and the Logtail agent supports manual installation, perform the following steps:

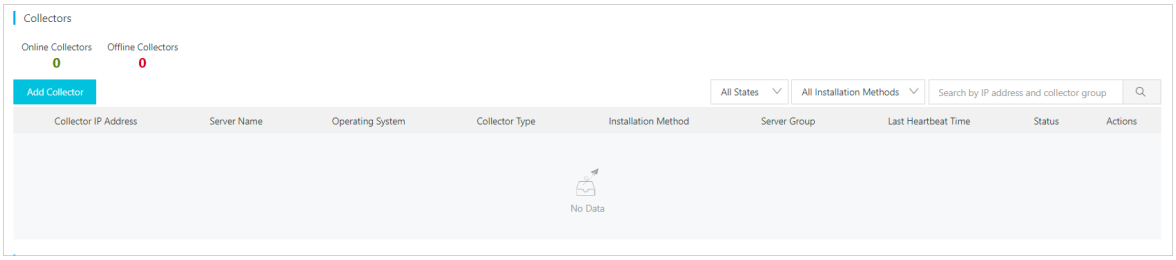
a. In the **Linux Operating Systems** section, click **Download Installation Package** and determine whether to download a 32-bit or 64-bit installation file to your computer.
b. Log on to the server as an administrator. Then, run the installation command to install the Logtail agent.

- If the server runs a Windows operating system and the Logtail agent supports manual installation, perform the following steps:
  a. In the **Windows Operating Systems** section, click **Download Installation Package** to download the installation file to your computer.
  b. Upload the installation file to the server. For example, you can use an FTP client to upload the installation file to the server.
  c. Run the installation file on the server as an administrator.

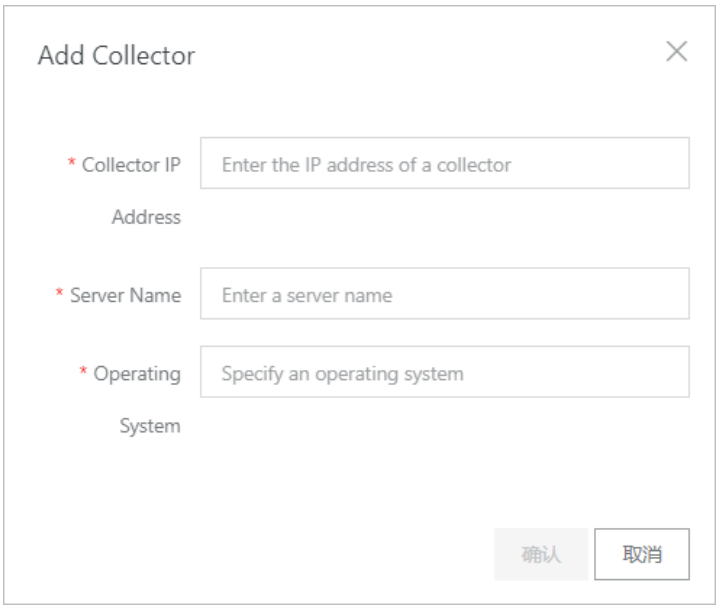
After the Logtail agent is installed, it automatically runs, and a startup application is added to the registry.

Add a log collector

1.
2.
3. click **Log Configurations**.
4. On the **Log Configurations** page, click the **Collectors** tab.
5. In the **Collectors** section, click **Add Collector**.



6. In the **Add Collector** dialog box, configure the following parameters and click **OK**.



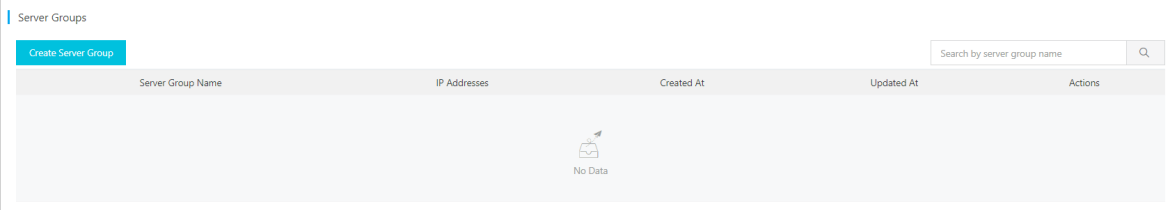
Parameter	Description
Collector IP Address	The IP address of the source whose logs you want to collect.
Server Name	The name of the server where the Logtail agent is installed.

Parameter	Description
Operating System	The operating system of the server where the Logtail agent is installed.

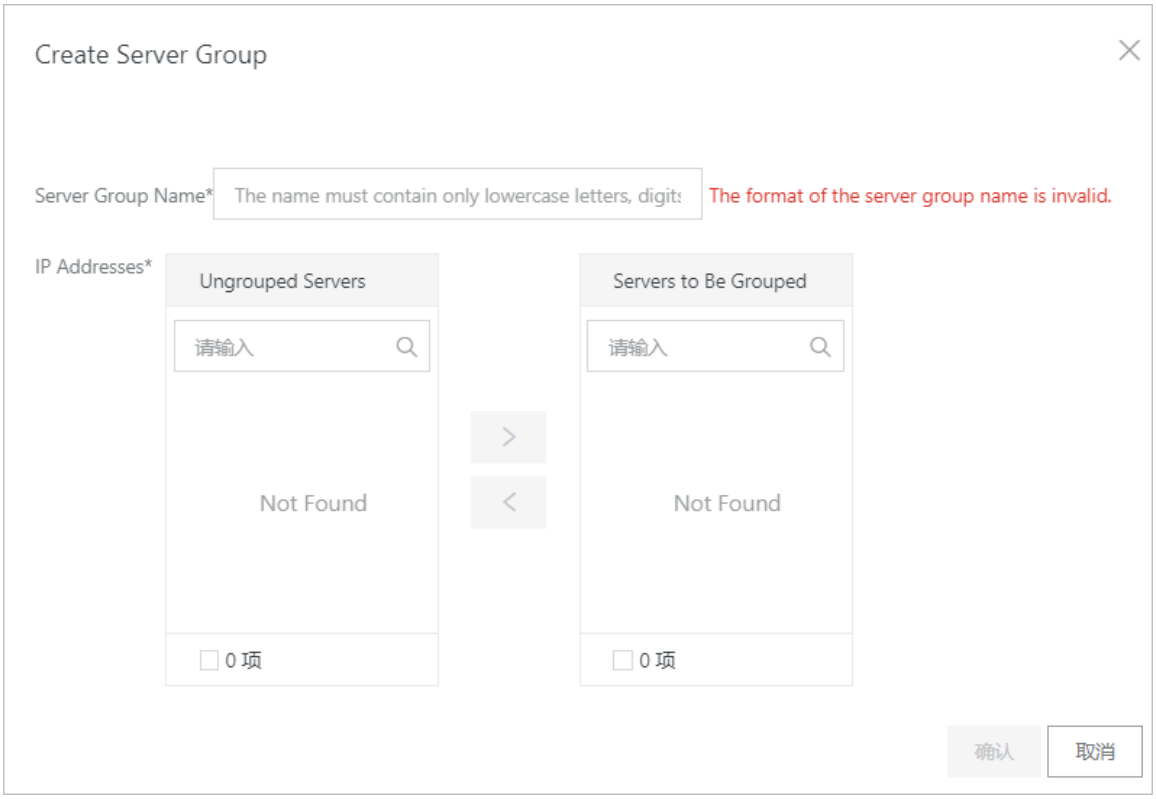
After the log collector is added, you can view the information of the log collector in the **Collectors** section. The information includes **Collector IP Address** and **Server Name**.

Create a server group

1.
2.
3. click **Log Configurations**.
4. On the **Log Configurations** page, click the **Collectors** tab.
5. In the **Server Groups** section, click **Create Server Group**.



6. In the **Create Server Group** dialog box, configure the following parameters and click **OK**.



Parameter	Description
Server Group Name	The name of the server group.

Parameter	Description
IP Addresses	The IP addresses of the servers that you want to add to the server group. To add a server to the server group, perform the following steps: i. In the **Ungrouped Servers** section, select the server that you want to add to the group. ii. Click the > icon to add the server to the **Servers to Be Grouped** section.

After the server group is created, you can view basic information of the server group in the **Server Groups** section. If you want to view detailed information of a server group or delete a server group, perform the following steps:

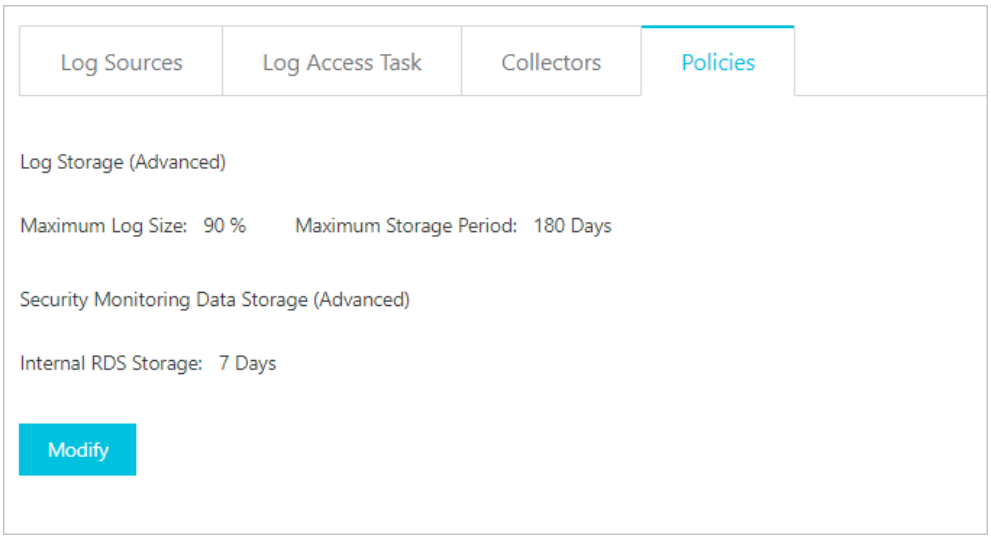
- Find the server group. Then, click **Details** in the Actions column to view the detailed information of the server group.
- Find the server group. Then, click **Delete** in the Actions column to delete the server group.

22.1.9.4.3.4. Manage storage policies

This topic describes how to view and configure storage policies for logs.

Procedure

1.
2.
3. click **Log Configurations**.
4. On the **Log Configurations** page, click the **Policies** tab.
5. On the **Policies** tab, view the storage policy for logs.



6. Click **Modify**.
7. On the **Policies** tab, configure parameters such as the storage periods for logs and monitoring data. Then, click **Modify**.

Section	Parameter	Description
	Maximum Log Size	The upper limit of the storage space that can be used to store logs.

Section	Parameter	Description
Log Storage (Advanced)	Maximum Storage Period	The maximum number of days during which logs can be stored.
Security Monitoring Data Storage (Advanced)	ApsaraDB RDS Storage Period	The maximum number of days during which monitoring data can be stored.

22.1.9.4.4. Security Audit

22.1.9.4.4.1. Overview

Apsara Stack Security provides the security audit feature to collect system security data, analyze weaknesses in system running, report audit events, and classify the events into three risk levels. The risk levels are high, medium, and low. Security administrators can use the feature to follow operations in the system, and view and analyze audit events. This way, security administrators can improve system performance.

Security auditing is a long-term security control activity, which is required throughout the lifecycles of cloud services.

22.1.9.4.4.2. View security audit overview

This topic describes how to view the summarized information about security audit.

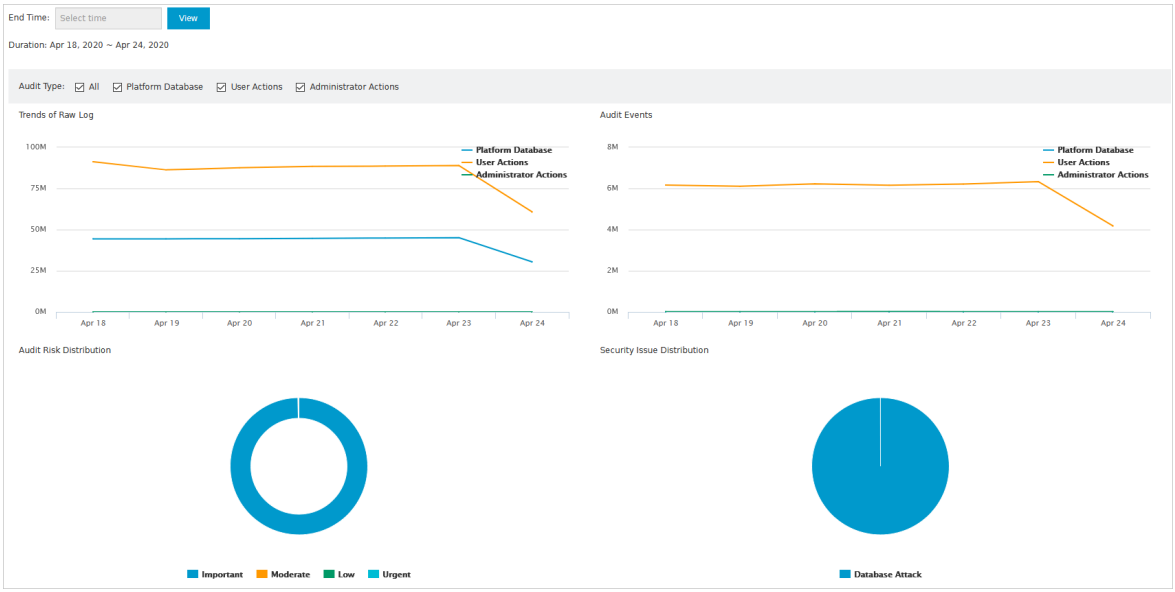
Context

The **Overview** tab provides reports on the raw log trend, audit event trend, audit risk distribution, and security event distribution. The reports are displayed in run charts or pie charts to help security administrators analyze the trend of risks in your cloud services.

On the **Overview** tab, security administrators can check the number of log entries and the storage usage in a specific time range.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Overview** tab.
5. On the **Overview** tab, view the audit summary for the last seven days.



- Trends of Raw Log

  This chart displays the trend of logs generated by physical servers, network devices, ApsaraDB RDS instances, Elastic Compute Service (ECS) instances, and API calls in the last seven days. Security administrators can analyze the trend to check whether the number of log entries is at a normal level.

- Audit Events

  This chart displays the trend of audit events that are generated by physical servers, network devices, ApsaraDB RDS instances, ECS instances, and API calls in the last seven days. Security administrators can analyze the trend to check whether the number of audit events is at a normal level.

- Audit Risk Distribution

  This chart displays the percentage distribution of audit events at different risk levels in the last seven days. Risk levels are important, moderate, and low. Security administrators can analyze the trend to check whether the audit events are at acceptable risk levels.

- Security Issue Distribution

  This chart displays the percentage distribution of different event types in the last seven days. Security administrators can analyze this chart to check for the most frequent audit events and identify high-risk events to improve security protection.

- Log Size

  This chart displays the volume of online logs and offline logs. If these logs consume many storage resources, we recommend that you back up required audit logs and delete unnecessary logs.

- Audit Log Size

  This chart displays the size of logs for each audit type.

6. View the audit summary in a specific time range.
   i. Specify **End Time** as the end of the time range to query.
   ii. In **Audit Type**, select the audit types to query.
   iii. Click **View** to view the audit summary in the last seven days before the specified end time.

22.1.9.4.4.3. Query audit events

This topic describes how to query audit events.

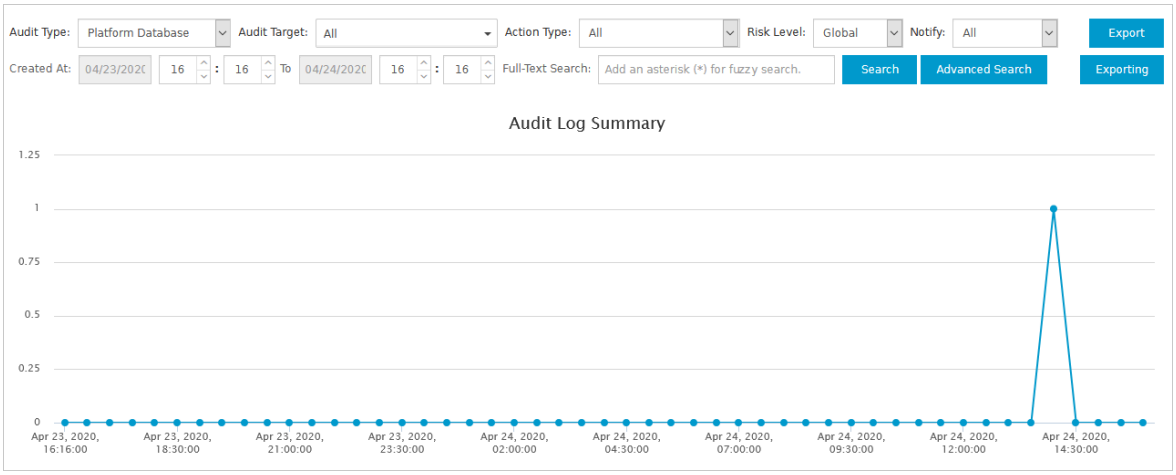
Context

On the **Audit Query** tab, you can view the details of audit events, including the log creation time, audit type, audit object, action type, risk level, and log content.

The system matches the logs that are collected by a security audit module with audit rules. If the log content matches the regular expression in an audit rule, an audit event is reported. For more information about audit rules, see Add an audit policy.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Audit Query** tab.
5. On the **Audit Query** tab, configure query conditions to view audit events within the specified time range.



- Basic query
  a. Configure **Audit Type**, **Audit Target**, **Action Type**, **Risk Level**, and **Notify**.
  b. Specify a time range to query.
  c. In the **Full-Text Search** search box, enter a keyword.
  d. Click **Search**.
- Advanced query

  You can also configure advanced query conditions.
  a. Configure basic query conditions.
  b. Click **Advanced Search**.
  c. In the **Filter Condition** section, configure **Filter Name**, **User**, **Target**, **Action**, **Result**, and **Cause**.
  d. Click **Save**.

6. Click **Export** to export the audit events.

   Download the exported file for analysis. For more information, see Manage export tasks.

22.1.9.4.4.4. View raw logs

This topic describes how to view raw audit logs.

Context

On the **Raw Log** tab, you can view the raw logs generated by a running audit object. Raw logs contain information that is required for debugging. Security administrators can use these raw logs to troubleshoot system failures.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Raw Log** tab.
5. On the **Raw Log** tab, configure query conditions to view the log summary chart and raw logs within a specific time range.
   i. Specify **Audit Type** and **Audit Target**.
   ii. Enter a keyword.
   iii. Specify a time range to query.
   iv. Click **Search**.

   If you want to compare log summary charts and raw logs in multiple time ranges, you must specify multiple time ranges and click **New Query** each time you specify a time range. Then, different query tabs are generated. You can click the query tabs to compare the log summary charts and raw logs in the time ranges.
6. Click **Export** to export the audit events.

   Download the exported file for analysis. For more information, see Manage export tasks.

22.1.9.4.4.5. Manage log sources

This topic describes how to view and manage log sources.

Context

You can view the number of log entries by log type or log source. You can also specify whether to display logs.

- The Log Types sub-tab provides the number of all log entries for a specific audit object of a specific device instance.
- The Log Sources sub-tab provides the number of log entries for all audit objects of a specific device instance.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Log Types** sub-tab to view the number of log entries by audit object.

   You can view the number of log entries that are recorded on the current day and the number of log entries that are recorded during the last 30 days for each audit object.

   If you do not want to display the log entries for an audit object, perform the following steps:

   i. Find the audit object and click **Hide** in the **Actions** column.
   ii. In the Note message, click **Confirm**.

   > **Note** The process that is used to display the log entries for an audit object is similar to the process that is used to hide the log entries.
5. Click the **Log Sources** sub-tab and view the number of log entries for each device instance.

   You can view the number of log entries that are recorded on the current day and the number of log entries that are recorded during the last 30 days for each device instance.

If you do not want to display the log entries for an audit object from a specific device instance, perform the following steps:

i. Find the device instance and click **Hide** in the **Actions** column.

ii. In the Note message, click **Confirm**.

> **Note** The process that is used to display the log entries for an audit object is similar to the process that is used to hide the log entries.

22.1.9.4.4.6. Policy settings

Manage audit rules

This topic describes how to create, modify, or delete an audit rule.

Context

If a log entry matches an audit rule, an audit event is reported. You can specify regular expressions in an audit rule to match log entries. A regular expression defines a matching pattern for character strings and can be used to check whether a string contains a specific substring. The following table provides examples about the pattern.

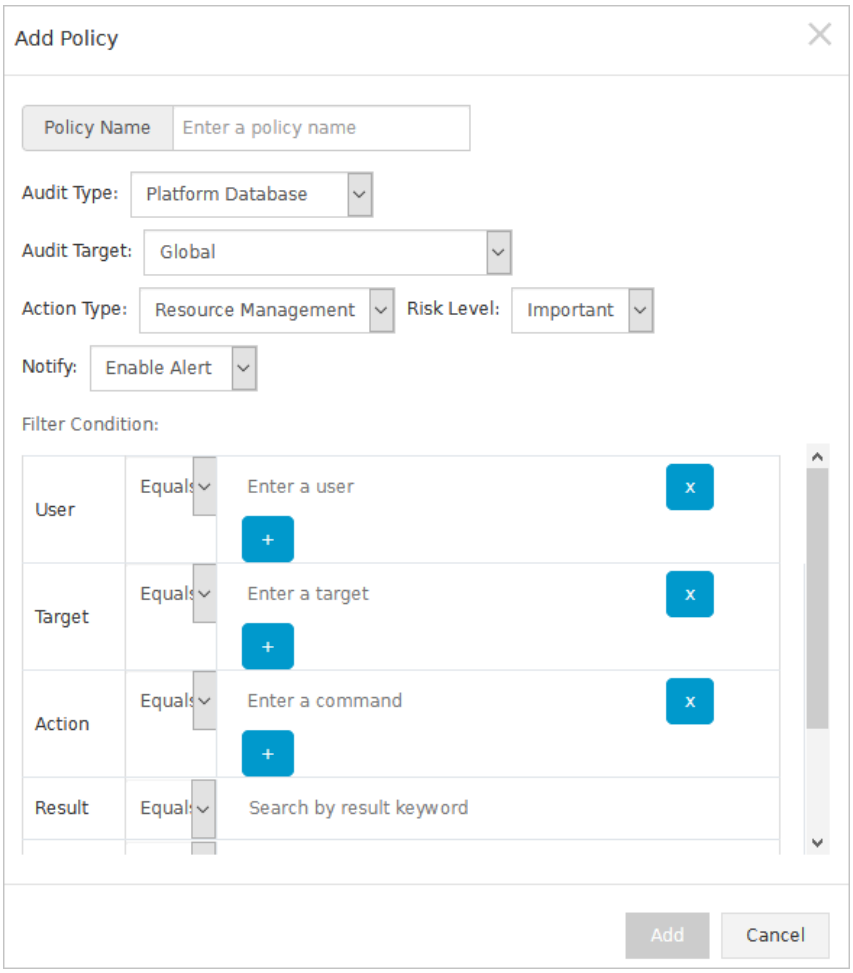
Regular expression	Description
`^\d{5,12}$`	Matches the consecutive numbers from the fifth number to the twelfth number.
`load_file\(`	Matches the "load_file(" string.

The security audit module defines the default audit rule based on the string that is generated in the log. This applies when an audit event is reported. The security administrator can also customize audit rules based on the string that is generated in the log. This applies when the system encounters an attack.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Policies** tab.
5. On the **Policies** tab, click the **Audit Rules** sub-tab.
6. Create an audit rule.
   i. Click **New** in the upper-right corner.

ii. In the **Add Policy** dialog box, configure parameters.



iii. Click **Add**.

The system sends an alert email to the specified alert recipient after you create an audit rule. This applies if one string in an audit log of the specified audit type, audit object, or risk level matches the regular expression of the audit rule.

7. Manage audit rules.

   You can create, query, disable, enable, and delete audit rules.

   - Query audit rules

     Specify **Audit Type** and **Audit Target**. Enter a keyword in the search box and click **Search**.

   - Disable an audit rule

     Find the audit rule that you want to disable and click **Disable** in the **Actions** column.

   - Enable an audit rule

     Find the audit rule that has been disabled and click **Enable** in the **Actions** column.

   - Delete an audit rule

     Find the audit rule that you want to delete and click **Delete** in the **Actions** column.

     > **Note** You can delete only custom rules.

Configure alert recipients

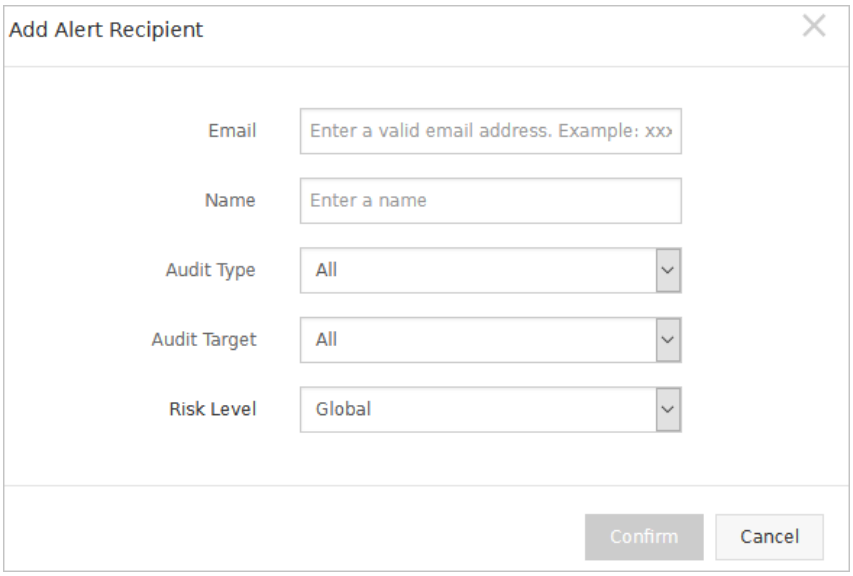
This topic describes how to configure the recipients of alerts on audit events.

Context

After you specify an alert recipient, the system sends a report to the email address of the alert recipient when an audit event occurs.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Policies** tab.
5. On the **Policies** tab, click the **Alert Settings** sub-tab.
6. Create an alert recipient.
   i. Click **New**.
   ii. In the **Add Alert Recipient** dialog box, configure parameters.



   iii. Click **Confirm**.
7. Manage alert recipients.
   - Search for alert recipients

     Specify **Audit Type**, **Audit Target**, and **Risk Level**, enter the keyword of the email address, and then click **Search**.
   - Delete alert recipients

     Find the email address that you want to delete and click **Delete** in the Actions column.

Manage archives of events and logs

This topic describes how to query and download the archives of audit events and raw logs.

Context

You can download the archives of events and logs to analyze audit events. This ensures the security of the Apsara Stack environment.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Policies** tab.
5. On the **Policies** tab, click the **Archiving** sub-tab.
6. Query the archives of events and logs.
   i. Specify **Audit Type** and **Archiving Type**.
   ii. Specify a time range to query.
   iii. Click **Search**.
7. Find the target file where the archive information is stored and click **Download** in the **Actions** column to save the archive file to your computer.

Manage export tasks

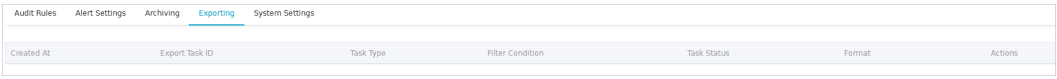
This topic describes how to download or delete exported audit events and logs.

Context

You can export audit events or logs on the **Audit Query** or **Raw Log** tab of the Security Audit page. After you export audit events or logs, you can manage the export tasks on the Exporting sub-tab.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Policies** tab.
5. On the **Policies** tab, click the **Exporting** sub-tab.
6. View the created export tasks.



7. Click **Download** to download audit events or logs to your on-premises server.
8. Click **Delete** to delete the export task.

Modify system settings

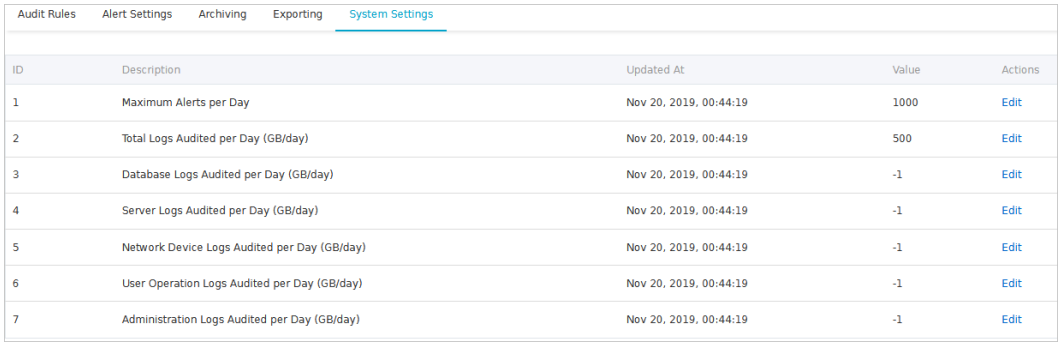
This topic describes how to configure system parameters for security audit.

Context

You can configure system parameters to specify the maximum number of system alerts per day and the maximum number of audits per day for raw logs.

Procedure

1.
2.
3. click **Security Audit**.
4. On the **Security Audit** page, click the **Policies** tab.
5. On the **Policies** tab, click the **System Settings** sub-tab.
6. Find the configuration item that you want to modify and click **Edit** in the **Actions** column.



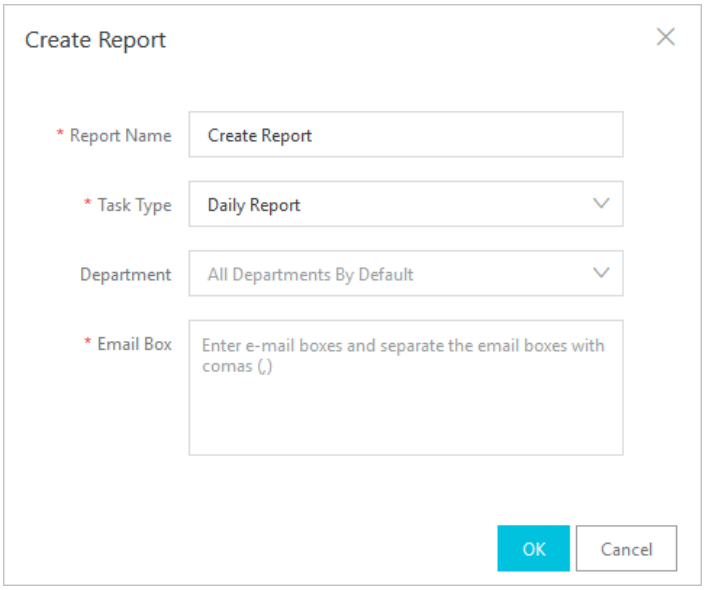
7. Enter a required value in the Value column and click **Confirm** in the Actions column.

22.1.9.5. Create a report task

This topic describes how to create a report task. After you create a report task, the system sends reports on a regular basis.

Procedure

1.
2.
3. , click **Report Management**.
4. On the **Report Management** page, click **Create Report**.
5. In the **Create Report** dialog box, configure parameters.



Parameter	Description
Report Name	The name of the report task. We recommend that you enter information such as the report purpose for easier identification and management.
Task type	The type of the task. Valid values: **Daily Report**, **Weekly Report**, and **Monthly Report**.
Department	The department related to the report.

Parameter	Description
Email Box	The email address of the report recipient. If you enter more than one email address, separate the email addresses with commas (,).

6. Click **Confirm**.

Result

In the report task list, you can view, edit, and delete the newly created report tasks.

22.1.10. Optional security products

22.1.10.1. Anti-DDoS settings

22.1.10.1.1. Overview

In DDoS attacks, attackers exploit the client/server architecture to combine multiple computers into a platform that can launch attacks on one or more targets. This significantly increases the threat of attacks.

The following section describes common DDoS attacks:

- **Network-layer attacks**: A typical example is UDP reflection attacks, such as NTP flood. When this type of attacks are launched, the network of the victim is congested by heavy traffic. As a result, the victim cannot respond to user requests.
- **Transport-layer attacks**: Typical examples include SYN flood and connection flood. When this type of attacks are launched, a large number of connection resources of the target server are consumed. As a result, the server rejects service requests.
- **Session-layer attacks**: A typical example is SSL flood. These attacks consume the SSL session resources of a server to cause denial of service.
- **Application-layer attacks**: Typical examples include DNS flood, HTTP flood, and NTP flood. When this type of attacks are launched, a large number of connection resources of the target server are consumed. As a result, the server rejects service requests.

Apsara Stack Security can redirect, scrub, and re-inject attack traffic to protect your server against DDoS attacks and ensure normal business operations.

> **Note** Apsara Stack Security cannot scrub the traffic between internal networks.

22.1.10.1.2. View and configure DDoS mitigation policies

This topic describes how to view and configure DDoS mitigation policies. Anti-DDoS provides default DDoS mitigation policies and DDoS traffic scrubbing policies.

Context

After an alert threshold of DDoS traffic for an IP address is set, an alert is triggered when traffic to the IP address reaches the threshold. The alert threshold for an IP address must be specified based on the traffic volume. An excessively large traffic volume may indicate DDoS attacks. We recommend that you set an alert threshold to a value that is slightly higher than the peak traffic volume.

Apsara Stack Security supports a global alert threshold, alert threshold for a specific CIDR block, and alert threshold for an IP address.

- Global alert threshold: You cannot specify a global alert threshold. The threshold is automatically specified when Apsara Stack Security is initialized.
- Alert threshold for a CIDR block: You can specify an alert threshold for a CIDR block based on the traffic volume

of the CIDR block. CIDR block-specific alert thresholds allow you to manage the traffic to each CIDR block.

Procedure

1. Log on to Apsara Stack Security Center.
2. choose **Security > Network Security > Traffic Scrubbing**.
3. In the left-side navigation pane, click **DDoS Defense Policy**. On the **Anti-DDoS Policy** page, view and customize DDoS mitigation policies.

Operation	Description
View the default policy	Move the pointer over the icon next to Built-in Default Policy to view the default DDoS mitigation policy.
Customize a policy	Click View to view CIDR block-specific policies. Click +Add to customize a DDoS mitigation policy for a CIDR block.

To customize a policy for a CIDR block, perform the following steps:

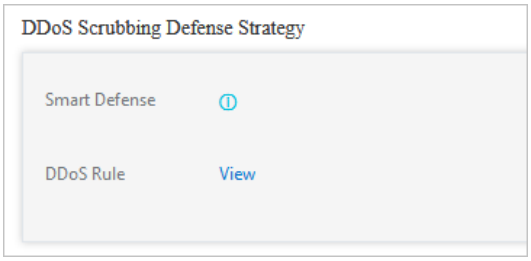
i. Click **+Add** to the right of **Custom Mode**.

ii. In the **Modify** dialog box, configure the parameters.

Parameter	Description
CIDR Block	The CIDR block for which you want to use the alert threshold.
Bandwidth Threshold (pps)	The alert threshold for bandwidth usage in a data center. When the sum of inbound and outbound traffic reaches the threshold specified by this parameter, DDoS detection is triggered. Set this parameter to a value that is slightly higher than the peak traffic volume. We recommend that you set the value to 100 or higher. Unit: Mbit/s.
Packet Threshold (pps)	The alert threshold for the packet rate in a data center. When the sum of inbound and outbound packet rates reaches the threshold specified by this parameter, DDoS detection is triggered. Set this parameter to a value that is slightly higher than the peak packet rate. We recommend that you set the value to 20000 or higher. Unit: packets per second (pps).

iii. Click **OK**.

4. In the **Policy for DDoS Attack Traffic Scrubbing** section, click **View** to view DDoS traffic scrubbing policies.



22.1.10.1.3. View DDoS traffic scrubbing events

This topic describes how to view DDoS traffic scrubbing events.

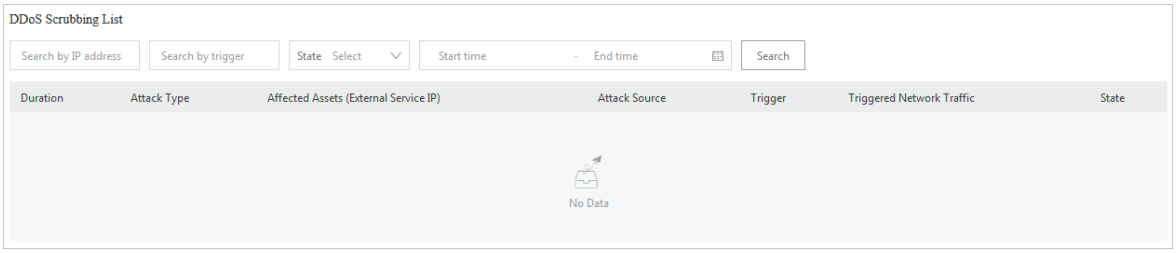
Context

Apsara Stack Security reports security events to Apsara Stack Security Center during and after traffic scrubbing.

Procedure

1. Log on to Apsara Stack Security Center.
2. choose **Security > Network Security > Traffic Scrubbing**.
3. In the left-side navigation pane, click **Network Protection**. On the **Anti-DDoS** page, view anti-DDoS statistics.
4. (Optional)In the **DDoS Attack Traffic Scrubbing** section, specify search conditions and click **Search**.

   Note If you want to view all traffic scrubbing events, skip this step.

Search condition	Description
Search by IP address	The IP address that was under a DDoS attack.
Search by trigger	The metric whose value exceeds the specified alert threshold in the DDoS attack traffic.
Status	The status of DDoS attack traffic scrubbing. Valid values: ○ **Local Scrubbing** ○ **Switching to Anti-DDoS Pro** ○ **Switching to Anti-DDoS Service** ○ **Local Scrubbing Completed** ○ **Cloud Scrubbing Completed** ○ **Under Blackhole Filtering**
Start time and end time	The start time and end time of DDoS attack traffic scrubbing.

5. In the DDoS Scrubbing List section, view details about DDoS traffic scrubbing events.

22.1.10.2. Sensitive Data Discovery and Protection

22.1.10.2.1. Grant access permissions

Before you use Sensitive Data Discovery and Protection (SDDP), you must grant access permissions to SDDP. This topic describes how to authorize SDDP to access the data of your department.

Prerequisites

The name and AccessKey pair of your department are obtained before you grant access permissions on the department. For more information, see the "Obtain the AccessKey pair of an organization" topic in *Apsara Uni-manager Management Console User Guide*. To find the topic, choose **Enterprise > Organizations > Obtain the AccessKey pair of an organization**.

Context

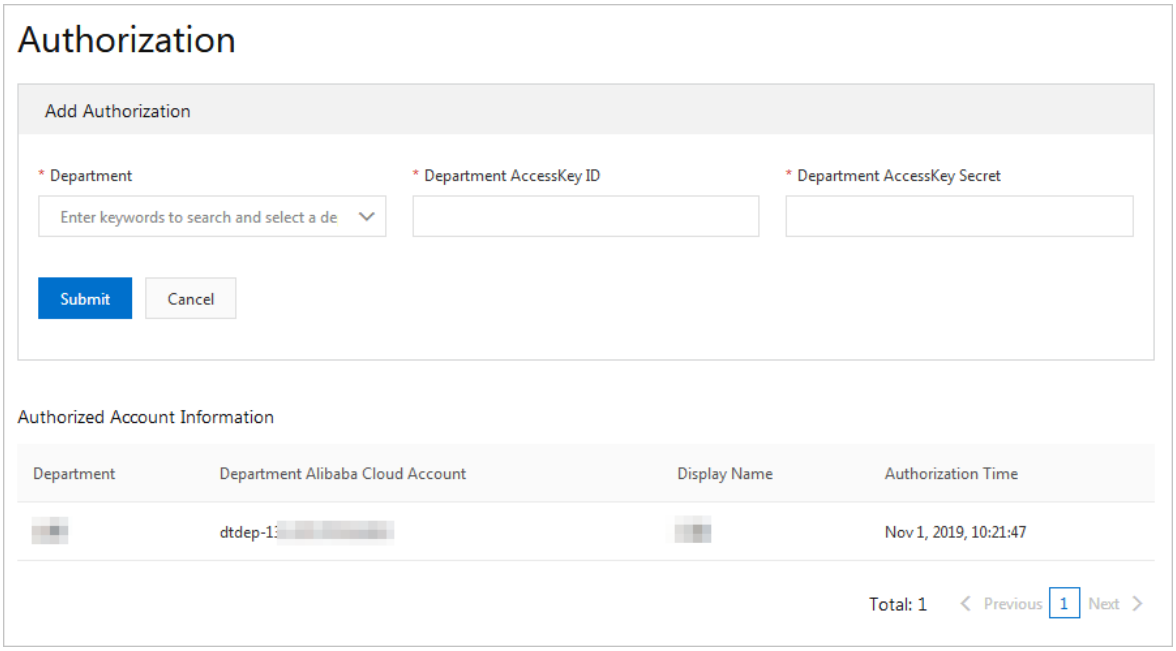
Before you use SDDP, you must complete the following operations:

- Authorize SDDP to access the data of your department.
- Authorize SDDP to access the data of Apsara Stack services of your department. The services include MaxCompute, Object Storage Service (OSS), and Tablestore.

Procedure

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane, click **Authorization**.

> **Note** If SDDP is not authorized to access the data of your department, the **Authorization** page appears. You must configure the parameters on this page.



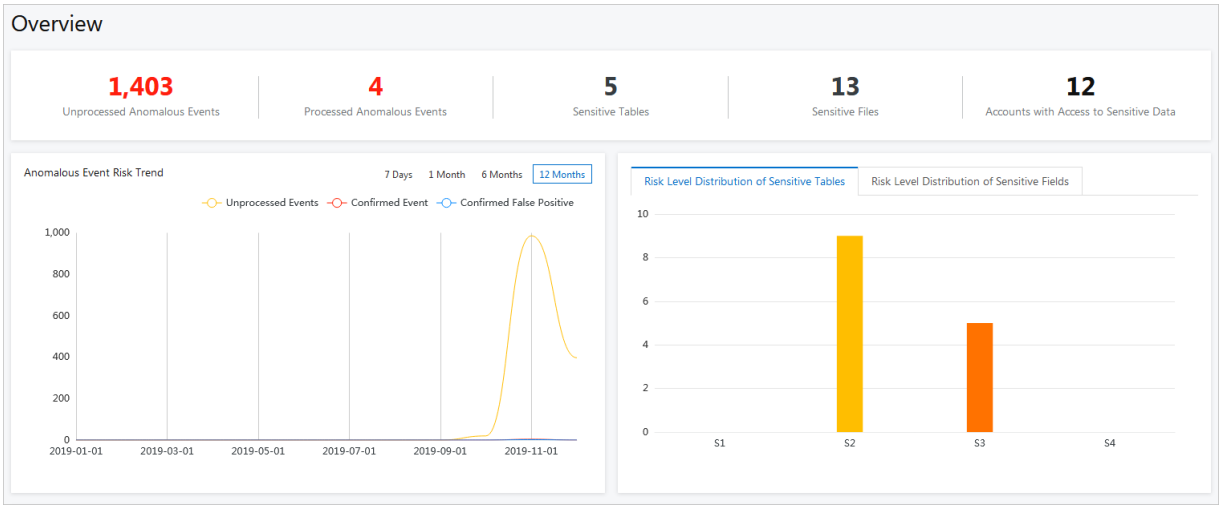
4. In the **Add Authorization** section, authorize SDDP to access the data of your department.
   i. In the **Department** drop-down list, enter a keyword and select the department.
   ii. Configure **Department AccessKey ID** and **Department AccessKey Secret**.
   iii. Click **Submit**.
5. In the **Authorized Account Information** section, view the departments that SDDP is authorized to access.

22.1.10.2.2. SDDP overview

This topic provides an overview about Sensitive Data Discovery and Protection (SDDP). The Overview page of SDDP displays the overall security status of data protected by SDDP. The information on this page allows security administrators to have an overview of sensitive data and take countermeasures in time.

SDDP can detect sensitive data in your data assets based on specific detection rules and track the use of sensitive data. SDDP also provides an overview of sensitive data to help you obtain the security status of your data assets in real time.

If you want to view the overall security status of the sensitive data, log on to Apsara Stack Security Center. In the top navigation bar, choose **Security > Data Security > Sensitive Data Discovery and Protection**. In the left-side navigation pane, click Overview.



- **Overview**: displays the overall information about sensitive data. The information includes **Unprocessed Anomalous Events**, **Processed Anomalous Events**, **Sensitive Tables**, **Sensitive Files**, and **Accounts with Access to Sensitive Data**.
- **Abnormal Event Risk Trend**: displays the trends of different events in a line chart. You can click **7 Days**, **1 Month**, **6 Months**, or **12 Months** to view the trends of different events, such as **Unprocessed Events**, **Confirmed Event**, and **Confirmed False Positive**.
- **Risk Level Distribution of Sensitive Tables**: displays the distribution of sensitive tables at each sensitivity level, including S3 (high sensitivity), S2 (moderate sensitivity), S1 (low sensitivity), and N/A (unknown sensitivity).
- **Risk Level Distribution of Sensitive Fields**: displays the distribution of sensitive fields at each sensitivity level, including S3 (high sensitivity), S2 (moderate sensitivity), S1 (low sensitivity), and N/A (unknown sensitivity).
- **Data Flow Status**:
  - Displays the dynamic statistics on core data flows in DataHub and Data Integration.
  - Provides a data flowchart. The flowchart dynamically shows data flows and abnormal output. You can click an anomalous event in the flowchart to go to the **Abnormal Data Flow** page.

    You can monitor the data links among different entities, such as data storage services, data transmission services, data stream processing services, external databases, and external files. The data storage services include MaxCompute, AnalyticDB for MySQL, Object Storage Service (OSS), and Tablestore. The data transmission services include DataHub and Data Integration. The data stream processing services include Blink.

22.1.10.2.3. Data asset authorization

22.1.10.2.3.1. Authorize SDDP to access data assets

Sensitive Data Discovery and Protection (SDDP) must be authorized to access your data assets before it can detect sensitive data in the data assets. Supported data assets include Object Storage Service (OSS) buckets, ApsaraDB RDS instances, PolarDB-X databases, Tablestore instances, self-managed databases hosted on Elastic Compute Service (ECS) instances, MaxCompute projects, AnalyticDB for MySQL clusters, ApsaraDB for OceanBase clusters, and AnalyticDB for PostgreSQL instances. This topic describes how to authorize SDDP to access your data assets.

Context

SDDP can access and scan specific data assets to detect and mask sensitive data only after you grant the required permissions to SDDP.

> **Notice** 已开启授权的OSS Bucket（OSS文件桶）会消耗您的OSS存储容量，已开启授权的数据库或项目会消耗您的数据库和项目数。只有在OSS存储容量、数据库和项目数量充足时，您才可以成功进行相应授权操作。您可以在**云上托管**页面查看剩余的OSS存储容量、数据库和项目数。

For more information about how to authorize SDDP to access supported data assets, see the following sections:

- Authorize SDDP to access OSS buckets
- Authorize SDDP to access ApsaraDB RDS instances
- Authorize SDDP to access PolarDB-X databases
- Authorize SDDP to access Tablestore instances
- Authorize SDDP to access self-managed databases hosted on ECS instances
- Authorize SDDP to access MaxCompute projects
- Authorize SDDP to access AnalyticDB for MySQL clusters or AnalyticDB for PostgreSQL instances
- Authorize SDDP to access ApsaraDB for OceanBase clusters

Authorize SDDP to access OSS buckets

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **OSS** tab, grant the required permissions on instances or buckets.
   - If you want to grant permissions on a single instance or bucket, turn on the switches in the **Identify permissions**, **Desensitization permissions**, **OCR Authority**, and **Audit permissions** columns of the instance or bucket. Then, configure the **Sensitive data sampling** and **Audit log archiving** parameters.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Desensitization permissions	The permissions to mask sensitive data in selected data assets.
OCR Authority	The permissions to detect sensitive data in text on images.
Audit permissions	The permissions to audit data in selected data assets.
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**. **Note** You do not need to activate Log Service to archive the audit logs generated by SDDP.

- If you want to grant permissions on multiple instances or buckets at the same time, perform the following steps:
  a. Select the required instances or buckets and click **Batch operation**.
  b. In the **Batch processing for selected assets** dialog box, turn on the switches of detection, audit, masking, and OCR permissions, and configure the parameters that remain.
  c. Click **Ok**.

After SDDP is authorized, SDDP scans the OSS buckets to detect sensitive data. If SDDP scans an OSS bucket for the first time, SDDP automatically performs a full scan.

In the list of OSS buckets on which SDDP has access permissions, you can modify or revoke permissions on the OSS buckets. If you revoke permissions on an OSS bucket, SDDP no longer scans the OSS bucket.

> **Note** SDDP scans only the accessible OSS buckets and analyzes the risks of sensitive data detected in these OSS buckets.

Authorize SDDP to access ApsaraDB RDS instances

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **RDS** tab.
5. On the **RDS** tab, click **Not authorized**.
6. Find the instance that you want SDDP to access and enter the required database username and its password in the **Username** and **Password** columns.

   You can also click **Batch password import** to import logon information for multiple data assets at a time. For more information, see Import the logon information for multiple data assets at the same time.

   > **Notice** If the username or password is not correct, the authorization fails. Make sure that the information you enter is correct.

7. Select the databases that you want SDDP to access and click **Batch operation**.

   You can also click **One-click authorization** in the Actions column of an instance to grant all its permissions.
8. In the Batch processing for selected assets dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Audit permissions	The permissions to audit data in selected data assets. SDDP allows you to collect audit logs that cover the data generation, update, and use of your data assets. The log information includes the audit rule that is triggered for a data asset, the type of the data asset, the type of the operation that triggers the audit rule, and the operator account. SDDP安全审计功能的使用，请参见Create an audit rule。
Desensitization permissions	The permissions to mask sensitive data in selected data assets.

Parameter	Description
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

9. Click **Ok**.

   > **Note** If the authorization fails, check whether the username and password are correct.

   After SDDP is authorized, SDDP scans the databases to detect sensitive data.

   In the list of databases on which SDDP has access permissions, you can modify or revoke permissions on the databases. You can modify only the username and password of a valid database account. If you revoke permissions on a database, SDDP no longer scans the database.

Authorize SDDP to access PolarDB-X databases

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **DRDS** tab.
5. On the **DRDS** tab, click **Not authorized**.
6. Find the instance that you want SDDP to access and enter the required database username and its password in the **Username** and **Password** columns.

   You can also click **Batch password import** to import logon information for multiple data assets at a time. For more information, see Import the logon information for multiple data assets at the same time.

   > **Notice** If the username or password is not correct, the authorization fails. Make sure that the information you enter is correct.

7. Select the databases that you want SDDP to access and click **Batch operation**.

   You can also click **One-click authorization** in the Actions column of an instance to grant all its permissions.

8. In the Batch processing for selected assets dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.

Parameter	Description
Audit permissions	The permissions to audit data in selected data assets. SDDP allows you to collect audit logs that cover the data generation, update, and use of your data assets. The log information includes the audit rule that is triggered for a data asset, the type of the data asset, the type of the operation that triggers the audit rule, and the operator account. SDDP安全审计功能的使用，请参见Create an audit rule。
Desensitization permissions	The permissions to mask sensitive data in selected data assets.
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

9. Click **Ok**.

> **Note** If the authorization fails, check whether the username and password are correct.

After SDDP is authorized, SDDP scans the databases to detect sensitive data.

In the list of databases on which SDDP has access permissions, you can modify or revoke permissions on the databases. You can modify only the username and password of a valid database account. If you revoke permissions on a database, SDDP no longer scans the database.

Authorize SDDP to access Tablestore instances

You can authorize SDDP to access one or more Tablestore instances.

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **OTS** tab.
5. On the **OST** tab, grant the required permissions on instances or buckets.
   - If you want to grant permissions on a single instance or bucket, turn on the switches in the **Identify permissions**, **Desensitization permissions**, and **Audit permissions** columns of the instance or bucket. Then, configure the **Sensitive data sampling** and **Audit log archiving** parameters.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Desensitization permissions	The permissions to mask sensitive data in selected data assets.

Parameter	Description
Audit permissions	The permissions to audit data in selected data assets.
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

- If you want to grant permissions on multiple instances or buckets at the same time, perform the following steps:
  a. Select the required instances or buckets and click **Batch operation**.
  b. In the **Batch processing for selected assets** dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.
  c. Click **Ok**.

After SDDP is authorized, SDDP scans the instances to detect sensitive data.

Authorize SDDP to access self-managed databases hosted on ECS instances

A self-managed database hosted on an ECS instance must meet the following requirements before it can be scanned by SDDP:

- The ECS instance resides in a virtual private cloud (VPC).
- The database is a MySQL or SQL Server database.

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **ECS self-built database** tab.
5. On the **ECS self-built database** tab, click **Add data assets**.
6. In the **Asset authorization** dialog box, configure the parameters and click **Next**.

   The following table describes the parameters.

Parameter	Description
Region	The region of the self-managed database that you want to authorize SDDP to access.
ECS instance ID	The ID of the ECS instance on which the self-managed database is hosted.
Database type	The type of the self-managed database that you want to authorize SDDP to access. Valid values: MySQL and SQL Server.

Parameter	Description
Library name	The name of the self-managed database that you want to authorize SDDP to access. **Note** If you want to authorize SDDP to access other self-managed databases hosted on the same ECS instance, click **Add Database**.
Port	The port number used to connect to the self-managed database.
User name	The username of the account that you use to connect to the self-managed database.
Password	The password of the account that you use to connect to the self-managed database.

7. In the Batch processing for selected assets dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Audit permissions	The permissions to audit data in selected data assets. SDDP allows you to collect audit logs that cover the data generation, update, and use of your data assets. The log information includes the audit rule that is triggered for a data asset, the type of the data asset, the type of the operation that triggers the audit rule, and the operator account. SDDP安全审计功能的使用，请参见Create an audit rule。
Desensitization permissions	The permissions to mask sensitive data in selected data assets.
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

8. Click **Ok**.
   After SDDP is authorized, SDDP scans the databases to detect sensitive data.

Authorize SDDP to access MaxCompute projects

1.
2.

3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **MaxCompute** tab.
5. On the **MaxCompute** tab, grant the required permissions on instances or buckets.
   - If you want to grant permissions on a single instance or bucket, turn on the switches in the **Identify permissions**, **Desensitization permissions**, and **Audit permissions** columns of the instance or bucket. Then, configure the **Sensitive data sampling** and **Audit log archiving** parameters.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Desensitization permissions	The permissions to mask sensitive data in selected data assets.
Audit permissions	The permissions to audit data in selected data assets.
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

   - If you want to grant permissions on multiple instances or buckets at the same time, perform the following steps:
     a. Select the required instances or buckets and click **Batch operation**.
     b. In the **Batch processing for selected assets** dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.
6. Click **Ok**.

> **Note** If the authorization fails, check whether the permission parameters are correctly configured and whether the SDDP account is added to the project.

After SDDP is authorized, SDDP scans the projects to detect sensitive data.

In the list of projects on which SDDP has access permissions, you can revoke permissions on the projects. If you revoke permissions on a project, SDDP no longer scans the project.

Authorize SDDP to access AnalyticDB for MySQL clusters or AnalyticDB for PostgreSQL instances

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **ADS** or **GPDB** tab.
5. On the ADS or GPDB tab, click **Add data assets**.
6. In the **Add data assets** dialog box, configure the parameters and click **Ok**.

The following table describes the parameters used to authorize SDDP to access an AnalyticDB for MySQL cluster.

Parameter	Description
Region	The region of the AnalyticDB for MySQL database that you want to authorize SDDP to access.
Instance Name	The name of the cluster to which the AnalyticDB for MySQL database belongs.
Database Name	The name of the AnalyticDB for MySQL database.
User name	The username and password of the account that you use to connect to the AnalyticDB for MySQL database.
Password	
Automatic scanning	The switch of triggering scans on the AnalyticDB for MySQL database each time identification rule settings are modified.

7. On the ADS or GPDB tab, grant the required permissions on multiple instances or buckets at the same time.
   - If you want to grant permissions on a single instance or bucket, turn on the switches in the **Identify permissions**, **Desensitization permissions**, and **Audit permissions** columns of the instance or bucket. Then, configure the **Sensitive data sampling** and **Audit log archiving** parameters.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Desensitization permissions	The permissions to mask sensitive data in selected data assets.
Audit permissions	The permissions to audit data in selected data assets.
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

   - If you want to grant permissions on multiple instances or buckets at the same time, perform the following steps:
     a. Select the required instances or buckets and click **Batch operation**.
     b. In the **Batch processing for selected assets** dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.
     c. Click **Ok**.

After SDDP is authorized, SDDP scans the databases to detect sensitive data.

Authorize SDDP to access ApsaraDB for OceanBase clusters

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click the **OceanBase** tab.
5. On the **OceanBase** tab, click **Add data assets**.
6. In the **Add data assets** dialog box, configure the parameters and click **Next**.

   The following table describes the parameters.

Parameter	Description
Region	The region of the ApsaraDB for OceanBase database that you want to authorize SDDP to access.
Database type	The type of the ApsaraDB for OceanBase database. Valid values: MySQL and Oracle.
Cluster Name	The name of the cluster to which the ApsaraDB for OceanBase database belongs.
Tenant Name	The name of the tenant to which the ApsaraDB for OceanBase database belongs.
Database Name	The name of the ApsaraDB for OceanBase database. **Note** If you want to authorize SDDP to access other ApsaraDB for OceanBase databases hosted on the same ECS instance, click **Add Database**.
Link Address	The endpoint that you use to connect to the ApsaraDB for OceanBase database.
User name	The username and password of the account that you use to connect to the ApsaraDB for OceanBase database.
Password	

7. In the Batch processing for selected assets dialog box, turn on the switches of detection, audit, and masking permissions, and configure the parameters that remain.

Parameter	Description
Identify permissions	The permissions to detect sensitive data in selected data assets.
Audit permissions	The permissions to audit data in selected data assets. SDDP allows you to collect audit logs that cover the data generation, update, and use of your data assets. The log information includes the audit rule that is triggered for a data asset, the type of the data asset, the type of the operation that triggers the audit rule, and the operator account. SDDP安全审计功能的使用，请参见Create an audit rule。
Desensitization permissions	The permissions to mask sensitive data in selected data assets.

Parameter	Description
Sensitive data sampling	The number of sensitive data samples to reserve after SDDP detects sensitive data. If you turn on the switch in **Identify permissions**, you must also specify this parameter. If SDDP detects sensitive data in a data asset, SDDP collects samples of the detected data. You can use the sensitive data samples for further analysis. Valid values: **0**, **5**, and **10**.
Audit log archiving	The duration to retain the audit logs of selected data assets. If you turn on the switch in **Audit permissions**, you must also specify this parameter. Valid values: **30 days**, **90 days**, and **180 days**.

8. Click **Ok**.
   After SDDP is authorized, SDDP scans the databases to detect sensitive data.

Import the logon information for multiple data assets at the same time

SDDP allows you to upload an EXCEL file to import the logon information for multiple data assets, including RDS databases, PolarDB-X databases, and self-managed databases hosted on ECS instances, at the same time to improve authorization efficiency. The following procedure describes how to import the logon information for multiple data assets at the same time:

1.
2.
3. choose **Data protection authorization > Data Asset authorization**.
4. On the **Cloud hosting** page, click **Batch password import** in the upper-right corner.
5. In the **Batch password import** dialog box, click **SDDP Authorization File Template.xlsx**.
6. Open the downloaded file, enter the usernames and passwords used to access each data asset in the **username** and **password** columns, and then save the file.

   If you modify the existing usernames and passwords in the downloaded file and upload the file to SDDP, the logon information saved in SDDP is updated.
7. In the **Batch password import** dialog box, click **File Upload** to upload the template file that you have edited.
8. Click **Ok**.
   After you upload the Excel file, the usernames and passwords that you enter in the file are synchronized to the **Username** and **Password** columns for the relevant databases on the **RDS**, **DRDS**, and **ECS self-built database** tabs. Then, you can authorize SDDP to access these data assets without the need to manually enter the logon information on the **Cloud hosting** page.

22.1.10.2.3.2. Manage usernames and passwords of databases

Sensitive Data Discovery and Protection (SDDP) can detect sensitive data that is stored in a data source only after SDDP is authorized to access the data source. To authorize SDDP to access a data source, you must add the username and password that are used to connect to a database of the data source. This topic describes how to view and add the username and password that are used to connect to a database.

Context

SDDP allows you to manage the usernames and passwords that are used to connect to databases in ApsaraDB RDS and PolarDB-X.

View the username and password of a database

1.

2.
3. choose **Data protection authorization > Authorized account management**.
4. Click the tab that displays the required data source. In this example, click the **RDS** tab.
5. (Optional)Specify filter conditions to search for an ApsaraDB RDS instance.

   > **Note** If you want to view all ApsaraDB RDS instances, skip this step.

Filter condition	Description
Region	The region where the ApsaraDB RDS instance resides.
Instance/Bucket	The name of the ApsaraDB RDS instance.
Database type	The type of the database engine that is run by the ApsaraDB RDS instance. ApsaraDB RDS and PolarDB-X support the MySQL and SQL Server database engines.

6. In the instance list, view the usernames and passwords that are used to connect to the databases of the ApsaraDB RDS instance.

Add the username and password of a database

1.
2.
3. choose **Data protection authorization > Authorized account management**.
4. Click the tab that displays the required data source. In this example, click the **RDS** tab.
5. Find the ApsaraDB RDS instance and configure the **Username** and **Password** parameters that are used to connect to a database.

   > **Notice** If the username or password is invalid, SDDP fails to be authorized. Make sure that the information that you enter is valid.

6. Click **Add**.
   After the username and password of the database are added, **Status** of the ApsaraDB RDS instance changes to **Added successfully**.

What's next

If you want to change the username or password of a database, find the instance and click **Edit** in the **Actions** column.

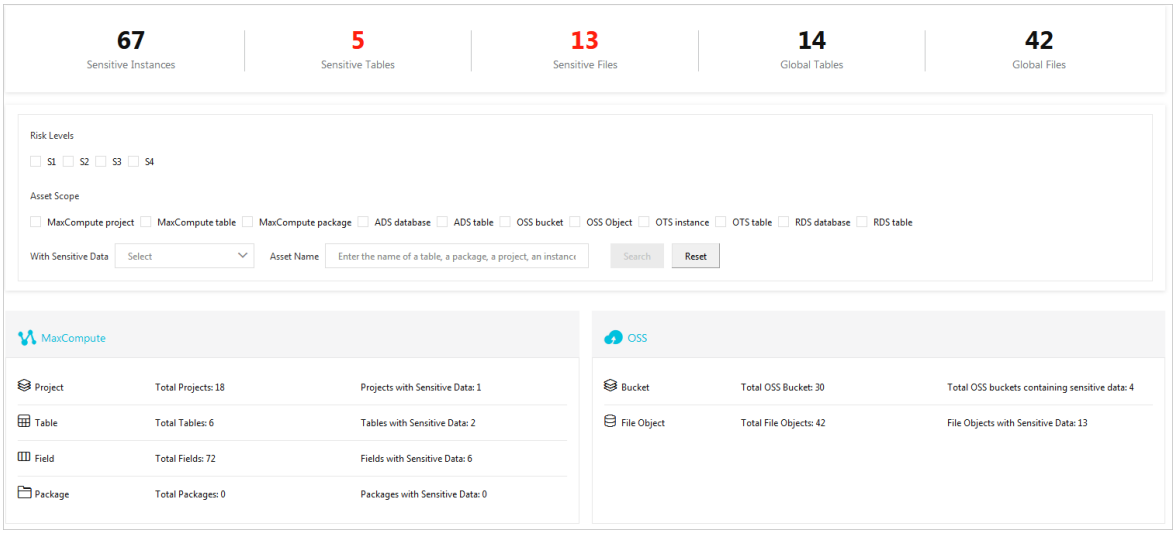
After you change the parameter values,**Save**.

22.1.10.2.4. Sensitive data discovery

22.1.10.2.4.1. Sensitive data overview

This topic describes the Sensitive Data Overview page that displays the overall security status of your data assets.

1. Log on to Apsara Stack Security Center.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Identification > Sensitive Data Overview**.
4. On the **Sensitive Data Overview** page, view the overall security status of sensitive data.



- You can view the overall information about sensitive data. The information includes **Total number of instances**, **Sensitive Tables**, **Sensitive Files**, **Global Tables**, and **Global Files**.
- You can search for sensitive data based on conditions such as the risk level, asset scope, sensitive data type, and asset name.
- You can view the statistics on the access information and sensitive data of cloud services, such as **MaxCompute**, **OSS**, **AnalyticDB for MySQL**, and **Tablestore**, in real time.

22.1.10.2.4.2. View statistics on sensitive data

Sensitive Data Discovery and Protection (SDDP) can detect sensitive data in data sources, such as Object Storage Service (OSS) buckets, ApsaraDB RDS instances, and MaxCompute projects. This topic describes how to view statistics on sensitive data that is detected by SDDP.

View statistics on sensitive data detected in OSS

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **OSS** tab, find the OSS bucket whose details you want to view and click **File details** in the Actions column.
5. In the **OSS object query** panel, view the proportions of sensitive objects at each sensitivity level, the top five sensitive data detection rules that are most frequently hit, and the list of objects in which the sensitive data is detected.
   - **Proportions of sensitive objects**

     In the **Proportions of sensitive objects** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of objects at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five sensitive data detection rules that are most frequently hit and the number of times that each rule is hit.
   - **List of objects in which the sensitive data is detected**

     In the object list, you can view the information about the objects in which the sensitive data is detected. The information includes the object name, size, type, and number of sensitive fields that are detected in the object. You can click **Hit details** in the Actions column of an object to view the details about the sensitive data detection rules that are hit by the object.

View statistics on sensitive data detected in ApsaraDB RDS

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **RDS** tab.
5. On the **RDS** tab, find the ApsaraDB RDS instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in MaxCompute

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **MaxCompute** tab.
5. On the **MaxCompute** tab, find the MaxCompute project whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in self-managed databases that are hosted on ECS instances

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **ECS self-built database** tab.
5. On the **ECS self-built database** tab, find the database instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in PolarDB-X

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **DRDS** tab.
5. On the **DRDS** tab, find the database instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in Tablestore

You can view statistics on sensitive data detected in Tablestore.

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **OTS** tab.
5. On the **OTS** tab, find the Tablestore instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in AnalyticDB for PostgreSQL

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **GPDB** tab.
5. On the **GPDB** tab, find the instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in AnalyticDB for MySQL

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **ADS** tab.
5. On the **ADS** tab, find the instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in DataWorks

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **DataWorks** tab.
5. On the **DataWorks** tab, find the DataWorks workspace whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

View statistics on sensitive data detected in ApsaraDB for OceanBase

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Sensitive Data Identification > Sensitive data assets** .
4. On the **Sensitive data assets** page, click the **OceanBase** tab.
5. On the **OceanBase** tab, find the instance whose details you want to view and click **Table Details** in the Actions column.
6. In the **Table Query** panel, view the proportions of sensitive tables, the top five sensitive data detection rules that are most frequently hit, and the list of tables in which the sensitive data is detected.
   - **Proportions of tables**

     In the **Proportion of tables** section, you can view the numbers of objects at the following sensitivity level: Highly sensitive, Medium sensitive, Low sensitivity, and Unrecognized. You can also view a pie chart that shows the proportions of tables at each level.
   - **Top five rules that are most frequently hit**

     In the **Hit Rule Top5** section, you can view the top five rules that are most frequently hit and the number of times that each rule is hit.
   - **List of tables in which the sensitive data is detected**

     In the table list, you can view the information about the tables in which the sensitive data is detected. The information includes the table name, total number of rows, total number of fields, number of sensitive fields, and rules that are hit. You can click **Column details** in the Actions column to view the field details, such as the fields that hit sensitive data detection rules and the sensitivity levels of the sensitive fields.

22.1.10.2.4.3. Query sensitive data

The Sensitive data search page displays all the sensitive data that is detected in your data assets. You can specify one or more types of sensitive data to query and view the distribution of the sensitive data across your data assets. This topic describes how to query sensitive data.

Procedure

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane, choose **Sensitive Data Identification > Sensitive data search**.
4. On the **Sensitive data search** page, specify filter conditions based on your business requirements.

   Sensitive Data Discovery and Protection (SDDP) provides the following filter conditions:
   - **Hit data**: the type of sensitive data. You can select multiple data types, such as email addresses and mobile phone numbers.
   - **Enter file name to search** or **Enter table name to search**: the name of the object or table in which the sensitive data is detected. Fuzzy match is supported.
   - **Region**: the region where the data assets reside.
   - **Bucket, Instance, or Project**: the name of the bucket, instance, or project in which the sensitive data is detected.

   > **Note** If you specify multiple filter conditions, SDDP returns the sensitive data that meets all the specified filter conditions.
5. Click **Search**.

   In the result list of the **Sensitive data search** page, you can view the information about the objects or tables in which the sensitive data is detected. You can group the objects or sort the tables by using the following

methods:

- **Group objects by sensitivity level**

  On the **OSS-file** tab, set the **Sensitivity level** parameter to S1, S2, or S3 to display the objects that contain sensitive data by sensitivity level.

- **Sort tables based on the total number of rows or sensitive fields in ascending or descending order**

  On a tab such as the **RDS-table** tab, click the icon to the right of **Total number of rows** or **Sensitive column**. This way, tables that contain sensitive data are sorted based on the total number of rows or sensitive fields in ascending or descending order. The first time you click the icon, the tables are sorted in descending order. The next time you click the icon, the tables are sorted in ascending order.

6. Find the object or table that contains sensitive data. To view the details of sensitive data in an object, click **Details** in the **Operation** column. To view the details of sensitive data in a table, click **Column details** in the **Operation** column.

   In the **Hit query** panel for a bucket or the **Column details** panel for a table, you can view the following details of all the sensitive data that is detected in the object or table:

   - **Column name**: the name of the sensitive field that is detected in the table.

     > **Note** This parameter is displayed only in the **Column details** panel for a table in an ApsaraDB RDS instance, MaxCompute project, self-managed database that is hosted on an Elastic Compute Service (ECS) instance, PolarDB-X database, Tablestore instance, AnalyticDB for MySQL cluster, or AnalyticDB for PostgreSQL instance. This parameter is not displayed in the **Hit query** panel for an OSS bucket.

   - **Hit Rule**: the type and name of the sensitive data detection rule that is hit.
   - **Sensitivity level**: the sensitivity level of the detected sensitive data.
   - **Number of hits**: the number of times that the sensitive data detection rule is hit in the object.

     > **Note** This parameter is displayed in the **Hit query** panel for an OSS bucket.

   - **Data Sampling**: the samples that are collected from the sensitive data. To configure the **Sensitive data sampling** parameter, perform the following operations: Choose **Security > Data Security > Sensitive Data Discovery and Protection**. Then, choose **Data protection authorization** > Data Asset authorization. On the **Cloud hosting** page, set the Sensitive data sampling parameter to 0, 5, or 10. The number of samples displayed in Data Sampling does not exceed the value of **Sensitive data sampling** that you configure when you authorize SDDP to protect your data assets.

22.1.10.2.4.4. Manage scan tasks

Sensitive Data Discovery and Protection (SDDP) automatically scans for sensitive data in the data assets that SDDP is authorized to access. On the Identify task monitoring page, you can view the details of scan tasks for the data assets and rescan the data assets.

Context

SDDP can monitor scan tasks that detect sensitive data in Object Storage Service (OSS), ApsaraDB RDS, MaxCompute, self-managed databases that are hosted on Elastic Compute Service (ECS) instances, PolarDB-X, Tablestore, ApsaraDB for OceanBase, AnalyticDB for MySQL, and AnalyticDB for PostgreSQL.

After you authorize SDDP to access specific data assets, SDDP creates and runs scan tasks for these data assets to detect sensitive data. By default, the **automatic scan** feature is enabled for the scan tasks. This feature allows SDDP to run a full scan on the data assets that SDDP is authorized to access and scan the data that is newly written to or modified in these data assets at an interval of 4 hours. In addition, after you create or modify a sensitive data detection rule, SDDP automatically reruns scan tasks for which the automatic scan feature is enabled.

View the details of scan tasks

On the **Identify task monitoring** page, you can view the details of each scan task. The details include the related data asset, task status, and time when the task is complete. To view the details of scan tasks, perform the following steps:

1.
2.
3. choose **Sensitive Data Identification > Identification task monitoring**
4. On the Identify task monitoring page, click the tab of the data source for which you want to view scan tasks.
5. (Optional)Select the region, enter the name of the data asset, specify the start and end of the time range to query, and then click **Search**. You can enter the name of a bucket or instance.
6. In the task list, view the details of each scan task. The details include the related data asset, task status, and time when the task is complete.

Rescan your data assets

You can rescan your data assets in the following scenarios:

- If the **automatic scan** feature is not enabled for a scan task, the scan task is not run after the task is created. In this case, you must rescan your data assets.
- If you enable the **automatic scan** feature for a scan task, SDDP automatically scans the data that is newly written to or modified in the specific data asset at an interval of 4 hours. If you want to immediately scan the specific data asset after you modify the data in the data asset, you can rescan the data asset.

To rescan a data asset for sensitive data, perform the following steps:

1.
2.
3. On the Identify task monitoring page, click the tab of the data source for which you want to rescan data assets.
4. Find the required data asset and click **Rescan** in the **Operation** column.
5. In the **Confirm rescan** dialog box, click **OK**.

   In most cases, the rescan process requires approximately 10 minutes to complete. Wait until the data asset is scanned.

   After the rescan is started, Scan Status of the asset changes to **Scanning** or **Waiting**. The percentage that appears in the **Scan Status** column indicates the progress of the scan task.

What's next

After the scan is complete, Scan Status of the asset changes to **Complete**. If you want to view the latest scan results, you can perform the following operations: In the top navigation bar, choose **Security > Data Security > Sensitive Data Discovery and Protection**. In the left-side navigation pane, choose **Sensitive Data Identification** > Identify task monitoring. Then, click the tab of the data source for which you want to view the scan results.

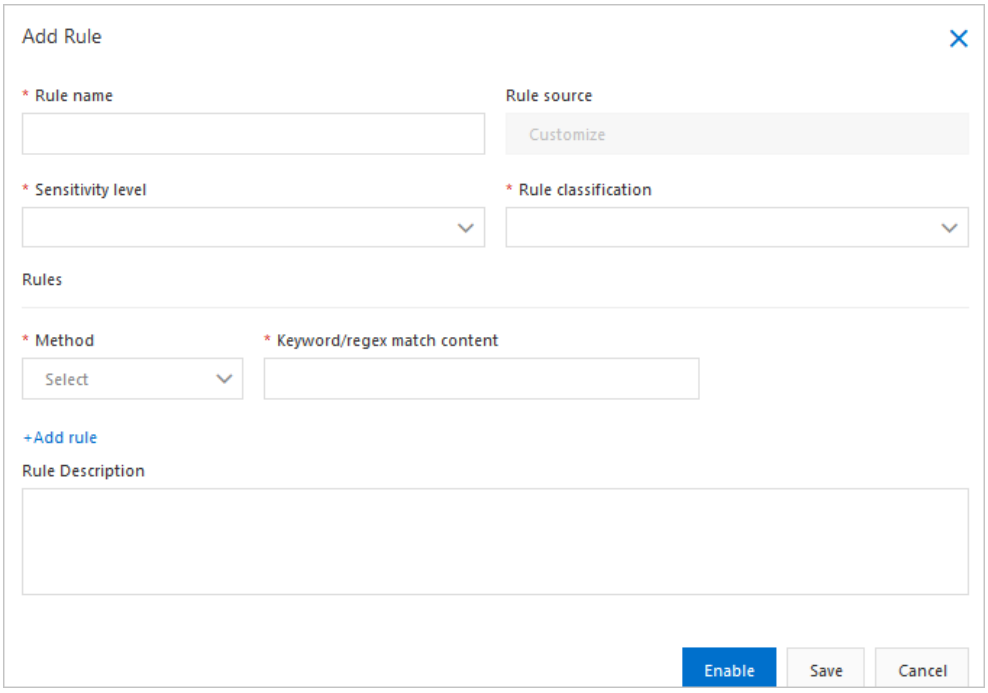
22.1.10.2.4.5. Manage detection rules

Sensitive Data Discovery and Protection (SDDP) allows you to customize the detection rules for classifying sensitive data. You can view and configure detection rules to detect sensitive data. This topic describes how to create and manage custom detection rules, view built-in detection rules, and modify sensitivity levels.

Create a custom detection rule

SDDP detects sensitive data in files or tables based on specified rules and generates alerts. You can customize detection rules to detect sensitive data based on your business requirements. To customize a detection rule, perform the following steps:

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Identification > Identification Rules**. On the **Identification Rules** page, click **Add rule**.
4. In the **Add Rule** dialog box, configure parameters.



The following table describes the parameters used to create a custom detection rule.

Parameter	Description
Rule name	The name of the detection rule.
Rule source	The source of the detection rule. The default value is **Customize** and cannot be changed.
Sensitivity level	The sensitivity level for the detection rule. Valid values: ○ **N/A: Public**: non-sensitive ○ **S1: Internal**: low sensitive ○ **S2: Secret**: moderately sensitive ○ **S3: Confidential**: highly sensitive

Parameter	Description
Rule classification	The class of the sensitive data that the detection rule can detect. Valid values: ○ Personal and sensitive information ○ Device sensitive information ○ Key sensitive information ○ Sensitive picture information ○ Sensitive corporate information ○ Location-sensitive information ○ Universal sensitive information
Rules	The content of the detection rule. The content is used to match sensitive fields or text. You must set the **Method** parameter and enter the keyword that is used to detect sensitive data in the **Keyword/regex match content** field. If you want to create a custom detection rule to detect the mobile phone number *133 1234*****, you must set the **Method** parameter to **Contains** and enter *1331234***** in the **Keywords/regex match content** field. **Note** The keyword must be a precise value, such as a specific mobile phone number, email address, or ID card number. After a detection rule is created, the detection rule appears in the rule list. However, the rule list does not display the details of the rule. You can click **Details** in the Operation column to view the details of the detection rule.

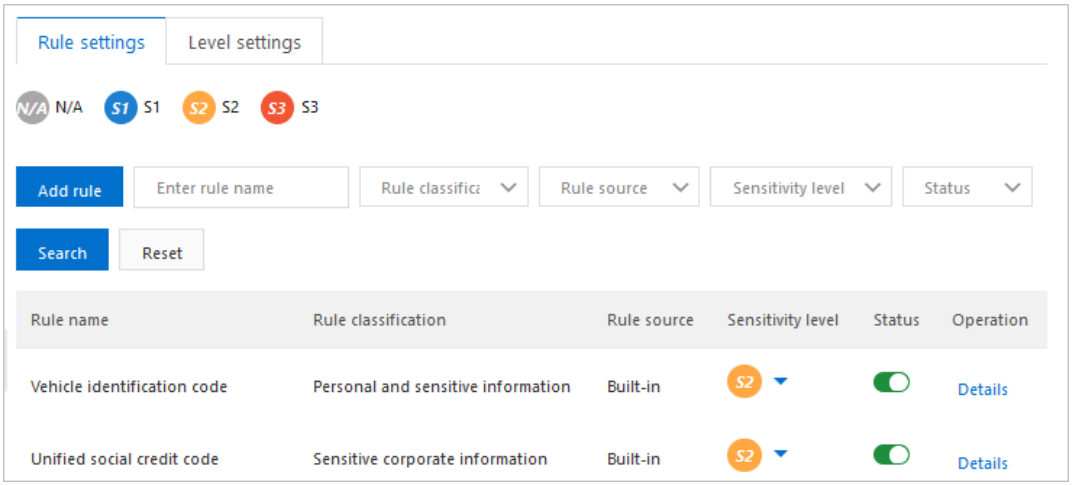
5. Click **Enable** or **Save**.
   - **Enable**: If you click **Enable**, the detection rule is created and enabled. SDDP starts to detect sensitive data based on the detection rule.
   - **Save**: If you click **Save**, the detection rule is created but is not enabled. To enable the detection rule, you must turn on the switch in the Status column for the detection rule in the rule list.

Note
- SDDP detects sensitive data based on all sensitive data detection rules that are enabled.
- A detection rule takes effect after it is created and enabled. If you want to temporarily exclude specific data from sensitive data, you can disable the specific detection rule. After you disable a detection rule, SDDP no longer detects sensitive data based on the detection rule. We recommend that you enable all detection rules to reduce risks.
- You can modify and delete custom detection rules. You can view built-in detection rules but cannot modify or delete them.

View built-in detection rules

The built-in detection rules that SDDP provides apply to various types of common sensitive data, such as mobile phone numbers and ID card numbers. You can view all information about a built-in detection rule, such as the rule type, rule name, and sensitivity level. You cannot view the rule definition. To view built-in detection rules, perform the following steps:

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Identification > Identification Rules**. On the **Identification Rules** page, select **Built-in** from the **Rule source** drop-down list.
4. View built-in detection rules in the list that appears.



You can view the information about each built-in detection rule, such as **Rule name**, **Rule classification**, and **Rule source**.

5. Find a built-in detection rule whose details you want to view and click **Details** in the Operation column.
6. In the **Rule details** dialog box, view the details of the built-in detection rule.

   You can view **Rule name**, **Rule source**, **Rule classification**, and **Sensitivity level** of a built-in detection rule.

Modify a sensitivity level

SDDP allows you to modify the name and description of a sensitivity level. To modify a sensitivity level, perform the following steps:

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Identification > Identification Rules**. On the Identification Rules page, click the **Level settings** tab.
4. Find the sensitivity level that you want to modify and click **Edit** in the Actions column.
5. In the **Sensitivity level** dialog box, modify the information in the Sensitivity level and Description fields.

   By default, SDDP marks sensitive data with the following sensitivity levels: **N/A**, **S1**, **S2**, and **S3**. **N/A** indicates an unknown risk level. The sensitivity levels of S1, S2, and S3 increase in sequence. You can customize the names and descriptions of the four sensitivity levels to classify the sensitive data detected in your data assets based on your business requirements. SDDP provides the following default descriptions for the S1, S2, and S3 levels:

   - **S1**: low risk.
   - **S2**: medium risk.
   - **S3**: high risk.
6. Click **Ok**.

The modification immediately takes effect after you submit it. Refresh the **Data Security > Sensitive Data Discovery and Protection > Sensitive Data Identification > Identification Rules** page. You can view the new sensitivity level on the **Level settings** tab.

22.1.10.2.5. Check data permissions

22.1.10.2.5.1. View permission statistics

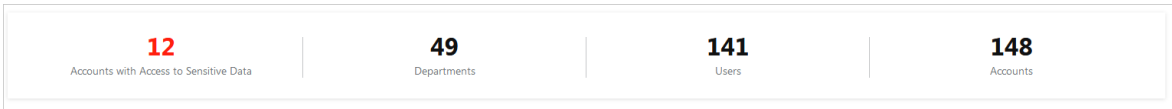
This topic describes how to view permission statistics.

Context

On the Permission Management page, you can view the overall permission distribution of Apsara Stack. You can also identify vulnerable accounts and users, and troubleshoot and handle security issues at your earliest opportunity.

Procedure

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Data Permissions > Permission Management**.
4. View the overall permission statistics.



   - **Accounts with Access to Sensitive Data**: the number of accounts that can access sensitive data.
   - **Departments**: the number of departments in Apsara Stack.
   - **Users**: the number of users in Apsara Stack.
   - **Accounts**: the number of accounts in Apsara Stack.
5. View the department-level permission statistics.

   You can view the statistics on the users, accounts, and anomalous events that are related to permissions for each department.

22.1.10.2.5.2. View the permissions of an account

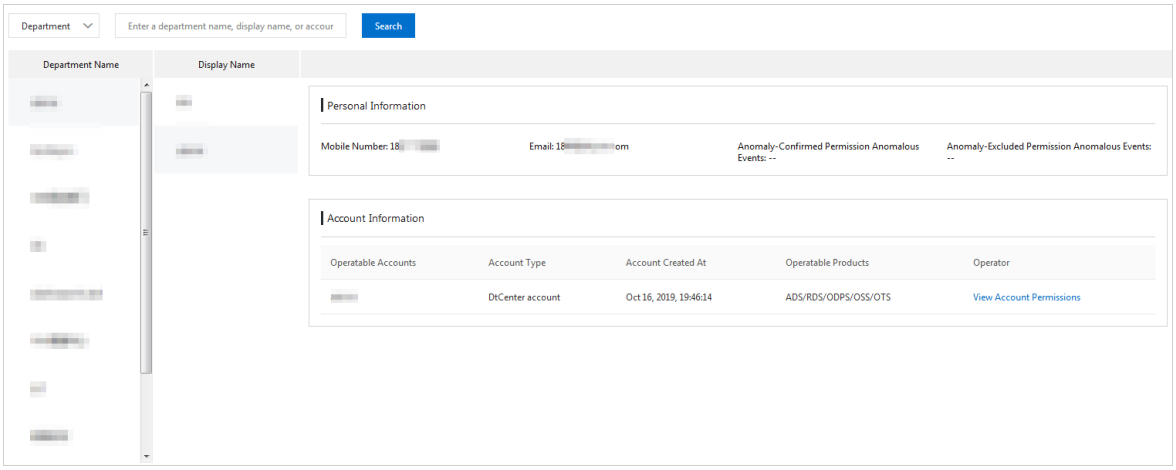
This topic describes how to view the permissions of an account.

Context

You can search for an account and view its information. This way, you can quickly find the owner of sensitive data.

Procedure

1.
2.
3. choose **Data Permissions > Permission Search**.



4. Search for a specific account.

   To search for an account, perform the following steps:

   i. Select **Department** or **Employee** from the drop-down list.

   ii. Enter a keyword, such as a department name or an account.

   iii. Click **Search**. You can view the search results in the **Display Name** column.

   > **Note** You can also click a department in the Department Name column. All accounts of the department are displayed in the Display Name column.

5. In the **Display Name** column, click the account whose details you want to view.
6. View information in the **Personal Information** and **Account Information** sections on the right.
   - **Personal Information**

     You can view the contact information about the account owner. You can also view the numbers of confirmed anomalous events that are related to permission access and excluded anomalous events that are related to permission access.
   - **Account Information**

     You can view the accounts that the owner can use and the details of each account. The details include the account type, time when the accounts are created, and Apsara Stack services that the accounts can access.

     You can click **View Account Permissions** in the Actions column of an account to view the resources, resource types, resource paths, and operation permissions.

22.1.10.2.6. Monitor data flows

22.1.10.2.6.1. View data flows in DataHub

This topic describes how to view data flows in DataHub.

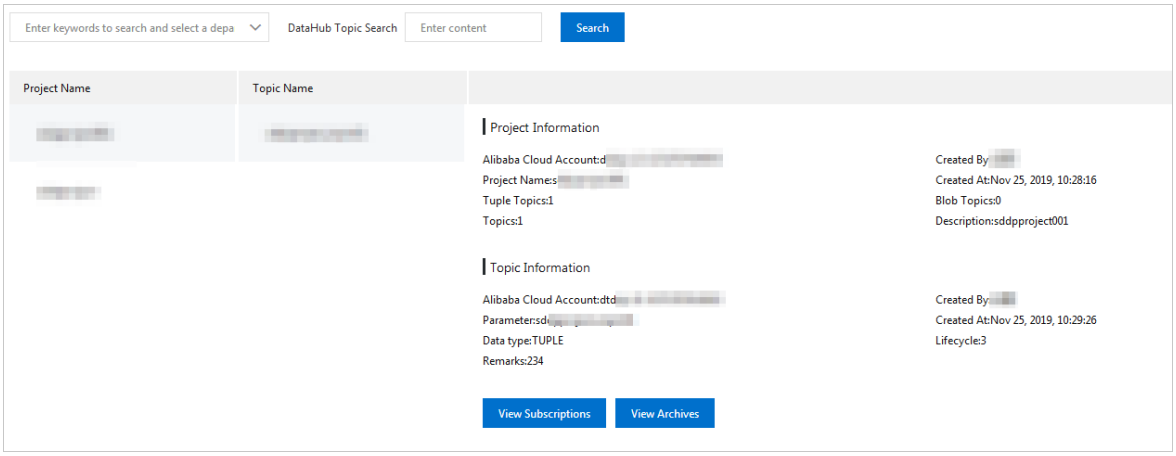
Context

DataHub is a platform that is designed to process streaming data. You can publish and subscribe to streaming data in DataHub. You can also distribute the data to other platforms. DataHub allows you to analyze streaming data and build applications based on the streaming data.

On the **DataHub** page, you can view the details of data flows in DataHub. The details include the relationships between DataHub projects and topics, and the relationships among topics, subscribed applications, and archive sources.

Procedure

1. Log on to Apsara Stack Security Center.
2. choose **Security > Data Security > Sensitive Data Discovery and Protection**.
3. In the left-side navigation pane of the Sensitive Data Discovery and Protection page, choose **Data Flow Monitoring > DataHub**.
4. In the search box, enter a keyword and select a department from the drop-down list. Enter a topic keyword in the **DataHub Topic Search** field and click **Search**.



> **Note**
>
> You can also click the required project in the **Project Name** column and click the required topic in the **Topic Name** column.

In the **Project Information** and **Topic Information** sections, you can view the information about the project and the topic.

- **Project Information**

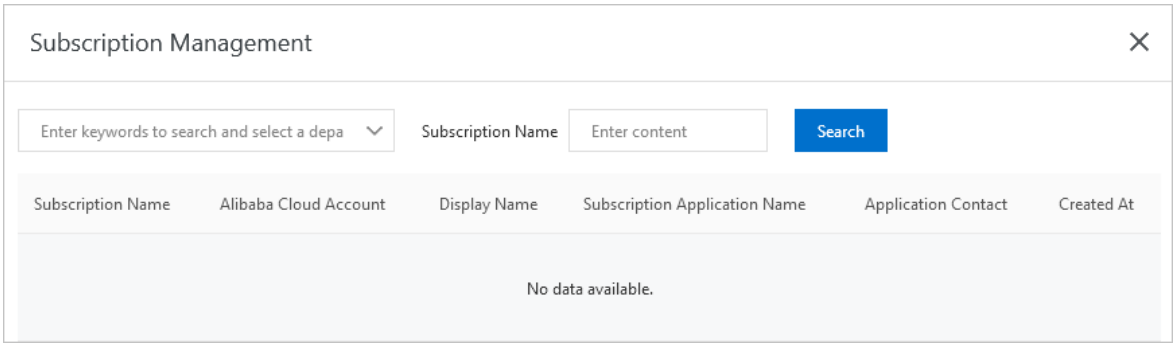
  Displays information such as the project name, Apsara Stack account, creator, creation time, and number of topics.

- **Topic Information**

  Displays information such as the topic name, Apsara Stack account, creator, creation time, and topic type.

5. Click **View Subscriptions** to view the subscription list.

   The subscription list provides information such as the subscription name, Apsara Stack account of the creator, display name, name of the subscribed application, and contact for the application.



   i. Enter a keyword and select a department from the drop-down list.
   ii. In the **Subscription Name** field, enter a keyword.
   iii. Click **Search** to search for the required DataHub topic.

6. Click **View Archives** to view the archive list.

The archive list provides information such as the name of the connected instance, Apsara Stack account of the creator, display name, source service, resource path, and risk level.

i. Enter a keyword and select a department from the drop-down list.
ii. In the **Instance Name** field, enter a keyword.
iii. Click **Search** to search for the required instance.

22.1.10.2.7. Sensitive data masking

22.1.10.2.7.1. Create a static masking task

This topic describes how to create a static masking task and run the task to mask sensitive data.

Procedure

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Static Desensitization**.
4. In the upper-right corner of the **Desensitization task configuration** tab, click **Add Desensitization Task**.
5. On the **Add Desensitization Task** page, configure parameters.
   i. In the **Basic Task Information** step, configure **Task Name** and **Task notes**. Then, click **Next**.
   ii. In the **Desensitization Source Configuration** step, specify the source of data that you want to mask and click **Next**.

   You can use Sensitive Data Discovery and Protection (SDDP) to create masking tasks for different data sources, including tables in ApsaraDB RDS, MaxCompute, and ApsaraDB for OceanBase, and objects in Object Storage Service (OSS). The following table describes the parameters used to create a masking task for each data source.

Data source	Parameter	Description
MaxCompute	**Types of data storage**	Select **RDS Table / DRDS Table / MaxCompute Table / PolarDB Table / ADS-Table / OceanBase Table**.
	Source Product	Select **MaxCompute**.
	Source Database/Project	Select the source database or project whose data you want to mask from the drop-down list.
	Source table name	Select the source table whose data you want to mask from the drop-down list.
	Source Partition	Enter the name of the source partition whose data you want to mask. You can configure partitions when you create a MaxCompute table. Partitions define different logical divisions of a table. When you query data, you can specify partitions to improve query efficiency. **Note** **Source Partition** is optional. If you leave this parameter unspecified, SDDP masks sensitive data in all partitions of the source table.

Data source	Parameter	Description
ApsaraDB RDS, PolarDB-X, and ApsaraDB for OceanBase	**Types of data storage**	Select **RDS Table / DRDS Table / MaxCompute Table / PolarDB Table / ADS-Table / OceanBase Table**.
	Source Product	Select **RDS**, **DRDS**, or **OceanBase**.
	Source Database/Project	Select the source database or project whose data you want to mask from the drop-down list.
	Source table name	Select the source table whose data you want to mask from the drop-down list.
	Sample SQL	Optional. Enter an SQL statement and specify the data that you want to mask.
AnalyticDB for MySQL and PolarDB	**Source Type**	Select **RDS Table / DRDS Table / MaxCompute Table / PolarDB Table / ADS-Table / OceanBase Table**.
	Source Product	Select **ADS** or **PolarDB**.
	Source Database/Project	Select the source project whose data you want to mask from the drop-down list.
	Source table name	Select the source table whose data you want to mask from the drop-down list.
OSS	**Types of data storage**	Select **OSS files**.
	File source	Upload a file from your computer or select a bucket. Valid values: ■ **Uploaded Local File**: If you select this option, click **Select a local file** and select a source file from your computer. ■ **OSS Bucket**: If you select this option, select the OSS bucket to which the source object belongs.
	Source file description	Enter an informative description for the source file to help identify the task. **Note** This parameter is required only when you set **File source** to **Uploaded Local File**.
	OSS Bucket where the source file is located	Select the OSS bucket to which the source object belongs. **Note** This parameter is required only when you set **File source** to **OSS Bucket**.

Data source	Parameter	Description
	Source file names	Optional. Enter the name of the source object. **Note** This parameter is required only when you set **File source** to **OSS Bucket**. If you want to use wildcards to specify objects, turn on **Open the pass**. After you turn on Open the pass, you can use asterisks (*) as wildcards to specify multiple OSS objects at a time. However, you can use asterisks only in object names. For example, enter test*.xls. After you specify an object name by using an asterisk, SDDP masks the data of the matched objects. Make sure that these objects use the same column structure.
	Separator selection	Optional. Select a column delimiter based on the format of the object that you specify. This parameter is required for objects in the CSV or TXT format. Valid values: ■ **Semicolon ";" (MacOS/Linux default)** ■ **Comma "," (Windows default)**
	Table contains header rows	Optional. Specify whether the data to be masked contains header rows.

iii. In the **Desensitization algorithm** step, specify the algorithm to mask data and click **Next**.

In this step, you must specify the algorithm type, select an algorithm, and turn on the masking switch for the source field of data that you want to mask.

iv. (Optional)In the **Data Watermark** step, turn on **Open data watermark**. Specify the following parameters: Please select the field to embed the watermark, Please select a watermark algorithm, and Please enter watermark information. Then, click **Next**.

The Please select a watermark algorithm parameter has the following values:

- **Space Algorithm**: If you want to add watermarks for fields of a string type, select this value.
- **Modify the least significant bit algorithm**: If you want to add watermarks for fields of a numeric type, select this value.

v. In the **Destination Location Configuration** step, specify the destination table to store the data after masking, test and make sure that you have write permissions on the destination table, and then click **Next**. The parameters for the destination table include **Types of data storage** and **Target**.

vi. In the **Confirm Process Logic** step, configure the processing logic of the task.

Parameter	Description
How the task is triggered	Select a method to run the masking task. Valid values: ■ **Manual Only**: You must manually run the masking task on the Static Desensitization page. ■ **Scheduled Only**: The masking task is automatically run at a specific point in time on an hourly, daily, or monthly basis. ■ **Manual + Scheduled**: You can manually run the masking task or enable automatic running of the masking task at a specific point in time on an hourly, daily, or monthly basis.
Turn on incremental desensitization	Optional. Enable incremental masking based on your business requirements. If you turn on this switch, SDDP masks only the data that is added after the last masking task is completed. You must specify a field whose value is increased over time as the incremental identifier. For example, you can specify the creation time field or the auto-increment ID field as the incremental identifier. **Note** SDDP supports incremental masking only for data in ApsaraDB RDS.
Shard field	Optional. Select a field based on which SDDP divides the source data into multiple shards and concurrently masks the data in these shards. In this case, data masking is more efficient. You can specify one or more shard fields based on your business requirements. **Note** ■ SDDP supports incremental masking only for data in ApsaraDB RDS. We recommend that you use a primary key or a field on which a unique index is created as the shard field. ■ If you leave this parameter unspecified, a primary key is used as the shard field. SDDP divides the source data based on the primary key and masks the data. If the source data does not have a primary key, you must specify a shard field. Otherwise, the masking task fails. ■ If you specify excessive shard fields, query performance and data accuracy may deteriorate. Proceed with caution.
Table name conflict resolution	Select a method to handle a table name conflict. Valid values: ■ **Delete the target table and create a new table with the same name** ■ **Attach data to the target table**: This method is recommended.
Row Conflict Resolution	Select a method to handle a row conflict. Valid values: ■ **Keep conflicting rows in the target table and discard the new data**: This method is recommended. ■ **Delete conflicting rows in the target table and insert the new data**

vii. Click **Submit**.

After you create the masking task, you can view the task in the list of masking tasks on the **Desensitization task configuration** tab.

6. In the list of masking tasks, turn on the switch and run the masking task.
7. On the **Task Execution Status** tab, view **Execution Progress** and **Status** of the masking task.

22.1.10.2.7.2. View dynamic data masking tasks

Sensitive Data Discovery and Protection (SDDP) provides the dynamic data masking feature. You can call the ExecDatamask operation to dynamically mask sensitive data.

Context

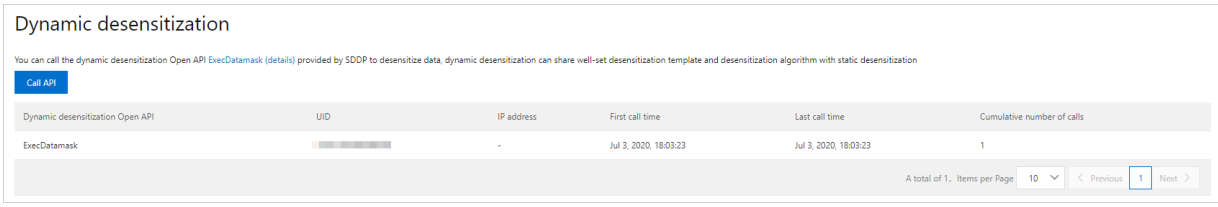
When you call this operation, you must specify the ID of the data masking template to use. Static data masking and dynamic data masking can use the same template. To obtain the template ID, perform the following operations: Log on to Apsara Stack Security Center. In the top navigation bar, choose **Security > Data Security > Sensitive Data Discovery and Protection**. In the left-side navigation pane, choose **Sensitive Data Desensitization** > Desensitization Template. You can also create custom data masking templates. For more information, see Create a data masking template.

Limits

Before you can call the ExecDatamask operation to dynamically mask sensitive data, make sure that the size of the sensitive data is less than 2 MB. The `Data` parameter specifies the size.

View the call history of the ExecDatamask operation

Log on to Apsara Stack Security Center. In the top navigation bar, choose **Security > Data Security > Sensitive Data Discovery and Protection**. In the left-side navigation pane, choose **Sensitive Data Desensitization** > Dynamic desensitization. On the page that appears, you can view the call history of the ExecDatamask operation. Each record includes the name of the operation, the UID of the Apsara Stack tenant account or RAM user that called the operation, the IP address from which the call is initiated, the points in time at which the operation was first and last called, and the total number of calls.



Note Only one record is generated for calls that are initiated by the same Apsara Stack tenant account or RAM user from the same IP address. In this case, the cumulative number of calls is recorded.

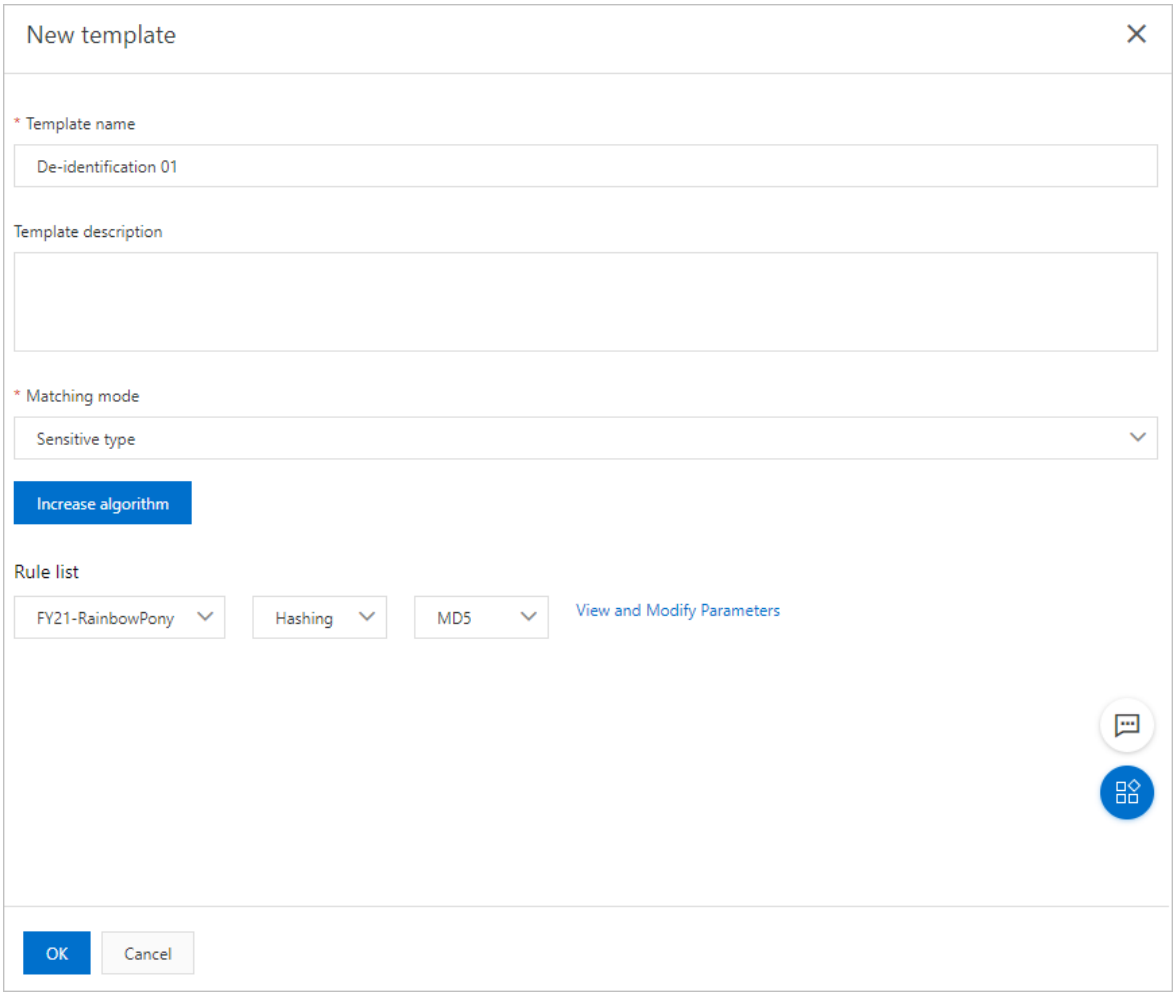
22.1.10.2.7.3. Create a data masking template

Sensitive Data Discovery and Protection (SDDP) allows you to create data masking templates. You can create a data masking template and add data masking algorithms that are frequently used in the same scenario to the template. This avoids repeated configuration of data masking algorithms and makes sensitive data processing more efficient. This topic describes how to create and manage data masking templates.

Create a data masking template

You can create an unlimited number of data masking templates.

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Desensitization Template**.
4. On the **Desensitization Template** page, click **New template**.
5. In the **New template** panel, configure the following parameters.



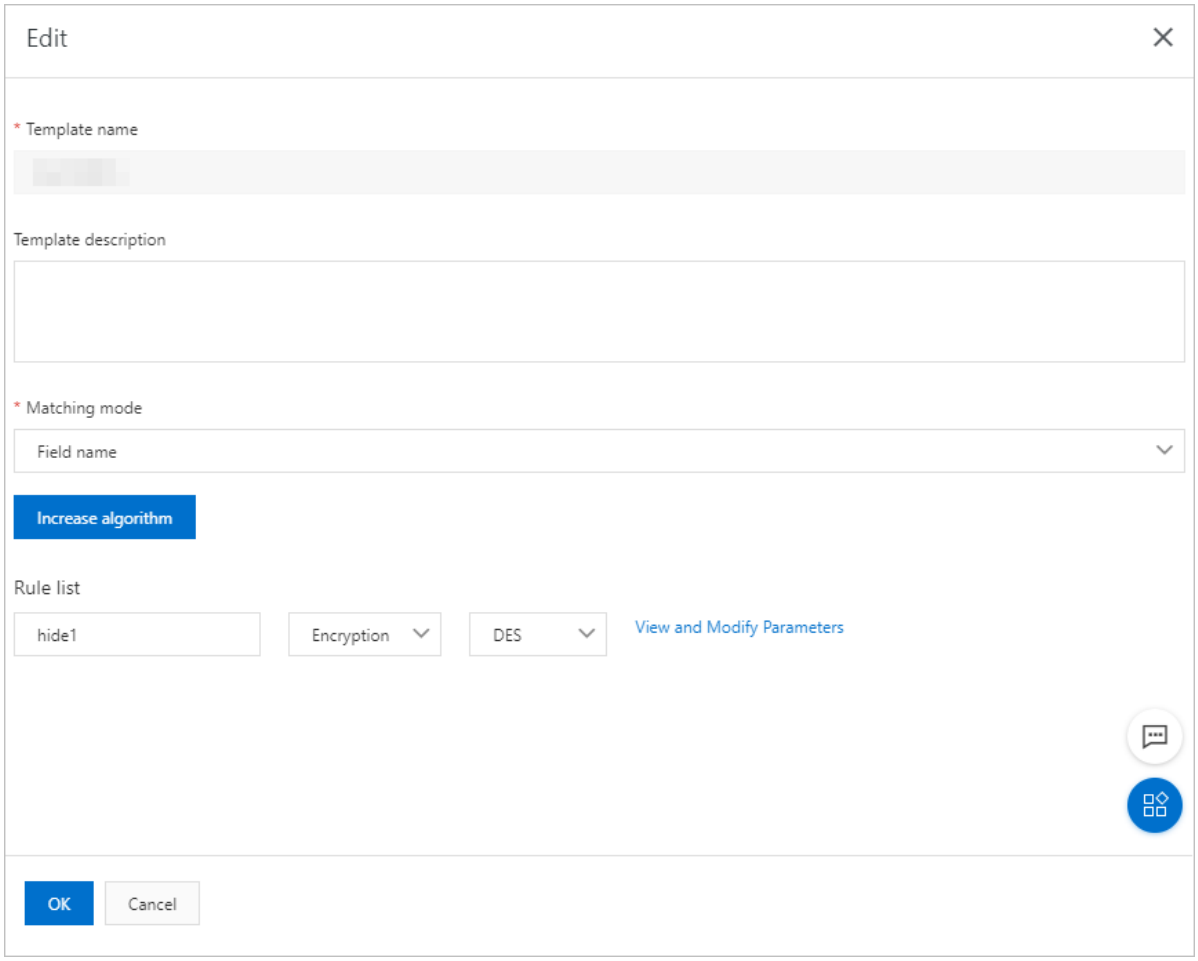
Parameter	Description
Template name	The name of the data masking template.
Template description	The description of the data masking template. You can enter information such as the scenario to which the template is applied.
Matching mode	The mode in which the data masking template handles its matched sensitive data. Valid values: ○ **Sensitive type**: If you select this option, select the types of sensitive data that you want to mask and the data masking algorithm for each type of sensitive data. The types of sensitive data include vehicle identification numbers and unified social credit codes. ○ **Field name**: If you select this option, specify the fields that you want to mask and the data masking algorithm for each field.

Parameter	Description
Rule list	The rules that are used to mask sensitive data. To configure a rule, select a sensitive data type or enter a field that you want to mask and specify a data masking algorithm. SDDP supports the following data masking algorithms: ○ Hashing ○ Redaction ○ Substitution ○ Rounding ○ Encryption ○ Shuffling ○ Data decryption For more information, see Configure data masking algorithms. You can configure multiple rules in a template. To configure more rules, click **Increase algorithm**.

Manage data masking templates

- **Edit a data masking template**

  To edit a data masking template, find the template on the **Desensitization Template** page and click **Edit** in the Actions column. In the **Edit** panel, modify the description or rules of the data masking template.



- **Delete a data masking template**

To delete a data masking template that you no longer use, find the template on the **Desensitization Template** page and click **Delete** in the Actions column.

> **Note** If you delete a data masking template, it cannot be restored. Proceed with caution.

22.1.10.2.7.4. Configure data masking algorithms

This topic describes how to configure data masking algorithms.

Context

The following table describes the data masking algorithms that are supported by Sensitive Data Discovery and Protection (SDDP).

Category	Description	Algorithm	Input	Suitable sensitive data and scenario
Hashing	This type of algorithm is irreversible. This type of algorithm is suitable for password masking and the scenarios in which you must check whether data is sensitive by comparison. You can use common hashing algorithms and specify a salt value.	MD5	Salt value	• Sensitive data: keys • Scenario: data storage
		SHA-1	Salt value	
		SHA-256	Salt value	
		HMAC	Salt value	

Redaction by using asterisks (*) or number signs (#)	This type of algorithm is irreversible.
	This type of algorithm is suitable for the scenarios in which you need to show sensitive data on a GUI or share sensitive data.
	This type of algorithm masks specific content in sensitive data by using asterisks (*) or number signs (#).
		Keeps the first N characters and the last M characters	Values of N and M	• Sensitive data: sensitive personal information
				• Scenarios:
				  ◦ Data usage
				  ◦ Data sharing

		Keeps characters from the Xth position to the Yth position	Values of X and Y	
		Masks the first N characters and the last M characters	Values of N and M	
		Masks characters from the Xth position to the Yth position	Values of X and Y	
		Masks characters that precede a special character when the special character appears for the first time	At sign (@), ampersand (&), or period (.)	

Category	Description	Algorithm	Input	Suitable sensitive data and scenario
		Masks characters that follow a special character when the special character appears for the first time	At sign (@), ampersand (&), or period (.)	
Substitution (customization supported)	Some of the algorithms are reversible. This type of algorithm can be used to mask fields in fixed formats. For example, you can use the algorithms to mask ID card numbers. This type of algorithm substitutes the entire value or a part of the value of a field with a mapped value by using a mapping table. In this case, data masking is reversible. This type of algorithm also substitutes the entire value or a part of the value of a field randomly based on a random interval. In this case, data masking is irreversible. SDDP provides multiple built-in mapping tables and allows you to customize substitution algorithms.	Substitutes specific content in ID card numbers with mapped values	Mapping table for randomly substituting IDs of administrative regions	• Sensitive data: ◦ Sensitive personal information ◦ Sensitive information of enterprises ◦ Sensitive information of devices • Scenarios: ◦ Data storage ◦ Data sharing
		Randomly substitutes specific content in ID card numbers	Mapping table for randomly substituting IDs of administrative regions	
		Randomly substitutes specific content in IDs of military officer cards	Mapping table for randomly substituting type codes	
		Randomly substitutes specified content in passport numbers	Mapping table for randomly substituting purpose fields	
		Randomly substitutes specific content in permit numbers of Exit-Entry Permit for Travelling to and from Hong Kong and Macao	Mapping table for randomly substituting purpose fields	
		Randomly substitutes specific content in bank card numbers	Mapping table for randomly substituting bank identification numbers (BINs)	
		Randomly substitutes specific content in landline telephone numbers	Mapping table for randomly substituting IDs of administrative regions	
		Randomly substitutes specific content in mobile phone numbers	Mapping table for randomly substituting mobile network codes	
		Randomly substitutes specific content in unified social credit codes	Mapping table for randomly substituting IDs of registration authorities, mapping table for randomly substituting type codes, and mapping table for randomly substituting IDs of administrative regions	

Category	Description	Algorithm	Input	Suitable sensitive data and scenario
		Substitutes specific content in general tables with mapped values	Mapping table for substituting uppercase letters, mapping table for substituting lowercase letters, mapping table for substituting digits, and mapping table for substituting special characters	
		Randomly substitutes specific content in general tables	Mapping table for randomly substituting uppercase letters, mapping table for randomly substituting lowercase letters, mapping table for randomly substituting digits, and mapping table for randomly substituting special characters	

Rounding	Some of the algorithms are reversible.
	This type of algorithm can be used to analyze and collect statistics on sensitive datasets.
	SDDP provides two types of rounding algorithms. One algorithm rounds numbers and dates, which is irreversible. The other algorithm bit-shifts text, which is reversible.
		Rounds down a number to the Nth digit before the decimal point	N	• Sensitive data: general sensitive information
				• Scenarios:
				  ◦ Data storage
				  ◦ Data usage

		Rounds dates	Date rounding level	
		Shifts characters	Number of places by which specific bits are moved and shift direction (left or right)	

Encryption	This type of algorithm is reversible.
	This type of algorithm can be used to encrypt sensitive fields that need to be retrieved after encryption.
	Common symmetrical encryption algorithms are supported.
		Data Encryption Standard (DES) algorithm	Encryption key	• Sensitive data:
				  ◦ Sensitive personal information
				  ◦ Sensitive information of enterprises
				• Scenario: data storage

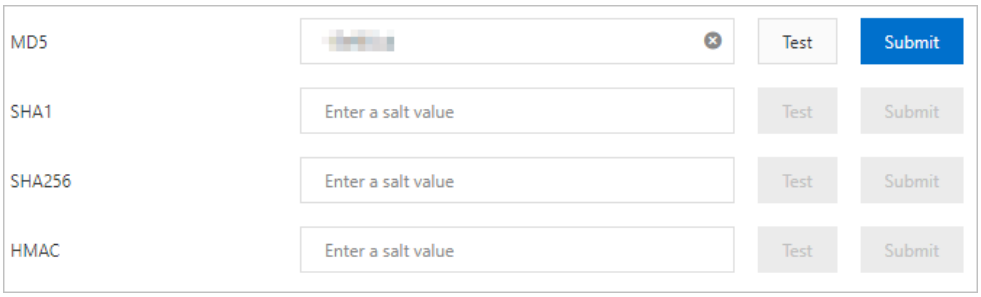
		Triple Data Encryption Standard (3DES) algorithm	Encryption key	
		Advanced Encryption Standard (AES) algorithm	Encryption key	

Category	Description	Algorithm	Input	Suitable sensitive data and scenario
Shuffling	This type of algorithm is irreversible. This type of algorithm can be used to mask structured data columns. This type of algorithm extracts values of a field in a specified range from the source table and rearranges the values in a specific column. Alternatively, this type of algorithm randomly selects values from a specific column within the value range and rearranges the selected values. This way, the values are mixed up and masked.	Randomly shuffles data	Rearranged values or randomly selected values	• Sensitive data: ◦ Sensitive information of devices ◦ Location-sensitive information • Scenario: data storage

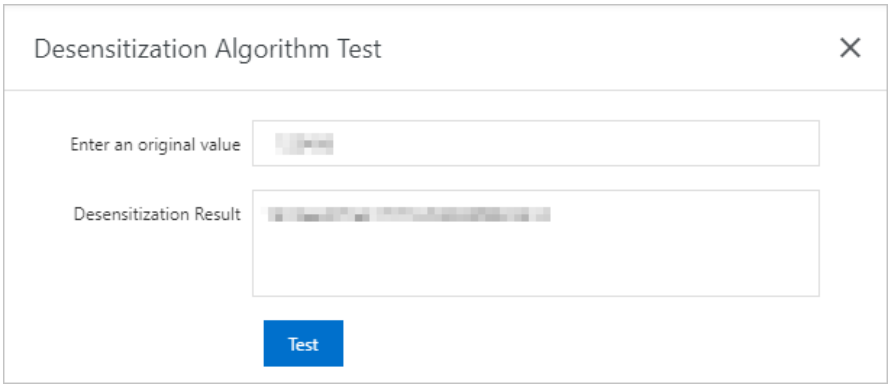
Hashing

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Desensitization algorithm**. On the **Desensitization algorithm** page, click the **Hashing** tab.
4. Specify a salt value for each algorithm.

   > **Note** In cryptography, you can insert a specific string to a fixed position of a password to generate a hash value that is different from that of the original password. This process is called salting.
   >
   > A salt value is the specific string that you insert.

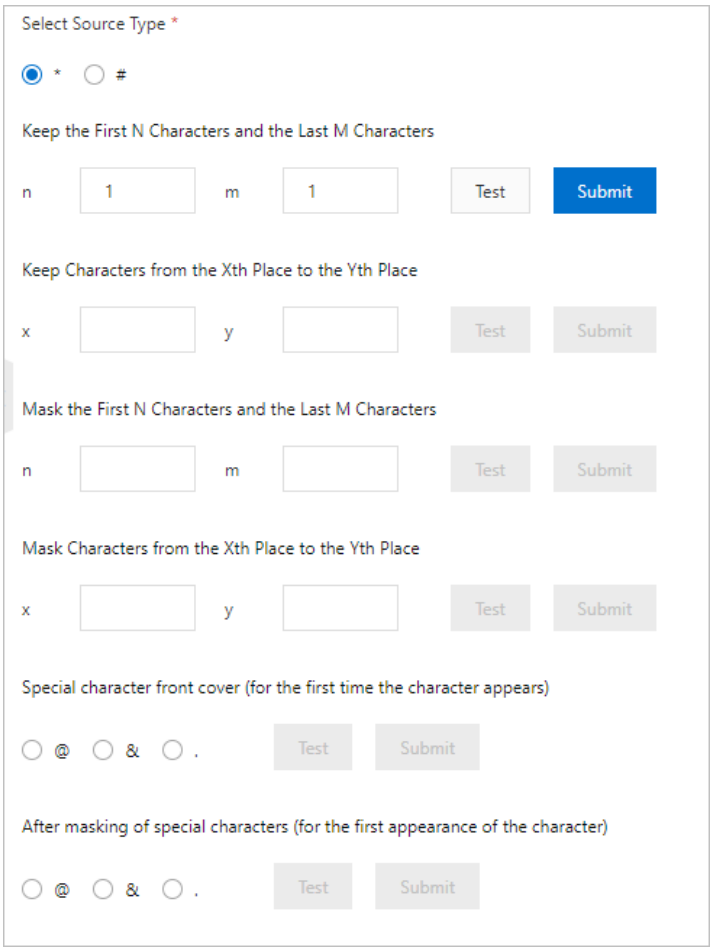
5. In the **Desensitization Algorithm Test** panel, enter the original value and click **Test** to check whether the algorithm works.



6. After the test is complete, click **Submit**.

Redaction

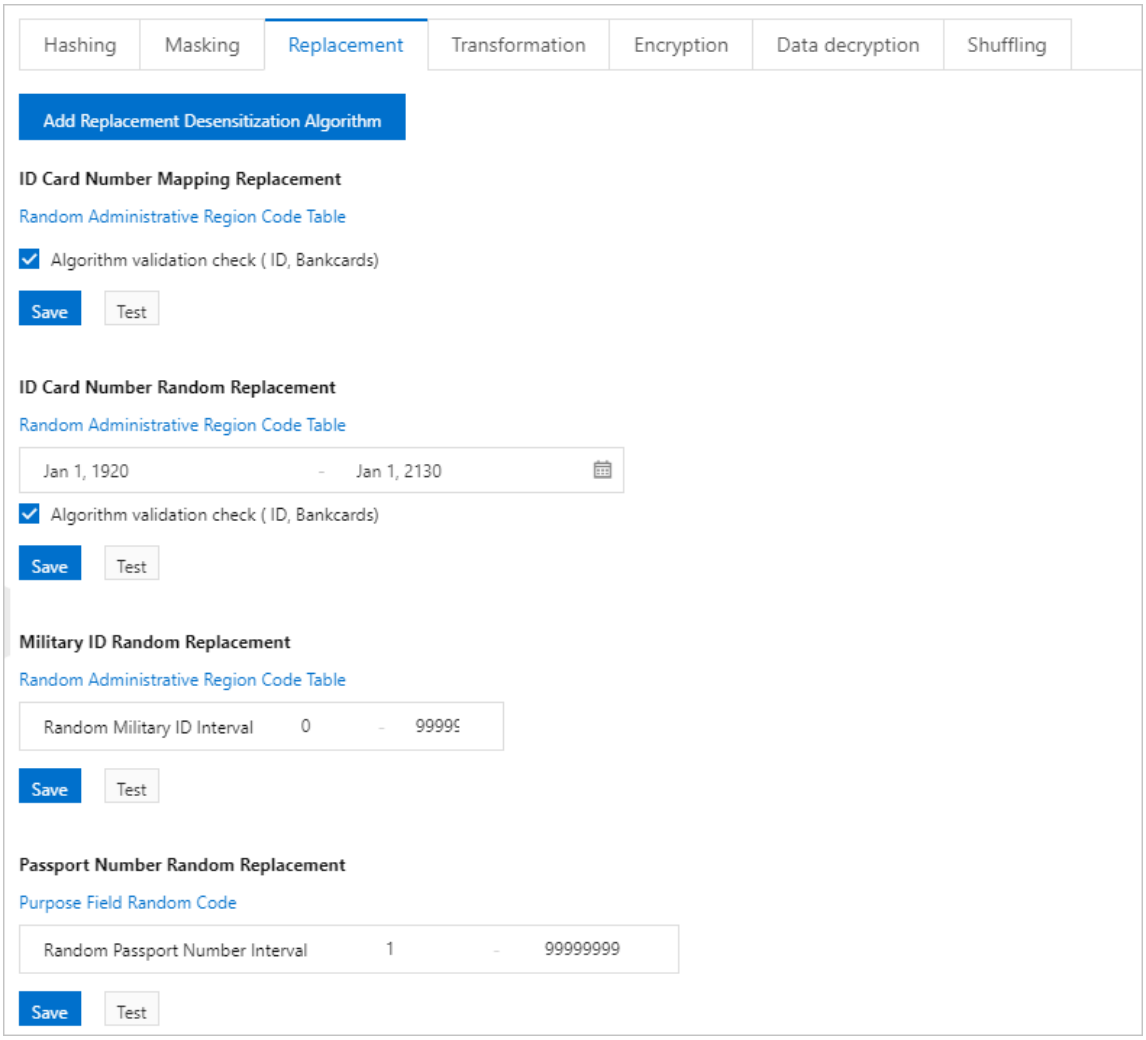
1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Desensitization algorithm**. On the **Desensitization algorithm** page, click the **Masking** tab.
4. Configure the parameters.



5. In the **Desensitization Algorithm Test** panel, enter the original value and click **Test** to check whether the algorithm works.
6. After the test is complete, click **Submit**.

Substitution

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Desensitization algorithm**. On the **Desensitization algorithm** page, click the **Replacement** tab.
4. Configure the parameters.



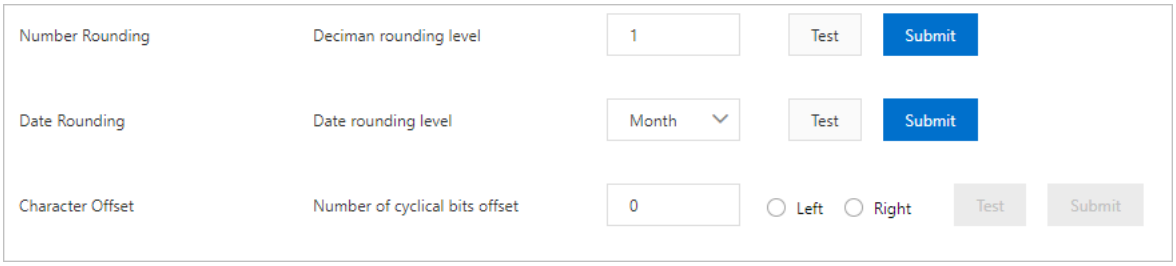
> **Note** By default, SDDP provides multiple common substitution algorithms, such as ID Card Number Mapping Replacement and Telephone Number Random Replacement.

   - If you want to customize a mapping table, click the required mapping table, replace the original content with your own mapping table, and then click **Save**.
   - If you want to customize an algorithm, click **Add Replacement Desensitization Algorithm** and specify the interval and mapping table.
5. In the **Desensitization Algorithm Test** panel, enter the original value and click **Test** to check whether the algorithm works.
6. After the configuration is complete, click **Save**.

Rounding

1.

2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Data Desensitization > Desensitization algorithm**. On the **Desensitization algorithm** page, click the **Transformation** tab.
4. Configure the parameters.



5. In the **Desensitization Algorithm Test** panel, enter the original value and click **Test** to check whether the algorithm works.
6. After the test is complete, click **Submit**.

Encryption

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Desensitization algorithm**. On the **Desensitization algorithm** page, click the **Encryption** tab.
4. Specify a key for an algorithm.
5. In the **Desensitization Algorithm Test** panel, enter the original value and click **Test** to check whether the algorithm works.
6. After the test is complete, click **Submit**.

Shuffling

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Desensitization algorithm**. On the **Desensitization algorithm** page, click the **Shuffling** tab.
4. Select a shuffling method.

5. Click **Submit**.

22.1.10.2.7.5. Extract watermarks

You can add watermarks when you create a data masking task. If data is leaked after it is distributed, you can use watermarks to trace the data flow process. This way, the impacts of data leaks are reduced. Sensitive Data Discovery and Protection (SDDP) extracts and identifies watermarks from the leaked data to trace the data flow process and identify the organization or user that is responsible for the data leaks. This topic describes how to extract watermarks.

Procedure

1.
2. Choose **Data Security > Sensitive Data Discovery and Protection**.
3. choose **Sensitive Data Desensitization > Extract watermarks**.

4. On the **Extract watermark** page, configure the **Source Product**, **Source database/project name**, and **Source table name** parameters. Then, click **Extract watermark**.

Parameter	Description
Source Product	The name of the data source to which the table containing watermarks belongs.
Source database/project name	The name of the database or project to which the table containing watermarks belongs.
Source table name	The name of the table that contains watermarks.

The extracted watermarks appear in the field below this parameter. If you want to copy the information, click **Copy Result**.

22.1.11. Apsara Stack Security configurations

22.1.11.1. Rules

22.1.11.1.1. Create an IPS rule for traffic monitoring

This topic describes how to create an intrusion prevention system (IPS) rule for traffic monitoring in Cloud Firewall. Cloud Firewall has built-in IPS rules. This topic describes how to create custom IPS rules based on your business requirements and network environment.

Procedure

1.
2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **Rules**.
4. On the Rules page, click the **Cloud Firewall IPS Rules** tab.
5. Click **Create Rule**.
6. In the **Create Rule** panel, configure the following parameters.

Parameter	Description
Rule Name	The name of the IPS rule. We recommend that you enter a name that can help you identify and manage the IPS rule in an efficient manner.
Rules Engine	The rules engine that you want to use. Valid values: **Basic Policies** and **Virtual Patches**.
Attack Type	The type of the attack that you want to detect by using the IPS rule
Severity	The severity of the attack. Valid values: **Low**, **Medium**, and **High**.
CVE	The Common Vulnerabilities and Exposures (CVE) ID of the vulnerability that you want to add to the rule. **Note** CVE provides a list of public security vulnerabilities. CVE IDs are allocated by a CVE Numbering Authority (CNA).

Parameter	Description
Application	The name of the attacked application.
Rule Mode	The mode of the IPS rule. Valid values: **Packet** and **Traffic**.
Direction	The direction of traffic that you want to monitor by using the IPS rule. Valid values: **Inbound and Outbound**, **Inbound**, and **Outbound**.
Rule Content	The content of the IPS rule. You must use the Snort syntax to specify the content. **Note** To prevent a negative impact on your business, make sure that the content you enter for the IPS rule is valid.
Rule Description	The description of the IPS rule. We recommend that you enter information that can help you identify the IPS rule, such as the purpose or impact of the rule.
Description	The additional description of the IPS rule. We recommend that you enter information that can help you identify the IPS rule, such as the purpose or impact of the rule.

7. Click **OK**.

22.1.11.1.2. Create an IDS rule for traffic monitoring

This topic describes how to create an intrusion detection system (IDS) rule for traffic monitoring.

Procedure

1.
2.
3. click **Rules**.
4. On the Rules page, click the **Traffic Monitoring IDS Rules** tab.
5. Click **Create Rule**.
6. In the **Create Rule** panel, configure parameters.

Parameter	Description
Rule Name	The name of the IDS rule. We recommend that you enter a name that can help you identify and manage the IDS rule in an efficient manner.
Rules Engine	The rules engine that you want to use. Valid values: **Basic Policies** and **Virtual Patches**.
Attack Type	The type of the attack that you want to detect by using the IDS rule
Severity	The severity of the attack. Valid values: **Low**, **Medium**, and **High**.
CVE	The Common Vulnerabilities and Exposures (CVE) ID of the vulnerability that you want to add to the rule. **Note** CVE provides a list of public security vulnerabilities. CVE IDs are allocated by a CVE Numbering Authority (CNA).

Parameter	Description
Application	The name of the attacked application.
Rule Mode	The mode of the IDS rule. Valid values: **Packet** and **Traffic**.
Direction	The direction of traffic that you want to monitor by using the IDS rule. Valid values: **Inbound and Outbound**, **Inbound**, and **Outbound**.
Rule Content	The content of the IDS rule. You must use the Snort syntax to specify the content. **Note** To prevent a negative impact on your business, make sure that the content you enter for the IDS rule is valid.
Rule Description	The description of the IDS rule. We recommend that you enter information that can help you identify the IDS rule, such as the purpose or impact of the rule.
Description	The additional description of the IDS rule. We recommend that you enter information that can help you identify the IDS rule, such as the purpose or impact of the rule.

7. Click **OK**.

22.1.11.1.3. Manage IDS rules for traffic monitoring

This topic describes how to view, enable, and disable intrusion detection system (IDS) rules for traffic monitoring.

Context

On the **Traffic Monitoring IDS Rules** tab, you can view the built-in and custom IDS rules. You can also enable or disable the rules based on your business requirements.

Procedure

1.
2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **Rules**.
4. On the Rules page, click the **Traffic Monitoring IDS Rules** tab.
5. Manage IDS rules for traffic monitoring.

   In the list of IDS rules, you can view rule details, enable rules, and disable rules.

   - View rule details

     Find the rule whose details you want to view and click **Details** in the **Actions** column to view the rule details.

   - Enable a rule

     Find the rule that you want to enable and turn on the switch in the **Enable or not** column to change the status of the rule from **Disable** to **Enable**.

   - Disable a rule

     If a rule is not suitable for your business, you can disable the rule.

     Find the rule that you want to disable and turn off the switch in the **Enable or not** column to change the status of the rule from **Enable** to **Disable**.

22.1.11.1.4. Specify custom thresholds for DDoS traffic scrubbing policies and traffic redirection

This topic describes how to specify custom thresholds for DDoS traffic scrubbing policies and traffic redirection. Default thresholds are provided. If you want to specify custom thresholds, perform the following steps:

Procedure

1.
2.
3. click **Rules**.
4. Specify a custom threshold for a DDoS traffic scrubbing policy.
   i. Click the **Anti-DDoS Service Rules** tab.
   ii. Then, click the **Scrubbing Policy** tab.
   iii. Find the policy for which you want to specify a custom threshold and click **Modify Threshold** in the **Actions** column.
   iv. In the **Modify Threshold** dialog box, enter a threshold value.
   v. Click **OK**.
5. Specify a custom threshold for traffic redirection.
   i. Click the **Anti-DDoS Service Rules** tab.
   ii. Then, click the **Scrubbing Policy** tab.
   iii. Find a rule for traffic redirection whose threshold you want to modify and click **Modify Threshold** in the **Actions** column.
   iv. In the **Modify Threshold** dialog box, enter a threshold value.
   v. Click **OK**.

22.1.11.1.5. View Server Guard rules

This topic describes how to view the operations of Server Guard rules. You can view vulnerabilities, baselines, and host exceptions.

Procedure

1.
2.
3. click **Rules**.
4. On the **Rules** page, click the **Server Guard Rules** tab.
5. In the overview section, you can view the total number of **vulnerability libraries**, number of **baselines**, number of **host exceptions**, and the available **engines**.
6. View the vulnerability list.
   i. Click the **Vulnerabilities** tab.
   ii. In the overview section, you can view the total number of **Linux vulnerabilities**, total number of **Windows vulnerabilities**, total number of **Web-CMS vulnerabilities**, and total number of **urgent vulnerabilities**.

iii. Specify search conditions to search for the vulnerabilities that you want to view.

> **Note** If you want to view all vulnerabilities, skip this step.

In the vulnerability list, you can view the **vulnerability name**, **CVE ID**, **vulnerability type**, **operating system**, **update time**, and **status**.

7. View the baseline list.
   i. Click the **Baselines** tab.
   ii. In the overview section, you can view the numbers of baseline types and the number of check items.
   iii. Specify search conditions to view the baselines that meet the search conditions

> **Note** If you want to view all baselines, skip this step.

In the baseline list, you can view the **baseline type**, **check item category**, **check item name**, **risk level**, **update time**, and **status**.

8. View the host exception list.
   i. Click the **Server Exceptions** tab.
   ii. In the overview section, you can view the number of **rule alert subcategories**, the number of webshells, and the number of malicious viruses.
   iii. Specify search conditions to search for the host exceptions that you want to view.

> **Note** If you want to view all exceptions, skip this step.

In the host exception list, you can view the **subcategory name**, **rule category**, **risk level**, **update time**, **source**, and **status**.

22.1.11.2. Threat intelligence

22.1.11.2.1. View the Overview page

This topic describes how to view the overall situation and statistics about threats to your assets over the last 30 days on the Overview page.

Prerequisites

The **service configuration** feature is enabled. For more information, see Enable the service configuration feature.

Procedure

1.
2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **Overview**.
4. On the **Overview** page, view the statistics and threats that are detected on Apsara Stack services by the threat intelligence module.

   On the **Overview** page, you can perform the following operations:

   - View **Total malicious metric intelligence**

     In the **Total malicious metric intelligence** section of the **Overview** page, view the information about the detected threats on Apsara Stack services. The information includes the number of malicious IP addresses, the number of malicious domain names, and the number of malicious URLs.

   - View **Threat trends in the last 30 days**

- Search for an IP address to check whether the IP address is malicious.

  In the upper-right corner of the search box, enter the IP address that you want to check and click the icon. Then, you are redirected to the details page of the IP address. For more information, see Search for an IP address.

- View **Top 10 active IP malicious addresses**

  In the **Top 10 active IP malicious addresses** section of the **Overview** page, view the information about the top 10 malicious IP addresses. The information includes **IP address**, **First malicious observation**, **Last malicious observation**, and **Malicious label**. Find the malicious IP address whose details you want to view and click **View** in the Actions column. Then, you are redirected to the details page of the IP address. For more information, see Search for an IP address.

22.1.11.2.2. Search for and view the information about a suspicious or malicious IP address

The threat intelligence module allows you to search for threat intelligence. This module helps you handle suspicious or malicious IP addresses at the earliest opportunity.

Prerequisites

The **service configuration** feature is enabled. For more information, see Enable the service configuration feature.

Procedure

1.
2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **IP Address Search**.
4. In the search box on the **Search** page, enter the suspicious or malicious IP address that you want to query and click the icon.
5. On the details page of the IP address, view the values of **Threat Level**, **Basic Information**, **Threat Overview**, **IP Details**, and **Analysis of Attack Risk Degree** of the IP address.

   You can view the following information on the details page of the IP address:

   - **Threat Level**: View the threat level of the suspicious or malicious IP address.

     The threat intelligence module classifies IP addresses into the following threat levels: normal, suspicious, and high-risk. If the IP address is identified as high-risk, we recommend that you handle the IP address at the earliest opportunity.

   - **Basic Information**: View the basic information about the suspicious or malicious IP address.

     The basic information includes the server in a data center, Abstract Syntax Notation One (ASN.1), the country and city to which the IP address belongs, and the number of domain names for the IP address.

   - View the statistics about the suspicious or malicious IP address.

     You can view **Threat Overview**, **IP Details**, and **Threat Details** of the IP address.

     - The **Threat Overview** tab displays **Top5 Target Preference**, **Number of Attacks (Classified by Threat)**, and **Analysis of Attack Risk Degree** of the IP address.
     - The **IP Details** tab displays **WHOIS Information** and **IP Reverse Check Information** of the IP address.
     - The **Threat Details** tab displays the **threat tags** of the IP address. The tags include the intelligence source, the time when the IP address is first detected, the time when the IP address is last active, and the threat tag.

22.1.11.2.3. Enable the service configuration feature

The threat intelligence module integrates threat monitoring and big data analysis. You can use these features to obtain information about the latest developments in the threat intelligence field. After you enable the service configuration feature, the system starts to monitor and collect threat intelligence. This topic describes how to enable the service configuration feature.

Procedure

1.
2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **Service Configurations**.
4. On the **Consumer product configuration** page, view the data types and descriptions on the Situation awareness and Web application firewall tabs.
5. Click the tab in which you want to enable threat monitoring and turn on **Activation status**.

   After you turn on **Activation status**, the system starts to monitor and collect threat intelligence for the data types listed on the tab.

What's next

After you enable the service configuration feature, you can view the overall situation and statistics of threats within the last 30 days on the **Overview** tab. For more information, see View the Overview page.

22.1.11.3. Alert settings

22.1.11.3.1. Configure alert contacts

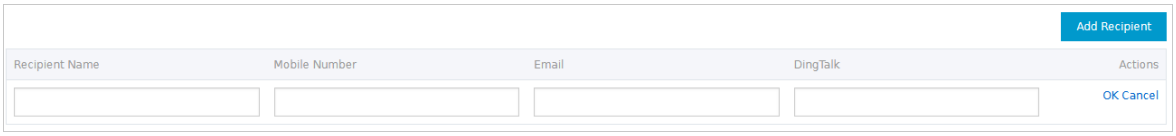
This topic describes how to configure and manage alert contacts.

Context

Apsara Stack Security sends alert notifications to alert contacts by text message or email. If the detected information matches an alert rule, Apsara Stack Security sends an alert notification to the alert contacts.

Procedure

1.
2.
3. click **Alert Settings**.
4. On the **Alert Settings** page, click the **Alert Recipient** tab.
5. Click **Add Recipient**.
6. Enter the contact information and click **OK**.



7. Manage alert contacts.

   In the contact list, find a contact whose information you want to modify and click **Edit** in the Actions column.

22.1.11.3.2. Configure alert notifications

This topic describes how to configure alert notification methods for security events on tenants or platforms.

Context

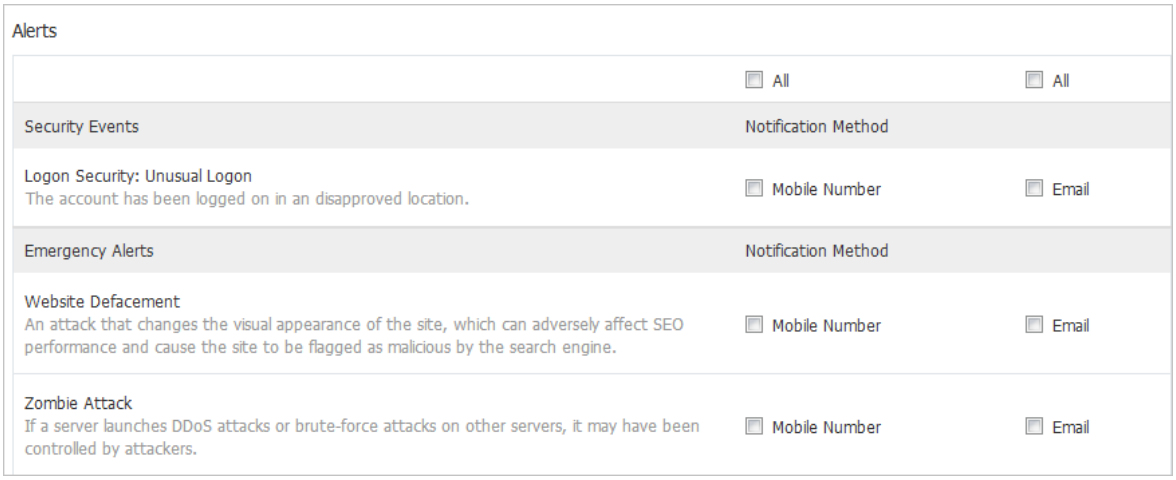
In the **Alerts** section, security administrators can configure the alert notification method for security events. When a security event occurs, the system notifies the alert contacts by email or text message. For more information about how to configure alert contacts, see Set alert recipients.

Alerts on tenants

1.
2.
3. click **Alert Settings**.
4. On the **Alert Settings** page, click **Tenant Alerts**.
5. In the **Alerts** section, select notification methods for each security event.
6. Click **Confirm**.

Alerts on the platform

1.
2.
3. click **Alert Settings**.
4. On the **Alert Settings** page, click **Platform Alerts**.
5. In the **Alerts** section, select notification methods for each security event.



6. Click **Confirm**.

22.1.11.4. Updates

22.1.11.4.1. Overview of the system updates feature

The system updates feature allows you to manually or automatically update the Apsara Stack Security and rule libraries for up-to-date protection.

The supported package import method depends on the Apsara Stack network environment.

- If Apsara Stack is connected to the Internet, you can choose **Automatically Download Update Packages**.
- If Apsara Stack is not connected to the Internet, you can choose **Manually Import Update Packages**.

The following table lists the update statuses of a rule library.

Update statuses of a rule library

Status	Description
To Be Updated	Indicates that a new version of the rule library is available for update.
Updating	Indicates that the rule library is being downloaded from Alibaba Cloud for update.
Updated	Indicates that the rule library has been updated.
Update Failed	Indicates that the rule library failed to be updated.

22.1.11.4.2. Enable automatic update check and update rule libraries

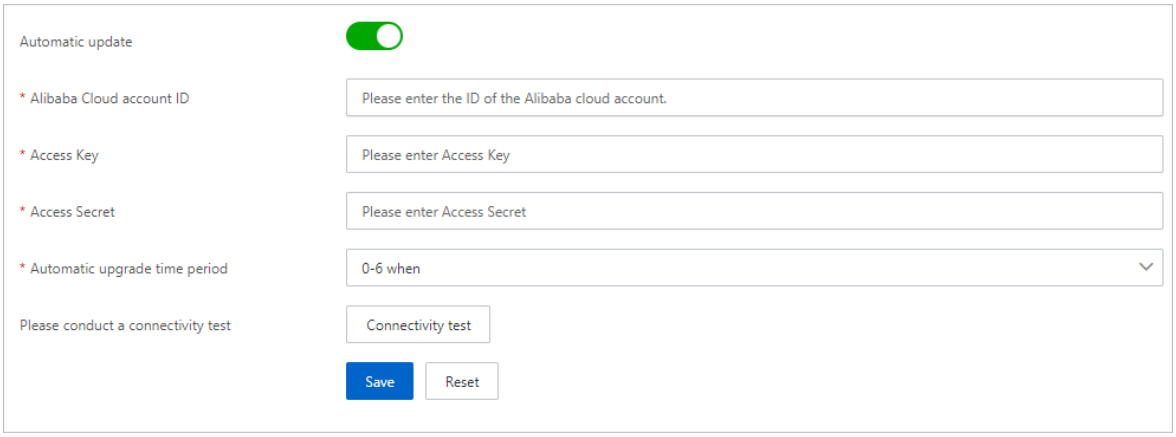
This topic describes how to enable automatic download of update packages and update rule libraries.

Context

If the Apsara Stack environment can connect to the Internet, you can enable automatic download of update packages to update the rule libraries.

Procedure

1.
2.
3. click **Updates**.
4. Click the **Upgrade configuration** tab.
5. On the **Upgrade configuration** tab, click the **Automatic update** tab.
6. Turn on **Automatic update** to enable automatic download of update packages and configure the following parameters.



Parameter	Description
Alibaba Cloud account ID	Enter the ID of the Alibaba Cloud account.
Access Key	Enter the AccessKey ID.
Access Secret	Enter the AccessKey secret.

Parameter	Description
Automatic upgrade time period	Select a time period for automatic updates. Valid values: ○ 0-6 when ○ 0-8 when ○ 0-24 when ○ 22-6 when

7. Click **Connectivity test**.
8. Click **Save** to enable the automatic check for update packages.

   After this switch is turned on, the system automatically downloads update packages on a regular basis.

22.1.11.4.3. Manually import an update package and update your service

This topic describes how to manually import an update package and update your service.

Prerequisites

The security administrator obtained the offline update package.

Context

If the Apsara Stack environment cannot connect to the Internet, you can manually import an update package to update the rule libraries.

Procedure

1.
2.
3. click **Updates**.
4. Click the **Upgrade configuration** tab.
5. Click the **Manual update** tab.
6. Manually import an update package.
   i. Click **Manually import offline upgrade packages**.
   ii. In the **Manually import offline upgrade packages** dialog box, click **Please select a file** to select an offline update package that is downloaded to your computer.
   iii. Click **Confirm**.

   After the update package is imported, the package appears on the **Version Update** tab. You can click Upgrade now in the Operation column of the update package to update your service.

22.1.11.4.4. Roll back a rule library

This topic describes how to roll back a rule library to a previous version.

Context

If an error occurs with an updated rule library, you can roll back the library to a previous version to prevent service interruption.

Procedure

1.
2.
3. click **Updates**.
4. Click the tab of the specific rule library. Example: **Server Security**.
5. In the Actions column for the rule library, choose **More > Roll Back**.
6. In the **Version Rollback** dialog box, click **Confirm**.

22.1.11.4.5. View the update history of a rule library

This topic describes how to view the update history of a rule library.

Context

You can view the update history of a rule library. This way, if an error occurs with the latest version, you can identify the issue and roll back the rule library to an earlier version.

Procedure

1.
2.
3. click **Updates**.
4. Click the tab of the specific rule library. Example: **Server Security**.
5. In the **Actions** column of a rule library, click **History**.

   On the **Previous Updates** page, you can view the update history of the rule library. Click **Details** to view the details about an update package.

22.1.11.5. Global configuration

22.1.11.5.1. Set CIDR blocks for traffic monitoring

22.1.11.5.1.1. Add a CIDR block for traffic monitoring

This topic describes how to add a CIDR block for traffic monitoring. Network Traffic Monitoring System of Apsara Stack Security monitors the traffic of a specific CIDR block.

Context

CIDR blocks are configured for Network Traffic Monitoring System. Security administrators can change the CIDR blocks for traffic monitoring based on business requirements. The settings of CIDR blocks apply only to a data center that is deployed in the region to which the specific CIDR block belongs.

> **Note**
> - Changes to CIDR block settings immediately take effect without the intervention of security administrators.
> - If you add the same CIDR block on the traffic collection CIDR block setting page and region setting page, make sure that you select the same region on both pages.

Procedure

1.

2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Traffic Collection IP Range** tab.
5. Click **Add**.
6. In the **Add CIDR Block for Monitoring** dialog box, configure parameters.

   - **CIDR Block**: Enter a CIDR block for traffic monitoring.

     Note Make sure the CIDR block that you enter is valid and unique.

   - **Region**: Select the region of the data center.
7. Click **OK**.

22.1.11.5.1.2. Manage CIDR blocks for traffic monitoring

This topic describes how to modify or delete CIDR blocks for traffic monitoring.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Traffic Collection IP Range** tab.
5. Select a region, enter the CIDR block that you want to query, and then click **Search**.

   View the information about the CIDR block for traffic monitoring and the region in the search result.
6. In the **Actions** column, manage a CIDR block for traffic monitoring.
   - Modify the CIDR block for traffic monitoring

     Click **Modify** to modify the region of the CIDR block for traffic monitoring.
   - Delete the CIDR block for traffic monitoring

     Click **Delete** to delete the CIDR block for traffic monitoring.

22.1.11.5.2. Region settings

22.1.11.5.2.1. Add a CIDR block for a region

This topic describes how to add CIDR blocks for regions that are detected and reported by using Server Guard.

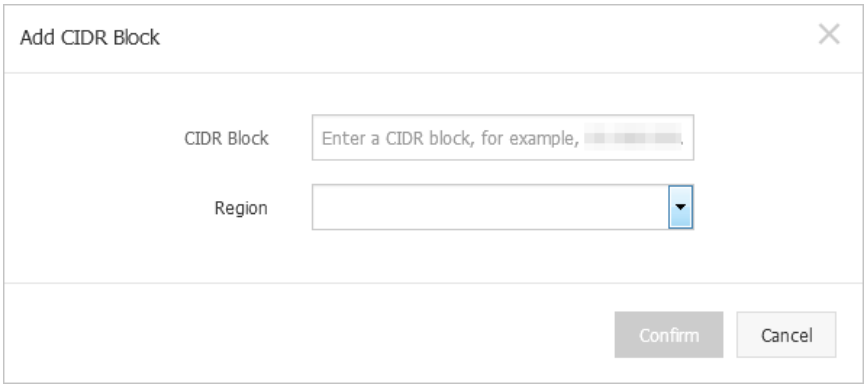
Context

Region settings are used for region detection of the Server Guard agent. Server Guard servers automatically detect and match the regions of servers based on the IP address information that is reported by the Server Guard agent.

> **Note** You can change the region of a CIDR block. After you change the region, you must also change the region for all assets in the CIDR block on the Asset Overview page.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Region** tab.
5. Click **Add**.
6. In the **Add CIDR Block** dialog box, configure parameters.



   - **CIDR Block**: Enter a CIDR block for the region.

     > **Note** Enter a valid CIDR block. You cannot enter a CIDR block that is configured for the region.

   - **Region**: Select a region.
7. Click **Confirm**.

22.1.11.5.2.2. Manage CIDR blocks for a region

This topic describes how to modify or delete CIDR blocks for a region.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Region** tab.
5. Select a region, enter the CIDR block that you want to modify or delete, and then click **Search**.

   You can view the information about the CIDR block for the region in the search result.
6. In the **Actions** column, click Modify or Delete to manage the CIDR block for the region.
   - Modify the CIDR block for the region

     Click **Modify** to modify the CIDR block for the region.
   - Delete the CIDR block for the region

     Click **Delete** to delete the CIDR block for the region.

22.1.11.5.3. Configure whitelists

This topic describes how to configure the whitelist for the feature that blocks brute-force attacks in Server Guard, and how to configure the whitelists in Threat Detection Service (TDS). These whitelists consist of IP addresses allowed by server brute-force attack blocking, IP addresses allowed by application attack blocking, and IP addresses allowed by web attack blocking.

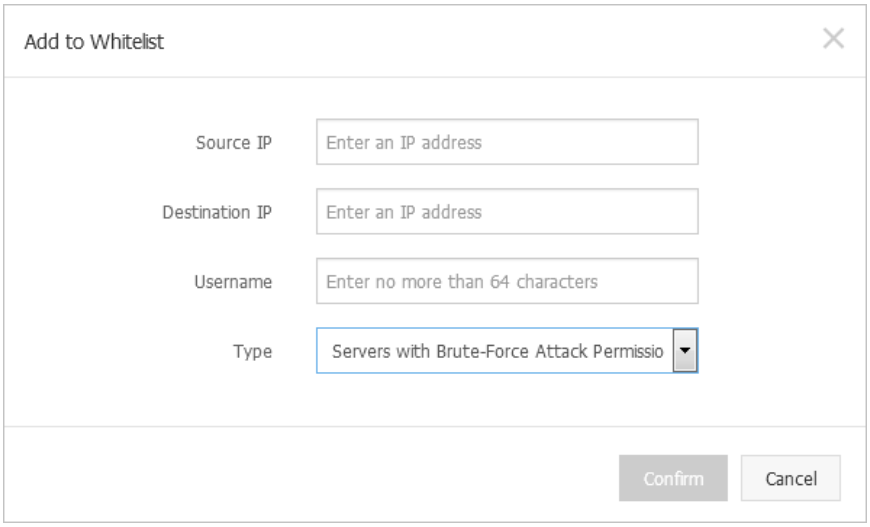
Context

If a normal request is identified as an attack by the attack blocking feature of TDS or the unusual logon detection feature of Server Guard, you can add the source IP address of the request to a whitelist to prevent further false positives.

> **Note** Make sure that the IP addresses in the whitelist are trusted.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Whitelist** tab.
5. Click **Add**.
6. In the **Add to Whitelist** dialog box, configure the parameters.



Parameter	Description
Type	○ **Global Login Whitelist**: Server Guard does not generate alerts for brute-force attacks or unusual logons from the IP addresses that are contained in this whitelist. ○ **Brute-Force Whitelist**: The attack blocking feature does not generate alerts for brute-force attacks from the IP addresses that are contained in this whitelist.
Source IP	Enter a source IP address or Classless Inter-Domain Routing (CIDR) block.
Destination IP	Enter a destination IP address or CIDR block.

7. Click **Confirm**.
   If you want to delete an existing whitelist, click **Delete** in the Actions column. In the **Delete Whitelist** message, click **Confirm**.

22.1.11.5.4. Configure policies that are used to block attacks

This topic describes how to enable web attack blocking and brute-force attack blocking.

Context

The attack blocking features protect your servers against web attacks and brute-force attacks.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Policy Configuration** tab.
5. Turn on or off the switches in the Actions column to enable or disable **Web Attack Blocking** or **Brute-Force Attack Blocking**.

> **Note**
>
> In the Actions column, a red switch indicates that a feature is disabled. A green switch indicates that a feature is enabled.
>
> After you disable the blocking feature for an attack type, Apsara Stack Security only generates alerts for this type of attacks.

22.1.11.5.5. Block IP addresses

This topic describes how to manually block requests for an IP address with a few clicks.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Block IP Addresses** tab.
5. In the upper-right corner of the tab, click **Add**.
6. In the **Add** dialog box, specify the IP address to block.

Parameter	Description
IP Protocol	Specify the type of the IP address that you want to block. Valid values: **IPv4** and **IPv6**.
Source IP	Enter the source IP address that you want to block.
Destination IP	Enter the destination IP address that you want to block.

Parameter	Description
Destination Port	Enter the destination port that is used together with the specified destination IP address.
Blocking Duration	Select a time range during which you want to block requests. Valid values: **1 Day**, **7 Days**, and **30 Days**.
Type	Select the blocking mode. Valid values: **Whitelist** and **Blacklist**.
Remarks	Enter the reason for the block.

7. Click **Confirm**.

22.1.11.5.6. Configure custom IP addresses and locations

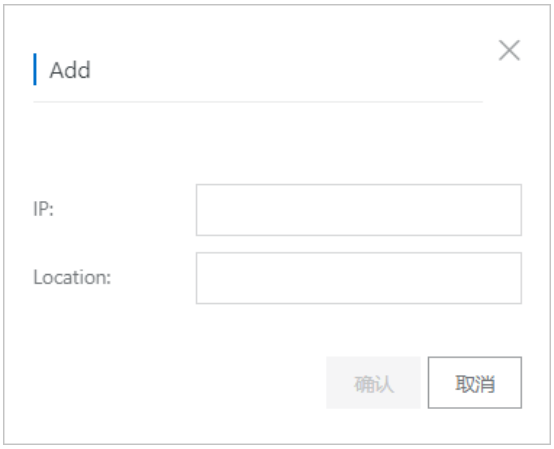
22.1.11.5.6.1. Add custom IP addresses and locations

This topic describes how to add custom IP addresses and locations. You can customize internal IP addresses based on your network plan. After you configure the internal IP addresses, IP addresses from the public address library do not match the addresses outside China.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Custom IP Location** tab.
5. Click **Add**.

   If you want to add multiple IP addresses and locations at a time, click **Batch Upload (.txt)**. Then, you can use Batch Upload (.txt) as a template to import multiple IP addresses and locations.
6. In the **Add** dialog box, configure the parameters.



7. Click **Ok**.

22.1.11.5.6.2. Manage custom IP addresses and locations

This topic describes how to modify and delete custom IP addresses and locations.

Procedure

1.
2.
3. click **Global Settings**.
4. On the **Global Settings** page, click the **Custom IP Location** tab.
5. In the **Actions** column, manage custom IP addresses and locations.
   - To change a custom IP address and a location:

     Click **Modify**. In the **Modify** dialog box, change the custom geographical location.
   - To delete a custom IP address and a location:

     Click **Delete**. In the **Delete** message, click **OK**.

22.1.11.6. System monitoring

22.1.11.6.1. Inspect services

This topic describes how to inspect services such as Cloud Firewall and Network Traffic Monitoring System in Apsara Stack Security Center. You can monitor the status and features of the services.

Procedure

1.
2.
3. click **System Monitoring**.
4. In the **System Inspection** section of the **Network Security** tab, inspect the services in the inspection list.

   To inspect a single service or multiple services at a time, perform the following operations:
   - Inspect multiple services at a time: In the **System Inspection** section, click **One-Click Inspection** to inspect all services in the inspection list.
   - Inspect a single service: In the **System Inspection** section, click **Inspect Now** in the **Actions** column of the service that you want to inspect.

   After the services are inspected, the status of the services changes to **Complete** in the **Inspection Status** column.
5. View the inspection results.

   You can view the following information about a service:
   - In the inspection list, view the service name, last inspection time, number of inspection items, number of inspection items whose status is normal, number of inspection items whose status is abnormal, and inspection status.
   - Click **Details** in the **Actions** column of a service. In the **Inspection Result Details** panel, view the number of inspection items whose status is normal, number of inspection items whose status is abnormal, and details of each item.
   - Click **Download** in the **Actions** column of a service. Download the inspection results to your computer as prompted for backup and reference.

22.1.11.7. Account management

22.1.11.7.1. View and modify an Apsara Stack tenant account

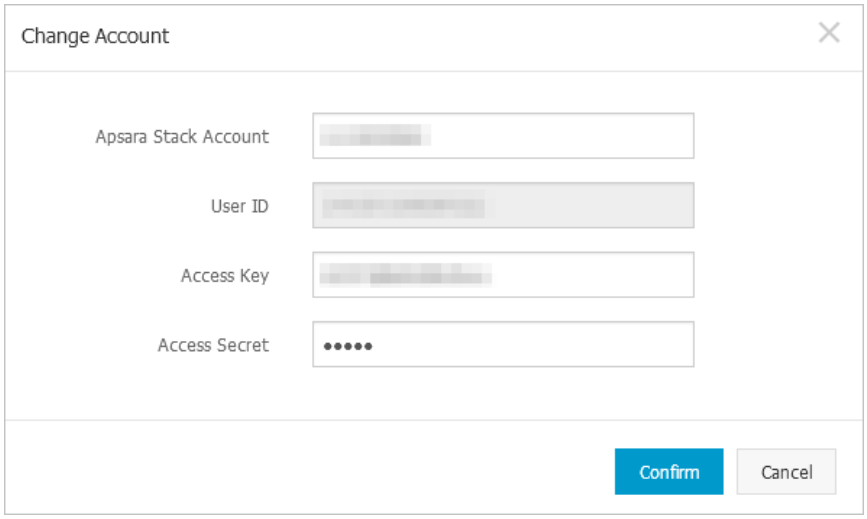
This topic describes how to view and modify the information about your Apsara Stack tenant account that is bound to the system.

Context

Note All assets in Apsara Stack Security are bound to your Apsara Stack tenant account. You can modify the account information. Proceed with caution.

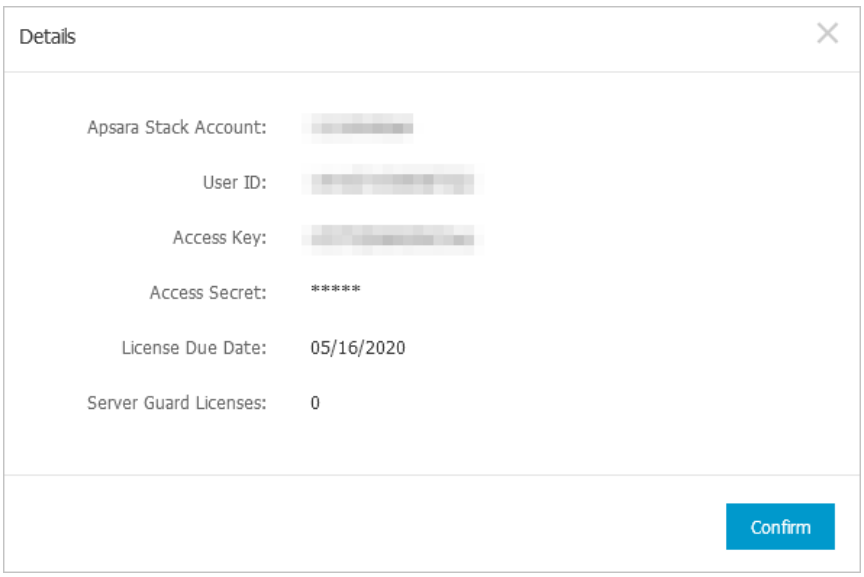
Procedure

1.
2.
3. click **Accounts**.
4. On the **Accounts** page, click the **Apsara Stack Account** tab.
5. Modify the information about your Apsara Stack tenant account.
   i. In the Actions column, click **Modify**.
   ii. In the **Change Account** dialog box, modify the account information.



   iii. Click **Confirm**.
6. View the details of your Apsara Stack tenant account.

   In the Actions column, click **Details** to view the details of your Apsara Stack tenant account.



22.1.11.7.2. Add an Alibaba Cloud account

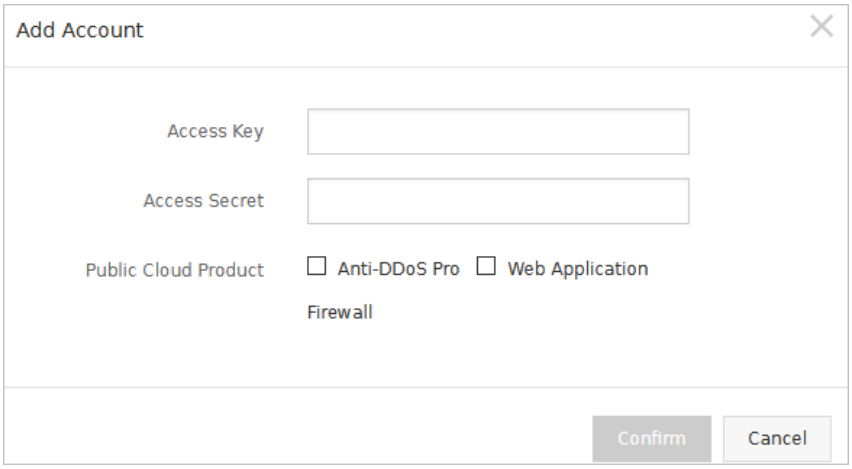
This topic describes how to add an Alibaba Cloud account in Apsara Stack Security Center. After you add the account, you can use features in a hybrid cloud.

Context

After you add an Alibaba Cloud account in Apsara Stack Security Center, you can manage the Anti-DDoS Pro, Anti-DDoS Premium, and Web Application Firewall (WAF) instances that belong to the Alibaba Cloud account in Apsara Stack Security Center. This way, you can use features in a hybrid cloud.

Procedure

1.
2.
3. click **Accounts**.
4. On the **Accounts** page, click the **Public Cloud Account** tab.
5. On the **Feature Integration** tab, click **Add**.
6. In the **Add Account** dialog box, enter the information about your Alibaba Cloud account and select Alibaba Cloud services to use.



   - Enter the **AccessKey ID** and **AccessKey secret** of your Alibaba Cloud account.
   - Select Alibaba Cloud services to use for Public Cloud Product. Valid values: **Anti-DDoS Pro**, **Web Application Firewall**, and both.
7. Click **Confirm**.

Result

After the account is added, it is displayed on the **Public Cloud Account** tab. To modify or delete the account, you can click **Modify** or **Delete** in the Actions column.

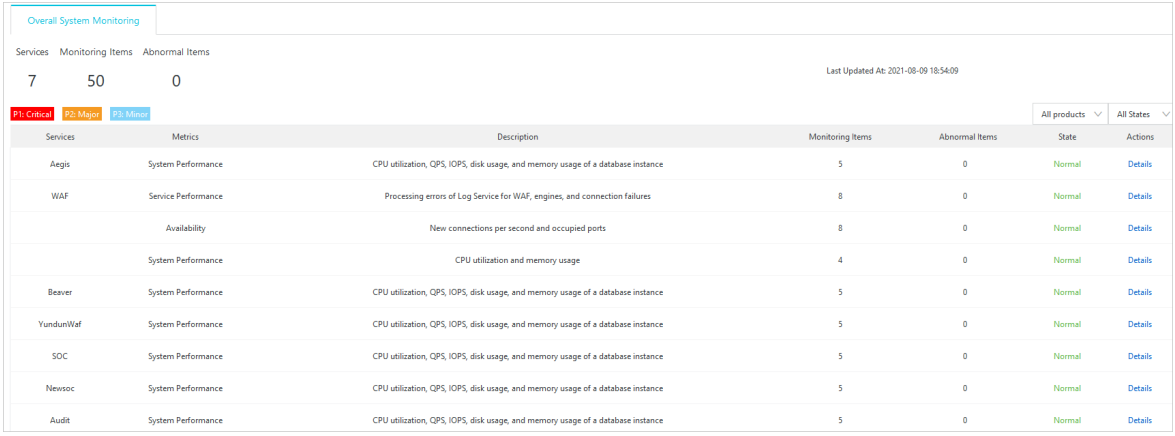
22.1.11.8. View and manage metrics

Apsara Stack Security Center allows you to monitor security services. This helps find performance bottlenecks at the earliest opportunity. Then, you can scale out, scale up, or downgrade services to prevent system failures. This topic describes how to view the information about security services in Apsara Stack Security Center and how to manage metrics.

View information about overall system monitoring

1.

2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **Security Monitoring**.
4. On the **Security monitoring** tab, view the overall information about security services in Apsara Stack Security Center and the list of monitored security services.



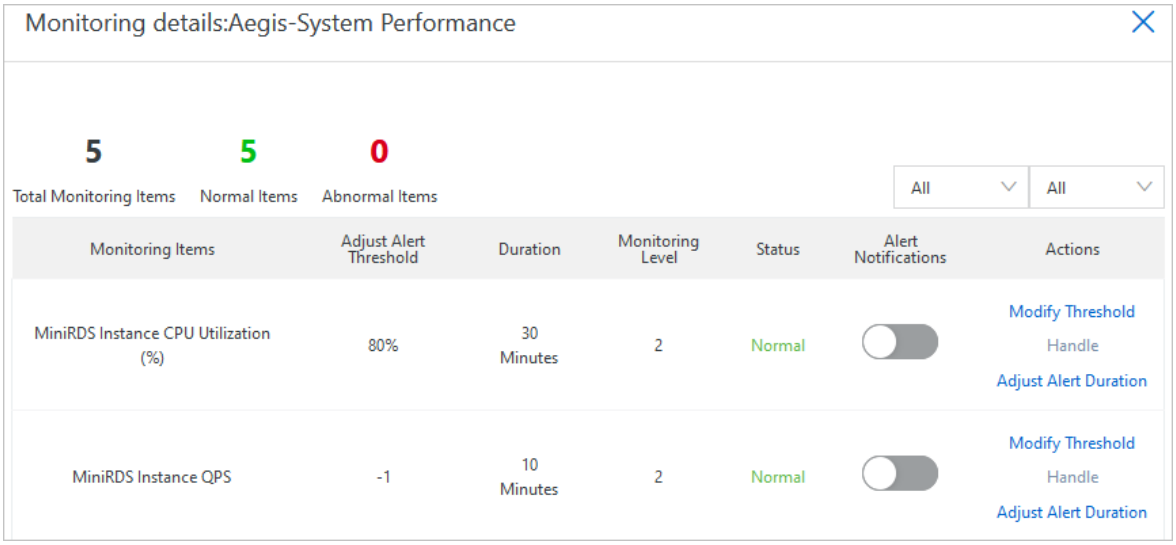
In the upper-left corner of the **Overall System Monitoring** tab, you can view the overall information about security services.

- **Services**: the total number of security services monitored in Apsara Stack Security Center.
- **Monitoring Items**: the total number of metrics.
- **Abnormal Items**: the number of metrics whose status is abnormal.

In the list of monitored security services, you can view the following information.

Parameter	Description
Services	The security service monitored in Apsara Stack Security Center.
Metrics	The monitoring indicator for the monitored security service.
Description	The description of the monitoring indicator.
Monitoring Items	The total number of metrics that belong to the monitoring indicator.
Abnormal Items	The number of metrics whose status is abnormal, and the number of metrics at each urgency level. The urgency levels are indicated by different colors. The red color indicates a critical exception, the orange color indicates an important exception, and the blue color indicates a moderate exception. 
State	The status of the monitoring indicator.

5. Click **Details** in the **Actions** column of a monitoring indicator. In the **Monitoring details** panel, view the details of metrics that belong to the monitoring indicator.



In the upper-left corner of the **Monitoring details** panel, you can view the overall information about the metrics.

- **Total Monitoring Items**: the total number of metrics that belong to the monitoring indicator.
- **Normal Items**: the number of metrics in the normal state.
- **Abnormal Items**: the number of metrics in the abnormal state.

In the metric list, you can view the following information.

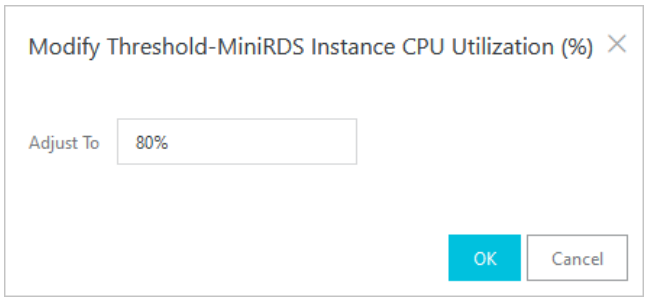
Parameter	Description
Monitoring Items	The name of the metric.
Adjust Alert Threshold	The threshold for the metric. If the value of the metric reaches the threshold and lasts for the specified period of time, the status of the metric becomes abnormal.
Duration	The period of time. If the value of the metric reaches the threshold and lasts for the specified period of time, the status of the metric becomes abnormal.
Monitoring Level	The urgency level displayed when the status of the metric becomes abnormal.
Status	The status of the metric.
Alert Notifications	The switch of the alert notification feature. You can turn on or off the switch in the **Alert Notifications** column based on your business requirements. If the On icon appears in the Alert Notifications column, a notification is sent when the status of the metric becomes abnormal.

Manage metrics

1.
2. Choose **Global Platform Security > Alibaba Cloud Security**.
3. click **Security Monitoring**.
4. On the **Security monitoring** tab, find a monitoring indicator and click **Details** in the **Actions** column.
5. In the **Monitoring details** panel, find a metric and click **Modify Threshold**, **Handle**, or **Adjust Alert Duration** in the **Actions** column to manage the metric.

   You can perform the following operations on the metric:

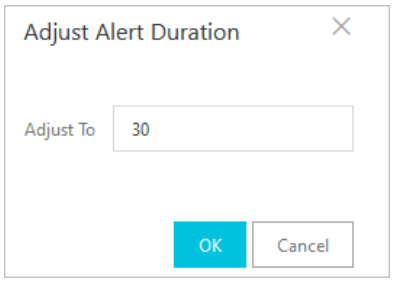
- **Modify Threshold**: In the **Modify Threshold** dialog box, modify the threshold and click **OK**.



  After the threshold is modified, an alert is generated when the value of the metric reaches the new threshold.

- **Handle**: You can configure the status of the metric based on your business requirements.
  - If you do not want to handle the alert generated from the metric, select **Ignore** from the drop-down list. The status of the metric becomes **ignored**.
  - After you handle the alert generated from the metric, select **Handled** from the drop-down list. The status of the metric becomes **handled**.

- **Adjust Alert Duration**: In the **Adjust Alert Duration** dialog box, modify the period of time and click **OK**.



  When the value of the metric reaches the threshold and lasts for the specified period of time, the status of the metric becomes abnormal.

22.2. Security Administrator Guide

22.2.1. Restrictions

Before logging on to Apsara Stack Security Center, make sure that your local PC meets the requirements.

Configuration requirements

Item	Requirements
Browser	• Internet Explorer: 11 or later • Google Chrome (recommended): 42.0.0 or later • Mozilla Firefox: 30 or later • Safari: 9.0.2 or later
Operating system	• Windows XP, Windows 7, or later • Mac

22.2.2. Log on to Cloud Security Operations Center

This topic describes how to log on to Cloud Security Operations Center.

Prerequisites

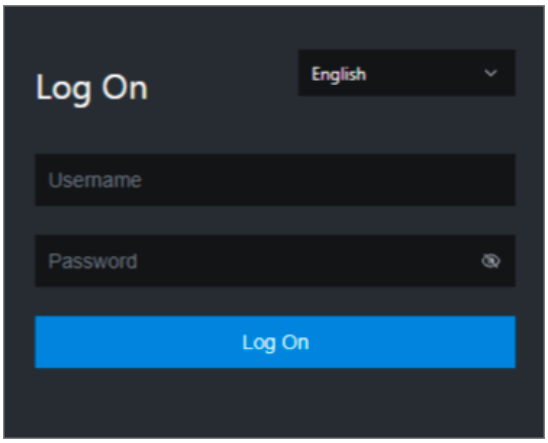
- The endpoint of the Apsara Uni-manager Operations Console and the username and password used to log on to the console are obtained from the deployment personnel or an administrator.

  The endpoint of the Apsara Uni-manager Operations Console is in the following format: *region-id*.ops.console.*intranet-domain-id*.

- A browser is available. We recommend that you use Google Chrome.

Procedure

1. Open your Chrome browser.
2. In the address bar, enter the endpoint of the Apsara Uni-manager Operations Console. Press the Enter key.

   > **Note** You can select a language from the drop-down list in the upper-right corner of the page.

3. Enter your username and password.

   > **Note** Obtain the username and password used to log on to the Apsara Uni-manager Operations Console from the deployment personnel or an administrator.

   When you log on to the Apsara Uni-manager Operations Console for the first time, you must change the password of your username.

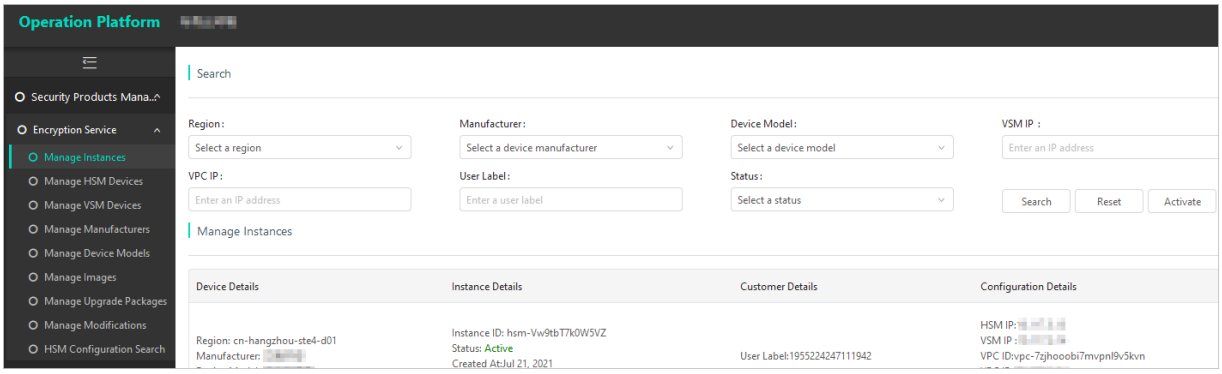
   For security reasons, your password must meet the following requirements:

   - The password contains uppercase and lowercase letters.
   - The password contains digits.
   - The password contains the following special characters: ! @ # $ %
   - The password must be 10 to 20 characters in length.

4. Click **Log On**.
5. In the top navigation bar of the Apsara Uni-manager Operations Console, click **O&M**. In the left-side navigation pane, choose **Product Management > Products**. On the page that appears, click **Cloud Security Operation Center** in the **Security Services** section.

Result

On the **Operations Platform** page, you can view the following information.



22.2.3. Services

22.2.3.1. Data Encryption Service

22.2.3.1.1. Manage Data Encryption Service instances

22.2.3.1.1.1. Create an instance

This topic describes how to create a Data Encryption Service instance.

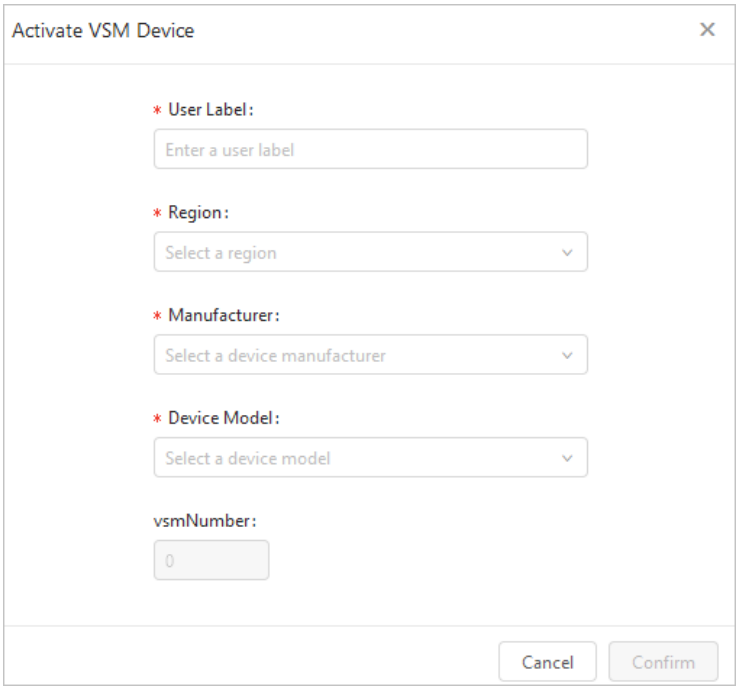
Context

After you create an instance, you can view the instance in the **Manage Instances** section. The instance is in the *Not Configured* state.

> **Note** Before you use Data Encryption Service, you must configure a virtual private cloud (VPC) for the instance. For more information, see Configure VPC.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Instances**.
3. On the page that appears, click **Activate** in the **Search** section.
4. In the **Activate VSM Device** dialog box, configure the following parameters.



> **Note** You can view the user ID in **Apsara Stack Security Center**. To view the user ID, log on to Apsara Stack Security Center, choose **System Configuration > Accounts**. On the page that appears, view the user ID in the **User ID** column.

5. Click **Yes**.

22.2.3.1.1.2. Configure a VPC

This topic describes how to configure a virtual private cloud (VPC) for a Data Encryption Service instance.

Prerequisites

A VPC is created.

Context

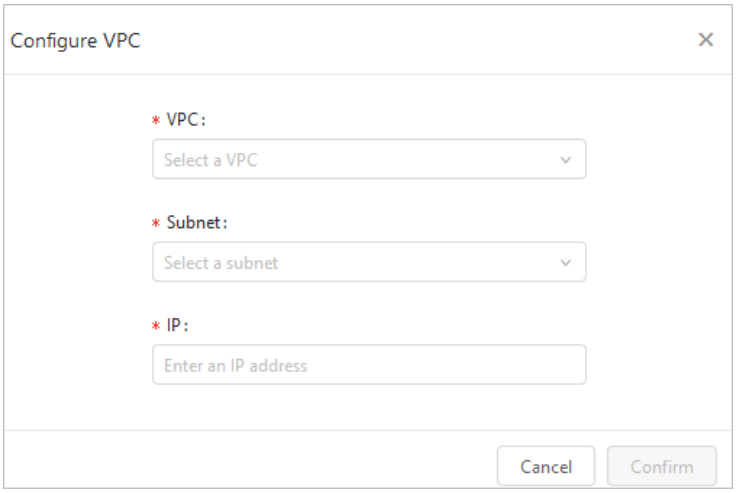
Before you can use a Data Encryption Service instance, you must configure a VPC for the instance.

> **Note** After you create an instance, you must configure a VPC for the instance.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Instances**.
3. On the page that appears, find the instance for which you want to configure a VPC in the **Manage Instances** section and click **Configure VPC** in the Actions column.

   In the **Search** section, you can specify the search conditions to find an instance.
4. In the Configure VPC dialog box, configure the following parameters for a VPC.



Parameter	Description
vpc	The VPC that you want to bind with the instance.
Subnet	The subnet of the VPC that you want to bind with the instance.
IP	The IP address of the instance. The IP address must be within the subnet of the VPC.

5. Click **Confirm**.

22.2.3.1.1.3. Manage an instance

This topic describes how to unbind a virtual private cloud (VPC) from a Data Encryption Service instance. This topic also describes how to release and delete the instance.

Context

You can perform the following operations on an existing instance.

Operation	Description
Unbind a VPC from the instance.	If the VPC or the subnet of the VPC to which the instance belongs is changed, you can unbind the VPC by performing the steps described in this topic. After you unbind the VPC, the **state** of the instance is changed to *Not Configured*. If you want to use the instance again, you must configure a VPC for the instance.
Release the instance.	If you no longer need the instance, you can release the instance. You can activate and reuse the instance that you released.
Delete the instance.	If the instance is in the *Released* **state**, you can delete the instance.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Instances**.
3. On the page that appears, find the instance that you want to manage in the **Manage Instances** section and perform the following operations based on your business requirements.
   - Click **Unbind VPC** to unbind the VPC from the instance.
   - Click **Release** to release the instance.

- Click **Delete** to delete the instance that you released.

4. In the dialog box that appears, click **OK**.

22.2.3.1.2. Manage HSMs

22.2.3.1.2.1. Add an HSM

This topic describes how to add a hardware security module (HSM).

Prerequisites

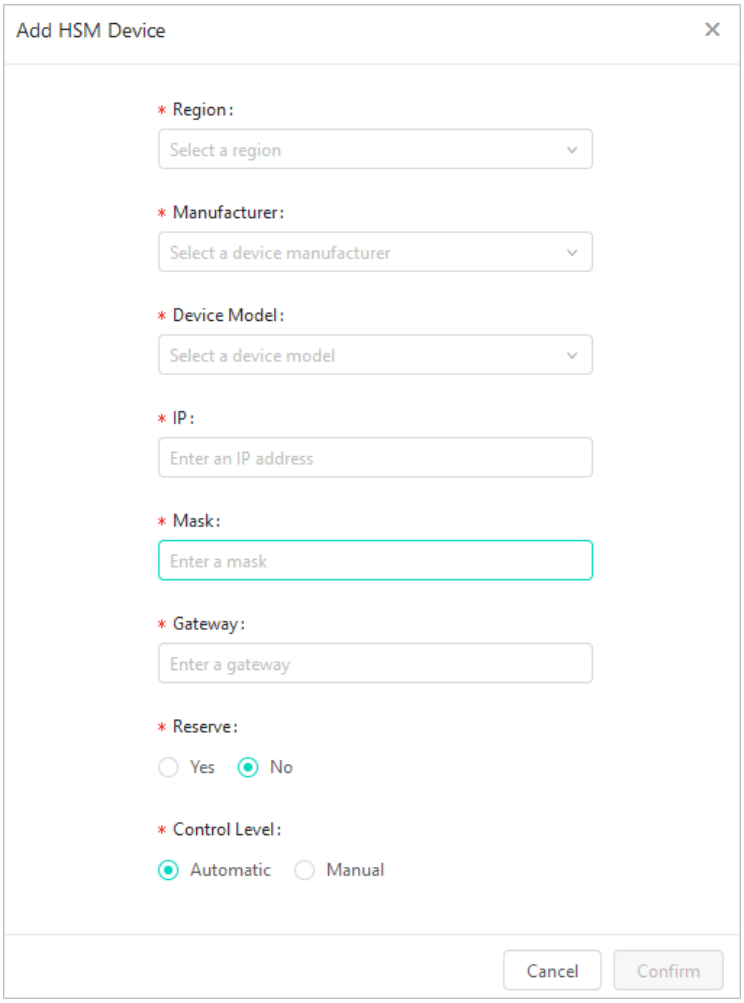
- The HSM can communicate with Cloud Security Operations Center.
- The HSM and all the virtual security modules (VSMs) on the HSM are in the running state.

Context

An HSM is a physical cryptographic device provided by a manufacturer. An HSM can host multiple VSMs.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage HSM Devices**.
3. On the page that appears, click **Add** in the **Search** section.
4. In the Add HSM Device dialog box, configure the following parameters.



Parameter	Description
Region	Select the region of the HSM.
Manufacturer	Select the manufacturer of the HSM.
Device Model	Select the model of the HSM.
IP	Specify the IP address of the HSM.
Mask	Specify the mask of the HSM.
Gateway	Specify the IP address of the gateway.
Reserve	Determine whether to reserve the HSM. All the VSMs on the reserved HSM are allocated and used only when you want to migrate or update an HSM.
Control Level	The control level of the HSM. Valid values: ○ Automatic ○ Manual

5. Click **Confirm**.

   The task of adding the HSM requires some time to complete. You can go to the **Manage Modifications** page to view the task.

22.2.3.1.2.2. Configure the network information for an HSM

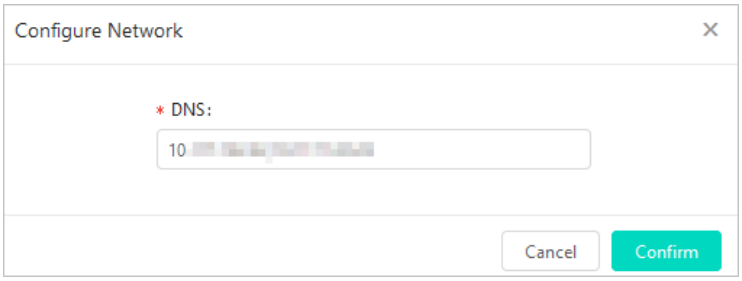
This topic describes how to configure the Domain Name System (DNS) information for a hardware security module (HSM).

Context

The network information for an HSM is configured. For more information about how to configure the network information, see Add an HSM. If you want to modify the network information of the HSM, find the HSM and click **Configure Network** in the Actions column. Then, modify the DNS information.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage HSM Devices**.
3. In the **Manage HSM Devices** section, find the HSM whose network information you want to modify and click **Configure Network**.
4. In the Configure Network dialog box, configure the DNS parameter.



5. Click **Confirm**.

22.2.3.1.2.3. Migrate an HSM

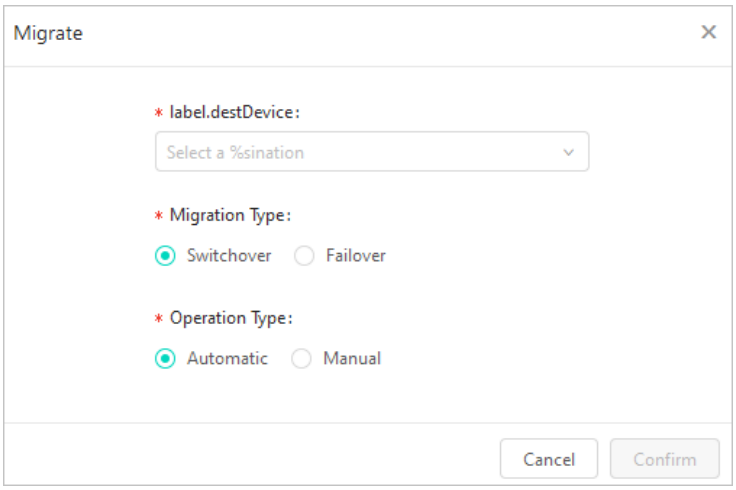
This topic describes how to migrate a source hardware security module (HSM) to a destination HSM.

Context

If a source HSM requires to be maintained or fails, you can migrate the source HSM to a destination HSM to ensure service continuity. After the source HSM recovers, the service is switched back to the source HSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage HSM Devices**.
3. On the page that appears, find the source HSM in the **Manage HSM Devices** section and click **Migrate**.
4. In the Migrate dialog box, configure the following parameters.



Parameter	Description
Destination HSM	Specify the destination HSM.
Migration Type	Select a scenario in which you want to migrate the source HSM to the destination HSM. Valid values: ○ Switchover: Select Switchover for daily maintenance. For example, you can select Switchover when you want to update the source HSM. After the update is complete, you can migrate the service back to the source HSM. ○ Failover: Select Failover if the source HSM fails.
Operation Type	Select a migration type. Valid values: ○ Automatic The source HSM is automatically migrated without manual operations. In most cases, we recommend that you select Automatic. ○ Manual The source HSM must be manually migrated. If you select Manual, you are redirected to the **Job Details** page. On this page, you must manually migrate the source HSM.

5. Click **Confirm**.
6. Determine whether to view the migration process.

- If you do not want to view the migration process, click **Cancel** to stay in the **Manage HSM Devices** section.
- If you want to view the migration process, click **Confirm** and click OK in the message that appears to go to the **Job Details** page. On this page, you can view the migration process.

22.2.3.1.2.4. Update an HSM

This topic describes how to update a hardware security module (HSM).

Context

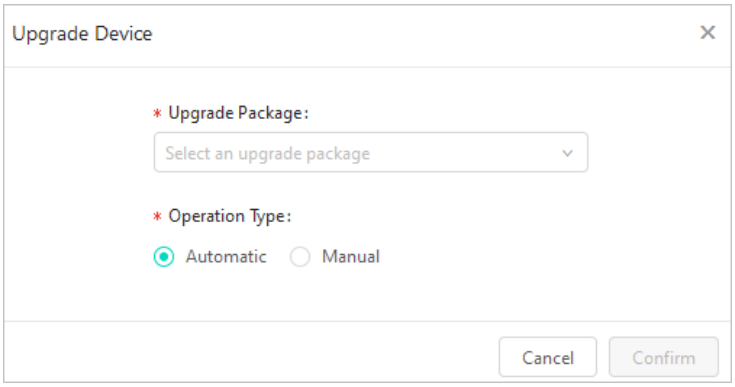
Before you update an HSM, you must migrate the HSM to another HSM that is of the same model and is running.

You must update or roll back an HSM and the VSMs in the following order:

- If you want to update an HSM and the VSMs at the same time, you must first update the HSM and then the VSMs.
- If you want to roll back an HSM and the VSMs at the same time, you must first roll back the VSMs and then the HSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage HSM Devices**.
3. On the page that appears, find the HSM that you want to update in the **Manage HSM Devices** section and click **Upgrade**.
4. In the **Upgrade Device** dialog box, configure the Upgrade Package and Operation Type parameters.



5. Click **Confirm**.

22.2.3.1.2.5. Manage an HSM

This topic describes how to manage a hardware security module (HSM), such as view details about the HSM, reload the HSM, and configure Network Time Protocol (NTP) for the HSM.

Context

You can perform the following operations on an existing HSM.

Operation	Description
Refresh the URL that is used to export snapshots. A snapshot contains the keys and configuration data on the HSM.	If the snapshots fail to be exported, the specified export URL may be invalid. Click Refresh Export Image URL to refresh the URL. Then, export the snapshots again.
View the details about the HSM.	View the details about the HSM.

Operation	Description
Reserve the HSM.	A reserved HSM is allocated and used only when you want to migrate or update an HSM. If you want to update an HSM or migrate an HSM, the HSM is migrated to the reserved HSM first.
Refresh the information about Log Service.	Refresh the information about Log Service.
Reload the information about the HSM.	If the information in the dialog box that appears after you click **Device Details** is different from the information in the database that is connected to Data Encryption Service, click **Reload** to synchronize the information.
Enable or disable monitoring on the HSM.	Enable or disable monitoring on the HSM.
Configure NTP for the HSM.	Configure NTP to synchronize the clock between the HSM and the NTP server.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage HSM Devices**.
3. On the page that appears, find the HSM that you want to manage in the **Manage HSM Devices** section and perform the following operations based on your business requirements:
   - Click **Refresh Image Export URL** to refresh the URL that is used to export snapshots.
   - Click **Device Details** to view the details about the HSM.
   - Click **Reserve** to reserve the HSM.
   - Click **Refresh SLS** to refresh the information about Log Service.
   - Click **Reload** to reload the information about the HSM.
   - Click **Modify** to enable or disable monitoring on the HSM.
   - Click **Configure NTP** to configure the NTP server and synchronization frequency.
4. In the message or dialog box that appears, click **OK** or Confirm.

22.2.3.1.3. Manage VSMs

22.2.3.1.3.1. Configure the network information for a VSM

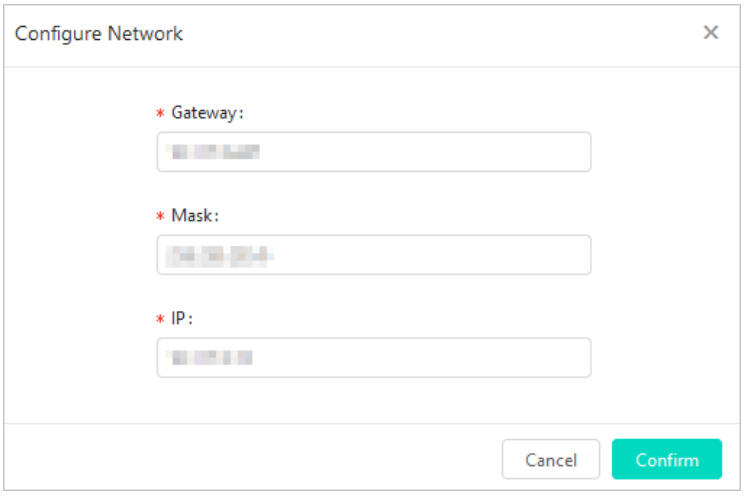
This topic describes how to configure the gateway, mask, and IP address of a virtual security module (VSM).

Context

If you want to modify the network information of the VSM, find the VSM and click **Configure Network** in the Actions column. Then, change the gateway, the mask, or the IP address of the VSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage VSM Devices**.
3. In the **Manage VSM Devices** section, find the VSM whose network information you want to modify and click **Configure Network**.
4. In the Configure Network dialog box, configure the Gateway, Mask, and IP parameters.



5. Click **Confirm**.

22.2.3.1.3.2. Update a VSM

This topic describes how to update a virtual security module (VSM).

Context

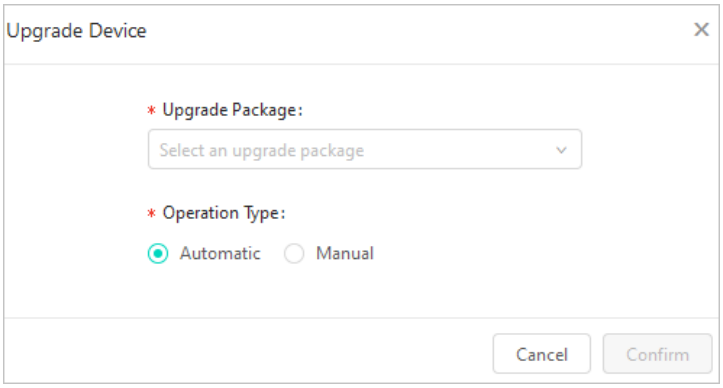
Before you update a VSM, you can migrate the VSM to another VSM that is of the same type and is running.

You must update and roll back an HSM and the VSMs in the following order:

- If you want to update an HSM and the VSMs at the same time, you must first update the HSM and then the VSMs.
- If you want to roll back an HSM and the VSMs at the same time, you must first roll back the VSMs and then the HSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage VSM Devices**.
3. On the page that appears, find the VSM that you want to update in the **Manage VSM Devices** section and click **Upgrade**.
4. In the **Upgrade Device** dialog box, configure the Upgrade Package and Operation Type parameters.



5. Click **Confirm**.

22.2.3.1.3.3. Export snapshots

This topic describes how to export the snapshots of a virtual security module (VSM). A snapshot contains keys and configuration data on the VSM.

Context

The exported snapshots can be used to restore the VSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage VSM Devices**.
3. On the page that appears, find the VSM whose snapshots you want to export in the **Manage VSM Devices** section and click **Export Snapshot**.
4. In the message that appears, click **OK**.

   After the snapshots are exported, you can view the information about the snapshots. To view the information, click Manage Images in the left-side navigation pane. On the page that appears, view the information in the **Manage Images** section.

22.2.3.1.3.4. Manage a VSM

This topic describes how to manage a virtual security module (VSM), such as view details about the VSM, reserve the VSM, and restart the VSM.

Context

You can perform the following operations on the VSM.

Operation	Description
View the details about the VSM.	View the details about the VSM.
Reserve the VSM.	A reserved VSM is allocated and used only when you want to migrate or update a VSM. If you want to update or migrate a VSM, the VSM is migrated to the reserved VSM first.
Reset the VSM.	If the VSM is released but data on the VSM is not deleted, reset the VSM.
Stop the VSM.	Stop the running VSM.
Restart the VSM.	Restart the VSM.
Enable or disable monitoring on the VSM.	Enable or disable monitoring on the VSM.
Delete the VSM.	Delete the VSM that you no longer need.
Change the state of the VSM to Running.	After the VSM that was in the *System Unavailable* state becomes available, click Activate to change the state of the VSM to *Running*. Then, the system can allocate the VSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage VSM Devices**.
3. On the page that appears, find the VSM that you want to manage in the **Manage VSM Devices** section and perform the following operations based on your business requirements:
   - Click **Device Details** to view the details about the VSM.

- Click **Reserve** to reserve the VSM.
- Click **Reset** to reset the VSM.
- Click **Stop** to stop the running VSM.
- Click **Restart** to restart the VSM.
- Click **Modify** to enable or disable monitoring on the VSM.
- Click **Delete** to delete the VSM.
- Click **Activate** to change the state of the VSM from *System Unavailable* to *Running*.

4. In the message or dialog box that appears, click **OK** or Confirm.

22.2.3.1.4. Manage manufacturers

22.2.3.1.4.1. Add a manufacturer

This topic describes how to add a manufacturer.

Context

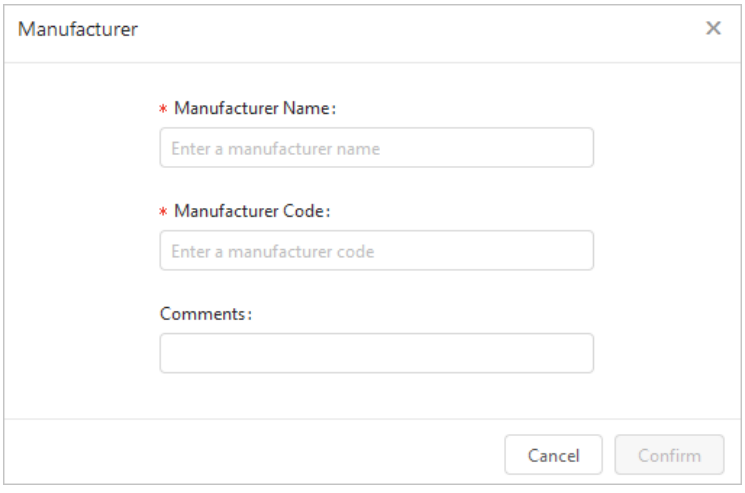
Accurate manufacturer information is required for Data Encryption Service instances and hardware security modules (HSMs). Before you use an HSM that you purchased, you must configure the manufacturer information.

The following manufacturer is supported.

Manufacturer	Code
TASS	jnta

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Manufacturers**.
3. In the upper-right corner of the page that appears, click **Add**.
4. In the Manufacturer dialog box, configure the parameters.



5. Click **Confirm**.

22.2.3.1.4.2. Manage a manufacturer

This topic describes how to modify the information about a manufacturer and delete a manufacturer.

Context

Accurate manufacturer information is required for Data Encryption Service instances and hardware security modules (HSMs). Before you use an HSM that you purchased, you must configure the manufacturer information.

The following manufacturer is supported.

Manufacturer	Code
TASS	jnta

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Manufacturers**.
3. On the page that appears, delete a manufacturer or modify the information about a manufacturer.
   - Click **Delete** to delete a manufacturer.
   - Click **Modify** to modify the information about a manufacturer.
4. Click OK or **Confirm**.

22.2.3.1.5. Manage HSM models

22.2.3.1.5.1. Add an HSM model

This topic describes how to add a hardware security module (HSM) model.

Context

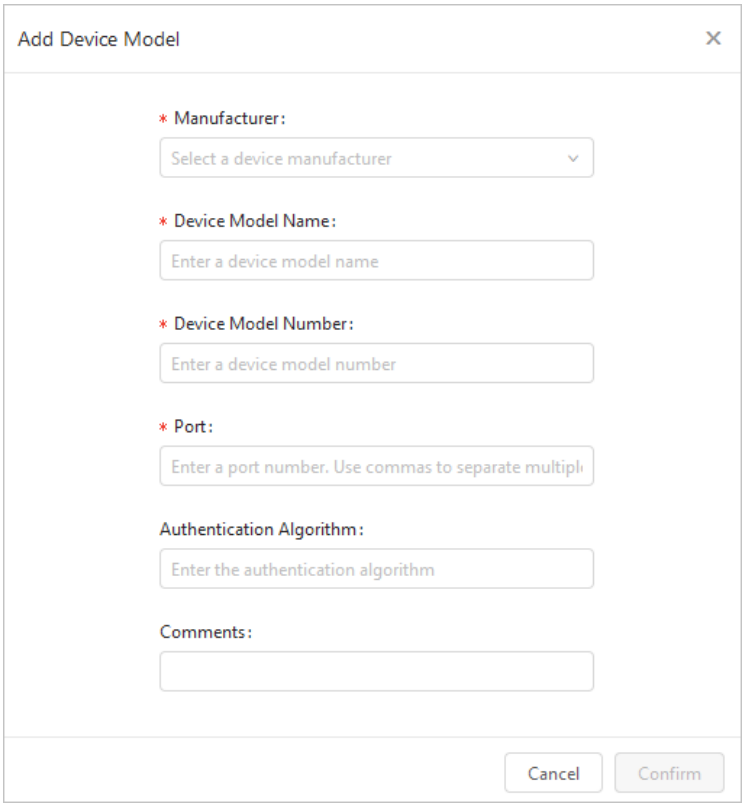
Accurate information about the HSM model is required for Data Encryption Service instances and HSMs. Before you use an HSM that you purchased, you must configure the HSM model information.

The following HSM models are supported.

Manufacturer	HSM model	HSM model number
TASS	Financial HSM	jnta.EVSM
TASS	Server HSM	jnta.GVSM
TASS	Signature verification HSM	jnta.SVSM

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Device Models**.
3. In the upper-right corner of the page that appears, click **Add**.
4. In the Add Device Model dialog box, configure the parameters.



5. Click **Confirm**.

22.2.3.1.5.2. Manage an HSM model

This topic describes how to modify the information about a hardware security module (HSM) model or delete an HSM model.

Context

Accurate information about the HSM model is required for Data Encryption Service instances and HSMs. Before you use an HSM that you purchased, you must configure the HSM model information.

The following HSM models are supported.

Manufacturer	HSM model	HSM model number
TASS	Financial HSM	jnta.EVSM
TASS	Server HSM	jnta.GVSM
TASS	Signature verification HSM	jnta.SVSM

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Device Models**.
3. On the page that appears, find the HSM model that you want to manage and perform the following operations based on your business requirements:
   - Click **Delete** to delete the HSM model.
   - Click **Modify** to modify the information about the HSM model.
4. Click **Confirm**.

22.2.3.1.6. View the information about snapshots

This topic describes how to view the snapshots of a virtual security module (VSM). A snapshot contains keys and configuration data on the VSM.

Context

You can use the snapshots that are displayed in the **Manage Images** section to restore VSMs. For more information about how to export snapshots, see Export snapshots.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Images**.
3. View the information about the snapshots.

22.2.3.1.7. Manage update files

22.2.3.1.7.1. Upload an update file

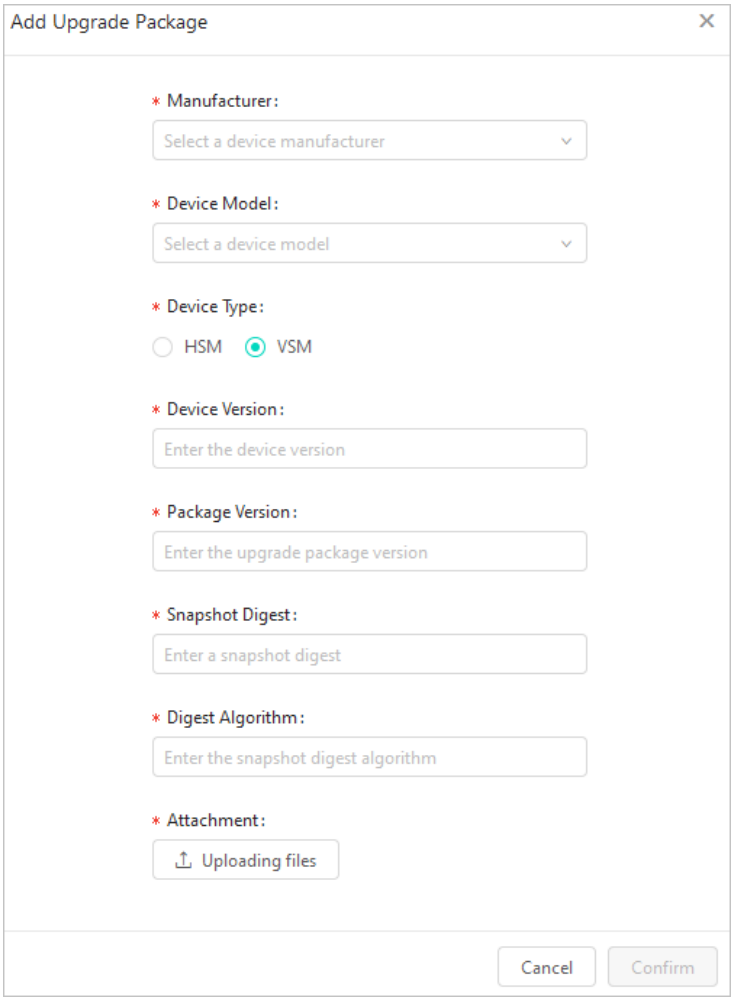
This topic describes how to upload an update file for a hardware security module (HSM) or a virtual security module (VSM).

Context

You can upload an update file based on your business requirements. If an update file fails to be uploaded, you must import the required HTTPS certificate on your browser. To import the required HTTPS certificate, you can open a new browser window, import the certificate, enter the upload URL in the address bar, and then re-upload the file.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Upgrade Packages**.
3. On the page that appears, click **Add** in the **Search** section.
4. In the Add Upgrade Package dialog box, configure the parameters.



5. Click **Uploading files**. Then, select and upload a update file from your computer.
6. Click **Confirm**.

22.2.3.1.7.2. Delete an update file

This topic describes how to delete an update file of a hardware security module (HSM) or a virtual security module (VSM).

Context

You can delete update files that you no longer need to free up storage space.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Upgrade Packages**.
3. On the page that appears, find the update file that you want to delete in the Manage Upgrade Packages section and click **Delete** in the Operation column.

22.2.3.1.8. Manage tasks

22.2.3.1.8.1. View task details

This topic describes how to view task details.

Context

If you perform an operation in Data Encryption Service, a task is generated and displayed on the Modification Details page. The task details show how the task is executed. If the task fails, you can view the task details to identify the issues.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Modifications**.
3. On the page that appears, find the task whose details you want to view and click **Job Details** in the Operation column.
4. View the details about the task.

22.2.3.1.8.2. Terminate a task

This topic describes how to terminate a running task.

Context

You can terminate running tasks that are performed by mistake or no longer needed.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > Manage Modifications**.
3. On the page that appears, find the running task that you want to terminate and click **Stop**.

22.2.3.1.9. Query the configurations of an HSM

This topic describes how to query the configurations of a hardware security module (HSM).

Context

You can query the configurations of an HSM. The configurations include the manufacturer, status, and network information about an HSM, and the virtual security modules (VSMs) on the HSM.

Procedure

1. Log on to Cloud Security Operations Center.
2. Choose **Security Products Management > Encryption Service > HSM Configuration Search**.
3. On the page that appears, configure **Manufacturer** and **HSM IP** in the **Search** section and click **Search**
4. View the configurations of the HSM and VSMs.

23.Log Service

23.1. User Guide

23.1.1. What is Log Service?

Log Service is a one-stop logging service developed by Alibaba Cloud. Log Service is widely used by Alibaba Group in big data scenarios.

You can use Log Service to collect, query, and consume log data without the need to invest in in-house data collection and processing systems. This enables you to focus on your business, improving business efficiency and helping your business to expand.

Log Service provides the following features:

- Log collection: Log Service allows you to collect events, binary logs, and text logs in real time by using multiple methods, such as Logtail and JavaScript.
- Query and analysis: Log Service allows you to query and analyze the collected log data and view analysis results on charts and dashboards.
- Status alert: Log Service can automatically run query statements at regular intervals after you create an alert task. If the query results meet the conditions of the alert task, Log Service sends an alert to the specified recipients in real time.
- Real-time consumption: Log Service provides real-time consumption interfaces through which log consumers can consume log data.

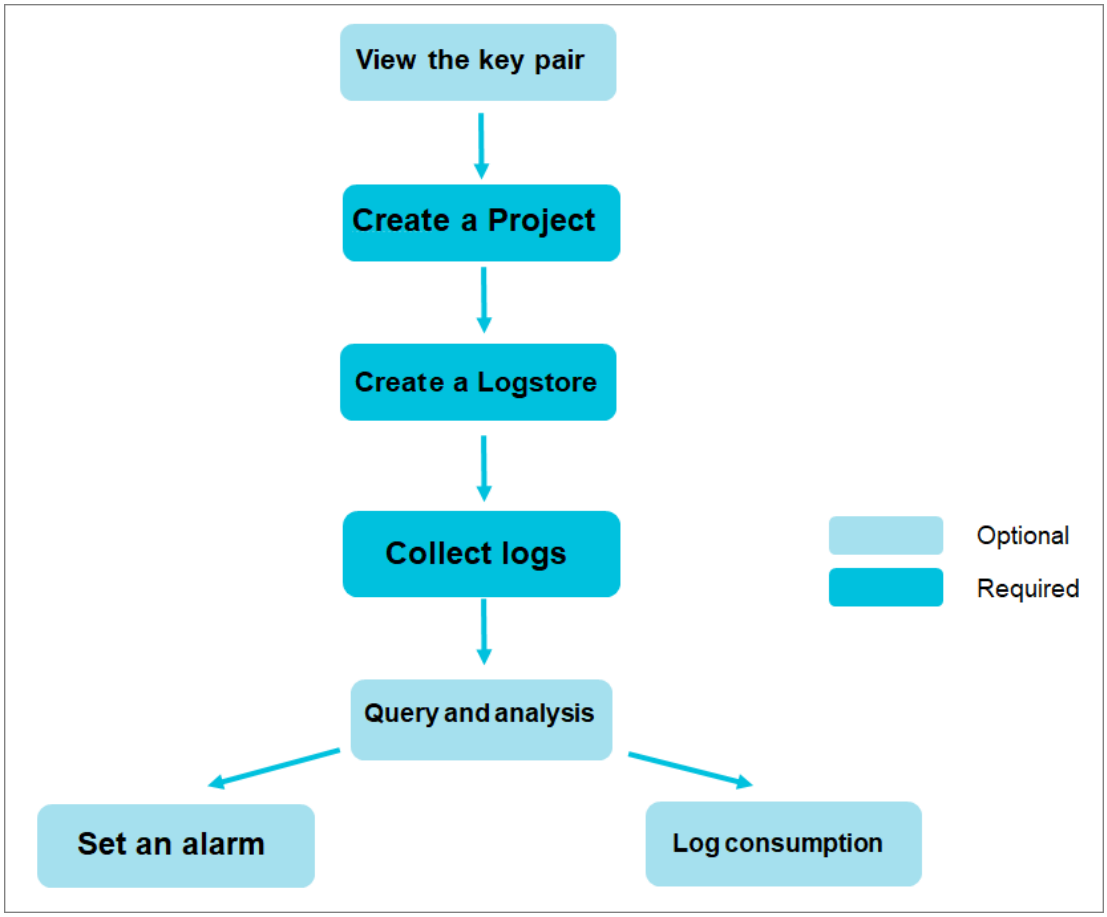
23.1.2. Quick start

23.1.2.1. Procedure

This topic describes the basic procedure to use Log Service. You can follow this procedure to create projects, create Logstores, and collect log data.

The following figure shows the procedure.

Procedure



1. Optional. Obtain an AccessKey pair.

   An AccessKey pair is a secure identity credential that you can use to call API operations and access your Alibaba Cloud resources. You can use the AccessKey pair to sign API requests and pass security authentication.

2. Create a project.

   Create a project in a specified region. For more information, see Create a project.

3. Create a Logstore.

   Create a Logstore for the project and specify the number of shards. For more information, see Create a Logstore.

4. Collect text logs.

   Select a method to collect log data based on your business requirements. For more information, see Collect text logs.

5. Configure indexes, and query and analyze log data.

   - Before you can query and analyze log data in Log Service, you must enable the indexing feature and configure indexes. For more information, see Enable the index feature and configure indexes for a Logstore.
   - After you enable the indexing feature and configure indexes, you can query and analyze log data in real time. Log Service allows you to query and analyze large amounts of log data in real time. For more information, see Query and analysis.
   - After you query and analyze log data, you can configure charts and dashboards to display query and analysis results. For more information, see Overview and Dashboard overview.

6. Configure alert rules.

   Log Service allows you to configure alert rules based on query and analysis results. Log Service sends alert notifications based on the notification methods that you specify. For more information, see Configure alerts.

7. Consume logs.

   Log Service allows you to consume logs by using multiple methods, such as a Spark Streaming client, Storm spout, and Flink connector. For more information, see Real-time consumption.

23.1.2.2. Log on to the Log Service console

This topic describes how to log on to the Log Service console.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Log Service**.
6. On the page that appears, select an organization and region, and then click **Access as Administrator**. The home page of the Log Service console is displayed.

23.1.2.3. Obtain an AccessKey pair

An AccessKey pair consists of an AccessKey ID and an AccessKey secret. The AccessKey pair is used to implement symmetric encryption to verify the identity of the requester. The AccessKey ID is used to identify a user. The AccessKey secret is used to encrypt the signature string. This topic describes how to obtain an AccessKey pair.

Prerequisites

Only the operation administrators or level-1 organization administrators can obtain the AccessKey pair of an organization.

Context

To call Apsara Uni-manager and cloud service APIs, we recommend that you use the AccessKey pair of a personal account. If you use the AccessKey pair of a personal account, you must configure header parameters as described in the following table for access control.

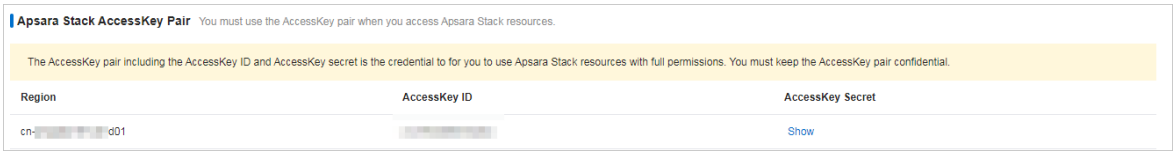
Parameter	Description
x-acs-regionid	The region ID, such as cn-hangzhou-*.
x-acs-organizationid	The ID of the organization in the Apsara Uni-manager Management Console.
x-acs-resourcegroupid	The ID of the resource set in the Apsara Uni-manager Management Console.
x-acs-instanceid	The ID of the instance on which you want to perform operations.

Warning The AccessKey pairs of personal accounts are under control of the Apsara Uni-manager permission system. AccessKey pairs of organization accounts have higher permissions. For security purposes, organization operations must be approved by administrators.

Obtain the AccessKey pair of a personal account

To obtain the AccessKey pair of a personal account, perform the following operations:

1. Log on to the Apsara Uni-manager Management Console.
2. In the upper-right corner of the homepage, move the pointer over the profile picture and click **Personal Information**.
3. In the **Apsara Stack AccessKey Pair** section, view your AccessKey pair.



Note The AccessKey pair consists of an AccessKey ID and an AccessKey secret. AccessKey pairs allow you to access Apsara Stack resources with full permissions for your account. You must keep your AccessKey pair confidential.

Obtain the AccessKey pair of an organization

To obtain the AccessKey pair of an organization, perform the following operations:

1. Log on to the Apsara Uni-manager Management Console as an administrator.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **Organizations**.
4. In the organization navigation tree, click an organization name.
5. In the Current Organization section, click **Management Accesskey**.
6. In the Management AccessKey, view the AccessKey pair of the organization.

> **Note** An AccessKey pair is automatically allocated to each level-1 organization. Subordinate organizations use the same AccessKey pair of their level-1 organization.

23.1.2.4. Manage a project

This topic describes how to create and delete a project in the Log Service console.

Context

A project in Log Service is the resource management unit that is used to separate and manage different resources. You can use a project to manage all logs and related log sources of an application. You can also use a project to manage Logstores and Logtail configurations. A project serves as an endpoint to access the resources of Log Service.

Projects provide the following features:

- Projects allow you to manage different Logstores. Logs that you collect and store in Log Service are generated from different projects, services, and environments. You can specify different projects for these logs to facilitate data consumption, exporting, and indexing. You can also grant permissions on these projects to different users.
- Projects serve as endpoints that allow authorized access to the resources of Log Service. Log Service allocates an exclusive endpoint to each project. You can use the endpoint to read, write, and manage the log data in the project.

Create a project.

> **Note**
> - You can create a project only by using the Log Service console.
> - You can create a maximum of 50 projects for each Apsara Stack tenant account.

1. Log on to the Log Service console.
2. In the Projects section, click **Create Project**.
3. In the Create Project panel, set the parameters. The following table describes the parameters.

Parameter	Description
Project Name	The name of the project. The name must be unique in a region. The name must meet the following requirements: ○ The name can contain only lowercase letters, digits, and hyphens (-). ○ The name must start and end with a lowercase letter or a digit. ○ The name must be 3 to 63 characters in length. **Note** The name cannot be modified after the project is created.
Description	The description of the project. After the project is created, the description is displayed in the **Projects** section. If you need to modify the description after the project is created, find the project in the **Projects** section, and click **Edit** in the Actions column. The description must be 0 to 64 characters in length and cannot contain the following characters: `<>'\"\\` .
Region	The region where the project resides. We recommend that you select a region based on the log source. After you create a project, you cannot modify the region where the project resides or migrate the project to another region.

Parameter	Description

4. Click **OK**.

Modify a project

To modify the description of a project, perform the following steps:

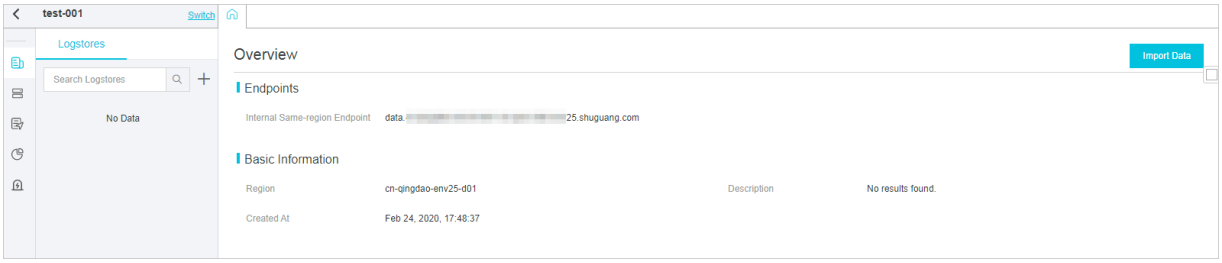
1. In the Projects section, find the project that you want to modify.
2. Click **Edit** in the **Actions** column.
3. In the **Modify Project** panel, modify the description of the project.

   > **Note** You cannot modify the project name or region.

4. Click **OK**.

View the information of a project

In the Projects section, click the name of the project that you want to view. On the Overview page, you can view the endpoint and region of the project.



Delete a project

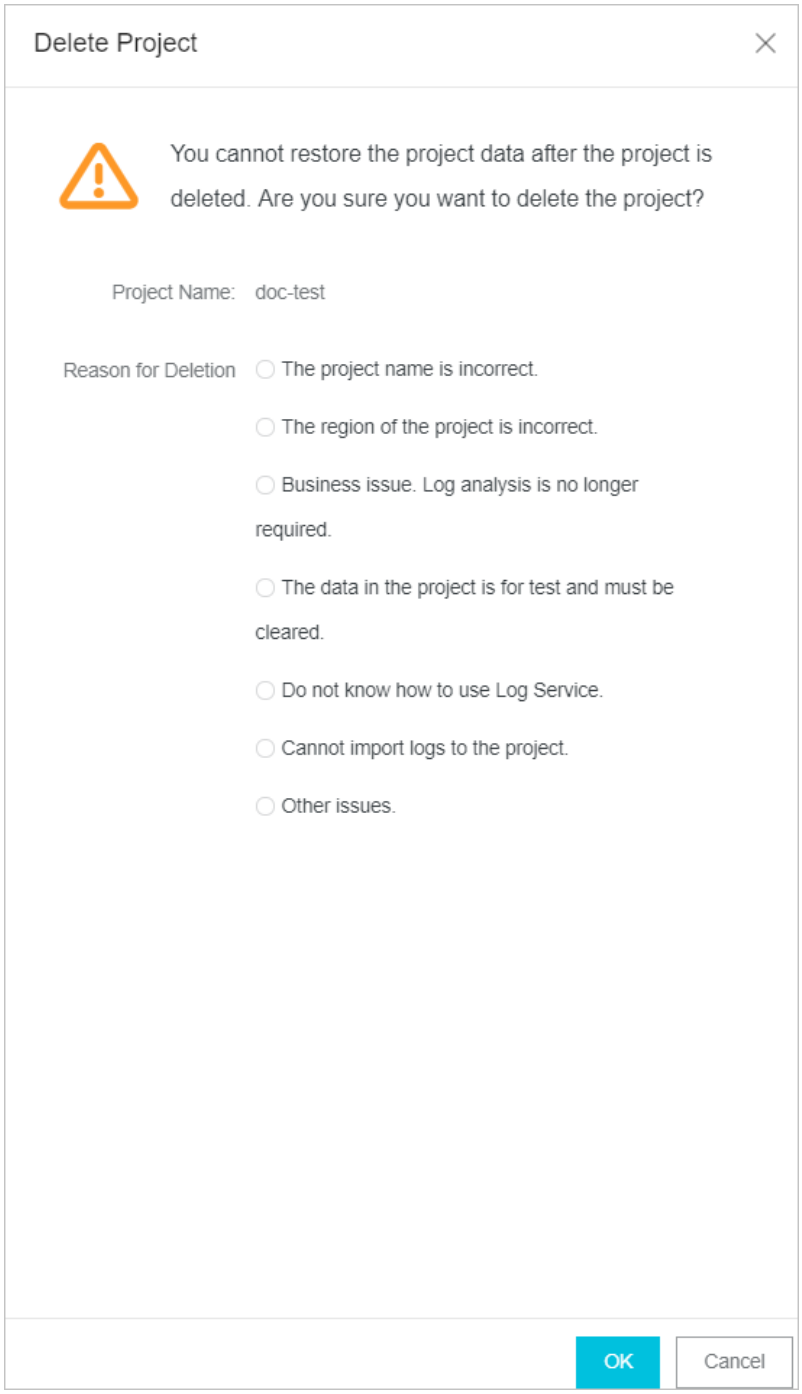
If you want to delete all logs from a project, you can delete the project. To delete a project, perform the following steps:

> **Warning** After you delete a project, all log data stored in the project and the configurations of the project are deleted and cannot be restored. Proceed with caution.

1. In the **Projects** section, find the project that you want to delete.
2. Click **Delete** in the **Actions** column.
3. Delete the project as prompted.

   If you select **Other issues**, enter the reason in the text box.

   Delete a project



23.1.2.5. Manage a Logstore

This topic describes how to create, modify, and delete a Logstore in the Log Service console.

Context

You can create multiple Logstores for a project. We recommend that you create a Logstore for each log type of an application.

Logstores provide the following features:

- Log collection: Logstores support real-time logging.
- Log storage: Logstores support real-time consumption.

- Index creation: Logstores support real-time log queries.

Create a Logstore

> **Note** You can create up to 200 Logstores for each project.

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to create a Logstore.
3. Choose **Log Storage > Logstores**. On the Logstores tab, click the Plus icon.
4. In the **Create Logstore** panel, set the parameters. The following table describes the parameters.

Parameter	Description
Logstore Name	The name of the Logstore. The name must be unique in the project to which the Logstore belongs. After the Logstore is created, you cannot modify the name. ○ The name can contain only lowercase letters, digits, hyphens (-), and underscores (_). ○ The name must start and end with a lowercase letter or a digit. ○ The name must be 3 to 63 characters in length.
WebTracking	If you turn on **WebTracking**, you can collect data from HTML, HTML5, iOS, or Android platforms and send the data to Log Service by using web tracking.
Permanent Storage	If you turn on **Permanent Storage**, Log Service permanently stores the collected logs. **Note** If you set the data retention period to 3650 by calling the related API operation, the data is permanently stored.
Data Retention Period	The retention period of the logs in the Logstore. Valid values: 1 to 3000. Unit: days. When the retention period of the logs in the Logstore expires, the logs are automatically deleted. If you do not turn on **Permanent Storage**, you must set the **Data Retention Period** parameter. **Note** If you shorten the data retention period, the new retention period takes effect after 1 hour. Then, Log Service deletes data based on the new retention period. The volume of data that is displayed on the **Storage Size(Log)** card on the homepage of the Log Service console is updated the next day. For example, if you modify the data retention period from five days to one day, Log Service deletes the logs of the previous four days after 1 hour.
Shards	Log Service provides shards to read and write data. Each shard supports a write speed of 5 MB/s and 500 write operations per second and a read speed of 10 MB/s and 100 read operations per second. You can create up to 10 shards for each Logstore. You can create up to 200 shards for each project.

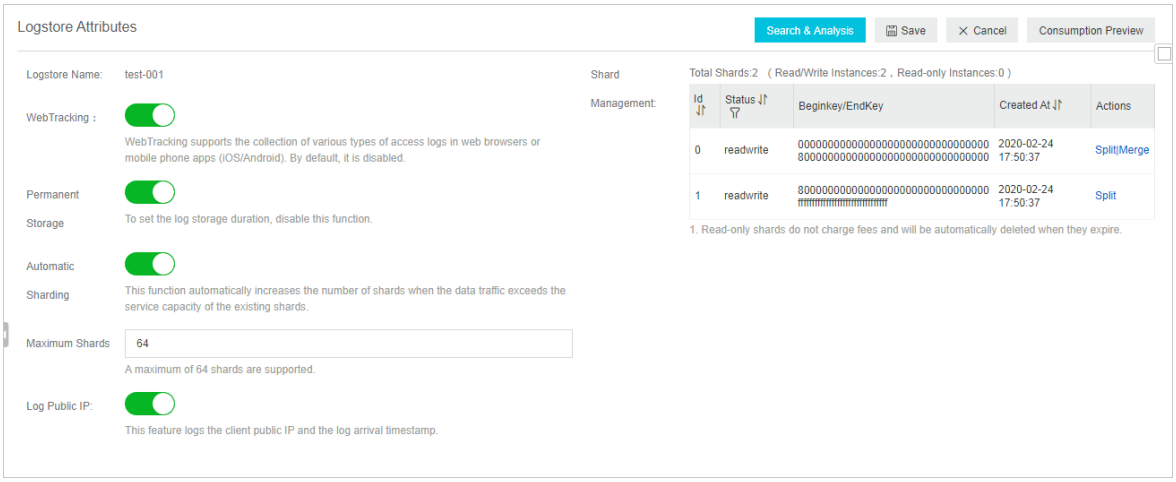
Parameter	Description
Automatic Sharding	Specifies whether to enable the automatic sharding feature. By default, this feature is enabled. If you turn on **Automatic Sharding** and the read or write speed does not meet your requirements, Log Service increases the number of shards. For more information, see Manage shards.
Maximum Shards	If you turn on **Automatic Sharding**, you must set the Maximum Shards parameter. The maximum number of shards is 64.
Log Public IP	If you turn on **Log Public IP**, Log Service adds the following information to the Tag field of the collected logs. ○ __client_ip__: the public IP address of the log source. ○ __receive_time__: the time when Log Service receives the log. The timestamp follows the UNIX time format. It is the number of seconds that have elapsed since 00:00:00 UTC, Thursday, January 1, 1970.

Modify the configurations of a Logstore

To modify the configurations of a Logstore, perform the following steps:

1. In the Projects section, click the name of the project in which you want to modify a Logstore.
2. In the left-side navigation pane, choose **Log Storage > Logstores**. On the Logstores tab, find the Logstore that you want to modify, click the icon, and then select **Modify**.
3. On the **Logstore Attributes** page, click **Modify**.

   You can modify the **Data Retention Period**, **WebTracking**, **Automatic Sharding**, **Maximum Shards**, and **Log Public IP** parameters. You can also split or merge the existing shards.



4. Click **Save**.

Delete a Logstore

To delete a Logstore, perform the following steps:

> **Notice**
> - Before you delete a Logstore, you must delete all related Logtail configurations.
> - If the log shipping feature is enabled for the Logstore, we recommend that you stop writing data to the Logstore before you delete the Logstore and make sure that all data in the Logstore is shipped.

1. In the left-side navigation pane, choose **Log Storage > Logstores**. On the Logstores tab, find the Logstore that you want to delete, click the icon, and then select **Delete**.

   Warning After you delete a Logstore, all log data in the Logstore is deleted and cannot be restored. Proceed with caution.

2. In the **Delete** dialog box, click **OK**.

23.1.2.6. Manage shards

This topic describes how to split, merge, and delete shards in the Log Service console. Logs are stored on shards in a Logstore. Each Logstore can have multiple shards. When you create a Logstore, you must specify the number of shards in the Logstore. After a Logstore is created, you can split or merge the shards.

Hash key

Log Service uses 128-bit MD5 hashes as the hash key of a Logstore. The entire MD5 hash range is [00000000000000000000000000000000,ffffffffffffffffffffffffffffffff). The hash key range of a Logstore falls within the entire MD5 hash range. When you create a Logstore, you must specify the number (N) of shards in the Logstore. The hash key range of the Logstore is evenly divided into N parts. Each part is assigned to a shard.

The hash key range of a shard is a left-closed and right-open interval that is specified by the following parameters:

- BeginKey: the start of the hash key range. The value of this parameter is included in the range.
- EndKey: the end of the hash key range. The value of this parameter is excluded from the range.

If you split a shard, the hash key range of the shard is evenly split. If you merge two shards, the hash key ranges of the shards are also merged. A hash key range determines the scope of a shard. When you push log data to a Logstore, you can specify a hash key for the log data. Log Service then writes the log data to the shard whose hash key range includes the specified hash key. This is called the hash key mode. If you do not specify a hash key for log data, the load balancing mode is used and Log Service writes the log data to a random available shard. However, when you pull log data from a Logstore, you must specify the shard where the log data is stored.

For example, a Logstore is divided into four shards and the hash key range of the Logstore is [00,FF). Example shards lists the hash key range of each shard.

Example shards

Shard	Hash key range
Shard0	[00,40)
Shard1	[40,80)
Shard2	[80,C0)
Shard3	[C0,FF)

If you set the hash key of log data to 5F, Log Service writes the log data to shard 1 because the hash key range of shard 1 includes 5F. If you set the hash key to 8C, the log data is written to shard 2 because the hash key range of shard 2 includes 8C.

Read/write capacity

Each shard provides an identical read/write capacity. Therefore, the read/write capacity of a Logstore depends on the number of shards in the Logstore. We recommend that you adjust the capacity of a Logstore based on the data traffic. For a Logstore, if the data traffic exceeds the read/write capacity, you can split shards to increase the Logstore capacity. If the data traffic is much less than the read/write capacity, you can merge shards to reduce the Logstore capacity and save costs.

For example, a Logstore consists of two read/write shards and the shards provide a maximum write capacity of 10 MB/s. If log data is written to the Logstore at a rate of 14 MB/s, we recommend that you split one of the shards into two shards. However, if log data is written at a rate of 3 MB/s, you can merge the two shards because the capacity of one shard already meets the read/write requirements.

> Note
> - If an API operation that writes data to a Logstore constantly returns 403 or 500 errors, you can check the data traffic metrics that are provided by Log Service and determine whether to split shards.
> - If the data traffic of a Logstore exceeds the read/write capacity of the Logstore, Log Service provides the best possible service but does not guarantee the service quality.

Shard status

A shard can be in one of the following states:

- Read/write
- Read-only

After a shard is created, the default status of the shard is read/write. If you split or merge shards, the status of the original shards changes to read-only and the new shards are in the read/write state. You can write data to and read data from a read/write shard. However, you can only read data from a read-only shard and cannot write data to the shard.

If you need to split a shard in a Logstore, you must specify the ID of the shard and an MD5 hash. The shard must be in the read/write state. The MD5 hash must be greater than the value of the BeginKey parameter of the shard and less than the value of the EndKey parameter of the shard. After the shard is split, the Logstore has two more shards. The status of the original shard changes from read/write to read-only. You can consume the log data in the original shard but cannot write log data to the shard. The new shards are in the read/write state and are listed below the original shard. The hash key ranges of the new shards cover that of the original shard.

If you need to merge shards in a Logstore, you must specify a read/write shard. The shard cannot be the last read/write shard in the shard list. Log Service finds the shard whose hash key range follows the hash key range of the specified shard, and merges the two shards into a new shard. The status of the original shards changes from read/write to read-only. You can consume the log data in the original shards but cannot write log data to the shards. The new shard is in the read/write state. The hash key range of the new shard covers those of the original shards.

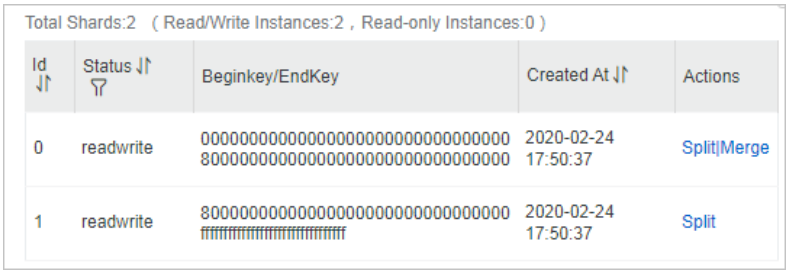
You can perform the following operations on shards in the Log Service console:

- Split a shard.
- Merge shards.

Split a shard

Each shard provides a write capacity of 5 MB/s and a read capacity of 10 MB/s. For a Logstore, if the data traffic exceeds the total read/write capacity of existing shards, we recommend that you split shards to increase the capacity.

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to create a Logstore.
3. In the left-side navigation pane, choose **Log Storage > Logstores**. On the Logstores tab, find the Logstore that you want to modify, click the icon, and then select **Modify**.
4. In the upper-right corner of the **Logstore Attributes** page, click **Modify**.
5. In the **Shard Management** section, find the shard that you want to split, and click **Split** in the **Actions** column.


Total Shards:2 (Read/Write Instances:2 , Read-only Instances:0)

Id	Status	Beginkey/EndKey	Created At	Actions
0	readwrite	00000000000000000000000000000000 80000000000000000000000000000000	2020-02-24 17:50:37	Split|Merge
1	readwrite	80000000000000000000000000000000 ffffffffffffffffffffffffffffffff	2020-02-24 17:50:37	Split

6. In the drop-down list, click the number of new shards.
7. In the **Actions** column, click **OK**.
8. In the upper-right corner of the **Logstore Attributes** page, click **Save**.

Enable the automatic sharding feature

Log Service provides the automatic sharding feature.

If you enable the automatic sharding feature for a Logstore, shards are automatically split if the following conditions are met:

- The data traffic of the Logstore exceeds the total capacity of the shards in the Logstore for more than five minutes.
- The number of read/write shards does not exceed the maximum number of shards that is allowed in the Logstore after an automatic shard split.

> **Note** After a shard is split into multiple shards, the new shards will not be automatically split within 15 minutes.

When you create or modify a Logstore, you can enable the automatic sharding feature. If this feature is enabled, you must specify the maximum number of shards that is allowed in the Logstore after an automatic shard split. For more information, see Manage Logstores.

- Automatic Sharding

  Specifies whether to enable the automatic sharding feature. If you turn on this switch, the shards are automatically split when the preceding conditions are met.

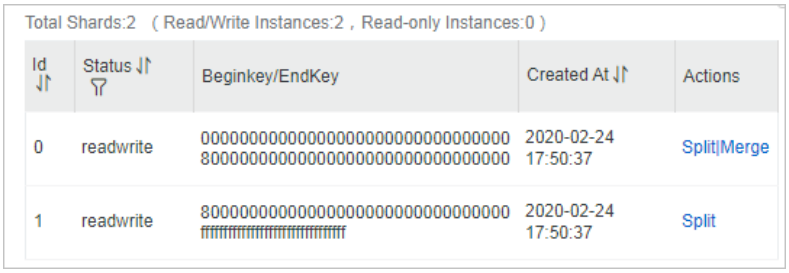
- Maximum Shards

  The maximum number of shards that is allowed in the Logstore after an automatic shard split. The maximum value is 64.

Merge shards

You can merge shards in a Logstore to reduce the capacity of the Logstore.

1. In the upper-right corner of the **Logstore Attributes** page, click **Modify**.
2. In the **Shard Management** section, find the shard that you want to merge with another shard, and click **Merge** in the **Actions** column.



3. In the upper-right corner of the **Logstore Attributes** page, click **Save**.

Delete a shard

- Automatic deletion

  When you create a Logstore, you can set a retention period for log data. When the retention period expires, the shards in the Logstore and the log data in the shards are automatically deleted.

- Manual deletion

  If you turn on the **Permanent Storage** switch when you create a Logstore, the shards in the Logstore and the log data in the shards are not automatically deleted. If you need to delete the shards and the log data in the shards, you can delete the Logstore. For more information, see Delete a Logstore.

23.1.3. Data collection

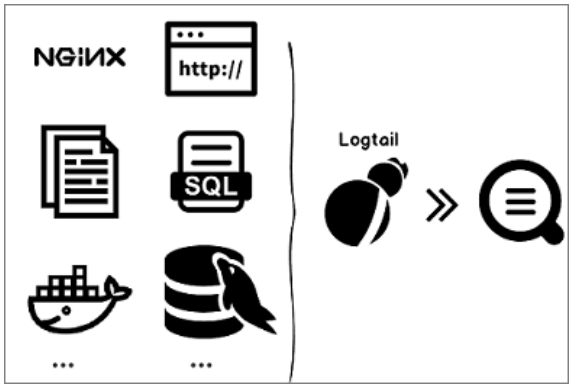
23.1.3.1. Collection by Logtail

23.1.3.1.1. Overview

23.1.3.1.1.1. Logtail overview

Logtail is a log collection agent that is provided by Log Service. You can use Logtail to collect logs from multiple data sources in real time. These sources include Elastic Compute Service (ECS) instances, data centers, and servers that belong to third-party cloud service providers. This topic describes the features, benefits, limits, and configuration process of Logtail.

Logtail-based log collection



Benefits

- Supports non-intrusive log collection based on log files. You do not need to modify your application code. Your applications are not affected when Logtail collects logs.
- Allows you to collect text logs, binary logs, HTTP data, and container logs.
- Allows you to collect logs from standard containers, swarm clusters, and Kubernetes clusters.
- Handles exceptions during log collection. If a network or server exception occurs, Logtail retries log collection

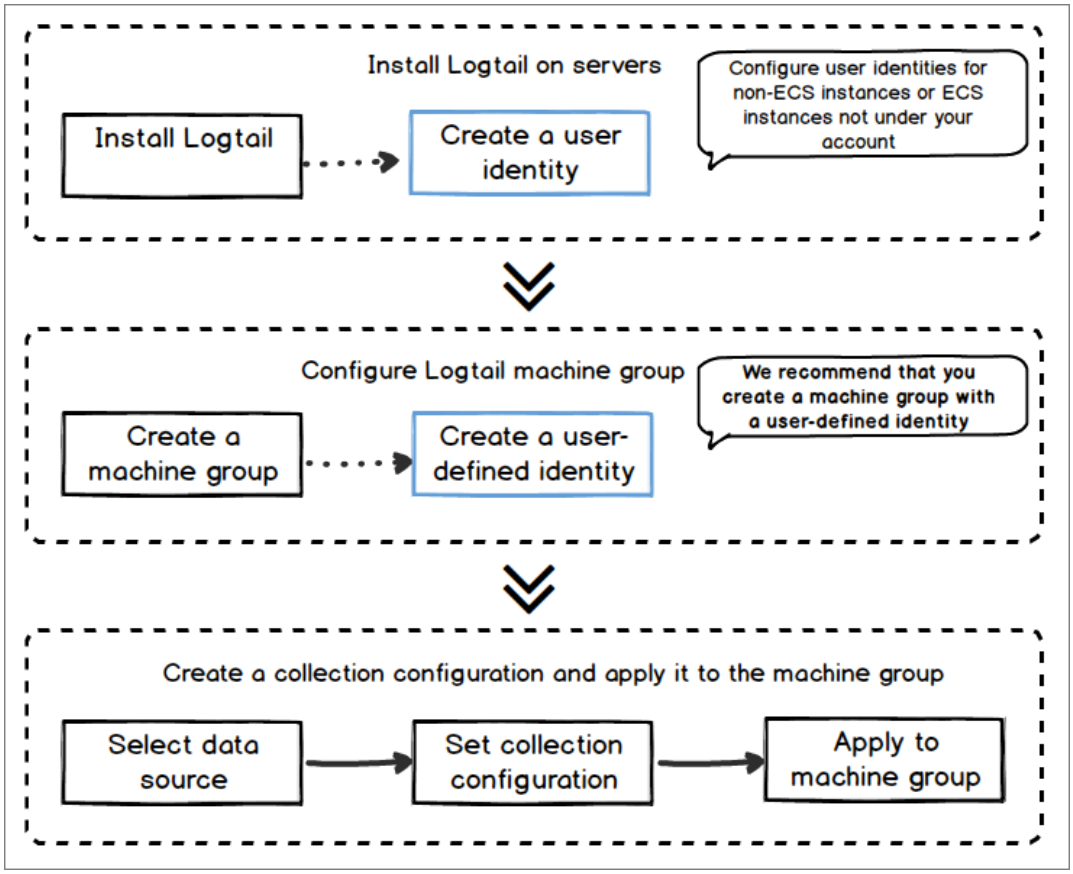
and caches logs on local servers to ensure data security.

- Provides centralized management based on Log Service. After you install Logtail on servers and create a machine group and Logtail configurations, Logtail collects logs from the servers and sends the logs to Log Service.
- Provides a comprehensive self-protection mechanism. The CPU, memory, and network resources that Logtail can use are limited. This ensures that Logtail does not affect the performance of other services on the server.

Limits

For more information about the limits of Logtail, see Limits.

Configuration process



To collect logs from servers by using Logtail, perform the following steps:

1. Install Logtail.

   Install Logtail on servers from which you want to collect logs. For more information, see Install Logtail in Linux and Install Logtail in Windows.

2. Create a machine group.

   Log Service allows you to create a custom ID-based machine group or an IP address-based machine group. For more information, see Create a machine group based on a server IP address and Create a machine group based on a custom ID.

3. Create a Logtail configuration and apply it to the machine group.

After you complete the preceding procedure, Logtail collects logs from your server and sends the logs to the specified Logstore. You can use the Log Service console, call API operations, or use SDKs to query logs.

Terms

- Machine group: A machine group contains one or more servers from which logs of a specific type are collected.

After you apply Logtail configurations to a machine group, Log Service collects logs from the servers in the machine group based on the configurations.

You can set an IP address-based identifier or a custom identifier for a machine group. Then, you can manage the servers in the machine group based on the identifier. You can create and delete a machine group, add servers to a machine group, and remove servers from a machine group in the Log Service console.

- Logtail: Logtail is a log collection agent that is provided by Log Service. Logtail runs on servers to collect logs from the servers. For more information, see Install Logtail in Linux or Install Logtail in Windows.
  - For a Linux-based server, Logtail is installed in the */usr/local/ilogtail* directory. Logtail initiates the following processes whose names start with ilogtail: a log collection process and a daemon process. The logs of Logtail are stored in the */usr/local/ilogtail/ilogtail.LOG* file.
  - For a Windows-based server, Logtail is installed in the *C:\Program Files\Alibaba\Logtail* directory (32-bit system) or *C:\Program Files (x86)\Alibaba\Logtail* directory (64-bit system). Choose **Control Panel > Administrative Tools > Services**. On the Services window, you can view the LogtailDaemon service. The logs of Logtail are stored in the `ilogtail.LOG` file.
- Logtail configurations: Logtail configurations are a set of policies that Logtail uses to collect logs. You can specify the data source and collection mode to create custom Logtail configurations for log collection. The configurations specify how to collect logs from servers, parse the logs, and send the logs to a specified Logstore.

Features

Feature	Description
Real-time log collection	Logtail monitors log files, and reads and parses incremental logs in real time. In most cases, logs are sent to Log Service within 3 seconds after they are generated. **Note** Logtail does not collect historical data. If Logtail reads a log later than 12 hours after the log was generated, Logtail drops the log.
Automatic log rotation	Multiple applications rotate log files based on the file size or date. The original log file is renamed and an empty log file is created during the rotation process. For example, the *app.LOG* file is renamed *app.LOG.1* and *app.LOG.2* during log rotation. You can specify the file to which collected logs are written, for example, *app.LOG*. Logtail monitors the log rotation process to ensure that no logs are lost.
Multiple data sources	Logtail can collect text logs, syslogs, HTTP logs, and MySQL binlogs.
Compatibility with open source collection agents	You can use open source agents such as Logstash and Beats to collect data. Then, you can use Logtail to collect data from the agents and send the data to Log Service.
Automatic exception handling	If data fails to be sent to Log Service due to exceptions, Logtail retries to collect logs based on the scenario. The exceptions include server errors, network errors, and quota exhaustion. If the retry fails, Logtail writes the data to the local cache and resends the data after 3 seconds.

Feature	Description
Flexible collection policy configuration	Logtail allows you to create configurations for log collection in a flexible manner. You can specify the directories and files from which logs are collected. You can also specify an exact match or a wildcard match based on your business requirements. You can also specify the log collection mode and customize the fields that you want to extract. You can use a regular expression to extract fields from logs. Log data in Log Service must have the timestamp information. Logtail allows you to customize log time formats and then extract the required timestamps from the time information based on different formats.
Automatic synchronization of Logtail configurations	After you create or update Logtail configurations in the Log Service console, the configurations are synchronized to the servers in which Logtail is installed and take effect within 3 minutes. Logs are collected based on the original configurations during the synchronization.
Status monitoring	Logtail monitors the CPU and memory resources that are consumed in real time. This ensures that Logtail does not consume an excessive number of resources or affect other services. If the resource consumption exceeds the limit, Logtail is automatically restarted. Logtail also monitors the network bandwidth resources that are consumed. This ensures that Logtail does not consume an excessive amount of bandwidth.
Data transmission with a signature	Logtail retrieves the AccessKey pair of your Apsara Stack tenant account and uses the pair to sign all log data that is sent to Log Service. This way, data tampering is prevented during data transmission. **Note** Logtail obtains the AccessKey pair of your Apsara Stack tenant account by using the HTTPS protocol to ensure the security of your AccessKey pair.

Data collection reliability

Logtail stores checkpoints that are periodically collected to the local server during log collection. If an exception such as an unexpected server shutdown or a process failure occurs, Logtail restarts and then collects data from the last checkpoint. This process avoids incomplete data collection. Logtail runs based on the startup parameters that are specified in the startup configuration file. If the usage of a resource exceeds the limit for more than 5 minutes, Logtail is forcibly restarted. After the restart, a small amount of duplicate data may be collected to the specified Logstore.

To improve log collection reliability, Logtail uses multiple internal mechanisms. However, logs may fail to be collected in the following scenarios:

- Logtail is not running, but logs are rotated multiple times.
- The log rotation rate is high, for example, one rotation per second.
- The log collection rate is lower than the log generation rate for a long period of time.

23.1.3.1.1.2. Log collection process of Logtail

This topic describes how Logtail collects logs. The log collection process consists of the following steps: monitor log files, read log files, process logs, filter logs, aggregate logs, and send logs.

Monitor log files

After you install Logtail on a server and create a Logtail configuration that is used to collect logs in the Log Service console, Log Service delivers the Logtail configuration to Logtail in real time. Then, Logtail monitors log files of the server based on the Logtail configuration. Logtail scans log directories and files based on the log file path and the maximum directory depth that you specify for monitoring in the Logtail configuration.

If the log files of the server in a machine group are not updated after you apply the Logtail configuration to the machine group, the log files are considered historical log files. Logtail does not collect historical log files. If log files are updated, Logtail reads and collects the files, and then sends the log files to Log Service. For more information about how to collect historical log files, see Import historical logs.

Logtail registers event listeners to monitor directories from which log files are collected. The event listeners pool the log files in the directories on a regular basis. This ensures that logs are collected at the earliest opportunity in a stable manner. For Linux-based servers, Inotify is used to monitor the directories and pool log files.

Read log files

After Logtail detects updated log files, Logtail reads the log files.

- The first time Logtail reads a log file, Logtail checks the size of the file.
  - If the file size is less than 1 MB, Logtail reads data from the beginning of the file.
  - If the file size is greater than 1 MB, Logtail reads from the last 1 MB of data in the file.
- If a log file is read before, Logtail reads the file from the previous checkpoint.
- Logtail can read up to 512 KB of data at the same time. Make sure that the size of each log in a log file does not exceed 512 KB.

> **Notice** If you change the system time on the server, you must restart Logtail. Otherwise, the log time becomes invalid and logs are dropped.

Process logs

After Logtail reads a log file, Logtail splits each log in the file into multiple lines, parses the log, and then configures the time field for the log.

- Split a log into multiple lines

  If you specify a regular expression to match the start part in the first line of a log, Logtail splits the log into multiple lines based on the regular expression. If you do not specify a regular expression, a single log line is processed as a log.

- Parse logs

  Logtail parses each log based on the collection mode that you specify in the Logtail configuration.

  > **Note** If you specify complex regular expressions, Logtail may consume an excessive amount of CPU resources. We recommend that you specify regular expressions that allow Logtail to parse logs in an efficient manner.

  If Logtail fails to parse a log, Logtail handles the failure based on the setting of the Drop Failed to Parse Logs parameter in the Logtail configuration.

  - If you turn on **Drop Failed to Parse Logs**, Logtail drops the log and reports an error.
  - If you turn off **Drop Failed to Parse Logs**, Logtail uploads the log. The key of the log is set to **raw_log** and the value is set to the log content.

- Configure the time field for a log
  - If you do not configure the time field for a log, the log time is set to the time when the log is parsed.

- If you configure the time field for a log, Logtail processes the log based on the following conditions:
  - If the difference between the time when the log is generated and the current time is within 12 hours, the log time is extracted from the parsed log fields.
  - If the difference between the time when the log is generated and the current time is greater than 12 hours, the log is dropped and an error is reported.

Filter logs

After logs are processed, Logtail filters the logs based on the specified filter conditions.

- If you do not specify filter conditions in the **Filter Configuration** field, the logs are not filtered.
- If you specify filter conditions in the **Filter Configuration** field, the fields in each log are traversed.

  Logtail collects only the logs that meet the filter conditions.

Aggregate logs

To reduce the number of network requests, Logtail caches the processed and filtered logs for a specified period of time. Then, Logtail aggregates the logs and sends the logs to Log Service.

If one of the following conditions is met during the cache process, logs are aggregated and sent to Log Service:

- The aggregation duration exceeds 3 seconds.
- The number of aggregated logs exceeds 4,096.
- The total size of aggregated logs exceeds 512 KB.

Send logs

Logtail sends the aggregated logs to Log Service. You can set the max_bytes_per_sec and send_request_concurrency parameters in the Logtail startup configuration file to specify the maximum transmission rate of log data and concurrent requests. For more information, see Set Logtail startup parameters.

If a log fails to be sent, Logtail retries or no longer sends the log based on the error code.

Error code	Description	Handling method
401	Logtail is not authorized to collect data.	Logtail drops the log packets.
404	The project or Logstore that is specified in the Logtail configuration does not exist.	Logtail drops the log packets.
403	The shard quota is exhausted.	Logtail tries again after 3 seconds.
500	A server exception occurs.	Logtail tries again after 3 seconds.

23.1.3.1.1.3. Logtail configuration files and record files

This topic describes the basic configuration files and record files of Logtail. When Logtail is active, Logtail uses the configuration files and generates record files.

Startup configuration file (ilogtail_config.json)

The *ilogtail_config.json* file is used to set the startup parameters of Logtail. For more information, see Set Logtail startup parameters.

> **Note**
> - The file must be a valid JSON file. Otherwise, Logtail cannot be started.
> - If you modify the file, you must restart Logtail for the modifications to take effect.

After you install Logtail on a server, you can perform the following operations on the *ilogtail_config.json* file:

- Modify the runtime parameters of Logtail.
- Check whether the installation commands are correct.

  The values of the config_server_address and data_server_list parameters in the *ilogtail_config.json* file vary based on the installation command that you select. If the region in the command is different from the region where the Log Service project resides or the address in the command cannot be accessed, the command is incorrect. If the command is incorrect, Logtail cannot collect logs and must be reinstalled.
- File path
  - Linux: The file is stored in */usr/local/ilogtail/ilogtail_config.json*.
  - Windows:
    - 64-bit: The file is stored in *C:\Program Files (x86)\Alibaba\Logtail\ilogtail_config.json*.
    - 32-bit: The file is stored in *C:\Program Files\Alibaba\Logtail\ilogtail_config.json*.

    > **Note** You can run both 32-bit and 64-bit applications in a 64-bit Windows operating system. To ensure compatibility, the operating system stores 32-bit applications in a separate x86 directory.
    >
    > Logtail for Windows is a 32-bit application. Therefore, Logtail is installed in the Program Files (x86) directory in 64-bit Windows. If Logtail for 64-bit Windows is available, you can install Logtail in the Program Files directory.
  - Containers: The file is stored in a Logtail container. The file path is specified in the environment variable *ALIYUN_LOGTAIL_USER_ID* of the Logtail container. You can run the `docker inspect ${logtail_container_name} | grep ALIYUN_LOGTAIL_CONFIG` command to view the file path. Example: */etc/ilogtail/conf/cn-hangzhou/ilogtail_config.json*.
- Sample file

  If you run the `cat /usr/local/ilogtail/ilogtail_config.json` command, the following output is returned:

```
{
    "config_server_address" : "http://logtail.cn-hangzhou-intranet.log.aliyuncs.com",
    "data_server_list" :
    [
        {
            "cluster" : "cn-hangzhou",
            "endpoint" : "cn-hangzhou-intranet.log.aliyuncs.com"
        }
    ],
    "cpu_usage_limit" : 0.4,
    "mem_usage_limit" : 100,
    "max_bytes_per_sec" : 2097152,
    "process_thread_count" : 1,
    "send_request_concurrency" : 4,
    "streamlog_open" : false
}
```

User identifier file

The user identifier file contains the ID of your Apsara Stack tenant account. The file specifies that the account is authorized to collect logs from the server on which Logtail is installed. For more information, see Configure an account ID on a server.

> **Note**
> - If the server is an Elastic Compute Service (ECS) instance that belongs to another Apsara Stack tenant account, a server that is deployed in a self-managed data center, or a server that is provided by a third-party cloud service provider, you must specify the ID of your Apsara Stack tenant account as a user identifier for the server after you install Logtail. Then, you can use Logtail to collect logs from the server by using the account.
> - You must specify the ID of an Apsara Stack tenant account as a user identifier in the user identifier file. You cannot specify the ID of a RAM user as a user identifier.
> - You must specify the name of the user identity file. You do not need to specify the file extension.
> - You can specify multiple user identifiers on a server. However, you can specify only one user identifier for a Logtail container.

- File path
  - Linux: The file is stored in */etc/ilogtail/users/*.
  - Windows: The file is stored in *C:\LogtailData\users*.
  - Containers: The file is stored in a Logtail container. The file path is specified in the environment variable ALIYUN_LOGTAIL_USER_ID of the Logtail container. You can run the `docker inspect ${logtail_container_name} | grep ALIYUN_LOGTAIL_USER_ID` command to view the file path.
- Sample file

  If you run the `ls /etc/ilogtail/users/` command, the following output is returned:

```
782392********** 37292*************
```

Custom identifier file (user_defined_id)

The *user_defined_id* file is used to configure a custom ID for a machine group. For more information, see Create a machine group based on a custom ID.

> **Note** When you create a custom ID-based machine group, you must configure the *user_defined_id* file.

- File path
  - Linux: The file is stored in */etc/ilogtail/user_defined_id*.
  - Windows: The file is stored in *C:\LogtailData\user_defined_id*.
  - Containers: The file is stored in a Logtail container. The file path is specified in the environment variable ALIYUN_LOGTAIL_USER_DEFINED_ID of the Logtail container. You can run the `docker inspect ${logtail_container_name} | grep ALIYUN_LOGTAIL_USER_DEFINED_ID` command to view the file path.
- Sample file

  If you run the `cat /etc/ilogtail/user_defined_id` command, the following output is returned:

```
aliyun-ecs-rs1e16355
```

Logtail configuration file (user_log_config.json)

The *user_log_config.json* file records the information of a Logtail configuration received by Logtail from Log Service. The file is in the JSON format and is updated along with configuration updates. You can use the *user_log_config.json* file to check whether the Logtail configuration is delivered to the server on which Logtail is installed. If the Logtail configuration file exists and the configurations in the file are the same as the settings of the Logtail configuration in Log Service, the Logtail configuration is delivered.

> **Note** We recommend that you do not modify the Logtail configuration file unless you need to specify sensitive information, such as the AccessKey pair and database password.

- File path
  - Linux: The file is stored in */usr/local/ilogtail/ilogtail_config.json*.
  - Windows
    - 64-bit: The file is stored in *C:\Program Files (x86)\Alibaba\Logtail\user_log_config.json*.
    - 32-bit: The file is stored in *C:\Program Files\Alibaba\Logtail\user_log_config.json*.
  - Containers: The file is stored in */usr/local/ilogtail/user_log_config.json*.
- Sample file

  If you run the `cat /usr/local/ilogtail/user_log_config.json` command, the following output is returned:

```
{
   "metrics" : {
      "##1.0##k8s-log-c12ba2028*****939f0b$app-java" : {
         "aliuid" : "16542189*****50",
         "category" : "app-java",
         "create_time" : 1534739165,
         "defaultEndpoint" : "cn-hangzhou-intranet.log.aliyuncs.com",
         "delay_alarm_bytes" : 0,
         "enable" : true,
         "enable_tag" : true,
         "filter_keys" : [],
         "filter_regs" : [],
         "group_topic" : "",
         "local_storage" : true,
         "log_type" : "plugin",
         "log_tz" : "",
         "max_send_rate" : -1,
         "merge_type" : "topic",
         "plugin" : {
            "inputs" : [
               {
                  "detail" : {
                     "IncludeEnv" : {
                        "aliyun_logs_app-java" : "stdout"
                     },
                     "IncludeLable" : {
                        "io.kubernetes.container.name" : "java-log-demo-2",
                        "io.kubernetes.pod.namespace" : "default"
                     },
                     "Stderr" : true,
                     "Stdout" : true
                  },
                  "type" : "service_docker_stdout"
               }
            ]
         },
         "priority" : 0,
         "project_name" : "k8s-log-c12ba2028c*****ac1286939f0b",
         "raw_log" : false,
         "region" : "cn-hangzhou",
         "send_rate_expire" : 0,
         "sensitive_keys" : [],
         "tz_adjust" : false,
         "version" : 1
      }
   }
}
```


AppInfo record file (app_info.json)

The *app_info.json* file records the information of Logtail, such as the startup time and the IP address and hostname obtained by Logtail.

If the IP address of a server is associated with the hostname in the */etc/hosts* file of the server, Logtail obtains the IP address. If you do not associate the IP address of a server with the hostname, Logtail obtains the IP address of the first network interface controller (NIC) on the server.

> **Note**
> - The AppInfo record file records only the basic information of Logtail.
> - If you modify the hostname or other network settings of the server, you must restart Logtail to obtain a new IP address.

- File path
  - Linux: The file is stored in */usr/local/ilogtail/app_info.json*.
  - Windows
    - 64-bit: The file is stored in *C:\Program Files (x86)\Alibaba\Logtail\app_info.json*.
    - 32-bit: The file is stored in *C:\Program Files\Alibaba\Logtail\app_info.json*.
  - Containers: The file is stored in the */usr/local/ilogtail/app_info.json*.
- Sample file

  If you run the `cat /usr/local/ilogtail/app_info.json` command, the following output is returned:

```
{
   "UUID" : "",
   "hostname" : "logtail-ds-slpn8",
   "instance_id" : "E5F93BC6-B024-11E8-8831-0A58AC14039E_1**.***.***.***_1536053315",
   "ip" : "1**.***.***.***",
   "logtail_version" : "0.16.13",
   "os" : "Linux; 3.10.0-693.2.2.el7.x86_64; #1 SMP Tue Sep 12 22:26:13 UTC 2017; x86_64",
   "update_time" : "2018-09-04 09:28:36"
}
```

Field	Description
UUID	The serial number of the server.
hostname	The hostname.
instance_id	The unique identifier of Logtail. This identifier is randomly generated.
ip	The IP address that is obtained by Logtail. If Logtail does not obtain an IP address, the value of this parameter is null. Logtail cannot run as expected. You must specify an IP address for the server and then restart Logtail. **Note** If you create an IP address-based machine group, you must make sure that the IP address that you specify for the machine group is the same as the value of this field. If the two IP addressed are different, modify the IP address that you specify for the machine group in the Log Service console. Check the IP addresses again after 1 minute.
logtail_version	The version of Logtail.
os	The version of the operating system.
update_time	The last startup time of Logtail.

Logtail operational log file (ilogtail.LOG)

The *logtail_plugin.LOG* file records the operational logs of Logtail plug-ins. The levels of logs in ascending order include INFO, WARN, ERROR. You can ignore logs at the INFO level.

- File path
  - Linux: The file is stored in */usr/local/ilogtail/ilogtail.LOG*.
  - Windows
    - 64-bit: The file is stored in *C:\Program Files (x86)\Alibaba\Logtail\ilogtail.LOG*.
    - 32-bit: The file is stored in *C:\Program Files\Alibaba\Logtail\ilogtail.LOG*.
  - Containers: The file is stored in */usr/local/ilogtail/ilogtail.LOG*.
- Sample file

  If you run the `tail /usr/local/ilogtail/ilogtail.LOG` command, the following output is returned:

```
[2018-09-13 01:13:59.024679]    [INFO]    [3155]    [build/release64/sls/ilogtail/elogtail.cpp:123]
change working dir:/usr/local/ilogtail/
[2018-09-13 01:13:59.025443]    [INFO]    [3155]    [build/release64/sls/ilogtail/AppConfig.cpp:175
]    load logtail config file, path:/etc/ilogtail/conf/ap-southeast-2/ilogtail_config.json
[2018-09-13 01:13:59.025460]    [INFO]    [3155]    [build/release64/sls/ilogtail/AppConfig.cpp:176
]    load logtail config file, detail:{
   "config_server_address" : "http://logtail.ap-southeast-2-intranet.log.aliyuncs.com",
   "data_server_list" : [
      {
         "cluster" : "ap-southeast-2",
         "endpoint" : "ap-southeast-2-intranet.log.aliyuncs.com"
      }
]
```

Logtail plug-in log file (logtail_plugin.LOG)

The *logtail_plugin.LOG* file records the operational logs of Logtail plug-ins. The levels of logs in ascending order include INFO, WARN, ERROR. You can ignore logs at the INFO level.

If an exception such as **CANAL_RUNTIME_ALARM** occurs, you can troubleshoot the exception based on the *logtail_plugin.LOG* file.

- File path
  - Linux: The file is stored in */usr/local/ilogtail/logtail_plugin.LOG*.
  - Windows:
    - 64-bit: The file is stored in *C:\Program Files (x86)\Alibaba\Logtail\logtail_plugin.LOG*.
    - 32-bit: The file is stored in *C:\Program Files\Alibaba\Logtail\logtail_plugin.LOG*.
  - Containers: The file is stored in */usr/local/ilogtail/logtail_plugin.LOG*.
- Sample file

  If you run the `tail /usr/local/ilogtail/logtail_plugin.LOG` command, the following output is returned:

```
2018-09-13 02:55:30 [INF] [docker_center.go:525] [func1] docker fetch all:start
2018-09-13 02:55:30 [INF] [docker_center.go:529] [func1] docker fetch all:stop
2018-09-13 03:00:30 [INF] [docker_center.go:525] [func1] docker fetch all:start
2018-09-13 03:00:30 [INF] [docker_center.go:529] [func1] docker fetch all:stop
2018-09-13 03:03:26 [INF] [log_file_reader.go:221] [ReadOpen] [##1.0##sls-zc-test-hz-pub$docker-stdout-config,k8s-stdout]    open file for read, file:/logtail_host/var/lib/docker/containers/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624-json.log    offset:40379573    status:794354-64769-40379963
2018-09-13 03:03:26 [INF] [log_file_reader.go:221] [ReadOpen] [##1.0##k8s-log-c12ba2028cfb444238cd9ac1286939f0b$docker-stdout-config,k8s-stdout]    open file for read, file:/logtail_host/var/lib/docker/containers/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624-json.log    offset:40379573    status:794354-64769-40379963
2018-09-13 03:04:26 [INF] [log_file_reader.go:308] [CloseFile] [##1.0##sls-zc-test-hz-pub$docker-stdout-config,k8s-stdout]    close file, reason:no read timeout    file:/logtail_host/var/lib/docker/containers/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624-json.log    offset:40379963    status:794354-64769-40379963
2018-09-13 03:04:27 [INF] [log_file_reader.go:308] [CloseFile] [##1.0##k8s-log-c12ba2028cfb444238cd9ac1286939f0b$docker-stdout-config,k8s-stdout]    close file, reason:no read timeout    file:/logtail_host/var/lib/docker/containers/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624/7f46afec6a14de39b59ee9cdfbfa8a70c2fa26f1148b2e2f31bd3410f5b2d624-json.log    offset:40379963    status:794354-64769-40379963
2018-09-13 03:05:30 [INF] [docker_center.go:525] [func1] docker fetch all:start
2018-09-13 03:05:30 [INF] [docker_center.go:529] [func1] docker fetch all:stop
```

Container path mapping file (docker_path_config.json)

The *docker_path_config.json* file is created only when you collect container logs. The file records the path mappings between container log files and host log files. The file is in the JSON format.

If the **DOCKER_FILE_MAPPING_ALARM** message appears when you troubleshoot a log collection exception, Docker files fail to be mapped to host files. You can use the *docker_path_config.json* file to troubleshoot the exception.

> **Note** This file is an information record file. Modifications to this file do not take effect. If you delete this file, another file is automatically created without service interruptions.

- File path

  /usr/local/ilogtail/docker_path_config.json

- Sample file

  If you run the `cat /usr/local/ilogtail/docker_path_config.json` command, the following output is returned:

```
{
   "detail" : [
      {
         "config_name" : "##1.0##k8s-log-c12ba2028cfb444238cd9ac1286939f0b$nginx",
         "container_id" : "df19c06e854a0725ea7fca7e0378b0450f7bd3122f94fe3e754d8483fd330d10",
         "params" : "{\n   \"ID\" : \"df19c06e854a0725ea7fca7e0378b0450f7bd3122f94fe3e754d8483fd330
d10\",\n   \"Path\" : \"/logtail_host/var/lib/docker/overlay2/947db346695a1f65e63e582ecfd10ae1f5701
9a1b99260b6c83d00fcd1892874/diff/var/log\",\n   \"Tags\" : [\n      \"nginx-type\",\n      \"access
-log\",\n      \"_image_name_\",\n      \"registry.cn-hangzhou.aliyuncs.com/log-service/docker-log-
test:latest\",\n      \"_container_name_\",\n      \"nginx-log-demo\",\n      \"_pod_name_\",\n
\"nginx-log-demo-h2lzc\",\n      \"_namespace_\",\n      \"default\",\n      \"_pod_uid_\",\n
\"87e56ac3-b65b-11e8-b172-00163f008685\",\n      \"_container_ip_\",\n      \"172.20.4.224\",\n
\"purpose\",\n      \"test\"\n   ]\n}\n"
      }
   ],
   "version" : "0.1.0"
}
```


23.1.3.1.2. Installation

23.1.3.1.2.1. Install Logtail on a Linux server

This topic describes how to install, upgrade, and uninstall Logtail on a Linux server.

Supported operating systems

You can install Logtail on servers that run one of the following x86-64 Linux operating systems:

- Aliyun Linux 2
- Red Hat Enterprise Linux 6, Red Hat Enterprise Linux 7, and Red Hat Enterprise Linux 8
- CentOS Linux 6, CentOS Linux 7, and CentOS Linux 8
- Debian GNU/Linux 8, Debian GNU/Linux 9, and Debian GNU/Linux 10
- Ubuntu 14.04, Ubuntu 16.04, Ubuntu 18.04, and Ubuntu 20.04
- SUSE Linux Enterprise Server 11, SUSE Linux Enterprise Server 12, and SUSE Linux Enterprise Server 15
- openSUSE Leap 15.1, openSUSE Leap 15.2, and openSUSE Leap 42.3
- Glibc 2.5 or later

Procedure

> **Note** If you run the installation command on a server on which Logtail is installed, the installer uninstalls Logtail from the server, deletes the `/usr/local/ilogtail` directory, and then reinstalls Logtail. If the installation is successful, Logtail starts when the system reboots.

1. Run the following command to download the Logtail installer:

   ```
   wget http://${service:sls-backend-server:sls_data.endpoint}/logtail.sh -O logtail.sh; chmod 755 lo
   gtail.sh
   ```

   > **Note** You must replace `${service:sls-backend-server:sls_data.endpoint}` in the command with the actual endpoint. You can view the endpoint information on the Overview page of a Logstore.

2. Run the installation command.

   Start Linux PowerShell and run the following command as an administrator to install Logtail:

```
./logtail.sh install
```

3. Configure a user identifier.

View the version of Logtail

Go to the installation directory and open the `/usr/local/ilogtail/app_info.json` file. The value of the `logtail_version` field indicates the version of Logtail. Run the `cat /usr/local/ilogtail/app_info.json` command to view the version of Logtail. The following example shows a response:

```
{
   "UUID" : "0DF18E97-0F2D-486F-B77F-*********",
   "hostname" : "david*******",
   "instance_id" : "F4FAFADA-F1D7-11E7-846C-00163E30349E_*********_1515129548",
   "ip" : "**********",
   "logtail_version" : "0.16.0",
   "os" : "Linux; 2.6.32-220.23.2.ali1113.el5.x86_64; #1 SMP Thu Jul 4 20:09:15 CST 2013; x86_64",
   "update_time" : "2018-01-05 13:19:08"
}
```

Upgrade Logtail

You can use the Logtail installation script logtail.sh to upgrade Logtail. The installation script selects an upgrade method based on the configurations of the Logtail that is installed.

Note During the upgrade, Logtail is temporarily stopped, and all files, except the configuration files and checkpoint files, are overwritten.

Run the following command to upgrade Logtail:

```
# Download the Logtail installer.
wget http://${service:sls-backend-server:sls_data.endpoint}/logtail.sh -O logtail.sh;
chmod 755 logtail.sh
# Upgrade Logtail.
sudo ./logtail.sh upgrade
```

Check the upgrade result:

```
# The upgrade is successful.
Stop logtail successfully.
ilogtail is running
Upgrade logtail success
{
   "UUID" : "***",
   "hostname" : "***",
   "instance_id" : "***",
   "ip" : "***",
   "logtail_version" : "0.16.11",
   "os" : "Linux; 3.10.0-693.2.2.el7.x86_64; #1 SMP Tue Sep 12 22:26:13 UTC 2017; x86_64",
   "update_time" : "2018-08-29 15:01:36"
}
# The upgrade fails because the current version is the latest version.
[Error]:    Already up to date.
```

Start and stop Logtail

- Start Logtail

Run the following command as an administrator to start Logtail:

```
/etc/init.d/ilogtaild start
```

- Stop Logtail

  Run the following command as an administrator to stop Logtail:

```
/etc/init.d/ilogtaild stop
```

Uninstall Logtail

Run the following command as an administrator to uninstall Logtail in Linux PowerShell:

```
wget http://${service:sls-backend-server:sls_data.endpoint}/logtail.sh -O logtail.sh
chmod 755 logtail.sh
./logtail.sh uninstall
```

23.1.3.1.2.2. Install Logtail in Windows

This topic describes how to install Logtail on a Windows server.

Supported systems

Logtail supports the following Windows operating systems:

- Windows 7 (Client) 32-bit
- Windows 7 (Client) 64-bit
- Windows Server 2008 32-bit
- Windows Server 2008 64-bit
- Windows Server 2012 64-bit
- Windows Server 2016 64-bit

Procedure

1. Download the installation package.

   Run the following command to download the installation package:

```
wget http://${service:sls-backend-server:sls_data.endpoint}/windows/logtail_installer.zip
```

   > **Note** You must replace `${service:sls-backend-server:sls_data.endpoint}` in the command with the actual endpoint. For more information about endpoints, see View the information of a project.

2. Decompress the `logtail_installer.zip` package to the current directory.
3. Run the installation command.

   Run Windows PowerShell or Command Prompt as an administrator. Enter the `logtail_installer` directory, and then run the installation command based on the network type.

```
.\logtail_installer.exe install me-east-1
```

   > **Note** You must replace `${region}` in the command with the actual endpoint. For more information about endpoints, see View the information of a project.

4. Configure an account ID for a server.

Installation directory

After you run the installation command, Logtail is installed in the specified directory. The directory cannot be changed. In the directory, you can View the version of Logtail in the *app_info.json* file or Uninstall Logtail.

The installation directory is as follows:

- 32-bit Windows: *C:\Program Files\Alibaba\Logtail*
- 64-bit Windows: *C:\Program Files (x86)\Alibaba\Logtail*

> **Note** You can run 32-bit or 64-bit applications in a 64-bit Windows operating system. However, the operating system stores 32-bit applications in separate x86 folders to ensure compatibility.
>
> Logtail for Windows is a 32-bit application. Therefore, it is installed in the *Program Files (x86)* folder in 64-bit Windows. If Logtail for 64-bit Windows becomes available in the future, it will be installed in the *Program Files* folder.

View the version of Logtail

To view the version of Logtail, go to the default installation directory, and then use the notepad or another text editor to open the *app_info.json* file. The `logtail_version` field shows the version of Logtail.

In the following example, the version of Logtail is 1.0.0.0:

```
{
    "logtail_version" : "1.0.0.0"
}
```

Upgrade Logtail

- Automatic upgrade

  Logtail later than 1.0.0.0 is automatically upgraded in Windows.

- Manual upgrade

  Logtail earlier than 1.0.0.0 must be manually upgraded. The manual upgrade procedure is the same as the installation procedure.

  > **Note** During a manual upgrade, the files in the original installation directory are deleted. We recommend that you back up the files before you perform a manual upgrade.

Start and stop Logtail

Open the **Control Panel**, choose System and Security > **Administrative Tools**, and then double-click **Services**.

Find the service based on your Logtail version.

- Logtail 0.x.x.x: LogtailWorker.
- Logtail 1.0.0.0 and later: LogtailDaemon.

Perform the following operations as required:

- Start Logtail

  Right-click the service and select **Start** from the shortcut menu.

- Stop Logtail

  Right-click the service and select **Stop** from the shortcut menu.

- Restart Logtail

  Right-click the service and select **Restart** from the shortcut menu.

Uninstall Logtail

Run Windows PowerShell or Command Prompt as an administrator. Enter the `logtail_installer` directory, and then run the following command:

```
.\logtail_installer.exe uninstall
```

After Logtail is uninstalled, the installation directory is deleted. However, some residual configuration data is still maintained in the *C:\LogtailData* directory. You can manually delete the data. The residual configuration data includes the following information:

- *checkpoint*: checkpoints of all plug-ins, for example, the Windows event log plug-in.
- *logtail_check_point*: checkpoints of Logtail.
- *users*: IDs of Apsara Stack tenant accounts.

23.1.3.1.2.3. Set the startup parameters of Logtail

Log Service limits the collection performance of Logtail. This way, Logtail does not consume an excessive number of server resources and affect other services. If you want to improve the collection performance of Logtail, modify the startup parameters of Logtail.

Scenarios

You can modify the startup parameters of Logtail in the following scenarios:

- You want to collect a large number of log files and the log files occupy a large amount of memory. For example, you want to collect more than 100 files or monitor more than 5,000 files in specific directories at the same time.
- Log data is transmitted at a high speed, which causes high CPU utilization. For example, Logtail collects log data at a speed higher than 2 MB/s in simple mode and at a speed higher than 1 MB/s in full regex mode.
- Logtail sends data to Log Service at a speed higher than 20 MB/s.

Set the startup parameters of Logtail

1. On the server on which Logtail is installed, open the */usr/local/ilogtail/ilogtail_config.json* file.
2. Set the startup parameters of Logtail based on your business requirements.

   The following example shows how to configure startup parameters:

   ```
   {
       ...
       "cpu_usage_limit" : 0.4,
       "mem_usage_limit" : 100,
       "max_bytes_per_sec" : 2097152,
       "process_thread_count" : 1,
       "send_request_concurrency" : 4,
       "buffer_file_num" : 25,
       "buffer_file_size" : 20971520,
       "buffer_file_path" : "",
       ...
   }
   ```

   Notice

   - The following table describes the startup parameters that are commonly used for Logtail. You can use the default values for other startup parameters.
   - You can add and modify startup parameters based on your business requirements.

   The following table describes the startup parameters of Logtail.

Startup parameters of Logtail

Parameter	Type	Description	Example
cpu_usage_limit	double	The CPU utilization threshold for Logtail. The threshold is specified based on a single-core CPU. ○ Valid values: 0.1 to the number of CPU cores of the current server. ○ Default value: 2. For example, if you set the parameter to 0.4, Logtail automatically restarts if the CPU utilization remains higher than 40% for 5 minutes. In most cases, a single core can collect logs at a speed of 24 MB/s in simple mode and 12 MB/s in full regex mode.	`"cpu_usage_limit" : 0.4`
mem_usage_limit	int	The memory usage threshold for Logtail. Unit: MB. ○ Valid values: 128 to the available memory size of the current server. ○ Default value: 2048. For example, if you set the parameter to 100, the memory usage is limited to 100 MB. Logtail restarts when the memory usage exceeds 100 MB. If you want to collect more than 1,000 log files at the same time, you can increase the value of this parameter.	`"mem_usage_limit" : 1000`
max_bytes_per_sec	int	The maximum speed at which Logtail sends raw data. Unit: bytes per second. ○ Valid values: 1024 to 52428800. ○ Default value: 20971520. For example, if you set this parameter to 2097152, the speed is limited to 2 MB/s. **Notice** If you set this parameter to a value that is greater than 20971520, the speed is not limited. The value 20971520 indicates that the speed is 20 MB/s.	`"max_bytes_per_sec" : 2097152`
process_thread_count	int	The number of threads that are used by Logtail to process data. ○ Valid values: 1 to 64. ○ Default value: 1. In most cases, a thread provides a write speed of 24 MB/s in simple mode and 12 MB/s in full regex mode. We recommend that you use the default value of this parameter.	`"process_thread_count" : 1`

Parameter	Type	Description	Example
send_request_concurrency	int	The number of concurrent requests that are sent by Logtail to asynchronously send data. ○ Valid values: 1 to 1000. ○ Default value: 20. If write transactions per second (TPS) is high, you can set this parameter based on your business requirements. Each concurrent request supports a network throughput of 0.5 MB/s to 1 MB/s. The number of concurrent requests varies based on the network latency. **Note** If the value of this parameter is excessively large, concurrent requests may occupy an excessive number of network ports. In this case, you must adjust the TCP parameters.	`"send_request_concurrency" : 4`
buffer_file_num	int	The maximum number of files that can be cached. ○ Valid values: 1 to 100. ○ Default value: 25. If a network error occurs or the data that needs to be written exceeds the threshold that you specify, Logtail caches parsed logs to the on-premises files in the installation directory. After the network is recovered and data that needs to be written does not exceed the threshold, Logtail retries to send the cached logs.	`"buffer_file_num" : 25`
buffer_file_size	int	The maximum size of a cached file. Unit: bytes. ○ Valid values: 1048576 to 104857600. ○ Default value: 20971520. The maximum disk space that can be occupied by cached files is calculated by multiplying the value of the buffer_file_size parameter by the value of the buffer_file_num parameter.	`"buffer_file_size" : 20971520`
buffer_file_path	String	The directory in which you want to store cached files. Default value: null. This value indicates that cached files are stored in the */usr/local/ilogtail* directory in which Logtail is installed. If you modify this parameter, you must move the cached files whose names match *logtail_buffer_file_ *to the directory that is specified by the new value of the parameter. This way, Logtail can read, send, and delete the cached files.	`"buffer_file_path" : ""`

Parameter	Type	Description	Example
bind_interface	String	The name of the Network Interface Card (NIC) that you want to associate with the server. Default value: null. This value indicates that an available NIC is automatically associated with the server. If you set this parameter to eth1, Logtail uses the NIC named eth1 to upload logs. **Note** This parameter is available only if Logtail runs on the Linux operating system.	`"bind_interface" : ""`
check_point_filename	String	The path in which the checkpoint file of Logtail is stored. Default value: */tmp/logtail_check_point*. We recommend that you modify the path, and then mount the path on the host as a Docker user. If you perform the operations as another type of user, Logtail collects duplicate files when a Docker container is released due to the loss of checkpoint information. For example, if you set the check_point_filename parameter in the Docker container to `/data/logtail/check_point.dat`, and then add `-v /data/docker1/logtail:/data/logtail` to the startup command, the */data/docker1/logtail* directory of the host is mounted on the */data/logtail* directory of the Docker container.	`"check_point_filename" : /tmp/logtail_check_point`
user_config_file_path	String	The path of the Logtail configuration. By default, Logtail configurations are stored in the directory in which the binary process is stored, and the file name is *user_log_config.json*. We recommend that you modify the path, and then mount the path on the host as a Docker user. If you perform the operations as another type of user, Logtail collects duplicate files when a Docker container is released due to the loss of checkpoint information. For example, if you set the user_config_file_path parameter in the Docker container to */data/logtail/user_log_config.json*, and then add `-v /data/docker1/logtail:/data/logtail` to the startup command, the */data/docker1/logtail* directory of the host is mounted to the */data/logtail* directory of the Docker container.	`"user_config_file_path" : user_log_config.json`
discard_old_data	Boolean	Specifies whether to drop historical logs. Default value: true. This value indicates that the logs that were generated 12 hours or more before the current time are dropped.	`"discard_old_data" : true`
working_ip	String	The server IP address that is reported by Logtail to Log Service. Default value: null. This value indicates that Log Service automatically obtains the server IP address.	`"working_ip" : ""`
working_hostname	String	The server hostname that is reported by Logtail to Log Service. Default value: null. This value indicates that Log Service automatically obtains the server hostname.	`"working_hostname" : ""`

Parameter	Type	Description	Example
max_read_buffer_size	long	The maximum size of each log. Default value: 524288. Unit: bytes. If the size of a log exceeds 524288 bytes or 512 KB, you can change the value of this parameter.	`"max_read_buffer_size" : 524288`
oas_connect_timeout	long	The timeout period of the connection that is established by Logtail to send a request to obtain the Logtail configuration or AccessKey pair. Default value: 5. Unit: seconds. If the network is unstable or the connection fails to be established, you can change the value of this parameter.	`"oas_connect_timeout" : 5`
oas_request_timeout	long	The timeout period of the request that is sent by Logtail to obtain the Logtail configuration or AccessKey pair. Default value: 10. Unit: seconds. If the network is unstable or the connection fails to be established, you can change the value of this parameter.	`"oas_request_timeout" : 10`
data_server_port	long	If you set the data_server_port parameter to true, Logtail transmits data to Log Service by using the HTTPS protocol. This parameter is available only for Logtail V1.0.10 or later.	`"data_server_port": 443`
enable_log_time_auto_adjust	Boolean	If you set the enable_log_time_auto_adjust parameter to true, the log time is automatically changed to the local time of the server. To ensure data security, Log Service checks the time information in requests, such as the requests sent by Logtail. Log Service rejects requests that are sent 15 minutes earlier or 15 minutes later than the time in Log Service. The point in time at which Logtail sends a request is considered as the local time of the server. In some test scenarios, the local time must be set to a future point in time. If you change the local time of the server, Log Service rejects the request and Logtail fails to write data to Log Service. You can synchronize the log time to the local time of the server based on this parameter. This parameter is available only for Logtail V1.0.19 or later. **Notice** ◦ If you enable the adaptive feature, the offset between the time in Log Service and the local time of the server are added to the log time. The offset is updated only if Log Service rejects a request. Therefore, the log time in the query result of Log Service may be different from the point in time when the log is written. ◦ Part of the logic of Logtail changes based on the increment of the system time. We recommend that you restart Logtail after you change the local time of the server.	`"enable_log_time_auto_adjust": true`

Parameter	Type	Description	Example
accept_multi_config	Boolean	Specifies whether to allow multiple Logtail clients to collect data from the same file. Default value: false. This value indicates that only one Logtail client can collect the data of a file. By default, only one Logtail client can collect the data of a file. You can set this parameter to true to allow multiple clients to collect data from the same file. Each Logtail client has an independent collection process. If multiple Logtail clients are allowed to collect data from the same file, the CPU utilization and the memory usage increase. This parameter is available only for Logtail V0.16.26 or later.	`"accept_multi_config": true`

3. Restart Logtail for the modified configurations to take effect.

```
/etc/init.d/ilogtaild stop && /etc/init.d/ilogtaild start
```

After you restart Logtail, you can run the `/etc/init.d/ilogtaild status` command to check the status of Logtail.

Appendix: environment variables

The following table describes the relationships between environment variables and the startup parameters of Logtail. For information about the startup parameters of Logtail, see Startup parameters of Logtail.

Relationships between environment variables and the startup parameters of Logtail

Parameter	Environment variable	Priority	Supported versions
cpu_usage_limit	cpu_usage_limit	If you use environment variables and configuration files to modify the startup parameters of Logtail, the modifications from the environment variables take effect.	Logtail V0.16.32 or later
mem_usage_limit	mem_usage_limit		
max_bytes_per_sec	max_bytes_per_sec		
process_thread_count	process_thread_count		
send_request_concurrency	send_request_concurrency		
check_point_filename	ALIYUN_LOGTAIL_CHECK_POINT_PATH	If you use environment variables and configuration files to modify the startup parameters of Logtail, the modifications from the environment variables take effect.	Logtail V0.16.36 or later

23.1.3.1.3. Logtail machine group

23.1.3.1.3.1. Overview

Log Service uses machine groups to manage the servers from which you want to collect logs by using Logtail.

A machine group is a virtual group that contains multiple servers. If you want to use a Logtail configuration file to collect logs from multiple servers, you can add the servers to a machine group. Then, you can apply the Logtail configuration file to the machine group.

To define a machine group, you can use one of the following methods:

- IP address: Add the IP addresses of all servers to a machine group. Each server can be identified by using its unique IP address.
- Custom ID: Use a custom ID to identify the machine group and use the same ID for servers in the machine group.

> **Note** Windows and Linux servers cannot be added to the same machine group.

IP address-based machine groups

You can add multiple servers to a machine group by adding their IP addresses to the machine group. Then, you can create a Logtail configuration file for all the servers at the same time.

- If you use ECS instances and have not associated them with hostnames or changed their network types, you can add their private IP addresses to the machine group.
- In other cases, you must add the server IP addresses obtained by Logtail to a machine group. The IP address of each server is recorded in the IP address field of the *app_info.json* file on the server.

> **Note** The *app_info.json* file records the internal information of Logtail. This file includes the server IP addresses obtained by Logtail. If you modify the IP address field of the file, the IP addresses obtained by Logtail remain unchanged.

Logtail obtains a server IP address by using the following methods:

- If the IP address of a server is associated with the hostname in the */etc/hosts* file of the server, Logtail obtains this IP address.
- If the IP address of a server is not associated with the hostname, Logtail obtains the IP address of the first network interface controller (NIC) on the server.

For more information, see Create an IP address-based server group.

Custom ID-based machine groups

You can use custom IDs to dynamically define machine groups.

An application system consists of multiple modules. You can scale out each module by adding multiple servers to the module. If you want to collect logs by module, you can create a machine group for each module. Therefore, you must specify a custom ID for each server in each module. For example, a website consists of an HTTP request processing module, a caching module, a logic processing module, and a storage module. The custom IDs of these modules can be http_module, cache_module, logic_module, and store_module.

For more information, see Create a machine group based on a custom ID.

23.1.3.1.3.2. Create an IP address-based machine group

Log Service allows you to create an IP address-based machine group. This topic describes how to create an IP address-based machine group in the Log Service console.

Prerequisites

- A project and a Logstore are created.
- At least one server is available.
  - If you use an Elastic Compute Service (ECS) instance, you must make sure that the ECS instance and Log Service project belong to the same Apsara Stack tenant account and reside in the same region.
  - If the ECS instance belongs to another Apsara Stack tenant account, you must configure a user identifier for the ECS instance before you create an IP address-based machine group. If the server is provided by a third-party cloud service provider or is deployed in a self-managed data center, you must also configure a user identifier. For more information, see Configure an account ID on a server.

- Logtail is installed on the server. For more information, see Install Logtail in Linux and Install Logtail in Windows.

Procedure

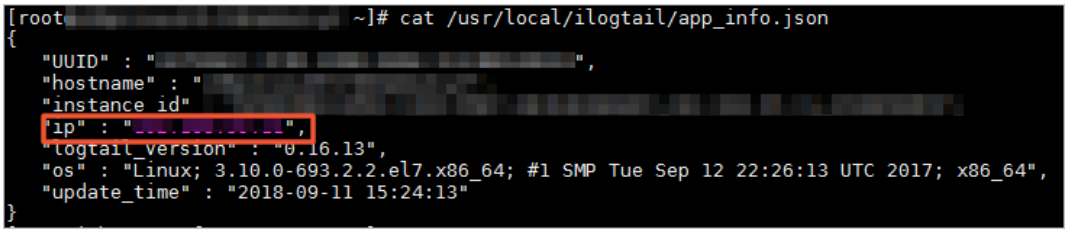
1. Obtain the IP addresses of the servers.

   The IP addresses that are obtained by Logtail are recorded in the ip field of the *app_info.json* file.

   On the servers where Logtail is installed, you can go to the following path to check the *app_info.json* file:

   - Linux: */usr/local/ilogtail/app_info.json*
   - 64-bit Windows: *C:\Program Files (x86)\Alibaba\Logtail\app_info.json*
   - 32-bit Windows: *C:\Program Files\Alibaba\Logtail\app_info.json*

   The following figure shows the IP address of the server in Linux.

2. Log on to the Log Service console.
3. In the Projects section, click the project in which you want to create a machine group.
4. In the left-side navigation pane, click the **Machine Groups** icon.
5. Click the icon next to Machine Groups and select **Create Machine Group**.

   You can also create a machine group in the Logtail configuration wizard.
6. In the **Create Machine Group** panel, set the parameters and click **OK**. The following table describes the parameters.

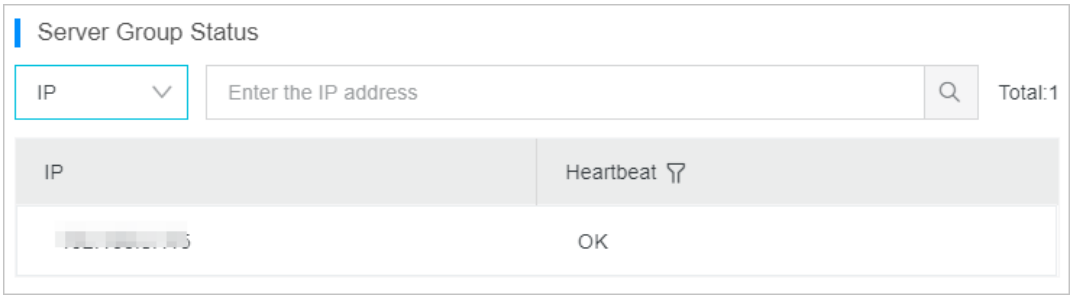
Parameter	Description
Name	The name of the machine group. The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Notice** After the machine group is created, you cannot modify its name. Proceed with caution.
Identifier	Select **IP Addresses**.
Topic	The topic of the machine group. This topic is used to differentiate log data that is generated on different servers. For more information, see Log topic.
IP address	Enter the IP addresses that are obtained in Step 1. **Note** - If a machine group contains multiple servers, you must separate the IP addresses with line feeds. - You cannot add Windows and Linux servers to the same machine group.

7. View the status of the machine group.

i. In the Machine Groups list, click the machine group that you create.

ii. On the **Machine Group Settings** page, view the status of the machine group.

If the **Heartbeat** status is **OK**, the server is connected to Logtail. If the status is **FAIL**, click **Automatic Retry**.



Result

You can view the created machine group in the Machine Groups list.

23.1.3.1.3.3. Create a custom ID-based machine group

Log Service allows you to create a custom ID-based machine group. This topic describes how to create a custom ID-based machine group in the Log Service console.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- At least one server is available.
  - If you use an Elastic Compute Service (ECS) instance, you must make sure that the ECS instance and Log Service project belong to the same Apsara Stack tenant account and reside in the same region. If the ECS instance belongs to another Apsara Stack tenant account, you must configure a user identifier for the ECS instance before you create a custom ID-based machine group. If the server is provided by a third-party cloud service provider or is deployed in a self-managed data center, you must also configure a user identifier. For more information, see Configure an account ID on a server.
  - Logtail is installed on the server. For more information, see Install Logtail on ECS instances.

Context

Custom ID-based machine groups offer significant benefits in the following scenarios:

- If your servers reside in multiple custom network environments such as virtual private clouds (VPCs), some IP addresses of the servers may conflict. In this case, Logtail cannot collect logs as expected. You can use a custom ID to prevent this issue.
- If you want to add multiple servers to a machine group, you can set the same custom ID for new servers as the machine group. Log Service identifies the custom ID and adds the servers with the same custom ID to the machine group.

Procedure

1. Create a file named *user_defined_id* in a specified directory.
   - Linux: Store the file in the */etc/ilogtail/user_defined_id* directory.
   - Windows: Store the file in the *C:\LogtailData\user_defined_ided_id* directory.

2. Set a custom ID for the server.

> **Note**
> - You cannot add Windows and Linux servers to the same machine group. You cannot set the same custom ID for Linux and Windows servers.
> - You can set one or more custom IDs for a single server and separate custom IDs with line feeds.
> - In the Linux server, if the */etc/ilogtail/* directory or the */etc/ilogtail/user_defined_id* file does not exist, you can create the directory and file. In the Windows server, if the *C:\LogtailData* directory or the *C:\LogtailData\user_defined_id* file does not exist, you can also create the directory and file.

   - Linux:

     Set the custom ID in the */etc/ilogtail/user_defined_id* file. For example, if you want to set the custom ID to `userdefined`, run the following command to edit the file. In the file, enter `userdefined`.

     ```
     vim /etc/ilogtail/user_defined_id
     ```

   - Windows:

     Set the custom ID in the *C:\LogtailData\user_defined_id* file. For example, if you want to set the custom ID to `userdefined_windows`, enter `userdefined_windows` in the *C:\LogtailData\user_defined_id* file.

3. Create a machine group.
   i. Log on to the Log Service console.
   ii. In the Projects section, click the name of the project in which you want to create a machine group.
   iii. In the left-side navigation pane, click the **Machine Groups** icon.
   iv. Click the icon next to **Machine Groups**, and then select **Create Machine Group**.
   v. In the Create Machine Group panel, set the parameters. The following table describes the parameters.

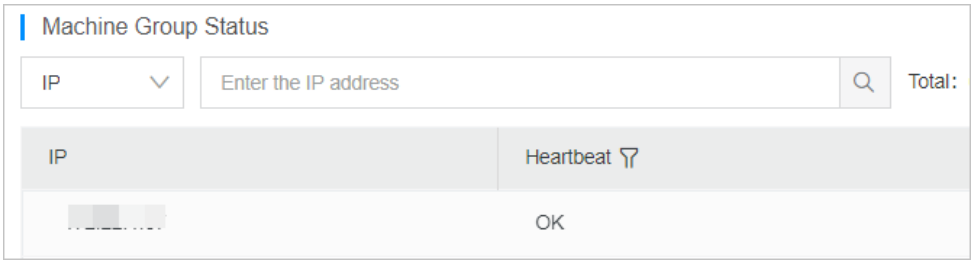
Parameter	Description
Name	The name of the machine group. The name must be 2 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Notice** After the machine group is created, you cannot modify its name. Proceed with caution.
Identifier	The identifier of the server. Select **Custom ID**.
Topic	The topic of the machine group. This topic is used to differentiate log data that is generated in different servers. For more information, see Log topic.
Custom Identifier	Enter the custom ID that is set in .

   vi. Click **OK**.

> **Note** If you need to add a server to the machine group, set a custom ID of the server to the custom ID of the machine group.

4. In the Machine Groups list, click the name of the machine group to view the status of the machine group.

   On the **Machine Group Settings** page, you can view the IP address list of the servers in the machine group and the heartbeat status in the **Machine Group Status** section.



- The Machine Group Status section lists the IP addresses of the servers whose custom ID is the same as the custom ID that you set for the machine group.

  For example, the custom ID of a machine group is userdefined and the IP addresses in the Machine Group Status section are 10.10.10.10, 10.10.10.11, and 10.10.10.12. This indicates that you specified the same custom ID for the servers in this machine group. If you want to add another server to the machine group and the IP address of the server is 10.10.10.13, set the custom ID to userdefined for the server. Then, you can view the IP address of the server that you added in the Machine Group Status section.
- If the Heartbeat status is OK, the server is connected to Log Service. If the status is FAIL, see What can I do if no heartbeat packet is received from a Logtail client?.

Disable a custom ID

If you want to set the Identifier parameter to IP Addresses, delete the user_defined_id file. The new configurations take effect within 1 minute.

- Linux:

```
rm -f /etc/ilogtail/user_defined_id
```

- In Windows, run the following command:

```
del C:\LogtailData\user_defined_id
```

Time to take effect

By default, after you add, delete, or modify the user_defined_id file, the new configurations take effect within 1 minute. If you want the configurations to immediately take effect, run the following command to restart Logtail.

- Linux:

```
/etc/init.d/ilogtaild stop
/etc/init.d/ilogtaild start
```

- Windows:
  i. Choose **Start Menu > Control Panel > Administrative Tools > Services**.
  ii. In the **Services** window, select the required service.
    - For Logtail V0.x.x.x, select LogtailWorker.
    - For Logtail V1.0.0.0 or later, select LogtailDaemon.
  iii. Right-click the service and then select **Restart** to validate the configurations.

23.1.3.1.3.4. View server groups

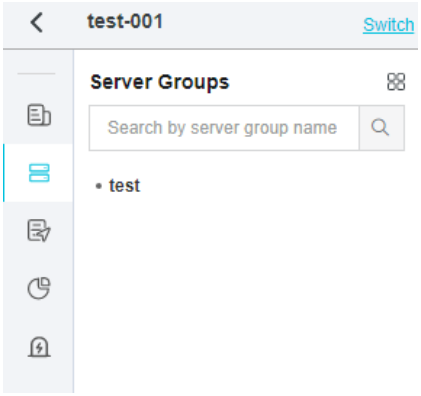
This topic describes how to view the server groups of a project on the **Server Groups** page in the Log Service console.

Procedure

1. Log on to the Log Service console.
2. Find the target project in the project list and click the project name.

3. In the left-side navigation pane of the page that appears, click the **Server Groups** icon to display the list of server groups.

   You can view all server groups of the project.



23.1.3.1.3.5. Modify a server group

This topic describes how to modify a server group in the Log Service console. After you create a server group, you can modify the parameters of the server group.

Procedure

1. Log on to the Log Service console.
2. Find the target project in the project list and click the project name.
3. In the left-side navigation pane of the page that appears, click the **Server Groups** icon to display the list of server groups.
4. Click the name of the server group to be modified. On the **Server Group Settings** page, click **Modify**.

   > **Note** The name of the server group cannot be modified.

5. Modify the parameters of the server group, and then click **Save**.

23.1.3.1.3.6. View the status of a server group

This topic describes how to view the status of a server group in the Log Service console. You can view the heartbeat information of Logtail to check whether Logtail is installed on the servers in a server group.

Procedure

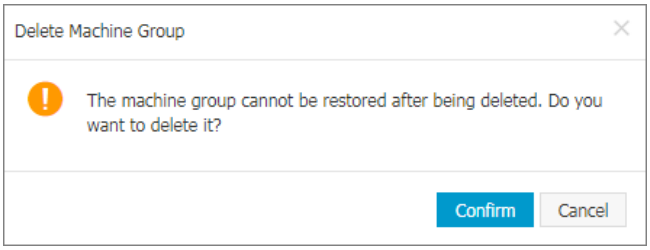
1. Log on to the Log Service console.
2. Find the target project in the project list and click the project name.
3. In the left-side navigation pane of the page that appears, click the **Server Groups** icon to display the list of server groups.
4. Click the name of the server group. On the **Server Group Settings** page, check the server group status.
   - If the heartbeat is OK, Logtail is installed on the servers in the server group and Logtail is connected to Log Service.
   - If the heartbeat status is FAIL, Logtail fails to connect to Log Service. If the FAIL state persists, perform troubleshooting based on the instructions provided in What can I do if no heartbeat packet is received from a Logtail client?

23.1.3.1.3.7. Delete a machine group

This topic describes how to delete a machine group in the Log Service console. You can delete a machine group if you no longer need to collect logs from the machine group.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to delete a machine group.
3. In the left-side navigation pane, click the icon. The Machine Groups list is displayed.
4. In the Machine Groups list, find the machine group that you want to delete, click the icon next to the machine group, and then select **Delete**.
5. In the message that appears, click **OK**.



23.1.3.1.3.8. Manage machine group configurations

This topic describes how to manage machine group configurations in the Log Service console. Log Service uses machine groups to manage the servers from which you collect logs by using Logtail. The servers can be Elastic Compute Service (ECS) instances.

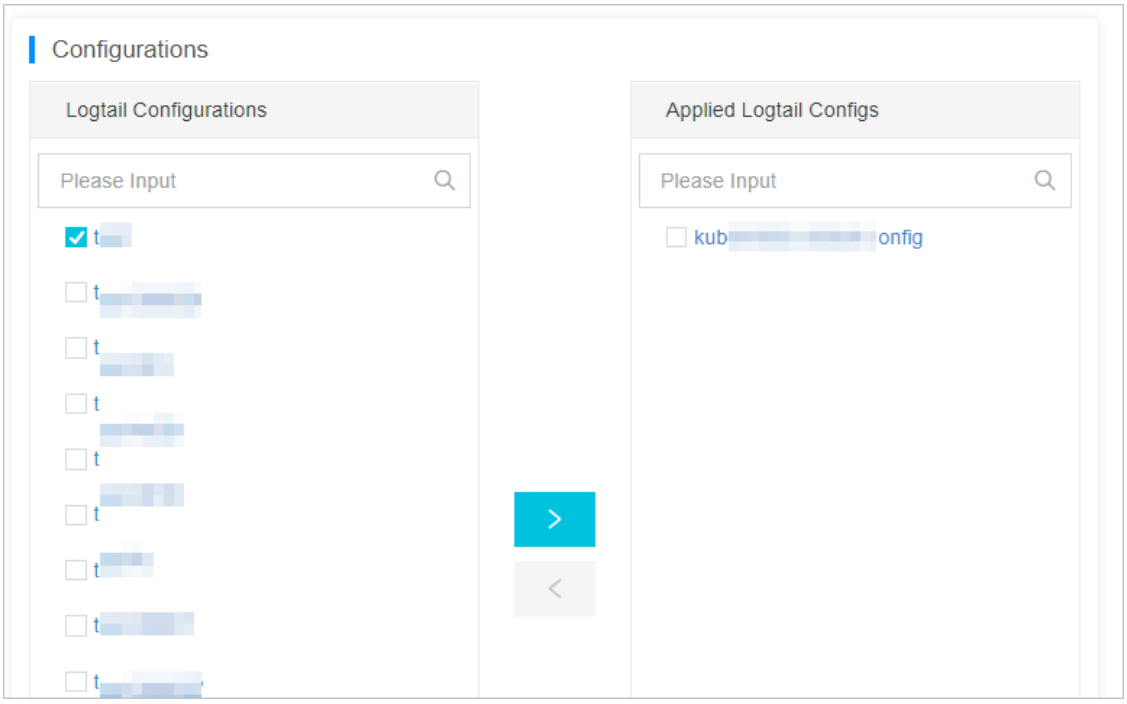
Context

Log Service allows you to manage the Logtail configurations that you create for Logtail installed on the servers in a machine group. You can apply or delete Logtail configurations to or from a machine group. The Logtail configurations specify what logs are collected on each server, how the logs are parsed, and which Logstore the logs are written to.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to manage a machine group.
3. In the left-side navigation pane, click the Machine Groups icon. The Machine Groups list is displayed.
4. In the Machine Groups list, click the machine group that you want to manage.
5. On the **Machine Group Settings** page, click **Modify**.
6. In the **Configurations** section, modify the Logtail configuration that you want to apply to the machine group and click **Save**.

   After you add a Logtail configuration, the Logtail configuration is delivered to Logtail on each server in the machine group. After you remove a Logtail configuration, the Logtail configuration is removed from Logtail.



23.1.3.1.3.9. Manage a Logtail configuration

This topic describes how to create, view, modify, and delete a Logtail configuration in the Log Service console.

View a list of Logtail configurations

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to view Logtail configurations.
3. Choose **Log Storage > Logstores**. On the Logstores tab, Click the **>** icon of the Logstore in which you want to view Logtail configurations. Then, choose **Data Import > Logtail Configurations**.
4. Click the Logtail configuration that you want to view.
5. On the **Logtail Config** page, view the details of the Logtail configuration.

Create a Logtail configuration

You can create a Logtail configuration in the Log Service console. For more information, see Configure text log collection.

Modify a Logtail configuration

Click the name of the Logtail configuration that you want to modify. On the **Logtail Config** page, click **Modify**.

You can also change the log collection mode of the Logtail configuration, and then apply the Logtail configuration to the related machine group again. The procedure to modify a Logtail configuration is the same as the procedure to create a Logtail configuration.

Delete a Logtail configuration

In the **Logtail Configurations** list, find the Logtail configuration that you want to delete, click the icon next to the Logtail configuration, and then select **Delete**.

Warning After you delete a Logtail configuration, the Logtail configuration is disassociated from the related machine group. Logtail no longer collects the logs that are specified by the Logtail configuration. Proceed with caution.

23.1.3.1.3.10. Configure a user identifier

This topic describes how to specify the ID of an Apsara Stack tenant account as a user identifier on a server.

Prerequisites

- A server is available.

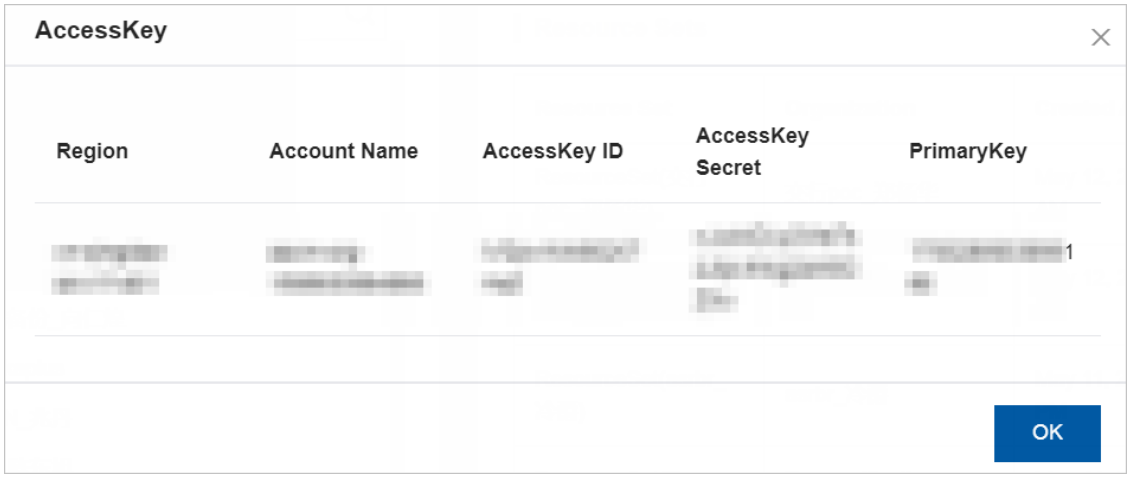
  The server can be an Elastic Compute Service (ECS) instance that belongs to another Apsara Stack tenant account, a server that is provided by a third-party cloud service provider, or a self-managed data center.
- Logtail is installed on the server from which you want to collect logs. For more information, see Install Logtail in Linux or Install Logtail in Windows.

Context

If your server is an ECS instance that belongs to another Apsara Stack tenant account, a server that is provided by a third-party cloud service provider, or a self-managed data center, you must specify the ID of your Apsara Stack tenant account as a user identifier on your server after Logtail is installed. Then, the Apsara Stack tenant account can use Logtail to collect logs from the server. If you do not configure a user identifier on your server, Log Service cannot receive the heartbeat of the server and Logtail cannot collect logs from the server.

Procedure

1. View the ID of your Apsara Stack tenant account.
   i. Log on to the Log Service console. For more information, see Log on to the Log Service console.
   ii. In the top navigation bar, click **Enterprise**.
   iii. In the left-side navigation pane, click **Organizations**.
   iv. Select the account that you want to view and click **Obtain an accesskey**
   v. In the **AccessKey** dialog box, view the ID of the Apsara Stack tenant account.



2. Log on to the server and specify the ID of the Apsara Stack tenant account as a user identifier.
   - Linux server:

     In the */etc/ilogtail/users* directory, create a file and set the name of the file to the ID of the Apsara Stack tenant account. Example:

```
touch /etc/ilogtail/users/1**************
touch /etc/ilogtail/users/1**************
```

- Windows server:

  In the *C:\LogtailData\users* directory, create a file and set the name of the file to the ID of the Apsara Stack tenant account. For example, you can set the file name to *C:\LogtailData\users\1**************.

> **Note**
> - If the */etc/ilogtail/users* directory does not exist, you must create the directory.
> - You can configure the IDs of multiple Apsara Stack tenant accounts on the same server.
> - After you configure or delete a user identifier, the change takes effect within 1 minute.
> - If you no longer need a user identity file on a server, we recommend that you delete the file from the server at the earliest opportunity.

23.1.3.1.4. Text logs

23.1.3.1.4.1. Configure text log collection

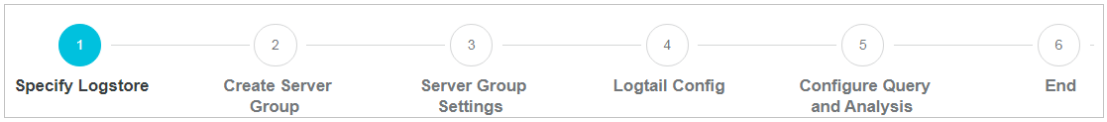
This topic describes the configuration process and collection modes when you use Logtail to collect text logs from servers.

Prerequisites

A project and a Logstore are created. For more information, see Create a project and Create a Logstore.

Configuration process

Log Service provides a configuration wizard that you can use to configure log collection.



Collection modes

Logtail supports various collection modes, such as simple mode, full regex mode, delimiter mode, JSON mode, NGINX configuration mode, IIS configuration mode, and Apache configuration mode.

- Collect logs in simple mode
- Collect logs in full regex mode
- Collect logs in delimiter mode
- Collect logs in JSON mode
- Collect logs in NGINX mode
- Collect logs in IIS mode
- Collect logs in Apache mode

Procedure

1. Log on to the Log Service console.
2. Select a data source.

   Select a data source based on your business requirements. Log Service supports the following data sources of text logs: **RegEx - Text Log**, **Single Line - Text Log**, **Multi-Line - Text Log**, **Delimiter Mode - Text Log**, **JSON - Text Log**, **Nginx - Text Log**, **IIS - Text Log**, and **Apache - Text Log**.

3. Select a destination project and Logstore, and then click **Next**.

   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.

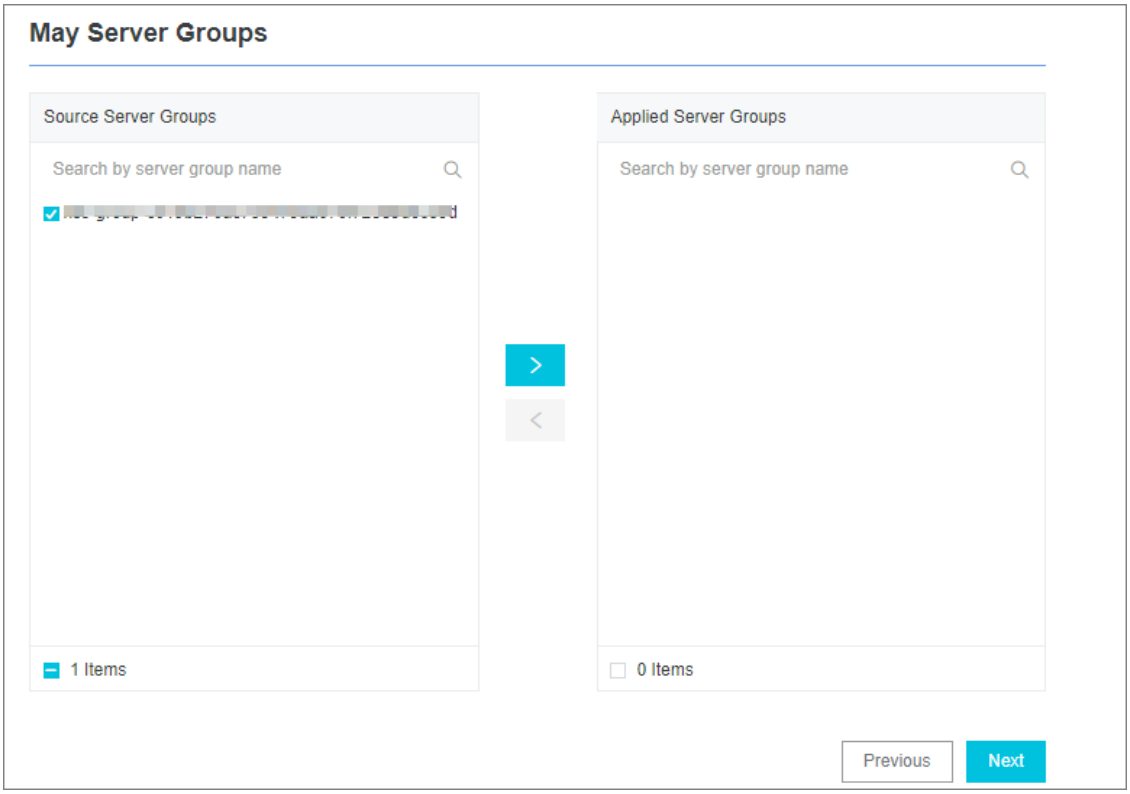
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.

5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



> **Notice** If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration and click **Next**.

   Logtail parameters vary based on collection modes. For more information, see the related parameters for specific collection methods in Collection modes.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be unique in a project. After the Logtail configuration is created, you cannot change the name of the Logtail configuration. You can also click **Import Other Configuration** to import a Logtail configuration from another project.

Parameter	Description
Log Path	The directory and name of the log file. The specified log file name can be a complete file name or a file name that contains wildcards. Log Service scans all levels of the specified directory to match log files. Examples: ◦ If you specify */apsara/nuwa/.../*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ◦ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ◦ By default, each log file can be collected by using only one Logtail configuration. ◦ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ◦ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Blacklist	If you turn on **Blacklist**, you can configure a blacklist to skip the specified directories or files when Logtail collects logs. You can use exact match or wildcard match to specify directories and files. Examples: ◦ If you select **Filter by Directory** from the Filter Type drop-down list and enter */tmp/mydir* in the Content column, all files in the directory are skipped. ◦ If you select **Filter by File** from the Filter Type drop-down list and enter */tmp/mydir/file* in the Content column, only the specified file is skipped.
Docker File	If you want to collect logs from Docker containers, you can turn on **Docker File** and specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs. For more information about container text logs, see Collect Kubernetes logs.
Mode	The default mode is **Simple Mode - Multi-line**. You can change the mode.

Parameter	Description

Log Sample	Enter a sample log that is retrieved from a log source in an actual scenario. Then, Log Service can automatically generate a regular expression to match the start part in the first line of the log. Example:
	`[2020-10-01T10:30:01,000] [INFO] java.lang.Exception: exception happened`
	`at TestPrintStackTrace.f(TestPrintStackTrace.java:3)`
	`at TestPrintStackTrace.g(TestPrintStackTrace.java:7)`
	`at TestPrintStackTrace.main(TestPrintStackTrace.java:16)`
	If you collect single-line text logs in simple mode, you do not need to set this parameter.

Regex to Match First Line	The regular expression that Logtail uses to match the start part in the first line of a log. The unmatched lines are collected as part of a log. You can specify a regular expression to match the start part in the first line of a log. You can also use the regular expression that is automatically generated by Log Service. ○ Automatically generate a regular expression to match the start part in the first line of a log. After you enter a sample log, click **Auto Generate**. Log Service automatically generates a regular expression to match the start part in the first line of the log. ○ Specify a regular expression to match the start part in the first line of a log. After you enter a sample log, click **Manual** and specify a regular expression to match the start part in the first line of the log. Then, click **Validate** to check whether the regular expression is valid. If you collect single-line text logs in simple mode, you do not need to set this parameter.
Drop Failed to Parse Logs	Specifies whether to drop logs that fail to be parsed. ○ If you turn on **Drop Failed to Parse Logs**, logs that fail to be parsed are not uploaded to Log Service. ○ If you turn off **Drop Failed to Parse Logs**, raw logs are uploaded to Log Service if the logs fail to be parsed.
Maximum Directory Monitoring Depth	The maximum depth at which the log directory is monitored. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.

Parameter	Description
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ○ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ○ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ○ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

> Note
> - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
> - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs.

23.1.3.1.4.2. Collect logs in simple mode

When you collect logs in simple mode, the logs are not parsed. Each log is collected and uploaded to Log Service as a whole. This simplifies the process of log collection. This topic describes how to create a Logtail configuration in simple mode in the Log Service console to collect logs.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- Ports 80 and 443 are enabled for the server from which you want to collect logs.

Context

The simple mode supports the following types of text logs:

- Single-line text log

  Each log line is collected as a log. Two logs in a log file are separated by a line feed. In single-line mode, you must specify the directories and names of log files. Then, Logtail collects logs by line from the specified files.

- Multi-line text log

  Multiple log lines are collected as a log by default. In multi-line mode, you must specify the directories and names of log files. In addition, you must enter a sample log and configure a regular expression to match the start part in the first line of a log. Logtail uses the regular expression to match the start part in the first line of a log and reckons unmatched lines as part of the log.

> Note If you collect logs in simple mode, the timestamp of a log indicates the system time of the server when the log is collected.

Procedure

1. Log on to the Log Service console.
2. Select a data source.

   Select **Single Line - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

   You can also click **Create Now** to create a project and a Logstore.

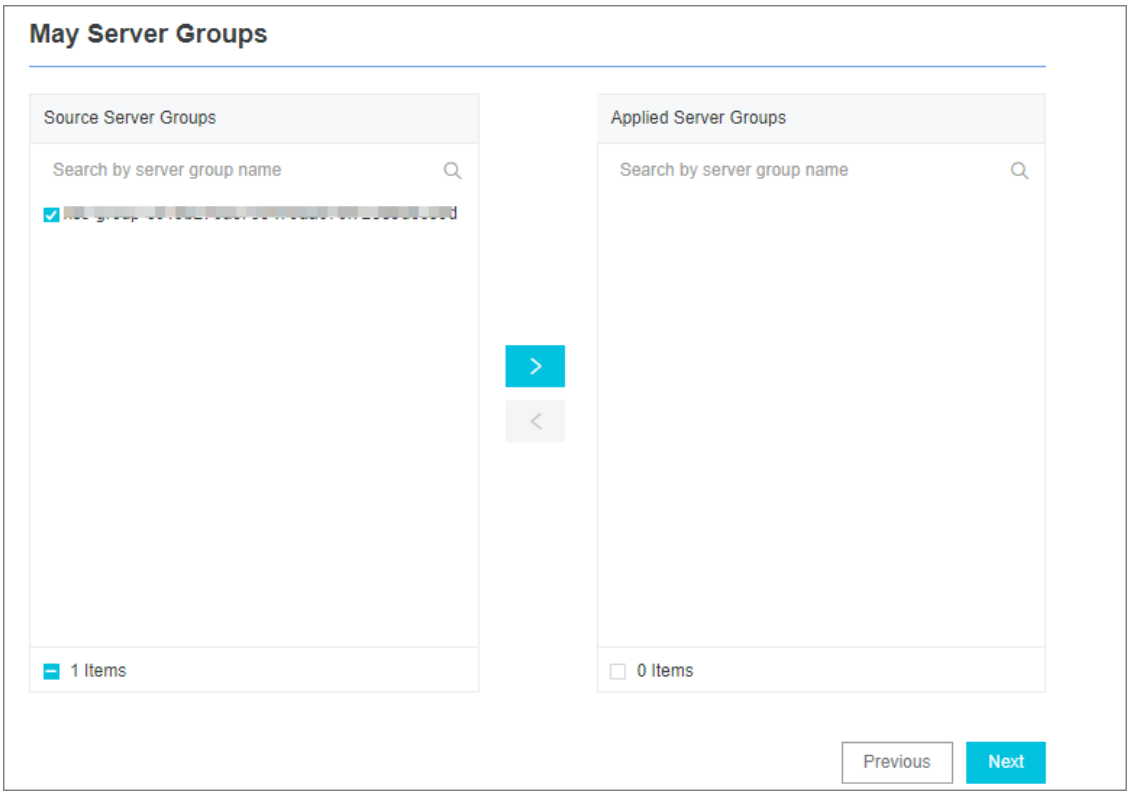
   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**,

and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration.

   The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Note** After the Logtail configuration is created, you cannot change the name of the Logtail configuration.

Parameter	Description
Log Path	The directory and name of the log file. ◦ The specified log file name can be a complete file name or a file name that contains wildcards. ◦ Recursive directory matching is used in the log file search. If this matching method is applied, all files that match the specified file name in the specified directory and its sub-directories are monitored. ▪ If you specify `/apsara/nuwa/ ** /*.log`, Log Service matches the files whose name is suffixed by `.log` in the `/apsara/nuwa` directory and its recursive subdirectories. ▪ If you specify `/var/logs/app_* … /*.log*`, Log Service matches the files that meet the following conditions: The file name contains `.log`. The file is stored in a subdirectory of the `/var/logs` directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the `app_*` pattern. **Note** ▪ By default, each log file can be collected by using only one Logtail configuration. ▪ You can use only asterisks (`*`) and question marks (`?`) as wildcards in the log path.
Docker File	If you want to collect logs from Docker containers, you can specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs.
Mode	If you have specified **Single Line - Text Log** for the data source, the default mode is **Simple Mode**. You can change the mode.
Maximum Directory Monitoring Depth	The maximum number of directory layers that can be recursively monitored when logs are collected from the data source. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.

Parameter	Description
Topic Generation Mode	The topic generation mode. ○ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ○ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ○ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

> Note
>
> - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
> - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in full regex mode.

23.1.3.1.4.3. Collect logs in full regex mode

You can use the full regex mode to extract custom fields from logs. This topic describes how to create a Logtail configuration in full regex mode in the Log Service console to collect logs.

Context

If you want to collect multi-line logs and extract fields from the logs, we recommend that you use regular expressions. Log Service can generate a regular expression based on a sample log that you specify in the Import Data wizard. However, you must modify a regular expression before it can match fields in the sample log as expected. For more information, see How do I test a regular expression?.

Procedure

1. Log on to the Log Service console.
2. Select a data source.

   Select **RegEx - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

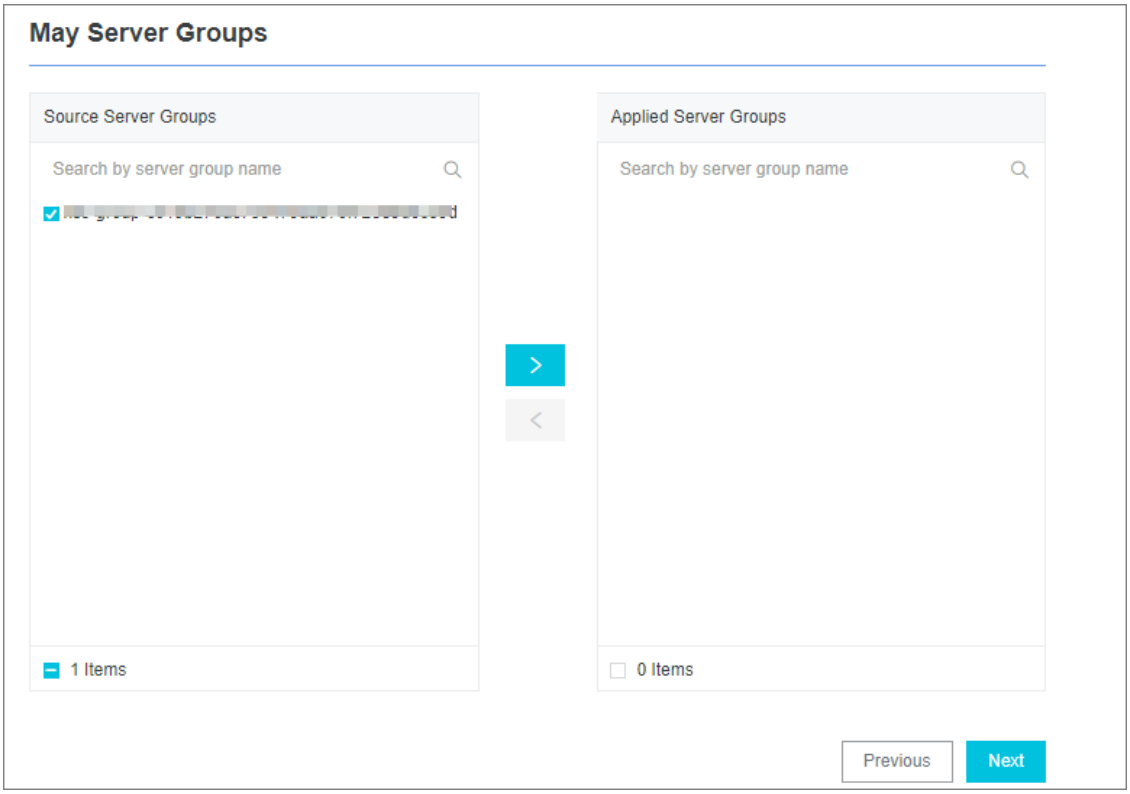
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration.

   The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be unique in a project. After the Logtail configuration is created, you cannot change the name of the Logtail configuration. You can also click **Import Other Configuration** to import a Logtail configuration from another project.

Parameter	Description
Log Path	The directory and name of the log file. ◦ The specified log file name can be a complete file name or a file name that contains wildcards. ◦ Log Services scans all levels of the specified directory to match log files. Examples: ■ If you specify */apsara/nuwa/**/*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ■ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ■ By default, each log file can be collected by using only one Logtail configuration. ■ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ■ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Docker File	If you collect logs from Docker containers, you can configure the paths and tags of the containers. Logtail monitors the containers when they are created and destroyed, filters the logs of the containers by tag, and collects the filtered logs.
Mode	If you have specified **RegEx - Text Log** for the data source, the default mode is **Full Regex Mode**. You can change the mode.
Singleline	The singleline mode is enabled by default. In this mode, logs are separated by line. To collect multi-line logs, such as Java program logs, you must disable the **Singleline** mode and configure **Regex to Match First Line**.
Log Sample	Enter a sample log that is retrieved from a log source in an actual scenario. Then, Log Service can automatically generate a regular expression.
Regex to Match First Line	You can click **Auto Generate** or **Manual**. After you enter a sample log and click **Auto Generate**, Log Service automatically generates a regular expression. If no regular expression is generated, you can switch to the manual mode and enter a regular expression for verification.

Parameter	Description
Extract Field	To analyze and process specific fields in logs, you can turn on **Extract Field**. Then, the specified fields are converted to key-value pairs and sent to Log Service. You must specify a regular expression to parse the log content.
Regular Expression	If you turn on Extract Field, you must specify this parameter. ○ Automatically generate a regular expression. You can select the fields to be extracted from the sample log and then click Generate Regular Expression. Log Service automatically generates a regular expression. ○ Specify a regular expression. You can also enter a regular expression. Click **Manual** to switch to the manual mode. After you enter a regular expression, click **Validate** to check whether Log Service can parse the log content by using the regular expression. For more information, see How do I test a regular expression?.
Extracted Content	If you turn on Extract Field, you must specify this parameter. After a regular expression is automatically generated or manually specified, you must specify the key name for each extracted field.
Use System Time	If you turn on Extract Field, you must specify this parameter. If you turn off Use System Time, you must specify a field as the time field and name the field `time`. After you specify the `time` field, click **Auto Generate** in the **Time Conversion Format** field to parse the time. For more information, see Configure the time format.
Drop Failed to Parse Logs	Specifies whether to upload logs to Log Service if the logs fail to be parsed. ○ If you turn on this switch, logs that fail to be parsed are not uploaded to Log Service. ○ If you turn off this switch, raw logs are uploaded to Log Service if the logs fail to be parsed.
Maximum Directory Monitoring Depth	The maximum number of directory levels that can be recursively monitored when Log Service collects logs from the data source. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.

Parameter	Description
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ○ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ○ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ○ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

> **Note**
>
> - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
> - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in full regex mode.

23.1.3.1.4.4. Collect logs in delimiter mode

Log Service allows you to collect logs in delimiter mode. After logs are collected, you can transform and ship the logs, and perform multidimensional log analysis. This topic describes how to create a Logtail configuration in delimiter mode in the Log Service console to collect logs.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- Ports 80 and 443 are enabled for the server from which you want to collect logs.

Procedure

1. Log on to the Log Service console.
2. In the **Import Data** section, select **Delimiter Mode - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

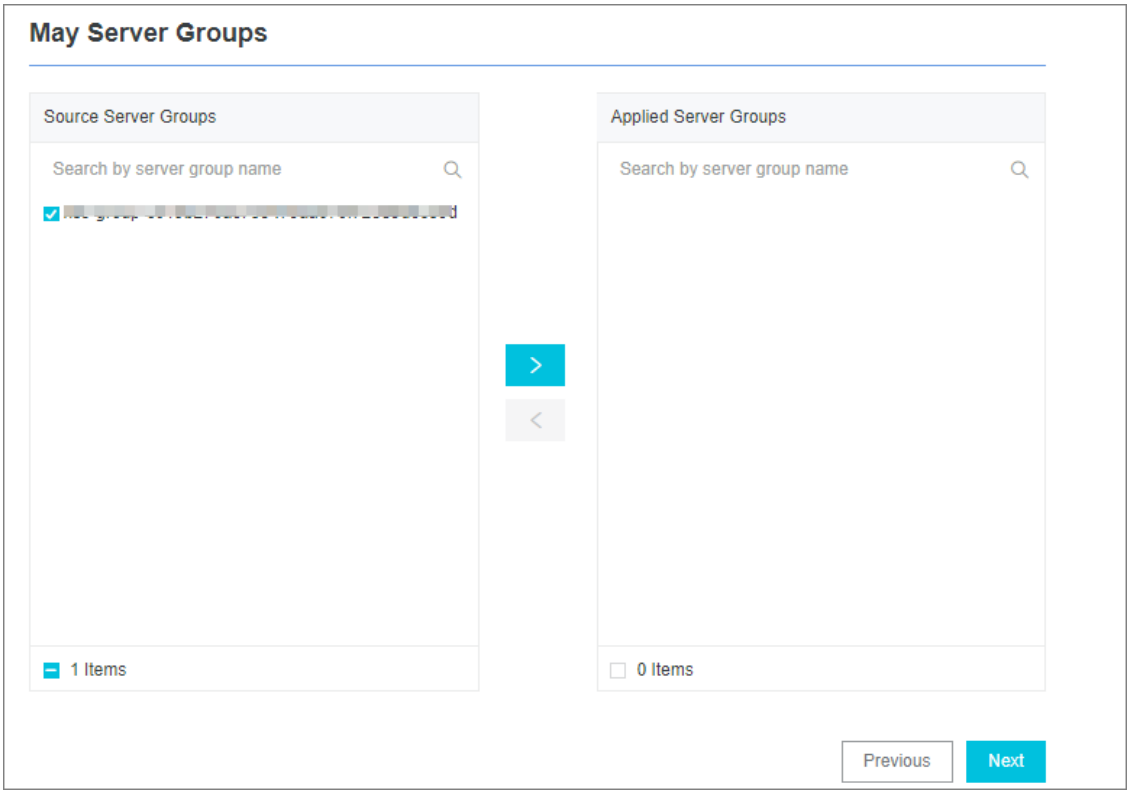
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be FAIL. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration and click **Next**.

   The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Note** After the Logtail configuration is created, you cannot change the name of the Logtail configuration.

Parameter	Description
Log Path	The directory and name of the log file. The specified log file name can be a complete file name or a file name that contains wildcards. For more information, see Wildcard matching. Log Services scans all levels of the specified directory to match log files. Examples: ○ If you specify */apsara/nuwa/**/*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ○ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ○ By default, each log file can be collected by using only one Logtail configuration. ○ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ○ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Docker File	If you want to collect logs from Docker containers, you can specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs.
Mode	The default mode is **Delimiter Mode**. You can change the mode.
Log Sample	Enter a sample log that is retrieved from a log source in an actual scenario. Example: `127.0.0.1|#|-|#|13/Apr/2020:09:44:41 +0800|#|GET /1 HTTP/1.1|#|0.000|#|74|#|404|#|3650|#|-|#|curl/7.29.0` **Note** The delimiter mode applies only to single-line logs. If you want to collect multi-line logs, we recommend that you select Simple Mode - Multi-line or Full Regex Mode.

Parameter	Description
Delimiter	Select a delimiter based on the log format. For example, if you select Vertical Line, a vertical bar (|) is used as the delimiter. For more information, see Appendix: delimiters and sample logs. **Note** If you set the Delimiter parameter to **Hidden Characters**, you must enter a character in the following format: `0x`*Hexadecimal ASCII code of the non-printable character*. For example, if you want to use the non-printable character whose hexadecimal ASCII code is 01, you must enter **0x01**.
Quote	If a log field contains delimiters, you must specify a pair of quotes to enclose the field. Log Service parses the content that is enclosed in a pair of quotes into a complete field. Select a quote based on the log format. **Note** If you set the Quote parameter to **Hidden Characters**, you must enter a character in the following format: `0x`*Hexadecimal ASCII code of the non-printable character*. For example, if you want to use the non-printable character whose hexadecimal ASCII code is 01, you must enter **0x01**.
Extracted Content	Log Service extracts the log content based on the sample log and delimiter that you specify. The extracted log content is delimited into values. You must specify a key for each value.
Incomplete Entry Upload	Specifies whether to upload a log whose number of parsed fields is less than the number of the specified keys. If you turn on this switch, the log is uploaded. If you turn off this switch, the log is dropped. For example, if you specify a vertical bar (|) as the delimiter, the log 11|22|33|44|55 is parsed into the following fields: 11, 22, 33, 44, and 55. You can set the keys to A, B, C, D, and E. ◦ If you turn on **Incomplete Entry Upload**, 55 is uploaded as the value of the D key when Log Service collects the log 11|22|33|55. ◦ If you turn off **Incomplete Entry Upload**, the log 11|22|33|55 is dropped because the number of fields parsed from the log does not match the number of the specified keys.
Use System Time	Specifies whether to use the system time. ◦ If you turn on **Use System Time**, the timestamp of a log indicates the system time of the server when the log is collected. ◦ If you turn off **Use System Time**, you must set the **Specify Time Key** and **Time Format** parameters based on the value of the time field that is specified in **Extracted Content**. For example, if you set the **Specify Time Key** parameter to **time_local** and the **Time Format** parameter to **%d/%b/%Y:%H:%M:%S**, the timestamp of a log is the value of the **time_local** field.
Drop Failed to Parse Logs	Specifies whether to upload logs to Log Service if the logs fail to be parsed. ◦ If you turn on **Drop Failed to Parse Logs**, logs that fail to be parsed are not uploaded to Log Service. ◦ If you turn off **Drop Failed to Parse Logs**, raw logs are uploaded to Log Service if the logs fail to be parsed.

Parameter	Description
Maximum Directory Monitoring Depth	The maximum depth at which the log directory is monitored. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ○ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ○ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ○ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.

Parameter	Description
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > ○ To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > ○ If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in delimiter mode.

Appendix: delimiters and sample logs

Logs that are in the delimiter-separated values (DSV) format use line feeds as boundaries. Each line indicates a log. Each log is parsed into multiple fields by using delimiters. Both single-character and multi-character delimiters are supported. If a field contains delimiters, you can enclose the field in a pair of quotes.

- Single-character delimiter

  The following example shows sample logs with single-character delimiters:

```
05/May/2016:13:30:28,10.10.*.*,"POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=**************
**&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=************************
******* HTTP/1.1",200,18204,aliyun-sdk-java
05/May/2016:13:31:23,10.10.*.*,"POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=**************
**&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=************************
******* HTTP/1.1",401,23472,aliyun-sdk-java
```

  If a log contains single-character delimiters, you must specify the delimiter. You can also specify a quote.

- Delimiter: Available single-character delimiters include the tab character (\t), vertical bar (|), space character, comma (,), and semicolon (;). You can also specify a non-printable character as the delimiter. You cannot specify a double quotation mark (") as the delimiter.

  However, a double quotation mark (") can be used as a quote. You can place the double quotation mark at the border of a field, or in the field. If a double quotation mark (") is included in a log field but is not used as a quote, it must be escaped as double quotation marks (""). When Log Service parses log fields, the double quotation marks ("") are automatically converted to a double quotation mark ("). For example, you can specify a comma (,) as the delimiter and a double quotation mark (") as the quote in a log field. You must enclose the field that contains commas (,) in a pair of quotes. In addition, you must escape the double quotation mark (") in the field to double quotation marks (""). If a processed log is in the 1999,Chevy,"Venture ""Extended Edition, Very Large""","",5000.00 format, the log can be parsed into the following five fields: 1999, Chevy, Venture "Extended Edition, Very Large", an empty field, and 5000.00.

- Quote: If a log field contains delimiters, you must specify a pair of quotes to enclose the field. Log Service parses the content that is enclosed in a pair of quotes into a new complete field.

  Available quotes include the tab character (\t), vertical bar (|), space character, comma (,), semicolon (;), and non-printable characters.

  For example, if you specify a comma (,) as the delimiter and a double quotation mark (") as the quote, the log 1997,Ford,E350,"ac, abs, moon",3000.00 is parsed into the following five fields: 1997, Ford, E350, ac, abs, moon, and 3000.00.

- Multi-character delimiter

  The following example shows sample logs with multi-character delimiters:

```
05/May/2016:13:30:28&&10.200.**.**&&POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=**********
******&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=pD12XYLmGxKQ%2Bmkd6x7
hAgQ7b1c%3D HTTP/1.1&&200&&18204&&aliyun-sdk-java
05/May/2016:13:31:23&&10.200.**.**&&POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=**********
******&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=**********************
********** HTTP/1.1&&401&&23472&&aliyun-sdk-java
```

  A multi-character delimiter can contain two or three characters, such as ||, &&&, and ^_^. Log Service parses logs based on delimiters. You do not need to use quotes to enclose log fields.

  > **Note** You must make sure that the delimiters in a field cannot be parsed into a new field. Otherwise, Log Service cannot parse the fields as expected.

  For example, if you specify && as the delimiter, the log 1997&&Ford&&E350&&ac&abs&moon&&3000.00 is parsed into the following five fields: 1997, Ford, E350, ac&abs&moon, and 3000.00.

23.1.3.1.4.5. Collect logs in JSON mode

Log Service allows you to collect JSON logs in JSON mode by using Logtail. After logs are collected, you can transform and ship the logs, and perform multidimensional log analysis. This topic describes how to create a Logtail configuration in JSON mode in the Log Service console to collect logs.

Context

JSON logs can be written in the object or array structure. A log in the object structure contains key-value pairs, and a log in the array structure contains an ordered list of values.

In JSON mode, Logtail can parse JSON logs in the object structure and extract the keys and values from the first layer of each object. The extracted keys are used as field names, and the extracted values are used as field values. Logtail cannot parse JSON logs in the array structure. If you want to parse JSON logs in the array structure, you can collect data from the JSON logs in full regex or simple mode. For more information, see Collect logs by line or Use regular expressions to collect logs.

Sample JSON logs:

{"url": "POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=U0Ujpek********&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=pD12XYLmGxKQ%2Bmkd6x7hAgQ7b1c%3D HTTP/1.1", "ip": "10.200.98.220", "user-agent": "aliyun-sdk-java", "request": {"status": "200", "latency": "18204"}, "time": "05/Jan/2020:13:30:28"}
{"url": "POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=U0Ujpek********&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=pD12XYLmGxKQ%2Bmkd6x7hAgQ7b1c%3D HTTP/1.1", "ip": "10.200.98.210", "user-agent": "aliyun-sdk-java", "request": {"status": "200", "latency": "10204"}, "time": "05/Jan/2020:13:30:29"}

Procedure

1. Log on to the Log Service console.
2. In the **Import Data** section, select **JSON - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

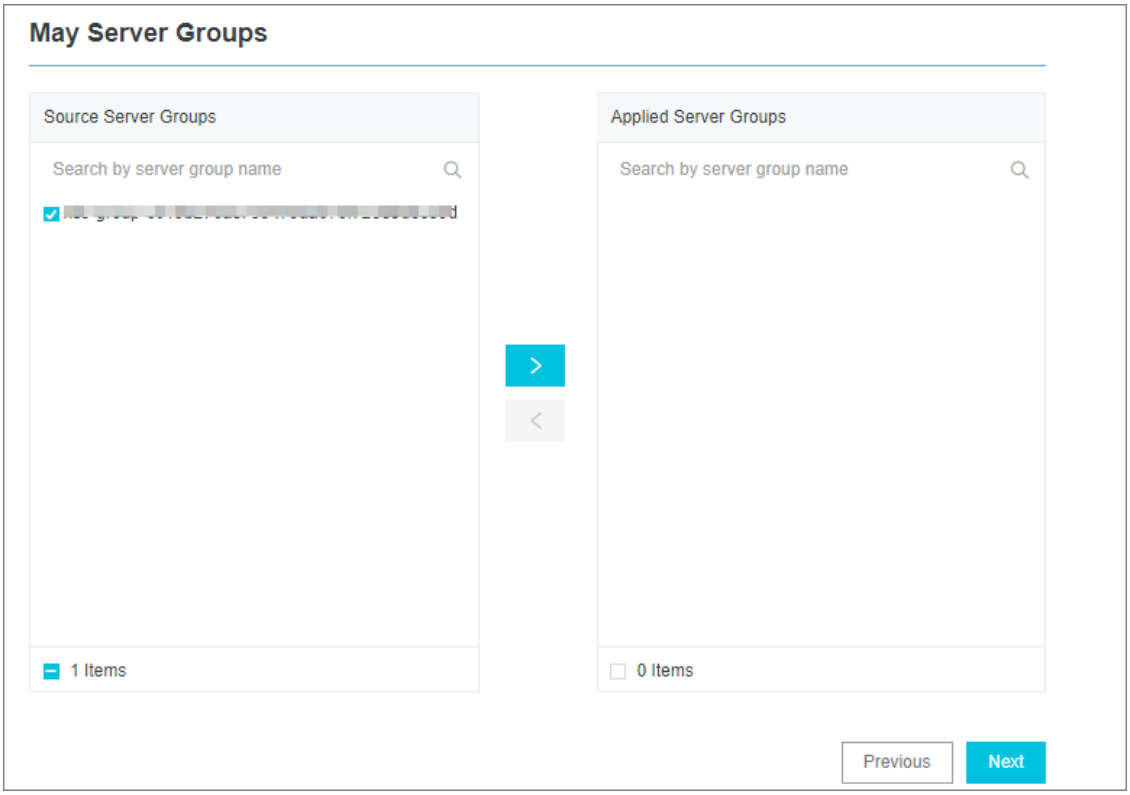
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



> **Notice** If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration and click **Next**.

   The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Note** After the Logtail configuration is created, you cannot change the name of the Logtail configuration.

Parameter	Description
Log Path	The directory and name of the log file. The specified log file name can be a complete file name or a file name that contains wildcards. For more information, see Wildcard matching. Log Services scans all levels of the specified directory to match log files. Examples: ◦ If you specify */apsara/nuwa/**/*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ◦ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ◦ By default, each log file can be collected by using only one Logtail configuration. ◦ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ◦ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Docker File	If you want to collect logs from Docker containers, you can specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs.
Mode	The default mode is **JSON Mode**. You can change the mode.

Parameter	Description
Use System Time	Specifies whether to use the system time. ◦ If you turn on **Use System Time**, the timestamp of a log indicates the system time of the server when the log is collected. ◦ If you turn off **Use System Time**, you must set the **Specify Time Key** and **Time Format** parameters based on the time field of JSON logs. For example, if the time information in a JSON log is **"time": "05/May/2016:13:30:28"**, you can set the **Specify Time Key** parameter to **time** and the **Time Format** parameter to **%d/%b/%Y:%H:%M:%S**.
Drop Failed to Parse Logs	Specifies whether to upload logs to Log Service if the logs fail to be parsed. ◦ If you turn on **Drop Failed to Parse Logs**, logs that fail to be parsed are not uploaded to Log Service. ◦ If you turn off **Drop Failed to Parse Logs**, raw logs are uploaded to Log Service if the logs fail to be parsed.
Maximum Directory Monitoring Depth	The maximum depth at which the log directory is monitored. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ◦ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ◦ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ◦ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.

Parameter	Description
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > ○ To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > ○ If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in JSON mode.

23.1.3.1.4.6. Collect logs in NGINX mode

Log Service allows you to collect NGINX logs and perform multidimensional log analysis. This topic describes how to create a Logtail configuration in NGINX mode in the Log Service console to collect logs.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- Ports 80 and 443 are enabled for the server from which you want to collect logs.

Procedure

1. Log on to the Log Service console.
2. In the **Import Data** section, select **Nginx - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

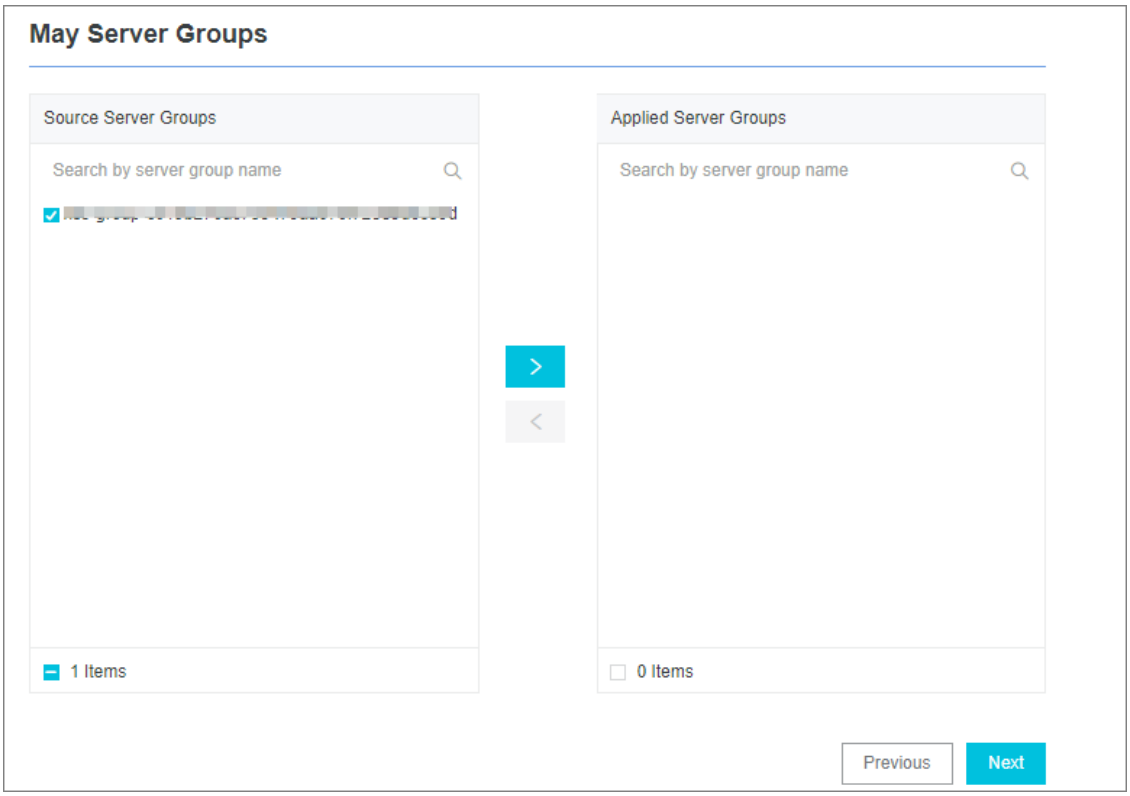
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



> **Notice** If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration and click **Next**.

   The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Note** After the Logtail configuration is created, you cannot change the name of the Logtail configuration.
Log Path	The directory and name of the log file. The specified log file name can be a complete file name or a file name that contains wildcards. For more information, see Wildcard matching. Log Services scans all levels of the specified directory to match log files. Examples: ◦ If you specify */apsara/nuwa/**/*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ◦ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ◦ By default, each log file can be collected by using only one Logtail configuration. ◦ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ◦ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Docker File	If you want to collect logs from Docker containers, you can specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs.
Mode	The default mode is **NGINX Configuration Mode**. You can change the mode.

Parameter	Description

NGINX Log Configuration	Enter the log configuration section that is specified in the NGINX configuration file. The section starts with **log_format**. Example:
	`log_format main '$remote_addr - $remote_user [$time_local] "$request" '`
	`'$request_time $request_length '`
	`'$status $body_bytes_sent "$http_referer" '`
	`'"$http_user_agent"';`
	For more information, see Appendix: log formats and sample logs.

NGINX Key	The NGINX keys and values are automatically generated based on the content of NGINX Log Configuration and Log Sample.
Drop Failed to Parse Logs	Specifies whether to upload logs to Log Service if the logs fail to be parsed. ◦ If you turn on **Drop Failed to Parse Logs**, logs that fail to be parsed are not uploaded to Log Service. ◦ If you turn off **Drop Failed to Parse Logs**, raw logs are uploaded to Log Service if the logs fail to be parsed.
Maximum Directory Monitoring Depth	The maximum depth at which the log directory is monitored. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ◦ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ◦ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ◦ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.

Parameter	Description
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > ○ To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > ○ If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in NGINX mode.

Appendix: log formats and sample logs

Before you collect NGINX access logs, you must specify log_format and access_log in the */etc/nginx/nginx.conf* file. The log_format parameter is used to specify the log format. The access_log parameter is used to specify the path in which the NGINX log files are stored.

- Log format

  The following sample code shows the default values of the log_format and access_log parameters:

```
log_format main  '$remote_addr - $remote_user [$time_local] "$request" '
                 '$request_time $request_length '
                 '$status $body_bytes_sent "$http_referer" '
                 '"$http_user_agent"';
access_log /var/logs/nginx/access.log main
```

  The following table describes the log fields.

Log field	Description
remote_addr	The IP address of the client.
remote_user	The username of the client.
time_local	The system time of the server. The value must be enclosed in brackets [].
request	The URI and HTTP protocol of a request.
request_time	The time that is required to process a request. Unit: seconds.
request_length	The length of a request. The length includes the request line, request header, and request body.
status	The status of a request.
body_bytes_sent	The number of bytes in a response that is sent to the client. The size of the response header is excluded.
http_referer	The URL of the source web page.
http_user_agent	The browser information of the client.

- Sample log

```
192.168.1.2 - - [10/Jul/2020:15:51:09 +0800] "GET /ubuntu.iso HTTP/1.0" 0.000 129 404 168 "-" "Wget
/1.11.4 Red Hat modified"
```

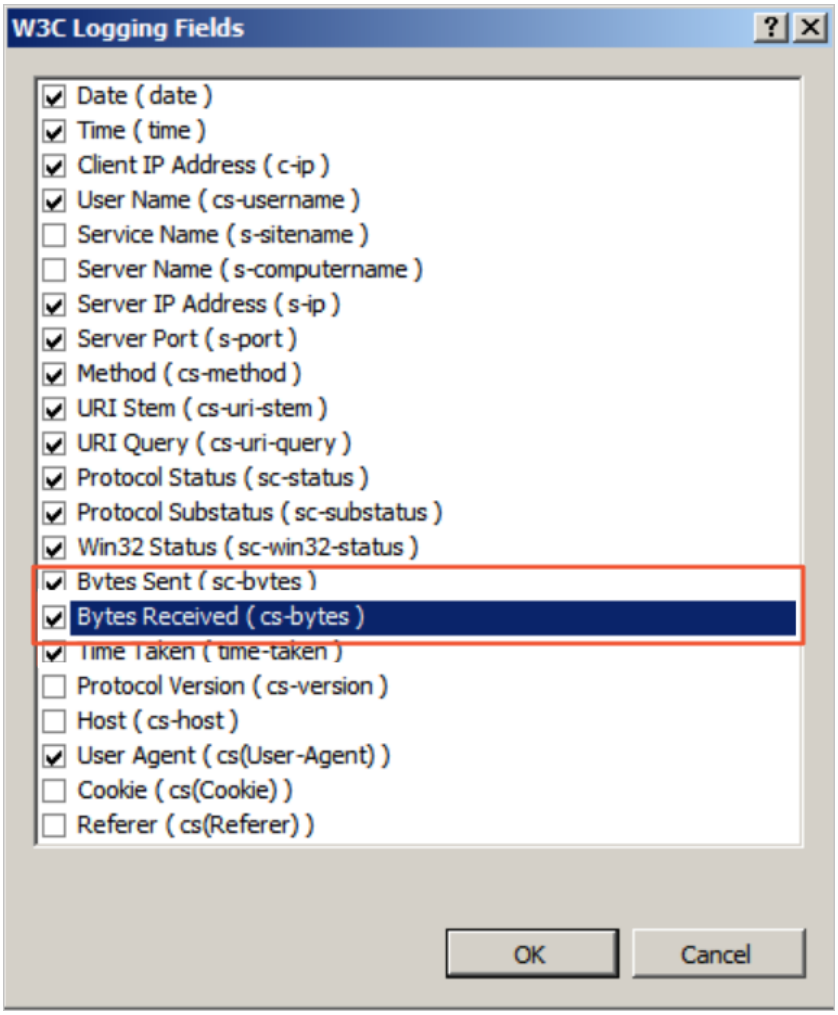
23.1.3.1.4.7. Collect logs in IIS mode

Log Service allows you to collect Internet Information Services (IIS) logs and perform multidimensional log analysis. This topic describes how to create a Logtail configuration in IIS mode in the Log Service console to collect logs.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- Ports 80 and 443 are enabled for the server from which you want to collect logs.
- Logs are generated on the server in the IIS, NCSA Common, or W3C Extended format.

We recommend that you use the W3C Extended log format. If you select the W3C Extended format, you must configure the fields in the W3C Logging Fields dialog box. To do so, you must select **Bytes Sent (sc-bytes)** and **Bytes Received (cs-bytes)** and use the default settings for other fields.



Procedure

1. Log on to the Log Service console.
2. Select a data source.

   In the Import Data section, select **IIS - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

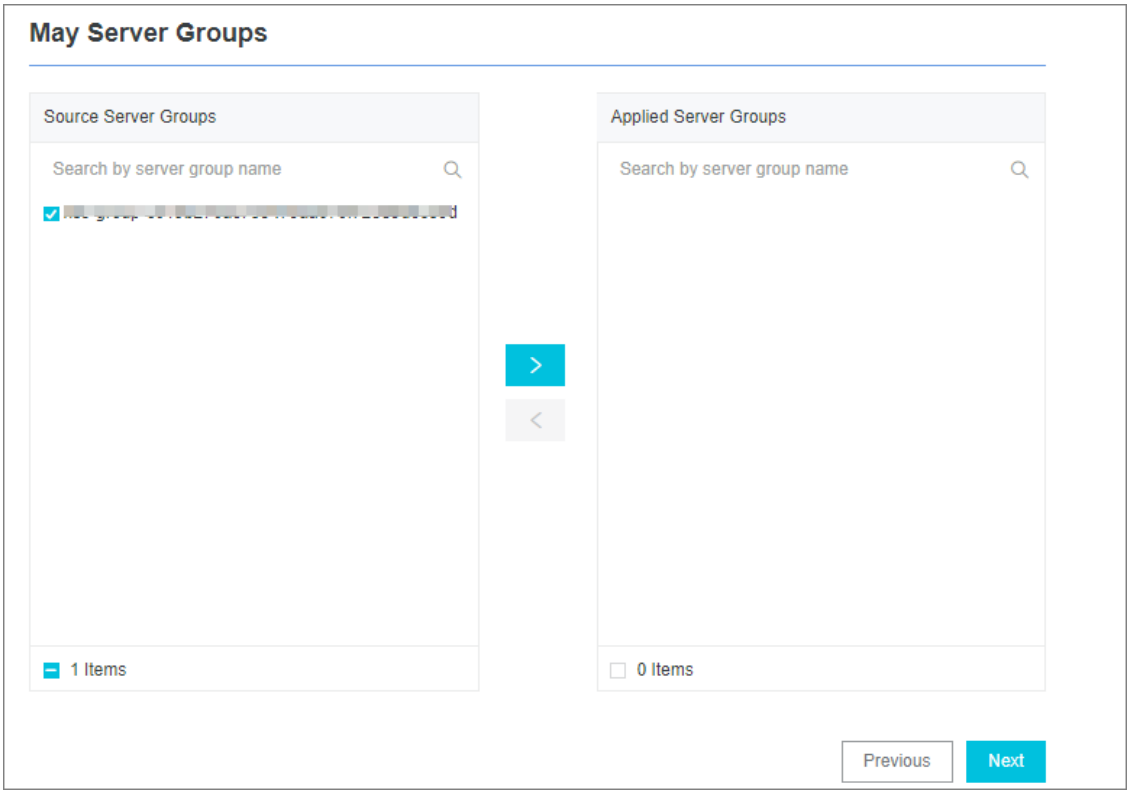
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration.

   The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Note** After the Logtail configuration is created, you cannot change the name of the Logtail configuration.

Parameter	Description
Log Path	The directory and name of the log file. The specified log file name can be a complete file name or a file name that contains wildcards. For more information, see Wildcard matching. Log Services scans all levels of the specified directory to match log files. Examples: ○ If you specify */apsara/nuwa/**/*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ○ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ○ By default, each log file can be collected by using only one Logtail configuration. ○ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ○ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Docker File	If you want to collect logs from Docker containers, you can specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs. For more information, see Collect container text logs.
Mode	If you have specified **IIS - Text Log** for the data source, the default mode is **IIS Configuration Mode**. You can change the mode.
Log format	Select the format of logs that are generated on the IIS server. ○ IIS: Microsoft IIS log file format ○ NCSA: NCSA Common log file format ○ W3C: W3C Extended log file format

Parameter	Description
IIS Configuration	Specify the IIS configuration fields. ◦ If you set the Log format parameter to IIS or NCSA, the IIS configuration fields are automatically generated. ◦ If you set the Log format parameter to W3C, enter the content that is specified in the logFile logExtFileFlags field of the IIS configuration file. `logExtFileFlags="Date, Time, ClientIP, UserName, SiteName, ComputerName, ServerIP, Method, UriStem, UriQuery, HttpStatus, Win32Status, BytesSent, BytesRecv, TimeTaken, ServerPort, UserAgent, Cookie, Referer, ProtocolVersion, Host, HttpSubStatus"` ■ Default path of the IIS5 configuration file: C:\WINNT\system32\inetsrv\MetaBase.bin ■ Default path of the IIS6 configuration file: C:\WINDOWS\system32\inetsrv\MetaBase.xml ■ Default path of the IIS7 configuration file: C:\Windows\System32\inetsrv\config\applicationHost.config
IIS Key Name	Log Service automatically extracts IIS keys based on the content of **IIS Configuration**.
Drop Failed to Parse Logs	◦ If you turn on **Drop Failed to Parse Logs**, logs that fail to be parsed are not uploaded to Log Service. ◦ If you turn off **Drop Failed to Parse Logs**, raw logs are uploaded to Log Service if the logs fail to be parsed.
Maximum Directory Monitoring Depth	The maximum depth at which the log directory is monitored. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.

Parameter	Description
Topic Generation Mode	The topic generation mode. ○ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ○ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ○ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

> **Note**
>
> - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
> - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in IIS mode.

Appendix: sample logs and field descriptions

The following example shows a sample IIS log:

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2020-09-08 09:30:26
#Fields: date time s-sitename s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs(User-Agent) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken
2009-11-26 06:14:21 W3SVC692644773 125.67.67.* GET /index.html - 80 - 10.10.10.10 Baiduspider+(+http://www.baidu.com)200 0 64 185173 296 0
```

- Field prefixes

Prefix	Description
s-	The server action.
c-	The client action.
cs-	The client-to-server action.
sc-	The server-to-client action.

- Fields

Log field	Description
date	The date on which the client sends the request.
time	The point in time at which the client sends the request.
s-sitename	The Internet service name and instance ID of the site that is visited by the client.
s-computername	The name of the server on which the log is generated.
s-ip	The IP address of the server on which the log is generated.
cs-method	The HTTP request method that is used by the client, for example, GET or POST.
cs-uri-stem	The URI resource that is requested by the client.
cs-uri-query	The query string that follows the question mark (?) in the HTTP request.
s-port	The port number of the server.

Log field	Description
cs-username	The authenticated domain name or username that is used by the client to access the server. ◦ Authenticated users are referenced as `domain\username`. ◦ Anonymous users are indicated by a hyphen (-).
c-ip	The real IP address of the client that sends the request.
cs-version	The protocol version that is used by the client, for example, HTTP 1.0 or HTTP 1.1.
cs(User-Agent)	The browser that is used by the client.
Cookie	The content of the sent or received cookie. If no cookie is sent or received, a hyphen (-) is displayed.
referer	The previous site that is visited by the user.
cs-host	The host information.
sc-status	The HTTP status code that is returned by the server.
sc-substatus	The HTTP substatus code that is returned by the server.
sc-win32-status	The Windows status code that is returned by the server.
sc-bytes	The number of bytes that are sent by the server.
cs-bytes	The number of bytes that are received by the server.
time-taken	The time that is required to process the request. Unit: milliseconds.

23.1.3.1.4.8. Collect logs in Apache mode

Log Service allows you to collect Apache logs and perform multidimensional log analysis. This topic describes how to create a Logtail configuration in Apache mode in the Log Service console to collect logs.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- Ports 80 and 443 are enabled for the server from which you want to collect logs.
- The print format, log path, and log file name are specified in the Apache configuration file. For more information, see Appendix: log formats and sample logs.

Procedure

1. Log on to the Log Service console.
2. Select a data source.

   In the Import Data section, select **Apache - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

   You can also click **Create Now** to create a project and a Logstore.

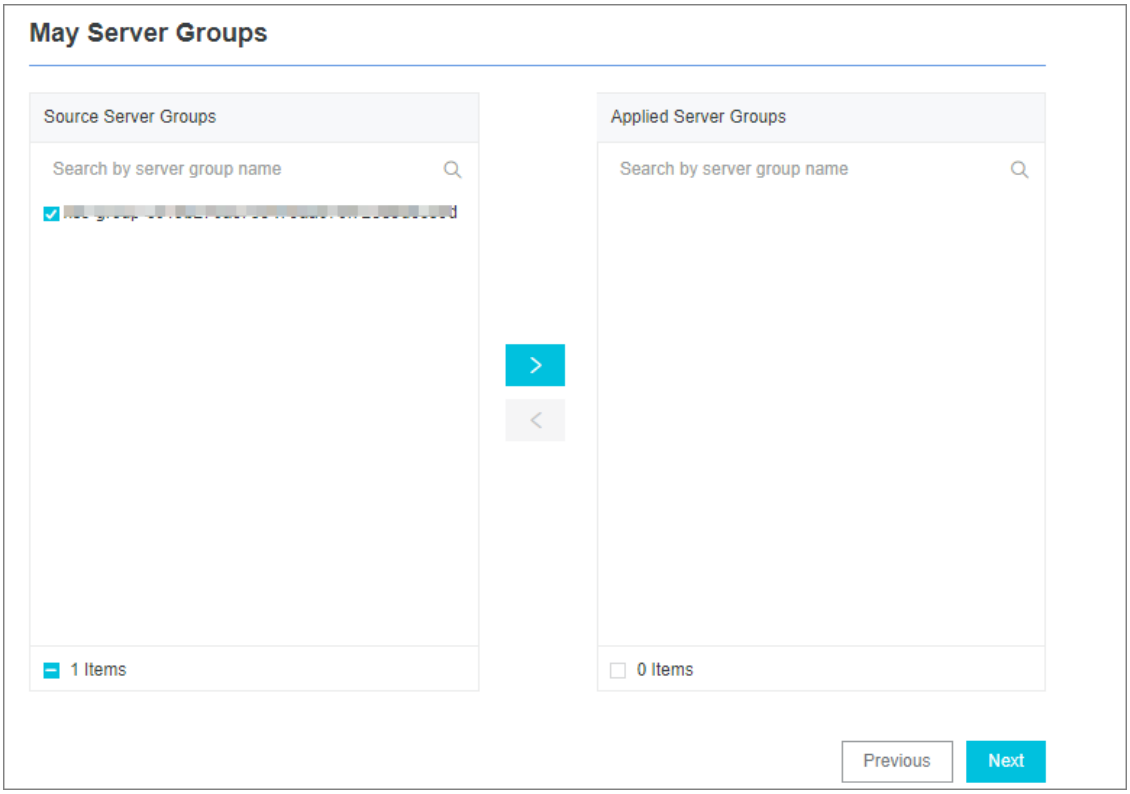
   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

Before you can create a machine group, you must install Logtail.

Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.

5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration.

The following table describes the Logtail parameters.

Parameter	Description
Config Name	The name of the Logtail configuration. The name must be 3 to 128 characters in length, and can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit. **Note** After the Logtail configuration is created, you cannot change the name of the Logtail configuration.

Parameter	Description
Log Path	The directory and name of the log file. The specified log file name can be a complete file name or a file name that contains wildcards. For more information, see Wildcard matching. Log Services scans all levels of the specified directory to match log files. Examples: ○ If you specify */apsara/nuwa/**/*.log*, Log Service matches the files whose name is suffixed by *.log* in the */apsara/nuwa* directory and its recursive subdirectories. ○ If you specify */var/logs/app_*/*.log*, Log Service matches the files that meet the following conditions: The file name contains *.log*. The file is stored in a subdirectory of the */var/logs* directory or in a recursive subdirectory of the subdirectory. The name of the subdirectory matches the *app_** pattern. **Note** ○ By default, each log file can be collected by using only one Logtail configuration. ○ To use multiple Logtail configurations to collect one log file, we recommend that you create a symbolic link that points to the directory where the file is located. For example, you want to collect two copies of the log.log file from the */home/log/nginx/log/log.log* directory. You can run the following command to create a symbolic link that points to the directory. When you configure the Logtail configurations, use the original path in one Logtail configuration and use the symbolic link in the other Logtail configuration. `ln -s /home/log/nginx/log /home/log/nginx/link_log` ○ You can use only asterisks (*) and question marks (?) as wildcards in the log path.
Docker File	If you want to collect logs from Docker containers, you can specify the paths and tags of the containers. Logtail monitors the containers to check whether the containers are created or destroyed. Then, Logtail filters the logs of the containers based on tags and collects the filtered logs. For more information about container text logs, see Collect container text logs.
Mode	If you have specified **Apache - Text Log** for the data source, the default mode is **Apache Configuration Mode**. You can change the mode.
Log format	Select a log format based on the format specified in the Apache configuration file. Valid values: common, combined, and Custom.
APACHE Logformat Configuration	Enter the log configuration section that is specified in the Apache configuration file. The section starts with LogFormat. For more information, see Appendix: log formats and sample logs. ○ If you set **Log format** to **common** or **combined**, the system automatically inserts a value into this field. Check whether the value is the same as the value specified in the Apache configuration file. ○ If you set **Log format** to **Custom**, specify a value based on your business requirements. For example, you can enter **LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\" %D %f %k %p %q %R %T %I %O" customized**.

Parameter	Description
APACHE Key Name	Log Service automatically reads Apache keys from the value of the **APACHE Logformat Configuration** field.
Drop Failed to Parse Logs	Specifies whether to drop logs that fail to be parsed. ◦ If you turn on **Drop Failed to Parse Logs**, logs that fail to be parsed are not uploaded to Log Service. ◦ If you turn off **Drop Failed to Parse Logs**, raw logs are uploaded to Log Service if the logs fail to be parsed.
Maximum Directory Monitoring Depth	The maximum depth at which the log directory is monitored. Valid values: 0 to 1000. The value 0 indicates that only the specified log file directory is monitored.

7. (Optional)Specify **Advanced Options** and click **Next**.

   Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ◦ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ◦ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ◦ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.
Log File Encoding	The encoding format of log files. Valid values: ◦ utf8: UTF-8 encoding format ◦ gbk: GBK encoding format

Parameter	Description
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > ○ To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > ○ If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect logs in Apache mode.

Appendix: log formats and sample logs

Before you collect Apache logs, you must specify the print format, log path, and log file name. For example, **CustomLog "/var/log/apache2/access_log" combined** indicates that the combined format is used when logs are printed. The log file path is */var/log/apache2/access_log*. Log Service supports the following log formats. A sample log is also provided.

- Log formats
  - The combined log format:

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined
```

- The common log format:

```
LogFormat "%h %l %u %t \"%r\" %>s %b"
```

- A custom log format:

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\" %D %f %k %p %q %R %T %I %O" customized
```

The following table describes the related log fields. For more information, see mod_log_config.

Format string	Log field	Description
%a	client_addr	The IP address of the client.
%A	local_addr	The local IP address.
%b	response_size_bytes	The number of bytes in a response. If no bytes are sent, a hyphen (-) is displayed for this field.
%B	response_bytes	The number of bytes in a response. If no bytes are sent, the digit 0 is displayed for this field.
%D	request_time_msec	The time required to process a request. Unit: microseconds.
%f	filename	The file name.
%h	remote_addr	The name of the remote host.
%H	request_protocol_supple	The request protocol.
%I	bytes_received	The number of bytes that are received by the server. This field is recorded in logs only after you enable the mod_logio module.
%k	keep_alive	The number of keep-alive requests handled on the connection.
%l	remote_ident	The information that is provided by the remote host for identification.
%m	request_method_supple	The HTTP request method.
%O	bytes_sent	The number of bytes that are sent by the server. This field is recorded in logs only after you enable the mod_logio module.
%p	remote_port	The port number of the server.
%P	child_process	The ID of the child process.
%q	request_query	The query string. If no query strings exist, an empty string is displayed.
%r	request	The content of the request. The content includes the method name, address, and HTTP protocol.
%R	response_handler	The type of the handler that generates a response on the server.

Format string	Log field	Description
%s	status	The initial HTTP status of a response.
%>s	status	The final HTTP status of a response.
%t	time_local	The point in time at which the server receives a request.
%T	request_time_sec	The time required to process a request. Unit: seconds.
%u	remote_user	The username of the client.
%U	request_uri_supple	The URI in a request. The URI does not include the query string.
%v	server_name	The name of the server.
%V	server_name_canonical	The name of the server. The name is specified by using the UseCanonicalName directive.
"%{User-Agent}i"	http_user_agent	The information of the client.
"%{Rererer}i"	http_referer	The URL of the web page. The URL is linked to the resource that is being requested.

- Sample log

```
192.168.1.2 - - [02/Feb/2020:17:44:13 +0800] "GET /favicon.ico HTTP/1.1" 404 209 "http://localhost/x1.html" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.97 Safari/537.36"
```

23.1.3.1.4.9. Configure parsing scripts

When Log Service collects logs, Log Service extracts some fields in raw logs as log content based on specific parsing methods. This way, you can collect logs based on your business requirements. This topic describes the parsing methods that are supported by Log Service.

Specify a method to separate log lines

A complete access log such as an NGINX access log occupies a line. Separate multiple logs with line feeds. The following example shows two access logs:

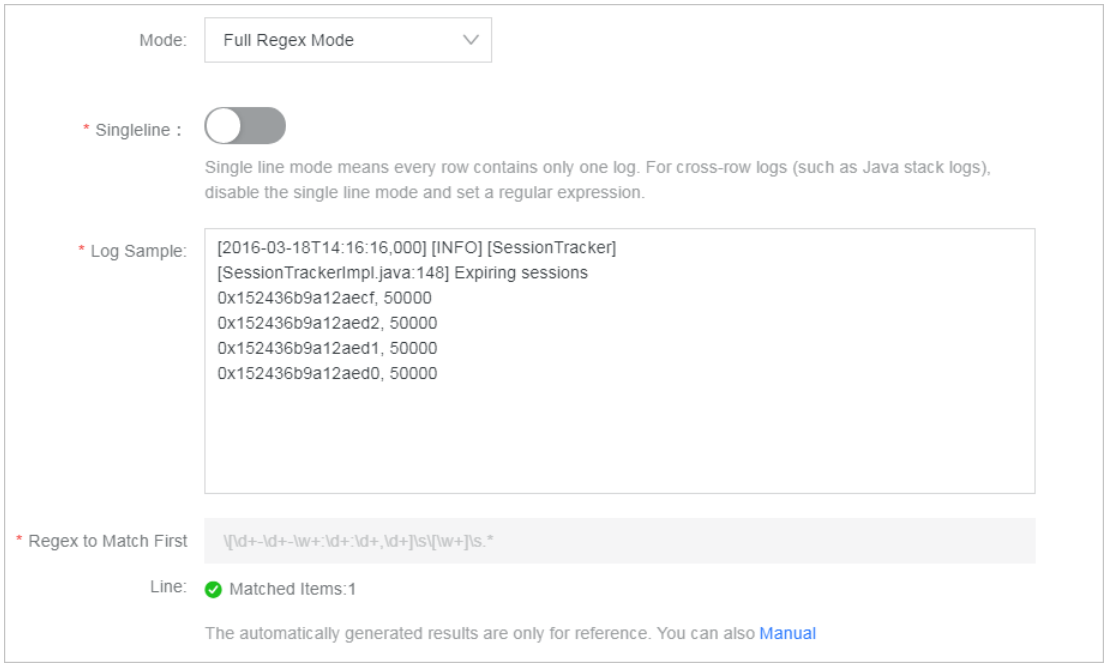
```
203.0.113.10 - - [13/Mar/2016:10:00:10 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"
203.0.113.10 - - [13/Mar/2016:10:00:11 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"
```

In most cases, logs for Java applications contain multiple lines. Therefore, logs are separated based on the start part in the first line of a log. The following example shows a Java application log:

```
[2016-03-18T14:16:16,000] [INFO] [SessionTracker] [SessionTrackerImpl.java:148] Expiring sessions
0x152436b9a12aecf, 50000
0x152436b9a12aed2, 50000
0x152436b9a12aed1, 50000
0x152436b9a12aed0, 50000
```

The preceding Java application log starts with a datetime value. To ensure the accuracy of log collection, you can specify a regular expression to match the start part in the first line of a log. In this example, the regular expression that is used to match the start part in the first line of a log is `[\d+-\d+-\w+:\d+:\d+,\d+]\s.*`. The following figure shows how to enter a sample log and specify a regular expression in the Log Service console.

Full regex mode



Extract log fields

A log contains one or more key-value pairs based on the data model of Log Service. If you want to extract specific fields for analysis, you must specify a regular expression to match the content that you want to extract. If you do not need to process the content of a log, you can process the log as a key-value pair.

The following example shows two access logs. You can use one of the following two methods to parse the logs.

```
203.0.113.10 - - [13/Mar/2016:10:00:10 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"
203.0.113.10 - - [13/Mar/2016:10:00:11 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"
```

- Extract specific fields.

  In this example, the regular expression is `(\S+)\s-\s-\s\[(\S+)\s[^]]+]\s"(\w+).*`. The extracted fields are `203.0.113.10`, `13/Mar/2016:10:00`, and `GET`.

- Extract all.

  In this example, the regular expression is `(.*)`. The extracted content is `203.0.113.10 - - [13/Mar/2016:10:00:10 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"`.

Specify the log time

A log must contain a time field whose value is a UNIX timestamp based on the data model of Log Service. You can use the system time when Logtail collects a log or the time in the log content as the log time.

The following example shows two access logs. You can use one of the following two methods to parse the logs.

```
203.0.113.10 - - [13/Mar/2016:10:00:10 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"
203.0.113.10 - - [13/Mar/2016:10:00:11 +0800] "GET / HTTP/1.1" 0.011 180 404 570 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; 360se)"
```

- Extract the time in the log content as the log time.

  In this example, the time in the log content is `13/Mar/2016:10:00:10`. To extract the time, use the following time expression: `%d/%b/%Y:%H:%M:%S`.

- Use the system time when the log was collected by Logtail as the log time.

  If you use the system time when the log was collected by Logtail as the log time, the time is converted to a timestamp.

23.1.3.1.4.10. Time formats

When you use Logtail to collect logs, you must specify time formats based on the time strings of raw logs. Logtail extracts a time string from a raw log and parses the string into a UNIX timestamp. This topic describes the commonly used time formats and provides related examples.

Commonly used time formats of logs

The following table describes the time formats that are supported by Logtail.

> **Note**
> - The timestamp of a log in Log Service is accurate to seconds. Therefore, you can specify the time format only to seconds.
> - You need to specify the time format only for the time in a time string. Other parameters such as the time zone are not required.
> - In Linux, Logtail supports all time formats provided by the strftime function. Logtail can parse and use all log time strings that can be formatted by using the strftime function.

Format	Description	Example
%a	The abbreviated day name.	Fri
%A	The full day name.	Friday
%b	The abbreviated month name.	Jan
%B	The full month name.	January
%d	The day of a month. Valid values: 01 to 31.	07, 31
%h	The abbreviated month name. The format is equivalent to %b.	Jan
%H	The hour in the 24-hour format.	22
%I	The hour in the 12-hour format.	11
%m	The month. Valid values: 01 to 12.	08
%M	The minute. Valid values: 00 to 59.	59

Format	Description	Example
%n	The line feed.	A line feed
%p	The abbreviation that indicates the morning or afternoon. Valid values: AM and PM.	AM or PM
%r	The time in the 12-hour format. The format is equivalent to %I:%M:%S %p.	11:59:59 AM
%R	The time expressed in hours and minutes. The format is equivalent to %H:%M.	23:59
%S	The second. Valid values: 00 to 59.	59
%t	The tab character.	None
%y	The two-digit year number. Valid values: 00 to 99.	04 or 98
%Y	The four-digit year number.	2004 or 1998
%C	The two-digit century number. Valid values: 00 to 99.	16
%e	The day of a month. Valid values: 1 to 31. Prefix a single-digit number with a space character.	7 or 31
%j	The day of a year. Valid values: 001 to 366.	365
%u	The day of a week as a number. Valid values: 1 to 7. The value 1 indicates Monday.	2
%U	The week of a year. Sunday is the first day of each week. Valid values: 00 to 53.	23
%V	The week of a year. Monday is the first day of each week. Valid values: 01 to 53. If a week that contains January 1 has four or more January days, the week is the first week of a year. Otherwise, the next week is considered the first week of a year.	24
%w	The day of a week as a number. Valid values: 0 to 6. The value 0 indicates Sunday.	5

Format	Description	Example
%W	The week of a year. Monday is the first day of each week. Valid values: 00 to 53.	23
%c	The date and time in the ISO 8601 format.	Tue Nov 20 14:12:58 2020
%x	The date in the ISO 8601 format.	Tue Nov 20 2020
%X	The time in the ISO 8601 format.	11:59:59
%s	The UNIX timestamp.	1476187251

Examples

The following table lists commonly used time formats. It also provides related examples and time expressions.

Example	Time expression	Time format
2017-12-11 15:05:07	%Y-%m-%d %H:%M:%S	Custom
[2017-12-11 15:05:07.012]	[%Y-%m-%d %H:%M:%S]	Custom
02 Jan 06 15:04 MST	%d %b %y %H:%M	RFC822
02 Jan 06 15:04 -0700	%d %b %y %H:%M	RFC822Z
Monday, 02-Jan-06 15:04:05 MST	%A, %d-%b-%y %H:%M:%S	RFC850
Mon, 02 Jan 2006 15:04:05 MST	%A, %d %b %Y %H:%M:%S	RFC1123
2006-01-02T15:04:05Z07:00	%Y-%m-%dT%H:%M:%S	RFC3339
2006-01-02T15:04:05.999999999Z07:00	%Y-%m-%dT%H:%M:%S	RFC3339Nano

23.1.3.1.4.11. Import historical log files

This topic describes how to import historical log files from a server to Log Service. By default, Logtail collects only incremental logs from servers. You can configure Logtail to collect historical logs.

Prerequisites

- Logtail V0.16.15 (Linux), Logtail V1.0.0.1 (Windows), or later is installed on the server. For more information, see Install Logtail in Linux and Install Logtail in Windows.
- A Logtail configuration is created and applied to a machine group. For more information, see Configure text log collection.

  If you use the Logtail configuration to import only historical files, you can specify a log collection path that does not exist.

Context

Logtail collects logs based on the modifications in the log files that are monitored. Logtail can also collect logs by loading events from local files. Logtail collects historical logs by loading local events.

You must import historical log files to the installation directory of Logtail. The directory varies based on the operating system.

- Linux: */usr/local/ilogtail*
- Windows:
  - 32-bit: *C:\Program Files\Alibaba\Logtail*
  - 64-bit: *C:\Program Files (x86)\Alibaba\Logtail*

> **Note**
> - The maximum interval between the time when a local event is generated and the time when the local event is imported is 1 minute.
> - If a local event is loaded, Logtail sends the `LOAD_LOCAL_EVENT_ALARM` message to the server.
> - If you want to import a large number of log files, we recommend that you modify the startup parameters of Logtail to increase the value of the cpu_usage_limit parameter to 2 or more and increase the value of the mem_usage_limit parameter to 512 MB or more. For more information, see Set Logtail startup parameters.

Procedure

1. Obtain the unique identifier of the Logtail configuration.

   Open the *user_log_config.json* file in the directory where Logtail is installed. You can obtain the unique identifier of the Logtail configuration from this file.

   For example, to obtain the unique identifier of the Logtail configuration in a Linux server, run the following command:

   ```
   grep "##" /usr/local/ilogtail/user_log_config.json | awk '{print $1}'
                       "##1.0##log-config-test$multi"
                       "##1.0##log-config-test$ecs-test"
                       "##1.0##log-config-test$metric_system_test"
                       "##1.0##log-config-test$redis-status"
   ```

2. Add a local event.
   i. Create the *local_event.json* file in the Logtail installation directory.

ii. Add the local event in the JSON format to the *local_event.json* file of the Logtail installation directory. The following example shows the format of the local event:

```
[
  {
    "config" : "${your_config_unique_id}",
    "dir" : "${your_log_dir}",
    "name" : "${your_log_file_name}"
   },
  {
   ...
   }
   ...
]
```

Note To prevent Logtail from loading invalid JSON files, we recommend that you save the configurations of the local event to a temporary file. Then, edit and copy the configurations to the *local_event.json* file.

Parameter	Description
config	Enter the unique identifier that is obtained in Step 1. Example: ##1.0##log-config-test$ecs-test.
dir	The directory in which historical log files are saved. Example: */data/logs*. **Note** The directory cannot end with a forward slash (/).
name	The name of the historical log file. The name can contain wildcards. Example: access.log.2018-08-08 and access.log*.

The following example shows how to configure a local event in Linux by using the `cat /usr/local/ilogtail/local_event.json` command:

```
[
  {
    "config": "##1.0##log-config-test$ecs-test",
    "dir": "/data/log",
    "name": "access.log*"
  },
  {
    "config": "##1.0##log-config-test$tmp-test",
    "dir": "/tmp",
    "name": "access.log.2017-08-09"
  }
]
```

FAQ

- How do I check whether Logtail loads a Logtail configuration?

  After you save the *local_event.json* file, Logtail loads the configurations of the local event to the memory within 1 minute. Then, the content of the *local_event.json* file is deleted.

  You can use the following methods to check whether the Logtail configuration is loaded.

i. If no content exists in the *local_event.json* file, Logtail reads the local event from the file.

ii. Check whether the *ilogtail.LOG* file in the Logtail installation directory contains the process local event parameter. If the content in the *local_event.json* file is cleared but the process local event parameter does not exist, the content of the *local_event.json* file may be invalid and filtered out.

- Why am I unable to collect a log file after a Logtail configuration is loaded?
  - The Logtail configuration is invalid.
  - The configurations of the local event in the *local_event.json* file are invalid.
  - The log file does not exist in the path that is specified in the Logtail configuration.
  - The log file has been collected by Logtail.

23.1.3.1.4.12. Log topics

Logs can be identified by log topics. When you collect logs, you can specify a topic for the logs.

You can specify a topic in the following scenarios: when you use Logtail to collect logs, when you call API operations, or when you use an SDK to upload log data. In the Log Service console, you can set the topic generation mode to Null - Do not generate topic, Machine Group Topic Attributes, or File Path RegEx.

- Null - Do not generate topic

  In this mode, the topic is an empty string. You can query logs without the need to specify a topic.

- Machine Group Topic Attributes

  You can use this mode to identify logs that are generated on different servers. If the logs are saved with the same file name or the logs are saved in the same directory, you can specify different topics to identify the logs.

  You can add servers to different machine groups, and configure different topic attributes for the machine groups. When you create a Logtail configuration, set **Topic Generation Mode** to **Machine Group Topic Attributes**. If Logtail sends the logs of a server in a machine group to Log Service, Logtail uploads the topic attributes of the machine group as topic names. You can use the topic attributes as filters to query logs.

- File Path RegEx

  You can use this mode to identify logs that are generated by different users or instances. Log Service stores logs in different directories for different users or instances. However, if duplicate sub-directory names or log file names exist in these directories, Log service cannot identify which user or instance generates the logs.

  To resolve this issue, you can set **Topic generation modes** to **File Path Regex** and enter the regular expression of the log file path when you create a Logtail configuration. The regular expression must match the log file path. When Logtail sends logs to Log Service, Logtail uploads the username or the instance name as the topic name. You can use the topic name as a filter to query logs.

  Logs that are generated by different users or instances may be stored in different files with the same name. Each file is stored in a different directory. For example, three log files are all named *service.log* and you only specify the *service.log* file in the */logs* directory as the log source when you collect logs from these files. After the logs are sent to Log Service, Log Service cannot identify which users or instances generate the logs. To resolve this issue, you can set **Topic Generation Mode** to **File Path RegEx**, and enter the `\/(.*)\/serviceA\/.*` regular expression. Then, Log Service generates the following topics for logs that are in different directories: userA, userB, and userC.

```
/logs
  | - /userA/serviceA
    | - service.log
  | - /userB/serviceA
    | - service.log
  | - /userC/serviceA
    | - service.log
```

> **Note** You must escape the forward slash (/) in the regular expression that is used to match file paths.

To extract multiple fields from a file path, you can use the `?P<key>` sub-expression to extract fields from the layers of the file path. The value of the key parameter can only contain lowercase letters and digits. Example:

```
/home/admin/serviceA/userB/access.log
\/home\/admin\/(?P<service>[^\/]+)/(?P<user>[^/]+)/.*
```

The following custom tags are created for logs:

```
"__tag__ : service : serviceA"
"__tag__ : user : userB"
```

- Static topic generation

  You can set **Topic Generation Mode** to **File Path RegEx**. In the **Custom RegEx** field, enter `customized:// + custom topic name`.

  > **Note** Static topic generation is supported by Logtail V0.16.21 (Linux) and later.

23.1.3.1.5. Custom plug-ins

Context

Log Service allows you to collect text logs and system logs by using Logtail. Logtail supports connections with multiple data sources, such as HTTP or MySQL query results and MySQL binary logs.

You can collect HTTP request data and upload the processing results to Log Service in real time to check service availability and perform continuous availability monitoring. You can configure MySQL query results as the data source, and then synchronize incremental data based on custom IDs or time. You can also configure an SQL data source to synchronize MySQL binary logs, subscribe to database changes, and query or analyze logs in real time.

> **Note** This feature supports only Logtail V0.16.0 or later that runs on Linux. For more information about Logtail versions and version updates, see Install Logtail in Linux.

Configuration process

1. Configure a collection method for a data source.

   Configure collection methods based on different data sources.

2. Configure a data processing method.

   Logtail provides multiple processing methods for binary logs, MySQL query results, NGINX monitoring data, and HTTP input sources. You can configure multiple processing methods for a single input source. Each input source supports all processing methods. Logtail runs the configured processing methods in sequence.

   For more information, see Configure data processing methods.

3. Apply the configurations to the specified machine group.

   Apply the log collection configurations and processing configurations to the specified machine group. Then, Logtail automatically pulls the configurations and starts to collect logs.

23.1.3.1.5.1. Collect MySQL binary logs

Logtail is used as a MySQL slave. It is used to collect binary logs from a MySQL master. Logtail collects binary logs by using a similar method to Alibaba Canal. This improves the efficiency of log collection.

Features

- Allows you to collect incremental data of databases in the form of binary logs to improve performance. Supports MySQL databases such as ApsaraDB RDS for MySQL.
- Supports multiple database filters, such as regular expressions.
- Allows you to set binary log file positions.
- Allows you to record synchronization statuses by using the checkpoint mechanism.

Limits

- MySQL binary logs are available only for Logtail 0.16.0 or later versions that you install on Linux. For more information about how to update Logtail and view Logtail versions, see Install Logtail in Linux.
- Binary logs in the ROW format must be enabled for MySQL databases. By default, binary logs in the ROW format are enabled for RDS instances.

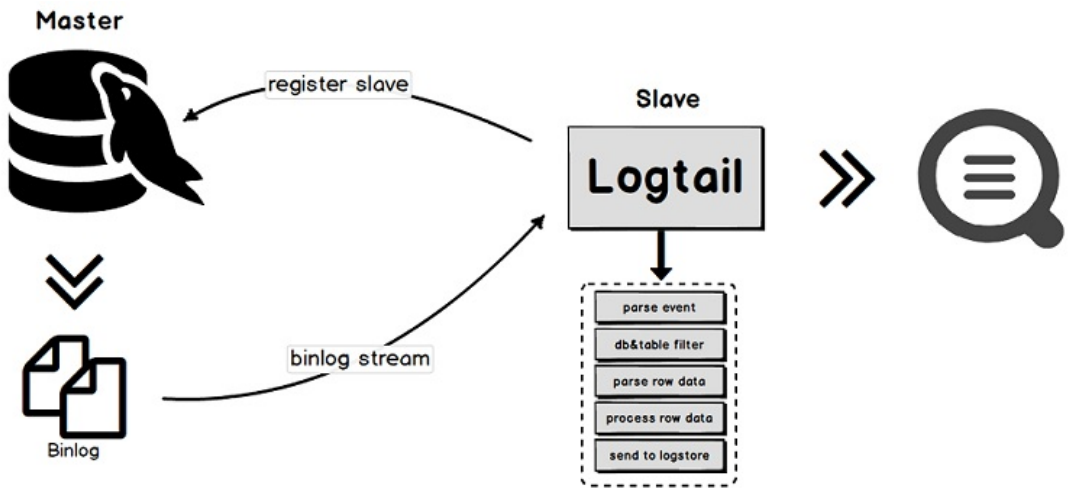
```
# Check whether binary logs are enabled.
mysql> show variables like "log_bin";
+---------------+-------+
| Variable_name | Value |
+---------------+-------+
| log_bin       | ON    |
+---------------+-------+
1 row in set (0.02 sec)
# View the format of binary logs.
mysql> show variables like "binlog_format";
+---------------+-------+
| Variable_name | Value |
+---------------+-------+
| binlog_format | ROW   |
+---------------+-------+
1 row in set (0.03 sec)
```

- Each server ID must be unique. Make sure that the ID of each slave to be synchronized is unique.
- Limits for RDS databases:
  - Logtail cannot be installed on an RDS instance. You must install Logtail on an ECS instance that can communicate with the destination RDS instance.
  - Secondary RDS databases cannot be used to collect binary logs. You must configure a primary RDS database to collect binary logs.

Implementation

Logtail enables communication between master and slave MySQL servers. The process of how master and slave MySQL servers communicate is provided in the following information:

1. Logtail is used as a MySQL slave. It can also be used to send dump requests to the MySQL master.
2. After the dump requests are received, the MySQL master delivers its binary logs to Logtail in real time.
3. Logtail parses and filters binary logs, and then uploads the results to Log Service.



Scenarios

The MySQL binary logging feature applies to scenarios in which you need to synchronize large amounts of data and meet high performance requirements.

- Query the incremental data of databases in real time.
- Audit operations that are performed on databases.
- Use Log Service to query data, visualize query results, transform data for stream processing, export data to MaxCompute for offline computing, and export log to Object Storage Service (OSS) for long-term storage.

Data reliability

We recommend that you enable the global transaction identifier (GTID) feature of the MySQL server and upgrade Logtail to version 0.16.15 or later. This prevents repeated data collection during a primary/secondary server switchover and ensures data reliability.

- Incomplete data collection: If the network between Logtail and the MySQL server is disconnected for a long period of time, some data may not be collected.

  A MySQL binary log plug-in is used as a MySQL slave to collect binary logs from the master server. Logtail establishes a connection with the master server to obtain data from the server. If the network between Logtail and the master node is disconnected, the master node still generates new binary logs and deletes expired binary logs. After the connection is reestablished and Logtail is reconnected to the master server, Logtail uses the last checkpoint to request binary log data from the master server. However, if the network is disconnected for a long period of time, the data generated after the checkpoint may be deleted. In this case, the recovery mechanism specifies the new point at which Logtail resumes collecting binary logs. The new point is the most recent binary log file position. If the network is disconnected for a long period of time, some data generated between the checkpoint and the new data collection point may not be collected.

- Repeated data collection: If the ordinal numbers of binary logs on the master and slave servers are different and a master/slave switchover occurs, repeated data collection may occur.

  If the MySQL master-slave synchronization is configured, the master server synchronizes the generated binary log data to the slave server. Then, the slave server stores the received binary log data to the local binary log file. If the ordinal numbers of binary logs on the master and slave servers are different, a master/slave switchover occurs. In this case, the mechanism that uses a binary log file name and an offset as the checkpoint causes repeated data collection.

For example, assume that a data entry ranges from `(binlog.100, 4)` to `(binlog.105, 4)` on the master server, and ranges from `(binlog.1000, 4)` to `(binlog.1005, 4)` on the slave. Logtail has obtained the data from the master server and updated the checkpoint to `(binlog.105, 4)`. In this case, if a master/slave switchover occurs without exception, Logtail continues to obtain binary logs from the new master server based on the local checkpoint `(binlog.105, 4)`. The new master server returns the data entries that range from `(binlog.1000, 4)` to `(binlog.1005, 4)` to Logtail. This is because the ordinal numbers of these data entries on the new master server are greater than the ordinal numbers of data entries requested by Logtail. As a result, log data is repeatedly collected.

Parameter

The type of input sources is `service_canal`.

Parameter	Type	Required	Description
Host	string	No	The IP address of the host where the database resides. Default value: 127.0.0.1.
Port	int	No	The port number that you can use to connect with the database. Default value: 3306.

User	string	No	The database username. Default value: root.
			The configured user must have the read permissions on the source database and the MySQL REPLICATION permission. Example:
			`CREATE USER canal IDENTIFIED BY 'canal';`
			`GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT`
			`ON *. * TO 'canal'@'%';`
			`-- GRANT ALL PRIVILEGES ON *. * TO 'canal'@'%' ;`
			`FLUSH PRIVILEGES;`

Password	string	No	The database password. By default, this parameter is unspecified. If you require a high level of data security, we recommend that you set both the username and the password to xxx. After your configurations are synchronized to the on-premises server, find the Password parameter in the */usr/local/ilogtail/user_log_config.json* file and modify the value. **Note** • After you modify the password, use the **sudo /etc/init.d/ilogtaild stop; sudo /etc/init.d/ilogtaild start** command to restart Logtail. • If you modify the value of the Password parameter on the web and synchronize your configurations to the on-premises server, the configurations on the on-premises server are overwritten. You can change the configurations on the on-premises server later.

Parameter	Type	Required	Description
ServerID	int	No	The ID of a MySQL slave whose role is assumed by Logtail. Default value: 125. **Note** In a MySQL database, each ID must be unique. Otherwise, synchronization fails.
IncludeTables	String array	Yes	The names of matched tables. Each value contains a database name and a table name, for example, `test_db.test_table`. You must specify a regular expression for the parameter. Logtail does not collect incremental data from tables whose names do not match the regular expression. To collect incremental data from all tables of a database, set the value of the IncludeTables parameter to `.*\\..*`. **Note** If an exact match is required, make sure that the regular expression is prefixed by `^`. In this case, you must also make sure that the regular expression is suffixed by `$`. Example: `^test_db\\.test_table$`.
ExcludeTables	String array	No	The names of excluded tables expressed as a regular expression. The name of a table must include the name of the database to which the table belongs, for example, `test_db.test_table`. If a table meets one of the conditions specified in the parameter, the table is not collected. If you do not specify this parameter, incremental data from all tables is collected. **Note** If an exact match is required, make sure that the regular expression is prefixed by `^`. In this case, you must also make sure that the regular expression is suffixed by `$`. Example: `^test_db\\.test_table$`.

Parameter	Type	Required	Description

StartBinName	string	No	The name of the first binary log file that is collected by Logtail. If you do not specify this parameter, Logtail starts to collect binary log files that are generated from the current time.
			To collect data from a specific location, view the name of the current binary log file and the file offset. Then, set `StartBinName` and `StartBinLogPos` to actual values. Example:
			`# Set StartBinName to mysql-bin.000063 and StartBinLogPos to 0.`
			`mysql> show binary logs;`
			`+------------------+-----------+`
			`| Log_name | File_size |`
			`+------------------+-----------+`
			`| mysql-bin.000063 | 241 |`
			`| mysql-bin.000064 | 241 |`
			`| mysql-bin.000065 | 241 |`
			`| mysql-bin.000066 | 10778 |`
			`+------------------+-----------+`
			`4 rows in set (0.02 sec)`
			Note If you set the `StartBinName` parameter, a large amount of traffic is generated during the first collection.

StartBinLogPos	int	No	The offset of the first binary log file that is collected. Default value: 0.
EnableGTID	bool	No	Specifies whether to add GTID. Default value: true. If the value is false, no GTID is added to uploaded data.
EnableInsert	bool	No	Specifies whether to collect log events triggered by INSERT operations. Default value: true. If the value is false, INSERT events are not collected.
EnableUpdate	bool	No	Specifies whether to collect UPDATE events. Default value: true. If the value is false, UPDATE events are not collected.
EnableDelete	bool	No	Specifies whether to collect DELETE events. Default value: true. If the value is false, DELETE events are not collected.
EnableDDL	bool	No	Specifies whether to collect data definition language (DDL) events. Default value: false. If the value is false, DDL events are not collected. **Note** This parameter does not support the `IncludeTables` or `ExcludeTables` filtering methods.
Charset	string	No	The encoding method. Default value: `utf-8`.
TextToString	bool	No	Specifies whether to convert data of the text type into a string. Default value: false.

Parameter	Type	Required	Description
PackValues	bool	No	Specifies whether to encapsulate event data into the JSON format. Default value: false. If the value is false, event data is not encapsulated. If this feature is enabled, Logtail encapsulates event data into the data and old_data fields in the JSON format. The old_data field is available only for ROW_UPDATE events. For example, assume that a table has three fields named c1, c2, and c3. If this feature is disabled, the ROW_INSERT event data contains three fields c1, c2, and c3. If this feature is enabled, c1, c2, and c3 are encapsulated into one data field and the value is `{"c1":"...", "c2": "...", "c3": "..."}`. **Note** This parameter is available only for Logtail V0.16.19 and later.
EnableEventMeta	bool	No	Specifies whether to collect event metadata. Default value: false. If the value is false, event metadata is not collected. The metadata of binary log events includes `event_time`, `event_log_position`, `event_size`, and `event_server_id`. **Note** This parameter is available only for Logtail V0.16.21 and later.

Procedure

Synchronize data from tables whose names do not end with `_inner` in the `user_info` RDS database.

1. Log on to the Log Service console.
2. Select a data source.

   Click **Import Data**. On the **Import Data** page, select **MYSQL BinLog**.
3. Select a destination project and Logstore, and then click **Next**.

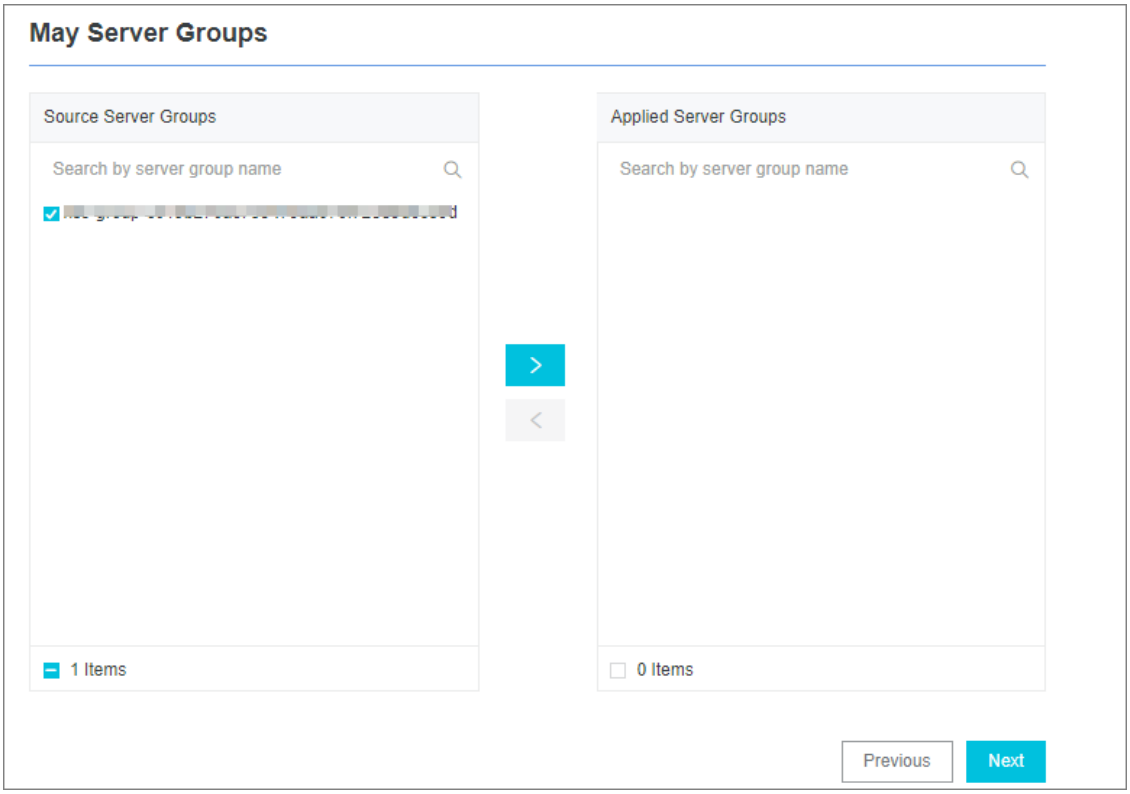
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Configure the data source.

   Set the Config Name and Plug-in Config fields.

   In the **Plug-in Config** field, modify the parameter settings in the default configuration template based on your business requirements.

```
{
 "inputs": [
     {
         "type": "service_canal",
         "detail": {
             "Host": "************.mysql.rds.aliyuncs.com",
             "Port": 3306,
             "User" : "root",
             "ServerID" : 56321,
             "Password": "*******",
             "IncludeTables": [
                 "user_info\\.. *"
             ],
             "ExcludeTables": [
                 ". *\\. \\S+_inner"
             ],
             "TextToString" : true,
             "EnableDDL" : true
         }
     }
 ]
}
```

- *inputs*: specifies the collection configurations. This parameter is required. You must configure statements to collect data based on your data source.
- *processors*: specifies the processing method. This parameter is optional. For more information about how to set a processing method, see Configure data processing methods.

7. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

8. (Optional)Modify the configurations on the server.

   If you do not enter the actual URL, username, or password in **Plug-in Config**, you must replace them with actual values after the configurations are synchronized to the server.

   > **Note** If you have entered the actual information, skip this step.

   i. Log on to the server where Logtail is installed, find the `service_canal` keyword in the `/usr/local/ilogtail/user_log_config.json` file, and then set related fields. These fields include `Host`, `User`, and `Password`.

   ii. Run the following command to restart Logtail:

   ```
   sudo /etc/init.d/ilogtaild stop; sudo /etc/init.d/ilogtaild start
   ```

The configurations that are used to collect binary logs are completed. If changes are made to your database, Logtail immediately collects the updated data and uploads the data to Log Service.

> **Note** Logtail collects incremental binary logs. If no data is collected, check whether changes are made to the table in your database after the configurations are updated.

Metadata fields

When you collect binary logs, some metadata is also uploaded. The following table lists the fields of uploaded metadata.

Parameter	Description	Example
`_host_`	The name of the host where the database resides.	`*********.mysql.rds.aliyuncs.com`
`_db_`	The name of the RDS database.	`my-database`
`_table_`	The name of the table.	`my-table`
`_event_`	The type of the event. Valid values:	`row_update`, `row_insert`, and `row_delete`

Parameter	Description	Example
`_id_`	The ID of the current collection. The value starts from 0 and increments by 1 each time a binary log event is collected.	`1`
`_gtid_`	The GTID.	`7d2ea78d-b631-11e7-8afb-00163e0eef52:536`
`_filename_`	The name of the binary log file.	`binlog.001`
`_offset_`	The offset of the binary log file. The value is updated after each COMMIT operation.	`12876`

Example

After you completed the preceding steps to set a processing method, perform `INSERT`, `UPDATE`, and `DELETE` operations on the `SpecialAlarm` table in the `user_info` database. The following information shows the schema, database operations, and sample logs that are collected by Logtail.

- Schema

```
CREATE TABLE `SpecialAlarm` (
`id` int(11) unsigned NOT NULL AUTO_INCREMENT,
`time` datetime NOT NULL,
`alarmtype` varchar(64) NOT NULL,
`ip` varchar(16) NOT NULL,
`count` int(11) unsigned NOT NULL,
PRIMARY KEY (`id`),
KEY `time` (`time`) USING BTREE,
KEY `alarmtype` (`alarmtype`) USING BTREE
) ENGINE=MyISAM AUTO_INCREMENT=1;
```

- Database operations

  Perform the INSERT, DELETE, and UPDATE operations on the database.

```
insert into specialalarm (`time`, `alarmType`, `ip`, `count`) values(now(), "NO_ALARM", "10.10. **.
***", 55);
delete from specialalarm where id = 4829235  ;
update specialalarm set ip = "10.11. ***.**" where id = "4829234";
```

  Create an index for `zc.specialalarm`.

```
ALTER TABLE `zc`.`specialalarm`
ADD INDEX `time_index` (`time` ASC);
```

- Sample logs

  On the data preview or Search & Analysis page, you can view a sample log that corresponds to each operation.

- INSERT statement

```
__source__:  10.30. **.**
__tag__:__hostname__:  iZbp145dd9fccu*****
__topic__:
_db_:  zc
_event_:  row_insert
_gtid_:  7d2ea78d-b631-11e7-8afb-00163e0eef52:536
_host_:  *********.mysql.rds.aliyuncs.com
_id_:  113
_table_:  specialalarm
alarmtype:  NO_ALARM
count:  55
id:  4829235
ip:  10.10. ***.***
time:  2017-11-01 12:31:41
```

- DELETE statement

```
__source__:  10.30. **.**
__tag__:__hostname__:  iZbp145dd9fccu****
__topic__:
_db_:  zc
_event_:  row_delete
_gtid_:  7d2ea78d-b631-11e7-8afb-00163e0eef52:537
_host_:  *********.mysql.rds.aliyuncs.com
_id_:  114
_table_:  specialalarm
alarmtype:  NO_ALARM
count:  55
id:  4829235
ip:  10.10. **.***
time:  2017-11-01 12:31:41
```

- UPDATE statement

```
__source__:  10.30. **.**
__tag__:__hostname__:  iZbp145dd9fccu****
__topic__:
_db_:  zc
_event_:  row_update
_gtid_:  7d2ea78d-b631-11e7-8afb-00163e0eef52:538
_host_:  *********.mysql.rds.aliyuncs.com
_id_:  115
_old_alarmtype:  NO_ALARM
_old_count:  55
_old_id:  4829234
_old_ip:  10.10.22.133
_old_time:  2017-10-31 12:04:54
_table_:  specialalarm
alarmtype:  NO_ALARM
count:  55
id:  4829234
ip:  10.11. ***.***
time:  2017-10-31 12:04:54
```

- DDL statement

```
__source__:  10.30. **.**
__tag__:__hostname__:  iZbp145dd9fccu****
__topic__:
_db_:  zc
_event_:  row_update
_gtid_:  7d2ea78d-b631-11e7-8afb-00163e0eef52:539
_host_:  *********.mysql.rds.aliyuncs.com
ErrorCode:  0
ExecutionTime:  0
Query:  ALTER TABLE `zc`.`specialalarm`
ADD INDEX `time_index` (`time` ASC)
StatusVars:
```

Usage notes

We recommend that you increase resource limits on Logtail to process traffic surges and prevent data security risks. If the limits are exceeded, Logtail may be forcibly restarted.

You can modify the resource limits in the */usr/local/ilogtail/ilogtail_config.json* file. Then, you can run the **sudo /etc/init.d/ilogtaild stop;sudo /etc/init.d/ilogtaild start** command to restart Logtail.

The following example shows how to set the CPU limit to two and memory limit to 2,048 MB:

```
{
    ...
    "cpu_usage_limit":2,
    "mem_usage_limit":2048,
    ...
}
```

23.1.3.1.5.2. Collect MySQL query results

This topic describes how to create a Logtail configuration in the Log Service console to collect MySQL query results.

Prerequisites

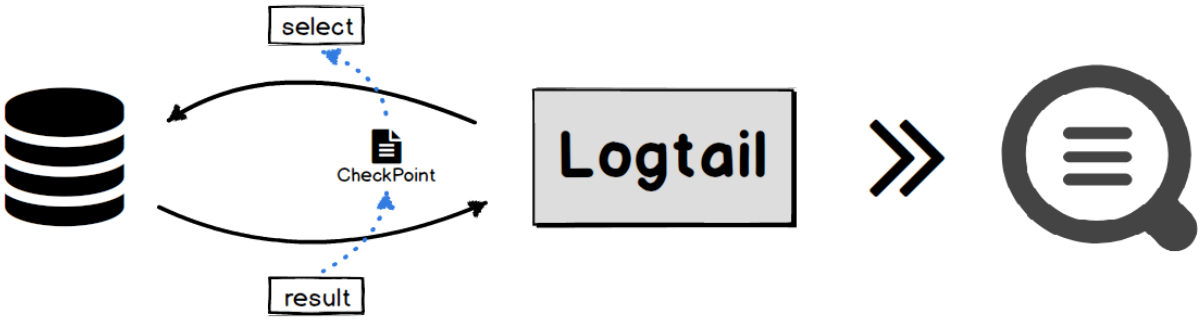
Logtail is installed on the server from which you want to collect MySQL query results. For more information, see Install Logtail in Linux or Install Logtail in Windows.

> **Note** This feature applies only to Logtail V0.16.0 or later that runs on Linux and Logtail V1.0.0.8 or later that runs on Windows.

Implementation

Logtail executes the specified SELECT statement on a regular basis based on the Logtail configuration that you create, and then uploads the query results to Log Service.

When Logtail obtains a query result, Logtail saves the value of the CheckPoint field on the on-premises server. When Logtail executes the SELECT statement again, Logtail adds the value of the CheckPoint field to the SELECT statement. This way, Logtail can collect the incremental data of MySQL databases.



Features

- Supports MySQL databases.
- Allows you to paginate query results.
- Allows you to specify time zones.
- Allows you to specify timeout periods.
- Allows you to use checkpoints to record data synchronization status.
- Supports SSL.
- Allows you to specify the maximum size of data that can be collected at a time.

Scenarios

- Collect incremental data based on specific marks such as an auto-increment ID or a point in time.
- Customize data synchronization based on specified filtering conditions.

Procedure

The following procedure describes how to synchronize incremental data from a MySQL database to Log Service. In this procedure, the logtail.VersionOs field is synchronized every 10 seconds and the value of the count parameter in this field is greater than 0. The value of the initial checkpoint is 2017-09-25 11:00:00. Logs are paginated and each page contains 100 logs. The checkpoint of each page is saved. The procedure includes the following steps:

1. Log on to the Log Service console.
2. Select a data source.

   Click **Import Data**. On the **Import Data** page, select **MySQL Query Result - Plug-in**.
3. Select a destination project and Logstore, and then click **Next**.

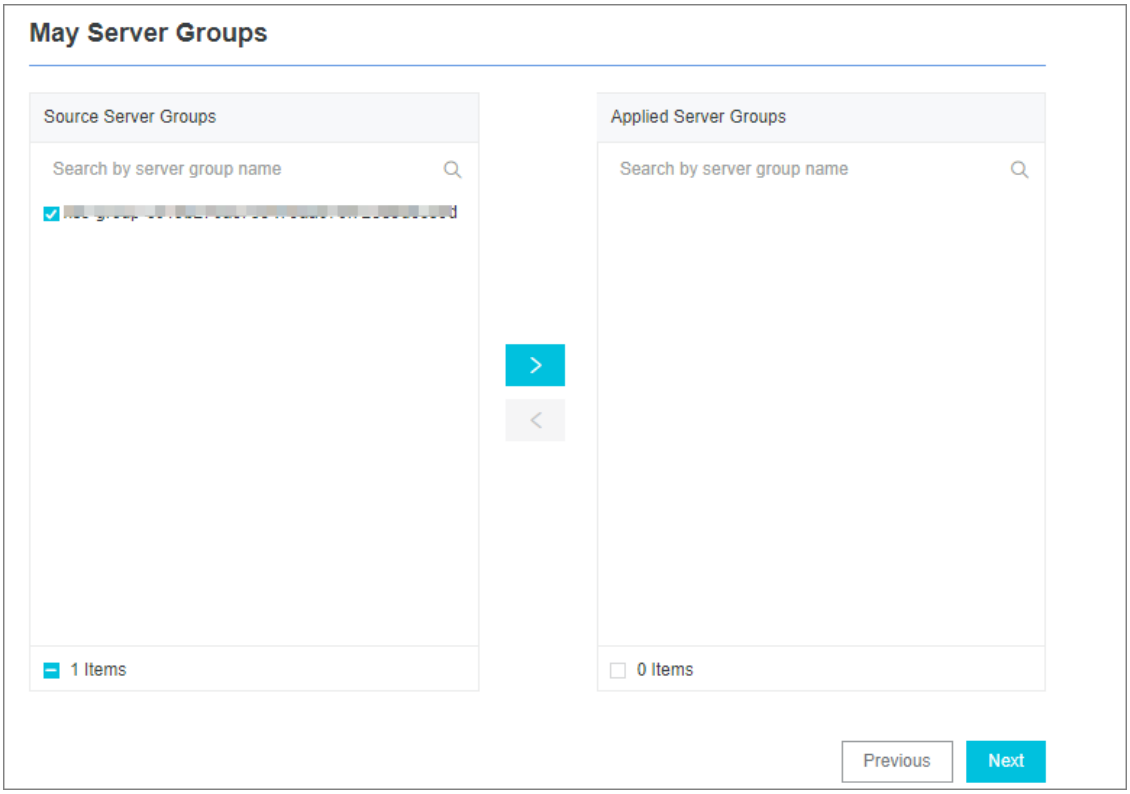
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. In the **Specify Data Source** step, set the **Config Name** and **Plug-in Config** parameters.
   - In the **Plug-in Config** field, modify the parameter settings in the default configuration template based on your business requirements.
   - `inputs`: specifies the collection configurations. This parameter is required. `processors`: specifies the processing method. This parameter is optional. You must specify statements to collect data based on your data source. For more information, see Configure data processing methods.

Note If you have high requirements for data security, we recommend that you set the username and the password to xxx. After your configurations are synchronized to the on-premises server, find the Password parameter in the /usr/local/ilogtail/user_log_config.json file and change the value.

The following example shows the configurations:

```
{
  "inputs": [
    {
      "type": "service_mysql",
      "detail": {
        "Address": "************.mysql.rds.aliyuncs.com",
        "User": "****",
        "Password": "*******",
        "DataBase": "****",
        "Limit": true,
        "PageSize": 100,
        "StateMent": "select * from db.VersionOs where time > ?",
        "CheckPoint": true,
        "CheckPointColumn": "time",
        "CheckPointStart": "2018-01-01 00:00:00",
        "CheckPointSavePerPage": true,
        "CheckPointColumnType": "time",
        "IntervalMs": 60000
      }
    }
  ]
}
```

Parameter	Type	Required	Description
type	string	Yes	The type of the data source. Set the value to service_canal.
Address	string	No	The address of the MySQL database. Default value: 127.0.0.1:3306.
User	string	No	The username of the MySQL database. Default value: root.
Password	string	No	The password of the MySQL database. By default, this parameter is left empty. If you have high requirements for data security, we recommend that you set the username and the password to xxx. After your configurations are synchronized to the on-premises server, find the Password parameter in the */usr/local/ilogtail/user_log_config.json* file and change the value. **Note** If you modify this parameter in the console, the on-premises configuration is overwritten after the modification is synchronized to the on-premises server.
DialTimeOutMs	int	No	The timeout period for the database connection. Unit: milliseconds. Default value: 5000.
ReadTimeOutMs	int	No	The timeout period for data reading. Unit: milliseconds. Default value: 5000.

Parameter	Type	Required	Description
StateMent	string	No	The SQL statement. If you set the CheckPoint parameter to true, you must include the CheckPointColumn parameter in the WHERE clause of the SQL statement that you specify for the StateMent parameter. You must also set the CheckPointColumn parameter to ?. For example, if you set the CheckPointColumn parameter to id, you must set the StateMent parameter in format of `SELECT * from ... where id > ?` .
Limit	bool	No	Specifies whether to paginate query results. Default value: false. This value indicates that query results are not paginated. We recommend that you set the Limit parameter to true. If you set the Limit parameter to true, a LIMIT clause is automatically appended to the SQL statement that you specify for the StateMent parameter when you execute the SQL statement.
PageSize	int	No	The number of logs to return on each page. If you set the Limit parameter to true, you must specify this parameter.
MaxSyncSize	int	No	The maximum number of logs that can be synchronized at a time. Default value: 0. This value indicates that no limit is placed on the size of data that can be synchronized at a time.
CheckPoint	bool	No	Specifies whether to use checkpoints during data collection. Default value: false. This value indicates that checkpoints are not used during data collection.
CheckPointColumn	string	No	The name of the checkpoint column. If you set the CheckPoint parameter to true, you must specify this parameter.
CheckPointColumnType	string	No	The type of the checkpoint column. Valid values: int and time. If you set this parameter to int, the values in the checkpoint column are stored as 64-bit integers. If you set this parameter to time, the values in the checkpoint column can be of the date, time, or datetime type supported by MySQL. If you set the CheckPoint parameter to true, you must specify this parameter.
CheckPointStart	string	No	The initial value of the checkpoint. If you set the CheckPoint parameter to true, you must specify this parameter.
CheckPointSaveEachPage	bool	No	If you set this parameter to true, a checkpoint is saved after each pagination. If you set this parameter to false, a checkpoint is saved after each synchronization.

Parameter	Type	Required	Description
IntervalMs	int	Yes	The synchronization interval. Unit: milliseconds.

7. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

Modify the configurations on the server where Logtail is installed.

If you do not enter the actual address, username, or password in the **Plug-in Config** field, you can modify the parameters after the configurations are synchronized to the server where Logtail is installed.

1. Log on to the server where Logtail is installed.
2. Find the service_canal keyword in the */usr/local/ilogtail/user_log_config.json* file and modify the Address, User, and Password parameters.
3. Run the following command to restart Logtail:

```
sudo /etc/init.d/ilogtaild stop; sudo /etc/init.d/ilogtaild start
```

Example

After Logtail sends MySQL query results to Log Service, you can view the results in the Log Service console. This section shows the schema and data that are collected by Logtail.

- Schema

```
CREATE TABLE `VersionOs` (
  `id` int(11) unsigned NOT NULL AUTO_INCREMENT COMMENT 'id',
  `time` datetime NOT NULL,
  `version` varchar(10) NOT NULL DEFAULT '',
  `os` varchar(10) NOT NULL,
  `count` int(11) unsigned NOT NULL,
  PRIMARY KEY (`id`),
  KEY `timeindex` (`time`)
)
```

- Sample log

```
"count":  "4"
"id:  "721097"
"os:  "Windows"
"time:  "2017-08-25 13:00:00"
"version":  "1.3.0"
```

23.1.3.1.5.3. Collect syslogs

This topic describes how to use the syslog plug-in of Logtail to collect syslogs from a server.

Prerequisites

Logtail 0.16.13 or a later version is installed on the server.

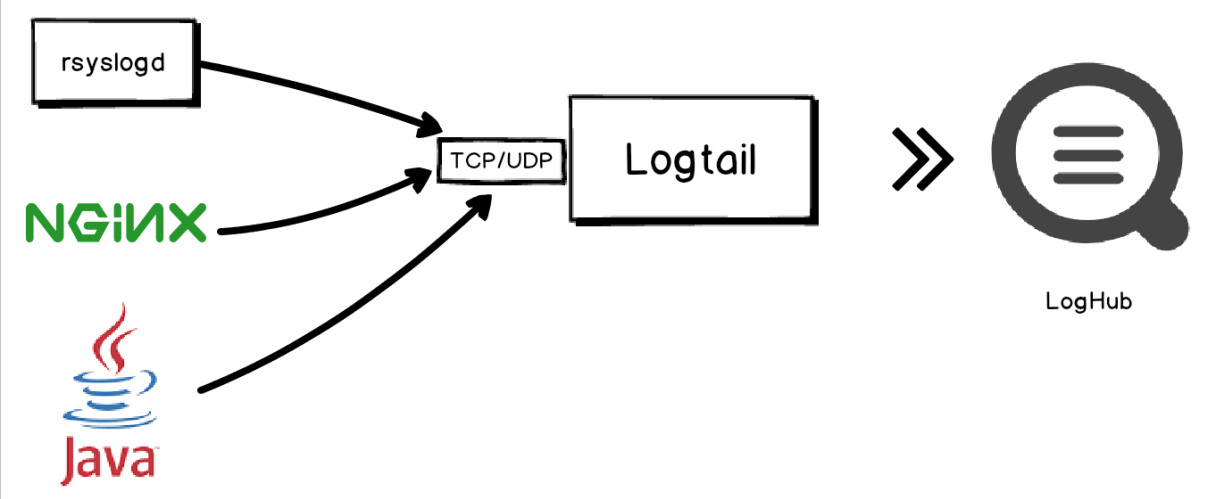
Overview

On a Linux server, local syslogs can be forwarded to the IP address and port of a specified server by using syslog agents such as rsyslog. After you create a Logtail configuration for the specified server, the syslog plug-in of Logtail receives the syslogs over TCP or UDP. In addition, the syslog plug-in parses the received syslogs and extract log fields such as facility, tag (program), severity, and content based on the specified syslog protocol. The syslog protocol can be RFC 3164 or RFC 5424.

> **Note**
>
> - Logtail installed on a Windows server does not support the syslog plug-in.
> - You can configure multiple syslog plug-ins for Logtail. For example, you can use both TCP and UDP to listen on 127.0.0.1:9999.

Implementation

After the syslog plug-ins start to listen on a specified IP address and port, Logtail can act as a syslog server to collect syslogs from various data sources. These syslogs include system logs collected by rsyslog, access or error logs forwarded by NGINX, and logs forwarded by syslog clients in languages such as Java.



Logtail parameters

The following table describes Logtail parameters. The type of the input is `service_syslog`.

Parameter	Type	Required	Description

Parameter	Type	Required	Description
Address	String	No	The protocol, address, and port on which the syslog plug-in listens. The syslog plug-in obtains logs based on the value of this parameter. Format: `[tcp/udp]://[ip]:[port]`. Default value: `tcp://127.0.0.1:9999`. **Note** • The specified protocol, address, and port must be the same as those specified for the forwarding rule in the rsyslog configuration file. • If the server on which Logtail is installed has multiple IP addresses, you can set the IP address to 0.0.0.0. It means that the syslog plug-in listens on all IP addresses of the server.
ParseProtocol	String	No	The protocol that is used to parse logs. This parameter is not specified by default, indicating that logs are not parsed. Valid values: • rfc3164: The RFC 3164 protocol is used to parse logs. • rfc5424: The RFC 5424 protocol is used to parse logs. • auto: The syslog plug-in selects a protocol based on the log content.
IgnoreParseFailure	Boolean	No	Specifies whether to ignore a parsing failure. Default value: true. Valid values: • true: Logs that fail to be parsed are included in the returned content field. • false: Logs that fail to be parsed are dropped.

Default fields

Field	Type	Description
hostname	String	The hostname. If a hostname is not provided in the log, the hostname of the current host is obtained.
program	String	The tag field in the protocol.
priority	String	The priority field in the protocol.
facility	String	The facility field in the protocol.
severity	String	The severity field in the protocol.

Field	Type	Description
unixtimestamp	String	The timestamp of the log.
content	String	The log content. If the log fails to be parsed, this field contains the complete content of the log.
_ip_	String	The IP address of the current host.

Configure the plug-in of Logtail to collect syslogs

1. Add a forwarding rule for rsyslog.

   Modify the */etc/rsyslog.conf* rsyslog configuration file on the server from which syslogs are collected. Add a forwarding rule at the end of the configuration file. Then, rsyslog forwards syslogs to the specified IP address and port.

   - If you want to collect syslogs of the server by using Logtail on this server, set the forwarding address to 127.0.0.1 and the port to an idle port.
   - If you want to collect syslogs of the server by using Logtail on a second server (Server B), set the forwarding address to the public IP address of the second server and port to an idle port.

   For example, the following forwarding rule indicates that logs are forwarded to 127.0.0.1:9000 over TCP.

   ```
   *. * @@127.0.0.1:9000
   ```

2. Run the following command to restart rsyslog and validate the log forwarding rule:

   ```
   sudo service rsyslog restart
   ```

3. Log on to the Log Service console.
4. Select the data source **Custom Data Plug-in**.

   You can use one of the following three methods to select a data source:

   - On the homepage of the Log Service console, select a data source in the **Import Data** section.
   - In the **Projects** section, click a project name. On the **Overview** page, click **Import Data**, and then select a data source.
   - On the Logstores tab in the left-side navigation pane, find a Logstore and click the closing angle bracket (>) in front of the Logstore name. Click the plus sign (+) next to **Data Import**, and then select a data source.

5. Select a destination project and Logstore, and then click **Next**.

   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.

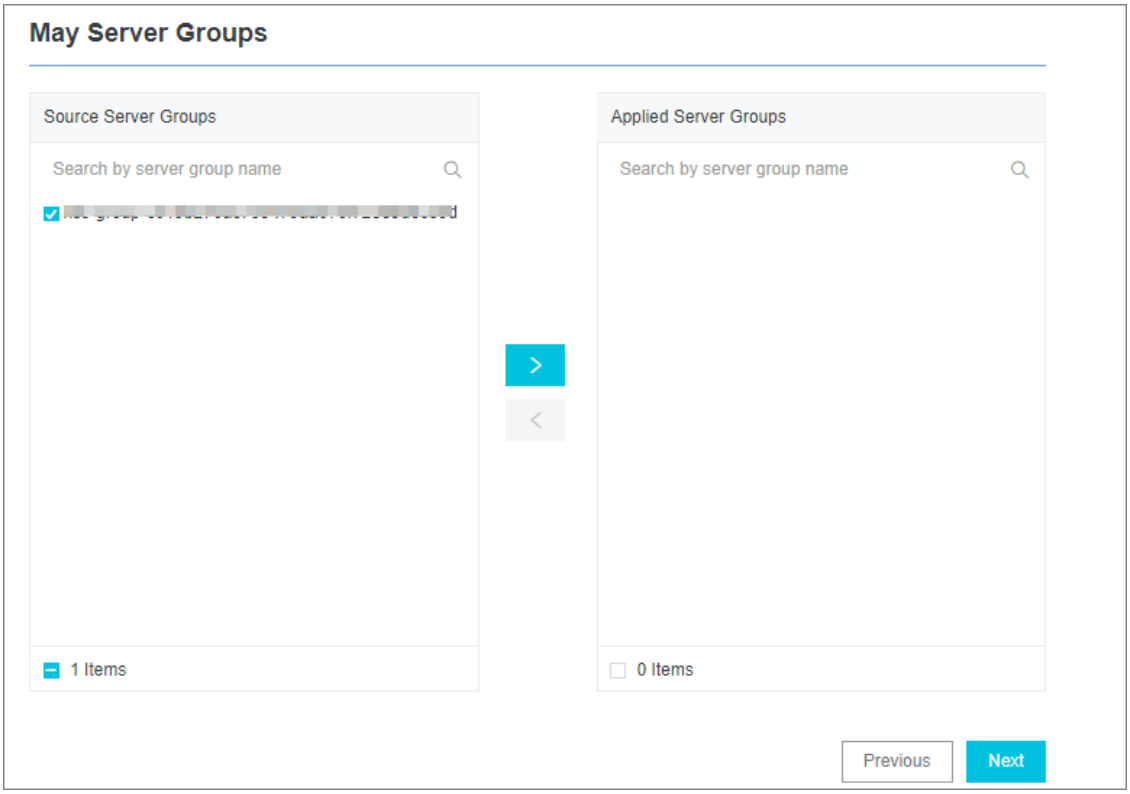
6. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.

7. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

8. Configure the data source.

   Set **Config Name** and **Plug-in Config**.

   The `inputs` section is required. It specifies the collection configuration. The `processors` section is optional. It specifies the processing configuration. You must specify a collection statement for the collection configuration based on the data source. You can specify one or more processing methods for the processing configuration. For more information, see Configure data processing methods.

   The following sample code shows how to use UDP and TCP to listen on 127.0.0.1:9000:

```
{
    "inputs": [
        {
            "type": "service_syslog",
            "detail": {
                "Address": "tcp://127.0.0.1:9000",
                "ParseProtocol": "rfc3164"
            }
        },
        {
            "type": "service_syslog",
            "detail": {
                "Address": "udp://127.0.0.1:9001",
                "ParseProtocol": "rfc3164"
            }
        }
    ]
}
```

9. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

Configure the plug-in of Logtail to collect NGINX logs

NGINX access logs can be forwarded to specified addresses and ports over the syslog protocol. To deliver NGINX access logs as syslogs from a server to Log Service, you can create a Logtail configuration and apply it to the server group to which the server belongs.

1. Add a forwarding rule to the *nginx.conf* configuration file on the NGINX server.

   For example, add the following content to the configuration file.

   ```
   http {
       ...
       # Add this line.
       access_log syslog:server=127.0.0.1:9000,facility=local7,tag=nginx,severity=info combined;
       ...
   }
   ```

2. Run the following command to restart the NGINX service and validate the configuration.

   ```
   sudo service nginx restart
   ```

3. Create a Logtail configuration and apply it to the server group to which the server belongs.

   For more information, see Configure the plug-in of Logtail to collect syslogs.

4. Check whether the Logtail configuration takes effect.

   Run the `curl http://127.0.0.1/test.html` command in shell to generate an access log. If the Logtail configuration takes effect, you can view the log information on the query page of the Log Service console.

23.1.3.1.5.4. Customize Logtail plug-ins to process data

If you have complex logs that cannot be parsed in basic modes such as full regex, NGINX, and JSON, you can use Logtail plug-ins to parse the logs. You can configure Logtail plug-ins for one or more processing methods. Then, Logtail executes the processing methods in sequence.

Limits

- Performance limits

  If a plug-in is used to process data, Logtail consumes more resources. Most of these resources are CPU resources. You can modify the Logtail parameter settings based on your business requirements. For more information, see Set Logtail startup parameters.

- Limits on text logs

  Log Service allows you to process text logs in basic modes such as full regex, NGINX, or JSON. Log Service also allows you to use Logtail plug-ins to process text logs. However, Logtail plug-ins have the following limits on text logs:

- If you enable the plug-in processing feature, some advanced features of the specified mode become unavailable. For example, you cannot configure the filter, upload raw logs, specify the system time zone, drop logs that fail to be parsed, or upload incomplete logs in delimiter mode.
- Plug-ins use the line mode to process text logs. In this mode, file-level metadata such as __tag__:__path__ and __topic__ is stored in each log. If you use Logtail plug-ins to process data, the following limits apply to tag-related features:
  - You cannot use the contextual query and LiveTail features because these features depend on fields such as __tag__:__path__.
  - The name of the __topic__ field is renamed to __log_topic__.
  - Fields such as __tag__:__path__ no longer have original field indexes. You must configure indexes for these fields.

Usage notes

When you configure data processing methods, you must set the key in the configuration file to processors and set the value to an array of JSON objects. Each object of the array contains the details of a processing method.

Each processing method contains the type and detail fields. The type field specifies the type of the processing method and the detail field contains configuration details.

```
"processors" : [
      {
          "type" : "processor_split_char",
          "detail" : {"SourceKey" : "content",
              "SplitSep" : "|",
              "SplitKeys" : ["method", "type", "ip", "time", "req_id", "size", "detail"]
          }
      },
      {
          "type" : "processor_anchor",
          "detail" : "SourceKey" : "detail",
              "Anchors" : [
                  {
                      "Start" : "appKey=",
                      "Stop" : ",env=",
                      "FieldName" : "appKey",
                      "FieldType" : "string"
                  }
              ]
          }
  ]
```

The following table describes the Logtail plug-ins that are available and the operations that you can perform by using these plug-ins.

Logtail plug-in	Description
processor_regex	You can use the processor_regex plug-in to extract the fields that match a specified regular expression. For more information, see Extract log fields by using a regular expression.
processor_anchor	You can use the processor_anchor plug-in to anchor strings and extract fields based on the start and stop keywords that you specify. For more information, see Extract log fields by using start and stop keywords.
processor_split_char	You can use the processor_split_char plug-in to extract fields based on a specified single-character delimiter. For more information, see Extract log fields by using a single-character delimiter.

Logtail plug-in	Description
processor_split_string	You can use the processor_split_string plug-in to extract fields based on a specified multi-character delimiter. For more information, see Extract log fields by using a multi-character delimiter.
processor_split_key_value	You can use the processor_split_key_value plug-in to extract fields based on key-value pairs. For more information, see Extract log fields by splitting key-value pairs.
processor_add_fields	You can use the processor_add_fields plug-in to add fields to a log. For more information, see Add log fields.
processor_drop	You can use the processor_drop plug-in to drop specified fields. For more information, see Drop log fields.
processor_rename	You can use the processor_rename plug-in to rename specified fields. For more information, see Rename log fields.
processor_packjson	You can use the processor_packjson plug-in to encapsulate one or more fields into a field in the JSON format. For more information, see Encapsulate log fields (JSON).
processor_json	You can use the processor_json plug-in to expand JSON fields. For more information, see Expand JSON fields.
processor_filter_regex	You can use the processor_filter_regex plug-in to filter logs. For more information, see Filter logs by using regular expressions.
processor_gotime	You can use the processor_gotime plug-in to extract time information from a field in a time format that is supported by Golangand, and then configure the time information as the log time. For more information, see Extract log time (Go).
processor_strptime	You can use the processor_strptime plug-in to extract time information from a field in a time format that is supported by strptime, and then configure the time information as the log time. For more information, see Extract log time (strptime).
processor_geoip	You can use the processor_geoip plug-in to convert IP addresses in logs to geographical locations. A geographical location includes the following information: country, province, city, longitude, and latitude. For more information, see Convert an IP address to a geographical location.

You can also create a custom method that includes one or more of the preceding methods. For more information, see Custom methods.

Extract log fields by using a regular expression

You can use a regular expression to extract log fields.

The type of the plug-in is `processor_regex`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_regex.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
Regex	String	Yes	The regular expression. Enclose the fields that you want to extract in parentheses `()`.

Parameter	Type	Required	Description
Keys	String array	Yes	The array of fields that are extracted, for example, ["ip", "time", "method"].
NoKeyError	Boolean	No	Specifies whether to report an error if a field is not matched. Default value: false. This value indicates that no error is reported if a field is not matched.
NoMatchError	Boolean	No	Specifies whether to report an error if the regular expression does not match the value of a specified field. Default value: false. This value indicates that no error is reported if a field is not matched.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: false. This value indicates that the source field is not retained.
FullMatch	Boolean	No	Default value: true. This value indicates that exact match is performed when the regular expression specified in the Regex parameter is used to match field values. If you set the value to false, partial match is performed when the regular expression is used to match field values.

- Configuration example

  The following example shows how to extract the value of the content field. Then, you can set the names of the destination fields to ip, time, method, url, request_time, request_length, status, length, ref_url, and browser.

  - Raw log

    ```
    "content" : "203.0.113.10 - - [10/Aug/2017:14:57:51 +0800] \"POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=<yourSignature> HTTP/1.1\" 0.024 18204 200 37 \"-\" \"aliyun-sdk-java"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
        "type" : "processor_regex",
        "detail" : {"SourceKey" : "content",
             "Regex" : "([\\d\\.]+) \\S+ \\S+ \\[(\\S+) \\S+\\] \"(\\w+) ([^\\\"]*)\" ([\\d\\.]+) (\\d+) (\\d+) (\\d+|-) \"([^\\\"]*)\" \"([^\\\"]*)\" (\\d+)",
             "Keys"   : ["ip", "time", "method", "url", "request_time", "request_length", "status", "length", "ref_url", "browser"],
             "NoKeyError" : true,
             "NoMatchError" : true,
             "KeepSource" : false
        }
    }
    ```

- Result

```
"ip" : "203.0.113.10"
"time" : "10/Aug/2017:14:57:51"
"method" : "POST"
"url" : "/PutData?Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun
%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=<yourSignature>"
"request_time" : "0.024"
"request_length" : "18204"
"status" : "200"
"length" : "27"
"ref_url" : "-"
"browser" : "aliyun-sdk-java"
```

Extract log fields by using start and stop keywords

You can use start and stop keywords to anchor strings and extract log fields.

The type of the plug-in is `processor_anchor`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_anchor.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
Anchors	Anchor array	Yes	The list of the parameters that are set to anchor strings.
NoAnchorError	Boolean	No	Specifies whether to report an error if no keyword is found. Default value: false. This value indicates that no error is reported if no keyword is found.
NoKeyError	Boolean	No	Specifies whether to report an error if a field is not matched. Default value: false. This value indicates that no error is reported if a field is not matched.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: false. This value indicates that the source field is not retained.

  The following table describes the parameters of the Anchors parameter.

Parameter	Type	Required	Description
Start	String	Yes	The keyword that anchors the start of a substring in a string. If you do not specify the parameter, the start of the string is matched.
Stop	String	Yes	The keyword that anchors the end of a substring in a string. If you do not specify the parameter, the end of the string is matched.

Parameter	Type	Required	Description
FieldName	String	Yes	The name of the field that you want to extract.
FieldType	String	Yes	The type of the field that you want to extract. Valid values: string and json.
ExpondJson	Boolean	No	Specifies whether to expand a JSON substring that is anchored. Default value: false. This value indicates that a JSON substring that is anchored is not expanded. This parameter is available only if the value of the FieldType parameter is set to json.
ExpondConnecter	String	No	The character that is used to connect expanded keys. Default value: _.
MaxExpondDepth	Int	No	The maximum depth of JSON expansion. Default value: 0. This value indicates that the depth of JSON expansion is unlimited.

- Configuration example

  The following example shows how to extract the value of the content field. Then, you can set the names of the destination fields to time, val_key1, val_key2, val_key3, value_key4_inner1, and value_key4_inner2.

  - Raw log

    ```
    "content" : "time:2017.09.12 20:55:36\tjson:{\"key1\" : \"xx\", \"key2\": false, \"key3\":123.456
    , \"key4\" : { \"inner1\" : 1, \"inner2\" : false}}"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
       "type" : "processor_anchor",
       "detail" : {"SourceKey" : "content",
          "Anchors" : [
              {
                  "Start" : "time",
                  "Stop" : "\t",
                  "FieldName" : "time",
                  "FieldType" : "string",
                  "ExpondJson" : false
              },
              {
                  "Start" : "json:",
                  "Stop" : "",
                  "FieldName" : "val",
                  "FieldType" : "json",
                  "ExpondJson" : true
              }
          ]
      }
    }
    ```

- Result

```
"time" : "2017.09.12 20:55:36"
"val_key1" : "xx"
"val_key2" : "false"
"val_key3" : "123.456"
"value_key4_inner1" : "1"
"value_key4_inner2" : "false"
```

Extract log fields by using a single-character delimiter

You can use a specified single-character delimiter to extract fields. This processing method allows you to specify a quote to enclose the delimiter.

The type of the plug-in is `processor_split_char` .

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_split_char.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
SplitSep	String	Yes	The delimiter. The delimiter must be a single character. You can specify a non-printable character as a single-character delimiter, for example, \u0001.
SplitKeys	String array	Yes	The names of the delimited fields, for example, ["ip","time","method"].
QuoteFlag	Boolean	No	Specifies whether to use a quote to enclose the specified delimiter. Default value: false. This value indicates that a quote is not used to enclose the specified delimiter.
Quote	String	No	The quote. The quote must be a single character. You can specify a non-printable character as a quote, for example, \u0001. This parameter is available only if the value of QuoteFlag is set to true.
NoKeyError	Boolean	No	Specifies whether to report an error if a field is not matched. Default value: false. This value indicates that no error is reported if a field is not matched.
NoMatchError	Boolean	No	Specifies whether to report an error if a delimiter is not matched. Default value: false. This value indicates that no error is reported if a delimiter is not matched.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: false. This value indicates that the source field is not retained.

- Configuration example

The following example shows how to use a vertical bar (|) as a delimiter to extract the value of the content field. Then, you can set the names of the destination fields to ip, time, method, url, request_time, request_length, status, length, ref_url, and browser.

- Raw log

```
"content" : "203.0.113.10|10/Aug/2017:14:57:51 +0800|POST|PutData?
Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun%202013%2006%3A53%
3A30%20GMT&Topic=raw&Signature=<yourSignature>|0.024|18204|200|37|-|
aliyun-sdk-java"
```

- Logtail plug-in configurations for data processing

```
{
   "type" : "processor_split_char",
   "detail" : {"SourceKey" : "content",
      "SplitSep" : "|",
      "SplitKeys" : ["ip", "time", "method", "url", "request_time", "request_length", "status", "
length", "ref_url", "browser"]
  }
}
```

- Result

```
"ip" : "203.0.113.10"
"time" : "10/Aug/2017:14:57:51 +0800"
"method" : "POST"
"url" : "/PutData?Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun
%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=<yourSignature>"
"request_time" : "0.024"
"request_length" : "18204"
"status" : "200"
"length" : "27"
"ref_url" : "-"
"browser" : "aliyun-sdk-java"
```

Extract log fields by using a multi-character delimiter

You can use a specified multi-character delimiter to extract fields. You cannot specify a quote to enclose the delimiter.

The type of the plug-in is `processor_split_string`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_split_string.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
SplitSep	String	Yes	The delimiter. The delimiter contains multiple characters. You can specify non-printable characters in the delimiter, for example, \u0001\u0002.
SplitKeys	String array	Yes	The names of the delimited fields, for example, ["key1","key2"].

Parameter	Type	Required	Description
PreserveOthers	Boolean	No	Specifies whether to retain excess fields if the number of fields is greater than the number of fields that are specified by the SplitKeys parameter. Default value: false. This value indicates that excess fields are not retained.
ExpandOthers	Boolean	No	Specifies whether to parse excess fields. Default value: false. This value indicates that excess fields are not parsed.
ExpandKeyPrefix	String	No	The name prefix of excess fields. For example, if you specify expand_ for the parameter, the first two excess fields are named expand_1 and expand_2.
NoKeyError	Boolean	No	Specifies whether to report an error if a field is not matched. Default value: false. This value indicates that no error is reported if a field is not matched.
NoMatchError	Boolean	No	Specifies whether to report an error if a delimiter is not matched. Default value: false. This value indicates that no error is reported if a delimiter is not matched.
KeepSource	Boolean	No	Specifies whether to retain the source field. This value indicates that the source field is not retained.

- Configuration example

  The following example shows how to use a delimiter (|#|) to extract the value of the content field. Then, you can set the names of the destination fields to ip, time, method, url, request_time, request_length, status, expand_1, expand_2, and expand_3.

  - Raw log

    ```
    "content" : "203.0.113.10|#|10/Aug/2017:14:57:51 +0800|#|POST|#|PutData?
    Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun%202013%2006%3A53%
    3A30%20GMT&Topic=raw&Signature=<yourSignature>|#|0.024|#|18204|#|200|#|27|#|-|#|
    aliyun-sdk-java"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
       "type" : "processor_split_string",
       "detail" : {"SourceKey" : "content",
          "SplitSep" : "|#|",
          "SplitKeys" : ["ip", "time", "method", "url", "request_time", "request_length", "status"],
          "PreserveOthers" : true,
          "ExpandOthers" : true,
          "ExpandKeyPrefix" : "expand_"
      }
    }
    ```

- Result

```
"ip" : "203.0.113.10"
"time" : "10/Aug/2017:14:57:51 +0800"
"method" : "POST"
"url" : "/PutData?Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun
%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=<yourSignature>"
"request_time" : "0.024"
"request_length" : "18204"
"status" : "200"
"expand_1" : "27"
"expand_2" : "-"
"expand_3" : "aliyun-sdk-java"
```

Extract log fields by splitting key-value pairs

You can split key-value pairs to extract log fields.

The type of the plug-in is `processor_split_char`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_split_key_value.

  > **Note** Only Logtail V0.16.26 or later supports the plug-in.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
Delimiter	String	No	The delimiter between key-value pairs. Default value: `\t`.
Separator	String	No	The delimiter that is used to separate the key and the value in a single key-value pair. A colon (:) is used by default.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: true. This value indicates that the source field is retained.
ErrIfSourceKeyNotFound	Boolean	No	Specifies whether to trigger an alert if a field is not matched. Default value: true. This value indicates that an alert is triggered if a field is not matched.
DiscardWhenSeparatorNotFound	Boolean	No	Specifies whether to drop the key-value pair if a field is not matched. Default value: false. This value indicates that the key-value pair is not dropped if a field is not matched.
ErrIfSeparatorNotFound	Boolean	No	Specifies whether to trigger an alert if the delimiter specified by the Separator parameter does not exist. Default value: true. This value indicates that an alert is triggered if the specified delimiter does not exist.

- Configuration example

The following example shows how to split the key-value pairs in the value of the content field. The delimiter that is used to separate key-value pairs is a tab character (`/t`). The delimiter that is used to separate the key and the value in a single key-value pair is a colon (:).

- Raw log

```
"content": "class:main\tuserid:123456\tmethod:get\tmessage:\"wrong user\""
```

- Logtail plug-in configurations for data processing

```
{
  "processors":[
    {
      "type":"processor_split_key_value",
      "detail": {
        "SourceKey": "content",
        "Delimiter": "\t",
        "Separator": ":",
        "KeepSource": true
      }
    }
  ]
}
```

- Result

```
"content": "class:main\tuserid:123456\tmethod:get\tmessage:\"wrong user\""
"class": "main"
"userid": "123456"
"method": "get"
"message": "\"wrong user\""
```

Convert an IP address to a geographical location

This processing method converts IP addresses in logs to geographical locations. A geographical location includes the following information: country, province, city, longitude, and latitude.

The type of the plug-in is `processor_geoip` .

> **Note**
> - GeoIP databases are not included in the Logtail installation package. You must download and configure a GeoIP database on the server where Logtail is installed. We recommend that you download a database that provides the city information of an IP address.
> - Make sure that the database format is MMDB.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_geoip.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field that you want to convert.
DBPath	String	Yes	The absolute path of the GeoIP database, for example, */user/data/GeoLite2-City_20180102/GeoLite2-City.mmdb*.

Parameter	Type	Required	Description
NoKeyError	Boolean	No	Specifies whether to report an error if a field is not matched. Default value: false. This value indicates that no error is reported if a field is not matched.
NoMatchError	Boolean	No	Specifies whether to report an error if an IP address is invalid or is not matched in the database. Default value: false. This value indicates that no error is reported if an IP address is invalid or is not matched in the database.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: true. This value indicates that the source field is retained.
Language	String	No	The language of the GeoIP database. Default value: zh-CN. Make sure that your GeoIP database can be displayed in a language that is suitable for your business.

- Configuration example

  The following example shows how to configure the processing method to convert IP addresses in logs to geographical locations.

  - Raw log

    ```
    "source_ip" : "203.0.113.10"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
       "type": "processor_geoip",
        "detail": {
             "SourceKey": "ip",
             "NoKeyError": true,
             "NoMatchError": true,
             "KeepSource": true,
             "DBPath" : "/user/local/data/GeoLite2-City_20180102/GeoLite2-City.mmdb"
        }
    }
    ```

  - Result

    ```
    "source_ip_city_" : "**.**.**.**"
    "source_ip_province_" : "Zhejiang"
    "source_ip_city_" : "Hangzhou"
    "source_ip_province_code_" : "ZJ"
    "source_ip_country_code_" : "CN"
    "source_ip_longitude_" : "120.********"
    "source_ip_latitude_" : "30.********"
    ```

Filter logs by using regular expressions

This method uses regular expressions to filter logs. You can specify conditions in the `Include` and `Exclude` parameters.

The type of the plug-in is `processor_filter_regex` .

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_filter_regex.

  > **Note** A log is collected only if the log exactly matches the regular expression that is specified in the Include parameter and does not match the regular expression that is specified in the Exclude parameter.

Parameter	Type	Required	Parameters
Include	JSON object that conatins key-value pairs	No	A map that includes key-value pairs. In each key-value pair, the key specifies a field and the value specifies a regular expression that the value of the same field in each log must match. If the values of all fields in a log match the regular expressions that are specified in the Include parameter, the log is collected.
Exclude	JSON object that conatins key-value pairs	No	A map that includes key-value pairs. In each key-value pair, the key specifies a field and the value specifies a regular expression that the value of the same field in each log must match. If the values of all fields in a log match the regular expressions that are specified in the Exclude parameter, the log is not collected.

- Configuration example

  The following example shows how to use regular expressions to filter logs.

  - Raw logs
    - Log 1

      ```
      "ip" : "203.0.113.10"
      "method" : "POST"
      ...
      "browser" : "aliyun-sdk-java"
      ```

    - Log 2

      ```
      "ip" : "203.0.113.20"
      "method" : "POST"
      ...
      "browser" : "chrome"
      ```

    - Log 3

      ```
      "ip" : "198.51.100.10"
      "method" : "POST"
      ...
      "browser" : "ali-sls-ilogtail"
      ```

- Logtail plug-in configurations for data processing

```
{
   "type" : "processor_filter_regex",
    "detail" : {
         "Include" : {
            "ip" : "203\\..*",
            "method" : "POST"
         },
         "Exclude" : {
            "browser" : "aliyun.*"
         }
    }
}
```

- Result

Log	Collected	Reason
Log 1	No	The value of the browser parameter matches the regular expression that is specified in the Exclude parameter.
Log 2	Yes	All the filter conditions are met.
Log 3	No	The value of the ip parameter does not match the regular expression that is specified in the Include parameter.

Add log fields

You can use this method to add multiple fields to a log.

The type of the plug-in is `processor_add_fields`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_add_fields.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
Fields	Map	No	The key-value pairs that you want to add. You can specify multiple key-value pairs in the parameter.
IgnoreIfExist	Boolean	No	Specifies whether to retain key-value pairs that have the same key. Default value: false. This value indicates that a key-value pair is not retained if the key is the same as another specified key.

- Configuration example

  The following example shows how to add the aaa2 and aaa3 fields to a log.

  - Raw log

```
"aaa1":"value1"
```

- Logtail plug-in configurations for data processing

```
{
  "processors":[
    {
      "type":"processor_add_fields",
      "detail": {
        "Fields": {
          "aaa2": "value2",
          "aaa3": "value3"
        }
      }
    }
  ]
}
```

- Result

```
"aaa1":"value1"
"aaa2":"value2"
"aaa3":"value3"
```

Drop log fields

You can use this method to drop specified fields from a log.

The type of the plug-in is `processor_drop`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_drop.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
DropKeys	String array	No	The fields that you want to drop. You can drop one or more fields from a log.

- Configuration example

  The following example shows how to drop the aaa1 and aaa2 fields from a log.

  - Raw log

```
"aaa1":"value1"
"aaa2":"value2"
"aaa3":"value3"
```

- Logtail plug-in configurations for data processing

```
{
  "processors":[
    {
      "type":"processor_drop",
      "detail": {
        "DropKeys": ["aaa1","aaa2"]
      }
    }
  ]
}
```

- Result

```
"aaa3":"value3"
```

Extract log time (Go)

You can use this method to extract time information from a specified field, and then convert the time format.

The type of the plug-in is `processor_gotime`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_gotime.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
SourceFormat	String	Yes	The format of the time information in the source field.
SourceLocation	Int	Yes	The source time zone. If you do not specify the parameter, the current time zone of the server where Logtail is installed is used.
DestKey	String	Yes	The name of the destination field.
DestFormat	String	Yes	The format of the time information in the destination field.
DestLocation	Int	No	The destination time zone. If you do not specify the parameter, the current time zone of the server where Logtail is installed is used.
SetTime	Boolean	No	Specifies whether to configure the time information as the log time. Default value: true. This value indicates that the time information is configured as the log time.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: true. This value indicates that the source field is retained.

Parameter	Type	Required	Description
NoKeyError	Boolean	No	Specifies whether to report an error if a field is not matched. Default value: true. This value indicates that an error is reported if the source field is not matched.
AlarmIfFail	Boolean	No	Specifies whether to trigger an alert if the time information fails to be extracted. Default value: true. This value indicates that an alert is triggered if the time information fails to be extracted.

- Configuration example

  In this example, the time information `2006-01-02 15:04:05 (UTC+8)` is extracted from the s_key field, converted to `2006/01/02 15:04:05 (UTC+9)`, and then added to the d_key field.

  - Raw log

    ```
    "s_key":"2019-07-05 19:28:01"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
      "processors":[
        {
          "type":"processor_gotime",
          "detail": {
            "SourceKey": "s_key",
            "SourceFormat":"2006-01-02 15:04:05",
            "SourceLocation":8,
            "DestKey":"d_key",
            "DestFormat":"2006/01/02 15:04:05",
            "DestLocation":9,
            "SetTime": true,
            "KeepSource": true,
            "NoKeyError": true,
            "AlarmIfFail": true
          }
        }
      ]
    }
    ```

  - Result

    ```
    "s_key":"2019-07-05 19:28:01"
    "d_key":"2019/07/05 20:28:01"
    ```

Expand JSON fields

You can use this method to expand a JSON field.

The type of the plug-in is `processor_json`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_json.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
NoKeyError	Boolean	No	Specifies whether to report an error if the source field is not matched. Default value: true. This value indicates that an error is reported if the source field is not matched.
ExpandDepth	Int	No	The depth of JSON expansion. Default value: 0. This value indicates that the depth of JSON expansion is unlimited. If the value is n, the depth of JSON expansion is n.
ExpandConnector	String	No	The character that is used to connect expanded keys. You can leave this parameter empty. Default value: _.
Prefix	String	No	The prefix that is added to expanded keys. You can leave this parameter empty.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: true. This value indicates that the source field is retained.
UseSourceKeyAsPrefix	Boolean	No	Specifies whether to add the name of the source field as a prefix to all expanded keys. Default value: false. This value indicates that the name of the source field is not added.

- Configuration example

  The following example shows how to expand the JSON field s_key, and then add j and the name of the source field s_key as a prefix to the expanded keys.

  - Raw log

    ```
    "s_key":"{\"k1\":{\"k2\":{\"k3\":{\"k4\":{\"k51\":\"51\",\"k52\":\"52\"},\"k41\":\"41\"}}})"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
      "processors":[
        {
          "type":"processor_json",
          "detail": {
            "SourceKey": "s_key",
            "NoKeyError":true,
            "ExpandDepth":0,
            "ExpandConnector":"-",
            "Prefix":"j",
            "KeepSource": false,
            "UseSourceKeyAsPrefix": true
          }
        }
      ]
    }
    ```

- Result

```
"s_key":"{\"k1\":{\"k2\":{\"k3\":{\"k4\":{\"k51\":\"51\",\"k52\":\"52\"},\"k41\":\"41\"}}})"
"js_key-k1-k2-k3-k4-k51":"51"
"js_key-k1-k2-k3-k4-k52":"52"
"js_key-k1-k2-k3-k41":"41"
```

Encapsulate log fields (JSON)

You can use this method to encapsulate one or more fields into a field in the JSON format.

The type of the plug-in is `processor_packjson`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_packjson.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
SourceKeys	String array	Yes	The field that you want to encapsulate. The field is in the string array format.
DestKey	String	No	The destination field in the JSON format.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: true. This value indicates that the source field is retained.
AlarmIfIncomplete	Boolean	No	Specifies whether to trigger an alert if the source field does not exist. Default value: true. This value indicates that an alert is triggered if the source field does not exist.

- Configuration example

  The following example shows how to encapsulate the a and b fields into the d_key field.

  - Raw log

```
"a":"1"
"b":"2"
```

  - Logtail plug-in configurations for data processing

```
{
  "processors":[
    {
      "type":"processor_packjson",
      "detail": {
        "SourceKeys": ["a","b"],
        "DestKey":"d_key",
        "KeepSource":true,
        "AlarmIfEmpty":true
      }
    }
  ]
}
```

- Result

```
"a":"1"
"b":"2"
"d_key":"{\"a\":\"1\",\"b\":\"2\"}"
```

Rename log fields

You can use this method to rename multiple fields.

The type of the plug-in is `processor_rename`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_rename.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
NoKeyError	Boolean	Yes	Specifies whether to report an error if a field that you want to rename is not matched. Default value: false. This value indicates that no error is reported if a field that you want to rename is not matched.
SourceKeys	String array	Yes	The source fields that you want to rename.
DestKeys	String array	Yes	The fields that are renamed.

- Configuration example

  The following example shows how to rename the aaa1 field to bbb1 and the aaa2 field to bbb2.

  - Raw log

```
"aaa1":"value1"
"aaa2":"value2"
"aaa3":"value3"
```

  - Logtail plug-in configurations for data processing

```
{
  "processors":[
    {
      "type":"processor_rename",
      "detail": {
        "SourceKeys": ["aaa1","aaa2"],
        "DestKeys": ["bbb1","bbb2"],
        "NoKeyError": true
      }
    }
  ]
}
```

  - Result

```
"bbb1":"value1"
"bbb2":"value2"
"aaa3":"value3"
```

Extract log time (strptime)

You can use this method to extract time information from a field, and then configure the time information as the log time.

The type of the plug-in is `processor_strptime`.

- Parameters

  The following table describes the parameters that you can specify in the detail parameter if you set the type parameter to processor_strptime.

  > **Note** Only Logtail V0.16.28 or later supports the plug-in.

Parameter	Type	Required	Description
SourceKey	String	Yes	The name of the source field.
Format	String	Yes	The format of the time information in the source field.
AdjustUTCOffset	Boolean	No	Specifies whether to modify the time zone. Default value: false. This value indicates that the time zone is not modified.
UTCOffset	Int	No	The offset that is used to modify the time zone. For example, the value 28800 indicates that the time zone is modified to UTC+8.
AlarmIfFail	Boolean	No	Specifies whether to trigger an alert if the time information fails to be extracted. Default value: true. This value indicates that an alert is triggered if the time information fails to be extracted.
KeepSource	Boolean	No	Specifies whether to retain the source field. Default value: true. This value indicates that the source field is retained.

- Configuration examples

  The following examples show how to parse the value of the log_time field into the `%Y/%m/%d %H:%M:%S` format. The current time zone of the server where Logtail is installed is used.

- Example 1: The time zone is UTC+8.
  - Raw log

    ```
    "log_time":"2016/01/02 12:59:59"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
      "processors":[
        {
          "type":"processor_strptime",
          "detail": {
            "SourceKey": "log_time",
            "Format": "%Y/%m/%d %H:%M:%S"
          }
        }
      ]
    }
    ```

  - Result

    ```
    "log_time":"2016/01/02 12:59:59"
    Log.Time = 1451710799
    ```

- Example 2: The time zone is UTC+7.
  - Raw log

    ```
    "log_time":"2016/01/02 12:59:59"
    ```

  - Logtail plug-in configurations for data processing

    ```
    {
      "processors":[
        {
          "type":"processor_strptime",
          "detail": {
            "SourceKey": "log_time",
            "Format": "%Y/%m/%d %H:%M:%S",
            "AdjustUTCOffset": true,
            "UTCOffset": 25200
          }
        }
      ]
    }
    ```

  - Result

    ```
    "log_time":"2016/01/02 12:59:59"
    Log.Time = 1451714399
    ```

Custom methods

You can use multiple processing methods to process logs. The following example shows how to use a single-character delimiter to split a log into several fields and then specify anchor points to extract content from the `detail` field.

- Raw log

```
"content" :
"ACCESS|QAS|203.0.113.10|1508729889935|52460dbed4d540b88a973cf5452b1447|1238|appKey=ba,env=pub,requestTime=1508729889913,latency=22ms,
request={appKey:ba,optional:{\\domains\\:\\daily\\,\\version\\:\\v2\\},rawQuery:{\\query\\:\\The route to Location A\\,\\domain\\:\\Navigation\\,\\intent\\:\\navigate\\,\\slots\\:\\to_geo:level3=Location A\\,\\location\\:\\Location B\\},
requestId:52460dbed4d540b88a973cf5452b1447},
response={answers:[],status:SUCCESS}|"
```

- Logtail plug-in configurations for data processing

```
"processors" : [
      {
          "type" : "processor_split_char",
          "detail" : {"SourceKey" : "content",
              "SplitSep" : "|",
              "SplitKeys" : ["method", "type", "ip", "time", "req_id", "size", "detail"]
          }
      },
      {
          "type" : "processor_anchor",
          "detail" : "SourceKey" : "detail",
              "Anchors" : [
                  {
                          "Start" : "appKey=",
                      "Stop" : ",env=",
                      "FieldName" : "appKey",
                      "FieldType" : "string"
                  },
                  {
                      "Start" : ",env",
                      "Stop" : ",requestTime=",
                      "FieldName" : "env",
                      "FieldType" : "string"
                  },
                  {
                      "Start" : ",requestTime=",
                      "Stop" : ",latency",
                      "FieldName" : "requestTime",
                      "FieldType" : "string"
                  },
                  {
                      "Start" : ",latency=",
                      "Stop" : ",request=",
                      "FieldName" : "latency",
                      "FieldType" : "string"
                  },
                  {
                      "Start" : ",request=",
                      "Stop" : ",response=",
                      "FieldName" : "request",
                      "FieldType" : "string"
                  },
                  {
                      "Start" : ",response=",
                      "Stop" : "",
                      "FieldName" : "response",
                      "FieldType" : "json"
                  }
              ]
          }
      }
  ]
```

- Result

```
"method" : "ACCESS"
"type" : "QAS"
"ip" : "203.0.113.10"
"time" : "1508729889935"
"req_id" : "52460dbed4d540b88a973cf5452b1447"
"size" : "1238"
"appKey" : "ba"
"env" : "pub"
"requestTime" : "1508729889913"
"latency" : "22ms"
"request" : "{appKey:nui-banma,optional:{\\domains\\:\\daily-faq\\,\\version\\:\\v2\\},rawQuery:{\\
query\\:\\\345\216\273\344\271\220\345\261\261\347\232\204\350\267\257\347\272\277\\,\\domain\\:\\\
345\257\274\350\210\252\\,\\intent\\:\\navigate\\,\\slots\\:\\to_geo:level3=\344\271\220\345\261\26
1\\,\\location\\:\\\345\214\227\344\272\254\\},requestId:52460dbed4d540b88a973cf5452b1447}"
"response_answers" : "[]"
"response_status" : "SUCCESS"
```

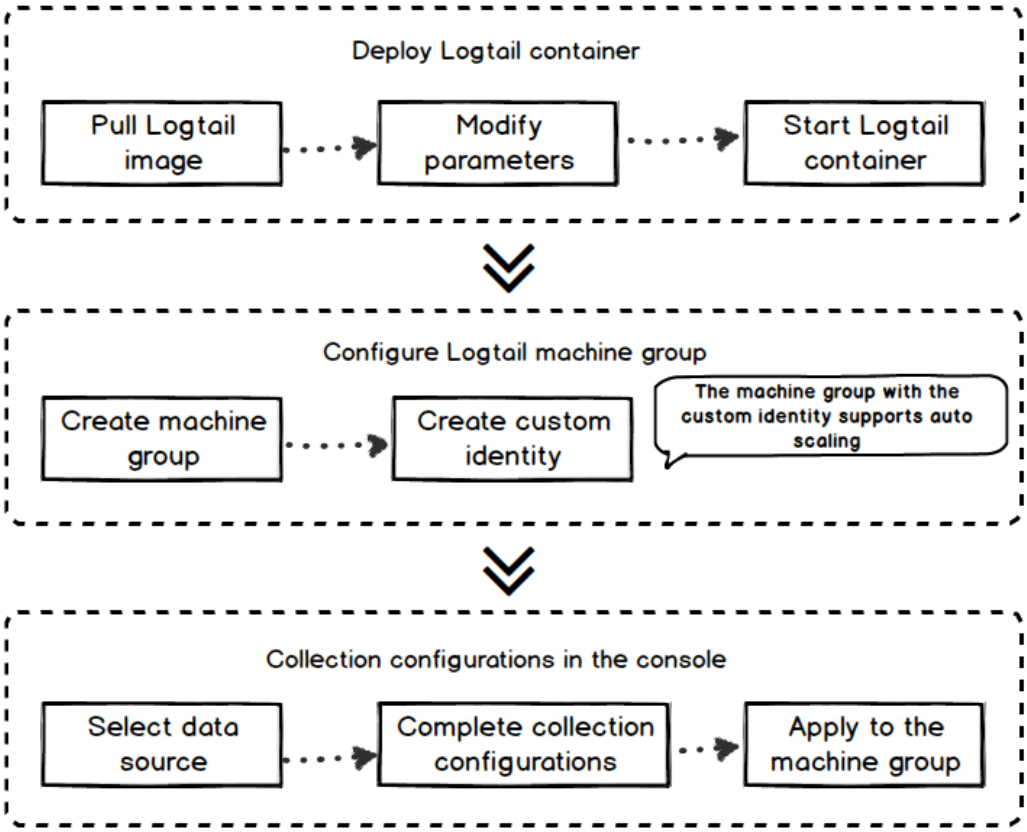

23.1.3.1.6. Collect container logs

23.1.3.1.6.1. Collect standard Docker logs

This topic describes how to use Logtail to collect standard Docker logs and upload these logs together with the container metadata to Log Service.

Procedure

Procedure



1. Deploy a Logtail container.

2. Configure a Logtail server group.

   Create a server group with a custom ID in the Log Service console. The container cluster can automatically scale up or down based on data traffic.

3. Create a Logtail configuration.

   Create a Logtail configuration in the Log Service console. The Logtail configuration process is completed in the Log Service console. No local configuration is needed.

Deploy a Logtail container

1. Run the following command to pull the Logtail image.

```
docker pull registry.cn-hangzhou.aliyuncs.com/log-service/logtail
```

2. Start a Logtail container.

   Set the `${your_region_name}`, `${your_aliyun_user_id}`, and `${your_machine_group_user_defined_id}` parameters in the startup template.

```
docker run -d -v /:/logtail_host:ro -v /var/run:/var/run --env
ALIYUN_LOGTAIL_CONFIG=/etc/ilogtail/conf/${your_region_name}/ilogtail_config.json
--env ALIYUN_LOGTAIL_USER_ID=${your_aliyun_user_id} --env
 ALIYUN_LOGTAIL_USER_DEFINED_ID=${your_machine_group_user_defined_id} registry.cn-hangzhou.aliyuncs.com/log-service/logtail
```

   Notice Before you set the parameters, you must complete one of the following configurations. Otherwise, the `container text file busy` error may occur when you delete another container.

   - For CentOS 7.4 and later versions, set fs.may_detach_mounts to 1. For more information, see Bug 1468249, Bug 1441737, and Issue 34538.
   - Grant the `privileged` permission to Logtail by adding the `--privileged` flag to the startup parameters. For more information, see Docker run reference.

Parameter	Description
`${your_region_name}`	The region of the project. For more information, see View the information of a project.
`${your_aliyun_user_id}`	The user ID. Set this parameter to the ID of your Alibaba Cloud account, which is a string. For information about how to view the ID, see Step 1 in Configure an account ID for a server.
`${your_machine_group_user_defined_id}`	The custom ID of your server group. For information about how to set the custom ID, see Step 1 in Create a machine group based on a custom ID.

   After you set the parameters, run the following command to start the Logtail container.

```
docker run -d -v /:/logtail_host:ro -v /var/run:/var/run
--env ALIYUN_LOGTAIL_CONFIG=/etc/ilogtail/conf/cn_hangzhou/ilogtail_config.json --env
 ALIYUN_LOGTAIL_USER_ID=1654218******--env ALIYUN_LOGTAIL_USER_DEFINED_ID=log-docker-demo registry.cn-hangzhou.aliyuncs.com/log-service/logtail
```

> **Notice**
>
> You can customize the startup parameters of the Logtail container if the following conditions are met:
>
> - The following environment variables exist before you start the Logtail container: `ALIYUN_LOGTAIL_USER_DEFINED_ID`, `ALIYUN_LOGTAIL_USER_ID`, and `ALIYUN_LOGTAIL_CONFIG`.
> - The */var/run* directory is mounted on the */var/run* directory of the Logtail container.
> - To collect container standard output, container logs, or host files, you must mount the root directory on the `/logtail_host` directory of the Logtail container.
> - If an error showing *The parameter is invalid : uuid=none* occurs in the `/usr/local/ilogtail/ilogtail.LOG` Logtail log file, create a file named *product_uuid* on the host. Add a valid UUID such as `169E98C9-ABC0-4A92-B1D2-AA6239C0D261` to the file, and mount the file on the */sys/class/dmi/id/product_uuid* directory of the Logtail container.

Configure a Logtail server group

1. Log on to the Log Service console.
2. Click a project name.
3. In the left-side navigation pane, click the **Server Groups** icon to show the server group list.
4. Click the icon next to Server Groups, and then select **Create Server Group**.

   You can also create a server group when you import data to Log Service.
5. In the dialog box that appears, select **Custom ID** from the Identifier drop-down list. Enter the value of `ALIYUN_LOGTAIL_USER_DEFINED_ID` set in the previous step in the **Custom Identifier** field.

Click OK. One minute later, click the name of the server group in the **Server Groups** list. On the **Server Group Settings** page that appears, you can view the heartbeat status of the Logtail container. For more information, see View the status of a server group.

Create a Logtail configuration

Create a Logtail configuration in the console.

- For more information about Docker logs, see Collect container text logs.
- For more information about Docker standard output, see Collect stdout and stderr logs from containers.
- Host text logs.

  The root directory of a host is mounted on the `/logtail_host` directory of the Logtail container by default. You must add the /logtail_host prefix to the log path. For example, if you want to collect data from the `/home/logs/app_log/` directory of the host, you must set the log path as `/logtail_host/home/logs/app_log/`.

What to do next

- View the status of the Logtail container.

  You can run the `docker exec ${logtail_container_id} /etc/init.d/ilogtaild status` command to view the status of Logtail.
- View the version number, IP address, and startup time of Logtail.

  You can run the `docker exec ${logtail_container_id} cat /usr/local/ilogtail/app_info.json` command to view Logtail information.
- View the operations logs of Logtail.

  The operations logs of Logtail are stored in the `ilogtail.LOG` file in the `/usr/local/ilogtail/` directory. If the log file is rotated and compressed, it is stored as a file named `ilogtail.LOG.x.gz`.

  For example:

```
[root@iZbp17enxc2us3624wexh2Z ilogtail]# docker exec a287de895e40 tail -n 5 /usr/local/ilogtail/ilogtail.LOG
[2018-02-06 08:13:35.721864]    [INFO]    [8]    [build/release64/sls/ilogtail/LogtailPlugin.cpp:104]    logtail plugin Resume:start
[2018-02-06 08:13:35.722135]    [INFO]    [8]    [build/release64/sls/ilogtail/LogtailPlugin.cpp:106]    logtail plugin Resume:success
[2018-02-06 08:13:35.722149]    [INFO]    [8]    [build/release64/sls/ilogtail/EventDispatcher.cpp:369]    start add existed check point events, size:0
[2018-02-06 08:13:35.722155]    [INFO]    [8]    [build/release64/sls/ilogtail/EventDispatcher.cpp:511]    add existed check point events, size:0    cache size:0    event size:0    success count:0
[2018-02-06 08:13:39.725417]    [INFO]    [8]    [build/release64/sls/ilogtail/ConfigManager.cpp:3776]    check container path update flag:0    size:1
```

Ignore the following standard output:

```
start umount useless mount points, /shm$|/merged$|/mqueue$
umount: /logtail_host/var/lib/docker/overlay2/3fd0043af174cb0273c3c7869500fbe2bdb95d13b1e110172ef57fe840c82155/merged: must be superuser to unmount
umount: /logtail_host/var/lib/docker/overlay2/d5b10aa19399992755de1f85d25009528daa749c1bf8c16edff44beab6e69718/merged: must be superuser to unmount
umount: /logtail_host/var/lib/docker/overlay2/5c3125daddacedec29df72ad0c52fac800cd56c6e880dc4e8a640b1e16c22dbe/merged: must be superuser to unmount
...
xargs: umount: exited with status 255; aborting
umount done
start logtail
ilogtail is running
logtail status:
ilogtail is running
```

- Restart Logtail.

  To restart Logtail, use the following sample code:

```
[root@iZbp17enxc2us3624wexh2Z ilogtail]# docker exec a287de895e40 /etc/init.d/ilogtaild stop
kill process Name: ilogtail pid: 7
kill process Name: ilogtail pid: 8
stop success
[root@iZbp17enxc2us3624wexh2Z ilogtail]# docker exec a287de895e40 /etc/init.d/ilogtaild start
ilogtail is running
```

23.1.3.1.6.2. Collect Kubernetes logs

This topic describes how to install and use Logtail to collect logs from Kubernetes clusters.

Configuration procedure

Perform the following steps to collect logs from Kubernetes clusters:

1. Install the alibaba-log-controller Helm package.
2. Use the Log Service console to manage log collection configurations.

Step 1: Install Logtail

- Install Logtail in an Alibaba Cloud Container Service for Kubernetes cluster.

  If Log Service components are not installed in your cluster, you must manually install the components.

  i. Connect to the Kubernetes cluster by using CloudShell.

  ii. Run the following command in CloudShell to obtain the ID of your Apsara Stack tenant account.

```
echo $ALIBABA_CLOUD_ACCOUNT_ID
```

iii. After you set the `${your_k8s_cluster_id}`, `${your_ali_uid}`, and `${your_k8s_cluster_region_id}` parameters, run the following command:

```
wget https://acs-logging.oss-cn-hangzhou.aliyuncs.com/alicloud-k8s-log-installer.sh -O alicloud-k8s-log-installer.sh; chmod 744 ./alicloud-k8s-log-installer.sh; ./alicloud-k8s-log-installer.sh --cluster-id ${your_k8s_cluster_id} --ali-uid ${your_ali_uid} --region-id ${your_k8s_cluster_region_id}
```

- Install Logtail in a user-created Kubernetes cluster.

> **Notice**
> - The version of the Kubernetes cluster must be 1.8 or later.
> - Helm 2.6.4 or later must be installed.

i. In the Log Service console, create a project whose name starts with `k8s-log-custom-`.

ii. Replace the parameters in the following command based on your business requirements:

```
wget http://logtail-release-cn-hangzhou.oss-cn-hangzhou.aliyuncs.com/kubernetes/alicloud-log-k8s-custom-install.sh; chmod 744 ./alicloud-log-k8s-custom-install.sh; sh ./alicloud-log-k8s-custom-install.sh {your-project-suffix} {region-id} {aliuid} {access-key-id} {access-key-secret}
```

The following table lists the parameters in the preceding command.

Parameter	Description
{your-project-suffix}	The portion of the project name at the end of `k8s-log-custom-`. For example, if you create a project whose name is `k8s-log-custom-xxxx`, set this parameter to `xxxx`.
{regionId}	The ID of the region where the project resides. For more information, see View the information of a project.
{aliuid}	The user ID. Set this parameter to the ID of your Apsara Stack tenant account. **Note** The ID of an Apsara Stack tenant account is a string. For more information about how to obtain the ID, see Configure an account ID for a server.
{access-key-id}	The AccessKey ID of your Apsara Stack tenant account.
{access-key-secret}	The AccessKey secret of your Apsara Stack tenant account.

After Logtail is installed in the Kubernetes cluster, Log Service automatically creates a machine group named `k8s-group-${your_k8s_cluster_id}` for the project.

> **Note**
>
> - A Logstore named `config-operation-log` is automatically created in the project. Do not delete the Logstore.
> - When you install Logtail in a user-created Kubernetes cluster, Logtail is granted `privileged` permissions by default. This prevents the `container text file busy` error when you delete a pod. For more information, visit Bug 1468249, Bug 1441737, and Issue 34538.

The following example shows a successful installation:

```
[root@iZbp1dsxxxxxqfbiaZ ~]#  wget http://logtail-release-cn-hangzhou.oss-cn-hangzhou.aliyuncs.com/
kubernetes/alicloud-log-k8s-custom-install.sh; chmod 744 ./alicloud-log-k8s-custom-install.sh; sh .
/alicloud-log-k8s-custom-install.sh xxxx cn-hangzhou 165xxxxxxxx050 LTAxxxxxxxxxxx AIxxxxxxxxxxxxxx
xxxxxxxxxxxxxxe
....
....
....
NAME:   alibaba-log-controller
LAST DEPLOYED: Fri May 18 16:52:38 2018
NAMESPACE: default
STATUS: DEPLOYED
RESOURCES:
==> v1beta1/ClusterRoleBinding
NAME                    AGE
alibaba-log-controller  0s
==> v1beta1/DaemonSet
NAME        DESIRED  CURRENT  READY  UP-TO-DATE  AVAILABLE  NODE SELECTOR  AGE
logtail-ds  2        2        0      2           0          <none>         0s
==> v1beta1/Deployment
NAME                    DESIRED  CURRENT  UP-TO-DATE  AVAILABLE  AGE
alibaba-log-controller  1        1        1           0          0s
==> v1/Pod(related)
NAME                                    READY  STATUS             RESTARTS  AGE
logtail-ds-7xf2d                        0/1    ContainerCreating  0         0s
logtail-ds-9j4bx                        0/1    ContainerCreating  0         0s
alibaba-log-controller-796f8496b6-6jxb2  0/1    ContainerCreating  0         0s
==> v1/ServiceAccount
NAME                    SECRETS  AGE
alibaba-log-controller  1        0s
==> v1beta1/CustomResourceDefinition
NAME                                  AGE
aliyunlogconfigs.log.alibabacloud.com  0s
==> v1beta1/ClusterRole
alibaba-log-controller  0s
[INFO] your k8s is using project : k8s-log-custom-xxx, region : cn-hangzhou, aliuid : ************
**, accessKeyId : LTA**********
[SUCCESS] install helm package : alibaba-log-controller success.
```

To check the status of each Log Service component in the Kubernetes cluster, run the `helm status alibaba-log-controller` command. If all pods are in the Running state, Logtail is installed.

Log on to the Log Service console to find the project. If you have multiple projects, search for the project by using the `k8s-log` keyword.

Step 2: Configure log collection

Create Logtail configurations for log collection in the console as required.

- For information about how to collect Kubernetes text logs, see Collect container text logs.

- For information about how to collect Kubernetes stdout logs, see Collect stdout and stderr logs from containers.
- Host text logs.

  By default, the root directory of a host is mounted to the `/logtail_host` directory of the Logtail container. You must add the /logtail_host prefix to the log path. For example, if you want to collect data from the `/home/logs/app_log/` directory of the host, you must set the log path to `/logtail_host/home/logs/app_log/`.

Other common commands

- Store logs of multiple clusters in one project

  You can collect logs from multiple Kubernetes clusters. If you want to store these logs in the same project, you can specify the same cluster ID for the `${your_k8s_cluster_id}` parameter when you install Log Service components on multiple Kubernetes clusters.

  For example, if you have three Kubernetes clusters whose IDs are abc001, abc002, and abc003, specify `abc001` for the `${your_k8s_cluster_id}` parameter when you install Log Service components for each Kubernetes cluster.

  > **Notice** This feature is unavailable for Kubernetes clusters that reside in different regions.

- Logtail container logs

  Logtail log files named `ilogtail.LOG` and `logtail_plugin.LOG` are stored in the `/usr/local/ilogtail/` directory of a Logtail container. Ignore the following standard output:

  ```
  start umount useless mount points, /shm$|/merged$|/mqueue$
  umount: /logtail_host/var/lib/docker/overlay2/3fd0043af174cb0273c3c7869500fbe2bdb95d13b1e110172ef57
  fe840c82155/merged: must be superuser to unmount
  umount: /logtail_host/var/lib/docker/overlay2/d5b10aa19399992755de1f85d25009528daa749c1bf8c16edff44
  beab6e69718/merged: must be superuser to unmount
  umount: /logtail_host/var/lib/docker/overlay2/5c3125daddacedec29df72ad0c52fac800cd56c6e880dc4e8a640
  b1e16c22dbe/merged: must be superuser to unmount
  ......
  xargs: umount: exited with status 255; aborting
  umount done
  start logtail
  ilogtail is running
  logtail status:
  ilogtail is running
  ```

- View the status of each Log Service component in a Kubernetes cluster

  ```
  helm status alibaba-log-controller
  ```

- Troubleshoot alibaba-log-controller startup failures

  Make sure that the following conditions are met:

  - Log Service components are installed on the master node of the Kubernetes cluster.
  - The Kubernetes cluster ID that you specified is valid when you install Log Service components.

  If Log Service components fail to be installled because the preceding conditions are not met, run the `helm del --purge alibaba-log-controller` command to delete the installation package and install Log Service components again.

- View the status of Logtail DaemonSet in a Kubernetes cluster

  Run the `kubectl get ds -n kube-system` command.

> **Note** The default namespace of Logtail is `kube-system`.

- View the version number, IP address, and startup time of Logtail.

  Example:

```
[root@iZbp1dsu6v77zfb40qfbiaZ ~]# kubectl get po -n kube-system | grep logtail
NAME            READY     STATUS    RESTARTS   AGE
logtail-ds-gb92k   1/1       Running   0          2h
logtail-ds-wm7lw   1/1       Running   0          4d
[root@iZbp1dsu6v77zfb40qfbiaZ ~]# kubectl exec logtail-ds-gb92k -n kube-system cat /usr/local/ilogt
ail/app_info.json
{
   "UUID" : "",
   "hostname" : "logtail-ds-gb92k",
   "instance_id" : "0EBB2B0E-0A3B-11E8-B0CE-0A58AC140402_172.20.4.2_1517810940",
   "ip" : "172.20.4.2",
   "logtail_version" : "0.16.2",
   "os" : "Linux; 3.10.0-693.2.2.el7.x86_64; #1 SMP Tue Sep 12 22:26:13 UTC 2017; x86_64",
   "update_time" : "2018-02-05 06:09:01"
}
```

- View the operational logs of Logtail

  Logtail operational logs are stored in the `ilogtail.LOG` file in the `/usr/local/ilogtail/` directory. If the log file is rotated and compressed, it is stored as a file named `ilogtail.LOG.x.gz`.

  Example:

```
[root@iZbp1dsu6v77zfb40qfbiaZ ~]# kubectl exec logtail-ds-gb92k -n kube-system tail /usr/local/ilog
tail/ilogtail.LOG
[2018-02-05 06:09:02.168693] [INFO] [9] [build/release64/sls/ilogtail/LogtailPlugin.cpp:104] logtai
l plugin Resume:start
[2018-02-05 06:09:02.168807] [INFO] [9] [build/release64/sls/ilogtail/LogtailPlugin.cpp:106] logtai
l plugin Resume:success
[2018-02-05 06:09:02.168822] [INFO] [9] [build/release64/sls/ilogtail/EventDispatcher.cpp:369] star
t add existed check point events, size:0
[2018-02-05 06:09:02.168827] [INFO] [9] [build/release64/sls/ilogtail/EventDispatcher.cpp:511] add
existed check point events, size:0 cache size:0 event size:0 success count:0
```

- Restart Logtail in a pod

  Example:

```
[root@iZbp1dsu6v77zfb40qfbiaZ ~]# kubectl exec logtail-ds-gb92k -n kube-system /etc/init.d/ilogtail
d stop
kill process Name: ilogtail pid: 7
kill process Name: ilogtail pid: 9
stop success
[root@iZbp1dsu6v77zfb40qfbiaZ ~]# kubectl exec logtail-ds-gb92k -n kube-system /etc/init.d/ilogtail
d start
ilogtail is running
```

23.1.3.1.6.3. Collect container text logs

Logtail can collect and upload container text logs together with container metadata to Log Service.

Features

Compared with basic log file collection, Docket file collection by using Logtail has the following features:

- Allows you to specify the log path of a container without the need to manually map the log path of the container to a path on the host.
- Uses labels to specify containers for log collection.
- Uses labels to exclude containers from log collection.
- Uses environment variables to specify containers for log collection.
- Uses environment variables to exclude containers from log collection.
- Supports multi-line logs such as Java Stack logs.
- Supports automatic labeling for Docker container logs.

> **Note**
> - The preceding labels are retrieved by using the docker inspect command. These labels are different from the labels that are specified in a Kubernetes cluster.
> - The preceding environment variables are the same as the environment variables that are specified to start containers.

Limits

- Stop policy: If a container is stopped and Logtail detects the `die` event on the container, Logtail stops collecting logs from the container. In this case, if a collection delay occurs, some logs that are generated before the stop action may be lost.
- Docker storage driver: Only overlay and overlay2 are supported. For other storage drivers, you must mount the log directory on the on-premises host.
- Logtail running mode: Logtail must run in a container and must be deployed based on Logtail deployment solutions.

Step 1: Deploy and configure Logtail

- Kubernetes

  For more information about how to collect Kubernetes logs, see Logtail deployment solution for collecting Kubernetes logs.

- Methods used to manage other containers

  For more information about the methods used to manage other containers, such as Swarm and Mesos, see Common deployment solution for collecting Docker logs.

Step 2: Create a Logtail configuration for log collection

1. Log on to the Log Service console.
2. Click **Import Data**. On the **Import Data** page, select **Docker File - Container**.
3. Select a destination project and Logstore, and then click **Next**.

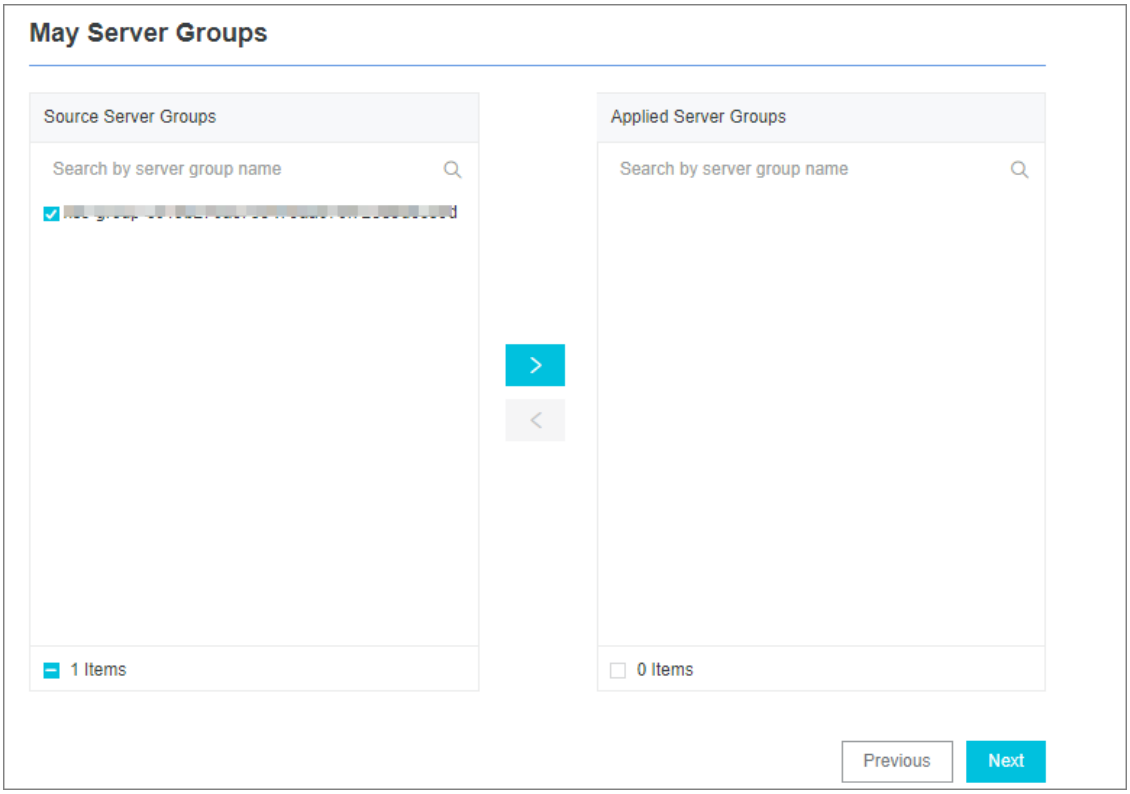
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration.

   The following table describes the specific parameters of data sources. For information about common parameters, see Configure text log collection.

Parameter	Description
Docker File	Checks whether the file that is collected from the specified data source is a Docker file.
Label Whitelist	If you want to specify a label whitelist, you must specify the LabelKey parameter. If the value of the LabelValue parameter is not empty, logs are collected from the containers whose label key-value pairs match the specified key-value pairs. If the value of the LabelValue parameter is empty, logs are collected from the containers whose label keys match the specified keys. **Note** ○ Key-value pairs are connected by the OR operator. If a label key-value pair of a container matches one of the specified key-value pairs, the logs of the container are collected. ○ The labels that are described in this topic refer to the label information in docker inspect.

Parameter	Description
Label Blacklist	If you want to specify a label blacklist, you must specify the LabelKey parameter. If the value of the LabelValue parameter is not empty, logs are not collected from the containers whose label key-value pairs match the specified key-value pairs. If the value of the LabelValue parameter is empty, logs are not collected from the containers whose label keys match the specified keys. **Note** ◦ Key-value pairs are connected by the OR operator. If a label key-value pair of a container matches one of the specified key-value pairs, the logs of the container are not collected. ◦ The labels that are described in this topic refer to the label information in docker inspect.
Environment Variable Whitelist	If you want to specify an environment variable whitelist, you must specify the EnvKey parameter. If the value of the EnvValue parameter is not empty, logs are collected from the containers whose environment variable key-value pairs match the specified key-value pairs. If the value of the EnvValue parameter is empty, logs are collected from the containers whose environment variable keys match the specified keys. **Note** ◦ Key-value pairs are connected by the OR operator. If an environment variable key-value pair of a container matches one of the specified key-value pairs, the logs of the container are collected. ◦ The environment variable refers to the environment information configured in container startup.
Environment Variable Blacklist	If you want to specify an environment variable blacklist, you must specify the EnvKey parameter. If the value of the EnvValue parameter is not empty, logs are not collected from the containers whose environment variable key-value pairs match the specified key-value pairs. If the value of the EnvValue parameter is empty, logs are not collected from the containers whose environment variable keys match the specified keys. **Note** ◦ Key-value pairs are connected by the OR operator. If an environment variable key-value pair of a container matches one of the specified key-value pairs, the logs of the container are not collected. ◦ The environment variable refers to the environment information configured in container startup.

> **Note**
>
> - Labels in a whitelist and a blacklist are different from the labels that are defined in Kubernetes. The labels that are described in this topic refer to the label information in docker inspect.
> - A namespace and a container name of a Kubernetes cluster can be mapped to a Docker label. The value of the LabelKey parameter for a namespace is `io.kubernetes.pod.namespace`. The value of the LabelKey parameter for a container name is `io.kubernetes.container.name`. For example, the namespace of the pod that you created is backend-prod and the container name is worker-server. In this case, you can set the key-value pair of a whitelist label to `io.kubernetes.pod.namespace : backend-prod` or `io.kubernetes.container.name : worker-server`. Then, you can collect logs from only the worker-server container.
> - In a Kubernetes cluster, we recommend that you specify only the `io.kubernetes.pod.namespace` and `io.kubernetes.container.name` labels. You can also specify the Environment Variable Whitelist parameter or the Environment Variable Blacklist parameter based on your business requirements.

7. Configure indexes in the Configure Query and Analysis step. Click **Next**.

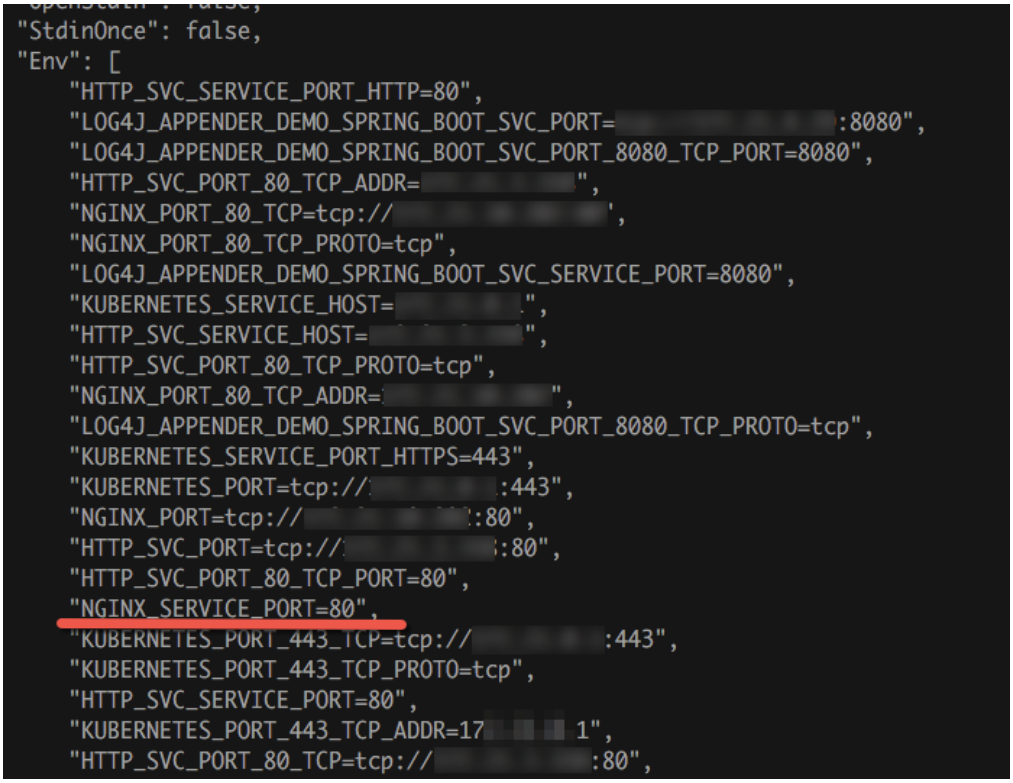
   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   >
   > - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

Configuration examples

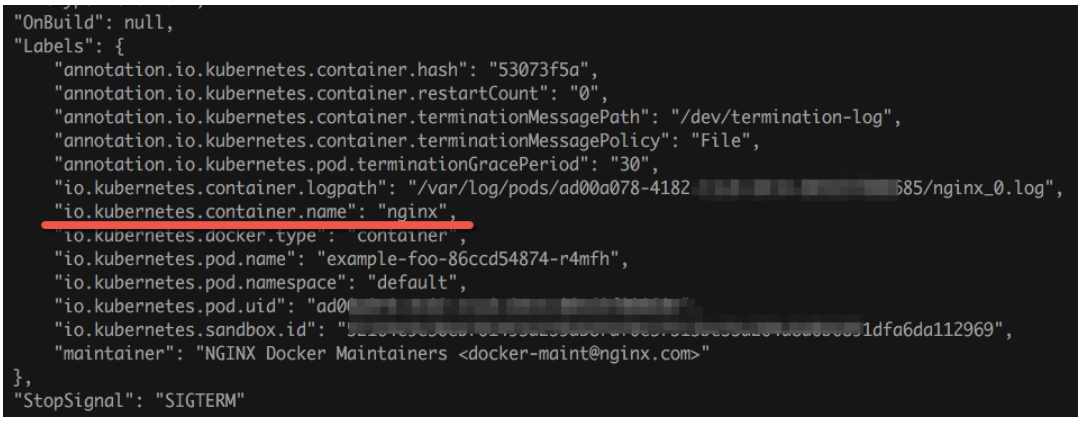
- Configure environment variables

  Collect the logs of the containers whose environment variables include `NGINX_PORT_80_TCP_PORT=80` and exclude `POD_NAMESPACE=kube-system`. The log file path is `/var/log/nginx/access.log` and logs are parsed in simple mode.



- Configure labels

  Collect the logs of the containers whose container labels include `io.kubernetes.container.name=nginx`. The log file path is `/var/log/nginx/access.log` and logs are parsed in simple mode.



Default fields

The following table describes the fields that are uploaded by default for each log of a common Docker container.

Log field	Description
`_image_name_`	The name of the image.
`_container_name_`	The name of the container.
`_container_ip_`	The IP address of the container.

The following table describes the fields that are uploaded by default for each log of a Kubernetes cluster.

Log field	Description
_image_name_	The name of the image.
_container_name_	The name of the container.
_pod_name_	The name of the pod.
namespace	The namespace where the pod resides.
_pod_uid_	The unique identifier of the pod.
_container_ip_	The IP address of the pod.

23.1.3.1.6.4. Collect stdout and stderr logs from containers

Logtail can collect and upload container standard output (stdout) and standard error (stderr) logs together with container metadata to Log Service. This topic describes how to create a Logtail configuration in the Log Service console to collect Kubernetes stdout and stderr logs.

Prerequisites

- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- The Helm package alibaba-log-controller is installed. For more information, see Install Logtail.

Features

Logtail can collect and upload container stdout and stderr logs together with container metadata to Log Service. Logtail has the following features when it collects Kubernetes stdout and stderr logs:

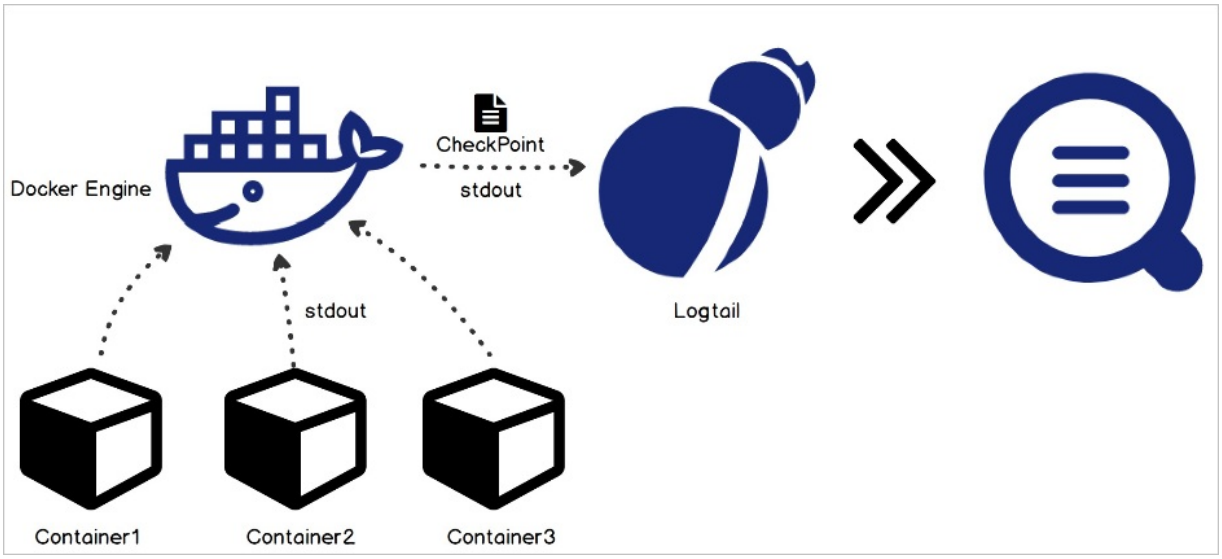
- Collects stdout and stderr logs in real time.
- Uses labels to specify containers for log collection.
- Uses labels to exclude containers from log collection.
- Uses environment variables to specify containers for log collection.
- Uses environment variables to exclude containers from log collection.
- Supports multi-line logs such as Java Stack logs.
- Supports automatic labeling for Docker container logs.
- Supports automatic labeling for Kubernetes container logs.

> **Note**
> - The preceding labels are retrieved by using the docker inspect command. These labels are different from the labels that are specified in a Kubernetes cluster.
> - The preceding environment variables are the same as the environment variables that are specified to start containers.

Implementation

A Logtail container uses a UNIX domain socket to communicate with the Docker daemon. The Logtail container queries all Docker containers and finds the specified Docker containers based on the specified labels and environment variables. Logtail uses the `docker logs` command to collect the logs of the specified Docker containers.

When Logtail collects the stdout and stderr logs of a Docker container, Logtail periodically stores checkpoints to a checkpoint file. If Logtail is restarted, Logtail collects logs from the last checkpoint.



Limits

- Logtail version: Only Logtail V0.16.0 or later that runs on Linux can be used to collect stdout and stderr logs. For more information, see Install Logtail in Linux.
- Permissions: By default, Logtail uses the `/var/run/docker.sock` socket to access the Docker engine. Make sure that a UNIX domain socket is available and the Logtail container has permissions to access the Docker engine.
- Multi-line logs: To ensure that a multi-line log is not split into multiple logs due to output latency, the multi-line log that is last collected is cached for 3 seconds by default. You can set the cache time by specifying the `BeginLineTimeoutMs` parameter. The value of the `BeginLineTimeoutMs` parameter cannot be less than 1,000 ms. Otherwise, an error may occur.
- Stop policy: If a container is stopped and Logtail detects the `die` event on the container, Logtail stops collecting the stdout and stderr logs of the container. In this case, if a collection delay occurs, some stdout or stderr logs that are generated before the stop action may be lost.
- Docker logging driver: The logging driver collects stdout and stderr logs only in JSON files.
- Context: By default, logs that are collected from different containers by using a Logtail configuration are in the same context. If you want the logs of each container to be in different contexts, create a Logtail configuration for each container.
- Data processing: The collected data is contained in the `content` field. You can process the data by using a common processing method. For more information, see Configure data processing methods.

Create a Logtail configuration

1. Log on to the Log Service console.
2. Click Import Data. On the **Import Data** page, select **Docker Standard Output - Container**.
3. Select a destination project and Logstore, and then click **Next**.

   You can also click **Create Now** to create a project and a Logstore.

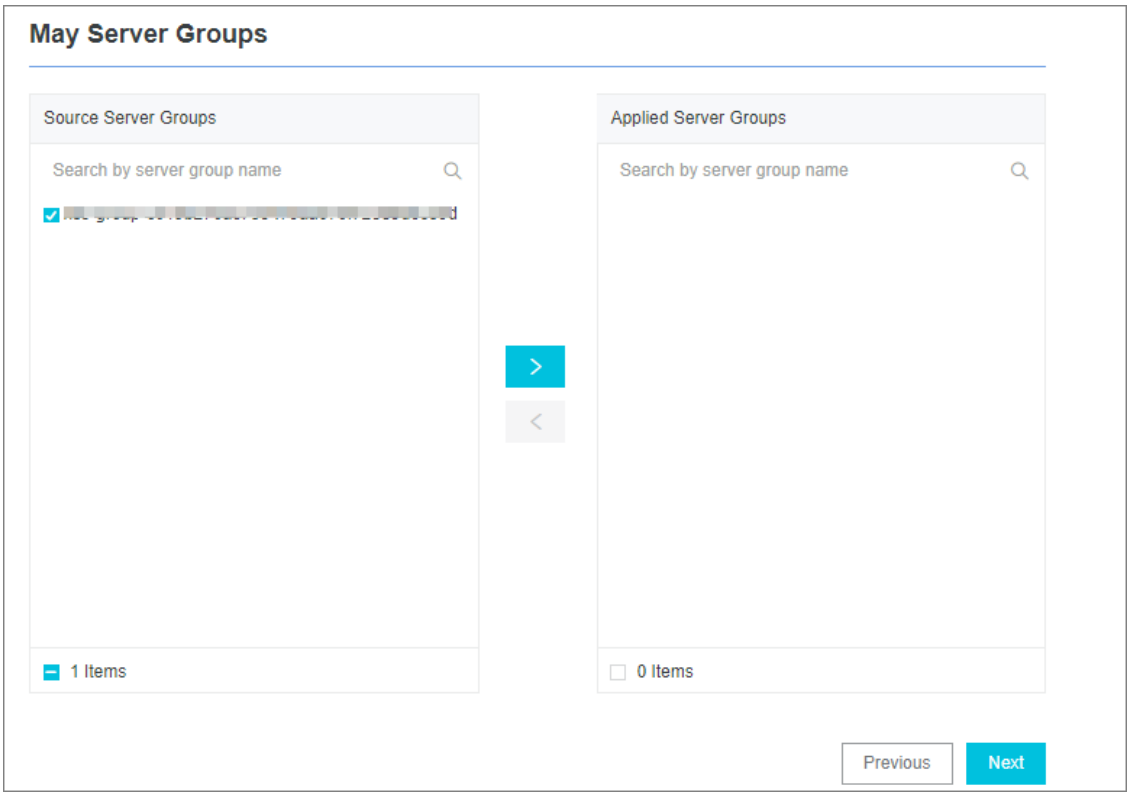
   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.

5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



> **Notice** If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be **FAIL**. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. In the Specify Data Source step, specify the data source and click **Next**.

   Enter configurations for log collection in the **Plug-in Config** field. The following example shows the parameters:

```
{
 "inputs": [
     {
         "type": "service_docker_stdout",
         "detail": {
             "Stdout": true,
             "Stderr": true,
             "IncludeLabel": {
                 "io.kubernetes.container.name": "nginx"
             },
             "ExcludeLabel": {
                 "io.kubernetes.container.name": "nginx-ingress-controller"
             },
             "IncludeEnv": {
                 "NGINX_SERVICE_PORT": "80"
             },
             "ExcludeEnv": {
                 "POD_NAMESPACE": "kube-system"
             }
         }
     }
 ]
}
```

The type of the input source is `service_docker_stdout`.

Parameter	Type	Required	Description
IncludeLabel	Map	Yes	The value of the IncludeLabel parameter is a map. The keys and values of the map are strings. The default value of this parameter is an empty map. This default value indicates that logs from all containers are collected. If the keys are not empty and the values are empty, logs are collected from the containers whose label keys match the specified keys. **Note** ○ Key-value pairs are connected by the OR operator. If a label key-value pair of a container matches one of the specified key-value pairs, the logs of the container are collected. ○ By default, the values in the map are strings. Logs are collected from the containers whose names match the values. If you use a regular expression to specify a value, logs are collected from the containers whose names match the regular expression. If you specify a value that starts with a caret (^) and ends with a dollar sign ($), for example, ^(kube-system\|istio-system)$, logs are collected from a container named kube-system and a container named istio-system.

Parameter	Type	Required	Description
ExcludeLabel	Map	No	The value of the ExcludeLabel parameter is a map. The keys and values of the map are strings. The default value of this parameter is an empty map and indicates that logs from all containers are collected. If the keys are not empty and the values are empty, logs are not collected from the containers whose label keys match the specified keys.
IncludeEnv	Map	No	The value of the IncludeEnv parameter is a map. The keys and values of the map are strings. The default value of this parameter is an empty map and indicates that logs from all containers are collected. If the keys are not empty and the values are empty, logs are collected from the containers whose environment variable keys match the specified keys. **Note** ◦ Key-value pairs are connected by the OR operator. If an environment variable key-value pair of a container matches one of the specified key-value pairs, the logs of the container are collected. ◦ By default, the values in the map are strings. Logs are collected from the containers whose names match the values. If you use a regular expression to specify a value, logs are collected from the containers whose names match the regular expression. If you specify a value that starts with a caret (^) and ends with a dollar sign ($), for example, ^(kube-system\|istio-system)$, logs are collected from a container named kube-system and a container named istio-system.

Parameter	Type	Required	Description
ExcludeEnv	Map	No	The value of the ExcludeEnv parameter is a map. The keys and values of the map are strings. The default value of this parameter is an empty map and indicates that logs from all containers are collected. If keys are not empty and values are empty, logs are not collected from the containers whose environment variable keys match the specified keys. **Note** ◦ Key-value pairs are connected by the OR operator. If an environment variable key-value pair of a container matches one of the specified key-value pairs, the logs of the container are not collected. ◦ By default, the values in the map are strings. Logs are not collected from the containers whose names match the values. If you use a regular expression to specify a value, logs are not collected from the containers whose names match the regular expression. If you specify a value that starts with a caret (^) and ends with a dollar sign ($), for example, ^(kube-system\|istio-system)$, logs are not collected from a container named kube-system or a container named istio-system.
Stdout	bool	No	Default value: true. If you set the value of this parameter to false, stdout logs are not collected.
Stderr	bool	No	Default value: true. If you set the value of this parameter to false, stderr logs are not collected.
BeginLineRegex	string	No	The regular expression that is used to match the start part in the first line of a log. The default value of this parameter is an empty string. If a line matches the specified regular expression, the line is recorded to be the first line of a new log. Otherwise, the line is recorded to be a part of the last log.
BeginLineTimeoutMs	int	No	The timeout period for the specified regular expression to match the start part in the first line of a log. Default value: 3000. Unit: ms. If no new log is generated within 3 seconds, the last log is uploaded.

Parameter	Type	Required	Description
BeginLineCheckLength	int	No	The size of the start part in the first line of a log that matches the specified regular expression. Default value: 10 × 1,024. Unit: bytes. You can specify this parameter to check whether the start part in the first line of a log matches the regular expression. This improves match efficiency.
MaxLogSize	int	No	The maximum size of a log. Default value: 512 × 1,024. Unit: bytes. If the size of a log exceeds the specified value, the log is uploaded.

> **Note**
> - The preceding IncludeLabel and ExcludeLabel parameters are included in the label information that is retrieved by using the `docker inspect` command.
> - A namespace and a container name of a Kubernetes cluster can be mapped to a Docker label. The value of the LabelKey parameter for a namespace is `io.kubernetes.pod.namespace`. The value of the LabelKey parameter for a container name is `io.kubernetes.container.name`. For example, the namespace of the pod that you created is backend-prod and the container name is worker-server. In this case, if you set the key-value pair of a whitelist label to `io.kubernetes.pod.namespace : backend-prod`, the logs of all containers in the pod are collected. If you set the key-value pair of a whitelist label to `io.kubernetes.container.name : worker-server`, the logs of the container are collected.
> - In a Kubernetes cluster, we recommend that you specify only the `io.kubernetes.pod.namespace` and `io.kubernetes.container.name` labels. You can also specify the IncludeEnv or ExcludeEnv parameter based on your business requirements.

7. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

> **Note**
> - To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
> - If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

Default fields

- Common Docker containers

  The following table describes the fields that are uploaded by default for each log of a common Docker container.

Log field	Description
`_time_`	The point in time when data is uploaded. Example: `2018-02-02T02:18:41.979147844Z`.
`_source_`	The type of the input source. Valid values: stdout and stderr.

Log field	Description
_image_name_	The name of the image.
_container_name_	The name of the container.
_container_ip_	The IP address of the container.

- Kubernetes

  The following table describes the fields that are uploaded by default for each log of a Kubernetes cluster.

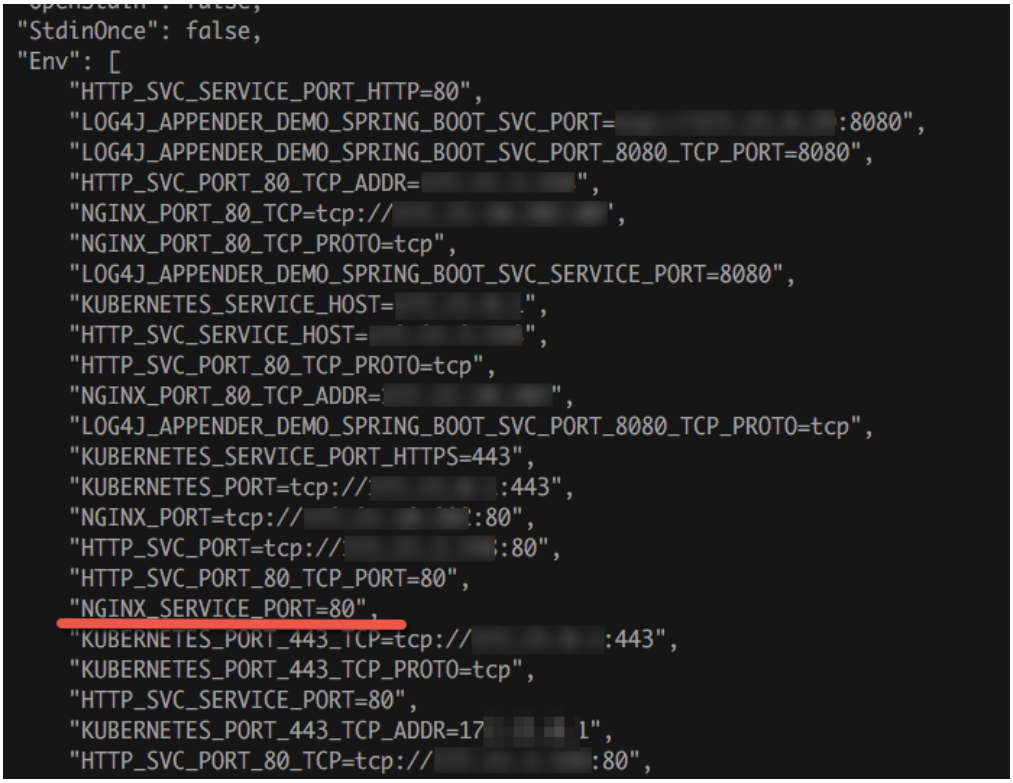
Log field	Description
time	The point in time when data is uploaded. Example: 2018-02-02T02:18:41.979147844Z.
source	The type of the input source. Valid values: stdout and stderr.
_image_name_	The name of the image.
_container_name_	The name of the container.
_pod_name_	The name of the pod.
namespace	The namespace where the pod resides.
_pod_uid_	The unique identifier of the pod.
_container_id_	The IP address of the pod.

Configuration examples of single-line log collection

- Configure environment variables

  Collect the stdout and stderr logs of the containers whose environment variables include NGINX_PORT_80_TCP_PORT=80 and exclude POD_NAMESPACE=kube-system.

  Configuration example of environment variables



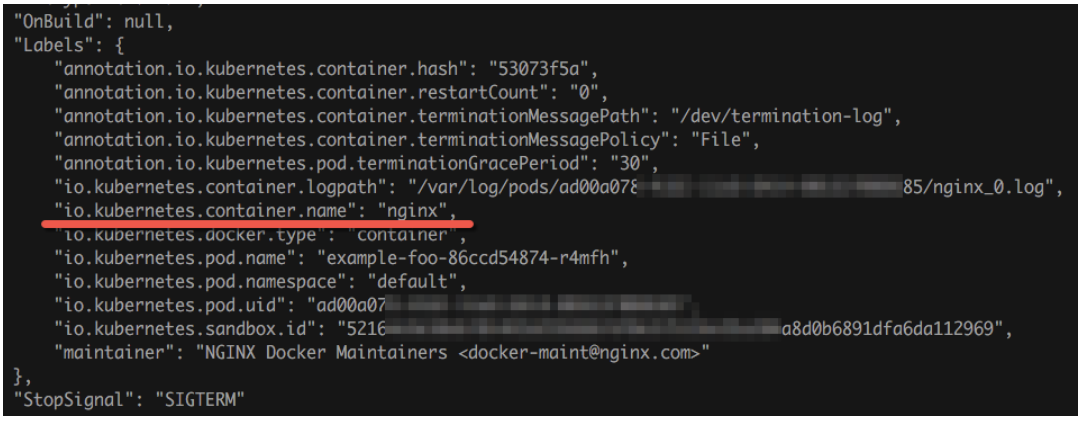
The following script shows the configurations of the environment variables:

```
{
    "inputs": [
        {
            "type": "service_docker_stdout",
            "detail": {
                "Stdout": true,
                "Stderr": true,
                "IncludeEnv": {
                    "NGINX_PORT_80_TCP_PORT": "80"
                },
                "ExcludeEnv": {
                    "POD_NAMESPACE": "kube-system"
                }
            }
        }
    ]
}
```

- Configure labels

  Collect the stdout and stderr logs of the containers whose labels include `io.kubernetes.container.name=nginx` and exclude `type=pre`.

  Configuration example of labels



The following script shows the label configurations:

```
{
    "inputs": [
        {
            "type": "service_docker_stdout",
            "detail": {
                "Stdout": true,
                "Stderr": true,
                "IncludeLabel": {
                    "io.kubernetes.container.name": "nginx"
                },
                "ExcludeLabel": {
                    "type": "pre"
                }
            }
        }
    ]
}
```

Configuration examples of multi-line log collection

Before you can collect Java exception stack logs, you must configure multi-line log collection. The following section describes how to collect stdout and stderr logs of standard Java applications.

- Sample log

```
2018-02-03 14:18:41.968  INFO [spring-cloud-monitor] [nio-8080-exec-4] c.g.s.web.controller.DemoCon
troller : service start
2018-02-03 14:18:41.969 ERROR [spring-cloud-monitor] [nio-8080-exec-4] c.g.s.web.controller.DemoCon
troller : java.lang.NullPointerException
at org.apache.catalina.core.ApplicationFilterChain.internalDoFilter(ApplicationFilterChain.java:193
)
at org.apache.catalina.core.ApplicationFilterChain.doFilter(ApplicationFilterChain.java:166)
at org.apache.catalina.core.StandardWrapperValve.invoke(StandardWrapperValve.java:199)
at org.apache.catalina.core.StandardContextValve.invoke(StandardContextValve.java:96)
...
2018-02-03 14:18:41.968  INFO [spring-cloud-monitor] [nio-8080-exec-4] c.g.s.web.controller.DemoCon
troller : service start done
```

- Log collection configuration

  Collect the logs of the containers whose labels include `app=monitor` and the specified first bytes of a line is of a fixed-format date type. To improve match efficiency, only the first 10 bytes of each line are checked.

```
{
"inputs": [
  {
    "detail": {
      "BeginLineCheckLength": 10,
      "BeginLineRegex": "\\d+-\\d+-\\d+.*",
      "IncludeLabel": {
        "app": "monitor"
      }
    },
    "type": "service_docker_stdout"
  }
]
}
```

Data processing examples

Logtail can process the collected Docker standard output. For more information, see Common data processing methods.

- Collect the logs of the containers whose labels include `app=monitor` and the specified first bytes of a line is of a fixed-format data type. To improve match efficiency, only the first 10 bytes of each line are checked. Regular expressions are used to parse logs into the values of the time, level, module, thread, and message. The following script shows the configurations of log collection and data processing:

```
{
"inputs": [
  {
    "detail": {
      "BeginLineCheckLength": 10,
      "BeginLineRegex": "\\d+-\\d+-\\d+.*",
      "IncludeLabel": {
        "app": "monitor"
      }
    },
    "type": "service_docker_stdout"
  }
],
"processors": [
    {
        "type": "processor_regex",
        "detail": {
            "SourceKey": "content",
            "Regex": "(\\d+-\\d+-\\d+ \\d+:\\d+:\\d+\\.\\d+)\\s+(\\w+)\\s+\\[([^]]+)]\\s+\\[([^]]+)
]\\s+:\\s+([\\s\\S]*)",
            "Keys": [
                "time",
                "level",
                "module",
                "thread",
                "message"
            ],
            "NoKeyError": true,
            "NoMatchError": true,
            "KeepSource": false
        }
    }
]
}
```

The collected log `2018-02-03 14:18:41.968 INFO [spring-cloud-monitor] [nio-8080-exec-4] c.g.s.web.controller.DemoController : service start done` is processed, as shown in the following script:

```
__tag__:__hostname__:logtail-dfgef
_container_name_:monitor
_image_name_:registry.cn-hangzhou.aliyuncs.xxxxxxxxxxxxxx
_namespace_:default
_pod_name_:monitor-6f54bd5d74-rtzc7
_pod_uid_:7f012b72-04c7-11e8-84aa-00163f00c369
_source_:stdout
_time_:2018-02-02T14:18:41.979147844Z
time:2018-02-02 02:18:41.968
level:INFO
module:spring-cloud-monitor
thread:nio-8080-exec-4
class:c.g.s.web.controller.DemoController
message:service start done
```

- Collect the JSON logs of the containers whose labels include `app=monitor`. The following script shows the configurations of log collection and data processing:

```
{
"inputs": [
  {
    "detail": {
      "IncludeLabel": {
        "app": "monitor"
      }
    },
    "type": "service_docker_stdout"
  }
],
"processors": [
    {
        "type": "processor_json",
        "detail": {
            "SourceKey": "content",
            "NoKeyError":true,
            "KeepSource": false
        }
    }
]
}
```

23.1.3.1.7. Limits

This topic describes the limits of Logtail. These limits apply when you collect files, manage resources, and resolve errors.

Limits on file collection

Item	Description
File encoding	Log files can be encoded in UTF-8 and GBK. To improve processing performance, we recommend that you encode log files in UTF-8. If log files are encoded in other formats, errors such as garbled characters and data loss may occur.

Item	Description
Log file size	Unlimited.
Log file rotation	Supported. Both `.log*` and `.log` are supported for file names.
Log collection behavior when log parsing is blocked	When log parsing is blocked, Logtail keeps the log file descriptor (FD) open. If log file rotation occurs multiple times during the blocking period, Logtail attempts to parse new log files in sequence. If the number of new log files that are not parsed exceeds 20, Logtail does not process the excess log files.
Symbolic link	Monitored directories can be symbolic links.
Size of a single log	The maximum size of a single log is 512 KB. If a regular expression is used to split a multi-line log to match the start part in the first line of the log, the maximum size of each log after splitting is still 512 KB. If the size of a log exceeds 512 KB, the log is forcibly split into multiple parts and collected. For example, if the size of a log is 1,025 KB, the log is split into three parts of the following sizes: 512 KB, 512 KB, and 1 KB. Then, the log parts are collected in sequence.
Regular expression	Perl-based regular expressions can be used.
Multiple Logtail configurations for the same log file	Not supported. We recommend that you collect and store log files to one Logstore, and then configure multiple subscriptions. If this feature is required, configure symbolic links for log files to bypass this limit.
File opening behavior	When Logtail collects data from a log file, Logtail keeps the log file open. If the log file is not updated for more than 5 minutes and log rotation does not occur, Logtail closes the log file.
First log collection behavior	Logtail collects data only from incremental log files. If the size of a log file exceeds 1 MB the first time an update to the log file is detected, Logtail collects data from the last 1 MB. If the log file size does not exceed 1 MB, Logtail collects data from the beginning of the log file. If the log file is not updated after the Logtail configuration is delivered, Logtail does not collect data from the log file.
Non-standard text logs	If a log contains `'\0'` in multiple lines, the log is truncated at the first `'\0'`.

Limits on checkpoints

Item	Description
Checkpoint timeout period	If a log file is not updated for more than 30 days, the checkpoint of the log file is deleted.
Checkpoint storage policy	Checkpoints are saved every 15 minutes and are automatically saved when you exit Logtail.
Checkpoint storage path	By default, checkpoints are stored in the `/tmp/logtail_checkpoint` directory. You can modify the values of the related parameters. For more information, see Set Logtail startup parameters.

Limits on configurations

Item	Description
Configuration update	A custom configuration update requires approximately 30 seconds to take effect.
Dynamic loading of Logtail configurations	Supported. The update of a Logtail configuration does not affect other Logtail configurations.
Number of Logtail configurations	Unlimited. However, we recommend that you create a maximum of 100 Logtail configurations on a server.
Multi-tenant isolation	Logtail configurations for different tenants are isolated.

Limits on resources and performance metrics

Item	Description
Throughput for log processing	The default transmission speed of raw logs is limited to 2 MB/s. Log data is uploaded after it is encoded and compressed. The compression ratio ranges from 5:1 to 10:1. If the transmission speed exceeds the limit, log data may be lost. You can modify the values of the related parameters. For more information, see Set Logtail startup parameters.
Maximum processing speed	Single-core processing speed: The maximum processing speed is 100 MB/s for logs in simple mode, 40 MB/s for logs in delimiter mode, and 30 MB/s for logs in JSON mode. By default, the maximum processing speed is 20 MB/s for logs in full regex mode based on the complexity of regular expressions. If multiple processing threads are enabled, the performance can be improved by 1.5 to 3 times.
Number of monitored directories	Logtail limits the depth of monitored directories to reduce the consumption of your resources. If the upper limit is reached, Logtail stops monitoring additional directories or log files. Logtail can monitor a maximum of 3,000 directories, including subdirectories.
Number of monitored files	By default, you can use a Logtail configuration on each server to monitor a maximum of 10,000 files. By default, a Logtail client on each server can monitor a maximum of 100,000 files. Excessive files are not monitored. If the upper limit is reached, you can perform the following operations: • Improve the depth of the monitored directory in each Logtail configuration. • Increase the value of the mem_usage_limit parameter to raise the threshold of memory resources that are available for Logtail. For more information, see Set Logtail startup parameters. You can raise the threshold to a maximum of 2 GB. This way, the maximum number of files that can be monitored by using each Logtail configuration is increased to 100,000, and the maximum number of files that the Logtail client on each server can monitor is increased to 1,000,000.
Default resources	By default, Logtail occupies a maximum of 40% of the CPU and 256 MB of memory. If logs are generated at a high speed, you can modify the values of the related parameters. For more information, see Set Logtail startup parameters.
Processing policy of threshold-crossing resources	If the resources that are occupied by Logtail exceed the upper limit and this issue lasts for 5 minutes or more, Logtail is forcibly restarted. The restart may cause data loss or duplication.

Limits on error handling

Item	Description
Network error handling	If a network error occurs, Logtail automatically retries and adjusts the retry interval.
Processing policy of threshold-crossing resources	If the data transmission speed exceeds the quota of the Logstore, Logtail restricts the log collection speed and retries the log collection.
Maximum retry period before timeout	If data fails to be transmitted and the issue lasts for more than six consecutive hours, Logtail discards the data.
Status self-check	Logtail restarts if an exception occurs, for example, an application unexpectedly exits or the resource usage exceeds the quota.

Other limits

Item	Description
Log collection latency	A latency of less than 1 second exists between the point in time when a log is written to a disk and the point in time when Logtail collects the log. However, if the log collection speed is restricted, the latency increases.
Log upload policy	Before Logtail uploads logs, Logtail aggregates the logs in the same file. The log upload starts if the number of logs exceeds 2,000, the total size of logs exceeds 2 MB, or the log collection duration exceeds 3 seconds.

23.1.3.2. Other collection methods

23.1.3.2.1. Use the web tracking feature to collect logs

Log Service provides the web tracking feature that you can use to collect logs from the HTML, HTML5, iOS, and Android platforms. You can also customize dimensions and metrics to collect logs. This topic describes how to use the web tracking feature to collect logs.

Context

You can use the web tracking feature to collect user information from browsers, iOS apps, or Android apps. The information includes:

- Browsers, operating systems, and resolutions that are used by users.
- User browsing behavior, such as the number of clicks and purchases on a website.
- The amount of time that users spend on an app and whether users are active users.

Usage notes

- After you enable the web tracking feature for a Logstore, the write permissions on the Logstore are granted to anonymous users from the Internet. This may generate dirty data.
- The HTTP body of each GET request cannot exceed 16 KB.
- You can use the POST method to call the PutLogs API operation and write a maximum of 3 MB or 4,096 log entries to Log Service.

Step 1: Enable the web tracking feature

You can use the Log Service console or an SDK to enable the web tracking feature.

- Enable the web tracking feature in the Log Service console.

i. Log on to the Log Service console.
ii. In the **Projects** section, click the project in which you want to enable the web tracking feature for a Logstore.
iii. Find the Logstore for which you want to enable the web tracking feature and choose > **Modify**.
iv. In the upper-right corner of the **Logstore Attributes** page, click **Modify**.
v. Turn on **WebTracking** and click **Save**.

- Use an SDK to enable the web tracking feature.

  The following script shows how to use Log Service SDK for Java to enable the web tracking feature:

```
import com.aliyun.openservices.log.Client;
import com.aliyun.openservices.log.common.LogStore;
import com.aliyun.openservices.log.exception.LogException;
public class WebTracking {
  static private String accessId = "your accesskey id";
  static private String accessKey = "your accesskey";
  static private String project = "your project";
  static private String host = "log service data address";
  static private String logStore = "your logstore";
  static private Client client = new Client(host, accessId, accessKey);
  public static void main(String[] args) {
      try {
          // Enable the web tracking feature for an existing Logstore.
          LogStore logSt = client.GetLogStore(project, logStore).GetLogStore();
          client.UpdateLogStore(project, new LogStore(logStore, logSt.GetTtl(), logSt.GetShardCount
(), true));
          // Disable the web tracking feature.
          //client.UpdateLogStore(project, new LogStore(logStore, logSt.GetTtl(), logSt.GetShardCou
nt(), false));
          // Create a Logstore for which you want to enable the web tracking feature.
          //client.UpdateLogStore(project, new LogStore(logStore, 1, 1, true));
      }
      catch (LogException e){
          e.printStackTrace();
      }
  }
}
```

Step 2: Collect logs

After you enable the web tracking feature for a Logstore, you can upload logs to a Logstore by using the following methods:

- Use SDK for JavaScript to upload logs.
  i. Install the dependency.

```
npm install --save js-sls-logger
```

  ii. Import the application module.

```
import SlsWebLogger from 'js-sls-logger'
```

  iii. Set the opts parameter. The following table describes the parameters.

```
const opts = {
  host: 'cn-qingdao-env12-d01.sls-pub.cloud.env12.shuguang.com',
  project: 'my_project_name',
  logstore: 'my_logstore_name',
  time: 10,
  count: 10,
}
```

Parameter	Required	Description
host	Yes	The endpoint of the region where Log Service resides. In this example, the endpoint of the China (Hangzhou) region is used. Replace the value of the parameter with the actual endpoint. For more information, see the **Obtain an endpoint** topic in *Log Service Developer Guide*.
project	Yes	The name of the project.
logstore	Yes	The name of the Logstore.
time	No	The time interval at which logs are sent. Default value: 10. Unit: seconds.
count	No	The number of logs that are sent. Default value: 10.

iv. Create SlsWebLogger.

```
const logger = new SlsWebLogger(opts)
```

v. Upload logs.

```
logger.send({
    customer: 'zhangsan',
    product: 'iphone 12',
    price: 7998
})
```

- Use the GET method to upload logs.

  Run the following command to upload logs. Replace the values of the parameters based on your business requirements. The following table describes the parameters.

```
curl --request GET 'http://${project}.${host}/logstores/${logstore}/track?APIVersion=0.6.0&key1=val1&key2=val2'
```

Parameter	Required	Description
${project}	Yes	The name of the project.
${host}	Yes	The endpoint of the region where Log Service resides. For more information, see the **Obtain an endpoint** topic in *Log Service Developer Guide*.
${logstore}	Yes	The name of the Logstore.
APIVersion=0.6.0	Yes	A reserved parameter.
__topic__=yourtopic	No	The topic of the log that you want to upload.

Parameter	Required	Description
key1=val1&key2=val2	Yes	The key-value pairs that you want to upload to Log Service. Make sure that the data size is less than 16 KB.

- Use HTML <img> tags to upload logs.

```
<img src='http://${project}.${host}/logstores/${logstore}/track.gif?APIVersion=0.6.0&key1=val1&key2=val2'/>
<img src='http://${project}.${host}/logstores/${logstore}/track_ua.gif?APIVersion=0.6.0&key1=val1&key2=val2'/>
```

  The *track_ua.gif* file contains custom parameters that you want to upload to Log Service. If you use this method to upload logs, Log Service records the custom parameters and the User-Agent and Referer HTTP headers as log fields.

  > **Note** To collect the Referer HTTPS header, make sure that the URL in the preceding <img> tag uses the HTTPS protocol.

- Use the POST method to upload logs.

  You can send an HTTP POST request to upload a large amount of data. For more information, see the "PutWebtacking" topic of **API Reference** in *Log Service Developer Guide*.

23.1.3.2.2. Use SDKs to collect logs

23.1.3.2.2.1. Producer Library

The Aliyun LOG Java Producer supports Java applications that run in big data processing scenarios with high concurrency. The library is easy to use and highly customizable.

For more information about the related GitHub project, visit Aliyun LOG Java Producer.

23.1.3.2.2.2. Log4j Appender

This topic describes Alibaba Cloud Log4j Appender.

Log4j is an open source project of Apache. Log4j allows you to specify the output destination and format of logs. You can also specify the severity level of each log for fine-grained control on log generation. Log4j consists of the following three components:

- Loggers

  The severity levels of logs are classified into ERROR, WARN, INFO, and DEBUG in descending order.

- Appenders

  An appender specifies that logs are sent to the Log Service console or files.

- Layouts

  A layout specifies the output format of logs.

You can use Alibaba Cloud Log4j Appender to send logs to Log Service. For more information about Alibaba Cloud Log4j Appender, visit Log4j Appender.

23.1.3.2.2.3. Logback Appender

This topic describes how to write logs to Log Service by using Aliyun Log Logback Appender.

Logback is an open source project that is developed by the founder of Log4j. Logback allows you to write logs to multiple destinations. These destinations include the Log Service console, files, graphical user interface (GUI) components, socket servers, NT kernel loggers, and UNIX syslog daemons. You can specify the output format of each log. You can also specify the severity level of each log for fine-grained control on log generation.

The following example shows the format of a log that is written to Log Service by using Aliyun Log Logback Appender:

```
level: ERROR
location: com.aliyun.openservices.log.logback.example.LogbackAppenderExample.main(LogbackAppenderExamp
le.java:18)
message: error log
throwable: java.lang.RuntimeException: xxx
thread: main
time: 2018-01-02T03:15+0000
log: 2018-01-02 11:15:29,682 ERROR [main] com.aliyun.openservices.log.logback.example.LogbackAppenderE
xample: error log
__source__: xxx
__topic__: yyy
```

For more information about Aliyun Log Logback Appender, see Logback Appender.

23.1.3.2.2.4. Golang Producer Library

The Aliyun LOG Go Producer Library supports Go applications that run in big data processing scenarios with high concurrency. The library is easy to use and highly customizable. You can use the library to create producers that allow you to resend failed logs. Before Go applications send log data to Log Service, you can use these producers to compress the log data. This improves write performance.

For more information about the related GitHub project, visit Aliyun Log Go Producer.

23.1.3.2.2.5. Python logging

This topic describes how to use the Python logging module to collect log data.

Configurations

For more information about the configurations that are related to the Python logging module, see Logging configuration.

The Python logging module allows you to use code or a configuration file to configure logging. The following example shows how to use the `logging.conf` configuration file to configure logging.

```
[loggers]
keys=root,sls
[handlers]
keys=consoleHandler, slsHandler
[formatters]
keys=simpleFormatter, rawFormatter
[logger_root]
level=DEBUG
handlers=consoleHandler
[logger_sls]
level=INFO
handlers=consoleHandler, slsHandler
qualname=sls
propagate=0
[handler_consoleHandler]
class=StreamHandler
level=DEBUG
formatter=simpleFormatter
args=(sys.stdout,)
[handler_slsHandler]
class=aliyun.log.QueuedLogHandler
level=INFO
formatter=rawFormatter
args=(os.environ.get('ALIYUN_LOG_SAMPLE_ENDPOINT', ''), os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSID', '
'), os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSKEY', ''), os.environ.get('ALIYUN_LOG_SAMPLE_TMP_PROJECT',
''), "logstore")
[formatter_simpleFormatter]
format=%(asctime)s - %(name)s - %(levelname)s - %(message)s
[formatter_rawFormatter]
format=%(message)s
```

Two handlers named `root` and `sls` are created. The `sls` handler is an object of the `aliyun.log.QueuedLogHandler` class. The following script shows the parameters that you can specify for the sls handler. For more information, see Parameters.

```
args=(os.environ.get('ALIYUN_LOG_SAMPLE_ENDPOINT', ''), os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSID', '
'), os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSKEY', ''), os.environ.get('ALIYUN_LOG_SAMPLE_TMP_PROJECT',
''), "logstore")
```

Note In this case, the `os.environ` function is used to obtain configurations from environment variables. You can also specify values for these parameters based on your business requirements.

Upload logs

If you want to upload logs to Log Service, you can use the configuration file.

```
import logging
import logging.config
# Configurations
logging.config.fileConfig('logging.conf')
logger = logging.getLogger('sls')
# Use the logger
logger.info("test1")
try:
    1/0
except ZeroDivisionError as ex:
    logger.exception(ex)
```

Then, logs are automatically uploaded to Log Service. If you want to use the query and analysis feature, you must enable the indexing feature for the related Logstore.

Configure indexes for a Logstore

Enable the indexing feature for the Logstore that receives logs and configure indexes for specific fields. We recommend that you use the Log Service command-line interface (CLI) to configure indexes. For more information, see python_logging_handler_index.json.

```
aliyunlog log update_index --project_name="project1" --logstore_name="logstore1" --index_detail="file:
///Users/user1/loghandler_index.json"
```

Specify log fields that you want to collect

The following table describes the log fields that you can collect.

Field	Description
message	The content of a log.
record_name	The name of a handler. In the preceding example, `sls` is used.
level	The severity level of a log, such as INFO and ERROR.
file_path	The full path of a configuration file.
func_name	The name of a function.
line_no	The number of a log line.
module	The name of a module where the function resides.
thread_id	The ID of the thread that runs the function.
thread_name	The name of the thread that runs the function.
process_id	The ID of the process that runs the function.
process_name	The name of the process that runs the function.

You can specify log fields that you want to collect based on the `fields` parameter of a class. For more information, see aliyun.log.LogFields.

The following example shows how to modify the preceding configuration file and collect several fields, such as module and func_name.

```
[handler_slsHandler]
class=aliyun.log.QueuedLogHandler
level=INFO
formatter=rawFormatter
args=('cn-beijing.log.aliyuncs.com', 'ak_id', 'ak_key', 'project1', "logstore1", 'mytopic', ['level',
'func_name', 'module', 'line_no']  )
```

> **Note**
> - The message field is collected regardless of your configurations.
> - If you want to add a prefix and suffix to the names of these fields, use the `buildin_fields_prefix` and `buildin_fields_suffix` parameters. Example: `__level__`.

Use a JSON text to configure logging

If you want to create flexible logging configurations, you can use a JSON text.

```
#encoding: utf8
import logging, logging.config, os
# Configurations
conf = {'version': 1,
        'formatters': {'rawformatter': {'class': 'logging.Formatter',
                                        'format': '%(message)s'}
                       },
        'handlers': {'sls_handler': {'()':
                                     'aliyun.log.QueuedLogHandler',
                                     'level': 'INFO',
                                     'formatter': 'rawformatter',
                                     # custom args:
                                     'end_point': os.environ.get('ALIYUN_LOG_SAMPLE_ENDPOINT', ''),
                                     'access_key_id': os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSID', '')
,
                                     'access_key': os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSKEY', ''),
                                     'project': 'project1',
                                     'log_store': "logstore1"
                                     }
                     },
        'loggers': {'sls': {'handlers': ['sls_handler', ],
                                  'level': 'INFO',
                                  'propagate': False}
                    }
        }
logging.config.dictConfig(conf)
# Use the logger
logger = logging.getLogger('sls')
logger.info("Hello world")
```

> **Note** If you want to instantiate an object of the `aliyun.log.QueuedLogHandler` class, pass named parameters to the constructor. For more information, see Parameters.

23.1.3.2.3. Collect common logs

23.1.3.2.3.1. Collect Log4j logs

Log Service allows you to use LogHub Log4j Appender or Logtail to collect Log4j logs.

Log format

Log4j is an open source project of Apache. Log4j allows you to specify the output destination and format of logs. You can also specify the severity level of logs. The severity levels of logs are classified into ERROR, WARN, INFO, and DEBUG in descending order. The output destination specifies whether logs are sent to the console or files. The output format specifies the format of logs. The following example shows the default configurations of Log4j:

```
<Configuration status="WARN">
  <Appenders>
    <Console name="Console" target="SYSTEM_OUT">
      <PatternLayout pattern="%d{yyyy-MM-dd HH:mm:ss:SSS zzz} [%t] %-5level %logger{36} - %msg%n"/>
    </Console>
  </Appenders>
  <Loggers>
    <Logger name="com.foo.Bar" level="trace">
      <AppenderRef ref="Console"/>
    </Logger>
    <Root level="error">
      <AppenderRef ref="Console"/>
    </Root>
  </Loggers>
</Configuration>
```

The following example shows a sample log:

```
2013-12-25 19:57:06,954 [10.10.10.10] WARN impl.PermanentTairDaoImpl - Fail to Read Permanent Tair,key
:e:470217319319741_1,result:com.example.tair.Result@172e3ebc[rc=code=-1, msg=connection error or timeo
ut,value=,flag=0]
```

Regular expression that matches each IP address that indicates the start of a line:

```
\d+-\d+-\d+\s.*
```

Regular expression that is used to extract log information:

```
(\d+-\d+-\d+\s\d+:\d+:\d+,\d+)\s\[([^\]]*)\]\s(\S+)\s+(\S+)\s-\s(.*)
```

Time conversion format:

```
%Y-%m-%d %H:%M:%S
```

The following table lists the extraction results of the sample log.

Key	Value
time	2013-12-25 19:57:06,954
ip	203.0.113.2
level	WARN
class	impl.PermanentTairDaoImpl

Key	Value
message	Fail to Read Permanent Tair,key:e:470217319319741_1,result:com.example.tair.Result@172e3ebc[rc=code=-1, msg=connection error or timeout,value=,flag=0]

Use LogHub Log4j Appender to collect Log4j logs

For more information about how to collect Log4j logs by using LogHub Log4j Appender, see Log4j Appender.

Use Logtail to collect Log4j logs

The procedure when you use Logtail to collect Log4j logs is similar to that when you use Logtail to collect Python logs. Configure Logtail based on the actual network deployment and your business requirements. For more information, see Python logs.

23.1.3.2.3.2. Collect Python logs

Log Service allows you to use the Python logging module to collect Python logs.

The Python logging module provides a general logging system, which can be used by third-party modules or applications. The logging module defines multiple log severity levels and logging methods. The logging module consists of four components: loggers, handlers, filters, and formatters.

To collect Python logs, we recommend that you use logging handlers. For more information, see the following topics:

- Use logging handlers to automatically upload Python logs
- Use logging handlers to automatically upload and parse logs in the key-value format
- Use logging handlers to automatically parse logs in the JSON format

Log format

Formatters specify the output format of logs. The fields in the configurations of a formatter are in the %(key)s format.

```
import logging
import logging.handlers
LOG_FILE = 'tst.log'
handler = logging.handlers.RotatingFileHandler(LOG_FILE, maxBytes = 1024*1024, backupCount = 5) # Crea
te a handler object.
%(asctime)s - %(filename)s:%(lineno)s - %(levelno)s %(levelname)s %(pathname)s %(module)s %(funcName)s
%(created)f %(thread)d %(threadName)s %(process)d %(name)s - %(message)s // Define the output format o
f logs.
formatter = logging.Formatter(fmt)   # Create a formatter object.
handler.setFormatter(formatter)      # Add the formatter to the handler.
logger = logging.getLogger('tst')    # Retrieve a logger that is named tst.
logger.addHandler(handler)           # Add the handler to the logger.
logger.setLevel(logging.DEBUG)
logger.info('first info message')
logger.debug('first debug message')
```

The following table describes the fields in the formatter configurations.

Field	Description
%(name)s	The name of the logger that generates a log.
%(levelno)s	The severity level of a log in the numeric format.

Field	Description
%(levelname)s	The severity level of a log in the text format. Valid values: DEBUG, INFO, WARNING, ERROR, and CRITICAL.
%(pathname)s	The full path name of the source file where the logging call is initiated.
%(filename)s	The name of the source file.
%(module)s	The name of the module where the logging call is initiated.
%(funcName)s	The name of the function from which the logging call is initiated.
%(lineno)d	The line number in the source file where the logging call is initiated.
%(created)f	The time when a log is created. The value is a UNIX timestamp. It is the number of seconds that have elapsed since 00:00:00 UTC, Thursday, January 1, 1970.
%(relativeCreated)d	The difference between the time when a log is created and the time when the logging module is loaded. Unit: milliseconds.
%(asctime)s	The time when a log is created. Example: 2003-07-08 16:49:45,896. The digits after the comma (,) indicate the millisecond portion of the time.
%(msecs)d	The millisecond portion of the time when a log is created.
%(thread)d	The ID of the thread.
%(threadName)s	The name of the thread.
%(process)d	The ID of the process.
%(message)s	The log content.

The following example shows sample logs:

```
2015-03-04 23:21:59,682 - log_test.py:16 - tst - first info message
2015-03-04 23:21:59,682 - log_test.py:17 - tst - first debug message
```

Configure Logtail to collect Python logs

1. Log on to the Log Service console.
2. In the **Import Data** section, select **RegEx - Text Log**.
3. Select a destination project and Logstore, and then click **Next**.

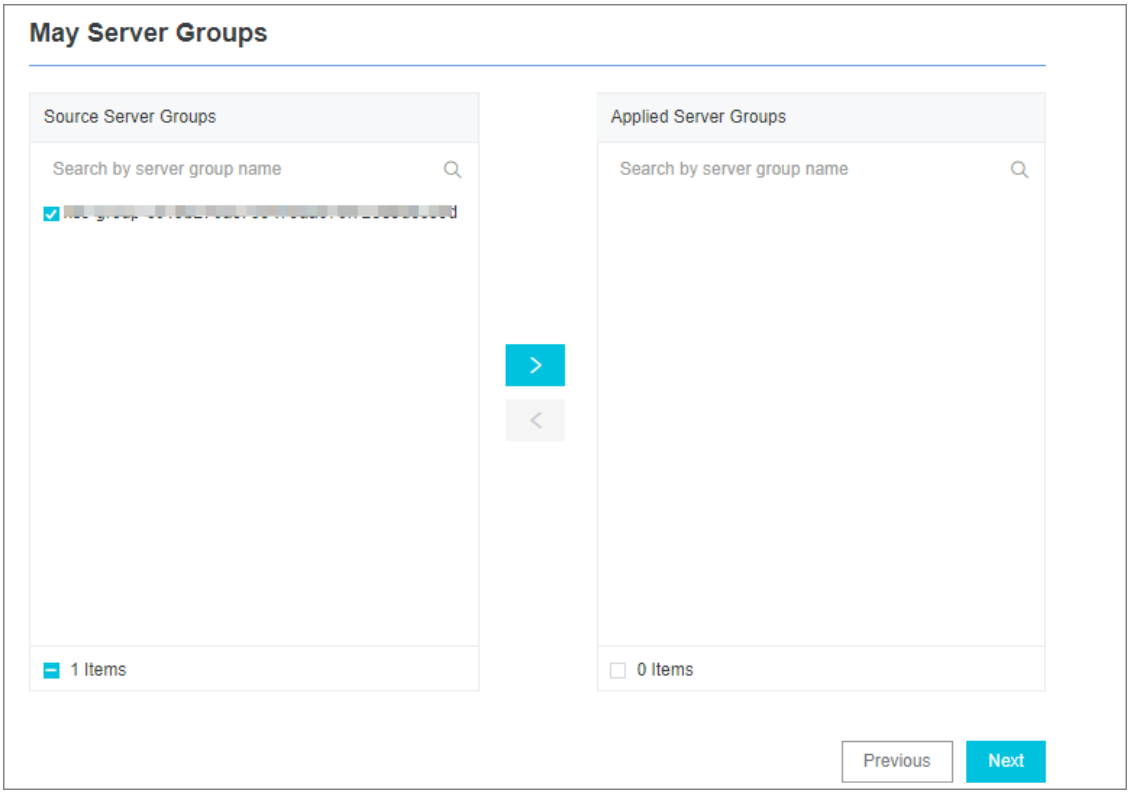
   You can also click **Create Now** to create a project and a Logstore.

   If you enter the configuration procedure by clicking the plus sign (+) next to **Data Import** under a Logstore on the Logstores tab, the system skips this step.
4. Create a machine group and click **Next**.

   Before you can create a machine group, you must install Logtail.

   Install Logtail as prompted. For more information, see Install Logtail in Linux or Install Logtail in Windows.

   After you install Logtail, click **Complete Installation** to create a machine group. If a machine group is created, click **Use Existing Machine Groups** to select the machine group.
5. Select a machine group, move the machine group from **Source Machine Groups** to **Applied Server Groups**, and then click **Next**.



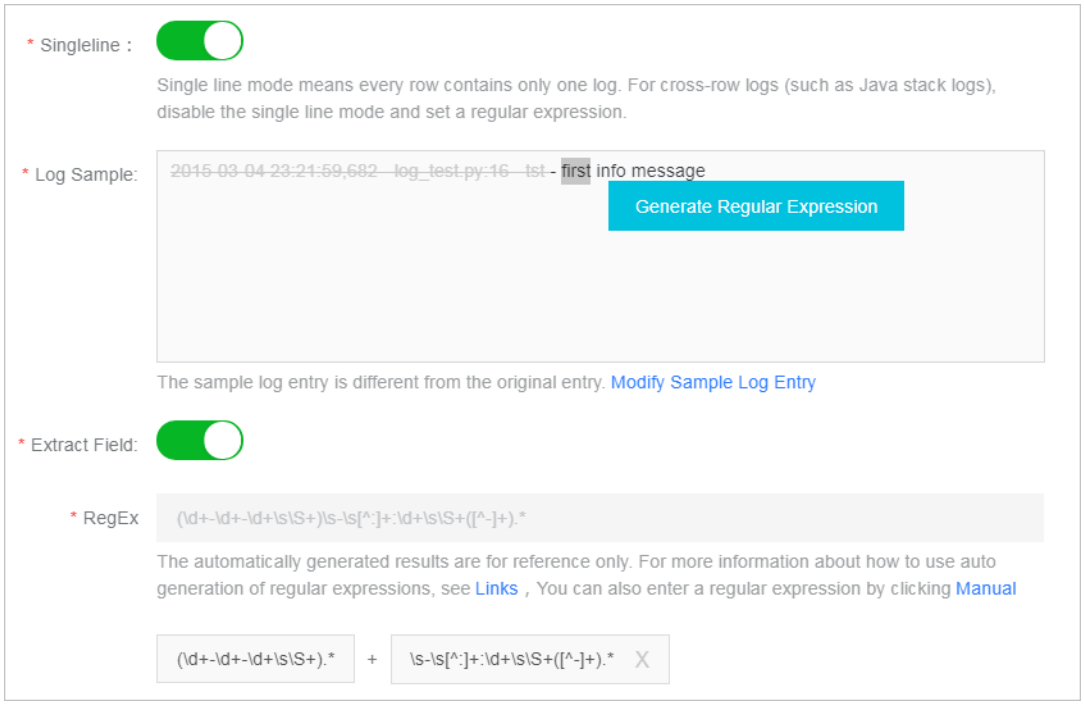
Notice If you want to apply a machine group immediately after it is created, the heartbeat status of the machine group may be FAIL. This issue occurs because no servers in the machine group are connected to Log Service. In this case, you can click **Automatic Retry**.

6. Create a Logtail configuration.
   i. Set the **Config Name** and **Log Path** parameters and set the Mode parameter to **Full Regex Mode**.
   ii. Turn on **Singleline**.
   iii. Enter a sample log in the **Log Sample** field.
   iv. Turn on **Extract Field**.

v. Set a regular expression in the **RegEx** field.

a. Select fields to generate a regular expression.

If the regular expression that is automatically generated does not match your sample log, you can select fields in the sample log to generate a regular expression. Log Service can automatically parse the highlighted fields of the sample log to generate a regular expression. In the **Log Sample** field, select the required fields, and click **Generate Regular Expression**. The regular expression of the selected field is displayed in the **RegEx** field. To obtain a full regular expression for the sample log, generate regular expressions for each log field.



b. Modify the regular expression.

Actual data formats may vary. In this case, click **Manual** under the RegEx field to adjust the regular expression that is automatically generated based on your business requirements. This ensures that the regular expression is suitable for all formats of the collected logs.

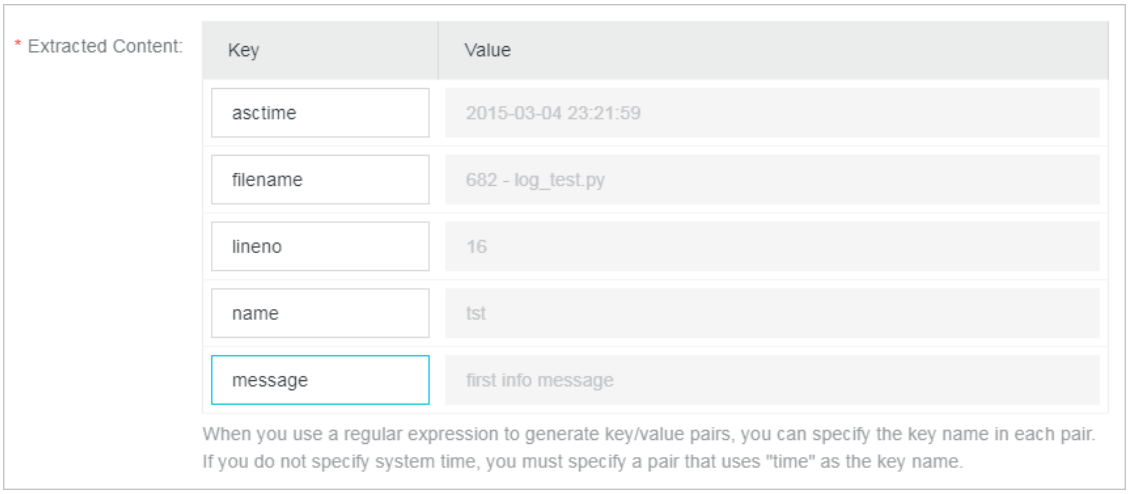
c. Verify the regular expression.

After you modify the regular expression, click **Validate** next to the RegEx field. If the regular expression is valid, the extraction results are displayed. If the regular expression is invalid, modify the regular expression again.

vi. Verify the result in the **Extracted Content** field.

View the extraction results of log fields and specify keys for the extracted fields.

Specify an informative name for each log field in the extraction results. For example, you can use time as the name for a time field. If you do not use the system time, you must specify the name of a time field in the Value field and time in the Key field.



7. (Optional)Specify **Advanced Options** and click **Next**.

Set the parameters in the Advanced Options section based on your business requirements. We recommend that you do not modify the settings. The following table describes the parameters in the Advanced Options section.

Parameter	Description
Enable Plug-in Processing	Specifies whether to enable plug-in processing. If you turn on this switch, you can use the plug-in of Logtail to process text logs. **Note** If you turn on **Enable Plug-in Processing**, specific parameters such as Upload Raw Log, Timezone, Drop Failed to Parse Logs, Filter Configuration, and Incomplete Entry Upload (Delimiter mode) become unavailable.
Upload Raw Log	If you turn on **Upload Raw Log**, each raw log is uploaded to Log Service as a value of the __raw__ field together with the parsed log.
Topic Generation Mode	The topic generation mode. ◦ **Null - Do not generate topic**: This mode is selected by default. In this mode, the topic field is set to an empty string. You do not need to enter a topic to query logs. ◦ **Machine Group Topic Attributes**: This mode is used to differentiate logs that are generated by different servers. ◦ **File Path RegEx**: In this mode, you must specify a regular expression in the **Custom RegEx** field. The part of a log path that matches the regular expression is used as the topic. This mode is used to differentiate logs that are generated by different users or instances.
Custom RegEx	If you set the Topic Generation Mode parameter to **File Path RegEx**, you must enter a custom regular expression.

Parameter	Description
Log File Encoding	The encoding format of log files. Valid values: ○ utf8: UTF-8 encoding format ○ gbk: GBK encoding format
Timezone	The time zone where logs are collected. Valid values: ○ System Timezone: This option is selected by default. It indicates that the time zone where logs are collected is the same as the time zone to which the server belongs. ○ Custom: If you select this value, you must select a time zone.
Timeout	The timeout period of log files. If a log file is not updated within the specified period, Logtail reckons the file to be timed out. Valid values: ○ Never: All log files are continuously monitored and never time out. ○ 30 Minute Timeout: If a log file is not updated within 30 minutes, Logtail considers the log file to be timed out and no longer monitors the file. If you select **30 Minute Timeout**, you must specify the **Maximum Timeout Directory Depth** parameter. Valid values: 1 to 3.
Filter Configuration	Only logs that **meet all filter conditions** are collected. Examples: ○ Collect logs that meet specified conditions: If you set **Key** to **level** and **Regex** to **WARNING|ERROR**, only WARNING-level and ERROR-level logs are collected. ○ Filter out logs that do not meet specified conditions. ■ If you set **Key** to **level** and **Regex** to **^(?!.*(INFO|DEBUG)).***, INFO-level or DEBUG-level logs are not collected. ■ If you set **Key** to **url** and **Regex** to **.*^(?!.*(healthcheck)).***, logs whose URLs contain healthcheck are not collected. For example, the log is not collected if the value of the Key field is url and the value of the Value field is /inner/healthcheck/jiankong.html.

8. Configure indexes in the Configure Query and Analysis step. Click **Next**.

   By default, Log Service enables Full Text Index to query and analyze logs. You can manually or automatically configure Field Search. For more information, see Enable the index feature and configure indexes for a Logstore.

   > **Note**
   > ○ To query and analyze logs, you must enable Full Text Index or Field Search. If you enable both Full Text Index and Field Search, the settings of Field Search prevail.
   > ○ If the data type of the index is long or double, the Case-Sensitive switch and the Delimiter field are unavailable.

After you complete the settings, you can start to collect Python logs.

23.1.3.2.3.3. Collect Node.js logs

Node.js logs are displayed in the Log Service console by default. This affects your data collection and troubleshooting efficiency. Log4js is a tool used to manage Node.js logs. You can use Log4js to send Node.js logs to files and customize the log format. Log4js allows you to collect and consolidate data in an efficient manner.

The following code shows how to configure Log4js to send logs to a file:

```
var log4js = require('log4js');
log4js.configure({
  appenders: [
    {
      type: 'file', // Output to a file
      filename: 'logs/access.log',
      maxLogSize: 1024,
      backups:3,
      category: 'normal'
    }
  ]
});
var logger = log4js.getLogger('normal');
logger.setLevel('INFO');
logger.info("this is a info msg");
logger.error("this is a err msg");
```

Log format

After you use Log4js to write logs to text files, the logs are displayed in the following format:

```
[2016-02-24 17:42:38.946] [INFO] normal - this is a info msg
[2016-02-24 17:42:38.951] [ERROR] normal - this is a err msg
```

Log4js classifies log severities into the following six levels in ascending order: TRACE, DEBUG, INFO, WARN, ERROR, and FATAL.

Use Logtail to collect Node.js logs

The procedure when you configure Logtail to collect Node.js logs is similar to that when you configure Logtail to collect Python logs. For more information, see Python logs. Set related parameters based on the actual network deployment and your business requirements.

The regular expression that is automatically generated is based on the sample log and may not apply to other logs. Therefore, you must modify the regular expression based on your business requirements before you use it. You can use the following sample Node.js logs to configure regular expressions for your logs.

Sample Node.js logs and regular expressions:

- Example 1
  - Sample log

    ```
    [2016-02-24 17:42:38.946] [INFO] normal - this is a info msg
    ```

  - Regular expression

    ```
    \[([^]]+)]\s\[([^\]]+)]\s(\w+)\s-(.*)
    ```

  - Extracted fields

    `time`, `level`, `loggerName`, and `message`

- Example 2
  - Sample log

    ```
    [2016-01-31 12:02:25.844] [INFO] access - 42.120.73.203 - - "GET /user/projects/ali_sls_log?ignor
    eError=true HTTP/1.1" 304 - "http://
    aliyun.com/" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_3) AppleWebKit/537.36 (KHTML, like Gec
    ko) Chrome/48.0.2564.97 Safari/537.36"
    ```

- Regular expression

```
\[([^]]+)]\s\[(\w+)]\s(\w+)\s-\s(\S+)\s-\s-\s"([^"]+)"\s(\d+)[^"]+("[^"]+)"\s"([^"]+).*
```

- Extracted fields

`time`, `level`, `loggerName`, `ip`, `request`, `status`, `referer`, and `user_agent`

23.1.3.2.3.4. Collect WordPress logs

This topic describes the format of WordPress logs and extraction results of a sample log.

Log format

Sample log:

```
172.64.0.2 - - [07/Jan/2016:21:06:39 +0800] "GET /wp-admin/js/password-strength-meter.min.js?ver=4.4 H
TTP/1.0" 200 776 "http://wordpress.c4a1a0aecdb1943169555231dcc4adfb7.cn-hangzhou.alicontainer.com/wp-a
dmin/install.php" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_5) AppleWebKit/537.36 (KHTML, like Gec
ko) Chrome/47.0.2526.106 Safari/537.36"
```

Configure Logtail to collect WordPress logs

If you use Logtail to collect WordPress logs, you must configure the following settings:

- Regular expression that matches each IP address that indicates the start of a line

```
\d+\.\d+\.\d+\.\d+\s-\s.*
```

- Regular expression that is used to extract log information

```
(\S+) - - \[([^\]]*)] "(\S+) ([^"]+)" (\S+) (\S+) "([^"]+)" "([^"]+)"
```

- Time conversion format

```
%d/%b/%Y:%H:%M:%S
```

- The following table lists the extraction results of the sample log.

Key	Value
ip	10.10.10.1
time	07/Jan/2016:21:06:39 +0800
method	GET
url	/wp-admin/js/password-strength-meter.min.js?ver=4.4 HTTP/1.0
status	200
length	776
ref	http://wordpress.c4a1a0aecdb1943169555231dcc4adfb7.cn-hangzhou.alicontainer.com/wp-admin/install.php
user-agent	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_5) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36

23.1.3.2.3.5. Collect Unity3D logs

This topic describes how to use the web tracking feature of Log Service to collect Unity3D logs.

Context

Unity3D is a cross-platform game engine that is developed by Unity Technologies. You can use the engine to create 3D video games, VR buildings, real-time 3D animation, and other interactive content.

In this example, Unity Debug.Log is used to describe how to collect Unity3D logs.

Procedure

1. Enable the web tracking feature.

   For more information, see WebTracking.

2. Create a Unity3D logging handler.

   In the Unity editor, create a C# file named *LogOutputHandler.cs*, add the following code to the file, and then modify the following variables:

   - project: the name of the Log Service project.
   - logstore: the name of the Logstore.
   - serviceAddr: the endpoint of the Log Service project. For more information, see **Obtain an endpoint** in *Log Service Developer Guide*.

```
using UnityEngine;
using System.Collections;
public class LogOutputHandler : MonoBehaviour
{
    //Register the HandleLog function on scene start to fire on debug.log events
    public void OnEnable()
    {
        Application.logMessageReceived += HandleLog;
    }
    //Remove callback when object goes out of scope
    public void OnDisable()
    {
        Application.logMessageReceived -= HandleLog;
    }
    string project = "your project name";
    string logstore = "your logstore name";
    string serviceAddr = "http address of your log service project";
    //Capture debug.log output, send logs to Loggly
    public void HandleLog(string logString, string stackTrace, LogType type)
    {
        string parameters = "";
        parameters += "Level=" + WWW.EscapeURL(type.ToString());
        parameters += "&";
        parameters += "Message=" + WWW.EscapeURL(logString);
        parameters += "&";
        parameters += "Stack_Trace=" + WWW.EscapeURL(stackTrace);
        parameters += "&";
        //Add any User, Game, or Device MetaData that would be useful to finding issues later
        parameters += "Device_Model=" + WWW.EscapeURL(SystemInfo.deviceModel);
        string url = "http://" + project + "." + serviceAddr + "/logstores/" + logstore + "/track?
APIVersion=0.6.0&" + parameters;
        StartCoroutine(SendData(url));
    }
    public IEnumerator SendData(string url)
    {
        WWW sendLog = new WWW(url);
        yield return sendLog;
    }
}
```

You can use the preceding code to asynchronously send logs to Log Service. You can also specify other fields in the code to collect the fields.

3. Generate Unity3D logs.

   Create a file named *LogglyTest.cs* and add the following code to the file:

```
using UnityEngine;
using System.Collections;
public class LogglyTest : MonoBehaviour {
    void Start () {
        Debug.Log ("Hello world");
    }
}
```

4. View logs in the Log Service console.

   After you run the Unity3D application, logs are generated and sent to Log Service. You can view the logs in the Log Service console.

23.1.4. Query and analysis

23.1.4.1. Log search overview

Log Service allows you to search for 1 billion to hundreds of billions of rows of log data within seconds. This topic describes the syntax and limits of the log search feature and provides examples.

Syntax

Each query statement consists of a search statement and an analytic statement. The search statement and the analytic statement are separated by a vertical bar (|). For more information about the query statement, see Search syntax.

> **Note**
> - A search statement can be executed alone. However, an analytic statement must be executed together with a search statement. The log analysis feature is based on search results or all data in a Logstore.
> - If you need to search for tens of billions of rows of data, you can repeatedly execute a search statement up to 10 times to obtain the complete result.

- Syntax

```
Search statement|Analytic statement
```

Statement	Description
Search statement	A search statement specifies one or more search conditions and returns the logs that meet the specified conditions. A search statement can be a keyword, a value, a value range, a space character, or an asterisk (*). If you specify a space character or an asterisk (*) as the search statement, no conditions are specified and all logs are returned. For more information, see Search syntax.
Analytic statement	An analytic statement is used to aggregate or analyze all log data or the log data that meets the specified search conditions in a Logstore. For more information, see Log analysis overview.

- Example

```
* | SELECT status, count(*) AS PV GROUP BY status
```

Limits

- Each project supports a maximum of 1,000 concurrent search statements at the same time.

  For example, 1,000 users can concurrently search for data in all Logstores of a project at the same time.
- You can specify a maximum of 30 keywords for each search statement.
- The maximum size of a field value is 10 KB. If the size of a field value exceeds 10 KB, the excess content is not queried.
- The returned logs are displayed on multiple pages. Each page displays a maximum of 100 search results.
- Log Service performs the DOM operation only on the first 10,000 characters of a log.
- If you perform a fuzzy search, Log Service searches for 100 words that meet the specified conditions. Logs that contain one or more of the 100 words and meet the search conditions are returned.

Search methods

Notice Before you search for logs, you must make sure that logs are collected and indexes are configured for the fields. Indexes are used in a storage structure to sort one or more columns of log data. For more information, see Enable the indexing feature and configure indexes for a Logstore.

- Use the Log Service console

  Log on to the Log Service console. On the Search & Analysis page of a Logstore, specify a time range and execute a search statement. For more information, see Query logs and Search syntax.

- Call API operations

  Call the GetLogs and GetHistograms operations to search for log data. For more information, see the GetLogs and GetHistograms topics of API Reference in *Developer Guide*.

23.1.4.2. Log analysis overview

Log Service provides the log analysis feature. This feature allows you to search for log data and use SQL functions to analyze the data. This topic describes the syntax and limits of the analytic statements. This topic also provides the SQL functions that you can call when you use the log analysis feature.

Note If you want to use the log analysis feature, you must turn on **Enable Analytics** when you configure indexes for log fields. For more information, see Enable the indexing feature and configure indexes for a Logstore. If you turn on **Enable Analytics**, you can analyze log data within seconds without additional costs.

Syntax

Each query statement consists of a search statement and an analytic statement. The search statement and the analytic statement are separated by a vertical bar (|). You can execute a search statement alone. However, you must execute an analytic statement together with a search statement. You can use the log analysis feature to analyze data that meets specified search conditions in a Logstore. You can also use the feature to analyze all data in a Logstore.

Note

- You do not need to specify a FROM or WHERE clause in an analytic statement. By default, all data of the current Logstore is analyzed.
- You do not need to add a semicolon (;) at the end of an analytic statement to end the statement.
- Analytic statements are case-insensitive.

- Syntax

```
Search statement|Analytic statement
```

Statement	Description
Search statement	A search statement specifies one or more search conditions. A search statement can be a keyword, a value, a value range, a space character, or an asterisk (*). If you specify a space character or an asterisk (*) as the search statement, no conditions are specified and all logs are returned. For more information, see Search syntax.

Statement	Description
Analytic statement	An analytic statement is used to aggregate or analyze all log data or the log data that meets the specified search conditions in a Logstore.

- Example

```
* | SELECT status, count(*) AS PV GROUP BY status
```

Limits

- Each project supports a maximum of 15 concurrent analytic statements at the same time.

  For example, 15 users can concurrently execute analytic statements in all Logstores of a project at the same time.
- You can analyze only the data that is written to Log Service after the log analysis feature is enabled.
- By default, an analytic statement returns a maximum of 100 rows of data.

  If you want to view more data, use a LIMIT clause. For more information, see LIMIT syntax.
- The maximum size of a field value is 16 KB. If the size of a field value exceeds 16 KB, the excess content is not analyzed.
- The maximum timeout period for an analytic statement is 55 seconds.
- Each shard supports only 1 GB of data for an analytic statement.
- The value of a double-type field can contain a maximum of 52 digits after the decimal point.

  If the number of digits after the decimal point is greater than 52, the accuracy of the field value is compromised.

SQL functions and syntax

This section lists the SQL functions and syntax that Log Service supports.

- The following aggregate functions are available for SELECT statements:
  - General aggregate functions
  - Security check functions
  - Map functions
  - Approximate functions
  - Mathematical statistics functions
  - Mathematical calculation functions
  - String functions
  - Date and time functions
  - URL functions
  - Regular expression functions
  - JSON functions
  - Type conversion functions
  - IP functions
  - Array functions
  - Binary string functions
  - Bitwise operations
  - Interval-valued comparison and periodicity-valued comparison functions
  - Comparison functions and operators
  - Lambda functions
  - Logical functions

  - Geospatial functions
  - Geography functions
  - Machine learning syntax and functions
- GROUP BY syntax
- Window functions
- HAVING syntax
- ORDER BY syntax
- LIMIT syntax
- Syntax for CASE statements and if() functions
- UNNEST function
- Field aliases
- Nested subqueries

23.1.4.3. Configure indexes

Indexes are used in a storage structure to sort one or more columns of log data. You can query and analyze log data only after you configure indexes. Query and analysis results vary based on index configurations. Therefore, you must configure indexes based on your business requirements.

Prerequisites

Logs are collected. For more information, see Data collection.

Index types

The following table describes the index types that are supported by Log Service.

Index type	Description
Full-text index	Log Service splits an entire log into multiple words based on specified delimiters to create indexes. In a search statement, the field names (keys) and field values (values) are both plain text. For example, the search statement `error` returns the logs that contain the keyword `error`.
Field index	After you configure field indexes, you can specify field names and field values in the key:value format to search for logs. For example, the search statement `level:error` returns the logs in which the value of the `level` field contains `error`. If you want to use the analysis feature, you must configure field indexes and turn on Enable Analytics for the required fields. The analysis feature does not generate index traffic or occupy storage space.

Notice

- The indexing feature is applicable only to the log data that is written to the current Logstore after you configure indexes.
- If you configure both full-text indexes and field indexes, the configurations of the field indexes take precedence.

Configure full-text indexes

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to query and analyze logs.

3. Choose **Log Storage > Logstores**. On the Logstores tab, click the Logstore in which logs are stored.
4. On the Search & Analysis page of the Logstore, choose **Index Attributes > Attributes**.

   If the indexing feature is not enabled, click **Enable**.
5. Configure indexes in the Search & Analysis panel.

Parameter	Description
LogReduce	If you turn on **LogReduce**, Log Service automatically clusters text logs that have the same pattern during log collection. This way, you can obtain the overall information of the logs. For more information, see LogReduce.
Full Text Index	If you turn on **Full Text Index**, the full-text indexing feature is enabled.
Case Sensitive	Specifies whether searches are case-sensitive. ◦ If you turn on **Case Sensitive**, searches are case-sensitive. For example, if a log contains `internalError`, you can search for the log only by using the keyword `internalError`. ◦ If you turn off **Case Sensitive**, searches are not case-sensitive. For example, if a log contains `internalError`, you can search for the log by using the keyword `INTERNALERROR` or `internalerror`.
Include Chinese	Specifies whether to distinguish between Chinese content and English content in searches. ◦ If you turn on **Include Chinese** and a log contains Chinese characters, the Chinese content is split based on the Chinese grammar. The English content is split based on specified delimiters. **Notice** When the Chinese content is split, the data write speed is reduced. Proceed with caution. ◦ If you turn off **Include Chinese**, all content is split based on specified delimiters.
Delimiter	The delimiters that are used to split the content of a log into multiple words. The following delimiters are supported: `, '";=()[]{}?@&<>/:\n\t\r`. `\n` indicates a line feed, `\t` indicates a tab character, and `\r` indicates a carriage return. For example, the content of a log is `/url/pic/abc.gif`. ◦ If you do not specify a delimiter, the log is processed as a single word `/url/pic/abc.gif`. You can search for the log only by using the keyword `/url/pic/abc.gif`. You can also perform a fuzzy search by using the keyword `/url/pic/*`. ◦ If you set the delimiter to a forward slash (/), the content of the log is split into the following three words: `url`, `pic`, and `abc.gif`. You can search for the log by using the keyword `url`, `abc.gif`, or `/url/pic/abc.gif`. You can also perform a fuzzy search by using the keyword `pi*`. ◦ If you set the delimiter to a forward slash (/) and a period (.), the content of the log is split into the following four words: `url`, `pic`, `abc`, and `gif`.

6. Click **OK**.

   The index configurations take effect within 1 minute.

Configure field indexes

1. Log on to the Log Service console.

2. In the Projects section, click the project in which you want to query and analyze logs.
3. Choose **Log Storage > Logstores**. On the Logstores tab, click the Logstore in which logs are stored.
4. On the Search & Analysis page of the Logstore, choose **Index Attributes > Attributes**.

   If the indexing feature is not enabled, click **Enable**.
5. Configure indexes in the **Search & Analysis** panel.

Parameter	Description
Key Name	The name of the log field. Example: client_ip. **Note** ◦ When you configure an index for a tag field, you must specify the value of the **Key Name** parameter in the __tag__:KEY format. For example, you can set this parameter to __tag__:__receive_time__. Different tag fields are supported. For example, a tag field can indicate a public IP address or a UNIX timestamp. ◦ When you configure an index for a tag field, you must set the **Type** parameter for each tag field to **text**. Numeric data types are not supported.
Type	The data type of the log field value. Valid values: text, long, double, and json. For more information, see Data types. **Note** If a field is of the long or double type, you cannot set the **Case Sensitive**, **Include Chinese**, or **Delimiter** parameter.
Alias	The alias of the field. Example: ip. An alias is used only in analytic statements. You must use the original field name in search statements. For more information, see Field aliases.
Case Sensitive	Specifies whether searches are case-sensitive. ◦ If you turn on **Case Sensitive**, searches are case-sensitive. For example, if a log contains `internalError`, you can search for the log only by using the keyword `internalError`. ◦ If you turn off **Case Sensitive**, searches are not case-sensitive. For example, if a log contains `internalError`, you can search for the log by using the keyword `INTERNALERROR` or `internalerror`.

Parameter	Description
Delimiter	The delimiters that are used to split the content of a log into multiple words. The following delimiters are supported: `, '";=()[]{}?@&<>/:\n\t\r`. `\n` indicates a line feed, `\t` indicates a tab character, and `\r` indicates a carriage return. For example, the content of a log is `/url/pic/abc.gif`. ○ If you do not specify a delimiter, the log is processed as a single word `/url/pic/abc.gif`. You can search for the log only by using the keyword `/url/pic/abc.gif`. You can also perform a fuzzy search by using the keyword `/url/pic/*`. ○ If you set the delimiter to a forward slash (/), the content of the log is split into the following three words: `url`, `pic`, and `abc.gif`. You can search for the log by using the keyword `url`, `abc.gif`, or `/url/pic/abc.gif`. You can also perform a fuzzy search by using the keyword `pi*`. ○ If you set the delimiter to a forward slash (/) and a period (.), the content of the log is split into the following four words: `url`, `pic`, `abc`, and `gif`.
Include Chinese	Specifies whether to distinguish between Chinese content and English content in searches. ○ If you turn on **Include Chinese** and a log contains Chinese characters, the Chinese content is split based on the Chinese grammar. The English content is split based on specified delimiters. **Notice** When the Chinese content is split, the data write speed is reduced. Proceed with caution. ○ If you turn off **Include Chinese**, all content is split based on specified delimiters.
Enable Analytics	To use the analysis feature, you must turn on **Enable Analytics**.

6. Click **OK**.

   The index configurations take effect within 1 minute.

23.1.4.4. Query and analyze logs

This topic describes how to query and analyze logs in a Logstore. After you enable the indexing feature and configure indexes for a Logstore, you can query and analyze the logs that are stored in the Logstore in real time.

Prerequisites

- Logs are collected and stored in a Logstore. For more information, see Data collection.
- The indexing feature is enabled and indexes are configured for the Logstore. For more information, see Enable the indexing feature and configure indexes for a Logstore.

Query and analyze logs

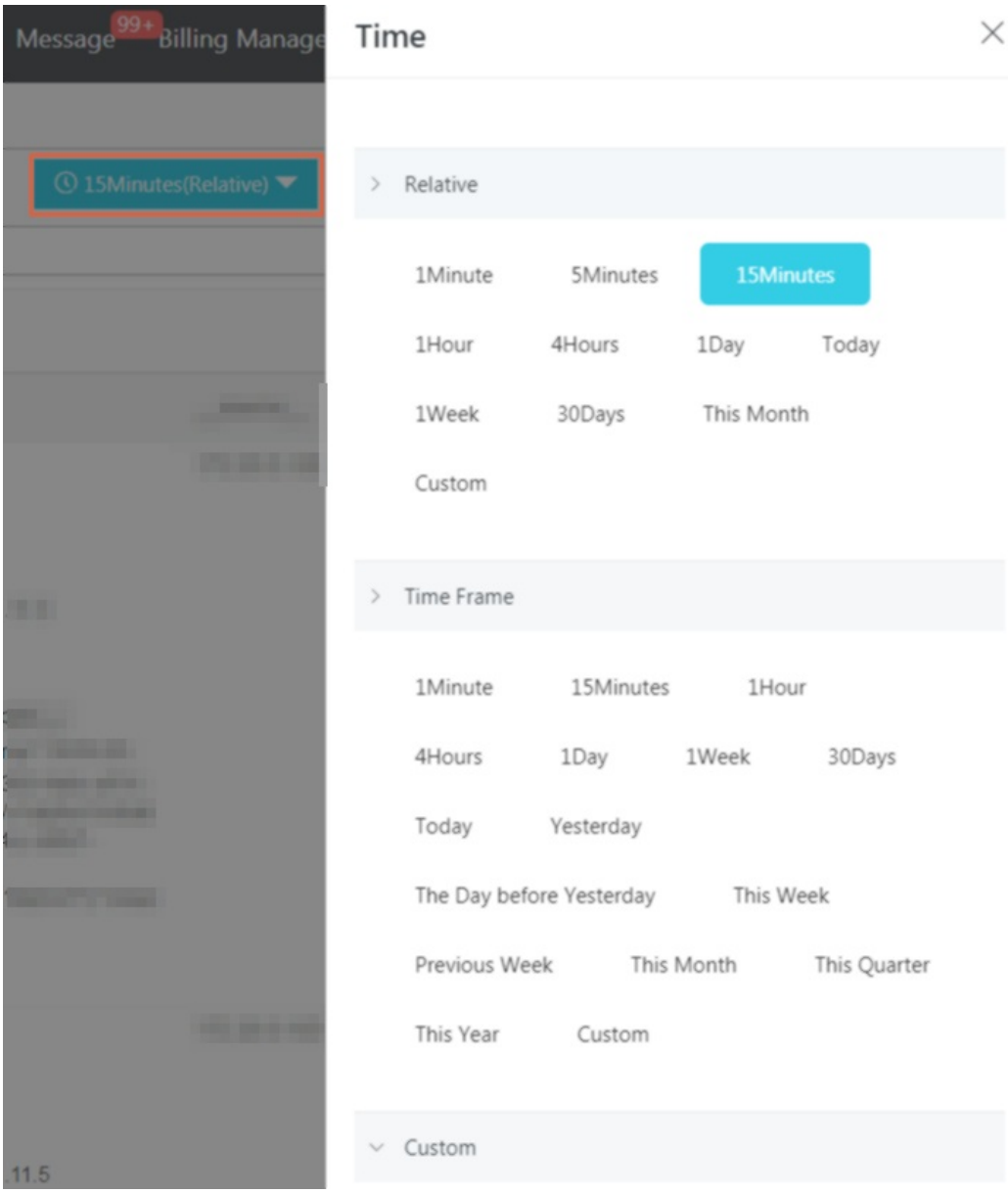
1. Log on to the Log Service console.
2. In the Projects section, click the project to which the Logstore belongs.
3. Click the icon next to the name of the Logstore and select **Search & Analysis**.
4. Enter a query statement in the search box.

   A query statement consists of a search statement and an analytic statement in the format of `Search statement|Analytic statement`. For more information, see Log search overview and Log analysis overview.

5. Click **15 Minutes(Relative)** to specify a time range.

   You can select a relative time or a time frame. You can also specify a custom time range.

   > **Note** The query results may contain logs that are generated 1 minute earlier or later than the specified time range.



6. Click **Search & Analytics**.

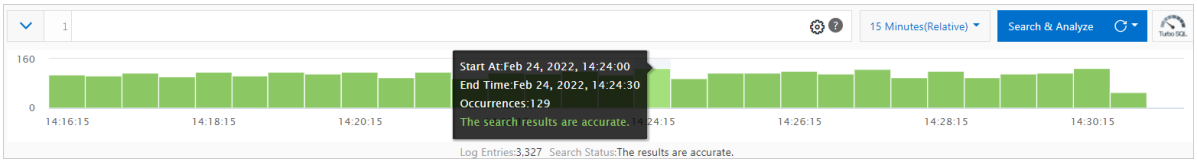
Manage query and analysis results

You can view the query and analysis results in a log distribution histogram, on the Raw Logs tab, or in a chart that is displayed on the Graph tab. You can also configure alerts and saved searches.

> **Note** By default, only 100 rows of data are returned after you execute a query statement. You can use a LIMIT clause to change the number of returned rows. For more information, see LIMIT syntax.

- Log distribution histogram

The log distribution histogram shows the distribution of query and analysis results in different time ranges.



- Move the pointer over a green rectangle to view the time range that is represented by the rectangle. You can also view the number of logs that are obtained within the time range.
- Click the green rectangle to view a more fine-grained log distribution. You can also view the query and analysis results on the **Raw Logs** tab.

- Raw Logs tab

  On the **Raw Logs** tab, you can view the logs that match your search conditions.

  - Quick analysis: Use this feature to analyze the distribution of a specific field within a specific period of time. For more information, see Quick analysis.
  - Contextual query: Click the icon of a log and select **Context View** to view the context of the log. For more information, see Contextual query.

    > **Note** The contextual query feature supports only the log data that is collected by Logtail.

  - LiveTail: Click the icon of a log and select **LiveTail** to monitor the log in real time and extract key information. For more information, see LiveTail.

    > **Note** LiveTail can monitor and extract only the log data that is collected by Logtail.

  - Log download: Click the download icon in the upper-right corner of the tab, select a time range, and then click **OK**. For more information, see Export logs.
  - Column settings: Click **Column Settings** in the upper-right corner of the tab, select fields from the section on the left, and then click **Add** to add the fields to the section on the right. The columns of the fields that you added appear on the tab. The field names are the column names, and the column values are the field values.

    > **Note** To view the log content on the tab, you must select **Content**.

- Content column settings: By default, if the content of a field exceeds 3,000 characters, excess characters are hidden and the message "The character string is too long and has been truncated" is displayed before the Key field. Click **Display Content Column**. In the dialog box that appears, set the **Key-Value Arrangement** and **Truncate Character String** parameters.

  > **Note** If you set the content limit to 10,000 characters, no delimiter is used for the excess characters.

Parameter		Description
Key-Value Pair Arrangement		You can set this parameter to **New Line** or **Full Line**.
Truncate Character String	**Key**	If a field value contains more than 3,000 characters, the field value is truncated by default. If no field value exceeds 3,000 characters, you can leave this parameter empty. The value of this parameter is the key of the truncated value.
	Status	Specifies whether to enable the value truncation feature. By default, this feature is enabled. ■ **Enable**: If the value in a key-value pair exceeds the specified **Truncate Step**, excess characters are truncated. You can click the Show button at the end of the value to show the truncated characters. The increment per click is the specified truncate step. ■ **Disable**: If the value in the key-value pair exceeds the specified **Truncate Step**, excess characters are not truncated.
	Truncate Step	Specifies the maximum number of characters that can be displayed for a field value. This parameter also specifies the number of incremental characters that are displayed each time you click the Show button. Valid values: 500 to 10000. Default value: 3000.

- Charts

  If you enable the analysis feature when you configure indexes for fields and use query statements to query logs, you can view the analysis results on the **Graph** tab.

  - Log Service provides multiple chart types, such as tables, line charts, and bar charts. You can select a chart type to display analysis results. For more information, see Chart overview.
  - Log Service allows you to create dashboards for real-time data analysis. Click **Add to New Dashboard** to save query statements as charts to a dashboard. For more information, see Dashboard overview.
  - Drill-down analysis allows you to view more detailed analysis results. You can set the drill-down parameters and add a chart to the dashboard. Then, you can click the values in the chart to view the analysis results in multiple dimensions. For more information, see Drill-down analysis.

- LogReduce

  On the **LogReduce** tab, click **Enable LogReduce** to cluster similar logs. For more information, see LogReduce.

- Alerts

  On the Search & Analysis page, click **Save as Alarm** to create an alert rule for query results. For more information, see Alert overview.

- Saved searches

  On the Search & Analysis page, click **Save Search** to save a query statement as a saved search. For more information, see Saved search.

23.1.4.5. Download logs

This topic describes how to download logs from Log Service to an on-premises host.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to download logs.
3. Click the icon next to the name of the Logstore whose logs you want to download and select **Search & Analysis**.
4. In the upper-right corner of the **Raw Logs** tab, click the icon.
5. In the **Log Download** dialog box, select a method that you can use to download logs and click **OK**.
   - **Download Log in Current Page**: Download the logs that are displayed on the current page to a file in the comma-separated values (CSV) format.
   - **Download All Logs Using Command Line Tool**: Download all logs as prompted.

23.1.4.6. Index data type

23.1.4.6.1. Overview

When you configure indexes, you can set the data type of a field to text, long, double, or JSON. This topic describes the index data types that are supported by Log Service.

Data types

The following table describes the supported data types.

Query type	Index data type	Description	Example
Basic query	text	The text type. You can use keywords and fuzzy matches to query logs.	`uri:"login*" and method:"post"`
	long	The numeric type. You can specify numeric ranges to query indexes of this type.	`status in [200, 500]`
	double	The floating-point type.	`price>28.95`
Combined query	JSON	Indicates that the index is a JSON field that supports nested queries. By default, the data type of the field is text. You can configure indexes of the text, long, and double types for the b elements at layer a in the a.b path format.	`level0.key>29.95 and level0.key2:"action"`
	text	Creates indexes for all fields in a log except the time field. The data type of the indexes is text.	`error and "login fail"`

23.1.4.6.2. Text type

This topic describes how to query text data.

Usage notes

Similar to search engines, Log Service queries text data based on terms. Therefore, you must set the Delimiter and Case Sensitive parameters when you configure indexes.

- Case Sensitive switch

  You can specify whether searches are case-sensitive. For example, you want to query a log entry that contains `internalError`.

  - If you turn off Case Sensitive, searches are case-insensitive, and you can find the log entry by using the `INTERNALERROR` or `internalerror` keyword.
  - If you turn on Case Sensitive, searches are case-sensitive, and you can find the log entry only by using the `internalError` keyword.

- Delimiter parameter

  You can use delimiters to split the content of a log entry into multiple words. For example, you want to query a log entry that contains the following content:

  ```
  /url/pic/abc.gif
  ```

  - If you do not specify a delimiter, the entire string is processed as a single word in the `/url/pic/abc.gif` format. In this case, you can find the log entry by using the entire string as a keyword for exact match or by using the `/url/pic/*` keyword for fuzzy match.
  - If you set the delimiter to a forward slash (/), the content is divided into the following three words: `url`, `pic`, and `abc.gif`. You can find the log entry by using one of the three words. You can also use part of each word to search for the log entry in fuzzy match mode.

    For example, you can find the log entry by using the `url`, `abc.gif`, or `pi*` keyword. You can also find the log entry by using the `/url/pic/abc.gif` keyword. The /url/pic/abc.gif keyword is split into the following search conditions: `url and pic and abc.gif`.
  - If you set the delimiter to a forward slash (/) and a period (.), the content is split into the following four words: `url`, `pic`, `abc`, and `gif`.

  > **Note** You can specify appropriate delimiters to extend query ranges.

- Full Text Index switch

  By default, after you turn on Full Text Index, the data type of all fields, except the time field, is set to text. You do not need to specify keys. For example, you want to query a log entry that consists of the following four fields:

  ```
  time:2018-01-02 12:00:00
  level:"error"
  status:200
  message:"some thing is error in this field"
  ```

  ```
  [20180102 12:00:00],200,error,some thing is error in this field
  ```

Note

- Prefixes are not required for full-text indexes. If you use error as a keyword, the level and message field values that contain error match the keyword.
- You must specify delimiters for full-text indexes. For example, if you specify a comma (,) as a delimiter, the `status:200` string is processed as a single word. If you specify a colon (:) as a delimiter, the string is split into the following two words: `status` and `200`.
- Numbers are processed as text data. For example, you can find the log entry by using the keyword 200. The time field is not processed as text data.
- If the query statement is a key, for example, `status`, the log entry is matched.

23.1.4.6.3. Numeric type

When you configure indexes, you can set the data type of a field to a numeric type. Then, you can query the value of the field by value range.

Usage notes

You can query the value of a field by using a numeric range only after you set the data type of the field to long or double.

- If the value of a log field is an integer, we recommend that you set the data type of the field to long when you configure indexes.
- If the value of a log field is a floating-point number, we recommend that you set the data type of the field to double when you configure indexes.

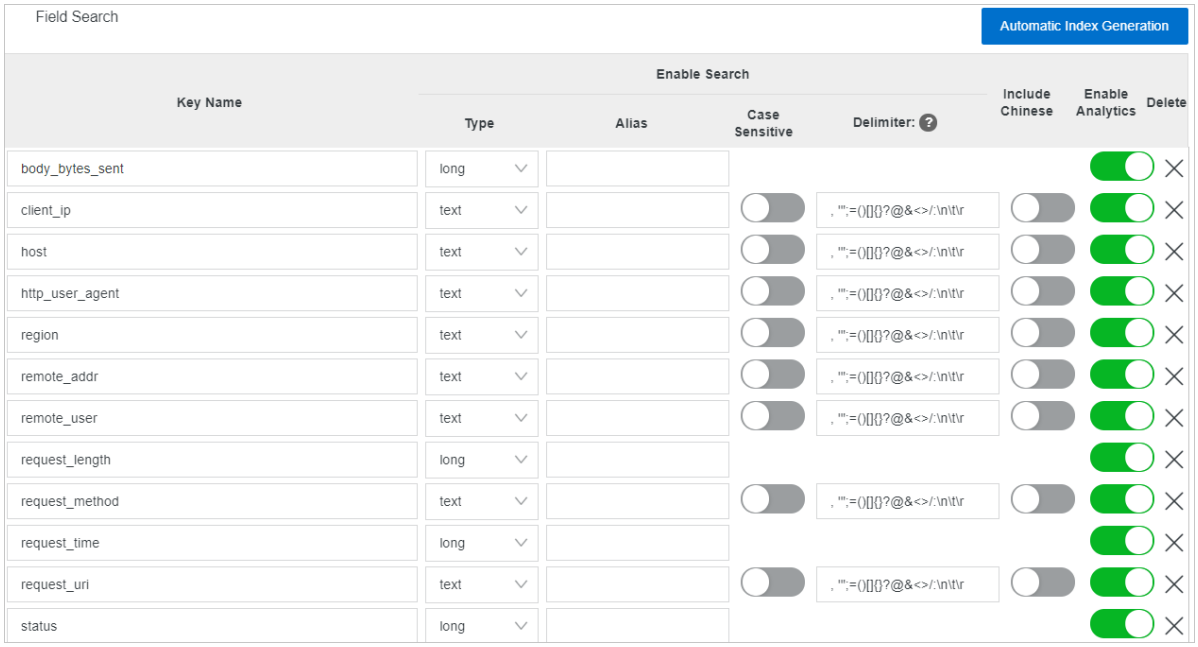
Notice

- If you set the data type of a field to long but the value of the field is a floating-point number, you cannot query the value of the field.
- If you set the data type of a field to long or double but the value of the field is a string, you cannot query the value of the field.
- If you set the data type of a field to long or double, you cannot use asterisks (*) or question marks (?) to query the value of the field in fuzzy match mode.
- If the value of a field is an invalid numeric value, you can query data by using the **not key > -1000000** search statement. The not key > -1000000 search statement returns the log entries in which a field value is an invalid numeric value. -100000 can be replaced by a valid value that is less than or equal to the smallest valid value of the field in your log entries.

Sample search statements

- Sample log entry

- Index configurations



- Query statements
  - To query the log entries in which the request duration is greater than 60 seconds, execute the following search statement:

    ```
    request_time > 60
    ```

  - To query the log entries in which the request duration is greater than or equal to 60 seconds and less than 200 seconds, execute one of the following search statements:

    ```
    request_time in [60 200)
    ```

    ```
    request_time >= 60 and request_time < 200
    ```

  - To query the log entries in which the response status code is 200, execute the following search statement:

    ```
    status = 200
    ```

23.1.4.6.4. JSON type

If the value of a field is in the JSON format, you can set the data type of the field to JSON when you configure indexes. This topic describes how to set the data type of a field to JSON and provides some examples.

Usage notes

- You can set the data type of a field in JSON objects to long, double, or text based on the field value, and turn on Enable Analytics to enable the analysis feature. After you turn on Enable Analytics, Log Service allows you to query and analyze fields in JSON objects.
- For partially valid JSON-formatted data, only the valid parts can be parsed in Log Service.

  The following example shows an incomplete JSON log entry. Log Service can parse the conctent.remote_addr, content.request.request_length, and content.request.request_method fields.

```
content: {
     remote_addr:"192.0.2.0"
     request: {
             request_length:"73"
             request_method:"GE
```

> **Notice**
>
> - Log Service allows you to configure indexes for leaf nodes in JSON objects. However, you cannot configure indexes for child nodes that contain leaf nodes.
> - You cannot configure indexes for fields whose values are JSON arrays or configure indexes for the fields in a JSON array.
> - If the value of a field is of the Boolean type, you can set the data type of the field to text when you configure indexes.
> - The format of a query statement in Log Service is `Search statement | Analytic statement`. In an analytic statement, you must enclose a field name by using double quotation marks ("") and enclose a string by using single quotation marks ('').

Examples

The following table lists the keys included in the sample log entry. The data type of the `message` field is JSON.

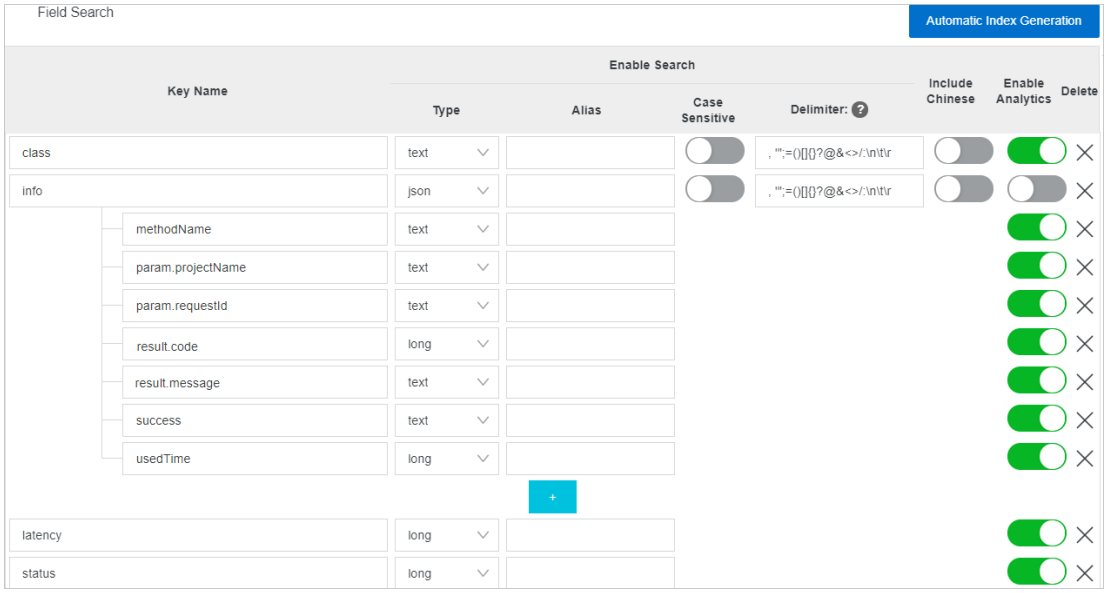
Serial number	Key	Type
0	time	N/A
1	class	text
2	status	long
3	latency	double
4	message	json

Sample log entry:

```
0. time:2018-01-01 12:00:00
1. class:central-log
2. status:200
3. latency:68.75
4. message:
  {
      "methodName": "getProjectInfo",
      "success": true,
      "remoteAddress": "203.0.113.10:11111",
      "usedTime": 48,
      "param": {
              "projectName": "ali-log-test-project",
              "requestId": "d3f0c96a-51b0-4166-a850-f4175dde7323"
      },
      "result": {
          "message": "successful",
          "code": "200",
          "data": {
              "clusterRegion": "ap-southeast-1",
              "ProjectName": "ali-log-test-project",
              "CreateTime": "2017-06-08 20:22:41"
          },
          "success": true
      }
  }
```

The following figure shows an example on how to configure indexes.

Index configurations



The following settings are configured in the preceding figure:

- ① specifies that Log Service can query data of the string and Boolean types in JSON fields.
- ② specifies that Log Service can query data of the long type.
- ③ enables SQL analysis for specified fields.
- Query log data of the string and Boolean types.

> **Note**
> - You do not need to configure JSON fields.
> - JSON maps and arrays are automatically expanded and can contain multiple layers. You must separate multiple layers with periods (.).

Query statement:

```
message.traceInfo.requestId : 92.137_1518139699935_5599
message.param.projectName : ali-log-test-project
message.success : true
message.result.data.ProjectStatus : Normal
```

- Query log data of the double and long types.

  > **Note** You must configure each JSON field. A JSON field cannot be contained in an array.

  Query statement:

  ```
  message.usedTime > 40
  ```

- Use SQL statements to analyze fields.

  > **Note**
  > - You must configure each JSON field. A JSON field cannot be contained in an array.
  > - You must enclose a field name by using double quotation marks ("") or specify an alias for the field.

  Query statement:

  ```
  * | select avg("message.usedTime") as avg_time ,"message.methodName"  group by "message.methodName"
  ```

23.1.4.7. Query syntax and functions

23.1.4.7.1. Search syntax

This topic describes the search syntax of Log Service.

Search types

A search statement specifies one or more search conditions and returns the log entries that meet the specified conditions. Searches are classified by indexing method into full-text searches and field-specific searches, or classified by precision into exact searches and fuzzy searches.

> **Note**
>
> - If you configure both full-text indexes and field indexes, the configurations of the field indexes take precedence.
> - You must set the data type of a field to double or long before you specify a value range to query log entries of the field. If you do not set the data type of a field to double or long, or the syntax of the value range is invalid, Log Service performs a full-text search and the search result may be different from the expected result. For example, if the `owner_id>100` search statement is executed and the data type of the owner_id field is not double or long, log entries that contain owner_id, > (non-delimiter), and 100 are returned.
> - If you change the data type of a field from text to double or long, you can use only the equal sign (=) to query the log entries that are collected before the change.

- Full-text searches and field-specific searches

Search type	Description	Example
Full-text search	After you configure full-text indexes, Log Service splits a log entry into multiple words by using the delimiters that you specify. You can specify keywords and rules in a search statement to query log entries. The keywords can be field names or field values.	`PUT and cn-shanghai`: returns the log entries that contain the keywords PUT and cn-shanghai.
Field-specific search	After you configure field indexes, you can query log entries. To query log entries, specify field names and field values in the format of key:value. You can perform basic searches or combined searches based on the data types of the fields in the field indexes. For more information, see Data types.	`request_time>60 and request_method:Ge*`: returns the log entries in which the value of the request_time field is greater than 60 and the value of the request_method field starts with Ge.

- Exact searches and fuzzy searches

Search type	Description	Example
Exact search	Complete words are used for queries.	○ `host:www.yl.mock.com`: returns the log entries in which the value of the host field is www.yl.mock.com. ○ `PUT`: returns the log entries that contain the keyword PUT.

Search type	Description	Example
Fuzzy search	You can add an asterisk (*) or a question mark (?) as a wildcard in the middle or at the end of a keyword to perform a fuzzy search. Each keyword must be 1 to 64 characters in length. If a keyword contains a wildcard, Log Service searches all log entries and obtains up to 100 words that match the keyword. Then, Log Service returns the log entries that contain one or more of these words. The more accurate a keyword is, the more accurate the search results are. **Note** ◦ A keyword cannot start with an asterisk (*) or a question mark (?). ◦ The long and double data types do not support asterisks (*) or question marks (?) in fuzzy searches. You can specify a numeric range, for example, status in [200 299], to perform a fuzzy search. A fuzzy search is a sample-based search that uses the following mechanism: ◦ If you enable the field indexing feature and specify a search field, Log Service randomly obtains samples from the index data of the specified field and returns part of the search results. ◦ If you enable the full-text indexing feature and do not specify a search field, Log Service randomly takes samples from the full-text index data and returns part of the search results.	◦ `addr*` : searches for 100 words that start with addr from log entries, and returns the log entries that contain one or more of these words. ◦ `host:www.yl*` : searches for 100 words that start with www.yl from the value of the host field. Then, Log Service returns the log entries in which the value of the host field contains one or more of these words.

Operators

The following table describes the operators that are supported by search statements.

Note

- The in operator is case-sensitive. Other operators are not case-sensitive.
- Log Service uses the following operators: **sort**, **asc**, **desc**, **group by**, **avg**, **sum**, **min**, **max**, and **limit**. If you use these operators as keywords, enclose the operators in double quotation marks ("").
- The following list shows the precedence of the operators in descending order:
  i. Colons (:)
  ii. Double quotation marks ("")
  iii. Parentheses ()
  iv. and
  v. not
  vi. or

Operator	Description

Operator	Description
and	The and operator. Example: `request_method:GET and status:200`. If no syntax keyword exists among multiple keywords, the keywords are joined by using the and operator by default. For example, `GET 200 cn-shanghai` is equivalent to `GET and 200 and cn-shanghai`.
or	The or operator. Example: `request_method:GET or status:200`.
not	The not operator. Example: `request_method:GET not status:200` or `not status:200`.
( )	This operator is used to increase the priority of the search conditions that are enclosed in parentheses (). Example: `(request_method:GET or request_method:POST) and status:200`.
:	This operator is used for field-specific searches (key:value). Example: `request_method:GET`. If a field name or field value contains reserved characters such as spaces and colons (:), enclose the field name or field value in double quotation marks (""). Example: `"file info":apsara`.
""	To convert the keyword to an ordinary character, you can use double quotation marks ("") to enclose a syntax keyword. For example, `"and"` returns log entries that contain and. In this case, and is not an operator. In a field-specific search, the words in the double quotation marks ("") are considered as a whole.
\	The escape character. This character is used to escape double quotation marks (""). Escaped double quotation marks ("") indicate double quotation marks (""). For example, if the log content is `instance_id:nginx"01"`, you can execute the `instance_id:nginx\"01\"` statement to query log entries.
*	The wildcard character. This character is used to match zero, one, or multiple characters. Example: `host:www.yl.mo*k.com`. **Note** Log Service searches all log entries and obtains up to 100 words that meet the specified conditions. Then, Log Service returns log entries that contain one or more of the 100 words and meet the search conditions.
?	The wildcard character. This character is used to match a single character. Example: `host:www.yl.mo?k.com`.
>	Queries the log entries in which the value of a specified field is greater than a specified number. Example: `request_time>100`.
>=	Queries the log entries in which the value of a specified field is greater than or equal to a specified number. Example: `request_time>=100`.
<	Queries the log entries in which the value of a specified field is smaller than a specified number. Example: `request_time<100`.

Operator	Description
<=	Queries the log entries in which the value of a specified field is smaller than or equal to a specified number. Example: `request_time<=100`.
=	Queries the log entries in which the value of a specified field is equal to a specified number. Equal signs (=) and colons (:) have the same effect on fields of the double or long data type. For example, `request_time=100` is equivalent to `request_time:100`.
in	Queries the log entries in which the value of a specified field is within a numerical range. Brackets [] indicate a closed interval and parentheses () indicate an open interval. A space character is used to separate two numbers. Example: `request_time in [100 200]` or `request_time in (100 200]`. **Note** The characters in must be in lowercase.
__source__	Queries the log entries of a specified log source. Wildcard characters are supported. Example: `__source__:192.0.2.*`. **Notice** The __source__ field is a reserved field in Log Service. This field can be abbreviated to source. If you customize a field in the source format, the custom field conflicts with the reserved source field of Log Service. If you want to search for the custom field, you must use Source or SOURCE in a search statement.
__tag__	Queries log entries based on metadata. Example: `__tag__:__receive_time__:1609837139`.
__topic__	Queries the log entries of a specified log topic. Example: `__topic__:nginx_access_log`.

Sample search statements

Expected search result	Search statement
Log entries that contain successful GET requests (status codes: 200 to 299)	`request_method:GET and status in [200 299]`
Log entries that contain GET requests but do not contain the China (Shanghai) region	`request_method:GET not region:cn-shanghai`
Log entries that contain GET requests or POST requests	`request_method:GET or request_method:POST`
Log entries that do not contain GET requests	`not request_method:GET`
Log entries that contain successful GET requests or successful POST requests	`(request_method:GET or request_method:POST) and status in [200 299]`

Expected search result	Search statement
Log entries that contain failed GET requests or failed POST requests	`(request_method:GET or request_method:POST) not status in [200 299]`
Log entries that contain successful GET requests (status codes: 200 to 299) and in which the request duration is less than 60 seconds	`request_method:GET and status in [200 299] not request_time>=60`
Log entries in which the request duration is equal to 60 seconds	`request_time:60` `request_time=60`
Log entries in which the request duration is greater than or equal to 60 seconds and less than 200 seconds	`request_time>=60 and request_time<200` `request_time in [60 200)`
Log entries in which the value of the http_user_agent field contains Firefox	`http_user_agent:Firefox`
Log entries in which the value of the http_user_agent field contains Linux and Chrome	`http_user_agent:"Linux Chrome"` `http_user_agent:Linux and http_user_agent:Chrome`
Log entries that contain and	`"and"` In this search statement, and is a common string but not an operator.
Log entries in which the value of the http_user_agent field contains Firefox or Chrome	`http_user_agent:Firefox or http_user_agent:Chrome`
Log entries in which the value of the file info field contains apsara	`"file info":apsara`
Log entries that start with cn	`cn*`
Log entries in which the value of the region field starts with cn	`region:cn*`
Log entries in which the value of the region field contains cn*	`region:"cn*"`

Expected search result	Search statement
Log entries in which the value of the region field ends with hai	Not supported.
Log entries that start with mo, end with la, and contain one character between mo and la	`mo?la`
Log entries that start with mo, end with la, and contain zero, one, or more characters between mo and la	`mo*la`
Log entries that contain words that start with Moz and words that start with Sa	`Moz* and Sa*`
Log entries whose log topics are https or http	`__topic__:https or __topic__:http`
Log entries that are collected from the 192.0.2.1 host	`__tag__:__client_ip__:192.0.2.1` `__tag__:__client_ip__` indicates the IP address of the host where log entries reside.
Log entries in which the value of the remote_user field is not empty	`not remote_user:""`
Log entries in which the value of the remote_user field is empty	`remote_user:""`
Log entries that do not contain the remote_user field	`not remote_user:*`
Log entries that contain the remote_user field	`remote_user:*`
Log entries in which the value of the request_uri field is /request/path-2	`request_uri:/request/path-2`

23.1.4.7.2. LiveTail

This topic describes how to use LiveTail to monitor and analyze logs.

Prerequisites

Logs are collected by Logtail. For more information, see Use Logtail to collect logs.

> **Note** LiveTail can monitor and analyze only the log data that is collected by Logtail.

Context

In online O&M scenarios, you may need to monitor log data in real time and extract key information from the latest log data to identify causes of errors. In the traditional O&M model, you must run the `tail -f` command on the server if you want to monitor logs in real time. LiveTail that is provided in the Log Service console allows you to monitor and analyze online log data in real time. This helps you reduce your O&M workloads.

Features

- Log entries are monitored in real time and filtered by keyword.
- Log entries are collected and indexed based on the configurations of log collection.
- Log fields are delimited. This allows you to query contextual logs that contain delimiters.
- A log file can be found based on a log entry in the log file. This allows you to monitor the log file in real time without the need to log on to a server.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project that you want to manage.
3. Click the icon next to the name of the Logstore in which logs are stored, and then select **Search & Analysis**.
4. On the **Raw Logs** tab, click the icon of a log entry and select **LiveTail**.
5. View logs on LiveTail.

   After LiveTail is started, log data collected by Logtail is displayed on the page in real time. By default, the latest log data is displayed at the bottom of the list. You can view the latest log data without the need to scroll down. Up to 1,000 log entries can be displayed on the page. If more than 1,000 log entries are collected, the page is automatically refreshed to show the latest 1,000 log entries.
6. If exceptions are detected in log data, click **Stop LiveTail**.

   After you stop LiveTail, log entries in the log monitoring list are no longer updated. You can analyze and fix errors that are found when you monitor logs.

23.1.4.7.3. LogReduce

This topic describes how to use the LogReduce feature of Log Service. You can enable the LogReduce feature, view raw logs and log clustering results, adjust the precision of log clustering, and compare the number of log entries in different time ranges.

Context

The LogReduce feature allows you to cluster similar logs and extract patterns from the logs. The LogReduce feature supports text logs of multiple formats. You can use the feature to perform O&M operations in DevOps scenarios, for example, identify the causes of errors, detect anomalies, and roll back versions. You can also use the feature to detect network intrusions to ensure data security. In addition, you can use charts to display the results of log clustering, add the charts to a dashboard, and then view the clustered data in real time.

Benefits

- You can cluster logs in multiple formats, such as Log4j logs, JSON-formatted logs, and single-line logs.
- You can cluster hundreds of millions of log entries within seconds.
- You can cluster logs in a variety of modes.
- You can retrieve raw log entries based on pattern signatures.
- You can compare log patterns of different time ranges.

- You can adjust the precision of log clustering.

Enable the LogReduce feature of a Logstore

By default, the LogReduce feature is disabled.

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to enable the LogReduce feature.
3. Click the icon next to the name of the Logstore in which you want to enable the LogReduce feature, and then select **Search & Analysis**.
4. Enable the LogReduce feature.
   i. Choose **Index Attributes > Attributes**.
   If the indexing feature is not enabled, click **Enable**.
   ii. In the **Search & Analysis** panel, turn on **LogReduce**.
   iii. (Optional)Configure a whitelist or a blacklist to filter fields.
   You can filter logs based on keywords. Logs that are filtered based on keywords are automatically clustered.
   iv. Click **OK**.

View raw logs and the log clustering results

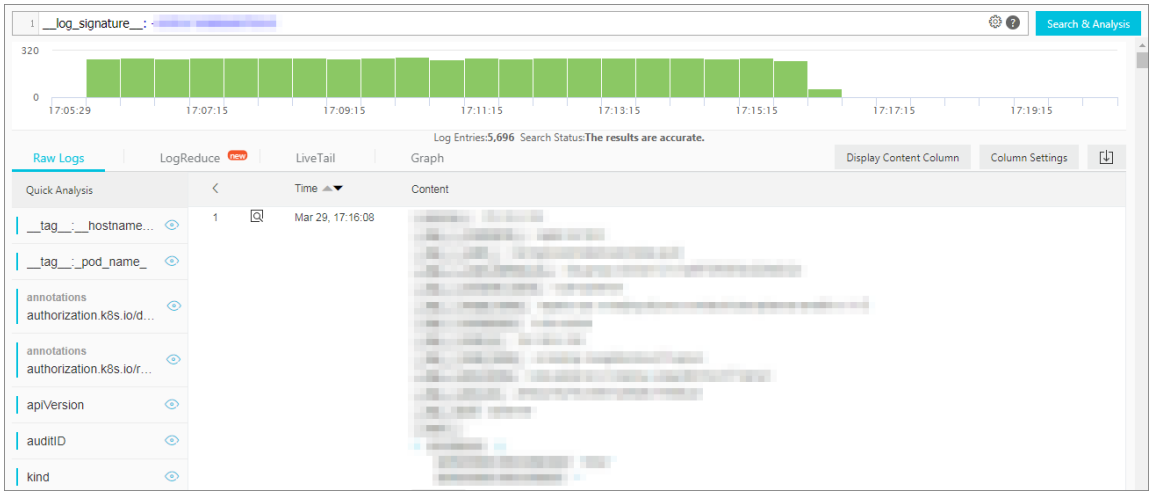
1. On the Search & Analysis page, enter a search statement in the search box, specify a time range, and then click **Search & Analytics**.

   Note You can use only search statements to filter logs. You cannot use analytic statements to analyze logs because the LogReduce feature cannot cluster analysis results.

2. Click the **LogReduce** tab to view the log clustering results.
   On the **LogReduce** tab, you can view the filtered log clustering results.

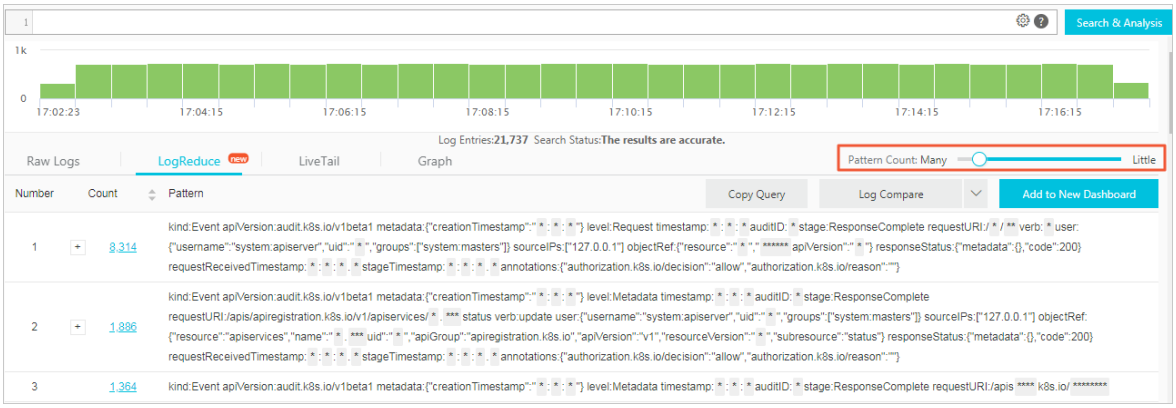
Parameter	Description
Number	The sequence number of a log cluster.
Count	The number of log entries in a pattern. The log entries are obtained in the current time range.
Pattern	The log pattern. Each log cluster has one or more sub-patterns.

   - Move the pointer over a value in the **Count** column to view the sub-patterns of the corresponding log cluster. You can also view the percentage of each sub-pattern in the log cluster. Click the plus sign (+) next to the value to expand the sub-pattern list.
   - Click a value in the Count column to view the raw log entries of the corresponding pattern.



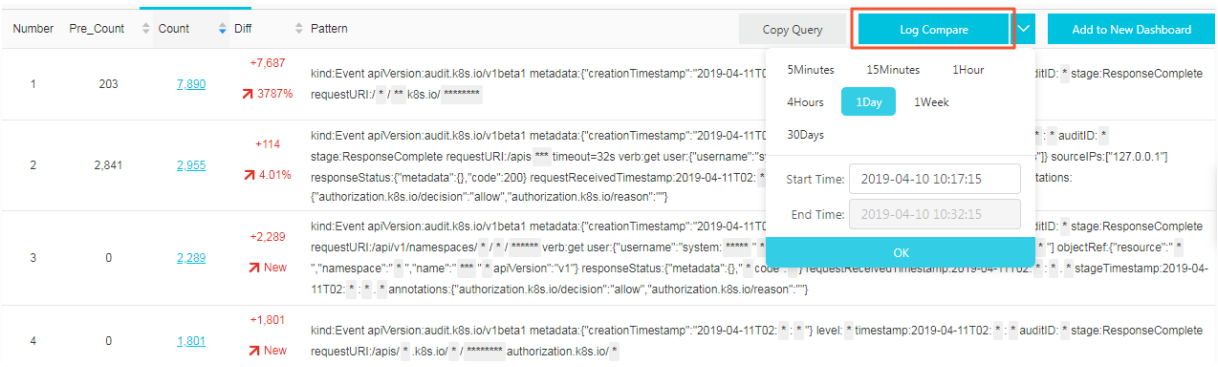
Adjust the precision of log clustering

1. On the Search & Analysis page, click the **LogReduce** tab.
2. In the upper-right corner of the tab, drag the **Pattern Count** slider to adjust the precision of log clustering.
   - If you drag the slider towards **Many**, you can obtain a more precise log clustering result with more detailed patterns.
   - If you drag the slider towards **Little**, you can obtain a less precise log clustering result with fewer detailed patterns.



Compare the number of log entries that are clustered in different time ranges

On the **LogReduce** tab, click **Log Compare**, select a time range, and then click **OK**.



Display item	Description
Number	The sequence number of a log cluster.
Pre_Count	The number of log entries that are clustered based on the current pattern within the previous time range.
Count	The number of log entries that are clustered based on the current pattern in the current time range.
Diff	The difference between the value of Pre_Count and the value of Count.
Pattern	The pattern of a log cluster.

Examples

The following examples show SQL statements that you can use to obtain clustered log data.

- Obtain log clustering results.
  - SQL statement

    ```
    * | select a.pattern, a.count,a.signature, a.origin_signatures from (select log_reduce(3) as a fr
    om log) limit 1000
    ```

    > **Note** When you view log clustering results, you can click **Copy Query** to obtain the related SQL statement.

  - Input parameter: log_reduce (precision)

    precision: the precision of log clustering. The value is an integer that ranges from 1 to 16. A smaller value indicates a higher precision and more patterns. The default value is 3.

  - Returned fields:
    - pattern: the sub-patterns of log entries in a log cluster.
    - count: the number of log entries in a log cluster.
    - signature: the signature of a log cluster.
    - origin_signatures: the secondary signature of a log cluster. You can use this signature to query raw log entries.
- Compare the number of log entries that are clustered in different time ranges.
  - SQL statement

    ```
    * | select v.pattern, v.signature, v.count, v.count_compare, v.diff from (select compare_log_redu
    ce(3, 86400) as v from log) order by v.diff desc limit 1000
    ```

    > **Note** When you click **Log Compare** to compare log clustering results in different time ranges, you can click **Copy Query** to obtain the related SQL statement.

  - Input parameters: compare_log_reduce(precision, compare_interval)
    - precision: the precision of log clustering. The value is an integer that ranges from 1 to 16. A smaller value indicates a higher precision and more patterns. The default value is 3.
    - compare_interval: the time difference between the two time ranges when the number of log entries is compared. The value of this parameter must be a positive integer. Unit: seconds.

- Returned fields:
  - pattern: the sub-patterns of log entries in a log cluster.
  - signature: the signature of a log cluster.
  - count: the number of log entries in a log cluster.
  - count_compare: the number of log entries for a same-pattern log cluster within a specified time range.
  - Diff: the difference between the value of the count field and the value of the count_compare field.

23.1.4.7.4. Contextual query

This topic describes the contextual query feature provided in the Log Service console. You can use this feature to query the full context of the log file from which specified log entries are obtained.

Prerequisites

- Logs are collected by Logtail. For more information, see Use Logtail to collect logs.

  > **Note** The contextual query feature is supported only for log data that is collected by Logtail.

- The indexing feature is enabled and indexes are configured. For more information, see Enable the indexing feature and configure indexes for a Logstore.

Context

To perform a contextual query, you must specify a source server, a source file, and a log entry whose context you want to query. You can obtain the log entries that precede or follow the specified log entry collected from the log file of the server. This helps you identify and resolve errors.

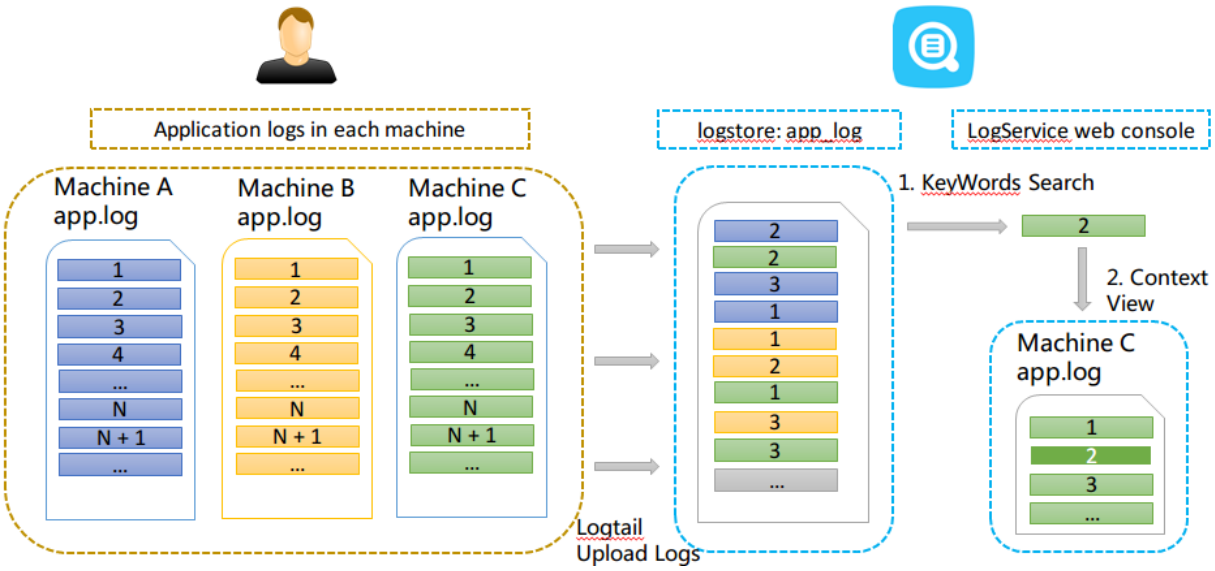
Scenarios

For example, a transaction on an online-to-offline (O2O) takeout website is recorded in an application log file on a server. You must perform the following steps to complete a transaction: logon to the website, browse products, select a product, add the product to the shopping cart, place an order, pay for the order, deduct the order amount, and generate the order.

If the order fails, the O&M engineers must identify the cause of the failure at the earliest opportunity. In traditional contextual queries, the O&M engineers must be authorized by an administrator before they can log on to each server on which the O2O application is deployed. After the authorization is complete, the O&M engineers can search application logs files by order ID to identify the cause of the failure.

In Log Service, the O&M engineers can perform the following steps to locate the cause of the failure:

1. Install Logtail on the server. Create a machine group and a Logtail configuration in the Log Service console. Then, enable Logtail to upload incremental log entries to Log Service.
2. On the search and analysis page of the Log Service console, specify a time range and find the log entry that records the failure based on the order ID.
3. After you find the log entry, scroll up until other related log entries are found, for example, a log entry that records a credit card payment failure.



Note The contextual query feature does not support syslog logs.

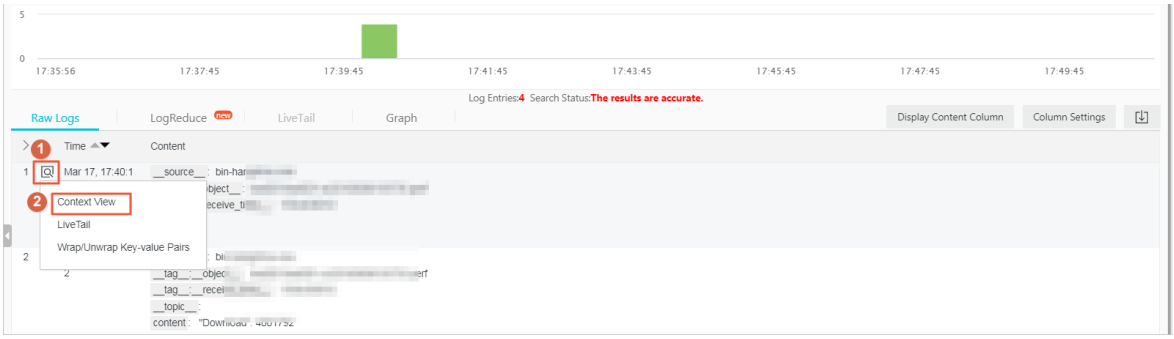
Benefits

- You can identify the causes of failures without the need to modify applications or log file formats.
- You can query the context of a log entry from a log file that is collected from a server in the Log Service console. You do not need to log on to the server to query the context.
- You can specify a time range to find suspicious log entries before you perform a contextual query in the Log Service console. This improves troubleshooting efficiency.
- You do not need to worry about data loss that is caused by insufficient server storage or log file rotation. You can view historical log data in the Log Service console at any time.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to perform contextual queries.
3. Click the icon next to the name of the Logstore in which you want to perform contextual queries, and then select **Search & Analysis**.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analytics**.
5. On the **Raw Logs** tab, find the log entry whose context you want to query and click the icon.

   On the query result page, if the **Context View** icon is available in the drop-down list of the icon to the left of a log entry, the log entry supports contextual query.



6. On the page that appears, scroll up and down to view the contextual log entries of the selected log entry.
   - To scroll up, click **Old**.
   - To scroll down, click **New**.
   - To configure the displayed fields, click **Filter by Field**.

23.1.4.7.5. Saved search

Log Service provides the saved search feature that can save the required data query and analysis operations. The saved search feature allows you to query and analyze log data in an efficient manner. This topic describes how to create a saved search in the Log Service console.

Prerequisites

The indexing feature is enabled and indexes are configured. For more information, see Enable the indexing feature and configure indexes for a Logstore.

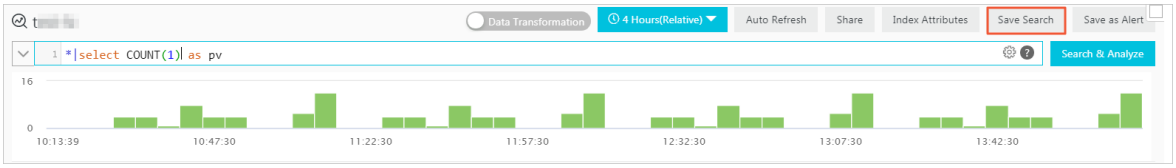
Context

If you need to frequently view the result of a query statement, you can save the query statement as a saved search. Then, you can click the name of the saved search on the left side of the search and analysis page to execute the query statement and view the result.

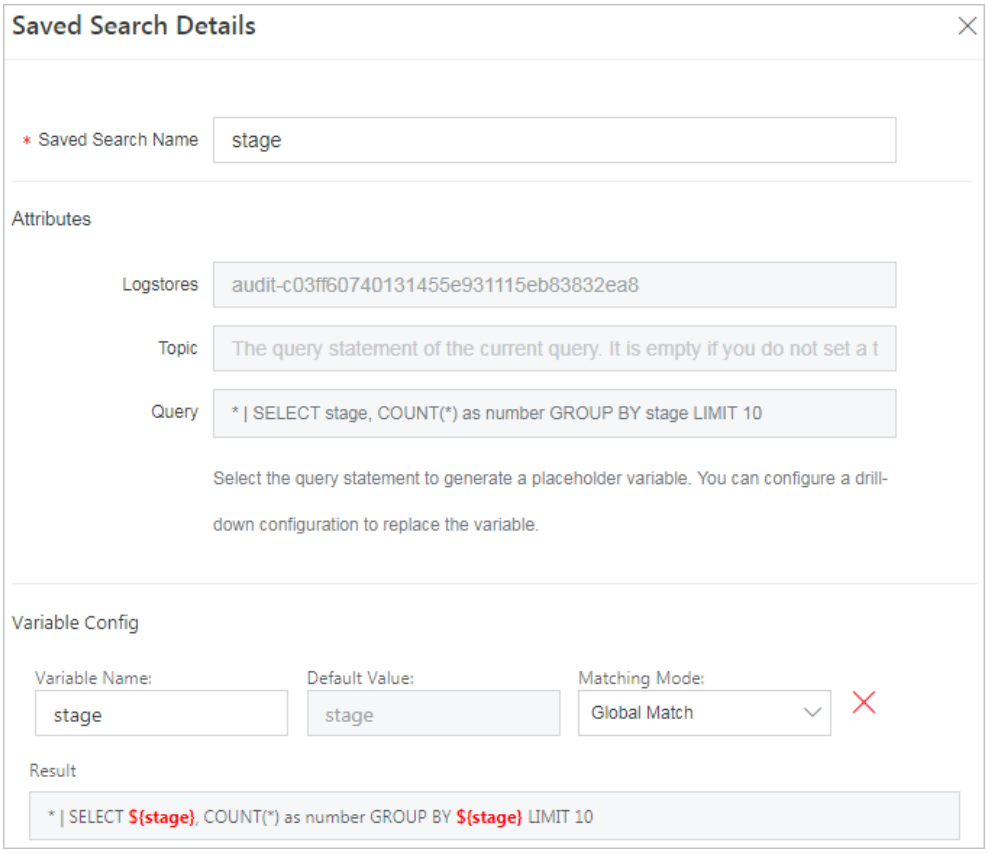
You can also use the saved search in alert rules. Log Service periodically executes the query statement of the saved search and sends alert notifications if the query result meets the preset condition.

Create a saved search

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to create a saved search.
3. Click the icon next to the name of the Logstore in which you want to query and analyze data, and then select **Search & Analysis**.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analytics**.
5. In the upper-right corner of the page, click **Save Search**.



6. In the **Saved Search Details** panel, set the required parameters. The following table describes the parameters.

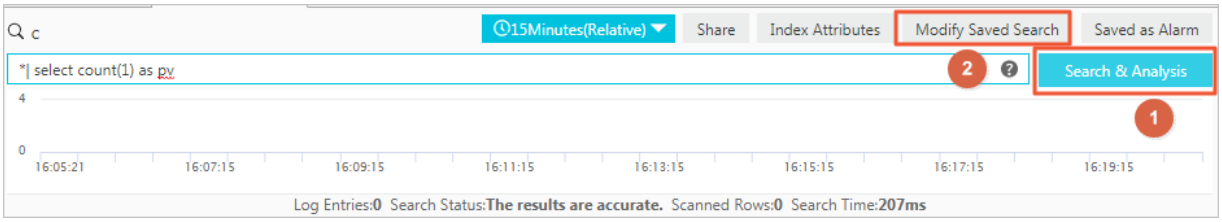


Parameter	Description
Saved Search Name	The name of the saved search. The name must be 3 to 63 characters in length.
Variable Config	Select the required content of the query statement in the **Query** field and click **Generate Variable** to generate a placeholder variable. ○ **Variable Name**: the name of the placeholder variable. ○ **Default Value**: the content that you select from the **Query** field. ○ **Matching Mode**: the match mode. You can use the match mode to replace the default value by triggering a drill-down event. Valid values: Global Match and Exact Match. For example, you set **Event Action** to **Open Saved Search** for a chart when you configure drill-down analysis for the chart, and specify the saved search. The **Variable** of the chart is the same as the **Variable Name** of the saved search. When you click the chart value, you are redirected to the saved search. The **Default Value** of the placeholder variable is replaced by the chart value that triggers the drill-down event. Then, the new query statement is executed. For more information, see Drill-down analysis. **Note** Before you can set **Event Action** to **Open Saved Search**, you must create a saved search and configure variables.

7. Click **OK**.

Modify a saved search

In the left-side navigation pane, click the Saved Search icon. In the Saved Search list, click the saved search that you want to modify. Enter a new query statement in the search box, click **Search & Analytics**, and then click **Modify Saved Search**. In the **Saved Search Details** panel, modify the settings and click OK.



23.1.4.7.6. Quick analysis

Log Service provides the quick analysis feature that allows you to analyze the distribution of a field within a specified time range in an efficient manner.

Prerequisites

Indexes are configured and the analysis feature is enabled for specified fields. For more information, see Enable the indexing feature and configure indexes for a Logstore.

Features

- Allows you to analyze the first 100,000 log entries that are returned for a query.

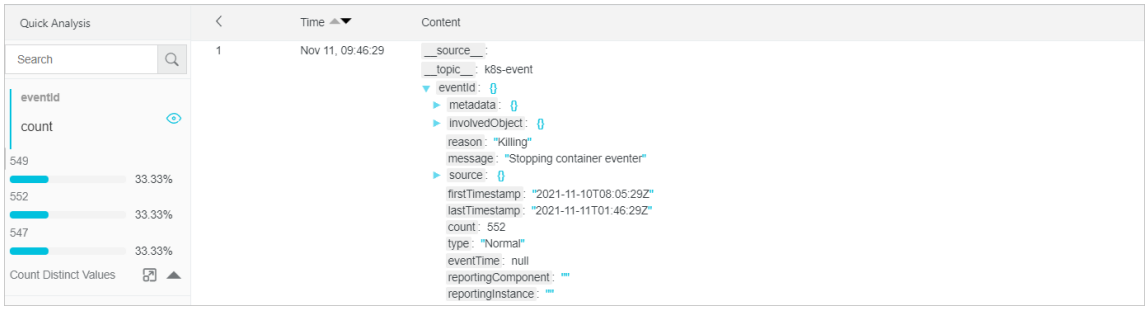
  > **Note** When you perform quick analysis on log entries within a specified time range, Log Service collects the first 100,000 log entries. If you use a saved search to query all data in a Logstore, you must delete the Limit 100000 clause.

- Groups fields of the text type and provides statistics about the top 10 groups.
- Generates **approx_distinct** statements for fields of the text type.
- Supports histogram-based statistics about the approximate distribution of fields of the long and double type.

  Histogram-based statistics group sampling data and calculate the average value of each group.
- Searches for the maximum, minimum, average, or sum of fields of the long and double type.
- Generates a query statement based on quick analysis.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to perform quick analysis.
3. Click the name of the Logstore in which you want to perform quick analysis.
4. On the **Raw Logs** tab, click the field that you want to analyze in the **Quick Analysis** column.
5. Click the icon to perform quick analysis based on the current query time range and view the distribution of the field in logs.



Text type

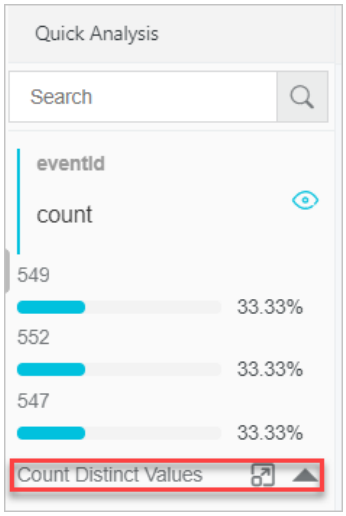
- Group and analyze log data by field of the text type.

  Click the Eye icon next to a field of the **text** type to group the first 100,000 log entries and obtain the percentages of the top 10 groups.

  The following example shows a query statement:

  ```
  $Search | select ${keyName} , pv, pv *1.0/sum(pv) over() as percentage from( select count(1) as pv , "${keyName}" from (select "${keyName}" from log limit 100000) group by "${keyName}" order by pv desc) order by pv desc limit 10
  ```

  The following figure shows the result that is returned after log entries are grouped by the `request_method` field. In this example, `GET` requests account for the majority of all returned requests.

- Calculate the number of unique values of a field.

  Under the specified field in the **Quick Analysis** column, click **Count Distinct Values** to calculate the number of unique values of the `${keyName}` field.

- Automatically copy the query statement that is used to group and analyze data to the Search & Analyze search box.

  Click the icon next to **Count Distinct Values**. The query statement that is used to group and analyze data is automatically copied to the Search & Analyze search box. You can modify the query statement.

Long and double types

- Display approximate distribution by using histograms.

  A large number of values exist for fields of the `long and double` types. Therefore, the preceding grouping and analysis method is not suitable for these fields. You can use the following query statement to divide field values into 10 buckets and display the approximate distribution of the values in a histogram:

  ```
  $Search | select numeric_histogram(10, ${keyName})
  ```

  The following figure shows the approximate distribution of the values in the `request_time` field. In this example, the largest percentage of request time is distributed at approximately 0.059 seconds.

- Perform quick analysis by using the `Max`, `Min`, `Avg`, and `Sum` clauses.

  You can click `Max` under a field to search for the maximum value, `Min` to search for the minimum value, `Avg` to calculate the average value, and `Sum` to calculate the sum of the values.

- Automatically copy the query statement that is used to calculate the approximate distribution to the Search &

Analyze search box.

Click the  icon next to `Sum`. The query statement that is used to calculate the approximate distribution is automatically copied to the Search & Analyze search box. You can modify the query statement.

23.1.4.8. SQL syntax and functions

23.1.4.8.1. General aggregate functions

An aggregate function is used to calculate a set of values and return a single value. This topic describes the syntax of aggregate functions. This topic also provides examples on how to use the functions.

Function	Description	Example
arbitrary(KEY)	Returns a random, non-null value from a specified column.	The following query statement returns an arbitrary value from the request_method column: `* | SELECT arbitrary(request_method) AS request_method`
avg(KEY)	Calculates the arithmetic mean of the values in a specified column.	The following query statement returns the projects whose average latency is greater than 1,000 microseconds. You can execute the statement to analyze the write latency of the projects. `method: PostLogstoreLogs | SELECT avg(latency) AS avg_latency, Project GROUP BY Project HAVING avg_latency > 1000`
checksum(KEY)	Calculates the checksum of a specified column and returns a result that is encoded in Base64.	The following query statement calculates the checksum of the request_method column: `* | SELECT checksum(request_method)` The returned result is `D2UmTL3octI=`.
count(*)	Calculates the number of log entries.	The following query statement calculates the number of page views (PVs): `* | SELECT count(*) AS PV`
count(KEY)	Calculates the number of the log entries that contain a specified field. If the field value of a log entry is null, the log entry is not counted.	The following query statement calculates the number of the log entries that contain the request_method field: `* | SELECT count(request_method)`

Function	Description	Example
count(1)	Calculates the number of log entries. This function is equivalent to `count(*)`.	The following query statement calculates the number of PVs: `* | SELECT count(1) AS PV`
count_if(KEY)	Calculates the number of log entries that meet a specified condition.	The following query statement calculates the number of requests for the value of the url field. The value ends with abc. `* | SELECT count_if(url like '%abc')`
geometric_mean(KEY)	Calculates the geometric mean of the values in a specified column.	The following query statement calculates the geometric mean of request durations: `* | SELECT geometric_mean(request_time)`
max_by(KEY_01,KEY_02)	Returns the value of KEY_01 that is associated with KEY_02 whose value is the maximum value.	The following query statement returns the point in time when the highest consumption occurs: `* | SELECT max_by(UsageEndTime, PretaxAmount) as time`
max_by(KEY_01,KEY_02,n)	Returns the n values of KEY_01 that is associated with KEY_02 whose values are the first n maximum values. The returned result is a JSON array.	The following query statement returns the three request methods that have the longest request durations: `* | SELECT max_by(request_method,request_time,3)` The returned result is `["GET","PUT","DELETE"]`.
min_by(KEY_01,KEY_02)	Returns the value of KEY_01 that is associated with KEY_02 whose value is the minimum value.	The following query statement returns the request method whose request duration is the shortest: `* | SELECT min_by(request_method,request_time)`
min_by(KEY_01,KEY_02,n)	Returns the n values of KEY_01 that is associated with KEY_02 whose values are the first n minimum values. The returned result is a JSON array.	The following query statement returns the three request methods that have the shortest request durations: `* | SELECT min_by(request_method,request_time,3)` The returned result is `["GET","PUT","DELETE"]`.

Function	Description	Example
max(KEY)	Queries the maximum value of a specified column.	The following query statement queries the longest request duration: `* | SELECT max(request_time)`
min(KEY)	Queries the minimum value of a specified column.	The following query statement queries the shortest request duration: `* | SELECT min(request_time)`
sum(KEY)	Calculates the total value of a specified column.	The following statement calculates the total size of daily NGINX traffic: `* | select date_trunc('day',__time__) AS time, sum(body_bytes_sent) AS body_bytes_sent GROUP BY time ORDER BY time`
bitwise_and_agg(KEY)	Returns the result of the bitwise AND operation for the values of a specified column. The returned result is in the two's complement format.	The following query statement performs a bitwise AND operation on all values of the request_time column: `* | SELECT bitwise_and_agg(request_time)`
bitwise_or_agg(KEY)	Returns the result of the bitwise OR operation for values of a specified column. The returned result is in the two's complement format.	The following query statement performs a bitwise OR operation on all values of the request_time column: `* | SELECT bitwise_or_agg(request_time)`

23.1.4.8.2. Security check functions

Security check functions in Log Services are designed based on the globally shared WhiteHat Security asset library. This topic describes security check functions that you can use to check whether an IP address, domain name, or URL in logs is secure.

Scenarios

- O&M personnel of enterprises and institutions in Internet, gaming, information, and other industries that require robust O&M services can use security check functions to identify suspicious requests or attacks. They can also use the functions to implement in-depth analysis and defend against potential attacks.
- O&M personnel of enterprises and institutions in banking, securities, e-commerce, and other industries that require strong protection for internal assets can use security check functions to identify requests to suspicious websites and downloads initiated by trojans. Then the O&M personnel can take immediate actions to prevent potential losses.

Features

- Reliability: built upon the globally shared WhiteHat Security asset library that is updated in a timely manner.

- Efficiency: capable of screening millions of IP addresses, domain names, and URLs within seconds.
- Ease of use: supports the analysis of network logs by using the security_check_ip, security_check_domain, and security_check_url functions.
- Flexibility: supports interactive queries, report creation, and alert configurations and subsequent actions.

Functions

Function	Description	Example
security_check_ip	Checks whether an IP address is secure. • The value 1 indicates that the specified IP address is suspicious. • The value 0 indicates that the specified IP address is secure.	`select security_check_ip(real_client_ip)`
security_check_domain	Checks whether a domain name is secure. • The value 1 indicates that the specified domain name is suspicious. • The value 0 indicates that the specified domain name is secure.	`select security_check_domain(site)`
security_check_url	Checks whether a URL is secure. • The value 1 indicates that the specified URL is suspicious. • The value 0 indicates that the specified URL is secure.	`select security_check_domain(concat(host, url))`

Examples

- Check external suspicious requests and generate reports

  For example, an e-commerce enterprise collects logs from its NGINX servers and wants to scan suspicious client IP addresses. To do this, the enterprise can pass the ClientIP field in logs that are collected from the NGINX servers to the `security_check_ip` function and filter out IP addresses associated with the returned value 1. Then the enterprise can query the countries where the IP addresses are located and ISPs to which the IP addresses belong.

  SQL statement for this scenario:

  ```
  * | select ClientIP, ip_to_country(ClientIP) as country, ip_to_provider(ClientIP) as provider, count(1) as PV where security_check_ip(ClientIP) = 1 group by ClientIP order by PV desc
  ```

  Display the ISPs and countries in a map.

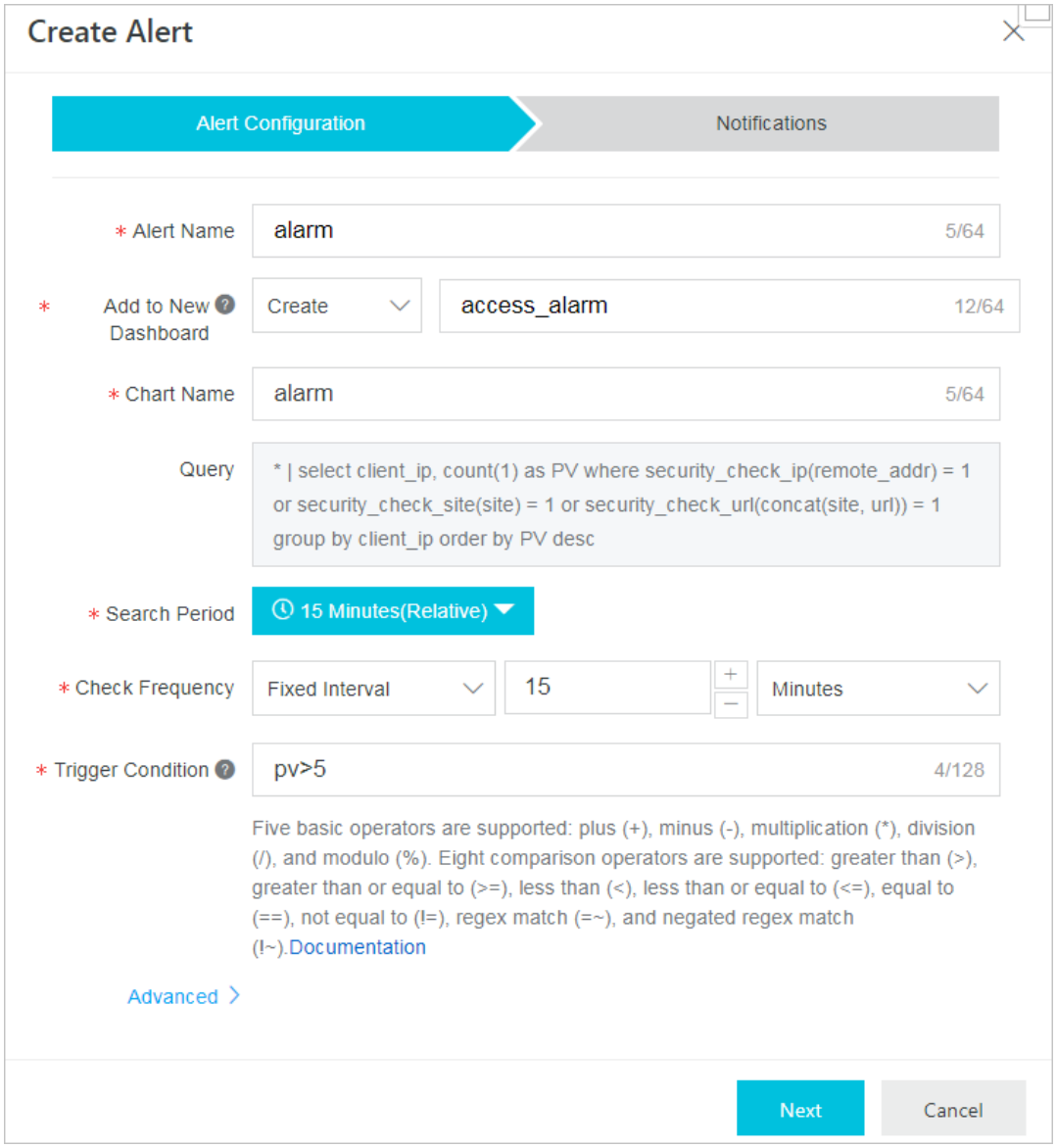
client_ip	country	provider	PV
180 3	CN		3
103	CN		3
180 7	CN		1

- Check internal suspicious requests and send alerts

  For example, a securities operator collects logs of its internal devices that access the Internet through gateways. To check requests to suspicious websites, the operator can run the following statement:

```
* | select client_ip, count(1) as PV where security_check_ip(remote_addr) = 1 or security_check_site(site) = 1 or security_check_url(concat(site, url)) = 1 group by client_ip order by PV desc
```

The operator can save this statement as a saved search and configure an alert. An alert is triggered when a client frequently accesses suspicious websites. The statement in the alert can be configured to run every five minutes to check if a client has frequently (more than five times) accessed suspicious websites in the past one hour. The following figure shows the configurations of an alert.



23.1.4.8.3. Map functions

This topic describes the syntax of map functions and provides examples on how to use the functions.

The following table describes the map functions that are supported by Log Service.

Function	Description	Example
Subscript operator []	Returns the value of a key from a map.	-

Function	Description	Example
histogram(x)	Groups the values of x and calculates the number of occurrences for each value. The syntax is equivalent to `select count group by x`. **Note** The return value is in the JSON format.	The statement `latency > 10 | select histogram(status)` is equivalent to the statement `latency > 10 | select count(1) group by status`.
histogram_u(x)	Groups the values of x and calculates the number of occurrences for each value. **Note** The return value is a table that contains multiple rows and columns.	The statement `latency > 10 | select histogram_u(status)` is equivalent to the statement `latency > 10 | select count(1) group by status`.
map_agg(Key,Value)	Returns a random value of the key in the format of a map that consists of key-value pairs.	`latency > 100 | select map_agg(method,latency)`
multimap_agg(Key,Value)	Returns all values of the key in the format of a map that consists of key-value pairs.	`latency > 100 | select multimap_agg(method,latency)`
cardinality(x) → bigint	Returns the size of a map.	-
element_at(map `<K, V>`, key) → V	Returns the value of a key from a map.	-
map() → map `<unknown, unknown>`	Returns an empty map.	-
map(array `<K>`, array `<V>`) → map `<K,V>`	Returns a map in which each key-value pair consists of two elements from two separate arrays.	`SELECT map(ARRAY[1,3], ARRAY[2,4]); — {1 -> 2, 3 -> 4}`
map_from_entries(array `<row<K, V>>`) → map `<K,V>`	Converts a multi-dimensional array to a map.	`SELECT map_from_entries(ARRAY[(1, 'x'), (2, 'y')]); — {1 -> 'x', 2 -> 'y'}`
map_concat(map1 `<K, V>`, map2 `<K, V>`, ..., mapN `<K, V>`) → map `<K,V>`	Returns a map that is the union of all specified maps. If a key is found in multiple maps, the value of the key in the returned map is the value of the key that occurs in the last map.	-
map_filter(map `<K, V>`, function) → map `<K,V>`	For more information, see the map_filter function in Lambda functions.	-
map_keys(x `<K, V>`) → array `<K>`	Returns an array that consists of the keys in the specified map.	-

Function	Description	Example
map_values(x <K, V>) → array <V>	Returns an array that consists of the values in the specified map.	-

23.1.4.8.4. Approximate functions

This topic describes the syntax of approximate functions. This topic also provides examples on how to use the functions.

Function	Description	Example
`approx_distinct(x)`	Estimates the number of unique values in the x field.	None
`approx_percentile(x,percentage)`	Sorts the values of the x field in ascending order and returns the value that is approximately at the percentage position.	`approx_percentile(x,0.5)`: returns the value that is approximately at the 50% position in the x field.
`approx_percentile(x,percentages)`	This function is similar to `approx_percentile(x,percentage)`. You can specify multiple percentages to return the values at the specified percentage positions.	`approx_percentile(x,array[0.1,0.2])`
`numeric_histogram(buckets,Value)`	Distributes all values of the *Value* field to multiple buckets. The *buckets* parameter specifies the number of buckets. The key of each bucket and the number of values in a bucket are returned. This function is equivalent to `select count group by`. **Note** The returned result is of the JSON type.	`method:POST | select numeric_histogram(10,latency)`: distributes the values of the latency field for POST requests to 10 buckets and calculates the number of latency field values in each bucket.
`numeric_histogram_u(buckets,Value)`	Distributes all values of the *Value* field to multiple buckets. The *buckets* parameter specifies the number of buckets. The key of each bucket and the number of values in a bucket are returned. This function is equivalent to `select count group by`. **Note** The returned result is a table that includes multiple rows and columns.	`method:POST | select numeric_histogram(10,latency)`: distributes the values of the latency field for POST requests to 10 buckets and calculates the number of latency field values in each bucket.

Note Buckets are evenly divided by aggregation degree. The returned result for each bucket includes the average value of the bucket and the number of values in the bucket.

23.1.4.8.5. Mathematical statistics functions

This topic describes the syntax of mathematical statistics functions. This topic also provides examples on how to use the functions.

Syntax

Function	Description
corr(*key1*, *key2*)	Calculates the correlation coefficient between two specific columns. The return value is in the range of [0,1].
covar_pop(*key1*, *key2*)	Calculates the population covariance of two specific columns.
covar_samp(*key1*, *key2*)	Calculates the sample covariance of two specific columns.
regr_intercept(*key1*, *key2*)	Returns the linear regression intercept of input values. *key1* is the dependent value and *key2* is the independent value.
regr_slope(*key1*, *key2*)	Returns the linear regression slope of input values. *key1* is the dependent value and *key2* is the independent value.
stddev(*key*)	Calculates the sample standard deviation of the *key* column. This function is equivalent to the stddev_samp function.
stddev_samp(*key*)	Calculates the sample standard deviation of the *key* column.
stddev_pop(*key*)	Returns the population standard deviation of the *key* column.
variance(*key*)	Calculates the sample variance of the *key* column. This function is equivalent to the var_samp function.
var_samp(*key*)	Calculates the sample variance of the *key* column.
var_pop(*key*)	Calculates the population variance of the *key* column.

Examples

- Example 1: Calculate the correlation coefficient of two specific columns.
  - Query statement

    ```
    * | SELECT corr(request_length, request_time)
    ```

  - Query and analysis result

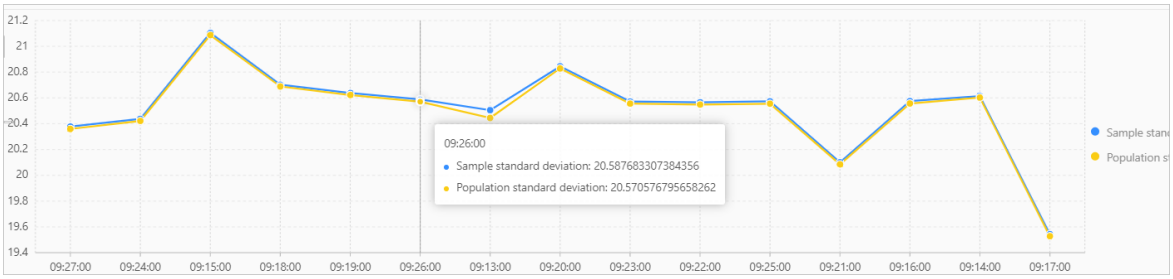
    

- Example 2: Query the sample standard deviation and population standard deviation of pre-tax income.

- Query statement

```
* | SELECT stddev(PretaxGrossAmount) as "sample standard deviation", stddev_pop(PretaxGrossAmount) as "population standard deviation", time_series(__time__, '1m', '%H:% I:%s', '0') AS time GROUP BY time
```

- Query and analysis result



23.1.4.8.6. Mathematical calculation functions

This topic describes the syntax of mathematical calculation functions. This topic also provides examples on how to use the functions.

Syntax

> **Note**
> - Mathematical calculation functions support the following operators: `+-*/%`.
> - In the following functions, *x* and *y* can be numbers, log fields, or expressions whose calculation result is a number.

Function	Description
abs(x)	Calculates the absolute value of a number.
cbrt(x)	Calculates the cube root of a number.
sqrt(x)	Calculates the square root of a number.
cosine_similarity(x,y)	Calculates the cosine similarity between x and y.
degrees(x)	Converts radians to degrees.
radians(x)	Converts degrees to radians.
e()	Returns Euler's number.
exp(x)	Returns Euler's number raised to the power of a number.
ln(x)	Calculates the natural logarithm of a number.
log2(x)	Calculates the base-2 logarithm of a number.
log10(x)	Calculates the base-10 logarithm of a number.
log(x,b)	Calculates the base-b logarithm of a number.
pi()	Returns the value of π to 14 decimal places.

Function	Description
pow(x,b)	Calculates the value of a number raised to the power of b.
rand()	Returns a random number.
random(0,n)	Returns a random number that is greater than or equal to 0 and less than n.
round(x)	Returns a number rounded to the nearest integer.
round(x, N)	Returns a number rounded to N decimal places.
floor(x)	Returns a number rounded down to the nearest integer. For example, when you execute the `* | SELECT floor(2.5)` statement, 2.0 is returned.
ceiling(x)	Returns a number rounded up to the nearest integer. For example, when you execute the `* | SELECT ceiling(2.5)` statement, 3.0 is returned.
from_base(varchar, bigint)	Converts a string to a base-encoded number.
to_base(x, radix)	Converts a number to a base-encoded string.
truncate(x)	Truncates the fractional part of a number.
acos(x)	Calculates the arc cosine of a number.
asin(x)	Calculates the arc sine of a number.
atan(x)	Calculates the arc tangent of a number.
atan2(y,x)	Calculates the arc tangent of the quotient of a number divided by another number.
cos(x)	Calculates the cosine of a number.
sin(x)	Calculates the sine of a number.
cosh(x)	Calculates the hyperbolic cosine of a number.
tan(x)	Calculates the tangent of a number.
tanh(x)	Calculates the hyperbolic tangent of a number.
infinity()	Returns a positive infinity value.
is_nan(x)	Checks whether a value is a non-numeric value.

Example

Compare the number of page views (PVs) of today with the number of PVs of the previous day, and show the comparison result as a percentage.

- Query statement

```
* | SELECT diff [1] AS today, round((diff [3] -1.0) * 100, 2) AS growth FROM (SELECT compare(pv, 86
400) as diff FROM (SELECT COUNT(*) as pv FROM website_log))
```

- Query and analysis result

today	growth
1564075.0	-22.11

23.1.4.8.7. String functions

This topic describes the syntax of string functions. This topic also provides examples on how to use the functions.

Syntax

> **Note**
> - If you want to use strings in analytic statements, you must enclose the strings in single quotation marks (''). Strings that are not enclosed or enclosed in double quotation marks ("") indicate field names or column names. For example, 'status' indicates the status string, and status or "status" indicates the status log field.
> - The key parameter in the following table indicates a log field.

Function	Description
chr(*number*)	Returns the characters that match the ASCII value of a specified parameter.
codepoint(*key*)	Converts a field of the ASCII type to a field value of the bigint type.
length(*key*)	Calculates the length of a string. The return value is of the integer type.
lower(*key*)	Converts the characters in a string to lowercase letters. The return value is of the varchar type in lowercase letters.
upper(*key*)	Converts the characters in a string to lowercase letters. The return value is of the varchar type in uppercase letters.
lpad(*key*, *length*, *lpad_string*)	Pads a string to a specified length from the left with a specified substring. The value of the *length* parameter is an integer that specifies the length of the result string. • If the length of the string is less than the value of the *length* parameter, the string is padded by the specified substring from the left. • If the length of the string is greater than the value of the *length* parameter, the function returns only the first *length* characters in the string. The return value is of the varchar type.

Function	Description
rpad(*key*, *length*,*rpad_string*)	Pads a string to a specified length from the right with a specified substring. The value of the *length* parameter is an integer that specifies the length of the result string. • If the length of the string is less than the value of the *length* parameter, the string is padded by the specified substring from the right. • If the length of the string is greater than the value of the *length* parameter, the function returns only the first *length* characters in the string. The return value is of the varchar type.
trim(*key*)	Deletes space characters from the start and the end of a string. The return value is of the varchar type.
ltrim(*key*)	Deletes space characters from the start of a string. The return value is of the varchar type.
rtrim(*key*)	Deletes space characters from the end of a string. The return value is of the varchar type.
replace(*key*,*substring*,*replace*)	Replaces matched characters in a string with specified characters. The return value is of the varchar type.
replace(*key*,*substring*)	Deletes matched characters from a string. The return value is of the varchar type.
reverse(*key*)	Reverses the characters in a string.
split(*key*,*delimeter*,*N*)	Splits a string with a specified delimiter and returns a set of N substrings. The return value is an array.
split_part(*key*,*delimeter*,*part*)	Splits a string with a specified delimiter and returns the substring at a specified position. The value of the *part* parameter is an integer that is greater than 0. The return value is of the varchar type.
split_to_map(*key*, *delimiter01*, *delimiter02*)	Splits a string with the first specified delimiter, and then splits the string with the second specified delimiter. The return value is a map.
position(*substring* IN *key*)	Returns the position of a specified substring in a string. The return value is of the integer type. The value starts from 1.
strpos(*key*, *substring*)	Returns the position of a specified substring in a string. This function is equivalent to the position(*substring* IN *key*) function. The return value is of the integer type. The value starts from 1.

Function	Description
substr(*key*, *start*)	Returns the substring at a specified position in a string. The *start* parameter specifies the position of the substring to be extracted. The value of the start parameter starts from 1. The return value is of the varchar type.
substr(*key*, *start*, *length*)	Returns the substring at a specified position in a string and specifies the length of the substring. The *start* parameter specifies the position of the substring to be extracted. The value of the start parameter starts from 1. The *length* parameter specifies the length of the substring. The return value is of the varchar type.
concat(*key01*,*key02*,*key03*)	Concatenates multiple strings into one string. The return value is of the integer type. The value starts from 1.
levenshtein_distance(*key01*, *key02*)	Returns the minimum edit distance between two strings.
hamming_distance (*string1*,*string2*)	Returns the Hamming distance between two strings.

Examples

Sample log:

```
http_user_agent:Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_4; en-us) AppleWebKit/528.4+ (KHTML, like Gecko) Version/4.0dp1 Safari/526.11.2
request_uri:/request/path-1/file-9?0457349059345
scheme:https
server_protocol:HTTP/2.0
region:cn-shanghai
time: upstream_response_time:"80", request_time:"40"
```

- Use a question mark (?) to split the value of the request_uri field and return the first substring. The returned substring indicates a file path. Then, calculate the number of requests that correspond to each path.

```
* | SELECT count(*) AS PV, split_part(request_uri, '?', 1) AS Path GROUP BY Path ORDER BY pv DESC LIMIT 3
```

PV	Path
4355	/request/path-1/file-5
4328	/request/path-1/file-1
4296	/request/path-2/file-3

- Extract the first four characters (HTTP) from the value of the server_protocol field and calculate the number of requests that use the HTTP protocol.

```
* | SELECT substr(server_protocol,1,4) AS protocol, count(*) AS count GROUP BY server_protocol
```

protocol	count
HTTP	9078

- Use commas (,) and colons (:) to split the value of the time field and return a value of the map type.

```
* | SELECT split_to_map(time,',',':')
```

_col0
{"request_time":"\"40\"","upstream_response_time":"\"80\""}
{"request_time":"\"40\"","upstream_response_time":"\"80\""}

- Check whether the value of the http_user_agent field starts with the letter M.

```
* | SELECT substr(http_user_agent, 1, 1)=chr(77)
```

_col0
true
true

- Return the position of the letter H in the value of the server_protocol field.

```
* | SELECT strpos(server_protocol,'H')
```

_col0
1
1

- Use a forward slash (/) to split the value of the server_protocol field into two substrings and return an array of the substrings.

```
* | SELECT split(server_protocol,'/',2)
```

_col0
["HTTP","2.0"]

- Replace cn in the region field with China.

```
* | select replace(region,'cn','China')
```

_col0
China-shanghai
China-shanghai

23.1.4.8.8. Date and time functions

Log Service provides the following types of date and time functions that you can use to analyze log data: date function, time function, truncation function, interval function, and time series padding function. You can use the functions to convert the date and time formats of log data. You can also use the functions to group and aggregate log data.

> **Note**
> - The timestamp of a log entry is accurate to seconds. Therefore, you can specify the time format only to seconds.
> - You need to specify the time format only for the time in a time string. Other parameters such as the time zone are not required.
> - Each log entry in Log Service contains the reserved __time__ field. The value of the field is a UNIX timestamp. For example, 1592374067 indicates 2020-06-17 14:07:47.

Date functions

Function	Description	Example
current_date	Returns the current date. • Return value format: `YYYY-MM-DD`, for example, 2021-01-12. • Return value type: date.	`* | select current_date`
current_time	Returns the current time. • Return value format: `HH:MM:SS.Ms Time zone`, for example, 01:14:51.967 Asia/Shanghai. • Return value type: time.	`* | select current_time`
current_timestamp	Returns the current date and time. • Return value format: `YYYY-MM-DD HH:MM:SS.Ms Time zone`, for example, 2021-01-12 17:16:09.035 Asia/Shanghai. • Return value type: timestamp.	`* | select current_timestamp`
current_timezone()	Returns the current time zone. Return value type: varchar, for example, Asia/Shanghai.	`* | select current_timezone()`
localtime	Returns the local time. • Return value format: `HH:MM:SS.Ms` • Return value type: time.	`* | select localtime`
localtimestamp	Returns the local date and time. • Return value format: `YYYY-MM-DD HH:MM:SS.Ms` • Return value type: timestamp.	`* | select localtimestamp`

Function	Description	Example
now()	Returns the current date and time. This function is equivalent to the current_timestamp function. • Return value format: `YYYY-MM-DD HH:MM:SS.Ms Time zone`. • Return value type: timestamp.	`* | select now()`
from_iso8601_timestamp(*ISO8601*)	Converts an ISO 8601-formatted datetime expression to a timestamp expression that contains a time zone. • Return value format: `YYYY-MM-DD HH:MM:SS.Ms Time zone`. • Return value type: timestamp.	`* | select from_iso8601_timestamp('2020-05-03T17:30:08')`
from_iso8601_date(*ISO8601*)	Converts an ISO 8601-formatted date expression to a date expression. • Return value format: `YYYY-MM-DD` • Return value type: date.	`* | select from_iso8601_date('2020-05-03')`
from_unixtime (*UNIX timestamp*)	Converts a UNIX timestamp to a timestamp expression. • Return value format: `YYYY-MM-DD HH:MM:SS.Ms` • Return value type: timestamp.	`* | select from_unixtime(1494985275)`
from_unixtime (*UNIX timestamp*,*time zone*)	Converts a UNIX timestamp to a timestamp expression that contains a time zone. • Return value format: `YYYY-MM-DD HH:MM:SS.Ms Time zone`. • Return value type: timestamp.	`* | select from_unixtime(1494985275,'Asia/Shanghai')`
to_unixtime(*timestamp*)	Converts a timestamp expression to a UNIX timestamp. Return value type: long, for example, 1494985500.848.	`*| select to_unixtime('2017-05-17 09:45:00.848 Asia/Shanghai')`

Time functions

Function	Description	Example
date_format(*timestamp*,*format*)	Converts a timestamp expression to a datetime format string.	`* | select date_format(date_parse('2017-05-17 09:45:00','%Y-%m-%d %H:%i:%S'), '%Y-%m-%d')`

Function	Description	Example
date_parse(*string*,*format*)	Represents a datetime format string, and then converts the datetime format string to a timestamp expression. The following table describes the formats.	`* | select date_format (date_parse(time,'%Y-%m-%d %H:%i:%S'), '%Y-%m-%d')`

Formats

format	Description
%a	The abbreviated day name, for example, Sun or Sat.
%b	The abbreviated month name, for example, Jan or Dec.
%c	The month. Valid values: 1 to 12.
%D	The day of the month, for example, 0th, 1st, 2nd, or 3rd.
%d	The day of the month, Valid values: 01 to 31.
%e	The day of the month, Valid values: 1 to 31.
%H	The hour in the 24-hour clock.
%h	The hour in the 12-hour clock.
%I	The hour in the 12-hour clock.
%i	The minutes. Valid values: 00 to 59.
%j	The day of the year. Valid values: 001 to 366.
%k	The hours. Valid values: 0 to 23.
%l	The hours. Valid values: 1 to 12.
%M	The full month name, for example, January or December.
%m	The month. Valid values: 01 to 12.
%p	The abbreviation that indicates the morning or afternoon. Valid values: AM and PM.
%r	The time in the 12-hour clock. The time is in the `hh:mm:ss AM/PM` format.
%S	The seconds. Valid values: 00 to 59.
%s	The seconds. Valid values: 00 to 59.
%T	The time in the 24-hour clock. The time is in the `hh:mm:ss` format.
%V	The week number of the year. Sunday is the first day of each week. Valid values: 01 to 53.
%v	The week number of the year. Monday is the first day of each week. Valid values: 01 to 53.

format	Description
%W	The full day name, for example, Sunday or Saturday.
%w	The day of the week as a number. The value 0 indicates Sunday.
%Y	The four-digit year number, for example, 2020.
%y	The two-digit year number, for example, 20.
%%	Escapes the percent sign (%).

Truncation function

The date_trunc() function truncates a datetime expression based on the specified part of a time. You can use a truncation function to truncate a time by second, minute, hour, day, month, or year. This function is suitable for time-based statistics.

- Syntax

```
date_trunc('unit',x)
```

- Parameters

  The value of the x parameter can be a datetime expression, for example, 2021-01-12 03:04:05.000 or 1610350836. The value of the x parameter can be a time field, for example, __time__. The valid values of the unit parameter are second, minute, hour, day, week, month, quarter, and year. The following table describes examples of this parameter.

Example	Result	Description
* | select date_trunc('second', 2021-01-12 03:04:05.000)	2021-01-12 03:04:05.000	None
* | select date_trunc('minute', 2021-01-12 03:04:05.000)	2021-01-12 03:04:00.000	None
* | select date_trunc('hour', 2021-01-12 03:04:05.000)	2021-01-12 03:00:00.000	None
* | select date_trunc('day', 2021-01-12 03:04:05.000)	2021-01-12 00:00:00.000	Returns 00:00:00.000 of the specified date.
* | select date_trunc('week', 2021-01-12 03:04:05.000)	2021-01-11 00:00:00.000	Returns 00:00:00.000 of the Monday of the specified week.
* | select date_trunc('month', 2021-01-12 03:04:05.000)	2021-01-01 00:00:00.000	Returns 00:00:00.000 of the first day of the specified month.
* | select date_trunc('quarter', 2021-01-11 03:04:05.000)	2021-01-01 00:00:00.000	Returns 00:00:00.000 of the first day of the specified quarter.
* | select date_trunc('year', 2021-01-11 03:04:05.000)	2021-01-01 00:00:00.000	Returns 00:00:00.000 of the first day of the specified year.

- Query and analysis examples

  To truncate the average request durations by minute, and group and sort the average request durations by time, execute the following query statement:

```
* | select  date_trunc('minute' ,  __time__)  as time,
       truncate (avg(request_time) ) as avg_time ,
       current_date as date
       group by  time
       order by time  desc
       limit 100
```

You can use the date_trunc('unit', *x*) function to truncate a time only by second, minute, hour, day, month, or year. To truncate a time based on specified intervals such as 5 minutes, you must use a GROUP BY clause based on the modulus method.

```
* | select count(1) as pv,  __time__ - __time__ %300 as time group by time limit 100
```

In the preceding statement, `%300` indicates that modulo and truncation are performed every 5 minutes.

Interval functions

You can use interval functions to perform the interval-related calculations. For example, you can add or subtract an interval based on a date, or calculate the interval between two dates.

Function	Description

Function	Description
date_add(*unit*, *N*,*timstamp*)	Adds N units to a `timestamp`. To subtract an interval, set the value of *N* to a negative value.

Function	Description

Function	Description
date_diff(*unit*, *timestamp1*, *timestamp2*)	Returns the time difference between two time expressions. For example, you can calculate the difference between `timestamp1` and `timestamp2` by unit.

The following table describes the valid values of the unit parameter.

unit	Description
millisecond	Unit: milliseconds.
second	Unit: seconds.
minute	Unit: minutes.
hour	Unit: hours.
day	Unit: days.
week	Unit: weeks.
month	Unit: months.
quarter	Unit: quarters.
year	Unit: years.

Time series padding function

You can use the time_series() function to add the missing data when you query in the time window.

Notice You must use the time_series() function together with GROUP BY and ORDER BY clauses. You cannot use the DESC keyword in an ORDER BY clause to sort data returned in descending order.

- Syntax

```
time_series(time_column, window, format, padding_data)
```

- Parameters

Parameter	Description
time_column	The sequence of time (KEY), for example, __time__. The value of this parameter can be a long datetime or timestamp expression.
window	The size of the window. Valid units: s (seconds), m (minutes), h (hours), and d (days). For example, you can set the window to 2h, 5m, or 3d.
format	The format in which you want the function to return the value.
padding_data	The content that you want to add. Valid values: ○ 0: adds 0. ○ null: adds null. ○ last: adds the value of the last point in time. ○ next: adds the value of the next point in time. ○ avg: adds the average value of the last and next values.

- Example

  To add missing data by 2 hours, execute the following query statement:

```
* | select time_series(__time__, '2h', '%Y-%m-%d %H:%i:%s', '0')  as time, count(*) as num from log group by time order by time
```

time	num
2021-07-20 00:00:00	11602
2021-07-20 02:00:00	63089
2021-07-20 04:00:00	36583
2021-07-20 06:00:00	11135
2021-07-20 08:00:00	62746
2021-07-20 10:00:00	18314

23.1.4.8.9. URL functions

This topic describes the syntax of URL functions. This topic also provides examples on how to use the functions.

URL functions extract fields from standard URLs. The following example shows the format of a URL:

```
[protocol:][//host[:port]][path][?query][#fragment]
```

The following table describes common URL functions.

Function	Description	Example	
		Query statement	Query result
`url_extract_fragment(url)`	Extracts the fragment from a URL. The return value is of the varchar type.	`*| select url_extract_fragment('https://sls.console.aliyun.com/#/project/dashboard-demo/categoryList')`	`/project/dashboard-demo/categoryList`
`url_extract_host(url)`	Extracts the host from a URL. The return value is of the varchar type.	`*|select url_extract_host('http://www.aliyun.com/product/sls')`	`www.aliyun.com`
`url_extract_parameter(url, name)`	Extracts the value of a specified parameter in the query string from a URL. The return value is of the varchar type.	`*|select url_extract_parameter('http://www.aliyun.com/product/sls?userid=testuser','userid')`	`testuser`
`url_extract_path(url)`	Extracts the path from a URL. The return value is of the varchar type.	`*|select url_extract_path('http://www.aliyun.com/product/sls?userid=testuser')`	`/product/sls`
`url_extract_port(url)`	Extracts the port number from a URL. The return value is of the bigint type.	`*|select url_extract_port('http://www.aliyun.com:80/product/sls?userid=testuser')`	`80`
`url_extract_protocol(url)`	Extracts the protocol from a URL. The return value is of the varchar type.	`*|select url_extract_protocol('http://www.aliyun.com:80/product/sls?userid=testuser')`	`http`

Function	Description	Example	
		Query statement	Query result
`url_extract_query(url)`	Extracts the query string from a URL. The return value is of the varchar type.	`*|select url_extract_query('http://www.aliyun.com:80/product/sls?userid=testuser')`	`userid=testuser`
`url_encode(value)`	Encodes a URL.	`*|select url_encode('http://www.aliyun.com:80/product/sls?userid=testuser')`	`http%3a%2f%2fwww.aliyun.com%3a80%2fproduct%2fsls%3fuserid%3dtestuser`
`url_decode(value)`	Decodes a URL.	`*|select url_decode('http%3a%2f%2fwww.aliyun.com%3a80%2fproduct%2fsls%3fuserid%3dtestuser')`	`http://www.aliyun.com:80/product/sls?userid=testuser`

23.1.4.8.10. Regular expression functions

This topic describes the available regular expression functions. You can use these functions when you query and analyze data in Log Service.

A regular expression function parses a string and returns the required substrings.

The following table lists common regular expression functions.

Function	Description	Example
`regexp_extract_all(string, pattern)`	Returns an array where each element is a substring that matches the regular expression. These substrings derive from the specified string.	The returned result of `*| SELECT regexp_extract_all('5a 67b 890m', '\d+')` is `['5','67','890']`. The returned result of `*|SELECT regexp_extract_all('5a 67a 890m', '(\d+)a')` is `['5a','67a']`.
`regexp_extract_all(string, pattern, group)`	Returns an array where each element is a part of a substring that matches the regular expression. This part is the content in the group parameter value of the `()` of a substring that derives from the specified string.	The returned result of `*| SELECT regexp_extract_all('5a 67a 890m', '(\d+)a',1)` is `['5','67']`.
`regexp_extract(string, pattern)`	Returns the first substring of the specified string that matches the regular expression.	The returned result of `*|SELECT regexp_extract('5a 67b 890m', '\d+')` is `'5'`.

Function	Description	Example
`regexp_extract(string, pattern,group)`	Returns a part of the first substring that matches the regular expression. This part is the content in the group parameter value of the `()` of the substring that derives from the specified string.	The returned result of `*|SELECT regexp_extract('5a 67b 890m', '(\d+)([a-z]+)',2)` is `'a'`.
`regexp_like(string, pattern)`	Returns a Boolean value. If the string and its substrings cannot match the regular expression, the value False is returned.	The returned result of `*|SELECT regexp_like('5a 67b 890m', '\d+m')` is True.
`regexp_replace(string, pattern, replacement)`	Replaces the substrings of the specified string that match the regular expression with the value of the replacement parameter.	The returned result of `*|SELECT regexp_replace('5a 67b 890m', '\d+','a')` is `'aa ab am'`.
`regexp_replace(string, pattern)`	Removes the substrings of the specified string that match the regular expression. This function is equivalent to `regexp_replace(string,pat term,'')`.	The returned result of `*|SELECT regexp_replace('5a 67b 890m', '\d+')` is `'a b m'`.
`regexp_split(string, pattern)`	Returns an array where each element is a substring of the specified string that is split based on the regular expression.	The returned result of `*|SELECT regexp_split('5a 67b 890m', '\d+')` is `['a','b','m']`.

23.1.4.8.11. JSON functions

This topic describes the syntax of JSON functions. This topic also provides examples on how to use the functions.

> **Note**
> - If a string fails to be parsed into JSON data, null is returned.
> - In analytic statements of Log Service, a JSON array that is enclosed in single quotation marks ('') indicates a string.
> - If the value of a log field is of the JSON type and needs to be expanded to multiple rows, we recommend that you use UNNEST clauses. For more information, see UNNEST function.

json_parse() function

The json_parse() function is used to convert a string to JSON data. The returned result is of the JSON type.

- Syntax

```
json_parse(string)
```

- Example

  Convert the [1, 2, 3] string to the [1,2,3] JSON array.

```
* | SELECT json_parse('[1, 2, 3]')
```

The returned result is [1,2,3].

json_format()

The json_format() function is used to convert JSON data to a string. The returned result is a string.

- Syntax

```
json_format(json)
```

- Example

  Convert the [1,2,3] JSON array to the [1, 2, 3] string.

```
* | SELECT json_format(json_parse('[1, 2, 3]'))
```

  The returned result is [1,2,3].

json_array_contains()

The json_array_contains() function is used to check whether a JSON array or a JSON string contains a specified value. The returned result is true or false.

- Syntax

```
json_array_contains(json , value)
```

- Examples
  - Check whether the [1, 2, 3] JSON array contains 2.

```
* | SELECT json_array_contains(json_parse('[1, 2, 3]'), 2)
```

    The returned result is true.

  - Check whether the [1, 2, 3] JSON string contains 2.

```
* | SELECT json_array_contains('[1, 2, 3]', 2)
```

    The returned result is true.

json_array_get()

The json_array_get() function is used to extract the element that corresponds to the subscript of a JSON array.

- Syntax

```
json_array_get(json_array, index)
```

- Example

  Extract the element that corresponds to the subscript 0 of the ["status", "request_time", "request_method"] JSON array.

```
* | SELECT json_array_get('["status", "request_time", "request_method"]', 0)
```

  The returned result is status.

json_array_length()

The json_array_length() function is used to calculate the number of elements in a JSON array.

- Syntax

```
json_array_length(json array)
```

- Example

Calculate the number of the elements in the ["status", "request_time", "request_method"] JSON array.

```
* | SELECT json_array_length('["status", "request_time", "request_method"]')
```

The returned result is 3.

json_extract()

The json_extract() function is used to extract the value of a specified field from a JSON object. The returned result is of the JSON type.

> **Note** If the JSON data is invalid when you use the json_extract() function, an error message appears. We recommend that you use the json_extract_scalar() function.

- Syntax

```
json_extract(json, json_path)
```

The format of json_path is `$.store.book[0].title`.

- Examples
  - Extract the value of the status field from the content field. The content field is a JSON object.

    ```
    * | SELECT json_extract(content, '$.status')
    ```

    The returned result is the value of the status field, for example, **"200"**.
  - Expand the value of the request_time field and use row to represent the expanded rows. The value of the request_time field is a JSON array. Then, extract and calculate the sum of the values of the status field from the rows.

    ```
    * | select sum(cast (json_extract_scalar(row, '$.status') as bigint) ) from  log, unnest(cast(json_parse(request_time)  as array(json) ) as  t(row)
    ```

    The returned result is the sum result.

json_extract_scalar()

The json_extract_scalar() function is used to extract the value of a specified field from a JSON object. The returned result is a string.

- Syntax

```
json_extract_scalar(json, json_path)
```

The format of json_path is `$.store.book[0].title`.

- Examples
  - Extract the value of the status field from the content field. The content field is a JSON object.

    ```
    * | SELECT json_extract_scalar(content, '$.status')
    ```

    The returned result is the value of the status field, for example, **"200"**.
  - Extract the value of the status field from the content field. The content field is a JSON object. Then, convert the value to the bigint type and calculate the sum.

    ```
    * | select sum( cast (json_extract_scalar(content, '$.status') as bigint) )
    ```

    The returned result is the sum result.

json_size()

The json_size() function is used to calculate the number of elements in a JSON object or JSON array.

- Syntax

```
json_size(json,json_path)
```

- Example

  Calculate the number of elements in the status field.

```
* | SELECT json_size('{"status":[1, 2, 3]}','$.status')
```

  The returned result is 3.

23.1.4.8.12. Type conversion functions

Type conversion functions convert the data type of a specified value or column in a query.

You can use the index attribute feature of Log Service to set the data type of a field to LONG, DOUBLE, TEXT, or JSON. You can also query fields of various data types, including BIGINT, DOUBLE, VARCHAR, and TIMESTAMP. To query fields of a specific data type, you can use type conversion functions to convert the data type configured in an index into the data type that you use in a query.

Syntax

Note We recommend that you use the try_cast() function if a log contains dirty data. Otherwise, a query may fail due to the dirty data.

- Convert the data type of a column of values or a specific value into the specified type in a query. If the data type of a value fails to be converted, the query is terminated.

```
cast([key|value] AS type)
```

- Convert the data type of a column of values or a specific value into the specified type in a query. If the data type of a value fails to be converted, NULL is returned for the value, and the query continues.

```
try_cast([key|value] AS type)
```

Parameter	Description
key	The key of a field whose value data type is to be converted.
value	The field value whose data type is to be converted into the specified type.

Example

- To convert the numeric value 123 to a value of the VARCHAR type, use the following statement:

```
cast(123 AS varchar)
```

- To convert the data type of the uid field values to the VARCHAR type, use the following statement:

```
cast(uid AS varchar)
```

23.1.4.8.13. IP functions

IP functions can be used to identify whether an IP address is an internal or external IP address. IP functions can also be used to identify the country, state, and city to which an IP address belongs. This topic describes the syntax of IP functions and provides examples on how to use the functions.

Note The KEY parameter in the following functions indicates a log field, for example, client_ip. The value of this parameter is an IP address.

Function	Description	Example
ip_to_domain(KEY)	Checks whether an IP address is an internal IP address or an external IP address. The returned result is **intranet** or **internet**. • **intranet**: an internal IP address. • **internet**: an external IP address.	`* | SELECT ip_to_domain(client_ip)`
ip_to_country(KEY)	Identifies the country or the region to which an IP address belongs. The returned result is the Chinese name of a country or a region.	`* | SELECT ip_to_country(client_ip)`
ip_to_country(KEY,'en')	Identifies the country or the region to which an IP address belongs. The returned result is the code of a country or a region.	`* | SELECT ip_to_country(client_ip,'en')`
ip_to_country_code(KEY)	Identifies the country or the region to which an IP address belongs. The returned result is the code of a country or a region.	`* | SELECT ip_to_country_code(client_ip)`
ip_to_province(KEY)	Identifies the state to which an IP address belongs. The returned result is the Chinese name of a state.	`* | SELECT ip_to_province(client_ip)`
ip_to_province(KEY,'en')	Identifies the state to which an IP address belongs. The returned result is the administrative region code of a state.	`* | SELECT ip_to_province(client_ip,'en')`
ip_to_city(KEY)	Identifies the city to which an IP address belongs. The returned result is the Chinese name of a city.	`* | SELECT ip_to_city(client_ip)`
ip_to_city(KEY,'en')	Identifies the city to which an IP address belongs. The returned result is the administrative region code of a city.	`* | SELECT ip_to_city(client_ip,'en')`

Function	Description	Example
ip_to_geo(KEY)	Identifies the longitude and latitude of the location to which an IP address belongs. The returned result is in the `latitude,longitude` format. For information about geohash functions, see Geography functions.	`* | SELECT ip_to_geo(client_ip)`
ip_to_city_geo(KEY)	Identifies the longitude and latitude of the city to which an IP address belongs. This function returns the longitude and latitude of a city. Each city has only one set of coordinates. The returned result is in the `latitude,longitude` format.	`* | SELECT ip_to_city_geo(client_ip)`
ip_to_provider(KEY)	Identifies the Internet service provider (ISP) of an IP address. The returned result is the name of an ISP.	`* | SELECT ip_to_provider(client_ip)`

23.1.4.8.14. GROUP BY clause

GROUP BY clauses are used together with aggregate functions to group analysis results based on one or more columns.

Syntax

```
* | SELECT column name, aggregate function GROUP BY [ column name | alias | serial number ]
```

Note If you use a GROUP BY clause in an SQL statement, you can perform aggregate calculations on only a column that is specified in the GROUP BY clause or a random column that is not a non-GROUP BY column. For example, `* | SELECT status, request_time, COUNT(*) AS PV GROUP BY status` is an invalid query statement because request_time is not a GROUP BY column.

A GROUP BY clause can be used to group data by column name, alias, and serial number. The following table describes the related parameters.

Parameter	Description
Column name	The name of a log field or the return column of an aggregate function. You can group data by log field name (key) or the result of an aggregate function. A GROUP BY clause supports single column or multiple columns.
Alias	Data is grouped by the alias of a log field name or the return column alias of an aggregate function. You can specify the alias of a log field in the Field Search section of the Search & Analysis panel. For more information, see Field aliases.

Parameter	Description

Serial number	The serial number of a column in a SELECT statement. The number starts from 1.
	For example, the serial number of the status column is 1. In this case, the following two statements are equivalent:
	`* | SELECT status, count(*) AS PV GROUP BY status`
	`* | SELECT status, count(*) AS PV GROUP BY 1`

Aggregate function	A GROUP BY clause can be used together with aggregate functions such as MIN, MAX, AVG, SUM, and COUNT. For more information, see General aggregate functions.

Examples

- To calculate the number of access requests of different HTTP status codes, you can execute the following query statement:

```
* | SELECT status, count(*) AS PV GROUP BY status
```

- To calculate the number of page views (PVs) by 1 hour, you can execute the following query statement:

```
* | SELECT count(*) AS PV , date_trunc('hour', __time__) AS time GROUP BY time ORDER BY time limit 1000
```

The __time__ field is a reserved field in Log Service. This field indicates the time column. time is the alias of date_trunc('hour', __time__). For more information about the date_trunc() function, see Truncation function.

> **Note**
> - The clause limit 1000 indicates that a maximum of 1,000 rows of data can be returned. If you do not use a LIMIT clause, you can obtain a maximum of 100 rows of data by default.
> - If you turn on Enable Analytics for a log field in the Search & Analysis panel, the analysis feature is automatically enabled for the __time__ field.

- To calculate the number of PVs by 5 minutes, you can execute the following query statement.

  The date_trunc() function can only collect statistics at a specified interval. If you want to perform statistical analysis by custom time, you can group data based on the modulus method. In this example, %300 in the following statement indicates that the modulo and truncation operations are performed every 5 minutes.

```
* | SELECT count(*) AS PV, __time__ - __time__%300 AS time GROUP BY time limit 1000
```

- To extract a column that is not grouped by using a GROUP BY clause, you can execute the following query statement.

  If you use a GROUP BY clause in an SQL statement, you can perform aggregate calculations on only a column that is specified in the GROUP BY clause or a random column that is not a non-GROUP BY column. For example, `* | SELECT status, request_time, COUNT(*) AS PV GROUP BY status` is an invalid query statement because request_time is not a GROUP BY column.

```
* | SELECT status, arbitrary(request_time), count(*) AS PV GROUP BY status
```

23.1.4.8.15. Window functions

This topic describes the syntax of window functions.

Window functions are used to perform calculations across rows of a log. Common SQL aggregate functions calculate the results of only one row or aggregate all rows into one row for calculation. Window functions support cross-row calculation and fill the calculation results in each row.

Syntax of window functions:

```
SELECT key1, key2, value,
       rank() OVER (PARTITION BY key2
                    ORDER BY value DESC) AS rnk
FROM orders
ORDER BY key1,rnk
```

`rank() OVER (PARTITION BY KEY2 ORDER BY value DESC)` indicates that PARTITION BY is first used to partition by KEY2 and sort by the value if KEY2 is the same, and then the rank() function is used to aggregate data.

Special aggregate functions used in windows

Function	Description
rank()	Returns the rank of a value within a group of values. The rank is one plus the number of preceding rows that are not peers of the current row.
row_number()	Returns a unique, sequential number for each row.
first_value(x)	Returns the first value in the window. In most cases, the function is used to sort all values in a window and then return the maximum value.
last_value(x)	Returns the last value in the window. In most cases, the function is used to sort all values in a window and then return the minimum value.
nth_value(x, offset)	Returns the value at the specified offset from the beginning of the window.
lead(x,offset,defaut_value)	Returns the value in offset rows that follow the current row in the window. If the target row does not exist, the default_value is returned.
lag(x,offset,defaut_value)	Returns the value in offset rows that precede the current row in the window. If the target row does not exist, the default_value is returned.

Examples

- To rank the salaries of employees in their departments, execute the following query statement:

```
* | select department, persionId, sallary , rank() over(PARTITION BY department order by sallary de
sc) as sallary_rank  order by department,sallary_rank
```

Query and analysis result

department	persionId	sallary	sallary_rank
dev	john	9000	1
dev	Smith	8000	2
dev	Snow	7000	3

department	persionId	sallary	sallary_rank
dev	Achilles	6000	4
Marketing	Blan Stark	9000	1
Marketing	Rob Stark	8000	2
Marketing	Sansa Stark	7000	3

- To calculate the percentages of salaries of employees in their departments, execute the following query statement:

```
* | select department, persionId, sallary *1.0 / sum(sallary) over(PARTITION BY department ) as sal
lary_percentage
```

Query and analysis result

department	persionId	sallary	sallary_percentage
dev	john	9000	0.3
dev	Smith	8000	0.26
dev	Snow	7000	0.23
dev	Achilles	6000	0.2
Marketing	Blan Stark	9000	0.375
Marketing	Rob Stark	8000	0.333
Marketing	Sansa Stark	7000	0.29

- To calculate the daily UV increase over the previous day, execute the following query statement:

```
* | select day ,uv, uv *1.0 /(lag(uv,1,0) over() ) as diff_percentage from
(
select approx_distinct(ip) as uv, date_trunc('day',__time__) as day from log group by day order by
day asc
)
```

Query and analysis result

day	uv	diff_percentage
2017-12-01 00:00:00	100	null
2017-12-02 00:00:00	125	1.25
2017-12-03 00:00:00	150	1.2
2017-12-04 00:00:00	175	1.16
2017-12-05 00:00:00	200	1.14
2017-12-06 00:00:00	225	1.125
2017-12-07 00:00:00	250	1.11

23.1.4.8.16. HAVING clause

This topic describes the syntax of HAVING clauses.

The query and analysis feature of Log Service supports the standard SQL HAVING clause. A HAVING clause is used together with a GROUP BY clause to filter GROUP BY results.

The following example shows the syntax of a HAVING clause:

```
method :PostLogstoreLogs |select avg(latency),projectName group by projectName  having avg(latency) > 100
```

Difference between HAVING and WHERE clauses

A HAVING clause is used to filter the aggregation and calculation results after you use a GROUP BY clause. A WHERE clause is used to filter the raw data during aggregation.

Example

To calculate the average rainfall of each province in which the temperature is higher than 10°C, and return only the provinces in which the average rainfall is greater than 100 ml, execute the following query statement:

```
* | select avg(rain) ,province where temperature > 10 group by province having avg(rain) > 100
```

23.1.4.8.17. ORDER BY clause

This topic describes the syntax of ORDER BY clauses.

An ORDER BY clause is used to sort query results based on only one column.

- Syntax

  ```
  order by column name [desc|asc]
  ```

- Example

  ```
  method :PostLogstoreLogs |select avg(latency) as avg_latency,projectName group by projectName
  HAVING avg(latency) > 5700000
  order by avg_latency desc
  ```

23.1.4.8.18. LIMIT syntax

The LIMIT clause is used to limit the number of returned rows.

Syntax formats

Log Service supports the following LIMIT syntax formats:

- Reads only the first N rows:

  ```
  limit N
  ```

- Reads N rows starting from the S-th row:

  ```
  limit S , N
  ```

Note

- If you use the LIMIT clause to paginate results, the final results rather than the intermediate results of the SQL query are obtained.
- You cannot apply the LIMIT clause to subqueries. For example, the following statement is not supported:

```
* | select count(1) from ( select distinct(url) from limit 0,1000)
```

- If you use the LIMIT clause for pagination, the offset value cannot exceed 1,000,000. For example, in the `limit S , N` clause, the sum of S and N cannot exceed 1,000,000, and the value of N cannot exceed 10,000.

Example

- To obtain the first 100 rows of results, run the following statement.

```
* | select distinct(url) from log limit 100
```

- To obtain a total of 1,000 results from row 0 to row 999, run the following statement.

```
* | select distinct(url) from log limit 0,1000
```

- To obtain a total of 1,000 results from row 1,000 to row 1,999, run the following statement:

```
* | select distinct(url) from log limit 1000,1000
```

23.1.4.8.19. Conditional expressions

This topic describes the syntax of conditional expressions and provides examples on how to use conditional expressions.

CASE WHEN statement

CASE WHEN statements are used to classify data.

- Syntax

```
CASE WHEN condition1 THEN result1
     [WHEN condition2 THEN result2]
     [ELSE result3]
END
```

- Examples

- Extract browser information from the value of the http_user_agent field, classify the information into Chrome, Safari, and unknown types, and then calculate the number of page views (PVs) for the three types.
  - Query statement

```
* | SELECT CASE
 WHEN http_user_agent like '%Chrome%' then 'Chrome'
 WHEN http_user_agent like '%Safari%' then 'Safari'
 ELSE 'unknown'
 END AS http_user_agent,
    count(*) AS pv
    GROUP BY http_user_agent
```

  - Query and analysis result

http_user_agent	pv
Chrome	5563
Safari	1842
unknown	1666

- Query the distribution of requests that are sent at different points in time.
  - Query statement

```
* | SELECT
 CASE
 WHEN request_time < 10 then 't10'
 WHEN request_time < 100 then 't100'
 WHEN request_time < 1000 then 't1000'
 WHEN request_time < 10000 then 't10000'
 ELSE 'large' END
 AS request_time,
 count(*) AS pv
 GROUP BY request_time
```

  - Query and analysis result

request_time	pv
t100	1563542
large	533

if() function

The if() function is used to classify data. This function is similar to CASE WHEN statements.

- Syntax
  - If the *condition* is true, the *true_value* column is returned. Otherwise, null is returned.

```
if(condition, true_value)
```

  - If the *condition* is true, the *true_value* column is returned. Otherwise, the *false_value* column is returned.

```
if(condition, true_value, false_value)
```

- Example

  Calculate the ratio of requests whose status code is 200 to all requests.

- Query statement

```
* | SELECT sum(if(status =200,1,0))*1.0 / count(*) AS status_200_percentage
```

- Query and analysis result

status_200_percentage
0.8846858366766299

coalesce() function

The coalesce() function is used to return the first non-null value in multiple columns.

- Syntax

```
coalesce(expression1, expression2, expression3, expression4)
```

- Example

  Calculate the ratio of the expenses of yesterday to the expenses of the same day last month.

  - Query statement

```
* | SELECT compare("expenses of yesterday", 604800) AS diff FROM (SELECT coalesce(sum(PretaxAmount), 0) AS "expenses of yesterday" FROM website_log)
```

  - Query and analysis result

diff
[6514393413.0,19578267596.0,0.33273594719539659]

    - The value 6514393413.0 indicates the expenses of yesterday.
    - The value 19578267596.0 indicates the expenses of the same day last month.
    - The value 0.33273594719539659 indicates the ratio of the expenses of yesterday to the expenses of the same day last month.

nullif() function

The nullif() function is used to check whether the values of two columns are the same. If the values are the same, null is returned. Otherwise, the value of expression1 is returned.

- Syntax

```
nullif(expression1, expression2)
```

- Example

  Check whether the values of the client_ip and host fields are the same.

  - Query statement

```
* | SELECT nullif(client_ip,host)
```

- Query and analysis result

  If the values of the client_ip and host fields are different, the value of the client_ip field is returned.

try() function

The try() function is used to capture errors to ensure that Log Service can continue to query and analyze data.

- Syntax

```
try(expression)
```

- Example

  If an error occurs when the regexp_extract function is invoked, the try() function captures the error and Log Service continues to query and analyze data. The query and analysis result is returned.

  - Query statement

```
* | SELECT try(regexp_extract(request_uri, '.*\/(file.*)', 1)) AS file, count(*) AS count GROUP B
Y file
```

  - Query and analysis result

file	count
file-5	851
file-7	928
file-3	837
file-4	863

23.1.4.8.20. Nested subquery

This topic describes how to use nested subqueries when you query logs.

You can use nested queries to perform more complicated queries.

You must specify a FROM clause in the SQL statement of each nested query. However, this rule does not apply to non-nested queries. You must specify the `from log` keyword in each SQL statement to read raw data from logs.

Example:

```
* | select sum(pv) from
(
select count(1) as pv from log group by method
)
```

23.1.4.8.21. Array functions

This topic describes the syntax of array functions. This topic also provides examples on how to use the functions.

Function	Description
Subscript operator []	Obtains an element in an array.
array_distinct	Removes duplicate elements from an array and returns the distinct elements in the array.
array_intersect(x,y)	Returns the intersection of the x and y arrays.
array_union(x,y)	Returns the union of the x and y arrays. The return value is an array.
array_except(x,y)	Returns the subtraction of the x and y arrays. The return value is an array.
array_join(x,delimiter, null_replacement)	Concatenates the elements in an array by using a specified delimiter into a string and replaces null elements with the value of null_replacement. The return value is of the varchar type. **Note** The maximum size of the result of the array_join function is 1 KB. If the returned result exceeds 1 KB, the excess data is truncated.
array_max(x)	Returns the maximum value in the x array.
array_min(x)	Returns the minimum value in the x array.
array_position(x,element)	Returns the subscript of the element in the x array. The subscript starts from 1. If no subscript exists, the value 0 is returned. The return value is of the bigint type.
array_remove(x,element)	Removes a specified element from an array. The return value is an array.
array_sort(x)	Sorts the elements in an array and moves null elements to the end of the array. The return value is of the bigint type.
cardinality(x)	Returns the number of the elements in an array. The return value is of the bigint type.
concat(array1, array2, ..., arrayN)	Concatenates arrays. The return value is an array.
contains(x, element)	Returns true if the x array contains a specified element. The return value is of the Boolean type.
filter(array, function)	A Lambda function. For more information, see filter() function. The return value is an array.
flatten(x)	Concatenates the arrays in a two-dimensional array into a one-dimensional array. The return value is an array.
reduce(array, initialState, inputFunction, outputFunction)	A Lambda function. For more information, see reduce() function.

Function	Description
reverse(x)	Reverses the elements in the x array. The return value is an array.
sequence(start, stop)	Generates a sequence of elements from start to stop with a step of 1. The return value is an array.
sequence(start, stop, step)	Generates a sequence of elements from start to stop with a specified step. The return value is an array.
sequence(start, stop, step)	Generates a sequence of timestamps from start to stop with a specified step. The value of the step parameter is of the interval type. The format can be INTERVAL DAY TO SECOND or INTERVAL YEAR TO MONTH. The return value is an array.
shuffle(x)	Shuffles an array. The return value is an array.
slice(x,start, length)	Returns a subset of the x array from the specified start position with the specified length. The return value is an array.
transform(array, function)	A Lambda function. For more information, see transform() function. The return value is an array.
zip(array1, array2[, ...])	Merges multiple arrays. The Nth field of the Mth element in the returned array is the Mth element of the Nth array that is specified in the function. The return value is an array. Example: `SELECT zip(ARRAY[1, 2], ARRAY['1b', null, '3b']); — [ROW(1, '1b'), ROW(2, null), ROW(null, '3b')]`.
zip_with(array1, array2, function)	A Lambda function. For more information, see zip_with() function. The return value is an array.
array_agg (key)	An aggregate function that returns an array. The returned array contains the values in the key column. Example: `* | select array_agg(key)`.
array_transpose(array[array[*x*,*y*,*z*], array[*a*,*b*,*c*]])	Transposes the values of a matrix from rows to columns.

23.1.4.8.22. Binary string functions

This topic describes the syntax of binary string functions. This topic also provides examples on how to use the functions.

Varbinary data is different from varchar data.

Function	Description
Concatenation operator (`||`)	The result of `a || b` is `ab`.
length(binary)	Returns the length of a binary string in bytes. The return value is of the bigint type.

Function	Description
concat(binary1, ..., binaryN)	Concatenates binary strings. This function is equivalent to ‖. The return value is of the varbinary type.
to_base64(binary)	Converts a binary string to a Base64 string. The return value is of the varchar type.
from_base64(string)	Converts a Base64 string to a binary string. The return value is of the varbinary type.
to_base64url(binary)	Converts a string to a URL-safe Base64 string. The return value is of the varchar type.
from_base64url(string)	Converts a URL-safe Base64 string to a binary string. The return value is of the varbinary type.
to_hex(binary)	Converts a binary string to a hexadecimal string. The return value is of the varchar type.
from_hex(string)	Converts a hexadecimal string to a binary string. The return value is of the varbinary type.
to_big_endian_64(bigint)	Converts a number to a binary string in big endian mode. The return value is of the varbinary type.
from_big_endian_64(binary)	Converts a binary string in big endian mode to a number. The return value is of the bigint type.
md5(binary)	Calculates the MD5 value of a binary string. The return value is of the varbinary type.
sha1(binary)	Calculates the SHA1 value of a binary string. The return value is of the varbinary type.
sha256(binary)	Calculates the SHA256 hash value of a binary string. The return value is of the varbinary type.
sha512(binary)	Calculate the SHA512 value of a binary string. The return value is of the varbinary type.
xxhash64(binary)	Calculates the xxhash64 value of a binary string. The return value is of the varbinary type.

23.1.4.8.23. Bitwise functions

This topic describes the syntax of bitwise functions. This topic also provides examples on how to use the functions.

Function	Description	Example
bit_count(x, bits)	Counts the number of 1s in x in two's complement. The x variable is a signed integer that includes the specified number of bits. The return value is of the bigint type.	• `SELECT bit_count(9, 64)` returns 2. • `SELECT bit_count(9, 8)` returns 2. • `SELECT bit_count(-7, 64)` returns 62. • `SELECT bit_count(-7, 8)` returns 6.

Function	Description	Example
bitwise_and(x, y)	Returns the bitwise AND of x and y in two's complement. The return value is of the bigint type.	None
bitwise_not(x)	Returns the bitwise NOT of x in two's complement. The return value is of the bigint type.	None
bitwise_or(x, y)	Returns the bitwise OR of x and y in two's complement. The return value is of the bigint type.	None
bitwise_xor(x, y)	Returns the bitwise XOR of x and y in two's complement. The return value is of the bigint type.	None

23.1.4.8.24. Interval-valued comparison and periodicity-valued comparison functions

Log Service supports interval-valued comparison and periodicity-valued comparison functions. You can use these functions to query and analyze log data.

compare() function

The compare() function is used to compare the calculation result of the current time period with the calculation result of a time period N seconds before. You can use this function to perform an interval-valued comparison or periodicity-valued comparison on data.

- Syntax

```
compare(column name,N)
```

> **Note** The compare() function can be used to compare the calculation results of multiple periods of time, for example, compare(*column name*,N1,*N2*,*N3*).

  - column name: the name of the specified column. The value of this parameter must be of the double type or long type.
  - N: the time window. Unit: seconds. Example: 3600 (1 hour), 86400 (one day), or 604800 (one week).

- Response

  The returned result is a JSON array in the following format: [the current value, the value before N seconds, the ratio of the current value to the value of N seconds before, the UNIX timestamp before N seconds]. Example: [1176.0,1180.0,0.9966101694915255,1611504000.0].

- Examples

- Calculate the ratio of the page views (PVs) of the current hour to the PVs of the same time period the day before.

  Set the time range to **1 Hour(Time Frame)** and execute the following query statement. 86400 indicates the current time minus 86400 seconds (one day). log indicates the Logstore name.

  ```
  * | SELECT compare(PV, 86400) FROM (SELECT count(*) AS PV FROM log)
  ```

  The following figure shows the returned result.

_col0
[3337.0,3522.0,0.947473026689381]

  - **3337.0** indicates the PVs of the current 1 hour, for example, Dec 25, 2020, 14:00:00 ~ Dec 25, 2020, 15:00:00.
  - **3522.0** indicates the PVs of the same time period the day before, for example, Dec 24, 2020, 14:00:00 ~ Dec 24, 2020, 15:00:00.
  - **0.947473026689381** indicates the ratio of the PVs of the current hour to the PVs of the same time period the day before.

  To display the analysis result in multiple columns, you can execute the following query statement:

  ```
  * | SELECT diff[1] AS today, diff[2] AS yesterday, diff[3] AS ratio FROM (SELECT compare(PV,86400
  ) AS diff FROM (SELECT count(*) AS PV FROM log))
  ```

  The following figure shows the returned result.

today	yesterday	ratio
3337.0	3522.0	0.947473026689381

  Note

  To compare the data of a specified year or week with the data of the previous year or week, you can use the query statements in the following examples:

  - For example, if you want to calculate the ratio of the PVs of November 2020 to the PVs of November 2019, you can set the time range to **Nov 1, 2020, 00:00~Dec 1, 2020, 00:00**, and execute the following query statement:

    ```
    * | SELECT compare(PV, 31622400) FROM (SELECT count(*) AS PV FROM log)
    ```

  - For example, if you want to calculate the ratio of the PVs of a Tuesday to the PVs of the previous Tuesday, you can set the time range to **Jan 18, 2021, 00:00~Jan 19, 2021, 00:00**, and execute the following query statement:

    ```
    * | SELECT compare(PV, 604800) FROM (SELECT count(*) AS PV FROM log)
    ```

- Calculate the ratio of the PVs of each hour of the current day to the PVs of the same time period the day before and two days before.

  Set the time range to **Today(Time Frame)** and execute the following query statement. 86400 indicates the current time minus 86400 seconds (one day). 172800 indicates the current time minus 172800 seconds (two days). log indicates the Logstore name. date_format(from_unixtime(__time__), '%H:00') indicates the format of the returned time.

```
* | SELECT time, compare(PV, 86400,172800) as diff from (SELECT count(*) as PV, date_format(from_unixtime(__time__), '%H:00') as time from log GROUP BY time) GROUP BY time ORDER BY time
```

  The following figure shows the returned result.

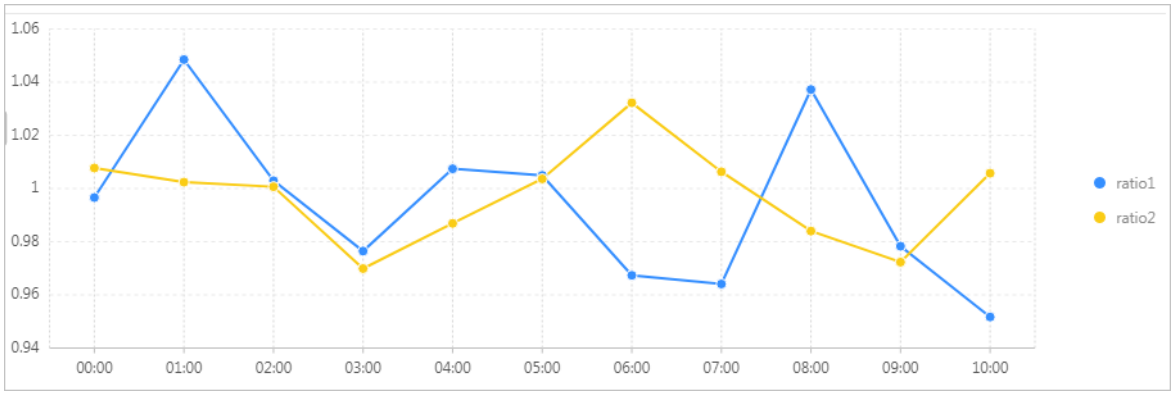
time	diff
00:00	[1176.0,1180.0,1167.0,0.9966101694915255,1.0077120822622108]
01:00	[10077.0,9611.0,10053.0,1.0484861096660077,1.0023873470605789]
02:00	[26921.0,26842.0,26903.0,1.002943148796662,1.0006690703638999]

  - **1176.0** indicates the PVs of the current time period, for example, Dec 25, 2020, 00:00 ~ Dec 25, 2020, 01:00.
  - **1180** indicates the PVs of the same time period the day before, for example, Dec 24, 2020, 00:00 ~ Dec 24, 2020, 01:00.
  - **1167.0** indicates the PVs of the same time period two days before, for example, Dec 23, 2020, 00:00:00 ~ Dec 23, 2020, 01:00:00.
  - **0.9966101694915255** indicates the ratio of the PVs of the current time period to the PVs of the same time period the day before.
  - **1.0077120822622108** indicates the ratio of the PVs of the current period to the PVs of the same period two days before.

  To display the analysis result in multiple columns, you can execute the following query statement:

```
* | SELECT time, diff[1] AS day1, diff[2] AS day2, diff[3] AS day3, diff[4] AS ratio1,  diff[5] AS ratio2 FROM (SELECT time, compare(PV, 86400,172800) as diff from (SELECT count(*) as PV, date_format(from_unixtime(__time__), '%H:00') as time from log GROUP BY time) GROUP BY time ORDER BY time)
```

  The following figure shows the returned result.



- Calculate the ratio of the PVs of December to the PVs of November in the same year.

  Set the time range to **This Month(Time Frame)** and execute the following query statement. 2592000 indicates the current time minus 2592000 seconds (one month). log indicates the Logstore name. date_trunc('month', __time__) indicates that the date_trunc function is used to truncate a point in time by month.

  ```
  *| SELECT time, compare(PV, 2592000) AS diff from (SELECT count(*) AS PV, date_trunc('month', __time__) AS time from log GROUP BY time) GROUP BY time ORDER BY time
  ```

  The following figure shows the returned result.

time	diff
2021-01-01 00:00:00.000	[11958378.0,448571.0,26.658829928818404]

ts_compare() function

The ts_compare() function is used to compare the calculation result of the current time period with the calculation result of a time period N seconds before. You can use this function to perform an interval-valued comparison or periodicity-valued comparison on data. The analysis results of the ts_compare() function must be grouped by the time column by using GROUP BY clauses.

- Syntax

  ```
  ts_compare(column name,N)
  ```

  > **Note** The ts_compare() function can be used to compare the calculation results of multiple periods of time, for example, ts_compare(*column name*,N1,*N2*,*N3*).

  - column name: the name of the specified column. The value of this parameter must be of the double type or long type.
  - N: the time window. Unit: seconds. Example: 3600 (1 hour), 86400 (one day), or 604800 (one week).

- Response

  The returned result is a JSON array in the following format: [the current value, the value before N seconds, the ratio of the current value to the value of N seconds before, the UNIX timestamp before N seconds]. Example: [1176.0,1180.0,0.9966101694915255,1611504000.0].

- Example

  Calculate the ratio of the PVs of every hour today to the PVs of the previous hour.

  Set the time range to **Today(Relative)** and execute the following query statement. 3600 indicates the current time minus 3600 seconds (1 hour). log indicates the Logstore name. date_trunc('hour',__time__) indicates that the date_trunc function is used to truncate the time by hour.

  ```
  * | SELECT time, ts_compare(PV, 3600) AS data FROM(SELECT date_trunc('hour',__time__ ) AS time, count(*) AS PV from log GROUP BY time ORDER BY time ) GROUP BY time
  ```

  The following figure shows the returned result.

time	data
2021-01-27 00:00:00.000	[1160.0,10034.0,0.11560693641618497,1611673200.0]
2021-01-27 01:00:00.000	[10177.0,1160.0,8.773275862068966,1611676800.0]
2021-01-27 02:00:00.000	[26804.0,10177.0,2.6337820575808195,1611680400.0]

23.1.4.8.25. Comparison functions and operators

This topic describes the comparison functions and operators in Log Service. You can use these functions and operators to query and analyze log data.

A comparison function compares the values of two parameters. The values can be one of the arbitrary comparable data types, such as integer, bigint, double, and text.

Comparison operators

A comparison operator is used to compare two values. If the statement is true, TRUE is returned. Otherwise, FALSE is returned.

Operator	Description
<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
=	Equal to
<>	Not equal to
!=	Not equal to

Range operator BETWEEN

The BETWEEN operator is used to check whether a value falls in a specified closed interval.

- If the value falls in the specified closed interval, TRUE is returned. Otherwise, FALSE is returned.

  Example: `SELECT 3 BETWEEN 2 AND 6;`. The statement is true, and TRUE is returned.

  The preceding statement is equivalent to `SELECT 3 >= 2 AND 3 <= 6;`.
- The BETWEEN operator can be specified after the NOT operator to check whether a value falls out of a specified closed interval.

  Example: `SELECT 3 NOT BETWEEN 2 AND 6;`. The statement is false, and FALSE is returned.

  The preceding statement is equivalent to `SELECT 3 < 2 OR 3 > 6;`.
- If one of the three values is NULL, the result is NULL.

IS NULL and IS NOT NULL

The IS NULL and IS NOT NULL operators are used to check whether a value is NULL.

IS DISTINCT FROM and IS NOT DISTINCT FROM

The IS DISTINCT FROM and IS NOT DISTINCT FROM operators are similar to the EQUAL TO and NOT EQUAL TO operators. The difference is that the IS DISTINCT FROM and IS NOT DISTINCT FROM operators can be used to check whether a NULL value exists.

Examples:

```
SELECT NULL IS DISTINCT FROM NULL; -- false
SELECT NULL IS NOT DISTINCT FROM NULL; -- true
```

You can use the DISTINCT operator to compare parameter values under multiple conditions. The following table describes the conditions.

a	b	a = b	a <> b	a DISTINCT b	a NOT DISTINCT b
1	1	TRUE	FALSE	FALSE	TRUE
1	2	FALSE	TRUE	TRUE	FALSE
1	NULL	NULL	NULL	TRUE	FALSE
NULL	NULL	NULL	NULL	FALSE	TRUE

GREATEST and LEAST

The GREATEST operator is used to obtain the maximum value from multiple columns. The LEAST operator is used to obtain the minimum value from multiple columns.

Example:

```
select greatest(1,2,3) ; -- Returns 3.
```

Quantified comparison predicates: ALL, ANY, and SOME

The ALL, ANY, and SOME quantifiers are used to check whether a parameter value meets specified conditions.

- ALL is used to check whether a parameter value meets all conditions. If the statement is true, TRUE is returned. Otherwise, FALSE is returned.
- ANY is used to check whether a parameter value meets one of the specified conditions. If the statement is true, TRUE is returned. Otherwise, FALSE is returned.
- SOME is used to check whether a parameter value meets one of the specified conditions. SOME is equivalent to ANY.
- ALL, ANY, and SOME must be specified after comparison operators.

You can use ALL and ANY to compare values under multiple conditions. The following table describes the conditions.

Expression	Description
A = ALL (...)	Returns TRUE if A matches all values.
A <> ALL (...)	Returns TRUE if A does not match all values.
A < ALL (...)	Returns TRUE if A is smaller than the smallest value.
A = ANY (...)	Returns TRUE if A is equal to a value. This statement is equivalent to A IN (...).
A <> ANY (...)	Returns TRUE if A does not match a value.
A < ANY (.....)	Returns TRUE if A is smaller than the largest value.

Examples:

```
SELECT 'hello' = ANY (VALUES 'hello', 'world'); -- true
SELECT 21 < ALL (VALUES 19, 20, 21); -- false
SELECT 42 >= SOME (SELECT 41 UNION ALL SELECT 42 UNION ALL SELECT 43); -- true
```

23.1.4.8.26. Lambda functions

This topic describes Lambda functions and provides some examples. You can use Lambda functions to analyze log data in Log Service

Lambda expressions

Lambda expressions use the arrow operator `->`.

Examples:

```
x -> x + 1
(x, y) -> x + y
x -> regexp_like(x, 'a+')
x -> x[1] / x[2]
x -> IF(x > 0, x, -x)
x -> COALESCE(x, 0)
x -> CAST(x AS JSON)
x -> x + TRY(1 / 0)
```

Most MySQL expressions can be used in Lambda functions.

filter(array<T>, function<T, boolean>) → ARRAY<T>

Returns an array whose elements are filtered from the specified array based on the Lambda expression.

Examples:

```
SELECT filter(ARRAY [], x -> true); -- []
SELECT filter(ARRAY [5, -6, NULL, 7], x -> x > 0); -- [5, 7]
SELECT filter(ARRAY [5, NULL, 7, NULL], x -> x IS NOT NULL); -- [5, 7]
```

map_filter(map<K, V>, function<K, V, boolean>) → MAP<K,V>

Returns a map whose elements are filtered based on the Lambda expression. The map is generated from the map function.

Examples:

```
SELECT map_filter(MAP(ARRAY[], ARRAY[]), (k, v) -> true); -- {}
SELECT map_filter(MAP(ARRAY[10, 20, 30], ARRAY['a', NULL, 'c']), (k, v) -> v IS NOT NULL); -- {10 -> a
, 30 -> c}
SELECT map_filter(MAP(ARRAY['k1', 'k2', 'k3'], ARRAY[20, 3, 15]), (k, v) -> v > 10); -- {k1 -> 20, k3
-> 15}
```

reduce(array<T>, initialState S, inputFunction<S, T, S>, outputFunction<S, R>) → R

The reduce function starts from the initial state, traverses each element in the array, and then calls inputFunction(S,T) to generate a new state. After all the elements in the array are traversed and the final state is generated, the reduce function calls outputFunction to assign the final state value to the result R and output the result. The procedure is described as follows:

1. Start from the initial state S.
2. Traverse each element T.
3. Calculate inputFunction(S,T) to generate a new state S.
4. Repeat steps 2 and 3 until the last element is traversed and has a new state.
5. Turn the final state S into the final result R.

Examples:

```
SELECT reduce(ARRAY [], 0, (s, x) -> s + x, s -> s); -- 0
SELECT reduce(ARRAY [5, 20, 50], 0, (s, x) -> s + x, s -> s); -- 75
SELECT reduce(ARRAY [5, 20, NULL, 50], 0, (s, x) -> s + x, s -> s); -- NULL
SELECT reduce(ARRAY [5, 20, NULL, 50], 0, (s, x) -> s + COALESCE(x, 0), s -> s); -- 75
SELECT reduce(ARRAY [5, 20, NULL, 50], 0, (s, x) -> IF(x IS NULL, s, s + x), s -> s); -- 75
SELECT reduce(ARRAY [2147483647, 1], CAST (0 AS BIGINT), (s, x) -> s + x, s -> s); -- 2147483648
SELECT reduce(ARRAY [5, 6, 10, 20], -- calculates arithmetic average: 10.25
             CAST(ROW(0.0, 0) AS ROW(sum DOUBLE, count INTEGER)),
             (s, x) -> CAST(ROW(x + s.sum, s.count + 1) AS ROW(sum DOUBLE, count INTEGER)),
             s -> IF(s.count = 0, NULL, s.sum / s.count));
```

transform(array<T>, function<T, U>) → ARRAY<U>

This Lambda function traverses each element in an array to generate a new result U.

Examples:

```
SELECT transform(ARRAY [], x -> x + 1); -- []
SELECT transform(ARRAY [5, 6], x -> x + 1); -- [6, 7] -- Increments each element by 1.
SELECT transform(ARRAY [5, NULL, 6], x -> COALESCE(x, 0) + 1); -- [6, 1, 7]
SELECT transform(ARRAY ['x', 'abc', 'z'], x -> x || '0'); -- ['x0', 'abc0', 'z0']
SELECT transform(ARRAY [ARRAY [1, NULL, 2], ARRAY[3, NULL]], a -> filter(a, x -> x IS NOT NULL)); -- [
[1, 2], [3]]
```

zip_with(array<T>, array<U>, function<T, U, R>) → array<R>

This Lambda function merges two arrays and generates the element R in the new array based on element T and element U.

Examples:

```
SELECT zip_with(ARRAY[1, 3, 5], ARRAY['a', 'b', 'c'], (x, y) -> (y, x)) --Transposes the elements of t
he two arrays to generate a new array. Result: [['a', 1], ['b', 3],['c', 5]]
SELECT zip_with(ARRAY[1, 2], ARRAY[3, 4], (x, y) -> x + y); -- Result: [4, 6]
SELECT zip_with(ARRAY['a', 'b', 'c'], ARRAY['d', 'e', 'f'], (x, y) -> concat(x, y)) -- Concatenates th
e elements of the two arrays to generate a new string. Result: ['ad', 'be', 'cf']
```

23.1.4.8.27. Logical functions

This topic describes the available logical functions in Log Service. You can use these functions to query and analyze log data.

Logical operators

Operator	Description	Example
AND	The result is TRUE if both values are TRUE.	a AND b
OR	The result is TRUE if either value is TRUE.	a OR b
NOT	The result is TRUE if the value is FALSE.	NOT a

Effect of NULL on logical operators

The following tables list the truth values when the values of a and b are TRUE, FALSE, and NULL, respectively.

Truth table 1

a	b	a AND b	a OR b
TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	FALSE	TRUE
TRUE	NULL	NULL	TRUE
FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE
FALSE	NULL	FALSE	NULL
NULL	TRUE	NULL	TRUE
NULL	FALSE	FALSE	NULL
NULL	NULL	NULL	NULL

Truth table 2

a	NOT a
TRUE	FALSE
FALSE	TRUE
NULL	NULL

23.1.4.8.28. Column aliases

This topic describes how to specify an alias for a column and provides some examples.

A column name in an SQL statement can contain only letters, digits, and underscores (_). The column name must start with a letter.

When you configure log collection, you may specify a column name that does not conform to the SQL standard, for example, User-Agent. In this case, you must specify an alias for the column in the Search & Analysis panel in which you can configure index attributes. The alias is used only if you execute an SQL statement to query and analyze logs. The original name of each column is stored. Therefore, you must search for columns by original name.

If the original name of a column is long, you can specify an alias for the column in an SQL statement.

Sample aliases

Original column name	Alias
User-Agent	ua
User.Agent	ua
123	col
abceefghijklmnopqrstuvw	a

23.1.4.8.29. Use a JOIN clause to query data from a Logstore and an RDS database

This topic describes how to use a JOIN clause to query data from a Logstore and an RDS database at the same time. This topic also describes how to save the query result to the RDS database.

Procedure

1. Create an RDS instance and a virtual private cloud (VPC), and configure a whitelist. For more information, see **ApsaraDB RDS for SQL Server > Quick start > Create an instance and configure a whitelist**.

   Specify a virtual private cloud (VPC) for the RDS instance. After you create an RDS instance, you can obtain the VPC ID and the RDS instance ID.

2. Configure a whitelist for the RDS instance.

   Add the following CIDR blocks to the whitelist: `100.104.0.0/16`, `11.194.0.0/16`, and `11.201.0.0/16`.

3. Create an external store.

   Run the following script to create an external store. Replace the parameter values based on your business requirements.

```
{
"externalStoreName":"storeName",
"storeType":"rds-vpc",
"parameter":
   {
   "region":"cn-qingdao",
   "vpc-id":"vpc-m5eq4irc1pucp*******"
   "instance-id":"i-m5eeo2whsn*******"
   "host":"localhost",
   "port":"3306",
   "username":"root",
   "password":"****",
   "db":"scmc"
   "table":"join_meta"
   }
}
```

   Parameters

Parameter	Description
region	The region where your RDS instance resides.
vpc-id	The ID of the VPC.
instance-id	The ID of the RDS instance.
host	The ID of the Elastic Compute Service (ECS) instance.
port	The port of the ECS instance.
username	The username.
password	The password.
db	The name of the database.
table	The name of the table.

> **Note** You can join a Logstore with an RDS table that resides only in one of the following regions: China (Beijing), China (Qingdao), and China (Hangzhou).

4. Use a JOIN clause to query data.

   Log on to the Log Service console. On the **Search & Analysis** page of the Logstore in which you want to query data, execute a query statement that includes a JOIN clause.

   Log Service supports the following syntax of JOIN clauses:

   - INNER JOIN
   - LEFT JOIN
   - RIGHT JOIN
   - FULL JOIN

   ```
   [ INNER ] JOIN
   LEFT [ OUTER ] JOIN
   RIGHT [ OUTER ] JOIN
   FULL [ OUTER ] JOIN
   ```

   > **Note**
   > - JOIN clauses apply only to Logstores and small tables.
   > - In a JOIN clause, you must write the name of a Logstore before the join parameter and the name of an external store after the join parameter.
   > - In a JOIN clause, you must specify the name of an external store. Log Service automatically replaces this name with a name that combines an RDS database name and a table name. You cannot specify only the name of an RDS table.

   The following example shows a query statement that includes a JOIN clause:

   ```
   method:postlogstorelogs | select count(1) , histogram(logstore) from log  l join join_meta m on l.projectid = cast( m.ikey as varchar)
   ```

5. Save the query result to the RDS database.

   Log Service allows you to use an INSERT statement to insert query results into an RDS database. The following example shows an INSERT statement:

   ```
   method:postlogstorelogs | insert into method_output  select cast(methodasvarchar(65535)),count(1)fromloggroupbymethod
   ```

Sample Python script

```
# encoding: utf-8
from __future__ import print_function
from aliyun.log import *
from aliyun.log.util import base64_encodestring
from random import randint
import time
import os
from datetime import datetime
    endpoint = os.environ.get('ALIYUN_LOG_SAMPLE_ENDPOINT', 'cn-chengdu.log.aliyuncs.com')
    accessKeyId = os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSID', '')
    accessKey = os.environ.get('ALIYUN_LOG_SAMPLE_ACCESSKEY', '')
    logstore = os.environ.get('ALIYUN_LOG_SAMPLE_LOGSTORE', '')
    project = "ali-yunlei-chengdu"
    client = LogClient(endpoint, accessKeyId, accessKey, token)
# Create an external store.
    res = client.create_external_store(project,ExternalStoreConfig("rds_store","region","rds-vpc","vpc
id","instance id","instance ip","instance port","username","password","database","table"));
    res.log_print()
    # Obtain the details of the external store.
    res = client.get_external_store(project,"rds_store");
    res.log_print()
    res = client.list_external_store(project,"");
    res.log_print();
    # Execute a query statement that includes a JOIN clause.
    req = GetLogStoreLogsRequest(project,logstore,From,To,"","select count(1) from  "+ logstore +"  s
join  meta m on  s.projectid = cast(m.ikey as varchar)");
    res = client.get_logs(req)
    res.log_print();
     # Write the query result to the RDS database.
    req = GetLogStoreLogsRequest(project,logstore,From,To,"," insert into rds_store select count(1) f
rom  "+ logstore );
    res = client.get_logs(req)
    res.log_print();
```

23.1.4.8.30. Geospatial functions

This topic describes the available geospatial functions in Log Service. You can use these functions to query and analyze log data.

Concept of geometry

Geospatial functions support geometries in the well-known text (WKT) format.

Geometry formats

Geometry	WKT format
Point	`POINT (0 0)`
LineString	`LINESTRING (0 0, 1 1, 1 2)`
Polygon	`POLYGON ((0 0, 4 0, 4 4, 0 4, 0 0), (1 1, 2 1, 2 2, 1 2, 1 1))`
MultiPoint	`MULTIPOINT (0 0, 1 2)`
MultiLineString	`MULTILINESTRING ((0 0, 1 1, 1 2), (2 3, 3 2, 5 4))`

Geometry	WKT format
MultiPolygon	`MULTIPOLYGON (((0 0, 4 0, 4 4, 0 4, 0 0), (1 1, 2 1, 2 2, 1 2, 1 1)), ((-1 -1, -1 -2,`
GeometryCollection	`GEOMETRYCOLLECTION (POINT(2 3), LINESTRING (2 3, 3 4))`

Constructors

Constructor description

Function	Description
ST_Point(double, double) → Point	Returns a geometry point instance with the specified coordinate values.
ST_LineFromText(varchar) → LineString	Returns a geometry LineString instance from a WKT representation.
ST_Polygon(varchar) → Polygon	Returns a geometry polygon instance from a WKT representation.
ST_GeometryFromText(varchar) → Geometry	Returns a geometry instance from a WKT representation.
ST_AsText(Geometry) → varchar	Returns the WKT representation of a geometry.

Operations

Function	Description
ST_Boundary(Geometry) → Geometry	Returns the closure of the combinatorial boundary of a geometry.
ST_Buffer(Geometry, distance) → Geometry	Returns the geometry that represents all points whose distance from the specified geometry is shorter than or equal to the specified distance.
ST_Difference(Geometry, Geometry) → Geometry	Returns the geometry value that represents the point set difference of the specified geometries.
ST_Envelope(Geometry) → Geometry	Returns the bounding rectangular polygon of a geometry.
ST_ExteriorRing(Geometry) → Geometry	Returns a line string that represents the exterior ring of the input polygon.
ST_Intersection(Geometry, Geometry) → Geometry	Returns the geometry value that represents the point set intersection of two geometries.
ST_SymDifference(Geometry, Geometry) → Geometry	Returns the geometry value that represents the point set symmetric difference of two geometries.

Relationship tests

Function	Description

Function	Description
ST_Contains(Geometry, Geometry) → boolean	Returns True if and only if no points of the second geometry lie in the exterior of the first geometry, and at least one point of the interior of the first geometry lies in the interior of the second geometry. Returns False if points of the second geometry are on the boundary of the first geometry.
ST_Crosses(Geometry, Geometry) → boolean	Returns True if the specified geometries share some, but not all, interior points in common.
ST_Disjoint(Geometry, Geometry) → boolean	Returns True if the specified geometries do not spatially intersect.
ST_Equals(Geometry, Geometry) → boolean	Returns True if the specified geometries represent the same geometry.
ST_Intersects(Geometry, Geometry) → boolean	Returns True if the specified geometries spatially intersect in two dimensions.
ST_Overlaps(Geometry, Geometry) → boolean	Returns True if the specified geometries share space in the same dimension, but are not completely contained by each other.
ST_Relate(Geometry, Geometry) → boolean	Returns True if the first geometry is spatially related to the second geometry.
ST_Touches(Geometry, Geometry) → boolean	Returns True if the specified geometries have at least one point in common, but their interiors do not intersect.
ST_Within(Geometry, Geometry) → boolean	Returns True if the first geometry is completely inside the second geometry. Returns False if the two geometries have points in common at the boundaries.

Accessors

Function	Description
ST_Area(Geometry) → double	Returns the two-dimensional Euclidean area of a geometry.
ST_Centroid(Geometry) → Geometry	Returns the point value that is the mathematical centroid of a geometry.
ST_CoordDim(Geometry) → bigint	Returns the coordinate dimension of a geometry.
ST_Dimension(Geometry) → bigint	Returns the inherent dimension of a geometry object, which must be less than or equal to the coordinate dimension.
ST_Distance(Geometry, Geometry) → double	Returns the minimum two-dimensional Cartesian distance (based on spatial ref) between two geometries in projected units.
ST_IsClosed(Geometry) → boolean	Returns True if the start and end points of the linestring are coincident.

Function	Description
ST_IsEmpty(Geometry) → boolean	Returns True if the specified geometry is an empty geometry, such as geometry collection, polygon, and point.
ST_IsRing(Geometry) → boolean	Returns True if and only if the line is closed and simple.
ST_Length(Geometry) → double	Returns the length of a LineString or multi-LineString by using Euclidean measurement on a two-dimensional plane (based on spatial ref) in projected units.
ST_XMax(Geometry) → double	Returns the X maximum of the bounding box of the geometry.
ST_YMax(Geometry) → double	Returns the Y maximum of the bounding box of the geometry.
T_XMin(Geometry) → double	Returns the X minimum of the bounding box of the geometry.
ST_YMin(Geometry) → double	Returns the Y minimum of the bounding box of the geometry.
ST_StartPoint(Geometry) → point	Returns the first point of a geometry LineString instance.
ST_EndPoint(Geometry) → point	Returns the last point of a geometry LineString instance.
ST_X(Point) → double	Returns the X coordinate of a point.
ST_Y(Point) → double	Returns the Y coordinate of a point.
ST_NumPoints(Geometry) → bigint	Returns the number of points in a geometry.
ST_NumInteriorRing(Geometry) → bigint	Returns the cardinality of the collection of interior rings of a polygon.

23.1.4.8.31. Geography functions

This topic describes the syntax of geography functions and provides some examples.

IP functions identify the country, province, city, Internet service provider (ISP), and longitude and latitude of a specific IP address. For more information, see IP functions.

Function	Description	Example
geohash(string)	Returns the geohash value of a specified geographical coordinate. The geographical coordinate is represented by a string in the format of "<latitude>, <longitude>". The return value is a string. Example: `geohash('34.1,120.6')`.	`* | select geohash('34.1,120.6')='wwjcbrdnzs'`

Function	Description	Example
geohash(lat,lon)	Returns the geohash value of a specified geographical coordinate. The geographical coordinate is represented by two separate parameters that indicate the latitude and longitude. The return value is a string.	`* | select geohash(34.1,120.6)= 'wwjcbrdnzs'`

23.1.4.8.32. JOIN clause

You can use JOIN clauses in SQL statements to join multiple tables by using fields that are shared by the tables. In Log Service, you can join one or more Logstores. You can also join Logstores with ApsaraDB RDS instances. This topic describes how to join different Logstores.

Procedure

1. Download the latest version of the Log Service SDK for Python.
2. Call the GetProjectLogs operation to query logs.

Sample SDK

```
#!/usr/bin/env python
#encoding: utf-8
import time,sys,os
from aliyun.log.logexception import LogException
from aliyun.log.logitem import LogItem
from aliyun.log.logclient import LogClient
from aliyun.log.getlogsrequest import GetLogsRequest
from aliyun.log.getlogsrequest import GetProjectLogsRequest
from aliyun.log.putlogsrequest import PutLogsRequest
from aliyun.log.listtopicsrequest import ListTopicsRequest
from aliyun.log.listlogstoresrequest import ListLogstoresRequest
from aliyun.log.gethistogramsrequest import GetHistogramsRequest
from aliyun.log.index_config import *
from aliyun.log.logtail_config_detail import *
from aliyun.log.machine_group_detail import *
from aliyun.log.acl_config import *
if __name__=='__main__':
    token = None
    endpoint = "http://cn-hangzhou.log.aliyuncs.com"
    accessKeyId  = '******'
    accessKey = '******'
    client = LogClient(endpoint, accessKeyId, accessKey,token)
    logstore = "meta"
    # In the query statement, specify two Logstores, the query time ranges of both Logstores, and the
key to join the Logstores.
    req = GetProjectLogsRequest(project,"select count(1) from sls_operation_log  s join  meta m on s._
_date__ >'2018-04-10 00:00:00' and s.__date__ < '2018-04-11 00:00:00' and m.__date__ >'2018-04-23 00:0
0:00' and m.__date__ <'2018-04-24 00:00:00' and s.projectid = cast(m.ikey as varchar)");
    res = client.get_project_logs(req)
    res.log_print();
    exit(0)
```

Note For more information about the syntax and usage examples of JOIN clauses, see JOIN clause.

23.1.4.8.33. UNNEST clause

This topic describes the syntax of UNNEST clauses.

Scenario

The value of a column in log data is stored as a primitive data type, such as string or number. In some cases, the value of a column may be of a complex data type, such as array, map, or JSON. When you query and analyze logs that contain fields whose values are of the preceding types, you can use an UNNEST clause to expand the field values into multiple rows for analysis.

Example:

```
__source__: 1.1.1.1__tag__:__hostname__: vm-req-170103232316569850-tianchi111932.tc__topic__: TestTopi
c_4array_column: [1,2,3]double_column: 1.23map_column: {"a":1,"b":2}text_column: Product
```

The value of the `array_column` field is an array. To obtain the sum of all elements in the value of the `array_column` field, you must traverse all elements of each array.

Syntax of UNNEST clauses

Syntax	Description
`unnest(array) as table_alias(column_name)`	Expands an array into multiple rows. `column_name` specifies the column name of the rows.
`unnest(map) as table(key_name, value_name)`	Expands a map into multiple rows. `key_name` specifies the column name of the keys and `value_name` specifies the column name of the values.

Note An UNNEST clause is applicable only to arrays or maps. If you want to expand a string, you must convert the string to JSON data. Then, you can use the `cast(json_parse(array_column) as array(bigint))` syntax to convert the JSON data to an array or a map.

Traverse the elements of an array

Use an UNNEST clause to expand an array into multiple rows. The rows are stored in a table named t. The column name of the rows is referenced as a.

```
* | select array_column, a from log, unnest(cast(json_parse(array_column) as array(bigint))) as t(a)
```

When the elements in an array are traversed, you can also use other SQL syntax to query and analyze data. Examples:

- Calculate the sum of the elements in an array:

  ```
  * | select sum(a) from log, unnest(cast(json_parse(array_column) as array(bigint))) as t(a)
  ```

- Use a GROUP BY clause to group the elements in an array by column name:

  ```
  * | select a, count(1) from log, unnest(cast(json_parse(array_column) as array(bigint))) as t(a) gr
  oup by a
  ```

Traverse the elements of a map

- Traverse the elements of a map:

```
* | select map_column, a, b from log, unnest(cast(json_parse(map_column) as map(varchar, bigint)))
as t(a,b)
```

- Use a GROUP BY clause to group the elements in a map by key:

```
* | select key, sum(value) from log, unnest(cast(json_parse(map_column) as map(varchar, bigint))) a
s t(key, value) GROUP BY key
```

Visualize the results of the histogram and numeric_histogram functions

- histogram

  The histogram function is similar to the count group by syntax. For more information, see Map functions.

  The histogram function returns JSON data that cannot be visualized. The following example shows a query statement:

```
* | select histogram(method)
```

  To visualize the logs that contain the method field, you can use an UNNEST clause to expand the JSON data that is returned by the histogram function into multiple rows. The following example shows a query statement:

```
* | select key, value from (select histogram(method) as his from log), unnest(his) as t(key, value)
```

- numeric_histogram

  The numeric_histogram function is used to compute the approximate histogram of a specified field based on the number of histogram columns specified by the bucket parameter. This function is equivalent to the GROUP BY clause that is used to group data by numeric value column. For more information, see Approximate functions.

```
* | select numeric_histogram(10, Latency)
```

  To visualize the result of the numeric_histogram function, execute the following query statement:

```
* | select key, value from (select numeric_histogram(10, Latency) as his from log), unnest(his) as
t(key, value)
```

23.1.4.9. Machine learning syntax and functions

23.1.4.9.1. Overview

Log Service provides the machine learning feature that supports multiple algorithms and calling methods. You can use SELECT statements and machine learning functions to call machine learning algorithms and analyze the characteristics of one or more fields within a specific period of time.

Log Service offers various time series analysis algorithms. You can call these algorithms to solve problems that are related to time series data. For example, you can predict time series, detect time series anomalies, decompose time series, and cluster multiple time series. In addition, the algorithms are compatible with standard SQL functions. This simplifies the usage of the algorithms and improves the efficiency of troubleshooting.

Features

- Supports various smooth operations on single-time series data.
- Supports algorithms that are used for the prediction, anomaly detection, change point detection, inflection point detection, and multi-period estimation of single-time series data.
- Supports decomposition operations on single-time series data.
- Supports various clustering algorithms for multi-time series data.
- Supports multi-field pattern mining based on the sequence of numeric data or text.

Limits

When you use the machine learning feature of Log Service, take note of the following limits:

- The specified time series data must be sampled based on the same interval.
- The specified time series data cannot contain data that is repeatedly sampled from the same point in time.
- The processing capacity cannot exceed the maximum capacity. The following table describes the limits.

Item	Description
Capacity of the time-series data processing	Data can be collected from a maximum of 150,000 consecutive points in time. If the data volume exceeds the processing capacity, you must aggregate the data or reduce the sampling amount.
Capacity of the density-based clustering algorithm	A maximum of 5,000 time series curves can be clustered at a time. Each curve cannot contain more than 1,440 points in time.
Capacity of the hierarchical clustering algorithm	A maximum of 2,000 time series curves can be clustered at a time. Each curve cannot contain more than 1,440 points in time.

Machine learning functions

Type	Function	Description
Smooth functions	ts_smooth_simple	Uses the Holt-Winters forecasting algorithm to filter time series data.
	ts_smooth_fir	Uses a finite impulse response (FIR) filter to filter time series data.
	ts_smooth_iir	Uses an infinite impulse response (IIR) filter to filter time series data.
Multi-period estimation functions	ts_period_detect	Estimates time series data by period.
Change point detection functions	ts_cp_detect	Detects the intervals in which data has different statistical features. The interval endpoints are change points.
	ts_breakout_detect	Detects the points in time at which data dramatically changes.
Maximum value detection function	ts_find_peaks	Detects the local maximum value of time series data in a specified window.
Prediction and anomaly detection functions	ts_predicate_simple	Uses default parameters to model time series data, predict time series data, and detect anomalies.
	ts_predicate_ar	Uses an autoregressive (AR) model to model time series data, predict time series data, and detect anomalies.
	ts_predicate_arma	Uses an autoregressive moving average (ARMA) model to model time series data, predict time series data, and detect anomalies.
	ts_predicate_arima	Uses an autoregressive integrated moving average (ARIMA) model to model time series data, predict time series data, and detect anomalies.

Type	Function	Description
	ts_regression_predict	Predicts the trend for a single periodic time series.
	ts_anomaly_filter	Filters the anomalies that are detected from multiple time series curves based on the custom anomaly mode. The anomalies are detected during the anomaly detection. This function helps you find abnormal curves at the earliest opportunity.
Time series decomposition function	ts_decompose	Uses the Seasonal and Trend decomposition using Loess (STL) algorithm to decompose time series data.
Time series clustering functions	ts_density_cluster	Uses a density-based clustering method to cluster multiple time series.
	ts_hierarchical_cluster	Uses a hierarchical clustering method to cluster multiple time series.
	ts_similar_instance	Queries time series curves that are similar to a specified time series curve.
Frequent pattern statistics function	pattern_stat	Mines representative combinations of attributes among the given multi-attribute field samples to obtain the frequent pattern in statistical patterns.
Differential pattern statistics function	pattern_diff	Identifies the pattern that causes differences between two collections in specified conditions.
Root cause analysis function	rca_kpi_search	Analyzes the subdimension attributes that cause the anomalies of a monitoring metric.
Correlation analysis functions	ts_association_analysis	Identifies the metrics that are correlated to a specified metric among multiple observed metrics in the system.
	ts_similar	Identifies the metrics that are correlated to specified time series data among multiple observed metrics in the system.
Kernel density estimation function	kernel_density_estimation	Uses the smooth peak function to fit the observed data points. In this way, the function simulates the real probability distribution curve.

23.1.4.9.2. Smooth functions

This topic describes the smooth functions that you can use to smooth and filter specified time series curves. Filtering is the first step to discover the shapes of time series curves.

Functions

Function	Description
ts_smooth_simple	Uses the Holt-Winters forecasting algorithm to filter time series data. This function is the default smooth function.
ts_smooth_fir	Uses a finite impulse response (FIR) filter to filter time series data.

Function	Description
ts_smooth_iir	Uses an infinite impulse response (IIR) filter to filter time series data.

ts_smooth_simple

- Syntax

```
select ts_smooth_simple(x, y)
```

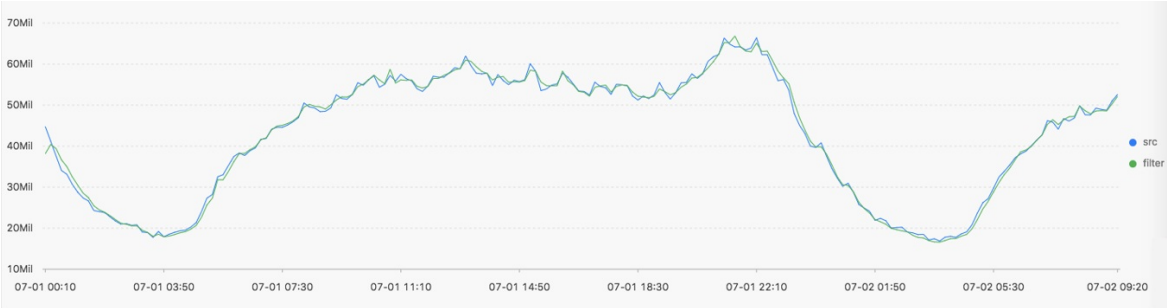
- The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None

- Example
  - Query statement

```
* | select ts_smooth_simple(stamp, value) from ( select __time__ - __time__ % 120 as stamp, avg(v
) as value from log GROUP BY stamp order by stamp )
```

  - Query result



- The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The UNIX timestamp of the data. Unit: seconds.
Vertical axis	src	The unfiltered data.
	filter	The filtered data.

ts_smooth_fir

- Syntax
  - If you cannot determine filter parameters, use the built-in window parameters in the following statement:

```
select ts_smooth_fir(x, y,winType,winSize)
```

  - If you can determine filter parameters, you can specify the parameters as needed in the following statement:

```
select ts_smooth_fir(x, y,array[])
```

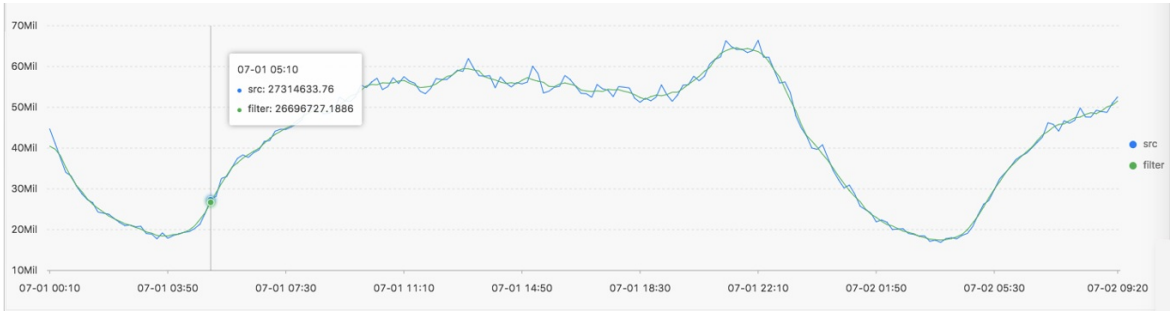
- The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
winType	The type of the window that you want to use to filter data.	Valid values: ○ rectangle: rectangle window ○ hanning: hanning window ○ hamming: hamming window ○ blackman: blackman window **Note** We recommend that you set the winType parameter to rectangle for better display effects.
winSize	The length of the filter window.	The value is of the long type. Valid values: 2 to 15.
array[]	The parameter that you want to use for FIR filtering.	The value is an array and the sum of the elements in the array is 1. Example: array[0.2, 0.4, 0.3, 0.1].

- Example 1
  - Query statement

```
* | select ts_smooth_fir(stamp, value, 'rectangle', 4) from ( select __time__ - __time__ % 120 as stamp, avg(v) as value from log GROUP BY stamp order by stamp )
```

  - Query result



- Example 2
  - Query statement

```
* | select ts_smooth_fir(stamp, value, array[0.2, 0.4, 0.3, 0.1]) from ( select __time__ - __time__ % 120 as stamp, avg(v) as value from log GROUP BY stamp order by stamp )
```

- Query result

- The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The UNIX timestamp of the data. Unit: seconds.
Vertical axis	src	The unfiltered data.
	filter	The filtered data.

ts_smooth_iir

- Syntax

```
select ts_smooth_iir(x, y, array[], array[] )
```

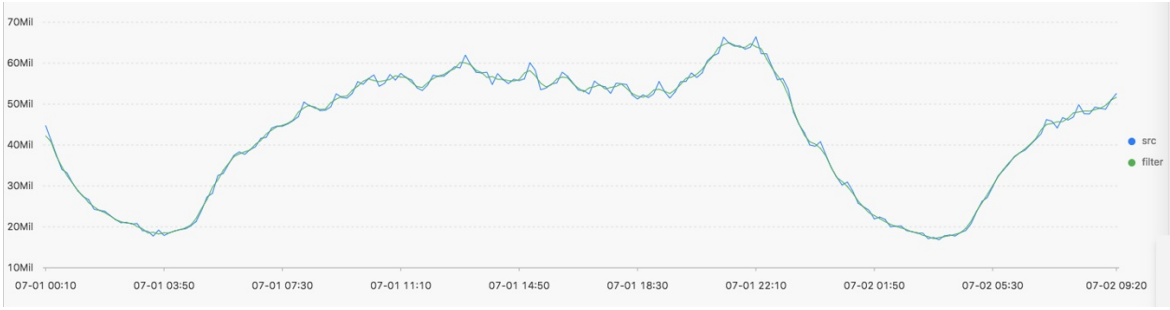
- The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
array[]	The parameter that you want to use for IIR filtering in terms of x_i.	The value is an array and the sum of the elements in the array is 1. The length of the array ranges from 2 to 15. Example: array[0.2, 0.4, 0.3, 0.1].
array[]	The parameter that you want to use for IIR filtering in terms of y_{i-1}.	The value is an array and the sum of the elements in the array is 1. The length of the array ranges from 2 to 15. Example: array[0.2, 0.4, 0.3, 0.1].

- Example
  - Query statement

```
* | select ts_smooth_iir(stamp, value, array[0.2, 0.4, 0.3, 0.1], array[0.4, 0.3, 0.3]) from ( se
lect __time__ - __time__ % 120 as stamp, avg(v) as value from log GROUP BY stamp order by stamp )
```

- Query result



- The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The UNIX timestamp of the data. Unit: seconds.
Vertical axis	src	The unfiltered data.
	filter	The filtered data.

23.1.4.9.3. Multi-period estimation functions

This topic describes multi-period estimation functions that you can use to estimate the periodicity of time series data distributed in different time intervals. This topic also describes how to extract the periodicity by using a series of operations such as Fourier transform (FT).

Functions

Function	Description
ts_period_detect	Estimates the periodicity of time series data that is distributed in different time intervals.
ts_period_classify	Uses FT to calculate the periodicity of specified time series curves. This function can be used to identify periodic curves.

ts_period_detect

Syntax

```
select ts_period_detect(x,y,minPeriod,maxPeriod)
```

The following table describes the parameters in the function.

Parameter	Description
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis. Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.
minPeriod	The ratio of the minimum length of the estimated period to the total length of the time series data. The value of this parameter is of the float type. Valid values: (0,1].

Parameter	Description
maxPeriod	The ratio of the maximum length of the estimated period to the total length of the time series data. The value of this parameter is of the float type. Valid values: (0,1]. **Note** The value of the *maxPeriod* parameter must be greater than the value of the *minPeriod* parameter.

Example

- Query statement

```
* | select ts_period_detect(stamp, value, 0.2, 1.0) from ( select __time__ - __time__ % 120 as stamp, avg(v) as value from log GROUP BY stamp order by stamp )
```

- Query result

The following table describes the display items.

Display item	Description
period_id	An array with a length of 1. The element in the array indicates the sequence number of the period. The array [0] indicates the original time series curve.
time_series	The sequence of timestamps.
data_series	The sequence of data at each timestamp. • If the value of period_id is 0, the function returns the original time series data. • If the value of period_id is not 0, the function returns filtered time series data.

ts_period_classify

Syntax

```
select ts_period_classify(stamp,value,instanceName)
```

The following table describes the parameters in the function.

Parameter	Description
stamp	The time sequence. Points in time are sorted in ascending order along the horizontal axis. Each point in time is a UNIX timestamp. Unit: seconds.
value	The sequence of numeric data at a specific point in time.

Parameter	Description
instanceName	The name of the time series curve.

Example

- Query statement

```
* and h : nu2h05202.nu8 | select ts_period_classify(stamp, value, name) from log
```

- Query result

line_name	prob	type
asg-2zxjqjn6zf5ewg188pg5	1.0	-1.0
asg-bp1j8snc92p6v5pptgpj	0.07203669207039314	0.0
asg-wz99hse7u4ubopo5dt9o	0.0	0.0
asg-bp18oqni0gq96vy85te4	0.05590892692207093	0.0

The following table describes the display items.

Display item	Description
line_name	An array with a length of 1. The element in the array indicates the sequence number of the period. The array [0] indicates the original time series curve.
prob	The ratio of the number of values within the primary period to the total number of values on the time series curve. Valid values: [0, 1]. You can set the value to 0.15 for testing.
type	The type of the curve. Valid values: • -1: The time series curve has a length of less than 64 points. • -2: The time series curve has a failure rate of higher than 20%. • 0: The time series curve is periodic.

23.1.4.9.4. Change point detection functions

This topic describes the change point detection functions that you can use to detect the change points in time series data.

Change point detection functions can detect the following two kinds of change points:

- Changes in statistical features within a specific period of time
- Anomalies in time series data

Functions

Function	Description
ts_cp_detect	Detects the intervals in which data has different statistical features. The interval endpoints are change points.
ts_breakout_detect	Detects the points in time at which data dramatically changes.

ts_cp_detect

Syntax

```
select ts_cp_detect(x, y, minSize)
```

The following table describes the parameters in the function.

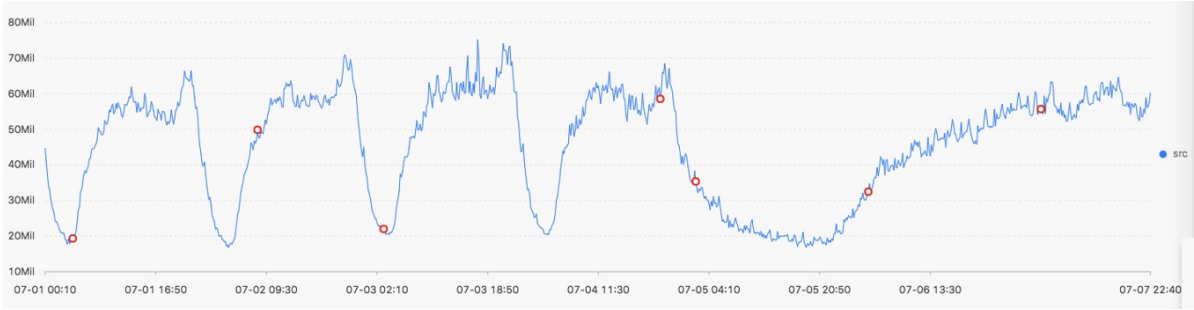
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
minSize	The minimum length of time series data within a consecutive interval.	The minimum length is 3. The maximum length cannot exceed one tenth of the length of the specified time series data. Default value: 10.

Example

- Query statement

```
* | select ts_cp_detect(stamp, value, 3) from (select __time__ - __time__ % 10 as stamp, avg(v) as value from log GROUP BY stamp order by stamp)
```

- Query result



The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The timestamp of the data. Unit: seconds. Example: 1537071480.
Vertical axis	src	The unfiltered data. Example: 1956092.7647745228.
	prob	The probability that a point in time is a change point. Valid values: 0 to 1.

ts_breakout_detect

Syntax

```
select ts_breakout_detect(x, y, winSize)
```

The following table describes the parameters in the function.

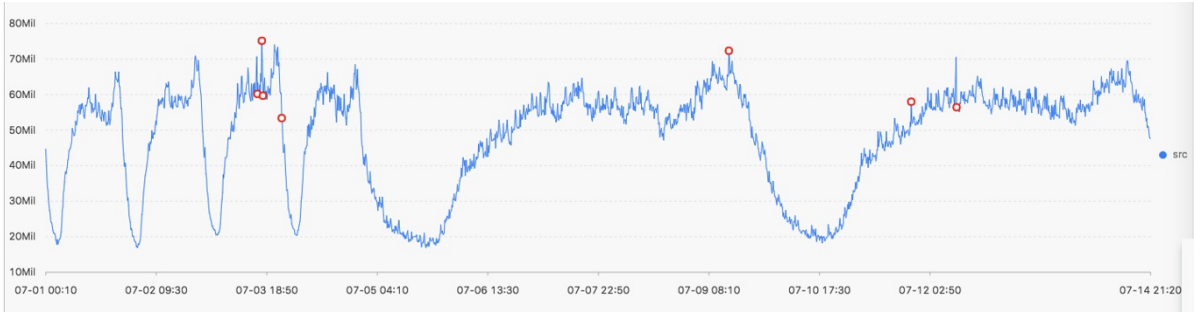
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
winSize	The minimum length of time series data within a consecutive interval.	The minimum length is 3. The maximum length cannot exceed one tenth of the length of the specified time series data. Default value: 10.

Example

- Query statement

```
* | select ts_breakout_detect(stamp, value, 3) from (select __time__ - __time__ % 10 as stamp, avg(
v) as value from log GROUP BY stamp order by stamp)
```

- Query result



The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The timestamp of the data. Unit: seconds. Example: 1537071480.
Vertical axis	src	The unfiltered data. Example: 1956092.7647745228.
	prob	The probability that a point in time is a change point. Valid values: 0 to 1.

23.1.4.9.5. Maximum value detection function

This topic describes the maximum value detection function that you can use to detect the local maximum value of time series data in a specified window.

ts_find_peaks

Syntax

```
select ts_find_peaks(x, y, winSize)
```

The following table describes the parameters in the function.

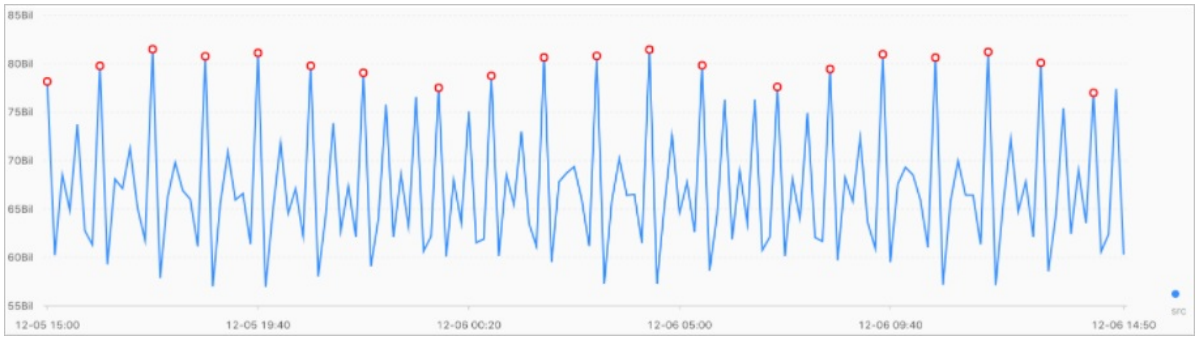
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
winSize	The minimum length of the detection window.	The value of this parameter is of the long type and ranges from 1 to the length of time series data. We recommend that you set this parameter to one tenth of the actual data length.

Example

- Query statement

```
* and h : nu2h05202.nu8 and m: NET |  select ts_find_peaks(stamp, value, 30) from (select __time__
- __time__ % 10 as stamp, avg(v) as value from log GROUP  BY  stamp order by stamp)
```

- Query result



The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The timestamp of the data. Unit: seconds. Example: 1537071480.
Vertical axis	src	The unfiltered data. Example: 1956092.7647745228.
	peak_flag	Indicates whether the numeric value at a point in time is the maximum value. Valid values: • 1.0: The numeric value at the point in time is the maximum value. • 0.0: The numeric value at the point in time is not the maximum value.

23.1.4.9.6. Prediction and anomaly detection functions

Prediction and anomaly detection functions predict the trend of time series curves and identify the Ksigma and quantiles of the errors between a predicted curve and an actual curve. You can use the functions to detect anomalies.

Functions

Function	Description
ts_predicate_simple	Uses default parameters to model time series data, predict time series data, and detect anomalies.
ts_predicate_ar	Uses an autoregressive (AR) model to model time series data, predict time series data, and detect anomalies.
ts_predicate_arma	Uses an autoregressive moving average (ARMA) model to model time series data, predict time series data, and detect anomalies.
ts_predicate_arima	Uses an autoregressive integrated moving average (ARIMA) model to model time series data, predict time series data, and detect anomalies.
ts_regression_predict	Predicts the trend for a single periodic time series. Scenario: You can use this function to predict metering data, network traffic, financial data, and different business data that follows certain rules.
ts_anomaly_filter	Filters the anomalies that are detected from multiple time series curves based on the custom anomaly mode. The anomalies are detected during the anomaly detection. This function helps you find abnormal curves at the earliest opportunity.

ts_predicate_simple

Syntax

```
select ts_predicate_simple(x, y, nPred, isSmooth)
```

The following table describes the parameters in the function.

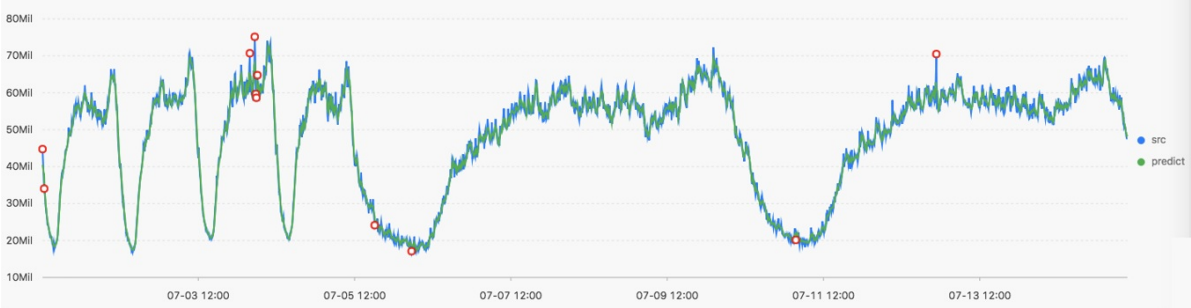
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
nPred	The number of points for prediction.	The value is of the long type. This value must be equal to or greater than 1.
isSmooth	Specifies whether to filter the raw data. • true: The raw data is filtered. • false: The raw data is not filtered.	The value is of the Boolean type. Default value: true.

Example

- Query statement

```
* | select ts_predicate_simple(stamp, value, 6) from (select __time__ - __time__ % 60 as stamp, avg
(v) as value from log GROUP BY stamp order by stamp)
```

- Query result



The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The UNIX timestamp of the data. Unit: seconds.
Vertical axis	src	The raw data.
	predict	The predicted data.
	upper	The upper limit of the prediction. The confidence level is 0.85. This value cannot be modified.
	lower	The lower limit of the prediction. The confidence level is 0.85. This value cannot be modified.
	anomaly_prob	The probability that the point is an anomaly. Valid values: 0 to 1.

ts_predicate_ar

Syntax

```
select ts_predicate_ar(x, y, p, nPred, isSmooth)
```

The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
p	The order of the AR model.	The value is of the long type. Valid values: 2 to 8.
nPred	The number of points for prediction.	The value is of the long type. Valid values: 1 to 5*p*.
isSmooth	Specifies whether to filter the raw data. • true: The raw data is filtered. • false: The raw data is not filtered.	The value is of the Boolean type. Default value: true.

Query statement

```
* | select ts_predicate_ar(stamp, value, 3, 4) from (select __time__ - __time__ % 60 as stamp, avg(v) as value from log GROUP BY stamp order by stamp)
```

Note The result is similar to the result that is returned by the ts_predicate_simple function. For more information, see ts_predicate_simple.

ts_predicate_arma

Syntax

```
select ts_predicate_arma(x, y, p, q, nPred, isSmooth)
```

The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
p	The order of the AR model.	The value is of the long type. Valid values: 2 to 100.
q	The order of the ARMA model.	The value is of the long type. Valid values: 2 to 8.
nPred	The number of points for prediction.	The value is of the long type. Valid values: 1 to 5*p*.
isSmooth	Specifies whether to filter the raw data. • true: The raw data is filtered. • false: The raw data is not filtered.	The value is of the Boolean type. Default value: true.

Query statement

```
* | select ts_predicate_arma(stamp, value, 3, 2, 4) from (select __time__ - __time__ % 60 as stamp, avg(v) as value from log GROUP BY stamp order by stamp)
```

Note The result is similar to the result that is returned by the ts_predicate_simple function. For more information, see ts_predicate_simple.

ts_predicate_arima

Syntax

```
select ts_predicate_arima(x, y, p, d, q, nPred, isSmooth)
```

The following table describes the parameters in the function.

Parameter	Description	Value

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
p	The order of the AR model.	The value is of the long type. Valid values: 2 to 8.
d	The order of the ARIMA model.	The value is of the long type. Valid values: 1 to 3.
q	The order of the ARMA model.	The value is of the long type. Valid values: 2 to 8.
nPred	The number of points for prediction.	The value is of the long type. Valid values: 1 to 5*p*.
isSmooth	Specifies whether to filter the raw data. • true: The raw data is filtered. • false: The raw data is not filtered.	The value is of the Boolean type. Default value: true.

Query statement

```
* | select ts_predicate_arima(stamp, value, 3, 1, 2, 4) from (select __time__ - __time__ % 60 as stamp
, avg(v) as value from log GROUP BY stamp order by stamp)
```

Note The result is similar to the result that is returned by the ts_predicate_simple function. For more information, see ts_predicate_simple.

ts_regression_predict

Syntax

```
select ts_regression_predict(x, y, nPred, algotype, processType)
```

The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
nPred	The number of points for prediction.	The value is of the long type. Valid values: 1 to 500.

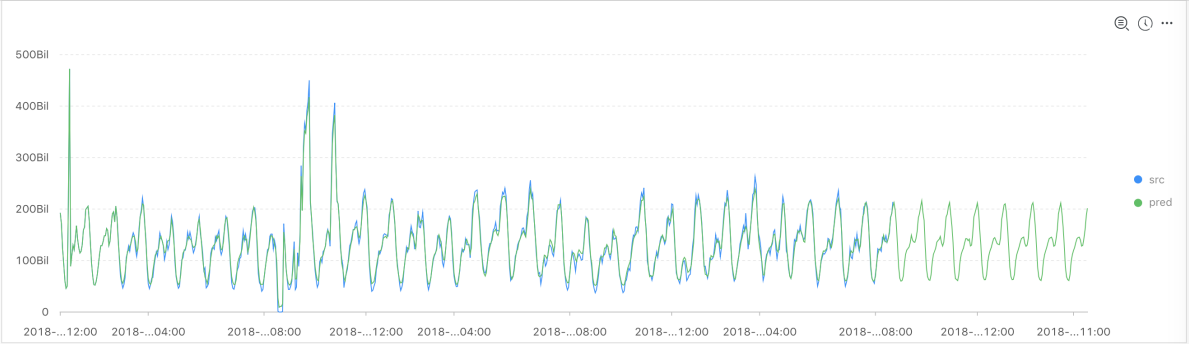
Parameter	Description	Value
algotype	The type of the algorithm used for prediction. Valid values: • origin: uses the Gradient Boosted Regression Tree (GBRT) algorithm for prediction. • forest: uses the GBRT algorithm for prediction based on the trend component decomposed by Seasonal and Trend decomposition using Loess (STL), and then uses the additive model to calculate the sum of the decomposed components and obtains the predicted data. • linear: uses the Linear Regression algorithm for prediction based on the trend components decomposed by STL, and then uses the additive model to calculate the sum of the decomposed components and obtains the predicted data.	None
processType	Specifies whether to preprocess the data. Valid values: • 0: No additional data preprocessing is performed. • 1: Abnormal data is removed before prediction.	None

Example

- Query statement

```
* and h : nu2h05202.nu8 and m: NET |  select ts_regression_predict(stamp, value, 200, 'origin') fro
m (select __time__ - __time__ % 60 as stamp, avg(v) as value from log GROUP  BY  stamp order by sta
mp)
```

- Query result



The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The UNIX timestamp of the data. Unit: seconds.
Vertical axis	src	The raw data.
	predict	The predicted data.

ts_anomaly_filter

Syntax

```
select ts_anomaly_filter(lineName, ts, ds, preds, probs, nWatch, anomalyType)
```

The following table describes the parameters in the function.

Parameter	Description	Value
lineName	The name of each curve. The value is of the varchar type.	None
ts	The time sequence of the curve, which indicates the time of the current curve.	The value of this parameter is an array of points in time of the double type. The points in time are sorted in ascending order.
ds	The actual value sequence of the curve.	The value of this parameter is an array of data points of the double type. The length of the value is the same as the length of the value of the ts parameter.
preds	The predicted value sequence of the curve.	The value of this parameter is an array of data points of the double type. The length of the value is the same as the length of the value of the ts parameter.
probs	The sequence of anomaly detection results of the curve.	The value of this parameter is an array of data points of the double type. The length of the value is the same as the length of the value of the ts parameter.
nWatch	The number of the actual values that are recently observed on the curve. The value is of the long type. This value must be less than the number of points in time on the curve.	The value is of the long type.
anomalyType	The type of anomaly that you want to filter. Valid values: • 0: all anomalies • 1: positive anomalies • -1: negative anomalies	The value is of the long type.

Example

- Query statement

```
* | select res.name, res.ts, res.ds, res.preds, res.probs
     from (
         select ts_anomaly_filter(name, ts, ds, preds, probs, cast(5 as bigint), cast(1 as bigint))
as res
       from (
         select name, res[1] as ts, res[2] as ds, res[3] as preds, res[4] as uppers, res[5] as lowe
rs, res[6] as probs
     from (
         select name, array_transpose(ts_predicate_ar(stamp, value, 10)) as res
         from (
           select name, stamp, value from log where name like '%asg-%') group by name)) );
```

- Query result

```
| name                      | ts                                                                | ds          | p
reds     | probs       |
| ------------------------- | ----------------------------------------------------------------- | ----------- | -
-------- | ----------- |
| asg-bp1hylzdi2wx7civ0ivk | [1.5513696E9, 1.5513732E9, 1.5513768E9, 1.5513804E9] | [1,2,3,NaN] | [
1,2,3,4] | [0,0,1,NaN] |
```

23.1.4.9.7. Time series decomposition function

This topic describes the time series decomposition function that you can use to decompose time series curves and show the trend and periodicity of curves.

ts_decompose

Syntax

```
select ts_decompose(x, y)
```

The following table describes the parameters in the function.

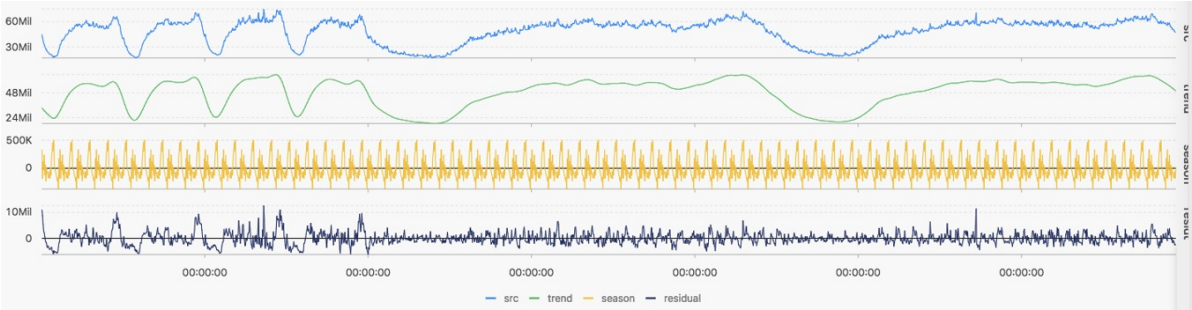
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None

Example

- Query statement

```
* | select ts_decompose(stamp, value) from (select __time__ - __time__ % 60 as stamp, avg(v) as val
ue from log GROUP BY stamp order by stamp)
```

- Query result



The following table describes the display items.

Display item		Description
Horizontal axis	unixtime	The UNIX timestamp of the data. Unit: seconds.
	src	The raw data.
	trend	The decomposed data that indicates the trend of the time series data.

Vertical axis Display item		Description
	season	The decomposed data that indicates the periodicity of the time series data.
	residual	The residual data that is decomposed from the time series data.

23.1.4.9.8. Time series clustering functions

You can use time series clustering functions to cluster data of multiple time series and obtain different curve shapes. Then, you can use the data to find the cluster center and identify curves with shapes that are different from other curve shapes in the cluster.

Functions

Function	Description
ts_density_cluster	Uses a density-based clustering method to cluster multiple time series.
ts_hierarchical_cluster	Uses a hierarchical clustering method to cluster multiple time series.
ts_similar_instance	Queries time series curves that are similar to a specified time series curve.

ts_density_cluster

Syntax

```
select ts_density_cluster(x, y, z)
```

The following table describes the parameters in the function.

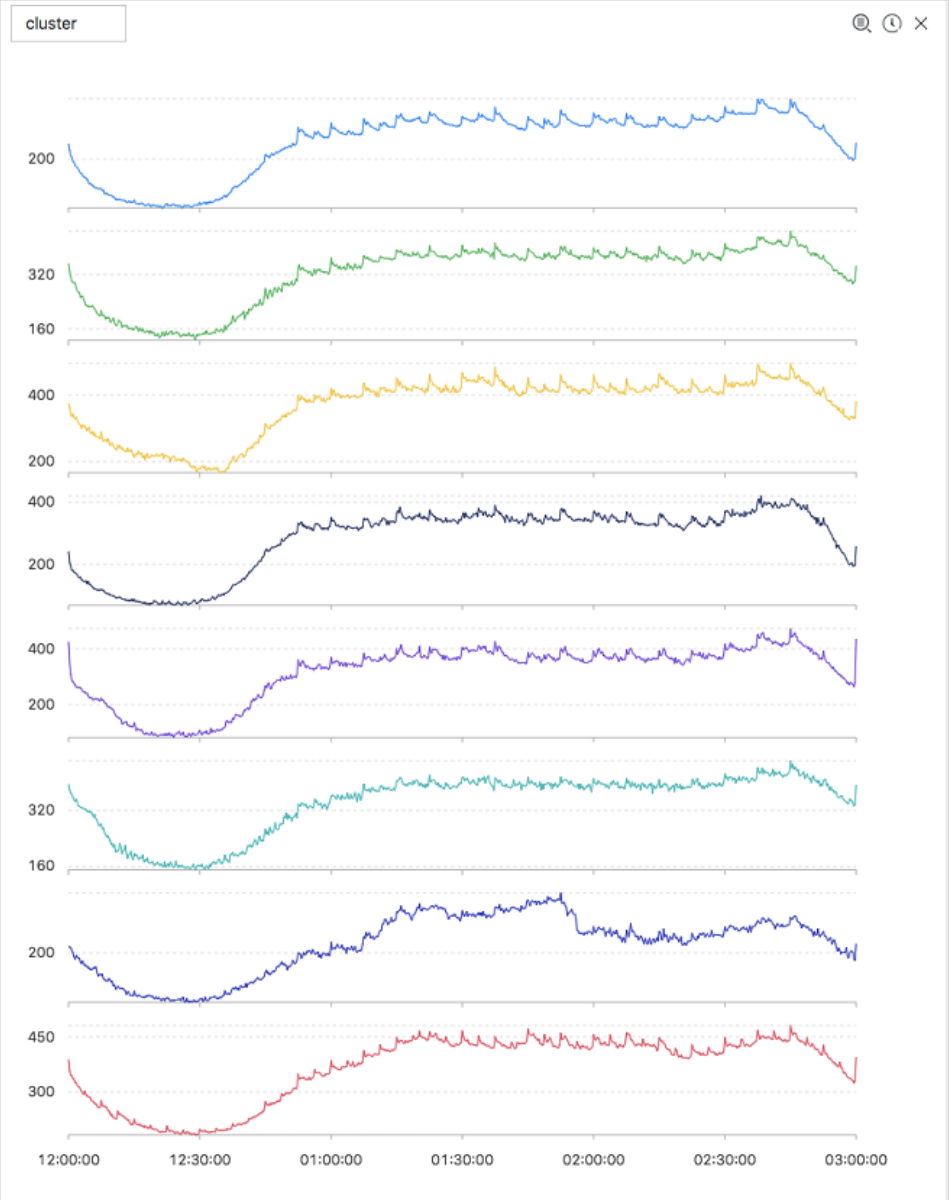
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
z	The name of the curve that corresponds to the data at a specified point in time.	The value of this parameter is a string. Example: machine01.cpu_usr.

Example

- Query statement

```
* and (h: "machine_01" OR h: "machine_02" OR h : "machine_03") | select ts_density_cluster(stamp, m
etric_value,metric_name ) from ( select __time__ - __time__ % 600 as stamp, avg(v) as metric_value,
h as metric_name from log GROUP BY stamp, metric_name order BY metric_name, stamp )
```

- Query result



The following table describes the display items.

Display item	Description
cluster_id	The category of the cluster. The value -1 indicates that the cluster is not categorized in a cluster center.
rate	The proportion of instances in the cluster.
time_series	The timestamp sequence of the cluster center.
data_series	The data sequence of the cluster center.
instance_names	The instances that are included in the cluster center.
sim_instance	The name of an instance in the cluster.

ts_hierarchical_cluster

Syntax

```
select ts_hierarchical_cluster(x, y, z)
```

The following table describes the parameters in the function.

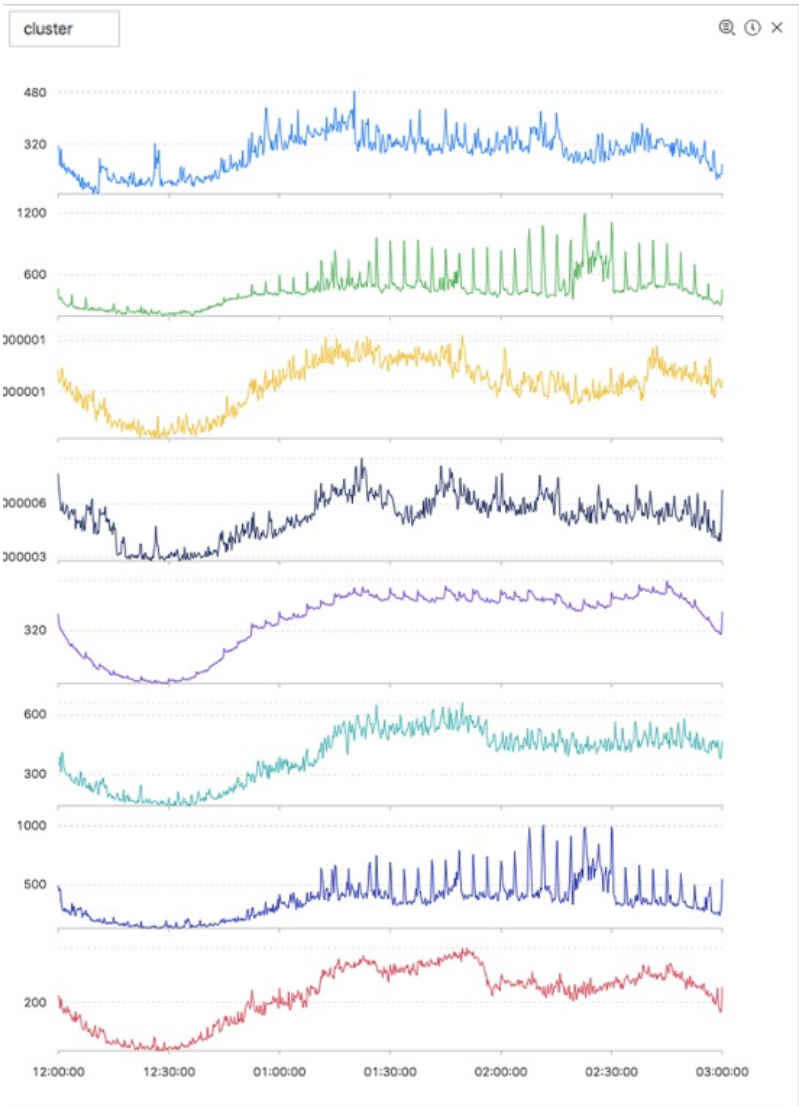
Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
z	The name of the curve that corresponds to the data at a specified point in time.	The value of this parameter is a string. Example: machine01.cpu_usr.

Example

- Query statement

```
* and (h: "machine_01" OR h: "machine_02" OR h : "machine_03") | select ts_hierarchical_cluster(stamp, metric_value, metric_name) from ( select __time__ - __time__ % 600 as stamp, avg(v) as metric_value, h as metric_name from log GROUP BY stamp, metric_name order BY metric_name, stamp )
```

- Query result



The following table describes the display items.

Display item	Description
cluster_id	The category of the cluster. The value -1 indicates that the cluster is not categorized in a cluster center.
rate	The proportion of instances in the cluster.
time_series	The timestamp sequence of the cluster center.
data_series	The data sequence of the cluster center.
instance_names	The instances that are included in the cluster center.
sim_instance	The name of an instance in the cluster.

ts_similar_instance

Syntax

```
select ts_similar_instance(x, y, z, instance_name, topK, metricType)
```

The following table describes the parameters in the function.

Parameter	Description	Value
x	The time sequence. Points in time are sorted in ascending order along the horizontal axis.	Each point in time is a UNIX timestamp. Unit: seconds.
y	The sequence of numeric data at a specific point in time.	None
z	The name of the curve that corresponds to the data at a specified point in time.	The value of this parameter is a string. Example: machine01.cpu_usr.
instance_name	The name of a specified curve that you want to query.	The value is of this parameter is a string. Example: machine01.cpu_usr. **Notice** You must specify an existing curve.
topK	The maximum number of curves that are similar to the specified curve can be returned.	None
metricType	`{'shape', 'manhattan', 'euclidean'}`. The metric used to measure the similarity between time series curves.	None

Query statement

```
* and m: NET and m: Tcp and (h: "nu4e01524.nu8" OR  h: "nu2i10267.nu8" OR  h : "nu4q10466.nu8") | sele
ct ts_similar_instance(stamp, metric_value, metric_name, 'nu4e01524.nu8' ) from ( select __time__ - __
time__ % 600 as stamp, sum(v) as metric_value, h as metric_name from log GROUP BY stamp, metric_name o
rder BY  metric_name, stamp )
```

The following table describes the display items.

Display item	Description
instance_name	The list of metrics that are similar to the specified metric.
time_series	The timestamp sequence of the cluster center.
data_series	The data sequence of the cluster center.

23.1.4.9.9. Frequent pattern statistics function

The frequent pattern statistics function combines representative attributes in a specified multi-attribute field sample.

pattern_stat

Syntax:

```
select pattern_stat(array[col1, col2, col3], array['col1_name', 'col2_name', 'col3_name'], array[col5,
col6], array['col5_name', 'col6_name'], support_score, sample_ratio)
```

The following table lists the parameters of the function.

Parameter	Description	Value
array[col1, col2, col3]	A column of character values.	An array of values, for example, array[clientIP, sourceIP, path, logstore].
array['col1_name', 'col2_name', 'col3_name']	The field names of the character values.	An array of field names, for example, array['clientIP', 'sourceIP', 'path', 'logstore'].
array[col5, col6]	A column of numeric values.	An array of values, for example, array[Inflow, OutFlow].
array['col5_name', 'col6_name']	The field names of the numeric values.	An array of field names, for example, array['Inflow', 'OutFlow'].
support_score	The support ratio of samples for pattern mining.	The value is of the DOUBLE data type. Value range: (0,1].
sample_ratio	The sampling ratio. The default value is 0.1, which indicates that only 10% of the total samples are used.	The value is of the DOUBLE data type. Value range: (0,1].

Example:

- Query statement

```
* | select pattern_stat(array[ Category, ClientIP, ProjectName, LogStore, Method, Source, UserAgent ], array[ 'Category', 'ClientIP', 'ProjectName', 'LogStore', 'Method', 'Source', 'UserAgent' ], array[ InFlow, OutFlow ], array[ 'InFlow', 'OutFlow' ], 0.45, 0.3) limit 1000
```

- Display item

Display item	Description
count	The number of samples in the current pattern.
support_score	The score of the current pattern. The score indicates the degree to which the current pattern is supported.
pattern	The content of the pattern. The pattern is organized in the format that is defined by the query conditions.

23.1.4.9.10. Differential pattern statistics function

The differential pattern statistics function analyzes differential patterns of specified multi-field samples based on the specified condition. The function helps you identify the causes of the differences under the current condition at the earliest opportunity.

pattern_diff

Syntax

```
select pattern_diff(array_char_value, array_char_name, array_numeric_value, array_numeric_name, condition, supportScore,posSampleRatio,negSampleRatio )
```

The following table describes the parameters in the function.

Parameter	Description	Value
array_char_value	A column of values of the character data type.	The value of this parameter is an array. Example: array[clientIP, sourceIP, path, logstore].
array_char_name	The column names of the values of the character data type.	The value of this parameter is an array. Example: array['clientIP', 'sourceIP', 'path', 'logstore'].
array_numeric_value	A column of numeric values.	The value of this parameter is an array. Example: array[Inflow, OutFlow].
array_numeric_name	The column names of the numeric values.	The value of this parameter is an array. Example: array['Inflow', 'OutFlow'].
condition	The condition that is used to filter data. The value True indicates positive samples and the value False indicates negative samples.	Example: Latency <= 300.
supportScore	The support ratio of positive and negative samples for pattern mining.	The value of this parameter is of the double type. Valid values: (0,1].
posSampleRatio	The sampling ratio of positive samples. Default value: 0.5. This value indicates that 50% of positive samples are collected.	The value of this parameter is of the double type. Valid values: (0,1].
negSampleRatio	The sampling ratio of negative samples. Default value: 0.5. This value indicates that 50% of negative samples are collected.	The value of this parameter is of the double type. Valid values: (0,1].

Example

- Query statement

```
* | select pattern_diff(array[ Category, ClientIP, ProjectName, LogStore, Method, Source, UserAgent ], array[ 'Category', 'ClientIP', 'ProjectName', 'LogStore', 'Method', 'Source', 'UserAgent' ], array[ InFlow, OutFlow ], array[ 'InFlow', 'OutFlow' ], Latency > 300, 0.2, 0.1, 1.0) limit 1000
```

- Display item

Display item	Description
possupport	The support ratio of positive samples for the mined patterns.
posconfidence	The confidence level of the mined patterns in positive samples.
negsupport	The support ratio of negative samples for the mined patterns.
diffpattern	The content of the mined patterns.

23.1.4.9.11. Root cause analysis function

Log Service provides alerting and analysis capabilities that allow you to analyze and identify anomalies in specific subdimensions of a metric at the earliest opportunity. You can use the root cause analysis function to identify and analyze the subdimension attributes that cause the anomalies.

rca_kpi_search

Syntax

```
select rca_kpi_search(varchar_array, name_array, real, forecast, level)
```

The following table describes the parameters in the function.

Parameter	Description	Value
varchar_array	The array of subdimension attributes.	Example: array[col1, col2, col3].
name_array	The array of subdimension attribute names.	Example: array['col1', 'col2', 'col3'].
real	The actual value of each subdimension attribute that is specified by the varchar_array parameter. The value of this parameter is of the double type.	Valid values: all real numbers.
forecast	The predicted value of each subdimension attribute that is specified by the varchar_array parameter. The value of this parameter is of the double type.	Valid values: all real numbers.
level	The number of subdimension attributes identified in the returned root cause set. The value 0 indicates that the function returns all root causes that are found. The value of this parameter is of the double type.	Valid values: [0, number of analyzed subdimensions]. The number of analysis dimensions is based on the length of the array that is specified by the varchar_array parameter.

Example

- Query statement

  Use a subquery to obtain the actual value and predicted value of each subdimension attribute, and then call the rca_kpi_search function to analyze the root causes of anomalies.

```
* not Status:200 |
select rca_kpi_search(
 array[ ProjectName, LogStore, UserAgent, Method ],
 array[ 'ProjectName', 'LogStore', 'UserAgent', 'Method' ], real, forecast, 1)
from (
select ProjectName, LogStore, UserAgent, Method,
 sum(case when time < 1552436040 then real else 0 end) * 1.0 / sum(case when time < 1552436040
then 1 else 0 end) as forecast,
 sum(case when time >=1552436040 then real else 0 end) *1.0 / sum(case when time >= 1552436040
then 1 else 0 end) as real
 from (
select __time__ - __time__ % 60 as time, ProjectName, LogStore, UserAgent, Method, COUNT(*) as real

from log GROUP by time, ProjectName, LogStore, UserAgent, Method )
GROUP BY ProjectName, LogStore, UserAgent, Method limit 100000000)
```

- Query result

The following figure shows the structure of the query result.

```
{
  "rcSets": [
    {
      "rcItems": [
        {
          "kpi": [("attr": "xxx", "val": "xxx")],
          "nleaf": 100,
          "change": 0.524543,
          "score": 0.1454543
        }
      ]
    }
  ]
}
```

The following table describes the display items.

Display item	Description
rcSets	The root cause sets. Each value is an array.
rcItems	A root cause set.
kpi	An item in the root cause set. Each item is formatted in an array where each element is of the JSON type. The attr parameter indicates the name of a subdimension. The val parameter indicates the attribute name that corresponds to the subdimension.
nleaf	The number of leaf nodes that a kpi in the root cause set covers in the raw data. **Note** A leaf node is a log entry that contains the finest-grained attributes.
change	The ratio of the number of anomaly changes in the leaf nodes that are covered by a kpi to the total number of anomaly changes in the root cause set at the same point in time.
score	The abnormality score of the current kpi. Valid values: [0,1].

The following example shows the query result that is in the JSON format:

```
{
  "rcSets": [
  {
    "rcItems": [
    {
      "kpi": [
      {
        "attr": "country",
        "val": "*"
      },
      {
        "attr": "province",
        "val": "*"
      },
      {
        "attr": "provider",
        "val": "*"
      },
      {
        "attr": "domain",
        "val": "download.huya.com"
      },
      {
        "attr": "method",
        "val": "*"
      }
      ],
      "nleaf": 119,
      "change": 0.3180687806279939,
      "score": 0.14436007709620113
    }
    ]
  }
  ]
}
```


23.1.4.9.12. Correlation analysis functions

You can use a correlation analysis function to find the metrics that are correlated with a specified metric or time series data among multiple observed metrics in the system.

Functions

Function	Description
ts_association_analysis	Identifies the metrics that are correlated to a specified metric among multiple observed metrics in the system.
ts_similar	Identifies the metrics that are correlated to specified time series data among multiple observed metrics in the system.

ts_association_analysis

Syntax

```
select ts_association_analysis(stamp, params, names, indexName, threshold)
```

The following table describes the parameters in the function.

Parameter	Description	Value
stamp	The UNIX timestamp that is of the long type.	None
params	The metrics that you want to analyze. The value of this parameter is an array. Each element in the array is of the double type.	Example: Latency, QPS, and NetFlow.
names	The names of the metrics that you want to analyze. The value of this parameter is an array. Each element in the array is of the varchar type.	Example: Latency, QPS, and NetFlow.
indexName	The name of the target metric. The value of this parameter is of the varchar type.	Example: Latency.
threshold	The threshold of correlation between the target metric and the metrics that you want to analyze.	Valid values: [0,1].

- Query statement

```
* | select ts_association_analysis(
              time,
              array[inflow, outflow, latency, status],
              array['inflow', 'outflow', 'latency', 'status'],
              'latency',
              0.1) from log;
```

- Query result

```
| results               |
| --------------------- |
| ['latency', '1.0']    |
| ['outflow', '0.6265'] |
| ['status', '0.2270']  |
```

- Description of the query result
  - name: the name of the metric that meets the specified correlation condition of the target metric.
  - score: the value of correlation between the returned metric and the target metric. Valid values: [0, 1].

ts_similar

Syntax 1

```
select ts_similar(stamp, value, ts, ds)
select ts_similar(stamp, value, ts, ds, metricType)
```

The following table describes the parameters in the function.

Parameter	Description	Value
stamp	The UNIX timestamp that is of the long type.	None

Parameter	Description	Value
value	The value of the metric that you want to analyze. The value of this parameter is of the double type.	None
ts	The time sequence of the specified time series curve. The value of this parameter is an array. Each element in the array is of the double type.	None
ds	The sequence of numeric data of the specified time series curve. The value of this parameter is an array. Each element in the array is of the double type.	None
metricType	The type of correlation between the measured curves. The value of this parameter is of the varchar type. Valid values: SHAPE, RMSE, PEARSON, SPEARMAN, R2, and KENDALL.	Example: SHAPE.

Syntax 2

```
select ts_similar(stamp, value, startStamp, endStamp, step, ds)
select ts_similar(stamp, value, startStamp, endStamp, step, ds, metricType )
```

The following table describes the parameters in the function.

Parameter	Description	Value
stamp	The UNIX timestamp that is of the long type.	None
value	The value of the metric that you want to analyze. The value of this parameter is of the double type.	None
startStamp	The start timestamp of the specified time series curve. The value of this parameter is of the long type.	None
endStamp	The end timestamp of the specified time series curve. The value of this parameter is of the long type.	None
step	The time interval between two adjacent data points in a time series. The value of this parameter is of the long type.	None
ds	The sequence of numeric data of the specified time series curve. The value of this parameter is an array. Each element in the array is of the double type.	None

Parameter	Description	Value
metricType	The type of correlation between the measured curves. The value of this parameter is of the varchar type. Valid values: SHAPE, RMSE, PEARSON, SPEARMAN, R2, and KENDALL.	Example: SHAPE.

- Query statement

```
* | select vhost, metric, ts_similar(time, value, 1560911040, 1560911065, 5, array[5.1,4.0,3.3,5.6,4.0,7.2], 'PEARSON') from log  group by vhost, metric;
```

- Query result

```
| vhost  | metric          | score               |
| ------ | --------------- | ------------------- |
| vhost1 | redolog         | -0.3519082537204182  |
| vhost1 | kv_qps          | -0.15922168009772697 |
| vhost1 | file_meta_write | NaN                 |
```

- Description of the query result

  score: the correlation between the analyzed metric and the specified time series curve. Valid values: [-1, 1].

23.1.4.9.13. Kernel density estimation function

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable.

The Kernel density estimation function uses the smooth peak function to fit the observed data points. In this way, the function simulates the real probability distribution curve.

- Syntax

```
select kernel_density_estimation(bigint stamp, double value, varchar kernelType)
```

- Parameters

Parameter	Description
stamp	The Unix timestamp of observed data. Unit: second.
value	The observed value.
kernelType	○ box: rectangle window. ○ epanechniov: Epanechnikov curve. ○ gausener: Gaussian curve.

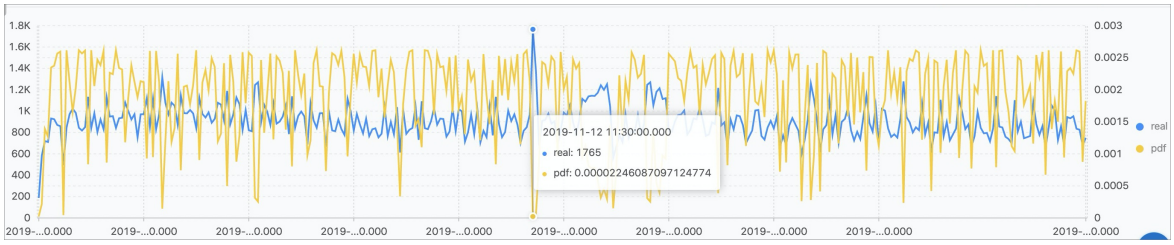
- Response

Display item	Description
unixtime	The Unix timestamp of observed data.
real	The observed value.
pdf	The probability of each observed data point.

- Example
  - Sample statement

```
* |
select
    date_trunc('second', cast(t1[1] as bigint)) as time, t1[2] as real, t1[3] as pdf from (
        select kernel_density_estimation(time, num, 'gaussian') as res from (
            select __time__ - __time__ % 10 as time, COUNT(*) * 1.0 as num from log group by time
order by time)
        ), unnest(res) as t(t1)  limit 1000
```

  - Response



23.1.4.10. Advanced analysis

23.1.4.10.1. Optimize queries

This topic describes how to optimize queries to improve query efficiency.

Increase the number of shards

More shards indicate more computing resources and faster computing speed. You can increase the number of shards to ensure that the average number of log entries that are scanned in each shard does not exceed 50 million. You can increase the number of shards by splitting shards. For more information, see Split a shard.

Reduce the query time range and data volume

- A larger time range means a slower query. If you query data within a year or a month, data is computed on a daily basis. To improve the computing speed, you can reduce the query time range.
- Larger data volumes slow down queries. Reduce the amount of data that you want to query as much as possible.

Repeat queries multiple times

If you find that the result of a query is inaccurate, you can repeat the query multiple times. The underlying acceleration mechanism ensures that each query uses the previous query result to analyze data. This way, multiple queries can improve the accuracy of the query result.

Optimize SQL statements for queries

A time-consuming query statement has the following characteristics:

- Uses GROUP BY clauses to group string-formatted columns.
- Use GROUP BY clauses to group more than five columns.
- Includes operations that generate strings.

We recommend that you use the following methods to optimize SQL statements for queries:

- Avoid operations that generate strings if possible.

- If you use the date_format function to generate a formatted timestamp, the query is inefficient.

  ```
  * | select date_format(from_unixtime(__time__) , '%H_%i') as t, count(1) group by t
  ```

- If you use the substr() function, strings are generated. We recommend that you use the date_trunc or time_series function in a query statement.

- Avoid using GROUP BY clauses to group string-formatted columns if possible.

  If you use a GROUP BY clause to group strings, a large number of hash calculations are required. The number of the hash calculations account for more than 50% of the total number of calculations. The following example shows two query statements:

  ```
  * | select count(1) as pv , date_trunc('hour',__time__) as time group by time
  * | select count(1) as pv , from_unixtime(__time__-__time__%3600) as time group by __time__-__time_
  _%3600
  ```

  Query 1 is less efficient than query 2 because query 1 needs to hash strings.

  - Query 1 and query 2 calculate the total number of log entries per hour.
  - Query 1 converts the time to a string, for example, 2017-12-12 00:00:00, and then uses a GROUP BY clause to group the string.
  - Query 2 calculates the on-the-hour time value, uses a GROUP BY clause to group the result, and then converts the value to a string.

- List fields alphabetically based on the initial letter when you use a GROUP BY clause to group multiple columns.

  For example, you need to query 100 million users who are from 13 provinces.

  ```
  Fast: * | select province,uid,count(1)groupby province,uid
  Slow: * | select province,uid,count(1) group by uid,province
  ```

- Use estimating functions.

  Estimating functions provide better performance than accurate calculation. In estimation, accuracy is compromised to an acceptable level to achieve fast calculation.

  ```
  Fast: * |select approx_distinct(ip)
  Slow: * | select count(distinct(ip))
  ```

- Specify only the required columns in an SQL statement if possible.

  When you use an SQL statement to query data, specify only the required columns to speed up the calculation.

  ```
  Fast: * | select a,b c
  Slow: * | select *
  ```

- Specify columns that do not need to be grouped in an aggregate function if possible.

  For example, a user ID is associated with a username. Therefore, you can use a GROUP BY clause to group data by userid.

  ```
  Fast: * | select userid, arbitrary(username), count(1)group by userid
  Slow: * | select userid, username, count(1) group by userid,username
  ```

- Avoid using the IN operator if possible.

  Use the OR operator in SQL statements instead of the IN operator if possible.

  ```
  Fast: key : a or key :b or key:c | select count(1)
  Slow: * | select count(1) where key in ('a','b')
  ```

23.1.4.10.2. Use cases

This topic describes some use cases of data analysis in Log Service.

Cases

- Trigger an alert if the error rate exceeds 40% over the last 5 minutes
- Calculate the amount of transferred data and configure alerts
- Calculate the average latency of traffic data in different sizes
- Obtain the percentages of different results
- Calculate the number of log entries that meet the query condition

Trigger an alert if the error rate exceeds 40% over the last 5 minutes

Calculate the percentage of 500 Internal Server Error every minute. If the error rate exceeds 40% over the last 5 minutes, an alert is triggered.

```
status :500 |
select
  __topic__,
  max_by(error_count, window_time) / 1.0 / sum(error_count) as error_ratio,
  sum(error_count) as total_error
FROM  (
    select
      __topic__,
      count(*) as error_count,
      __time__ - __time__ % 300 as window_time
    FROM      log
    group by
      __topic__,
      window_time
  )
group by
  __topic__
having
  max_by(error_count, window_time) / 1.0 / sum(error_count) > 0.4
  and sum(error_count) > 500
order by
  total_error desc
limit
  100
```

- You can use the following clause to calculate the error rate: `max_by(error_count,window_time)/1.0/sum(error_count) as error_ratio` .
- You can use the following clause to calculate the total number of 500 Internal Server Error: `sum(error_count) as total_error` .
- You can use the following clause to query the number of errors every 5 minutes: `select __topic__, count(*) as error_count , __time__ - __time__ % 300 as window_time from log group by __topic__, window_time` .

Calculate the amount of transferred data and configure alerts

Calculate the amount of transferred data every minute. If the amount of transferred data sharply decreases, an alert is triggered. Transferred data counted in the last minute does not cover a full minute. The `(max(time) - min(time))` clause is used for normalization to count the average traffic per minute.

```
* |
SELECT
  SUM(inflow) / (max(__time__)-min(__time__)) as inflow_per_minute,
  date_trunc('minute', __time__) as minute
group by
  minute
```

Calculate the average latency of traffic data in different sizes

Distribute traffic data to multiple buckets based on the data size and calculate the average latency of the data in each bucket.

```
* |
select
  avg(latency) as latency,
  case
    when originSize < 5000 then 's1'
    when originSize < 20000 then 's2'
    when originSize < 500000 then 's3'
    when originSize < 100000000 then 's4'
    else 's5'
  end as os
group by
  os
```

Obtain the percentages of different results

Obtain the number and percentage of each count result for different departments. The following query statement includes subqueries and window functions. The `sum(c) over()` clause is used to calculate the sum of values in all rows.

```
* |
select
  department,
  c * 1.0 / sum(c) over ()
from(
    select
      count(1) as c,
      department
    FROM      log
    group by
      department
  )
```

Calculate the number of log entries that meet the query condition

Use the count_if clause to calculate the number of URLs that meet specified conditions and obtain the number of URLs that meet each condition by minute.

```
* |
select
  count_if(uri like '%login') as login_num,
  count_if(uri like '%register') as register_num,
  date_format(date_trunc('minute', __time__), '%m-%d %H:%i') as time
group by
  time
order by
  time
limit
  100
```

- You can use the following clause to calculate the number of URLs that end with login: `count_if(uri like '%login')`.
- You can use the following clause to calculate the number of URLs that end with register: `count_if(uri like '%register')`.

23.1.4.10.3. Examples of time field conversion

In most cases, you need to process the time fields in log data when you query and analyze the log data. For example, you need to convert a timestamp to a specified time format. This topic provides some examples on how to convert the values of time fields.

A log may contain multiple fields that record points in time for different events. Examples:

- `__time__`: records the time when you call the API or use an SDK to write log data. You can use this field when you ship, query, and analyze log data.
- Original time field in log data: records the time when the log data is generated. This field exists in raw logs.

Time fields in different formats are difficult to read. To simplify the read process, you can convert the time format when you query and analyze log data. Examples:

1. Convert the value of __time__ to a timestamp
2. Display the value of __time__ in a specified format
3. Convert the time in a log to a specified format

Convert the value of __time__ to a timestamp

You can use the from_unixtime function to convert the value of the `__time__` field to a timestamp.

```
* | select from_unixtime(__time__)
```

Display the value of __time__ in a specified format

To display the value of the `__time__` field in the format of `YYYY-MM-DD HH:MM:SS`, you can use the date_format function.

```
* | select date_format(__time__, '%Y-%m-%d %H:%i:%S')
```

Convert the time in a log to a specified format

To convert the value of the time field in a log to a specified format, such as `YYYY-MM-DD HH:MM:SS`, and then perform the GROUP BY operation on the `YYYY-MM-DD` part, you can use the date_format function.

- Sample log

```
__topic__:
body_byte_sent:  307
hostname:  www.host1.com
http_user_agent:  Mozilla/5.0 (iPhone; CPU iPhone OS 10_3_3 like Mac OS X) AppleWebKit/603.3.8 (KHT
ML, like Gecko) Mobile/14G60 QQ/7.1.8.452 V1_IPH_SQ_7.1.8_1_APP_A Pixel/750 Core/UIWebView NetType/
WIFI QBWebViewType/1
method:  GET
referer:  www.host0.com
remote_addr:  36.63.1.23
request_length:  111
request_time:  2.705
status:  200
upstream_response_time:  0.225582883754
url:  /?k0=v9&
time:2017-05-17 09:45:00
```

- Sample SQL statement

```
* | select date_format (date_parse(time,'%Y-%m-%d %H:%i:%S'), '%Y-%m-%d') as day, count(1) as uv gr
oup by day order by day asc
```

23.1.4.11. Visual analysis

23.1.4.11.1. Analysis graph

23.1.4.11.1.1. Chart overview

Log Service allows you to render query and analysis results into visualized charts.

Prerequisites

Indexes are configured and the analysis feature is enabled. To enable the analysis feature, turn on **Enable Analytics** for the fields in the **Search & Analysis** panel. For more information, see Enable the index feature and configure indexes for a Logstore.

> **Note**
> - Before you configure charts, we recommend that you are familiar with the log analysis feature. For more information, see Log analysis overview.
> - You must specify an analytic statement in a query statement. Log Service cannot display charts based on query results.

Usage notes

When you execute multiple query statements in sequence, the **Value Column**, **X Axis**, or **Y Axis** configurations are not automatically modified based on the current query statement. The X-axis and Y-axis configurations may remain the same as the configurations in the previous query statement. In this case, the query and analysis result of the current query statement cannot be automatically displayed on a chart. If the following errors occur, you must modify the **Properties** settings based on the current query statement:

- The dimensions that you selected are not in the query results. Check and modify the Properties settings.
- X Axis or Y Axis is unavailable. Check and modify the Properties settings.

Chart configurations

To go to the Graph tab, perform the following steps:

1. Log on to the Log Service console.

2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Click **Graph** to go to the Graph tab.

On the **Graph** tab, multiple charts are provided to display query and analysis results. You can select a type of chart from the chart bar based on your business requirements.

- On the **Graph** tab, query and analysis results are displayed in the **Chart Preview** and **Data Preview** sections. The **Chart Preview** section shows the query and analysis results that are displayed in the specified type of chart. The **Data Preview** section shows the data of the related chart in a table.
- On the **Graph** tab, you can also configure the following settings:
  - On the **Data Source** tab, you can specify placeholder variables. For example, you can configure the drill-down event of Chart A to redirect to the dashboard on which Chart B is located. After you configure the drill-down event of Chart A, the placeholder variable is replaced by the variable that you click to trigger the drill-down event and execute the query statement of Chart B. This way, to trigger the drill-down event, you must click the placeholder variable that you configured for Chart B. For more information, see Drill-down analysis.

    This feature applies if you need to configure drill-down events to redirect to destination dashboards.
  - On the **Properties** tab, you can configure the chart properties that you want to display. You can set the X-axis, left Y-axis and right Y-axis, margins, and other parameters. Different types of charts have different properties. For more information, see the related topic of each chart.

    This feature is applicable to all query and analysis scenarios.
  - On the **Interactive Behavior** tab, you can configure drill-down events for a chart. Then, you can click the variable value in the chart to trigger the specified drill-down event. For more information, see Drill-down analysis.

    This feature applies if you need to trigger drill-down events for charts.

Supported chart types

- Table
- Line chart
- Column chart
- Bar chart
- Pie chart
- Area chart
- Individual value plot
- Progress bar
- Map
- Flow diagram
- Sankey diagram
- Word cloud
- Treemap chart

23.1.4.11.1.2. Display query results in a table

Tables are used to sort and display data for quick reference and analysis. All query results that match specified query statements can be rendered into visualized charts. By default, query results are displayed in a table.

Components

- Table header
- Row

- Column

Where:

- The number of columns can be specified by using a `SELECT` statement.
- The number of rows is calculated based on the number of log entries in a specified time range. The default clause is `LIMIT 100`.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. View the query result.

   By default, the query result is displayed in a table on the **Graph** tab.
6. On the **Properties** tab, configure the properties of the table.

   If you want to modify the rows and columns of the table or the entries to return on each page, you can set the parameters on the Properties tab.

Parameter	Description
Items per Page	The number of entries to return on each page.
Zebra Striping	Specifies whether to display the query result in a zebra-striped table.
Transpose Rows and Columns	Specifies whether to transpose rows and columns.
Hide Reserved Fields	Specifies whether to hide reserved fields, such as `__time__` and `__source__`.
Disable Sorting	Specifies whether to disable the sorting feature.
Disable Search	Specifies whether to disable the search feature.
Highlight Settings	If you turn on Highlight Settings, you can create rules to highlight matched rows or columns.
Sparkline	If you turn on **Sparkline**, you can add an area chart, a line chart, or a column chart for columns in the table.

Example

To query and analyze the distribution of page views (PVs) for different users based on status, execute the following query statement:

```
* |select Status,AlertDisplayName as name,COUNT(*) as count group by Status,name
```

23.1.4.11.1.3. Display query results on a line chart

This topic describes how to configure a line chart to display query results.

Background information

Line charts are used to analyze the changes of field values based on an ordered data type. In most cases, the analysis is based on a specified time range. You can use a line chart to analyze the following change characteristics of field values over a specified period of time:

- Increment or decrement
- Increment or decrement rate
- Increment or decrement pattern, for example, periodicity
- Peak value and bottom value

You can use line charts to analyze the changes of field values over a specified period of time. You can also use line charts to analyze the changes of multiple field values in multiple lines over the same period. This way, you can analyze the relationship between different fields. For example, the values of a field are proportional or inversely proportional to the values of another field.

Each line chart consists of the following elements:

- X-axis
- Left Y-axis
- Right Y-axis (optional)
- Data point
- Line of trend changes
- Legend

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the line chart.

Parameter	Description
X Axis	The sequential data. In most cases, a time series is selected.
Left Y Axis	The numeric data. You can select one or more fields for the left Y-axis.
Right Y Axis	The numeric data. You can select one or more fields for the right Y-axis. The layer of the right Y-axis is higher than the layer of the left Y-axis.
Column Marker	The column on the left or right Y-axis. The column is displayed as a histogram.
Legend	The position of the legend in the chart. Valid values: Top, Bottom, Left, and Right.
Format Left Y-axis	The format in which the data on the left and right Y-axis is displayed.
Format Right Y-axis	
Line Type	The type of line that is displayed in the line chart. Valid values: **Straight Line** and **Curve**.

Parameter	Description
Anomaly Point Column	The column where anomaly points are located. You can set the **Anomaly Point Lower Limit** and **Anomaly Point Upper Limit** parameters for a column. ○ **Anomaly Point Lower Limit**: Values that are less than the lower limit are highlighted in red. ○ **Anomaly Point Upper Limit**: Values that exceed the upper limit are highlighted in red.
Upper Limit Column	The area that is formed based on the values.
Lower Limit Column	
Time Series	A series of data points that are listed in chronological order.
Time Format	The time format of the time series fields.
Margin	The distance between the axis and the borders of the chart. The parameters include Top Margin, Bottom Margin, Left Margin, and Right Margin.

Note Each line in a line chart must contain more than two data points. Otherwise, the data trend cannot be generated. We recommend that you select no more than five lines for a line chart.

Example of a simple line chart

To query the page views (PVs) of the IP address `203.0.113.10` in the previous 24 hours, execute the following query statement:

```
remote_addr: 203.0.113.10 | select date_format(date_trunc('hour', __time__), '%m-%d %H:%i')
as time, count(1) as PV group by time order by time limit 1000
```

Select `time` for X Axis and `PV` for Left Y Axis. Set the Legend parameter and adjust the margins based on your business requirements.

Example of a dual Y-axis line chart

To query the PVs and number of unique visitors (UVs) in the previous 24 hours, execute the following query statement:

```
* | select date_format(date_trunc('hour', __time__), '%m-%d %H:%i') as time, count(1) as PV, approx_di
stinct(remote_addr) as UV group by time order by time limit 1000
```

Select `time` for X Axis, `PV` for Left Y Axis, `UV` for Right Y Axis, and `PV` for Column Marker.

23.1.4.11.1.4. Display query results on a column chart

This topic describes how to configure a column chart to display query results.

Background information

A column chart uses vertical or horizontal bars to show the values of different categories. You can use a column chart to count the number of values in each category.

Each column chart consists of the following elements:

- X-axis (horizontal)

- Y-axis (vertical)
- Rectangular bar
- Legend

By default, column charts in Log Service use vertical bars. Each rectangular bar has a fixed width and a variable height that indicates a value. If you select multiple columns of data for the Y-axis, a grouped column chart is used to display the data.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the column chart.

> **Note** If a query statement returns no more than 20 log entries, you can use a column chart to display the query results. You can use a `LIMIT` clause to limit the number of rectangular bars. If the chart contains a large number of rectangular bars, the analysis results may not be displayed as expected. We recommend that you select no more than five fields for the Y-axis.

Parameter	Description
X Axis	The categorical data.
Y Axis	The numeric data. You can select one or more fields for the left Y-axis.
Legend	The position of the legend in the chart. Valid values: Top, Bottom, Left, and Right.
Format	The format in which the data on the Y-axis is displayed.
Margin	The distance between the axis and the borders of the chart. The parameters include Top Margin, Bottom Margin, Left Margin, and Right Margin.

Example of a simple column chart

To query the number of visits for each `http_referer` in the specified time range, execute the following query statement:

```
* | select http_referer, count(1) as count group by http_referer
```

Select `http_referer` for X Axis and `count` for Y Axis.

Example of a grouped column chart

To query the number of visits and the average bytes for each `http_referer` in the specified time range, execute the following query statement:

```
* | select http_referer, count(1) as count, avg(body_bytes_sent) as avg group by http_referer
```

Select `http_referer` for X Axis. Select `count` and `avg` for Y Axis.

23.1.4.11.1.5. Display query results on a bar chart

This topic describes how to configure a bar chart to display query results.

Background information

A bar chart is a horizontal column chart that is used to analyze the top N values of fields. You can configure a bar chart in a similar manner in which you configure a column chart.

Each bar chart consists of the following elements:

- X-axis (vertical)
- Y-axis (horizontal)
- Rectangular bar
- Legend

Each rectangular bar has a fixed height and a variable width. The variable width indicates a value. If multiple columns of data are mapped to the Y-axis, you can use a grouped bar chart to display the data.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the bar chart.

> **Note**
> - If a query statement returns no more than 20 log entries , you can use a bar chart to display the query results. You can also use a LIMIT clause to limit the number of rectangular bars. If the chart contains a large number of rectangular bars, analysis results may not be displayed as expected. You can also use an `ORDER BY` clause to analyze Top N values of fields. We recommend that you select no more than five fields for the Y-axis.
> - You can use a grouped bar chart to display query results. However, the values represented by each rectangular bar in a group must be positively or negatively associated with each other.

Parameter	Description
X Axis	The categorical data.
Y Axis	The numeric data. You can select one or more fields for the Y-axis.
Legend	The position of the legend in the chart. Valid values: Top, Bottom, Left, and Right.
Format X-axis	The format in which the data on the X-axis is displayed.
Margin	The distance between the axis and the borders of the chart. The parameters include Top Margin, Bottom Margin, Left Margin, and Right Margin.

Example

To analyze the top 10 request URIs (`request_uri`) that are most frequently visited, execute the following query statement:

```
* | select request_uri, count(1) as count group by request_uri order by count desc limit 10
```

23.1.4.11.1.6. Display query results on a pie chart

This topic describes how to configure a pie chart to display query results.

Background information

A pie chart is used to show the percentages of different categories of data. The arc length of each segment in a pie chart is proportionate to the quantity that is represented by each category. A pie chart is divided into multiple segments based on the percentages of categories. Each segment shows the percentage of a category. The sum of all percentages is equal to 100%.

Each pie chart consists of the following elements:

- Segment
- Percentage in the text format
- Legend

Types

Log Service provides the following types of standard pie charts: pie chart, donut chart, and polar area chart.

- Donut chart

  A donut chart is a variant of a pie chart that has a hollow center. Compared with a pie chart, a donut chart provides the following advantages:

  - Displays more information, such as the total number of occurrences of all field values.
  - Allows you to compare data between two donut charts based on ring lengths. Data across different pie charts is difficult to compare.

- Polar area chart

  A polar area chart is a column chart in the polar coordinate system. Each category of data is represented by a segment with the same angle, and the radius of each segment varies based on the value. Compared with a pie chart, a polar area chart provides the following advantages:

  - If a query statement returns no more than 10 log entries, you can use a pie chart to display the query results. If a query statement returns 10 to 30 log entries, you can use a polar area chart to display the query results.
  - Enlarges the differences among the values of categories because the area of the segment correlates with the square of the radius. Therefore, the polar area chart is suitable for the comparison of similar values.
  - A circle can be used to display a periodic pattern. Therefore, you can use a polar area chart to analyze value changes in different periods, such as weeks and months.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the pie chart.

> **Note**
>
> - If a query statement returns no more than 10 log entries, you can use a pie chart or donut chart to display the query results. You can use a `LIMIT` clause to limit the number of segments. If a chart contains a large number of segments with different colors, the analysis results may not be displayed as expected.
> - If the number of log entries exceeds 10, we recommend that you use a polar area chart or column chart.

Parameter	Description
Chart Types	The type of the chart. Valid values: Pie Chart, Donut Chart, and Polar Area Chart.
Legend Filter	The categorical data.
Value Column	The values that correspond to different categories of data.
Legend	If you turn on **Show Legend**, you can set this parameter to adjust the position of the legend in the chart.
Format	The format in which data is displayed.
Tick Text Format	Valid values: **Percentage** and **Category: Percentage**.
Margin	The distance between the axis and the borders of the chart. The parameters include Top Margin, Bottom Margin, Left Margin, and Right Margin.

Example of a pie chart

To analyze the percentages of the `requestURI` field values, execute the following query statement:

```
* | select requestURI as uri , count(1) as c group by uri limit 10
```

Example of a donut chart

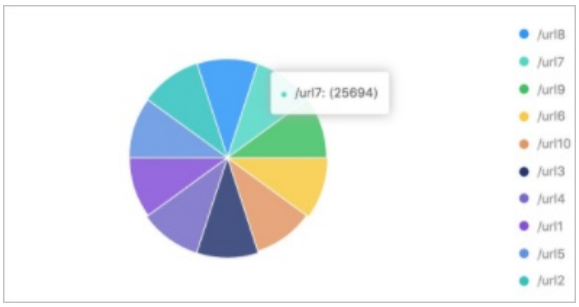
To analyze the percentages of the `requestURI` field values, execute the following query statement:

```
* | select requestURI as uri , count(1) as c group by uri limit 10
```

Example of a polar area chart

To analyze the percentages of the `requestURI` field values, execute the following query statement:

```
* | select requestURI as uri , count(1) as c group by uri limit 10
```



23.1.4.11.1.7. Display query results on an area chart

This topic describes how to configure an area chart to display query results.

Background information

An area chart is built based on a line chart. The colored section between a line and the axis is an area. The color is used to highlight the trend. An area chart is similar to a line chart and shows the changes of numeric values over a specified period of time. An area chart is used to highlight the overall data trend. Both line charts and area charts display the trend and relationship between numeric values instead of specific values.

Each area chart consists of the following elements:

- X-axis (horizontal)
- Y-axis (vertical)
- Area segment

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the icon.
6. On the **Properties** tab, configure the properties of the area chart.

> **Note** In an area chart, a single area segment must contain more than two data points. If a single area segment contains two or fewer data points, the data trend cannot be analyzed. We recommend that you select less than five area segments in an area chart.

Parameter	Description
X Axis	The sequential data. In most cases, a time series is selected.
Y Axis	The numeric data. You can select one or more fields for the Y-axis.
Legend	The position of the legend in the chart. Valid values: Top, Bottom, Left, and Right.
Format	The format in which data is displayed.

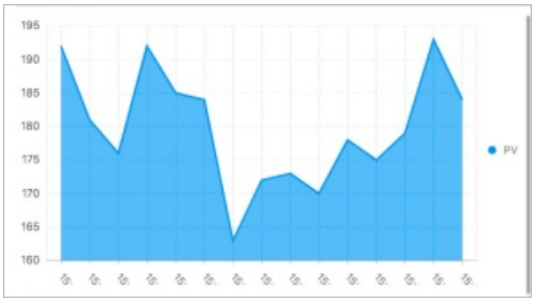
Parameter	Description
Margin	The distance between the axis and the borders of the chart. The parameters include Top Margin, Bottom Margin, Left Margin, and Right Margin.

Example of a simple area chart

To query the page views (PVs) of the IP address `203.0.113.10` in the previous 24 hours, execute the following query statement:

```
remote_addr: 203.0.113.10 | select date_format(date_trunc('hour', __time__), '%m-%d %H:%i') as time, count(1) as PV group by time order by time limit 1000
```

Select `time` for X Axis and `PV` for Y Axis.

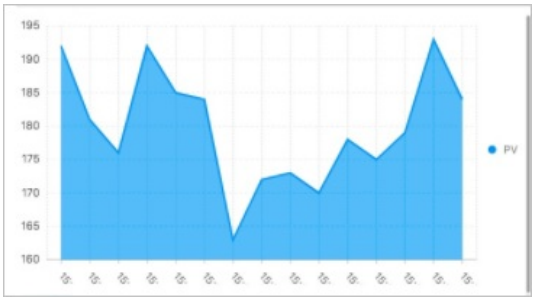


Example of a cascade chart

To query the number of PVs and the number of unique visitors (UVs) for each hour within one day, execute the following query statement:

```
* | select date_format(date_trunc('hour', __time__), '%m-%d %H:%i') as time, count(1) as PV, approx_distinct(remote_addr) as UV group by time order by time limit 1000
```

Select `time` for X Axis. Select `PV` and `UV` for Y Axis.



23.1.4.11.1.8. Display query results on a single value chart

This topic describes how to configure a single value chart to display query results.

Background information

A single value chart highlights a single value. Log Service supports the following types of single value charts:

- Rectangle Frame: shows a general value.
- Dial: shows the difference between the current value and a specified threshold value.
- Compare Numb Chart: shows the SQL query results of interval-valued comparison and periodicity-valued comparison functions. For more information, see Interval-valued comparison and periodicity-valued comparison

functions.

By default, Rectangle Frame is selected. Rectangle Frame is the most basic type of single value chart to display data at a specified point. In most cases, this chart type is used to show the key information at a specified point in time. To display a proportional metric, you can select Dial.

Each single value chart consists of the following elements:

- Numeric value
- Unit (optional)
- Description (optional)
- Chart type

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the 123 icon.
6. On the **Properties** tab, configure the properties of the single value chart.

   > **Note** Log Service normalizes data in charts that contain numeric values. For example, `230000` is processed as `230K`. You can include mathematical calculation functions in query statements to customize numeric formats. For more information, see Mathematical calculation functions.

   - The following table describes the parameters of a rectangle frame.

Parameter	Description
Chart Types	The type of the single value chart. If you select **Rectangle Frame**, query results are displayed in a rectangle frame.
Value Column	The value that is displayed in the chart. By default, the data in the first row of the specified column is displayed.
Unit	The unit of the data.
Unit Font Size	The font size of the unit. You can drag the slider to adjust the font size. Valid values: 10 to 100. Unit: pixels.
Description	The description of the value.
Description Font Size	The font size of the description. You can drag the slider to adjust the font size. Valid values: 10 to 100. Unit: pixels.
Format	The format in which data is displayed.
Font Size	The font size of the value. You can drag the slider to adjust the font size. Valid values: 10 to 100. Unit: pixels.
Font Color	The color of the value, unit, and description in the chart. You can use the default color or select a color.
Background Color	The color of the background. You can use the default color or select a color.

- The following table describes the parameters of a dial.

Parameter	Description
Chart Types	The type of the single value chart. If you select **Dial**, query results are displayed on a dial.
Actual Value	The value that is displayed in the chart. By default, the data in the first row of the specified column is displayed.
Unit	The unit of the value on the dial.
Font Size	The font size of the value and unit. Valid values: 10 to 100. Unit: pixels.
Description	The description of the value.
Description Font Size	The font size of the description. You can drag the slider to adjust the font size. Valid values: 10 to 100. Unit: pixels.
Dial Maximum	The maximum value of the scale on the dial. Default value: 100.
Use Query Results	If you turn on Use Query Results, Dial Maximum is replaced by Maximum Value Column. Then, you can select the maximum value from query results for this parameter.
Format	The format in which data is displayed.
Colored Regions	The number of segments that divide the dial. Each segment is displayed in a different color. Valid values: 2, 3, 4, and 5. Default value: 3.
Region Max Value	The maximum value of the scale in each colored segment of the dial. By default, the maximum value in the last segment is the maximum value on the dial. You do not need to specify this value. **Note** By default, a dial is evenly divided into three colored segments. If you change the value of **Colored Regions**, Region Max Value is not automatically adjusted. You can manually set the maximum value for each colored segment based on your business requirements.
Font Color	The color of the value on the dial.
Region	The colored segments that divide the dial. By default, a dial is evenly divided into three segments. The segments are displayed in blue, yellow, and red. If you set **Colored Regions** to a value greater than 3, the added segments are displayed in blue by default. You can change the color of each segment.

- The following table describes the parameters of a compare numb chart.

Parameter	Description
Chart Types	The type of the single value chart. If you select Compare Numb Chart, query results are displayed on a compare numb chart.

Parameter	Description
Show Value	The value that is displayed in the center of the compare numb chart. In most cases, this value is set to the statistical result that is calculated by the related comparison function in the specified time range.
Compare Value	The value that is compared with the threshold. In most cases, this value is set to the result of the comparison between the statistical results that are calculated by the related comparison function in the specified time range and in the previously specified time range.
Font Size	The font size of the show value. Valid values: 10 to 100. Unit: pixels.
Unit	The unit of the show value.
Unit Font Size	The font size of the unit for the show value. Valid values: 10 to 100. Unit: pixels.
Compare Unit	The unit of the compare value.
Compare Font Size	The font size of the compare value and unit. Valid values: 10 to 100. Unit: pixels.
Description	The description of the show value and its growth trend.
Description Font Size	The font size of the description. Valid values: 10 to 100. Unit: pixels.
Trend Comparison Threshold	The value that is used to measure the variation trend of the compare value. For example, the compare value is -1. ■ If you set **Trend Comparison Threshold** to 0, a down arrow that indicates a value decrease is displayed on the page. ■ If you set **Trend Comparison Threshold** to -1, the value remains unchanged. The system does not display the trend on the page. ■ If you set **Trend Comparison Threshold** to -2, an up arrow that indicates a value increase is displayed on the page.
Format	The format in which data is displayed.
Font Color	The color of the show value and its description.
Growth Font Color	The font color of the compare value that is greater than the threshold.
Growth Background Color	The background color that is displayed when the compare value is greater than the threshold.
Decrease Font Color	The font color that is displayed when the compare value is less than the threshold.
Decrease Background Color	The background color that is displayed when the compare value is less than the threshold.
Equal Background Color	The background color that is displayed when the compare value is equal to the threshold.

Examples

To view the number of page views (PVs), execute the following query statements. The analysis results are displayed in charts.

- Rectangle frame

  To view the number of PVs in the previous 15 minutes, execute the following query statement:

  ```
  * | select count(1) as pv
  ```

- Dial

  To view the number of PVs in the previous 15 minutes, execute the following query statement:

  ```
  * | select count(1) as pv
  ```

- Compare numb chart

  To view and compare the PVs on the current day and in the previous day, execute the following query statement:

  ```
  * | select diff[1],diff[2], diff[1]-diff[2] from (select compare( pv , 86400) as diff from (select count(1) as pv from log))
  ```

23.1.4.11.1.9. Display query results on a progress bar

This topic describes how to configure a progress bar to display query results.

Background information

A progress bar shows the percentage of the actual value of a field to the maximum value of the field. You can configure the properties of a progress bar to change the style and configure display rules for the progress bar.

Each progress bar consists of the following elements:

- Actual value
- Unit (optional)
- Total value

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the progress bar.

Parameter	Description
Actual Value	The actual value in the chart. By default, data in the first row of the specified column is displayed.
Unit	The unit of the value in the progress bar.
Total Value	The maximum value indicated by the progress bar. Default value: 100.

Parameter	Description
Maximum Value Column	The maximum value in the specified column. If you turn on **Use Query Results**, **Total Value** is replaced by **Maximum Value Column**. Then, you can select the maximum value from the query results for this parameter.
Use Query Results	If you turn on Use Query Results, Total Value is replaced by Maximum Value Column. Then, you can select the maximum value from the query results for this parameter.
Edge Shape	The edge shape of the progress bar.
Vertical Display	Specifies whether to display the progress bar in vertical display mode.
Font Size	The font size of the value in the progress bar.
Thickness	The thickness of the progress bar.
Background Color	The background color of the progress bar.
Font color	The font color of the value in the progress bar.
Default Color	The default color of the progress bar.
Color Display Mode	The display mode of the progress bar.
Start Color	The start color of the progress bar. This parameter is available if you select **Gradient** for **Color Display Mode**.
End Color	The end color of the progress bar. This parameter is available if you select **Gradient** for **Color Display Mode**.
Display Color	The display color of the progress bar. This parameter is available if you select **Display by Rule** for **Color Display Mode**. **Note** The value of **Actual Value** is compared with the value of **Threshold** based on the condition specified by **Operator**. If the actual value matches the condition specified by **Operator**, the progress bar is displayed in the color specified by Display Color. If the actual value does not match the condition, the progress bar is displayed in the default color.
Operator	The condition that is used to determine whether to display the progress bar in the color specified by Display Color. This parameter is available if you select **Display by Rule** for **Color Display Mode**.
Threshold	The threshold based on which the color of the progress bar is determined. This parameter is available if you select **Display by Rule** for **Color Display Mode**.

Example

To calculate the ratio of the page views (PVs) of the current hour to the PVs of the same period of time on the previous day, execute the following query statement:

```
* | SELECT diff[1] AS today, diff[2] AS yesterday, diff[3] AS ratio FROM (SELECT compare(PV,86400) AS diff FROM (SELECT count(*) AS PV FROM log))
```

23.1.4.11.1.10. Display query results on a map

This topic describes how to configure a map to display query results.

Background information

You can color and mark a map to display geographic data. Log Service provides the map of China. The display modes of an AMap include the anchor point and heat map. You can use specific functions in query statements to display analysis results as maps.

Each map consists of the following elements:

- Map canvas
- Colored area

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analytics**.
5. On the **Graph** tab, use the map of China, click the  icon..
6. On the **Properties** tab, configure the properties of the map.

   In this example, the map of China is selected.

Parameter	Description
Provinces	The location information that is recorded in logs. The information is displayed in one of the following dimensions based on the map type:
Value Column	The amount of data at the location.
Show Legend	If you turn on Show Legend, the legend information is displayed.

Example of a map of China

To display query results on a map of China, you can execute the following query statement in which the ip_to_province function is used:

```
* | select ip_to_province(remote_addr) as address, count(1) as count group by address order by count desc limit 10
```

Select *address* for Provinces and *count* for Value Column.

23.1.4.11.1.11. Display query results on a flow chart

This topic describes how to configure a flow chart to display query results.

Background information

A flow chart, also known as ThemeRiver, is a stacked area chart around a central axis. The banded branches with different colors indicate different categorical data. The width of the band indicates the numeric value. By default, the time information of the data is mapped to the X-axis. A flow chart can display the data in three dimensions.

You can select Line Chart or Column Chart for the Chart Types parameter. If you select Column Chart, a stacked column chart is displayed. In a stacked column chart, each category of data starts from the top of the last column of categorical data.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the flow chart.

Parameter	Description
Chart Types	The type of the chart. Valid values: Line Chart, Area Chart, and Column Chart. Default value: Line Chart.
X Axis	The sequential data. In most cases, a time series is selected.
Y Axis	The numeric data. You can select one or more fields for the Y-axis.
Aggregate Column	The field information that must be aggregated as the third point for comparison.
Legend	The position of the legend in the chart. Valid values: Top, Bottom, Left, and Right.
Format	The format in which data is displayed.
Margin	The distance between the axis and the borders of the chart. The parameters include **Top Margin**, **Bottom Margin**, **Right Margin**, and **Left Margin**.

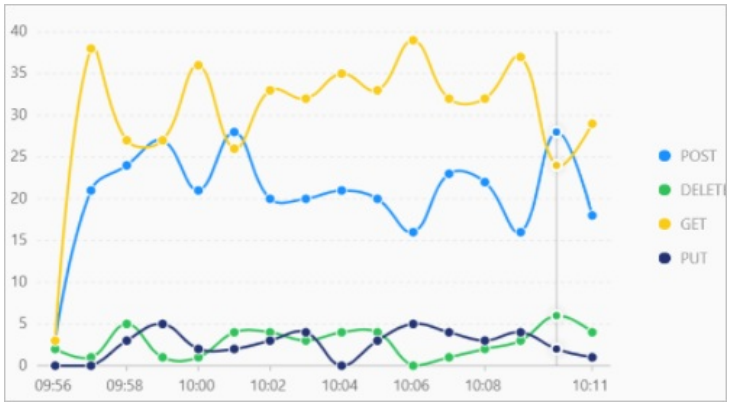
Example

A flow chart is suitable for displaying data in three dimensions, for example, time, category, and numeric value. In this example, the following query statement is executed:

```
* | select date_format(from_unixtime(__time__ - __time__% 60), '%H:%i:%S') as minute, count(1) as c,
request_method group by minute, request_method order by minute asc limit 100000
```

Select **minute** for X Axis, **c** for Y Axis, and **request_method** for Aggregate Column.

Flow chart



23.1.4.11.1.12. Display query results in a Sankey diagram

This topic describes how to configure a Sankey diagram to display query results.

Background information

A Sankey diagram is a type of flow chart. A Sankey diagram shows the flow from one set of values to another set of values. You can use Sankey diagrams to collect statistics about network traffic flows. A Sankey diagram contains the values of the `source`, `target`, and `value` fields. The `source` and `target` fields describe the source and target nodes, and the `value` field describes the flows from the `source` node to the `target` node.

Each Sankey diagram consists of the following elements:

- Node
- Edge

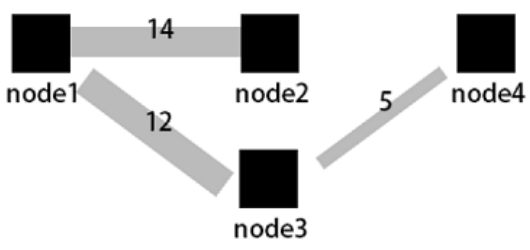
A Sankey diagram has the following features:

- The start flow is equal to the end flow. The sum of the widths of all main edges is equal to the sum of the widths of all branch edges. This allows you to manage and maintain a balanced flow of all traffic.
- The edge width in a row represents the volume of traffic in a specific status.
- The width of an edge between two nodes represents the flow volume in a status.

The following table describes the data that can be displayed in a Sankey diagram.

source	target	value
node1	node2	14
node1	node3	12
node3	node4	5

The following figure shows the data relationships in a Sankey diagram.



Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the Sankey diagram.

Parameter	Description
Start Column	The start node.
End Column	The end node.
Value Column	The volume of traffic between the start node and end node.
Margin	The distance between an axis and the borders of the chart. The parameters include **Top Margin**, **Bottom Margin**, **Right Margin**, and **Left Margin**.

Example

If a log contains the `source`, `target`, and `value` fields, you can use a nested subquery to obtain the sum of all `steamValue` values.

```
* | select sourceValue, targetValue, sum(streamValue) as streamValue from (select sourceValue, targetV
alue,
 streamValue, __time__ from log group by  sourceValue, targetValue, streamValue, __time__ order by __t
ime__ desc) group by sourceValue,
 targetValue
```

23.1.4.11.1.13. Display query results on a word cloud

This topic describes how to configure a word cloud to display query results.

Background information

A word cloud shows text data. A word cloud is a cloud-like and colored image composed of words. You can use a word cloud to display a large amount of text data. The font size or color of a word indicates the significance of the word. This allows you to identify whether a word is significant in an efficient manner.

The words in a word cloud are sorted.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the  icon.
6. On the **Properties** tab, configure the properties of the word cloud.

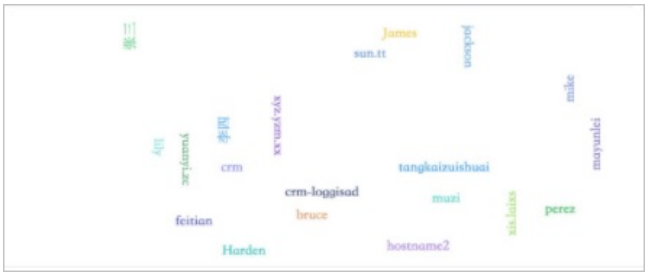
Parameter	Description
Word Column	The words that you want to display.
Value Column	The numeric value that corresponds to a word.
Font Size	The font size of a word. ○ The minimum font size ranges from 10 pixels to 24 pixels. ○ The maximum font size ranges from 50 pixels to 80 pixels.

Example

To query the distribution of hostnames in NGINX logs, execute the following query statement:

```
* | select hostname, count(1) as count group by hostname order by count desc limit 1000
```

Select `hostname` for Word Column and `count` for Value Column.



23.1.4.11.1.14. Display query results on a treemap chart

This topic describes how to configure a treemap chart to display query results.

Background information

A treemap chart consists of multiple rectangles that represent the data volumes. A larger rectangle area represents a larger proportion of the categorical data.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you want to configure a chart.
3. Click the name of the Logstore in which you want to configure a chart.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analyze**.
5. On the **Graph** tab, click the icon.
6. On the **Properties** tab, configure the properties of the treemap chart.

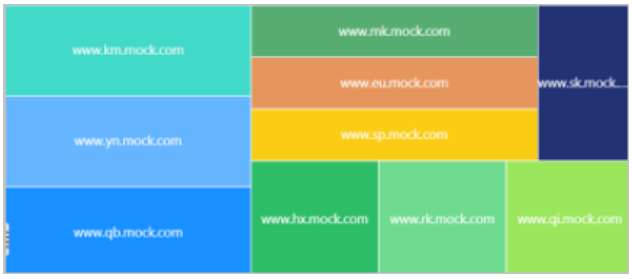
Parameter	Description
Legend Filter	The field that includes categorical data.
Value Column	The numeric value of a field. A greater field value represents a larger rectangle.

Example

To query the distribution of hostnames in NGINX logs, execute the following query statement:

```
* | select host, count(1) as count group by host order by count desc limit 1000
```

Select `host` for **Legend Filter** and select `count` for **Value Column**.



23.1.4.11.2. Dashboard

23.1.4.11.2.1. Overview

A dashboard provided by Log Service is a platform where you can analyze data in real time. You can add multiple charts to a dashboard for data analysis. Each chart is a visualized search and analytic statement.

A dashboard allows you to view the charts of multiple search and analytic statements at one time. When you open or refresh the dashboard, the statements of the charts run automatically.

After you add a chart to a dashboard, you can configure Drill-down analysis for the chart. Then you can click the chart on the dashboard to further analyze data and obtain more fine-grained analysis results.

Limits

- You can create a maximum of 50 dashboards for a project.
- Each dashboard can contain a maximum of 50 analysis charts.

Features

A dashboard has two modes: display mode and edit mode.

- Configure the display mode of a dashboard

  In the display mode, you can configure multiple display settings on the dashboard page.

  - Dashboard: You can specify the time range, the automatic refresh interval, full screen, and the display mode of the title for the dashboard, configure alerts for all charts on the dashboard, and filter chart data based on the Configure and use a filter on a dashboard of a Logstore.
  - Chart: You can view the analysis details of a specified chart, specify the time range and configure alerts for the chart, download logs and the chart, and check whether drill-down analysis is configured for the chart.

- Edit mode

  In the edit mode, you can change the configurations of the dashboard and charts.

  - Dashboard: You can use a dashboard as a canvas and add Markdown chart, custom charts, text, icons, and other chart elements to the dashboard. You can also add lines between chart elements that are self-adaptive to the positions of the charts. You can also add Configure and use a filter on a dashboard of a Logstore, which can be used to filter chart data in the display mode. In addition, you can configure display gridlines to help arrange chart elements such as icons in an orderly manner.
  - Chart: You can also edit a chart on the dashboard. You can modify the statement, properties, and interactive behavior such as drill-down analysis of the chart.

23.1.4.11.2.2. Create and delete a dashboard

This topic describes how to create a dashboard. After you create a dashboard, you can follow, view, and delete the dashboard.

Prerequisites

The indexing feature is enabled and indexes are configured. For more information, see Enable the indexing feature and configure indexes for a Logstore.

Create a dashboard

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to create a dashboard.
3. In the left-side navigation pane, click the icon.
4. Click the plus sign (+) to create a dashboard.
5. In the **Add to New Dashboard** dialog box, enter a name for the dashboard in the Dashboard Name field and click **OK**.

Related operations

After you create a dashboard, you can follow, view, and delete the dashboard.

- In the Dashboard list, find the dashboard that you created and choose > **Delete** to delete the dashboard.

  Notice After you delete a dashboard, the dashboard cannot be restored. Proceed with caution.

- In the Dashboard list, find the dashboard that you created and choose > **Details** to view the dashboard.

23.1.4.11.2.3. Manage a dashboard in display mode

This topic describes how to configure a dashboard in display mode. By default, a dashboard shows all charts in display mode. You can perform multiple operations on a dashboard in display mode.

Specify a time range for a dashboard

By default, the time range that you specify for a dashboard applies to all charts on the dashboard. After you specify a time range for a dashboard, all charts on the dashboard display the query and analysis results of the time range. For information about how to specify a time range for a single chart, see Specify a time range for a chart.

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to manage a dashboard.
3. In the left-side navigation pane, click the icon.
4. In the Dashboard list, click the dashboard that you want to manage.
5. Click **Please Select** to specify a time range.

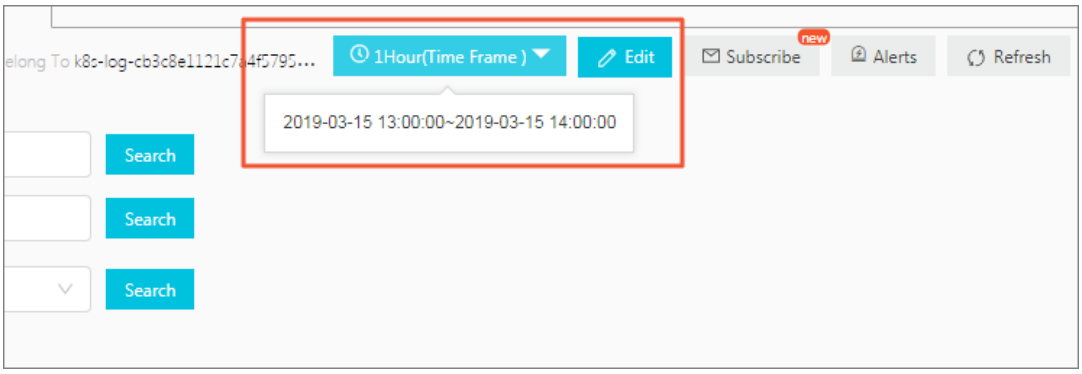
   Log Service supports the following types of time ranges:

   - Relative: queries log data that is generated within a time range that ends with the current time, such as the previous 1, 5, or 15 minutes. The time range is accurate to seconds. For example, if the current time is 19:20:31 and you select 1Hour(Relative) as the time range, the charts on the dashboard display the log data that is generated from 18:20:31 to 19:20:31.
   - Time Frame: queries log data that is generated within a time range that ends with the current time, such as the previous 1 or 15 minutes. The time range is accurate to minutes, hours, or days. For example, if the

current time is 19:20:31 and you select 1Hour(Time Frame) as the time range, the charts on the dashboard display the log data that is generated from 18:00:00 to 19:00:00.

- Custom: queries log data that is generated within a custom time range.

6. Mover the pointer over the **Please Select** button to confirm the specified time range.



Enter the edit mode

In the Dashboard list, click the dashboard that you want to modify. On the page that appears, click **Edit** to enter the edit mode. In edit mode, you can perform multiple operations on the dashboard and the charts on the dashboard. For more information, see Edit mode.

Configure an alert rule

In the upper-right corner of the dashboard page, click **Alerts** and select **Create** to create an alert rule for the charts on the dashboard. For more information, see Configure alerts.

Configure a refresh method

You can manually refresh a dashboard or select an interval to automatically refresh the dashboard.

- In the upper-right corner of the dashboard page, choose **Refresh > Once**. The dashboard is immediately refreshed.
- In the upper-right corner of the dashboard page, choose **Refresh > Auto Refresh** and select an interval at which the dashboard is automatically refreshed.

  The interval can be 15 seconds, 60 seconds, 5 minutes, or 15 minutes.

> **Note** If your browser is inactive, the dashboard may not refresh at the specified interval as expected.

Share a dashboard

In the upper-right corner of the dashboard page, click **Share** to copy the link of the dashboard to the clipboard. Then, you can send the link to authorized users. The shared dashboard page uses the settings of the dashboard at the point in time when you share the dashboard. The settings include the time range of charts and the display format of chart titles.

> **Note** Before you share a dashboard with other users, you must grant the read permissions to the users.

Display a dashboard in full screen

In the upper-right corner of the dashboard page, click **Full Screen**. Then, the dashboard enters the full-screen mode. This mode is suitable for scenarios such as presentations and reporting.

Select a display format for chart titles

In the upper-right corner of the dashboard page, click **Title Configuration** to select a display format for chart titles. Valid values:

- Single-line Title and Time Display
- Title Only
- Time Only

Reset the time range

In the upper-right corner of the dashboard page, click **Reset Time** to restore the saved time range of all charts on the dashboard. You can use this feature to restore time settings.

Configure charts

You can select a chart and perform the following operations on the chart.

> **Note** Different types of charts on a dashboard can display different information. You cannot view the analysis details of non-statistical charts such as custom charts and Markdown charts.

- View analysis details

  Find the chart whose details you want to view, click the ⋮ icon and select **View Analysis Details**. On the page that appears, you can view the query statement and the properties of the chart.

- Specify a time range for a chart

  Find the chart that you want to manage, click the ⋮ icon and select **Select Time Range** to specify a time range for the chart.

- Configure an alert rule for a chart

  Find the chart for which you want to configure an alert rule, click the ⋮ icon and select **Create Alert** to create an alert rule for the chart. For more information, see Configure alerts.

- Download log data

  Find the chart whose log data you want to download, click the ⋮ icon and select **Download Log**. The log data that is returned by the query statement of the chart within the current time range is downloaded in a comma-separated values (CSV) file.

- Check whether a drill-down event is configured for a chart

  Find the chart that you want to check, click the ⋮ icon and move the pointer over the icon to check whether a drill-down event is configured for the chart. If the icon is red, a drill-down event is configured for the chart.

- Preview the query statement of a chart

  Find the chart whose query statement you want to preview, click the ⋮ icon, and then click the icon. In the Preview Query Statement dialog box, you can view the query statement of the chart.

23.1.4.11.2.4. Manage a dashboard in edit mode

You can manage a dashboard in edit mode. For example, you can add chart elements, adjust chart layouts, edit charts, and change the dashboard name.

Add chart elements

1. In the Projects section, click the project in which you want to modify a dashboard.
2. In the left-side navigation pane, click the icon.
3. In the Dashboard list, click the dashboard that you want to manage.

4. In the upper-right corner of the dashboard page, click **Edit**.

   You can add the following chart elements on a dashboard in edit mode.

   > **Notice** If you modify a dashboard in edit mode, you must save the modifications before the modifications can take effect. To save modifications, click **Save** in the upper-right corner of the dashboard page.

   - Rectangles and diamonds

     Drag the rectangular icon or the diamond icon to a position. Then, double-click the icon and enter text. You can also modify the text properties and the border properties of rectangles and diamonds.

   - Common icons

     Log Service allows you to display common icons on a dashboard page. You can drag an icon to a position.

   - Text

     Drag the text icon to a position. Then, double-click the text box and enter text. You can also modify the properties of the text. The properties include the font size, font style, alignment, and font color.

   - Markdown chart

     Drag the Markdown icon to a position. Then, double-click the text box and insert elements such as text, charts, and videos. For more information, see Markdown chart.

   - Filter

     Click the filter icon to add a filter. For more information, see Add a filter.

     After you add a filter to a dashboard, you can use the filter to refine search results or replace placeholder variables in query statements.

   - Custom SVG

     Click the SVG icon. In the Customize SVG dialog box, click the box or drag a Scalable Vector Graphics (SVG) file to the box to upload the file.

     > **Note** The size of an SVG file cannot exceed 10 KB.

   - Custom image's HTTP link

     Click the Customize image's HTTP link icon in the menu bar. On the page that appears, enter the HTTP link of an image and click OK.

Adjust chart layouts

On a dashboard in edit mode, all charts and chart elements are displayed on a canvas. You can drag and scale each chart. The width of the canvas cannot exceed the width of your browser. The height of the canvas is unlimited and is measured in pixels.

On the canvas, you can perform the following operations:

- Adjust the position of a chart.
  - You can drag a chart to a position.
  - You can select a chart and set the **L** and **T** parameters to adjust the chart position.
- Adjust the width and height of a chart.
  - You can select a chart and drag the lower-right corner of the chart to resize the chart.
  - You can select a chart and set the **W** and **H** parameters to resize the chart.
- Add lines to connect charts.

You can add a directional line between two charts. When you adjust the position or size of the charts, the line automatically moves to show the relative position between the two charts.

- Configure chart levels.

  You can select a chart and click the Move Layer to Top icon or the Move Layer to Down icon in the menu bar to move the chart to the upper part or lower part of the dashboard.

Configure charts

You can modify, copy, and delete a chart on a dashboard in edit mode.

- Modify the query statement, properties, data source, and interactive behavior for a chart.
  i. In the upper-right corner of the dashboard page, click Edit to enter the edit mode. Find the chart that you want to modify and choose > **Edit**.
  ii. Modify the query statement, properties, data source, and interactive behavior for the chart.

  For information about how to configure interactive behavior for a chart, see Drill-down analysis.
  iii. Click **Preview** to check the configuration results.
  iv. Click **OK**.
  v. In the upper-right corner of the dashboard page, click **Save**.
- Create a copy of a chart. The copy uses the same configurations as the chart.
  i. In the upper-right corner of the dashboard page, click Edit to enter the edit mode. Find the chart that you want to copy and choose > **Copy**.
  ii. Drag the copy to a position. Then, specify the margins and size of the copy.
  iii. In the upper-right corner of the dashboard page, click **Save**.
- Delete a chart.
  i. In the upper-right corner of the dashboard page, click Edit to enter the edit mode. Find the chart that you want to delete and choose > **Delete**.
  ii. In the upper-right corner of the dashboard page, click **Save**.

23.1.4.11.2.5. Configure a drill-down event

Log Service allows you to configure a drill-down event for a chart to obtain more detailed analysis results. This topic describes how to configure a drill-down event in the Log Service console.

Prerequisites

- The indexing feature is enabled and indexes are configured. For more information, see Configure indexes.
- A Logstore is created. This prerequisite must be met if the drill-down event that you want to configure is to open a Logstore. For more information, see Create a Logstore.
- A saved search is created. This prerequisite must be met if the drill-down event that you want to configure is to open a saved search. For more information, see Saved search.

  Placeholder variables are configured in the query statement of the saved search. This prerequisite must be met if you want to configure variables.
- A dashboard is created. This prerequisite must be met if the drill-down event that you want to configure is to open a dashboard. For more information, see Create a dashboard.

  Placeholder variables are configured in the related chart on the dashboard. This prerequisite must be met if you want to configure variables.
- If the drill-down event that you configure is to open a custom HTTP link, you must create the HTTP link.

Context

Drilling is required for data analysis. This feature allows you to analyze data in a fine-grained or coarse-grained manner. You can use this feature to roll up or drill down data. Drill-down allows you to obtain more detailed analysis results. This way, you extract more value from data and make better decisions for your business.

Procedure

1. Log on to the Log Service console.
2. Click the name of the project in which you want to configure a drill-down event.
3. Click the icon next to the name of the Logstore in which you want to query and analyze data, and then select **Search & Analysis**.
4. Enter a query statement in the search box, specify a time range, and then click **Search & Analytics**.
5. On the **Graph** tab, select a chart type. On the Properties tab, set the parameters.

   For more information about the parameters of a chart, see Chart configurations.
6. Click the **Interactive Behavior** tab. On this tab, configure a drill-down event for the chart.

   You can set the Event Action parameter to Disable, Open Logstore, Open Saved Search, Open Dashboard, Open Dashboard, or Custom HTTP Link.

   - **Disable**: disables the drill-down feature.
   - **Open Logstore**: configures the drill-down event to open a Logstore. The following table describes the parameters that you can specify if you set the Event Action parameter to Open Logstore.

Parameter	Description
Open in New Tab	If you turn on this switch, the Logstore that you specify is opened on a new tab when the drill-down event is triggered.
Select Logstore	The name of the Logstore to which you want to be redirected. When a drill-down event is triggered, you are redirected to the Search & Analysis page of the Logstore.
Time Range	The time range. The system queries the data that is generated within the time range. Valid values: ■ Default: queries data in the Logstore to which you are redirected based on the default time range. The default time range is 15 minutes (relative) and accurate to seconds. ■ Inherit table time: queries data in the Logstore to which you are redirected based on the time range specified for the chart when the drill-down event is triggered. ■ Relative: queries data in the Logstore to which you are redirected based on the time range that you specify. The time range is accurate to seconds. ■ Time Frame: queries data in the Logstore to which you are redirected based on the time range that you specify. The time range is accurate to minutes, hours, or days.
Inherit Filtering Conditions	If you turn on **Inherit Filtering Conditions**, the filter conditions that are added to the dashboard are synchronized to the Search & Analysis page of the Logstore to which you are redirected when the drill-down event is triggered. The filter conditions are added to the start of the query statement by using the `AND` operator.

Parameter	Description
Filter	On the **Filter** tab, you can enter a filter statement in the Filter Statement field. When the drill-down event is triggered, the filter statement is synchronized to the Search & Analysis page of the Logstore to which you are redirected. The filter statement is added to the start of the query statement by using the `AND` operator. The filter statement can contain fields that you specify in the **Optional Parameter Fields** field.

- **Open Saved Search**: configures the drill-down event to open a saved search. The following table describes the parameters that you can specify if you set the Event Action parameter to Open Saved Search.

Description	Description
Open in New Tab	If you turn on this switch, the saved search that you specify is opened on a new tab when the drill-down event is triggered.
Select Saved Search	The name of the saved search to which you want to be redirected.
Time Range	The time range for the saved search. Valid values: ■ Default: queries data by using the saved search based on the default time range. The default time range is 15 minutes (relative) and accurate to seconds. ■ Inherit table time: queries data by using the saved search based on the time range specified for the chart when the drill-down event is triggered. ■ Relative: queries data by using the saved query statement based on the time range that you specify. The time range is accurate to seconds. ■ Time Frame: queries data by using the saved search based on the time range that you specify. The time range is accurate to minutes, hours, or days.
Inherit Filtering Conditions	If you turn on **Inherit Filtering Conditions**, the filter conditions that are added to the dashboard are synchronized to the saved search that you want to execute when the drill-down event is triggered. The filter conditions are added to the start of the saved search by using the `AND` operator.
Inherit Variables	If you turn on **Inherit Variables** and the variable that you configure on the dashboard is the same as the variable in the saved search, the variable value on the dashboard replaces the variable in the saved search. **Note** If you want to inherit variables, you must configure a placeholder variable in the saved search.
Filter	On the **Filter** tab, you can enter a filter statement in the Filter Statement field. When the drill-down event is triggered, the filter statement is added to the start of the saved search by using the `AND` operator. The filter statement can contain fields that you specify in the **Optional Parameter Fields** field.

Description	Description
Variable	Log Service allows you to modify a saved search by using variables. If you configure a variable that is the same as the variable in the saved search, the variable value that you click to trigger the drill-down event replaces the variable in the saved search. You can add variables on the **Variable** tab. **Note** ■ If you want to configure a variable, you must configure a placeholder variable for the saved search to which you want to be redirected. ■ You can add up to five dynamic variables and up to five static variables. ■ Dynamic variables ■ **Variable**: the name of the variable. ■ **Variable Value Column**: the column in which the variable values are located. The values are used to dynamically replace the variable in the saved search. ■ Static variables ■ **Variable**: the name of the variable. ■ **Value**: The value of the static variable. The value is used to replace the placeholder variable in the saved search.

- **Open Dashboard**: configures the drill-down event to open a dashboard. The following table describes the parameters that you can specify if you set the Event Action parameter to Open Dashboard.

Parameter	Description
Open in New Tab	If you turn on this switch, the dashboard that you specify is opened on a new tab when the drill-down event is triggered.
Select Dashboard	The name of the dashboard to which you want to be redirected.
Time Range	The time range to query data for the dashboard. Valid values: ■ Default: queries data for the dashboard to which you are redirected based on the default time range. The default time range is 15 minutes (relative) and accurate to seconds. ■ Inherit table time: queries data for the dashboard to which you are redirected based on the time range specified for the chart when the drill-down event is triggered. ■ Relative: queries data for the dashboard to which you are redirected based on the time range that you specify. The time range is accurate to seconds. ■ Time Frame: queries data for the dashboard to which you are redirected based on the time range that you specify. The time range is accurate to minutes, hours, or days.
Inherit Filtering Conditions	If you turn on **Inherit Filtering Conditions**, the filter conditions that are added to the current dashboard are synchronized to the dashboard to which you are redirected when the drill-down event is triggered.
Inherit Variables	If you turn on **Inherit Variables**, the variables that you configure on the current dashboard are synchronized to the dashboard to which you are redirected.

Parameter	Description
Filter	On the **Filter** tab, you can enter a filter statement in the Filter Statement field. When the drill-down event is triggered, the filter statement can be synchronized to the dashboard to which you are redirected. The filter statement can contain fields that you specify in the **Optional Parameter Fields** field.
Variable	The variables that you configure are synchronized to the dashboard to which you are redirected when the drill-down event is triggered. You can add variables on the **Variable** tab. **Note** ■ If you want to configure a variable for the chart, you must configure a placeholder variable for the chart on the dashboard to which you are redirected. ■ You can add up to five dynamic variables and up to five static variables. ■ Dynamic variables ■ **Variable**: the name of the variable. ■ **Variable Value Column**: the column in which the variable values are located. The values are dynamically synchronized to the dashboard to which you are redirected. ■ Static variables ■ **Variable**: the name of the variable. ■ **Value**: the static value of the variable. The value is synchronized to the dashboard to which you are redirected.

- **Custom HTTP Link**: configures the drill-down event to open a custom HTTP link.

  The path in the custom HTTP link is the path of the file that you want to access. You can add an optional parameter to the path. If you click a variable value on a chart to trigger the drill-down event, the parameter is replaced by the value, and you are redirected to the custom HTTP link.

Parameter	Description
Enter Link	The address to which you want to be redirected.
Use System Variables	If you turn on **Use System Variables**, you can insert variables that are provided by Log Service to the HTTP link. The variables include ${sls_project}, ${sls_dashboard_title}, ${sls_chart_name}, ${sls_chart_title}, ${sls_region}, ${sls_start_time}, ${sls_end_time}, ${sls_realUid}, and ${sls_aliUid}.
Transcoding	If you turn on **Transcoding**, the HTTP link is encoded.
Optional Parameter Fields	If you add an optional parameter to the path, the parameter is replaced by the value that you click to trigger the drill-down event.

7. Click **Add to New Dashboard**.
8. In the dialog box that appears, specify a dashboard name and a chart name, and then click **OK**.

Example

This section provides an example on how to store NGINX access logs in a Logstore named accesslog and how to create two dashboards named RequestMethod and destination_drilldown for drill-down analysis. Before you can perform drill-down analysis, you must add a table of request methods to the RequestMethod dashboard, and configure a drill-down event for the table to open the destination_drilldown dashboard. Then, you must add a line chart to the destination_drilldown dashboard. The line chart displays the trend of page views (PVs) over a specified period of time. After you complete the settings, you can click a request method on the RequestMethod dashboard. Then, you are redirected to the destination_drilldown dashboard and you can view the trend of PVs over a specified period of time on the destination_drilldown dashboard.

1. Create a dsahboard named destination_drilldown.

   Before you configure a drill-down event for the table of request methods, you must create a dashboard to which you want to be redirected and add a line chart to the dashboard. The line chart displays the trend of PVs over a specified period of time. You need to configure the following settings. For more information, see Create a dashboard.

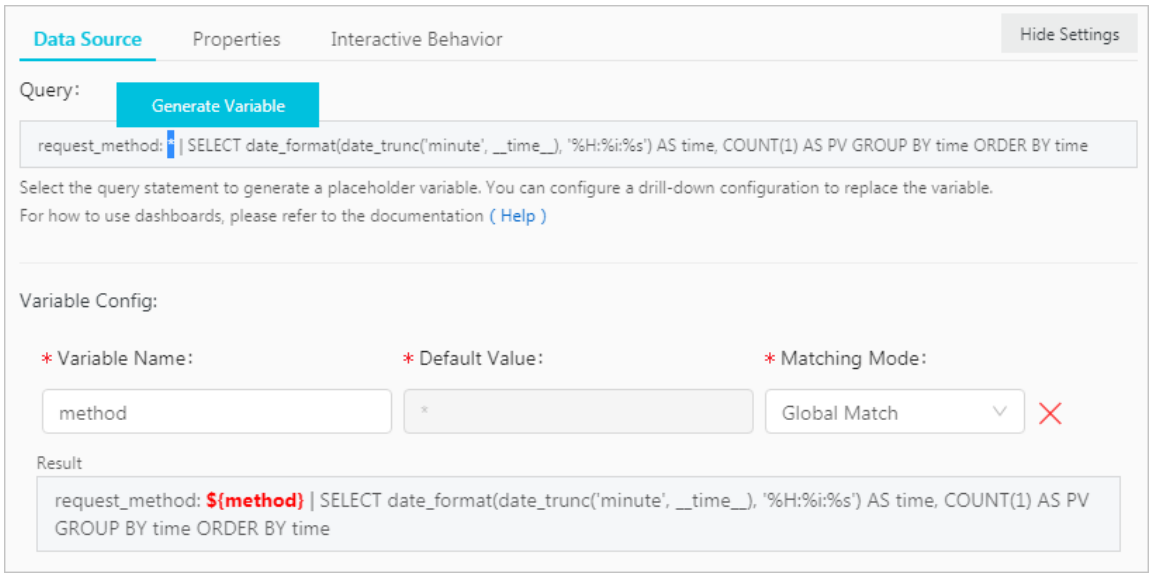
   - Specify a query statement.

     The query statement queries logs by request type. You can view the trend of PVs over a specified period of time.

     ```
     request_method: * | SELECT date_format(date_trunc('minute', __time__), '%H:%i:%s') AS time, COUNT(1) AS PV GROUP BY time ORDER BY time
     ```

   - Configure a placeholder variable.

     Specify the asterisk (*) to generate a placeholder variable and set the variable name to method.

2. Configure a drill-down event for the table of request methods and add the table to the RequestMethod dashboard.

   You need to configure the following settings. For more information, see Procedure.

   - Specify a query statement.

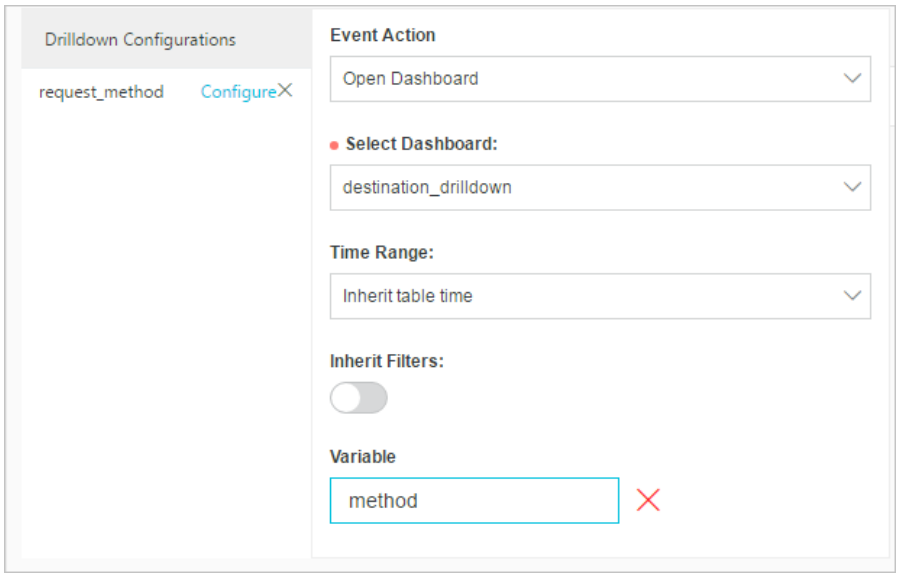
     The query statement queries the logs that are generated for each request method among the NGINX access logs.

     ```
     *|SELECT request_method, COUNT(1) AS c GROUP BY request_method ORDER BY c DESC LIMIT 10
     ```

   - Select a chart type.

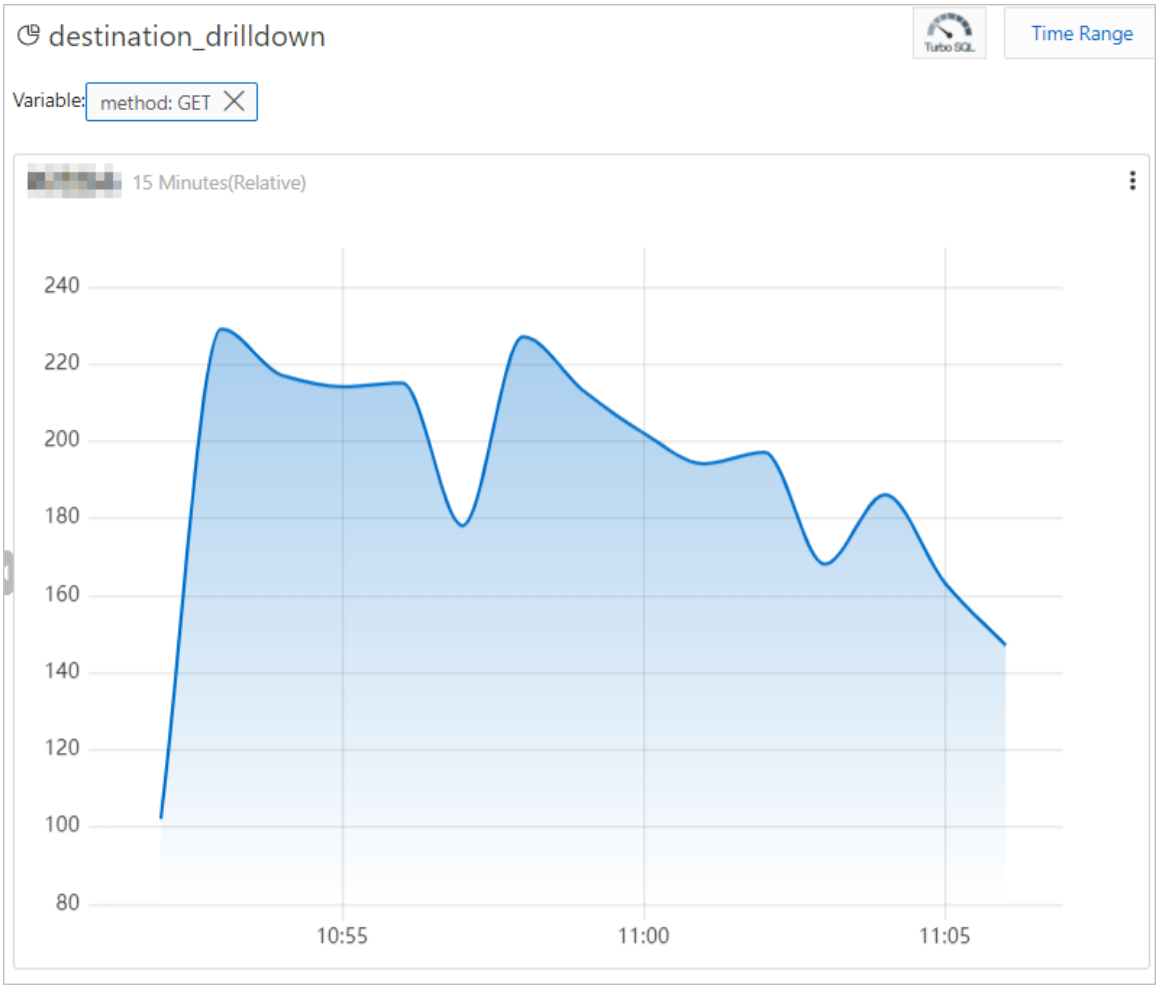
     In this example, a table is selected.

- Configure a drill-down event for the table.
  - Configure a drill-down event for the request_method column in the table.
  - Set **Select Dashboard** to **destination_drilldown**.
  - Set **Variable** to **method**.



3. View drill-down results.

   On the RequestMethod dashboard, click **GET**. You are redirected to the destination_drilldown dashboard. The asterisks (*) in the query statement is replaced by the value **GET**. The trend of PVs for GET requests over a specified period of time is displayed in a line chart.



23.1.4.11.2.6. Add a filter

You can add a filter to a dashboard. Then, you can use the filter to refine search results or replace placeholder variables with specified values. This topic describes how to add a filter to a dashboard.

Prerequisites

- The indexing feature is enabled and indexes are configured. For more information, see Enable the indexing feature and configure indexes for a Logstore.
- Charts are added to a dashboard. For more information, see Add a chart to a dashboard.

  If you set the Type parameter to Replace Variable when you add a filter to a dashboard, you must specify a placeholder variable when you add a chart to the dashboard.

Context

A filter is used to modify the query statements or replaces placeholder variables for all charts on a dashboard. Each chart displays the query and analysis results after you execute a query statement in the [search query] | [sql query] format. After you add a filter to a dashboard, the filter condition or variables that you specify for the filter apply to the query statement that corresponds to each chart on the dashboard. The following types of filters are supported:

- Filter: uses key-value pairs as a filter condition.

The filter condition is added to the start of a query statement by using the AND or NOT operator. For example, the **Key: Value AND [search query] | [sql query]** statement queries logs that contain **Key:Value** in the query result of the [search query] | [sql query] statement. For the Filter type, you can select or enter multiple key-value pairs. If you specify multiple key-value pairs, the logical OR operator is used between the pairs.

- Replace Variable: specifies a variable and the value of the variable.

  If the variable that you specify for the filter is configured for existing charts on the dashboard, the variable in the query statement of each chart is automatically replaced by the variable value that you specify for the filter. This applies to all charts for which the same variable is configured.

Procedure

1. Log on to the Log Service console.
2. Click the name of the project in which you want to manage a dashboard.
3. In the left-side navigation pane, click the Dashboard icon.
4. In the Dashboard list, click the name of the dashboard that you want to manage.
5. In the upper-right corner of the dashboard page, click **Edit**.
6. Click the  icon and set the following parameters. Then, click **OK**.

Parameter	Description
Filter Name	The name of the filter.
Display Settings	Valid values: ○ **Title**: specifies whether to add a title for the filter. You can turn on Title to add a title for the filter. ○ **Border**: specifies whether to add borders to the filter. You can turn on Border to add borders to the filter. ○ **Background**: specifies whether to add a white background to the filter. You can turn on Background to add a white background to the filter.
Type	The type of the filter. ○ **Filter**: uses key-value pairs to filter data. The key-value pairs are used as a filter condition and added to the start of a query statement by using the AND or NOT operator. By default, AND is used. ■ AND: **Key:Value AND [search query]	[sql query]** ■ NOT: **Key:Value NOT [search query]	[sql query]** You can specify multiple key-value pairs in the **Static List Items** field. ○ **Replace Variable**: specifies a variable and the value of the variable. If the variable that you specify for the filter is configured for existing charts on the dashboard, the variable in the query statement of each chart is automatically replaced by the variable value that you specify for the filter. You can specify multiple values of variables in the **Static List Items** field.

Parameter	Description
Key	○ If you select **Filter**, enter the key that you want to use to filter data in the **Key** field. ○ If you select **Replace Variable**, enter the variable that you want to use to filter data in the **Key** field. **Note** If you select **Replace Variable**, you must specify a placeholder variable when you add a chart to the dashboard. The placeholder variable must be the same as the variable that you specify in the Key field.
Alias	The alias of the key. This parameter is available only when you select **Filter**.
Global filter	The parameter is available only when you select **Filter**. ○ If you want to filter the specified values in all fields, turn on **Global filter**. ○ If you want to filter the specified values in specified keys, turn off **Global filter**.
Static List Items	Click the plus sign (+) and enter a value for the key in the text box. The value for **Key** is used to filter data. You can click the plus sign (+) to add more values for the specified key. If you turn on **Select by Default** for a value, the value is used to filter data each time you open a dashboard.
Add Dynamic List Item	Dynamic list items are dynamic values that are obtained by executing the specified query statement. If you turn on **Add Dynamic List Item**, you must perform the following steps before you can preview dynamic values: Select a Logstore, configure the Inherit Filtering Conditions parameter to specify whether the filter condition on the dashboard is inherited, enter a query statement, specify a time range, and then click **Search**.

Example 1: Use different time granularities to analyze logs

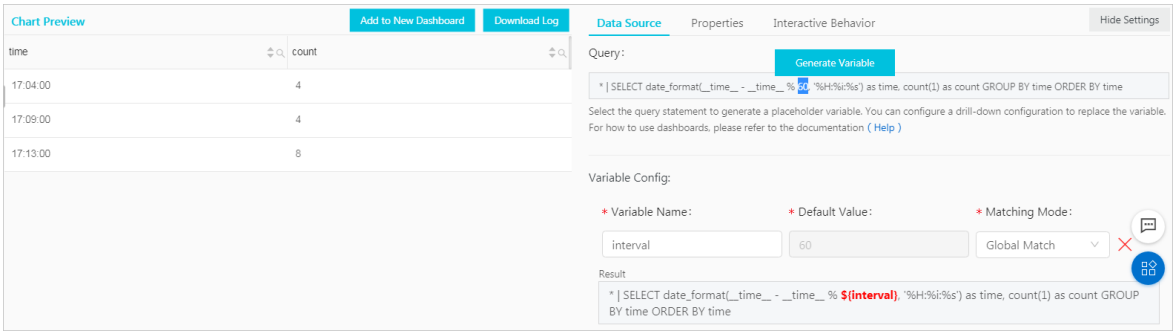
After you collect NGINX access logs, you can query and analyze these logs in real time.

You can use a query statement to view the number of page views (PVs) per minute. If you want to view the number of PVs per second, you must change the value of `__time__ - __time__ % 60` in the query statement. To simplify this operation, you can use a filter to replace variables in the query statement.

1. Execute the following query statement to view the number of PVs per minute:

```
* | SELECT date_format(__time__ - __time__ % 60, '%H:%i:%s') as time, count(1) as count GROUP BY time ORDER BY time
```

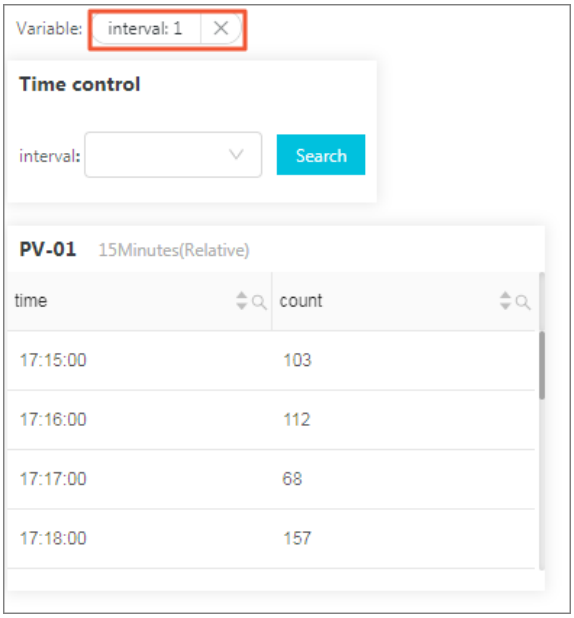
2. Add the chart to a dashboard and select `60` to generate a variable named `interval`.



3. Add a filter on the dashboard page.
   - **Type**: Select **Replace Variable**.
   - **Key**: Enter `interval`.
   - **Static List Items**: Add `1` and `120` as values of the key. Unit: seconds.
4. In the Filter section of the dashboard, select `1` from the Interval drop-down list to view statistics by second.

   The following example shows the query statement in which the placeholder variable is replaced:

```
* | SELECT date_format(__time__ - __time__ % 1, '%H:%i:%s') as time, count(1) as count GROUP BY ti
me ORDER BY time
```



Example 2: Switch between request methods

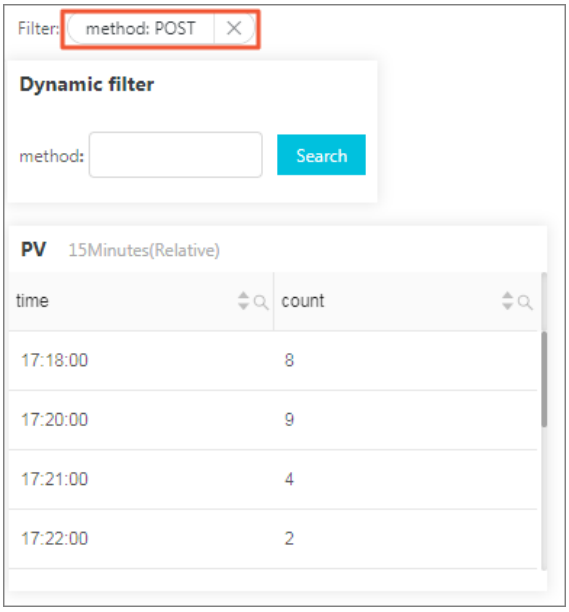
You can add dynamic values to a filter to dynamically switch between request methods. In example 1, the query statement starts with an asterisk (`*`), which means no condition is specified to filter the query results and all logs are queried. You can add another filter to view the PV data of each request `method`.

1. Add a filter on the dashboard and turn on **Add Dynamic List Item**.

   You can configure the following settings:
   - **Type**: Select Filter.
   - **Key**: Enter `method`.
   - **Select Logstore**: Select the Logstore to which the dashboard belongs.
   - **Add Dynamic List Item**: Enter a query statement to obtain dynamic values.
2. In the Filter section of the dashboard, select `POST` from the drop-down list.

   Only the PV data whose `method` is `POST` is displayed in the chart. In this example, the query statement is changed to the following format:

```
(*) and (method: GET) | SELECT date_format(__time__ - __time__ % 60, '%H:%i:%s') as time, count(1)
as count GROUP BY time ORDER BY time
```



23.1.4.11.2.7. Manage a Markdown chart

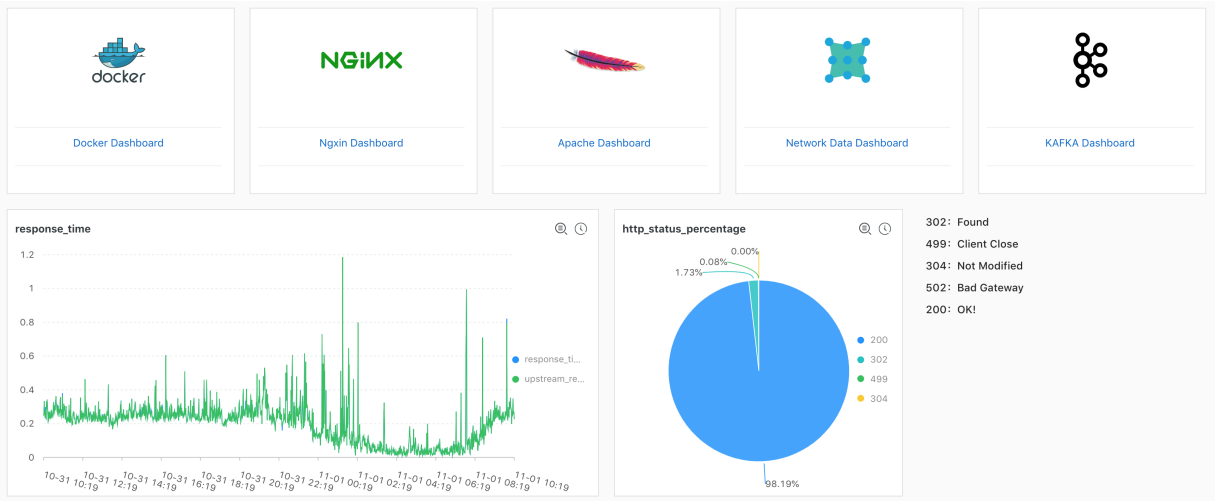
Log Service allows you to add a Markdown chart to a dashboard. In the Markdown chart, you can insert images, links, videos, and other elements to make your dashboard page more intuitive to use.

Context

You can add multiple analysis charts to a dashboard. This allows you to view multiple analysis results and monitor the status of multiple applications on a single dashboard. You can also add Markdown charts to a dashboard. A Markdown chart is edited by using the Markdown syntax.

You can create different Markdown charts based on your business requirements. Markdown charts can make a dashboard more intuitive to use. You can insert text such as background information, chart description, notes, and extension information into a Markdown chart. You can insert custom images and videos into a Markdown chart. You can also insert saved searches or the dashboard links of other projects to redirect to other query pages.

You can insert links into a Markdown chart to redirect to the other dashboard pages of the current project. You can also insert an image that corresponds to each link. In addition, you can use a Markdown chart to describe the parameters of analysis charts.



Add a Markdown chart

1. Log on to the Log Service console.

2. In the Projects section, click the name of the project in which you want to manage a dashboard.
3. In the left-side navigation pane, click the icon.
4. In the Dashboard list, click the dashboard that you want to manage.
5. In the upper-right corner of the dashboard page, click **Edit**.
6. In edit mode, drag the icon from the menu bar and drop the icon on a specified position to create a Markdown chart.
7. Double-click the Markdown chart.
8. In the **Markdown Edit** dialog box, set the parameters, and then click **OK**.

Parameter	Description
Chart Name	The name of the Markdown chart.
Show Border	Specifies whether to show the borders of the Markdown chart. You can turn on **Show Border** to show the borders of the Markdown chart.
Show Title	Specifies whether to show the title of the Markdown chart. You can turn on **Show Title** to show the title of the Markdown chart.
Show Background	Specifies whether to show the background of the Markdown chart. You can turn on **Show Background** to show the background of the Markdown chart.
Query Binding	Specifies whether to associate a query statement with a Markdown chart. You can turn on **Query Binding** and associate a query statement with a Markdown chart. Then, dynamic query results are displayed in the Markdown chart. i. Select a Logstore whose data you want to query. ii. Enter a query statement in the search box, specify a time range, and then click **Search**. For more information, see Overview. **Note** The query results may contain logs that are generated 1 minute earlier or later than the specified time range. The first returned log is displayed. iii. Click the plus sign next to a field to insert the corresponding query result into the **Markdown Content** column.
Markdown Content	Enter Markdown content in the **Markdown content** column on the left. The data preview is displayed in real time in the **Show Chart** column on the right. You can modify the Markdown content based on the data preview. For more information, see Common Markdown syntax.

9. Click **Save**.

Modify a Markdown chart

1. In the upper-right corner of the dashboard page, click **Edit**.
2. Modify the position and size of a Markdown chart

   Drag the Markdown icon to a position on the dashboard and drag the lower-right corner of the chart to adjust the size of the chart.
3. Modify the properties of a Markdown chart
   i. Double-click the Markdown chart that you want to modify.

ii. In the **Markdown Edit** dialog box, modify the parameters, and then click **OK**.

You can modify the chart name, display settings, query settings, and Markdown content. For more information, see Add a Markdown chart.

Delete a Markdown chart

1. In the upper-right corner of the dashboard page, click **Edit**.
2. Find the Markdown chart that you want to delete and choose > **Delete**.
3. In the upper-right corner of the dashboard page, click **Save**.

Common Markdown syntax

- Heading
  - Markdown syntax

```
# Level 1 heading
## Level 2 heading
### Level 3 heading
```

  - Result

- Link

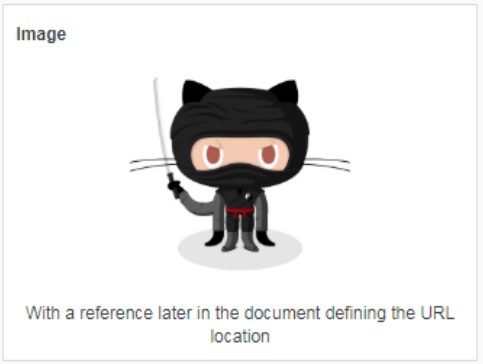
  Markdown syntax

```
### Contents
[Test](https://www.alibabacloud.com/)
```

- Image
  - Markdown syntax

```
<div align=center>
![Alt txt][id]
With a reference later in the document defining the URL location
[id]: https://octodex.github.com/images/dojocat.jpg  "The Dojocat"
```

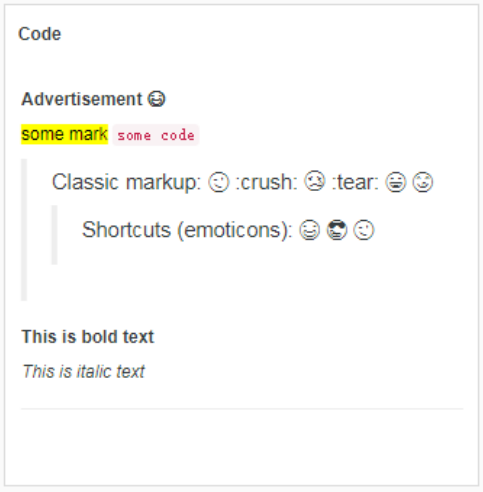
- Preview



- Special tag
  - Markdown syntax

```
---
__Advertisement :)__
==some mark==   `some code`
> Classic markup: :wink: :crush: :cry: :tear: :laughing: :yum:
>> Shortcuts (emoticons): :-)  8-) ;)
__This is bold text__
*This is italic text*
---
```

  - Result



23.1.5. Alerts

23.1.5.1. Overview

Log Service provides the alerting feature. You can configure alert rules to trigger alerts based on query and analysis results. After you create an alert rule, Log Service checks related query and analysis results on a regular basis. If a query and analysis result meets the trigger condition that you specify in the alert rule, Log Service sends an alert notification. This way, you can monitor the service status in real time.

Limits

The following table describes the limits of the alerting feature in Log Service.

Item	Description
Associated query statements	You can associate an alert rule with a maximum of three query statements.
Field value size	If the number of characters that are included in a field exceeds 1,024, Log Service extracts only the first 1,024 characters for data processing.
Trigger condition	Trigger conditions have the following limits: • Each trigger condition must be 1 to 128 characters in length. • If a query result includes more than 100 rows, Log Service only checks whether the first 100 rows meet the specified trigger condition. • Log Service checks whether a trigger condition is met for a maximum of 1,000 times for the specified query statements.
Query time range	The maximum time range that you can specify for each query is 24 hours.
Voice calls	If a voice call is not answered, Log Service sends an SMS notification. You are charged only once for the voice call regardless of whether the call is answered. You are not charged for SMS notifications.

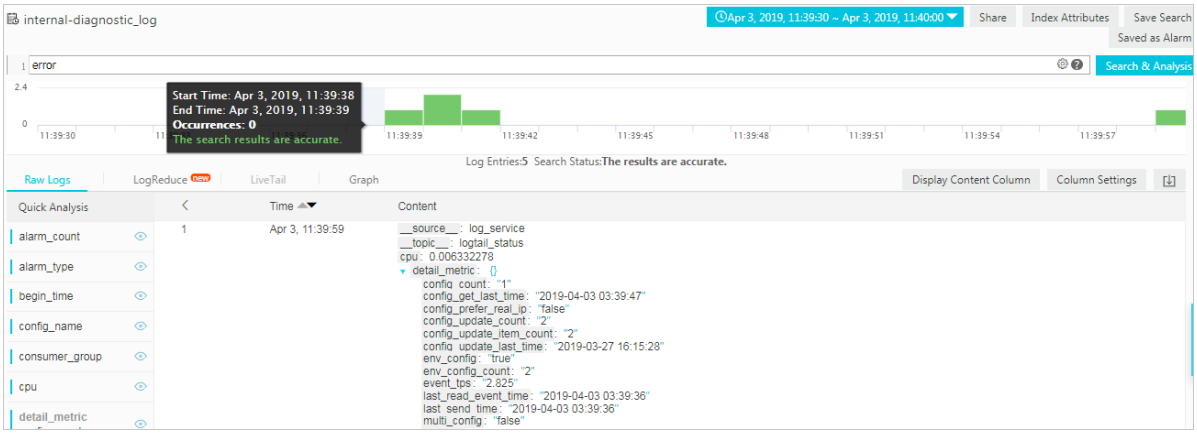
Query statements in alert rules

An alert rule is associated with one or more charts in a dashboard. Each chart displays the result of a query statement. You can associate an alert rule with one or more search statements or query statements.

- A search statement returns the log entries that meet the specified search condition.

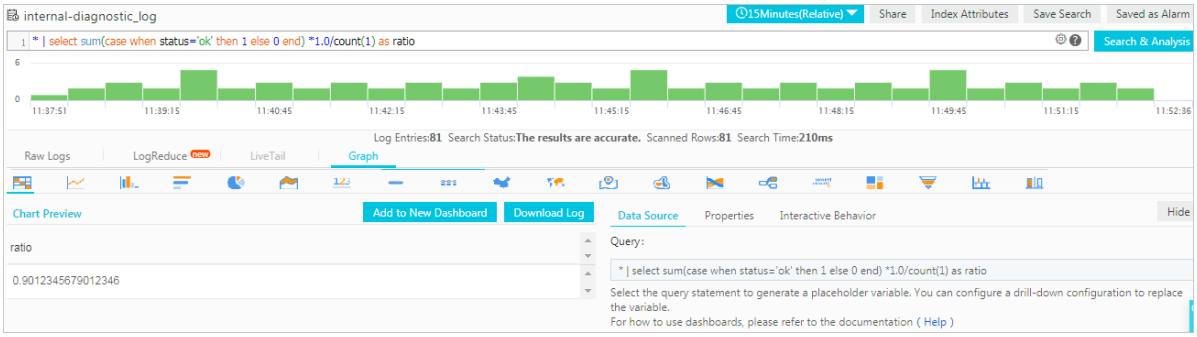
  For example, you can execute the error statement to search for the log entries that are generated in the previous 15 minutes and contain error. A total of 154 log entries are returned. Each log entry consists of key-value pairs. You can specify a trigger condition based on the value of a key.

  > **Note** If the number of returned log entries exceeds 100, Log Service checks only the first 100 log entries. If one of the log entries meets the specified condition, an alert is triggered.



- A query statement consists of a search statement and an analytic statement. The analytic statement analyzes the log entries that meet the search condition and returns a result.

For example, the * | select sum(case when status='ok' then 1 else 0 end) *1.0/count(1) as ratio statement returns the percentage of the log entries in which the value of the status field is ok. If you set the trigger condition of an alert rule to ratio < 0.9, an alert is triggered if the percentage of the log entries whose status code is ok is less than 90%.



23.1.5.2. Configure an alarm

23.1.5.2.1. Configure an alert rule

You can create an alert rule on the Search & Analysis page of a Logstore or on a dashboard page in the Log Service console. After you create an alert rule, Log Service sends alert notifications when the trigger condition in the alert rule is met. This topic describes how to create an alert rule in the Log Service console.

Prerequisites

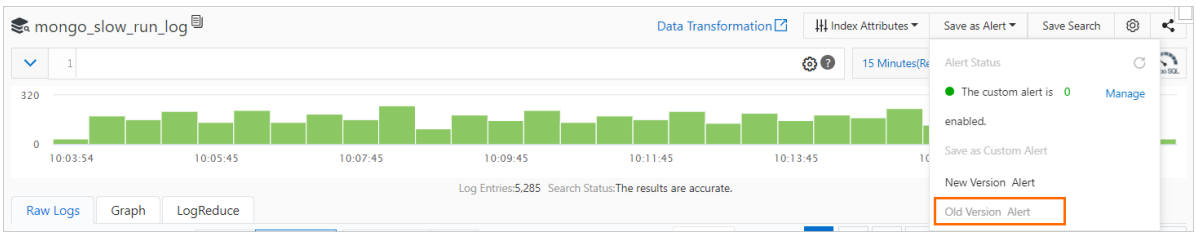
- Logs are collected and stored in a Logstore.
- The indexing feature is enabled and indexes are configured. For more information, see Enable the index feature and configure indexes for a Logstore.

Context

Log Service allows you to configure alerts based on charts. You can create an alert rule for a query statement on the Search & Analysis page. After you create the alert rule, a chart that shows the query result of the query statement is automatically created in the specified dashboard. You can also create an alert rule for one or more existing charts in a dashboard.

- Create a chart and configure an alert rule for the chart

  After you create an alert rule for a query statement, a chart that shows the query result of the query statement is automatically created in the specified dashboard. Therefore, when you create an alert rule for a query statement on the Search & Analysis page, you must specify a dashboard and name the chart.



- Create an alert rule for existing charts on a dashboard

  If you create an alert rule for multiple existing charts, you can specify one or more charts with which you want to associate the alert rule. You can specify a conditional expression for each chart, and then combine the conditional expressions into a trigger condition of the alert rule.

The following section describes how to configure an alert rule for multiple existing charts on a dashboard.

> **Note** If you modify the query statement of a chart with which an alert rule is associated, you must update the query statement in the alert rule. For more information, see Modify an alert.

For information about common configurations for alert rules, see FAQ about alerts.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to create a Logstore.
3. In the left-side navigation pane, click the icon.
4. In the Dashboard list, click the dashboard that you want to manage.
5. In the upper-right corner of the dashboard page, choose **Alerts > Create**.
6. In the **Alert Configuration** step, set the parameters and click **Next**.

   The following table describes the parameters.

Parameter	Description
Alert Name	The name of the alert rule. The name must be 1 to 64 characters in length.
Associated Chart	The chart with which you want to associate the alert rule. You can add up to three charts. You can configure an alert rule for up to three query statements at the same time. The number before the chart name is the serial number of the chart. The serial number of the chart is valid in the alert rule. You can use the serial number to specify a chart in the **Trigger Condition** parameter. You can click the icon next to the Query field to modify the query statement. The **Search Period** parameter specifies the time range of each query. You can select a relative time or a time frame. For example, the current time is 14:30:06. ○ If you set the **Search Period** parameter to 15 Minutes(Relative), the time range of the query is 14:15:06-14:30:06. ○ If you set the **Search Period** parameter to 15 Minutes(Time Frame), the time range of the query is 14:15:00-14:30:00.
Frequency	The frequency at which query results are checked.
Trigger Condition	The trigger condition of an alert. If the specified trigger condition is met, an alert is triggered and alert notifications are sent based on the values of the **Frequency** and **Notification Interval** parameters. For example, you can set the trigger condition to `pv%100 > 0 && uv > 0`. **Note** You can use `$serial number` to specify a chart in a trigger condition. For example, `$0` indicates chart 0.

Parameter	Description
Advanced	
Notification Trigger Threshold	An alert is triggered only if the specified trigger condition is met during continuous check periods. If the number of continuous triggers reaches the specified threshold, alert notifications are sent at the specified notification interval. If the trigger condition is not met, no alert is triggered. Default value: 1. This value indicates that alert notifications are sent if the trigger condition is met. You can set the Notification Trigger Threshold parameter to an integer that is greater than 1. In this case, alert notifications are sent only if the number of continuous triggers reaches the threshold. For example, you set the **Notification Trigger Threshold** parameter to 100. In this case, if the trigger condition is met for 100 times during continuous check periods, the value of **Notification Trigger Threshold** is reached. If the interval between the current time and the last time when alert notifications are sent exceeds the specified value of the **Notification Interval** parameter, an alert notification is sent. After an alert notification is sent, Log Service resets the number of continuous triggers to zero. If a check fails due to network exceptions, the check is not counted.
Notification Interval	The interval at which Log Service sends alert notifications. If the trigger condition is met in a check, Log Service checks whether the number of continuous triggers reaches the specified value of the **Notification Trigger Threshold** parameter. Log Service also checks whether the interval between the current time and the last time when alert notifications are sent exceeds the specified value of the **Notification Interval** parameter. If you set the Notification Interval parameter to 5 minutes, only one alert notification is received once every 5 minutes. **Note** You can use the Notification Trigger Threshold and Notification Interval parameters to control the number of alert notifications that you receive.

7. In the **Notifications** step, configure notification methods and click **Submit**.

   Log Service supports the following notification methods: WebHook-Custom and WebHook-DingTalk Bot. For more information, see Notification methods.

23.1.5.2.2. Authorize a RAM user to manage alert rules

You can use your Apsara Stack tenant account to authorize a RAM user to manage alert rules. This topic describes how to create a RAM user and authorize the RAM user to manage alert rules.

Procedure

1.
2. Create a RAM role.
3. Create a permission policy.

   Replace the content in the Policy Document field with the following script. Replace *Project name* in the script with the name of your Log Service project.

```
{
  "Version": "1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "log:CreateLogStore",
        "log:CreateIndex",
        "log:UpdateIndex"
      ],
      "Resource": "acs:log:*:*:project/Project name/logstore/internal-alert-history"
    },
    {
      "Effect": "Allow",
      "Action": [
        "log:CreateDashboard",
        "log:CreateChart",
        "log:UpdateDashboard"
      ],
      "Resource": "acs:log:*:*:project/Project name/dashboard/*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "log:*"
      ],
      "Resource": "acs:log:*:*:project/Project name/job/*"
    }
  ]
}
```

4. Create a user.
5. Create a RAM user group.
6. Add a RAM user to a RAM user group
7. Grant permissions to a RAM role.

23.1.5.2.3. Configure notification methods

Log Service supports the following alert notification methods: custom webhooks and DingTalk chatbot webhooks. This topic describes how to configure notification methods.

Custom webhooks

If you set the notification method to WebHook-Custom, Log Service sends alert notifications to a specified webhook URL.

> **Note** If Log Service does not receive a response within 5 seconds after a notification is sent, the request times out.

1. When you configure an alert rule, select **WebHook-Custom** from the **Notifications** drop-down list. For more information, see Configure alerts.
2. Set the parameters as described in the following table.

Parameter	Description
Request URL	The custom webhook URL.

Parameter	Description
Request Method	The method that is used to send the notification. The following request methods are supported: GET, POST, DELETE, PUT, and OPTIONS. The default request header is Content-Type: application/json;charset=utf-8. If you need to add request headers, click **Add Request Headers**.
Request content	The content of the alert notification. Log Service provides a default content. The content must be 1 to 500 characters in length. You can customize the content. You can also use template variables in the content. For more information, see Template variables.

3. Click **Submit**.

DingTalk chatbot webhooks

If you set the notification method to Webhook-DingTalk Bot, Log Service sends alert notifications by using a DingTalk chatbot to the DingTalk group to which a specified webhook URL points. The chatbot can also remind the specified contacts of the alert notifications.

> **Note** Each DingTalk chatbot can send up to 20 alert notifications per minute.

1. Create a DingTalk chatbot.
   i. Open DingTalk and go to a DingTalk group.
   ii. In the upper-right corner of the chat window, click the **Group Settings** icon and choose **Group Assistant > Add Robot**.
   iii. In the **ChatBot** dialog box, click the **+** icon in the **Add Robot** section.
   iv. In the Robot details dialog box, select **Custom (Custom message services via Webhook)** and click **Add**.
   v. In the **Add Robot** dialog box, enter a chatbot name in the **Chatbot name** field and select security options in the **Security Settings** section based on your business requirements. Then, select the **I have read and accepted DingTalk Custom Robot Service Terms of Service** check box and click **Finished**.

      > **Note** We recommend that you set the **Security Settings** parameter to **Custom Keywords**. You can specify up to 10 keywords. The chatbot sends only messages that contain at least one of the specified keywords. We recommend that you specify **Alert** as a keyword.

   vi. Click **Copy** to copy the webhook URL.
2. Configure a notification method in the Log Service console.
   i. When you configure an alert rule, select **WebHook-DingTalk Bot** from the **Notifications** drop-down list. For more information, see Configure alerts.

ii. Set the parameters as described in the following table.

Parameter	Description
Request URL	The webhook URL of the DingTalk chatbot. Paste the webhook URL that you copied in Step 1.
Title	The title of the alert notification. The title must be 1 to 100 characters in length. You can customize the title. You can also use template variables in the title. For more information, see Template variables.
Recipients	The group members whom you want to remind of the alert notification. Valid values: None, All, and Specified Members. If you select **Specified Members**, enter the mobile phone numbers of the group members in the **Tagged List** field. Separate multiple mobile phone numbers with commas (,).
Content	The content of the alert notification. Log Service provides a default content. You can modify the content based on your business requirements. The content must be 1 to 500 characters in length. You can customize the content. You can also use template variables in the content. For more information, see Template variables. **Note** If you want to remind a group member of the alert notification, use the `@<Mobile phone number of the group member>` syntax in the **Content** field.

iii. Click **Submit**.

Template variables

You can use template variables when you configure a notification method for an alert rule. When you specify the **Content** and **Subject** parameters, you can use the *${fieldName}* syntax to reference a template variable. When Log Service sends an alert notification, Log Service replaces the template variables that are referenced in the **Content** and **Subject** parameters with the actual values. For example, Log Service replaces *${Project}* with the name of the project to which the alert rule belongs.

Notice You must reference valid variables. If a referenced variable does not exist or is invalid, Log Service processes the variable as an empty string. If the value of a referenced variable is of the object type, the value is converted and displayed as a JSON string.

The following table describes the supported template variables and the methods that are used to reference the variables.

Variable	Description	Example	Reference example
Aliuid	The ID of the Apsara Stack tenant account to which the project belongs.	1234567890	An alert is triggered for the Apsara Stack tenant account ${Aliuid}.
Project	The project to which an alert rule belongs.	my-project	An alert is triggered in the ${Project} project.
AlertID	The ID of an alert.	0fdd88063a611aa114938f9371daeeb6-1671a52eb23	The ID of the alert is ${AlertID}.
AlertName	The ID of an alert rule. The ID is unique in a project.	alert-1542111415-153472	An alert is triggered based on the ${AlertName} alert rule.

Variable	Description	Example	Reference example
AlertDisplayName	The display name of an alert rule.	My alert	An alert is triggered based on the ${AlertDisplayName} alert rule.
Condition	The conditional expression that triggers an alert. In an alert notification, a variable is replaced by an actual value that is enclosed in a pair of brackets [].	[5] > 1	The conditional expression that triggers an alert is ${Condition}.
RawCondition	The original conditional expression that triggers an alert.	count > 1	The original conditional expression that triggers an alert is ${RawCondition}.
Dashboard	The name of the dashboard that is associated with an alert rule.	mydashboard	The alert rule is associated with the ${Dashboard} dashboard.
DashboardUrl	The URL of the dashboard that is associated with an alert rule.	https://sls.console.aliyun.com/next/project/myproject/dashboard/mydashboard	The URL of the dashboard that is associated with the alert rule is ${DashboardUrl}.
FireTime	The time when an alert is triggered.	2018-01-02 15:04:05	The alert is triggered at ${FireTime}.
FullResultUrl	The URL that is used to query the details of an alert.	https://sls.console.aliyun.com/next/project/my-project/logsearch/internal-alert-history?endTime=1544083998&queryString=AlertID%3A9155ea1ec10167985519fccede4d5fc7-1678293caad&queryTimeType=99&startTime=1544083968	Click ${FullResultUrl} to view the alert details.

Variable	Description	Example	Reference example

Results	The parameters and results of a query. The value is of the array type. For more information, see Fields in alert log entries.
	Note The Results variable can contain the information of up to 100 alerts.
		<pre>[
		  {
		    "EndTime": 1542507580,
		    "FireResult": {
		      "__time__": "1542453580",
		      "count": "0"
		    },
		    "LogStore": "test-logstore",
		    "Query": "* | SELECT COUNT(*) as count",
		    "RawResultCount": 1,
		    "RawResults": [
		      {
		        "__time__": "1542453580",
		        "count": "0"
		      }
		    ],
		    "StartTime": 1542453580
		  }
		]</pre>
			The start time of the first query is ${Results[0].StartTime}. The end time is ${Results[0].EndTime}. The value of count is ${Results[0].FireResult.count}.
			Note In this example, 0 is the serial number of a chart. For more information, see How can I view the serial number of a chart?

23.1.5.3. Modify and view an alarm

23.1.5.3.1. Modify an alert rule

This topic describes how to modify an alert rule. After you create an alert rule, you can modify the chart that is associated with the alert rule, and then update the alert rule. If you associate a search statement with an alert rule, you can modify the search statement to modify the alert rule.

Usage notes

- You can modify only search statements with which alert rules are associated. You cannot change a search

statement to a query statement that is in the `Search statement|Analytic statement` format.

For example, after you associate the search statement `request_method: GET` with an alert rule, you can change the search statement to `error`. You cannot change the search statement to `error| select count(1) as c`.

- You can modify an alert rule on the **Alert Overview** page of the alert rule. You can also modify an alert rule on the dashboard that is associated with the alert rule.

Modify the search statement that is associated with an alert rule

1. Log on to the Log Service console.
2. In the Projects section, click the name of the project in which you want to modify an alert rule.
3. In the left-side navigation pane, click the icon.
4. In the Dashboard list, click the name of the dashboard that you want to manage.
5. In the upper-right corner of the page, choose **Alerts > Modify**.
6. Find the search statement that you want to modify and click .

   You can modify only search statements with which alert rules are associated. You cannot change a search statement to a query statement that is in the `Search statement|Analytic statement` format.
7. In the dialog box that appears, enter a new search statement, click **Preview**, and then click **OK** after the search statement is verified.
8. Modify other parameters based on your business requirements, such as **Frequency** and **Trigger Condition**, and then click **Next**.
9. Configure notification methods and click **Submit**.

Modify the chart that is associated with an alert rule

1. In the Dashboard list, click the name of the dashboard that you want to manage.
2. In the upper-right corner of the page, choose **Alerts > Modify**.
3. Find the chart that you want to modify and then click next to **Query**.
4. In the dialog box that appears, enter a new query statement, click **Preview**, and then click **OK** after the query statement is verified.
5. Modify the other parameters based on your business requirements, such as **Frequency** and **Trigger Condition**, and then click **Next**.
6. Configure notification methods.
7. Click **Submit**.

23.1.5.3.2. View alert statistics

This topic describes how to view the alert statistics of alert rules in the Log Service console. Log Service creates a Logstore to store alert statistics as log data. Log Service also creates a dashboard to display the details of alert statistics, such as the execution results and notification results of alert rules.

Context

- You can view alert logs in the dedicated Logstore.

  After you create an alert rule in a project, Log Service creates a Logstore named **internal-alert-history**. The Logstore is used to store the alert logs of the alert rules in the project. A log is generated and written to the Logstore each time an alert rule is executed in the project, regardless of whether an alert is triggered. For more information, see Fields in alert log entries.

- You can view alert statistics on the dedicated dashboard.

  After you create an alert rule in a project, Log Service creates a dashboard named **Alert History Statistics**. The dashboard displays the alert statistics of the alert rules in the project. The alert statistics include the number of triggered alerts, the percentage of alert rule executions that are successful, the percentage of notifications that are sent for successful executions, and the top 10 alert rules that are sorted based on the total number of executions.

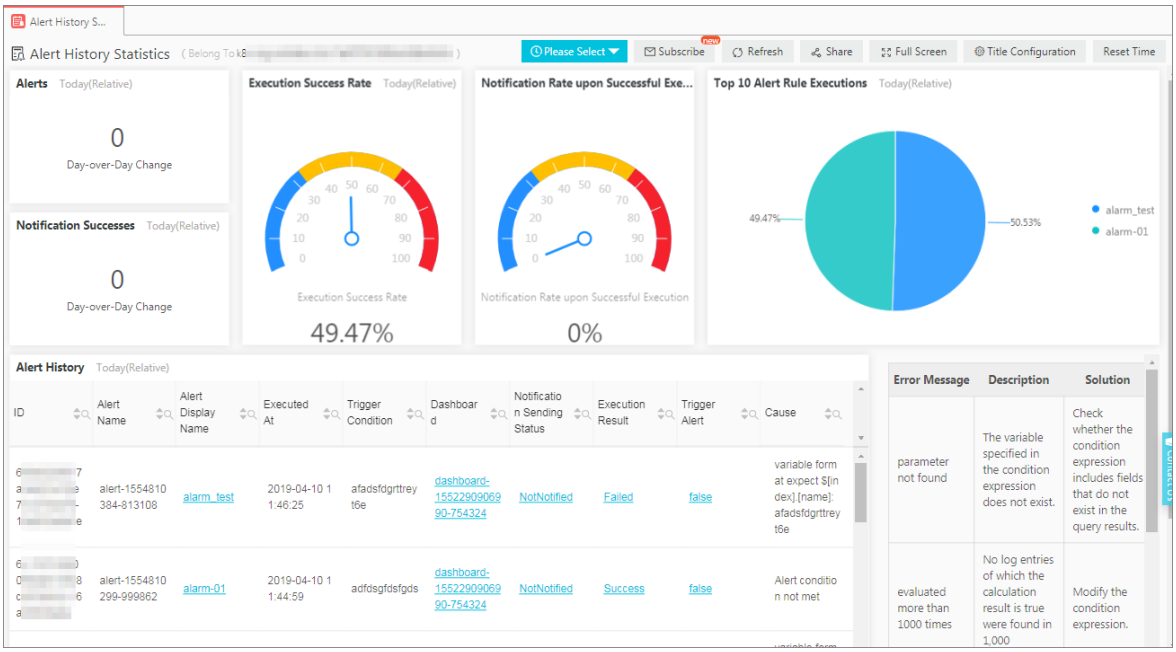
View alert logs in the dedicated Logstore

1. Log on to the Log Service console.
2. Click the name of the project in which you want to view alert logs.
3. Find the **internal-alert-history** Logstore and choose > **Search & Analysis**.
4. On the page that appears, query alert logs based on your business requirements.

View alert statistics on the dedicated dashboard

1. In the left-side navigation pane, click the icon.
2. In the Dashboard list, click **Alert History Statistics**.

   The **Alert History Statistics** dashboard displays the details of alert statistics for alert rules, such as whether an alert is triggered, the reason why the alert is triggered, and the error information that is related to the alert.



23.1.5.3.3. Manage an alert rule

After you create an alert rule, you can view the status and other information of the alert rule. You can also modify and delete the alert rule. This topic describes how to manage an alert rule.

View the information of an alert rule

1. Log on to the Log Service console.
2. In the Projects section, click the project in which you created an alert rule.
3. In the left-side navigation pane, click the alerts icon.
4. In the Alerts list, click the alert rule that you want to manage.

5. View the information of the alert rule.

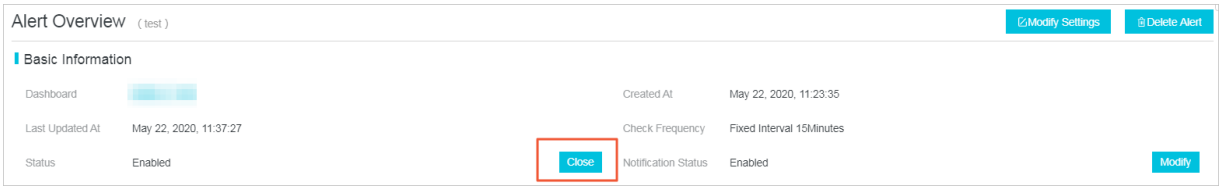
   On the **Alert Overview** page, view the information of the alert rule. You can view the name of the dashboard that is associated with the rule. You can also view the point in time when the rule was created, the last update time, the check frequency, the rule status, and the alert notification status.

Disable or enable an alert rule

After you create an alert rule, you can disable or enable the alert rule at any time.

> **Note** After you disable the alert rule, Log Service no longer checks the related data or sends alert notifications.

On the **Alert Overview** page, click the **Enable** or **Close** button next to the **Status** parameter.



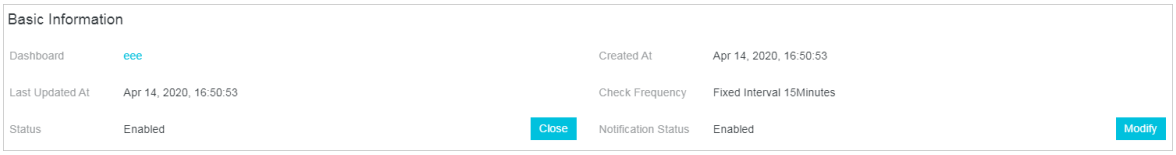
Disable or enable alert notifications for an alert rule

If an alert rule is in the **Enabled** state, you can disable or enable alert notifications for the alert rule.

> **Note** During the disabled duration that you have specified, Log Service checks data at the specified frequency based on the alert rule. However, Log Service does not send alert notifications even if the specified trigger condition is met.

1. On the **Alert Overview** page, click the **Modify** button next to the **Notification Status** field.
2. In the Disable Alert Notifications panel, set the Disabled Duration parameter and click **OK**.

   After you disable alert notifications, the point in time at which the feature is automatically enabled is displayed in the **Notification Status** parameter. If you want to enable alert notifications before the displayed time, click the **Modify** button next to the **Notification Status** parameter. In the message that appears, click OK.



Delete an alert rule

> **Warning** After you delete an alert rule, the rule cannot be restored. Proceed with caution.

In the upper-right corner of the **Alert Overview** page, click **Delete Alert**.

23.1.5.4. Relevant syntax and fields for reference

23.1.5.4.1. Syntax of conditional expressions in alert rules

Log Service checks whether the specified trigger conditions are met based on the execution results of conditional expressions specified in alert rules. The result of a specified query statement is used as input, and the fields in the result of set operations are used as variables. If the condition that is specified in a conditional expression is met, an alert is triggered.

Limits

The conditional expressions that you can specify in an alert rule have the following limits:

- Negative numbers must be enclosed in parentheses (), for example, x+(-100)<100.
- Numeric values are converted to 64-bit floating-point numbers. If a comparison operator such as equal-to (==) is used, errors may occur.
- Variable names can contain only letters and digits, and must start with a letter.
- A conditional expression must be 1 to 128 characters in length.
- You can specify up to 1,000 conditions in a conditional expression. If the evaluation result of each condition in a conditional expression is false, the evaluation result of the conditional expression is false.
- An alert rule can be associated with a maximum of three charts or query statements.
- An alert is triggered only if the result of the specified conditional expression is true. For example, if a conditional expression is 100+100, the result is 200 and is not true, and no alert is triggered.
- Log Service reserves the words true and false. Log Service also reserves the special characters dollar sign ($) and period (.). You cannot use the reserved words and special characters as variables.

Syntax

The following table describes the types of syntax that is supported for the conditional expressions of an alert rule.

Syntax type	Description	Example
Arithmetic operators	The addition (+), subtraction (-), multiplication (*), division (/), and modulus (%) operators are supported. +-*/%	• `x * 100 + y > 200` • `x % 10 > 5`
Comparison operators	The following eight comparison operators are supported: greater-than (>), greater-than-or-equal-to (>=), less-than (<), less-than-or-equal-to (<=), equal-to (==), not-equal-to (!=), regex match (=~), and regex not match (!~). **Note** • Backslashes (\) in regular expressions must be escaped. • The Regular Expression 2 (RE2) syntax is used.	• `x >= 0` • `x < 100` • `x <= 100` • `x == 100` • `x == "foo"` • Regex match: `x =~ "\\w +"`
Logical operators	The AND (&&) and OR (||) operators are supported.	• `x >= 0 && y <= 100` • `x > 0 || y >0`
Not operator	The not operator (!) is supported.	`!(a < 1 && a > 100)`
Numeric constants	Numeric constants are supported. Log Service converts numeric constants to 64-bit floating-point numbers.	`x > 100`
String constants	String constants are supported. The string constants are in the format of 'String', for example, 'string'.	`foo == 'string'`

Syntax type	Description	Example
Boolean constants	Boolean constants are supported. Valid values: true and false.	`(x > 100) == true`
Parentheses	Parentheses () can be used to override the standard precedence order and force Log Service to evaluate the enclosed part of a trigger condition before an unenclosed part.	`x * (y + 100) > 100`
contains function	The contains function can be used to check whether a string contains a substring. For example, if you invoke contains(foo,'hello') and true is returned, this indicates that the foo string contains the hello substring.	`contains(foo, 'hello')`

Evaluate the results of multiple query statements

- Syntax

  An alert rule can be associated with multiple query statements. If you want to use a variable in the trigger condition to reference a field from the result of a query statement, you must prefix the variable with the serial number of the query statement in the $N.fieldname format. The serial number of a query statement is the same as the serial number of the chart that shows the result of the query statement. Each alert rule can be associated with up to three query statements. Therefore, the value range of N is 0 to 2. For example, $0.foo references the value of the foo field from the result of the first query statement. If an alert rule is associated with only one query statement, you do not need to specify the prefix in the trigger condition.

- Evaluate a conditional expression

  If multiple query results are returned, the variables specified in the conditional expression specify the results that need to be used to evaluate the conditional expression. For example, three query statements are specified and three sets of query results are returned. The number of log entries in the first set is x, the number of log entries in the second set is y, and the number of log entries in the third set is z. If the conditional expression that you specify is $0.foo > 100 && $1.bar < 100, only the first two sets are used to evaluate the conditional expression. Up to x × y times of evaluation ,or 1,000 if x × y is greater than 1,000, is performed. If the conditional expression is met within the maximum number of times of evaluation, true is returned. Otherwise, false is returned.

Operation methods

> **Note**
>
> - Log Service converts all numeric values to 64-bit floating-point numbers.
> - A string constant must be enclosed in single quotation marks ('') or double quotation marks (""), for example, 'String' or "String".
> - Boolean values include true and false.

Operator	Operation method		
	Operation between variables	Operation between a non-string constant and a variable	Operation between a string constant and a variable
Arithmetic operators: addition (+), subtraction (-), multiplication (*), division (/), and modulus (%)	Before an arithmetic operator is applied, the left and right operands are converted to 64-bit floating-point numbers.	Before an arithmetic operator is applied, the left and right operands are converted to 64-bit floating-point numbers.	Not supported.
Comparison operators: greater-than (>), greater-than-or-equal-to (>=), less-than (<), less-than-or-equal-to (<=), equal-to (==), and not-equal-to (!=)	Log Service uses the following comparison rules that are sorted in the precedence order: 1. The left and right operands are converted to 64-bit floating-point numbers, and then compared based on the numerical order. If the conversion fails, the operation of the next priority is performed. 2. The left and right operands are converted to strings, and then compared based on the lexicographic order.	The left and right operands are converted to 64-bit floating-point numbers, and then compared based on the numerical order.	The left and right operands are converted to strings, and then compared based on the lexicographic order.
Matching operators: regex match (=~) and regex not match (!~)	Before a matching operator is applied, the left and right operands are converted to strings.	Not supported.	Before a matching operator is applied, the left and right operands are converted to strings.
Logical operators: AND (&&) and OR (||)	A logical operator cannot be applied to log fields. The left and right operands must be sub-expressions and the result of the operation must be a Boolean value.	A logical operator cannot be applied to log fields. The left and right operands must be sub-expressions and the result of the operation must be a Boolean value.	A logical operator cannot be applied to log fields. The left and right operands must be sub-expressions and the result of the operation must be a Boolean value.

Operator	Operation method		
	Operation between variables	Operation between a non-string constant and a variable	Operation between a string constant and a variable
Not operator (!)	The not operator cannot be applied to log fields. The specified operand must be a sub-expression and the result of the operation must be a Boolean value.	The not operator cannot be applied to log fields. The specified operand must be a sub-expression and the result of the operation must be a Boolean value.	The not operator cannot be applied to log fields. The specified operand must be a sub-expression and the result of the operation must be a Boolean value.
contains function	Before the contains function is run, the left and right operands are converted to strings.	Not supported.	Before the contains function is run, the left and right operands are converted to strings.
Parentheses ()	Parentheses () are used to override the standard precedence order and force Log Service to evaluate the enclosed part of a trigger condition before an unenclosed part.	Parentheses () are used to override the standard precedence order and force Log Service to evaluate the enclosed part of a trigger condition before an unenclosed part.	Parentheses () are used to override the standard precedence order and force Log Service to evaluate the enclosed part of a trigger condition before an unenclosed part.

23.1.5.4.2. Fields in alert logs

After you configure an alert rule, Log Service automatically creates a Logstore to store log entries that are generated when the alert rule is executed and alert notifications are sent. This topic describes the fields in alert logs.

Fields in the logs that are generated when an alert rule is executed

Log field	Description	Example
AlertDisplayName	The display name of an alert rule.	Test alert rule
AlertID	The ID of an alert. The ID is unique.	0fdd88063a611aa114938f9371daeeb6-1671a52eb23
AlertName	The name of the alert rule. The name is unique within a project.	alert-1542111415-153472

Log field	Description	Example
Condition	The conditional expression of an alert rule.	$0.count > 1
Dashboard	The dashboard that is associated with the alert rule.	my-dashboard
FireCount	The cumulative number of triggers since the last alert notification was sent.	1
Fired	Indicates whether an alert was triggered. Valid values: true and false.	true
LastNotifiedAt	The time when the last alert notification was sent. The value is a UNIX timestamp.	1542164541
NotifyStatus	The notification status of an alert. Valid values: • Success: Alert notifications were sent. • Failed: Alert notifications failed to be sent. • NotNotified: No alert notification was sent. • PartialSuccess: Some of the alert notifications were sent.	Success
Reason	The reason why alert notifications failed to be sent or no alert notification was sent.	result type is not bool

Results	The parameters and results of each log query. The value is of the array type. For more information, see Results field.	[
		{
		"EndTime": 1542334900,
		"FireResult": null,
		"LogStore": "test-logstore",
		"Query": "* | select count(1) as count",
		"RawResultCount": 1,
		"RawResults": [
		{
		"__time__": "1542334840",
		"count": "0"
		}
		],
		"StartTime": 1542334840
		}
		]

Status	The execution result of an alert. Valid values: Success and Failed.	Success

Results field

Log field	Description	Example
Query	The query statement.	* | select count(1) as count
LogStore	The Logstore in which data is queried.	my-logstore

Log field	Description	Example
StartTime	The beginning of the time range for a query.	2019-01-02 15:04:05
StartTimeTs	The beginning of the time range for a query. The value is a UNIX timestamp.	1542334840
EndTime	The end of the time range for a query.	2019-01-02 15:19:05
EndTimeTs	The end of the time range for a query. The value is a UNIX timestamp. The actual query time range is `[StartTime, EndTime)`.	1542334900
RawResults	The query result that is formatted in an array. Each element in the array is a log entry. An array can contain a maximum of 100 elements.	`[ { "__time__": "1542334840", "count": "0" } ]`
RawResultsAsKv	The query result that is formatted in key-value pairs. **Note** This field can be used as a template variable. No data is stored to a Logstore for this field.	[foo:0]
RawResultCount	The number of raw log entries that are returned.	1
FireResult	The log entry that records the triggers of an alert. If no alert is triggered, the value is NULL.	`{ "__time__": "1542334840", "count": "0" }`
FireResultAsKv	The log entry that records the triggers of an alert. The log entry is formatted in key-value pairs. **Note** This field can be used as a template variable. No data is stored to a Logstore for this field.	[foo:0]

23.1.6. Real-time consumption

23.1.6.1. Overview

Log Service provides the real-time log consumption feature that allows you to read and write full data in the first-in, first-out (FIFO) order. This feature is similar to the features provided by Kafka. This topic describes the types of real-time consumption.

The following table describes the methods that you can use to process log data after the log data is sent to Log Service.

Method	Scenario	Timeliness	Retention period
Real-time log consumption	Stream computing and real-time computing	Real time	You can specify a retention period based on your business requirements.
Log query	Online query of recent hot data	Near real time with a latency of no more than 3 seconds in all cases and a latency of no more than 1 second in 99.9% cases	You can specify a retention period based on your business requirements.
Log shipping	Storage of full log data for offline analysis	A latency of 5 to 30 minutes	The retention period is based on the storage system.

Real-time log consumption

Log Service allows you to pull log data and consume the data in real time. The following procedure describes how log data is consumed from a shard:

1. Obtain a cursor based on the start time and end time of data consumption.
2. Read log data based on the cursor and step parameters and return the position of the next cursor.
3. Move the cursor to continuously consume log data.

- Use an SDK to consume log data

  Log Service provides SDKs in multiple programming languages, such as Java, Python, and Go. You can use an SDK to consume log data.

- Use consumer groups to consume log data

  Log Service provides an advanced method that allows you to consume log data by using consumer groups. A consumer group is a lightweight computing framework that allows multiple consumers to consume data from a Logstore at the same time. Shards are automatically allocated to consumers in a consumer group. Data is consumed in sequence based on the time when data is written to the Logstore. After a breakpoint, consumers can continue to consume data by using checkpoints. You can use consumer group SDKs in multiple programming languages, such as Go, Python, and Java, to consume log data.

- Use real-time computing systems to consume log data
  - Use Spark Streaming to consume log data.
  - Use Storm to consume log data
  - Use open source Flink to consume log data

- Use open source services to consume log data

  Use Flume to consume log data

23.1.6.2. Consume log data

This topic describes how to use an SDK to consume log data and preview log data in the Log Service console.

Use an SDK to consume log data

Log Service provides SDKs in various programming languages. You can use an SDK to consume log data. For more information, see the SDK Reference topic in *Log Service Developer Guide*. The following example shows how to use Log Service SDK for Java to consume log data from a shard:

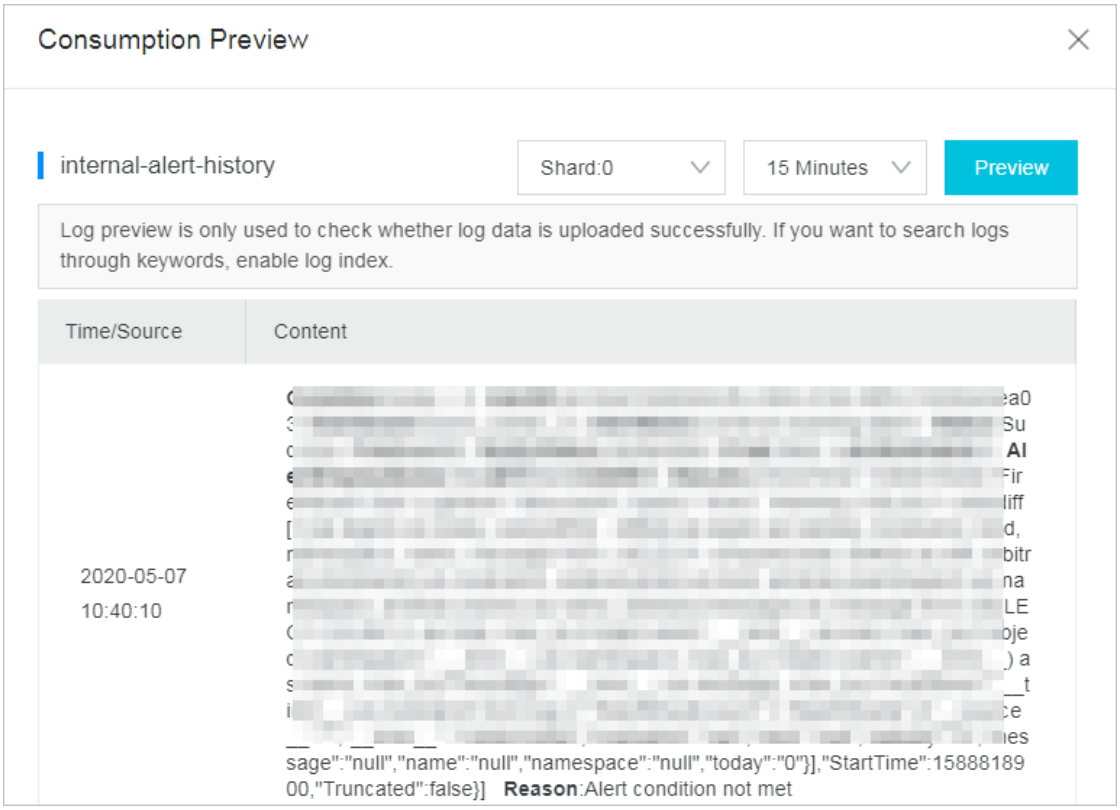
```
Client client = new Client(host, accessId, accessKey);
    String cursor = client.GetCursor(project, logStore, shardId, CursorMode.END).GetCursor();
    System.out.println("cursor = " +cursor);
    try {
      while (true) {
        PullLogsRequest request = new PullLogsRequest(project, logStore, shardId, 1000, cursor);
        PullLogsResponse response = client.pullLogs(request);
        System.out.println(response.getCount());
        System.out.println("cursor = " + cursor + " next_cursor = " + response.getNextCursor());
        if (cursor.equals(response.getNextCursor())) {
            break;
                }
        cursor = response.getNextCursor();
        Thread.sleep(200);
      }
    }
    catch(LogException e) {
      System.out.println(e.GetRequestId() + e.GetErrorMessage());
    }
```

Preview log data in the Log Service console

Consumption preview is a type of log data consumption. The consumption preview feature allows you to preview specific log data that is stored in a Logstore in the Log Service console.

1. Log on to the Log Service console.
2. In the Projects section, click the project from which you want to consume log data.
3. In the Logstores list, find the Logstore from which you want to consume log data, click the icon next to the Logstore, and then select **Consumption Preview**.
4. In the **Consumption Preview** panel, select a shard and a time range, and then click **Preview**.

   The Consumption Preview panel displays the log data of the first 10 packets in the specified time range.



23.1.6.3. Consumption by consumer groups

23.1.6.3.1. Use consumer groups to consume log data

If you use consumer groups to consume log data, you do not need to focus on factors such as Log Service implementation, load balancing among consumers, and failovers that may occur. This way, you can focus on the business logic during log data consumption.

Terms

Term	Description
consumer group	A consumer group consists of multiple consumers. Each consumer in a consumer group consumes different data in a Logstore. You can create a maximum of 30 consumer groups for a Logstore.
consumer	The consumers in a consumer group consume data. The name of each consumer in a consumer group must be unique.

A Logstore has multiple shards. A consumer library allocates shards to the consumers in a consumer group based on the following rules:

- You can allocate a shard to only one consumer.
- Each consumer can consume data from multiple shards.

After you add a consumer to a consumer group, shards that are allocated to the consumer group are reallocated to each consumer for load balancing. The shards are reallocated based on the preceding rules.

A consumer library stores checkpoints. This way, consumers can resume data consumption from a checkpoint and do not consume data after a program fault is resolved.

Procedure

You can use Java, Python, or Go to create consumers and consume data. The following procedure uses Java as an example:

1. Add Maven dependencies.

```
<dependency>
  <groupId>com.google.protobuf</groupId>
  <artifactId>protobuf-java</artifactId>
  <version>2.5.0</version>
</dependency>
<dependency>
  <groupId>com.aliyun.openservices</groupId>
  <artifactId>loghub-client-lib</artifactId>
  <version>0.6.33</version>
</dependency>
```

2. Create a file named Main.java.

```
import com.aliyun.openservices.loghub.client.ClientWorker;
import com.aliyun.openservices.loghub.client.config.LogHubConfig;
import com.aliyun.openservices.loghub.client.exceptions.LogHubClientWorkerException;
public class Main {
    // The endpoint of Log Service. Set the parameter based on your business requirements.
    private static String sEndpoint = "cn-hangzhou.log.aliyuncs.com";
    // The name of a Log Service project. Set the parameter based on your business requirements.
    private static String sProject = "ali-cn-hangzhou-sls-admin";
    // The name of a Logstore. Set the parameter based on your business requirements.
    private static String sLogstore = "sls_operation_log";
    // The name of a consumer group. Set the parameter based on your business requirements.
    private static String sConsumerGroup = "consumerGroupX";
    // The AccessKey pair that is used to access Log Service. Specify the AccessKey ID and AccessK
ey secret based on your business requirements.
    private static String sAccessKeyId = "";
    private static String sAccessKey = "";
    public static void main(String[] args) throws LogHubClientWorkerException, InterruptedExceptio
n {
        // consumer_1 is the name of the consumer. The name of each consumer in a consumer group m
ust be unique. If different consumers start multiple processes on multiple servers to consume the
data of a Logstore, you can use a server IP address to identify a consumer.
        The maxFetchLogGroupSize parameter specifies the maximum number of log groups that you can
obtain from the server at the same time. Valid values: (0,1000]. We recommend that you use the def
ault value.
        LogHubConfig config = new LogHubConfig(sConsumerGroup, "consumer_1", sEndpoint, sProject,
sLogstore, sAccessKeyId, sAccessKey, LogHubConfig.ConsumePosition.BEGIN_CURSOR);
        ClientWorker worker = new ClientWorker(new SampleLogHubProcessorFactory(), config);
        Thread thread = new Thread(worker);
        // After you execute the thread, the ClientWorker instance automatically runs and extends
the Runnable interface.
        thread.start();
        Thread.sleep(60 * 60 * 1000);
        // The shutdown function of the ClientWorker instance is called to exit the consumption in
stance. The associated thread is automatically stopped.
        worker.shutdown();
        // Multiple asynchronous tasks are generated when the ClientWorker instance is running. To
ensure that all running tasks exit after shutdown, we recommend that you set Thread.sleep to 30 se
conds.
        Thread.sleep(30 * 1000);
    }
}
```

3. Create a file named SampleLogHubProcessor.java.

```
import com.aliyun.openservices.log.common.FastLog;
import com.aliyun.openservices.log.common.FastLogContent;
import com.aliyun.openservices.log.common.FastLogGroup;
import com.aliyun.openservices.log.common.FastLogTag;
import com.aliyun.openservices.log.common.LogGroupData;
import com.aliyun.openservices.loghub.client.ILogHubCheckPointTracker;
import com.aliyun.openservices.loghub.client.exceptions.LogHubCheckPointException;
import com.aliyun.openservices.loghub.client.interfaces.ILogHubProcessor;
import com.aliyun.openservices.loghub.client.interfaces.ILogHubProcessorFactory;
import java.util.List;
public class SampleLogHubProcessor implements ILogHubProcessor {
    private int shardId;
    // The point in time when the last persistent checkpoint was saved.
    private long mLastCheckTime = 0;
    public void initialize(int shardId) {
        this.shardId = shardId;
```

```
    }
    // The main logic of data consumption. You must include the code to handle all exceptions that
may occur during log data consumption.
    public String process(List<LogGroupData> logGroups,
                          ILogHubCheckPointTracker checkPointTracker) {
        // Display the data that you obtained.
        for (LogGroupData logGroup : logGroups) {
            FastLogGroup flg = logGroup.GetFastLogGroup();
            System.out.println(String.format("\tcategory\t:\t%s\n\tsource\t:\t%s\n\ttopic\t:\t%s\n
\tmachineUUID\t:\t%s",
                    flg.getCategory(), flg.getSource(), flg.getTopic(), flg.getMachineUUID()));
            System.out.println("Tags");
            for (int tagIdx = 0; tagIdx < flg.getLogTagsCount(); ++tagIdx) {
                FastLogTag logtag = flg.getLogTags(tagIdx);
                System.out.println(String.format("\t%s\t:\t%s", logtag.getKey(), logtag.getValue()
));
            }
            for (int lIdx = 0; lIdx < flg.getLogsCount(); ++lIdx) {
                FastLog log = flg.getLogs(lIdx);
                System.out.println("--------\nLog: " + lIdx + ", time: " + log.getTime() + ", GetC
ontentCount: " + log.getContentsCount());
                for (int cIdx = 0; cIdx < log.getContentsCount(); ++cIdx) {
                    FastLogContent content = log.getContents(cIdx);
                    System.out.println(content.getKey() + "\t:\t" + content.getValue());
                }
            }
        }
        long curTime = System.currentTimeMillis();
        // Write a checkpoint to the server every 30 seconds. If the ClientWorker instance unexpec
tedly stops within 30 seconds, a newly started ClientWorker instance continues to consume data fro
m the last checkpoint. A small amount of data may be repeatedly consumed.
        if (curTime - mLastCheckTime > 30 * 1000) {
            try {
                // If you set the parameter to true, checkpoints are immediately synchronized to t
he server. If you set the parameter to false, checkpoints are locally cached. The default value of
a synchronization interval of checkpoints is 60 seconds.
                checkPointTracker.saveCheckPoint(true);
            } catch (LogHubCheckPointException e) {
                e.printStackTrace();
            }
            mLastCheckTime = curTime;
        }
        return null;
    }
    // The ClientWorker instance calls this function when the instance exits. You can delete the c
heckpoints.
    public void shutdown(ILogHubCheckPointTracker checkPointTracker) {
        // Save checkpoints to the server.
        try {
            checkPointTracker.saveCheckPoint(true);
        } catch (LogHubCheckPointException e) {
            e.printStackTrace();
        }
    }
}
class SampleLogHubProcessorFactory implements ILogHubProcessorFactory {
    public ILogHubProcessor generatorProcessor() {
        // Generate a consumption instance.
        return new SampleLogHubProcessor();
    }
```

```
}
```

For more information, see Java, Python, and Go.

View the status of a consumer group

You can use the Log Service console, call the API, or use an SDK to view the progress of your data consumption. For more information, see View the status of a consumer group.

Related operations

- Handle exceptions.

  We recommend that you configure Log4j for the consumer program to return error messages in consumer groups. This way, you can handle exceptions at the earliest opportunity. The following code shows a configuration file of log4j.properties:

  ```
  log4j.rootLogger = info,stdout
  log4j.appender.stdout = org.apache.log4j.ConsoleAppender
  log4j.appender.stdout.Target = System.out
  log4j.appender.stdout.layout = org.apache.log4j.PatternLayout
  log4j.appender.stdout.layout.ConversionPattern = [%-5p] %d{yyyy-MM-dd HH:mm:ss,SSS} method:%l%n%m%n
  ```

  After you configure Log4j, you can view the information of exceptions that occur when you run the consumer program. The following example shows an error message:

  ```
  [WARN ] 2018-03-14 12:01:52,747 method:com.aliyun.openservices.loghub.client.LogHubConsumer.sampleL
  ogError(LogHubConsumer.java:159)
  com.aliyun.openservices.log.exception.LogException: Invalid loggroup count, (0,1000]
  ```

- Use a consumer group to consume data that is generated from a certain point in time.

  ```
  // consumerStartTimeInSeconds indicates that the data generated after the point in time is consumed
  .
  public LogHubConfig(String consumerGroupName,
                      String consumerName,
                      String loghubEndPoint,
                      String project, String logStore,
                      String accessId, String accessKey,
                      int consumerStartTimeInSeconds);
  // The value of the position parameter is an enumeration variable. LogHubConfig.ConsumePosition.BEG
  IN_CURSOR indicates that the consumption starts from the earliest data. LogHubConfig.ConsumePositio
  n.END_CURSOR indicates that the consumption starts from the latest data.
  public LogHubConfig(String consumerGroupName,
                      String consumerName,
                      String loghubEndPoint,
                      String project, String logStore,
                      String accessId, String accessKey,
                      ConsumePosition position);
  ```

  Note
  - You can use different constructors based on your business requirements.
  - If a checkpoint is stored on the server, data consumption starts from this checkpoint.

- Reset a checkpoint.

```
public static void updateCheckpoint() throws Exception {
        Client client = new Client(host, accessId, accessKey);
        long timestamp = Timestamp.valueOf("2017-11-15 00:00:00").getTime() / 1000;
        ListShardResponse response = client.ListShard(new ListShardRequest(project, logStore));
        for (Shard shard : response.GetShards()) {
            int shardId = shard.GetShardId();
            String cursor = client.GetCursor(project, logStore, shardId, timestamp).GetCursor();
            client.UpdateCheckPoint(project, logStore, consumerGroup, shardId, cursor);
        }
    }
```

Authorize a RAM user to access consumer groups

Before you use a RAM user to access consumer groups, you must grant the required permissions to the RAM user. For more information, see Grant permissions to a RAM role.

The following table describes the actions that a RAM user can perform.

Action	Description	Resource
log:GetCursorOrData (GetCursor and PullLogs)	Obtains a cursor based on the point in time when log data is generated.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*
log:CreateConsumerGroup	Creates a consumer group for a specified Logstore.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*/consumergroup/*
log:ListConsumerGroup	Queries all consumer groups in a specified Logstore.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*/consumergroup/*
log:UpdateCheckPoint	Updates the consumption checkpoint in a shard of a specified consumer group.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*/consumergroup/*${consumerGroupName}*
log:ConsumerGroupHeartBeat	Sends a heartbeat packet to Log Service for a specified consumer.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*/consumergroup/*${consumerGroupName}*
log:UpdateConsumerGroup	Modifies the attributes of a specified consumer group.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*/consumergroup/*${consumerGroupName}*
log:ConsumerGroupUpdateCheckPoint	Retrieves the consumption checkpoints in one or all shards of a specified consumer group.	acs:log:*${regionName}*:*${projectOwnerAliUid}*:project/*${projectName}*/logstore/*${logstoreName}*/consumergroup/*${consumerGroupName}*

For example, the project-test project resides in the China (Hangzhou) region. The ID of the Apsara Stack tenant account to which the project belongs is 174649****602745. The name of the Logstore from which you want to consume log data is logstore-test, and the consumer group name is consumergroup-test. To allow a RAM user to access the consumer group, you must grant the following permissions to the RAM user:

```
{
  "Version": "1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "log:GetCursorOrData"
      ],
      "Resource": "acs:log:cn-hangzhou:174649****602745:project/project-test/logstore/logstore-test"
    },
    {
      "Effect": "Allow",
      "Action": [
        "log:CreateConsumerGroup",
        "log:ListConsumerGroup"
      ],
      "Resource": "acs:log:cn-hangzhou:174649****602745:project/project-test/logstore/logstore-test/consumergroup/*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "log:ConsumerGroupUpdateCheckPoint",
        "log:ConsumerGroupHeartBeat",
        "log:UpdateConsumerGroup",
        "log:GetConsumerGroupCheckPoint"
      ],
      "Resource": "acs:log:cn-hangzhou:174649****602745:project/project-test/logstore/logstore-test/consumergroup/consumergroup-test"
    }
  ]
}
```

23.1.6.3.2. View the status of a consumer group

This topic describes how to view the status of a consumer group.

View the consumption progress in the Log Service console

1. Log on to the Log Service console.
2. In the Projects section, click the project that you want to manage.
3. In the Logstores list, find the Logstore that you want to manage and choose > **Data Consumption**.
4. Click the consumer group whose data consumption progress you want to view. The data consumption progress of each shard in the Logstore is displayed.

Call the API or use an SDK to view the data consumption progress

The following code shows how to call the API to view the data consumption progress. In this example, Log Service SDK for Java is used.

```
package test;
import java.util.ArrayList;
import com.aliyun.openservices.log.Client;
import com.aliyun.openservices.log.common.Consts.CursorMode;
import com.aliyun.openservices.log.common.ConsumerGroup;
import com.aliyun.openservices.log.common.ConsumerGroupShardCheckPoint;
import com.aliyun.openservices.log.exception.LogException;
```

```
public class ConsumerGroupTest {
    static String endpoint = "";
    static String project = "";
    static String logstore = "";
    static String accesskeyId = "";
    static String accesskey = "";
    public static void main(String[] args) throws LogException {
        Client client = new Client(endpoint, accesskeyId, accesskey);
        // Obtain all consumer groups that are created for the Logstore. If no consumer group exists,
an empty string is returned.
        List<ConsumerGroup> consumerGroups = client.ListConsumerGroup(project, logstore).GetConsumerGr
oups();
        for(ConsumerGroup c: consumerGroups){
            // Print the properties of consumer groups. The properties of a consumer group include the
name, heartbeat timeout, and whether the consumer group consumes data in order.
            System.out.println("Name: " + c.getConsumerGroupName());
            System.out.println("Heartbeat timeout: " + c.getTimeout());
            System.out.println("Consumption in order: " + c.isInOrder());
            for(ConsumerGroupShardCheckPoint cp: client.GetCheckPoint(project, logstore, c.getConsumer
GroupName()).GetCheckPoints()){
                System.out.println("shard: " + cp.getShard());
                // The consumption time. The time is a long integer and is accurate to milliseconds.
                System.out.println("The last time when data was consumed: " + cp.getUpdateTime());
                System.out.println("Consumer name: " + cp.getConsumer());
                String consumerPrg = "";
                if(cp.getCheckPoint().isEmpty())
                    consumerPrg = "Consumption not started";
                else{
                    // The UNIX timestamp. Unit: seconds. Format the output value of the timestamp.
                    try{
                        int prg = client.GetPrevCursorTime(project, logstore, cp.getShard(), cp.getChe
ckPoint()).GetCursorTime();
                        consumerPrg = "" + prg;
                    }
                    catch(LogException e){
                        if(e.GetErrorCode() == "InvalidCursor")
                            consumerPrg = "Invalid. The previous point in time when data was consumed
is out of the retention period of the data in the Logstore";
                        else{
                            //internal server error
                            throw e;
                        }
                    }
                }
                System.out.println("Consumption progress: " + consumerPrg);
                String endCursor = client.GetCursor(project, logstore, cp.getShard(), CursorMode.END).
GetCursor();
                int endPrg = 0;
                try{
                    endPrg = client.GetPrevCursorTime(project, logstore, cp.getShard(), endCursor).Get
CursorTime();
                }
                catch(LogException e){
                    //do nothing
                }
                // The UNIX timestamp. Unit: seconds. Format the output value of the timestamp.
                System.out.println("The point in time when the last data entry was received: " + endPr
g);
            }
        }
```

```
    }
}
```

23.1.6.4. Use Storm to consume log data

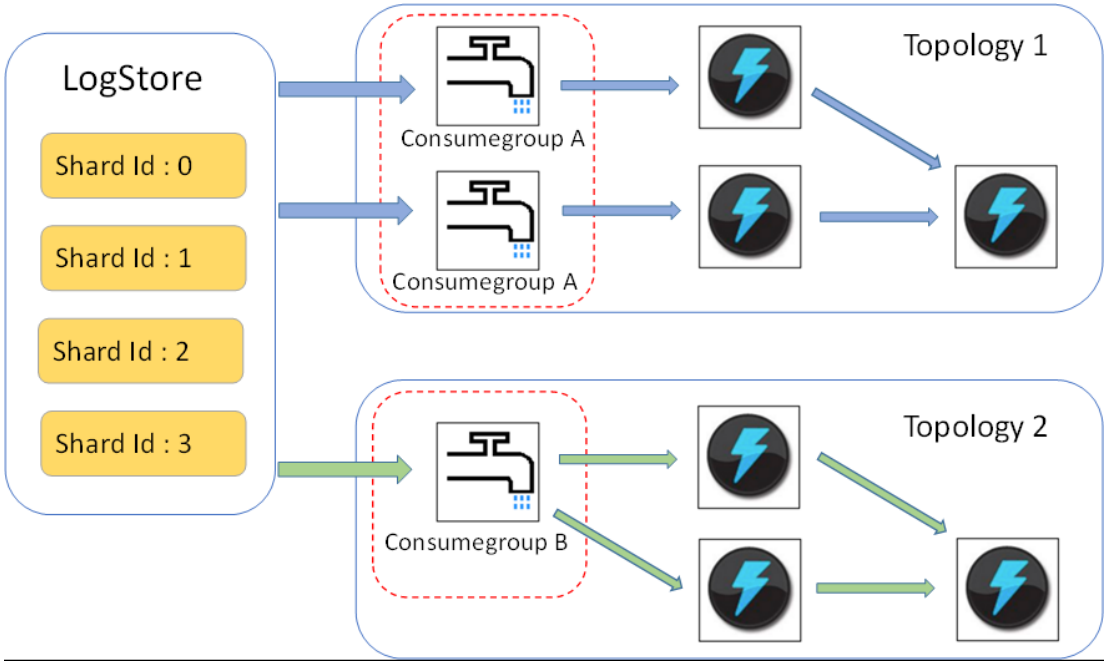
Log Service LogHub provides efficient and reliable log collection and consumption channels. You can use services such as Logtail or SDKs to collect log data in real time. After log data is collected and sent to Log Service, you can use stream computing systems such as Spark Streaming and Apache Storm to consume the log data.

To reduce the costs of data consumption, Log Service provides LogHub Storm spouts to read data from Log Service in real time.

Architecture and implementation

- In the following figure, LogHub Storm spouts are enclosed in red dashed-line boxes. Each Storm topology has a group of spouts that work together to read data from a Logstore. Spouts in different topologies are independent of each other.
- Each topology is identified by the unique name of a LogHub consumer group. Spouts in the same topology use a consumer group to perform load balancing and automatic failover. For more information, see Use consumer groups to consume log data.
- Spouts in a topology read data from a Logstore in real time, and then send the data to bolts in the topology. The spouts save consumption checkpoints to the LogHub server on a regular basis.

Architecture and implementation



Limits

- You can create a maximum of 10 consumer groups to consume log data from a Logstore. If you no longer need a consumer group, you can call the DeleteConsumerGroup operation of the SDK for Java to delete the consumer group.
- We recommend that you configure the same number of spouts for a Logstore as the number of shards in the Logstore. This is because a single spout may not be able to process a large amount of data from multiple shards.
- If the data volume in a shard exceeds the processing capacity of a single spout, you can split the shard to reduce the data volume in each shard.
- LogHub Storm spouts and bolts must use the ack method to check whether log data is sent from spouts to

bolts and whether the data is processed by the bolts.

Examples

- The following code provides an example to show how to construct a Storm topology:

```
     public static void main( String[] args )
    {
        String mode = "Local"; // Use the local test mode.
           String consumser_group_name = ""; // The name of the consumer group of a topology. The name must be unique. The name cannot be an empty string. The name must be 3 to 63 characters in length and can contain lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or a digit.
        String project = ""; // The Log Service project.
        String logstore = ""; // The Log Service Logstore.
        String endpoint = ""; // The endpoint of Log Service.
        String access_id = ""; // The AccessKey ID.
        String access_key = "";
        // Configure a LogHub Storm spout.
        LogHubSpoutConfig config = new LogHubSpoutConfig(consumser_group_name,
                endpoint, project, logstore, access_id,
                access_key, LogHubCursorPosition.END_CURSOR);
        TopologyBuilder builder = new TopologyBuilder();
        // Create a LogHub Storm spout.
        LogHubSpout spout = new LogHubSpout(config);
        // In actual scenarios, we recommend that you create the same number of spouts for a Logstore as the number of shards in the Logstore.
        builder.setSpout("spout", spout, 1);
        builder.setBolt("exclaim", new SampleBolt()).shuffleGrouping("spout");
        Config conf = new Config();
        conf.setDebug(false);
        conf.setMaxSpoutPending(1);
        // If you use Kryo to serialize and deserialize data, configure the serialization method of LogGroupData by using the LogGroupDataSerializSerializer class.
        Config.registerSerialization(conf, LogGroupData.class, LogGroupDataSerializSerializer.class);
        if (mode.equals("Local")) {
            logger.info("Local mode...");
            LocalCluster cluster  = new LocalCluster();
            cluster.submitTopology("test-jstorm-spout", conf, builder.createTopology());
            try {
                Thread.sleep(6000 * 1000);    //waiting for several minutes
            } catch (InterruptedException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
            }
            cluster.killTopology("test-jstorm-spout");
            cluster.shutdown();
        } else if (mode.equals("Remote")) {
            logger.info("Remote mode...");
            conf.setNumWorkers(2);
            try {
                StormSubmitter.submitTopology("stt-jstorm-spout-4", conf, builder.createTopology());
            } catch (AlreadyAliveException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
            } catch (InvalidTopologyException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
            }
```

```
            }
        } else {
            logger.error("invalid mode: " + mode);
        }
    }
}
```

- The following code provides an example to show how to consume log data, and then display the content of each log entry by using bolts:

```
public class SampleBolt extends BaseRichBolt {
    private static final long serialVersionUID = 4752656887774402264L;
    private static final Logger logger = Logger.getLogger(BaseBasicBolt.class);
    private OutputCollector mCollector;
    @Override
    public void prepare(@SuppressWarnings("rawtypes") Map stormConf, TopologyContext context,
            OutputCollector collector) {
        mCollector = collector;
    }
    @Override
    public void execute(Tuple tuple) {
        String shardId = (String) tuple
                .getValueByField(LogHubSpout.FIELD_SHARD_ID);
        @SuppressWarnings("unchecked")
        List<LogGroupData> logGroupDatas = (ArrayList<LogGroupData>) tuple.getValueByField(LogHubSp
out.FIELD_LOGGROUPS);
        for (LogGroupData groupData : logGroupDatas) {
            // Each log group consists of one or more log entries.
            LogGroup logGroup = groupData.GetLogGroup();
            for (Log log : logGroup.getLogsList()) {
                StringBuilder sb = new StringBuilder();
                // Each log entry contains a time field and other fields that are formatted in key-
value pairs.
                int log_time = log.getTime();
                sb.append("LogTime:").append(log_time);
                for (Content content : log.getContentsList()) {
                    sb.append("\t").append(content.getKey()).append(":")
                            .append(content.getValue());
                }
                logger.info(sb.toString());
            }
        }
        // LogHub spouts and bolts must use the ack method to check whether log data is sent from s
pouts to bolts and whether the data is processed by the bolts.
        mCollector.ack(tuple);
    }
    @Override
    public void declareOutputFields(OutputFieldsDeclarer declarer) {
        //do nothing
    }
}
```

Maven

The following code provides an example to show how to add Maven dependencies for Storm 1.0 or earlier, such as Storm 0.9.6:

```
<dependency>
  <groupId>com.aliyun.openservices</groupId>
  <artifactId>loghub-storm-spout</artifactId>
  <version>0.6.6</version>
</dependency>
```

The following code provides an example to show how to add Maven dependencies for Storm 1.0 or later:

```
<dependency>
  <groupId>com.aliyun.openservices</groupId>
  <artifactId>loghub-storm-1.0-spout</artifactId>
  <version>0.1.3</version>
</dependency>
```

23.1.6.5. Use Flume to consume log data

This topic describes how to use Flume to consume log data. You can use the aliyun-log-flume plug-in to connect Log Service to Flume and write log data to Log Service or consume log data from Log Service.

Context

The aliyun-log-flume plug-in connects Log Service to Flume. When Log Service is connected to Flume, you can use Flume to connect Log Service to other systems such as Hadoop distributed file system (HDFS) and Kafka. The aliyun-log-flume plug-in provides sinks and sources to connect Log Service to Flume.

- Sink: reads data from other data sources and writes the data to Log Service.
- Source: consumes log data from Log Service and writes the log data to other systems.

For more information, visit GitHub.

Procedure

1. Download and install Flume. For more information, see Flume.
2. Download the aliyun-log-flume plug-in and save the plug-in in the *cd/***/flume/lib* directory. To download the plug-in, click aliyun-log-flume-1.3.jar.
3. In the *cd/***/flume/conf* directory, create a configuration file named flumejob.conf.
   - For information about how to configure a sink, see Sink.
   - For information about how to configure a source, see Source.
4. Start Flume.

Sink

You can configure a sink for Flume to write data from other data sources to Log Service. Data can be parsed into the following two formats:

- SIMPLE: A Flume event is written to Log Service as a field.
- DELIMITED: A Flume event is parsed into fields based on the configured column names and written to Log Service.

The following table describes the parameters of a sink.

Parameter	Required	Description
type	Yes	Default value: `com.aliyun.Loghub.flume.sink.LoghubSink`.
endpoint	Yes	The endpoint of the region where the Log Service project resides.

Parameter	Required	Description
project	Yes	The name of the project.
logstore	Yes	The name of the Logstore.
accessKeyId	Yes	The AccessKey ID that is used to access Log Service. For more information, see Obtain an AccessKey pair.
accessKey	Yes	The AccessKey secret that is used to access Log Service. For more information, see Obtain an AccessKey pair.
batchSize	No	The number of data entries that are written to Log Service at a time. Default value: 1000.
maxBufferSize	No	The maximum number of data entries in the buffer. Default value: 1000.
serializer	No	The serialization format of the Flume event. Default value: SIMPLE. Valid values: • **DELIMITED**: delimiter mode. • **SIMPLE**: single-line mode. • Custom **serializer**: custom serialization mode. In this mode, you must specify the full names of columns.
columns	No	The names of colums. If you set the serializer parameter to **DELIMITED**, you must specify this parameter. Separate multiple columns with commas (,). The columns are sorted in the same order that the columns are sorted in the data entries.
separatorChar	No	The delimiter. If you set the serializer parameter to **DELIMITED**, you must specify a single character for this parameter. The default value is a comma (,).
quoteChar	No	The quote character. If you set the serializer parameter to **DELIMITED**, you must specify this parameter. The default value is double quotation marks (").
escapeChar	No	The escape character. If you set the serializer parameter to **DELIMITED**, you must specify this parameter. The default value is double quotation marks (").
useRecordTime	No	Specifies whether to use the value of the timestamp field in the data entries as the log time when data is written to Log Service. Default value: false. This value indicates that the current time is used as the log time.

For more information about how to configure a sink, visit GitHub.

Source

You can configure a source for Flume to ship data from Log Service to other data sources. Data can be parsed into the following two formats:

- DELIMITED: Log data is written to Flume in delimiter mode.
- JSON: Log data is written to Flume in the JSON format.

The following table describes the parameters of a source.

Parameter	Required	Description
type	Yes	Default value: `com.aliyun.loghub.flume.source.LoghubSource`.
endpoint	Yes	The endpoint of the region where the Log Service project resides.
project	Yes	The name of the project.
logstore	Yes	The name of the Logstore.
accessKeyId	Yes	The AccessKey ID that is used to access Log Service. For more information, see Obtain an AccessKey pair.
accessKey	Yes	The AccessKey secret that is used to access Log Service. For more information, see Obtain an AccessKey pair.
heartbeatIntervalMs	No	The heartbeat interval between the client and Log Service. Default value: 30000. Unit: milliseconds.
fetchIntervalMs	No	The interval at which data is pulled from Log Service. Default value: 100. Unit: milliseconds.
fetchInOrder	No	Specifies whether to consume log data in the order that the log data is written to Log Service. Default value: false.
batchSize	No	The number of log entries that are read at a time. Default value: 100.
consumerGroup	No	The name of the consumer group that reads log data.
initialPosition	No	The start point from which data is read. Valid values: **begin**, **end**, and **timestamp**. Default value: **begin**. **Note** If a checkpoint exists on the server, the checkpoint is preferentially used.
timestamp	No	The UNIX timestamp. If you set the initialPosition parameter to **timestamp**, you must specify this parameter.
deserializer	Yes	The deserialization format of the Flume event. Default value: DELIMITED. Valid values: • **DELIMITED**: delimiter mode. • **JSON**: JSON format. • Custom **deserializer**: custom deserialization mode. In this mode, you must specify the full names of the columns.
columns	No	The names of colums. If you set the deserializer parameter to **DELIMITED**, you must specify this parameter. Separate multiple columns with commas (,). The columns are sorted in the same order that the columns are sorted in the log entries.

Parameter	Required	Description
separatorChar	No	The delimiter. If you set the deserializer parameter to **DELIMITED**, you must specify a single character for this parameter. The default value is a comma (,).
quoteChar	No	The quote character. If you set the deserializer parameter to **DELIMITED**, you must specify this parameter. The default value is double quotation marks (").
escapeChar	No	The escape character. If you set the deserializer parameter to **DELIMITED**, you must specify this parameter. The default value is double quotation marks (").
appendTimestamp	No	Specifies whether to append the timestamp as a field to the end of each log entry. If you set the deserializer parameter to **DELIMITED**, you must specify this parameter. Default value: false.
sourceAsField	No	Specifies whether to add the log source as a field named __source__. If you set the deserializer parameter to **JSON**, you must specify this parameter. Default value: false.
tagAsField	No	Specifies whether to add the log tag as a field named __tag__:{name of the tag}. If you set the deserializer parameter to **JSON**, you must specify this parameter. Default value: false.
timeAsField	No	Specifies whether to add the log time as a field named __time__. If you set the deserializer parameter to **JSON**, you must specify this parameter. Default value: false.
useRecordTime	No	Specifies whether to use the value of the timestamp field in the log entries as the log time when log data is read from Log Service. Default value: false. This value indicates that the current time is used as the log time.

For more information about how to configure a source, visit GitHub.

23.1.6.6. Use open source Flink to consume log data

Log Service provides the flink-log-connector agent to connect with Flink. This topic describes how to connect Log Service with Flink to consume log data.

Prerequisites

- An AccessKey pair is obtained. For more information, see Obtain an AccessKey pair.
- A project and a Logstore are created. For more information, see Create a project and Create a Logstore.
- A RAM user is authorized to access the Logstore from which you want to consume data. For more information, see Grant a RAM user the permissions to consume data from a specified Logstore.

Context

The flink-log-connector agent consists of the flink-log-consumer and flink-log-producer agents. The two agents have the following differences:

- The flink-log-consumer agent reads data from Log Service. This agent supports the exactly-once semantics and load balancing among shards.
- The flink-log-producer agent writes data to Log Service.

Before you can use the flink-log-producer agent to write data to Log Service, you must add Maven dependencies.

The following example shows sample Maven dependencies:

```
<dependency>
   <groupId>com.aliyun.openservices</groupId>
   <artifactId>flink-log-connector</artifactId>
   <version>0.1.13</version>
</dependency>
<dependency>
   <groupId>com.google.protobuf</groupId>
   <artifactId>protobuf-java</artifactId>
   <version>2.5.0</version>
</dependency>
```

For more information, visit GitHub.

Flink Log Consumer

The flink-log-consumer agent can consume log data from a Logstore based on the exactly-once semantics. The flink-log-consumer agent detects the change in the number of shards in a Logstore.

Each Flink subtask consumes data from some shards in a Logstore. If the shards in a Logstore are split or merged, the shards from which the subtask consumes data also change.

If you use the flink-log-consumer agent to consume data from Log Service, you can call the following API operations:

- GetCursorOrData

  You can call this operation to pull log data from a shard. If you frequently call this operation, the amount of data that is transferred may exceed the capacity of shards. You can use the ConfigConstants.LOG_FETCH_DATA_INTERVAL_MILLIS parameter to specify the interval of API calls. You can use the ConfigConstants.LOG_MAX_NUMBER_PER_FETCH parameter to specify the number of log entries pulled by each call.

  Example:

  ```
  configProps.put(ConfigConstants.LOG_FETCH_DATA_INTERVAL_MILLIS, "100");
  configProps.put(ConfigConstants.LOG_MAX_NUMBER_PER_FETCH, "100");
  ```

- ListShards

  You can call this operation to view all shards in a Logstore and the status of each shard. If the shards are frequently split and merged, you can adjust the call interval to detect the changes in the number of shards at the earliest opportunity. Example:

  ```
  // Call the ListShards operation every 30,000 milliseconds.
  configProps.put(ConfigConstants.LOG_SHARDS_DISCOVERY_INTERVAL_MILLIS, "30000");
  ```

- CreateConsumerGroup

  You can call this operation to create a consumer group that is used to synchronize checkpoints.

- ConsumerGroupUpdateCheckPoint

  You can call this operation to synchronize snapshots of Flink to a consumer group.

1. Set startup parameters.

   The following example shows how to consume log data. The java.util.Properties class is used as a configuration tool. All constants must be configured in the ConfigConstants class.

```
Properties configProps = new Properties();
// Specify the endpoint of Log Service.
configProps.put(ConfigConstants.LOG_ENDPOINT, "cn-hangzhou.log.aliyuncs.com");
// Specify the AccessKey ID and AccessKey secret.
configProps.put(ConfigConstants.LOG_ACCESSKEYID, "");
configProps.put(ConfigConstants.LOG_ACCESSKEY, "");
// Specify the project.
configProps.put(ConfigConstants.LOG_PROJECT, "ali-cn-hangzhou-sls-admin");
// Specify the Logstore.
configProps.put(ConfigConstants.LOG_LOGSTORE, "sls_consumergroup_log");
// Specify the start position of log consumption.
configProps.put(ConfigConstants.LOG_CONSUMER_BEGIN_POSITION, Consts.LOG_END_CURSOR);
// Specify the data deserialization method.
RawLogGroupListDeserializer deserializer = new RawLogGroupListDeserializer();
final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
DataStream<RawLogGroupList> logTestStream = env.addSource(
        new FlinkLogConsumer<RawLogGroupList>(deserializer, configProps));
```

Note The number of Flink subtasks is independent of the number of shards in a Logstore. If the number of shards is greater than the number of subtasks, each subtask consumes one or more shards. If the number of shards is less than the number of subtasks, some subtasks remains idle until new shards are generated. The data of each shard is consumed by only one subtask.

2. Specify the start position of log consumption.

   If you use the flink-log-consumer agent to consume data from a Logstore, you can use the ConfigConstants.LOG_CONSUMER_BEGIN_POSITION parameter to specify the start position of log consumption. You can start to consume data from the earliest entry, the latest entry, or from a specific point in time. The flink-log-consumer agent also allows you to resume consumption from a specific consumer group. You can set the parameter to one of the following values:

   - Consts.LOG_BEGIN_CURSOR: starts to consume data from the earliest entry.
   - Consts.LOG_END_CURSOR: starts to consume data from the latest entry.
   - Consts.LOG_FROM_CHECKPOINT: starts to consume data from a checkpoint that is stored in a specified consumer group. You can use the ConfigConstants.LOG_CONSUMERGROUP parameter to specify the consumer group.
   - UnixTimestamp: a string of the integer data type. The timestamp follows the UNIX time format. It is the number of seconds that have elapsed since 00:00:00 Thursday, January 1, 1970. The value indicates that the data in a shard is consumed from this point in time.

   Example:

   ```
   configProps.put(ConfigConstants.LOG_CONSUMER_BEGIN_POSITION, Consts.LOG_BEGIN_CURSOR);
   configProps.put(ConfigConstants.LOG_CONSUMER_BEGIN_POSITION, Consts.LOG_END_CURSOR);
   configProps.put(ConfigConstants.LOG_CONSUMER_BEGIN_POSITION, "1512439000");
   configProps.put(ConfigConstants.LOG_CONSUMER_BEGIN_POSITION, Consts.LOG_FROM_CHECKPOINT);
   ```

   Note If you configure to resume consumption from a state backend of Flink when you start a Flink job, the flink-log-connector agent uses checkpoints that are stored in the state backend.

3. (Optional)Configure consumption progress monitoring.

   The flink-log-connector agent allows you to monitor the consumption progress. You can use the monitoring feature to obtain the consumption position of each shard in real time. The consumption position is indicated by a timestamp. For more information, see View the status of a consumer group.

   Example:

```
configProps.put(ConfigConstants.LOG_CONSUMERGROUP, "your consumer group name");
```

> **Note** This setting is optional. If you configure consumption progress monitoring and no consumer group exists, the flink-log-connector agent creates a consumer group. If a consumer group is available, the agent synchronizes snapshots to the consumer group. You can view the consumption progress of the agent in the Log Service console.

4. Configure consumption resumption and the exactly-once semantics.

   If the checkpointing feature of Flink is enabled, the flink-log-consumer agent periodically saves the consumption progress of each shard. If a subtask fails, Flink restores the subtask and starts to consume data from the latest checkpoint.

   When you specify the checkpoint period, you can specify the maximum amount of data that can be re-consumed if a subtask fails. You can use the following code to specify the checkpoint period:

   ```
   final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
   // Configure the exactly-once semantics.
   env.getCheckpointConfig().setCheckpointingMode(CheckpointingMode.EXACTLY_ONCE);
   // Save checkpoints every 5 seconds.
   env.enableCheckpointing(5000);
   ```

   For more information, see Checkpoints.

Flink Log Producer

The flink-log-producer agent writes data to Log Service.

> **Note** The flink-log-producer agent supports only the Flink at-least-once semantics. If a subtask fails, duplicate data may be written to Log Service. However, no data is lost.

If you use the flink-log-producer agent to write data to Log Service, you can call the following API operations:

- PostLogStoreLogs
- ListShards

1. Initialize the flink-log-producer agent.
   i. Set the initialization parameters of the flink-log-producer agent.

   The flink-log-producer agent is initialized the same way as the flink-log-consumer agent. The following example shows how to configure the initialization parameters of the flink-log-producer agent. In most cases, you can use the default values of the parameters. Example:

   ```
   // The number of I/O threads that are used to send data. Default value: 8.
   ConfigConstants.LOG_SENDER_IO_THREAD_COUNT
   // The time that is required to send cached logs. Default value: 3000.
   ConfigConstants.LOG_PACKAGE_TIMEOUT_MILLIS
   // The number of logs in the cached packet. Default value: 4096.
   ConfigConstants.LOG_LOGS_COUNT_PER_PACKAGE
   // The size of the cached packet. Default value: 3. Unit: MB.
   ConfigConstants.LOG_LOGS_BYTES_PER_PACKAGE
   // The total size of memory that can be used by the job. Default value: 100. Unit: MB.
   ConfigConstants.LOG_MEM_POOL_BYTES
   ```

ii. Reload LogSerializationSchema and define the method that is used to serialize data into raw log groups.

A raw log group is a collection of log entries.

If you want to write data to a specific shard, you can use the LogPartitioner parameter to generate hash keys for log data. LogPartitioner is an optional parameter. If you do not specify this parameter, data is written to a random shard.

Example:

```
FlinkLogProducer<String> logProducer = new FlinkLogProducer<String>(new SimpleLogSerializer(),
configProps);
logProducer.setCustomPartitioner(new LogPartitioner<String>() {
      // Generate a 32-bit hash value.
      public String getHashKey(String element) {
          try {
              MessageDigest md = MessageDigest.getInstance("MD5");
              md.update(element.getBytes());
              String hash = new BigInteger(1, md.digest()).toString(16);
              while(hash.length() < 32) hash = "0" + hash;
              return hash;
          } catch (NoSuchAlgorithmException e) {
          }
          return  "00000000000000000000000000000000000000000000000000000000000000000";
      }
  });
```

2. Write simulated data to Log Service, as shown in the following example:

```
// Serialize data into the format of raw log groups.
class SimpleLogSerializer implements LogSerializationSchema<String> {
    public RawLogGroup serialize(String element) {
        RawLogGroup rlg = new RawLogGroup();
        RawLog rl = new RawLog();
        rl.setTime((int)(System.currentTimeMillis() / 1000));
        rl.addContent("message", element);
        rlg.addLog(rl);
        return rlg;
    }
}
public class ProducerSample {
    public static String sEndpoint = "cn-hangzhou.log.aliyuncs.com";
    public static String sAccessKeyId = "";
    public static String sAccessKey = "";
    public static String sProject = "ali-cn-hangzhou-sls-admin";
    public static String sLogstore = "test-flink-producer";
    private static final Logger LOG = LoggerFactory.getLogger(ConsumerSample.class);
    public static void main(String[] args) throws Exception {
        final ParameterTool params = ParameterTool.fromArgs(args);
        final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment(
);
        env.getConfig().setGlobalJobParameters(params);
        env.setParallelism(3);
        DataStream<String> simpleStringStream = env.addSource(new EventsGenerator());
        Properties configProps = new Properties();
        // Specify the endpoint of Log Service.
        configProps.put(ConfigConstants.LOG_ENDPOINT, sEndpoint);
        // Specify the AccessKey ID and AccessKey secret.
        configProps.put(ConfigConstants.LOG_ACCESSKEYID, sAccessKeyId);
        configProps.put(ConfigConstants.LOG_ACCESSKEY, sAccessKey);
        // Specify the project to which logs are written.
        configProps.put(ConfigConstants.LOG_PROJECT, sProject);
        // Specify the Logstore to which logs are written.
        configProps.put(ConfigConstants.LOG_LOGSTORE, sLogstore);
        FlinkLogProducer<String> logProducer = new FlinkLogProducer<String>(new SimpleLogSerialize
r(), configProps);
        simpleStringStream.addSink(logProducer);
        env.execute("flink log producer");
    }
    // Simulate log generation.
    public static class EventsGenerator implements SourceFunction<String> {
        private boolean running = true;
        @Override
        public void run(SourceContext<String> ctx) throws Exception {
            long seq = 0;
            while (running) {
                Thread.sleep(10);
                ctx.collect((seq++) + "-" + RandomStringUtils.randomAlphabetic(12));
            }
        }
        @Override
        public void cancel() {
            running = false;
        }
    }
}
```

23.1.6.7. Use Logstash to consume log data

Log Service provides Logstash that you can use to consume log data. You can configure the Logstash input plug-in to read log data from Log Service, and then write the data to other systems, such as Kafka and Hadoop Distributed File System (HDFS).

Features

- Distributed collaborative consumption: You can configure multiple servers to consume log data from a Logstore at the same time.
- High performance: If you use a Java consumer group, the consumption speed of a single-core CPU can reach 20 MB/s.
- High reliability: Log Service saves consumption checkpoints. This mechanism resumes log consumption from the last checkpoint after a consumption exception is resolved.
- Automatic load balancing: Shards are automatically allocated based on the number of consumers in a consumer group. If you add a consumer to a consumer group or remove a consumer from the consumer group, shards are automatically reallocated.

Procedure

1. Install Logstash.
   i. Download the Logstash installation package.
   ii. Decompress the package that you downloaded to the specified directory.
2. Install the Logstash input plug-in.
   i. Download the input plug-in logstash-input-sls.
   ii. Install the Logstash input plug-in.

   ```
   logstash-plugin install logstash-input-sls
   ```

   Note For information about the causes of installation failures and solutions, see Plug-in installation and configuration.
3. Start Logstash.

   ```
   logstash -f logstash.conf
   ```

   The following table describes the parameters of the Logstash input plug-in.

Parameter	Type	Required	Description
endpoint	String	Yes	The endpoint of the region where the Log Service project resides.
access_id	String	Yes	The AccessKey ID of an Apsara Stack tenant account or RAM user that is used to access the project. The Apsara Stack tenant account or RAM user must have the permissions to consume log data by using consumer groups. For more information, see Permission to consume data of a specified Logstore.
access_key	String	Yes	The AccessKey secret of an Apsara Stack tenant account or RAM user that is used to access the project. The Apsara Stack tenant account or RAM user must have the permissions to consume log data by using consumer groups. For more information, see Permission to consume data of a specified Logstore.

Parameter	Type	Required	Description
project	String	Yes	The name of the Log Service project.
logstore	String	Yes	The name of the Log Service Logstore.
consumer_group	String	Yes	The name of the consumer group.
consumer_name	String	Yes	The name of the consumer. The name of each consumer in a consumer group must be unique.
position	String	Yes	The position where data consumption starts. Valid values: ◦ **begin**: Data is consumed from the first log entry that is written to the Logstore. ◦ **end**: Data is consumed from the current point in time. ◦ **yyyy-MM-dd HH:mm:ss**: Data is consumed from the specified point in time.
checkpoint_second	Number	No	The interval at which checkpoints are recorded. We recommend that you set the interval to a value between 10 and 60. Minimum value: 10. Default value: 30. Unit: seconds.
include_meta	Boolean	No	Specifies whether input log data contains metadata, such as the log source, time, tags, and topic fields. Default value: true. ◦ true: The log data contains metadata. ◦ false: The log data does not contain metadata.
consumer_name_with_ip	Boolean	No	Specifies whether to include an IP address in a consumer name. Default value: true. You must set this parameter to true if you want to apply distributed collaborative consumption. ◦ true: The name of the consumer contains an IP address. ◦ false: The name of the consumer does not contain an IP address.

Example

The following script shows how to configure Logstash to consume log data from a Logstore, and then print the data in stdout logs:

```
input {
  logservice{
  endpoint => "your project endpoint"
  access_id => "your access id"
  access_key => "your access key"
  project => "your project name"
  logstore => "your logstore name"
  consumer_group => "consumer group name"
  consumer_name => "consumer name"
  position => "end"
  checkpoint_second => 30
  include_meta => true
  consumer_name_with_ip => true
  }
}
output {
  stdout {}
}
```

23.1.6.8. Use Spark Streaming to consume log data

After Log Service collects log data, you can use Spark Streaming to consume the data.

The Spark SDK provided by Alibaba Cloud allows you to consume log data from Log Service in Receiver or Direct mode. You must add the following Maven dependency:

```
<dependency>
  <groupId>com.aliyun.emr</groupId>
  <artifactId>emr-logservice_2.11</artifactId>
  <version>1.7.2</version>
</dependency>
```

Consume log data in Receiver mode

In Receiver mode, a consumer group consumes data from Log Service and temporarily stores the data in a Spark executor. After a Spark Streaming job is started, the consumer group reads and processes data from the Spark executor. Each log entry is returned as a JSON string. The consumer group periodically saves checkpoints to Log Service. You do not need to update checkpoints. For more information, see Use consumer groups to consume log data.

- Parameters

Parameter	Type	Description
project	String	The name of the Log Service project.
logstore	String	The name of the Log Service Logstore.
consumerGroup	String	The name of the consumer group.
endpoint	String	The endpoint of the region where the Log Service project resides.
accessKeyId	String	The AccessKey ID that is used to access Log Service.

Parameter	Type	Description
accessKeySecret	String	The AccessKey secret that is used to access Log Service.

- Example

> **Note** In Receiver mode, data loss may occur if the default configurations are used. To avoid data loss, you can enable the Write-Ahead Logs feature. This feature is available in Spark 1.2 or later. For more information, see Spark.

```
import org.apache.spark.storage.StorageLevel
import org.apache.spark.streaming.aliyun.logservice.LoghubUtils
import org.apache.spark.streaming.{Milliseconds, StreamingContext}
import org.apache.spark.SparkConf
object TestLoghub {
  def main(args: Array[String]): Unit = {
    if (args.length < 7) {
      System.err.println(
        """Usage: TestLoghub <project> <logstore> <loghub group name> <endpoint>
          |         <access key id> <access key secret> <batch interval seconds>
        """.stripMargin)
      System.exit(1)
    }
    val project = args(0)
    val logstore = args(1)
    val consumerGroup = args(2)
    val endpoint = args(3)
    val accessKeyId = args(4)
    val accessKeySecret = args(5)
    val batchInterval = Milliseconds(args(6).toInt * 1000)
    def functionToCreateContext(): StreamingContext = {
      val conf = new SparkConf().setAppName("Test Loghub")
      val ssc = new StreamingContext(conf, batchInterval)
      val loghubStream = LoghubUtils.createStream(
        ssc,
        project,
        logstore,
        consumerGroup,
        endpoint,
        accessKeyId,
        accessKeySecret,
        StorageLevel.MEMORY_AND_DISK)
      loghubStream.checkpoint(batchInterval * 2).foreachRDD(rdd =>
        rdd.map(bytes => new String(bytes)).top(10).foreach(println)
      )
      ssc.checkpoint("hdfs:///tmp/spark/streaming") // set checkpoint directory
      ssc
    }
    val ssc = StreamingContext.getOrCreate("hdfs:///tmp/spark/streaming", functionToCreateContext _
)
    ssc.start()
    ssc.awaitTermination()
  }
}
```

Consume log data in Direct mode

In Direct mode, you can consume log data from Log Service without the need of consumer groups. You can call API operations to request data from Log Service. Consuming log data in Direct mode has the following benefits:

- Simplified concurrency. The number of Spark partitions is the same as the number of shards in a Logstore. You can split shards to improve the concurrency of tasks.
- Increased efficiency. You no longer need to enable the Write-Ahead Logs feature to prevent data loss.
- Exactly-once semantics. Data is directly read from Log Service. Checkpoints are submitted after a task is successful.

  In some cases, data may be repeatedly consumed if a task ends due to an unexpected exit of Spark.

If you want to consume data in Direct mode, you must configure the ZooKeeper service to store intermediate data. In addition, you must specify a checkpoint directory in the ZooKeeper service. Intermediate data is stored in the checkpoint directory. To re-consume data after you restart a task, you must delete the checkpoint directory from ZooKeeper and change the name of the consumer group.

- Parameters

Parameter	Type	Description
project	String	The name of the Log Service project.
logstore	String	The name of the Log Service Logstore.
consumerGroup	String	The name of the consumer group. This name is used only to save consumption checkpoints.
endpoint	String	The endpoint of the region where the Log Service project resides.
accessKeyId	String	The AccessKey ID that is used to access Log Service.
accessKeySecret	String	The AccessKey secret that is used to access Log Service.
zkAddress	String	The connection URL of the ZooKeeper service.

- Throttling configuration

  Spark Streaming consumes data from each shard in a single batch. You must specify the number of log entries that are consumed in each batch.

  In the underlying storage model of Log Service, a log group serves as the basic storage unit. Each log group corresponds to a write request. For example, a write request may contain multiple log entries. These log entries are stored and consumed as a log group. When you use web tracking to write logs, each write request contains only one log entry. In this case, the log group that corresponds to the request contains only one log entry. You can specify parameters to limit the amount of log data in a single batch. The following table describes the two parameters.

Parameter	Description	Default
spark.loghub.batchGet.step	The maximum number of log groups that are returned for a single consumption request.	100

Parameter	Description	Default
spark.streaming.loghub.maxRatePerShard	The maximum number of log entries that are consumed from each shard in a single batch.	10000

You can set the spark.streaming.loghub.maxRatePerShard parameter to specify the maximum number of log entries that are consumed from each shard in each batch. The Spark SDK obtains the number of log groups from the spark.loghub.batchGet.step parameter before it consumes log data from Log Service, and accumulates the number of log entries in these log groups during the consumption. When the accumulated number reaches or exceeds the specified number in the spark.streaming.loghub.maxRatePerShard parameter, the Spark SDK stops consuming log data. The spark.streaming.loghub.maxRatePerShard parameter does not precisely control the number of consumed log entries in each batch. The number of consumed log entries in each batch varies based on the value of the spark.loghub.batchGet.step parameter and the number of log entries in each log group.

- Example

```
import com.aliyun.openservices.loghub.client.config.LogHubCursorPosition
import org.apache.spark.SparkConf
import org.apache.spark.streaming.{Milliseconds, StreamingContext}
import org.apache.spark.streaming.aliyun.logservice.{CanCommitOffsets, LoghubUtils}
object TestDirectLoghub {
  def main(args: Array[String]): Unit = {
    if (args.length < 7) {
      System.err.println(
        """Usage: TestDirectLoghub <project> <logstore> <loghub group name> <endpoint>
          |         <access key id> <access key secret> <batch interval seconds> <zookeeper host:po
rt=localhost:2181>
        """.stripMargin)
      System.exit(1)
    }
    val project = args(0)
    val logstore = args(1)
    val consumerGroup = args(2)
    val endpoint = args(3)
    val accessKeyId = args(4)
    val accessKeySecret = args(5)
    val batchInterval = Milliseconds(args(6).toInt * 1000)
    val zkAddress = if (args.length >= 8) args(7) else "localhost:2181"
    def functionToCreateContext(): StreamingContext = {
      val conf = new SparkConf().setAppName("Test Direct Loghub")
      val ssc = new StreamingContext(conf, batchInterval)
      val zkParas = Map("zookeeper.connect" -> zkAddress,
        "enable.auto.commit" -> "false")
      val loghubStream = LoghubUtils.createDirectStream(
        ssc,
        project,
        logStore,
        consumerGroup,
        accessKeyId,
        accessKeySecret,
        endpoint,
        zkParas,
        LogHubCursorPosition.END_CURSOR)
      loghubStream.checkpoint(batchInterval).foreachRDD(rdd => {
        println(s"count by key: ${rdd.map(s => {
          s.sorted
          (s.length, s)
        }).countByKey().size}")
        loghubStream.asInstanceOf[CanCommitOffsets].commitAsync()
      })
      ssc.checkpoint("hdfs:///tmp/spark/streaming") // set checkpoint directory
      ssc
    }
    val ssc = StreamingContext.getOrCreate("hdfs:///tmp/spark/streaming", functionToCreateContext _
)
    ssc.start()
    ssc.awaitTermination()
  }
}
```

For more information, visit GitHub.

23.1.6.9. Use Realtime Compute to consume log data

You can use Realtime Compute (Blink) to create a schema for data in Log Service and consume the data. This topic describes how to use Realtime Compute to create a schema for data in Log Service. This topic also describes the attribute fields and data type mapping that you can configure when you create a schema.

Create a schema for data in Log Service

Log Service stores streaming data. Realtime Compute can use the streaming data as input data. The following example is a sample log entry:

```
__source__:  203.0.113.10
__tag__:__receive_time__:  1562125591
__topic__:  test-topic
a:  1234
b:  0
c:  hello
```

The following example is a DDL statement that is used to create a schema for data in Log Service:

```
create table sls_stream(
  a int,
  b int,
  c varchar
) with (
  type ='sls',
  endPoint ='your endpoint',
  accessId ='your AccessKey ID',
  accessKey ='your AccessKey Secret',
  startTime = '2017-07-05 00:00:00',
  project ='ali-cloud-streamtest',
  logStore ='stream-test',
  consumerGroup ='consumerGroupTest1'
);
```

The following table describes the parameters in the WITH clause.

Parameter	Required	Description
endPoint	Yes	The endpoint of Log Service. For more information, see **Obtain an endpoint** in *Log Service Developer Guide*.
accessId	Yes	The AccessKey ID that is used to access Log Service.
accessKey	Yes	The AccessKey secret that is used to access Log Service.
project	Yes	The name of the Log Service project.
logStore	Yes	The name of the Log Service Logstore.
consumerGroup	No	The name of the consumer group.
startTime	No	The point in time when Realtime Compute starts to consume log data.
heartBeatIntervalMills	No	The heartbeat interval of the client that consumes log data. Default value: 10. Unit: seconds.
maxRetryTimes	No	The maximum number of retries to read data. Default value: 5.

Parameter	Required	Description
batchGetSize	No	The number of log groups that you want to read at a time. Default value: 10. If the version of Blink is 1.4.2 or later, the default value is 100 and the maximum value is 1000. **Note** If the size of a single log entry and the number of log groups in a batch are large, the Java system may frequently recycle the data that is stored in the memory.
columnErrorDebug	No	Specifies whether to enable debugging. If you set the value to true, debugging is enabled and log entries that fail to be parsed are displayed. Default value: false. This value indicates that debugging is not enabled.

Attribute fields

Realtime Compute can extract fields from log data. Realtime Compute can also extract three attribute fields and custom tag fields. The following table describes the three attribute fields.

Attribute field	Description
`__source__`	The source of the log entry.
`__topic__`	The topic of the log entry.
`__timestamp__`	The point in time when the log entry is generated.

To extract the three attribute fields, you must add HEADERs in the DDL statement. Example:

```
create table sls_stream(
  __timestamp__ bigint HEADER,
  __receive_time__ bigint HEADER
  a int,
  b int,
  c varchar
) with (
  type ='sls',
  endPoint ='your endpoint',
  accessId ='your AccessKey ID',
  accessKey ='your AccessKey Secret',
  startTime = '2017-07-05 00:00:00',
  project ='ali-cloud-streamtest',
  logStore ='stream-test',
  consumerGroup ='consumerGroupTest1'
);
```

Data type mapping

The string data type in Log Service is mapped to the varchar data type in Realtime Compute. We recommend that you declare the mapping in a DDL statement. If you specify another data type to convert data in Log Service, Realtime Compute attempts to automatically convert the data. For example, you can specify bigint as the data type to convert the string `1000` and specify timestamp as the data type to convert the string `2018-01-12 12:00:00`.

Usage notes

- Blink 2.2.0 or earlier versions do not support shard scaling. If you split or merge shards when a job is reading data from a Logstore, the job fails and cannot continue. In this case, you must restart the job.
- You cannot delete or recreate a Logstore whose log data is being consumed, regardless of the Blink version.
- In Blink version 1.6.0 and earlier, the read performance may be affected if you specify a consumer group to consume log data from a Logstore that contains a large number of shards.
- You cannot define the map data type in Realtime Compute when you create a schema for data in Log Service.
- Fields that do not exist are set to null.
- Fields can be converted in a random order. However, we recommend that you convert the fields in the same order as the fields in the schema.
- If no new data is written to a shard, the latency of a job increases. In this case, you must change the number of concurrent tasks in the job to the number of shards in which data is read and written.
- To extract fields from tags such as `__tag__:__hostname__` and `__tag__:__path__`, you can delete the __tag__: prefix and follow the method used to extract attribute fields.

  Note You cannot extract this type of data during debugging. We recommend that you use the on-premises debugging method and the print method to display data in logs.

23.1.7. Data shipping

23.1.7.1. Ship logs to OSS

23.1.7.1.1. Overview

Log Service provides the data shipping feature. You can use this feature to ship logs to Object Storage Service (OSS) in real time by using the Log Service console. This topic describes the benefits and scenarios of the data shipping feature.

In the Log Service console, you can ship logs to other Apsara Stack services. Then, you can store or consume the log data by using other systems such as E-MapReduce. After you enable the log shipping feature, Log Service ships the collected logs to the specified cloud service at regular intervals.

Scenarios

The data shipping feature can be used to connect Log Service with data warehouses.

Benefits

The data shipping feature of Log Service has the following benefits:

- Ease of use

  You only need to complete a few settings in the Log Service console before you can ship logs from Logstores to other Apsara Stack services such as OSS.

- High efficiency

  Log Service stores logs that are collected from multiple servers. This improves efficiency when you ship log data to Apsara Stack services such as OSS.

- Effective management

  You can ship logs from different projects or Logstores to different OSS buckets. This way, you can efficiently manage the logs by log type or log source.

Log shipping destinations

For information about how to ship logs to OSS, see Ship log data from Log Service to OSS.

23.1.7.1.2. Ship log data from Log Service to OSS

You can use Log Service to collect log data and ship the log data to Object Storage Service (OSS) for storage and analysis. This topic describes how to ship log data from Log Service to OSS.

Prerequisites

- Log data is collected. For more information, see Log collection methods.
- OSS is activated. A bucket is created in the region where the Log Service project resides. For more information, see the **Create buckets** section in the *Service User Guide -Ojbect Storage Service(OSS)*.
- A Resource Access Management (RAM) role is created for the level-1 organization. For more information, see Obtain the ARN of a RAM role.

Context

Log Service can automatically ship log data from a Logstore to an OSS bucket.

- You can set a custom retention period for the log data in the OSS bucket. Permanent retention is supported.
- You can use data processing platforms such as E-MapReduce and Data Lake Analytics (DLA) or use custom programs to consume log data from the OSS bucket.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project from which you want to ship log data to OSS.
3. On the Logstores tab, click the > icon on the left of the specific Logstore and choose **Data Transformation > Export > Object Storage Service(OSS)**.
4. On the OSS Shipper page, click **Enable**.
5. In the **OSS LogShipper** pane, configure the shipping rules.

   The following table describes the required parameters.

Parameter	Description
OSS Shipper Name	The name of the shipping rule. The name can contain only lowercase letters, digits, hyphens (-), and underscores (_). The name must start and end with a lowercase letter or digit and must be 2 to 128 characters in length.
OSS Bucket	The name of the OSS bucket to which you want to ship log data. **Notice** You must specify the name of an existing OSS bucket. The specified OSS bucket must reside in the same region as the Log Service project.
OSS Prefix	The directory to which log data is shipped in the OSS bucket.
Partition format	The partition format of the bucket directory for the shipping task. The directory is automatically generated based on the time when the shipping task is created. The default format is %Y/%m/%d/%H/%M. The partition format cannot start with a forward slash (/). For information about partition format examples, see Partition format. For more information about parameters, see strptime API.

Parameter	Description
(Resource Access Management) RAM Role	The Alibaba Cloud Resource Name (ARN) of the RAM role. The RAM role is the identity that the OSS bucket owner creates for access control. Example: acs:ram::45643:role/aliyunlogdefaultrole. For information about how to obtain the ARN, see Obtain the ARN of a RAM role.
Shipping Size	The maximum size of raw log data that can be shipped to the OSS bucket in a shipping task. Valid values: 5 to 256. Unit: MB. If the size of shipped data exceeds the specified value, a new shipping task is automatically created.
Compress	Specifies whether to compress log data that is shipped to OSS. Valid values: ○ No Compress: The log data that is shipped to OSS is not compressed. ○ Compress (snappy): The snappy utility is used to compress the log data that is shipped to OSS. This way, the log data occupies less storage space of the OSS bucket.
Storage Format	The storage format of the log data that is shipped to OSS. Valid values: JSON, CSV, and Parquet. For more information, see Storage Formats.
Ship Tags	Specifies whether to ship log tags.
Shipping Time	The time period during which a shipping task runs. Valid values: 300 to 900. Default value: 300. Unit: seconds. If the specified time period expires, another shipping task is created.

6. Click **OK**.

> **Note**
> - After you configure a shipping rule, multiple shipping tasks can concurrently run. If the size of the data shipped from a shard reaches the specified threshold or the specified time period expires, another task is created.
> - After you create a shipping task, you can check whether the shipping rule satisfies your business requirements based on the task status and the data shipped to OSS.

View OSS data

After log data is shipped to OSS, you can access the log data in the OSS console, or by using the OSS API, an SDK, or another method. For more information, see the **Objects > Search for objects** section of the *Service User Guide -Ojbect Storage Service(OSS)*.

The following script shows a sample OSS directory:

```
oss://OSS-BUCKET/OSS-PREFIX/PARTITION-FORMAT_RANDOM-ID
```

OSS-BUCKET is the name of the OSS bucket. OSS-PREFIX is the prefix of the directory in the OSS bucket. PARTITION-FORMAT is the partition format of the directory for a shipping task. The partition format is calculated based on the time when the shipping task is created. For more information, see strptime API. RANDOM-ID is the unique identifier of the shipping task.

> **Note** The directory in the OSS bucket is created based on the time when the shipping task is created. For example, the shipping task is created at 00:00:00 on June 23, 2016 to ship data to OSS. The data is written to Log Service after 23:55:00 on June 22, 2016. The shipping interval is 5 minutes. To retrieve all logs shipped on June 22, 2016, you must check all objects in the *2016/06/22* directory. You must also check the *2016/06/23/00/* directory for the objects that are generated in the first 10 minutes of June 23, 2020.

Partition format

For each shipping task, log data is written to a directory of an OSS bucket. The directory is in the *oss://OSS-BUCKET/OSS-PREFIX/PARTITION-FORMAT_RANDOM-ID* format. A partition format is obtained by formatting the time when a shipping task is created. The following table describes the partition formats and directories that are obtained when a shipping task is created at 19:50:43 on January 20, 2017.

OSS Bucket	OSS Prefix	Partition format	OSS directory
test-bucket	test-table	%Y/%m/%d/%H/%M	*oss://test-bucket/test-table/2017/01/20/19/50_1484913043351525351_2850008*
test-bucket	log_ship_oss_example	year=%Y/mon=%m/day=%d/log_%H%M%s	*oss://test-bucket/log_ship_oss_example/year=2017/mon=01/day=20/log_195043_1484913043351525351_2850008.parquet*
test-bucket	log_ship_oss_example	ds=%Y%m%d/%H	*oss://test-bucket/log_ship_oss_example/ds=20170120/19_1484913043351525351_2850008.snappy*
test-bucket	log_ship_oss_example	%Y%m%d/	*oss://test-bucket/log_ship_oss_example/20170120/_1484913043351525351_2850008* **Note** This format may prevent platforms such as Hive from parsing the log data in the OSS bucket. We recommend that you do not use this format.
test-bucket	log_ship_oss_example	%Y%m%d%H	*oss://test-bucket/log_ship_oss_example/2017012019_1484913043351525351_2850008*

You can use Hive, MaxCompute, or Data Lake Analytics (DLA) to analyze OSS data. In this case, if you want to use partition information, you can set PARTITION-FORMAT in the key=value format. For example, you can set the partition format to *oss://test-bucket/log_ship_oss_example/year=2017/mon=01/day=20/log_195043_1484913043351525351_2850008.parquet*. In this example, year, mon, and day are specified as three partition keys.

What to do next

After shipping tasks are created based on a shipping rule, you can modify the shipping rule. You can also disable the data shipping feature, view the statuses and error messages of the tasks, and retry failed tasks on the **OSS Shipper** page of a Logstore.

- Modify the shipping rule.

  Click **Settings** to modify the shipping rule. For information about the parameters, see Procedure.

- Disable the data shipping feature.

  Click **Disable**. The data in the Logstore is no longer shipped to OSS.

- View the statuses and error messages of the tasks.

  You can view the log shipping tasks of the last two days and their statuses.

  - Statuses of a shipping task

Status	Equivalent
Succeeded	The shipping task has succeeded.
Running	The shipping task is running. Check whether the task succeeds later.
Failed	The shipping task has failed. If the task cannot be restarted due to external causes, troubleshoot the failure based on the error message and retry the task.

- Error messages

  If a shipping task fails, an error message is returned for the task.

Error message	Error cause	Solution
UnAuthorized	The error message returned because the AliyunLogDefaultRole role does not have the required permissions.	To fix the error, check the following configurations: ■ Check whether the AliyunLogDefaultRole role is created by the OSS bucket owner. ■ Check whether the specified ID of the Alibaba Cloud account in the permission policy is valid. ■ Check whether the AliyunLogDefaultRole role is granted the write permissions on the OSS bucket. ■ Check whether the ARN of the AliyunLogDefaultRole role that you entered in the RAM Role field is valid.
ConfigNotExist	The error message returned because the task does not exist.	Check whether the data shipping feature is disabled. If the feature is disabled, enable the feature, configure a shipping rule, and then retry the shipping task.
InvalidOssBucket	The error message returned because the specified OSS bucket does not exist.	To fix the error, check the following configurations: ■ Check whether the OSS bucket resides in the same region as the Log Service project. ■ Check whether the specified bucket name is valid.
InternalServerError	The error message returned because an internal error has occurred in Log Service.	Retry the failed shipping task.

- Retry a shipping task

  By default, if a shipping task fails, Log Service retries the task based on the retry policy. You can also manually retry the task. By default, Log Service retries all tasks of the last two days. The minimum interval between two consecutive retries is 15 minutes. If a task fails for the first time, Log Service retries the task 15 minutes later. If the task fails for the second time, Log Service retries the task 30 minutes later. If the task fails for the third time, Log Service retries the task 60 minutes later. A similar method is used for subsequent attempts.

  To immediately retry a failed task, you can click **Retry All Failed Tasks** or **Retry** on the right of the task. You can also use the Log Service API or an SDK to retry a task.

23.1.7.1.3. Obtain the ARN of a RAM role

When you use a RAM user to ship data from Log Service to Object Storage Service (OSS), you must first create a Resource Access Management (RAM) role and specify the ARN of the RAM role. This topic describes how to create a RAM role and obtain the ARN of a RAM role.

Procedure

1. Log on to the Log Service console.
2. In the top navigation bar, click **Configurations**.

3. On the **Service-Linked Roles** page, click **Create RAM Role**.
4. In the **Organization Name** drop-down list, select the organization that you created. In the **Service Name** drop-down list, select **Log Service**, and click **OK**.
5. On the **RAM Service Role** page, enter `AliyunLogDefaultRole` in the **Role Name** search box and click **Search**.
6. Obtain the ARN of the RAM role.

   In the search results, the value in the **role identifier** column is the ARN of the RAM role.

23.1.7.1.4. Storage Formats

Different storage formats are supported when Log Service ships logs to OSS, including JSON, CSV, and Parquet. This topic describes the field details of the formats.

JSON format

You can set the storage format for the data that is shipped to OSS. The following table shows how to set the **storage format** to **JSON**. For more information, see Configure a data shipping rule.

Compression type	File extension	Example file address	Description

Uncompressed	N/A	oss://oss-shipper-shenzhen/ecs_test/2016/01/26/20/54_1453812893059571256_937	You can download the raw JSON object to the local host and open each object as a text file. The following example is the content of a sample file:
			`{"__time__":1453809242,"__topic__":"","__source__":"10.170.***.***","ip":"10.200.***.***","time":"26/Jan/2016:19:54:02 +0800","url":"POST /PutData?Category=YunOsAccountOpLog&AccessKeyId=<yourAccessKeyId>&Date=Fri%2C%2028%20Jun%202013%2006%3A53%3A30%20GMT&Topic=raw&Signature=<yourSignature>`
			
			`HTTP/1.1","status":"200","user-agent":"aliyun-sdk-java"}`

snappy	.snappy	oss://oss-shipper-shenzhen/ecs_test/2016/01/26/20/54_1453812893059571256_937.snappy	JSON objects are compressed by using Snappy. For more information, see Decompression tools for Snappy-compressed files.

CSV-format

You can set the storage format for the data that is shipped to OSS. The following table shows how to set the **storage format** to **CSV**. For more information, see Configure a data shipping rule.

The following table describes the parameters. For more information, see Common Format and MIME Type for Comma-Separated Values (CSV) Files and PostgreSQL 9.4.26 Documentation.

Parameter	Description
CSV Fields	The names of the log fields that you want to ship to OSS. You can view log fields on the **Raw Logs** tab of a Logstore and enter the names of the fields that you want to ship to OSS in the Key Name column. The log fields that you can ship to OSS include the fields in the log content and the reserved fields such as __time__, __topic__, and __source__. **Note** The keys that you enter in the **CSV Fields** section must be unique.
Delimiter	You can use commas (,), vertical bars (|), spaces, or tabs to delimit fields.
Escape Character	If a field contains a delimiter, you must use an escape character to enclose the field. This ensures that the field is not delimited.
Invalid Fields	If a key that you specify in the **CSV Fields** section does not exist, enter the value of the key in the Invalid Fields field.
Shipped Fields	If you turn on the **Shipped Fields** switch, field names are written in a CSV file.

The following table lists the directories in OSS buckets that store the data shipped from Log Service.

Compression type	File extension	Example	Description
No	.csv	oss://oss-shipper-shenzhen/ecs_test/2016/01/26/20/54_1453812893059571256_937.csv	You can download the raw JSON object to the local host and open the object as a text file.
snappy	.snappy.csv	oss://oss-shipper-shenzhen/ecs_test/2016/01/26/20/54_1453812893059571256_937.snappy.csv	Decompression tools for Snappy compressed files For more information, see Decompression tools for Snappy-compressed files.

Parquet-format

You can set the storage format for the data that is shipped to OSS. The following figure shows how to set the **storage format** to **Parquet**. For more information, see Configure a data shipping rule.

The following table describes the related parameters.

Parameter	Description

Parameter	Description
Key Name	The name of the log field that you want to ship to OSS. You can view log fields on the **Raw Logs** tab of a Logstore. You can also enter the names of the fields that you want to ship to OSS in the Key Name column. When the fields are shipped to OSS, they are stored in the Parquet format in the order that the field names are entered. The names of the fields are the column names in OSS. The log fields that you can ship to OSS include the fields in the log content and the reserved fields such as __time__, _topic__, and __source__. The value of a field in the Parquet format is null in the following two scenarios: • The field does not exist in logs. • The value of the field fails to be converted from the string type to a non-string type, for example, double or Int64. **Note** The keys that you enter in the **Parquet Keys** field must be unique.
Type	The Parquet storage format supports six data types: string, Boolean, Int32, Int64, float, and double. Log fields are converted from the string type to a data type that the Parquet storage format supports. If the data type of a log field fails to be converted, the value of the log field is null.

The following table lists the directories in OSS buckets that store data shipped from Log Service.

Compression type	File extension	Example	Description
Uncompressed	.parquet	oss://oss-shipper-shenzhen/ecs_test/2016/01/26/20/54_1453812893059571256_937.parquet	You can download the OSS buckets to the local host and use the parquet-tools utility to open the objects. For more information about the parquet-tools utility, visit parquet-tools.
Snappy	.snappy.parquet	oss://oss-shipper-shenzhen/ecs_test/2016/01/26/20/54_1453812893059571256_937.snappy.parquet	You can download the OSS buckets to the local host and use the parquet-tools utility to open the objects. For more information about the parquet-tools utility, visit parquet-tools.

23.1.7.1.5. Decompress Snappy compressed files

When you ship data from Log Service to Object Storage Service (OSS), you can use Snappy to compress OSS objects. After the data is shipped to OSS, you can decompress OSS objects by using the C++ library, Java library, Python library, and decompression tool for Linux.

Use the C++ library to decompress OSS objects

Download the C++ library from the snappy page and use the Snappy.Uncompress method to decompress Snappy compressed OSS objects.

Use the Java library to decompress OSS objects

Download the Java library from the xerial snappy-java page and use the Snappy.Uncompress or Snappy.SnappyInputStream method to decompress Snappy compressed OSS objects. The SnappyFramedInputStream method is not supported.

> **Note** If you use Java Library 1.1.2.1, some Snappy compressed OSS objects may fail to be decompressed. For more information, see Bad handling of the MAGIC HEADER. To fix this issue, you can use Java Library 1.1.2.6 or later.

```
<dependency>
<groupId>org.xerial.snappy</groupId>
<artifactId>snappy-java</artifactId>
<version>1.0.4.1</version>
<type>jar</type>
<scope>compile</scope>
</dependency>
```

- Snappy.Uncompress

```
String fileName = "C:\\Downloads\\36_1474212963188600684_4451886.snappy";
RandomAccessFile randomFile = new RandomAccessFile(fileName, "r");
int fileLength = (int) randomFile.length();
randomFile.seek(0);
byte[] bytes = new byte[fileLength];
int byteread = randomFile.read(bytes);
System.out.println(fileLength);
System.out.println(byteread);
byte[] uncompressed = Snappy.uncompress(bytes);
String result = new String(uncompressed, "UTF-8");
System.out.println(result);
```

- Snappy.SnappyInputStream

```
String fileName = "C:\\Downloads\\36_1474212963188600684_4451886.snappy";
SnappyInputStream sis = new SnappyInputStream(new FileInputStream(fileName));
byte[] buffer = new byte[4096];
int len = 0;
while ((len = sis.read(buffer)) != -1) {
    System.out.println(new String(buffer, 0, len));
}
```

Use the Python Library to decompress OSS objects

1. Download and install the Python library.
2. Run the decompression script.

   The following example is a sample decompression script:

```
import snappy
compressed = open('/tmp/temp.snappy').read()
snappy.uncompress(compressed)
```

> **Note** The following two commands cannot be used to decompress Snappy compressed OSS objects. These commands can be used only in Hadoop mode (hadoop_stream_decompress) or streaming mode (stream_decompress).

```
python -m snappy -c uncompressed_file compressed_file.snappy
python -m snappy -d compressed_file.snappy uncompressed_file
```

Use decompression tools for Linux to decompress OSS buckets

Log Service allows you to decompress Snappy compressed files by using the decompression tool for Linux. Click snappy_tool to download the tool. Replace 03_1453457006548078722_44148.snappy and 03_1453457006548078722_44148 in the following code with the values specific to your environment and then run the following code:

```
./snappy_tool 03_1453457006548078722_44148.snappy 03_1453457006548078722_44148
compressed.size: 2217186
snappy::Uncompress return: 1
uncompressed.size: 25223660
```

23.1.8. RAM

23.1.8.1. Overview

Resource Access Management (RAM) is a resource access control service provided by Apsara Stack.

You can use RAM to manage users, including employees, systems, and applications. You can also use RAM to grant users permissions to access resources.

RAM provides the following features:

- RAM Role

  To authorize a cloud service in a level-1 organization to use other resources in the organization, you must create a RAM role. This role specifies the operations that the cloud service can perform on the resources.

  Only system administrators and level-1 organization administrators can create RAM roles.

- User group

  You can create multiple RAM users for an organization and grant the users different permissions on the same cloud resources in the organization.

  You can create RAM user groups to classify and authorize RAM users within your Apsara Stack tenant account. This simplifies the management of RAM users and their permissions.

  You can create RAM policies to grant permissions to different user groups.

23.1.8.2. Create a RAM role

To authorize a cloud service in a level-1 organization to use other resources in the organization, you must create a RAM role. This RAM role specifies the operations that the cloud service can perform on the resources.

Procedure

1. Log on to the Apsara Uni-manager console as an administrator.

   For more information, see Log on to the Log Service console.

2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **Roles**.

4. In the upper-right corner of the page, click **Create RAM Role**.
5. On the **Roles - Create RAM Role** page, set the **Role Name** and **Description** parameters.
6. Click **Create**.

23.1.8.3. Create a user

You can create a user as an administrator and then assign different roles to the user to meet different requirements for system access control.

Procedure

1.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **Users**.
4. On the Users page, click **Create**.
5. In the **Create User** dialog box, set the parameters.

Parameter	Description
Username	The username.
Display Name	The display name of the user.
Roles	The roles that you want to assign to the user.
Organization	The organization to which the user belongs.
Logon Policy	The logon policy that restricts the logon time and IP address of the user. If you do not specify this parameter, the default policy is attached to the created user.
Mobile Number	The mobile phone number of the user. If you want to send text messages about the resource requests and usage to the mobile phone number, make sure that the specified mobile phone number is valid.
Landline Number	Optional. The landline number of the user.
Email	The email address of the user. If you want to send emails about the resource requests and usage to the email address, make sure that the specified email address is valid.
DingTalk Key	Optional. The DingTalk key.
Notify User by Email	If you select the **Notify User by Email** check box, emails about the resource requests and usage are sent to the specified email address.
Notify User by DingTalk	If you select the **Notify User by DingTalk** check box, messages about the resource requests and usage are sent to the specified DingTalk user.

6. Click **OK**.

23.1.8.4. Create a RAM user group

This topic describes how to create a RAM user group in an organization and grant permissions to RAM users in the RAM user group.

Prerequisites

An organization is created.

Context

The relationships between RAM user groups and RAM users:

- A RAM user group can contain zero or more RAM users.
- A RAM user does not need to belong to a RAM user group.
- You can add a RAM user to multiple RAM user groups.

The relationships between RAM user groups and organizations:

- A RAM user group belongs to only one organization.
- You can create multiple RAM user groups in an organization.

The relationships between RAM user groups and RAM roles:

- Only one RAM role can be assigned to each RAM user group.
- A RAM role can be assigned to multiple RAM user groups.
- When a RAM role is assigned to a RAM user group, the permissions that the RAM role has are automatically granted to RAM users in the RAM user group.

The relationships between RAM user groups and resource sets:

- You can add zero or more user groups to a resource set.
- A user group can be added to multiple resource sets.

Procedure

1.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **User Groups**.
4. In the upper-right corner of the page, click **Create User Group**.
5. In the dialog box that appears, set the **User Group Name** and **Organization** parameters.
6. Click **OK**.

23.1.8.5. Add a RAM user to a RAM user group

This topic describes how to add a RAM user to a RAM user group.

Procedure

1.
2. In the top navigation bar, click **Enterprise**.
3.
4. Find the user group to which you want to add a RAM user, and click **Add User** in the **Actions** column.
5. In the dialog box that appears, select a RAM user from the left pane, and click the right arrow to move the RAM user to the right pane.
6. Click **OK**.

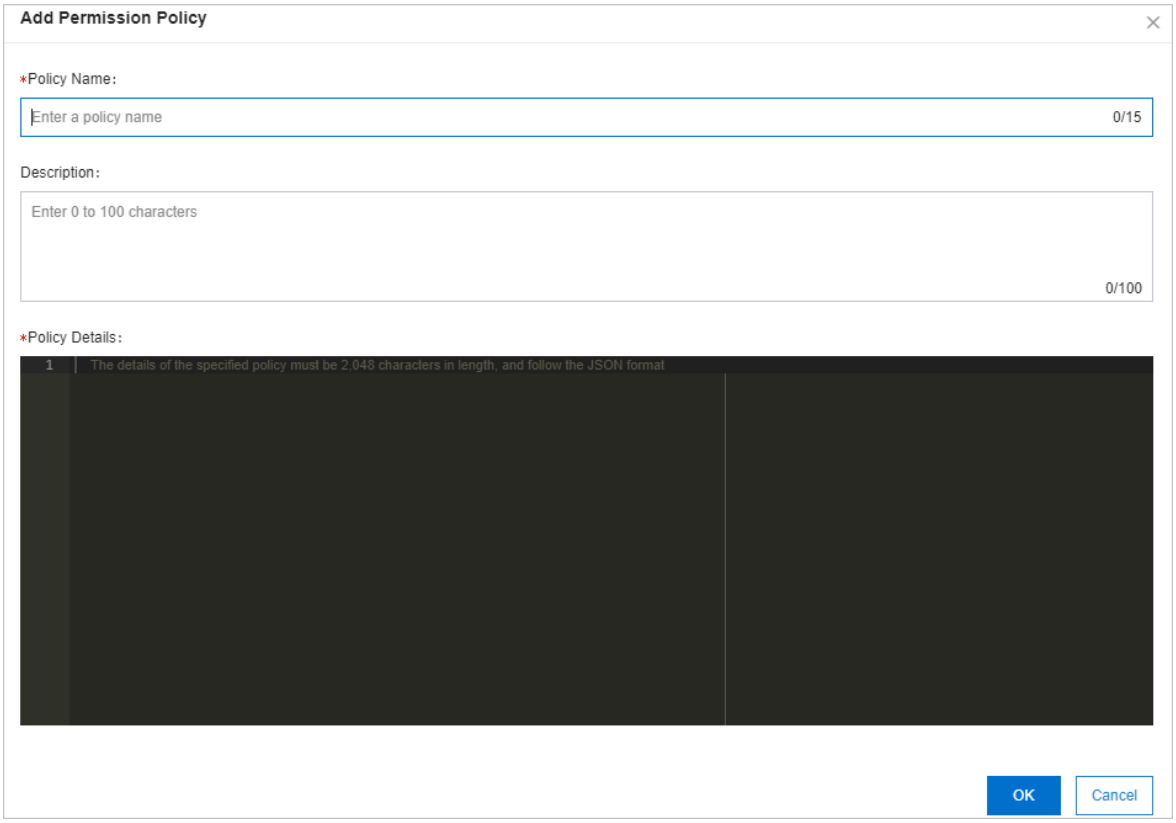
23.1.8.6. Create a permission policy

If you want to use a cloud service to access the resources of other cloud services, you must create a permission policy for a RAM role. Then, the policy is automatically attached to the RAM user group to which the RAM role is assigned.

Procedure

1.

2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **Roles**.
4. In the list of role names, find the RAM role for which you want to create a permission policy and choose **More > Modify**.
5. Click **Permissions**.
6. Click **Add Permission Policy**.
7. In the **Add Permission Policy** dialog box, enter the policy information.



For more information about how to specify the policy information, see Use custom policies to grant RAM user the required permissions.

23.1.8.7. Grant a RAM user the permissions to manage a project

This topic describes how to grant a Resource Access Management (RAM) user the permissions to manage a specified project.

Prerequisites

- A RAM user is created. For more information, see Create a user.
- A resource set is created and the RAM user is added to the resource set. For more information, see **Enterprise Center > Resource Sets** in *Apsara Uni-manager Management Console User Guide*.

Procedure

1.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **Data Permissions**.

4. Click the resource set that you want to manage, for example, ResourceSet(appstreaming).
5. In the **Product Type** section, click **Log Service**.
6. Find the instance that you want to manage and click **Authorize** in the Actions column.

   In this example, the instance is a Log Service project whose name is test-project.
7. Grant the RAM user the permissions to manage the project.

   If you turn on the Action switch of the **Update Project** permission, the RAM user can modify the project that is selected in Step .

23.1.8.8. Grant permissions to a RAM role

This topic describes how to grant permissions to a RAM role. After a RAM role is granted permissions, the RAM users in the associated RAM user group inherit the permissions.

Procedure

1.
2. In the top navigation bar, click **Enterprise**.
3. In the left-side navigation pane of the **Enterprise** page, click **Roles**.
4.
5. On the **Permissions** tab, click **Select Existing Permission Policy**.
6. In the **Select Existing Permission Policy** dialog box, select a permission policy and click **OK**.

   If no policies are available, create a policy. For more information, see Create a permission policy.

23.1.8.9. Use custom policies to grant permissions to a RAM user

This topic describes how to use custom policies to grant permissions to a RAM user. In the Resource Access Management (RAM) console, you can grant permissions to the RAM users that belong to your Apsara Stack tenant account.

Context

In terms of data security, we recommend that you follow the principle of least privilege (PoLP) when you grant permissions to RAM users. You must grant the read-only permissions on the project list to RAM users. Otherwise, the RAM users cannot view the projects in the project list.

Use the RAM console to grant permissions to a RAM user

- The read-only permissions on projects

  For example, you want to use your Apsara Stack tenant account to grant the following permissions to a RAM user:

  - The permissions to view the project list of the Apsara Stack tenant account
  - The read-only permissions on the projects that are specified by the Apsara Stack tenant account

  Use the following policy:

```
{
   "Version": "1",
   "Statement": [
     {
       "Action": ["log:ListProject"],
       "Resource": ["acs:log:*:*:project/*"],
       "Effect": "Alow"
      },
     {
       "Action": [
         "log:Get*",
         "log:List*"
       ],
       "Resource": "acs:log:*:*:project/Project name/*",
       "Effect": "Allow"
     }
   ]
 }
```

- The read-only permissions on a specified Logstore and the permissions to save a search and use the saved search

  For example, you want to use your Apsara Stack tenant account to grant the following permissions to a RAM user:

  - The permissions to view the project list of the Apsara Stack tenant account
  - The read-only permissions on a specified Logstore and the permissions to save a search and use the saved search

  Use the following policy.

  > **Note** If the content of the Resource element in a policy does not end with an asterisk (*), the RAM user can access only the specified resource of the current resource type. If the content of the Resource element ends with an asterisk (*), the RAM user can access all resources of the current resource type. Other resources are represented by an asterisk (*).

```
{
  "Version": "1",
  "Statement": [
    {
      "Action": [
        "log:ListProject"
      ],
      "Resource": "acs:log:*:*:project/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:List*"
      ],
      "Resource": "acs:log:*:*:project/Project name/logstore/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:Get*",
        "log:List*"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/logstore/Logstore name>"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:List*"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/dashboard",
        "acs:log:*:*:project/Project name/dashboard/*"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:Get*",
        "log:List*",
        "log:Create*"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/savedsearch",
        "acs:log:*:*:project/Project name/savedsearch/*"
      ],
      "Effect": "Allow"
    }
  ]
}
```

- The read-only permissions on a specified Logstore and the permissions to view all saved searches and dashboards in a project

  For example, you want to use your Apsara Stack tenant account to grant the following permissions to a RAM user:

  - The permissions to view the project list of the Apsara Stack tenant account

- The read-only permissions on a specified Logstore and the permissions to view all saved searches and dashboards in a project

  Use the following policy:

```
{
  "Version": "1",
  "Statement": [
    {
      "Action": [
        "log:ListProject"
      ],
      "Resource": "acs:log:*:*:project/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:List*"
      ],
      "Resource": "acs:log:*:*:project/Project name/logstore/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:Get*",
        "log:List*"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/logstore/Logstore name"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:Get*",
        "log:List*"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/dashboard",
        "acs:log:*:*:project/Project name/dashboard/*"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "log:Get*",
        "log:List*"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/savedsearch",
        "acs:log:*:*:project/Project name/savedsearch/*"
      ],
      "Effect": "Allow"
    }
  ]
}
```

Use API operations to grant permissions to a RAM user

- The permissions to write data to a specified project

To grant a RAM user only the permissions to write data to a specified project, use the following policy:

```
{
  "Version": "1",
  "Statement": [
    {
      "Action": [
        "log:Post*"
      ],
      "Resource": "acs:log:*:*:project/Project name/*",
      "Effect": "Allow"
    }
  ]
}
```

- The permissions to consume data from a specified project

  To grant a RAM user only the permissions to consume data from a specified project, use the following policy:

```
{
  "Version": "1",
  "Statement": [
    {
      "Action": [
        "log:ListShards",
        "log:GetCursorOrData",
        "log:GetConsumerGroupCheckPoint",
        "log:UpdateConsumerGroup",
        "log:ConsumerGroupHeartBeat",
        "log:ConsumerGroupUpdateCheckPoint",
        "log:ListConsumerGroup",
        "log:CreateConsumerGroup"
      ],
      "Resource": "acs:log:*:*:project/Project name/*",
      "Effect": "Allow"
    }
  ]
}
```

- The permissions to consume data from a specified Logstore

  To grant a RAM user only the permissions to consume data from a specified Logstore, use the following policy:

```
{
  "Version": "1",
  "Statement": [
    {
      "Action": [
        "log:ListShards",
        "log:GetCursorOrData",
        "log:GetConsumerGroupCheckPoint",
        "log:UpdateConsumerGroup",
        "log:ConsumerGroupHeartBeat",
        "log:ConsumerGroupUpdateCheckPoint",
        "log:ListConsumerGroup",
        "log:CreateConsumerGroup"
      ],
      "Resource": [
        "acs:log:*:*:project/Project name/logstore/Logstore name",
        "acs:log:*:*:project/Project name/logstore/Logstore name/*"
      ],
      "Effect": "Allow"
    }
  ]
}
```

23.1.9. FAQ

23.1.9.1. Log collection

23.1.9.1.1. How do I troubleshoot errors that occur when I use Logtail to collect logs?

If the preview page is blank or the No Data message appears on the query page after you create a Logtail configuration to collect logs, perform the steps that are described in the topic to troubleshoot the issue.

Procedure

1. Check whether Log Service receives heartbeats from the machine group.

   You can view the Logtail heartbeat status in the Log Service console. For more information, see View the status of a server group.

   If the heartbeat status is OK, perform the next step. If the heartbeat status is FAIL, identify the cause of the failure. For more information, see What can I do if no heartbeat packet is received from a Logtail client?.

2. Check whether the Logtail configuration is created.

   If the heartbeat status of Logtail is OK, check whether the Logtail configuration is created. Make sure that the path and name of monitored logs match those of the files stored on the server. The path can be a full path or a path that includes wildcards.

3. Make sure that the Logtail configuration is applied to the machine group.

   For more information, see Manage server group configurations.

4. Check collection errors.

   If the Logtail configuration is valid, check whether new logs are generated in real time. Logtail collects only incremental log data. Logtail does not read log files that are not updated. If a log file is updated but the updated data cannot be found in Log Service, you can use the following method to troubleshoot the issue:

   - View the logs of the Logtail client.

Client logs include key INFO logs, all WARNING logs, and all ERROR logs. To view complete and real-time error information, view the client logs in the following paths:

- Linux: */usr/local/ilogtail/ilogtail.LOG*.
- Linux: */usr/local/ilogtail/logtail_plugin.LOG*. The file contains the logs such as HTTP logs, MySQL binary logs, and MySQL query results.
- Windows x64 : *C:\Program Files (x86)\Alibaba\Logtail\logtail_*.log*
- Windows x32 : *C:\Program Files\Alibaba\Logtail\logtail_*.log*

- Check whether the amount of log data exceeds the limit.

  To collect large amounts of log data, you may need to modify the startup parameters of Logtail to increase the log collection throughput. For more information, see Set Logtail startup parameters.

23.1.9.1.2. What can I do if Log Service does not receive heartbeats from a Logtail client?

If Log Service does not receive heartbeats from a Logtail client, perform the steps that are described in the topic to troubleshoot the issue.

Context

After Logtail is installed on a server, the Logtail client sends heartbeats to Log Service. If the status page of the machine group shows that Log Service does not receive heartbeats from a Logtail client, the Logtail client is not installed or disconnected from the server.

Step 1: Check whether Logtail is installed

Use the following method to check whether Logtail is installed:

- On a Linux server, run the following command:

  ```
  sudo /etc/init.d/ilogtaild status
  ```

  If Logtail is installed, the following result appears:

  ```
  ilogtail is running
  ```

- On a Windows server, perform the following steps:
  i. On **Control Panel**, click **Administrative Tools**, and then click **Services**.
  ii. In the Services window, check the status of the LogtailDaemon and LogtailWorker services. If the services are in the Running state, Logtail is installed.

If Logtail is not installed, install Logtail. For more information, see Install Logtail in Linux or Install Logtail in Windows. Make sure that you install Logtail based on the region where your Log Service project resides. If Logtail is running, go to the next step.

Step 2: Check the Log Service endpoint in the Logtail installation command

When you install Logtail, you must specify a Log Service endpoint based on the region where your Log Service project resides. If the endpoint is incorrect or the Logtail installation command is invalid, Log Service cannot receive heartbeats from the Logtail client.

You can view the Log Service endpoint and the installation method in the Logtail configuration file named *ilogtail_config.json*. The file is stored in the following path:

- Linux: */usr/local/ilogtail/ilogtail_config.json*
- 64-bit Windows: *C:\Program Files (x86)\Alibaba\Logtail\ilogtail_config.json*
- 32-bit Windows: *C:\Program Files\Alibaba\Logtail\ilogtail_config.json*

In the *ilogtail_config.json* Logtail configuration file, check the endpoint that is specified for the `config_server_address` parameter. Then, check whether the Logtail client can use the endpoint to connect to Log Service. For example, if the endpoint that is recorded in the *ilogtail_config.json* Logtail configuration file is `logtail.cn-qingdao-env25-d01.sls-pub.inter.env25.shuguang.com`, you can run the following command to check the connection:

- Linux:

```
curl logtail.cn-qingdao-env25-d01.sls-pub.inter.env25.shuguang.com
```

- Windows:

```
telnet logtail.cn-qingdao-env25-d01.sls-pub.inter.env25.shuguang.com 80
```

If the Log Service endpoint in the Logtail installation command is incorrect, re-install Logtail. For more information, see Install Logtail in Linux or Install Logtail in Windows.

If the Log Service endpoint in the Logtail installation command is correct, go to the next step.

Step 3: Check the server IP addresses in the machine group

The server IP address that is obtained by a Logtail client must be configured in the machine group. Otherwise, Log Service cannot receive heartbeats or collect logs from the Logtail client. Logtail uses the following methods to obtain the IP address of a server:

- If the server is not bound with a hostname, Logtail obtains the IP address of the first network interface controller (NIC) card of the server.
- If the server is bound with a hostname, Logtail obtains the IP address that corresponds to the hostname. You can view the hostname and IP address in the */etc/hosts* file.

> **Note** You can run the **hostname** command to view the hostname.

Perform the following steps to check whether the server IP address that is obtained by the Logtail client is configured in the machine group.

1. Check the server IP address that is obtained by Logtail.

   The value of the `ip` field in the *app_info.json* file is the server IP address that is obtained by Logtail. The file is stored in the following path:

   - Linux: */usr/local/ilogtail/app_info.json*
   - 64-bit Windows: *C:\Program Files (x86)\Alibaba\Logtail\app_info.json*
   - 32-bit Windows: *C:\Program Files\Alibaba\Logtail\app_info.json*

   > **Note**
   > - If the `ip` field in the `app_info.json` file is empty, Logtail cannot work. In this case, you must configure an IP address for the server and restart Logtail.
   > - The `app_info.json` file is used only to record information. If you modify the IP address in the file, the server IP address that is obtained by Logtail is not updated.

2. Check the server IP addresses in the machine group.

   Log on to the Log Service console. In the Projects section, click the project to which the machine group belongs. In the left-side navigation pane, click **Machine Groups**, and then click the name of the machine group. In the Machine Group Status section of the **Machine Group Settings** page, check the server IP addresses.

   If no server IP address in the machine group is the same as the IP address that is obtained by Logtail, perform the following step to modify the IP address configurations in the Log Service console:

- If a server IP address in the machine group is incorrect, change the IP address to the IP address that is obtained by Logtail. Then, check the heartbeat status 1 minute after you save the change.
- If you modify the IP address of the server where Logtail is installed, for example, the */etc/hosts* file, restart Logtail. After Logtail obtains the new server IP address, set a server IP address in the machine group to the value of the `ip` field in the `app_info.json` file.

You can use the following method to restart Logtail:

- On a Linux server, run the following commands:

```
sudo /etc/init.d/ilogtaild stop
sudo /etc/init.d/ilogtaild start
```

- On a Windows server, perform the following steps:

  On **Control Panel**, choose **Administrative Tools > Services**. In the list that appears, find LogtailWorker and restart LogtailWorker.

23.1.9.1.3. How do I query the status of local log collection?

You can use the status query feature of Logtail to query the health status of Logtail and the log collection status. You can also use this feature to troubleshoot log collection issues and customize status monitoring for log collection.

Usage notes

After you install a Logtail client that supports the status query feature, you can query the local log collection status by running commands on the client. For more information about how to install Logtail, see Install Logtail in Linux.

You can run the `/etc/init.d/ilogtaild -h` command on a client to check whether the client supports the status query feature. If the result includes the `logtail insight, version` keyword, the client supports the status query feature.

```
/etc/init.d/ilogtaild -h
Usage: ./ilogtaild { start | stop (graceful, flush data and save checkpoints) | force-stop | status |
-h for help}$
logtail insight, version : 0.1.0
commond list :
       status all [index]
             get logtail running status
       status active [--logstore | --logfile] index [project] [logstore]
             list all active logstore | logfile. if use --logfile, please add project and logstore. de
fault --logstore
       status logstore [--format=line | json] index project logstore
             get logstore status with line or json style. default --format=line
       status logfile [--format=line | json] index project logstore fileFullPath
             get log file status with line or json style. default --format=line
       status history beginIndex endIndex  project logstore [fileFullPath]
             query logstore | logfile history status.
index :   from 1 to 60. in all, it means last $(index) minutes; in active/logstore/logfile/history, it
means last $(index)*10 minutes
```

Logtail supports multiple query commands. The following table describes the query commands, command functionalities, time ranges, and time windows for query results.

Command	Functionality	Maximum time range that can be queried	Time window

Command	Functionality	Maximum time range that can be queried	Time window
all	Queries the status of Logtail.	Last 60 minutes	1 minute
active	Queries the active Logstores that are collecting logs and the active log files from which logs are being collected.	Last 600 minutes	10 minutes
logstore	Queries the collection status of a Logstore.	Last 600 minutes	10 minutes
logfile	Queries the collection status of a log file.	Last 600 minutes	10 minutes
history	Queries the collection status of a Logstore or log file within a specified period of time.	Last 600 minutes	10 minutes

Note

- The `index` parameter in the preceding commands specifies the index of the time window. Valid values: 1 to 60. The index of the latest time window is 1. The time window ends at the current system time. If you specify a 1-minute time window, the status in the previous interval of `(index, index-1]` minutes is returned. If you specify a 10-minute time window, the status in the previous interval of `(10*index, 10*(index-1)]` minutes is returned.
- All commands in the preceding table are subcommands of the status command.

all command

- Syntax

```
/etc/init.d/ilogtaild status all [ index ]
```

Note The all command is used to query the status of Logtail. The index parameter is optional. Default value: 1.

- Examples

```
/etc/init.d/ilogtaild status all 1
ok
/etc/init.d/ilogtaild status all 10
busy
```

- Response

Status	Description	Priority	Troubleshooting
ok	Logtail runs as expected.	N/A	No action is required.

Status	Description	Priority	Troubleshooting
busy	The collection speed is high, and Logtail runs as expected.	N/A	No action is required.
many_log_files	A large number of log files are being collected by Logtail.	Low	You can check whether Logtail is configured to collect log files that do not need to be collected.
process_block	Log parsing is blocked.	Low	You can check whether a large number of logs are generated in a short period of time. If you use the all command multiple times and the returned value is always process_block, you can modify the limit of CPU utilization or the limit of concurrent packet sending. For more information, see Set Logtail startup parameters.
send_block	The process of packet sending is blocked.	High	You can check whether a large number of logs are generated in a short period of time and whether the network is stable. If you use the all command multiple times and the returned value is always send_block, you can modify the limit of CPU utilization or the limit of concurrent packet sending. For more information, see Set Logtail startup parameters.

active command

- Syntax

  Note The active command is used to query log files. We recommend that you query active Logstores before you query the active log files in the Logstores.

  ```
  /etc/init.d/ilogtaild status active [--logstore] index
  ```

  You can use the `active [--logstore] index` command to query all active Logstores. The `--logstore` parameter is optional.

  ```
  /etc/init.d/ilogtaild status active --logfile index project-name logstore-name
  ```

  You can use the `active --logfile index project-name logstore-name` command to query all active log files in the Logstore of a project.

- Examples

  ```
  /etc/init.d/ilogtaild status active 1
  sls-zc-test : release-test
  sls-zc-test : release-test-ant-rpc-3
  sls-zc-test : release-test-same-regex-3
  ```

  If you run the `active --logstore index` command, the names of the active Logstores are returned in the following format: `project-name : logstore-name`.

  ```
  /etc/init.d/ilogtaild status active --logfile 1 sls-zc-test release-test
  /disk2/test/normal/access.log
  ```

- If you run the `active --logfile index project-name logstore-name` command, the full paths of active log files are returned.
- The status of inactive Logstores or inactive log files in the query time window is not returned.

logstore command

- Syntax

```
/etc/init.d/ilogtaild status logstore [--format={line|json}] index project-name logstore-name
```

> **Note**
>
> - The logstore command is used to query the collection status of the specified project and Logstore in the `LINE` or `JSON` format.
> - The default value of the `--format=` parameter is `--format=line`. This value indicates that the status is returned in the LINE format.
> - If the Logstore specified in the preceding command does not exist or is not active in the query time window, an empty response in the `LINE` format or the `null` value in the `JSON` format is returned.

- Examples

```
/etc/init.d/ilogtaild status logstore 1 sls-zc-test release-test-same
time_begin_readable : 17-08-29 10:56:11
time_end_readable : 17-08-29 11:06:11
time_begin : 1503975371
time_end : 1503975971
project : sls-zc-test
logstore : release-test-same
status : ok
config : ##1.0##sls-zc-test$same
read_bytes : 65033430
parse_success_lines : 230615
parse_fail_lines : 0
last_read_time : 1503975970
read_count : 687
avg_delay_bytes : 0
max_unsend_time : 0
min_unsend_time : 0
max_send_success_time : 1503975968
send_queue_size : 0
send_network_error_count : 0
send_network_quota_count : 0
send_network_discard_count : 0
send_success_count : 302
send_block_flag : false
sender_valid_flag : true
/etc/init.d/ilogtaild status logstore --format=json 1 sls-zc-test release-test-same
{
   "avg_delay_bytes" : 0,
   "config" : "##1.0##sls-zc-test$same",
   "last_read_time" : 1503975970,
   "logstore" : "release-test-same",
   "max_send_success_time" : 1503975968,
   "max_unsend_time" : 0,
   "min_unsend_time" : 0,
   "parse_fail_lines" : 0,
   "parse_success_lines" : 230615,
   "project" : "sls-zc-test",
   "read_bytes" : 65033430,
   "read_count" : 687,
   "send_block_flag" : false,
   "send_network_discard_count" : 0,
   "send_network_error_count" : 0,
   "send_network_quota_count" : 0,
   "send_queue_size" : 0,
   "send_success_count" : 302,
   "sender_valid_flag" : true,
   "status" : "ok",
   "time_begin" : 1503975371,
   "time_begin_readable" : "17-08-29 10:56:11",
   "time_end" : 1503975971,
   "time_end_readable" : "17-08-29 11:06:11"
}
```

- Response

Parameter	Description	Unit

Parameter	Description	Unit
status	The status of the Logstore. For information about the different status of Logstore and the actions that are required to handle each status, see the following table.	N/A
time_begin_readable	The time when logs become readable.	N/A
time_end_readable	The time when logs become unreadable.	N/A
time_begin	The time when statistics collection starts.	UNIX timestamp in seconds
time_end	The time when statistics collection ends.	UNIX timestamp in seconds
project	The name of the project.	N/A
logstore	The name of the Logstore.	N/A
config	The name of the Logtail configuration. The name is globally unique. The format of the name is `##1.0##` + project + `$` + config.	N/A
read_bytes	The amount of log data that is read in the query time window.	Bytes
parse_success_lines	The number of log lines that are parsed in the query time window.	Lines
parse_fail_lines	The number of log lines that fail to be parsed in the query time window.	Lines
last_read_time	The time when logs are last read in the query time window.	UNIX timestamp in seconds
read_count	The number of times that the log file is read in the query time window.	N/A
avg_delay_bytes	The average of the difference between the actual file size and the offset value that is generated each time log data is read in the query time window.	Bytes
max_unsend_time	The maximum waiting period for an unsent packet in the sending queue. An unsent packet refers to a packet that is still in the sending queue at the end of the query time window. If no packets exist in the queue, the value is 0.	UNIX timestamp in seconds

Parameter	Description	Unit
min_unsend_time	The maximum waiting period for an unsent packet in the sending queue. An unsent packet refers to a packet that is still in the sending queue at the end of the query time window. If no packets exist in the queue, the value is 0.	UNIX timestamp in seconds
max_send_success_time	The maximum waiting period for an unsent packet in the sending queue.	UNIX timestamp in seconds
send_queue_size	The number of unsent packets in the sending queue at the end of the query time window.	N/A
send_network_error_count	The number of packets that cannot be sent due to network errors in the query time window.	N/A
send_network_quota_count	The number of packets that cannot be sent due to quota limit in the query time window.	N/A
send_network_discard_count	The number of packets that are discarded due to data errors or lack of permissions.	N/A
send_success_count	The number of packets that are sent in the query time window.	N/A
send_block_flag	Indicates whether the sending queue is blocked at the end of the query time window.	N/A
sender_valid_flag	Indicates whether the sender flag of the Logstore is valid. The value true indicates that the sender flag is valid. The value false indicates that the sender flag is invalid and disabled due to a network error or quota error.	N/A

Logstore status

Status	Description	Troubleshooting
ok	Logtail runs as expected.	No action is required.
process_block	Log parsing is blocked.	You can check whether a large number of logs are generated in a short period of time. If you use the all command multiple times and the returned value is always process_block, you can modify the limit of CPU utilization or the limit of concurrent packet sending. For more information, see Set Logtail startup parameters.
parse_fail	Logtail fails to parse logs.	You can check whether the format of logs is the same as the format that you specify in the Logtail configuration.

Status	Description	Troubleshooting
send_block	The process of packet sending is blocked.	You can check whether a large number of logs are generated in a short period time and whether the network is stable. If you use the all command multiple times and the returned value is always send_block, you can modify the limit of CPU utilization or the limit of concurrent packet sending. For more information, see Set Logtail startup parameters.

logfile command

- Syntax

```
/etc/init.d/ilogtaild status logfile [--format={line|json}] index project-name logstore-name fileFullPath
```

> **Note**
>
> - The logfile command is used to query the collection status of the specified log files in the `LINE` or `JSON` format.
> - The default value of the `--format=` parameter is `--format=line`. This value indicates that the status is returned in the LINE format.
> - If the log file specified in the command does not exist or is not active in the query time window, an empty response in the `LINE` format or the `null` value in the `JSON` format is returned.
> - The `--format` parameter is placed after the `logfile` parameter.
> - The value of the `filefullpath` parameter must be set to the full path of the log file.

- Examples

```
/etc/init.d/ilogtaild status logfile 1 sls-zc-test release-test-same /disk2/test/normal/access.log
time_begin_readable : 17-08-29 11:16:11
time_end_readable : 17-08-29 11:26:11
time_begin : 1503976571
time_end : 1503977171
project : sls-zc-test
logstore : release-test-same
status : ok
config : ##1.0##sls-zc-test$same
file_path : /disk2/test/normal/access.log
file_dev : 64800
file_inode : 22544456
file_size_bytes : 17154060
file_offset_bytes : 17154060
read_bytes : 65033430
parse_success_lines : 230615
parse_fail_lines : 0
last_read_time : 1503977170
read_count : 667
avg_delay_bytes : 0
/etc/init.d/ilogtaild status logfile --format=json 1 sls-zc-test release-test-same /disk2/test/norm
al/access.log
{
   "avg_delay_bytes" : 0,
   "config" : "##1.0##sls-zc-test$same",
   "file_dev" : 64800,
   "file_inode" : 22544456,
   "file_path" : "/disk2/test/normal/access.log",
   "file_size_bytes" : 17154060,
   "last_read_time" : 1503977170,
   "logstore" : "release-test-same",
   "parse_fail_lines" : 0,
   "parse_success_lines" : 230615,
   "project" : "sls-zc-test",
   "read_bytes" : 65033430,
   "read_count" : 667,
   "read_offset_bytes" : 17154060,
   "status" : "ok",
   "time_begin" : 1503976571,
   "time_begin_readable" : "17-08-29 11:16:11",
   "time_end" : 1503977171,
   "time_end_readable" : "17-08-29 11:26:11"
}
```

- Response

Parameter	Description	Unit
status	The collection status of the log file in the query time window. For more information, see the status parameter in the logstore command section.	N/A
time_begin_readable	The time when logs become readable.	N/A
time_end_readable	The time when logs become unreadable.	N/A

Parameter	Description	Unit
time_begin	The time when statistics collection starts.	UNIX timestamp in seconds
time_end	The time when statistics collection ends.	UNIX timestamp in seconds
project	The name of the project.	N/A
logstore	The name of the Logstore.	N/A
file_path	The path of the log file.	N/A
file_dev	The ID of the device from which the log file is collected.	N/A
file_inode	The inode of the log file.	N/A
file_size_bytes	The size of the last log file that is scanned in the query time window.	Bytes
read_offset_bytes	The parsing offset of the log file.	Bytes
config	The name of the Logtail configuration. The name is globally unique. The format of the name is `##1.0##` + project + `$` + config.	N/A
read_bytes	The amount of log data that is read in the query time window.	Bytes
parse_success_lines	The number of log lines that are parsed in the query time window.	Lines
parse_fail_lines	The number of log lines that fail to be parsed in the query time window.	Lines
last_read_time	The time when logs are last read in the query time window.	UNIX timestamp in seconds
read_count	The number of times that the log file is read in the query time window.	N/A
avg_delay_bytes	The average of the difference between the actual file size and the offset value that is generated each time log data is read in the query time window.	Bytes

history command

- Syntax

```
/etc/init.d/ilogtaild status history beginIndex endIndex project-name logstore-name [fileFullPath]
```

> **Note**
> - The history command is used to query the collection status of a Logstore or log file in the query time window.
> - The `beginIndex` and `endIndex` parameters specify the start and end indexes of the time windows that you want to query. You must ensure that beginIndex must be less than or equal to endIndex (`beginIndex <= endIndex`).
> - The `fileFullPath` parameter is optional. If you specify the path of a log file, the collection status of the log file is queried. Otherwise, the collection status of the Logstore is queried.

- Examples

  Query the collection status of a Logstore.

  - Command

    ```
    /etc/init.d/ilogtaild status history 1 3 sls-zc-test release-test-same /disk2/test/normal/access.
    log
    ```

  - Response

    ```
            begin_time          status      read  parse_success  parse_fail      last_read_time  read
    _count  avg_delay    device     inode  file_size  read_offset
     17-08-29 11:26:11              ok   62.12MB         231000           0   17-08-29 11:36:11
    671         0B     64800  22544459    18.22MB      18.22MB
     17-08-29 11:16:11              ok   62.02MB         230615           0   17-08-29 11:26:10
    667         0B     64800  22544456    16.36MB      16.36MB
     17-08-29 11:06:11              ok   62.12MB         231000           0   17-08-29 11:16:11
    687         0B     64800  22544452    14.46MB      14.46MB
    ```

  Query the collection status of a log file.

  - Command

    ```
    $/etc/init.d/ilogtaild status history 2 5 sls-zc-test release-test-same
    ```

  - Response

    ```
            begin_time          status      read  parse_success  parse_fail      last_read_time  read
    _count  avg_delay  send_queue  network_error  quota_error  discard_error  send_success  send_bloc
    k  send_valid          max_unsend          min_unsend    max_send_success
     17-08-29 11:16:11              ok   62.02MB         230615           0   17-08-29 11:26:10
    667         0B           0              0            0              0           300       false
    true   70-01-01 08:00:00   70-01-01 08:00:00   17-08-29 11:26:08
     17-08-29 11:06:11              ok   62.12MB         231000           0   17-08-29 11:16:11
    687         0B           0              0            0              0           303       false
    true   70-01-01 08:00:00   70-01-01 08:00:00   17-08-29 11:16:10
     17-08-29 10:56:11              ok   62.02MB         230615           0   17-08-29 11:06:10
    687         0B           0              0            0              0           302       false
    true   70-01-01 08:00:00   70-01-01 08:00:00   17-08-29 11:06:08
     17-08-29 10:46:11              ok   62.12MB         231000           0   17-08-29 10:56:11
    692         0B           0              0            0              0           302       false
    true   70-01-01 08:00:00   70-01-01 08:00:00   17-08-29 10:56:10
    ```

- Response
  - The collection status of the Logstore or log file in each query time window is listed in a line.
  - For more information about the response parameters, see the `logstore` command and `logfile` command sections.

Response status codes

- Success code

  If the parameters that you specify in a command is valid even if the queried Logstore or log file is not found, the code 0 is returned. Examples:

```
/etc/init.d/ilogtaild status logfile --format=json 1 error-project error-logstore /no/this/file
null
echo $?
0
/etc/init.d/ilogtaild status all
ok
echo $?
0
```

- Error codes

```
/etc/init.d/ilogtaild status nothiscmd
invalid param, use -h for help.
echo $?
10
/etc/init.d/ilogtaild status/all 99
invalid query interval
echo $?
1
```

  If a non-zero code is returned, an error occurs. The following table describes the possible non-zero codes.

Code	Description	Response	Troubleshooting
10	The command is invalid or the required parameters in the command are not specified.	`invalid param, use -h for help.`	You can run the `-h` command to obtain help information.
1	The value of the index parameter is not within the range of 1 to 60.	`invalid query interval`	You can run the `-h` command to obtain help information.
1	The collection status in the specified query time window cannot be queried.	`query fail, error: $(error)` . For more information, see errno.	The startup time of Logtail is earlier than the query time window. For more information, submit a ticket.
1	The start time of the query falls out of the query time window.	`no match time interval, please check logtail status`	You can check whether Logtail runs as expected. For more information, submit a ticket.
1	No data exists in the query time window that you specify.	`invalid profile, may be logtail restart`	You can check whether Logtail runs as expected. For more information, submit a ticket.

Scenarios

You can use the status query feature of Logtail to query the overall status of Logtail. You can also obtain specific metrics based on the collection status during log collection. You can customize a mechanism to monitor the log collection status based on the queried information.

Monitor the status of Logtail

You can monitor the status of Logtail by using the `all` command.

For example, you can run the command every minute to query the status of Logtail. If the `process_block`, `send_block`, or `send_error` value is returned for 5 consecutive minutes, an alert is triggered.

You can adjust the alert duration and monitoring scope based on the priorities of the collected log files.

Monitor the log collection status

You can monitor the log collection status of a Logstore by using the `logstore` command.

For example, you can run the `logstore` command every 10 minutes to query the status of the Logstore. If the value of the `avg_delay_bytes` parameter exceeds 1 MB (1024 × 1024 bytes) or the value of the `status` parameter is not `ok`, an alert is triggered.

You can adjust the alert threshold for the `avg_delay_bytes` metric based on the size of data that is generated during log collection.

Check whether Logtail has finished collecting log files

You can check whether Logtail has finished collecting log files by using the `logfile` command.

If Logtail no longer collects log files, you can run the `logfile` command every 10 minutes to query the status of the log file. If the value of the `read_offset_bytes` parameter is the same as the value of the `file_size_bytes` parameter, the log file is collected.

Troubleshoot log collection issues

If latency occurs on a server during log collection, you can use the `history` command to query the status history of log collection.

1. The value of the `send_block_flag` parameter is true. This indicates that log collection is blocked due to unstable network connections.
   - If the value of the `send_network_quota_count` parameter is greater than 0, split the shards in the Logstore. For more information, see Split a shard.
   - If the value of the `send_network_error_count` parameter is greater than 0, check the network connections.
   - If no network error occurs, adjust the limit of concurrent packet sending and the data transfer speed of Logtail. For more information, see Set Logtail startup parameters.
2. The parameters for packet sending are set to appropriate values. However, the value of the `avg_delay_bytes` parameter is large.
   - Use the value of the `read_bytes` parameter to calculate the average speed at which logs are parsed. You can determine whether a large amount of data is transferred during log collection based on the average speed.
   - Adjust the limits on resource usage for Logtail. For more information, see Set Logtail startup parameters.
3. The value of the `parse_fail_lines` parameter is greater than 0.

   Check whether the regular expression for log parsing can match all required log fields.

23.1.9.1.4. How do I debug a regular expression?

If you select the full regex mode when you configure Logtail to collect and parse text logs, you must specify a regular expression based on your sample log entries. This topic describes how to debug a regular expression.

Context

To debug the regular expression that you specified in the Log Service console, you can click **Validate** in the console and check the following results:

- If the regular expression is used to match the start part of the first line in a log entry, check whether the regular expression can match the expected number of log entries.
- If the fields are extracted by the regular expression, check whether the value of each field meets your requirements.

You can use online tools such as regex101.com and regextester.com to debug a regular expression. You can copy and paste the regular expression that is generated by Log Service to an online tool, and specify a sample log entry as the test string.

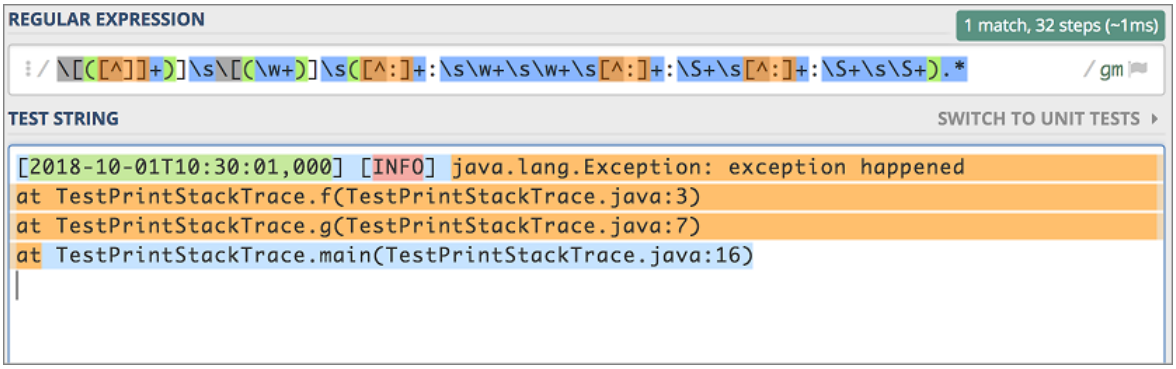
If you use the full regex mode, Log Service automatically generates a regular expression based on the sample log entry that you specify. However, the regular expression may fail to match the message field in multi-line log entries as expected. The following example shows how to use the regex101.com tool to debug a regular expression.

Procedure

1. Log on to the Log Service console.
2. In the Projects section, click the project that you want to manage.
3. In the Logstores list, find the Logstore that you want to manage and choose **Data Import > Logtail Configurations**.
4. Click the name of the Logtail configuration that you want to manage.
5. On the **Logtail Config** page, copy the regular expression that is automatically generated by Log Service based on sample log entries.
6. Visit the regex101.com website.
7. In the **REGULAR EXPRESSION** field, paste the regular expression.

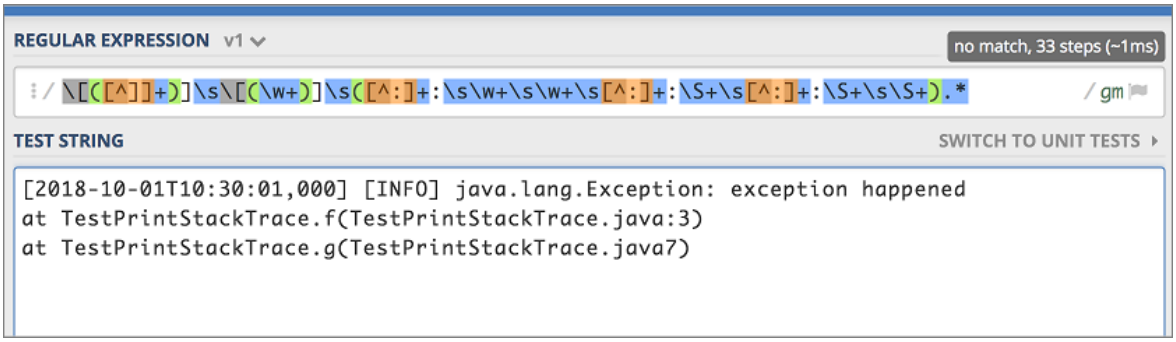
   On the right side of the page, you can view the explanation of the regular expression.
8. In the **TEST STRING** field, paste a sample log entry.

   In the following figure, the log content that is included in the message field is highlighted in orange, and the log content that is not included is highlighted in blue. The figure shows that the substring that follows the `at` word is not included in the `message` field. Therefore, the regular expression does not match fields in the sample log entry as expected and cannot be used to collect log data.

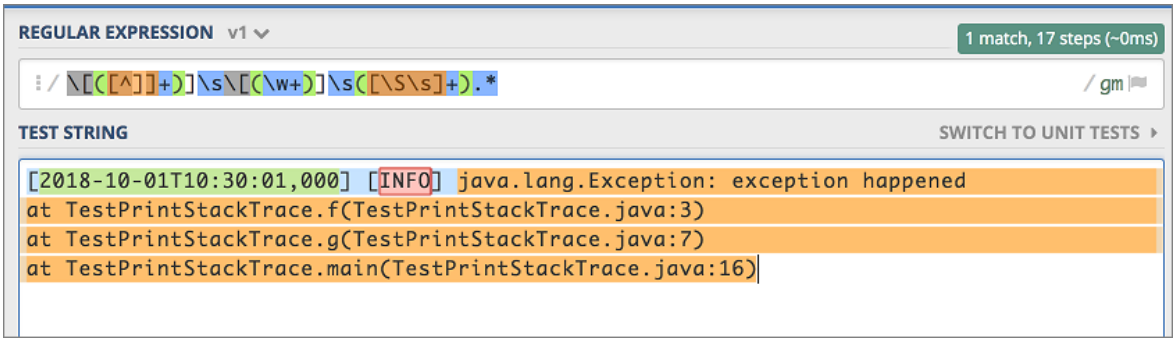
9. Check whether the regular expression can match fields in the sample log entry that contains two colons (::).

   The following figure shows that the regular expression fails to match fields in the sample log entry.

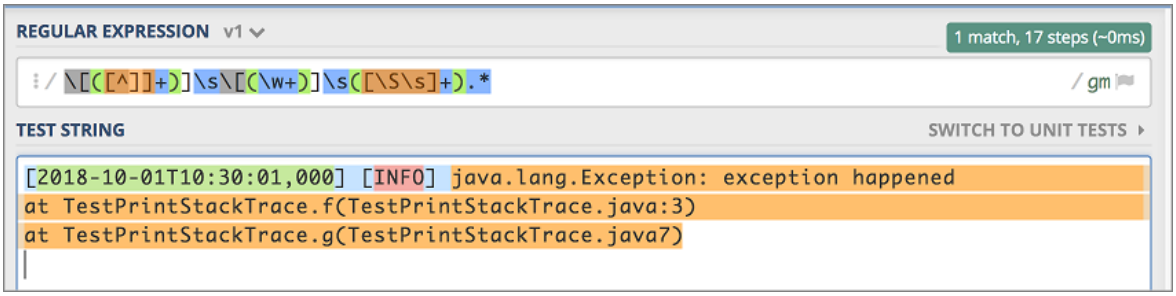


10. Replace the last subexpression in the regular expression with `[\S\s]+`, and check whether the regular expression can match fields in the sample log entries as expected.

    The following figure shows how the modified regular expression matches the substring that follows the `at` word.



    The following figure shows how the modified regular expression matches the sample log entry that contains two colons (::).



    You can perform the preceding steps to debug your regular expression. After you validate the regular expression, you can apply the expression to a Logtail configuration.

23.1.9.1.5. How do I optimize regular expressions?

A regular expression that accurately matches data with a high match rate can improve the performance of log collection.

When you optimize regular expressions, we recommend that you conform to the following rules:

- Use precise characters.

  We recommend that you do not use wildcard characters `.*` in a regular expression to match fields in log entries. Wildcard characters may cause mismatches that reduce the matching performance. For example, if you want to extract a field that consists of only letters, use `[A-Za-z]`.

- Use appropriate quantifiers.

  We recommend that you do not use `+,*`. For example, if you want to query data in a more efficient and accurate manner, use `\d` instead of `\d+` or `\d{1,3}` to match IP addresses.

- Debug regular expressions.

  If an error occurs when you use a regular expression, you can use online tools such as **regex101.com** to debug the regular expression. This way, you can troubleshoot the issue and optimize the regular expression in an efficient manner.

23.1.9.1.6. How do I use the full regex mode to collect log entries in multiple formats?

The full regex mode requires that log entries to be collected be in the same format. Therefore, if you want to collect log entries that are in multiple formats, you must use the schema-on-write or schema-on-read solution.

Taking Java logs as an example, the following section lists the types of error log entry and normal log entry.

- Multi-line WARNING log entries
- Simple text INFO log entries
- Key-value DEBUG log entries

```
[2018-10-01T10:30:31,000] [WARNING] java.lang.Exception: another exception happened
    at TestPrintStackTrace.f(TestPrintStackTrace.java:3)
    at TestPrintStackTrace.g(TestPrintStackTrace.java:7)
    at TestPrintStackTrace.main(TestPrintStackTrace.java:16)
[2018-10-01T10:30:32,000] [INFO] info something
[2018-10-01T10:30:33,000] [DEBUG] key:value key2:value2
```

To collect log entries of these types, you can use the following solutions:

- Schema-on-write: To extract log fields, you must apply multiple Logtail configurations with different regular expressions to a log file.

  > **Note** However, Logtail cannot apply multiple Logtail configurations directly to the same log file. Therefore, you must set up multiple symbolic links for the directory in which the log file resides. Each Logtail configuration applies to a symbolic link to collect log entries in a specific format.

- Schema-on-read: you can use a common regular expression to collect log entries in different formats.

  For example, if you want to collect log entries in multiple formats, you can configure a regular expression that matches the time and log level fields as the first line, and specify the rest of the log entries as the log message. If you want to parse the message, create an index for the message, specify a regular expression to extract log messages, and then extract target fields.

  > **Note** We recommend that you use this solution only for scenarios in which tens of millions of log entries are collected, or fewer.

23.1.9.1.7. How do I specify time formats for logs?

If you configure Logtail to collect logs, you must specify a common time format for the time field of the logs.

- The timestamp of a log is accurate to seconds. Therefore, you can specify the time format only to seconds.
- You need to specify the time format only for the time in the time field.

The following examples show time formats that are commonly used:

```
Custom1 2017-12-11 15:05:07
%Y-%m-%d %H:%M:%S
Custom2 [2017-12-11 15:05:07.012]
[%Y-%m-%d %H:%M:%S]
RFC822     02 Jan 06 15:04 MST
%d %b %y %H:%M
RFC822Z    02 Jan 06 15:04 -0700
%d %b %y %H:%M
RFC850      Monday, 02-Jan-06 15:04:05 MST
%A, %d-%b-%y %H:%M:%S
RFC1123     Mon, 02 Jan 2006 15:04:05 MST
%A, %d-%b-%y %H:%M:%S
RFC3339     2006-01-02T15:04:05Z07:00
%Y-%m-%dT%H:%M:%S
RFC3339Nano 2006-01-02T15:04:05.999999999Z07:00
%Y-%m-%dT%H:%M:%S
```

23.1.9.1.8. How do I configure non-printable characters in a sample log?

This topic describes how to configure non-printable characters in a sample log.

Context

If you collect logs in delimiter mode, Log Service allows you to specify a non-printable character as the delimiter or quote. Non-printable characters are characters whose decimal ASCII codes are within the range of 1 to 31 and 127. If you use a non-printable character as a delimiter or quote, you must find the hexadecimal ASCII code of the character and enter the character in the following format: `0x`*Hexadecimal ASCII code of the non-printable character*. For example, a sample log is `123456780`. You can specify 0x01 as the delimiter and 0x02 as the quote, and then enter a non-printable character 0x01 between the digits 5 and 6.

Procedure

1. Log on to the Log Service console.
2. Right-click the blank space on the browser and select **Inspect** from the shortcut menu.
3. On the page that appears, click the **Console** tab.
4. Enter `"\x01"` in the code editor and press Enter.
5. Copy the returned result.

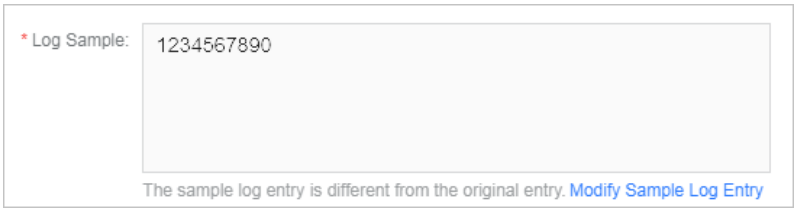
   A non-printable character is enclosed in double quotation marks (").

6. Paste the returned result between the digits 5 and 6.

   In the **Logtail Config** step, paste the result in the **Log Sample** field. For more information, see Collect DSV formatted logs.

7. Delete the double quotation marks (") between the digits 5 and 6.

   A non-printable character is configured in a sample log.



23.1.9.1.9. How do I troubleshoot errors that occur when I collect logs from containers?

If an error occurs when you use Logtail to collect logs from Docker containers, self-managed Kubernetes, or Container Service for Kubernetes (ACK), you can perform the steps that are described in this topic to troubleshoot the issue.

Troubleshoot an error if Log Service does not receive heartbeats from a Logtail client

To check whether Logtail is installed, perform the following steps:

1. View the heartbeat status of servers in a machine group.
   i. Log on to the Log Service console.
   ii. In the Projects section, click the project that you want to manage.
   iii. In the left-side navigation pane, click the **Machine Groups** icon.
   iv. In the Machine Groups list, click the name of the machine group that you want to view.

   In the **Machine Group Status** section, count the number of servers whose heartbeat status is **OK**.
2. Count the number of worker nodes in the related cluster.
   i. Log on to a master node in the Kubernetes cluster. For more information, see **Use kubectl to connect to a Kubernetes cluster** in *User Guide for Container Service for Kubernetes*.
   ii. Run the following command to view the number of worker nodes in the cluster:

   ```
   kubectl get node | grep -v master
   ```

   The following output is expected:

   ```
   NAME                                STATUS    ROLES     AGE       VERSION
   cn-hangzhou.i-bp17enxc2us3624wexh2   Ready     <none>    238d      v1.10.4
   cn-hangzhou.i-bp1ad2b02jtqd1shi2ut   Ready     <none>    220d      v1.10.4
   ```

3. Check whether the number of servers whose heartbeat status is **OK** in the machine group is equal to the number of worker nodes in the cluster. Troubleshoot the error based on the check result.
   - The heartbeat status of all servers in the machine group is **Failed**.

- If you use Logtail to collect standard Docker logs, check whether the values of the *${your_region_name}*, *${your_aliyun_user_id}*, and *${your_machine_group_user_defined_id}* parameters are valid. For more information, see the Parameters section in Collect standard Docker logs.
- If you use Logtail to collect ACK logs, submit a ticket.
- If you use Logtail to collect logs from self-managed Kubernetes, check whether the values of the *{your-project-suffix}*, *{aliuid}*, *{access-key-id}*, and *{access-key-secret}* parameters are valid. For more information, see the Parameters section in Collect Kubernetes logs.

  If the value of a parameter is invalid, run the `helm del --purge alibaba-log-controller` command to delete the installation package and re-install Logtail.

- The number of servers whose heartbeat status is**OK** is less than the number of worker nodes in the cluster.
  a. Check whether a DaemonSet is manually deployed by using a YAML file.

     Run the `kubectl get po -n kube-system -l k8s-app=logtail` command to perform the check. If the command returns pod information, a DaemonSet is manually deployed by using a YAML file.

  b. Download the latest version of the Logtail DaemonSet template.
  c. Set the *${your_region_name}*, *${your_aliyun_user_id}*, and *${your_machine_group_name}* parameters based on your business requirements.
  d. Run the `kubectl apply -f ./logtail-daemonset.yaml` command to update the DaemonSet YAML file.

  If the error persists, submit a ticket to contact Log Service technical support.

Troubleshoot an error if Log Service does not collect logs from containers

If no log is displayed in the **Consumption Preview** panel or on the **Search & Analysis** page of a Logstore, Log Service does not collect logs from the machine group of the Logstore. Check the status of the containers that correspond to the servers in the machine group. If the containers are working as expected, perform the following steps to troubleshoot the error:

1. Check the heartbeat status of the servers in the machine group. For more information, see View the heartbeat status of servers in a machine group.
2. Check whether the parameter settings in the related Logtail configuration are correct.

   Check whether the values of the IncludeLabel, ExcludeLabel, IncludeEnv, and ExcludeEnv parameters in the Logtail configuration meet your business requirements.

   > **Note** The `IncludeLabel` or ExcludeLabel parameter specifies whether to include the container images to which specified labels are attached. You can run the docker inspect command to retrieve a list of container image labels. These labels are not the labels that are defined by using Kubernetes. To check whether the parameter settings are valid in a Logtail configuration, delete the IncludeLabel, ExcludeLabel, IncludeEnv, and ExcludeEnv parameters of the Logtail configuration. If Log Service can collect logs from the containers after you delete the parameters, the parameter settings are invalid.

3. Check other items.

   Log Service does not collect logs from containers in the following scenarios:

   - Log files are not updated.
   - The log files of a container are not stored in the default storage or a storage attached to the container.

Other O&M operations

- Log on to a Logtail container
- View the operational logs of Logtail
- Ignore the stdout logs of a Logtail container
- View the status of Log Service components in a Kubernetes cluster

- View the version number, IP address, and startup time of Logtail

Log on to a Logtail container

Use one of the following methods based on your business requirements.

- Docker
  i. Log on to the host and run the following command to view and record the ID of the Logtail container:

  ```
  docker ps | grep logtail
  ```

  ii. Run the following command to log on to the Logtail container:

  ```
  docker exec -it [$ID] bash
  ```

  > **Note** [$ID] is the ID of the Logtail container.

- Kubernetes
  i. Run the following command to view and record the pod where the Logtail container resides:

  ```
  kubectl get po -n kube-system | grep logtail
  ```

  ii. Run the following command to log on to the pod:

  ```
  kubectl exec -it -n kube-system [$Pod_ID] bash
  ```

  > **Note** [$Pod_ID] is the ID of the pod.

View the operational logs of Logtail

The operational logs of Logtail are stored in the files named *ilogtail.LOG* and *logtail_plugin.LOG* in the */usr/local/ilogtail/* directory of a Logtail container.

1. Log on to a Logtail container. For more information, see Log on to a Logtail container.
2. Run the following command to go to the /usr/local/ilogtail/ directory:

   ```
   cd /usr/local/ilogtail
   ```

3. Run the following commands in sequence to view the ilogtail.LOG and logtail_plugin. LOG files:

   ```
   cat ilogtail.LOG
   cat logtail_plugin.LOG
   ```

Ignore the stdout logs of a Logtail container

The standard output of the container is irrelevant to this case. Ignore the following standard output:

```
start umount useless mount points, /shm$|/merged$|/mqueue$
umount: /logtail_host/var/lib/docker/overlay2/3fd0043af174cb0273c3c7869500fbe2bdb95d13b1e110172ef57fe840c82155/merged: must be superuser to unmount
umount: /logtail_host/var/lib/docker/overlay2/d5b10aa19399992755de1f85d25009528daa749c1bf8c16edff44beab6e69718/merged: must be superuser to unmount
umount: /logtail_host/var/lib/docker/overlay2/5c3125daddacedec29df72ad0c52fac800cd56c6e880dc4e8a640b1e16c22dbe/merged: must be superuser to unmount
......
xargs: umount: exited with status 255; aborting
umount done
start logtail
ilogtail is running
logtail status:
ilogtail is running
```

View the status of Log Service components in a Kubernetes cluster

1. Log on to a master node in the Kubernetes cluster. For more information, see **Use kubectl to connect to a Kubernetes cluster** in *User Guide for Container Service for Kubernetes*.
2. Run the following command to view the status of Log Service components in the Kubernetes cluster:

   ```
   helm status alibaba-log-controller
   ```

View the version number, IP address, and startup time of Logtail

1. Log on to a Logtail container. For more information, see Log on to a Logtail container.
2. Run the following command to view the version number, IP address, and start time of Logtail:

   ```
   kubectl exec logtail-ds-gb92k -n kube-system cat /usr/local/ilogtail/app_info.json
   ```

   The following output is expected:

   ```
   {
      "UUID" : "",
      "hostname" : "logtail-gb92k",
      "instance_id" : "0EBB2B0E-0A3B-11E8-B0CE-0A58AC140402_10.10.10.10_1517810940",
      "ip" : "203.0.113.10",
      "logtail_version" : "0.16.2",
      "os" : "Linux; 3.10.0-693.2.2.el7.x86_64; #1 SMP Tue Sep 12 22:26:13 UTC 2017; x86_64",
      "update_time" : "2018-02-05 06:09:01"
   }
   ```

23.1.9.2. Log search and analysis

23.1.9.2.1. FAQ about log query

This topic provides answers to some frequently asked questions (FAQ) about log query in Log Service.

How do I identify the source server from which Logtail collects logs during a query?

If a machine group uses IP addresses as its identifiers when logs are collected by using Logtail, the servers in the machine group are distinguished by internal IP addresses. When you query logs, you can use the hostname and custom IP address to identify the source server from which logs are collected.

For example, you can use the following query statement to calculate the number of occurrences of each hostname.

> **Note** You must configure an index for the __tag__:__hostname__ field and enable the analysis feature.

```
* | select "__tag__:__hostname__" , count(1) as count group by "__tag__:__hostname__"
```

How do I query IP addresses in logs?

You can use the exact match method to query IP addresses in logs. You can search for log data by IP address. For example, you can specify whether to include or exclude an IP address. However, you cannot use the partial match method to query the IP addresses in logs. This is because decimal points contained in an IP address are not default delimiters in Log Service. You can also filter data by using other methods. For example, you can use an SDK to download data and then use a regular expression or the string.indexof() method to search for results.

For example, if you execute the following query statement, the logs that contain the 203.0.113 CIDR block are still returned.

```
not ip:203.0.113 not status:200 not 360jk not DNSPod-Monitor not status:302 not jiankongbao
        not 301 and status:403
```

How do I query log data by using a keyword that contains a space character?

If you use a keyword that contains a space character to query log data, log data that contains a part of the keyword on the left or right of the space character is returned. You can enclose the keyword that contains a space character in double quotation marks (""). The entire enclosed content is regarded as a keyword to query log data as expected.

For example, you want to query log data that contains the keyword `POS version` from the following log data:

```
post():351]:&amp;nbsp;device_id:&amp;nbsp;BTAddr&amp;nbsp;:&amp;nbsp;B6:xF:xx:65:xx:A1&amp;nbsp;IMEI&amp;nbsp;:&amp;nbsp;35847xx22
xx81x9&amp;nbsp;WifiAddr&amp;nbsp;:&amp;nbsp;4c:xx:0e:xx:4e:xx&amp;nbsp;|&amp;nbsp;user_id:&amp;nbsp;bb07263xxd2axx43xx9exxea2
6e39e5f&amp;nbsp;POS&amp;nbsp;version:903
```

If you use `POS version` as the keyword, log data that contains `POS` and `version` is returned. This query result does not meet your requirements. If you use `"POS version"` as the keyword, log data that contains the keyword `POS version` is returned.

How do I use two conditions to query log data?

You can specify two conditions in a query statement to query log data.

For example, if you want to query logs in which the value of the status field is not 200 and the value of the request_method field is not GET in a Logstore, you can execute the `not status:200 not request_method:GET` statement to query logs as expected.

How do I query collected logs in Log Service?

You can use one of the following methods to query logs in Log Service:

- Use the Log Service console.
- Use an SDK.
- Use the Restful API.

23.1.9.2.2. What can I do if I cannot obtain the required results from a log query?

If you cannot find the required log data by using the query feature of Log Service, perform the steps that are described in this topic to troubleshoot the issue.

Log collection failure

If log data fails to be collected by Log Service, you cannot query the log data. Check whether log data is available on the consumption preview page of your Logstore.

If log data is available, log data is collected by Log Service.

If log data is unavailable, check whether the issue occurs due to the following causes:

- The log source does not generate log data.

  If no log data is generated by the log source, no log data can be sent to Log Service. Check your log source.

- Logtail has no heartbeat.

  On the **Machine Group Settings** page, check whether the heartbeat status of the related server is OK in the Machine Group Status section. For more information about how to troubleshoot the issue if Log Service does not receive heartbeats from a Logtail client, see What can I do if no heartbeat packet is received from a Logtail client?.

- Data is not written to the file that is monitored in real time.

  If data is not written to the file that is monitored in real time, you can view error messages in the */usr/local/ilogtail/ilogtail.LOG* file. Common error messages:

  - parse delimiter log fail: The error message returned because an error occurs when Log Service collects logs in delimiter mode.
  - parse regex log fail: The error message returned because an error occurs when Log Service collects logs in full regex mode.

Delimiter setting errors

View the specified delimiters and check whether you can use a keyword to find a log after the log content is split by using the delimiters. In this example, the default delimiters `,;=()[]{}?@&<>/:'` are used. If a log contains abc"defg,hij, the log is split into the following two words: abc"defg and hij. If the log is split, you cannot use the keyword abc to find the log.

Fuzzy match is supported. For more information, see Query syntax.

> **Note**
>
> - The indexing feature of Log Service is optimized. If you configure both full-text indexes and field indexes, the configurations of the field indexes take precedence. This helps you reduce indexing costs. For example, you configure an index for a log field whose key is message and specify a space character as a delimiter. To use a space character as a delimiter, you must specify the space character in the middle of the delimiters that you specify for an index. You can find a log that contains "message: this is a test message" by using the keyword message:this in the key:value format. However, you cannot find the log by using the keyword this because you have configured a field index for the key and the full-text indexing feature does not take effect for the log field.
> - You can create indexes or modify existing indexes. However, new or modified indexes take effect only for new data.
>
>   You can click Index Attributes to check whether the specified delimiters meet your business requirements.

Other reasons

If log data is generated, modify the query time range and perform a query operation again. Log Service allows you to preview log data in real time. The maximum latency of the query feature is 1 minute. We recommend that you query log data at least 1 minute after logs are generated.

If the issue persists, submit a ticket.

23.1.9.2.3. What are the differences between log consumption and log query?

Log Service provides the log consumption and log query features that allow you to read log data from Log Service.

Log consumption

The log consumption feature allows you to read and write full data in the first-in, first-out (FIFO) order. This feature is similar to the features provided by Kafka.

- Each Logstore has one or more shards. Data is written to a random shard.
- You can read multiple logs at a time from a specified shard based on the order in which the logs were written to the shard.
- You can specify a start position (cursor) to pull logs from shards based on the time when Log Service receives the logs.

Log query

Log Service allows you to query and analyze a large amount of log data based on specific conditions.

- You can specify query conditions to find required log data.
- You can use multiple operators such as AND, NOT, and OR to specify query conditions and perform SQL analysis on query results.
- The log query feature is independent of shards.

Differences

Item	Log query	Log consumption
Search by keyword	Supported.	Not supported.
Data read (a small amount of data)	Fast.	Fast.
Data read (full data)	Slow. Log Service reads 100 logs in 100 milliseconds. We recommend that you do not use this method.	Fast. Log Service reads 1 MB of log data in 10 milliseconds. We recommend that you use this method.
Data read by topic	Yes.	No. Data is identified only by shard.
Data read by shard	No. Data in all shards of a Logstore is queried.	Yes. You need to specify a shard each time to read data.
Fee	Medium.	Low.
Scenario	Monitoring, issue troubleshooting, and analysis.	Full data processing scenarios, such as stream computing and batch processing.

23.1.9.2.4. How do I resolve common errors that occur when I query log data?

This topic describes the common error messages that are returned when you query log data in the Log Service console and provides related solutions.

Error messages

Error message	Cause	Solution
line 1:44: Column 'XXX' cannot be resolved;please add the column in the index attribute	The XXX key cannot be specified in the query statement because the key does not exist.	Click Index Attributes to configure an index for the field. For more information, see Enable the indexing feature and configure indexes for a Logstore.
ErrorType:QueryParseError.ErrorMessage:syntax error error position is from column:10 to column:11,error near < : >	The query statement contains unnecessary colons (:).	Delete the unnecessary colons (:) from the query statement, and then execute the query statement.
Column 'XXX' not in GROUP BY clause;please add the column in the index attribute	You use a GROUP BY clause and specify a non-GROUP BY field in a SELECT statement. For example, you do not specify the key1 field in the `select key1, avg(latency) group by key2` statement in the GROUP BY clause.	You must specify the same field that you specified in the SELECT statement in the GROUP BY clause. Example: `* | select key1,avg(latency) group by key1,key2`
sql query must follow search query,please read syntax doc	The syntax of the query statement is invalid because a search statement is not specified.	Invalid query statement: `select ip,count(*) group by ip`. Valid query statement: `*|select ip,count(*) group by ip`.
line 1:10: identifiers must not start with a digit; surround the identifier with double quotes	The column name or variable name that is referenced in an SQL statement cannot start with a digit.	Change the column name or variable name to a name that starts with a letter.
line 1:9: extraneous input '' expecting	One or more words are misspelled.	Correct the misspelled words.
key (XXX) is not config as key value config,if symbol : is in your log,please wrap : with quotation mark "	The XXX field cannot be referenced in the analytic statement because no field index is configured for the field.	Click Index Attributes to configure an index for the field. For more information, see Enable the indexing feature and configure indexes for a Logstore.
Query exceeded max memory size of 3GB	The size of the memory that is used by the query statement exceeds 3 GB. The issue occurs because a large number of values are returned in the query result after you use a `GROUP BY` clause to remove duplicates.	Optimize the `GROUP BY` clause. Reduce the number of keys that is specified in the `GROUP BY` clause.

Error message	Cause	Solution
ErrorType:ColumnNotExists.ErrorPosition,line:0,column:1.ErrorMessage:line 1:123: Column 'XXX' cannot be resolved; it seems XXX is wrapper by ";if XXX is a string ,not a key field, please use 'XXX'	XXX is not an indexed field. In an SQL statement, you must enclose an indexed field in double quotation marks ("") and you must enclose a string in single quotation marks ('').	If you want to reference the XXX field, make sure that you index the field and enable the analysis feature for the field. For more information, see Enable the indexing feature and configure indexes for a Logstore. **Note** If XXX is a string, you must replace the double quotation marks (") with single quotation marks (').
user can only run 15 query concurrently	More than 15 concurrent search statements are executed. A maximum of 15 concurrent search statements can be executed by a user in a project.	Reduce the number of search statements based on your business requirements.
unclosed string quote	The double quotation marks (") in the query statement are incomplete.	Check the query statement, specify double quotation marks (") in pairs, and then execute the query statement.
error after :.error detail:error after :.error detail:line 1:147: mismatched input 'in' expecting {<EOF>, 'GROUP', 'ORDER', 'HAVING', 'LIMIT', 'OR', 'AND', 'UNION', 'EXCEPT', 'INTERSECT'}	The syntax of the query statement is valid.	Modify the SQL statement as prompted and execute the SQL statement.
Duplicate keys (XXX) are notallowed	Indexes are not case-sensitive. For example, if the aBc index exists, an error message is returned when you create the abc index.	Check whether duplicate indexes exist.
only support * or ? in the middle or end of the query	You can specify only asterisks (*) and question marks (?) in the middle or end of a field value to perform a fuzzy match.	You can use the SQL LIKE operator in a query statement based on your business requirements. For example, you cannot use `Msg: *xxx` to search for logs that contains the Msg field and the field value ends with XXX. You can use the SQL LIKE operator to perform a fuzzy match, as shown in the following query statement: `Msg: * | SELECT Msg WHERE Msg LIKE '%xxx'`
logstore (xxx) is not found	The XXX Logstore does not exist or the analysis feature is not enabled.	Check whether the Logstore exists. If the Logstore exists, you must index at least one field and enable the analysis feature for the Logstore.

Error message	Cause	Solution
condition number 43 is more than 30	A maximum of 30 fields can be referenced by a user in a query statement.	Modify the query statement to reduce the number of the referenced fields, and then execute the query statement.

23.1.9.2.5. Why data queries are inaccurate?

This topic describes the causes for inaccurate data queries. It also includes solutions to these issues.

When you search and analyze log data, the message **The results are inaccurate** may prompt in the console. This indicates that the returned result is inaccurate because some log data in a Logstore was not queried.

Possible causes include:

The time range for queries is excessive.

- Cause

  The time range for a query is excessively wide, for example, three months or a year. In this case, Log Service cannot scan all log data generated within this time period for one query.

- Solution

  Narrow down the query time range and perform multiple queries.

Query statements are complex.

- Cause

  The query statement is exceedingly complex or contains multiple frequently used words. In this case, Log Service cannot scan all related log data or read the query results at one time.

- Solution

  Narrow down the query scope and perform multiple queries.

The SQL computing needs to read an excessively large amount of data.

- Cause

  The SQL computing needs to read an excessively large amount of data. In this case, query results are likely to become inaccurate. A maximum of 1 GB of data can be read from each shard. For example, if the SQL computing needs to read strings from multiple columns, which exceed the threshold data volume, inaccurate query results will be returned.

- Solution

  Narrow down the query scope and perform multiple queries.

23.1.9.3. Alarm

23.1.9.3.1. FAQ about alerts

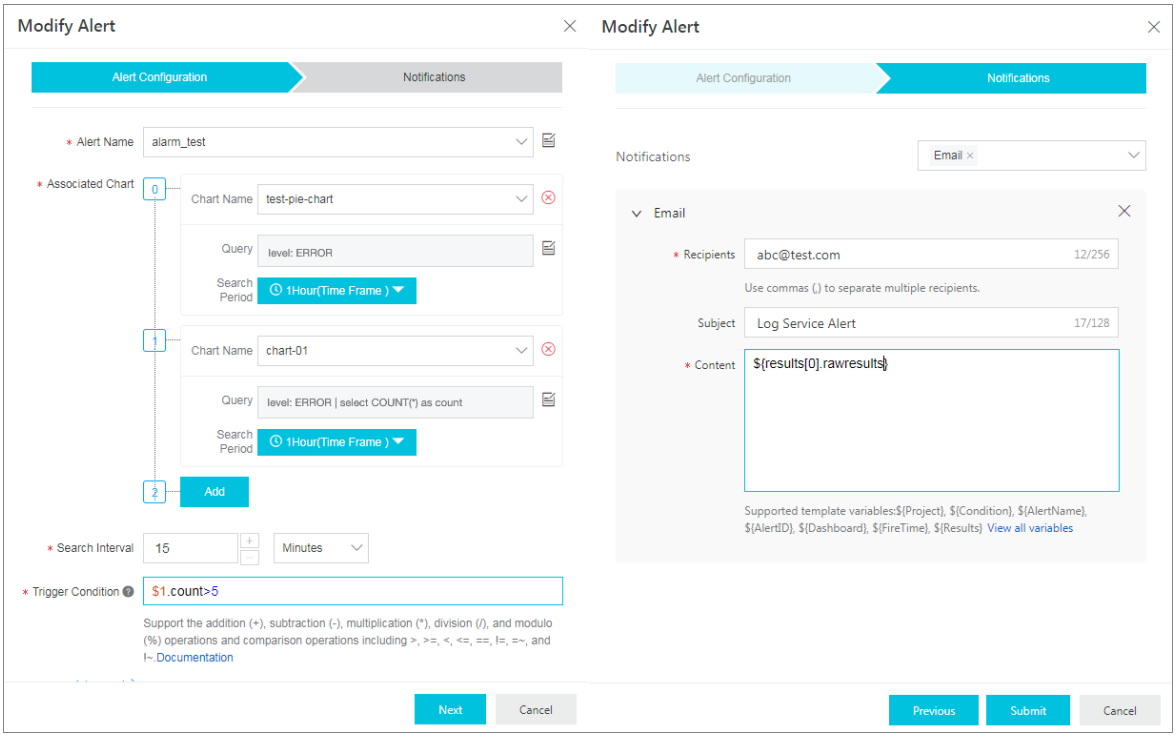
This topic provides answers to some frequently asked questions (FAQ) about alerts in the Log Service console.

How do I add raw logs to an alert notification?

For example, if more than five error logs are detected in the previous 5 minutes, an alert is triggered and an alert notification is sent. To add the raw logs to the alert notification, perform the following steps:

1. Log on to the Log Service console.

2. In the Projects section, click the project that you want to manage.
3. In the left-side navigation pane, click the icon.
4. In the Dashboard list, click the dashboard for which you want to configure alert rules.
5. In the upper-right corner of the dashboard page, choose **Alerts > Create**.
6. In the **Create Alert** panel, set the parameters. The following figure shows the parameter settings.



The following example shows how to set the parameters:

- Query
  - Association Chart 0: **level:ERROR**
  - Association Chart 1: **level: ERROR | select COUNT(*) as count**
- Trigger Condition: **$1.count > 5**
- Content: **${results[0].rawresults}**

7. Click **Submit**.

What can I do if the DingTalk chatbot fails to send a notification?

For example, after you set the notification method to WebHook-DingTalk Bot, the following error message is returned when the DingTalk chatbot sends a notification:

```
{"errcode":310000,"errmsg":"sign not match"}
{"errcode":310000,"errmsg":"keywords not in content"}
```

This error message is returned because the security settings of the latest chatbot are invalid. You can reconfigure the security settings. For more information, see DingTalk chatbot webhooks. If the issue persists or other error messages are returned, submit a ticket.

24.Apsara Stack DNS

24.1. User Guide

24.1.1. What is Apsara Stack DNS?

Apsara Stack DNS is a service that runs on Apsara Stack to resolve domain names. You can configure rules to map domain names to IP addresses. Apsara Stack DNS then distributes domain name requests from clients to cloud resources, business systems on your internal networks, or the business resources of Internet service providers (ISPs).

Apsara Stack DNS provides DNS resolution in VPCs. You can perform the following operations on your VPC by using Apsara Stack DNS:

- Manage internal domain names.
- Manage DNS records of internal domain names.
- Manage forwarding configurations.
- Manage recursive resolution configurations.

24.1.2. User roles and permissions

Role	Permission
System administrator	A user of this role has read, write, and execute permissions on all level-1 organization resources, global resources, and system configurations.
Level-1 organization administrator	A user of this role has read, write, and execute permissions on level-1 organization resources to which the user belongs, but does not have permissions on level-1 organization resources, global resources, or system configurations to which other users belong.
Lower-level organization administrator	A user of this role does not have permissions on Apsara Stack DNS. The user does not have permissions on level-1 organization resources to which the user belongs, and does not have permissions on level-1 organization resources, global resources, or system configurations to which other users belong.
Resource user	A user of this role does not have permissions on Apsara Stack DNS. The user does not have permissions on level-1 organization resources to which the user belongs, and does not have permissions on level-1 organization resources, global resources, or system configurations to which other users belong.
Other roles	A user of this role does not have permissions on Apsara Stack DNS. The user does not have permissions on level-1 organization resources to which the user belongs, and does not have permissions on level-1 organization resources, global resources, or system configurations to which other users belong.

24.1.3. Log on to the Apsara Stack DNS console

This topic describes how to log on to the Apsara Stack DNS console by using Google Chrome.

Prerequisites

- The URL of the Apsara Uni-manager Management Console is obtained from the deployment personnel before you log on to the Apsara Uni-manager Management Console.
- We recommend that you use the Google Chrome browser.

Procedure

1. In the address bar, enter the URL of the Apsara Uni-manager Management Console. Press the Enter key.
2. Enter your username and password.

   Obtain the username and password that you can use to log on to the console from the operations administrator.

   > **Note** When you log on to the Apsara Uni-manager Management Console for the first time, you must change the password of your username. Your password must meet complexity requirements. The password must be 10 to 32 characters in length and must contain at least two of the following character types:
   >
   > - Uppercase or lowercase letters
   > - Digits
   > - Special characters, which include ! @ # $ %

3. Click **Log On**.
4. If your account has multi-factor authentication (MFA) enabled, perform corresponding operations in the following scenarios:
   - It is the first time that you log on to the console after MFA is forcibly enabled by the administrator.
     a. On the Bind Virtual MFA Device page, bind an MFA device.
     b. Enter the account and password again as in Step 2 and click **Log On**.
     c. Enter a six-digit MFA verification code and click **Authenticate**.
   - You have enabled MFA and bound an MFA device.

     Enter a six-digit MFA authentication code and click **Authenticate**.

     > **Note** For more information, see the *Bind a virtual MFA device to enable MFA* topic in *Apsara Uni-manager Operations Console User Guide*.

5. In the top navigation bar, choose **Products > Networking > Apsara Stack DNS**.

24.1.4. Internal DNS resolution management

Internal DNS resolution management allows you to manage global internal domain names, global forwarding configurations, and global recursive resolution configurations that you have created in Apsara Stack.

24.1.4.1. Global internal domain names

24.1.4.1.1. Overview

All the operations of this feature require administrator privileges.

24.1.4.1.2. View an internal domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domains > Global Internal Domains**.
3. In the **Domain Name** search box, enter the domain name that you want to view.
4. Click **Search**.

   The search result is displayed.

24.1.4.1.3. Add a domain name

This topic describes how to add a domain name in the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Global Internal Domain Name**.
3. Click **Add Domain Name**.
4. In the dialog box that appears, enter **Global Internal Domain Name**.
5. Click **OK**.

24.1.4.1.4. Add a description for a domain name

This topic describes how to add a description for a domain name in the Apsara Uni-manager Management Console.

Context

You can add a description for a domain name to help you identify it. For example, you can add a hostname or internal system information to describe a domain name.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Global Internal Domain Name**.
3. Find the domain name for which you want to add a description, click the  icon in the Actions column, and then select **Description**.
4. In the dialog box that appears, enter a description.
5. Click **OK**.

24.1.4.1.5. Delete a domain name

This topic describes how to delete a domain name in the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Global Internal Domain Name**.
3. Find the domain name that you want to delete, click the  icon in the **Actions** column, and then select **Delete**.
4. In the message that appears, click **OK**.

24.1.4.1.6. Delete multiple domain names

This topic describes how to delete unnecessary domain names at a time in the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Global Internal Domain Name**.
3. Select one or more domain names that you want to delete and click **Delete** in the upper-right corner.
4. In the message that appears, click **OK**.

24.1.4.1.7. Configure DNS records

This topic describes how to configure DNS records in the Apsara Uni-manager Management Console.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Global Internal Domain Name**.
3. Find the domain name for which you want to configure DNS records, click the icon in the Actions column, and then select **Configure DNS Records**.
4. On the **Configure DNS Records** page, click **Add DNS Record** in the upper-right corner.
5.

24.1.4.1.8. View a resolution policy

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domains > Global Internal Domains**.
3. Find the target domain name, click the icon in the Actions column, and then select **Configure DNS Records**.
4. On the page that appears, select the domain name for which you want to configure DNS records, and click **Weight** in the **Resolution Policy** column.
5. On the page that appears, view the details of **Resolution Policy**.

24.1.4.2. Global forwarding configurations

24.1.4.2.1. Global forwarding domain names

24.1.4.2.1.1. Overview

All operations of this feature require administrator privileges.

Apsara Stack DNS forwards specific domain names to other DNS servers for resolution.

Two forwarding modes are available: forward all requests without recursion and forward all requests with recursion.

- Forward all requests without recursion: Only a specified DNS server is used to resolve domain names. If the specified DNS server cannot resolve the domain names or the request times out, a message is returned to the DNS client to indicate that the query failed.

- Forward all requests with recursion: A specified DNS server is preferentially used to resolve domain names. If the specified DNS server cannot resolve the domain names, the local DNS is used instead.

24.1.4.2.1.2. View global forwarding domain names

This topic describes how to view global forwarding domain names in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Forwarding Domain Names**.
3. In the **Domain Name** search box, enter the domain name that you want to query and click **Search**.

24.1.4.2.1.3. Add a domain name

This topic describes how to add a domain name in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Forwarding Domain Names**.
3. Click **Add Domain Name**.
4. In the dialog box that appears, configure *Global Forwarding Domain*, *Forwarding Mode*, and *Forwarder IP Addresses*. Then, click **OK**.

24.1.4.2.1.4. Add a description for a domain name

This topic describes how to add a description for a domain name in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Context

You can add a description for a domain name to help you identify it. For example, you can describe a domain name by using a hostname or internal system information.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Forwarding Domain Names**.
3. Select the domain name for which you want to add a description, click  in the Actions column, and then select **Description**.
4. In the dialog box that appears, enter a description and click **OK**.

24.1.4.2.1.5. Modify the forwarding configurations of a domain name

This topic describes how to modify the forwarding configurations of a domain name in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Forwarding Domain Names**.
3. Find the domain name whose forwarding configurations you want to modify, click the icon in the Actions column, and then select **Modify**.
4. In the dialog box that appears, change the value of *Forwarding Mode* or *Forwarder IP Addresses*, and click **OK**.

24.1.4.2.1.6. Delete a domain name

This topic describes how to delete a domain name in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Forwarding Domain Names**.
3. Find the domain name that you want to delete, click the icon in the Actions column, and then select **Delete**.
4. Click **OK**.

24.1.4.2.1.7. Delete multiple domain names

This topic describes how to delete multiple domain names at a time in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Domains > Forwarding Settings > Global Forwarding Domain Names**.
3. Select one or more domain names that you want to delete and click **Delete** in the upper-right corner.
4. In the message that appears, click **OK**.

24.1.4.2.2. Global default forwarding configurations

24.1.4.2.2.1. Enable default forwarding

This topic describes how to enable default forwarding in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Default Forwarding**.
3. Click the icon in the Actions column and select **Enable**.
4. In the dialog box that appears, configure *Default Forwarding Mode* and *Forwarder IP Addresses*. Then, click **OK**.

Make sure that Enable Default Forwarding is set to ON.

24.1.4.2.2.2. Modify default forwarding configurations

This topic describes how to modify default forwarding configurations in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Default Forwarding**.
3. Click the  icon in the Actions column and select **Modify**.
4. In the dialog box that appears, configure *Forwarding Mode* and *Forwarder IP Addresses*. Then, click **OK**.

24.1.4.2.2.3. Disable default forwarding

This topic describes how to disable default forwarding in the Apsara Uni-manager Management Console. This operation requires administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Internal Domain Names > Forwarding Settings > Global Default Forwarding**.
3. Click the  icon in the Actions column and select **Disable**.
4. In the message that appears, click **OK**.

24.1.4.3. Global recursive resolution

24.1.4.3.1. Enable global recursive resolution

Prerequisites

You have administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. Choose **Internal Domains > Global Recursive Resolution**.
3. Click the  icon in the Actions column and select **Enable**.
4. In the dialog box that appears, click **OK**.

24.1.4.3.2. Disable global recursive resolution

Prerequisites

You have administrator permissions.

Procedure

1. Log on to the Apsara Stack DNS console.
2. Choose **Internal Domains > Global Recursive Resolution**.
3. Click the icon in the Actions column and select **Disable**.
4. In the dialog box that appears, click **OK**.

24.1.5. PrivateZone (DNS Standard Edition only)

The PrivateZone feature allows you to create VPC-specific tenant domain names. You can bind the domain names to VPCs as required to achieve tenant isolation.

24.1.5.1. Tenant internal domain name

24.1.5.1.1. View a domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. In the **Domain Name** search box, enter the domain name that you want to view.
4. Click Search.

   The search result is displayed.

24.1.5.1.2. Add a domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Click **Add Domain Name**.
4. In the dialog box that appears, set *Tenant Internal Domain Name*.
5. Click **OK**.

24.1.5.1.3. Bind an organization to a VPC

Tenant domain names are isolated based on VPCs. To ensure that the DNS forwarding configurations take effect, you must bind the organization of domain names to a VPC.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Find the target domain name, click the icon in the Actions column, and select **Associate VPCs**.
4. Select one or more VPCs from the list of VPCs to Select, click the right arrow to add them to the list of VPCs Selected, and then click **OK**.

24.1.5.1.4. Unbind a domain name from a VPC

This topic describes how to unbind a domain name from a VPC.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Find the target domain name and click the number in the **VPCs Associated** column.
4. On the **VPCs Associated** page, find the target VPC, click the icon in the **Actions** column, and then select **Disassociate**.
   Make sure that the unbound VPC is no longer displayed on the VPCs Associated page.

24.1.5.1.5. Add a description for a domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Find the target domain name, click the icon in the Actions column, and then select **Description**.
4. In the dialog box that appears, enter a description.
5. Click **OK**.

24.1.5.1.6. Delete a domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Find the target domain name, click the icon in the Actions column, and then select **Delete**.
4. In the message that appears, click **OK**.

24.1.5.1.7. Delete multiple domain names

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Select one or more domain names that you want to delete and click **Delete** in the upper-right corner.
4. In the message that appears, click **OK**.

24.1.5.1.8. Configure DNS records

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Find the target domain name, click the icon in the Actions column, and then select **Configure DNS Records**.

4. In the upper-right corner of the **Configure DNS Records** page, click **Add DNS Record**.
5. In the **Add DNS Record** dialog box, configure *Host*, *Type*, *TTL*, *Resolution Policy*, and *Record Set*. Then, click **OK**.

   The following tables describe the types of DNS records.

   - A record

Resolution policy	Formatting rule
None	You can enter up to 100 unique IPv4 addresses, each in a separate row. Make sure that the IPv4 addresses are valid. Example: ■ 192.168.1.1 ■ 192.168.1.2 ■ 192.168.1.3
Weight	You can enter up to 100 unique IPv4 addresses, each in a separate row. Format: ■ [IPv4 address] [Weight] (The IPv4 address and weight are separated with a space.) ■ Make sure that the IPv4 addresses are valid. ■ The weight value is an integer ranging from 0 to 999. A larger value indicates a greater weight. Example: ■ 192.168.1.1 20 ■ 192.168.1.1 30 ■ 192.168.1.1 50

   - AAAA record

Resolution policy	Formatting rule
None	You can enter up to 100 unique IPv6 addresses, each in a separate row. Make sure that the IPv6 addresses are valid. Example: ■ 2400:3200::6666 ■ 2400:3200::6688 ■ 2400:3200::8888

Resolution policy	Formatting rule
Weight	You can enter up to 100 unique IPv6 addresses, each in a separate row. Format: ■ [IPv6 address] [Weight] (The IPv6 address and weight are separated with a space.) ■ Make sure that the IPv6 addresses are valid. ■ The weight value is an integer ranging from 0 to 999. A larger value indicates a greater weight. Example: ■ 2400:3200::6666 20 ■ 2400:3200::6688 20 ■ 2400:3200::8888 60

- CNAME record

Resolution policy	Formatting rule
None	You can enter only one domain name. The domain name must be a fully qualified domain name (FQDN) that ends with a dot (.). It must be 1 to 255 characters in length. Example: www.example.com.
Weight	You can enter up to 100 unique domain names, each in a separate row. Format: ■ [Domain name] [Weight] (The domain name and weight are separated with a space.) ■ The domain name must be an FQDN that ends with a period (.). It must be 1 to 255 characters in length. ■ The weight value is an integer ranging from 0 to 999. A larger value indicates a greater weight. Example: ■ www1.example.com. 20 ■ www2.example.com. 20 ■ www3.example.com. 60

- MX record

Resolution policy	Formatting rule

Resolution policy	Formatting rule
None	You can enter 100 unique email server hostnames, each in a separate row. Format: ■ [Priority] [Email server hostname] (The priority and hostname are separated with a space.) ■ The priority value is an integer ranging from 0 to 999. A smaller value indicates a higher priority. ■ The email server hostname must be an FQDN that ends with a period (.). It must be 1 to 255 characters in length. Example: ■ 10 mailserver1.example.com. ■ 20 mailserver2.example.com.

- TXT record

Resolution policy	Formatting rule
None	You can enter up to 100 unique character strings, each in a separate row. A string must be 1 to 255 characters in length. No row can be left blank. Example: "v=spf1 ip4:192.168.0.1/16 ip6:2001::1/96 ~all"

- PTR record

Resolution policy	Formatting rule
None	You can enter up to 100 unique domain names, each in a separate row. The DNS server address must be an FQDN that ends with a period (.). It must be 1 to 255 characters in length. Example: ■ www1.example.com. ■ www2.example.com. ■ www3.example.com.

- SRV record

Resolution policy	Formatting rule

Resolution policy	Formatting rule
None	You can enter up to 100 unique application server hostnames, each in a separate row. Format: ■ [Priority] [Weight] [Port number] [Application server hostname] (Every two items are separated with a space.) ■ The priority value is an integer ranging from 0 to 999. A smaller value indicates a higher priority. ■ The weight value is an integer ranging from 0 to 999. A larger value indicates a greater weight. ■ The port number is an integer ranging from 0 to 65535. It indicates the TCP or UDP port used for network communications. ■ The application server hostname must be an FQDN that ends with a period (.). It must be 1 to 255 characters in length. Example: ■ 1 10 8080 www1.example.com. ■ 2 20 8081 www2.example.com.

- NAPTR record

Resolution policy	Formatting rule
None	You can enter up to 100 unique NAPTR record values, each in a separate row. Format: ■ [Serial number] [Priority] [Flag] [Service information] [Regular expression] [Substitute domain name] (Every two items are separated with a space.) ■ The serial number is an integer ranging from 0 to 999. A smaller value indicates a higher priority. ■ The priority value is an integer ranging from 0 to 999. A smaller value indicates a higher priority. If two records have the same serial number, the one with a higher priority takes effect first. ■ The flag value can be left blank or be a character from A to Z, a to z, or 0 to 9. It is not case-sensitive and must be enclosed in double quotation marks (""). ■ The service information can be left blank or be a string of 1 to 32 characters. It must start with a letter and be enclosed in double quotation marks (""). ■ The regular expression can be left blank or be a string of 1 to 255 characters enclosed in double quotation marks (""). ■ The substitute domain name must be an FQDN that ends with a period (.). It must be 1 to 255 characters in length. Example: ■ 100 50 "S" "Z3950+I2L+I2C" "" _z3950._tcp.example.com. ■ 100 50 "S" "RCDS+I2C" "" _rcds._udp.example.com. ■ 100 50 "S" "HTTP+I2L+I2C+I2R" "" _http._tcp.example.com.

- CAA record

Resolution policy	Formatting rule
None	You can enter up to 100 unique CAA records, each in a separate row. Format: ■ [Certificate authority flag] [Certificate property tag] [Authorization information] (Every two items are separated with a space.) ■ The certification authority flag is an integer ranging from 0 to 255. ■ The certificate property tag can be issue, issuewild, or iodef. ■ The authorization information must be 1 to 255 characters in length and enclosed in double quotation marks (""). Example: ■ 0 issue "caa.example.com" ■ 0 issuewild ";" ■ 0 iodef "mailto:example@example.com"

- NS record

Resolution policy	Formatting rule
None	You can enter up to 100 unique DNS server addresses, each in a separate row. The DNS server address must be an FQDN that ends with a period (.). It must be 1 to 255 characters in length. Wildcard domain names are not allowed. Example: ■ ns1.example.com. ■ ns2.example.com.

6. After you add DNS records, perform the following operations as required:
   - Add a description for a DNS record.

     Find the target DNS record, click the  icon in the Actions column, and then select **Description**. In the dialog box that appears, enter a description and click **OK**.
   - Delete a DNS record.

     Find the target DNS record, click the  icon in the Actions column, and then select **Delete**. In the message that appears, click **OK**.
   - Modify a DNS record.

     Find the target DNS record, click the  icon in the Actions column, and then select **Modify**. In the dialog box that appears, modify the required parameters and click **OK**.
   - Delete multiple DNS records.

     Select the DNS records that you want to modify and click **Delete** in the upper-right corner. In the message that appears, click **OK**.

24.1.5.1.9. View a resolution policy

This topic describes how to view the details of a resolution policy.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Internal Domains**.
3. Find the target domain name, click the  icon in the Actions column, and then select Configure DNS Records.
4. View the resolution policy in the DNS Records list.

24.1.5.2. Tenant forwarding configurations

24.1.5.2.1. Tenant forwarding domain names

24.1.5.2.1.1. View a tenant forwarding domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. In the **Domain Name** search box, enter the domain name that you want to view.
4. Click **Search**.

   The search result is displayed.

24.1.5.2.1.2. Add a tenant forwarding domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Click **Add Domain Name**.
4. In the dialog box that appears, configure parameters such as *Domain Name*, *Forwarding Mode*, and *Forwarder IP Addresses*.

Parameter	Description
Domain Name	The domain name, which must meet the following formatting rules: ◦ The domain name must be 1 to 255 characters in length. This includes the period (.) at the end of the domain name. ◦ The domain name can contain multiple domain name segments that are separated with periods (.). A domain name segment must be 1 to 63 characters in length. It cannot contain consecutive periods (.) or be left blank. ◦ The domain name can only contain letters (a to z, A to Z), digits (0 to 9), hyphens (-), and underscores (_). ◦ The domain name must start with a letter, digit, or underscore (_) and end with a letter, digit, or period (.). ◦ The domain name is not case-sensitive. The system saves the domain name in lowercase letters. ◦ The period (.) at the end of the domain name is optional. The system adds a period (.) to the end of the domain name.

Parameter	Description
Forwarding Mode	For both domain name-based forwarding and default forwarding, the following two forwarding modes are supported: ○ Forward All Requests without Recursion: forwards DNS requests to the target DNS server. If the target DNS server cannot resolve the domain names, a message is returned to the DNS client indicating that the query failed. ○ Forward All Requests with Recursion: A specified DNS server is preferentially used to resolve domain names. If the specified DNS server cannot resolve the domain names, the local DNS is used instead. If you enter internal IP addresses in the Forwarder IP Addresses field, unexpected results may occur during recursive resolution. For example, a domain name used for internal network services may be resolved to a public IP address.
Forwarder IP Addresses	A list of destination IP addresses. **Note** Multiple IP addresses are separated with semicolons (;).

5. Click **OK**.

24.1.5.2.1.3. Bind an organization to a VPC

Tenant domain names are isolated based on VPCs. You must bind the organization of domain names to a VPC before the DNS forwarding configurations can take effect.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Find the target domain name, click the  icon in the Actions column, and then select **Associate VPCs**.
4. Select one or more VPCs from the list of VPCs to Select, click the right arrow to add them to the list of VPCs Selected, and then click **OK**.

24.1.5.2.1.4. Unbind a domain name from a VPC

This topic describes how to unbind a domain name from a VPC.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Find the target domain name and click the number in the **VPCs Associated** column.
4. On the VPCs Associated page, find the target VPC, click the  icon in the Actions column, and then select **Disassociate**.

   Make sure that the unbound VPC is no longer displayed on the VPCs Associated page.

24.1.5.2.1.5. Modify the forwarding configurations of a domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Find the target domain name, click the  icon in the Actions column, and then select **Modify**.
4. In the dialog box that appears, change the value of **Forwarding Mode** or **Forwarder IP Addresses**.
5. Click **OK**.

24.1.5.2.1.6. Add a description for a tenant forwarding domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Find the target domain name, click the  icon in the Actions column, and then select **Description**.
4. In the dialog box that appears, enter a description.
5. Click **OK**.

24.1.5.2.1.7. Delete a tenant forwarding domain name

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Find the target domain name, click the  icon in the Actions column, and then select **Delete**.
4. In the message that appears, click **OK**.

24.1.5.2.1.8. Delete multiple tenant forwarding domain names

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Forwarding Domains**.
3. Select one or more domain names that you want to delete and click **Delete** in the upper-right corner.
4. In the message that appears, click **OK**.

24.1.5.2.2. Tenant default forwarding configurations

24.1.5.2.2.1. View default forwarding configurations

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.

24.1.5.2.2.2. Add a default forwarding configuration

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Click **Add Settings**.
4. In the dialog box that appears, configure parameters such as *Forwarding Mode* and *Forwarder IP Addresses*.

Parameter	Description
Forwarding Mode	For both domain name-based forwarding and default forwarding, the following two forwarding modes are available: ◦ Forward All Requests without Recursion: Only a specified DNS server is used to resolve domain names. If the specified DNS server cannot resolve the domain names, a message is returned to the DNS client to indicate that the query failed. ◦ Forward All Requests with Recursion: A specified DNS server is preferentially used to resolve domain names. If the specified DNS server cannot resolve the domain names, a local DNS server is used instead. If you enter internal IP addresses in the Forwarder IP Addresses field, unexpected results may occur during recursive resolution. For example, a domain name used for internal network services may be resolved to a public IP address.
Forwarder IP Addresses	A list of destination IP addresses. **Note** Multiple IP addresses are separated with semicolons (;).

5. Click **OK**.

24.1.5.2.2.3. Bind an organization to a VPC

Tenant domain names are isolated based on VPCs. You must bind the organization of domain names to a VPC before the DNS forwarding configurations can take effect.

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Find the target organization, click the  icon in the Actions column, and then select **Associate VPCs**.
4. Select one or more VPCs from the list of VPCs to Select, click the right arrow to add them to the list of VPCs Selected, and then click **OK**.

24.1.5.2.2.4. Unbind a domain name from a VPC

This topic describes how to unbind a domain name from a VPC.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Find the target domain name and click the number in the **VPCs Associated** column.
4. On the VPCs Associated page, find the target VPC, click the  icon in the Actions column, and then select **Disassociate**.

   Make sure that the unbound VPC is no longer displayed on the VPCs Associated page.

24.1.5.2.2.5. Modify a default forwarding configuration

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Find the target organization, click the  icon in the Actions column, and then select **Modify**.
4. In the dialog box that appears, change the value of **Forwarding Mode** or **Forwarder IP Addresses**.
5. Click **OK**.

24.1.5.2.2.6. Add a default forwarding configuration

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Find the target organization, click the  icon in the Actions column, and then select **Description**.
4. In the dialog box that appears, enter **Description**.
5. Click **OK**.

24.1.5.2.2.7. Delete a default forwarding configuration

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Find the target organization, click the  icon in the Actions column, and then select **Delete**.
4. In the dialog box that appears, click **OK**.

24.1.5.2.2.8. Delete multiple default forwarding configurations

Prerequisites

You have the permissions of a system administrator or level-1 organization administrator.

Procedure

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, choose **Forwarding Configurations > Tenant Forwarding Settings > Tenant Default Forwarding**.
3. Select one or more domain names that you want to delete and click **Delete** in the upper-right corner.
4. In the message that appears, click **OK**.

24.1.6. Internal Global Traffic Manager (internal GTM Standard Edition only)

Internal Global Traffic Manager (GTM) supports multi-cloud disaster recovery for domain names of customers. This feature manages traffic loads between multiple Apsara Stack networks.

24.1.6.1. Scheduling instance management

24.1.6.1.1. Scheduling Instance

The Scheduling Instance tab displays all existing scheduling instances. You can add, delete, modify, and configure scheduling instances on this tab. When you create a scheduling instance, you must associate an address pool and scheduling domain with the instance.

24.1.6.1.1.1. Create a scheduling instance

After you create a scheduling instance, you can associate the scheduling instance with a scheduling domain and address pool.

1. In the left-side navigation pane, choose **Internal Global Traffic Manager** > **Scheduling Instances** > Scheduling Instance.

2. Click **Create Scheduling Instance** in the upper-right corner of the instance list.

3. In the dialog box that appears, configure Scheduling Instance Name, CNAME Access Domain Name, and Global TTL. Then, click **OK**.

24.1.6.1.1.2. Modify a scheduling instance

1. In the left-side navigation pane, choose **Internal Global Traffic Manager** > **Scheduling Instances** > Scheduling Instance.

2. Find the instance that you want to modify and click **Modify** in the Actions column.

3. Modify the parameter settings as prompted and click **OK**.

24.1.6.1.1.3. Configure a scheduling instance

You can create, delete, modify, and query access policies of scheduling instances.

1. In the left-side navigation pane, click **Internal Global Traffic Manager**. On the **Scheduling Instances** tab of the page that appears, click the **Scheduling Instances** tab.
2. Find the scheduling instance whose access policies you want to view and click **Configure** in the **Actions** column.
3. On the Access Policy Configuration page, view information about all the access policies of the scheduling instance. The information includes **Access Policy Name**, **DNS Request Source**, **Address Type**, **Effective Address Pool**, and **Last Modified At**.
4. Click the closing angle bracket (>) next to an access policy to view the details, including information about the **primary and secondary address pools**.
5. View the setting of **Address Pool Switchover Policy**. The default value is **Automatic**. You can change the value to **Manual**. If Address Pool Switchover Policy is set to Automatic, the system automatically selects an available address pool. If Address Pool Switchover Policy is set to Manual, you must manually specify whether to use the primary address pool or secondary address pool.

> **Note**
>
> Whether an address pool is available is determined based on the number of normal addresses in the address pool and Min. Number of Available Addresses that you specified when you configure the access policy. If the number of normal addresses in the address pool is less than the value of Min. Number of Available Addresses, the address pool is considered unavailable. You can perform a health check to obtain the number of normal addresses in the address pool.

Processing logic for automatic switchover

State of the primary address pool	State of the secondary address pool	Comparison between the numbers of normal addresses in the primary and secondary address pools	Effective address pool (list of available addresses)
Available	Available	-	Primary address pool (including the normal addresses that are intelligently returned and the addresses that are always online. Abnormal addresses that are intelligently returned are deleted, or their weight values are set to 0.)
Available	Unavailable	-	Primary address pool (including the normal addresses that are intelligently returned and the addresses that are always online. Abnormal addresses that are intelligently returned are deleted, or their weight values are set to 0.)
Unavailable	Available	-	Secondary address pool (including the normal addresses that are intelligently returned and the addresses that are always online. Abnormal addresses that are intelligently returned are deleted, or their weight values are set to 0.)
Unavailable	Unavailable	Number of normal addresses in the primary address pool > Number of normal addresses in the secondary address pool	Primary address pool (including the normal addresses that are intelligently returned and the addresses that are always online. Abnormal addresses that are intelligently returned are deleted, or their weight values are set to 0.)
Unavailable	Unavailable	Number of normal addresses in the primary address pool < Number of normal addresses in the secondary address pool	Secondary address pool (including the normal addresses that are intelligently returned and the addresses that are always online. Abnormal addresses that are intelligently returned are deleted, or their weight values are set to 0.)

Unavailable	Unavailable	Number of normal addresses in the primary address pool = Number of normal addresses in the secondary address pool > 0	Primary address pool (including the normal addresses that are intelligently returned and the addresses that are always online. Abnormal addresses that are intelligently returned are deleted, or their weight values are set to 0.)
Unavailable	Unavailable	Number of normal addresses in the primary address pool = Number of normal addresses in the secondary address pool = 0	If the DNS request source is a custom line, the system clears the DNS configurations of the line instead of selecting an address pool. The configurations of the custom line are deleted. If the DNS request source is the global default line, the system selects the primary address pool. The system returns all the configured addresses without considering their return mode.

When the system compares the numbers of normal addresses between the primary and secondary address pools, normal addresses include normal addresses that are intelligently returned and all addresses that are always online (with the health status ignored). The abnormal addresses that are intelligently returned and all addresses that are always offline (with the health status ignored) are not normal addresses.

The following table describes the processing logic for address types if a line is selected in two access policies.

Scenario	Address type for the effective address pool in two access policies		Processing logic
Same DNS request source: Scenario 1	IPv4	IPv6	IPv4 and IPv6 addresses take effect at the same time.
Same DNS request source: Scenario 2	IPv4	Domain name	Addresses of the domain name type take effect.
Same DNS request source: Scenario 3	IPv6	Domain name	Addresses of the domain name type take effect.

Create an access policy for a scheduling instance

You can create multiple access policies to resolve different address pools based on different DNS request sources.

1. Log on to the Apsara Stack DNS console. In the left-side navigation pane, click Internal Global Traffic Manager.

On the Scheduling Instances tab of the Global Traffic Management on the Internal Network page, find the scheduling instance for which you want to create an access policy and click Configure in the Actions column. On the page that appears, click **Create Access Policy**.

2. In the Create Access Policy dialog box, specify **Access Policy Name**, select items in the **DNS Request Source** section, and then configure parameters in the **Primary/Secondary Address Pool Configuration** section. In the DNS Request Source section, if you select Global default, you cannot select other items. The parameters on the Primary Address Pool tab must be configured, and the parameters on the Secondary Address Pool tab can be left empty.
3. In the **Primary/Secondary Address Pool Configuration** section, you can set **Address Type** to IPv4, IPv6, or Domain Name to select different types of address pools. You can also set **Min. Number of Available Addresses**. If the number of healthy addresses in an address pool is less than the value of this parameter, the address pool is determined to be unavailable. The value of Min. Number of Available Addresses must be an integer ranging from 1 to 100.
4. Click **OK**.

The following table describes the limits on address type conflicts between the primary and secondary address pools for two access policies that have the same DNS request source.

Scenario	Address type of the primary address pool (access policy 1)	Address type of the secondary address pool (access policy 2)	Processing logic
Same DNS request source: Scenario 1	IPv4	IPv4	The address pools are allowed to be added.
Same DNS request source: Scenario 2	IPv4	IPv6	The address pools are not allowed to be added.
Same DNS request source: Scenario 3	IPv4	Domain name	The address pools are allowed to be added.
Same DNS request source: Scenario 4	IPv6	IPv4	The address pools are not allowed to be added.
Same DNS request source: Scenario 5	IPv6	IPv6	The address pools are allowed to be added.
Same DNS request source: Scenario 6	IPv6	Domain name	The address pools are allowed to be added.
Same DNS request source: Scenario 7	Domain name	IPv6	The address pools are allowed to be added.
Same DNS request source: Scenario 8	Domain name	IPv4	The address pools are allowed to be added.
Same DNS request source: Scenario 9	Domain name	Domain name	The address pools are allowed to be added.

The following tables describe the limits on address type conflicts between the primary address pools and those between the secondary address pools for two access policies that have the same DNS request source.

Scenario	Address type of the primary address pool (access policy 1)	Address type of the primary address pool (access policy 2)	Processing logic
Same DNS request source: Scenario 1	IPv4	IPv6	The address pools are allowed to be added and they can coexist.

Same DNS request source: Scenario 2	IPv4	IPv4	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 3	IPv4	Domain name	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 4	IPv6	IPv6	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 5	Domain name	IPv6	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 6	Domain name	Domain name	The address pools are allowed to be added and they can coexist. However, the following conditions must be met: (1) The two primary address pools are the same. (2) Both of the secondary address pools exist. In addition, one secondary address pool is of the IPv4 type and the other is of the IPv6 type.

Scenario	Address type of the secondary address pool (access policy 1)	Address type of the secondary address pool (access policy 2)	Processing logic
Same DNS request source: Scenario 1	IPv4	IPv6	The address pools are allowed to be added and they can coexist.
Same DNS request source: Scenario 2	IPv4	IPv4	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 3	IPv4	Domain name	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 4	IPv6	IPv6	The address pools are not allowed to be added and they cannot coexist.
Same DNS request source: Scenario 5	Domain name	IPv6	The address pools are not allowed to be added and they cannot coexist.

Same DNS request source: Scenario 6	Domain name	Domain name	The address pools are allowed to be added and they can coexist. However, the following conditions must be met: (1) The two secondary address pools are the same. (2) Both of the two primary address pools exist. In addition, one primary address pool is of the IPv4 type and the other is of the IPv6 type.

Modify the access policy of a scheduling instance

> **Notice**
>
> If you do not change the primary or secondary address pool when you modify an access policy, no primary/secondary switchover is triggered. If you change one of the address pools, the system complies with the following processing rules:
>
> 1. Manual switchover mode: If the secondary address pool is deleted, the system forcibly switches services to the primary address pool. If the secondary address pool is not deleted, the effective address pool does not change.
> 2. Automatic switchover mode: The system determines the address pool that takes effect based on the status of the newly selected address pools. Exercise caution when you change the address pools.

1. On the Access Policy Configuration page, find the access policy that you want to modify and click **Modify** in the Actions column.
2. In the Modify Access Policy dialog box, modify **Access Policy Name**, select items in the **DNS Request Source** section, and then configure parameters in the **Primary/Secondary Address Pool Configuration** section. In the DNS Request Source section, if you select Global default, you cannot select other items. The parameters on the Primary Address Pool tab must be configured, and the parameters on the Secondary Address Pool tab can be left empty.
3. In the Primary/Secondary Address Pool Configuration section, you can set **Address Type** to IPv4, IPv6, or Domain Name to select address pools of different types. You can also set Min. Number of Available Addresses. If the number of healthy addresses in an address pool is less than the value of this parameter, the address pool is determined to be unavailable. The value of this parameter must be an integer ranging from 1 to 100.
4. Click **OK**.

Delete the access policy of a scheduling instance

1. On the Access Policy Configuration page, find the target access policy and click Delete in the Actions column.

2. In the dialog box that appears, click OK after you verify that the displayed information is correct.

24.1.6.1.1.4. Delete a scheduling instance

1. Log on to the Apsara Stack DNS console and choose Recursion Configurations > Scheduling Instances > Scheduling Instance.

2. Find the target instance and click Delete in the Actions column.

3. In the dialog box that appears, click OK.

Note: After you delete the instance, its configuration data is also deleted.

24.1.6.1.2. Address Pool

The Address Pool tab allows you to manage address pools. You can associate address pools with scheduling instances. The address pools are classified into three types: IPv4 address pools, IPv6 address pools, and domain name address pools. The load balancing policy of an address pool can be set to polling or weight.

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, click **Internal Global Traffic Manager**. On the **Scheduling Instance** tab of the Global Traffic Management on the Internal Network page, click the **Address Pool** tab.
3. Find the address pool whose information you want to view and click the closing angle bracket (>) next to the name of the address pool to view its detailed information. The information includes Address Pool ID, Address Pool Name, Address Type, Load Balancing Policy (Among Addresses), Created At, Last Modified At, Health Check, and Health Check Status.

24.1.6.1.2.1. Create an address pool

You can define a list of addresses that form an address pool, which can be associated with access policies of scheduling instances when you configure the access policies.

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, click **Internal Global Traffic Manager**. On the Scheduling Instance tab of the Global Traffic Management on the Internal Network page, click the Address Pool tab.
3. On the Address Pool tab, click **Create Address Pool**.
4. In the Create Address Pool dialog box, specify **Address Pool Name**, **Address Type**, and **Load Balancing Policy (Among Addresses)**, and add addresses one by one in the **Address List** section. You can also click **Batch Add** to add multiple addresses at a time. After you enter the required information, click **OK**.

Parameter	Description
Address Pool Name	The name can contain a maximum of 20 characters.
Address Type	You can select IPv4, IPv6, or Domain Name from the drop-down list of this parameter. This configuration cannot be changed.
Load Balancing Policy (Among Addresses)	You can select Polling or Weight from the drop-down list of this parameter. This configuration cannot be changed.
Mode	Valid values: • Automatically Returned: The system determines whether the address is available based on the health check result of the address. • Always online: The system ignores the health check result of the address and sets the address to be always available. The health check task is still running. • Always Offline: The system ignores the health check result of the address and sets the address to be always unavailable. The health check task is still running.

24.1.6.1.2.2. Modify the configurations of an address pool

Notice After you modify the configurations of an address pool, the health check results of all addresses are reset to normal if health check is enabled. If the address pool has been associated with access policies and automatic switchover is enabled, a primary/secondary switchover may be triggered. Proceed with caution.

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, click **Internal Global Traffic Manager**. On the **Scheduling Instance** tab of the Global Traffic Management on the Internal Network page, click the **Address Pool** tab.

3. Find the address pool that you want to modify and click **Modify** in the Actions column.
4. In the Modify Address Pool dialog box, you can change only **Address Pool Name** and the addresses in **Address (One in Each Row)**.
5. Click **OK**.

24.1.6.1.2.3. Delete an address pool

1. Log on to the Apsara Stack DNS console and choose Recursion Configurations > Scheduling Instances > Address Pool.

2. Find the target address pool and click Delete in the Actions column.

3. In the dialog box that appears, click OK after you verify that the displayed information is correct.

24.1.6.1.2.4. Enable health check

You can enable health check to check the status of the addresses in an address pool. Only the addresses whose health check status is normal can be returned.

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, click **Internal Global Traffic Manager**. On the **Scheduling Instances** tab of the Global Traffic Management on the Internal Network page, click the **Address Pool** tab.
3. Find the address pool for which you want to configure a health check and click **Health Check** in the Actions column.
4. In the Health Check dialog box, turn on or turn off **Health Check Switch**, specify the parameters in the **Protocol Settings**, **Health Check Settings**, and **Node Settings** sections, and then click **OK**.

Section	Parameter	Description	Protocol
Protocol Settings	Port	The port number that is used for health checks on the destination address. The value must be an integer in the range of 1 to 65535. This parameter cannot be empty.	HTTP, HTTPS, TCP, or UDP
	Path	The HTTP or HTTPS path that is used for health checks on the destination address. This path is used to check whether the HTTP or HTTPS service of the destination address is normal. If the HTTP status code returned from this path is 2xx or 3xx, the HTTP or HTTPS service is normal. The system automatically adds a forward slash (/) before the path name. The path can be empty. The default value is /. The path name can be up to 255 characters in length.	HTTP or HTTPS
	Host Configuration	The host configuration that is used for health checks. If you do not specify this parameter, the primary domain name is used.	HTTP or HTTPS
	Returned Error Code Greater Than or Equal to	The minimum value of the HTTP status code when the health check result is abnormal. The HTTP status code returned must be greater than or equal to the value of this parameter.	HTTP or HTTPS

Section	Parameter	Description	Protocol
	ICMP Packages Sent	The number of Internet Control Message Protocol (ICMP) packets sent each time an ICMP health check is performed.	ICMP
	Packet Loss Rate	The threshold of the packet loss rate. The threshold is used to determine whether the result of an ICMP health check is abnormal.	ICMP
Health Check Settings	Check Interval	The time interval at which health checks are performed on the destination address.	
	Timeout Duration	The timeout duration for which the system waits after an exception occurs during a health check.	
	Failure Threshold	The minimum number of consecutive health check failures when the status of the destination address is abnormal during a health check.	
Node Settings	Physical Network	The nodes that initiate health checks on the destination address.	
	VPC		

24.1.6.1.3. Scheduling Domain

The Scheduling Domain tab allows you to add, delete, and query scheduling domains.

You can log on to the Apsara Stack DNS console and choose Internal Global Traffic Manager > Scheduling Instances > Scheduling Domain to go to the scheduling domain list.

24.1.6.1.3.1. Create a scheduling domain

1. Log on to the Apsara Stack DNS console and choose Recursion Configurations > Scheduling Instances > Scheduling Domain. Then, click Create Scheduling Domain in the upper-right corner of the scheduling domain list.

2. In the dialog box that appears, enter the custom domain name and click OK.

24.1.6.1.3.2. Add a description for a scheduling domain

1. In the left-side navigation pane, choose Internal Global Traffic Manager > Scheduling Instances > Scheduling Domain.

2. Find the scheduling domain for which you want to add a description and click Edit in the Actions column.

3. In the dialog box that appears, add a description in the Edit field and click OK.

24.1.6.1.3.3. Delete a scheduling domain

1. In the left-side navigation pane, choose Internal Global Traffic Manager > Scheduling Instances > Scheduling Domain.

2. Find the scheduling domain that you want to delete and click Delete in the Actions column.

3. In the message that appears, click OK after you verify that the displayed information is correct.

24.1.6.1.4. View alert logs

> Note
>
> By default, the system saves alert logs of the last 90 days.

1. Log on to the Apsara Stack DNS console.
2. In the left-side navigation pane, click **Internal Global Traffic Manager**. On the **Scheduling Instance** tab of the Global Traffic Management on the Internal Network page, click the **Alert Logs** tab.
3. In the alert log list, view address pool information, such as the status of the address pool, health check results of addresses, and switchover between primary and secondary address pools. You can query alert logs of address pools by time or behavior, or by using a keyword.

24.1.6.2. Scheduling line management

24.1.6.2.1. IP Address Line Configuration

The IP Address Line Configuration tab allows you to define lines based on IP addresses. The lines are used to group request sources to achieve intelligent load balancing.

24.1.6.2.1.1. Add a line

1. In the left-side navigation pane, choose **Internal Global Traffic Manager** > **Scheduling Line Management** > IP Address Line Configuration.

2. Click **Add Line** in the upper-right corner of the line list.

3. In the dialog box that appears, configure the parameters as prompted and click **OK**.

24.1.6.2.1.2. Sort lines

1. In the left-side navigation pane, choose **Internal Global Traffic Manager** > **Scheduling Line Management** > IP Address Line Configuration.

2. Find the line whose sequence you want to change and click **Sort** in the Actions column.

3. Specify Sort Behavior as prompted and click **OK**.

24.1.6.2.1.3. Modify the configurations of a line

1. In the left-side navigation pane, choose **Internal Global Traffic Manager** > **Scheduling Line Management** > IP Address Line Configuration.

2. Find the line whose configurations you want to modify and click **Modify** in the Actions column.

3. Modify the configurations as prompted and click **OK**.

24.1.6.2.1.4. Delete a line

1. In the left-side navigation pane, choose **Internal Global Traffic Manager** > **Scheduling Line Management** > **IP Address Line Configuration**.

2. Find the line that you want to delete and click **Delete** in the Actions column.

3. In the message that appears, click **OK**.

24.1.6.3. System management

The system management feature allows you to manage Global Traffic Manager (GTM) clusters and system configurations.

24.1.6.3.1. GTM cluster management

On the GTM Cluster Management tab, you can perform the following operations on Global Traffic Manager (GTM) clusters: **Set Emergency Group** and **Merge GTM Control Domain**.

To go to the **GTM Cluster Management** tab, perform the following steps: Log on to the Apsara Stack DNS console. In the left-side navigation pane, click Internal Global Traffic Manager. On the page that appears, click the System Management tab. The GTM Cluster Management tab appears.

Set Emergency Group

You can select some available service instances to form a temporary cluster to provide services.

- Enable the emergency group feature: If a GTM cluster becomes abnormal, click **Set Emergency Group**. In the Set Emergency Group dialog box, turn on Emergency Group Switch, select available service instances to form an emergency group, and then click **OK**.
- Disable the emergency group feature: If a GTM cluster has resumed normal operations, click **Set Emergency Group**. In the Set Emergency Group dialog box, turn off Emergency Group Switch and click **OK**.

Merge GTM Control Domain

In multi-cloud scenarios, you can click **Merge GTM Control Domain** and enter the IP address of the leader service instance of the merged GTM control domain to form a large GTM cluster.

View the service instances in a GTM cluster

You can view the following information of the service instances in the current GTM cluster:

Instance IP Address, Instance Role, Working Mode, Status, Latest Synchronization Log ID, IP Address, and Instance Description.

You can also perform the following steps to switch the role of a service instance in the GTM cluster from follower to leader:

1. Find the service instance with the follower role and click **Switch Primary Node** in the Actions column.
2. In the message that appears, click **OK**.