

阿里云 专有云敏捷版

产品简介

产品版本：2.1.0

文档版本：20190227

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

表 -1: 格式约定

格式	说明	样例
	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	 禁止： 重置操作将丢失用户配置数据。
	该类警示信息可能导致系统重大变更甚至故障，或者导致人身伤害等结果。	 警告： 重启操作将导致业务中断，恢复业务所需时间约10分钟。
	用于警示信息、补充说明等，是用户必须了解的内容。	 注意： 导出的数据中包含敏感信息，请妥善保管。
	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	 说明： 您也可以通过按 Ctrl + A 选中全部文件。
>	多级菜单递进。	设置 > 网络 > 设置网络类型
粗体	表示按键、菜单、页面名称等UI元素。	单击 确定 。
courier 字体	命令。	执行 <code>cd /d C:/windows</code> 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	<code>bae log list --instanceid Instance_ID</code>
[]或者[a b]	表示可选项，至多选择一个。	<code>ipconfig [-all] -t</code>
{ }或者{a b}	表示必选项，至多选择一个。	<code>swich {stand slave}</code>

目录

法律声明	1
通用约定	1
1 阿里云专有云平台简介	1
1.1 什么是专有云敏捷版	1
1.2 为什么选择专有云敏捷版	3
1.2.1 统一的分布式云操作系统	3
1.2.2 天基部署与管控系统	5
1.2.3 统一运维管理及自动化运维能力	5
1.2.4 开放的云服务接口	6
1.3 产品架构	6
1.3.1 专有云架构类型	6
1.3.2 系统架构	6
1.3.3 网络架构	8
1.3.4 安全架构	10
1.3.5 底座组件	11
1.4 产品全景	12
1.5 应用场景	13
2 容器服务	14
2.1 什么是容器服务	14
2.2 产品优势	14
2.3 产品架构	14
2.4 功能特性	15
2.5 应用场景	17
2.5.1 容器化应用上云	17
2.5.2 容器化DevOps	18
2.5.3 微服务支持	18
2.5.4 弹性伸缩	19
2.5.5 混合云	19
2.6 使用限制	20
2.7 基本概念	20
3 对象存储OSS	24
3.1 什么是对象存储OSS	24
3.2 产品优势	24
3.3 产品架构	25
3.4 功能特性	26
3.5 应用场景	28

3.6 使用限制.....	29
3.7 基本概念.....	30
4 表格存储Table Store.....	33
4.1 什么是表格存储.....	33
4.2 产品优势.....	33
4.3 产品架构.....	34
4.4 功能特性.....	36
4.5 应用场景.....	38
4.6 使用限制.....	43
4.7 基本概念.....	44
5 文件存储NAS.....	47
5.1 什么是文件存储NAS.....	47
5.2 功能特性.....	47
5.3 产品优势.....	47
5.4 应用场景.....	48
5.5 基本概念.....	48
6 块存储.....	50
6.1 什么是弹性块存储ABS.....	50
6.2 产品架构.....	50
6.2.1 技术架构.....	50
6.2.2 部署架构.....	52
6.3 功能特性.....	53
6.3.1 概述.....	53
6.3.2 分布式文件系统.....	53
6.3.3 自动精简装置.....	54
6.3.4 磁盘快照功能.....	55
6.3.5 数据重建.....	55
6.4 典型业务场景.....	56
6.4.1 概述.....	56
6.4.2 OpenStack场景.....	56
6.4.3 Docker场景.....	57
7 云数据库RDS版.....	58
7.1 什么是关系型数据库RDS.....	58
7.2 产品优势.....	58
7.2.1 易于使用.....	58
7.2.2 高性能.....	59
7.2.3 高安全性.....	59
7.2.4 高可靠性.....	60
7.3 产品架构.....	61
7.4 功能特性.....	61

7.4.1 数据链路服务.....	61
7.4.2 高可用服务.....	62
7.4.3 备份恢复服务.....	64
7.4.4 监控服务.....	64
7.4.5 调度服务.....	65
7.4.6 迁移服务.....	65
7.5 应用场景.....	66
7.5.1 数据多样化存储.....	67
7.5.2 读写分离.....	68
7.5.3 大数据分析.....	69
7.6 使用须知.....	69
7.7 基本概念.....	70
8 数据管理DMS.....	72
8.1 什么是数据管理DMS企业版.....	72
8.2 产品优势.....	72
8.3 产品架构.....	73
8.4 功能特性.....	73
8.4.1 实例管理.....	73
8.4.2 用户管理.....	74
8.4.3 任务管理.....	74
8.4.4 安全规则.....	74
8.4.5 配置管理.....	74
8.4.6 操作日志.....	75
8.4.7 访问IP白名单.....	75
8.4.8 结构设计配置.....	75
8.4.9 权限申请.....	75
8.4.10 数据变更.....	75
8.4.11 数据导出.....	76
8.4.12 库表同步.....	76
8.4.13 结构设计.....	76
8.4.14 SQL审核.....	76
8.4.15 工作台.....	76
8.4.16 顶栏搜索.....	77
8.4.17 SQLConsole.....	77
8.5 基本概念.....	77
9 数据传输服务DTS.....	86
9.1 什么是数据传输服务DTS.....	86
9.2 产品优势.....	86
9.3 产品架构.....	87
9.4 功能特性.....	92

9.4.1 数据迁移.....	92
9.4.2 数据同步.....	94
9.4.3 数据订阅.....	97
9.5 应用场景.....	98
9.6 基本概念.....	104
10 云数据库KVStore for Redis.....	107
10.1 什么是云数据库KVStore for Redis.....	107
10.2 产品优势.....	107
10.3 产品架构.....	108
10.4 功能特性.....	109
10.5 应用场景.....	110
10.6 使用限制.....	111
10.7 基本概念.....	112
10.8 实例规格.....	113
11 云数据库MongoDB.....	116
11.1 什么是云数据库MongoDB.....	116
11.2 产品优势.....	116
11.3 产品架构.....	117
11.4 功能特性.....	118
11.5 应用场景.....	119
11.6 使用限制.....	120
11.7 基本概念.....	120
11.8 实例规格.....	121
12 分布式关系型数据库DRDS.....	122
12.1 什么是分布式关系型数据库服务DRDS.....	122
12.2 产品优势.....	123
12.3 产品架构.....	123
12.4 功能特性.....	126
12.5 应用场景.....	127
12.6 使用限制.....	127
12.7 基本概念.....	128
13 云盾.....	131
13.1 什么是云盾.....	131
13.2 产品优势.....	131
13.3 产品架构.....	133
13.4 功能特性.....	134
13.5 使用限制.....	134
14 密钥管理服务KMS.....	135
14.1 什么是密钥管理服务.....	135

14.2 产品优势.....	136
14.3 产品架构.....	136
14.4 功能特性.....	137
14.5 应用场景.....	138
14.6 使用限制.....	141
14.7 基本概念.....	142
15 ZStack.....	143
15.1 产品概述.....	143
15.2 产品功能.....	144
15.2.1 专有云功能.....	145
15.3 产品优势.....	158
15.3.1 专有云优势.....	159
15.4 应用场景.....	160
16 消息队列MQ.....	161
16.1 什么是消息队列MQ.....	161
16.2 产品版本对比.....	161
16.3 产品优势.....	166
16.4 产品架构.....	167
16.5 功能特性.....	169
16.6 应用场景.....	170
16.7 使用限制.....	174
16.8 基本概念.....	175

1 阿里云专有云平台简介

1.1 什么是专有云敏捷版

专有云定义

专有云是云计算服务提供商为企业在其内部建设的专有云计算系统。专有云将云基础设施与硬件资源建立在防火墙内，以供机构或企业内各部门共享数据中心内的资源。管理者可能是组织本身，也可能是第三方；位置可能在组织内部，也可能在组织外部。专有云与公共云相比，具备更好的私密性和专属性。

按服务企业规模或业务需求的大小可以划分为两类：

- 面向行业云、大型集团云等的多租户、完整专有云架构，以超大规模数字化应用为业务驱动，以DevOps应用持续集成持续开发和生产环境的运营支撑为主要IT诉求，自上而下构建全栈云服务的专有云体系。
- 面向中小企业和中小型应用场景的单租户、简版专有云，向上衔接和承接大型SaaS应用、行业云及大型集团云等技术体系，同时完成本地计算任务。

阿里云专有云

过去，在企业IT架构实现向云演进的过程中，越来越多的企业出于安全合规、已有数据中心利旧、本地化体验等自身的建设要求，希望在自己的数据中心内也可以获得大规模云计算带来的服务体验。

阿里云专有云解决方案诞生自阿里云公共云，将公共云的技术带到专有云领域，即专有公共云，简称专有云。通过帮助企业在自己的数据中心交付完整的可定制的阿里云软件解决方案，让用户可以在本地运行同阿里云公共云提供的超大规模云计算和大数据产品相同的特性，为企业提供一致性的混合云体验，从而满足用户按需获得IT资源，保持业务持续性的要求。

专有云敏捷版

在专有云市场上，中小规模专有云占据了大部分专有云市场需求，典型用户的部署规模偏小。针对不同企业级市场的使用特点，为客户提供一个开放、统一、可信的企业级云平台，增强在云市场上的核心竞争力，满足更广泛的业务需求，阿里云专有云推出面向企业中小规模专有云场景的敏捷云应用平台。

专有云敏捷版可以直接部署在企业已有的x86等硬件基础之上并加以管理，提供真正的企业级安全和稳定服务保障。专有云敏捷版采用了底座、飞天、网络 and 存储组件混合部署的新技术，因此，减

少了物理机的需求数量，有效地提升了资源利用率，并且可扩展、可扩容。专有云敏捷版虽然缩小了管控，但仍然可以保障高可用及数据安全，同时也极大地降低了购买成本。

服务价值

阿里云专有云以丰富的产品和服务为依托，以阿里巴巴集团成功的数字化实践案例为基础，结合已在各个行业形成的成熟的解决方案和丰富的经验，能够帮助政府及企业级用户完成数字化转型。使用阿里云专有云计算所带来的服务价值也逐步体现在以下四个方面：

- **弹性**

资源打通，将所有资源变成一台超级计算机，通过弹性扩容，使得成本、性能及稳定性达到最优。

- **敏捷**

业务打通，通过互联网化、微服务化，让传统企业快速实现创新。

- **数据**

数据打通，通过数据化打通不同垂直业务的数据，形成数据中台，应对海量数据处理。

- **智能**

商业打通，通过全局智能化重塑商业模式。

平台特征

阿里云专有云作为企业级的云平台，它具备如下三个特性：

- 软件定义平台，屏蔽底层硬件差异，资源可根据需求进行横向或纵向的扩展，对上层应用无感知。
- 生产级的可靠性及安全合规，确保企业数据的连续性和安全性。
- 统一管理入口，对不同人员的角色进行权限隔离，便于用户后期的运维管理。

发展历程

阿里云专有云敏捷版从2017年11月发布V1.0版本后，到目前为止历经了两个版本的演进，未来阿里云专有云敏捷版会向着更开放、更可信、更可控的企业级云平台发展，持续为用户提供价值服务。

表 1-1: 专有云敏捷版发展历程

版本	发布时间	相关内容
V1.0	2017年12月	• 场景：存储场景

版本	发布时间	相关内容
		• 关键产品：块存储ABS，对象存储OSS，文件存储NAS，表格存储。 • 平台特性：提供高可用、高吞吐量、低时延和良好扩展性的存储服务小型化专有云。
V1.1	2018年9月	• 场景：互联网中台+数据库 • 关键产品：IAAS+数据库产品+中间件（含docker）+云盾 • 平台特性：统一运维管理，管控节点压缩，输出的服务器体量得到有效控制。

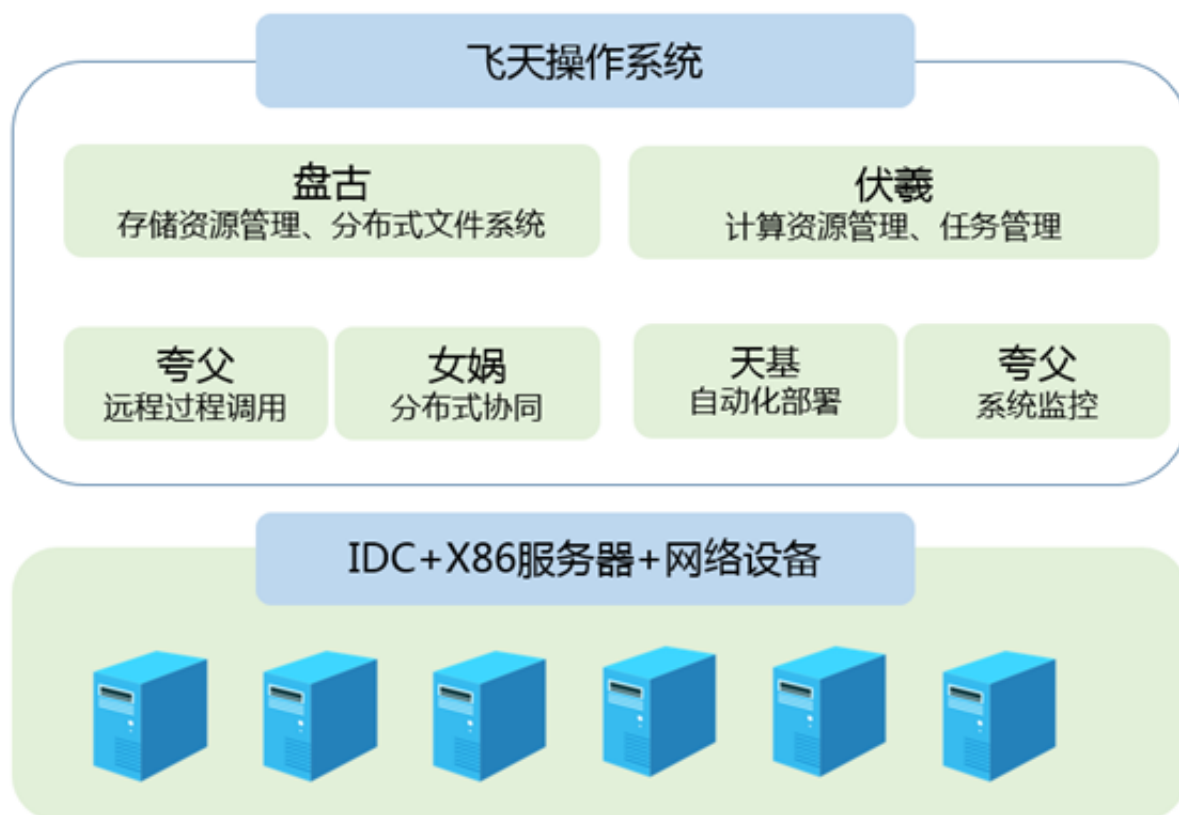
1.2 为什么选择专有云敏捷版

本节主要介绍专有云敏捷版的产品优势。

1.2.1 统一的分布式云操作系统

专有云敏捷版与公共云使用同一套底层架构——飞天大规模分布式计算系统内核，为上层的提供服务提供存储、计算和调度等方面的底层支持。它是由阿里云自主研发、服务全球的超大规模通用计算操作系统。飞天可以将遍布全球的百万级服务器连成一台超级计算机，以在线公共服务的方式为社会提供计算能力。飞天能够提供足够强大的计算能力，通用的计算能力，普惠的计算能力。

图 1-1: 飞天内核系统架构图



飞天平台内核包含的模块覆盖了以下主要的功能：

- **分布式系统底层服务**

提供分布式环境下所需要的协调服务、远程过程调用、安全管理和资源管理的服务。这些底层服务为上层的分布式文件系统、任务调度等模块提供支持。

- **分布式文件系统**

提供一个海量的、可靠的、可扩展的数据存储服务，将集群中各个节点的存储能力聚集起来，并能够自动屏蔽软硬件故障，为您提供不间断的数据访问服务；支持增量扩容和数据的自动平衡，提供类似于POSIX的用户空间文件访问API，支持随机读写和追加写的操作。

- **任务调度**

为集群系统中的任务提供调度服务，同时支持强调响应速度的在线服务和强调处理数据吞吐量的离线任务；自动检测系统中故障和热点，通过错误重试、针对长尾作业并发备份作业等方式，保证作业稳定可靠地完成。

- **集群监控和部署**

对集群的状态和上层应用服务的运行状态和性能指标进行监控，对异常事件产生警报和记录；为运维人员提供整个飞天平台以及上层应用的部署和配置管理，支持在线集群扩容、缩容和应用服务的在线升级。

1.2.2 天基部署与管控系统

天基提供了云服务产品的统一部署、验证、授权和管控能力，为云服务提供基础性的支撑。天基框架中包含了部署框架、资源库、元数据库、云盾、认证授权、接口网管、日志服务、管控服务等模块。

- 部署框架为所有的云服务提供了统一的接入平台部署和服务间的依赖关系管理功能。
- 资源库保存了所有云服务和依赖组件的执行文件。
- 云盾为云服务提供Web攻击防护功能。
- 认证授权组件为云服务提供访问控制能力。
- 接口网关为云服务提供统一的API管理平台。
- 日志服务为云服务提供了日志存储、检索、获取等功能。
- 管控模块监控各云服务的基础健康状态，支撑云平台的运维体系。

1.2.3 统一运维管理及自动化运维能力

专有云敏捷版提供统一的运维管理系统入口，可以针对不同的用户角色配置不同的管理权限。通过开放的OpenAPI，可以将管理运维的能力开放给用户，用户可以自定义云资源控制台。为了与企业现有IT系统完成同步的对接和整合，还提供了企业ITIL系统对接的能力。

图 1-2: 统一管理运维示意图



1.2.4 开放的云服务接口

云服务通过OpenAPI平台，提供丰富的SDK包和RESTful API接口。用户可以使用开放接口来灵活访问专有云提供的各种云服务。您还可以通过OpenAPI获取云平台的基础管控信息，将专有云平台接入到用户统一的管控系统。

1.3 产品架构

本节主要介绍专有云敏捷版的系统架构、网络架构、安全架构及底座组件。

1.3.1 专有云架构类型

专有云架构可以分为原生云架构和集成云架构两类：

- **原生云架构**

由互联网的开放架构演进而来，以分布式系统框架为基础，大数据和Web应用先行，然后扩展到各种基础服务。

- **集成云架构**

以计算虚拟化为主，基于传统的架构做突破，目前已被OpenStack开源逐步成为主流。

专有云采用原生云架构，以自研分布式技术和产品为基础，一套体系支撑所有云产品和服务，提供完整的云平台开放能力，具备完善企业级服务特性，具备完善的容灾和备份能力，具备完全自主可控的能力。

1.3.2 系统架构

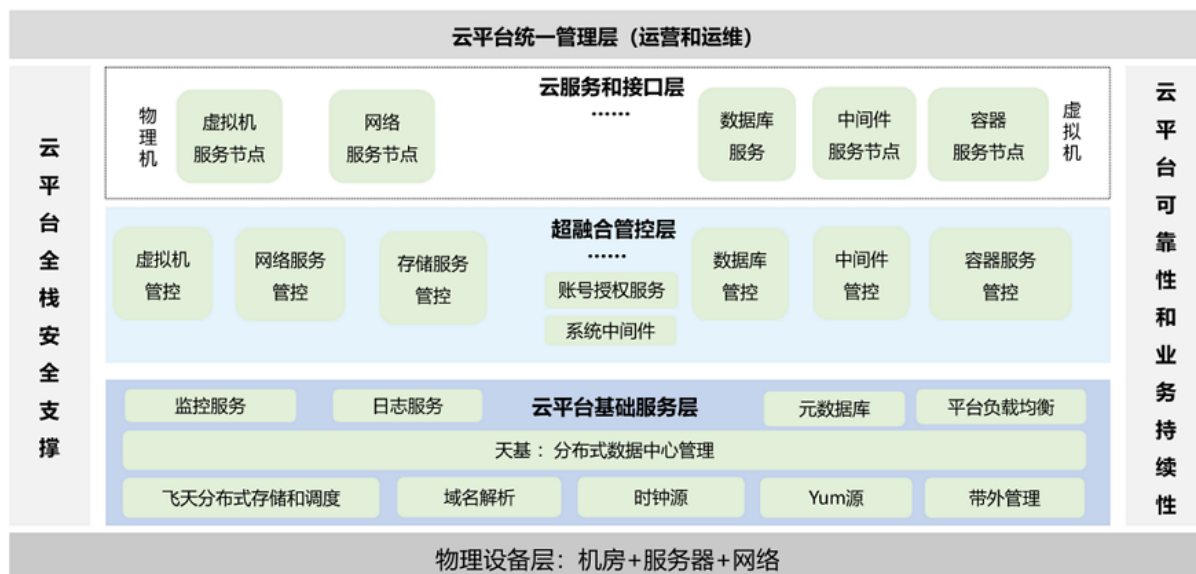
专有云敏捷版基于开放API模型，为您提供企业级云安全架构，与统一的管理运维体验。

专有云敏捷版系统架构主要分为如下几个部分，如图 1-3: 专有云系统架构图示，其中

- 物理设备层，主要包括用于云计算的物理机房、服务器、网络等硬件设备。
- 云平台基础服务层，在基础物理环境之上，为上层应用提供基础服务。
- 超融合管控层，利用超融合的管控架构，为上层应用或服务提供统一的调度。
- 云服务与接口层，一是通过融合的服务节点管理，对虚拟机和物理机提供统一管理和运维，二是通过开放的API管理平台，统一接口并支持定制化开发。
- 云平台统一管理层，提供统一的运营和运维管理入口。

此外，专有云敏捷版还提供了全栈的安全支撑，云平台可靠性和业务持续性的保障。

图 1-3: 专有云系统架构图

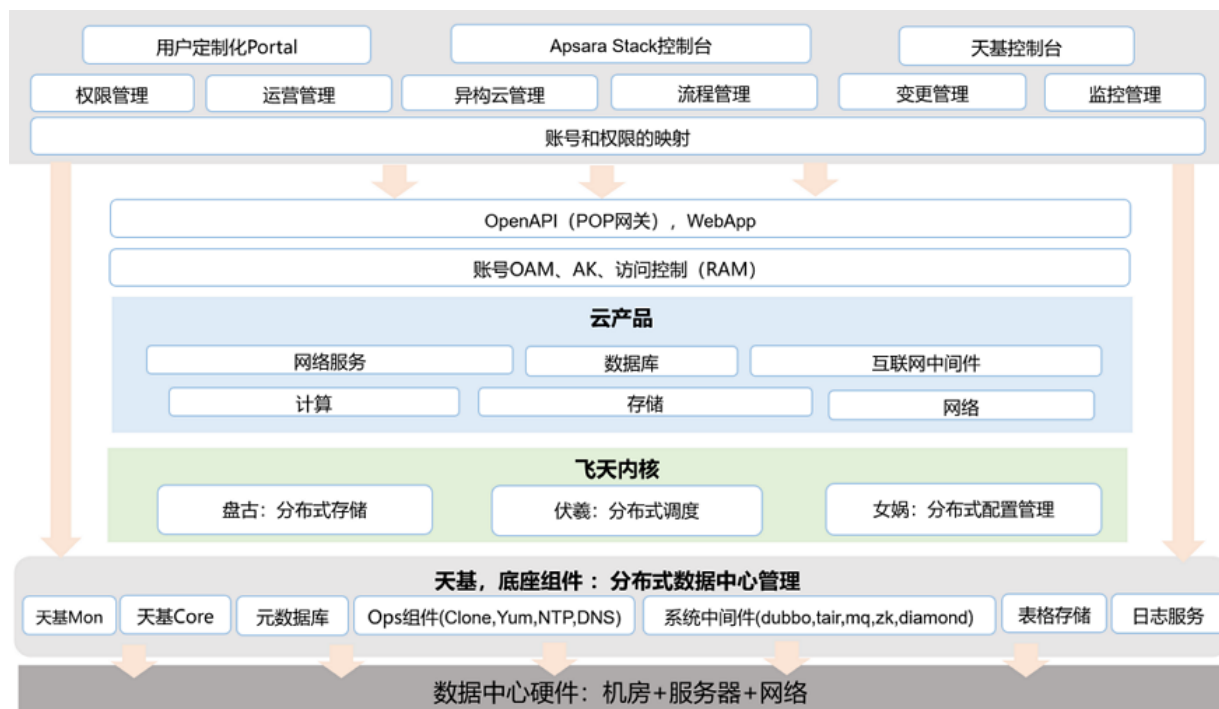


逻辑架构

专有云敏捷版通过将物理服务器的计算和存储能力以及网络设备虚拟化成虚拟计算、分布式存储和软件定义网络，并在此基础上提供云数据库、分布式中间件服务，为用户的应用系统提供IT基础服务的支撑能力，同时可以和用户现有的账号体系，监控运维系统进行对接。专有云逻辑架构有如下特点：

- 硬件基础是IDC+X86服务器+网络设备；
- 飞天内核（分布式引擎），基于飞天提供了各种云产品；
- 所有云产品都要求遵从一个统一的API框架，管理与运维（账号、授权、监控、日志）体系及安全体系；
- 保证所有产品遵从一致性的使用体验。

图 1-4: 专有云逻辑架构图



1.3.3 网络架构

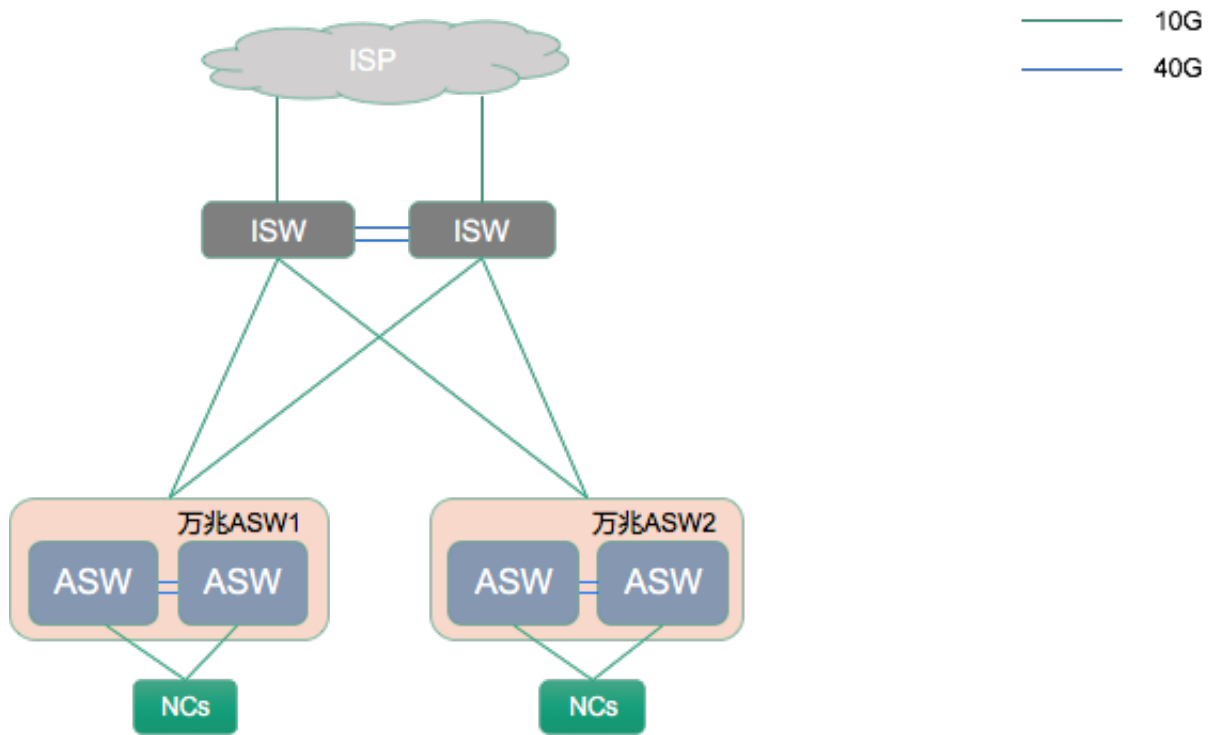
专有云敏捷版是Apsara Stack的最小化产品输出，网络架构经过优化和精简，仅包含ISW和ASW两种设备角色，最大支持96台服务器规模，采用MiniLVS代替SLB，支持MiniLVS扩展BGP协议。

各层交换机的角色和作用如下表所示。

表 1-2: 角色定义

角色名称	作用
ISW	互联交换机。对外互联ISP，对内互联ASW。
ASW	接入交换机。接入云服务器，上行互联ISW。
OOB	带外管理（Out of Band）。
OMR	带外管理核心交换机。
OASW	带外管理接入交换机，互联服务器IPMI控制卡。
ACS	串口服务器，互连网络设备console口，用于设备管理。

图 1-5: 网络架构逻辑分区

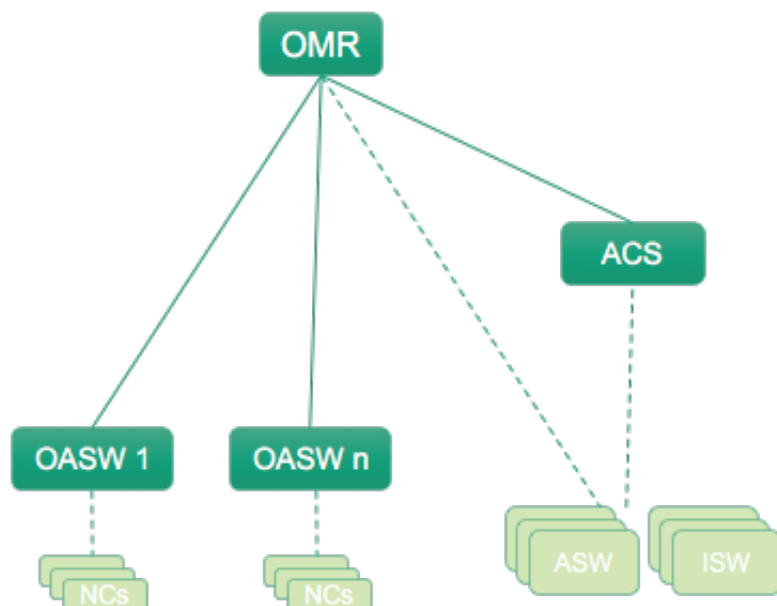


专有云Agility版网络架构中ISW对外互联ISP运营商（外网带宽取决于用户实际需求），对内互联ASW接入交换机。ISW固定输出2台，ISW之间互联带宽为2*40G，下行互联每组ASW带宽为320G。

专有云Agility版网络架构中ASW为接入交换机，互联服务器，为提供所有云业务的网络承载。ASW采用堆叠设计，两台为一组，敏捷版最多支持2组ASW，最大接入服务器能力为96台，每组ASW互联ISW带宽为320G，互联服务器带宽为960G，网络收敛比为1:3。

所有服务器配置双网卡，通过bond双链路上行ASW交换机，单台服务器出口达到20G。

图 1-6: 带外管理网



专有云Agility版OOB管理网提供集群内部服务器及交换机的管理能力，是天基系统对物理服务器进行装机、重启等操作的必要途径。

服务器管理网络

每台服务器的IPMI口（控制卡管理口）上联GE网线至OASW交换机。每台OASW交换机配置48千兆网口，OASW交换机二层透传，上联至OMR交换机，网关配置在OMR交换机。

网络设备管理网络

每台网络设备MGMT口（management管理口）上联GE网线直接连至OMR交换机。

网络设备Console口网络

每台网络设备Console口上联GE网线至ACS，ACS上联至OMR交换机。

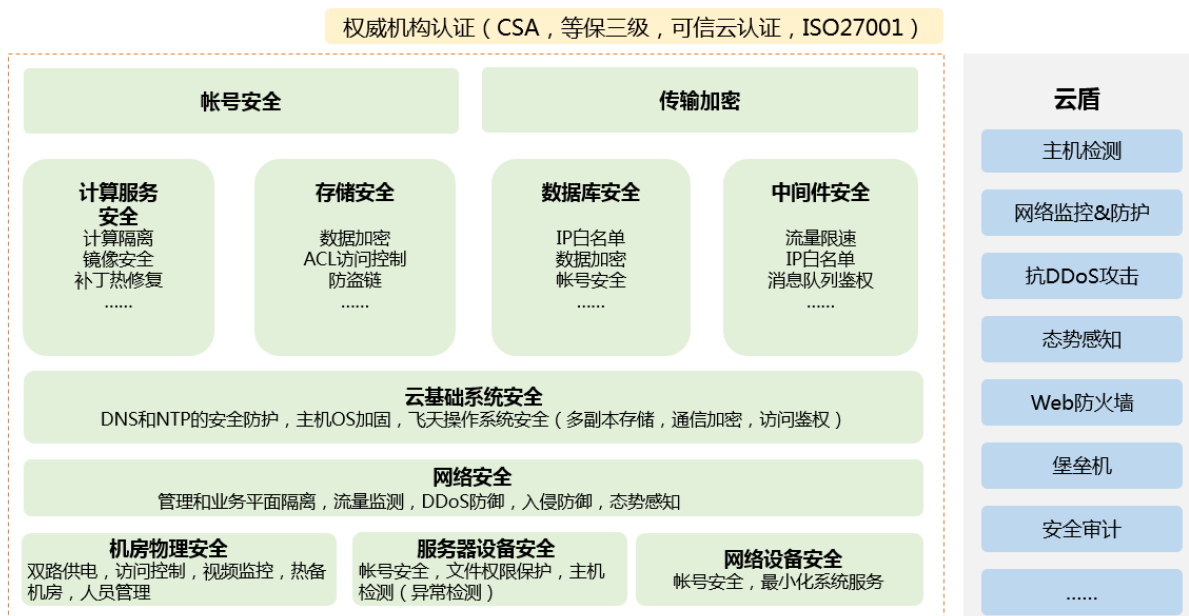
1.3.4 安全架构

云产品本身既有前台服务，又有后台系统，所以阿里云专有云安全架构分为两层：平台层和用户层。

专有云敏捷版提供从底层通信协议到上层应用的全方位的安全能力，保证用户的访问和数据安全。所有控制台均需要通过HTTPS证书的方式访问。专有云提供完善的角色授权机制，保证多租户模式下，资源访问的安全可控。支持不同的安全角色，包括安全管理员、系统管理员、安全审计员。

此外，专有云敏捷版支持云盾产品，为用户提供多层面、一体化的云安全防护解决方案。

图 1-7: 阿里云专有云分层安全框架



1.3.5 底座组件

专有云敏捷版底座有三类组件，共同为云平台的部署和运维提供支撑。

表 1-3: 底座组件

组件		功能说明
Ops组件	Yum	安装软件包 软件源在初始装机阶段已经部署完毕，主要用于物理机上安装操作系统、部署飞天等专有云的应用软件包及其依赖的组件。
	Clone	机器克隆服务
	NTP	时钟源服务 部署在专有云的物理机，从标准NTP时间源同步，并授时给其他宿主机。
	DNS	域名解析服务 为专有云内部环境提供域名的正解和反解服务，在两个 OPS 机器上各运行一个 bind 实例，通过keepalived 提供高可用服务，在一台失效的情况下，另外一台能够主动接管。
底座中间件	dubbo	分布式RPC服务
	tair	缓存服务

组件		功能说明
	mq	消息队列服务
	ZooKeeper	分布式协同
	Diamond	配置管理服务
	SchedulerX	定时任务服务
底座基础组件	天基	数据中心管理
	天基Mon	数据中心监控
	OTS-inner	表格存储服务
	SLS-inner	云平台日志服务
	元数据库	元数据库
	POP	云平台开放接口Open API
	OAM	账号系统
	RAM	认证授权系统
	WebApps	运维控制台支撑

1.4 产品全景

专有云敏捷版提供丰富多样的产品，满足不同用户的多种需求。

- **基础架构**

提供丰富的基础虚拟资源。主要产品有ZStack、容器服务、密钥管理服务KMS。

- **存储产品**

针对不同的存储对象提供多样化的存储产品。主要产品有对象存储OSS、文件存储NAS、表格存储（Table Store）、块存储。

- **互联网中间件和应用**

提供了应用中间件服务，支持各类客户应用程序，实现应用服务化。主要产品包括分布式关系型数据库DRDS、消息队列MQ。

- **数据库**

提供多样化的数据引擎，并且能够实现多数据引擎间互通。主要产品包括云数据库MySQL、数据传输DTS、数据管理DMS。

- **安全**

提供从底层通信协议到上层应用的全方位防护，保证用户的访问和数据安全，如云盾。

1.5 应用场景

阿里云专有云能够为不同规模、同行业的用户提供灵活的、可扩展的行业解决方案。针对各行业业务特性，打造个性化方案，为用户提供一站式的产品与服务。包括工业、农业、交通、政务、金融、交通、教育等各个方面，本章节将重点介绍存储小型化专有云应用场景。

存储小型化专有云应用场景

存储小型化专有云针对传统存储市场，提供高可用、高吞吐量、低时延和良好扩展性的存储服务小型化输出。包括对象存储OSS，文件存储NAS，表格存储Table Store。企业的数据量日益激增，多样的结构化及非结构化数据需要管理，大量的各式数据需要可靠高效的存储，但现有存储系统的扩容能力已接近极限，而且读、写带宽和响应时延不能满足要求

价值特点

- 一体化存储平台，资源利用率高、部署以及运营成本大幅降低。
- 总带宽随节点的扩展而线性提升，并保障弹性规模扩展下的系统性能，适应未来业务的发展趋势。
- 相比原来的传统存储在并发处理能力、读写速度上大幅度提升数据库小型化专有云。

越来越多的企业，尤其是银行、证券、保险、基金等金融机构，希望建设分布式数据库，以此对互联网形态业务提供支撑，并且这种小型化的飞天输出，可与专有云企业版组合，形成低成本的异地灾备方案。

2 容器服务

2.1 什么是容器服务

容器服务提供高性能可伸缩的容器应用管理服务，支持用 Kubernetes 进行容器化应用的生命周期管理，提供多种应用发布方式和持续交付能力并支持微服务架构。可以直接部署在企业已有的x86等硬件基础之上并加以管理，支持研发运维一体化、云原生应用架构和机器学习等场景，支持混合云管理，允许应用在公共云，和自有数据中心物理机统一部署管理，支持应用无缝迁云、弹性伸缩应对突发流量等场景。

2.2 产品优势

技术保障

- 经过双十一大规模线上业务验证。
- 提供专业化技术支持团队。

敏捷

- 应用秒级启动。
- 打造软件研发流水线，相对于常规应用开发流程，应用迭代速度提升13倍，加速企业业务迭代。
- 良好的隔离性，可以多种应用类型混合部署，系统利用率提升5-10倍。

弹性

- 密切监控应用性能，根据负载情况自动伸缩资源，从容应对突发流量。

开放

- 开放的产品技术路线。
- 活跃的生态。

2.3 产品架构

容器服务提供高性能可伸缩的容器应用管理服务，支持企业级Kubernetes容器化应用的生命周期管理。容器服务简化集群的搭建和扩容等运维工作，整合企业虚拟化、存储、网络和安全能力，打造最佳的企业级Kubernetes容器化应用运行环境。容器服务是Kubernetes认证服务供应商，全球首批通过Kubernetes一致性认证的平台。

容器服务部署在客户自有数据中心里，所有数据存储在客户本地，完全由客户控制。适配客户已有的资源，无需额外采购，充分利用现有硬件资源和软件资源。保证用户应用的敏捷上云，容器服务对外提供了标准开放的API，支持应用无缝迁云。为第三方能力扩展，实现了灵活可定制的扩展机制。

容器服务的产品架构图如下。



2.4 功能特性

资源调度

- 支持大规模集群的统一资源池化管理，支持管理用户的物理机、VMware等现有IAAS环境。
- 支持根据应用需求动态调度容器，可选择多种维度的调度策略，比如资源维度（CPU、内存、GPU等）、可用性要求维度、应用拓扑的亲亲和性维度等等。

微服务

- 内置通用的服务注册、发现、路由、负载均衡等机制，对开发语言和中间件无特殊需求。
- 提供声明式方式配置，无需编码。
- 支持Spring Cloud等开源微服务框架。

DevOps

- 内置容器化DevOps最佳实践，可以实现一键式从代码提交到应用变更上线的全自动流程。
- 支持与三方、开源CI/CD方案整合。
- 提供不间断发布和蓝绿发布、金丝雀发布等灰度发布机制，支持灵活、可控的服务更新。

日志与监控

- 提供企业级日志采集和输出方案。无缝集成容器日志采集，支持采集标准输出或指定目录的日志。
- 可以选择对接阿里云日志服务，支撑超万台服务器，IOPS超2万。也可以对接ELK等三方开源或企业已有日志框架。
- 提供容器级别、应用级别和宿主机级别的多维度监控，提供服务和应用视角聚合数据。
- 支持脚本、URL等自定义监控。
- 可以选择对接阿里云云监控，也可以对接三方开源监控解决方案。

安全

- 与企业用户目录无缝集成，支持统一用户认证管理。
- 支持基于角色的授权模型，对集群资源灵活控制。

镜像管理

- 支持高可用镜像仓库，支持高可用存储。
- 支持镜像安全扫描和数字证书签名，实现安全可控的协作和应用分发机制。

容器存储

- 支持本地和分布式数据卷管理，支持企业已有NAS和SAN存储。
- 内置支持阿里云存储能力，包括对象存储、块存储和分布式文件系统，提供标准FlexVolume驱动。
- 可以通过插件扩展机制对接更多存储实现。

容器网络

- 支持容器间高性能跨宿主机网络通信。
- 支持与企业现有网络方案对接。
- 支持打通云上云下，混合管理。
- 提供插件扩展机制支持更多网络方案。
- 支持容器访问策略和流控限制。

应用管理

- 支持应用灰度发布，蓝绿发布。
- 内置应用目录，支持Helm应用一键部署和升级

- 支持服务目录，简化服务集成。

混合云

- 与阿里云公共云无缝结合，实现混合云。
- 支持对自有数据中心和公共云容器集群的统一资源、镜像、应用和安全管理。
- 支持工作负载动态迁移。
- 支持云突发、云灾备、异地多中心管理等典型场景。

自动化运维与管理

- 基于阿里云飞天基础架构管理平台。
- 提供硬件资产管理、自动化监控、运维能力。
- 支持基础网络服务管理。

无缝对接阿里云公共云

- 提供通过专线、VPN等网络互通方式。
- 复用阿里云提供数据复制机制，可实现数据迁移、异地灾备等多种混合云业务应用场景。

2.5 应用场景

2.5.1 容器化应用上云

传统运行在数据中心的应用如果要向云上迁移面临的挑战是应用的运行环境的适配以及应用部署方式的改变。原有数据中心的环境和云环境的差异越大，迁移的成本越高。

由于容器技术具有跨平台移植的特性，可以解决不同环境下应用运行差异问题。无论用户的应用是何种语言，何种技术栈，都能够通过容器技术封装成为跨平台可用的Docker镜像。从而将原来繁重的应用迁移变成一系列标准化的操作。

一般情况下，在数据中心里运行的虚拟机镜像无法直接迁移到云上环境，原来的镜像构建和部署脚本都要重新开发。采用容器技术可以使用容器镜像替代虚拟机镜像。利用容器技术将原有的传统应用镜像化，可以在尽量减少或者没有改变原有代码的情况下快速上云，并保证部署的一致性。

利用Docker Compose部署模版，可以描述完整的应用栈和所需云资源。容器服务扩展了Compose模板支持服务治理：包括服务调度约束、服务注册/发现策略、弹性伸缩等；以及容器与外部资源的动态集成，如负载均衡，网络存储，RDS数据库服务等等。这样可以采用应用为中心的管理方式，大大提升了部署、运维效率。

这个方案的特点是：成本低、上线速度快，并且在完成迁移后可以利用容器服务的微服务支持进行进一步微服务化改造。

2.5.2 容器化DevOps

开发和运维中交接的代码，由于同一代码在不同环境中的差异性，往往导致很高的调试成本，软件的交付和测试周期长。开发运维团队需要花费很大的精力来屏蔽系统差异，标准化交付手段。

容器技术作为软件的交付手段，具有同一应用多个平台的可移植性，和不同应用在同一环境中的相互隔离的特点。利用这一特性，使容器成为开发环境、测试环境、生产环境的标准代码交付手段。在任何环境中运行的都是同一个容器，这样保证了由于环境不同所带来的差异被最大限度地屏蔽。利用这种标准交付手段，可以大大提高开发和运维的效率。

容器服务提供了容器化Jenkins的一键式部署模板，可以快速灵活地搭建一套CI/CD环境。同时还提供了Jenkins插件来充分发挥容器服务的能力，比如多种发布模式，保证新系统和服务上线的可靠性。比如通过灰度发布，可以灵活控制新老版本的流量切换，快速、可靠地回滚。

2.5.3 微服务支持

微服务作为云应用常见的架构模式正在被越来越多地接受。但是随着服务实例数量的快速增长，对服务治理的要求会越来越高。容器服务提供了和语言无关的微服务治理能力，您不限定语言和开发框架，不用改变应用逻辑，就可以轻松应对微服务的管理和伸缩。所有这些能力都可以通过在编排模版中以声明的方式指定，或者在容器服务控制台中可视化地指定。

从对SLA保证角度，容器服务提供了利用自定义命令/HTTP等健康检查的能力，可以细粒度地评估应用的健康状况。容器服务还可以按照资源约束调度和再平衡，当节点故障时自动重调度，按照服务亲和性和跨可用区调度约束等机制保证可用性。

服务发现是微服务平台的一个重要能力，所有服务的启动和停止都能够自动注册到平台，服务之间可以非常容易地进行服务发现。更进一步，配合Routing Mesh，简单路由服务可以为应用提供灵活的4层、7层路由方式。

对于任何运维团队来说，日志的集中和分析，云资源的集中监控都必不可少。容器服务提供了声明式日志、监控采集方案，只需要简单声明要采集的日志目录，就可以自动收集日志。支持对接多种日志、监控存储。

配合监控服务，在您自定义弹性伸缩策略后，容器服务对应用的指标进行监控，在触发条件时按照策略进行应用弹性扩容和缩容。容器服务同时也支持集群弹性伸缩。

2.5.4 弹性伸缩

根据业务流量自动对业务扩容/缩容

容器服务可以根据业务流量自动对业务扩容/缩容，不需要人工干预，避免流量激增扩容不及时导致系统挂掉，以及平时大量闲置资源造成浪费。

能够实现：

- 快速响应

业务流量达到扩容指标，秒级触发容器扩容操作。

- 全自动

整个扩容/缩容过程完全自动化，无需人工干预。

- 低成本

流量降低自动缩容，避免资源浪费

2.5.5 混合云

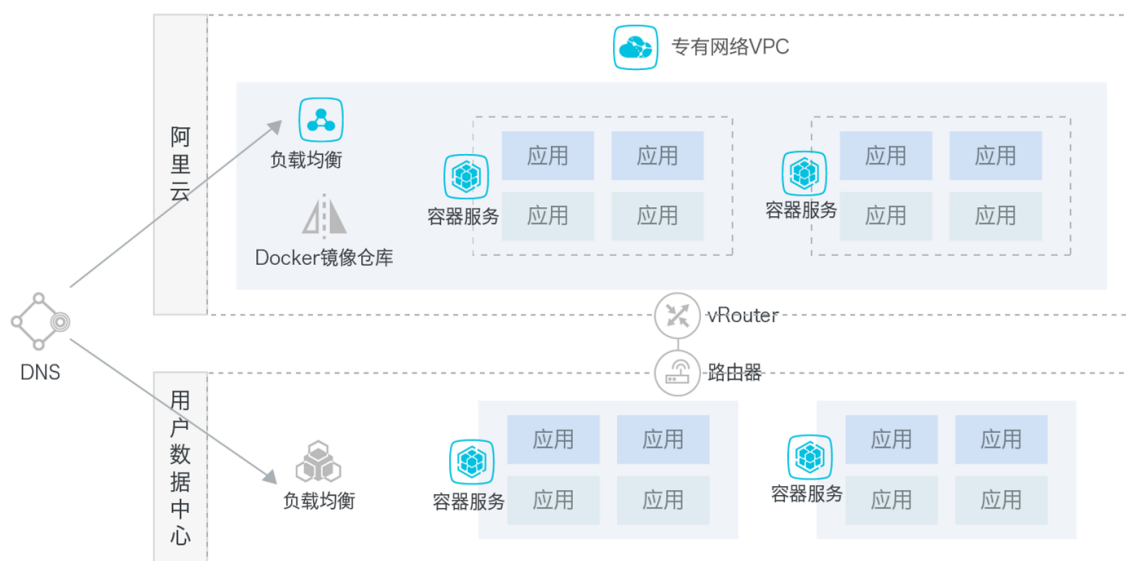
容器服务支持计算能力跨界迁移，通过容器服务可以管理本地专有云和公共云容器集群，提供了对所有环境完全一致的管理方式，您无需针对不同的其他云服务商或者IDC机器寻求特定的管理方案。使用同一套镜像和编排模板让应用可以无缝地在云间迁移。

混合云的典型场景包括：

- 场景1：应用快速在云端部署伸缩，应对大促突发流量峰值
- 场景2：异地、同城灾备
- 场景3：本地的机器开发测试，云端预发生产；或反之

针对混合云，为了提供最佳的网络实现，容器服务提供了不同的网络方案来适配不同环境。对于IDC与阿里公共云容器互通的需求，提供基于VPC的混合云互连方案，使用VPC+专线/VPN，可以实现IDC机器、IDC容器、ECS机器、ECS容器的全部互通。容器跨宿主机互连方案包括：Overlay (VXLAN) /Layer 3 (Calico) 等。

图 2-1: 混合云架构



2.6 使用限制

表 2-1: Kubernetes集群使用限制

限制项	限制说明
集群	- 目前 Kubernetes 集群只支持 Linux 容器。 - 目前 Kubernetes 集群只支持 VPC 网络。
ECS 实例	- 仅支持 CentOS 操作系统。 - 添加已有节点使用限制。 - 添加的云服务器必须与集群在同一专有网络。 - 容器服务不支持添加不同账号下的云服务器。
集群扩容和缩容	- Worker 节点数需大于等于 1。 - 目前，仅支持用户手动进行扩容和缩容，不支持自动弹性伸缩。 - 目前，不支持自动扩容容器服务集群中的 Master 节点。

2.7 基本概念

集群

一个集群指容器运行所需要的资源组合，关联了若干服务器节点等资源。

节点

一台服务器（可以是虚拟机实例或者物理服务器）已经安装了Docker Engine，可以用于部署和管理容器；容器服务的Agent程序会安装到节点上并注册到一个集群上。集群中的节点数量可以伸缩。

容器

一个通过Docker镜像创建的运行时实例，一个节点可运行多个容器。

镜像

Docker镜像是容器应用打包的标准格式，在部署容器化应用时可以指定镜像，镜像可以来自于Docker Hub，阿里云容器镜像服务，容器服务的DTR，或者用户的私有Registry。镜像ID可以由镜像所在仓库URI和镜像Tag（缺省为latest）唯一确认。

编排模板

编排模板包含了一组容器服务的定义和其相互关联，可以用于多容器应用的部署和管理。

Kubernetes 相关概念

Kubernetes是一个 Google 开源的，用于自动部署、扩展和管理容器化应用的大规模容器编排调度系统。具有可移植、可扩展和自动调度等特性。

节点 (Node)

Kubernetes 集群中的计算能力由 Node 提供，Kubernetes 集群中的 Node 是所有 Pod 运行所在的工作主机，可以是物理机也可以是虚拟机。工作主机的统一特征是上面要运行 kubelet 管理节点上运行的容器。

命名空间 (Namespace)

命名空间为 Kubernetes 集群提供虚拟的隔离作用。Kubernetes 集群初始有 3 个命名空间，分别是默认命名空间 default、系统命名空间 kube-system 和 kube-public，除此以外，管理员可以创建新的命名空间满足需要。

Pod

Pod是 Kubernetes 部署应用或服务的最小的基本单位。一个Pod 封装多个应用容器（也可以只有一个容器）、存储资源、一个独立的网络 IP 以及管理控制容器运行方式的策略选项。

副本控制器 (Replication Controller , RC)

RC 确保任何时候 Kubernetes 集群中有指定数量的 pod 副本(replicas)在运行。通过监控运行中的 Pod 来保证集群中运行指定数目的 Pod 副本。指定的数目可以是多个也可以是1个；少于指定数目，RC 就会启动运行新的 Pod 副本；多于指定数目，RC 就会终止多余的 Pod 副本。

副本集 (Replica Set , RS)

ReplicaSet (RS) 是 RC 的升级版，唯一区别是对选择器的支持，RS 能支持更多种类的匹配模式。副本集对象一般不单独使用，而是作为 Deployment 的理想状态参数使用。

部署 (Deployment)

部署表示用户对 Kubernetes 集群的一次更新操作。部署比 RS 应用更广，可以是创建一个新的服务，更新一个新的服务，也可以是滚动升级一个服务。滚动升级一个服务，实际是创建一个新的RS，然后逐渐将新 RS 中副本数增加到理想状态，将旧 RS 中的副本数减小到 0 的复合操作；这样一个复合操作用一个RS是不太好描述的，所以用一个更通用的 Deployment 来描述。不建议您手动管理利用 Deployment 创建的 RS。

服务 (Service)

Service 也是 Kubernetes 的基本操作单元，是真实应用服务的抽象，每一个服务后面都有很多对应的容器来提供支持，通过 Kube-Proxy 的 port 和服务 selector 决定服务请求传递给后端的容器，对外表现为一个单一访问接口，外部不需要了解后端如何运行，这给扩展或维护后端带来很大的好处。

标签 (labels)

Labels 的实质是附着在资源对象上的一系列 Key/Value 键值对，用于指定对用户有意义的对象的属性，标签对内核系统是没有直接意义的。标签可以在创建一个对象的时候直接赋予，也可以在后期随时修改，每一个对象可以拥有多个标签，但 key 值必须唯一。

存储卷 (Volume)

Kubernetes 集群中的存储卷跟 Docker 的存储卷有些类似，只不过 Docker 的存储卷作用范围为一个容器，而 Kubernetes 的存储卷的生命周期和作用范围是一个 Pod。每个 Pod 中声明的存储卷由 Pod 中的所有容器共享。支持使用 Persistent Volume Claim 即 PVC 这种逻辑存储，使用者可以忽略后台的实际存储技术，具体关于 Persistent Volumn(pv)的配置由存储管理员来配置。

持久存储卷 (Persistent Volume , PV) 和持久存储卷声明 (Persistent Volume Claim , PVC)

PV 和 PVC 使得 Kubernetes 集群具备了存储的逻辑抽象能力，使得在配置 Pod 的逻辑里可以忽略对实际后台存储技术的配置，而把这项配置的工作交给 PV 的配置者。存储的 PV 和 PVC 的这种关

系，跟计算的 Node 和 Pod 的关系是非常类似的；PV 和 Node 是资源的提供者，根据集群的基础设施变化而变化，由 Kubernetes 集群管理员配置；而 PVC 和 Pod 是资源的使用者，根据业务服务的需求变化而变化，由 Kubernetes 集群的使用者即服务的管理员来配置。

Ingress

Ingress 是授权入站连接到达集群服务的规则集合。你可以给 Ingress 配置提供外部可访问的 URL、负载均衡、SSL、基于名称的虚拟主机等。用户通过 POST Ingress 资源到 API server 的方式来请求 ingress。Ingress controller 负责实现 Ingress，通常使用负载均衡器，它还可以配置边界路由和其他前端，这有助于以 HA 方式处理流量。

3 对象存储OSS

3.1 什么是对象存储OSS

阿里云对象存储服务 (Object Storage Service , 简称 OSS) 是阿里云提供的海量、安全、低成本、高可靠的云存储服务。

OSS可以被理解成一个即开即用, 无限大空间的存储集群。相比传统自建服务器存储, OSS 在可靠性、安全性、成本和数据处理能力方面都有着突出的优势。使用 OSS, 您可以通过网络随时存储和调用包括文本、图片、音频和视频等在内的各种非结构化数据文件。

OSS 将数据文件以对象/文件 (Object) 的形式上传到存储空间 (Bucket) 中。OSS 提供的是一个 Key-Value 键值对形式的对象存储服务。用户可以根据Object的名称 (Key) 唯一地获取该Object的内容。

您可以进行以下 OSS 相关的操作：

- 创建存储空间, 并向存储空间中上传文件。
- 获取已上传文件的地址, 进行文件的分享和下载。
- 修改存储空间或文件的属性或元信息, 为其设置相应的权限。
- 在对象存储 OSS 控制台进行基础和高级 OSS 操作。
- 通过开发工具包 SDK 或直接在应用程序中调用 RESTful API, 进行基础和高级 OSS 操作。

3.2 产品优势

OSS与自建存储对比的优势

对比项	对象存储 OSS	自建服务器存储
可靠性	• 规模自动扩展, 不影响对外服务。 • 数据自动多重冗余备份。	• 受限于硬件可靠性, 易出问题, 一旦出现磁盘坏道, 容易出现不可逆转的数据丢失。 • 人工数据恢复困难、耗时、耗力。
安全	• 提供企业级多层次安全防护。 • 多用户资源隔离机制, 支持同城容灾。	• 需要另外购买清洗和黑洞设备。 • 需要单独实现安全机制。

对比项	对象存储 OSS	自建服务器存储
	提供多种鉴权和授权机制，以及白名单、防盗链、主子账号、STS临时授权访问功能。	
数据处理能力	提供图片处理功能。	需要额外采购，单独部署。

OSS具备的其他各项优势

- 方便、快捷的使用方式

提供标准的 RESTful API 接口（部分接口与Amazon S3 API兼容）、丰富的 SDK 包、客户端工具、控制台。您可以像使用文件一样方便地上传、下载、检索、管理用于 Web 网站或者移动应用的海量数据。

- 不限文件数量和大小。您可以根据所需存储量无限扩展存储空间，解决了传统硬件存储扩容问题。
 - 支持流式写入和读出。特别适合视频等大文件的边写边读业务场景。
 - 支持数据生命周期管理。您可以自定义将到期数据批量删除。

- 强大、灵活的安全机制

灵活的鉴权、授权机制。提供 STS 和 URL 鉴权和授权机制，以及白名单、防盗链、主子账号功能。

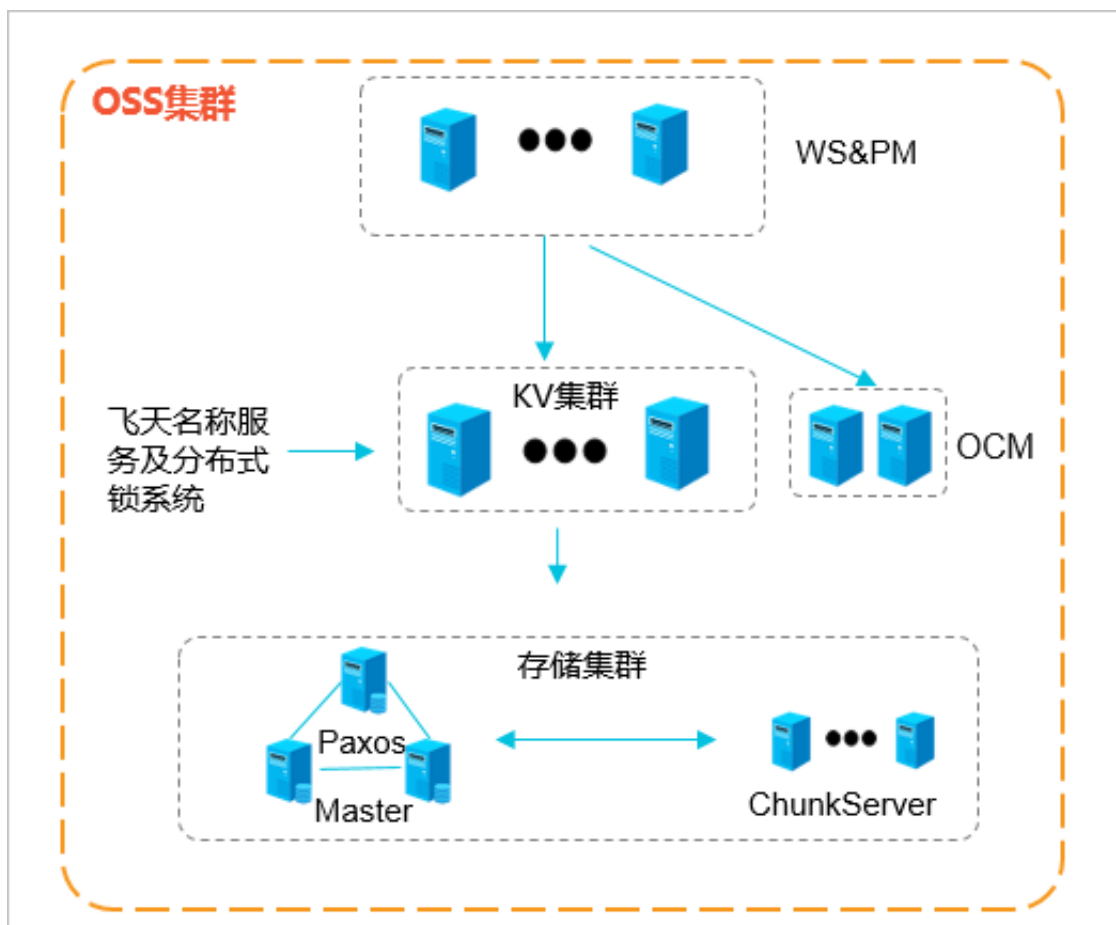
- 丰富的图片处理服务

支持 jpg、png、bmp、gif、webp、tiff 等多种图片格式的转换，以及缩略图、剪裁、水印、缩放等多种操作。

3.3 产品架构

对象存储 OSS 是构建在阿里云飞天平台上的一种存储解决方案。其基础是飞天平台的分布式文件系统，分布式任务调度等基础设施。这些基础设施提供了 OSS 以及其他阿里云服务所需的分布式调度、高速网络、分布式存储等重要特性。OSS 的架构如下图所示：

图 3-1: OSS 架构图



- WS&PM（协议接入层）：负责接收用户使用 RESTful 协议发来的请求，进行安全认证。如果认证通过，用户的请求将被转发到 Key-Value 引擎继续处理。如果认证失败，直接返回错误信息给用户。
- KV 集群：负责数据结构化处理，即按照 Key（对象名称）来查找或存储数据，并支持大规模并发的请求。当协调服务集群变更导致服务被迫改变运行物理位置时，可以快速协调找到接入点。
- 存储集群：元数据存储于 Master 上，Master 之间采用分布式消息一致性协议（Paxos）保证元数据的一致性。从而实现高效的文件分布式存储和访问。

3.4 功能特性

OSS 提供以下主要功能：

表 3-1: OSS主要功能

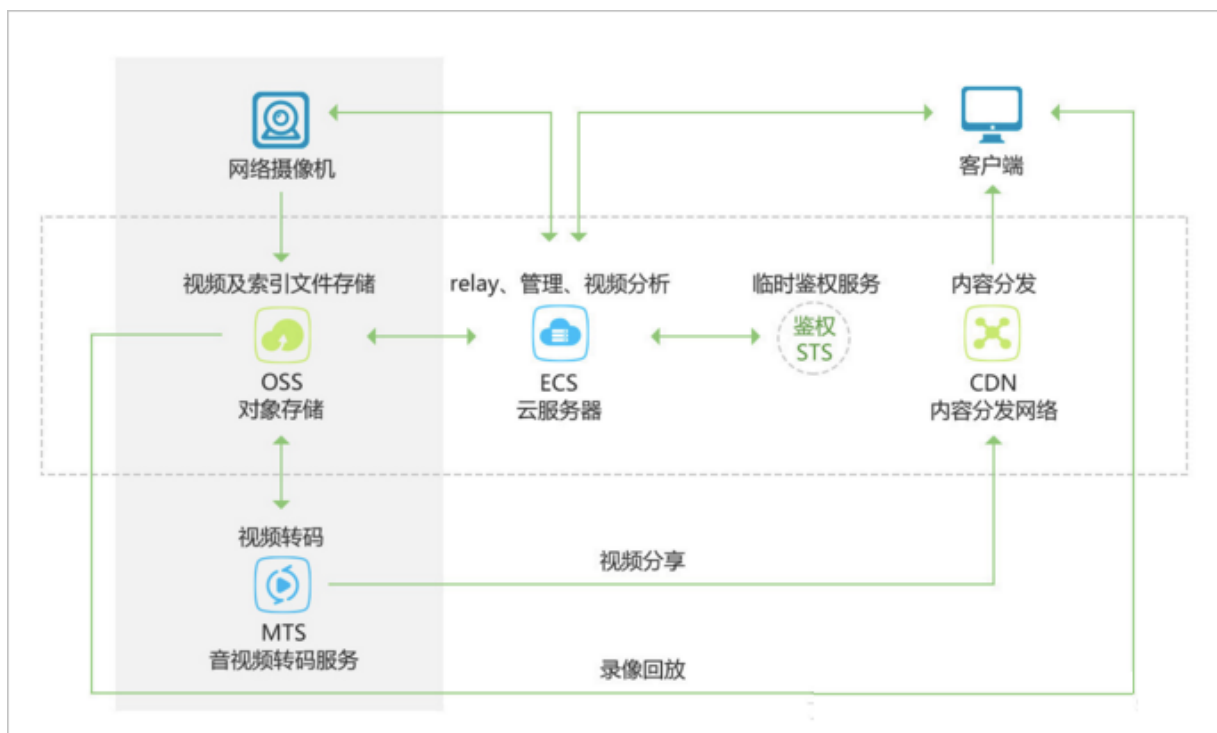
类别	功能	描述
存储空间	创建存储空间	在上传任何文件到 OSS 之前，您需要首先创建存储空间来存储文件。
	删除存储空间	如果您不再需要存储空间，请将其删除以免进一步产生费用。
	修改存储空间读写权限	OSS 提供权限控制 ACL (Access Control List)，您可以在创建存储空间的时候设置相应的 ACL 权限控制，也可以在创建之后修改 ACL。
	设置静态网站托管	将存储空间配置成静态网站托管模式，并通过存储空间域名访问该静态网站。
	设置防盗链	为了减少您存储于 OSS 的数据被其他人盗链而产生额外费用，OSS 支持设置基于 HTTP header 中表头字段 referer 的防盗链方法。
	管理跨域资源共享	OSS 提供 HTML5 协议中的跨域资源共享 CORS 设置，帮助您实现跨域访问。
	设置生命周期	定义和管理存储空间内所有对象或对象的某个子集的生命周期。设置生命周期一般用于文件的批量管理和自动碎片删除等操作。
对象（文件）	上传文件	您可以上传任意类型文件到存储空间中。
	新建文件夹	您可以像管理 Windows 文件夹一样管理 OSS 文件夹。
	搜索文件	在存储空间或文件夹中搜索具有相同的名称前缀的文件。
	获取文件访问地址	通过获取已上传文件的地址进行文件的分享和下载。
	删除文件	删除单个文件或批量删除文件。
	删除文件夹	删除单个文件夹或批量删除文件夹。
	修改文件读写权限	您可以在上传文件的时候设置相应的 ACL 权限控制，也可以在上传之后修改 ACL。
	管理碎片	删除存储空间内的全部或部分碎片文件。
图片处理	图片处理	对保存在OSS上的图片进行格式转换、剪裁、缩放、旋转、水印、样式封装等各种处理。
API	API	提供 OSS支持的 RESTful API 操作和相关示例。

类别	功能	描述
SDK	SDK	提供主流语言 SDK 的开发操作和相关示例。

3.5 应用场景

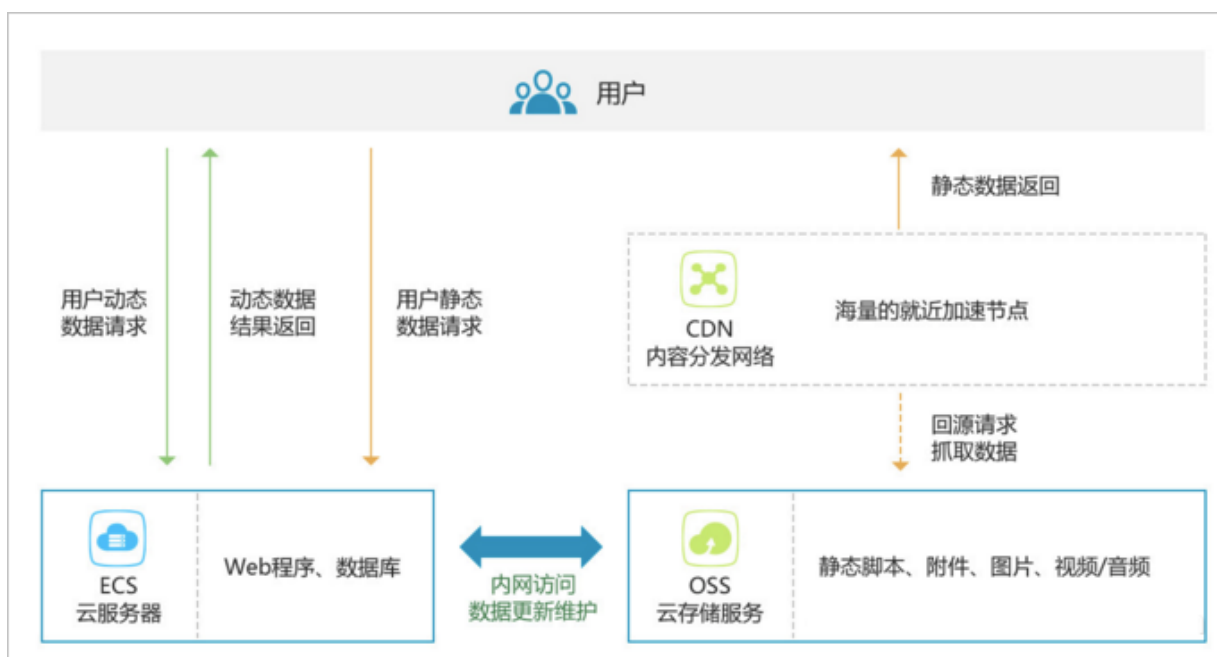
图片和音视频等应用的海量存储

OSS可用于图片、音视频、日志等海量文件的存储。各种终端设备、Web网站程序、移动应用可以直接向OSS写入或读取数据。OSS支持流式写入和文件写入两种方式。



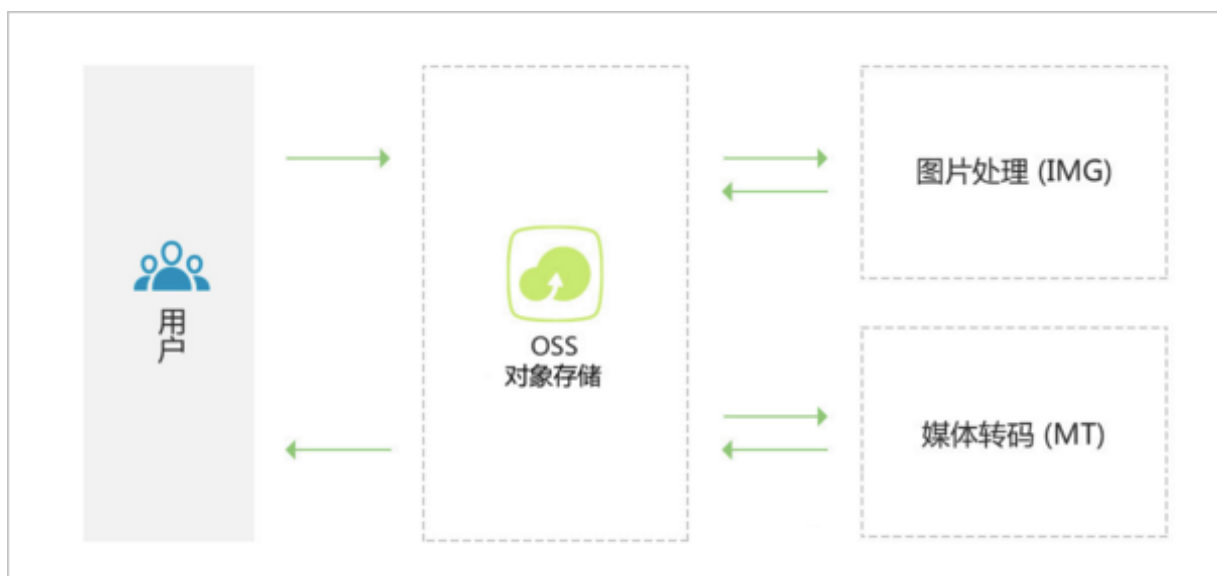
网页或者移动应用的静态和动态资源分离

利用BGP带宽，OSS可以实现超低延时的数据直接下载。



离线数据归档存储

依靠低成本、高可用的OSS对象存储，可以将企业内部长期需要离线归档的数据转存至OSS。



3.6 使用限制

限制项	说明
存储空间 (bucket)	- 同一用户创建的存储空间总数不能超过100个。 - 存储空间一旦创建成功，名称和区域不能修改。
上传文件	通过控制台上传、简单上传、表单上传、追加上传的文件大小不能超过5GB，要上传大小超过5GB的文件必须使用

限制项	说明
	分片上传(Multipart Upload) 上传方式。分片上传方式上传的文件大小不能超过48.8TB。 • OSS支持上传同名文件，但会覆盖已有文件。
删除文件	• 文件删除后无法恢复。 • 控制台批量删除文件的上限为50个，更大批量的删除必须通过API或SDK实现。
生命周期	每个存储空间的生命周期配置最多可容纳1000条规则。
图片处理	• 对于原图： ▪ 图片格式只能是：jpg、png、bmp、gif、webp、tiff。 ▪ 文件大小不能超过20MB。 ▪ 使用图片旋转或裁剪时图片的宽或者高不能超过4096 PX。 • 对于缩略后的图： ▪ 宽与高的乘积不能超过4096x4096。 ▪ 单边长度不能超过4096 PX。

3.7 基本概念

本部分将向您介绍 OSS 中涉及的几个基本概念，以便于您更好地理解 OSS 产品。

对象/文件 (Object)

对象是 OSS 存储数据的基本单元，也被称为 OSS 的文件。对象由元信息 (Object Meta)，用户数据 (Data) 和文件名 (Key) 组成。对象由存储空间内部唯一的 Key 来标识。对象元信息是一个键值对，表示了对象的一些属性，比如最后修改时间、大小等信息，同时用户也可以在元信息中存储一些自定义的信息。

对象的生命周期是从上传成功到被删除为止。在整个生命周期内，对象信息不可变更。重复上传同名的对象会覆盖之前的对象，因此，OSS 不支持修改文件的部分内容等操作。



说明：

如无特殊说明，本文档中的对象、文件称谓等同于 Object。

存储空间 (Bucket)

存储空间是您用于存储对象 (Object) 的容器，所有的对象都必须隶属于某个存储空间。您可以设置和修改存储空间属性用来控制访问权限、生命周期等，这些属性设置直接作用于该存储空间内所有对象，因此您可以通过灵活创建不同的存储空间来完成不同的管理功能。

- 同一个存储空间的内部是扁平的，没有文件系统的目录等概念，所有的对象都直接隶属于其对应的存储空间。
- 每个用户可以拥有多个存储空间。
- 存储空间的名称在 OSS 范围内必须是全局唯一的，一旦创建之后无法修改名称。
- 存储空间内部的对象数目没有限制。

强一致性

Object 操作在 OSS 上具有原子性，操作要么成功要么失败，不会存在有中间状态的Object。OSS 保证用户一旦上传完成之后读到的 Object 是完整的，OSS 不会返回给用户一个部分上传成功的 Object。

Object 操作在 OSS 上同样具有强一致性，用户一旦收到了一个上传 (PUT) 成功的响应，该上传的 Object 就已经立即可读，并且数据的三份副本已经写成功。不存在一种上传的中间状态，即 read-after-write 却无法读取到数据。对于删除操作也是一样的，用户删除指定的 Object 成功之后，该 Object 立即变为不存在。

强一致性方便了用户架构设计，可以使用跟传统存储设备同样的逻辑来使用OSS，修改立即可见，无需考虑最终一致性带来的各种问题。

OSS与文件系统的对比

OSS 是一个分布式的对象存储服务，提供的是一个 Key-Value 对形式的对象存储服务。用户可以根据 Object 的名称 (Key) 唯一地获取该Object的内容。虽然用户可以使用类似 test1/test.jpg 的名字，但是这并不表示用户的 Object 是保存在test1 目录下面的。对于 OSS 来说，test1/test.jpg 仅仅只是一个字符串，和 a.jpg 这种并没有本质的区别。因此不同名称的 Object 之间的访问消耗的资源是类似的。

文件系统是一种典型的树状索引结构，一个名为 test1/test.jpg 的文件，访问过程需要先访问到 test1 这个目录，然后再在该目录下查找名为 test.jpg 的文件。因此文件系统可以很轻易的支持文件夹的操作，比如重命名目录、删除目录、移动目录等，因为这些操作仅仅只是针对目录节点的操作。这种组织结构也决定了文件系统访问越深的目录消耗的资源也越大，操作拥有很多文件的目录也会非常慢。

对于 OSS 来说，可以通过一些操作来模拟类似的功能，但是代价非常昂贵。比如重命名目录，希望将 test1 目录重命名成 test2，那么 OSS 的实际操作是将所有以 test1/ 开头的 Object 都重新复制成以 test2/ 开头的 Object，这是一个非常消耗资源的操作。因此在使用 OSS 的时候要尽量避免类似的操作。

OSS 保存的 Object 不支持修改（追加写 Object 需要调用特定的接口，生成的 Object 也和正常上传的 Object 类型上有差别）。用户哪怕是仅仅需要修改一个字节也需要重新上传整个 Object。而文件系统的文件支持修改，比如修改指定偏移位置的内容、截断文件尾部等，这些特点也使得文件系统拥有广泛的适用性。但另外一方面，OSS 能支持海量的用户并发访问，而文件系统会受限于单个设备的性能。

因此，将 OSS 映射为文件系统是非常低效的，也是不建议的做法。如果一定要挂载成文件系统的话，建议尽量只做写新文件、删除文件、读取文件这几种操作。使用 OSS 应该充分发挥其优点，即海量数据处理能力，优先用来存储海量的非结构化数据，比如图片、视频、文档等。

三份副本的原理

在阿里云数据存储系统中，有三类角色，分别称为 Master、Chunk Server 和 Client。您的一个写操作，经过层层转换，最终会交由 Client 来执行，执行过程简要说明如下：

1. Client 计算出这个写操作对应的 Chunk。
2. Client 向 Master 查询该 Chunk 的三份副本的存放位置。
3. Client 根据 Master 返回的结果，向这 3 个 Chunk Server 发出写请求。
4. 如果三份都写成功，Client 向您返回成功；反之，Client 向您返回失败。

4 表格存储Table Store

4.1 什么是表格存储

表格存储 (Table Store) 是阿里云自主研发的海量 NoSQL 数据存储服务, 具有国家颁发的软件著作权证书。表格存储构建在阿里云飞天分布式系统之上, 提供海量结构化数据的存储和实时访问。

表格存储提供如下功能:

- 单集群可支持数据量不低于 10 PB; 单表可支持数据量不低于 1PB 或者 1 万亿条记录。在数据存储模型上, 支持 schema free 的表结构, 属性列的使用不需要预先定义, 增加或者减少属性列不需要做表级别变更。同时支持表级别的数据生命周期, 过期数据自动删除。
- 采用三副本技术, 并将数据副本文件保存在不同机架上。单集群可以支持纯 SSD 存储实例或者 SSD 与 SATA 混合存储实例, 以满足对成本和性能上的不同需要。
- 采用全冗余架构, 无单点故障; 可支持在线平滑升级, 动态增加或者删除结点以及进行集群热升级, 不需要停机维护, 数据可以自动迁移; 读写并发和存储容量支持同步线性扩容, 单集群最大规模不低于 500 台。
- 支持高并发读写。读写并发能够随着机器数量水平扩展, 读写性能与单表数据量无直接关系。
- 支持身份验证, 支持多租户, 具备完善的权限认证与隔离机制, 保障用户数据的私密性, 支持 HTTPS 访问。支持 RAM 机制, 提供多种鉴权和授权机制及主子账号功能, 授权粒度达到表级别和 API 级别。

4.2 产品优势

扩展性

表格存储中表的数据量没有上限, 随着表数据量的不断增大, 表格存储会进行数据分区的调整从而为该表配置更多的存储并提供更高的并发访问能力。

高性能

高性能实例能够提供单个毫秒级的单行平均访问延时, 读写性能不受单表数据大小的影响。

数据可靠性

表格存储通过存储多个数据备份及备份失效时的快速恢复, 提供极高的数据可靠性。

高可用性

通过自动的故障检测和数据迁移, 表格存储对应用屏蔽了机器和网络的硬件故障, 提供高可用性。

管理便捷

应用程序无需关心数据分区的管理、软硬件升级、配置更新、集群扩容等繁琐的运维任务。

访问安全性

表格存储提供多种权限管理机制，并对应用的每一次请求都进行身份认证和鉴权，以防止未授权的数据访问，确保数据访问的安全性。

强一致性

表格存储保证数据写入强一致，写操作一旦返回成功，保证数据 3 副本均写入磁盘，并且应用就能立即读到最新的数据。

灵活的数据模型

表格存储的表无固定格式要求，每行的列数可以不相同，支持多种数据类型（Integer、Boolean、Double、String、Binary）。

按量付费

表格存储根据用户预留和实际使用的资源进行收费，起步门槛低。

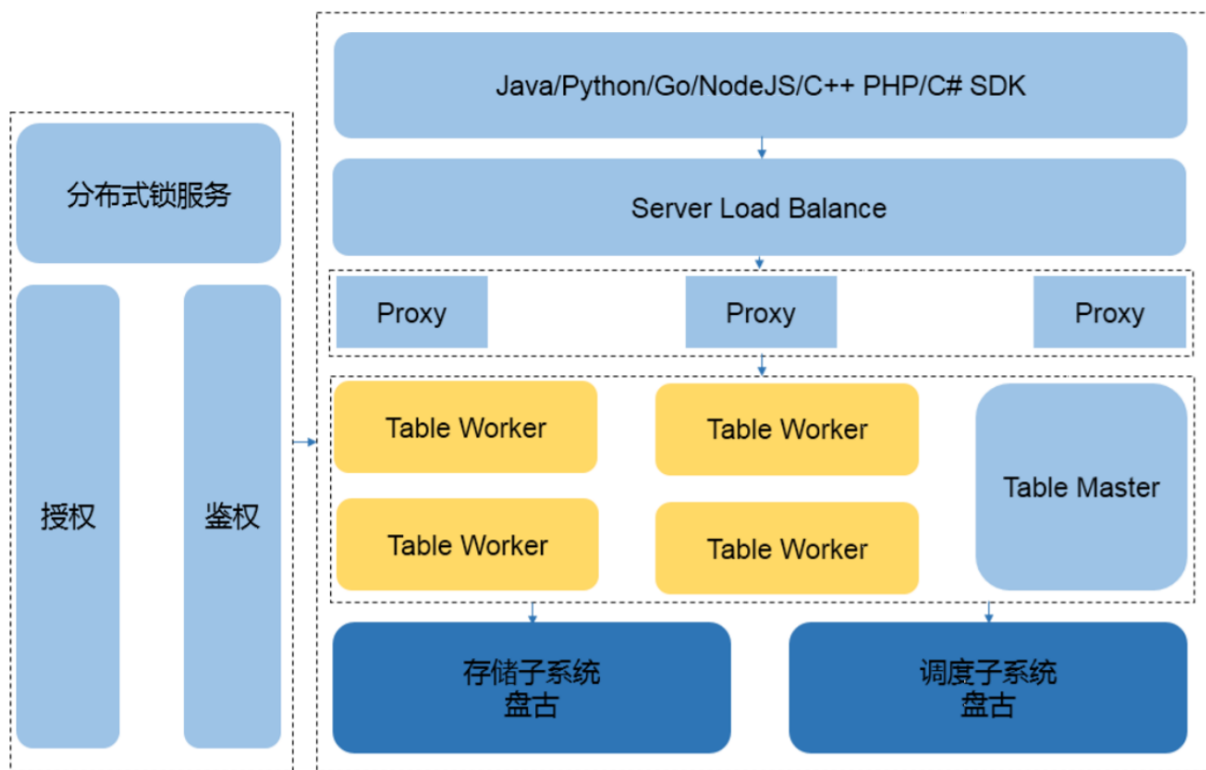
监控集成

用户可以从表格存储控制台实时获取每秒请求数、平均响应延时等监控。

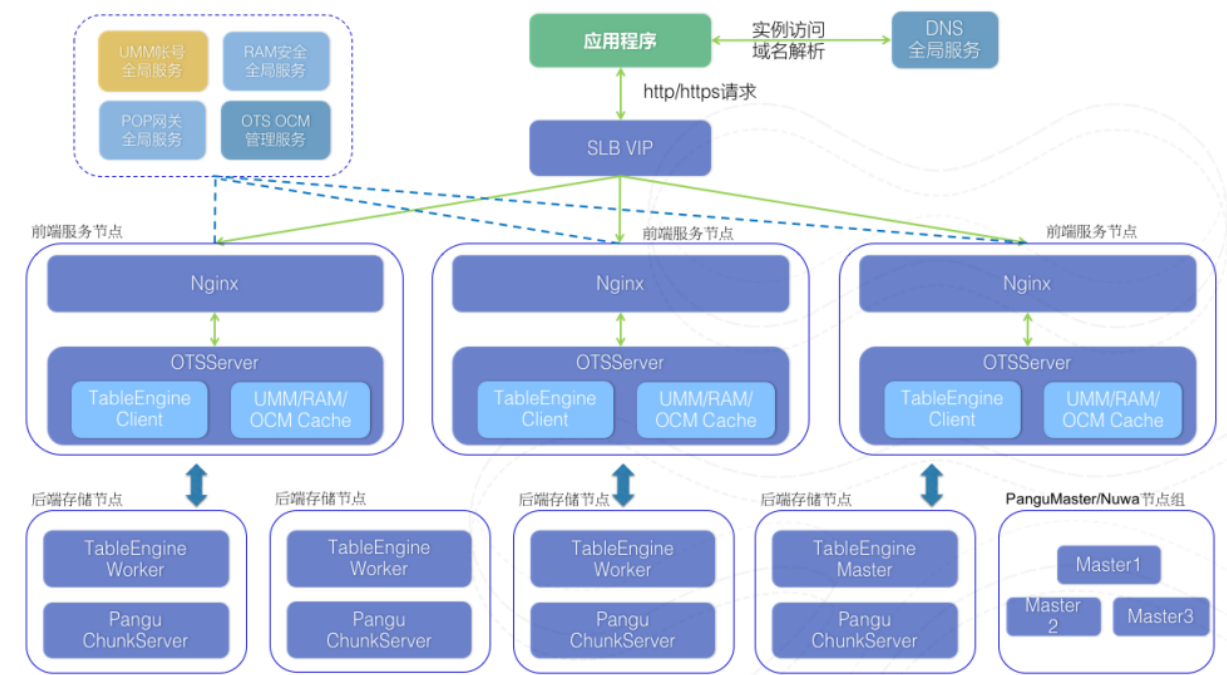
4.3 产品架构

表格存储的架构极大参考了 Google 的三驾马车之 BigTable，使用 LSM（Log-Structured Merge-Tree）存储引擎，能够提供极高的写性能，同时，基于主键 Key 的单行查询及范围查询性能稳定可预期，不受数据规模及访问并发的影响。

表格存储的基础架构如下：



表格存储的详细架构如下：



- 最上层是协议接入层，用户请求被 SLB 打散到不同的 Proxy 节点，Proxy 节点负责接收用户使用 RESTful 协议发来的请求，进行安全认证。如果认证通过，用户的请求将根据第一个主键的值转发到对应的数据引擎继续处理；如果认证失败，直接返回错误信息给用户。

- Table Worker 为数据引擎层，负责数据结构化处理，即按照主键 Key 来查找或存储数据，并支持大规模并发的请求。
- 最底层是持久存储层，即大规模分布式文件系统。元数据存储 Master 上，Master 之间采用分布式消息一致性协议（Paxos）保证元数据的一致性。从而实现高效的文件分布式存储和访问，保证数据在系统中有 3 个备份以及在软硬件错误发生以后的故障恢复。

4.4 功能特性

- **数据分区和负载均衡机制**

表中每一行主键的第一列称为数据分片键（Partition Key），系统根据数据分片键的范围将表切分为多个分区，这些分区被系统均匀的调度到不同的存储节点上。当单个数据分区内的数据不断增加到一定程度时，分区会自动分裂成两个更小的分区，从而将数据和访问负载分散到两个分区上，这两个分区会被调度到不同的节点上，从而将访问负载分散到不同的节点上，最终达到了单表数据规模和访问压力的线性扩展。由于使用了共享存储机制，数据分区是数据的逻辑组织结构，分区的分裂不涉及实际数据的迁移，负载均衡对该分区的理论影响时间在百毫秒级别。

- **单机故障自动恢复**

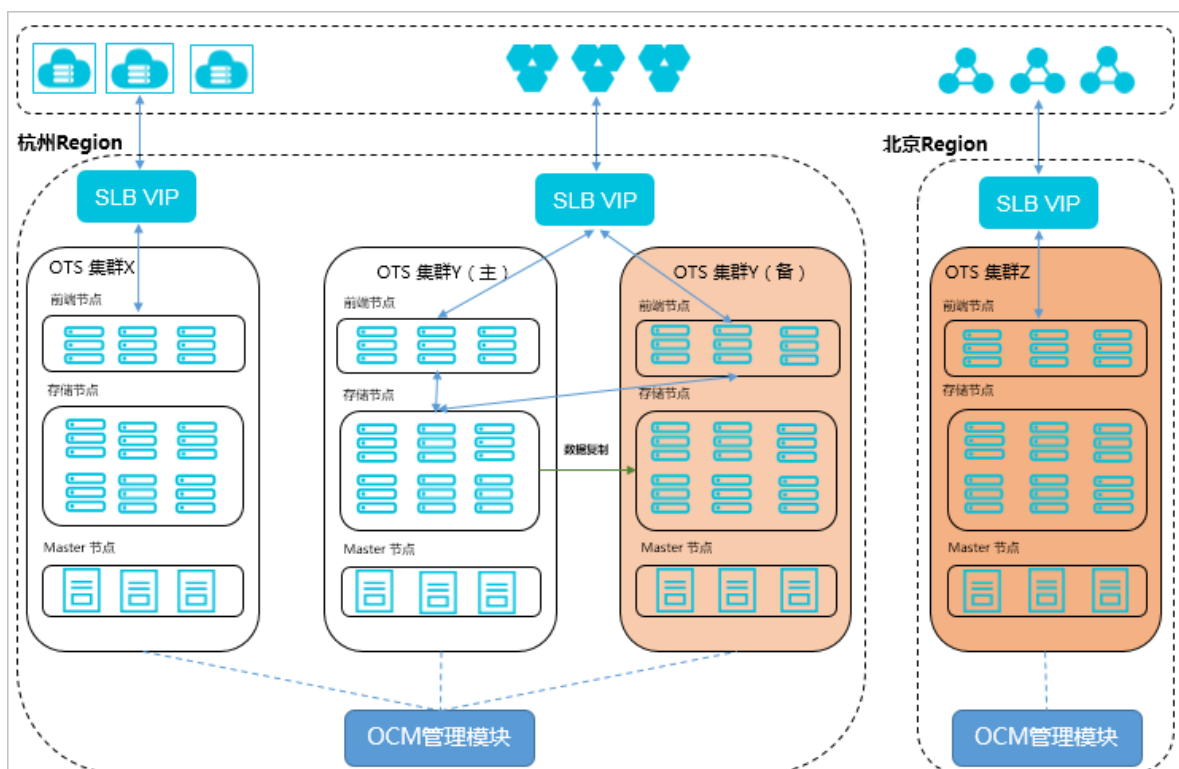
在表格存储的存储引擎中，每个节点都会服务一批不同表的数据分区，这些分区的分布和调度信息由一个 Master 节点负责来管理，并且 Master 节点也会监控每个服务节点的健康状态。当发现某个服务节点出现问题时，Master 就会将原先分配给这个服务节点的数据分区调度到其他健康的节点上。由于数据分区的迁移只是逻辑上的迁移，底层使用分布式文件系统，不涉及实际数据的移动，所以当出现单机故障时，服务能够在很短的时间内恢复。

- **同城及异地灾备**

为了满足业务对安全性和可用性的需求，表格存储提供了同城及异地双集群主备容灾，容灾粒度为实例级别。打开了容灾的实例下的数据表上的数据插入、更新、删除操作会异步的复制到备实例下的同名表，主备实例的数据保持一致的时间取决于主备集群的网络环境，在理想的网络环境下，为毫秒级别延时。所以在进行手动切换时，需要先停止对主集群的读写，并等待所有数据备份完成，再切换服务到备集群。主备切换之后，1个小时内不能再进行主备切换，且需要清理原集群数据并重新设置备集群信息。

同城双集群主备容灾中，应用访问主备集群的表格存储的服务域名不变，即发生切换之后，应用程序无需更改。异地双集群主备容灾中，主备集群的服务域名会不同，在发生切换情况下，应用程序需要更改访问表格存储的域名。

下图为包含了双集群主备容灾的多集群架构图，其中的 OTS 集群Y (主) 和 OTS 集群Y (备) 实现了主备容灾：



在单集群架构中，表格存储的所有数据均有3份副本，并保证数据在三份副本均写入磁盘之后再返回结果给用户，依次来保证单集群的高可靠。

在同城及异地灾备机制中，表格存储的服务节点会将数据按照写入的时间将数据以Replication方式异步发送到备集群的前端服务节点上，前端服务节点收到Replication请求之后，会按照正常的写入流程将数据写入备集群。

这种容灾方式中，最重要的两个内容就是数据同步和切换。在最初搭建双集群时，数据同步包括一次全量同步和实时的增量同步。在切换后，备集群开始提供服务，然后等主集群恢复后重新建立同步关系，并最终可以再切换回主集群。

需要额外说明的是，这里的数据同步是异步的，一条数据在主集群写入成功后即返回给客户端，此时可能还未同步给备集群。异步同步会带来一部分数据一致性问题，业务需要明确这种数据不一致带来的影响，并有相应的措施。

具体来说，在故障时刻进行主备切换时，需要假设主集群还有一部分数据未同步给备集群，那么：

- 原来的主集群上拥有切换前写入的所有数据，还包括切换前未同步给备集群的数据。

- 除了切换前未同步的一部分数据外，新的主集群(即备集群)上拥有所有写入的数据。

在正常的网络情况下，表格存储主备机制提供的RPO (Recovery Point Objective，复原点目标，最大数据丢失量) 为1分钟，服务恢复时间RTO (Recovery Time Objective，复原时间目标) 点为5分钟。

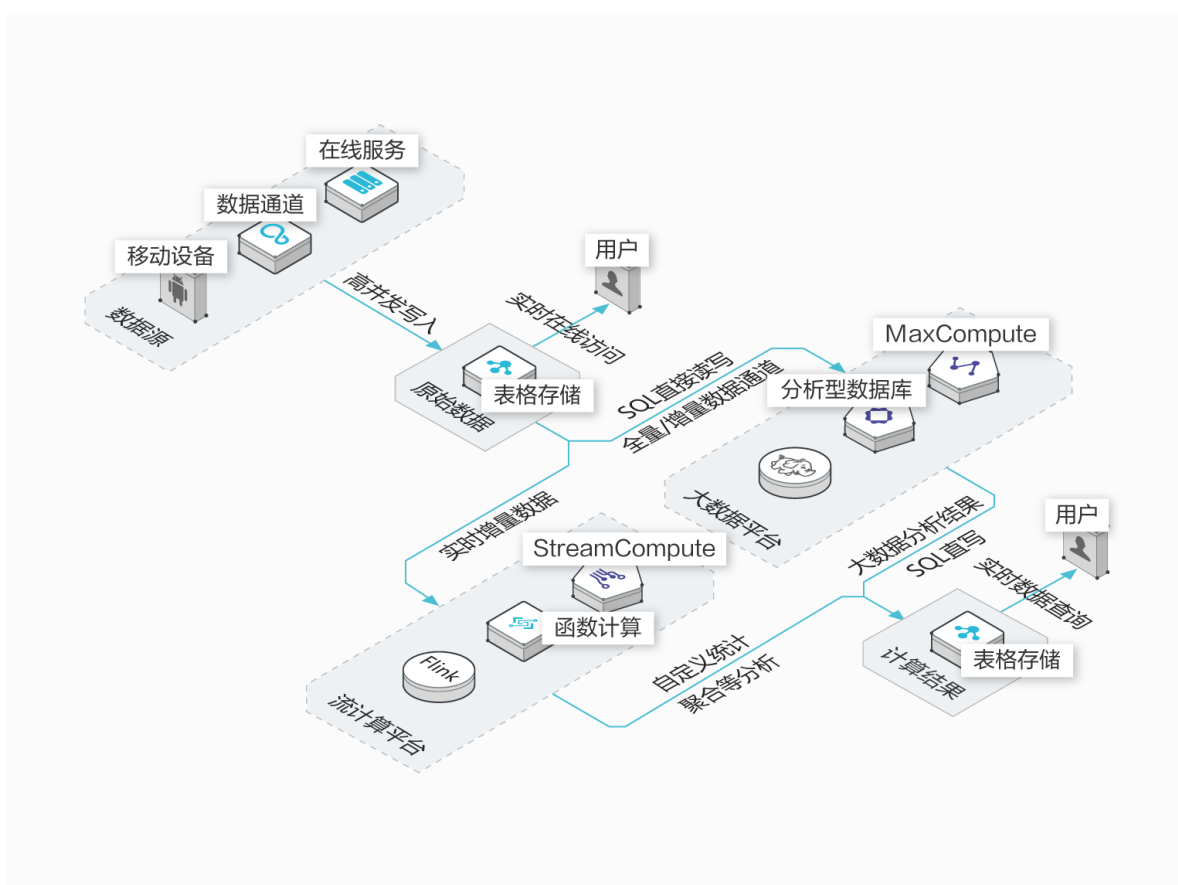
4.5 应用场景

• 场景一：海量数据存储与分析

表格存储提供低成本、高并发、低延时的海量数据存储与在线访问，提供增量以及全量数据通道并支持MaxCompute 等大数据分析平台的 SQL 直读直写，高效的增量流式读接口让数据轻松完成实时流计算。

能够提供：

- 单表 10PB 级数据量、万亿条记录数。
- 支持多种大数据计算平台、流计算以及实时计算服务。
- 弹性资源按量付费，提供高性能与容量型两种规格的实例，满足不同的业务需要。

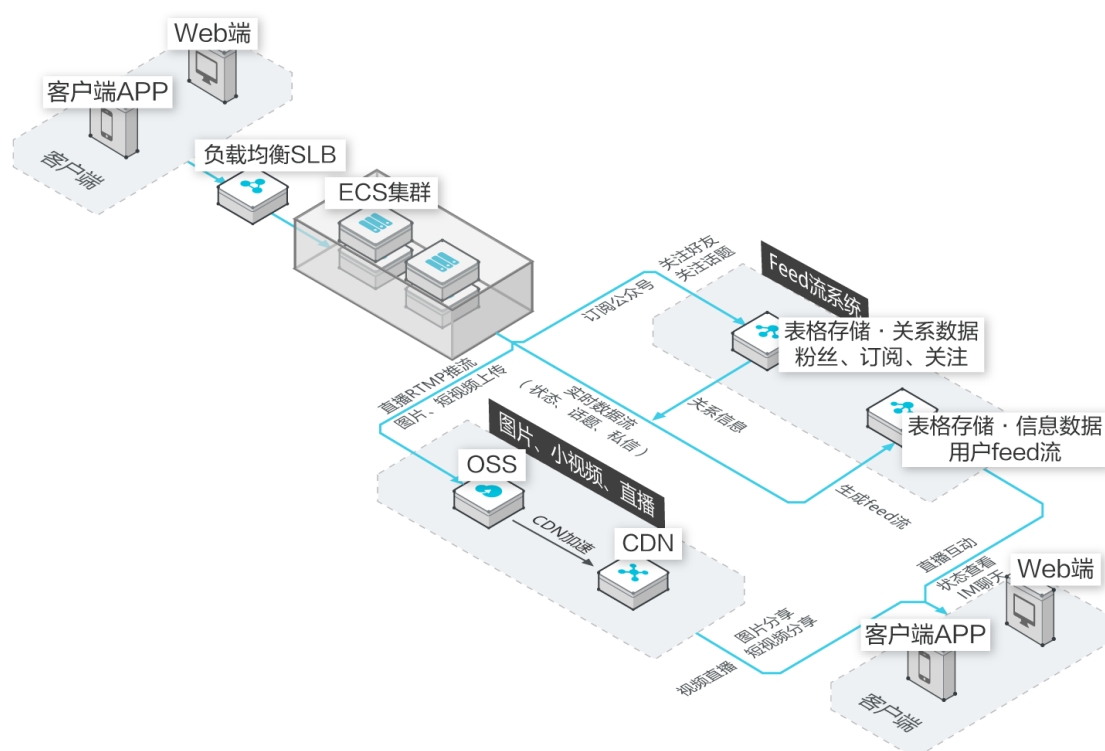


• 场景二：互联网社交 Feed 流

使用表格存储来存储大量的 IM 聊天、以及评论、跟帖和点赞等社交 Feed 流信息，表格存储的弹性资源按量付费能够以较低的成本满足访问波动明显大并发低延时的需要。

能够提供：

- 内置 PK 自增列，简化外部系统依赖。
- 高性能实例平均读写性能不受数据量大小影响。
- 高可靠的海量消息存储，消息多终端同步。

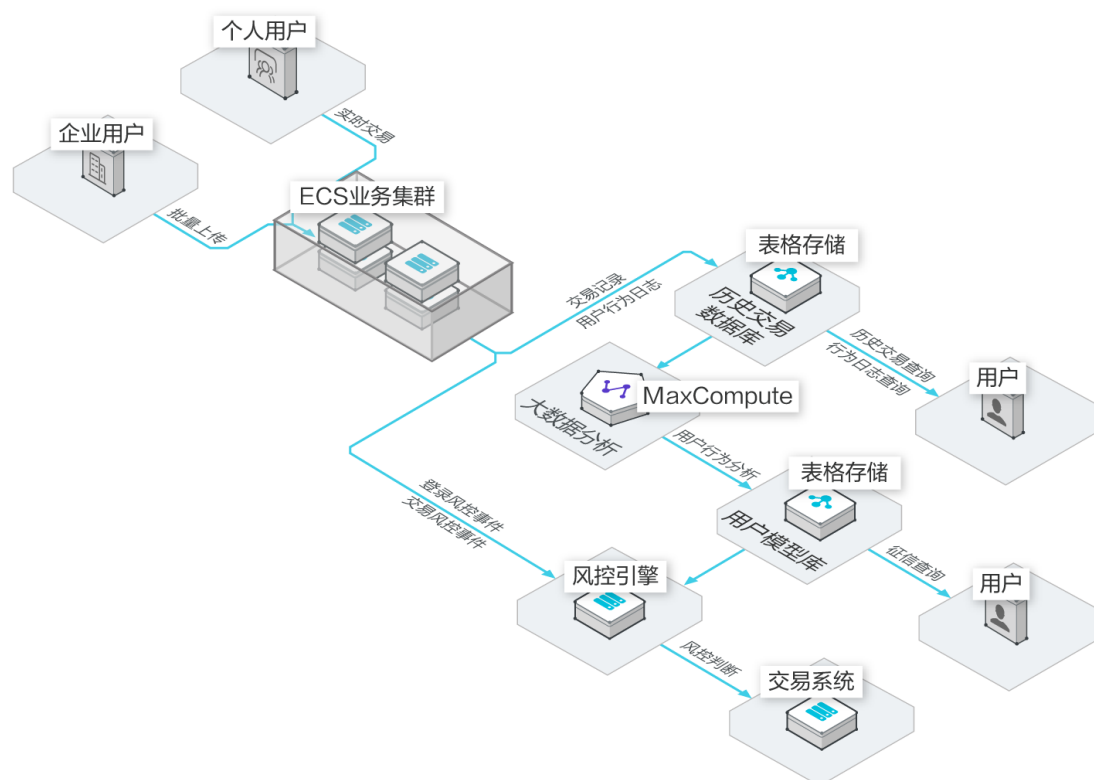


• 场景三：海量交易记录与用户模型的存储和实时查询

低延时、高并发，弹性资源按量付费让您的风控系统永远工作在最佳状态，牢牢控制交易风险，灵活的数据结构能够让业务模式跟随市场需求快速迭代。

能够提供：

- 单表万亿条记录数，轻松存储全量历史交易记录。
- 3副本强一致保证数据安全。
- Schema-free 模型，属性列字段随用随加，适应业务快速发展。

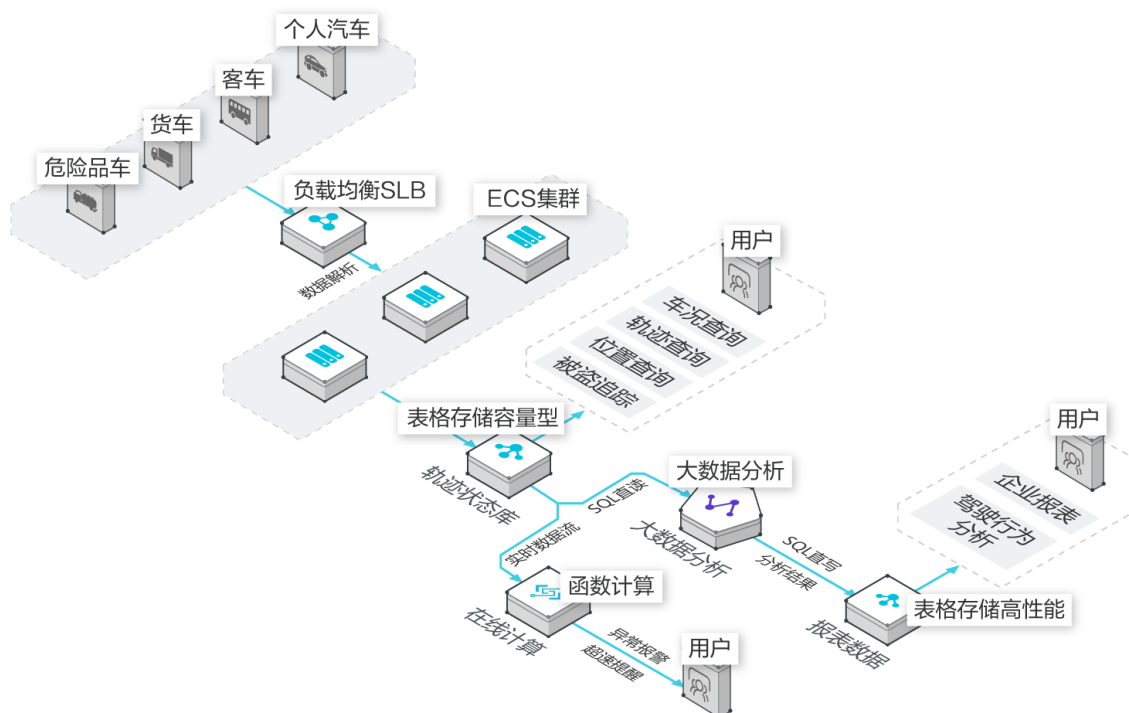


• 场景四：海量、高效、灵活的车联网数据存储

单表支撑 PB 级数据量，无需分库分表，简化业务逻辑，schemafree 数据模型轻松接入不同车辆设备的监控数据。与多种大数据分析平台、实时计算服务等无缝结合，轻松完成实时在线查询以及业务报表分析。

能够提供：

- 单表支撑 10PB 级数据量，无需分库分表，简化业务逻辑。
- 车况、轨迹查询性能稳定可预期。
- schemafree 数据模型轻松接入不同车辆设备的监控数据。

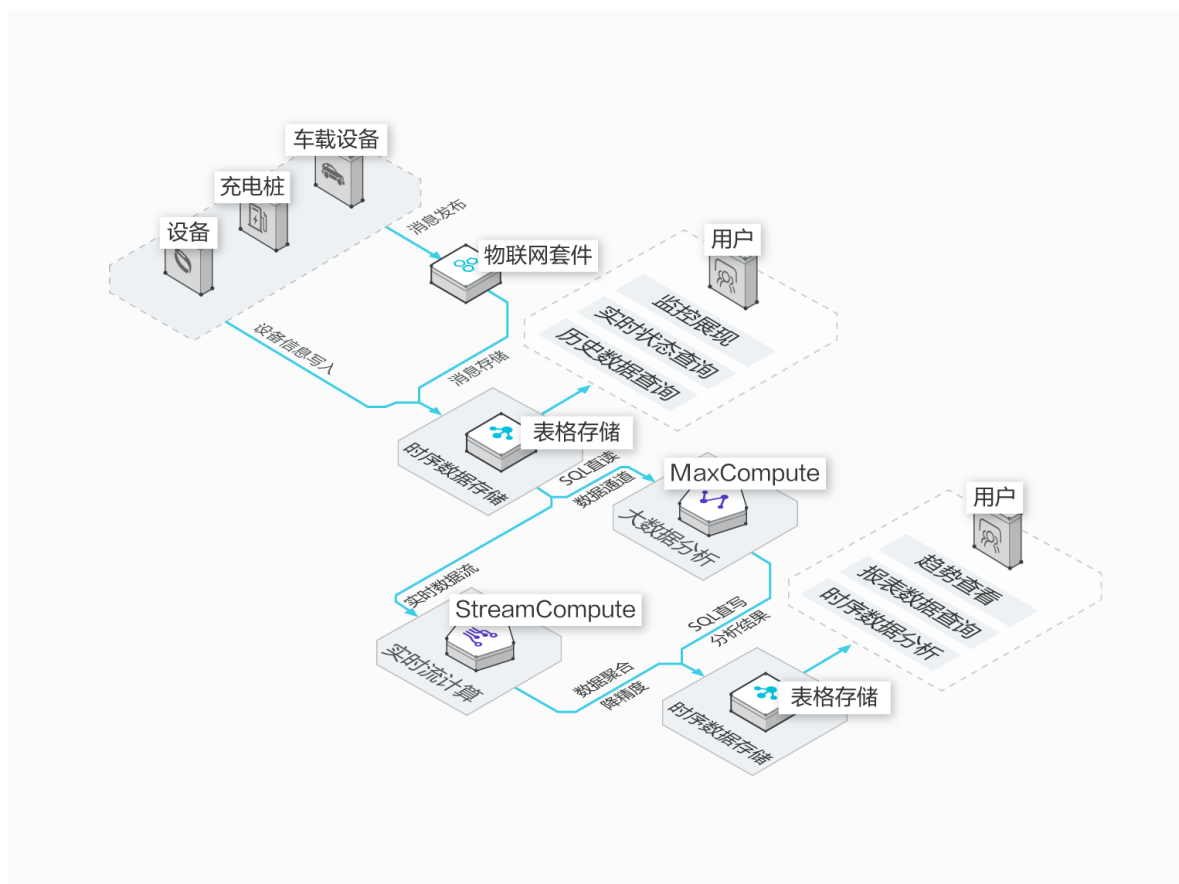


• 场景五：海量物联网数据存储，高效的查询与分析

单表 PB 级数据规模让表格存储轻松满足 IoT 设备、监控系统等时序数据的存储需求，大数据分析 SQL 直读以及高效的增量流式读接口让数据轻松完成离线分析与实时流计算。

能够提供：

- 单表支撑 10PB 级数据量满足超大规模 IoT 设备、监控系统的数据写入与存储需求。
- 对接多种离线、流式数据分析平台，一份数据同时满足不同业务的分析计算场景。
- 按量付费，超低使用成本，支持数据生命周期管理。

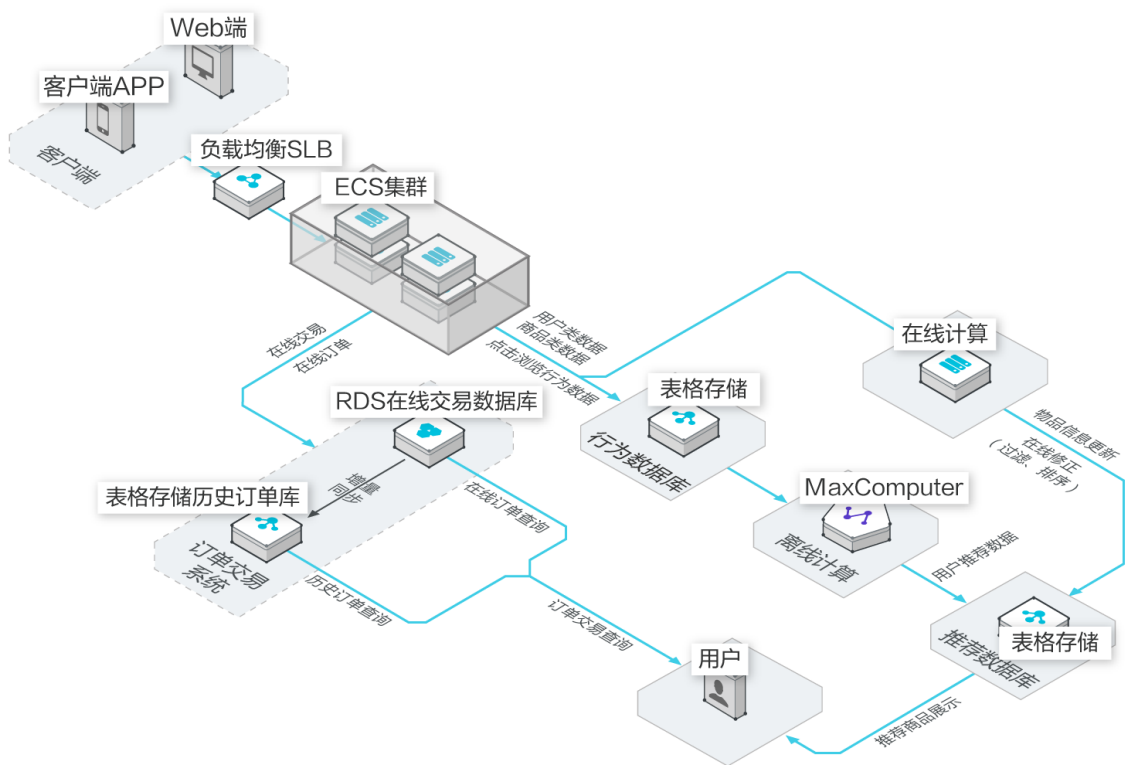


• 场景六：海量电商交易订单与用户推荐数据库

大量的历史交易订单使用表格存储让您无需担心数据规模与访问性能，配合大数据计算服务，轻松实现精准营销，弹性资源、按量付费，让您从容应对所有用户在线高峰时刻。

能够提供：

- 数据规模与访问并发自动扩展，满足波峰波谷各个时段的访问需求。
- 支持多种大数据分析平台直接分析用户行为。
- 海量交易订单毫秒级的查询延时。



4.6 使用限制

表 4-1: 使用限制为表格存储的使用限制汇总，部分限制范围表明用户能够使用的最大值，而不是建议值。为保证更好的性能，请合理设计表结构和单行数据大小。

表 4-1: 使用限制

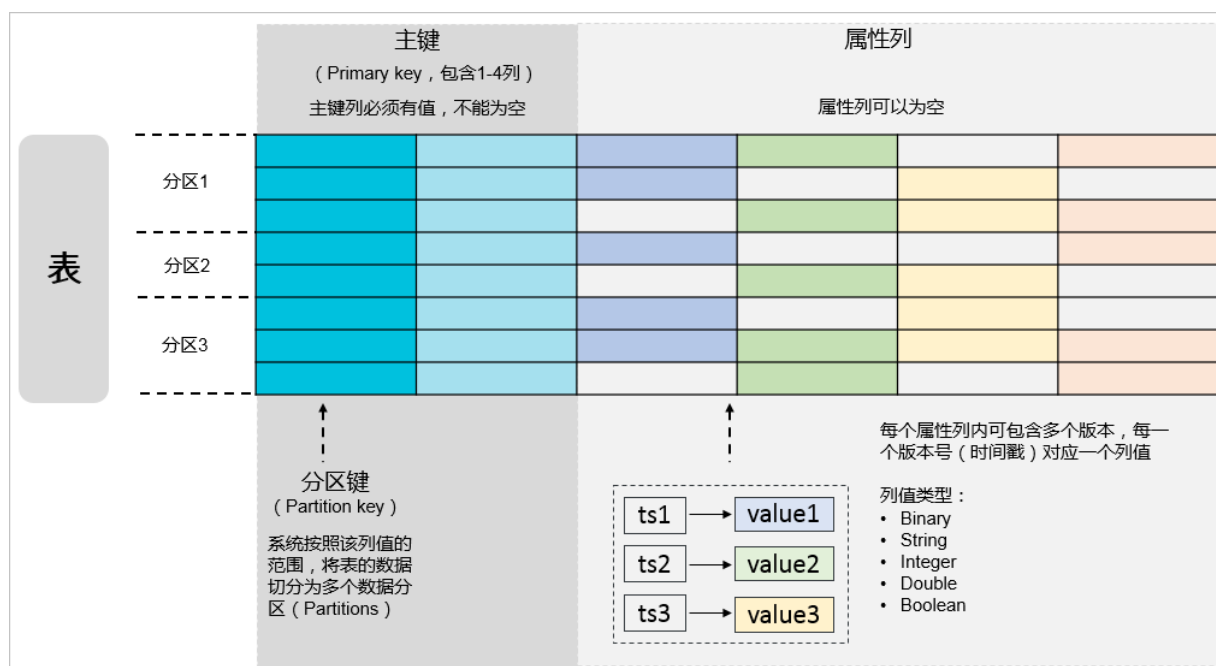
限制项	限制范围	说明
一个阿里云用户账号下可以保有实例数	1024	如有需求提高上限，请联系管理员。
一个实例中表的个数	1024	如有需求提高上限，请联系管理员。
主键包含的列数	1-4	至少 1 列，至多 4 列。
String 类型主键列的列值大小	1 KB	单一主键列 String 类型列的列值大小上限 1 KB。
String 类型属性列的列值大小	2 MB	单一属性列 String 类型列的列值大小上限 2 MB。

限制项	限制范围	说明
Binary 类型主键列的列值大小	1 KB	单一主键列 Binary 类型列的列值大小上限 1 KB。
Binary 类型属性列的列值大小	2 MB	单一属性列 Binary 类型列的列值大小上限 2 MB。
一行中属性列的个数	不限制	不限制单一行拥有的属性列个数。
一次请求写入的属性列的个数	1024 列	PutRow、UpdateRow或BatchWriteRow操作时，单行写入的属性列的个数不能超过 1024 列。
单行数据大小	不限制	不限制单一行中所有列名与列值总和大小。
读请求中 columns_to_get 参数的列的个数	0-128	读请求一行数据中获取的列的最大个数。
单表 UpdateTable 的次数	上调：无限制 下调：无限制	需要遵循单表的调整频率限制。
单表 UpdateTable 的频率	每 2 分钟 1 次	单表在 2 分钟之内，最多允许调整 1 次预留读/写能力值。
BatchGetRow 一次操作请求读取的行数	100	N/A
BatchWriteRow 一次操作请求写入行数	200	N/A
BatchWriteRow 一次操作的数据大小	4 MB	N/A
GetRange 一次返回的数据	5000 行或者 4 MB	一次返回数据的行数超过 5000 行，或者返回数据的数据大小大于 4 MB。以上任一条件满足时，超出上限的数据将会按行级别被截掉并返回下一行数据主键信息。
一次 HTTP 请求 Request Body 的数据大小	5 MB	N/A

4.7 基本概念

数据模型

表格存储的数据模型概念包括表、行、主键和属性。如下图所示：



最大版本数

最大版本数 (Max Versions) 是数据表的一个属性, 表示该数据表中的属性列能够保留多少个版本的数据。当一个属性列的版本个数超过 Max Versions 时, 最早的版本将被异步删除。

数据生命周期

数据生命周期 (Time To Live, 简称 TTL) 是数据表的一个属性, 即数据的存活时间, 单位为秒。表格存储会在后台对超过存活时间的数据进行清理, 以减少用户的数据存储空间, 降低存储成本。

有效版本偏差

有效版本偏差 (Max Version Offset) 是数据表的一个属性, 单位为秒。

为了防止非期望的写入, 服务端在处理写请求时会对属性列的版本号进行检查。当版本号小于当前写入时间减去 Max Version Offset, 或者大于等于当前写入时间加上 Max Version Offset 的值时, 该行数据写入失败。

属性列的有效版本范围为: [数据写入时间 - 有效版本偏差, 数据写入时间 + 有效版本偏差)。数据写入时间为 1970-01-01 00:00:00 UTC 时间到当前写入时间的秒数。属性列版本号为毫秒, 其除以 1000 换算成秒之后必须属于这个范围。

主键和属性

主键是表中每一行的唯一标识。主键由 1 到 4 个主键列组成。创建表的时候，必须明确指定主键的组成、每一个主键列的名字和数据类型以及它们的顺序。主键列的数据类型只能是 String、Integer 和 Binary。如果为 String 或者 Binary 类型，长度不超过 1 KB。

属性是属性存放行的数据。每一行包含的属性列个数没有限制。

读/写吞吐量

读/写吞吐量的单位为读服务能力单元和写服务能力单元，简称 CU（Capacity Unit），是数据读写操作的最小计费单位。

地域

地域（Region）是指物理的数据中心，表格存储服务会部署在多个阿里云地域中，用户可以根据自身的业务需求选择不同地域的表格存储服务。

实例

实例（Instance）是用户使用和管理表格存储服务的实体，对应于关系型数据库的 Database。实例是表格存储资源管理的基础单元，表格存储对应用程序的访问控制和资源计量都在实例级别完成。

服务地址

每个实例对应一个服务地址（EndPoint），应用程序在进行表和数据操作时需要指定服务地址。

Stream

Stream是数据表的一个属性，可用于增量数据流的实时增量分析和数据增量同步。

SATA

SATA（Serial ATA）口的硬盘又叫串口硬盘。Serial ATA采用串行连接方式，具备了更强的纠错能力，旨在提高数据传输的可靠性。

5 文件存储NAS

5.1 什么是文件存储NAS

阿里云文件存储 (Network Attached Storage , 简称 NAS) 是面向阿里云 ECS 实例、E-HPC 和容器服务等计算节点的文件存储服务, 提供标准的文件访问协议, 用户无需对现有应用做任何修改, 即可使用具备无限容量及性能扩展、单一命名空间、多共享、高可靠和高可用等特性的分布式文件系统。

用户创建 NAS 文件系统实例和挂载点后, 即可在 ECS、E-HPC 和容器服务等计算节点内通过标准的 NFS 协议挂载文件系统, 并使用标准的 Posix 接口对文件系统进行访问。多个计算节点可以同时挂载同一个文件系统, 共享文件和目录。

5.2 功能特性

无缝集成

NAS 支持 NFSv3 及 NFSv4 协议, 并使用标准的文件系统语义访问数据。主流的应用程序及工作负载无需任何修改即可无缝配合使用。

共享访问

一个 NAS 文件系统实例可以被多个计算节点同时访问, 非常适合跨多个 ECS、E-HPC 或容器服务实例部署的应用程序访问相同数据来源的应用场景。

安全控制

NAS 具有网络隔离 (专有网络) / 用户隔离 (经典网络)、文件系统标准权限控制、权限组访问控制和 RAM 主子账号授权等多种安全机制, 从而保证文件系统数据安全。

线性扩展的性能

NAS 能够为应用工作负载提供高吞吐量与高 IOPS、低时延的存储性能, 同时, 其性能与容量成线性关系, 可满足业务增长时对更高容量与存储性能的诉求。

5.3 产品优势

多共享

同一个文件系统可以同时挂载到多个计算节点上, 共享访问, 节约大量拷贝和同步成本。

高可靠

提供高数据可靠性，相比自建NAS存储，可以大量节约维护成本，降低数据安全风险。

弹性伸缩

单个文件系统容量上限达10 PB，轻松应对业务的随时扩容和缩容。

高性能

单个文件系统吞吐性能随存储量线性扩展，相比购买高端NAS存储设备，大幅降低成本。

易用性

支持NFSv3和NFSv4协议，无论是在ECS实例内，还是在E-HPC和容器服务等计算节点中，都可通过标准的Posix接口对文件系统进行访问操作。

5.4 应用场景

企业办公文件共享

企业员工办公需要访问和共享相同的数据集，管理员可创建NAS文件系统，为组织中的个人提供数据访问，并可设置文件或目录级别的用户和用户组权限。

数据备份

用户希望将线下机房的数据备份到云上，同时要求云上的存储服务兼容标准的文件访问接口。

服务器日志共享

将多个计算节点上的应用服务器日志存放在共享的文件存储上，方便后续的日志集中处理与分析。

5.5 基本概念

挂载点

挂载点是文件系统实例在专有网络或经典网络内的一个访问目标地址，每个挂载点都对应一个域名，用户mount时通过指定挂载点的域名来挂载对应的NAS文件系统到本地。

权限组

权限组是NAS提供的白名单机制，通过向权限组内添加规则来允许IP地址或网段以不同的权限访问文件系统。



说明：

每个挂载点都必须与一个权限组绑定。

授权对象

授权对象是权限组规则的一个属性，代表一条权限组规则被应用的目标。在专有网络内，授权对象可以是一个单独的IP地址或一个网段；在经典网络内，授权对象只能是一个单独的IP地址（一般为ECS实例的内网IP地址）。

6 块存储

6.1 什么是弹性块存储ABS

弹性块存储 (Apsara Block Storage , 简称ABS) 是基于分布式文件系统实现的一个高可靠、高可用的块设备存储服务。

它可以部署在通用X86架构的服务器上，将多台服务器的内置本地硬盘组成虚拟的存储资源池，对外提供块存储访问能力。

ABS提供基于标准接口协议的开放 API，满足客户对接Openstack云平台的业务需求，支持KVM、容器等计算虚拟化技术。

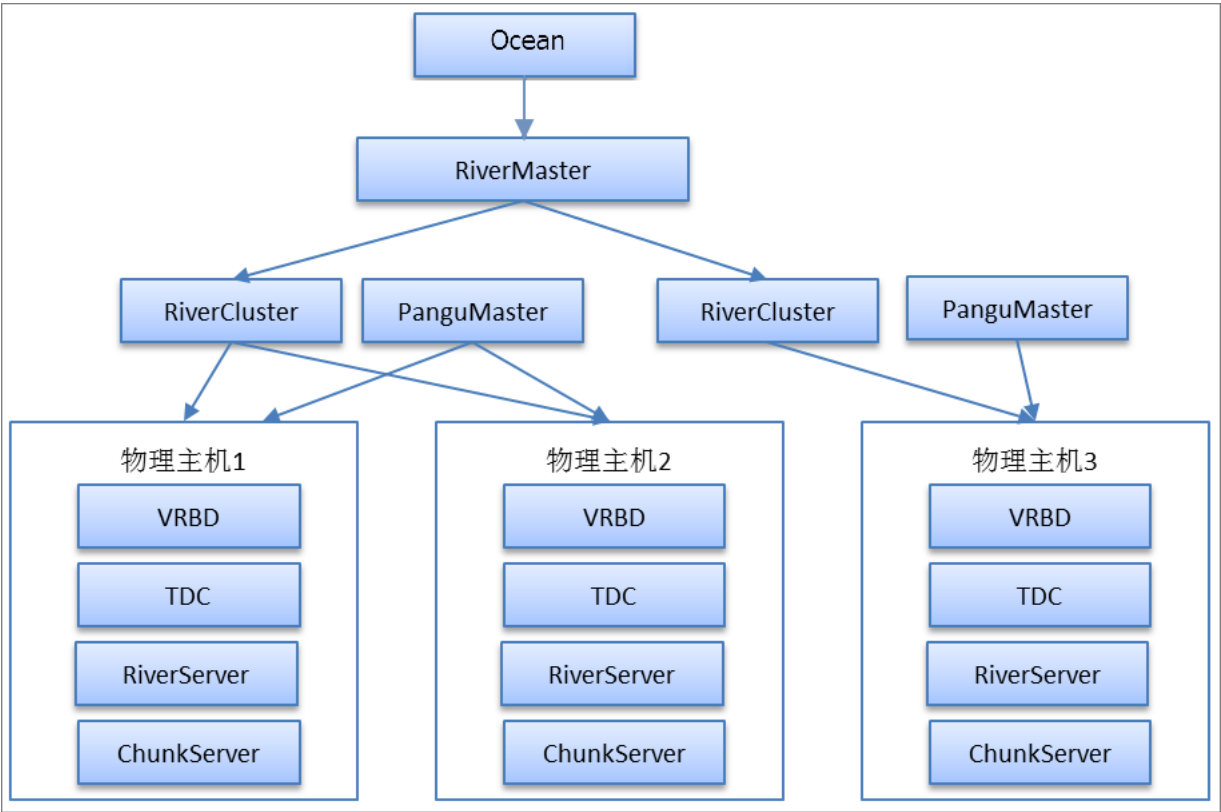
ABS支持快照、镜像、多租户、自动精简配置等高级数据服务功能，支持从最小规模无缝弹性扩展存储资源池，支持HDD机械硬盘和SSD闪存，满足不同业务模型下对存储性能和容量的多样性要求。

6.2 产品架构

6.2.1 技术架构

块存储的技术架构如[图 6-1: ABS软件架构](#)所示。

图 6-1: ABS软件架构



其中，各组件名称及其功能如表 6-1: 各组件功能说明所示。

表 6-1: 各组件功能说明

组件	功能
VRBD	虚拟块设备驱动，部署在计算节点，通过VRBD能创建虚拟的块设备，能够被挂载到物理主机上，能进行IO读写，完全和物理介质的磁盘一样的使用体验。
TDC	TDC负责将后端存储和前端组件如VRBD桥接起来，承载块设备IO，同时提供形态丰富的CLI接口，如磁盘的创建、销毁，快照的创建与回滚等。
PanguMaster	盘古中最重要的角色，维护了文件和数据块之间的映射quota数据、chunkserver元数据、和checkpoint等。盘古通过多master机制保证master的高可用性，同时使得盘古具备热升级的能力。master之间log的同步和主master的选举采用了Paxos算法来保证其一致性。
ChunkServer	运行在数据物理服务器上，主要的职责是管理本地的硬盘和支撑对硬盘的读写删操作。

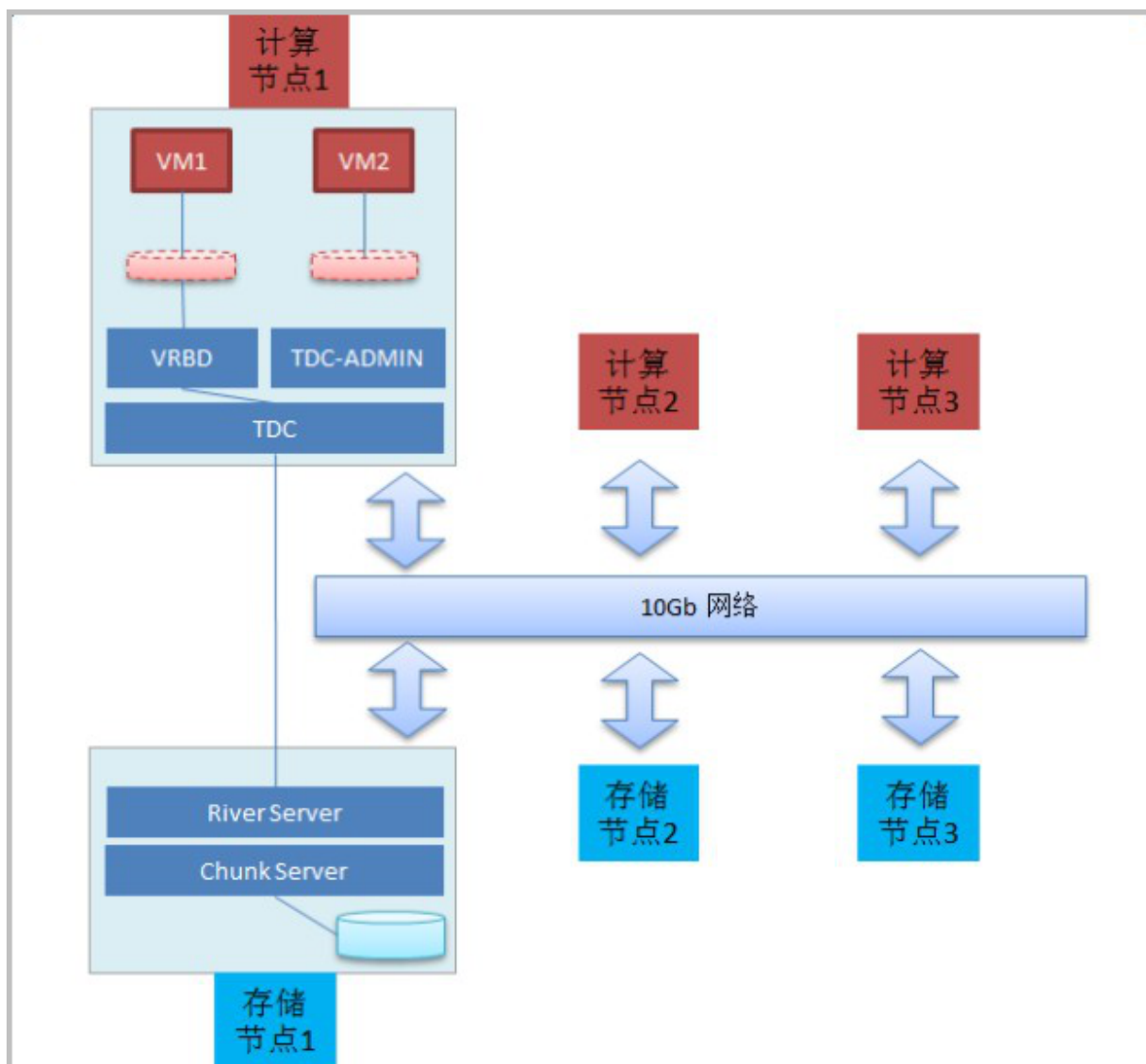
组件	功能
Ocean	Oceans是块存储的管理服务化平台，统筹块存储业务的整个控制流，并提供Restful API管理接口。通过这些接口，用户可以对ABS进行编程管理，实现软件定义的存储。
RiverMaster	在块存储系统中管控角色，负责磁盘在集群见的调度，包括磁盘创建的的调度、动态负载调度等，以及相关的管理工作，如存储库存水位管理等。
RiverCluster	负责管理一个集群，RiverCluster通过心跳维护RiverServer的状态。当RiverServer不可服务时，负责将RiverServer的磁盘动态迁移到另一台RiverServer上。
RiverServer	运行在每台物理主机上，RiverServer负责块层的IO读写处理，同时将IO请求转化后投递到Pangu文件层。

6.2.2 部署架构

阿里云ABS块存储系统采用分离部署模式，即计算端节点只部署计算端的软件组件，存储端节点只部署存储端的软件组件。如图 6-2: 分离部署模式所示。

分离部署模式中，计算节点仅部署与计算有关的软件组件，如VRBD、TDC和TDC_ADMIN等；存储节点仅部署与存储有关的软件组件如RiverServer和ChunkServer等。

图 6-2: 分离部署模式



6.3 功能特性

6.3.1 概述

本节主要介绍ABS的功能和特性。

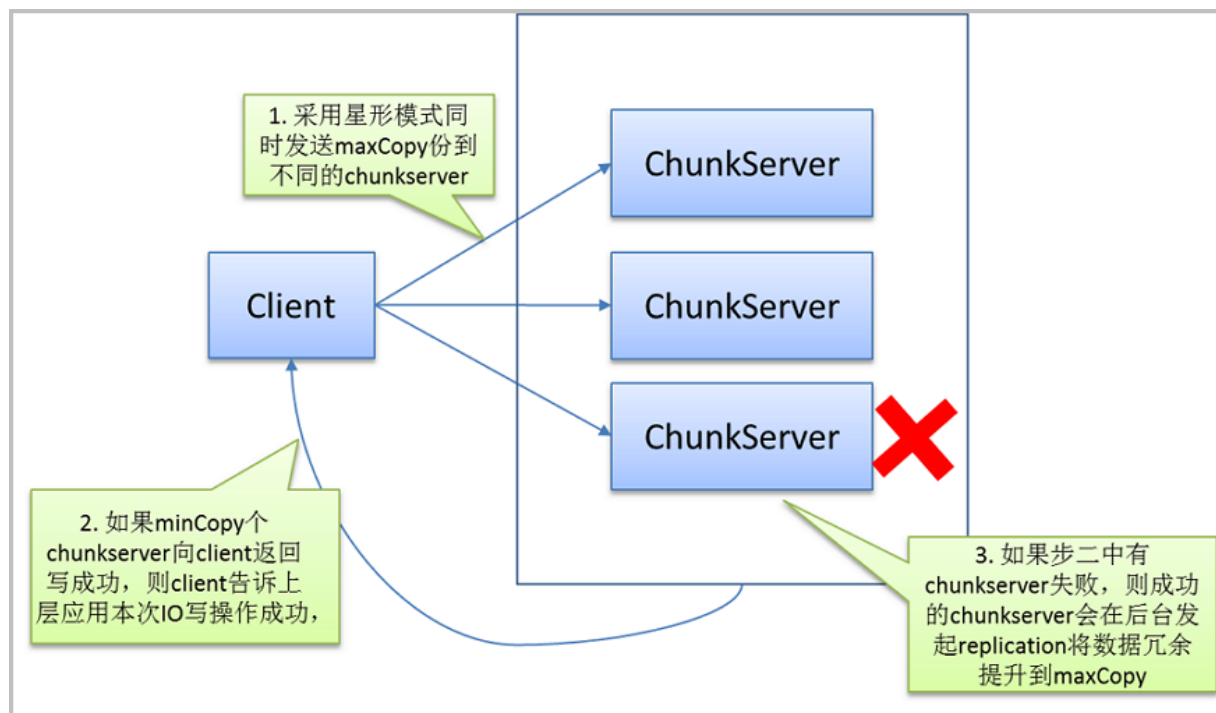
6.3.2 分布式文件系统

盘古采用多副本机制实现分布式文件系统，如图 6-3: 多副本机制所示，盘古多副本机制的内部实现原理如下：

1. Client采用星形模式，同时向服务器发送maxCopy份副本到不同的chunkserver。
2. 如果minCopy个chunkserver向Client返回写成功，则Client报告上层应用，本次IO写操作成功。

3. 如果上一步中有chunkserver失败，则成功的chunkserver会在后台发起replication，将数据冗余提升到maxCopy。

图 6-3: 多副本机制

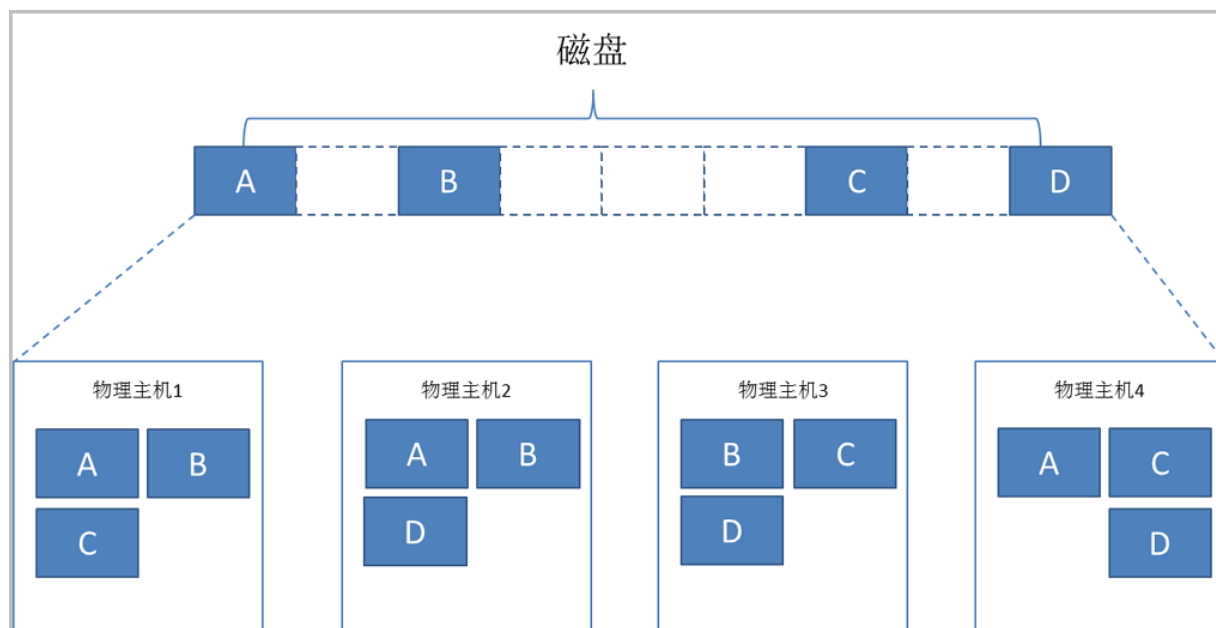


6.3.3 自动精简装置

ABS默认支持精简配置，以降低存储的成本，提高资源利用率，如图 6-4: 自动精简装置所示。

虽然磁盘有9个块的容量，用户磁盘真正写入的只有4个块。落在底层盘古，也并不是按照9个块空间进行分配与存储，而只是存储了真实有数据的块，并进行索引维护，从而实现了精简配置。

图 6-4: 自动精简装置



6.3.4 磁盘快照功能

当您对一个磁盘进行快照的时候，当前的所有数据都会被复制到对应的快照文件；在任何需要的时候，用户都可以回滚到这个保存的快照。Ocean不仅支持基本的快照管理功能，同时支持跨Azone的本地快照迁移，以提高快照的访问速度，提升用户体验。

ABS卷的快照具备如下特点：

- 快照通过增量技术实现，创建的第一个快照是卷的数据的全量备份，后续创建的快照均是自上次卷的数据量，进行增量备份。
- 快照是按照一定大小进行切割，并将数据有变化的数据块压缩后，备份到对象存储系统中。
- 快照的指令下发是秒级完成，快照数据同步是在系统后台进行的，不需要停止卷的运行状态。
- 快照删除后，不会影响后续快照的数据完备性，被删除的快照的数据块，只有引用计数为0时，才会被回收。

6.3.5 数据重建

盘古具备强大的数据保护和智能恢复机制。

数据存储时已被切片打散到多个节点上，这些分片数据通过算法被分配在不同的存储节点、不同机柜之间以提供最大数据安全保障，同时数据存储时采用多副本技术，支持两副本或三副本，数据会自动保存多份，每一个分片的不同副本也被分散保存到不同的存储节点上。

在硬件发生故障导致数据不一致时，盘古一方面通过异步机制来保证服务的可用性，另一方面通过优化的自检机制，能够快速比较不同节点上的副本分片，自动发现数据故障，并在发现故障后启动数据修复机制。

在后台修复数据，由于数据被分散到多个不同的存储节点上保存，数据修复时，在不同的节点上同时启动修复，每个节点上只需修复一小部分数据，多个节点并行工作，配合精准的流控机制不仅能充分利用剩余网络资源，还可以有效避免单个节点修复大量数据所产生的性能瓶颈，对上层业务的影响做到最小化。

6.4 典型业务场景

6.4.1 概述

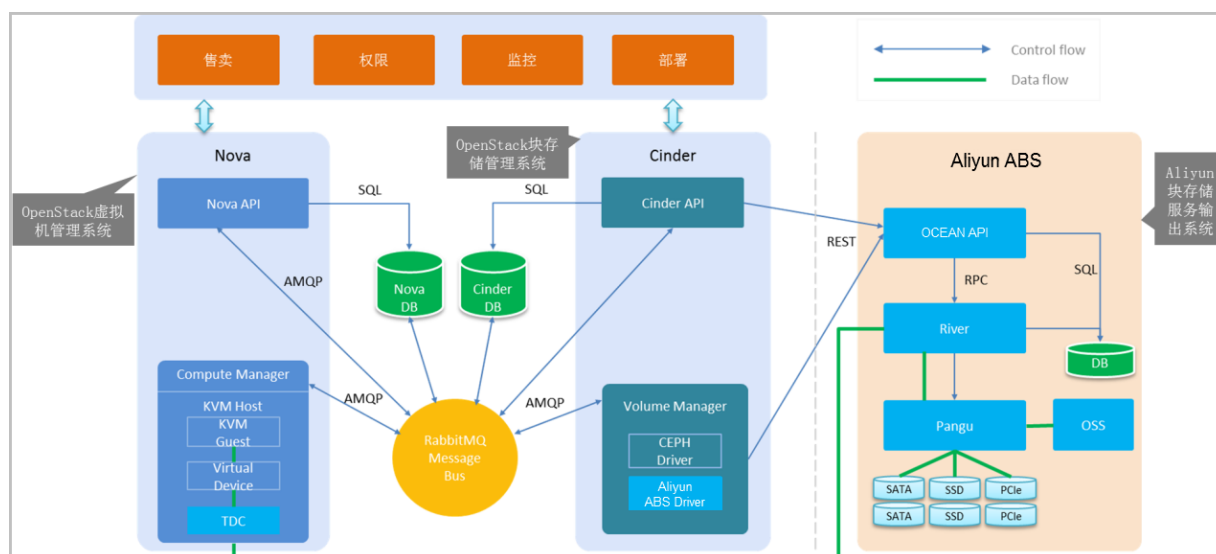
典型业务场景主要包括OpenStack场景和Docker场景。

6.4.2 OpenStack场景

阿里云ABS块存储服务作为一个基础的云存储基础设施，能很敏捷地与现有源云平台如OpenStack进行集成。

如图 6-5: *OpenStack场景* 所示，OpenStack系统中的存储子系统是Cinder，阿里云为Cinder开发了插件，将阿里云ABS块存储和OpenStack进行无缝集成。这样用户只需关注计算端的调度和管控，块存储服务由阿里云ABS来承载，并向上层提供丰富的功能，如磁盘类功能、快照类功能、镜像类功能、运维类功能及监控类功能。

图 6-5: OpenStack场景

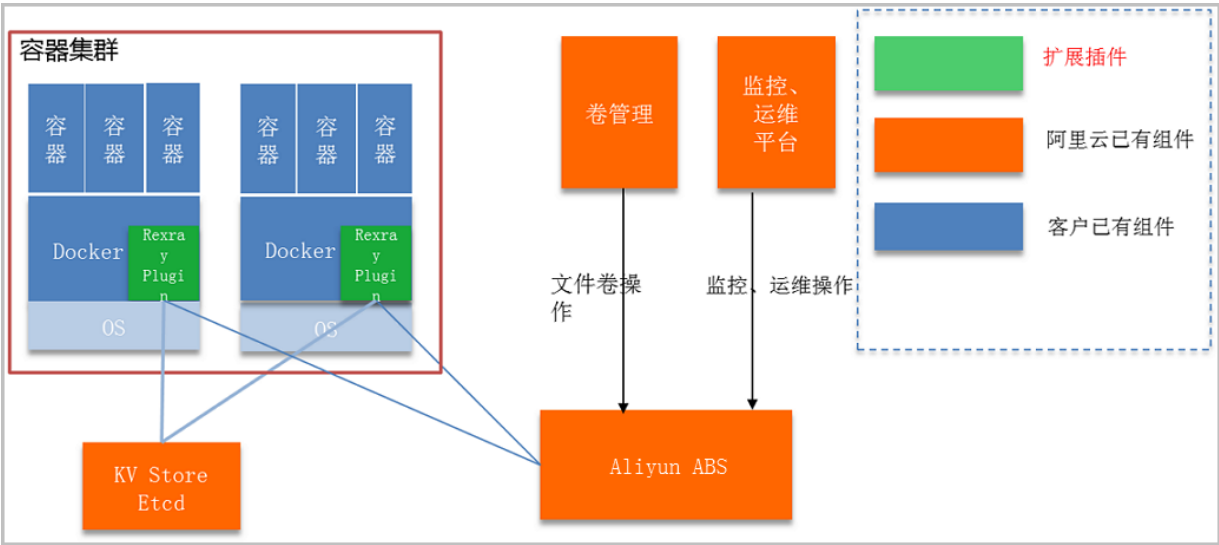


6.4.3 Docker场景

阿里云ABS同样能支持Docker容器技术。

如图 6-6: *Docker场景*所示，通过实现一个基于阿里云ABS的插件，即可将阿里云ABS和Docker系统无缝集成在一起。

图 6-6: Docker场景



7 云数据库RDS版

7.1 什么是关系型数据库RDS

阿里云关系型数据库RDS (Relational Database Service) 是一种稳定可靠、可弹性伸缩的在线数据库服务。基于阿里云分布式文件系统和高性能存储，提供了容灾、备份、恢复、监控、迁移等方面的全套解决方案，彻底解决数据库运维的烦恼。

云数据库MySQL版基于阿里巴巴的MySQL源码分支，经过双十一高并发、大数据量的考验，拥有优良的性能。云数据库MySQL版支持实例管理、帐号管理、数据库管理、备份恢复、白名单、透明数据加密以及数据迁移等基本功能。除此之外还提供如下高级功能：

- **只读实例**：在对数据库有大量读请求和少量写请求时，单个实例可能无法承受读取压力，为了实现读取能力的弹性扩展，减少单个实例的压力，云数据库MySQL 5.6版的实例支持只读实例，利用只读实例满足大量的数据库读取需求，以此增加应用的吞吐量。
- **CloudDBA数据库性能优化**：针对SQL语句的性能、CPU使用率、IOPS使用率、内存使用率、磁盘空间使用率、连接数、锁信息、热点表等，CloudDBA提供了智能的诊断及优化功能，能最大限度发现数据库存在的或潜在的健康问题。CloudDBA的诊断基于单个实例，该诊断会提供问题详情及相应的解决方案，可为您管理实例运行状况带来极大的便利。
- **数据压缩**：云数据库MySQL 5.6版支持通过TokuDB存储引擎压缩数据。经过大量测试表明，数据表从InnoDB存储引擎转到TokuDB存储引擎后，数据量可以减少80%到90%，即2T的数据量能压缩到400G甚至更低。除了数据压缩外，TokuDB存储引擎还支持事务和在线DDL操作，可以很好兼容运行于MyISAM或InnoDB存储引擎上的应用。

7.2 产品优势

7.2.1 易于使用

云数据库RDS拥有即开即用、按需升级、透明兼容和管理便捷的优点。

即开即用

您可以通过API进行RDS规格定制，创建后RDS实时生产出目标实例。

按需升级

随着数据库压力和数据存储量的变化，您可以灵活调整实例规格，且升级期间RDS不会中断数据链路服务。

透明兼容

RDS与原生数据库引擎的使用方法一致，您无需二次学习，上手即用。另外RDS兼容您现有的程序和工具。使用通用的数据导入导出工具即可将数据迁移到RDS，迁移过程中的人力开销非常低。

管理便捷

阿里云负责RDS的日常维护和管理，包括但不限于软硬件故障处理、数据库补丁更新等工作，保障RDS运转正常。您也可自行通过阿里云控制台完成数据库的增加、删除、重启、备份、恢复等管理操作。

7.2.2 高性能

云数据库RDS通过参数优化、SQL优化、高端硬件来实现高性能。

参数优化

所有RDS实例的参数都是经过多年的生产实践优化而得，在RDS实例的生命周期内，我们持续对其进行优化，确保RDS一直基于最佳实践在运行。

SQL 优化

针对您的应用场景特点，RDS会锁定效率低下的SQL语句并提出优化建议，以便您优化业务代码。

高端硬件投入

专有云中RDS所使用的服务器硬件经过多方评测，保证服务器运行稳定性。

7.2.3 高安全性

云数据库RDS通过防DDoS攻击、访问控制策略和系统安全实现高安全性。

防DDoS攻击



说明：

此功能需开通阿里云的安全产品。

当您使用外网连接和访问RDS实例时，可能会遭受DDoS攻击。当RDS安全体系认为您的实例正在遭受DDoS攻击时，会首先启动流量清洗的功能，如果流量清洗无法抵御攻击或者攻击达到黑洞阈值时，将会进行黑洞处理。

流量清洗和黑洞处理的方法及触发条件如下：

- 流量清洗：

只针对外网流入流量进行清洗，处于流量清洗状态的RDS实例可正常访问。

流量清洗的触发和结束由系统自动完成，单个RDS实例满足以下任一条件即触发流量清洗：

- PPS (Packets Per Second) 达到3万；
 - BPS (Bits Per Second) 达到180Mb；
 - 每秒新建并发连接达到1万；
 - 激活并发连接数达到1万；
 - 非激活并发连接数达到10万。
- 黑洞处理：

只针对外网流入流量进行黑洞处理，处于黑洞状态的RDS实例不可被外网访问，此时应用程序通常也处于不可用状态。黑洞处理是保证RDS整体服务可用性的一种手段。

黑洞触发条件如下：

- BPS (Bits Per Second) 达到2GB；
- 流量清洗无效。

黑洞结束条件为：黑洞在2.5小时后自动解除。

访问控制策略

您可以设置允许访问RDS的IP地址，其他的IP地址将被拒绝访问。

每个账号只能查询、操作自己的数据库。

系统安全

RDS处于多层防火墙的保护之下，可以有效抵抗各种恶意攻击，保证数据的安全。

RDS服务器不允许直接登录，只开放特定的数据库服务所需要的端口。

RDS服务器不允许主动向外发起连接，只能接受被动访问。

7.2.4 高可靠性

云数据库RDS通过双机热备、多副本冗余、数据备份和数据恢复实现高可靠性。

双机热备

RDS采用热备架构，物理服务器出现故障后服务秒级完成切换，整个切换过程对应用透明。

多副本冗余

RDS 服务器中的数据构建于RAID之上，数据备份存储在OSS上。

数据备份

RDS提供自动备份的机制，您可以自行选择备份周期，也可以根据自身业务特点随时发起临时备份。

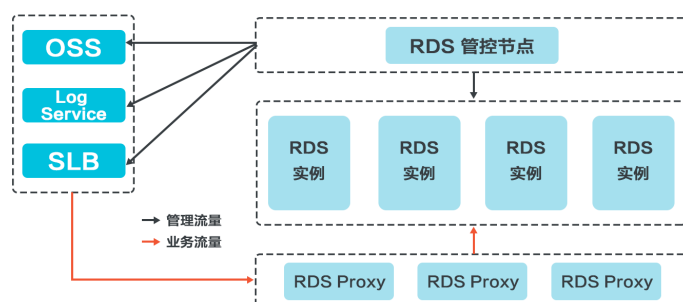
数据恢复

支持按备份集和指定时间点来创建克隆实例恢复数据，数据验证无误后即可将数据迁回RDS主实例，从而完成数据回溯。

7.3 产品架构

云数据库RDS版的系统架构如下。

图 7-1: RDS系统架构



7.4 功能特性

7.4.1 数据链路服务

云数据库RDS提供全数据链路服务，包括DNS、负载均衡等。

云数据库RDS使用原生的数据库引擎，对数据库的操作高度类似，基本没有学习成本，极大的方便了您访问使用数据库。

DNS

DNS模块提供域名到IP的动态解析功能，以便规避RDS实例IP地址改变带来的影响。在连接池中设置域名后，即使对应的IP地址发生了变化，仍然可以正常访问RDS实例。

例如，某RDS实例的域名为`test.rds.aliyun.com`，对应的IP地址为`10.10.10.1`。某程序连接池中设置为`test.rds.aliyun.com` 或 `10.10.10.1` 都可以正常访问RDS实例。

一旦该RDS实例发生了实例迁移或者版本升级后，IP地址可能变为**10.10.10.2**。如果程序连接池中设置的是域名**test.rds.aliyun.com**，则仍然可以正常访问RDS实例。但是如果程序连接池中设置的是IP地址**10.10.10.1**，就无法访问RDS实例了。

负载均衡

负载均衡模块提供实例IP地址（包括内网IP 和外网IP），以便屏蔽物理服务器变化带来的影响。

例如，某RDS实例的内网IP地址为**10.1.1.1**，对应的Proxy或者DB Engine运行在**192.168.0.1**上。在正常情况下，负载均衡模块会将访问**10.1.1.1**的流量重定向到**192.168.0.1**上。当**192.168.0.1**发生了故障，处于热备状态的**192.168.0.2**接替了**192.168.0.1**的工作。此时负载均衡模块会将访问**10.1.1.1**的流量重定向到 **192.168.0.2** 上，RDS实例仍旧正常提供服务。

Proxy

Proxy模块提供数据路由、流量探测和会话保持等功能。

- 数据路由功能：支持大数据场景下的分布式复杂查询聚合和相应的容量管理。
- 流量探测功能：降低SQL注入的风险，在必要情况下支持SQL日志的回溯。
- 会话保持功能：解决故障场景下的数据库连接中断问题。

7.4.2 高可用服务

高可用服务由Detection、Repair、Notice等模块组成。

高可用服务主要保障数据链路服务的可用性，除此之外还负责处理数据库内部的异常。

Detection

Detection模块负责检测DB Engine的主节点和备节点是否提供了正常的服务。通过间隔为8~10秒的心跳信息，HA节点可以轻易获得主节点的健康情况，结合备节点的健康情况和其它HA节点的心跳信息，Detection模块可以排除网络抖动等异常引入的误判风险，在30秒内完成异常切换操作。

Repair

Repair模块负责维护DB Engine的主节点和备节点之间的复制关系，还会修复主节点或者备节点在日常运行中出现的错误。

例如：

- 主备复制异常断开的自动修复。
- 主备节点表级别损坏的自动修复。
- 主备节点Crash的现场保存和自动修复。

Notice

Notice模块负责将主备节点的状态变动通知到负载均衡或者Proxy，保证您访问正确的节点。

例如：Detection模块发现主节点异常，并通知Repair模块进行修复。Repair模块进行了尝试后无法修复主节点，通知Notice进行流量切换。Notice模块将切换请求转发至负载均衡或者Proxy，此时您的流量全部指向备节点。与此同时，Repair在别的物理服务器上重建了新的备节点，并将变动同步给Detection模块。Detection模块开始重新检测实例的健康状态。

高可用策略

高可用策略是根据您自身业务的特点，采用服务优先级和数据复制方式之间的不同组合，以组合出适合自身业务特点的高可用策略。

服务优先级有以下两个级别：

- 复原时间目标RTO (Recovery Time Objective) 优先：数据库应该尽快恢复服务，即可用时间最长。如果您对数据库在线时间要求较高，您应该使用 RTO优先策略。
- 复原点目标RPO (Recovery Point Objective) 优先：数据库应该尽可能保障数据的可靠性，即数据丢失量最少。如果您对数据一致性要求较高，您应该使用RPO优先策略。

数据复制有以下三种方式：

- 异步复制 (Async)：应用发起更新（含增加、删除、修改操作）请求，Master完成相应操作后立即响应应用，Master向Slave异步复制数据。因此异步复制方式下，Slave不可用不影响主库上的操作，而Master不可用有较小概率会引起数据不一致。
- 强同步复制 (Sync)：应用发起更新（含增加、删除、修改操作）请求，Master完成操作后向Slave复制数据，Slave接收到数据后向Master返回成功信息，Master接到Slave的反馈后再响应应用。Master向Slave复制数据是同步进行的，因此Slave不可用会影响Master上的操作，而Master不可用不会引起数据不一致。
- 半同步复制 (Semi-sync)：正常情况下数据复制方式采用强同步复制方式，当Master向Slave复制数据出现异常的时候（Slave不可用或者双节点间的网络异常），Master会暂停对应用的响应，直到复制方式超时退化成异步复制。如果允许应用在此时更新数据，则Master不可用会引起数据不一致。当双节点间的数据复制恢复正常（Slave恢复或者网络恢复），异步复制会恢复成强同步复制。

您可以根据自身业务特点，选择服务优先级和数据复制方式的不同组合方式，提高可用性。

7.4.3 备份恢复服务

备份恢复服务主要提供数据的备份、转储和恢复。

云数据库RDS可以随时发起数据库的备份，能够根据备份策略将数据库恢复至任意时刻，提高了数据可回溯性。

备份

备份模块负责将主备节点上面的数据和日志压缩并上传。RDS默认将备份上传到OSS中，在特定场景下还支持将备份文件转储到更加廉价和持久的归档存储上。在备节点正常运作的情况下，备份总是在备节点上面发起，以避免对主节点提供的服务带来冲击；在备节点不可用或者损坏的情况下，备份模块会通过主节点创建备份。

恢复

恢复模块负责将OSS上面的备份文件恢复到目标节点上。

- 回滚主节点功能：您发起数据相关的误操作后可以通过回滚功能按时间点恢复数据。
- 修复备节点功能：在备节点出现不可修复的故障时自动新建备节点来降低风险。
- 创建只读实例功能：通过备份来创建只读实例。

转储

转储模块负责备份文件的上传、转储和下载。目前备份数据全部上传至OSS进行存储，您可以根据需要获取临时链接来下载。在某些特定场景下，转储模块支持将OSS上面的备份文件转储至归档存储来提供更长时间和更低费用的离线存储。

7.4.4 监控服务

云数据库RDS提供物理层、网络层、应用层等多方位的监控服务，保证业务可用性。

Service

Service模块负责服务级别的状态跟踪，监控负载均衡、OSS、归档存储和日志服务等RDS依赖的其他云产品是否正常，包括功能和响应时间等。对RDS内部的服务，Service也会通过日志来判定是否正常运作。

Network

Network模块负责网络层面的状态跟踪，包括ECS与RDS之间的连通性监控，RDS物理机之间的连通性监控，路由器和交换机的丢包率监控。

OS

OS模块负责硬件和OS内核层面的状态跟踪，包括：

- 硬件检修：不断检测CPU、内存、主板、存储等设备的工作状态，预判是否会发生故障，并提前进行自动报修。
- OS内核监控：跟踪数据库的所有调用，并从内核态分析调用缓慢或者出错的原因。

Instance

Instance模块负责RDS实例级别的信息采集，包括：

- 实例的可用信息；
- 实例的容量和性能指标；
- 实例的SQL执行记录。

7.4.5 调度服务

调度服务由Resource模块完成，主要提供资源调配服务。

Resource

Resource模块主要负责RDS底层资源的分配和整合，对您而言就是实例的开通和迁移。例如，您通过RDS控制台或者API创建实例，Resource模块会计算出最适合的物理服务器来承载流量。RDS实例迁移所需的底层资源也由Resource负责分配和整合。在经过长时间的实例创建、删除和迁移后，Resource模块会计算资源碎片化程度，并定期发起资源整合以提高服务承载量。

7.4.6 迁移服务

云数据库RDS提供了数据迁移服务DTS (Data Transmission Service) 工具，方便您快速迁移数据库。

迁移服务主要帮助您把数据从本地数据库迁移到阿里云数据库，或者把阿里云数据库的一个实例迁移到另一实例中。

DTS

DTS是一个云上的数据传输服务，能快速的将本地数据库或者RDS中的实例迁移到另一个RDS实例中。

DTS还提供了三种迁移模式，分别为结构迁移、全量迁移和增量迁移。

- 结构迁移

DTS会将迁移对象的结构定义迁移到目标实例，目前支持结构迁移的对象有表、视图、触发器、存储过程和存储函数。

- 全量迁移

DTS会将源数据库迁移对象已有数据全部迁移到目标实例中。



注意：

在全量迁移过程中，为了保证数据一致性，无主键的非事务表会被锁定。锁定期间这些表无法写入，锁定时长依赖于这些表的数据量大小。在这些无主键非事务表迁移完成后，锁才会释放。

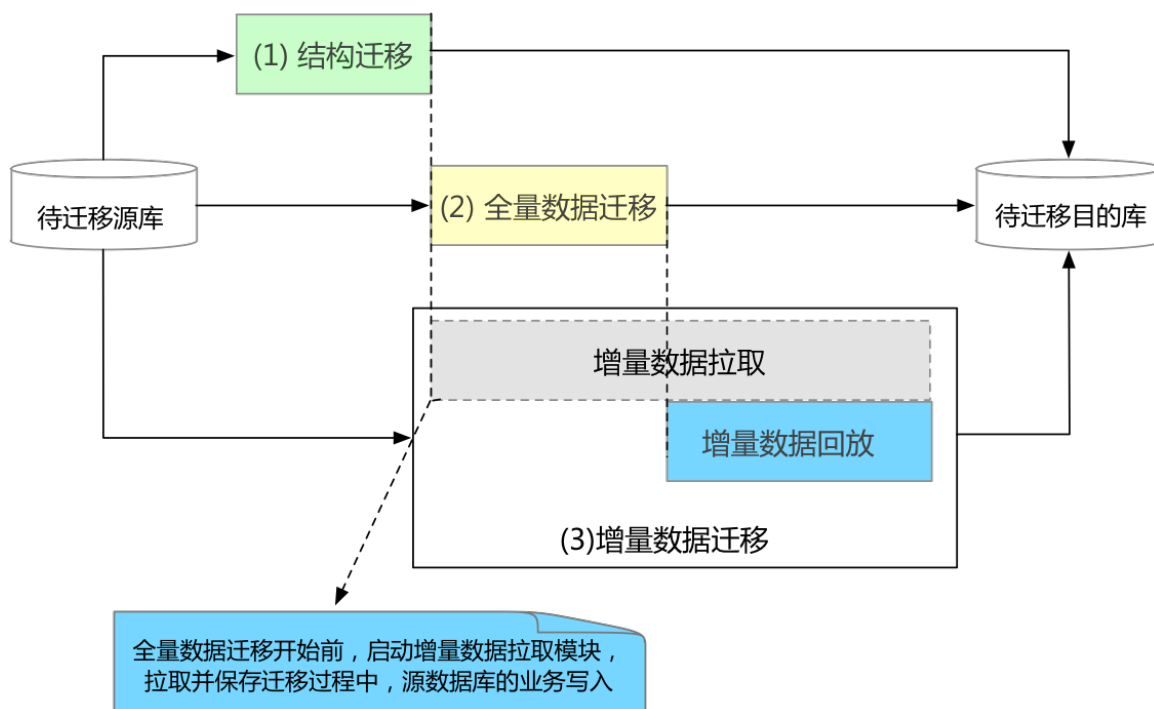
- 增量迁移

DTS会将迁移过程中数据变更同步到目标实例。



注意：

如果迁移期间进行了DDL操作，这些结构变更不会同步到目标实例。



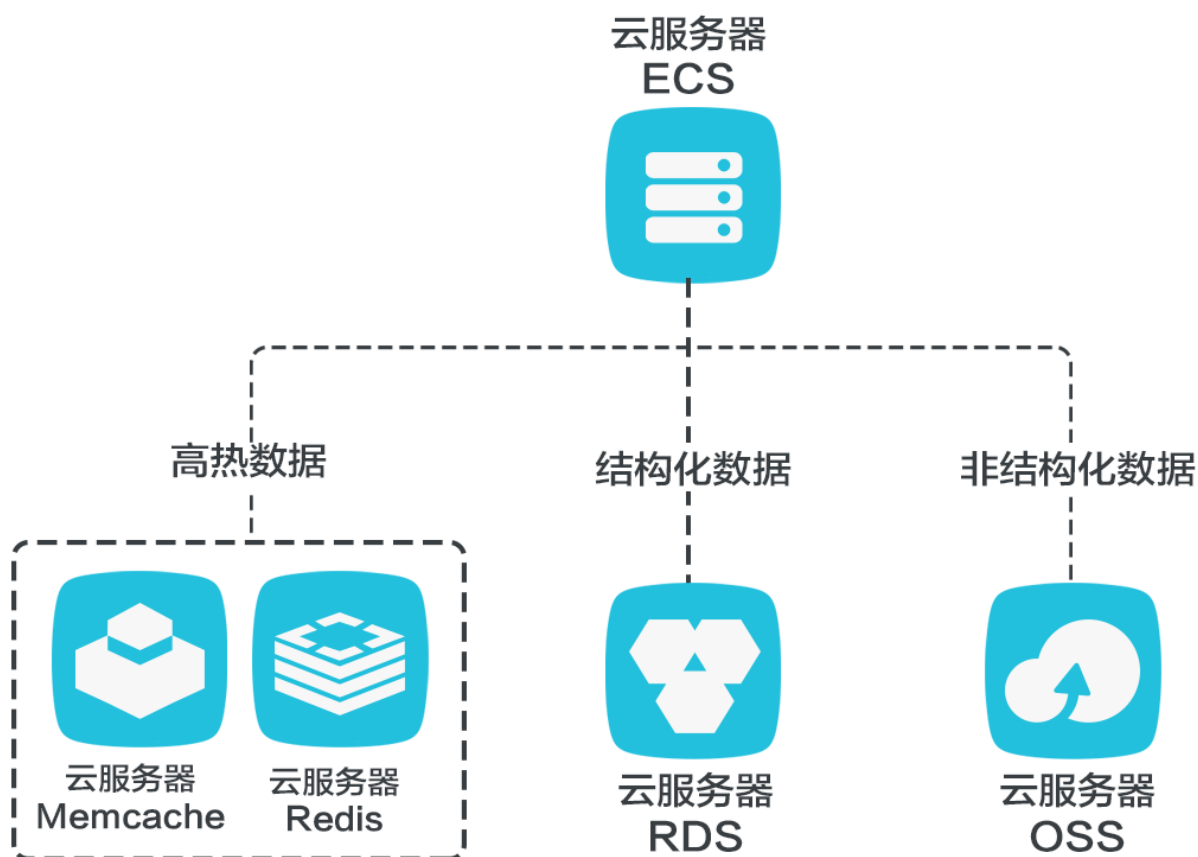
7.5 应用场景

7.5.1 数据多样化存储

云数据库RDS提供缓存数据持久化和多结构数据存储。

RDS支持搭配云数据库Memcache版、云数据库Redis版和对象存储OSS等存储产品使用，适用于多样化存储的场景，如图 7-2: 数据多样化存储所示。

图 7-2: 数据多样化存储



缓存数据持久化

RDS可以和云数据库Memcache版和云数据库Redis版搭配使用，组成高吞吐、低延迟的存储解决方案。与RDS相比，云数据库缓存产品有两个特性：

- 响应速度快，云数据库Memcache版和云数据库Redis版请求的时延通常在几毫秒以内。
- 缓存区能够支持比 RDS更高的每秒查询率QPS (Queries Per Second)。

多结构数据存储

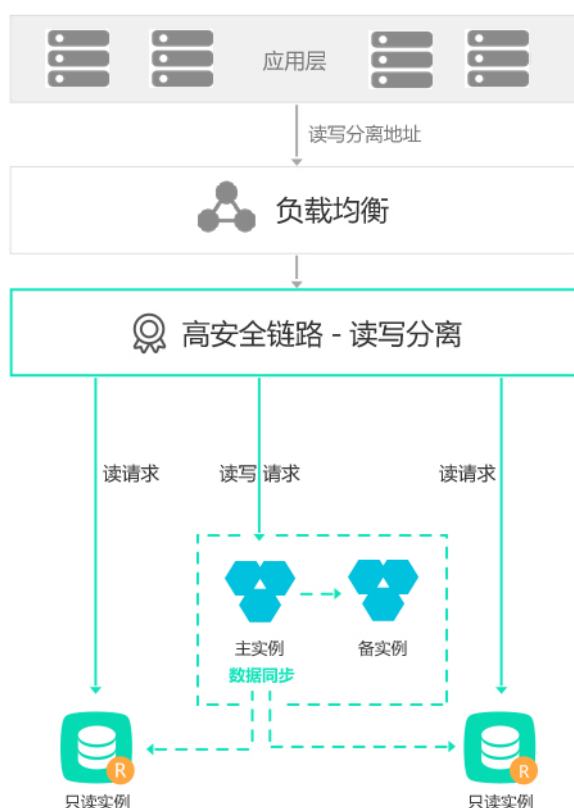
OSS是阿里云对外提供的海量、安全、低成本、高可靠的云存储服务。RDS可以和OSS搭配使用，组成多类型数据存储解决方案。例如，当业务场景为论坛时，RDS搭配OSS使用，注册用户的图像、帖子内容的图像等资源存储在OSS中，以减少RDS的存储压力。

7.5.2 读写分离

通过读写分离功能可以实现数据读取和写入操作的分离，扩展系统的处理能力。

阿里云数据库 MySQL 版支持直接挂载只读实例，分担主实例读取的压力。MySQL 版数据库的主实例和只读实例都具有独立的连接地址，当您开启读写分离功能后，系统就会额外提供一个读写分离地址，联动主实例及其下的所有只读实例，实现了自动的读写分离。应用程序只需连接同一个读写分离地址进行数据读取及写入操作，读写分离模块会自动将写入请求发往主实例，而将读取请求按照您设置的权重发往各个只读实例。您只需通过添加只读实例的个数，即可不断扩展系统的处理能力，应用程序上无需做任何修改，如[读写分离示意图](#)所示。

图 7-3: 读写分离示意图

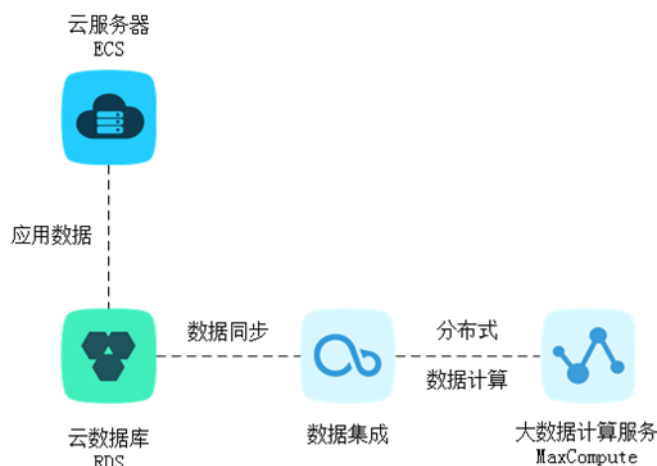


7.5.3 大数据分析

将RDS数据导入大数据计算服务，可以实现大规模的数据计算。

大数据计算服务（MaxCompute，原名ODPS），可服务于批量结构化数据的存储和计算，提供海量数据仓库的解决方案以及针对大数据的分析建模服务，如[大数据分析示意图](#)所示。

图 7-4: 大数据分析示意图



7.6 使用须知

在使用MySQL前，您需要了解使用过程中的一些注意事项或者使用限制。

为保障实例的稳定及安全，云数据库MySQL版有部分使用上的约束，如[表 7-1: MySQL使用约束](#)所示。

表 7-1: MySQL使用约束

操作	使用约束
修改数据库参数设置	大部分数据库参数须使用OpenAPI进行修改，同时出于安全性和稳定性考虑，部分参数不支持修改。
数据库的root权限	不提供root或者sa权限。
数据库备份	- 可使用命令行或图形界面进行逻辑备份。 - 仅限通过RDS管理控制台或API进行物理备份。
数据库还原	- 可使用命令行或图形界面进行逻辑还原。 - 仅限通过RDS管理控制台或API进行物理还原。
数据迁入	可使用命令行或图形界面进行逻辑导入。

操作	使用约束
	可使用mysql命令行工具、数据传输服务等方式迁入数据。
MySQL存储引擎	- 目前支持InnoDB、TokuDB两种引擎（MyISAM引擎由于自身缺陷，存在数据丢失的风险，因此仅部分存量实例暂时支持，新创建实例的MyISAM引擎表会自动转换为InnoDB引擎表。 - 出于性能和安全性考虑建议尽量采用InnoDB存储引擎。 - 不支持Memory引擎。如果您创建Memory引擎的表，我们将自动为您转换成InnoDB引擎的表。
搭建数据库复制	MySQL本身提供主备复制架构的双节点集群，无需您手动搭建。其中主备复制架构集群的备（slave）实例不对您开放，您的应用不可直接访问。
重启RDS实例	必须通过RDS管理控制台或API操作重启实例。
账号和数据库管理	MySQL默认需要通过RDS管理控制台进行账号和数据库管理。同时，MySQL也支持您通过创建高权限帐号的方式来自行管理用户、密码和数据库。
普通帐号	- 不可以自定义授权。 - 控制台有帐号管理和数据库管理的界面。 - 支持创建普通帐号的实例也支持创建高权限帐号。
高权限帐号	- 可以自定义授权。 - RDS控制台上帐号管理和数据库管理的界面消失，相关操作只能通过代码或数据管理DMS进行。 - 开通高权限帐号后无法回退。

7.7 基本概念

概念	解释
地域	地域（Region）指您创建的RDS实例的服务器所处的地理位置。您需要在创建RDS实例时指定，创建实例后暂不支持更改。在创建RDS实例时，需要搭配阿里云服务器ECS使用，RDS只支持内网访问，在地域选择时需要与ECS相同。
可用区	可用区是指在同一地域下，电力、网络隔离的物理区域。可用区之间内网互通，可用区内网络延时更小，不同可用区之间故障隔离。单可用区是指RDS实例副本集中的三个节点处于相同的可用区。如果ECS和RDS部署在相同的可用区，网络延迟更小。
实例	RDS实例，或简称实例，是创建RDS服务的基本单位。实例是阿里云数据库RDS版的运行环境，在主机上以单独的进程存在。您可以通过控制台来创建、修改和删除RDS实例。各实例之间相互独立、资源隔离，相互之间不存在CPU

概念	解释
	、内存、IO等抢占问题。每个实例拥有其自己的特性，例如数据库类型、版本等，系统有相应的参数来控制实例行为。
内存	云数据库RDS实例可以使用的内存上限。
磁盘容量	磁盘容量是您创建RDS实例时，所选择创建的磁盘大小。实例所占用的磁盘容量除集合数据外，还有实例正常运行所需要的空间，如系统数据库、数据库回滚日志、重做日志、索引等。请确保RDS实例具有足够的磁盘容量来存储数据，否则可能导致实例被锁定。若因磁盘容量不足导致实例被锁定，您可创建更大的磁盘容量来解锁实例。
IOPS	以4KB为单位，每秒进行块设备读写操作的次数上限。
CPU核	实例可以使用的计算能力上限。1个CPU拥有不低于2.3GHz 超线程（Intel Xeon 系列Hyper-Threading）的计算能力。
连接数	客户端和RDS实例之间的TCP连接。如果客户端使用了连接池，则客户端和RDS实例之间的连接为长连接，反之则为短连接。

8 数据管理DMS

8.1 什么是数据管理DMS企业版

数据管理服务 (Data Management Service , 简称DMS) 企业版是阿里云提供的支持MySQL、SQL Server、DRDS、PostgreSQL、ORACLE、Oceanbase等关系型数据库和MongoDB等NoSQL数据库的数据库管理产品。它是一种集数据管理、结构管理、研发流程、研发规范、用户管理、权限管理、访问安全于一体的数据管理服务。

- 提供研发从线下环境根据设计规范自助进行表结构设计、上线前代码内使用到增删改查的SQL审核、结构按需生产发布到指定环境的完整数据库研发流程支持。
- 提供库、表、字段级的细粒度操作权限管控，所有用户通过产品产生的数据库操作可在线化、可溯源审计支持。
- 提供根据业务需要可灵活配置结构设计、数据变更、数据导出、权限申请等各个模块的操作规范与审批流程支持。
- 统一研发与数据库交互的入口，任何用户都不再直接接触数据库账号密码也不需要频繁切换数据库链接进行管理。

8.2 产品优势

丰富的数据源支持

支持如下数据源：

- 关系型数据库：MySQL、SQL Server、DRDS、PostgreSQL、ORACLE、OceanBase
- NoSQL数据库：MongoDB

统一研发数据库操作入口，操作可审计

管理员配置好需要管理的数据库实例列表，所有需要使用数据库查询、变更结构、变更数据等的需求都可以在产品内完成；所有操作可按照人员、数据库、表、时间等多维度进行搜索审计。

细粒度权限管控

普通用户不再接触数据库账号密码，根据实际需要在产品内按需申请指定有效期范围内的库、表、字段的查询、导出、变更权限进行相应操作，到期后权限自动失效回收。无查询权限则无法对数据库表进行SQL查询，无导出权限则无法提交数据导出工单，无变更权限则无法提交数据变更工单。

细粒度研发流程可灵活定制

每个数据库实例可定义满足业务实际需要的各模块审批流程进行管控，可同时满足效率、安全等维度的要求。如测试环境轻管控，可减少流程或不设置流程；生产环境重管控，可配置具体操作经过指定人员逐层审批后方可生效执行等。

企业表结构设计规范

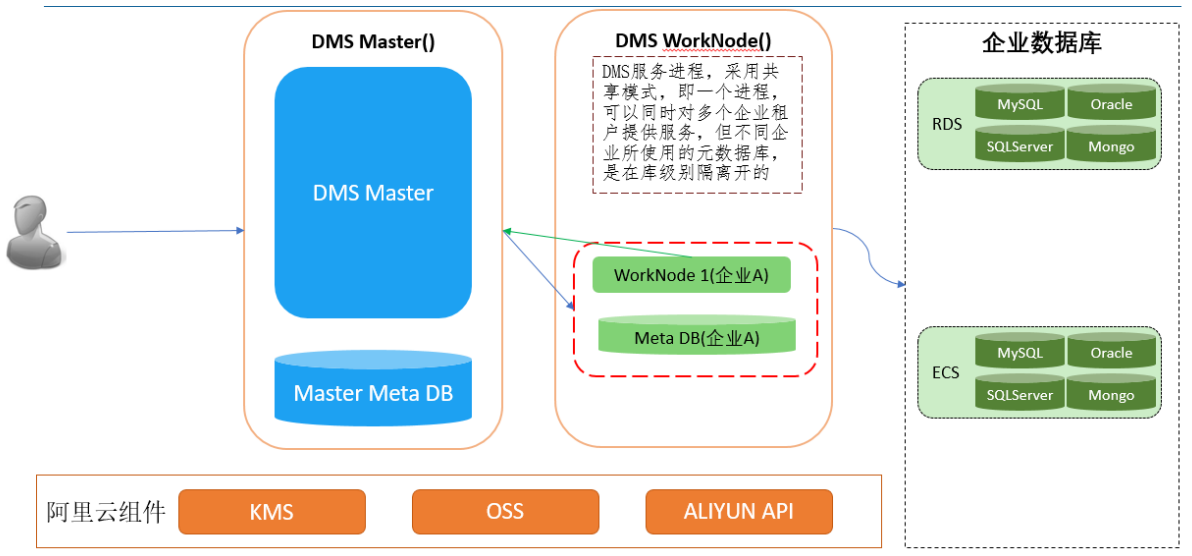
可按照企业内需要制定MySQL表结构设计规范，如字段类型、索引类型、索引个数、字段名长度、表名长度、研发发布流程等相关规范。

8.3 产品架构

DMS在架构上分为了两个大的节点：Master和Node。Master管理用户发起的登录、鉴权以及提供配置中心服务，并将用户发起的请求过滤后转给给Node；Node提供DMS企业版的业务服务。

如下图所示，用户访问的是Master节点，Master在对请求进行过滤后转发到Node节点，Node节点处理完访问数据库的业务逻辑后返回结果给客户。

图 8-1: DMS企业版架构



8.4 功能特性

8.4.1 实例管理

功能说明	专有云数据库实例的统一管理，支持的数据库包括MySQL、DRDS、SQLServer、PostgreSQL、ORACLE、OceanBase、MongoDB。
------	---

应用场景	维护需要管理的实例列表。
使用方法	打开 系统管理 > 实例管理 页面，为对应需要管理的目标数据库实例创建DMS企业版专用数据库账号密码、安全规则后进行录入。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.2 用户管理

功能说明	产品内用户的统一管理，支持RAM主账号、RAM子账号、普通云账号。
应用场景	维护需要使用DMS企业版的用户列表。
使用方法	打开 系统管理 > 用户管理 页面，为需要使用DMS企业版的用户进行维护，包括普通用户、DBA、管理员三个角色的设定。
使用限制	仅限产品内管理员角色可使用。

8.4.3 任务管理

功能说明	产品内任务的统一管理。
应用场景	新建任务、查看历史任务。
使用方法	打开 系统管理 > 任务管理 页面，可管理工单来源创建的各种任务、可直接新建变更任务。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.4 安全规则

功能说明	产品内实例级别细粒度的操作管控、流程管控基础。
应用场景	定义实例级别的审批流程。
使用方法	打开 系统管理 > 安全规则 页面，可按需定义不同的审批流程、审批人员等信息。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.5 配置管理

功能说明	产品内全局管控的一些参数配置。
应用场景	定义全局的一些管控策略。
使用方法	打开 系统管理 > 配置管理 页面，可按需定义不同的管控方式与管控阈值。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.6 操作日志

功能说明	产品内所有用户行为的日志记录汇总查看。
应用场景	审计某个数据库、某个数据表、某个人员在指定时间段内的操作情况。
使用方法	打开 系统管理 > 操作日志 页面，可按需检索。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.7 访问IP白名单

功能说明	可限制人员使用DMS企业版的服务器IP来源。
应用场景	管控产品使用范围。
使用方法	打开 系统管理 > 访问IP白名单 页面，可按需定义开启限制与否，限制具体公网IP地址。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.8 结构设计配置

功能说明	可定义MySQL-表结构设计规范、表结构变更的研发流程。
应用场景	定义符合业务实际情况的表结构设计规范。
使用方法	打开 系统管理 > 结构设计配置 页面，可按需定义符合本企业的设计规范。
使用限制	仅限产品内DBA、产品内管理员角色可使用。

8.4.9 权限申请

功能说明	可按需申请开通库、表、字段级别在一定有效期内的查询、导出、变更等权限。
应用场景	普通用户需要开通权限。
使用方法	打开 任务类型 > 权限申请 页面，可按需提交申请。
使用限制	产品内所有用户。

8.4.10 数据变更

功能说明	可对目标库表进行DDL、DML等变更的提交与执行。
应用场景	项目研发过程中、代码上线后的结构初始化、数据初始化、数据修复（数据订正）。
使用方法	打开 任务类型 > 数据变更 页面，可按需提交可执行的SQL脚本。

使用限制	产品内对目标库有变更权限的用户。
------	------------------

8.4.11 数据导出

功能说明	可对目标库表进行大量数据导出的提交与执行。
应用场景	大批量数据分析。
使用方法	打开 任务类型 > 数据导出 页面，可按需提交可执行的SQL脚本。
使用限制	产品内对目标库有导出权限的用户。

8.4.12 库表同步

功能说明	可对比源和目标库表之间的差异，提供差异DDL提交与执行。
应用场景	表结构同步。
使用方法	打开 任务类型 > 库表同步 页面，可按需提交对比任务与差异脚本的执行。
使用限制	产品内对源库有查询权限，对目标库有变更权限的用户；仅支持MySQL数据库类型。

8.4.13 结构设计

功能说明	可对目标库表进行符合研发规范的表结构设计。
应用场景	项目研发过程中的表结构设计。
使用方法	打开 任务类型 > 结构设计 页面，可按需编辑表结构产生变更脚本执行到目标数据库。
使用限制	产品内对目标库有变更权限的用户；仅支持MySQL数据库类型。

8.4.14 SQL审核

功能说明	可对目标库表进行代码中增删改查SQL的审核。
应用场景	项目上线前代码SQL的审核。
使用方法	打开 任务类型 > SQL审核 页面，可按需提交需要审核的SQL脚本。
使用限制	产品内对目标库有查询权限的用户；仅支持MySQL数据库类型。

8.4.15 工作台

功能说明	用户的个人中心，展现工单列表、最近使用的数据库、关注的数据库、有权限的数据库、owner的数据库入口，可进行相应的管理。
------	--

应用场景	查找工单、快速找数据库、管理个人权限、管理所owner数据库的人员权限开通情况等。
使用方法	打开工作台页面，可按需使用相关功能。
使用限制	产品内所有用户。

8.4.16 顶栏搜索

功能说明	可对产品内所有数据库、表元数据进行检索。
应用场景	快速寻找数据库、表进行相关操作。
使用方法	输入需要查找的数据库、表名关键字进行条件组合检索。
使用限制	产品内所有用户。

8.4.17 SQLConsole

功能说明	可对目标库表进行相关查询、变更等操作。
应用场景	数据查询分析等场景。
使用方法	在目标数据库下书写可执行SQL提交执行。
使用限制	产品内对目标库表有查询权限的用户。

8.5 基本概念

1. 路由算法

- a. 定义：分表/逻辑表，在SQLConsole或者数据变更、数据导出等场景需要用逻辑表进行数据操作时，为了更准确的进行条件定位查询，可以而配置路由算法，以提高操作的效率

A. 若未配置路由算法，对逻辑表的数据操作将会遍历逻辑表对应的每一个物理表，整体操作时间会成倍放大

B. 路由字段+算法组成路由算法

b. 用途

A. 分表数据查询

B. 分表数据变更

C. 分表数据导出

c. 配置

- 新增/修改：可以通过逻辑库找到对应的逻辑表，编辑修改路由算法

2. 字段安全级别

- a. 定义：字段安全等级根据存储的业务数据性质的差异，对于部分字段内容存储值如手机号码、身份证号码等属于敏感数据。非特殊需求外常规查询该表数据时此类字段的值不应该予以展现。
- b. 用途：
 - A. SQLConsole数据查询时，对无权限的敏感列、机密列默认显示为**（避免敏感数据的扩散）
 - B. 需要使用与查询、导出、数据变更则需要在权限申请时单独申请敏感列、机密列权限
 - C. 导出、数据变更涉及敏感列、机密列时，DBA、管理员可差别设置不同的审批流程
- c. 增量配置：
 - 表结构：查看时申请流程调整
 - 等级上调，提交即刻生效
 - 等级下调，由DBA、管理员设置对应的审批生效流程

3. 逻辑库

- a. 定义
 - A. 当业务达到一定规模后，需要通过分库分表来进行负载均衡从而达到大量业务压力场景下平滑支撑的能力
 - B. 库的数量 ≥ 1
 - 大于1时，库名个数、后缀格式需要有约束
 - 一般个数为2的指数幂
 - 一般后缀为_xxxx,即下划线4位数字从0开始递增补齐
 - 等于1时，**仅存在分表的单个数据库会抽成逻辑库**
- b. 典型的场景：
 - A. 单库-分表
 - B. 分库-分表【最常见】
 - C. 分库-不分表（即分库同表名）
- c. 用途：
 - A. SQL Console数据查询、结构设计、导出、数据变更等入口支持选择物理库或者逻辑库操作

- 选择逻辑库，则可便捷的进行分表/逻辑表 操作（将分表/逻辑表当做单个表一样写SQL）

B. 数据源权限申请，可直接申请逻辑库权限即开通其关联的所有物理库的对应权限

d. 配置：

- 通过搜索数据库列表，**更多**下拉选择【配置逻辑库】由数据owner或DBA进行配置

4. 逻辑表

a. 定义

A. 当业务达到一定规模后，需要通过分库分表来进行负载均衡从而达到大量业务压力场景下平滑支撑的能力。

B. 表的数量>1，按照逻辑库进行均分

- 一般表的个数为2的指数幂
- 一般表名后缀为_xxxx，即下划线4位数字从0开始递增补齐
- 表数量需要能被库的数量整除均分（如表的数量1024，库的数量32，则每个库均分32张物理表；若表的数量为1024，库的数量为33，则无法均分不支持创建）

b. 典型的场景：

A. 单库-分表

B. 单个物理库下面分表

C. 分库-分表【最常见】

- 按序逐个物理库内物理表数字递增均分
- 每个库相同表名，如每个库都是01-12的12张分表

D. 分库-不分表（即分库同表名）

- **常见的有：每个分库下有同一个物理表名**

c. 用途：

A. 满足以上三种情况的场景后非常核心的关键点，“**逻辑表**”一定要在【**逻辑库**】下进行操作

B. 数据源权限申请

- 直接申请逻辑表权限即开通其关联的所有物理表的对应权限

C. SQLConsole数据查询

D. 数据导出

E. 数据变更

F. 以上各入口支持选择具体【**逻辑库**】后，针对“**逻辑表**”可当做单表直接使用

d. 配置：**A. 初始化**

- 配置逻辑库后，根据逻辑库下物理表结构会自动抽取逻辑表。配置不支持用户自主调整

B. 抽取规则：

- 同一个物理表只能出现在一个逻辑表配置中
- 同一个逻辑表对应的所有物理表的结构（字段名、字段类型）需要严格一致，若不一致则无法聚合，**扩展价值可为用户提供不一致数据预警**

C. 新增

- 非平台新增的逻辑表，可以通过搜索到**逻辑库**结果列表，下拉操作-更多**抽取库下逻辑表**自助触发增量抽取

e. FAQ：

- Q：如果逻辑库下逻辑表已经存在【**不需要新建物理表**】，但是不在逻辑库显示的表列表中
- A：可以通过搜索找到逻辑库，操作-更多下拉抽取库下逻辑表

5. 元数据**a. 定义：**

A. 顾名思义，属于数据库本身的一些数据。目前采集的主要包含有以下内容

- 数据库名、数据库字符集
- 表名、表的大小、表的记录行数、表的字符集、表的字段、表的索引和表的描述
- 字段的类型、字段的精度、字段的描述

B. 说明：由于所有信息采集来源于information_schema等数据库内部采集数据，如表的大小、表的行数可能不是非常精准仅做数量级的参考

b. 应用场景**A. 结构查看与导出**

- 通过顶栏输入数据库名称的关键字后，**到达搜索结果页面的三种方式是**
 - 直接回车（光标不能下移到自动补齐页上表信息的记录行上）

- 单击自动补齐页上的搜索
- 单击输入框后的放大镜进行搜索
- 通过顶栏输入数据库名称的关键字后，**直达数据库表列表页面的方式是**
 - 在自动补齐页上的具体数据库对象行，鼠标单击
 - 在自动补齐页上的具体数据库对象行，移动光标到此行后按回车
- 通过搜索结果页面，**直达数据库表列表页面的方式是**
 - 单击对应数据库行末尾的详情
 - 选择对应数据库行末尾，单击更多-查看表详情
 - 双击对应数据库行

B. 通过上述方式找到数据库搜索结果页面，可以单击更多-下拉选择导出建表语句-create table语句、导出表结构的excel表格【**此处是全库所有表的数据**】

C. 通过上述方式找到数据库表列表页面

- 可以在页签内右上角单击导出-下拉选择整库建表语句-create table语句、整库表结构excel表格【**此处全库所有表的数据**】
- 可以在具体表记录行上单击更多-下拉选择导出建表语句-create table语句、导出表结构excel表格【**此处是单个物理表的数据**】（若只需要在线查看结构信息不需要导出，可单击具体表记录行，查看字段及属性、查看索引及属性）

c. 同步机制：

A. 注册到产品内时，会触发自动全量的字典采集；但如果DDL操作不在平台内发起，则平台目前不支持实时感知此类操作，需要主动发起一次同步元数据的操作

- 入口1：【普通用户】**数据库表列表页面**，右上角单击同步元数据
 - 物理库，则同步当前库
 - 逻辑库，则同步逻辑库下的每一个物理库（批量）
- 入口2：【DBA&管理员】系统管理-实例管理，找到对应目标实例行单击立即同步
- 入口3：【管理员】系统管理-配置管理，找到对应配置项进行**定时任务**部署支持每天按时全部实例的元数据自动同步
 - 参数名：全量元数据同步Cron表达式

16	全量元数据同步Cron表达式	0 0 5,12,18 * * ?	请输入全量元数据同步CRON表达式开启同步。关闭请输入：N。推荐"0 0 5,12,18 * * ?"每天三个时间点同步。	2017-11-02 16:18:46	修改历史
----	----------------	-------------------	---	---------------------	------

- B. 说明：**若新建数据库则只能用入口2进行处理，只是已有库内表的变更入口1和入口2均可。

6. 数据owner

- a. 定义：**作为数据库或者表的数据负责人，对其他用户使用其负责的数据库表进行管控。

b. 用途

- A. 用于系统管理-安全规则的各功能审批流程中的审批节点**
B. 管理【数据库/表-人员】的权限使用情况，进行主动授权或回收。

c. 配置

- A. 【被动】由DBA或原owner配置**
B. 【主动】业务方负责人自助申请
- 根据实例指定的不同的审批流程审批通过后进行自动配置

7. 安全规则

- a. 组成：**整个安全规则管理模块包括安全规则、审批流程线、审批节点三个核心概念

b. 审批节点

A. 系统会初始化几个动态节点（不可删除与编辑）

- Admin：系统管理员，如果系统有多个管理员则任意一个参与审批操作均可
- DBA：实例对应的DBA同学，在管理员或DBA录入实例时默认是当前录入者也可以指定给系统里的其他DBA角色
- Owner：数据库的数据owner，在DBA录入实例采集完数据字典时由DBA配置（也可以有已有的owner主动调整或自助流程申请成为某个库表的数据owner）

B. 新增节点（可按需删除与编辑）

- 在系统默认节点的基础上可以用户自定义审批节点及对应节点的审批人员，常见的有：
 - 测试人员
 - 数据安全员
 - 研发TL
 - 架构师
- 说明：这个节点的设置甚至可以是具体某个业务的相关人员，如：某个业务测试，某个业务研发TL。

c. 审批流程线

A. 系统会初始化几个动态流程线（不可删除与编辑）

- Admin：仅需管理员审批
- DBA：仅需DBA审批
- Owner：仅需Owner审批
- Owner->DBA：需要Owner审批后，DBA顺序审批
- Owner->DBA->Admin：需要Owner审批后，DBA顺序审批，最后Admin审批

B. 新增流程线（可按需删除与编辑）

- 利用系统的动态节点、手工添加的审批节点组合出特定的审批流程线，常见的有：
 - Owner->数据安全员：需要Owner审批后，数据安全员审批
 - Owner->测试人员：需要Owner审批后，测试人员审批
- 说明：这个流程线可以根据业务实际需要，用各个审批节点可以任意组成特定业务线的审批流程线

d. 安全规则**A. 系统会初始化高、中、低，三份安全规则，控制原则到具体模块的具体操作详情可在产品中查阅，此处列举具体模块支持的控制能力。（不可删除，但可编辑）**

- SQLConsole：
 - 是否允许DML（可设置影响行数阈值，超过阈值不允许执行）
 - 是否允许DDL（可设置表大小阈值，超过阈值不允许执行）
 - 是否允许执行高风险DDL（删表、删字段）
 - 是否允许执行其他SQL
- 数据变更
 - 是否允许DML（可设置影响行数阈值，超过阈值不允许执行）
 - 允许，则可以指定是否需要流程、需要怎么样的审批流程线
 - 是否允许DDL（可设置表大小阈值，超过阈值不允许执行）
 - 允许，则可以指定是否需要流程、需要怎么样的审批流程线
 - 是否允许执行高风险DDL（删表、删字段）
 - 允许，则可以指定是否需要流程、需要怎么样的审批流程线
 - 是否允许执行其他SQL

- 允许，则可以指定是否需要流程、需要怎么样的审批流程线
- 数据导出
 - 是否对导出进行审批控制
 - 是，则可以设置每次导出的阈值
 - 针对不同的阈值可设置是否需要流程，需要怎么样的审批流程线
 - 当涉及敏感数据时可设置审批流程线，同时此规则是最严格的规则，会忽略行数的流程
- 权限申请
 - 库表列权限
 - 是否对审批进行控制，控制时可设置审批流程线
 - 库、表权限
 - 敏感列权限
 - 机密列权限
 - 数据owner
 - 是否对审批进行控制，控制时可设置审批流程线
 - 有数据owner流程
 - 无数据owner流程
 - 敏感级别（升级将直接生效）
 - 是否对审批进行控制，控制时可设置审批流程线
 - 机密降为敏感
 - 机密降为内部
 - 敏感降为内部

B. 新增安全规则（可按需删除与编辑）

- 可通过审批流程线的组合，配置符合具体某个业务线诉求的安全规则。实现业务维度的灵活管控与操作审计



说明：

一个实例遵循一个安全规则，在测试环境可以严格管控、在后台系统的生产环境也可以松散管控，真正实现按需管理的灵活支持。

9 数据传输服务DTS

9.1 什么是数据传输服务DTS

数据传输服务DTS (Data Transmission Service) 是阿里云提供了一种支持关系型数据库及大数据等多种数据源之间数据交互的数据服务。

DTS提供了数据迁移、数据实时订阅及数据实时同步等多种数据传输能力。通过DTS可实现不停服数据迁移、数据异地灾备、跨境数据同步、缓存更新等多种业务应用场景，助您构建安全、可扩展、高可用的数据架构。

- DTS的目标是帮您承担复杂的数据交互工作，让您可以专注于上层的业务开发。
- DTS支持多种数据源类型，具体如下。
 - 关系型数据库：MySQL、DRDS、Oracle。
 - OLAP：MaxCompute、AnalyticDB。

9.2 产品优势

DTS支持关系型数据库、OLAP等数据源之间的数据传输。它提供了数据迁移、实时数据订阅及数据实时同步等功能。相对于第三方迁移同步工具，DTS提供更丰富多样、高性能、安全可靠的传输链路，同时它提供了诸多便利功能，极大地方便了传输链路的创建及管理。

丰富多样

DTS支持多种传输方式：数据迁移、实时数据订阅及数据实时同步。其中实时数据订阅及数据实时同步均为实时数据传输方式。

为了降低数据迁移对应用的影响，数据迁移功能支持不停服的迁移方式。不停服迁移，可实现在数据迁移过程中，应用停机时间降低到分钟级别。

高性能

DTS使用高规格服务器来保证每条迁移同步链路都能拥有良好的传输性能。

DTS底层使用了多种性能优化措施。

相对于传统的数据同步工具，DTS的实时同步功能能够将并发粒度缩小到事务级别，能够并发同步同张表的更新数据，从而极大提升同步性能。

安全可靠

DTS底层为服务集群，集群内任何一个节点宕机或发生故障，控制中心都能够将这个节点上的所有任务快速切换到其他节点上。

DTS内部对部分传输链路提供7×24小时的数据准确性校验，快速发现并纠正传输数据，保证传输数据可靠性。

DTS各模块间采用安全传输协议及安全token认证，有效得保证数据传输可靠性。

简单易用

DTS提供可视化管理界面，提供向导式的链路创建流程，用户可以在其控制台简单轻松地创建自己的传输链路。

DTS控制台展示了链路的传输状态及进度，传输性能等信息，用户可以方便管理自己的传输链路。

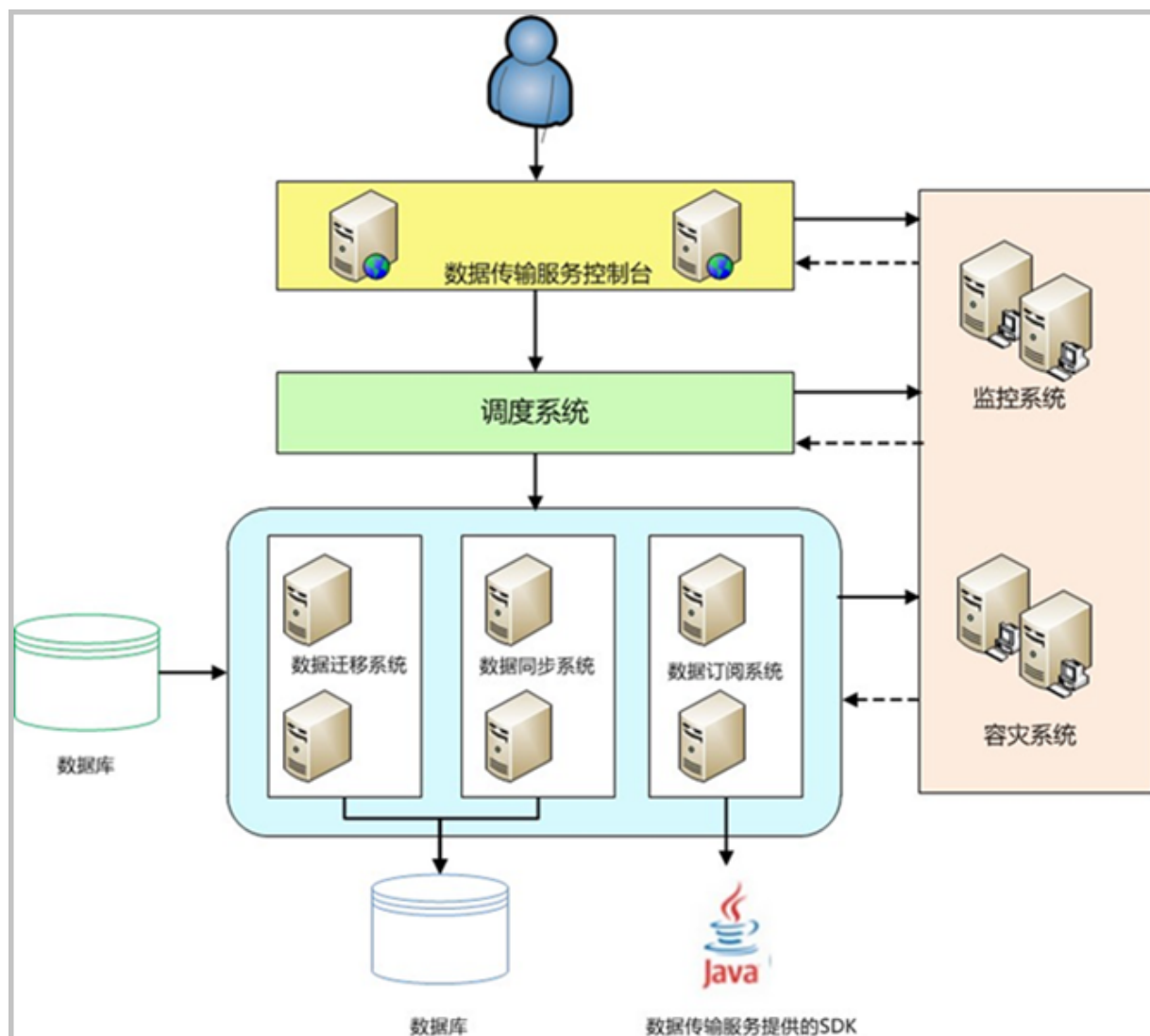
为了解决网络或系统异常等导致的链路中断问题，DTS提供链路断点续传功能，且定期监测所有链路的状态，一旦发现链路异常，先尝试自动修复重启，如果链路需要用户介入修复，那么用户可以直接在控制台修复，并重启链路。

9.3 产品架构

系统架构

如图 9-1: 系统架构图所示为DTS的系统架构图。

图 9-1: 系统架构图



- **系统高可用**

DTS内部每个模块都有主备架构，保证系统高可用。容灾系统实时检测每个节点的健康状况，一旦发现某个节点异常，会将链路秒级切换至其它节点。

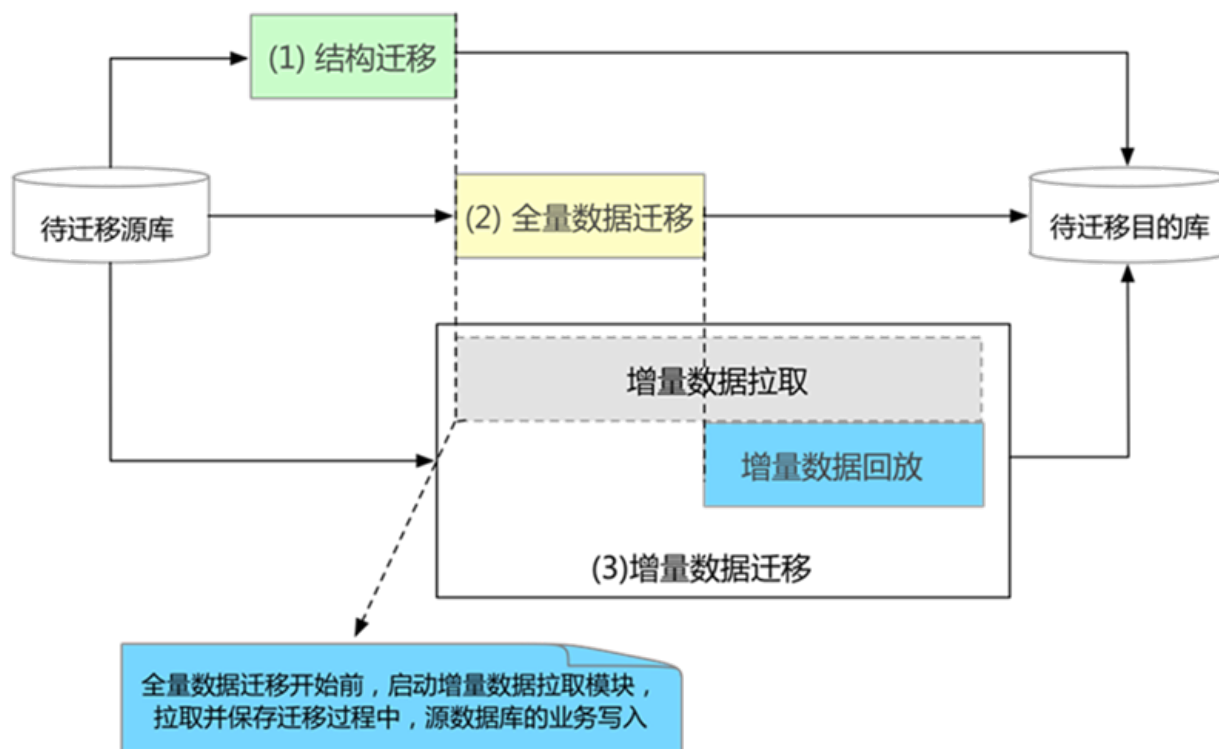
- **数据源地地址变更**

对于数据订阅及同步链路，容灾系统还会监测数据源的连接地址切换等变更操作，一旦发现数据源发生连接地址变更，它会动态适配数据源新的连接方式，保证链路的稳定性。

数据迁移基本原理

数据迁移基本原理如[图 9-2: 数据迁移基本原理图](#)所示。

图 9-2: 数据迁移基本原理图



数据迁移任务支持多种迁移类型：结构对象迁移、全量数据迁移以及增量数据迁移。如果需要通过不停服迁移，那么迁移过程需要经历：

1. 结构对象迁移
2. 全量迁移
3. 增量数据迁移

异构数据库之间进行结构迁移时，DTS会从源数据库读取结构定义语法，再根据目标数据库的语法定义，组装成目标数据库的语法定义格式，然后导入到目标实例中。

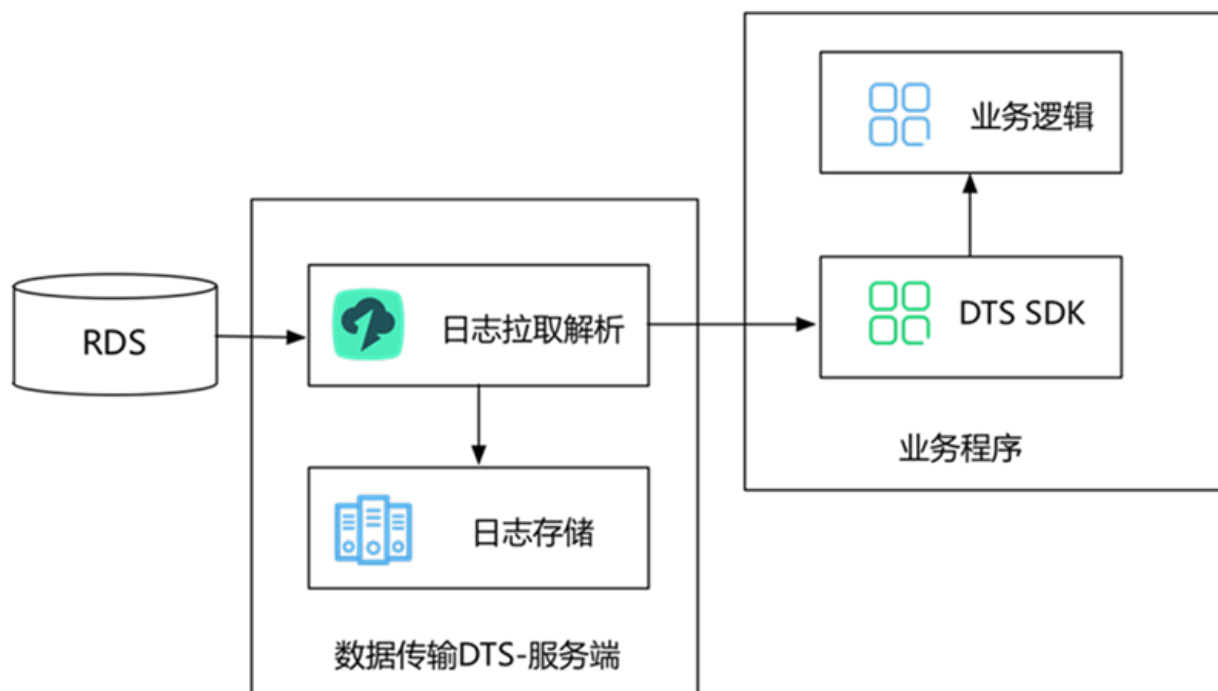
全量数据迁移过程持续较久，在这过程中，源实例不断有业务写入，为保证迁移数据的一致性，在全量数据迁移之前会启动增量数据拉取模块，增量数据拉取模块会拉取源实例的增量更新数据，并解析、封装、存储在本地存储中。

当全量数据迁移完成后，DTS会启动增量数据回放模块，增量数据回放模块会从增量拉取模块中获取增量数据，经过反解析、过滤、封装后同步到目标实例，从而实现源实例、目标实例数据实时同步。

数据订阅基本原理

数据订阅基本原理图如图 9-3: 数据订阅基本原理图所示。

图 9-3: 数据订阅基本原理图



如上图所示，数据订阅支持实时拉取RDS实例的增量日志，用户可以通过DTS SDK从数据订阅服务端订阅增量日志，根据业务需求，实现数据定制化消费。

DTS服务端的日志拉取模块从数据源抓取原始数据，并通过解析、过滤、标准格式化等流程，最终将增量数据在本地持久化。

日志抓取模块通过数据库协议连接并实时拉取源实例的增量日志。例如源实例为RDS for MySQL，那么数据抓取模块通过Binlog dump协议连接源实例。

DTS实现了日志拉取模块及下游消费SDK的高可用。

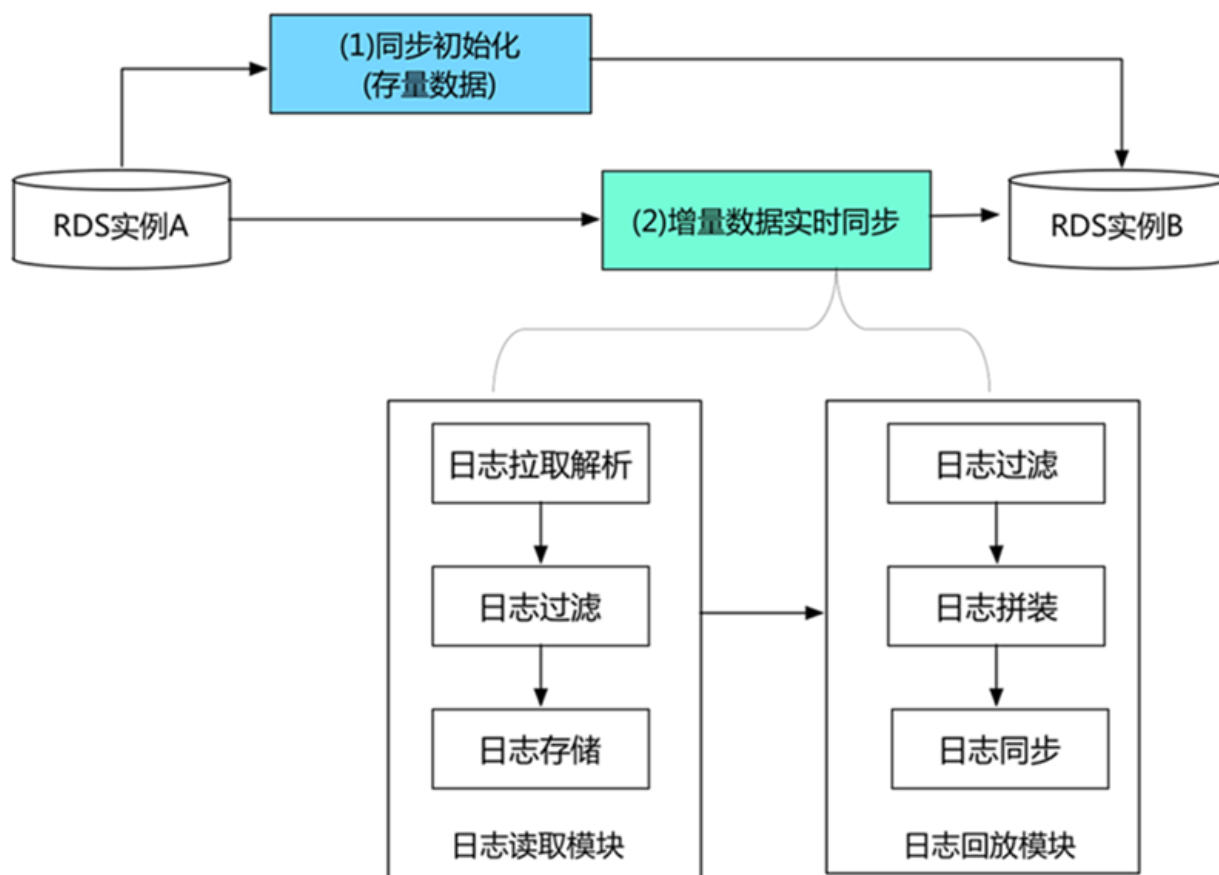
DTS容灾系统一旦检测到日志拉取模块出现异常，就会在健康服务节点上断点重启日志拉取模块，保证日志拉取模块的高可用。

DTS支持在服务端实现下游SDK消费进程的高可用。用户同时对一个数据订阅链路，启动多个下游SDK消费进程，服务端同时只向一个下游消费推送增量数据，当这个消费进程异常后，服务端会从其他健康下游中选择一个消费进程，向这个消费进程推送数据，从而实现下游消费的高可用。

实时同步基本原理

实时同步基础原理如图 9-4: 实时同步基本原理图所示。

图 9-4: 实时同步基本原理图



DTS的实时同步功能能够实现任何两个RDS实例之间的增量数据实时同步。

同步链路的创建过程包括：

- 同步初始化：在目标实例中初始化一份源实例的历史存量数据。
- 增量数据实时同步：当初始化完成后，进入两边增量数据实时同步阶段，在这个阶段DTS会实现源实例与目标实例之间的数据动态同步。

增量数据的实时同步过程中，DTS的底层实现模块主要包括：

- 日志读取模块

日志读取模块从源实例读取原始数据，经过解析、过滤及标准格式化，最终将数据在本地持久化。日志读取模块通过数据库协议连接并读取源实例的增量日志。如果源数据库为RDS for MySQL，那么数据抓取模块通过Binlog dump协议连接源库。

- 日志回放模块

日志回放模块从日志读取模块中请求增量数据，并根据用户配置的同步对象进行数据过滤，然后在保证事务时序性及事务一致性的前提下，将日志记录同步到目标实例。

DTS实现了日志读取模块、日志回放模块的高可用，DTS容灾系统一旦检测到链路异常，就会在健康服务节点上断点重启链路，从而有效保证同步链路的高可用。

9.4 功能特性

9.4.1 数据迁移

DTS的数据迁移功能旨在帮助用户方便、快速地实现各种数据源之间的数据迁移。实现数据迁移上云、阿里云内部跨实例数据迁移、数据库拆分扩容等业务场景。数据迁移功能能够支持同异构数据源之间的数据迁移，同时提供了库表列三级映射、数据过滤多种ETL特性。

多种数据源类型

数据迁移支持多种数据源之间的数据迁移，不同数据源的支持情况如[表 9-1: 不同数据源的支持情况](#)所示。

表 9-1: 不同数据源的支持情况

数据源	结构迁移	全量迁移	增量迁移
MySQL > RDS for MySQL	支持	支持	支持
MySQL > DRDS	不支持	支持	支持
Oracle > DRDS	不支持	支持	支持
Oracle > RDS for MySQL	支持	支持	支持

数据迁移支持的源实例类型包括：

- RDS实例
- Oracle数据库
- 本地自建数据库

数据迁移支持的目标实例包括：

- RDS实例
- DRDS实例

在线迁移方式

DTS默认使用在线迁移，在线迁移只要用户配置迁移的源、目标实例及迁移对象即可，DTS会自动完成整个数据迁移过程。在线迁移支持数据不停服迁移，但要求DTS服务器能够同时跟源实例、目标实例连通。

多种迁移步骤

数据迁移支持多种迁移步骤：结构迁移、全量数据迁移及增量数据迁移，具体说明如下。

- 结构迁移：将源实例中的结构对象定义一键迁移至目标实例。
- 全量数据迁移：将源实例中的历史存量数据迁移至目标实例。
- 增量数据迁移：将迁移过程中源实例产生的增量数据实时同步到目标实例。结构迁移 + 全量数据迁移 + 增量数据迁移可以简单实现业务不停服迁移。

多种ETL特性

数据迁移支持多种ETL特性，主要包括：

- 支持库表列三级对象名映射。库表列三级对象名映射是指可以实现源实例和目标实例的库名或表名，甚至列名不同的两个对象之间的数据迁移。
- 支持迁移数据过滤，用户可以对要迁移的表配置SQL条件以过滤要迁移的数据。例如用户可以配置时间条件，只迁移最新的数据。

报警机制

数据迁移提供迁移异常报警，一旦迁移任务出现异常，即会向任务的owner发送报警短信，让用户第一时间了解并处理异常任务。

迁移任务

迁移任务是DTS进行数据迁移的基本单元。如果需要进行数据迁移，必须创建一个迁移任务。当创建迁移任务时，需要配置待迁移源实例的连接方式、目标实例的连接方式、迁移对象及迁移类型等信息。用户可以在DTS控制台进行迁移任务的创建、管理、停止及删除等操作。

迁移任务在创建及运行过程中，不同阶段会处于不同的状态，具体如表 9-2: 迁移状态及说明所示。

表 9-2: 迁移状态及说明

迁移状态	状态说明	可进行操作
未启动	迁移任务已经完成任务配置，但是还没有进行迁移前的预检查。	• 预检查 • 删除
预检中	迁移任务正在进行前期的预检查。	删除
预检通过	迁移任务已经通过迁移前的预检查，但是还没有启动迁移。	• 启动 • 删除

迁移状态	状态说明	可进行操作
迁移中	迁移任务正在进行正常的数据迁移。	• 暂停 • 结束 • 删除
迁移失败	迁移任务异常，可以根据任务的进度确认具体是哪个阶段失败。	删除
暂停中	迁移任务已被暂停。	• 启动 • 删除
完成	迁移任务已完成，或者用户单击 结束 停止数据迁移。	删除

9.4.2 数据同步

数据实时同步功能旨在帮助用户实现两个数据源之间的数据实时同步。

通过数据实时同步功能可实现异地多活、数据异地灾备、本地数据灾备、数据异地多活、跨境数据同步、查询、报表分流、云BI及实时数据仓库等多种业务场景。

功能列表

DTS支持的同步功能如下表 9-3: 同步功能列表所示。

表 9-3: 同步功能列表

同步源实例	同步目标实例	单向/双向同步
MySQL	MySQL	• 单向同步 • 双向同步
DRDS	DRDS	单向同步

同步对象

- 数据同步的同步对象的选择粒度可以为：库、表、列。用户可以根据需要同步某几个表的数据。
- 数据同步支持库、表、列名映射，即用户可以进行两个不同库名的数据库之间的同步，或两个不同表名的表之间的数据同步。
- 数据同步支持列选择，即用户可以根据业务需求，只同步表中的某几列数据。

同步作业

同步作业是数据实时同步的基本单元。如果要进行两个实例间的数据同步，必须在DTS控制台创建同步作业。

同步作业在创建及运行过程中，不同阶段会处于不同的状态，具体如表 9-4: 作业状态及说明所示。

表 9-4: 作业状态及说明

作业状态	状态说明	可进行操作
预检中	同步作业正在进行启动前的预检查。	- 查看同步配置 - 删除同步 - 复制同步配置 - 配置监控报警
预检查失败	同步作业预检查没有通过。	- 预检查 - 查看同步配置 - 修改同步对象 - 修改同步速度 - 删除同步 - 复制同步配置 - 配置监控报警
未启动	迁移任务已经通过迁移之前的预检查，但是还没有启动。	- 预检查 - 开始同步 - 修改同步对象 - 修改同步速度 - 删除同步 - 复制同步配置 - 配置监控报警
同步初始化中	同步作业正在进行同步初始化。	- 查看同步配置 - 删除同步 - 复制同步配置 - 配置监控报警
同步初始化失败	同步作业在初始化过程中，迁移失败。	- 查看同步配置 - 修改同步对象 - 修改同步速度 - 删除同步 - 复制同步配置 - 配置监控报警
同步中	同步作业正常同步中。	查看同步配置

作业状态	状态说明	可进行操作
		• 修改同步对象 • 修改同步速度 • 暂停同步 • 删除同步 • 复制同步配置 • 配置监控报警
同步失败	同步作业同步异常。	• 查看同步配置 • 修改同步对象 • 修改同步速度 • 启动同步 • 删除同步 • 复制同步配置 • 配置监控报警
暂停中	同步作业被暂停，处于暂停状态。	• 查看同步配置 • 修改同步对象 • 修改同步速度 • 启动同步 • 删除同步 • 复制同步配置 • 配置监控报警

高级特性

数据订阅支持多种特性，有效降低用户使用门槛，主要包括：

- 动态增减同步对象

在数据同步过程中，用户可以随时增加或减少需要同步的对象。

- 完善性能查询体系

数据同步提供同步延迟、同步性能（RPS、流量）趋势图，用户可以方便查看同步链路的性能趋势。

- 完善监控体系

数据同步提供同步作业状态、同步延迟的报警监控功能。用户可以根据业务敏感度，自定义同步延迟报警阈值。

9.4.3 数据订阅

实时数据订阅功能旨在帮助用户获取RDS/DRDS的实时增量数据，用户能够根据自身业务需求自由消费增量数据，例如实现缓存更新、业务异步解耦、异构数据源数据实时同步及含复杂ETL的数据实时同步等多种业务场景。

功能列表

- 支持经典网络、专有网络下RDS For MySQL实例的数据订阅。

数据源类型

实时数据订阅支持的数据源类型包括：

- RDS For MySQL
- DRDS

其中，DRDS不记录事务日志，如果需要订阅DRDS的实时增量数据，则需要通过订阅DRDS底层挂载的RDS实例的增量日志来实现。

订阅对象

数据订阅的订阅对象可以为：库和表。用户可以根据需要订阅某几个表的增量数据。

数据订阅将增量数据细分为数据变更（Data Manipulation Language 简称DML）和结构变更（Data Definition Language，简称DDL），配置数据订阅时，可以选择需要订阅的具体数据变更类型。

订阅通道

订阅通道是进行增量数据订阅与消费的基本单元。如果要订阅RDS的增量数据，必须在DTS控制台创建一个针对这个RDS实例的订阅通道。订阅通道会实时拉取RDS的增量数据，并将最新一段时间的增量数据保存在订阅通道中，用户可以使用DTS的SDK从这个订阅通道中订阅增量数据，并进行相应的消费。同时，用户可以在DTS控制台进行订阅通道的创建、管理及删除等操作。

一个订阅通道同时只能被一个下游SDK订阅消费，如果用户有多个下游需要订阅同一个RDS实例，需要创建多个订阅通道。这些订阅通道订阅的RDS实例均为同一个实例ID。

订阅通道在创建及运行过程中，不同阶段会处于不同的状态，具体如[表 9-5: 通道状态及说明](#)所示。

表 9-5: 通道状态及说明

通道状态	状态说明	可进行操作
预检中	订阅通道已经完成任务配置，正在进行启动前的预检查。	删除订阅
未启动	迁移任务已经通过迁移前的预检查，但是还没有启动订阅。	- 开始订阅 - 删除订阅
初始化	订阅通道正在进行启动初始化，一般需要1分钟左右。	删除订阅
正常	订阅通道正在正常拉取RDS实例的增量数据。	- 查看示例代码 - 查看订阅数据 - 删除订阅
异常	订阅通道拉取RDS实例增量数据异常。	- 查看示例代码 - 删除订阅

高级特性

数据订阅支持多种特性，有效降低用户使用门槛，主要包括：

- 动态增减订阅对象

在数据订阅过程中，用户可以随时增加或减少需要订阅的对象。

- 在线查看订阅数据

数据传输DTS控制台支持在线查看订阅通道中的增量数据。

- 修改消费时间点

数据订阅支持用户随时修改需要消费数据对应的时间点。

- 完善监控体系

数据订阅提供订阅通道状态、下游消费延迟的报警监控功能。用户可以根据业务敏感度，自定义消费延迟报警阈值。

9.5 应用场景

DTS支持数据迁移、数据实时订阅及数据实时同步等多种功能，可以满足以下多种典型应用场景。

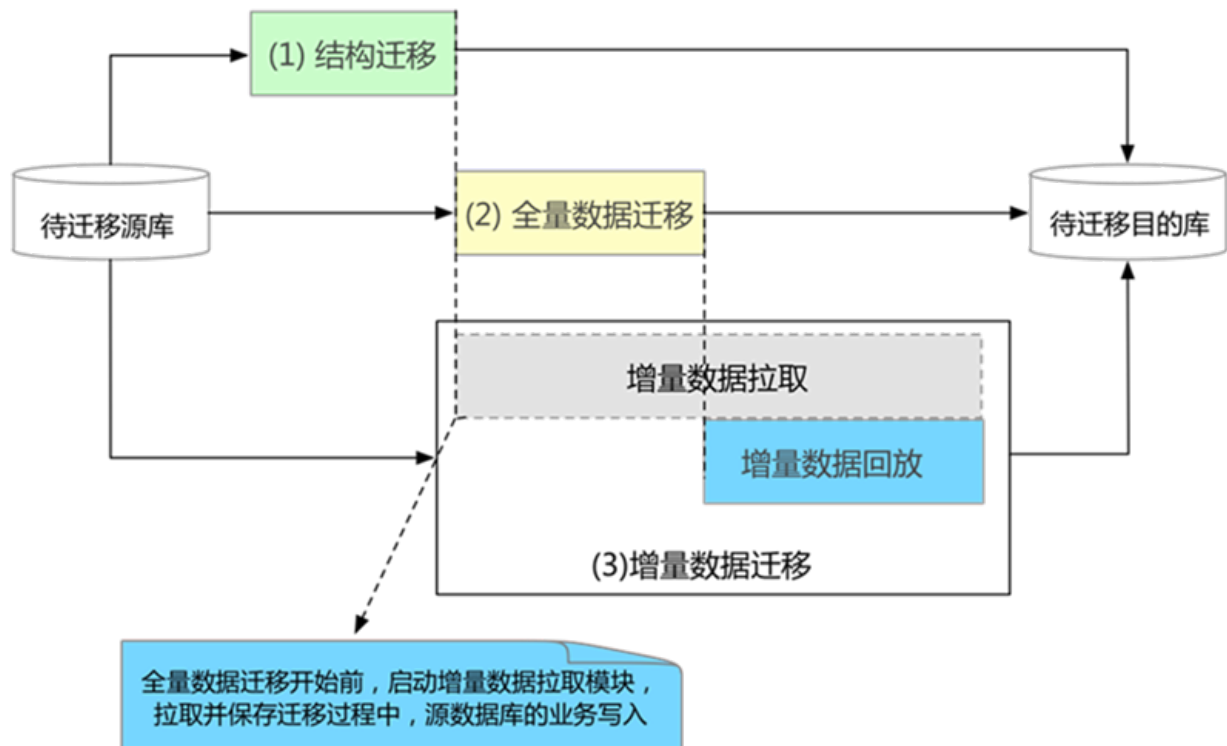
不停服迁移，迁移时业务停机时间降低到分钟级别

很多用户希望系统迁移时，尽可能不影响业务。然而在迁移过程中，如果业务不停服，迁移数据就会发生变化，无法保证迁移数据的一致性。为了保证迁移数据一致性，很多第三方迁移工具，要求在数据迁移期间，应用停止服务。整个迁移过程，业务可能需要停机数小时甚至数天，对业务影响极大。

为了降低数据库迁移门槛，DTS提供不停服迁移解决方案，让数据迁移过程中的业务停机时间降低到分钟级别。

不停服迁移的实现原理如图 9-5: 不停服迁移所示。

图 9-5: 不停服迁移



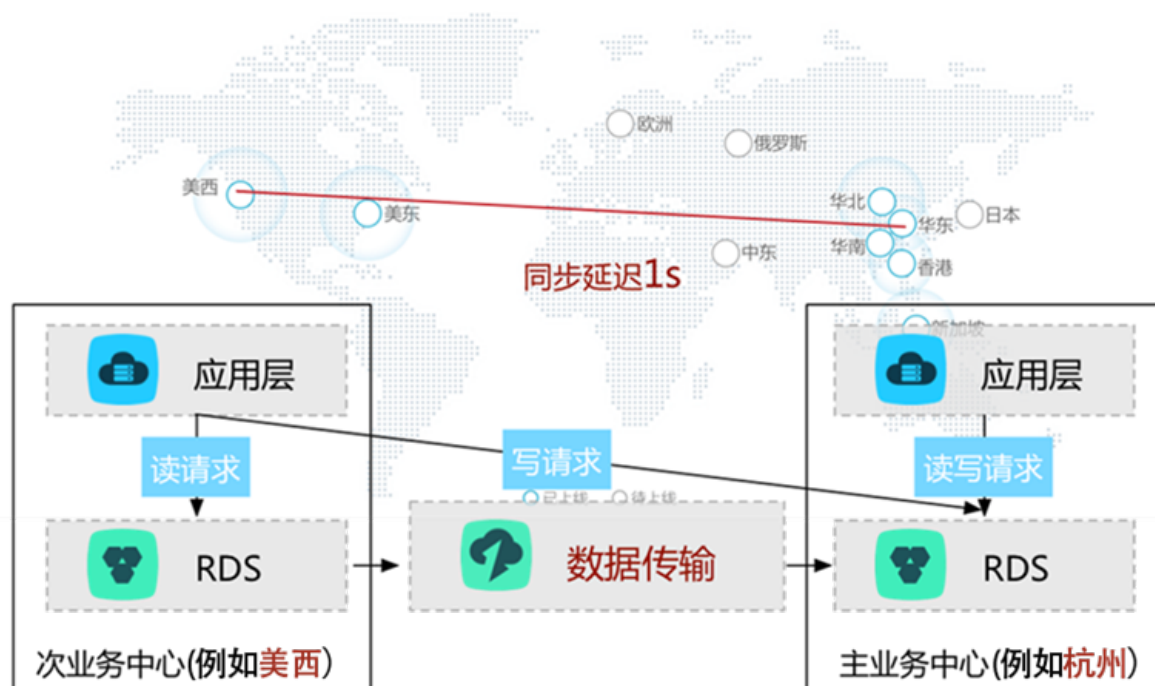
不停服迁移的迁移类型需包含结构迁移、全量数据迁移及增量数据迁移三个阶段。当进入增量数据迁移阶段时，目标实例会保持跟源数据库之间的数据实时同步，用户可以在目标数据库进行业务验证，当验证通过后，直接将业务切换到目标数据库，从而实现整个系统迁移。

由此可见，在整个迁移过程中，只有当业务从源实例切换到目标实例期间，会产生业务闪断，其他时间业务均能正常服务。

加速全球化业务访问速度，为跨境业务赋能

对于用户分布比较广的业务，例如全球化业务，如果按照传统架构，只在单地区部署服务，那么其他地区的用户需要跨地区远距离访问服务，导致访问延迟大、用户体验差的问题。为了加速全球化业务访问速度，优化访问体验，可以将架构调整为如图 9-6: 降低跨地区访问延迟所示。

图 9-6: 降低跨地区访问延迟



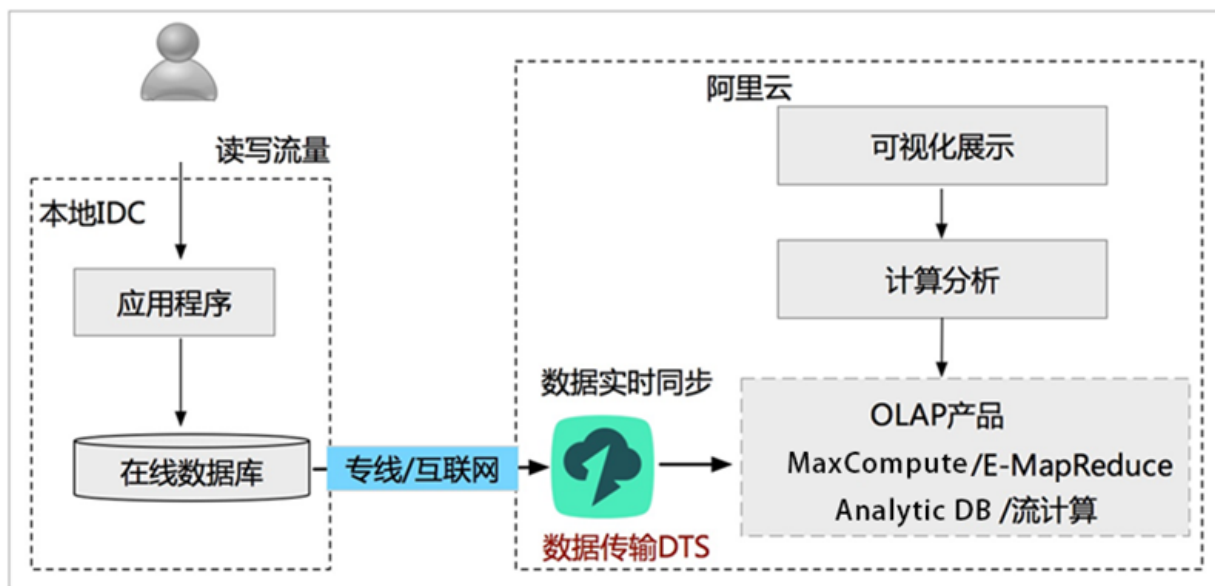
在这个架构中，我们定义了中心和单元的概念，所有地区用户的写请求全部路由回中心。通过DTS将中心的数据实时同步到各个单元，各个地区的用户的读请求，可以路由到就近的单元，从而避免远距离访问，降低访问延迟，加速全球化访问速度。

云BI，快速搭建定制化BI系统

自建BI系统无法满足越来越高的实时性要求，且操作复杂。而阿里云提供的BI体系，可以在不影响现有架构的情况下快速搭建BI系统，所以越来越多的用户选择在阿里云上搭建满足自身业务定制化要求的BI系统。

DTS可以帮助用户将本地自建数据库的数据实时同步到阿里云的BI存储系统（例如MaxCompute、Analytic数据库或流计算），然后用户使用各种计算引擎进行后续的数据分析，同时可以通过可视化工具进行计算结果的实时展示，或通过迁移工具将计算结果同步回本地IDC，具体实现架构如图 9-7: 云BI所示。

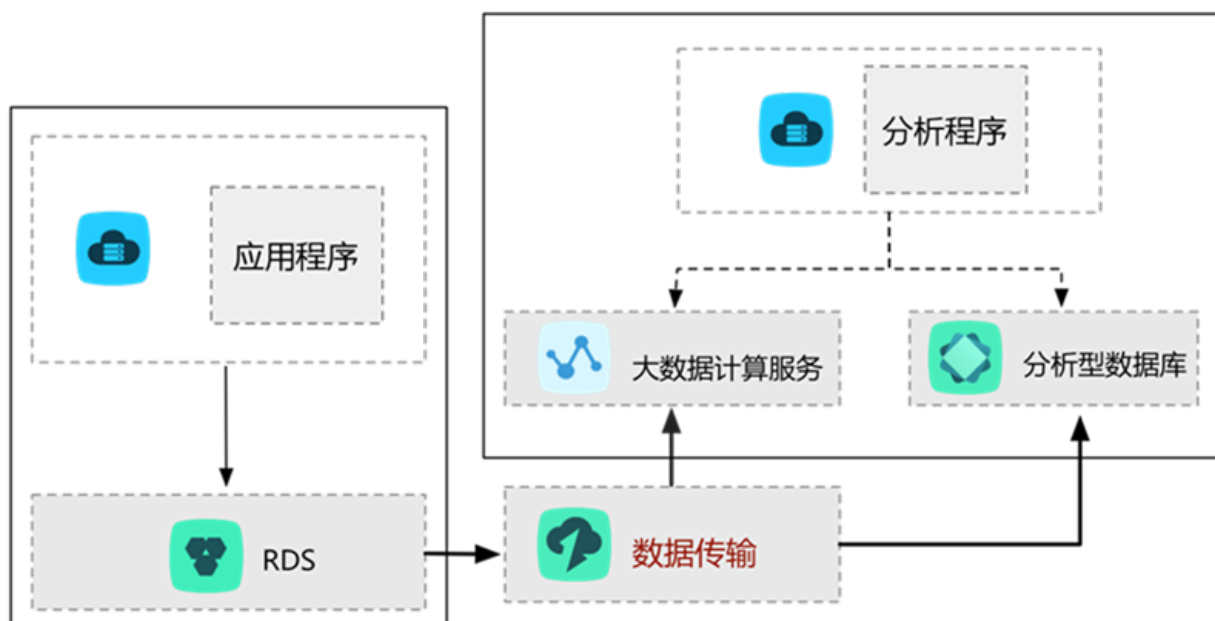
图 9-7: 云BI



数据实时分析，快速抢占商业先机

数据分析在提高企业洞察力和用户体验方面发挥着举足轻重的作用，且实时数据分析能够让企业更快速、灵活地调整市场策略，适应快速变化的市场方向及消费者体验。为了在不影响线上业务的情况下实现实时数据分析，需要将业务数据实时同步到分析系统中，由此可见，实时获取业务数据必不可少。DTS提供的数据订阅功能，可以在不影响线上业务的情况下，帮助您获取业务的实时增量数据，通过SDK可将其同步至分析系统中进行实时数据分析，如图 9-8: 数据实时分析所示。

图 9-8: 数据实时分析

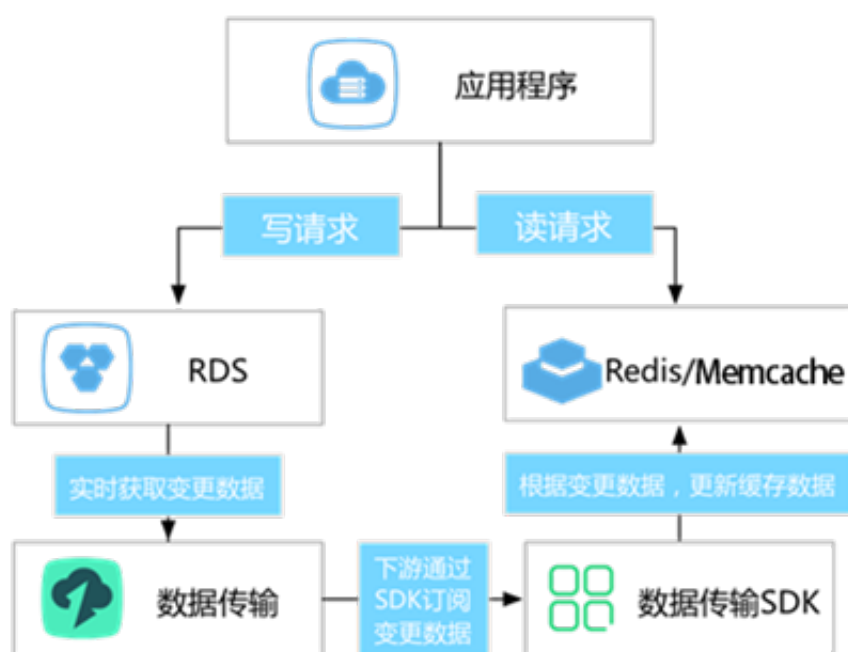


轻量级缓存更新策略，让核心业务更简单可靠

为了提高业务访问速度，提升业务读并发，很多用户都会在业务架构中引入缓存层。业务所有读请求全部路由到缓存层，通过缓存的内存读取机制大大提升业务读取性能。缓存中的数据不能持久化，一旦缓存异常退出，那么内存中的数据就会丢失，所以为了保证数据完整，业务的更新数据会落地到持久化存储中，例如数据库。

如上所述，业务会遇到缓存跟持久化数据库数据同步的问题。DTS提供的数据订阅功能，可以通过异步订阅数据库的增量数据，并更新缓存的数据，实现轻量级的缓存更新策略。这种策略的架构如图 9-9: 缓存更新策略所示。

图 9-9: 缓存更新策略



这种缓存更新策略的优势在于：

- 更新路径短，延迟低

缓存失效为异步流程，业务更新数据库完成后直接返回，不需要关心缓存失效流程，整个更新路径短，更新延迟低。

- 应用简单可靠

应用无需实现复杂双写逻辑，只需启动异步线程监听增量数据，更新缓存数据即可。

- 应用更新无额外性能消耗

因为数据订阅是通过解析数据库的增量日志来获取增量数据，获取数据的过程对业务、数据库性能无损。

业务异步解耦，让核心业务更简单可靠

通过数据订阅，可以将深耦合业务优化为通过实时消息通知实现的异步耦合，让核心业务逻辑更简单可靠。这个应用场景在阿里巴巴内部得到了广泛的应用，目前淘宝订单系统每天有上万个下游业务，通过数据订阅获取订单系统的实时数据更新，触发自身的变更逻辑。

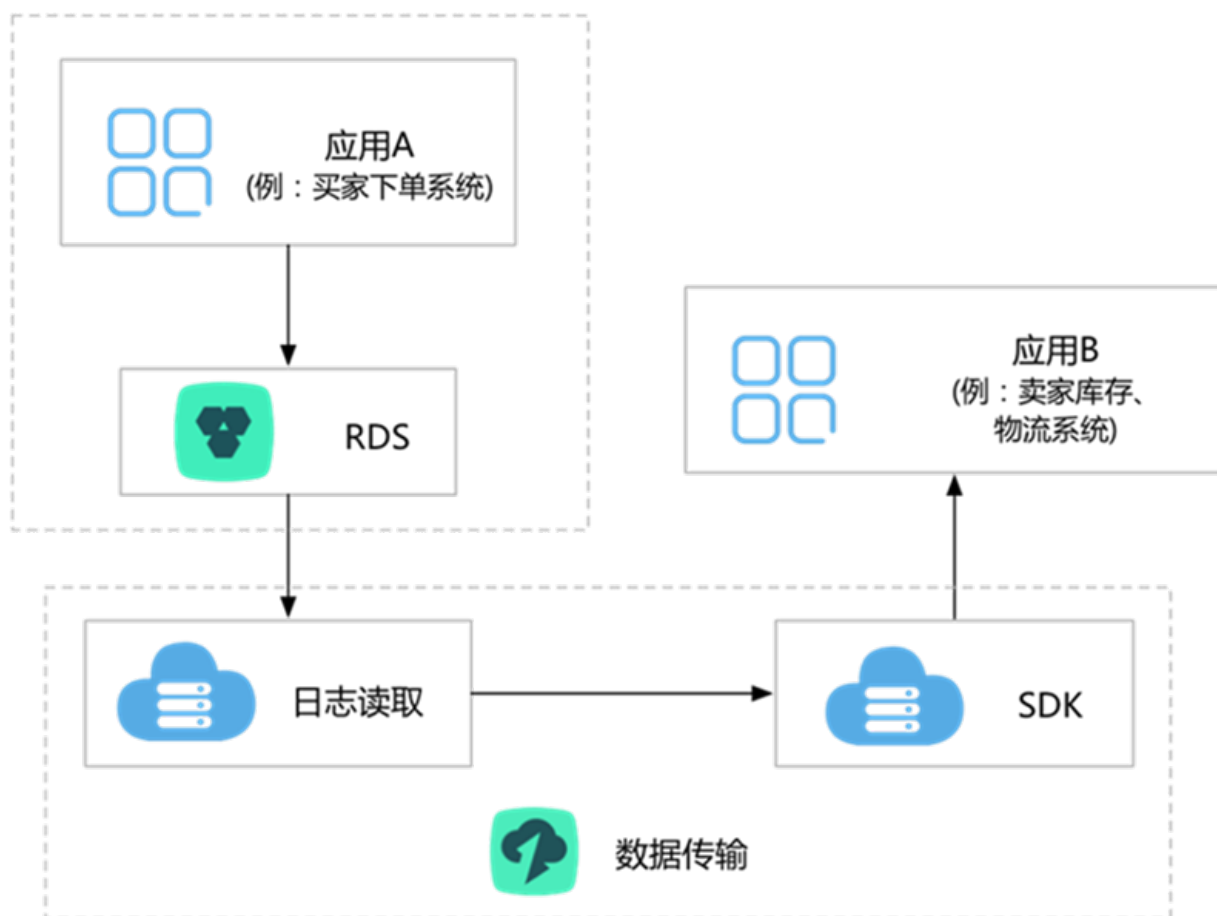
下面举个简单的逻辑，描述下整个应用场景的优势。

例如电商行业，涉及下单系统、卖家库存、物流发货等多个业务逻辑。如果将这些逻辑全部放在下单流程中，那么下单流程为：用户下单，系统通知卖家库存，物流发货等下游业务进行逻辑变更，当全部变更完成后，返回下单结果。这种下单逻辑存在如下问题：

- 下单流程长，用户体验差。
- 系统稳定性差，任何一个下游发生故障，直接影响下单系统的可用性。

为了提升核心应用用户体验，提高稳定性，可以将核心应用和依赖的下游业务异步解耦。让核心应用更稳定可靠。具体调整如[图 9-10: 业务异步解耦](#)所示。

图 9-10: 业务异步解耦



下单系统只要下完单就直接返回，底层通过DTS实时获取订单系统的变更数据，业务通过SDK订阅这个变更数据，并触发库存、物流等下游业务逻辑。由此，保证下单系统的简单可靠。

读能力横向扩展，快速适应业务发展

对于有大量读请求的应用场景，单个RDS实例可能无法承担全部的读取压力，甚至可能对主流程业务产生影响。为了实现读取能力的弹性扩展，分担数据库压力，可以使用DTS的实时同步功能构建只读实例，利用这些只读实例承担大量的数据库读取工作负载，从而方便地扩展了应用的吞吐量。

9.6 基本概念

预检查

预检查是迁移任务启动前的必经阶段，主要是对影响迁移成功的前置条件进行检查。例如源目标实例的连通性，迁移账号的权限等的检查。如果预检查失败，那么可以根据修复方法修复，然后重新进行预检查。

结构迁移

结构迁移是迁移任务的一种迁移类型。在数据库迁移中，它是指进行结构对象定义语法的迁移，包括表、视图、触发器、存储过程、存储函数、同义词等结构对象的语法迁移。对于异构数据库之间的迁移，在结构迁移阶段进行数据类型的映射，并根据源实例和目标实例的语法定义，对对象定义语法进行调整。

全量数据迁移

全量数据迁移是迁移任务的一种迁移类型。它是指将源实例数据库中的所有数据，不包括结构语法定义，迁移到目标实例。如果创建迁移任务时，只选择全量数据迁移，而不选增量数据迁移，那么迁移过程中源实例的新增数据不会迁移到目标实例。

增量数据迁移

增量数据迁移是迁移任务的一种迁移类型。它是指将迁移过程中，将源实例写入的增量数据同步到目标实例。如果创建迁移任务时，选择了全量数据迁移及增量数据迁移，那么DTS会先在源实例实现静态快照，先将快照数据迁移到目标实例，再将迁移过程中源实例的增量数据同步到目标实例。增量数据迁移是一个保持目标实例与源实例数据实时同步的过程，不会自动结束，如果需要结束迁移，需要在控制台手动结束任务。

同步初始化

同步初始化是指在同步链路增量数据同步前，将同步对象的历史数据初始化到目标实例。

同步初始化分为结构初始化和全量数据初始化。结构初始化是进行同步对象的结构定义的初始化。

全量数据初始化是进行同步对象数据的初始化。

同步性能

同步性能是指每秒同步到目标实例的记录数，单位为RPS。

同步延迟

同步延迟是指同步到目标实例的最新数据在源数据库执行的时间戳，和源实例当前时间戳的差值。

同步延迟反映了目标实例和源实例的数据时间差。当同步延迟为0时，表示源实例和目标实例数据完全一致。

订阅通道ID

订阅通道ID是订阅通道的唯一标识，购买完订阅通道，DTS会自动生成订阅通道ID。用户使用SDK消费增量数据时，需要配置对应的订阅通道ID。在DTS控制台的订阅列表中，显示每个订阅通道对应的订阅通道ID。

数据更新

DTS将数据库中的更新数据类型分为：数据更新和结构更新。数据更新是指只修改数据，但是不修改结构对象定义。例如insert、update、delete等。

结构更新

DTS将数据库中的更新数据类型分为：数据更新和结构更新。结构更新是指修改了结构对象定义的语法。例如create table、alter table、drop view 等。用户可以在创建订阅通道时，选择是否订阅结构更新。

数据范围

数据范围是指订阅通道中存储的增量数据时间戳的范围，增量数据对应的时间戳是这条增量数据在RDS实例中应用完并写入事务日志的时间戳。默认情况下订阅通道中只保留最新一天的增量数据。DTS会定期清理过期的增量数据，同时更新订阅通道的数据范围。

消费时间点

消费时间点是指下游SDK订阅数据且已经被消费掉的最新一条增量数据对应的时间戳。SDK每消费一条数据都向DTS服务端汇报ACK，服务端会更新并保存这个SDK对应的消费时间点，当SDK异常重启时，服务端会自动从最后的消费位点推送订阅数据。

10 云数据库KVStore for Redis

10.1 什么是云数据库KVStore for Redis

阿里云数据库KVStore for Redis (KVStore for Redis) 是兼容开源Redis协议的在线Key-Value存储服务。它支持字符串 (String)、链表 (List)、集合 (Set)、有序集合 (SortedSet)、哈希表 (Hash) 等多种数据类型；它还支持事务 (Transactions)、消息订阅与发布 (Sub/Pub) 等高级功能。通过内存加硬盘的存储方式，云数据库KVStore for Redis在提供高速数据读写能力的同时满足数据持久化需求。

除此之外，云数据库KVStore for Redis作为云计算服务，其硬件和数据部署在云端，有完善的基础设施、网络安全保障和系统维护服务，确保用户可以专注于业务创新。

10.2 产品优势

性能卓越

- 支持集群功能，提供128 GB及以上集群实例规格，可满足大容量和高性能需求。
- 提供32 GB及以下的主-从双节点实例，满足一般用户的容量和性能需求。

弹性扩容

- 存储容量一键扩容：您可根据业务需求通过控制台对实例存储容量进行调整。
- 在线扩容不中断服务：可在线进行调整实例存储容量，无需停止服务，不影响您的业务。

资源隔离

针对实例级别的资源隔离，可以更好地保障单个用户服务的稳定性。

数据安全

- 数据持久化存储：采用内存+硬盘的存储方式，在提供高速数据读写能力的同时满足数据持久化需求。
- 数据主从双备份：所有数据在主从节点上进行双备份，确保数据不丢失。
- 访问控制：支持密码认证方式以确保访问安全可靠。
- 数据传输加密：支持安全套接层协议 (Secure Sockets Layer，简称SSL) 和安全传输层协议 (Transport Layer Security，简称TLS) 的安全加密，保障数据传输的安全性。

高可用

- 每个实例均有主从双节点：避免单点故障引起的服务中断。
- 硬件故障自动检测与恢复：自动侦测硬件故障并在数秒内切换，尽量减少突发硬件故障对服务的影响。

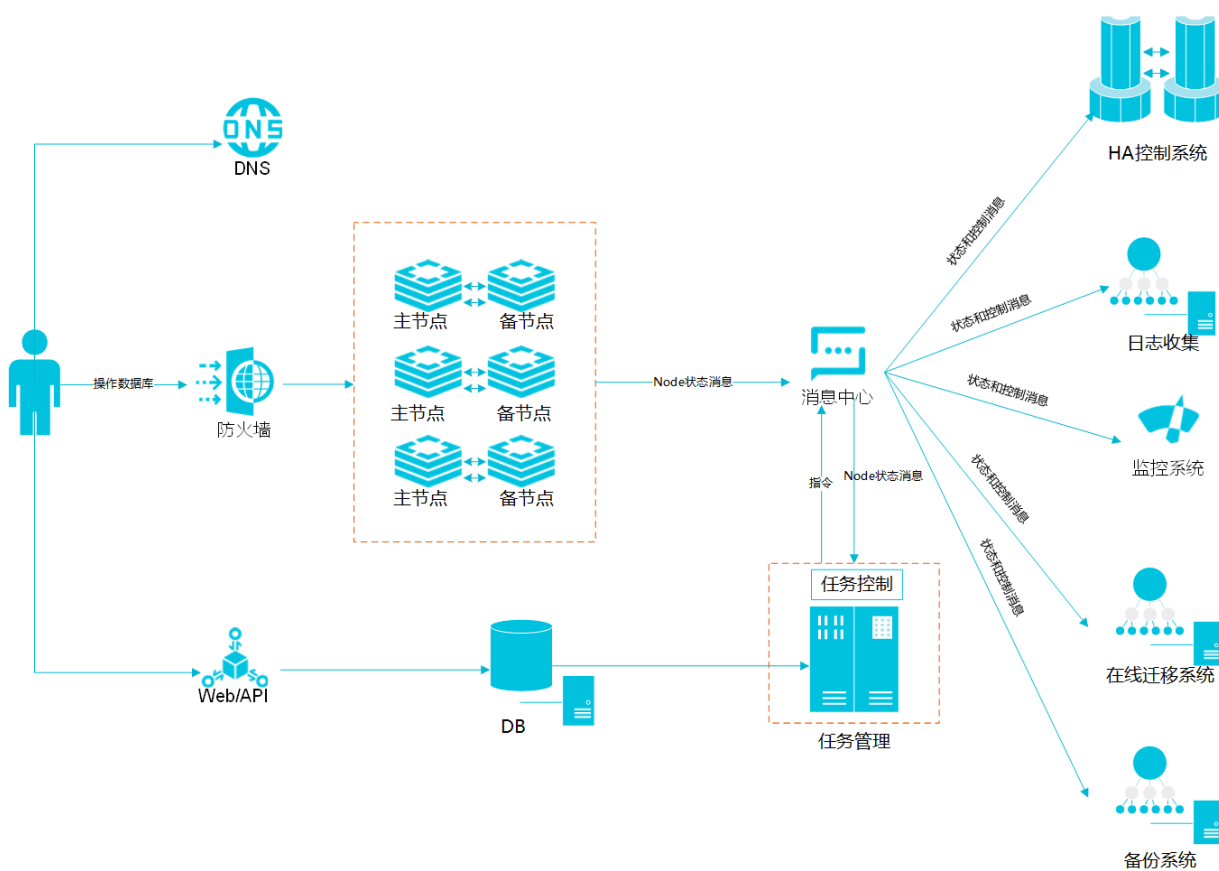
简单易用

- 服务开箱即用：支持即开即用的方式，购买之后即可使用，方便业务快速部署。
- 兼容开源Redis：兼容Redis命令，任何Redis客户端都可以轻松地与云数据库KVStore for Redis建立连接进行数据操作。

10.3 产品架构

云数据库KVStore for Redis的基础架构图，如图 10-1: 架构图所示。

图 10-1: 架构图



云数据库KVStore for Redis自动搭建好主备双节点结构供用户使用。

- HA 控制系统

实例高可用探测模块，用于探测监听Redis实例运行状况。如果断定主节点实例不可用，则自动进行主备节点的切换操作，保证Redis实例的高可用。

- **日志收集**

进行Redis运行情况的日志收集，包括实例慢查询日志，访问控制日志等。

- **监控系统**

进行Redis实例性能监控信息的收集工作，目前包括基本信息组监控，keys组信息监控，String信息组监控等核心信息。

- **在线迁移系统**

当实例所运行的物理机出现故障，在线迁移系统会根据备份系统中的备份文件进行实例重新搭建，保证业务不受影响。

- **备份系统**

针对Redis实例进行备份处理，并且将生成的备份文件存储至OSS系统上进行保存。目前Redis备份系统支持用户自定义备份设置，临时备份并且保存7天内的备份文件。

- **任务控制**

云数据库KVStore for Redis实例支持多种管理控制任务，如创建实例，变更配置，备份实例等，任务系统会根据用户下发的操作指令，进行控制并且进行任务跟踪及出错管理。

10.4 功能特性

- **采用高可用技术，保障业务顺畅进行。**

系统工作时主节点（Master）和备节点（Slave）数据实时同步，主节点故障时系统自动秒级切换，备节点接管业务，故障转移过程不影响业务，保障系统服务的高可用性。

集群（Cluster）实例采用分布式架构，每个节点都采用一主一从的高可用架构，自动进行容灾切换和故障迁移，保障服务的高可用性。

- **支持一键备份和恢复，可定制备份策略。**

在控制台可即时执行备份操作，并且可以进一步定制自动备份策略。系统自动保留7天内的备份数据，支持一键恢复，有效防范数据误操作，将业务损失降到最低。

- **采取多种网络安全防护措施，保障数据安全。**

VPC私有网络在TCP层直接进行网络隔离保护；DDOS防护实时监测并清除大流量攻击；支持1000个以下IP白名单配置，隔绝非法登录操作。

- **采用优化后的内核，避免漏洞攻击。**

阿里云专家团队对源码Redis进行深度内核优化，有效防止内存溢出，修复安全漏洞，为您保驾护航。

- **支持弹性扩容，突破容量和性能瓶颈。**

云数据库KVStore for Redis支持多种内存规格的产品配置，可根据业务量大小升级内存规格。

支持集群架构下弹性扩展数据库系统的存储空间及吞吐性能，突破性能瓶颈。

- **提供多种实例规格，灵活变配。**

支持单节点缓存架构和双节点存储架构，并且能根据不同业务的场景进行实例配置变更。

- **提供监控告警功能，用户可实时了解实例状态。**

提供CPU利用率、连接数、磁盘空间利用率等监控信息，并提供报警功能，用户可随时随地了解实例动态。

- **提供可视化运维平台，运维操作简便。**

在运维平台可对实例克隆、备份、数据恢复等高频高危操作进行一键式操作。

- **数据库内核版本自动升级，避免软件缺陷。**

实例主动升级，快速修复缺陷，免去日常版本管理的苦恼。

- **自定义参数配置，定制个性行为。**

在控制台可设置Redis参数，最大化地利用系统资源。

10.5 应用场景

游戏行业应用

游戏行业可以选择云数据库KVStore for Redis作为重要的部署架构组件。

场景一：Redis 作为存储数据库使用

游戏部署架构相对简单，主程序部署在ECS上，所有业务数据存储存储在Redis中，作为持久化数据库。云数据库KVStore for Redis支持持久化功能，主备双机冗余数据存储。

场景二：Redis作为缓存加速应用访问

Redis作为缓存层，加速应用访问，而数据则存储在后端的数据库（RDS）中。

Redis的服务可靠性至关重要，一旦Redis服务不可用，将导致后端数据库无法承载业务访问压力。云数据库KVStore for Redis提供双机热备的高可用架构，保障极高的服务可靠性。主节点对外提供服务，当主节点出现故障，系统自动切换备用节点接管服务，整个切换过程对用户全部透明。

视频直播类应用

视频直播类业务往往会重度依赖Redis业务。存储用户数据及好友互动关系。

双机热备保障高可用

云数据库KVStore for Redis提供双机热备的方式，可以极大地保障服务的可用性。

集群版解决性能瓶颈

云数据库 Redis 提供集群版实例，破除Redis单线程机制的性能瓶颈，可以有效地应对视频直播类流量突起，满足高性能的需求。

轻松扩容应对业务高峰

云数据库KVStore for Redis可支持一键扩容，整个升级过程对用户全透明，可以从容应对流量突发对业务产生的影响。

电商行业应用

电商行业大量采用Redis，多数应用在商品展示、购物推荐等模块。

场景一：秒杀类购物系统

大型促销秒杀系统，系统整体访问压力非常大，一般的数据库根本无法承载这样的读取压力。

云数据库KVStore for Redis支持持久化功能，可以直接选择 Redis 作为数据库系统使用。

场景二：带有计数系统的库存系统

底层用RDS存储具体数据信息，数据库字段中存储具体计数信息。数据库Redis来进行计数的读取，RDS存储计数信息。云数据库Redis版部署在物理机上，底层基于SSD高性能存储，可以提供极高的数据存储能力。

10.6 使用限制

项目	说明
List 数据类型	没有List个数限制，单个元素最大值为512 MB，推荐List的元素个数小于8192，value最大长度不超过1 MB。
Set 数据类型	没有Set个数限制，单个元素最大值为512 MB，推荐Set的元素个数小于8192，value最大长度不超过1 MB。
Sorted Set 数据类型	没有Sorted Set个数限制，单个元素最大值为512 MB，推荐Sorted Set的元素个数小于8192，value最大长度不超过1 MB。
Hash 数据类型	没有field个数限制，单个元素最大值为512 MB，推荐元素个数小于8192，value最大长度不超过1 MB。

项目	说明
DB 数限制	每个实例支持256 个DB。
Redis 命令支持	详情请参见《云数据库KVStore for Redis用户指南》中 Redis支持的命令 章节。
监控报警	云数据库Redis版未提供容量告警，需要用户到云监控中进行配置。建议设置如下监控的报警：实例故障、实例主备切换、已使用连接百分比、操作失败数、已用容量百分比、写入带宽使用率、读取带宽使用率。
数据过期删除策略	- 主动过期，系统后台会周期性的检测，发现已过期的key时，会将其删除。 - 被动过期，当用户访问某个key时，如果该key已经过期，则将其删除。
空闲连接回收机制	服务端不主动回收Redis空闲连接，由用户管理。
数据持久化策略	采用AOF_FSYNC_EVERYSEC方式，每秒fysnc。

10.7 基本概念

Redis

Redis是一款依据BSD开源协议发行的高性能Key-Value存储系统（cache and store）。

实例 ID

实例对应一个用户空间，是使用Redis的基本单位。

Redis对单个实例根据不同的容量规格有不同的连接数、带宽、CPU处理能力等限制。用户可在控制台看到自己购买的实例ID列表。Redis实例分为主-从双节点实例和高性能集群实例两种。

主-从双节点实例

指具备主-从架构的云数据库KVStore for Redis实例。主-从双节点能扩展的容量和性能有限。

高性能集群实例

指具有集群扩展性的云数据库KVStore for Redis实例。集群实例有更好的扩展性和性能，但是在功能上也有一定的限制。

连接地址

用于连接云数据库KVStore for Redis的Host地址。以域名方式展示，可在**实例信息**页面中的**连接信息**处查询。

逐出策略

是指在Redis内存达到`maxmemory`配置指令所设置的值时，Redis所采取的旧数据删除策略。云数据库KVStore for Redis与Redis的逐出策略保持一致，详情请参考[Redis逐出策略](#)。

DB

云数据库KVStore for Redis支持256个DB，默认写入到第0个DB中。

10.8 实例规格



说明：

带宽上限分别应用于上行带宽和下行带宽。

标准版-双副本

表 10-1: 标准套餐

规格	产品代码	连接数上限	带宽上限(Mb)	处理能力	说明	同城容灾部署情况
1G主从普通版同城容灾	redis.logic.sharding.drredisdb1g.1db.0rodb.4proxy.default	10000	10	单核	主-从双节点-同城容灾规格	同城双机房部署
2G主从普通版同城容灾	redis.logic.sharding.drredisdb2g.1db.0rodb.4proxy.default	10000	16	单核	主-从双节点-同城容灾规格	同城双机房部署
4G主从普通版同城容灾	redis.logic.sharding.drredisdb4g.1db.0rodb.4proxy.default	10000	24	单核	主-从双节点-同城容灾规格	同城双机房部署
8G主从普通版同城容灾	redis.logic.sharding.drredisdb8g.1db.0rodb.4proxy.default	10000	24	单核	主-从双节点-同城容灾规格	同城双机房部署
16G主从普通版同城容灾	redis.logic.sharding.drredisdb16g.1db.0rodb.4proxy.default	10000	32	单核	主-从双节点-同城容灾规格	同城双机房部署
32G主从普通版同城容灾	redis.logic.sharding.drredisdb32g.1db.0rodb.4proxy.default	10000	32	单核	主-从双节点-同城容灾规格	同城双机房部署

表 10-2: 定制套餐

规格	产品代码	连接数上限	带宽上限(Mb)	处理能力	说明	同城容灾部署情况
1G主从高配版	redis.master.small.special2x	20000	48	单核	主-从双节点实例	单机房部署
2G主从高配版	redis.master.mid.special2x	20000	48	单核	主-从双节点实例	单机房部署
4G主从高配版	redis.master.stand. special2x	20000	48	单核	主-从双节点实例	单机房部署
8G主从高配版	redis.master.large.special1x	20000	48	单核	主-从双节点实例	单机房部署
16G主从高配版	redis.master.2xlarge.special1x	20000	48	单核	主-从双节点实例	单机房部署
32G主从高配版	redis.master.4xlarge.special1x	20000	48	单核	主-从双节点实例	单机房部署

表 10-3: 集群版

规格	产品代码	连接数上限	带宽上限(Mb)	处理能力	说明
16G集群版	redis.sharding.small.default	80000	384	4核	高性能集群实例
32G集群版	redis.sharding.mid.default	80000	384	8核	高性能集群实例
64G集群版	redis.sharding.large.default	80000	384	8核	高性能集群实例
128G集群版	redis.sharding.2xlarge.default	160000	768	16核	高性能集群实例
256G集群版	redis.sharding.4xlarge.default	160000	768	16核	高性能集群实例

表 10-4: 集群版同城容灾

规格	产品代码	连接数上限	带宽上限(Mb)	处理能力	说明
16G集群版同城容灾	redis.logic.sharding.drredisdb16g.8db.0rodb.8proxy.default	80000	384	8核	集群版同城容灾规格
32G集群版同城容灾	redis.logic.sharding.drredisdb32g.8db.0rodb.8proxy.default	80000	384	8核	集群版同城容灾规格
64G集群版同城容灾	redis.logic.sharding.drredisdb64g.8db.0rodb.8proxy.default	80000	384	8核	集群版同城容灾规格
128G集群版同城容灾	redis.logic.sharding.drredisdb128g.16db.0rodb.16proxy.default	160000	768	16核	集群版同城容灾规格
256G集群版同城容灾	redis.logic.sharding.drredisdb256g.16db.0rodb.16proxy.default	160000	768	16核	集群版同城容灾规格

QPS能力参考

表 10-5: QPS能力参考

规格 (GB)	连接数上限 (个)	内网带宽上限 (MB)	CPU 处理能力	QPS 参考值
8	10000	24	单核	80000



说明：

非集群版实例的QPS能力参考范围为8万-10万，集群实例的QPS参考值为节点数目乘以8万-10万。

11 云数据库MongoDB

11.1 什么是云数据库MongoDB

阿里云数据库MongoDB是阿里云基于MongoDB专业打造的高性能分布式数据存储服务，100%完全兼容MongoDB协议，提供稳定可靠、弹性伸缩的数据库服务。同时提供容灾、备份、恢复、监控、报警等方面的全套数据库解决方案。

云数据库MongoDB版具有以下基本特点：

- 自动搭建基于三节点的MongoDB副本集供您使用。容灾切换、故障迁移等高级功能为您整体打包好，对您完全透明。
- 提供一键式的数据库备份、恢复功能。您可以通过控制台一键式地进行数据库的常规备份及数据库回溯功能。
- 提供多达20种的性能指标监控及报警功能，数据库性能数据尽收眼底。
- 提供可视化的数据管理工具，方便您运维。

11.2 产品优势

• 高可用

- 三节点副本集高可用架构，提供极高的业务可用性保障。

云数据库MongoDB服务采用三节点副本集的高可用架构，三个数据节点位于不同的物理服务器上，自动同步数据。Primary和Secondary节点提供服务，当Primary节点出现故障，系统自动选举新的Primary节点，当Secondary节点不可用，由备用节点接管服务。

- 自动备份，一键式数据恢复。

每天自动备份数据并上传至对象存储OSS，提高数据容灾能力的同时有效降低磁盘空间占用。通过备份文件将实例数据恢复至原实例，有效防范因误操作等原因对业务数据造成不可逆的影响。

• 高安全

- DDoS防护：在网络入口实时监测，当发现超大流量攻击时，对源IP进行清洗，清洗无效情况下可以触发黑洞机制。
- IP白名单配置：最多支持配置1000个允许连接MongoDB实例IP地址，从访问源进行直接的风险控制。

• 易用性

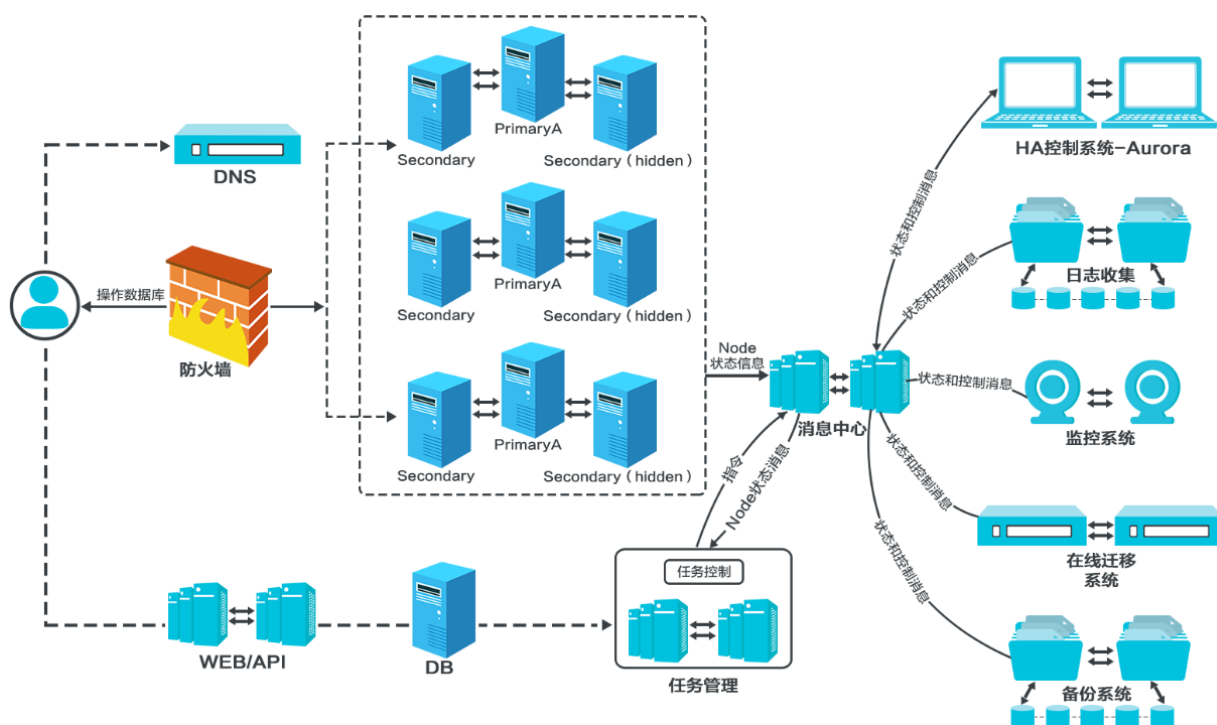
完善的性能监控。提供CPU利用率、IOPS、连接数、磁盘空间等实例信息，实时监控及报警，随时随地了解实例动态。

• 扩展性

副本集模式弹性扩容，云数据库MongoDB支持三节点的副本集模式，支持弹性扩容。当前实例配置无法满足应用的性能要求，或者当前实例的配置过高，您可以变更实例的配置。变更过程完全透明，对业务无影响。

11.3 产品架构

云数据库MongoDB版自动搭建好三节点的副本集供您使用，您可以直接操作Primary节点和一个Secondary节点。系统架构如下图所示：



- **HA控制系统**：实例高可用探测模块，用于探测监听MongoDB实例的运行状况。如果判断主节点实例不可用，进行主备节点的切换操作，保证MongoDB实例的高可用。
- **日志收集**：进行MongoDB运行情况的日志收集，包括实例慢日志查询以及访问控制日志等。
- **监控系统**：进行MongoDB实例性能监控信息的收集工作，包括基础指标、磁盘容量、网络请求以及操作次数等核心信息。
- **在线迁移系统**：当实例所运行的物理机出现故障，在线迁移系统会根据备份系统中的备份文件进行实例重新搭建，保证业务不受影响。

- **备份系统**：针对MongoDB实例进行备份处理，将生成的备份文件存储至OSS系统上。目前MongoDB备份系统支持您自定义备份设置和临时备份，保存7天内的备份文件。
- **任务控制**：云数据库MongoDB实例支持多种管理控制任务，如创建实例、变更配置以及备份实例等，任务系统会根据您下发的操作指令，进行灵活控制并进行任务跟踪及出错管理。

11.4 功能特性

架构灵活

云数据库MongoDB版自动搭建好三节点的副本集供您使用，您可以直接操作Primary节点和一个Secondary节点。如果判断主节点实例不可用，进行主备节点的切换操作，保证MongoDB实例的高可用。

弹性扩容

- **存储容量一键扩容**：您可根据业务需求通过控制台对实例存储容量进行调整。
- **在线扩容不中断服务**：在线调整实例存储容量，无需停止服务，不影响您的业务。

数据安全

- **自动备份**：云数据库MongoDB版支持您自行设置备份周期。备份开始时间可根据您的业务低峰灵活配置；所有备份文件免费保留7天。
- **临时备份**：您在需要时可以临时性发起备份操作；备份文件免费保留7天。
- **数据恢复**：利用备份文件，您可以直接覆盖型恢复至现有实例。
- **备份文件下载**：云数据库会将您的备份文件免费保留7天，在此期间您可以登录管理控制台，将备份文件下载至本地。
- **根据备份集创建实例**：根据备份文件在控制台上一键式地创建一个实例，实现快速部署的需求。
- **IP访问白名单**：提供对实例进行IP访问过滤功能，您可以登录云数据库MongoDB版管理控制台进行IP访问白名单设置，设置后便可实现最高级的访问安全保护，IP白名单最多可配置1000条。
- **多层网络安全防护**

VPC 私有网络在TCP层直接进行网络隔离保护；DDOS防护实时监测并清除大流量攻击；支持1000个以下IP白名单配置

智能运维

- **监控平台**

提供CPU利用率、连接数、磁盘空间利用率等实例信息实时监控及报警，随时随地了解实例动态。

- 可视化管理平台

管理控制平台对实例克隆、备份、数据恢复等高频高危操作可便捷地进行一键式操作。

- 数据库内核版本管理

主动升级，快速修复缺陷，免去日常版本管理苦恼；优化MongoDB参数配置，最大化利用系统资源。

11.5 应用场景

- 业务需要读写分离

MongoDB服务采用三节点副本集的高可用架构，三个数据节点位于不同的物理服务器上，自动同步数据。Primary和Secondary节点提供服务。两个节点分别提供独立域名，配合MongoDB Driver可自行实现读取压力分配。

- 适用于业务灵活多变的场景

由于MongoDB采用No-Schema的方式，对于初创型的业务非常适用，免去变更表结构的痛苦。您可以将模式固定的结构化数据存储存储在RDS中，模式灵活的业务存储在MongoDB，并将高热数据存储在云数据库Memcache或云数据库Redis中，实现对业务数据高效存取，并相应降低成本投入。

- 移动应用

云数据库MongoDB版支持二维空间索引，可以完美的支撑基于位置查询的移动类APP的业务需求。同时MongoDB动态模式存储方式非常适合存储多重系统的异构数据，满足移动APP应用的需求。

- 物联网应用

云数据库MongoDB版性能极高，具备异步数据写入功能，特定场景下可达到内存数据库的性能，非常适合物联网高并发写入的场景。同时MongoDB的map-reduce功能也可以将大量数据进行聚合分析。

- 核心日志系统

云数据库MongoDB版在异步落盘场景下可提供极高的插入性能，具备内存数据库的处理能力。MongoDB提供二级索引功能满足动态查询的需求，并且可以利用map-reduce的聚合框架进行多维度的数据分析。

11.6 使用限制

操作	约束
搭建数据库复制	系统自动搭建了三节点的副本集。其中对您提供了两个节点（Primary和Secondary），另外一个备份节点隐藏对您不可见。您暂时无法自行搭建Secondary节点。
重启数据库	必须通过控制台进行重启实例的操作。

11.7 基本概念

概念	解释
区域	区域指的是您创建的MongoDB实例的服务器所处的地理位置。您需要在创建MongoDB实例时指定，创建实例后暂不支持更改。在创建MongoDB实例时，需要搭配阿里云服务器ECS使用，MongoDB支持内网访问，在区域选择时需要与ECS相同。
实例	MongoDB实例，或简称实例，是您创建MongoDB服务的基本单位。实例是阿里云数据库MongoDB版的运行环境，在主机上以单独的进程存在。您可以通过控制台来创建、修改和删除MongoDB实例。各实例之间相互独立、资源隔离，相互之间不存在CPU、内存、IO等抢占问题；每个实例拥有其自己的特性，例如数据库类型、版本等，系统有相应的参数来控制实例行为。
内存	云数据库MongoDB实例可以使用的内存上限。
磁盘容量	磁盘容量是您创建MongoDB实例时，所选择创建的磁盘大小。实例所占用的磁盘容量除集合数据外，还有实例正常运行所需要的空间，如系统数据库、数据库回滚日志、重做日志、索引等。请确保MongoDB实例具有足够的磁盘容量来存储数据，否则可能导致实例被锁定。若因磁盘容量不足导致实例被锁定，您可以创建更大的磁盘容量来解锁实例。
IOPS	以4KB为单位，每秒进行块设备读写操作的次数上限。
CPU核	实例可以使用的计算能力上限。1个CPU拥有不低于2.3GHz 超线程（Intel Xeon系列Hyper-Threading）的计算能力。
连接数	客户端和MongoDB实例之间的TCP连接。如果客户端使用了连接池，则客户端和MongoDB实例之间的连接为长连接，反之则为短连接。
Mongos	MongoDB集群请求入口，所有的请求都通过mongos进行协调，mongos是一个请求分发中心，它负责把对应的数据请求转发到对应的 shard 服务器上。您可以选择多个mongos作为请求的入口，防止其中一个挂掉所有的MongoDB请求都没有办法操作。

概念	解释
configserver	配置服务器，存储所有数据库元信息（路由、分片）的配置mongos本身没有存储shard服务器和数据路由信息，只是缓存在内存里，配置服务器则实际存储这些数据。mongos第一次启动或者关掉重启就会从configserver加载配置信息，以后如果配置服务器信息变化会通知到所有的mongos更新自己的状态，这样 mongos就能继续准确路由。configserver存储了分片路由的元数据，服务可用性和数据可靠性要求极高，云数据库MongoDB采用三节点副本集的方式全方位保障ConfigSever的服务可靠性。

11.8 实例规格

表 11-1: MongoDB副本集规格参数

规格类型	规格信息	规格代码	最大连接数	最大IOPS
通用规格	1核2G	dds.mongo.mid	500	1000
	2核4G	dds.mongo.standard	1000	2000
	4核8G	dds.mongo.large	2000	4000
	8核16G	dds.mongo.xlarge	4000	8000
	8核32G	dds.mongo.2xlarge	8000	14000
	16核64G	dds.mongo.4xlarge	16000	16000
独享规格	2核16G	mongo.x8.medium	2500	4500
	4核32G	mongo.x8.large	5000	9000
	8核64G	mongo.x8.xlarge	10000	18000
	16核128G	mongo.x8.2xlarge	20000	36000
	32核256G	mongo.x8.4xlarge	40000	72000
独占物理机	60核440G	dds.mongo.2xmonopolize	100000	100000

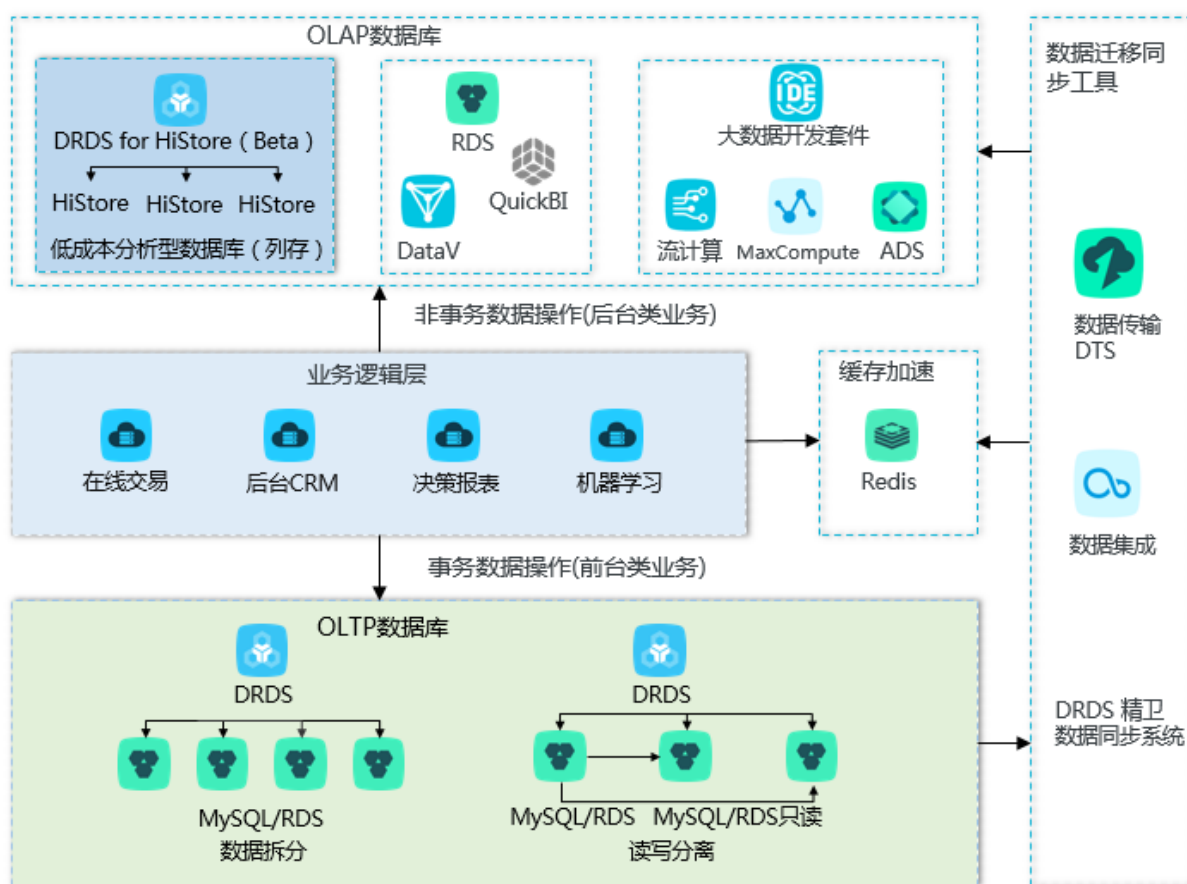
12 分布式关系型数据库DRDS

12.1 什么是分布式关系型数据库服务DRDS

分布式关系型数据库服务DRDS (Distributed Relational Database Service) 是阿里巴巴集团自主研发的中间件产品，专注于解决单机关系型数据库的扩展问题。DRDS兼容MySQL交互协议，支持大部分MySQL的DML和DDL语法，具备分库分表、平滑扩容、服务升降配、透明读写分离等分布式数据库核心能力和特性。产品具备轻量（无状态）、灵活、稳定和高效等特点，并且为您提供分布式数据库全生命周期的运维管控能力。

DRDS主要应用在大规模在线数据操作上。通过贴合业务的拆分方式，将操作效率提升到极致，有效满足地在线业务对关系型数据库的要求。

图 12-1: DRDS产品结构示意图



解决的问题

- **单机数据库容量瓶颈**：随着数据量和访问量的增长，传统单机数据库会遇到很大的挑战，依赖硬件升级并不能完全解决问题。分布式方案利用多个机器共同服务，可以有效解决数据存储容量和访问量的瓶颈问题。
- **关系型数据库扩展困难**：因为分布式数据库天然属性，可以通过数据平滑迁移改变数据分片存放节点，从而实现关系型数据库动态扩展。

12.2 产品优势

分布式架构

通过数据的水平拆分以及单一服务的集群化部署，解决了负载均衡、分布式数据库服务(DRDS)、数据存储 (RDS) 各层面的单点瓶颈问题，实现了分布式数据库的最大优点，即可扩展。

弹性伸缩

可进行DRDS实例以及数据存储 (RDS) 的动态升降配 (水平扩展)，实现服务能力的灵活选择。

高性能

DRDS for RDS (MySQL) 贴合业务进行数据拆分，尽可能让数据跟随主要业务操作进行聚簇，从而提升在线事务型操作的响应速度。DRDS for HiStore通过使用列存、知识网格，极大地提升大范围数据聚合、ad hoc查询等常用分析性操作的响应速度，同时还通过高度压缩控制存储成本。

安全可控

DRDS支持类单机数据库的账号权限体系，并且提供IP白名单、高风险SQL默认禁止等实用功能。即使需要集成到本地管控系统中，也有完整的API 体系作为支撑。同时我们也提供完整的产品支持和架构服务。

12.3 产品架构

DRDS输出方式有两种：阿里云整体输出和阿里中间件单独输出。两种输出方式，在DRDS 依赖的组件上会有所不同，特性上也有部分差异。

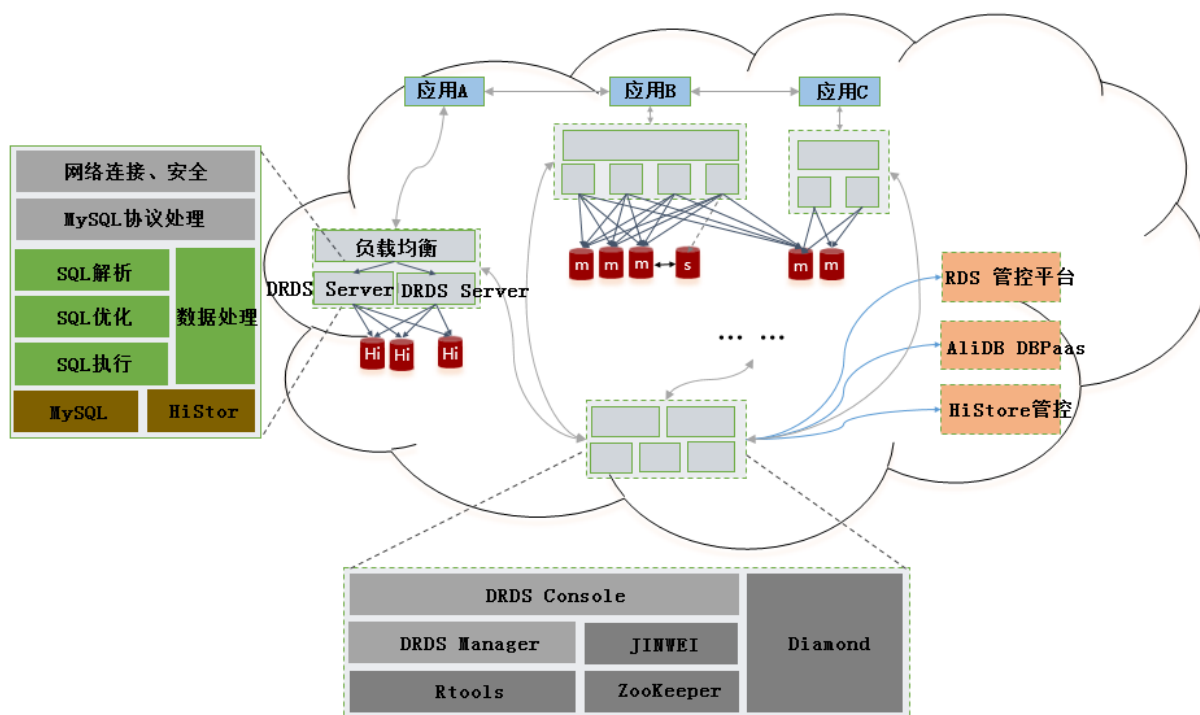
下表简要说明了两者的区别：

差异项	阿里云整体输出	阿里中间件单独输出
MySQL	RDS for MySQL	阿里巴巴集团数据库系统 (DBPaas)
负载均衡	集中式负载均衡 (SLB)	客户端负载均衡 (VIPServer)

差异项	阿里云整体输出	阿里中间件单独输出
特色存储支持	无	高压缩比列式存储（HiStore）

DRDS的系统架构如图 12-2: DRDS系统架构图所示。

图 12-2: DRDS系统架构图



DRDS Server

DRDS Server是DRDS的服务层，由多个服务节点组成服务集群，提供分布式数据库服务，包括读写分离、SQL路由执行、结果合并、动态数据库配置、全局唯一ID服务等功能。

RDS for MySQL（图中m，s表示）

RDS for MySQL主要是负责在线数据数据存储和操作，通过MySQL主备复制实现高可用，配合RDS的主备切换系统实现动态数据库故障转移。

RDS for MySQL的管理控制平台实现了RDS生命周期内的管理、监控、告警、资源管理等工作。

HiStore

当DRDS单独输出时（非阿里云整体输出），DRDS支持HiStore作为物理存储。HiStore是阿里巴巴自研的低成本高性能列存数据库。通过列式存储、知识网格、多核利用，做到相对于行存（如MySQL）更低的成本，更高的数据聚合和即席查询能力。

HiStore的管理控制平台实现了HiStore生命周期内的管理、监控、告警、资源管理等工作。

DBPaaS

当DRDS单独输出时（非阿里云整体输出），自带MySQL数据库运维平台DBPaaS，实现MySQL生命周期内的管理、监控、告警、资源管理等工作。

SLB（集中式负载均衡）

无需在用户侧安装客户端，用户的请求通过集中式负载均衡进行流量分配。当某个实例节点出现故障或者新增服务节点时，都能够通过负载均衡保证底层节点的流量相对均衡分配。

VIPServer（客户端负载均衡）

需要在用户侧安装客户端，弱依赖于中心管控服务（只有在发生负载配置变更时才会进行交互），用户的请求通过该模块进行流量分配，当某个实例节点出现故障或者新增服务节点时，都能够通过均衡负载保证底层节点的流量相对均衡分配。

Diamond配置中心

Diamond（配置中心）是负责DRDS配置存储和管理的系统，提供配置存储、查询、通知功能。在DRDS中主要存储数据库源数据、拆分规则、DRDS开关等配置。

精卫（JINGWEI）

DRDS数据迁移同步主要由精卫系统负责，核心能力包含全量数据迁移和增量数据同步，衍生出来的功能包括数据平滑导入、平滑扩容、全局二级索引能力。精卫数据迁移同步系统需要ZooKeeper和DRDS Rtools提供支撑。

DRDS Console（用户控制台）

DRDS 控制台主要面向业务DBA，按照用户隔离资源和操作，提供实例管理、库表管理、读写分离配置、平滑扩容、监控展现和IP白名单等功能。

DRDS Manager（运维控制台）

DRDS Manager主要面向全局运维和DBA，提供所有DRDS资源管理和系统监控，主要功能包括两个方面：

- RDS实例依赖的所有资源管理，包括虚拟机、负载均衡、域名等资源；
- DRDS实例状态监控，包括QPS、活跃线程、连接数、各节点网络I/O、各节点CPU占用等指标。

Rtools

Rtools是DRDS的运维支撑系统，支持进行数据库配置、读写权重、连接参数等管理以及库表拓扑、拆分规则等管理。

12.4 功能特性

本节介绍 DRDS 的主要功能。

分库分表

多种贴合OLTP业务的拆分方案，让操作聚焦少量数据，提升操作效率。利用分布式特性，让数据库系统每秒操作次数（Operations Per Second，OPS）达到极高的上限，并且可开启并行操作选项，超越单机关系型数据库的操作响应时间。

平滑扩容

通过数据迁移同步，改变数据分片位置，静默完成数据库扩容。

数据库账号体系

类单机数据库账号权限系统，具备多账号、读写DDL等权限分离能力。

读写分离

通过添加RDS或MySQL只读实例，线性地提高数据库读取数据的能力，帮您处理好事务、备机宕机、强制使用主实例或者只读实例的各种细节操作。

服务升降配

通过增减服务节点，让分布式数据库服务瞬间增减能力。

全局唯一数字序列

类似于单机数据库的写入，DRDS具备高性能填充自增主键的能力。无需关注主键在分布式环境下的全局唯一性，并且兼容主键值返回客户端，也支持应用端直接获取一个或者一批全局唯一数字ID。

分布式SQL引擎

提供稳定且性能强劲的分布式SQL引擎。SQL边界定义清晰明确，支持分布式Join和二次聚合排序计算，提供清晰易用的Explain、Outline功能。

分布式事务

提供稳定可靠的原生XA强一致事务与TCC柔性事务支持。

12.5 应用场景

本节介绍 DRDS 的典型应用场景。

DRDS非常适合用于前台高并发低延迟类业务。通过贴合业务场景的数据拆分，辅以分布式二级索引，让业务数据库保持极高的QPS上限。

DRDS目前也尝试支持阿里巴巴自研列式数据库以满足超大数据量低成本存储，高效数据聚合，以及ad hoc查询等需求。

图 12-3: DRDS应用场景

DRDS for MySQL/XKV	DRDS for MySQL	DRDS for HiStore/ES	DRDS for HiStore
前台业务 MySQL - 围绕数据主体操作，比如购物、纳税、车辆管理等； - QPS高（1000以上）； - 延迟10ms以下。	后台业务 (CRM、进销存、BOSS系统等) MySQL 搜索引擎/列存/MPP - 根据各种条件组合检索数据，通常带分页，然后做业务处理，比如打标； - QPS各异（比如一个企业后台QPS较低，一个ISV面向很多商户QPS较高）； - SQL延迟亚秒级。	报表和分析 列存内存、MPP或DAG MapReduce - 对数据做聚合、计算、分析； - QPS通常不高（1-100个QPS），但是涉及实时分析，有100-5000左右的QPS需求； - 延迟响应从实时分析的秒级到离线分析的分钟、小时级。	

可参考的业务场景举例：

- 面向客户类的互联网应用，围绕人或者用户为主体展开业务（DRDS for MySQL）；
- 前台高并发低延迟类数据业务，比如银行、医院等柜台类业务，车联网车辆数据操作、轨迹跟踪、油耗曲线等（DRDS for MySQL）；
- 归档不再改变（包括历史数据）的数据存储和聚合分析，比如已经完结的订单、流水、日志、操作行为记录等（DRDS for HiStore）。

12.6 使用限制

本节介绍 DRDS 的使用限制。

项目	限制
分表大小	建议单个分表的数据记录数不超过500W。
分表数目	每个分库里的分表数目理论上是没有限制的，受限于 DRDS 机器本身的硬件资源。分表数目的选择需要依据对业务数据量的评估，具体请参考 DRDS 最佳实践 如何选择分片数 。

项目	限制
单个 RDS 实例的默认分库数目	8 个，不可更改。
分布式JOIN	支持大部分的 JOIN 的语义，但对于比较复杂的情况，DRDS 做了一些限制。例如大表之间的 JOIN，由于执行代价过高，速度过慢容易导致性能或者系统不可用等情况，因此请尽量避免。

12.7 基本概念

本节主要对 DRDS 涉及的专有名词及术语进行定义和解析。

DRDS

DRDS (Distributed Relational Database Service) 是阿里巴巴自主研发，高度兼容 MySQL 协议和语法的分布式关系型数据库服务中间件。

DRDS 服务节点 (DRDS Server)

DRDS Server 是 DRDS 核心组件，提供 SQL 的解析、优化、路由和结果归并。

DRDS 实例

DRDS 实例是由一组 DRDS Server 节点组成的分布式数据库服务集群。各服务节点无状态，同时处理 SQL 请求。

DRDS 实例规格

DRDS 实例规格是 DRDS 实例处理能力体现，按照 CPU 和 内存提供不同的规格的实例，规格越高处理能力越强。如 4Core8G 和 8Core16G，在标准的 DRDS 测试场景下，后者的处理能力是前者的两倍。

实例升降配

DRDS 可以通过改变实例规格来改变处理能力，提升实例规格称为升配，降低实例规格称为降配。

水平拆分

水平拆分是将一个单机数据库拆分为多个物理分库，将数据库中的表数据按照拆分规则，拆分为多个物理分表，存储到不同的数据库分库上的过程。

拆分规则

水平拆分过程中将逻辑数据库表拆分为多个物理分表规则称为拆分规则。

拆分键

水平拆分过程中，生成拆分规则的数据库字段称为拆分键。

分库

DRDS 水平拆分后，逻辑数据库数据存储在多个物理存储实例上，每个存储实例上的物理库称为分库。

分表

DRDS 水平拆分后，每一个分库上的物理数据表称为分表。

逻辑 SQL

由应用端发送到 DRDS 的 SQL 称为逻辑 SQL。

物理 SQL

由 DRDS 对逻辑 SQL 进行解析之后发送到 RDS 上执行的 SQL 称为物理 SQL。

透明读写分离

DRDS 的单个存储实例节点遇到访问瓶颈时，可通过增加只读实例来分担主实例的压力。DRDS 的读写分离功能不需要修改任何应用代码，称为透明读写分离。

非拆分模式

DRDS 支持不进行数据库水平拆分而仅通过 DRDS 提供的透明读写分离来扩展数据库的服务能力。这种模式称为非拆分模式。

平滑扩容

DRDS 可通过增加存储实例节点完成数据库的扩容。扩容不影响原有数据的正常访问，称为平滑扩容。

小表广播

DRDS 将一些数据量小且更新频度不高的数据表存储为单表模式，这些数据表称为小表。通过数据同步将小表复制到与之 JOIN 的分库上进而提升 JOIN 效率的解决方案称为“小表广播”或者“小表复制”。

全表扫描

数据库拆分模式下，如果 SQL 语句中没有指定拆分键，DRDS 将在所有分表上执行 SQL 并归并结果返回，这个过程称为全表扫描。为避免影响性能，用户应尽量避免全表扫描。

全局唯一数字序列 (DRDS Sequence)

DRDS 全局唯一数字序列 (64位数字，对应 MySQL 中 BIGINT 类型) 的主要目标是为了保证所定义唯一字段中的数据的全局唯一 (比如 PRIMARY KEY，UNIQUE KEY 等) 和有序递增。

DRDS 自定义注释 (DRDS Hint)

DRDS 提供的自定义注释，用于指定一些特殊行为，通过相关的语法影响 SQL 的执行方式，从而对 SQL 进行特殊的优化。

13 云盾

13.1 什么是云盾

专有云云盾是防护云上安全的一套专有云安全解决方案。

在云计算环境下，随着技术的发展，传统以检测技术为主的边界安全防护不能充分保障云上业务的安全。专有云云盾结合云计算平台强大的数据分析能力以及专业的安全运营团队，从网络层、应用层、主机层等多个层面为用户提供一体化的安全防护服务。

13.2 产品优势

作为云上安全先行者，阿里云云盾拥有多项权威认证，通过成熟的安全体系、领先的安全技术，帮助专有云用户全面保障专有云环境的安全。

云上安全先行者

阿里安全团队从2005年起护航阿里巴巴集团内部所有业务系统的信息安全，不断积累安全经验。自2011年首次推出云盾产品，全方位保障云上安全，成为云上安全先行者。

云盾保护全国超过40%的网站，每天防御全中国一半以上的大流量攻击，每天抵挡36亿次恶意攻击，过去一年中帮助用户修复超过613万个漏洞。

权威认证，安全可靠

依靠平台自身的安全特性以及云盾为云上客户提供的攻击防御特性，先后取得了国内外多项云安全认证。

图 13-1: 获得的安全认证



- 全球首家获得云安全国际认证金牌（CSA STAR Certification）的云服务供应商。
- 全国首家获得ISO27001信息安全管理体系国际认证的云安全服务供应商。
- 全国首个通过公安部等级保护测评（DJCP）的云计算系统。
- 电子政务云平台首批通过党政部门云服务网络安全审查（增强级）。
- 全国首家云等保试点示范平台。
- 金融云高分通过等保四级测评，成为全国首个四级云平台。

体系完整，技术领先

十年攻防，一朝成盾。在经历了为阿里巴巴集团自身业务十年来的安全护航后，阿里巴巴积累了大量的安全研究成果、安全数据、安全运营和安全管理方法，形成了一支专业的云安全专家团队。云盾是集合这些安全专家多年攻防经验开发出来的面向云计算平台安全最佳实现的成熟体系，可有效保护专有云用户云平台、云网络环境和云业务系统的安全。

与传统安全产品对比

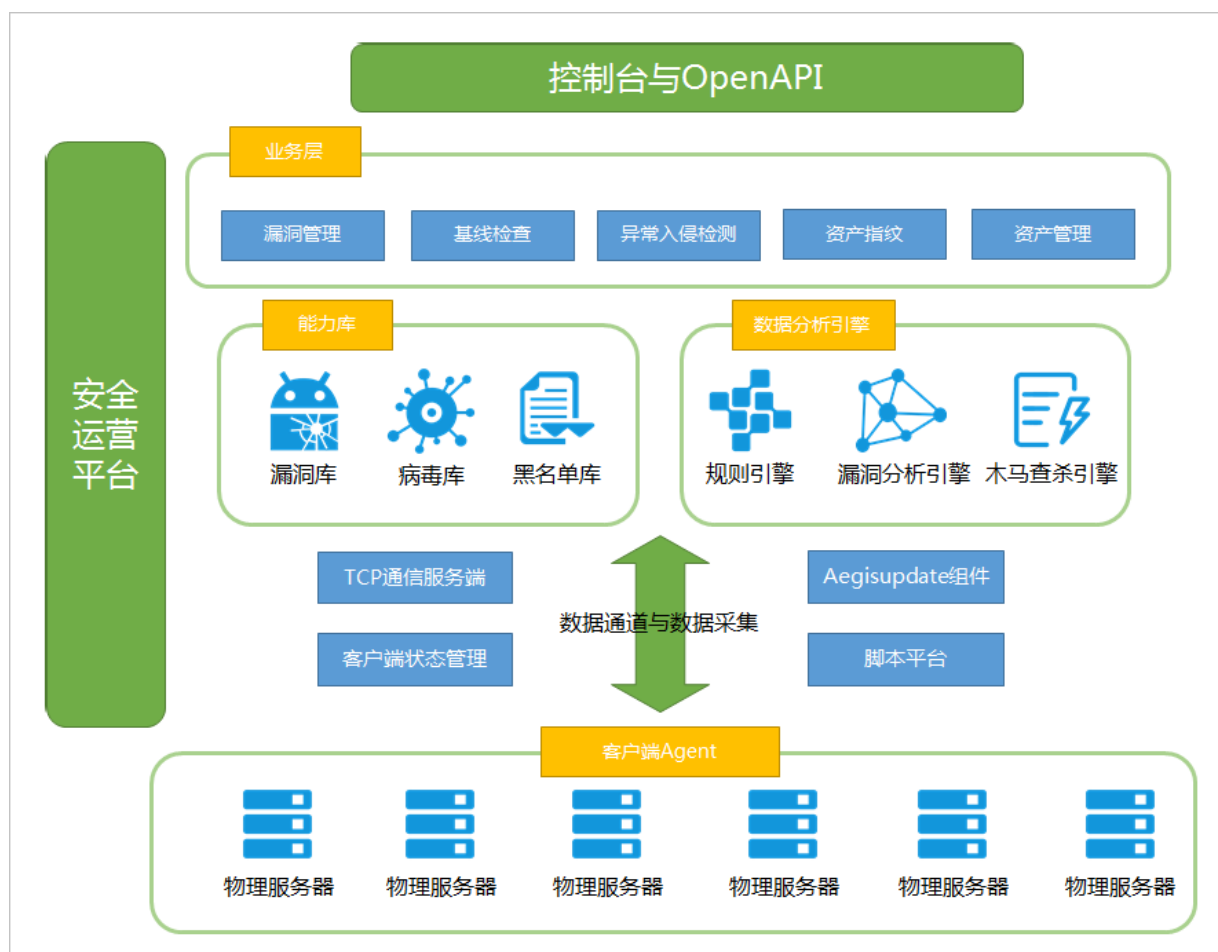
特点	传统安全产品	云盾
顶级互联网企业安全能力完整输出	传统安全厂商都有各自擅长的产品，但在其他产品上存在短板，无法形成整体安全防护体系。	阿里巴巴集团在与黑客攻击对抗过程中沉淀了大量的情报能力，及时发现流行的互联网攻击行为以及0 Day攻击手段，为用户提供完整的安全能力。
提前研判、预知风险爆发	传统安全厂商缺少完整的监控业务场景，无法准确研判风险的爆发。	对于重大漏洞、重大安全事件能够进行分析并及时响应，避免安全问题的爆发。
弹性扩容，与硬件解耦合	基于自身定制的硬件设备实现；软件化后的安全产品，也是基于虚拟化平台上的虚拟主机实现。	• 硬件解耦合：采用云架构设计，所有功能模块都基于通用的X86硬件平台，对硬件无依赖性。 • 弹性扩容：在性能不足时，无需改造网络结构，直接平滑地扩展硬件数量即可。
网络和主机的联动	通过设备的叠加来实现功能的完整覆盖。各设备之间不能有效的联动，一般只能通过管理	提供完整的网络安全、应用安全、主机安全的互联网防护能力，各防护组件互相联动形成

特点	传统安全产品	云盾
	平台将各个设备的日志、状态进行统一收集、展现。	整体防护体系，对已知的攻击行为达到最佳的防护效果。
兼容所有的IDC环境，和云平台解耦合	大部分的传统安全厂商仍然基于硬件盒子的形态提供安全产品。在SDN技术越来越普及的情况下，这样的形态无法与云环境兼容。	采用“网络出口检测 + 服务器操作系统联动”架构；采用数据分析方式发现安全威胁。通过这种架构及方式，避开了IDC内部复杂的网络结构，完全兼容所有的IDC环境。

13.3 产品架构

该章节介绍专有云云盾的产品架构。

产品结构如下所示。



安骑士采用Client-Server架构，在物理服务器上的客户端Agent与服务端通过TCP长连接进行通信，通过HTTP方式从服务端获取需要的脚本、规则、安装包等文件。基于客户端收集到的主机的

信息，通过日志监控、文件分析、特征扫描等手段，为专有云环境中的主机提供漏洞管理、基线检查、入侵检测、资产管理等安全防护措施。

13.4 功能特性

云盾不同于传统的软硬件安全产品，它采用纵深防御、多点联动的云安全架构，完全基于专有云的云计算环境研发，从网络层、应用层、主机层等多个层面为用户提供全面的、一体化的云安全防护能力。

主要功能

云盾所包含的详细功能如下表所示。

表 13-1: 云盾功能列表

功能项	功能说明
网站后门查杀	通过规则匹配对云服务器中存在的脚本后门进行精准查杀，并可手动对脚本后门进行隔离。
暴力破解攻击拦截	对黑客进行暴力破解的行为进行实时检测和拦截。
异常登录告警	通过分析和比对用户常用登录设置，对疑似的非常用登录行为进行告警。
漏洞扫描	对云服务器主机进行扫描，发现主机软件漏洞并提供漏洞修复方案。
漏洞修复	对云服务器主机中的应用和操作系统的高危漏洞提供一键修复，包括Web应用漏洞修复、系统文件修复等。
基线检查	对云服务器主机中的安全基线进行检查，包括账户安全检测、弱口令检查、以及配置项风险检测，以达到企业级服务器安全准入标准。
资产指纹	对云服务器主机的端口、账号、进程、应用软件等进行清点，全面了解主机资产的运行状态并有效进行回溯分析。
日志检索	将散落的云服务器主机的进程、网络、系统登录等日志集中管理，帮助在主机出现问题时一站式搜索相关日志，快速定位问题根源。

13.5 使用限制

无。

14 密钥管理服务KMS

14.1 什么是密钥管理服务

密钥管理服务KMS (Key Management Service) 是阿里云专有云提供的一款安全、易用的密钥托管服务。通过KMS，您可以轻松创建和管理密钥，使用密钥加密自己的数据。

KMS集成多款阿里云产品和服务，以帮助您保护云上数据。

KMS主要可以解决的问题如[KMS解决的问题](#)所示。

表 14-1: KMS解决的问题

角色	问题和要求	KMS如何解决
应用/网站开发者	- 我的程序需要使用密钥、证书用于加密或者签名，我希望密钥管理的功能安全且独立。 - 不论我的应用部署在哪里，都能安全的访问到密钥；我绝不接受把明文的密钥到处部署，这太危险了。	使用信封加密技术，主密钥存放在KMS服务中，只部署加密后的数据密钥，仅在需要使用时调用KMS服务解密数据密钥。
服务开发者	- 我不想承担客户密钥、数据的安全责任。 - 我希望用户自己管理他的密钥；在拥有授权的情况下，我会使用用户指定的密钥加密他的数据。这样，我就能专注于服务功能的开发。	基于信封加密技术以及KMS开放的API，服务能够集成KMS，使用用户指定的主密钥完成数据密钥的加解密，能够轻松的实现 明文不落盘 的要求，也不用为管理用户的密钥而犯愁。
首席安全官(CSO)	- 我希望公司的密钥管理能满足合规需求。 - 我需要确保密钥都被合理的授权，任何使用密钥的情况都必须被审计。	KMS服务接入RAM服务实现统一的授权管理。

14.2 产品优势

低成本

传统的密钥管理方案中，采购安全的密钥管理设备并建设安全的物理环境都需要很高的硬件成本，设计和执行安全的密钥管理规范还需要很高的软件成本。

使用密钥管理服务，您可以在云平台统一管理自己的密钥，节省硬件、软件的投入。

易用性

密钥管理服务使用统一、易用的 API 以及标准的 https 协议。

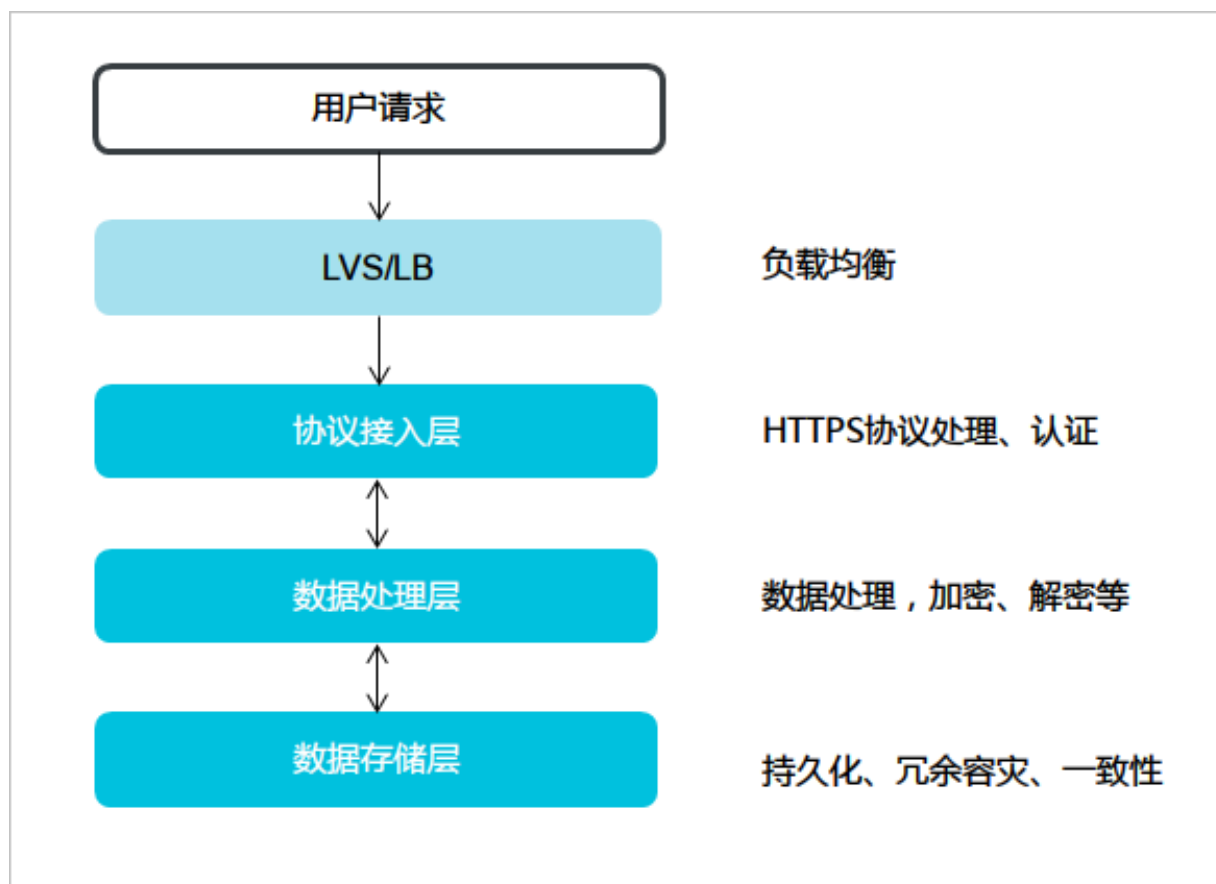
可靠性

密钥管理服务采用分布式系统与密码硬件设备结合，实现高可靠。

14.3 产品架构

KMS架构图如下。

图 14-1: KMS架构



- KMS的协议接入层负责接收用户向KMS发来的HTTPS请求，并且进行身份认证、权限识别。验证通过后，用户的请求会转发到数据处理层，并且接收数据处理层返回的结果再返回给用户；如果验证失败，则不处理数据直接返回错误信息。
- KMS的数据处理层负责请求数据的处理。KMS的数据处理主要是密码学相关的操作（如加密、解密等）。协议接入层与数据处理层之间采用基于TLS的RPC协议进行通信。数据处理层采用分布式的部署，各节点之间无状态，协议接入层的请求可以在任何一个数据处理层的节点中被正确处理。
- KMS的数据存储层存储KMS的核心根密钥数据，使用raft分布式一致性协议保证数据的一致性，并且使用基于TPM的加密方案对持久化的数据进行加密存储。

14.4 功能特性

KMS提供以下功能：

功能	描述
创建用户主密钥	在您使用KMS之前，至少需要创建一个用户主密钥（CMK）。主密钥可以直接用来加密少量的数据（少于4 k），但通常当做 产生数据密钥 API的输入用以生成可以加密数据的DataKey
创建数据密钥	使用指定的CMK创建一个数据密钥，用于KMS的信封加密场景。您可以用这个密钥进行本地数据的加密。
加密数据	使用指定的CMK加密数据。可以加密最多为4KB任意数据，如RSA密钥、数据库密码或其他敏感的客户信息。
解密数据	解密由KMS加密的数据。
查看密钥列表	获取您账户下当前区域的所有主密钥ID。
查看密钥详情	查看指定主密钥（CMK）的详细信息。包括创建的日期和时间、描述、全局唯一标识符、CMK的状态、用途、预计删除时间、创建者、CMK的密钥材料来源、密钥材料过期时间等。
启用密钥	将一个禁用状态的密钥恢复为启用状态。
禁用密钥	将一个密钥设置为禁用状态，被禁用的密钥不能用于加密，由该密钥加密的数据也不能被解密。
密钥预删除	为指定的密钥设置一个预删除时间，在到达预删除时间之后，密钥将会被彻底删除。
取消密钥预删除	取消密钥的预删除状态，使密钥恢复到启用状态。
API	提供KMS支持的HTTPS API操作。

功能	描述
SDK	提供主流语言的SDK。

14.5 应用场景

本文介绍KMS常见的两个使用场景：

- 使用KMS加密、解密数据。
- 使用信封加密在本地加密、解密数据。

表 14-2: 图例说明

图例	说明	图例	说明
	用户主密钥（CMK）		密文数据密钥
	明文证书		明文文件
	密文证书		密文文件
	明文数据密钥		
		-	-

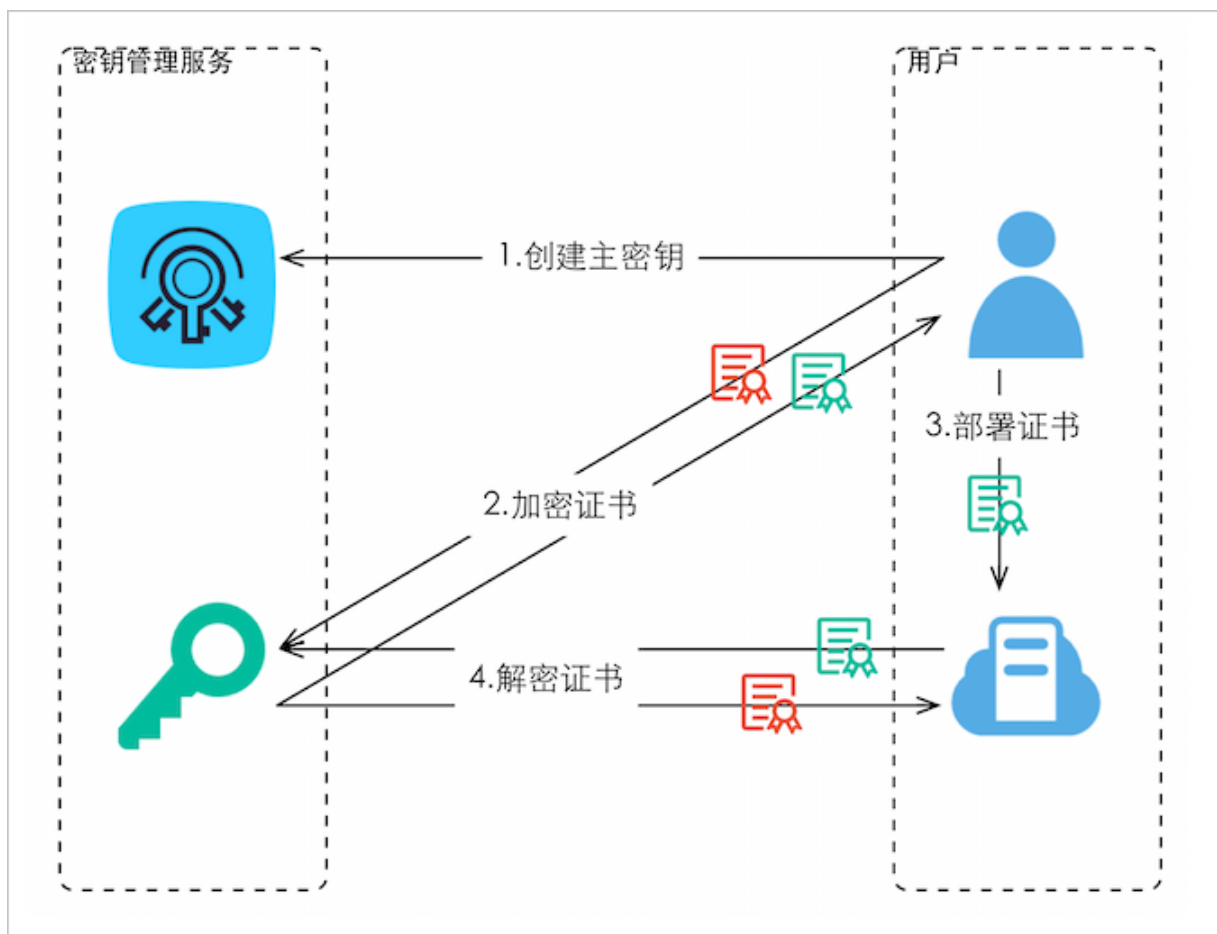
直接使用KMS加密、解密数据

用户可以直接调用KMS的API，使用指定的CMK来加密、解密数据。

此场景适用于少量（少于4KB）数据的加解密，用户的数据会通过安全信道传递到KMS服务端，对应的结果将在服务端完成加密、解密后通过安全信道返回给用户。

场景举例：保护服务器HTTPS证书，如[保护服务器HTTPS证书](#)所示。

图 14-2: 保护服务器HTTPS证书



流程如下：

1. 首先，用户需要创建一个主密钥。
2. 调用KMS服务的Encrypt接口，将明文证书加密为密文证书。
3. 用户在服务器上部署密文证书。
4. 当服务器启动需要使用证书时，调用KMS服务的Decrypt接口，将密文证书解密为明文证书。

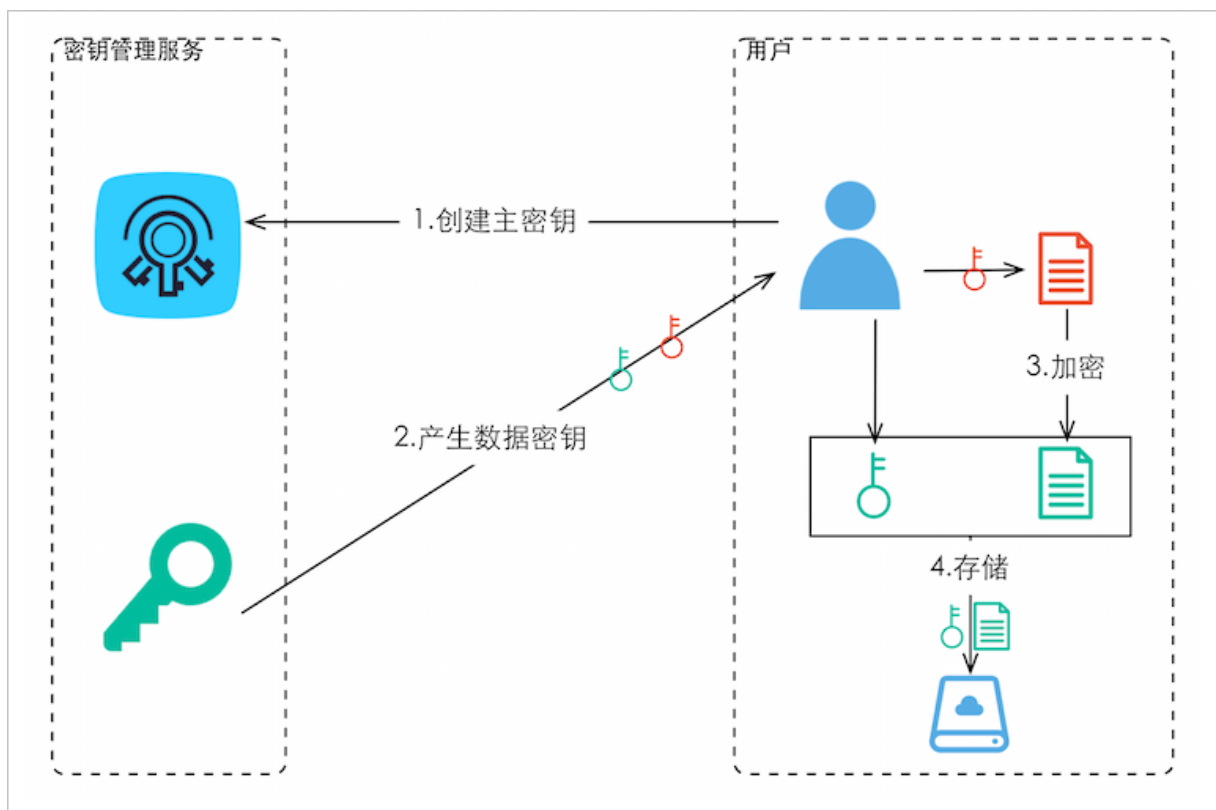
使用信封加密在本地加密、解密数据

用户可以直接调用KMS的API，使用指定的CMK来产生、解密数据密钥，自行使用数据密钥在本地加解密数据。

此场景适用于大量数据的加解密，用户的数据无需通过网络传输大量数据，可以低成本的实现大量数据的加解密。

场景举例：加密本地文件，如[加密本地文件](#)所示。

图 14-3: 加密本地文件

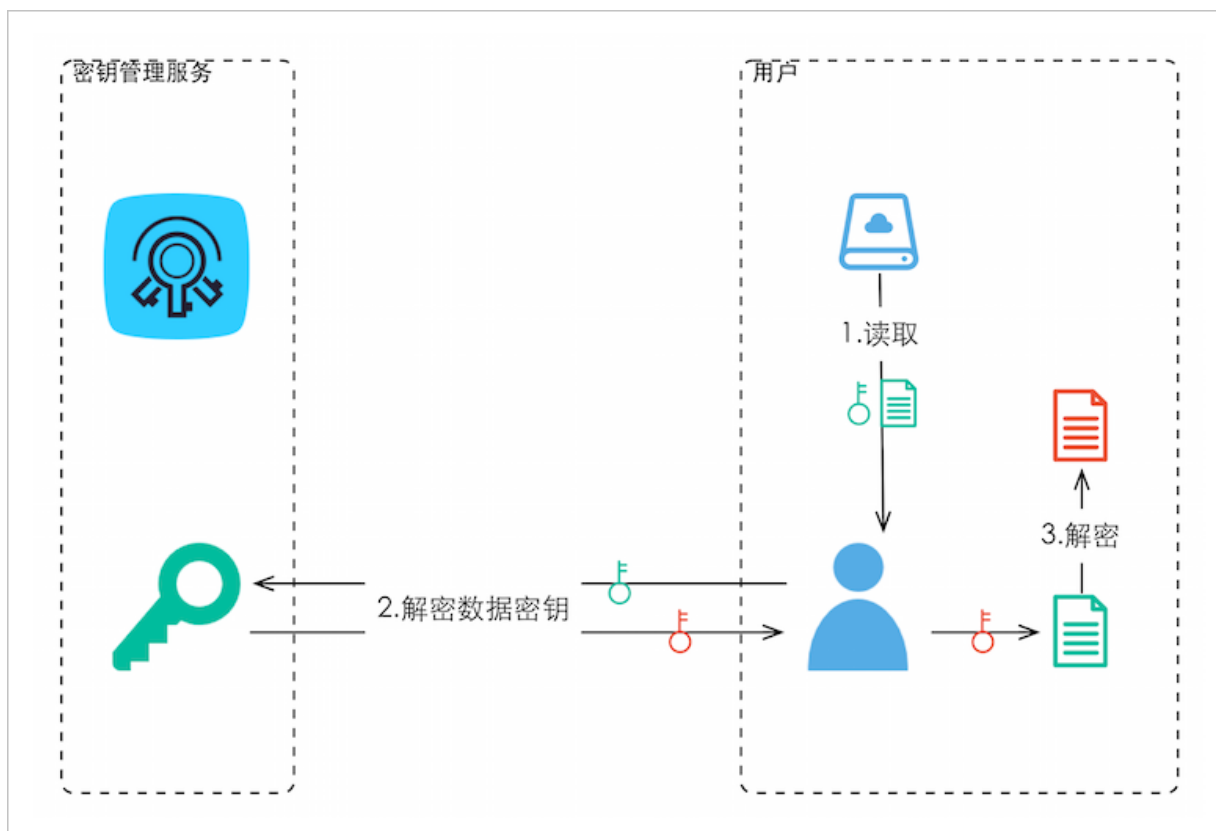


加密流程如下：

1. 首先，用户需要创建一个主密钥。
2. 调用KMS服务的GenerateDataKey接口，产生数据密钥。
这里用户能够得到一个明文的数据密钥和一个密文的数据密钥。
3. 用户使用明文的数据密钥，加密文件，产生密文文件。
4. 用户将密文数据密钥和密文文件一同存储到持久化存储设备或服务中。

上图的解密过程如[解密过程](#)所示。

图 14-4: 解密过程



解密流程如下：

1. 用户从持久化存储设备或服务中读取密文数据密钥和密文文件。
2. 调用KMS服务的Decrypt接口，解密数据密钥，取得明文数据密钥。
3. 使用明文数据密钥解密文件。

注意事项

1. 验证阿里云服务端的HTTPS证书，防止钓鱼者窃取您的信息。
2. 建议给不同密钥的使用者分别赋予不同的权限。

14.6 使用限制

用户主密钥(Customer Master Key)数量限制：一个部门最多创建200个用户主密钥。

14.7 基本概念

信封加密

信封加密 (envelope encryption)，为要加密的数据产生**一次一密**的对称密钥，使用特定的主密钥加密该对称密钥，使这个对称密钥处于一种**被密封的信封保护**的状态。在传输、存储等非安全的通信过程中直接传递**被密封保护的密钥**，当且仅当要使用该对称密钥时，打开信封取出密钥。

用户主密钥

用户主密钥 (Customer Master Key，简称CMK)，用户在密钥管理服务中创建的主密钥，主要用于加密保护数据密钥，产生信封。也可直接用于加密少量的数据。

信封数据密钥/数据密钥

信封数据密钥/数据密钥 (EnvelopedDataKey / DataKey，简称EDK/DK)，DK为加密数据使用的明文数据密钥，EDK为通过信封加密技术保密后的密文数据密钥。

15 ZStack

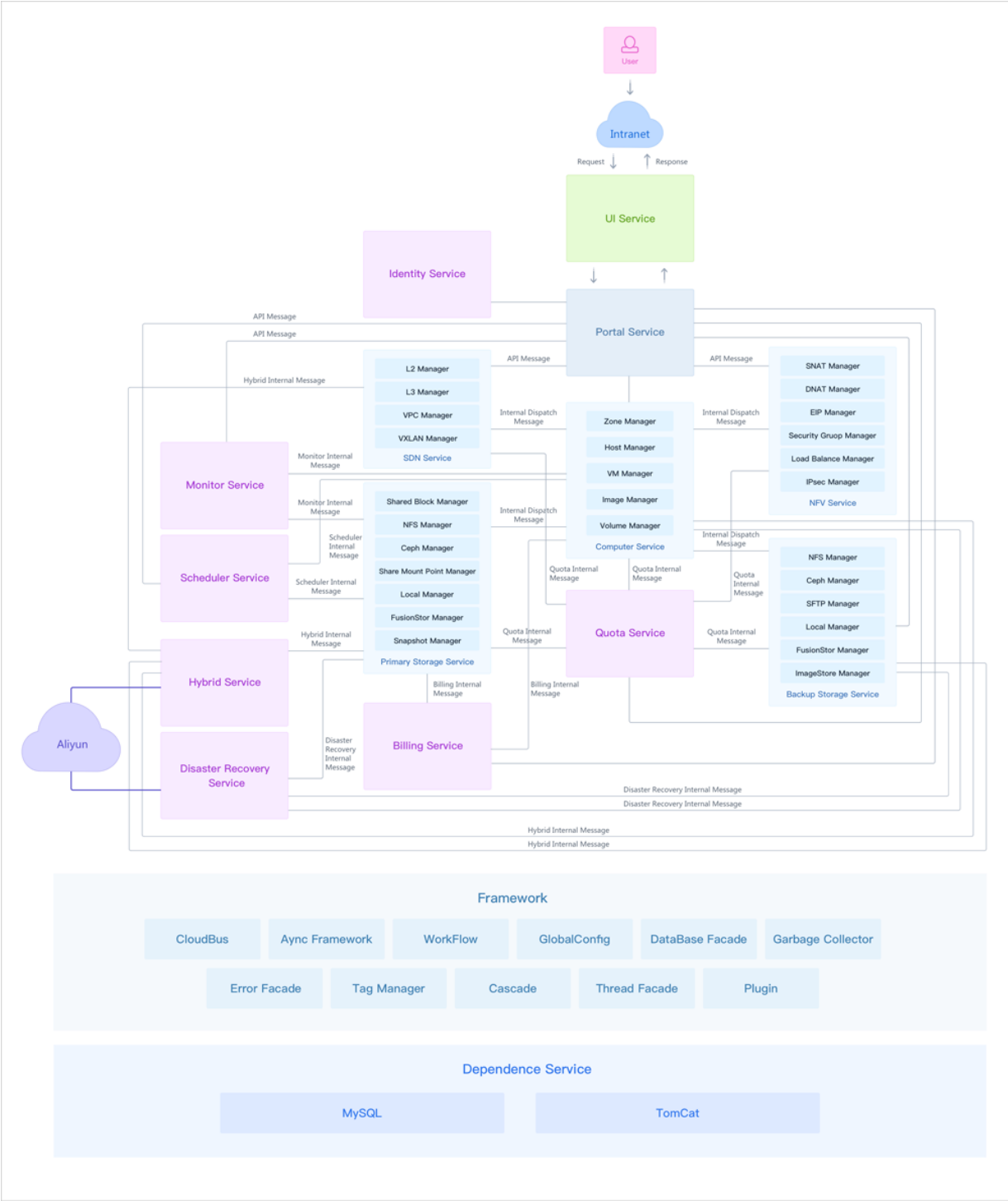
15.1 产品概述

ZStack是下一代开源的云计算IaaS（基础架构即服务）软件。它主要面向未来的智能数据中心，通过提供灵活完善的APIs来管理包括计算、存储和网络在内的数据中心资源。用户可以利用ZStack快速构建自己的智能云数据中心，也可以在稳定的ZStack之上搭建灵活的云应用场景，例如VDI（虚拟桌面基础架构）、PaaS（平台即服务）、SaaS（软件即服务）等。

通过对ZStack云引擎的深度定制，阿里云和ZStack联合推出了ZStack for Alibaba Cloud，其结合了ZStack专有云的简单、健壮、弹性、智能以及阿里云公共云的领先、安全、稳定等特点，以**云+端**的形式提供了一套无缝集成的混合云管理方案。

系统架构如[图 15-1: 系统架构示意图](#)所示：

图 15-1: 系统架构示意图



15.2 产品功能

15.2.1 专有云功能

ZStack作为产品级私有云平台，提供了对用户数据中心的计算、存储、网络等资源的管理和调度。用户使用ZStack可以快速配置私有云环境，并快速创建云主机、分配云盘和自动配置云主机网络。

ZStack企业版功能列表：

类别	特性	ZStack企业版
区域	管理多个区域	用户可以根据实际情况创建并管理多个区域，一般情况下可将一个物理数据中心归为一个Zone来管理；用户根据不同的业务需求，每个Zone内建立自己独立的集群、主存储、网络等资源
集群	存储架构	集群内使用同构存储服务，存储服务挂载到集群，提供云主机高可用
	物理机	集群内管理物理机，支持实时查看物理机全部CPU使用率、物理机全部内存使用百分比、物理机全部网卡出入速度和物理机全部磁盘读/写IOPS
	云主机	集群内管理云主机，支持实时查看云主机全部CPU使用率、云主机全部内存已用百分比、云主机全部网卡出入速度和云主机全部磁盘读/写IOPS
	集群功能	提供高可用特性，支持按照物理机CPU架构定义集群属性
	网络服务	支持VLAN、VXLAN网络加载到集群并统一管理、提供网络自助服务（IP池管理和弹性网络）、支持集群指定迁移网络、支持定义集群的CPU模式
物理机	虚拟化	支持KVM虚拟化技术，支持VMware虚拟化
	c74 ISO	- 支持使用最新英特尔® 至强® 可扩展处理器，例如支持部署在DELL EMC R740 14代服务器上，提升平台稳定性 - 支持灾备模块，进一步提升平台稳定性 - 初装用户推荐安装c74 ISO
	资源设定超分	支持CPU、内存和存储空间设定超分比例，适应云环境资源使用
	嵌套虚拟化	支持KVM/ESXi嵌套虚拟化，云主机内部开启CPU硬件虚拟化功能
	实时监控	采集物理机的CPU、内存、存储和网络运行数据，提供图形可视化
	停用与启用	对物理机设定可用属性，以便停止在该物理机上创建云主机

类别	特性	ZStack企业版
	维护模式	对物理机设定维护状态，设定维护模式后，物理机上的云主机将会迁移（共享存储）
	GPU透传	支持物理机GPU设备透传，让云主机拥有高性能计算和图形处理能力
	USB透传	支持USB透传，满足多种USB应用场景
	操作日志	展示物理机执行任务的事件审计
	导出CSV文件	支持物理机列表导出为CSV表格，方便统计分析处理
云主机	批量操作	批量管理云主机
	创建云主机	提供多种策略创建云主机，高效利用资源
	云主机生命周期	支持创建、停止、启动、重启、关闭电源、删除、暂停、恢复等基本生命周期控制
	根云盘在线扩容	支持云主机根云盘在线扩大容量，方便修改云主机配置
	数据云盘在线扩容	支持云主机数据云盘在线扩大容量，即时生效
	云主机控制台	用户可通过终端方式访问云主机，而不依赖云主机远程工具，支持控制台设置密码
	云主机快照	- 在云主机运行过程中进行快照 - 在线快照（支持ImageStore/Ceph类型的镜像服务器） - 关机快照（支持ImageStore/Sftp/Ceph类型的镜像服务器）
	云盘在线快照	在使用云盘的过程中进行快照
	CPU绑定	将云主机的逻辑CPU与计算节点的物理CPU绑定
	云主机在线修改密码	支持Windows/Linux的云主机在线修改密码
	云主机在线创建镜像	运行中的云主机在线创建镜像
	云主机QGA开关	灵活控制qemu guest agent的状态
	云主机RDP模式开关	针对VDI用户界面，启用后默认以RDP模式打开控制台
	云主机显卡切换	支持选择云主机显卡类型：qxl、cirrus、vga
	云主机显卡透传	支持英伟达和AMD GPU设备透传给云主机
	User Data导入	支持创建云主机时导入User Data
	云主机克隆（不带数据云盘）	基于云主机快速克隆若干个云主机

类别	特性	ZStack企业版
		- 在线克隆（支持ImageStore/Ceph类型的镜像服务器） - 关机克隆（支持ImageStore/Ceph类型的镜像服务器）
	整机克隆（带数据云盘）	- 同时克隆云主机的根云盘和数据云盘内容 - 仅支持ImageStore类型的镜像服务器 - LocalStorage/NFS/SMP类型的主存储，支持在线/暂停/关机克隆 - Ceph类型的主存储，支持在线/暂停/关机克隆 - Shared Block类型的主存储，支持暂停/关机克隆 - 挂载共享云盘的云主机不支持整机克隆
	更换系统盘	支持云主机关机状态下修改操作系统
	重置云主机	支持恢复云主机状态为模板初始状态
	根云盘扩容	支持在线/关机状态下的云主机根云盘扩容，方便修改云主机配置
	基于ISO部署	- 基于ISO系统光盘部署云主机，引导安装系统 - 允许一个云主机加载多个ISO，提升业务部署效率
	基于模板部署	基于系统模板创建云主机
	制作镜像模板	基于当前某个云主机制作模板
	创建镜像	- 云主机运行中在线创建镜像 - 在线创建镜像（支持ImageStore/Ceph类型的镜像服务器） - 关机创建镜像（支持ImageStore/Sftp/Ceph类型的镜像服务器）
	自定义MAC地址	- 支持创建云主机时指定MAC地址 - 支持云主机修改MAC地址
	云主机启动顺序	调整云主机的启动顺序，用于切换ISO引导
	动态加载、卸载云盘	云主机可动态加载和卸载云盘，支持优化驱动模型，支持SCSI WWN号唯一识别
	动态加载、卸载网卡	云主机可动态加载和卸载网卡，支持设置默认网卡
	加载GPU卡	支持创建云主机时加载GPU设备
	共享云盘	支持Ceph存储或Shared Block主存储下多云主机共享使用同一数据云盘

类别	特性	ZStack企业版
	实时性能监控	采集云主机的CPU、内存、存储和网络运行数据，提供图形可视化
	高可用特性	物理机故障，云主机自动重启，支持UI上展示恢复过程
	在线修改云主机CPU/内存	支持在线修改云主机配置，不用重启VM
	实时更新云盘和网络QoS	提供云盘和网络的限速能力，避免单个云主机占用过量资源
	SSH密钥注入	支持Linux和BSD操作系统SSH密钥注入，支持创建和删除密钥
	自定义计算规格	支持自定义计算规格，满足各种应用资源消耗特性
	自定义标签	支持自定义标签，满足查询和编写定时任务
	资源删除保护	云资源删除后，将移入回收站，提供恢复和确认销毁
	冷迁移	支持本地主存储类型上的云主机进行关机状态迁移
	在线迁移	支持所有主存储类型上的云主机进行在线迁移
	存储迁移	- 多NFS主存储之间的云主机支持跨存储设备冷迁移 - 多Ceph主存储之间的云主机支持跨存储设备冷迁移 - 多Shared Block主存储之间的云主机支持跨存储设备冷迁移，且支持同时迁移加载的云盘
	操作日志	展示云主机操作过程的事件审计
	Windows系统性能优化	提供Windows云主机性能优化加速
	USB重定向	支持将VDI客户端USB设备重定向至云主机
	导出CSV文件	支持云主机列表导出为CSV表格，方便统计分析处理
云盘	云盘管理	支持云盘的创建、启用、停用、加载、卸载、迁移、创建快照、创建镜像、扩容、更改所有者、存储迁移、删除
云盘规格	云盘规格管理	支持云盘规格的创建、启用、停用、全局共享、全局召回、云盘规格QoS、删除
计算规格	计算规格管理	- 支持计算规格的创建、启用、停用、磁盘QoS、网络QoS、全局共享、全局召回、删除 - 支持选择物理机分配策略 - 当物理机分配策略为CPU使用率最低/内存使用率最低，支持选择强制、非强制策略模式

类别	特性	ZStack企业版
镜像管理	系统模板	支持系统模板，支持QCOW2和RAW格式，自动匹配镜像类型
	ISO镜像	支持ISO镜像，支持从ISO镜像引导云主机
	系统镜像上传	支持URL上传和本地浏览器上传
	云盘镜像上传	支持URL上传和本地浏览器上传
	镜像迁移	支持Ceph主存储上的镜像跨存储设备迁移、支持NFS主存储上的镜像跨存储设备迁移
镜像仓库	镜像存放	存放镜像数据，包括ISO和系统模板
	镜像导出	支持镜像导出下载链接
	获取已有镜像	添加ImageStore类型的镜像服务器时，可获取该镜像服务器中URL路径下的已有镜像文件
	镜像同步	支持镜像仓库间的镜像互传，可以跨区域使用
	镜像仓库清理	支持清理镜像服务器中已被彻底删除的无效数据，释放存储空间
	标准系统镜像	支持标准的系统，支持Windows、红帽、Ubuntu和其他开源Linux系统
	预设运行镜像	支持众多的软件运行环境，支持Windows IIS和Dot Net Framework运行环境，支持Linux Tomcat、JAVA、Apache Web、Jboss、PHP、Node JS、Golang、Python等语言和运行环境，支持数据库Oracle、MySQL、Postgres、Mongodb、Influxdb、Cassandra和Redis等数据库服务；支持广泛的应用中间件
	预设应用镜像	支持众多的应用系统，论坛BBS、社交SNS、博客Blog、微博的常用应用系统；支持phpmyadmin等运维管理应用；支持厂商提供的应用镜像
	自定义镜像	支持管理员根据标准系统镜像和预设运行镜像，定义满足自身业务系统运行环境的镜像，以增量方式保存镜像内容，并实现智能去重功能
	存储支持	与本地存储、NFS、SMP、Ceph、Shared Block、阿里云NAS类型的主存储无缝支持
存储管理	本地存储	- 支持云盘存放到物理机本地 - 支持实时查看主存储已用容量百分比趋势图
	NFS存储	支持云盘存放到NFS协议存储，物理机共享访问

类别	特性	ZStack企业版
		- 共享文件系统管理节点高可用方案 - 支持指定存储网络，支持存储网络和管理网络分离，增强云主机高可用 - 支持实时查看主存储已用容量百分比趋势图
	共享挂载存储	- 支持云盘存放到POSIX兼容的共享存储，支持iSCSI/FC存储 - 共享文件系统管理节点高可用方案 - 支持指定存储网络，支持存储网络和管理网络分离，增强云主机高可用 - 支持实时查看主存储已用容量百分比趋势图
	Shared Block存储	- 支持添加iSCSI/FC协议存储，物理机共享访问 - 支持添加iSCSI服务器，支持自动在线扫描并发现磁盘和自动配置iSCSI发起等操作 - 支持共享云盘 - 支持添加多个LUN - 添加Shared Block主存储时，支持显示共享块设备候选列表 - 支持实时查看主存储已用容量百分比趋势图
	NAS存储	- 支持云盘存放到POSIX兼容的分布式网络文件存储 - 支持实时查看主存储已用容量百分比趋势图
	Ceph存储	- 支持共享云盘 - 超融合管理节点高可用方案 - 支持指定不同性能的磁盘卷创建云盘 - 支持云盘存放到Ceph分布式存储 - 支持数据冷迁移 - 支持指定存储网络，支持存储网络和管理网络分离，增强云主机高可用 - 支持创建Ceph pool，以pool计算容量并设置显示名，并设置显示名 - 支持实时查看主存储已用容量百分比趋势图
	多主存储支持	支持同一集群挂载多个主存储，包括：多个本地存储、多个NFS存储、多个Shared Block存储、一个本地存储和一个NFS/SMP/Shared Block存储

类别	特性	ZStack企业版
网络管理	VLAN二层隔离	支持VLAN 802.1q作为网络隔离手段
	VXLAN网络	支持VXLAN网络，有效解决云数据中心逻辑网段不足、上层交换机MAC地址溢出等问题、支持云主机的跨地域迁移
	分布式扁平网络	支持云主机直接使用真实网络IP资源
	分布式弹性网络	支持云主机使用虚拟网络地址，与真实网络映射
	分布式DHCP服务	支持云主机自动获取分配的IP地址
	网络地址空间预留	支持预留网络地址空间，以便与物理网络混合使用
	动态和静态分配IP	支持动态分配IP地址，支持指定使用某个IP地址
	多级网络管理	支持云主机接入多个网络，构建复杂场景的业务
	虚拟IP的QoS设置	支持对虚拟IP做QoS限制，对网络服务的高效分配管理
	MTU	自定义限制网络传输数据包的大小
	VPC路由器	支持创建VPC路由器的全生命周期管理，包括：创建、删除、修改、VPC网络的加载/卸载，东西向流量的设置、云路由网络的所有网络服务，集中在VPC路由器中配置DNS
	VPC网络	- 支持创建VPC网络、添加网络段、加载/卸载VPC路由器、删除 - VPC网络支持以下网络服务：安全组、虚拟IP、弹性IP、端口转发、负载均衡 - 负载均衡支持TCP/HTTP/HTTPS/UDP协议
	公有网络	- 支持创建云主机 - 支持为网络服务提供虚拟IP
	系统网络	可作为管理网络、存储网络、迁移网络等使用
网络管理	云路由网络	- 云路由支持以下网络服务：安全组、虚拟IP、弹性IP、端口转发、负载均衡 - 负载均衡支持TCP/HTTP/HTTPS/UDP协议 - 支持基于云路由的IPsec隧道服务 - 支持多个弹性IP绑定同一个云主机网卡 - 支持一个云路由器接多个公有网络 - 支持配置静态路由表 - 支持分布式DHCP提升服务性能
	网络拓扑	全局网络拓扑查看，支持高亮显示

类别	特性	ZStack企业版
		自定义选择资源展示拓扑图
定时任务	定时对象	支持云主机、云盘的定时操作
	定时操作	- 可对云主机关闭/重启，云盘快照等设置定时操作 - 创建云主机/云盘快照定时任务时，如果选择的所有云主机/云盘都使用Ceph主存储，支持设置保留快照数量
资源编排	资源栈	- 支持在线编辑方式和使用模板方式创建资源栈 - 支持预览/校验模板内容，支持云主机插入userdata - 支持删除资源栈和级联删除资源栈中所有资源
	自定义模板	支持通过文本编辑器方式和本地上传方式创建资源栈模板，并支持创建、查看、修改、删除、预览操作
	示例模板	云平台默认提供的资源栈模板示例，作为参考模板
	可视化资源编排	- 支持可视化拖拽资源方式创建资源栈模板 - 支持预览模板、生成资源栈、另存为资源栈模板 - 支持撤销、恢复、删除、清空画布操作
安全管理	三层安全策略	支持基于TCP/UDP端口的安全策略
	安全组统一管理	支持安全组统一管理云主机安全策略，实现组内互通，组间策略
性能TOP5和性能分析	性能TOP5	支持物理机、云主机、路由器、虚拟IP、三层网络等多种资源排序，并可自定义不同时间段查看
	云主机性能统计	支持自定义时间段查看，指定资源范围，对云主机CPU使用率、内存使用率、磁盘读速度、磁盘写速度、网卡入速度、网卡出速度、网卡入包率、网卡出包率、网卡入错误速率、网卡出错误速率进行过滤分析排序
	路由器性能分析	支持自定义时间段查看，指定资源范围，对路由器CPU使用率、内存使用率、磁盘读速度、磁盘写速度、网卡入速度、网卡出速度、网卡入包率、网卡出包率、网卡入错误速率、网卡出错误速率进行过滤分析排序
	物理机性能统计	支持自定义时间段查看，指定资源范围，对物理机CPU使用率、内存使用率、磁盘读速度、磁盘写速度、磁盘读IOPS、磁盘写IOPS、磁盘已用量百分比、网卡入速度、网卡出速度、网卡入包率、网卡出包率、网卡入错误速率、网卡出错误速率进行过滤分析排序

类别	特性	ZStack企业版
	三层网络性能分析	支持自定义时间段查看，指定资源范围，对三层网络已用IP数、已用IP百分比、可用IP数、可用IP百分比行过滤分析排序
	虚拟IP性能分析	支持自定义时间段查看，指定资源范围，对虚拟IP的下行网络流量、下行网络入包速率、上行网络流量、上行网络入包速率进行过滤分析排序
	镜像服务器性能分析	支持自定义时间段查看，指定资源范围，对镜像服务器的可用容量百分比进行过滤分析排序
ZWatch	物理机监控	对物理机运行实时监控，显示CPU、内存、磁盘和网络时序监控图
	云主机监控	对云主机运行实时监控，显示CPU、内存、磁盘和网络时序监控图
	监控	- 支持对系统时序数据进行监控，例如云主机内存使用率、物理机CPU使用率等 - 支持对系统事件进行监控，例如云主机状态变化事件、物理机失联事件等
	报警	对时序性数据和事件设置报警器，并通过SNS通知系统接收报警信息，支持邮件/钉钉/HTTP 应用方式接收报警信息
	多接收端	支持邮件/钉钉/HTTP应用等多种接收端
审计	资源审计	- 支持ZStack所有资源的审计查询，用户能对该资源的所有操作行为审计，有效保障用户在云环境下核心数据的安全 - 支持查看调用API名称、消耗时间、任务结果、操作员，任务创建/完成时间，以及API行为的消息详情，且支持CSV格式导出
操作日志	操作日志	支持查看操作描述、任务结果、操作员、登录IP、任务创建/完成时间，以及操作返回的消息详情，实现更细粒度管理，且支持CSV格式导出
账户管理	账户和用户管理	账户管理功能，分为账户和用户，其中账户是资源计量团体，用户可定义操作权限
	AD/LDAP账户	- 支持添加AD/LDAP账户，并绑定普通账户 - 支持自定义清除规则

类别	特性	ZStack企业版
	账户云资源配额	支持自定义分配账户最大可用资源，包括云主机运行数量、CPU、内存、云盘数量、云盘总容量、镜像数量、镜像总容量、弹性IP数量等
	用户组权限分配	支持用户组权限分配，统一编排用户权限
	用户操作权限分配	支持对用户进行权限分配
	云主机更改所有者	支持变更云主机所有者，指定云主机所属账户
	云盘更改所有者	支持变更云盘所有者，指定云盘所属账户
	计算规格指定分配	支持计算规格共享特性，可指定账户是否可使用
	镜像资源指定分配	支持镜像资源共享特性，可指定账户是否可使用
	云盘规格指定分配	支持云盘规格共享特性，可指定账户是否可使用
	网络资源指定分配	支持网络资源共享特性，可指定账户是否可使用
	全局配置	管理员可以直接在UI上对很多特性进行全局配置 • 所有的全局配置都有一个默认值 • 更新全局配置并不需要重启管理节点
	修改admin账户密码	忘记admin账户的登录密码，可以使用 <code>zstack-ctl reset_password</code> 还原默认值
计费	自定义计费单价	支持自定义CPU、GPU、内存、系统云盘和数据云盘的计费单价，其计费单价支持秒、分、小时和天；支持删除某时段的计费设置
	基于账户计费	基于账户进行计费，统计账户各项目消费情况
	灵活计费单价	动态可调的计费单价，满足周期性促销需求
访问	TUI	支持常用运维操作，定制化OS界面
	图形界面	• 支持以HTTP/HTTPS方式访问图形界面的云管理平台，账户（用户名密码方式或AD/LDAP方式）和用户支持图形界面登录访问 • 支持动态验证码验证。多次登陆（6次）失败触发验证码验证，防止恶意登陆 • 支持双因子认证。额外增加安全码身份验证，进一步增强账号安全
	命令行	支持通过命令行方式访问云管理平台，命令行支持全功能访问，账户和用户支持命令行登录访问

类别	特性	ZStack企业版
	API接口	支持全功能的API交付，API支持消息总线访问和HTTP接口访问
操作助手	智能提示	对ZStack的核心操作给出智能的环境检查和操作指导
亲和组	反亲和组	目前提供针对云主机与物理机的两种亲和组策略：反亲和组(非强制)、反亲和组(强制)，从而合理调度平台资源
UI强化	自定义产品信息	对UI上的产品Logo和产品名称等进行自定义
	首页大屏	华丽大屏展示平台整体情况
	加密访问	支持HTTPS安全访问登录平台
	过程展示	增加多个场景进度条
VDI	解决方案	- 通过定制客户端，支持SPICE，RDP，VNC等协议，并进行了优化 - 支持指定VDI网络 - 支持USB重定向，兼容多种USB设备 - 支持设置独立VDI网络 - 支持多屏显示 - 支持麦克风 - 支持SPICE流量优化
UI导航	快速入口	增加快速进入产品与服务的入口，并支持高亮标注
UI信息导出	列表信息CSV导出	导出云主机和物理机主列表的信息，离线管理便于图表编辑
应用中心	应用中心	支持添加包括存储、数据库、安全、IaaS、PaaS、SaaS类型在内的应用插件
License	云平台许可证	- 支持本地浏览器上传License - 支持License到期提醒 - 企业单机无限期试用版支持所有功能
	模块许可证	- 模块许可证 (Plus License) 为用户提供附加功能 - 依赖于平台许可证使用 - 已包括：企业管理模块、VMware管理模块、裸金属管理模块、灾备服务模块 - 支持本地浏览器上传License - License到期提醒

类别	特性	ZStack企业版
管理节点	管理节点高可用（基于超融合方案）	- 支持基于Ceph的超融合场景 - 支持基于NFS、SMP的共享文件系统场景 - 支持多网络灵活配置
	双物理机管理节点高可用	支持双物理机管理节点高可用。采用主备方案，某个管理节点故障后能迅速切换到另一个管理节点，保证业务正常运行

ZStack企业管理模块功能列表：

类别	特性	企业管理模块
组织架构	用户	- 用户是企业管理中的最基本单位 - admin/平台管理员可创建用户，并基于用户建立相应的组织架构 - 支持添加用户、删除用户、修改用户名、修改密码、修改个人信息、加入部门、从部门移除、加入项目、从项目移除 - 用户的个人信息包括姓名、手机号码、邮箱地址和编号
	组织	- 组织是企业管理中组织架构的基本单位 - 组织以组织架构树的方式呈现，分为顶级部门和部门，顶级部门是组织的一级部门，其下可添加多级部门，支持创建多个顶级部门 - 支持添加组织、删除组织、更改上级部门、更改部门负责人、创建子部门、删除子部门、添加用户、移除用户
项目管理	项目	- 用于表示在特定时间、资源、预算下指定相关人员完成特定目标的任务 - 企业管理以项目为导向进行资源规划，可为一个具体项目建立独立的资源池 - 支持创建项目、删除项目、启用项目、停用项目、更换项目负责人、生成项目模板、添加成员、移除成员、停用项目资源、恢复过期项目
	项目模板	- 用于标识各个资源配额的模板 - 在创建项目时，可直接使用模板定义的配额来快速创建项目 - 支持创建项目模板、删除项目模板

类别	特性	企业管理模块
	成员	- 成员作为项目的基本组成人员，一般由admin/平台管理员/项目负责人/项目管理员添加进入项目 - 项目成员的权限可由admin/平台管理员/项目负责人/项目管理员进行相应控制
	成员组	- 项目负责人/项目管理员可在项目中创建成员组，对成员进行分组管理 - 可以成员组为单位进行权限控制
工单管理	工单申请	- 项目成员可对云平台资源提出工单申请 - 支持项目成员创建、撤回、重新打开以及删除工单
	工单审批	- 支持admin创建工单流程，并按这种自定义流程审批工单 - 支持审批人通过、驳回工单，审批通过后将进入下一审批环节，最终由admin通过并部署、驳回工单 - admin可进行一键审批，审批通过后会自动部署，资源可自动部署成功并分发到项目中
	自定义流程审批	- admin可为不同项目设置各自的工单审批流程。支持将项目成员/项目管理员/项目负责人添加到审批环节 - 支持启用、停用、修改工单流程和删除操作
独立区域管理	平台管理员	- 平台管理员主要是带有区域属性的管理员 - admin可划分不同区域给不同平台管理员来管控不同区域的数据中心 - 支持创建/删除平台管理员、修改密码、添加区域和移除区域
	资源隔离	- 在对区域进行资源隔离的基础上，可对每个区域指定相应的区域管理员，实现各地机房的独立管理 - 同时admin可对所有区域进行巡查和管理

ZStack裸金属管理模块功能列表：

类别	特性	裸金属管理模块
裸金属管理	裸金属集群	- 通过创建裸金属集群，管理物理裸机 - 支持为裸金属集群加载二层网络
	部署服务器	通过部署服务器自动化对新上线裸金属设备进行系统安装部署

类别	特性	裸金属管理模块
	裸金属设备	- 通过IPMI网络批量部署裸金属设备 - 支持对裸机进行远程电源管理 - 支持vnc无人值守模式
	裸金属主机	- 支持使用ISO类型镜像为裸金属设备安装Linux操作系统 - 支持为裸金属主机添加网络配置

ZStack灾备服务模块功能列表：

类别	特性	灾备服务模块
灾备服务	备份	- 支持为云主机、云盘和数据库创建备份任务；支持用户按周/天/小时为任务设置备份策略 - 支持将数据备份在本地备份服务器并到远端服务器 - 支持使用已有镜像仓库或新建服务器作为本地备份服务器，主备服务器可无缝切换 - 支持将云主机、云盘和数据库的本地备份数据同步到远端数据库；远端备份服务器部署在异地机房或者阿里云，满足双重备份要求 - 支持云主机整机备份 - 新增支持使用Ceph主存储的资源进行备份 - 清理本地备份服务器/远端备份服务器中已被彻底删除的无效备份数据，释放存储空间。
	还原	- 从本地备份数据还原资源时，支持新建资源以及覆盖原始资源 - 支持云主机的整机还原 - 支持通过数据库的备份数据进行环境的还原，支持将备份数据导出进行手动还原 - 存在备份数据时，新建云平台支持通过Wizard引导界面还原数据中心

15.3 产品优势

ZStack for Alibaba Cloud结合了ZStack专有云的4S优势（Simple简单、Strong健壮、Scalable弹性、Smart智能）以及阿里云公共云强大的弹性支撑能力和多数据中心容灾备份等能力。

15.3.1 专有云优势

ZStack是基于专有云平台4S (Simple简单, Strong健壮, Scalable弹性, Smart智能) 标准设计的下一代云平台IaaS软件。

1. 简单 (Simple)

- 简单安装部署：提供安装文件网络下载，30分钟完成从裸机到云平台的安装部署。
- 简单搭建云平台：支持云主机的批量（生成，删除等）操作，提供列表展示和滑窗详情。
- 简单实用操作：详细的用户手册，足量的帮助信息，良好的社区，标准的API提供。
- 友好UI交互：设计精良的专业操作界面，精简操作实现强大的功能。

2. 健壮 (Strong)

- 稳定且高效的系统架构设计：拥有全异步的后台架构，进程内微服务架构，无锁架构，无状态服务架构，一致性哈希环，保证系统架构的高效稳定。目前已实现：单管理节点管理上万台物理主机、数十万台云主机；而多个管理节点构建的集群使用一个数据库、一套消息总线可管理十万台物理主机、数百万台云主机、并发处理数万个API。
- 支撑高并发的API请求：单ZStack管理节点可以轻松处理每秒上万个并发API调用请求。
- 支持HA的严格要求：在网络或节点失效情况下，业务云主机可自动切换到其它健康节点运行；利用管理节点虚拟化实现了单管理节点的高可用，故障时支持管理节点动态迁移。

3. 弹性 (Scalable)

- 支撑规模无限制：单管理节点可管理从一台到上万台物理主机，数十万台云主机。
- 全API交付：ZStack提供了全套IaaS API，用户可使用这些APIs完成全新跨地域的可用区域搭建、网络配置变更、以及物理服务器的升级。
- 资源可按需调配：云主机和云存储等重要资源可根据用户需求进行扩缩容。ZStack不仅支持对云主机的CPU、内存等资源进行在线更改，还可对云主机的网络带宽、磁盘带宽等资源进行动态调整。

4. 智能 (Smart)

- 自动化运维管理：在ZStack环境里，一切由APIs来管理。ZStack利用Ansible库实现全自动部署和升级，自动探测和重连，在网络抖动或物理主机重启后能自动回连各节点。其中定时任务支持定时开关云主机以及定时对云主机快照等轮询操作。
- 在线无缝升级：5分钟一键无缝升级，用户只需升级管控节点。计算节点、存储节点、网络节点在管控软件启动后自动升级。
- 智能化的UI交互界面：实时的资源计算，避免用户误操作。

- 实时的全局监控：实时掌握整个云平台当前系统资源的消耗情况，通过实时监控，智能化调配，从而节省IT的软硬件资源。

15.4 应用场景

云主机的应用场景非常广泛，这是因为传统x86服务器上的应用都可以封装到云主机中运行。当ZStack的各种丰富服务结合云主机本身的快速、灵活、易扩展性，就可以解决传统IT信息孤岛、资源利用率低、数据安全性差、用户管理不善以及监控维护薄弱等问题。具体应用场景包括但不限于以下几个方面：

1. 打破传统IT信息孤岛

将企业的业务整合上云，实现真正意义的互联网+。把企业信息服务系统从传统笨重的物理服务器搬迁到灵活的云主机之中，集中优化利用资源，减少重复投资。通过智能的负载均衡服务实现资源的合理化调度，同时具备的高可用能力从容面对各种突发异常，保障业务连续性。用户凭借ZStack云平台的强大能力，最大化保证企业信息服务系统长期稳定运行，同时释放运维压力，最小化TCO。

2. 提高企业开发、测试效率

在现代IT企业环境里，传统开发测试环境部署慢、审批周期长，严重制约业务上线周期。ZStack通过在线资源分配实现秒级开发测试环境的搭建和回收，从而加速企业产品开发测试节奏，加快业务上线速度。在同一个资源池中，企业可利用云主机封装的环境白天进行开发，晚上自动化测试；当前应用开发完毕，对应云主机所占用的资源可以快速释放并安排给其他项目；还可预先规划资源配置，按需申请立刻分配。

3. 企业PaaS/SaaS服务

对于无法将PaaS/SaaS服务迁移到公共云的企业，企业可以利用ZStack在数据中心中搭建一套类似阿里云ECS系统的中、大型云平台。将企业的PaaS/SaaS服务部署在ZStack之上，利用ZStack的弹性、稳定、高并发等特点确保企业对外服务安全稳定高效。

16 消息队列MQ

16.1 什么是消息队列MQ

本节主要介绍消息队列MQ的定义。

消息队列MQ (Message Queue) 是阿里巴巴集团自主研发的专业消息中间件。产品基于高可用分布式集群技术，提供消息订阅和发布、消息轨迹查询、定时（延时）消息、资源统计、监控报警等一系列消息云服务，是企业级互联网架构的核心产品。MQ历史超过9年，为分布式应用系统提供异步解耦、削峰填谷的能力，同时具备海量消息堆积、高吞吐、可靠重试等互联网应用所需的特性，是阿里巴巴双11使用的核心产品。

MQ是阿里云正式商用的产品，支持在多个地域（Region）部署高可用消息云服务，单个地域内支持多机房部署，可用性极高，即使整个机房都不可用，仍然可以为应用提供消息发布服务，产品稳定性及可用性完全按照阿里巴巴内部标准来实施，无单点。

MQ目前提供TCP和MQTT两种协议层面的接入方式，支持Java、C++以及.NET不同语言，方便不同编程语言开发的应用快速接入MQ消息云服务。用户可以将应用部署在阿里云ECS、企业自建云，或者嵌入到移动端、物联网设备中与MQ建立连接进行消息收发，同时本地开发者也可以通过公网接入MQ服务进行消息收发。

16.2 产品版本对比

下表详细罗列了MQ产品的各项功能，并标明了不同的产品版本对各项功能的支持情况。

表 16-1: MQ产品版本对比

一级功能	二级功能	三级功能	功能描述	公共云	专有云（专业版）	专有云（铂金版）
MQ Broker	消息类型	普通消息	普通的消息类型，最大支持4MB消息体，解决系统间异步解耦，削峰填谷，日志服务，大规模机器	支持	支持	支持

一级功能	二级功能	三级功能	功能描述	公共云	专有云（专业版）	专有云（铂金版）
			的Cache同步，实时计算分析等			
		顺序消息	严格按照消息的发布顺序进行顺序消费（FIFO），支持全局顺序与分区顺序；全局顺序（TPS<800）、分区顺序（高吞吐，无严格上限）	支持	不支持	支持
		事务消息	MQ提供类似 X/Open XA 的分布事务功能，通过 MQ事务消息能达到分布式事务的最终一致	支持	不支持	支持
		定时消息	将消息发送到MQ服务端，在消息发送时间（当前时间）之后的指定时间点进行投递，可精确到秒级，最大定时跨度支持40天	支持	不支持	支持

一级功能	二级功能	三级功能	功能描述	公共云	专有云（专业版）	专有云（铂金版）
	特性	低延迟（Low Latency）	经过多年双十一的演练，攻克了慢请求的业界难题，以及在高并发高吞吐量情况下系统负载过高导致的响应时间抖动问题，在2017年双十一万亿消息流转中99.6%的消息写入延迟在1ms以内，99.996%的消息写入延迟在10ms以内	支持	不支持	支持
MQTT套件	协议类型	MQTT协议	MQ for IoT，可通过MQTT协议进行端与云的双向通信	支持	支持（需要购买MQTT套件）	支持（需要购买MQTT套件）
		WebSocket	MQ for IoT，可通过WebSocket协议进行端与云的双向通信	支持	支持（需要购买MQTT套件）	支持（需要购买MQTT套件）
		国标808协议	MQ for IoT，可通过国标808协议进行端与云的双向通信	支持	不支持	不支持

一级功能	二级功能	三级功能	功能描述	公共云	专有云（专业版）	专有云（铂金版）
		新能源汽车GB/T32960协议	MQ for IoT，可通过新能源汽车GB/T32960协议进行端与云的双向通信	支持	不支持	不支持
MQ巡检服务	MQ监控巡检		提供MQ服务的在线实时健康巡检，提供服务异常告警以及产生巡检日报	支持	不支持	支持
MQ Console (用户控制台)	Topic管理	发布Topic	发布Topic	支持	支持	支持
		删除Topic	删除Topic	支持	支持	支持
		授权	为主账号或者RAM子账号授权	支持	不支持	支持
		查看授权	查看所有授权的账号信息	支持	不支持	支持
	发布订阅管理	发布管理	申请PID，同时可删除	支持	支持	支持
		订阅管理	申请CID，同时可删除	支持	支持	支持
		订阅管理——消费者状态	包括消费者状态检查和消费者连接信息	支持	支持	支持

一级功能	二级功能	三级功能	功能描述	公共云	专有云（专业版）	专有云（铂金版）
		订阅管理——重置消费位点	包括清除所有累积消息和按时间进行消费位点重置	支持	不支持	支持
	消息查询	普通消息查询	按照Topic、Message ID、Message Key查询	支持	支持	支持
		顺序消息查询	查询顺序消息，按照Topic、Message ID、Message Key查询	支持	不支持	支持
		事务消息查询	查询事务消息，按照Topic、Message ID、Message Key查询	支持	不支持	支持
		定时消息查询	查询定时消息，按照Topic、Message ID、Message Key查询	支持	不支持	支持
	MQTT管理	MQTT Group ID申请	申请MQTT Group ID，入口在Topic列表	支持	支持（需要购买MQTT套件）	支持（需要购买MQTT套件）
		MQTT Group ID管理	MQTT Group ID管理	支持	支持（需要购买MQTT套件）	支持（需要购买MQTT套件）

一级功能	二级功能	三级功能	功能描述	公共云	专有云（专业版）	专有云（铂金版）
		MQTT Topic 查询	按照一级 Topic和子 Topic查询	支持	支持（需要购买MQTT套件）	支持（需要购买MQTT套件）
		MQTT消息收发统计	按照查询条件统计	支持	支持（需要购买MQTT套件）	支持（需要购买MQTT套件）
	消息轨迹	新建查询	创建消息轨迹查询条件	支持	不支持	支持
		轨迹列表	多条消息查询结果列表	支持	不支持	支持
		轨迹可视化	单条消息的从发送到消费的整个轨迹可视化视图展现	支持	不支持	支持
	监控报警	新增监控项	进行监控配置；专有云对接客户短信接口后可以发短信	支持	支持	支持

16.3 产品优势

本文主要介绍消息队列MQ相比其他消息中间件所具备的优势。

专业

- 消息领域业内专业的消息中间件，产品历史超过9年，消息保证不丢，技术体系丰富成熟。
- 阿里巴巴内部产品名为MetaQ、Notify；开源社区产品名为RocketMQ；产品多次在国内外获奖。
- 阿里内部1000+核心应用使用，每天流转几千亿条消息，经过双11交易、商品等核心链路真实场景的验证，稳定可靠。

高可靠

- 一份消息多份落盘存储，经过严格的断电测试，消息依然保证不丢失。

- 允许海量消息堆积，单个Topic即使堆积100亿+条消息，系统高流量压力下依然可靠。
- 默认消息持久化存储3天，支持重置消费位点消费3天之内任何时间点的消息。

高性能

- 同一网络内，消息传输网络时延在10毫秒之内，性能测试下，网卡可被打满。
- 默认单Topic发送消息上限为每秒5000条，最高可申请扩展至10W以上。
- 支持4MB大消息。

多协议接入

- 支持MQTT 协议：支持主动推送模型，多级Topic模型支持一次触达1000万+ 终端，可广泛应用于物联网和社交即时通信场景。
- 支持TCP协议：提供更为专业、可靠、稳定的TCP协议的SDK接入。

独立部署

- 支持专有云独立输出，支持物理机和虚拟机，仅几台机器便可搭建完整消息云服务。
- 专有云配套mqadmin命令集和管理类Open API，方便运维人员实时监控系统状态。
- 支持混合云架构，允许通过专线的方式接入服务。

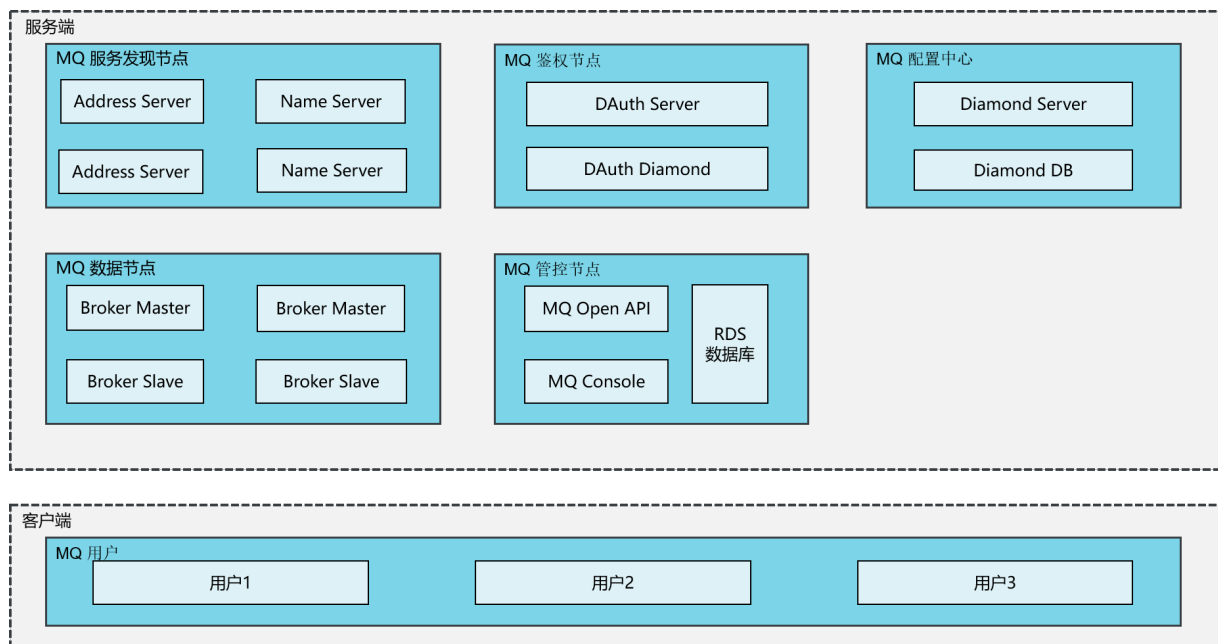
16.4 产品架构

本文向您介绍MQ的系统架构。

系统架构

MQ主要由MQ服务发现节点、MQ管控节点、MQ数据节点、MQ鉴权节点、MQ配置中心等系统模块组成。

图 16-1: MQ系统架构图



MQ服务发现节点

MQ服务发现节点由Name Server和Address Server组成。Name Server负责消息队列服务的注册与查找，是实现消息队列服务弹性部署和线性扩展的核心。

Name Server几乎无状态节点，节点之间不需要进行数据同步，可集群部署横向扩展。

Address Server，又称Cai，是消息队列MQ的地址服务器，主要负责Name Server的域名注册和发现。

MQ Broker

消息队列服务器（MQ Broker）是消息队列的核心处理模块，由多个服务节点组成服务集群，负责消息的收发以及消息的存储。MQBroker支持集群化、多副本部署，主备复制实现高可用。提供高可用、稳定高效、可线性扩容的消息服务能力。MQ Broker将消息队列Topic信息、订阅信息等注册到Name Server上，提供服务。

MQ Console

MQ Console（控制台）为用户提供消息的Topic管理、生产者管理、消费者管理、消息查询、消息轨迹查询以及监报告警等一整套完备的运维功能。用户可以通过MQ Console，快速接入消息队列、对用户资源进行管理、问题排查以及通过监报告警等服务帮助用户及时的发现问题。

MQ Open API

为方便用户接入、自主运维，MQ Open API通过HTTP/HTTPS接口的形式提供一整套API服务，用户可以通过Open API创建Topic、生产者信息、消费者信息、消息查询以及消费者状态查询等。

MQ鉴权节点

MQ鉴权节点（DAuth）为消息队列提供统一登录服务以及安全访问控制。包括资源的权限控制、跨账号与子账号访问控制、资源授权等功能。

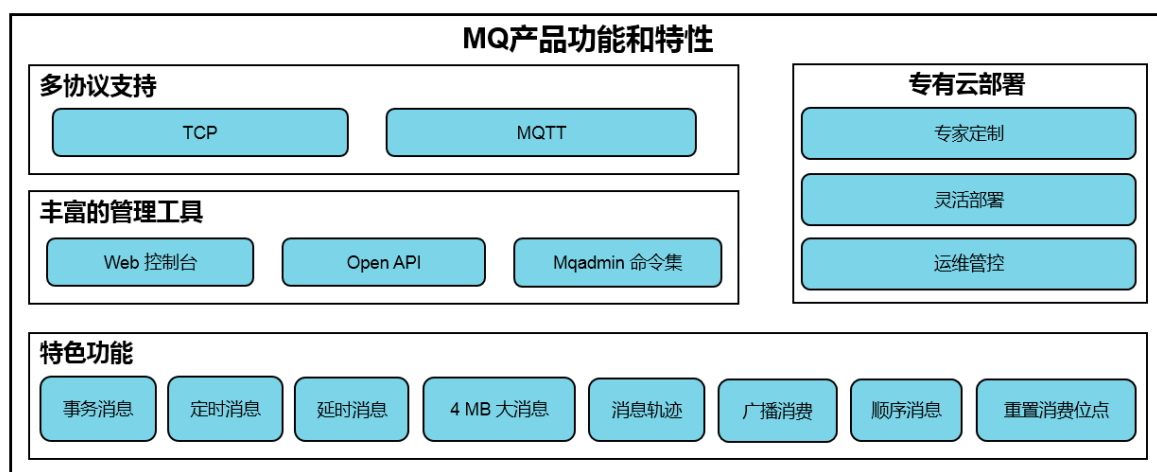
MQ配置中心

MQ配置中心（Diamond）主要负责存储消息队列的配置信息，如VIP转换规则、资源权限信息等数据。

16.5 功能特性

MQ提供了多种协议和开发语言的接入方式以及多维度的管理工具，同时针对不同的应用场景提供了一系列的特色功能。

图 16-2: 功能概览图



多协议接入

- 支持MQTT协议：支持主动推送模型，多级Topic模型支持一次触达1000万+ 终端，可广泛应用于物联网和社交即时通信场景。
- 支持TCP协议：提供更为专业、可靠、稳定的TCP协议的SDK接入。

管理工具

- Web控制台：支持Topic管理、生产者管理、消费者管理、消息查询、消息轨迹、资源报表以及监控报警管理。
- Open API：提供API允许您将MQ管理工具集成到自己的控制台。
- mqadmin命令集：专有云输出提供一套丰富的管理命令集，以命令方式对MQ服务进行管理。

特色功能

- 事务消息：实现类似X/Open XA的分布事务功能，以达到事务最终一致性状态。
- 定时（延时）消息：允许消息生产者指定消息进行定时（延时）投递，最长支持40天。
- 大消息：支持4MB大消息。
- 消息轨迹：通过消息轨迹，用户能清晰定位消息从发布者发出，经由MQ服务端，投递给消息订阅者的完整链路，方便定位排查问题。
- 广播消费：允许一个Consumer ID所标识的所有Consumer都会各自消费某条消息一次。
- 顺序消息：允许消息消费者按照消息发送的顺序对消息进行消费。
- 重置消费进度：根据时间重置消费进度，允许用户进行消息回溯或者丢弃堆积消息。

专有云部署

- 专家定制：提供技术方案设计；专家现场技术支持与培训。
- 灵活部署：支持专有云独立部署，同时支持混合云架构。
- 运维管控：专有云支持mqadmin命令集、Open API运维管理工具，方便管控平台集成以及统一运维。

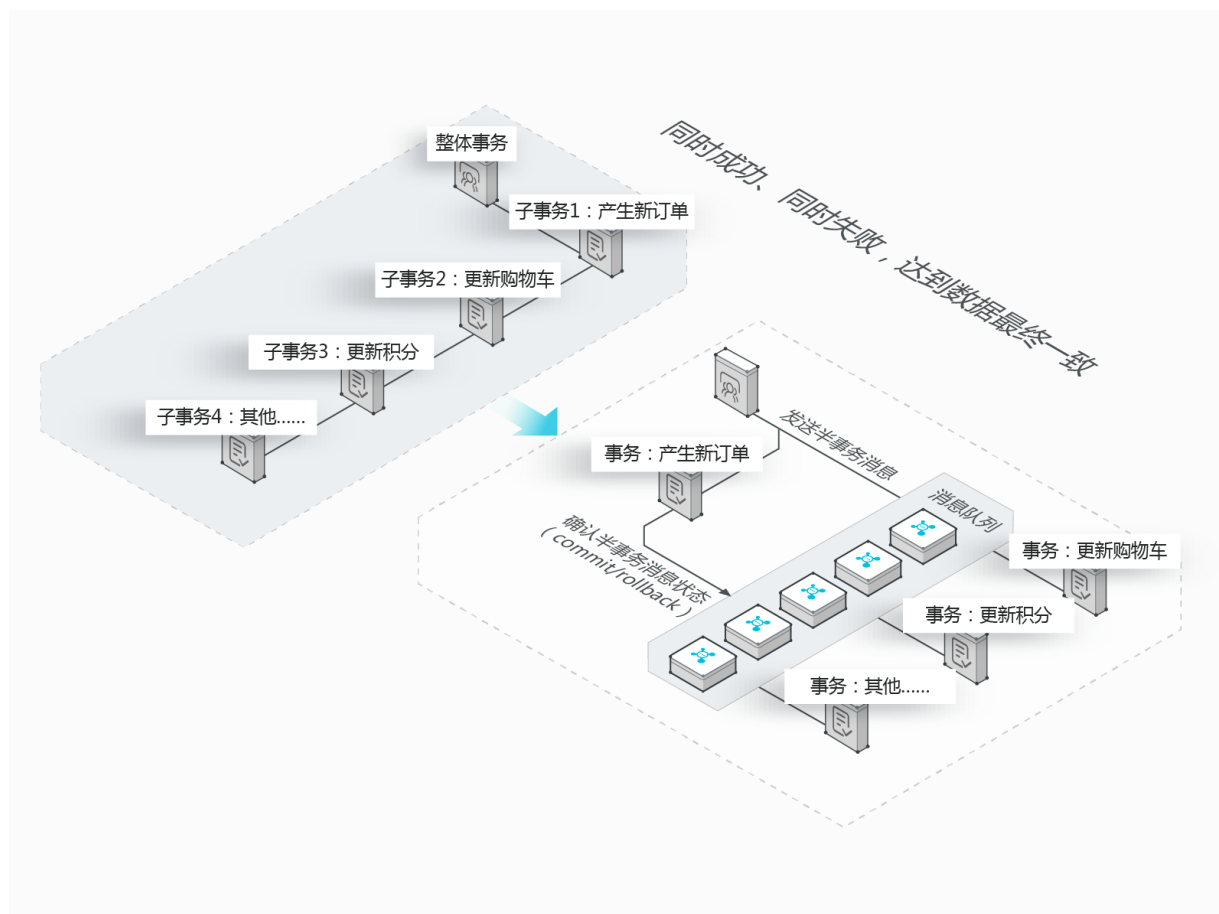
16.6 应用场景

MQ可应用于如下几个场景：

分布式事务

在传统的事务处理中，多个系统之间的交互耦合到一个事务中，响应时间长，影响系统可用性。引入分布式事务消息，交易系统和消息队列之间，组成一个事务处理，能保证分布式系统之间数据的最终一致；下游业务系统（购物车、积分、其他）相互隔离，并行处理。

图 16-3: 分布式事务场景

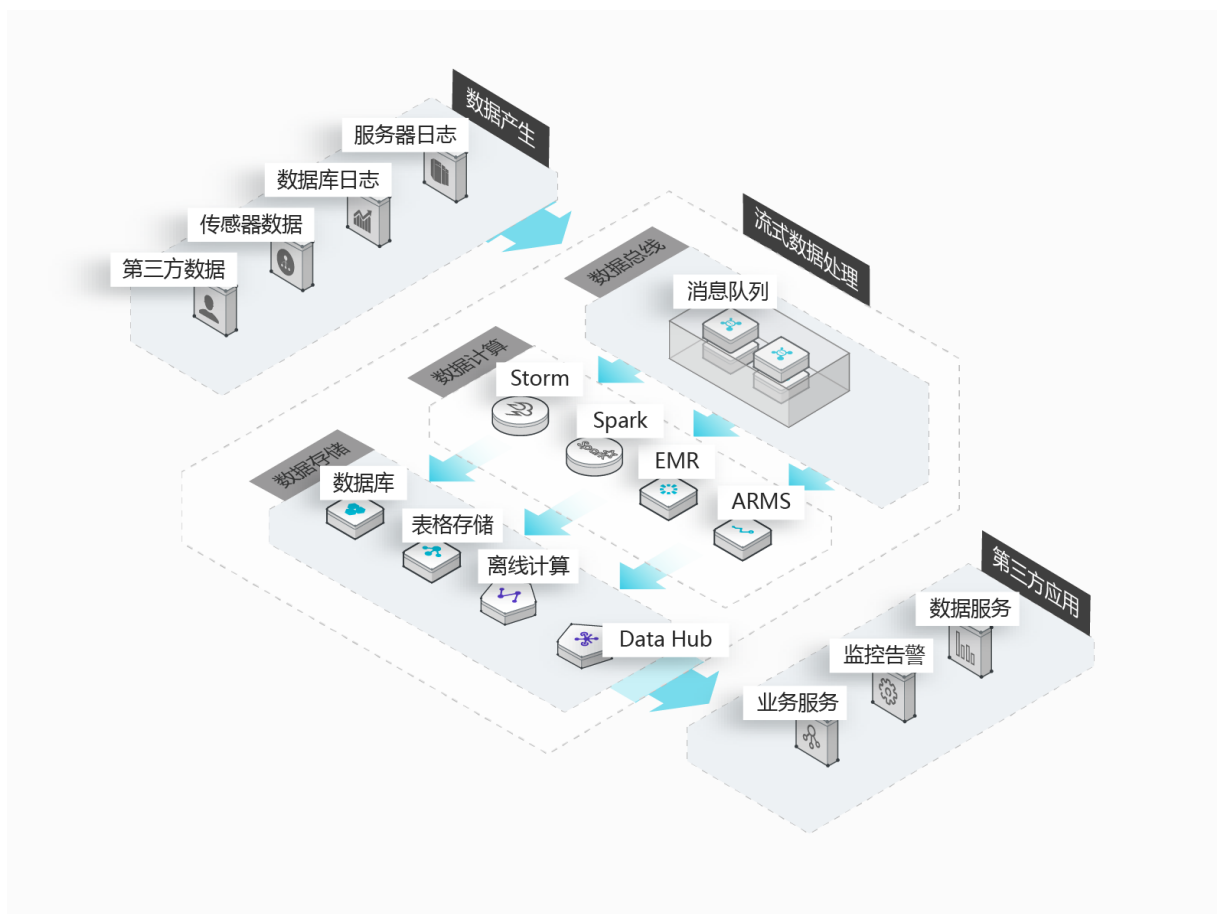


实时计算

通过MQ，将源端不停产生的数据实时流入到计算引擎，实现实时计算。您可采用如下计算引擎：

Spark / Storm / EMR / ARMS / BeamRunner。

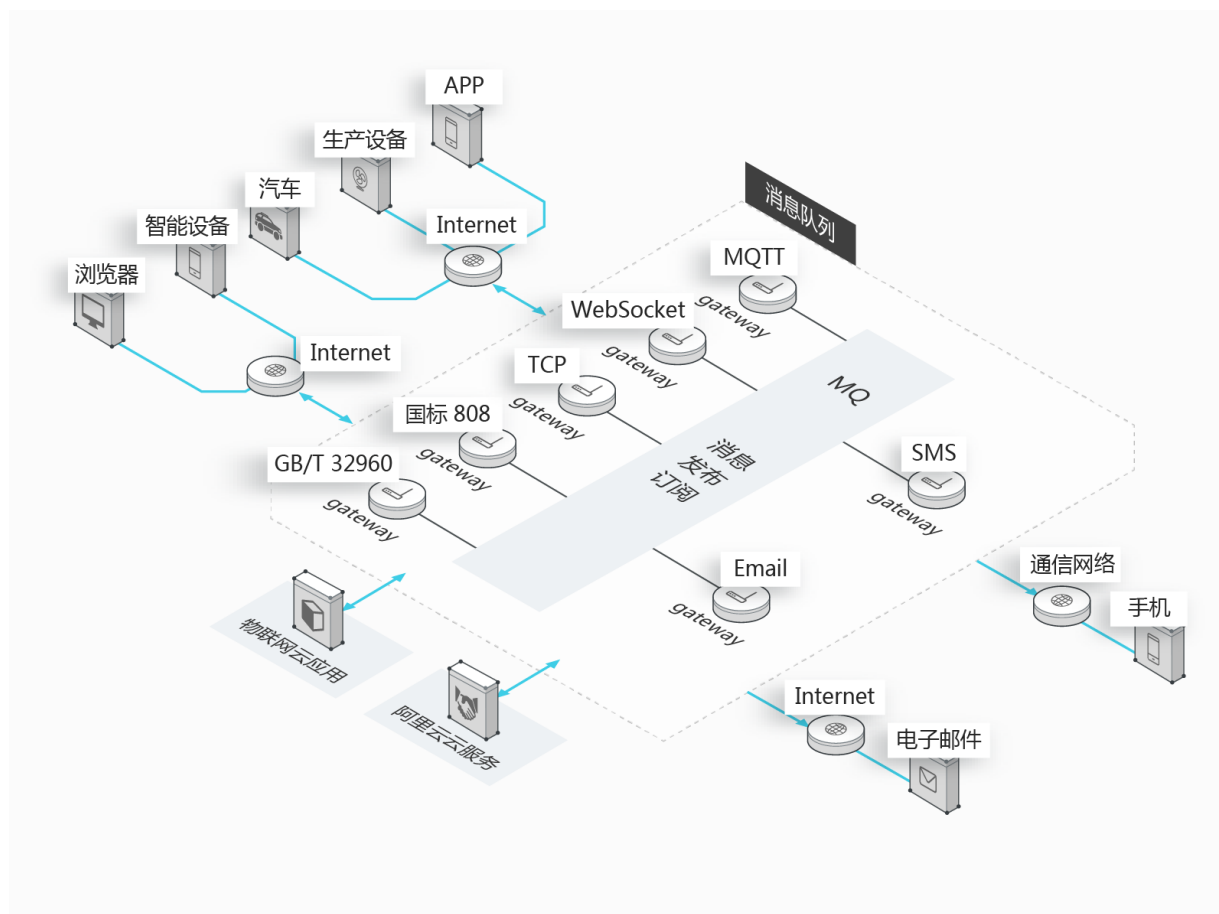
图 16-4: 实时计算场景



物联网应用

物联网设备通过微消息队列连接云端，双向通信，数据传输；设备数据通过消息队列MQ连接计算引擎，分析数据或者源数据实时高效写入到TSDB / HiStore / ODPS 等。

图 16-5: 物联网应用场景

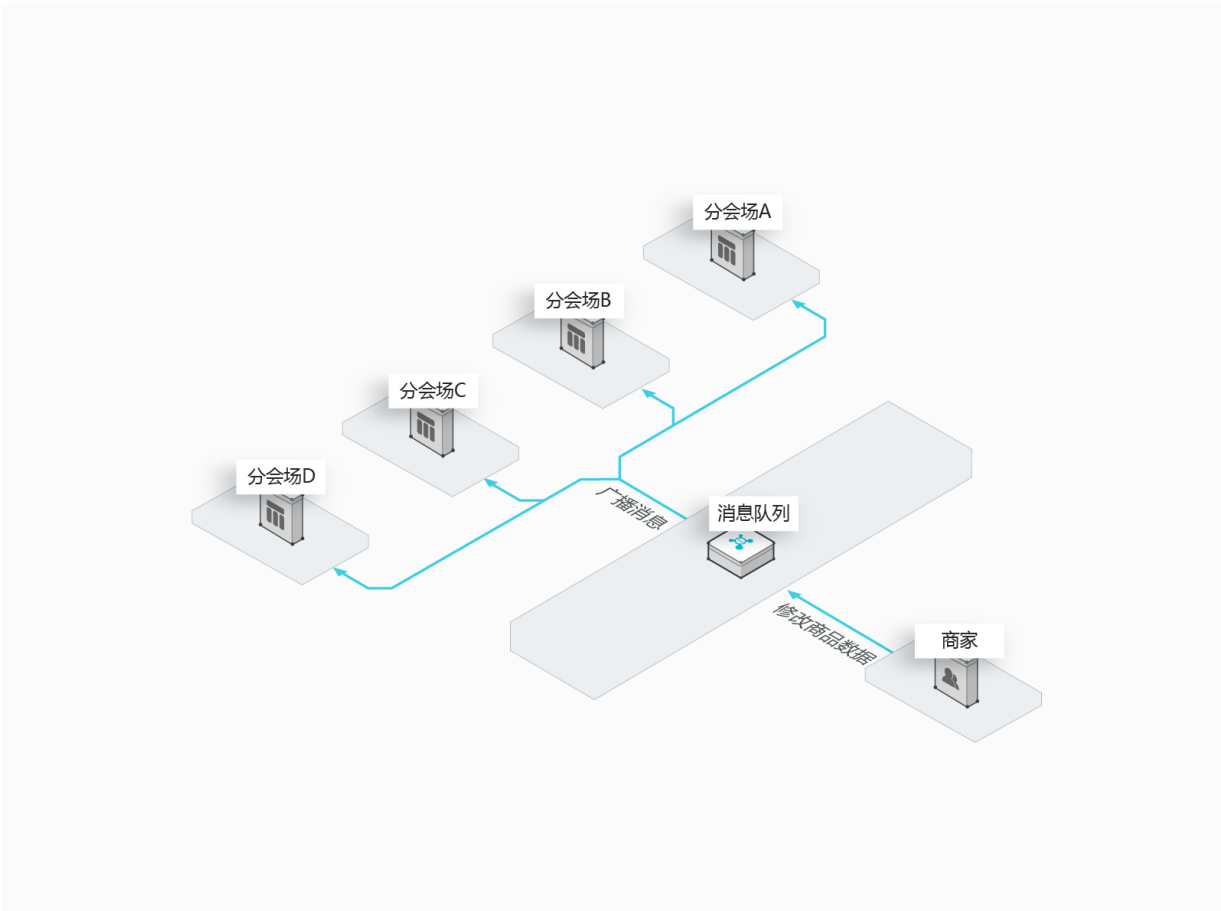


大规模缓存同步

在商业大促活动中，如“双11”大促，各个分会场会有琳琅满目的商品，每件商品的价格都会实时变化；同时，大量并发访问商品数据库，会场页面响应时间长。集中式缓存，带宽成瓶颈，无法满足对商品价格的访问需求。

MQ能够通过大规模缓存同步，减少页面响应时间；针对分会场的多缓存设计，满足客户对商品价格的访问需求。

图 16-6: 大规模缓存同步场景



16.7 使用限制

MQ对某些具体指标进行了约束，您在使用MQ时注意不要超过相应的限制值，以免系统出现异常。具体的限制项和限制值请参见下表。

表 16-2: MQ使用限制

限制项	限制值	说明
Topic名称长度	64个字符	Topic名称长度不得超过该限制，否则会导致无法发送或者订阅。
消息大小	4MB	消息大小不得超过该限制，否则消息会被丢弃。
消息保存时间	3天	消息最多保留3天，超过时间将自动滚动删除。

限制项	限制值	说明
消费位点重置	3天	支持重置消费3天之内任何时间点的消息。
Topic的消息收发TPS	按照2主2备的规格，以1K消息体为基准。 铂金版： - 消息发送：30,000条/秒 - 消息接收：30,000条/秒 专业版： - 消息发送：20,000条/秒 - 消息接收：20,000条/秒	无
定时/延时消息的延时时长	40天	<code>msg.setStartDeliverTime</code> 的参数可设置40天内的任何时刻（单位毫秒），超过40天消息发送将失败。

16.8 基本概念

本文主要对MQ涉及的专有名词及术语进行定义和解析，方便您更好地理解相关概念并使用MQ。

Message Queue	MQ，阿里云商用的专业消息中间件，是企业级互联网架构的核心产品，提供基于高可用分布式集群技术搭建的消息发布订阅、轨迹查询、资源统计、定时（延时）、监控报警等一系列消息云服务。
Message	消息，MQ中信息传递的载体。
Message ID	消息的全局唯一标识，由MQ系统自动生成，唯一标识某条消息。
Message Key	消息的业务标识，由消息生产者（Producer）设置，唯一标识某个业务逻辑。
Topic	消息主题，一级消息类型，通过Topic对消息进行分类。
Tag	消息标签，二级消息类型，用来进一步区分某个Topic下的消息分类。
Producer	消息生产者，也称为消息发布者，负责生产并发送消息。
Producer ID	一类Producer的标识。属于同一个Producer ID的消息生产者通常生产并发送一类消息，且发送逻辑一致。

Producer实例	Producer的一个对象实例，不同的Producer实例可以运行在不同进程内或者不同机器上。Producer实例线程安全，可在同一进程内多线程之间共享。
Consumer	消息消费者，也称为消息订阅者，负责接收并消费消息。
Consumer ID	一类Consumer的标识，这类Consumer通常接收并消费一类消息，且消费逻辑一致。
Consumer实例	Consumer的一个对象实例，不同的Consumer实例可以运行在不同进程内或者不同机器上。一个Consumer实例内配置线程池消费消息。
集群消费	一个Consumer ID所标识的所有Consumer平均分摊消费消息。例如某个Topic有9条消息，一个Consumer ID有3个Consumer实例，那么在集群消费模式下每个实例平均分摊，只消费其中的3条消息。
广播消费	一个Consumer ID所标识的所有Consumer都会各自消费某条消息一次。例如某个Topic有9条消息，一个Consumer ID有3个Consumer实例，那么在广播消费模式下每个实例都会各自消费9条消息。
定时消息	Producer将消息发送到MQ服务端，但并不期望这条消息立马投递，而是推迟到在当前时间点之后的某一个时间投递到Consumer进行消费，该消息即定时消息。
延时消息	Producer将消息发送到MQ服务端，但并不期望这条消息立马投递，而是延迟一定时间后才投递到Consumer进行消费，该消息即延时消息。
事务消息	MQ提供类似X/Open XA的分布事务功能，通过MQ事务消息能达到分布式事务的最终一致。
顺序消息	MQ提供的一种按照顺序进行发布和消费的消息类型，分为全局顺序消息和分区顺序消息。
顺序发布	对于指定的一个Topic，客户端将按照一定的先后顺序进行发送消息。
顺序消费	对于指定的一个Topic，按照一定的先后顺序进行接收消息，即先发送的消息一定会先被客户端接收到。
全局顺序消息	对于指定的一个Topic，所有消息按照严格的先入先出（FIFO）的顺序进行发布和消费。
分区顺序消息	对于指定的一个Topic，所有消息根据sharding key进行区块分区。同一个分区内的消息按照严格的FIFO顺序进行发布和消费。Sharding key是顺序消息中用来区分不同分区的关键字段，和普通消息的key是完全不同的概念。

消息堆积	Producer已经将消息发送到MQ服务端，但由于Consumer消费能力有限，未能在短时间内将所有消息正确消费掉，此时在MQ服务端保存着未被消费的消息，该状态即消息堆积。
消息过滤	订阅者可以根据消息标签（Tag）对消息进行过滤，确保订阅者最终只接收被过滤后的消息类型。消息过滤在MQ服务端完成。
消息轨迹	在一条消息从发布者发出到订阅者消费处理过程中，由各个相关节点的时间、地点等数据汇聚而成的完整链路信息。通过消息轨迹，用户能清晰定位消息从发布者发出，经由MQ服务端，投递给消息订阅者的完整链路，方便定位排查问题。
重置消费位点	以时间轴为坐标，在消息持久化存储的时间范围内（默认3天），重新设置消息订阅者对其订阅Topic的消费进度，设置完成后订阅者将接收设定时间点之后由消息发布者发送到MQ服务端的消息。