ALIBABA CLOUD

阿里云

云数据库 OceanBase

产品简介

文档版本：20210617

阿里云

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

格式	说明	样例
危险	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。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会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，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	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。	单击**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。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.什么是OceanBase

OceanBase是由蚂蚁集团、阿里巴巴完全自主研发的分布式关系型数据库，始创于2010年。

OceanBase具有数据强一致、高可用、高性能、在线扩展、高度兼容SQL标准和主流关系型数据库、低成本等特点。OceanBase至今已成功应用于支付宝全部核心业务：交易、支付、会员、账务等系统以及阿里巴巴淘宝（天猫）收藏夹、P4P广告报表等业务。除在蚂蚁集团和阿里巴巴业务系统中获广泛应用外，从2017年开始，OceanBase开始服务外部客户，客户包括南京银行、浙商银行、人保健康险等。

产品优势

- **高性能**：OceanBase采用了读写分离的架构，把数据分为基线数据和增量数据。其中增量数据放在内存里（MemTable），基线数据放在SSD盘（SSTable）。对数据的修改都是增量数据，只写内存。所以DML是完全的内存操作，性能非常高。
- **低成本**：OceanBase通过数据编码压缩技术实现高压缩。数据编码是基于数据库关系表中不同字段的值域和类型信息，所产生的一系列的编码方式，它比通用的压缩算法更懂数据，从而能够实现更高的压缩效率。
- **高兼容**：兼容常用MySQL/ORACLE功能及MySQL/ORACLE前后台协议，业务零修改或少量修改即可从MySQL/ORACLE迁移至OceanBase。
- **高可用**：数据采用多副本存储，少数副本故障不影响数据可用性。通过“三地五中心”部署实现城市级故障自动无损容灾。

产品介绍

2.产品优势与应用场景

OceanBase是一款金融级的分布式关系数据库，具备高性能、高可用、强一致、可扩展和兼容性高等典型优势，适用于对性能、成本和扩展性要求高的金融场景。

主要特性

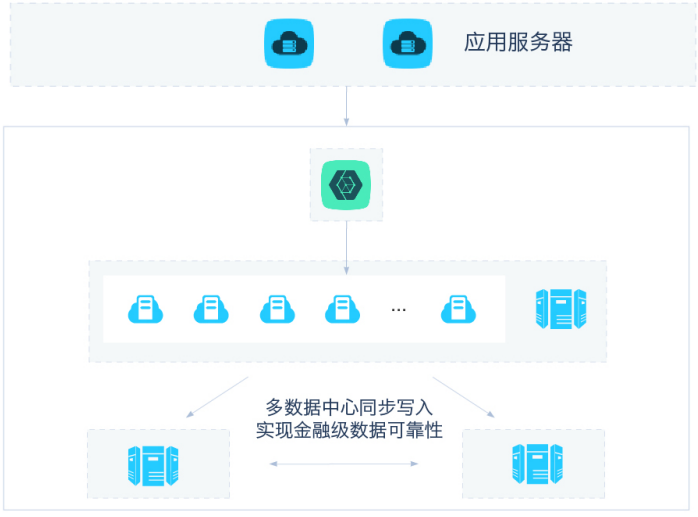
- **高性能**：存储采用读写分离架构，计算引擎全链路性能优化，准内存数据库性能。
- **低成本**：使用PC服务器和低端SSD，高存储压缩率降低存储成本，高性能降低计算成本，多租户混部充分利用系统资源。
- **高可用**：数据采用多副本存储，少数副本故障不影响数据可用性。通过“三地五中心”部署实现城市级故障自动无损容灾。
- **强一致**：数据多副本通过paxos协议同步事务日志，多数派成功事务才能提交。缺省情况下读、写操作都在主副本进行，保证强一致。
- **可扩展**：集群节点全对等，每个节点都具备计算和存储能力，无单点瓶颈。可线性、在线扩展和收缩。
- **兼容性**：兼容常用MySQL/ORACLE功能及MySQL/ORACLE前后台协议，业务零修改或少量修改即可从MySQL/ORACLE迁移至OceanBase。

应用场景

OceanBase的产品定位是一款分布式关系数据库，经过多年蚂蚁金服内部业务的打磨，目前已经支持蚂蚁金服100%核心交易系统，稳定支撑阿里、蚂蚁内部上百个关键业务以及浙商银行、南京银行等多个外部客户。OceanBase产品适用于金融、证券等涉及交易、支付和账务等对高可用、强一致要求特别高，同时对性能、成本和扩展性有需求的金融属性场景，以及各种关系型结构化存储的OLTP应用。OceanBase天然的Share-Nothing分布式架构对于各种OLAP型应用也有很好的支持，例如云数据库OceanBase适用于以下典型场景：

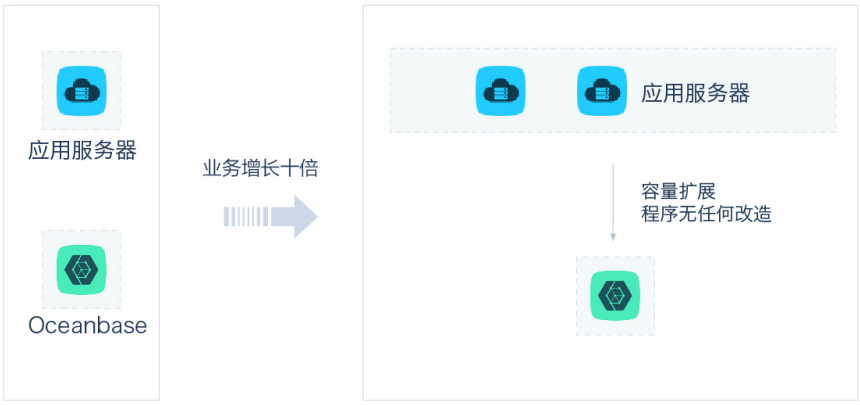
- 金融级数据可靠性需求

  金融环境下通常对数据可靠性有更高的要求，OceanBase每一次事务提交，对应日志总是会在多个数据中心实时同步，并持久化。即使是数据中心级别的灾难发生，总是可以在其他的数据中心恢复每一笔已经完成的交易，实现了真正金融级别的可靠性要求。



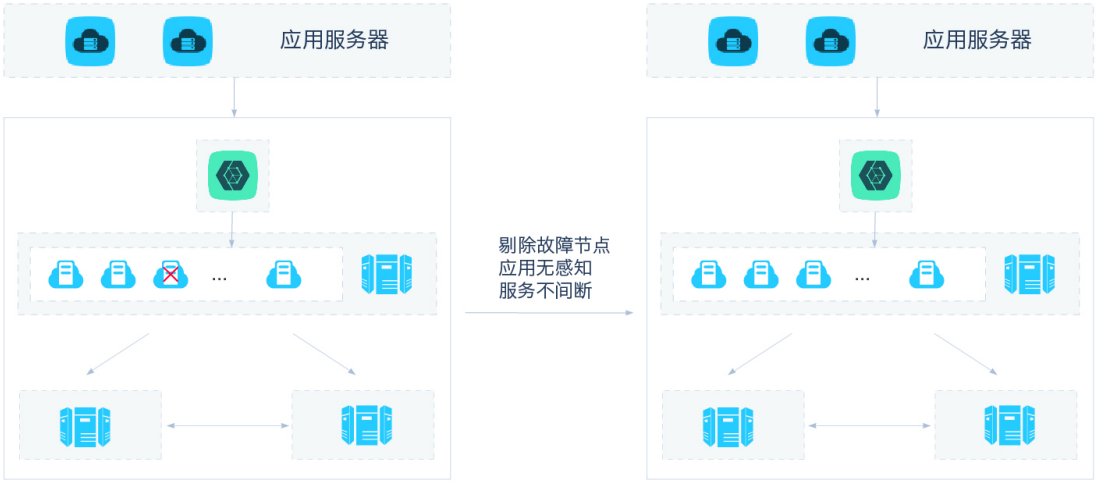
- 让数据库适应飞速增长的业务

业务的飞速成长，通常会给数据库带来成倍压力。OceanBase作为一款真正意义的分布式关系型数据库，由一个个独立的通用计算机作为系统各个节点，数据根据容量大小、可用性自动分布在各个节点，当数据量不断增长时，OceanBase可以自动扩展节点的数量，满足业务需求。



- 连续不间断的服务

  企业连续不间断的服务，通常意味着给客户最流畅的产品体验。分布式的OceanBase集群，如果某个节点出现异常时，可以自动剔除此服务节点，该节点对应的数据有多个其他副本，对应的数据服务也由其他节点提供。甚至当某个数据中心出现异常，OceanBase可以在短时间内将服务节点切换到其他数据中心，可以保证业务持续可用。



3.产品架构

OceanBase 数据库采用 Shared-Nothing 架构，各个节点之间完全对等，每个节点都有自己的 SQL 引擎、存储引擎，运行在普通 PC 服务器组成的集群之上，具备可扩展、高可用、高性能、低成本、云原生等核心特性。

OceanBase 数据库的整体架构如下图所示。

集群架构

OceanBase 数据库支持数据跨地域（Region）部署，每个地域可能位于不同的城市，距离通常比较远，所以 OceanBase 数据库可以支持多城市部署，也支持多城市级别的容灾。一个 Region 可以包含一个或者多个 Zone，Zone 是一个逻辑的概念，它包含了 1 台或者多台运行了 OBServer 进程的服务器（以下简称 OBServer）。每一个 Zone 上包含一个完整的数据副本，由于 OceanBase 数据库的数据副本是以分区为单位的，所以同一个分区的数据会分布在多个 Zone 上。每个分区的主副本所在服务器被称为 Leader，所在的 Zone 被称为 Primary Zone。如果不设定 Primary Zone，系统会根据负载均衡的策略，在多个全功能副本里自动选择一个作为 Leader。

每个 Zone 会提供两种服务：总控服务（RootService）和分区服务（PartitionService）。其中每个 Zone 上都会存在一个总控服务，运行在某一个 OBServer 上，整个集群中只存在一个主总控服务，其他的总控服务作为主总控服务的备用服务运行。总控服务负责整个集群的资源调度、资源分配、数据分布信息管理以及 Schema 管理等功能。 其中：

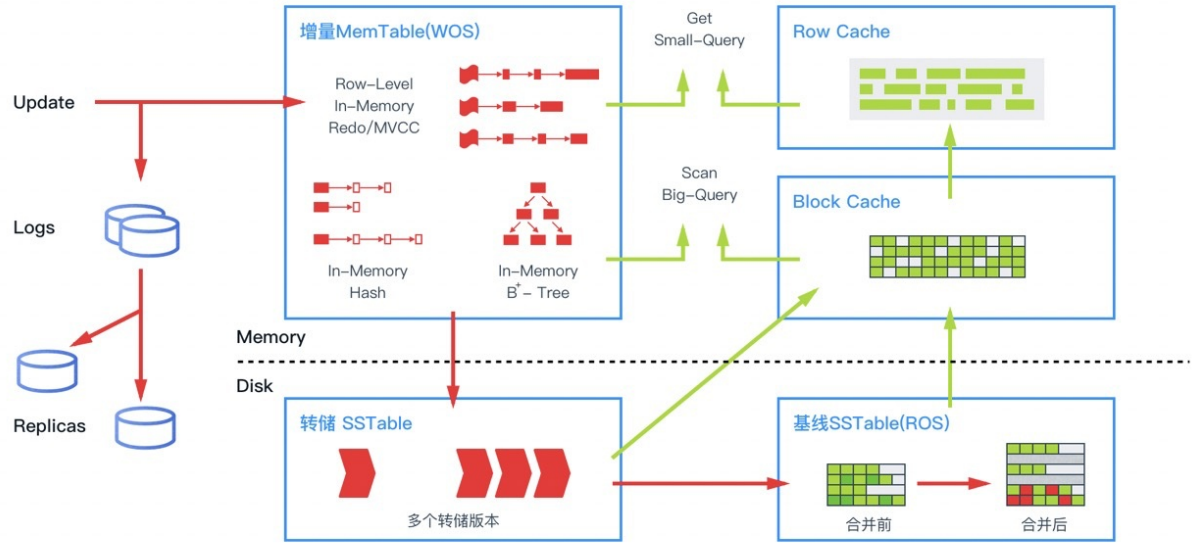
- 资源调度主要包含了向集群中添加、删除 OBServer，在 OBServer 中创建资源规格、Tenant 等供用户使用的资源；
- 资源均衡主要是指各种资源（例如：Unit）在各个 Zone 或者 OBServer 之间的迁移。
- 数据分布管理是指总控服务会决定数据分布的位置信息，例如：某一个分区的数据分布到哪些 OBServer 上。
- Schema 管理是指总控服务会负责调度和管理各种 DDL 语句。

分区服务用于负责每个 OBServer 上各个分区的管理和操作功能的模块，这个模块与事务引擎、存储引擎存在很多调用关系。

OceanBase 数据库基于 Paxos 的分布式选举算法来实现系统的高可用，最小的粒度可以做到分区级别。集群中数据的一个分区（或者称为副本）会被保存到所有的 Zone 上，整个系统中该副本的多个分区之间通过 Paxos 协议进行日志同步。每个分区和它的副本构成一个独立的 Paxos 复制组，其中一个分区为主分区（Leader），其它分区为备分区（Follower）。所有针对这个副本的写请求，都会自动路由到对应的主分区上进行。主分区可以分布在不同的 OBServer 上，这样对于不同副本的写操作也会分布到不同的数据节点上，从而实现数据多点写入，提高系统性能。

存储引擎

OceanBase 数据库的存储引擎采用了基于 LSM-Tree 的架构，把基线数据和增量数据分别保存在磁盘（SSTable）和内存（MemTable）中，具备读写分离的特点。对数据的修改都是增量数据，只写内存。所以 DML 是完全的内存操作，性能非常高。读的时候，数据可能会在内存里有更新过的版本，在持久化存储里有基线版本，需要把两个版本进行合并，获得一个最新版本。

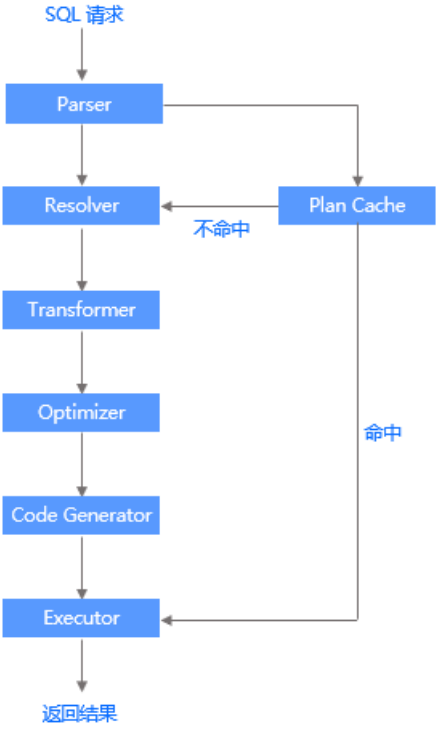


如上图所示，在内存中针对不同的数据访问行为，OceanBase 数据库设计了多种缓存结构。除了常见的数据块缓存之外，也会对行进行缓存，行缓存会极大加速对单行的查询性能。为了避免对不存在行的空查，OceanBase 数据库对行缓存构建了布隆过滤器，并对布隆过滤器进行缓存。OLTP 业务大部分操作为小查询，通过小查询优化，OceanBase 数据库避免了传统数据库解析整个数据块的开销，达到了接近内存数据库的性能。当内存的增量数据达到一定规模的时候，会触发增量数据和基线数据的合并，把增量数据落盘。同时每天晚上的空闲时刻，系统也会启动每日合并。另外，由于基线是只读数据，而且内部采用连续存储的方式，OceanBase 数据库可以根据不同特点的数据采用不同的压缩算法，既能做到高压缩比，又不影响查询性能，大大降低了成本。

SQL 引擎

OceanBase 数据库的 SQL 引擎是整个数据库的数据计算中枢，和传统数据库类似，整个引擎分为解析器、优化器、执行器三部分。当 SQL 引擎接受到了 SQL 请求后，经过语法解析、语义分析、查询重写、查询优化等一系列过程后，再由执行器来负责执行。所不同的是，在分布式数据库里，查询优化器会依据数据的分布信息生成分布式的执行计划。如果查询涉及的数据在多台服务器，需要走分布式计划，这是分布式数据库 SQL 引擎的一个重要特点，也是十分考验查询优化器能力的场景。OceanBase 数据库查询优化器做了很多优化，诸如算子下推、智能连接、分区裁剪等。如果 SQL 语句涉及的数据量很大，OceanBase 数据库的查询执行引擎也做了并行处理、任务拆分、动态分区、流水调度、任务裁剪、子任务结果合并、并发限制等优化技术。

下图描述了一条 SQL 语句的执行过程，并列出了 SQL 引擎中各个模块之间的关系。



- Parser（词法/语法解析模块）

  Parser 是整个 SQL 执行引擎的词法或语法解析器，在收到用户发送的 SQL 请求串后，Parser 会将字符串分成一个个的单词，并根据预先设定好的语法规则解析整个请求，将 SQL 请求字符串转换成带有语法结构信息的内存数据结构，称为语法树（Syntax Tree）。

  为了加速 SQL 请求的处理速度，OceanBase 数据库对 SQL 请求采用了特有的快速参数化，以加速查找执行计划的速度。

- Resolver（语义解析模块）

  当生成语法树之后，Resolver 会进一步将该语法树转换为带有数据库语义信息的内部数据结构。在这一过程中，Resolver 将根据数据库元信息将 SQL 请求中的 token 翻译成对应的对象（例如库、表、列、索引等），生成语句树。

- Transfomer（逻辑改写模块）

  在查询优化中，经常利用等价改写的方式，将用户 SQL 转换为与之等价的另一条 SQL，以便于优化器生成最佳的执行计划，这一过程称为查询改写。Transformer 在 Resolver 之后，分析用户 SQL 的语义，并根据内部的规则或代价模型，将用户 SQL改写为与之等价的其他形式，并将其提供给后续的优化器做进一步的优化。Transformer 的工作方式是在原 Statement Tree 上做等价变换，变换的结果仍然是一棵语句树。

- Optimizer（优化器）

  优化器是整个 SQL 优化的核心，其作用是为 SQL 请求生成最佳的执行计划。在优化过程中，优化器需要综合考虑 SQL 请求的语义、对象数据特征、对象物理分布等多方面因素，解决访问路径选择、联接顺序选择、联接算法选择、分布式计划生成等多个核心问题，最终选择一个对应该 SQL 的最佳执行计划。SQL 的执行计划是一棵由多个操作符构成的执行树。

- Code Generator（代码生成器）

  优化器负责生成最佳的执行计划，但其输出的结果并不能立即执行，还需要通过代码生成器将其转换为可执行的代码，这个过程由 Code Generator 负责。

- Executor（执行器）

  当 SQL 的执行计划生成后，Executor 会启动该 SQL 的执行过程。对于不同类型的执行计划，Executor 的逻辑有很大的不同：对于本地执行计划，Executor 会简单的从执行计划的顶端的算子开始调用，由算子自身的逻辑完成整个执行的过程，并返回执行结果；对于远程或分布式计划，Executor 需要根据预选的划分，将执行树分成多个可以调度的线程，并通过 RPC 将其发送给相关的节点执行。

- Plan Cache（执行计划缓存模块）

  执行计划的生成是一个比较复杂的过程，耗时比较长，尤其是在 OLTP 场景中，这个耗时往往不可忽略。为了加速 SQL 请求的处理过程，SQL 执行引擎会将该 SQL 第一次生成的执行计划缓存在内存中，后续的执行可以反复执行这个计划，避免了重复查询优化的过程。

4.产品部署方案

云数据库 OceanBase 分为高可用版本和基础版本，支持多机房部署、双机房部署和单机房部署三种部署方案。

高可用版本

云数据库 OceanBase 高可用版本为您提供机房级、主机级容灾等多种策略的高可用服务，支持多机房部署、双机房部署和单机房部署三种部署方案。

多机房部署

云数据库 OceanBase 多机房部署指将三个节点部署在三个不同可用区，实现跨可用区容灾，不额外收费。每个节点均为全功能副本，其中一个主副本提供读写服务，两个备副本提供只读服务。当主副本发生故障时，备副本将会升为主副本继续提供读写服务。

对性能和多机房可用性有着更高要求的客户建议选择多机房部署方案。

双机房部署

云数据库 OceanBase 双机房部署将两个节点部署在两个可用区，其中一个节点作为主副本提供读写服务，另外一个备节点可以提供只读服务。并在第三个可用区部署一个日志节点，该节点仅用于日志同步，不包含数据副本，不对外提供读写服务。双机房部署仍具备机房级容灾能力，与多机房部署相比在性价比上有较大提升。

单机房部署

云数据库 OceanBase 单机房部署指所有节点位于同一可用区，具备主机级别故障容灾能力，同时由于单机房部署的写请求无需进行跨机房同步，延时较小。

基础版本

云数据库 OceanBase 基础版本支持单机房单节点部署，容灾能力较弱。您购买时无需选择存储容量，存储费用根据实际使用量按小时计费。

云数据库 OceanBase 基础版本一般用于开发测试环境，实际生产环境建议使用高可用版本。

5.客户案例

南京银行

公司介绍

南京银行成立于 1996 年 2 月 8 日，是一家具有由国有股份、中资法人股份、外资股份及众多个人股份共同组成独立法人资格的股份制商业银行，实行一级法人体制。先后于 2001 年、2005 年引入国际金融公司和法国巴黎银行入股，在全国城商行中率先启动上市辅导程序并于 2007 年成功上市。入选英国《银行家》杂志公布的全球 1000 家大银行排行榜和全球银行品牌 500 强榜单，2017 年分列第 146 位和第 131 位。在互联网金融飞速发展的当下，南京银行积极转型，努力打造自己的互联网金融平台。

李勇

南京银行信息技术部副总经理

“OceanBase 数据库系统经过蚂蚁金服内部大量互联网金融场景验证，给了我们尝试使用的信心。实践证明，南京银行选择 OceanBase 数据库，给“鑫云 +”互金平台提供了更加坚实的保证。”

业务挑战

1. 在线水平扩展能力：能够在不中断业务的情况下，快速扩展硬件能力。
2. 高并发处理能力：能够应对类似双十一的瞬间高并发流量。
3. 软硬件和运维成本：能够在满足上述需求的同时，大幅降低成本。

优化结果

2017 年 9 月 28 日，南京银行、阿里云以及蚂蚁金服举行战略合作协议签约仪式，共同发布南京银行“鑫云 +”互金开放平台。南京银行“鑫云+”互金开放平台是阿里云、蚂蚁金融云合作整体输出的第一次努力，通过“鑫云”+平台的建设，南京银行互金核心系统在如下方面获得了质的提升：

1. 扩展能力：在平台建设期间和投产后，OceanBase 做过多次在线水平扩展。
2. 处理能力：从 10 万笔/日以下，增加到 100 万笔/日以上。
3. 成本降低：单账户的维护成本从 30~50 元/账户，降到 4 元/账户。

网商银行

公司介绍

网商银行定位为网商首选的金融服务商、互联网银行的探索者和普惠金融的实践者，为小微企业、大众消费者、农村经营者与农户、中小金融机构提供服务，是中国第一家将核心系统架构在金融云上的银行。基于金融云计算平台以及 OceanBase 的海量存储，网商银行拥有处理高并发金融交易、海量大数据和弹性扩容的能力，可以利用互联网和大数据的优势，给更多小微企业提供金融服务。

唐家才

网商银行 CTO

“网商银行选择 OceanBase 三地五中心部署架构，不仅在数据上从具备抵御同城机房故障提升到具备异地城市容灾的能力，同时内置的多租户隔离的能力，满足全行多应用系统的管理与使用需求，让应用系统多活架构设计上变的异常简单。”

业务挑战

1. 具备城市级别的容灾能力满足监管要求，同时最大限度地减少容灾上部署、运营和维护IT基础设施的工作量，从而降低系统运行和维护的成本。
2. 提供标准、安全和高效的数据库多租户隔离环境及管理工具，满足全行多应用系统（如存贷汇核心系统）的管理与使用需求。

优化结果

选择 OceanBase 三地五中心部署架构，实现了业务应用上杭州，上海异地多活的能力，极大的提升了全行的系统吞吐量。同时容灾上具备任意时间，任意服务器，任意机房，任意城市出现不可抗拒因素灾难时，完全无需人工接入的无损自适应容灾，RPO=0,RTO<30 秒，极大的减少了运营和维护 IT 基础设施的工作量，从而降低了运行和维护的成本。

1. 在平台建设期间和投产后，OceanBase 做过多次在线水平扩展，具备高扩展能力。
2. 借助 OceanBase 提供的多租户特性，在集群上按照业务重要程度与流量配比分配资源策略，在资源的共享与隔离上取得了最佳的平衡，极大的减少了 IT 基础设施的采购成本。同时通过 OceanBase 云平台运维管控产品，日常运营维护 100% 白屏化，大大的降低了维护运营成本。

支付宝

公司介绍

支付宝是国内领先的第三方支付平台，致力于提供“简单、安全、快速”的支付解决方案。在 2017 年双十一购物节，支付峰值最高达 25.6 万笔/秒。 支付宝的所有核心业务数据包括交易、账务、花呗、借呗等均存储在 OceanBase 上，相比传统的 Oracle 方案，OceanBase 使用更低的成本，实现了更高的扩展性，帮助支付宝平稳应对各种促销业务高峰。

程立

蚂蚁金服 CTO

“OceanBase 稳定支撑了支付宝的核心交易、支付与账务，经历了多次“双十一”的考验，形成了跨机房、跨区域部署的高可用架构，并在日常运行、应急演练和容灾切换中发挥了重要作用。”

业务挑战

1. 一致性，一致性是金融业务的生命线，为了应对硬件或者系统故障（IDC/OS/机器故障），传统的数据库在这方面为业务提供多种选择。最大可用模式在主库故障情况下可能造成数据丢失。最大保护模式会提高全年的不可用时间，并造成性能下降。
2. 扩展性，传统的基于硬件是 scale up 方案成本是非常高的，在蚂蚁内部采用 sharding 的方式，通过自研中间件 ZDAL 屏蔽分表信息，对业务提供单表视图。
3. 可用性，金融业务对系统的可用性要求非常高，通常在 99.99% 以上。一些金融机构通常采用数据库本身的特性来提供系统的可用性，以 Oracle 为例，为了保证高可用目前有两种方案：RAC 方案和 DataGuard 方案。在故障场景下恢复时间会比较长，因此业务上通常会实现一些高可用方案如Failover 等等提高故障恢复时间，同时也引入了大量的复杂度。
4. 成本和性能，对于传统数据库而言，成本分为机器成本和许可证（license）成本。不同于传统的金融企业，互联网金融服务的用户数非常大，传统的收费方式会带来非常高昂的成本。

优化结果

1. OceanBase 在一致性方面做了以下几个事情，架构层面引入 Paxos 协议，多重数据校验机制，完善支付宝业务模型，多重机制保障金融级别的一致性。
2. OceanBase 的高可用策略与传统的基于共享存储的方案有很大不同，OceanBase 采用 Share Nothing 架构，并且每个组件都有各自的持续可用方案。
3. 在部署架构上也引入了不同，支付宝的订单型业务采用了"同城三中心"的部署方式，具备单机和单 IDC 故障的容灾，通过 RFO 的方式提供异地容灾能力，在性能和可用性方面做到了极致的权衡。账务型业务采用"三地五中心"部署方式，除了具备单机，单 IDC 的容灾能力，还具备城市级故障自动容灾能力。在同城容灾和异地容灾场景下，RPO=0，RTO<30 秒。

淘宝网

公司介绍

阿里巴巴是全球最大的电子商务网站之一，2017 天猫双 11 整天成交金额 1682 亿元。淘宝(天猫)收藏夹是用户非常喜爱的功能之一，用户在浏览淘宝网站的时候会把自己喜欢的商品或者店铺加入收藏夹中，以便于以后能迅速的找到之前收藏过的商品。用户同时还能跟好友分享自己的收藏商品或者店铺。目前淘宝收藏夹已经达到几百 TB 规模，服务 8 亿淘宝用户。

林玉炳

淘宝技术部基础交易

“收藏夹服务集团内 50+ 业务方，总体收藏关系数将近千亿，并发量数十万，OceanBase 非常好的支持了收藏夹的读写场景，经历了多次大促高并发考验，运行稳定，吞吐量高，性能优异，成本低廉，非常好的满足了收藏夹的业务发展需求。”

业务挑战

1. 收藏夹每天写入量千万级的写入量，同时需要支持数万每秒的写入峰值。
2. 收藏夹的查询是收藏记录和商品信息的一个连接查询，平均每个查询都需要连接上百条记录，且双 11 的用户展示的峰值能达到数十万每秒左右。对数据库的性能提出了严苛要求。

优化结果

1. 利用 OceanBase 数据库先进的分布式的特性，把单表数据自动分布到数十台廉价微型服务器上，这数十台服务器同时支持每天的高强度写入，轻松化解写入压力。
2. 利用 OceanBase 出色的容灾特性，三个机房部署，即使某个机房整体异常，也不会影响用户访问。
3. 利用 OceanBase 提供的物化视图技术，消除实际查询中的连接操作，使得数据库的查询能力几十倍提升，保障了双 11 用户查询收藏夹的顺畅的用户体验。

阿里妈妈

公司介绍

阿里妈妈广告业务主要是一种 P4P（Pay for Performance）形式的广告业务系统，而报表中心作为阿里妈妈向广告主透出广告效果数据的唯一平台，在阿里巴巴大平台丰富多样的商业场景下，为客户提供优质，高效，可靠的数据服务，成为广告投放的风向标。报表平台将品类繁多的商业广告信息进行分类汇总，提炼出直通车，钻展，品效，一站式，原生内容，新单品等业务线的报表服务，为阿里巴巴商务平台上的卖家提供各种精确的，多维的广告效果分析服务。

张炜宇

阿里妈妈基础共享技术开发平台总监

“OceanBase 很好的满足了我们广告业务对于存储系统扩展性，并行计算，统计计算，高吞吐，低时延，资源隔离等大数据处理的需求，在报表业务的演进中帮助我们建立了一套业务和平台分离，面向效果指标开发的通用系统。”

业务挑战

1. 开发效率：报表平台承载了阿里巴巴商业平台上品类繁多的广告数据的汇总和对广告主的展示，不同业务线有不同的报表诉求，即使在相同的业务线下，基于不同的营销场景，也会有不同维度的数据抽象和封装。但在报表开发的演进过程中，报表平台逐步建立起业务与系统分离，由之前的面向报表的开发模式，转变为面向指标的通用解决方案，这就把报表开发的问题拆解为细粒度的指标组合，不同的指标依赖的计算存储模型会根据业务的特性会有极大的不同。而 OceanBase 提供的丰富的分区方式及 OLAP 能力有效地解决了不同场景下，业务指标的构建问题，这对于我们业务开发工作者来说可以更多的关注我需要什么样的指标，而不用考虑如何从存储系统中得到这些数据。
2. 大数据处理能力：随着阿里巴巴集团业务的高速发展，推广营销在商业引流上的重要性越发明显，报表作为营销产品的闭环，其诉求也越发的多样化、个性化，报表数据在近几年的发展中在量级上已经增长到TB甚至数十 TB 的规模。这个时候存储系统的扩展性就显得非常重要，如果一开始我们就预估 5-10 年的存储资源，在前期数据规模不大的情况下，必然存在严重的资源浪费，如果前期预估得太少，随着数据增长，MySQL+ 中间件的集群扩容带来的数据搬迁问题又费时费力。同时，为了让用户获得良好的数据展示体验，我们要求每一次数据计算的时间不能太长（通常不超过 10s），而对于一些大数据的读写请求，如果不使用并行计算能力，是很难达到这个要求的。然而大数据的并行查询不能拖垮系统中的高优先级的小请求，并且当 MySQL 单表数据规模超过 2000 万时，其查询性能就出现断崖式的下跌，这也是业务无法容忍的一大缺陷，因此，我们在系统选型上更倾向于 OceanBase 这样具有高吞吐，数据读写隔离，资源隔离能力的存储方案。

3. 易用性：广告业务是一种典型的线上分析型业务（OLAP），需要在庞大的买家数据和广告数据中分析两者的关联关系，然后精准的分析出广告主的广告投放效果。因此，报表平台中存在着较多的多维度的数据关联查询，以及大数据的分组汇总查询，同时也存在一些统计学上的专业函数计算。而广告业务领域目前比较流行的 ROLAP、MOLAP 的分析型数据查询方案 SQL 能力都不够友好。因此我们需要基于其提供的 API 做很重的业务抽象，封装成一套业务通用的 SDK，因此我们不得不投入更多的开发和维护人员在这套笨重的 SDK上，开发效率将大打折扣，所以我们还需要一个对 SQL 语言支持良好的存储系统。
4. 系统成本：另一种解决方案就是采用大多数商业公司使用的 Oracle 提供的 RAC 解决方案，通过共享存储的能力提供数据存储空间的扩容，通过在共享存储上增加计算节点来提供高速的并行处理能力。这套方案都是基于在昂贵的硬件基础和 Oracle 数据库 License 费用上的，这不符合我们打造低成本技术体系的初衷。

优化结果

1. OceanBase 作为一个通用的分布式关系数据库系统，其提供了丰富的分区方式（HASH, RANGE, RANGE+HASH 等），并且提供在线的业务无感知的动态分区能力，集群扩容只需要 DBA 简单的增加存储节点，以及做一些简单的 DDL 操作即可，完全对业务透明，解决了我们业务数据爆炸式增长的问题。
2. OceanBase 兼容 MySQL5.6 版本大部分功能，完全覆盖报表业务的需求，报表业务可以像使用 MySQL 那样去使用 OceanBase，不需要业务做过多的逻辑改造，同时作为分布式关系数据库，还能够提供复杂的跨多结点的分布式 JOIN 能力，以及并行的汇总排序能力和丰富的数学计算函数能力，友好的满足了我们大多数场景的计算需求。同时，OceanBase 还为报表平台量身定制了近似计算的功能，对于一些超大结果集的运算，OceanBase 会筛选出一些精度影响较大的数据，然后基于这些数据进行汇总计算，在超大的数据计算的情况下，能够快速的得出一个离正确结果相差不大的近似结果。
3. OceanBase 作为一个可水平扩展的分布式关系数据库系统，在集群中，每个节点的角色关系都是对等的，每个节点都可以提供读写能力，大大提高了系统整体的吞吐能力，这也满足了我们需要迅速导入数据的诉求（TPS 峰值需要在 10 万以上）。同时，每个节点都可以部署在廉价的 PC 服务器上，因此，系统成本上的性价比是 RAC 解决方案的数十倍。