云数据库 OceanBase

产品简介

产品简介

什么是OceanBase

OceanBase是由蚂蚁金服、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。除在蚂蚁金服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、印度Paytm、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。

产品优势与应用场景

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。

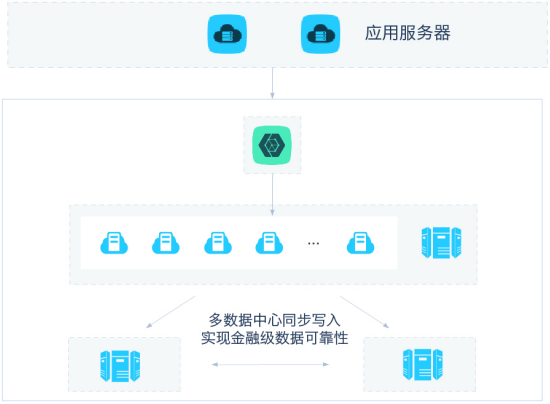
- **低成本**：使用PC服务器和低端SSD，高存储压缩率降低存储成本，高性能降低计算成本，多租户混部充分利用系统资源。
- **高可用**：数据采用多副本存储，少数副本故障不影响数据可用性。通过“三地五中心”部署实现城市级故障自动无损容灾。
- **强一致**：数据多副本通过paxos协议同步事务日志，多数派成功事务才能提交。缺省情况下读、写操作都在主副本进行，保证强一致。
- **可扩展**：集群节点全对等，每个节点都具备计算和存储能力，无单点瓶颈。可线性、在线扩展和收缩。
- **兼容性**：兼容常用MySQL/ORACLE功能及MySQL/ORACLE前后台协议，业务零修改或少量修改即可从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、PayTM等多个外部客户。OceanBase产品最适合于金融、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。在如下典型场景可以使用云数据库OceanBase。

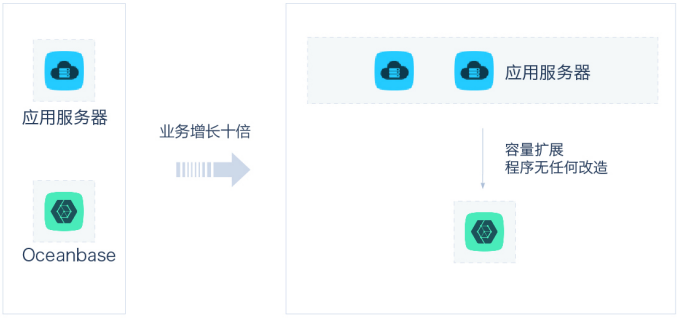
金融级数据可靠性需求

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。



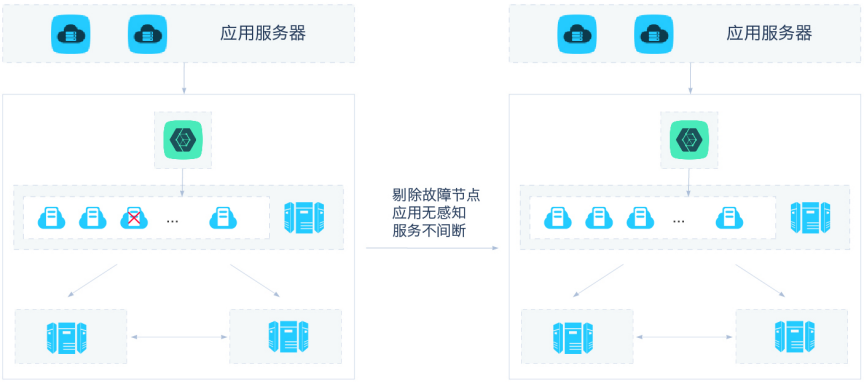
让数据库适应飞速增长的业务

业务的飞速成长，通常会成倍给数据库带来压力，OceanBase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。



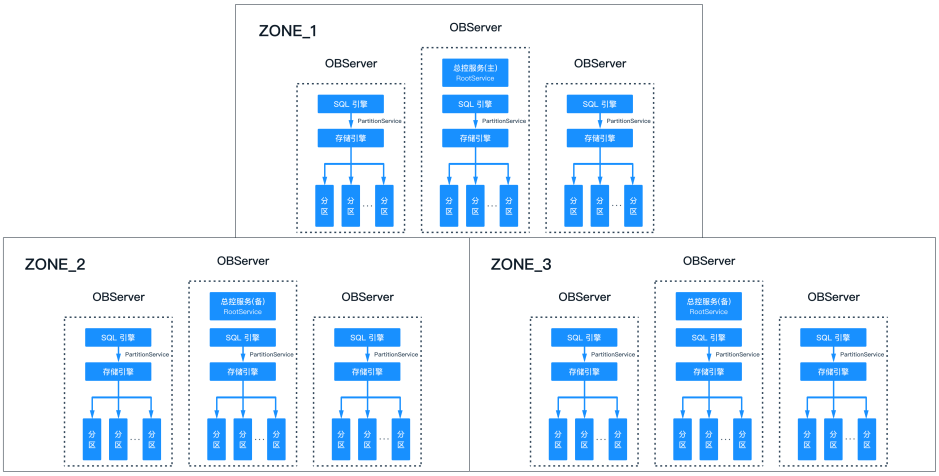
产品架构

OceanBase设计为一个Share-Nothing的架构，所以它是没有任何的共享存储结构的。至少需要部署三个以上的Zone，数据在每个Zone都存储一份。OceanBase的整个设计里面没有任何的单点，每个Zone有多个ObServer节点，这就从架构上解决了高可靠高可用的问题。

- 各个节点之间完全对等，各自有各自的SQL引擎和存储引擎。存储引擎只能访问本地数据，而SQL引擎可以访问到全局Schema，并生成分布式的查询计划。查询执行器可以访问各个节点的存储引擎，并在各个节点间做数据的分发和收集，完成分布式计划的执行，并把结果返回给用户。
- 其中一个节点还会额外担负起RootService服务，RootService同样会有多个备分布在各个Zone。主RootService和所有ObServer之间维持租约，当ObServer出现故障时，主RootService能够检测到并执行故障恢复操作。RootService是ObServer进程的一个功能模块，每台ObServer都具备

RootService功能。RootService的功能主要包括：服务器与Zone管理、分区管理、每日合并控制、系统自举、DDL操作等等。

OceanBase的架构图如下所示 ：



相关概念：

1. Zone（Availabilty Zone，区，可用区）

Zone是AvailabilityZone的简写。一个OceanBase集群，由若干个Zone组成。Zone的含义是可用区，通常指一个机房（数据中心，IDC）。为了数据的安全和高可用性，一般会把数据的多个副本分布在多个Zone上。这样，对于OceanBase来说，可以实现单个Zone的故障不影响数据库服务。一个Zone包括若干物理服务器。

2. OBServer（OceanBase服务器）

OBServer是一个OceanBase的服务进程，一般独占一台物理服务器。所以，通常也用OBServer指代其所在的物理机。在OceanBase内部，server由其IP地址和服务端口唯一标识。

3.表（Table）

最基本的数据库对象，OceanBase的表都是关系表。每个表由若干行记录组成，每一行有相同的预先定义的列。用户通过SQL语句对表进行增、删、查、改等操作。通常，表的若干列会组成一个主键，主键在整个表的数据集合内唯一。

4.分区（Partition）

分区是物理数据库设计技术，它的操作对象是表。实现分区的表，我们称之为分区表。表分布在多个分区上。当一个表很大的时候，可以水平拆分为若干个分区，每个分区包含表的若干行记录。根据行数据到分区的映射关系不同，分为hash分区，range分区（按范围），key分区等。每一个分区，还可以用不同的维度再分为若干分区，叫做二级分区。例如，交易记录表，按照用户ID分为若干hash分区，每个一级hash分区再按照交易时间分为若干二级range分区。

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。