

OCEANBASE

OceanBase 助力企业数字化转型

师文汇（虞舜）

OceanBase解决方案及产品总经理

蚂蚁集团数据库以及存储负责人



行业变革：千行百业的数字化转型，催生新的数据库架构与平台

政策：数字化政府驱动

提高数字政府建设水平，将数字技术广泛应用于政府管理服务，推动政府治理流程再造和模式优化，包括加强公共数据开放共享、推动政务信息化共建共用、提高数字化政务服务效能。

Source: 《国民经济和社会发展第十四个五年规划和2035年远景目标纲要》

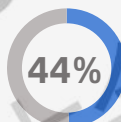
行业：数字化转型进入深水区

- 中国超过 **40%** 的企业已经成为数字化的支持者
- 中国Top1000大企业 **70%** 都把数字化转型作为战略核心
- 已经 **54.4%** 的企业在数字化转型的第3阶段， **14.9%** 达到第4阶段

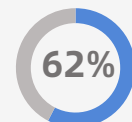
Source: IDC 2020

数字化转型，推动核心数据库升级

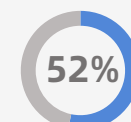
管理：行业客户在数据库使用中面临的新挑战



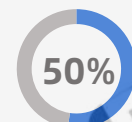
用不到：数据备份效率低，业务服务可用性降低



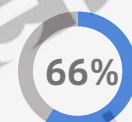
装不下：单机数据库已经不能满足海量数据规模



放不了：高并发的场景下，现有数据库性能已经很难支撑



难管理：数据库运维变得更加复杂和困难

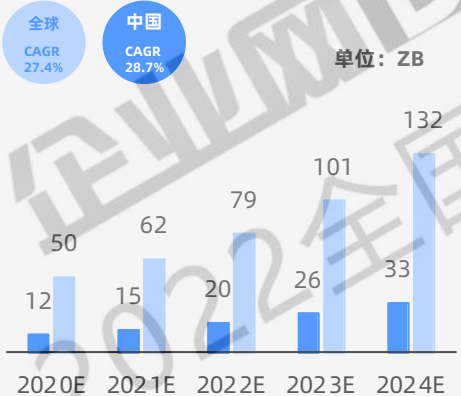


高风险：随着联网的系统增多，数据库安全风险提高

数据库作为承载企业业务与发展的核心信息化平台，需要兼顾**成本、性能、发展效率、安全**的所有新时代的需求

技术演进：数据的海量增长，业务复杂度增加，分布式数据库成为核心解决方案

数据量迎来爆发式增长



各行业客户数据管理需求日益复杂



新一代分布式数据库成为解决方案



传统主流数据库并非为应对今天的挑战而生

各行业面临一致痛点

- 性能瓶颈**：业务驱动数据规模增长，但传统数据库面对数据量增长难以维持性能。
- 分析能力不足**：传统方案需要构建独立AP系统，数据的传输、计算的时效性、单独建设分析系统。
- 成本高昂**：系统需按最大容量设计，不支持灵活水平扩展难，硬件投入大。集中式架构下，实现高可用、可靠性需要付出高额的成本代价。



企业数字化转型过程中，数据库选型的关键决策点

数据是否安全可靠

- 数据一致性
- 数据安全性
- 数据可靠性
- 国产化硬件适配

业务扩展性

- 业务增长是否会带来二次迁移和改造
- 当前选型是否足够弹性

ROI & TCO

- 利旧复用降低投入
- 极简运维
- 极致性价比带来的低TCO和高ROI

迁移改造复杂度

- 业务系统的改造时间和成本
- 已有系统和数据的迁移
- 迁移和切换是否影响业务

一库多芯

强一致性校验

全链路加密

企业级权限

四层安全防护

高压缩比

统一运维

集群多租户

HTAP

行列混存

OCEANBASE

原生分布式HTAP数据库

弹性扩展

无硬件绑定

多云部署

无感扩容

负载均衡

高度兼容
近乎0改造

全生命周期
管理工具

双向同步
平滑迁移

分布式HTAP数据库：满足企业数字化转型数据库平台

OceanBase 发展历程及构架演进



OceanBase 产品体系：全数据生命周期、全业务场景数据解决方案



四层安全防护、六大高可用技术，端到端保障客户数据



示例：磁盘静默错误一旦发生 在元数据中，将造成不可挽回的损失

丢数据案例

2018年国内某企业数据全部丢失原因:磁盘静默错误, 如果发生在涉及自己所在的企业, 后果难以想象!



挑战已客观存在

近些年, 固件门频发, 一旦出现, 影响重大

2009年希*固件门, 30款硬盘卡死

2. 2012年镁*固件门, 5200小时丢盘蓝屏门

3. 2020年惠*-*星, 32746小时数据全丢

4. 2021年西*降速门, 5GB降到3GB

- 经典的磁盘冗余方案, 对此无效
- 传统数据库, 对冷数据静默错误, 无效

示例:磁盘静默错误--OB支持坏块发现(热数据和冷数据)

- 1.模拟磁盘静默错误: dd if=/dev/zero of=/dev/vdc2
- 2.数据库主动检测, 若数据量大, 可配置每天检测1/N



svr_port	disk_id	store_file_path	m...	error_...	err...	check_time
2882	0	/home/admin/oceanbase/store/...	120	-4002	Bad data ...	2022-05-06 18:51:01.273577
2882	0	/home/admin/oceanbase/store/...	412	-4002	Bad data ...	2022-05-06 18:51:32.394198
2882	0	/home/admin/oceanbase/store/...	417	-4002	Bad data ...	2022-05-06 18:52:03.527675
2882	0	/home/admin/oceanbase/store/...	418	-4002	Bad data ...	2022-05-06 18:52:05.536986
2882	0	/home/admin/oceanbase/store/...	422	-4002	Bad data ...	2022-05-06 18:52:38.703522
2882	0	/home/admin/oceanbase/store/...	443	-4002	Bad data ...	2022-05-06 18:53:21.935361
2882	0	/home/admin/oceanbase/store/...	452	-4002	Bad data ...	2022-05-06 18:53:41.038521
2882	0	/home/admin/oceanbase/store/...	482	-4002	Bad data ...	2022-05-06 18:55:40.706822
2882	0	/home/admin/oceanbase/store/...	489	-4002	Bad data ...	2022-05-06 18:55:52.783175
2882	0	/home/admin/oceanbase/store/...	498	-4002	Bad data ...	2022-05-06 18:56:08.88008
2882	0	/home/admin/oceanbase/store/...	191	-4002	Bad data ...	2022-05-06 18:56:43.10563

示例：数据一致性——看得见(多副本之间、主备之间，数据自动校验)

检测分区副本的 checksum 是否相同来判断数据是否一致。

有两种方法，方法如下：

1、查看所有版本的数据一致性情况（存储在宏块上有表内容的表分区，数据版本包含多个）

```
select sum(s1*yzx) 数据一致分区数,sum(s1*abs(yzx-1)) 数据不一致分区数
from
(
select a.data_checksum,a.s1,
case when mod(s1,3)=0 then 1 else 0 end yzx
from
(select data_checksum,count(*) s1 from __all_virtual_partition_sstable_macro_info group by table_id,partition_id
)a
)
```

多版本数据一致分...	多版本数据不一致...
3690	0

2、查看所有表分区的数据一致性情况（包含无数据的所有分区，数据版本只有当前最新版本）

```
select sum(s1*yzx) 数据一致分区数,sum(s1*abs(yzx-1)) 数据不一致分区数
from
(
select a.report_data_checksum,a.s1,
case when mod(s1,3)=0 then 1 else 0 end yzx
from
(select report_data_checksum,count(*) s1 from __all_virtual_partition_store_info
)a
)group by report_data_checksum
)a
)
```

数据一致分区数	数据不一致分区数
5925	0

HTAP混合负载，一套系统完成TP+AP业务

一套系统海量交易+海量分析

OLTP+OLAP混合需求是TOP行业强诉求，OceanBase 基于分布式架构，向无限可扩展的Oracle 发展，做好交易处理场景的同时，完成分析、跑批等分析性场景

传统方式：高处理负载

HTAP引擎：混合需求完成

Step 1 OLTP 请求

OLTP 请求 + OLAP 请求

Step 2 OLAP 请求

OceanBase集群

混合负载引擎

复杂查询优化

- 自动计划演进、免手动调优

线性扩展的实时OLAP处理能力

- 水平线性扩展（千亿级数据关联查询）、秒级低时延响应



TP与AP一体化引擎

- 可同时处理TP与AP查询、数据插入立即可见

集群级并发资源管理

- 优化资源分配与流量控制、选择策略灵活

混合负载 + 数据库（DBaaS）服务实现低成本和简化运维

交易数据系统

分析数据系统

OLTP工作负载

OLAP工作负载

HTAP引擎

成本降低，简化运维

一套引擎支持 OLAP + OLTP 工作负载，同时实现两套系统功能，实现成本和复杂度大幅降低

行列混存、智能压缩，节省5-10倍客户存储成本

SSTable 由若干个宏块(Macro Block)构成, 宏块2M固定大小的长度不可更改。

在宏块内部数据被组织为多个大小为 16KB 左右的变长数据块, 称之为微块(Micro Block), 微块中包含若干数据行(Row), 微块是数据文件读 IO 的最小单位。每个数据微块在构建时都会根据用户指定的压缩算法进行压缩, 因此宏块上存储的实际是压缩后的数据微块。



行列混合的存储结构，为HTAP引擎提速

C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	C2	C3	C4	C5	C6	C7	C8	C9

平铺模式的微块 (压缩前)

C1	C1	C1	C1	C1	C1	C1	C2	C2
C2	C2	C2	C2	C2	C3	C3	C3	C3
C3	C3	C3	C4	C4	C4	C4	C4	C4
C4	C5	C5	C5	C5	C5	C5	C5	C6
C6	C6	C6	C6	C6	C6	C7	C7	C7
C7	C7	C7	C7	C8	C8	C8	C8	C8
C8	C8	C9	C9	C9	C9	C9	C9	C9

编码模式的微块 (压缩前)

SSTable由多个定长(2 MB)的宏块组成, 而每个宏块由多个变长的微块组成。微块是读IO的最小单位。

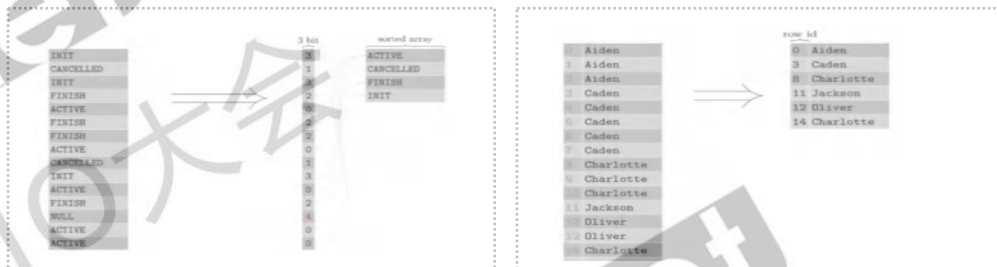
OceanBase 微块有两种存储方式, **平铺模式(Flat)** 或 **编码模式(Encoding)**。

平铺模式 即为通常意义上的行存形态, 微块内数据以行的顺序连续存储。

编码模式 在微块里仍旧存储完整的行数据, 但存储组织方式则按照列的形式。OceanBase 根据数据格式与语义选择多种编码手段, 以达到最佳的压缩效果。同时, 在执行过程中, 数据以列的组织方式载入内存提供给向量化引擎, 进一步提升HTAP处理性能。

高压压缩率背后的秘密：自适应编码压缩技术

OceanBase 数据库提供了多种按列进行压缩的编码格式, 根据实际数据定义进行选择, 包括列存数据库中常见的字典编码, 游程编码 (Run-Length Encoding), 整形差值编码 (Delta Encoding), 常量编码, 字符串前缀编码、Hex编码、列间等值编码、列间子串编码等等。几个典型编码模式:

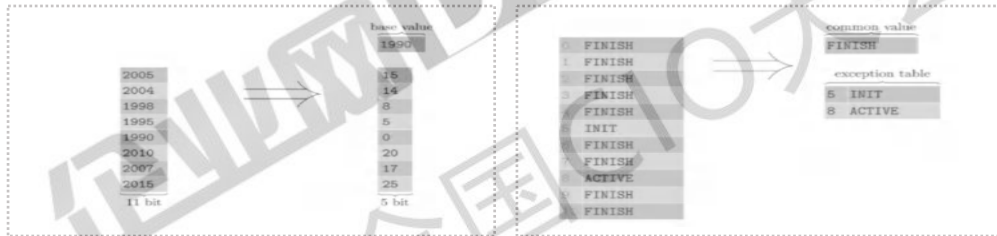


字典编码

当微块内的数据的基数(Cardinality)比较小时, 字典编码能通过微块内部构建字典, 存储每行的引用来进行压缩。

RLE编码

对于连续相等的数据, 只保留其起始行号和值。这种编码在数据库中通常用于处理有序数据, 例如: 索引前缀、后缀等。



差值编码

通过只存储每一行的值与微块中最小值的差值, 然后做 bit-packing 来减少实际存储的数据量。适用于存储的列是一个定长的数值, 且这个微块中的数据都分布在一个值域内。

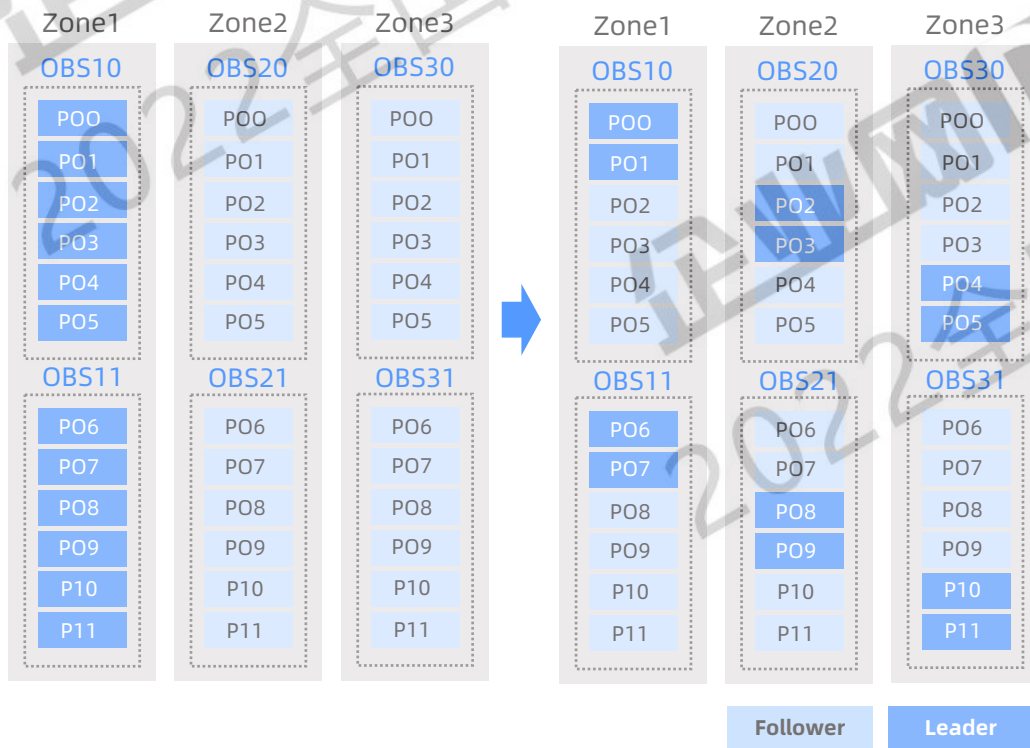
常量编码

一个微块内的一列可能基本都是相同的数据, 这时 OceanBase 数据库会通过常量编码(Const) 只存储常量和微块内不等于常量的值。

透明的扩展能力，从原型到超大规模

根据负载和容量自动均衡，发挥整个集群的最大效能

OceanBase 数据库通过 RootService 管理租户内各个资源单元间的负载均衡。不同类型的副本对资源的需求各不相同，RootService 在执行分区管理操作时需要考虑的因素包括每个资源单元的 CPU、磁盘使用量、内存使用量以及 IOPS 使用情况。经过负载均衡，最终会使得所有机器的各类型资源占用都处于一种比较均衡的状态，充分利用每台机器的所有资源。



自动负载均衡达到两个目标：



数据均衡

减少租户内数据副本在各个节点上数量和大小差值。



负载均衡

将数据副本的Leader平均调度到 Primary Zone 的全部节点。

分布式对应用透明，整个集群对外提供统一的数据库服务

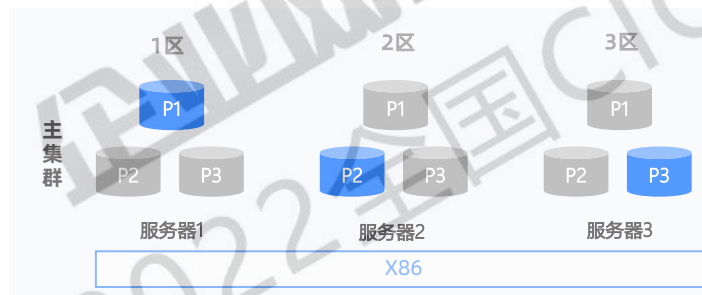


不同于其他分布式方案，OceanBase 每个节点均支持读写，且对外提供统一的数据库服务。无论节点内部资源或者租户资源如何变更，对业务应用没有影响。

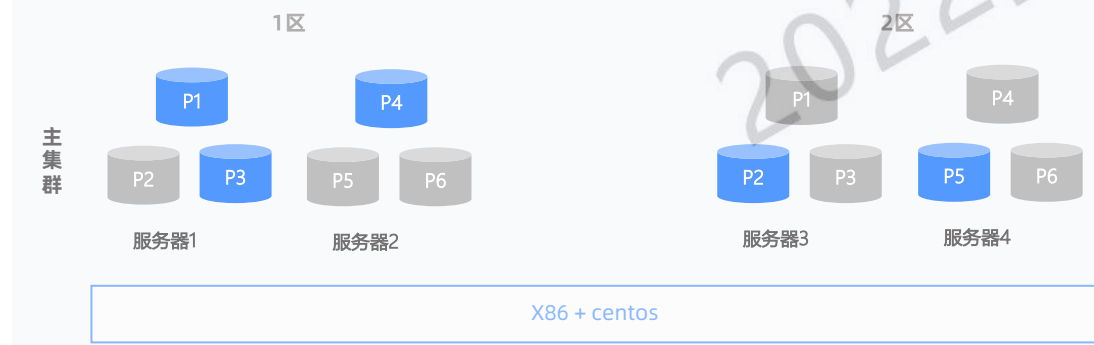
根据实际负载情况或者对未来负载的预测，管理员可以任意对租户使用的计算资源进行调整，实现平滑的扩缩容。

一库多芯，混部灰度：国产芯片灰度替换的解决方案

第一阶段：ARM芯验证



第二阶段：灰度/混合部署



方案特点

- **兼容开放**：支撑海光、鲲鹏、Intel等多种芯片及其生态，满足不同的用户和场景需求
- **功能一致**：数据库自动适配，用户不用区分底层的芯片形态，透明无感
- **数据一致**：支持异构芯片副本数据一致性校验，确保数据一致性和正确性，具备容灾切换能力
- **混合部署**：支持多副本混部、主备集群混部、机房混部等多种形态，支持长期混部运行
- **灰度切换**：支持按区（zone）切换，最细粒度支持按表分区灰度切换，支持平滑灰度迁移替换

经过金融级生产环境验证的原生分布式数据库扩展能力

以下所有数据来自于实际生产系统

6100

万次/秒

数据库峰值处理能力

>1000

台

集群节点数

>2

PB

单库存储容量

>3200

亿行

单表行数

RPO=0,
RTO<30

秒

少数副本故障时

生态发展：已与 40+ 产业上下游伙伴达成深度合作，推动产业生态建设

40+
产业端优秀伙伴深度合作

1w+
OceanBase 认证工程师

12家
技术/服务类伙伴

千万级
投入高校人才建设

Logo grid containing 40+ partner logos including: 恒生, DCITS 神州信息, 宇信科技, HOPAVUN 润和软件, Sunline 长亮科技, 同方软件, 易诚互动, DTEC 先进数通, 赞同科技, 中电金信 GienTech, 中科软科技 Sinosoft Co., Ltd, marsoft, Neusoft, eBaoTech 让保险更简单, 新大陆 Newland, 浩鲸科技, inspur 浪潮, 万达信息, 神玥软件 SHINEYUE SOFT, Dareway 山大软件, CS&S 中软国际, Chinasoft International, 华锐金融技术 ArhForce Financial Technology, THUNISOFT 华宇, 大汉软件, 中信网络科技股份有限公司 CITIC Application Service Provider Co., Ltd., 四川久远银海软件股份有限公司 SICHUAN JIUYUAN YINHAİ SOFTWARE CO., LTD., 用友 yonyou, Yondervision 华信永道, 新炬网络 SNC Net, 云和慧星 CHMC-ITTE, Apusic 金蝶天燕, 神州数码 Digital China, 网数科技 Wuhu IT-Group, 英方软件 infomobiz, HYGON 中科曙光, 统信软件 UNIONTECH, KYLINSOFT 麒麟软件, Kunpeng, Phytium 飞腾科技, DSG, 网数科技 WEIDATECH, 派喜动力 PockData, 用友 Tongfang, 中鼎科技 CHINDING SERVICE, OLM 联合易片



从金融走向全行业和全球，服务超过400家机构

银行



保险



证券



全球千行百业



面向不同客户核心业务系统OceanBase已经可以从容面对



大型客户

超大型企业、头部互联网

通过蚂蚁集团内部金融场景十年的积累，OceanBase所拥有的分布式架构、强一致性的事务处理，透明扩展能力、充分证明作为超大型企业核心系统的承载能力，已有超过100+金融、运营商等客户选择OceanBase



中型客户

中小企业、成长型互联网

商业化一年来，OceanBase在兼容性、性价比、高可靠、小型化等能力持续增强，满足了中型客户在TCO、数据主权等方面的核心诉求，越来越多的客户选择与OceanBase一起成长



小微客户

初创企业、个人电商、开发者

OceanBase Cloud云数据库为更多的小微客户提供成本低廉、架构先进、性能优越、免运维的数据库服务和产品，社区开源也吸引更多的数据库开发者加入OceanBase

多种部署形态

独立软件

私有云

混合云

公有云

多种产品形式

OCEANBASE 商业版

OCEANBASE 社区版

中石化新一代大集中加油卡系统实现一张油卡通全国

OceanBase 赋能 “数据+平台+应用” 的架构设计理念，建设大集中的实体卡系统

业务挑战

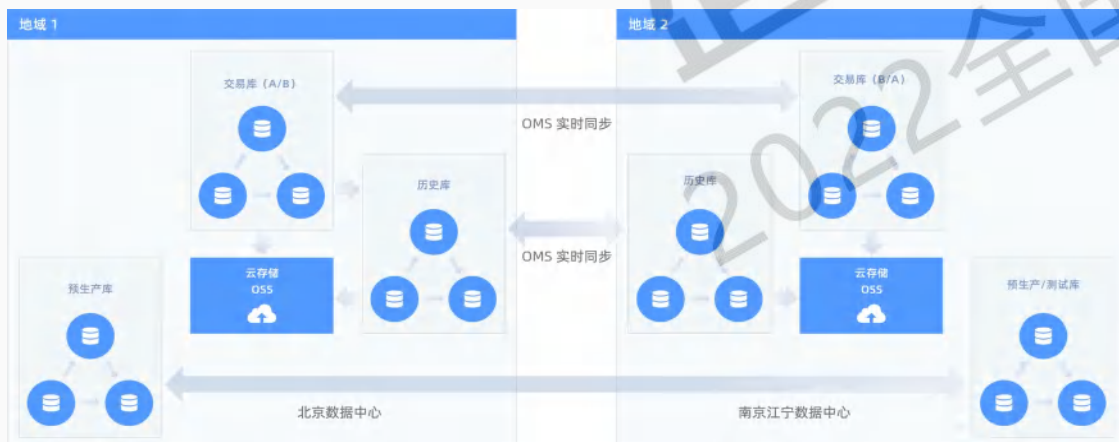
- 中石化基于新基建技术构建新一代智慧加油站，推进中石化生活综合服务商转型战略
- 原有加油卡系统无法适应互联网化客户营销服务体验和模式创新需求
- 异构分散式系统，无法满足业务转型所需低系统性风险、管理运维和自主创新。

解决方案

- 整合23套Sybase和Oracle数据库，基于原生分布式数据库架构实现省级和跨省分布式交易
- 分区、读写分离、异地双活等技术实现负载均衡和OLAP查询效果提升，LSM等技术提升OLTP事务效率，支持互联网化应用类型负载，体现强大的HTAP混合负载支持
- Oracle和Sybase数据库应用无损迁移，过渡方案确保柔性切割。
- Paxos /分区/OMS 等技术实现异地双中心容灾和高可用能力。

业务收益

- 外部支持全国近3万加油站的加油卡业务
- 内部支持交易流水由天级降低到秒级，实现一体化班日结和报表需求
- 电子券、返利实时化，单一支付方式向多种支付方式转变
- 23套分散系统运维降低至1套，8倍存储成本节约
- 数据查询时间由分钟级降低到秒级，支持每分钟5000笔业务交易
- 故障恢复时间从小时级降低到分钟级，业务连续性达到99.99%
- 安全级别达到等保2.0版要求、实现3级安全防护



8倍
节约存储成本

99.99%
业务连续性

5000笔
每分钟交易量

山东移动核心计费系统业务迁移实现大幅降本增效

OceanBase 高兼容性和 OMS 迁移服务保障了多个核心业务系统平滑迁移至 OceanBase

业务挑战

- 山东移动按用户规模在省级运营商中排名第二，计费系统是山东移动的核心业务系统，日处理各类详单数据130亿条，数据处理性能和准确性至关重要。过去使用的集中式单机数据库，面对互联网和5G时代不断激增的用户数和并发量，经常出现容量不足的情况。
- 应用迁移需要在复杂的业务逻辑中梳理Oracle数据库对象进行适配，需要一款产品可以自动做评估、转换、并支持在线搬库。

解决方案

- 整合多套分散系统，基于Paxos 协议和分区等技术，多机房部署实现高可用和容灾。
- 性能无损的数据高压缩比，分区、读写分离、LSM等技术提升OLTP事务效率
- 一站式数据库无损迁移，过渡方案确保柔性切割。



业务收益

- 详单处理率提升30%，存储成本降低90%，硬件和维保成本大幅降低。入选工信部“2020年网络安全技术应用试点示范项目”名单。
- 三数据中心的OceanBase集群部署，组成了跨越多数数据中心的分布式数据库，实现RPO=0的机房级别容灾能力，不再需要搭建灾备系统。
- 源系统数据类型、对象、存储过程仅少量修改达成应用适配，1小时完成应用切割，实现应用系统平滑迁移。
- 使用普通PC服务器，下线小机+集中式存储等传统架构降低硬件成本。

致欧家居实现跨境电商平台多云融合

OceanBase 集群汇聚MySQL 业务，简化运维管并提高安全稳定性，多云数据融合实现业务稳定线性增长

业务挑战

- 业务流量上升,难支撑高并发,单个实例扩容至瓶颈后,已无非垂直扩展
- 数据过于分散,聚合分析难,业务平台分布全球,数据需要融合汇聚,以做分析决策,急需一个能够跨多种云平台对数据时行融合汇聚的技术栈
- 全球业务站点增加,可用可靠要求增强
- 技术需自主可控 数据库种类多,实例数不断增加,运维复杂度增加,成本节节攀升,使用“OceanBase多云”解决方案,解决客户混合云场景下OceanBase统一管控的问题;

解决方案

- 通过OceanBase透明可扩展能力,无需业务分库分表,通过横向增加节点即可实现业务线性的性能拓展能力。
- 借助OB 产品家簇OMS 对欧洲aws, 北美aws, 日本aws , 数据进行融合汇聚



客户收益

- 基于OceanBase对MySQL的深度兼容,把现有业务系统迁移至OceanBase后,借助集群功能,把多个MySQL及RDS实例数据整合到一个集群,数据压缩比提高,极大的简化了运维,降低了成本
- OceanBase的水平扩展能力,为站点后续的快速增长提供弹性伸缩能
- 借助OceanBase产品家簇OMS 对欧洲aws, 北美aws, 日本aws , 数据进行融合汇聚
- 依托OceanBase支持 HTAP功能,简化实时报表数据同步困扰,简化分析数据流迁移

多云融合

可用可靠

弹性伸缩

理想汽车实现数据库的“自动驾驶”

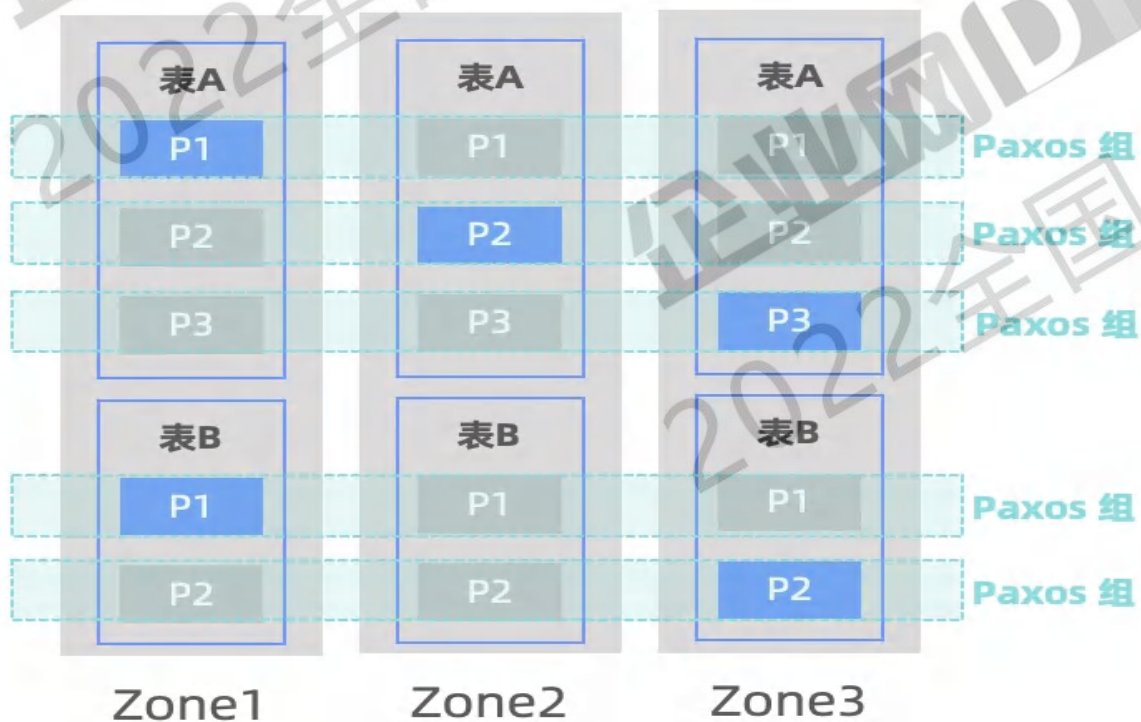


业务挑战

在科技制造业，生产流水线的平稳高效运转是企业的生命线。其中核心的产线执行系统如果出现故障，将直接导致停产，每一秒都意味着供应链和人力、资源、机会成本的巨大损失，甚至可以掀起蝴蝶效应。

解决方案

- OceanBase 通过基于 Paxos 协议实现了数据库服务“故障自动恢复”和“数据零丢失”，并且即使在网络条件复杂的情况下依然保持稳定的性能和可用性。
- OceanBase 智能运维体系，围绕监控、诊断、报告三个维度，为理想提供了功能丰富、简单易用的运维工具。



客户收益

- OceanBase 智能运维体系的护航下，异常 SQL 的诊断实时自动分析，DBA 在关键时刻只需看一眼可疑 SQL 列表，就能快速判断问题根因，并且获得合理的应急优化建议。
- 通过OCP将专家经验直接内嵌到每层监控中，从表层的响应 RT 一直下钻到单个物理节点的基础指标，用户只需在每层监控中点击关联的可疑指标，就能将问题层层定位。
- 迁移至 OceanBase 后，理想的产线执行系统数据库抖动频率平均下降约 80%，对于常见的故障事件真正做到“先恢复，后分析”。

面向未来-坚持“复杂留给数据库，简单留给客户”

专注做好数据库功能，帮应用屏蔽非功能复杂度

OceanBase原生分布式数据库



企业网DINet

相信

坚持的力量

攀登海量技术高峰



践行

长期主义的理念

突破海量数据管理难题



坚定

扎根行业的决心

深耕海量核心场景



OCEANBASE | 海量记录 笔笔算数

2022全国CIO大会

企业网DINet

2022全国CIO大会

企业网DINet

2022全国CIO大会