DTCC

第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

架构革新 高效可控

IT168.com ChinaUnix.net ITPUB

北京国际会议中心 | 2020/12/21-12/23

OceanBase 混合负载计算引擎与兼容性产品演进

高斌（艾伦）

蚂蚁集团 OceanBase 高级产品专家

用户对数据库的诉求

混合负载

- 按需平滑扩展存储与算力；
- 满足海量数据的处理需求；
- 一套产品解决绝大多数场景的负载；

自主研发

- 自主研发，降低安全风险；
- 适配国产化硬件和国产芯片；

弹性

- 快速部署，降低上线周期；
- 按需供应，避免资源浪费；
- 动态伸缩，满足市场运营；

安全可靠

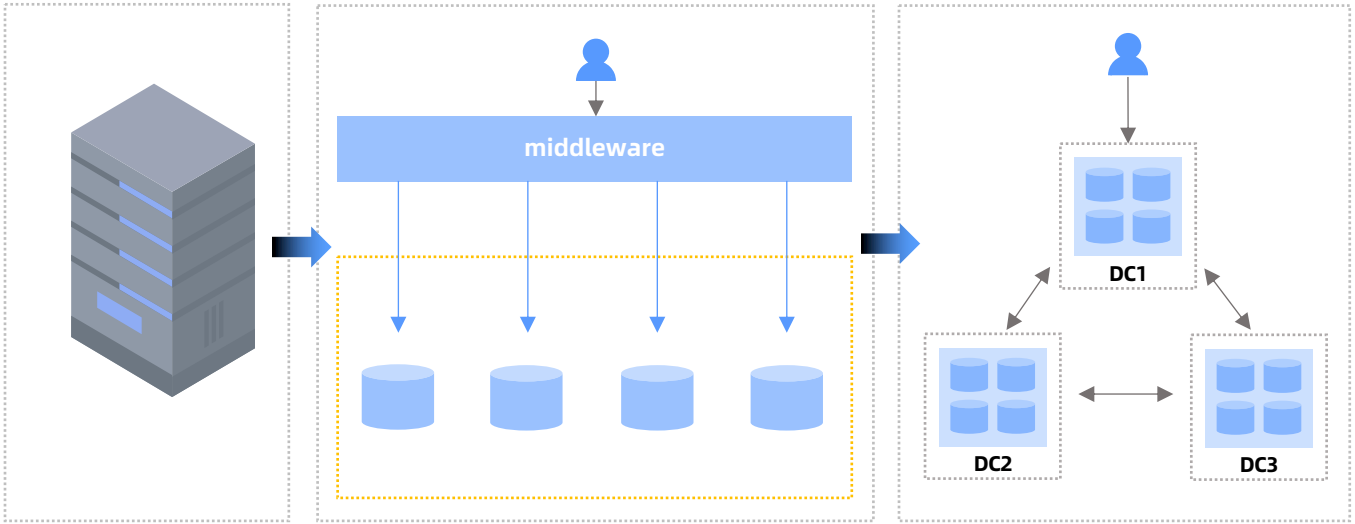
- 可靠的高可用和容灾方案；
- 完备的数据备份恢复解决方案
- 完备而可靠的数据安全保障

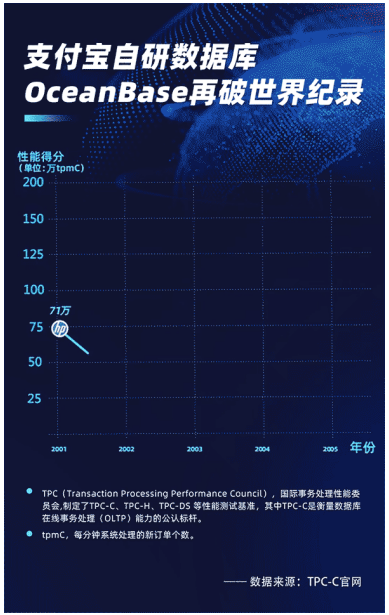
云服务

- 支持多租户部署；
- 租户间资源隔离；
- 租户管理便捷可靠；
- 即有应用能够平滑迁移；
- 自动化运维手段和告警设置；
- 完备的管理流程和最佳实践；

原生分布式数据库开启下一个时代

原生分布式数据库 —— 从开发阶段、运行阶段、运维阶段屏蔽复杂度，把简单留给应用设计、开发者和运维人员，把复杂留给基础设施





数据库兼容的几个层面

- 驱动

  OBCI、JDBC、ECOB

- 数据

  基础数据类型/表模式

- 功能

  SQL语法、语义、函数

  过程性语言（PL）

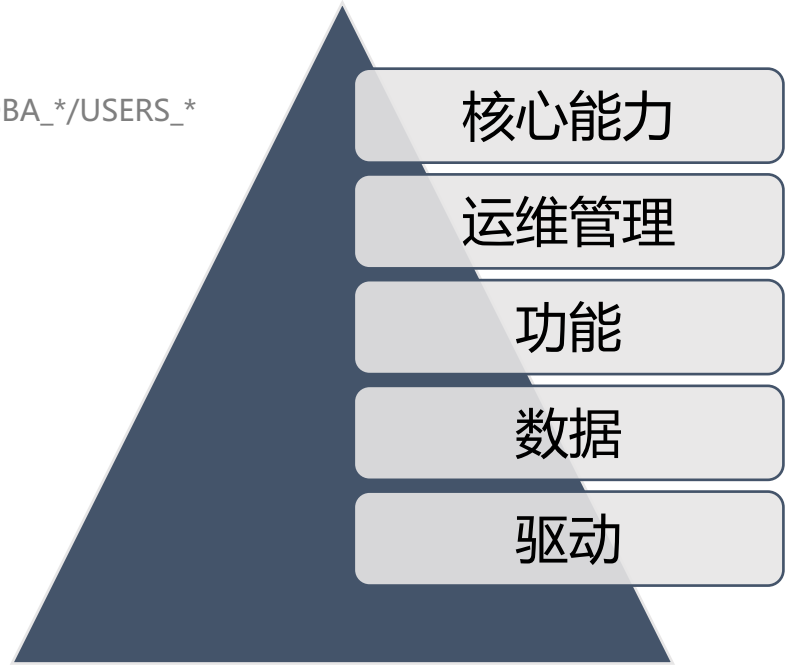
- 运维管理

  视图：ALL_*/DBA_*/USERS_*

- 核心能力

  性能、稳定性

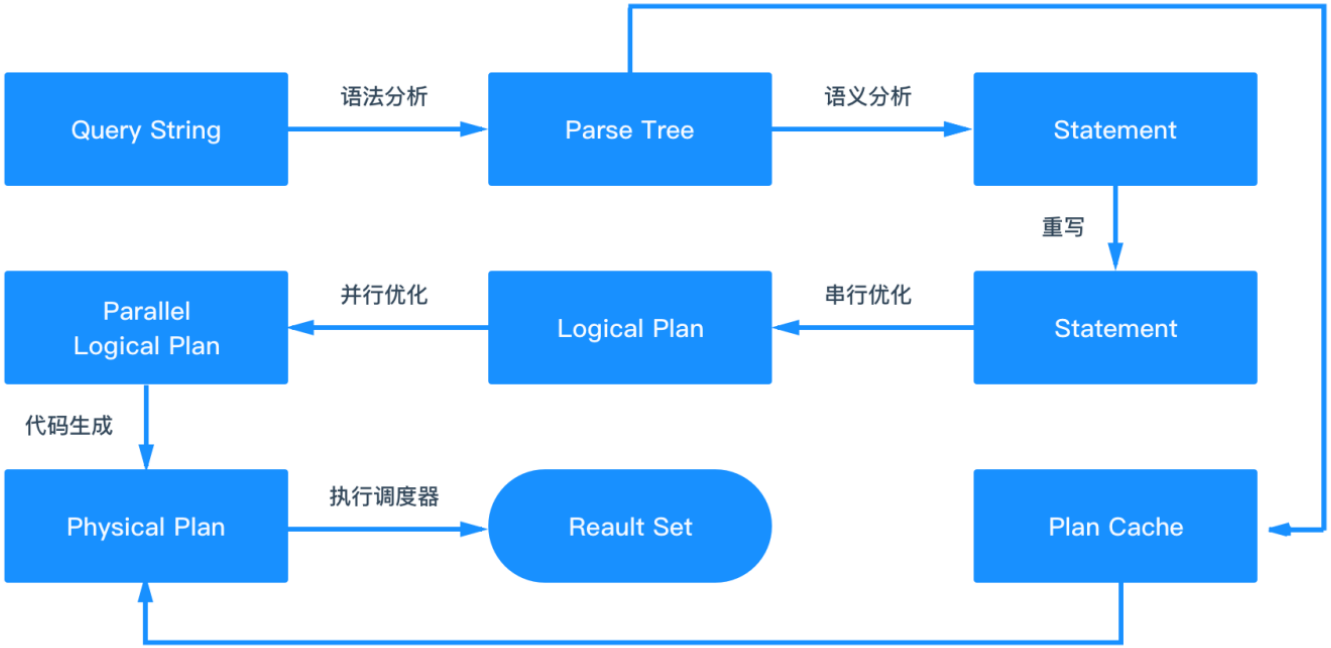
  -持续迭代



兼容性功能

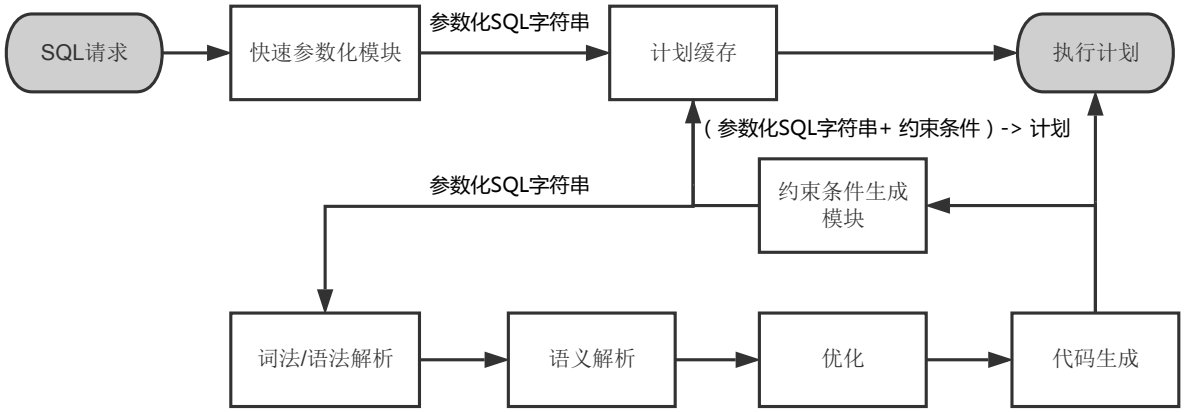
- **数据类型**
  number、varchar2、date、LOB、UDF
- **索引**
  本地索引、全局索引、函数索引
- **客户端**
  JDBC、ODBC、 OBCI、 ECOB
- **语法&函数**
  外连接(+)、层次查询、临时表、外键、集合运算、数据类型转换函数、窗口函数、分析函数等
- **分区**
  hash、range、list分区，模版化/非模版化二级分区，无主键表分区
- **事务层**
  多种隔离级别、flashback、XA
- **伪列**
  rownum、rowid，nextval, currval, level
- **内部视图**
  大部分All_/DBA_/USER_视图，gv$视图
- **PL**
  几乎全量的PL语法，大部分系统包

混合负载计算引擎（一套引擎解决多种负载）



混合负载计算引擎—计划缓存

- 跳过传统意义SQL的“硬解析”过程
- 参数化过程不涉及语义解析
- 基于约束条件的匹配，避免将语义不同的常量误识别为参数



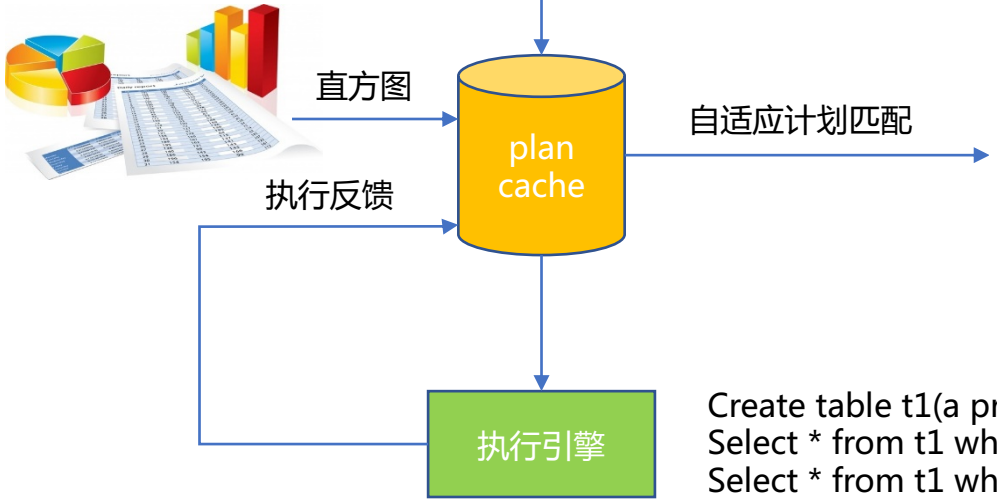

```
select c1, c2, c3
from t1
where c1 = 1 and c2 like
  'senior%'  and 1 = 1
order by 3 limit 1;
```

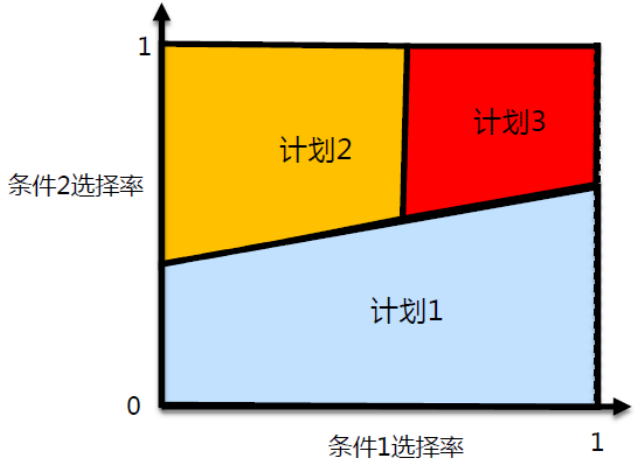
参数化

```
select c1, c2, c3
from t1
where c1 = @1 and c2 like @2
and @3 = @4
order by @5 limit @6;
```

约束：
@3 == @4
&& @5 == 3

混合负载计算引擎—计划缓存（ACS）

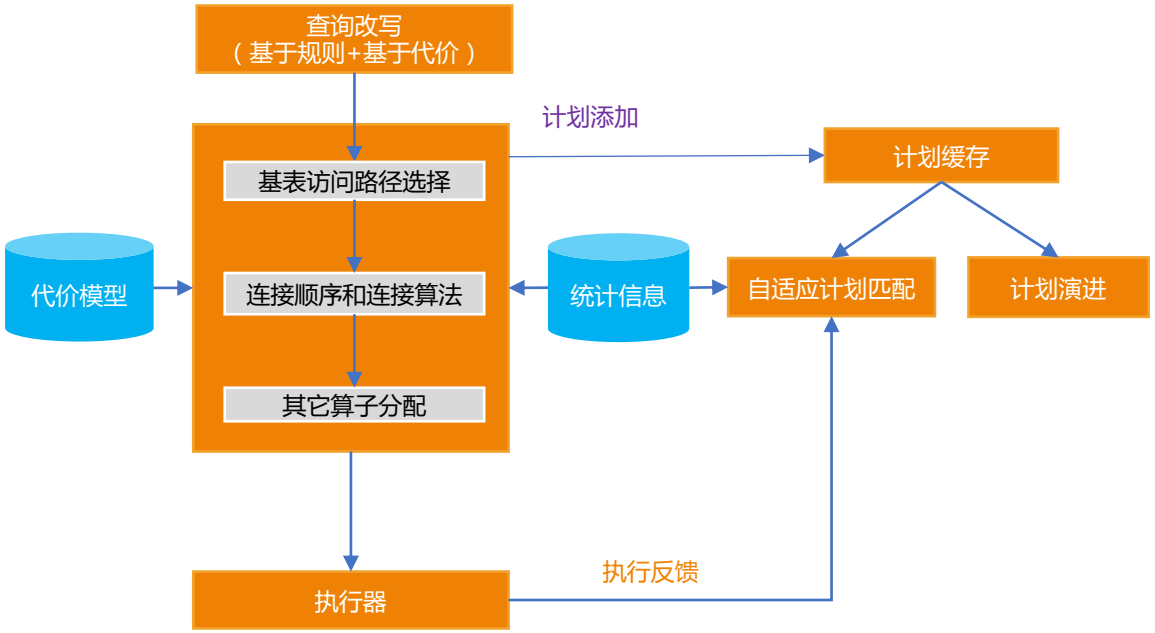




Create table t1(a primary key, b int, index k1(b));
Select * from t1 where b = 1; —— sel(b=1) = 1%， 选择使用索引K1扫描
Select * from t1 where b = 2; —— sel(b=2) = 50%， 选择使用主键a扫描

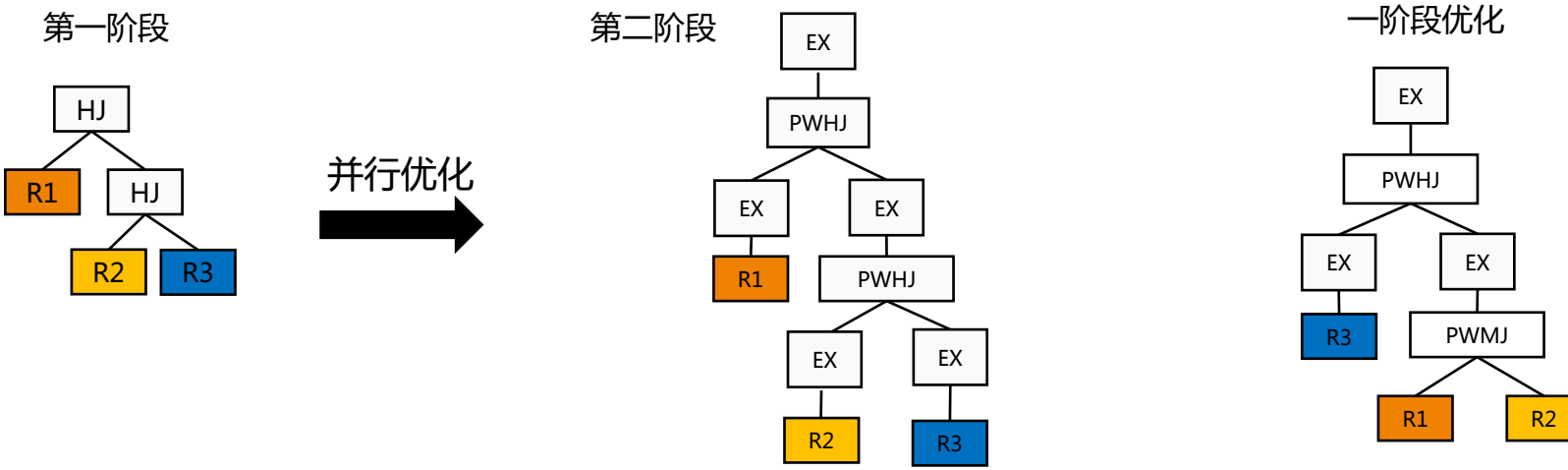
混合负载计算引擎—优化器

- 提前进行基于规则的改写
- 小查询倾向索引的选择
- 逻辑优化与物理优化分离
  - 引入大量复杂改写逻辑增强OLAP场景处理能力
  - 降低实现复杂度和优化开销
- 两阶段演进到一阶段



混合负载计算引擎—优化器

分布式查询优化策略—— 两阶段 v.s.一阶段



两阶段优化：

阶段一：根据表数据量、选择率等信息决定最佳的连接顺序
阶段二：针对数据分布生成必要数据重分布节点，生成分布式计划

一阶段优化：

整体考虑数据分布、数据量、选择率等信息直接生成分布式计划

混合负载计算引擎—执行引擎

	OLTP	OLAP
性能关键点	关键路径、计划open & close、数据结构复用	行级迭代、调用层次、局部性
分布式	RPC开销、移动数据、串行	面向流水线、并行执行
存储格式	行存	列存
算法依赖	不敏感	高度敏感

混合负载计算引擎—执行引擎演进

一期（强类型、内存预分配）

强数据类型运算

静态内存预分配

二期（向量化、批处理）

向量化处理逻辑

多行批处理

Cacheline友好

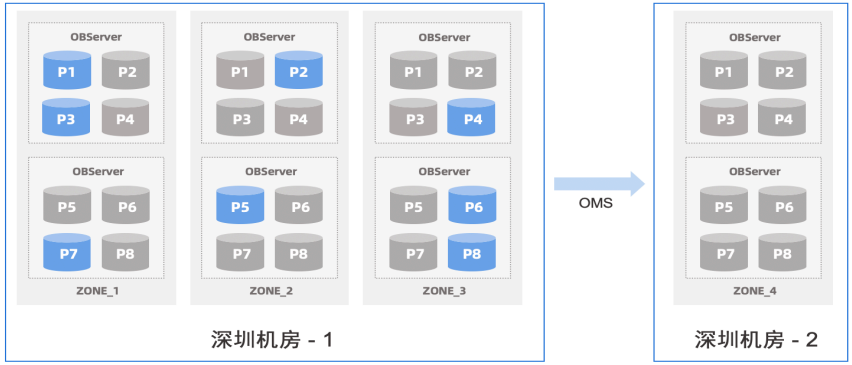
重点算子优化

三期（面向列存）

面向列存数据的算法重构

基于数据编码的运算

某大型能源行业客户加油卡项目



客户收益：

业务收益

- 外部支持近3万加油站，2.8亿会员和千万级APP日活业务负载
- 内部支持交易流水上传时间由天级降低到秒级，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，适应互联网化体验转型

信息化收益

- 23套分散系统运维降低至1套整合系统运维管理需求，8倍存储成本节约
- 3个月内完成Oracle和Sybase应用软件无损迁移
- 数据查询时间由分钟级降低到秒级，并发处理能力达到每秒5万笔
- 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%，降低系统性风险
- 安全级别达到等保2.0版要求、实现3级安全防护，符合信创要求

解决方案：

全自研原生分布式数据库方案

- 整合23套分散系统，基于Paxos 协议和分区等技术透明支持省级和跨省分布式交易
- 符合信创要求

强大的HTAP混合负载支持

- 分区、全局索引等技术实现负载均衡和OLAP查询效果提升
- LSM-Tree 存储引擎提升OLTP事务效率，支持互联网化应用类型负载

完善的应用迁移和运维

- Oracle和Sybase数据库应用无损迁移
- Paxos /分区/OMS 等技术实现异地双中心容灾和高可用能力
- 即有应用能够平滑迁移；
- 自动化运维手段和告警设置；
- 完备的管理流程和最佳实践；

某市公积金业务在线办理系统项目

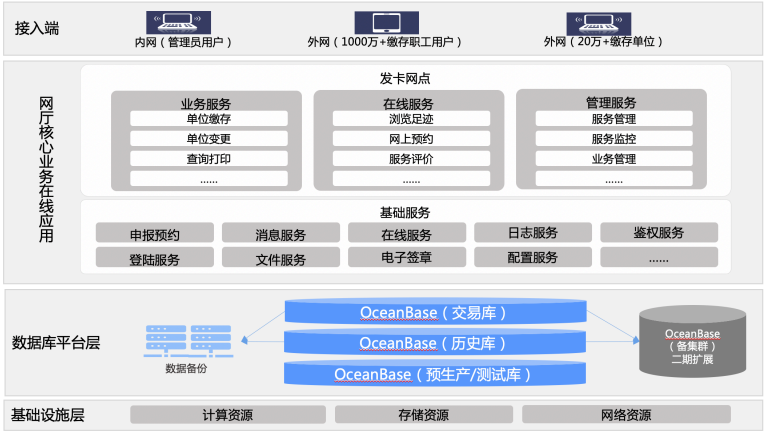
客户收益：

业务收益

- 集成公积金归集、提取、贷款、专办员学习等相关业务，对用户提供统一的互联网业务访问服务。
- 通过该系统处理的单位公积金业务量，超过单位公积金业务总量的40%，部分个人公积金业务使用频率也逐步增多。
- 全新分布式架构，降低运维成本的同时大幅提升业务创新效率。

信息化收益

- 交易库、历史库、开发测试库和完善数据备份机制，构建高效的开发测试和生产自动化运维。
- 一站式数据迁移，实现分钟级即时回滚、负载回放验证、秒级数据验证和一键完成迁移。
- 多个数据库的统一配置和运行监控，提升运营效率。
- 满足等保二级安全要求。



解决方案：

全自研原生分布式数据库方案

- 实现多套系统集中，透明支持政务云和本地数据中心分布式交易

强大的HTAP混合负载支持

- 分区、读写分离、全局索引等技术实现负载均衡和OLAP查询效果提升
- LSM-Tree 存储引擎提升OLTP事务效率，支持互联网化应用类型负载

完善的应用迁移和运维

- 一站式数据库无损迁移，过渡方案确保柔性切割
  - 即有应用能够平滑迁移；
  - 自动化运维手段和告警设置；
  - 完备的管理流程和最佳实践；

OceanBase 微信公众号

OceanBase 官网

THANKS