OceanBase：透明可扩展的企业级数据库

杨传辉 / 日照

蚂蚁金服 研究员

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE
全球技术领导力峰会

Geekbang. 极客邦科技 | TGO 鲲鹏會

500+高端科技领导者与你一起探讨
技术、管理与商业那些事儿

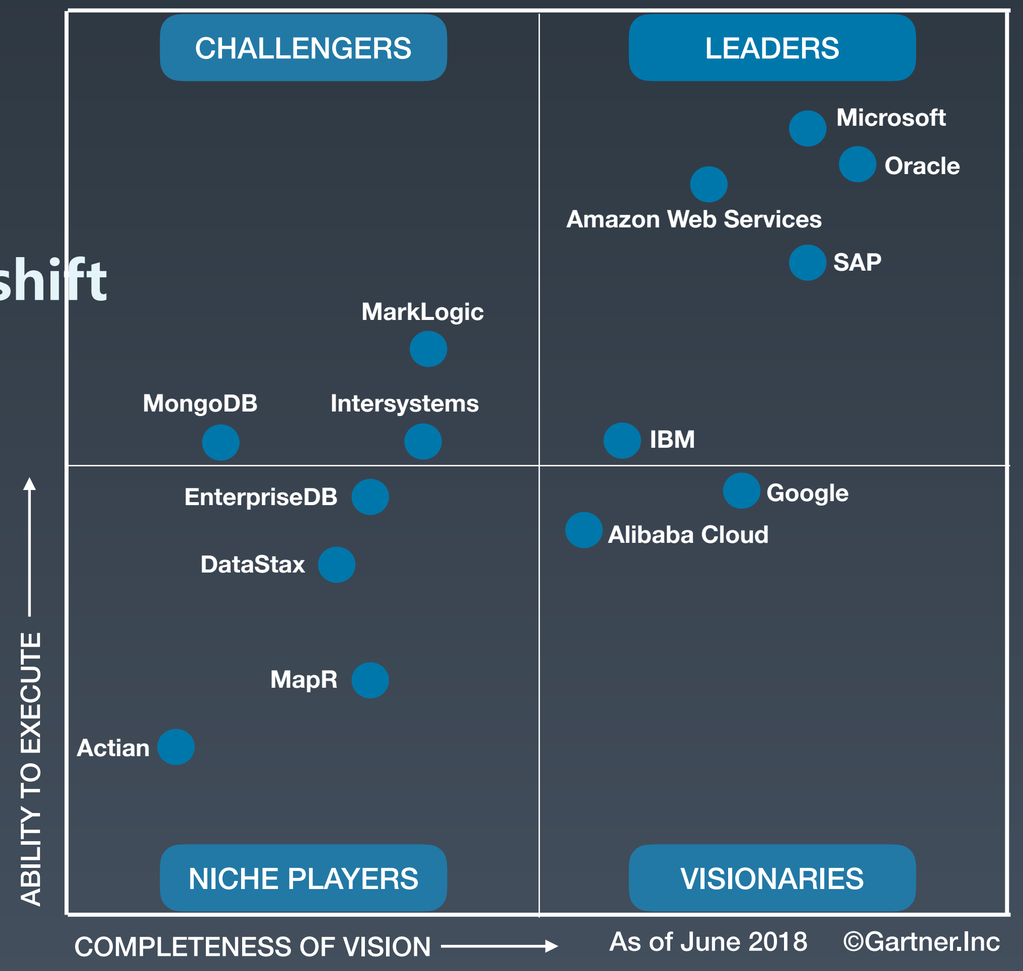
2019年6月14–15日 | 上海圣诺亚皇冠假日酒店

扫码了解更多信息

目录

行业现状

- 企业级数据库：Oracle、SQLServer、DB2
- 云数据库：Amazon Aurora、Amazon Redshift
- 魔力四象限



企业级数据库面临的问题

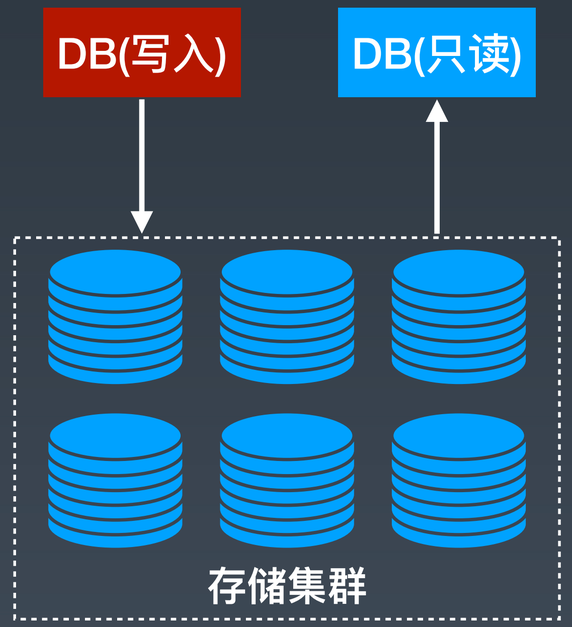
单机不可扩展

$$$

成本高

云数据库 != 透明可扩展

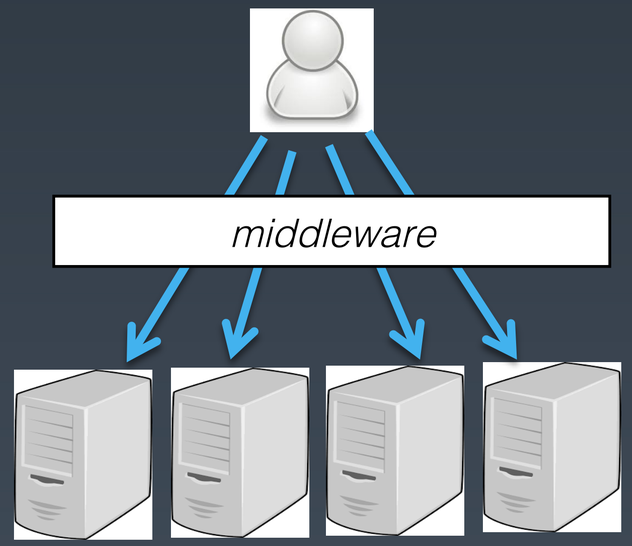
- 云数据库：开源数据库 + 存储计算分离
- 解决了存储可扩展问题，但事务和SQL不可扩展
- 开源数据库核心能力距离企业级数据库仍有较大差距



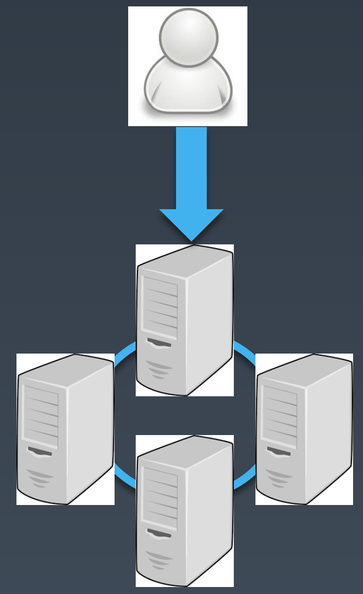
Hybrid clouds require excellent distributed OLTP DBMS, and the memory/storage architecture still requires a lot of work. In addition, data security and data management are both issues that need to be considered. —C Mohan@ICDE 2019, IBM Fellow

分库分表 != 透明可扩展

- 全局索引
- 全局快照
- 跨服务器复杂查询
- 跨服务器DML语句
- 带容错能力的分布式事务



中间件分库分表



分布式数据库

透明可扩展的企业级数据库

- 无需业务修改，按需扩容
  - 核心能力可扩展（存储、事务、SQL）
  - 线性可扩展
  - 持续可用，稳定
  - 企业级数据库功能
  - 通过核心业务和benchmark证明

目录

事务ACID

- 原子性（A）
  - 事务操作要么全部成功，要么全部失败
- 一致性（C）
  - 一个事务只能使数据库从一个一致的状态跳转到另一个一致的状态，不能破坏主键唯一或者所有列之和为固定值之类的约束
- 隔离性（I）
  - 多个并发事务互相不影响，就如同多个事务串行执行一般
- 持久性（D）
  - 一旦事务成功提交，它对数据库的影响是永久的

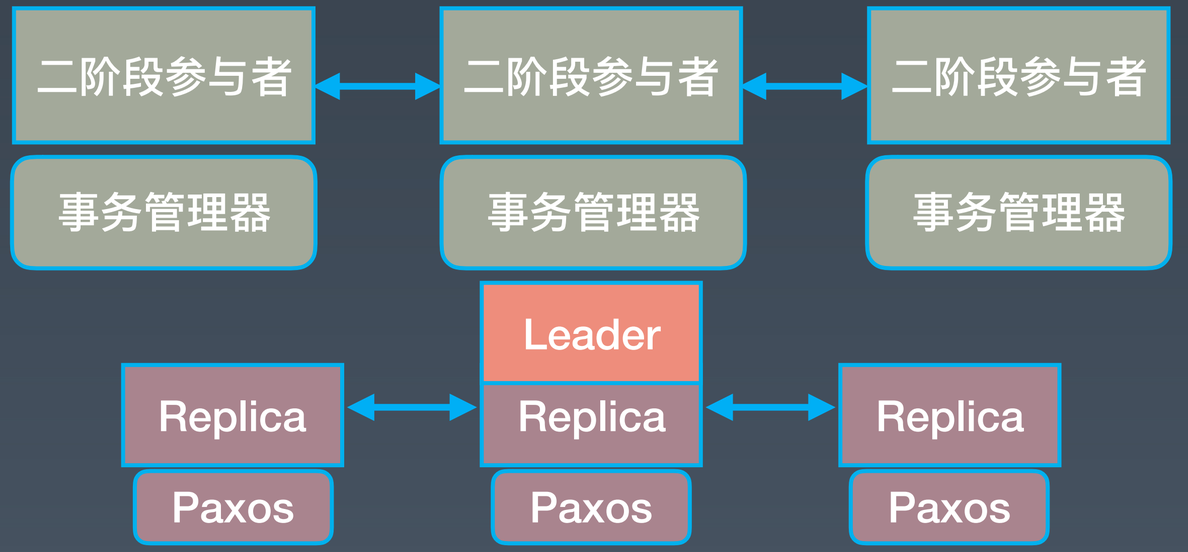
分布式事务：2PC协议的陷阱

- 1978年， Jim Gray
- 阻塞协议：参与者宕机/协调者宕机 ➡ 一台机器故障导致整个集群不可服务

分布式事务：Paxos + 2PC

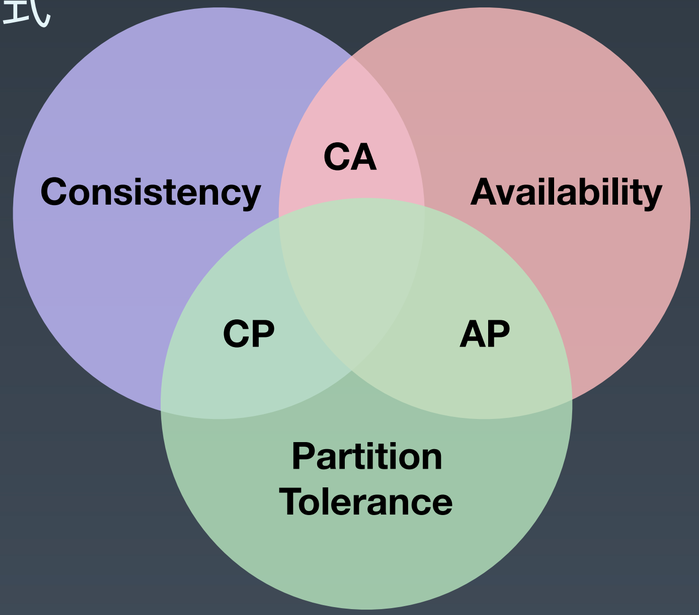
- 分布式事务的应对方案
  - 中间件XA：依赖数据库
  - NOSQL系统：CAP理论，回避一致性与分布式事务
- 云时代的架构选择：直面问题，采用Paxos + 2PC

Consensus on Transaction Commit
Jim Gray and Leslie Lamport
Microsoft Research
1 January 2004
Revised 19 April 2004,8 September 2005



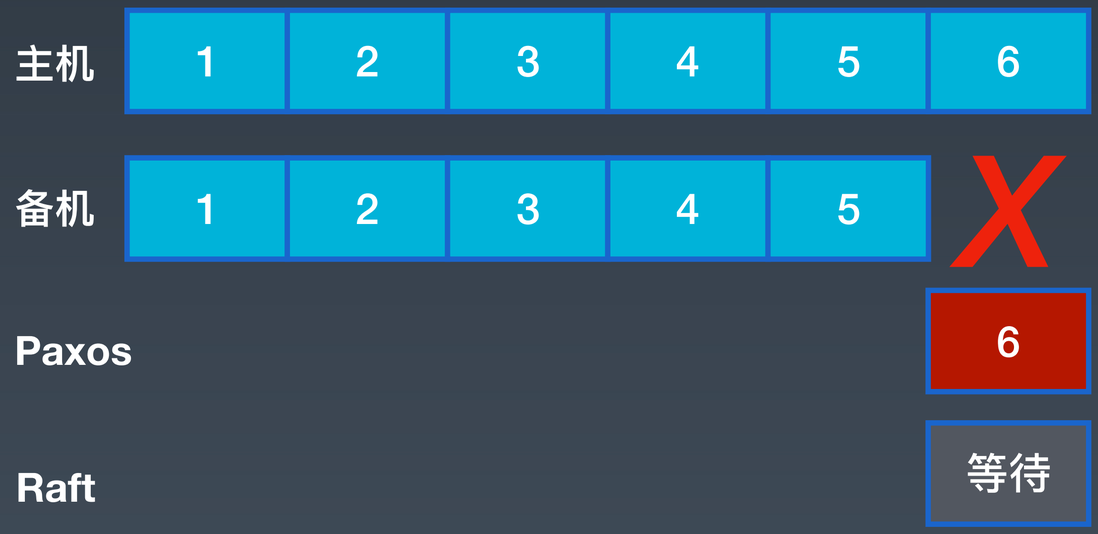
CAP与Paxos

- 主备同步模式：最高保护模式、最高性能模式、最高可用模式
- CAP：P无法规避，C与A不可兼得
- Paxos的高可用与CAP的可用性
  - **Paxos高可用：单点故障时多数派能否快速恢复**
  - **CAP可用性：单点故障时故障节点能否恢复**



Raft or Paxos

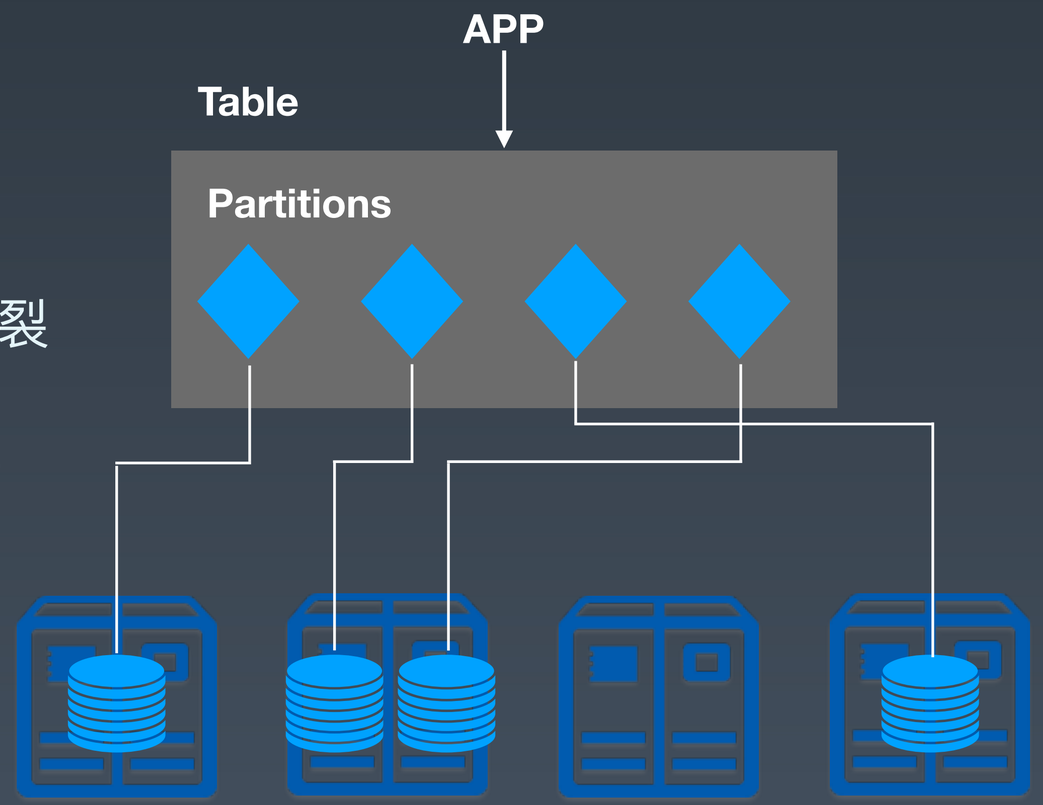
- Raft的得与失
  - **得：顺序提交日志，大大简化Paxos**
  - **失：并发能力更差，牺牲可用性，异地部署有风险**
- 常见系统做法
  - **Paxos阵营：Google Spanner，Ant Financial OceanBase 1.0，Amazon DynamoDB**
  - **Raft阵营：Ant Financial OceanBase 0.5，Tencent TDSQL，以及一系列开源系统**



目录

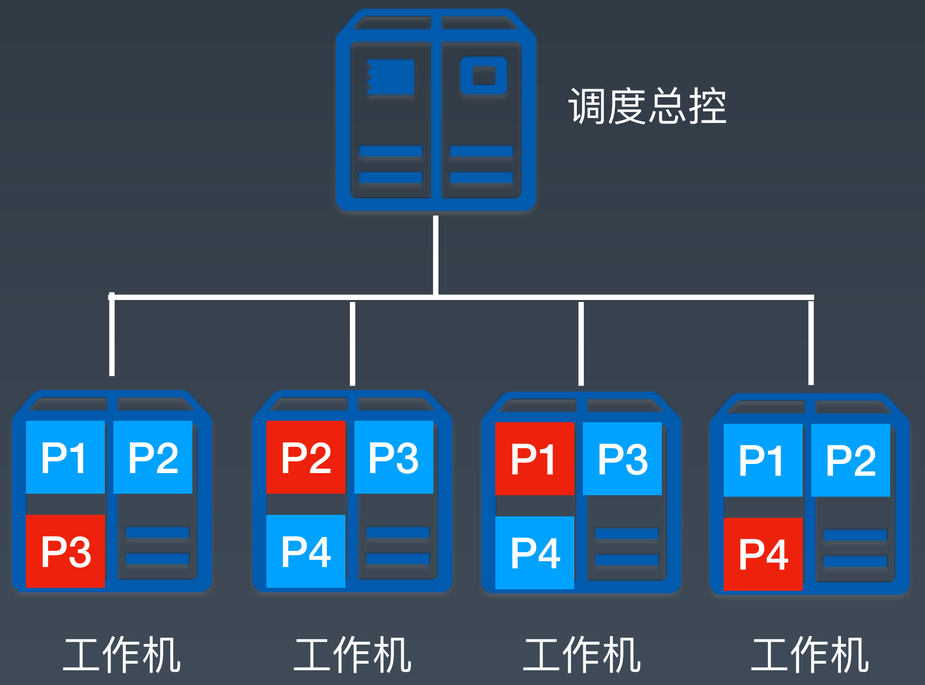
分布式分区表

- **全局一致性 / 强一致的全局索引**
- **多种数据分区，二级分区**
- 分区分裂：数据量太大或者load太高时自动分裂
- 分区合并：数据删除较多，相邻分区自动合并



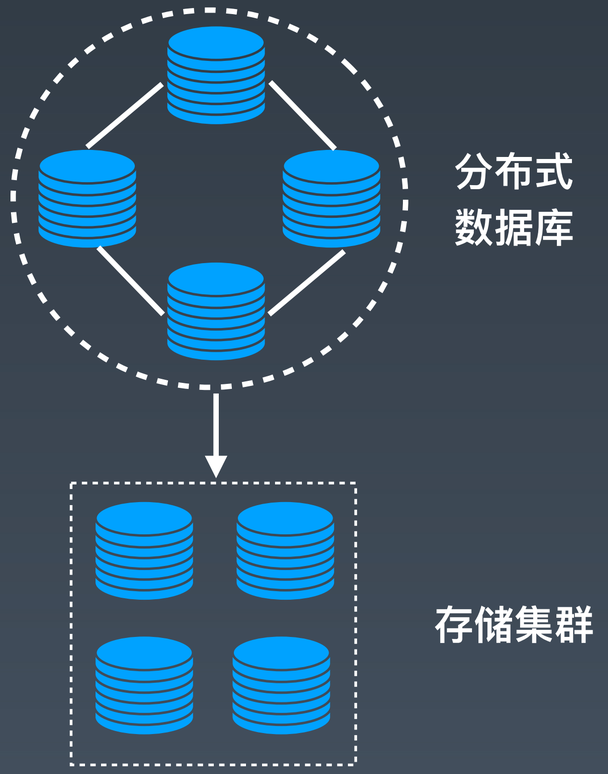
自动负载均衡

- 服务器自动上下线
- 负载重新均衡
- 逻辑复制与物理复制



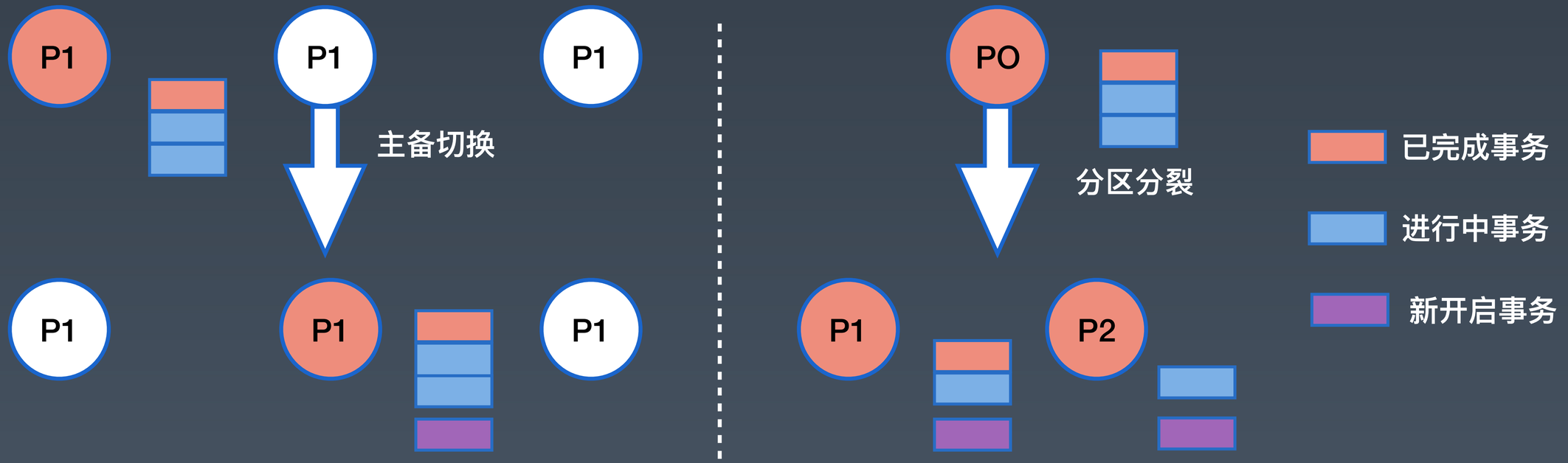
负载均衡的两难选择

- 多因子负载均衡
  - 计算均衡（CPU&内存），存储均衡（磁盘占用）
  - 计算存储资源配比和实际业务不匹配
  - 存储迁移耗时长，计算负载变化快
- 存储计算分离
  - 分布式数据库负责计算均衡，存储集群负责存储均衡



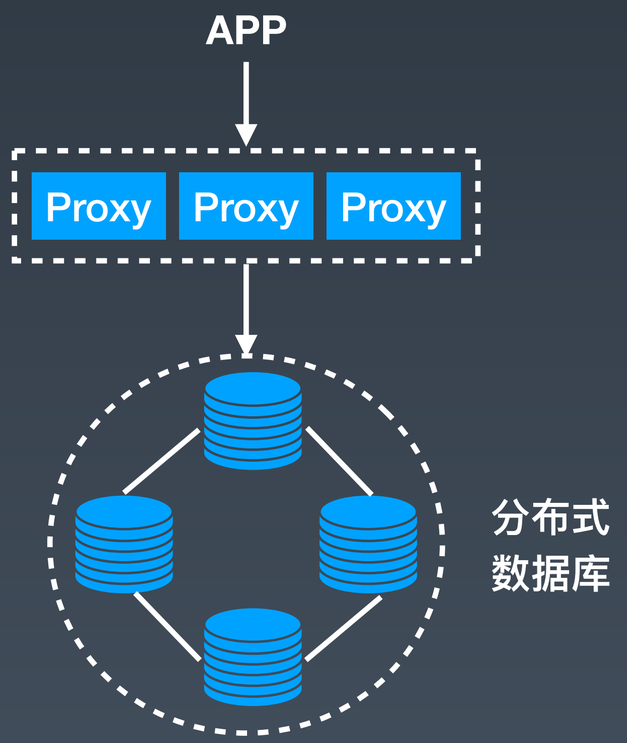
分区容错

- 主备切换不杀事务：新事务在新的主分区开启，进行中事务在线迁移
- 分区分裂不杀事务：新事务在分裂后的新分区开启，进行中事务在线迁移



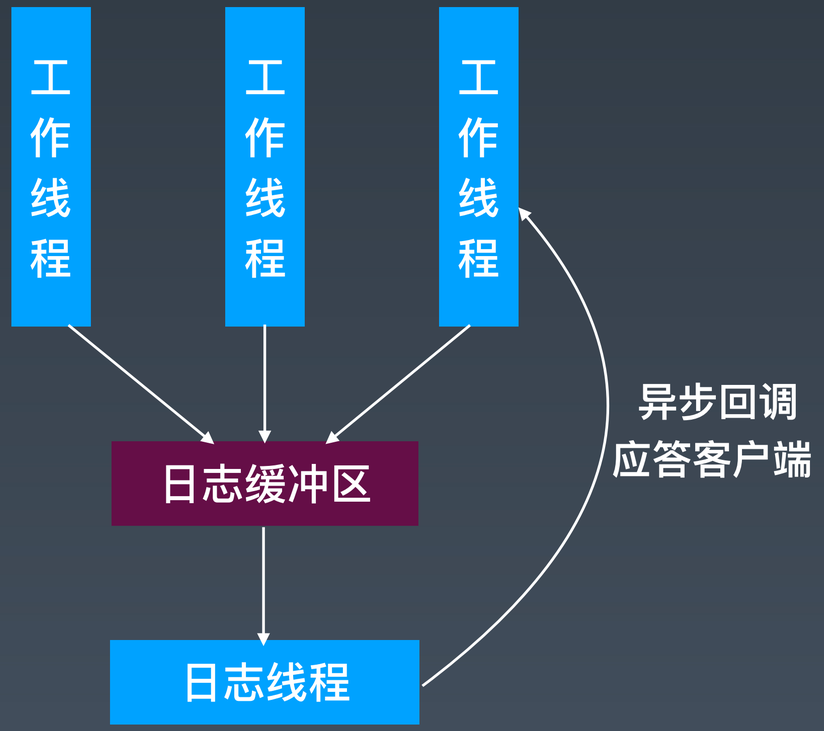
全链路请求容错

- 读写请求重试，防止重试风暴
- 任务级重试
- Proxy在线升级，数据库在线session迁移
- 异常处理：磁盘/服务器hung住，“半死不活”



分布式线程模型

- 分布式数据库跨机场景
  - 获取全局事务版本号（SCN）
  - 主备强同步
  - 两阶段提交
  - 分布式执行计划
- 异步执行不占用工作线程
- 协程降低线程切换开销



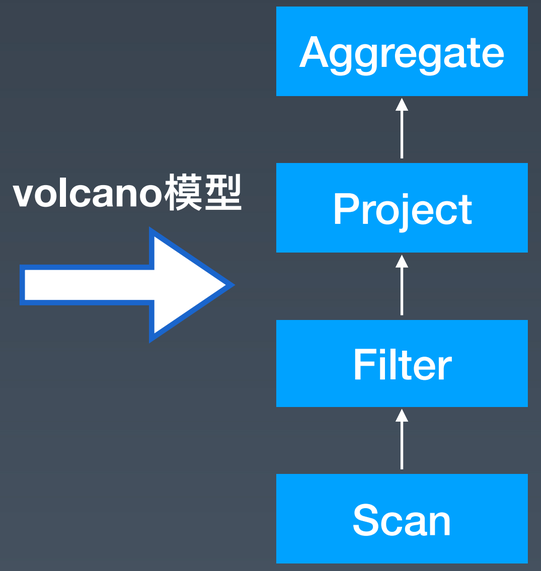
HTAP执行引擎

- **强类型系统**
- 行迭代批处理
- 推模型提升代码局部性
- 编译执行
- 并行执行

```
int compare(Key k1, Key k2)
{
  int ret = 0;
  if ( INT == k1.get_type() && INT == k2.get_type()) {
    ret = int_compare(k1.get_value(), k2.get_value());
  } else if {NUMBER == k1.get_type()
             && NUMBER == k2.get_type()) {
    ret = number_compare(k1.get_value(), k2.get_value());
  }
  ...
}
```

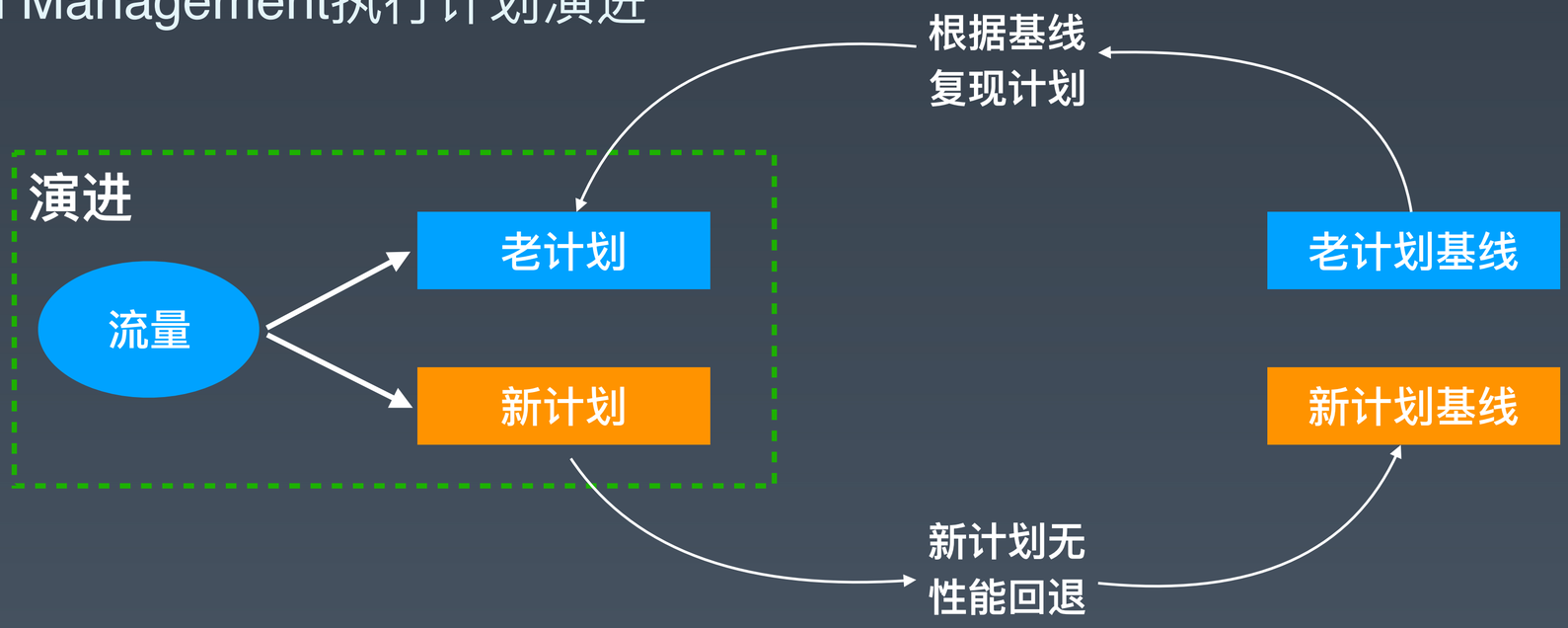
```
select count(*) from store_sales
where ss_item_sk = 1000;
```

volcano模型



企业级查询优化器

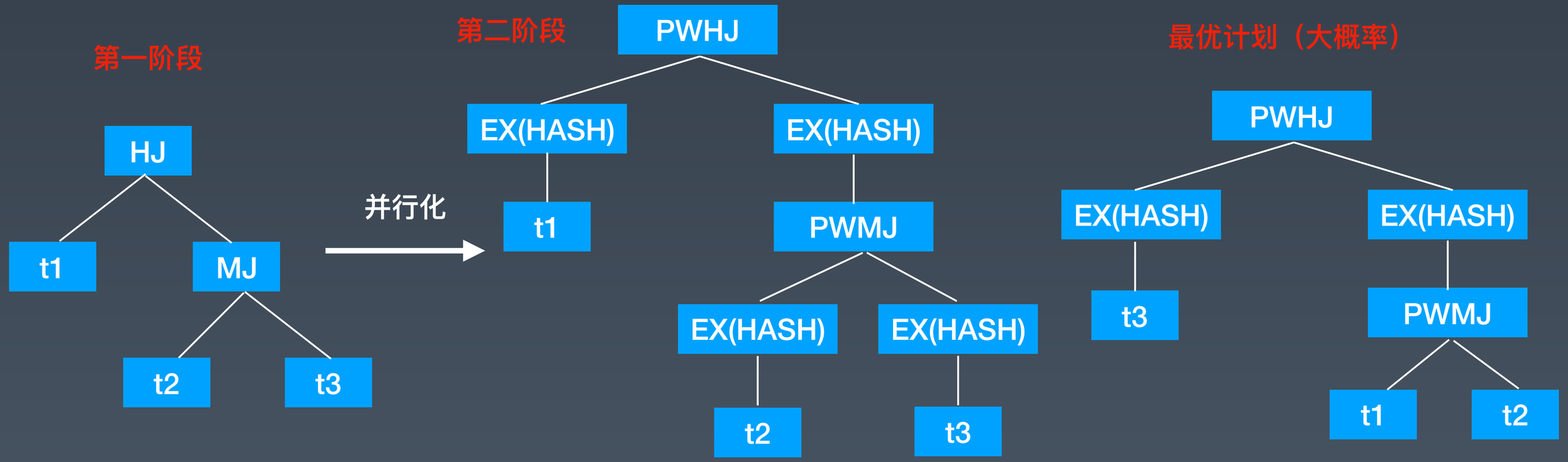
- **基于代价的查询优化器**
- **Adaptive Cursor Sharing解决大小账号问题**
- SQL Plan Management执行计划演进



并行优化

- 单机数据库：串行优化 => 算子局部并行化

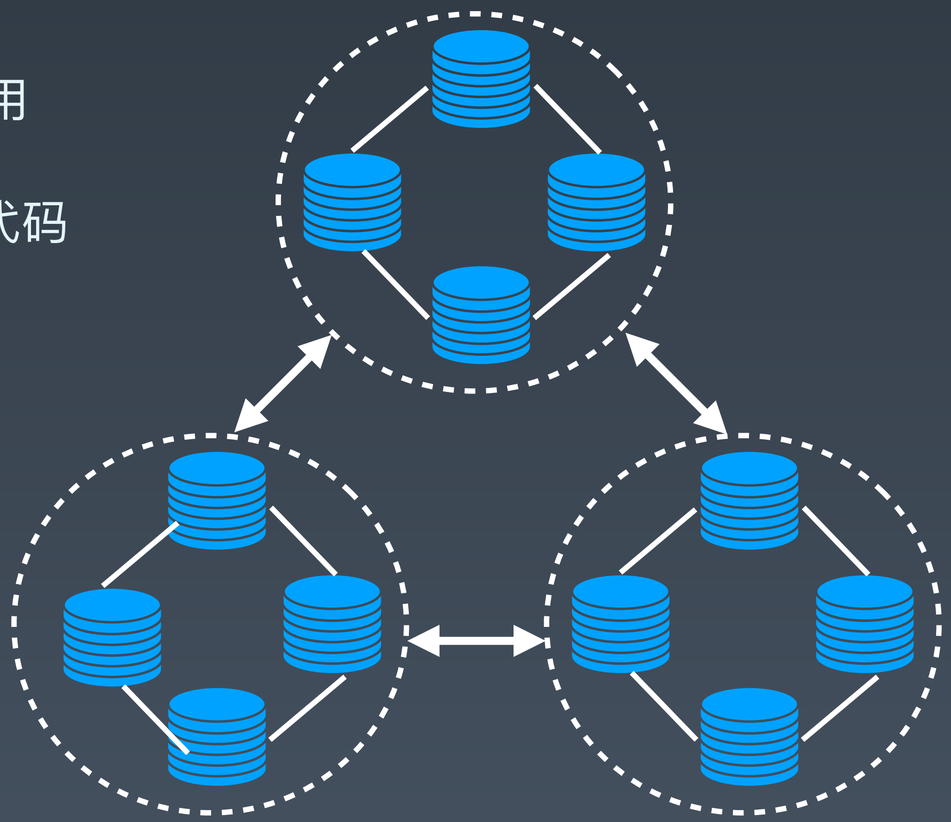
```
create table t1(a int primary key, b int, c int) partition by hash(a) partitions 4;
create table t2(a int primary key, b int, c int) partition by hash(a) partitions 4;
create table t3(a int primary key, b int, c int) partition by hash(a) partitions 5;
select * from t1, t2, t3 where t1.a = t2.a and t2.b = t3.b;
```



目录

关于OceanBase

- **阿里巴巴、蚂蚁金服自主研发的企业级分布式关系数据库**
- 第一次将Paxos协议引入到关系数据库领域，实现持续可用
- 工业级shared nothing分布式数据库架构，无需业务修改代码
  - **透明可扩展**
  - **全局一致的数据库视图**
  - **跨服务器复杂查询**
- **MySQL全兼容，Oracle部分兼容，原生多租户支持**



OceanBase使用情况

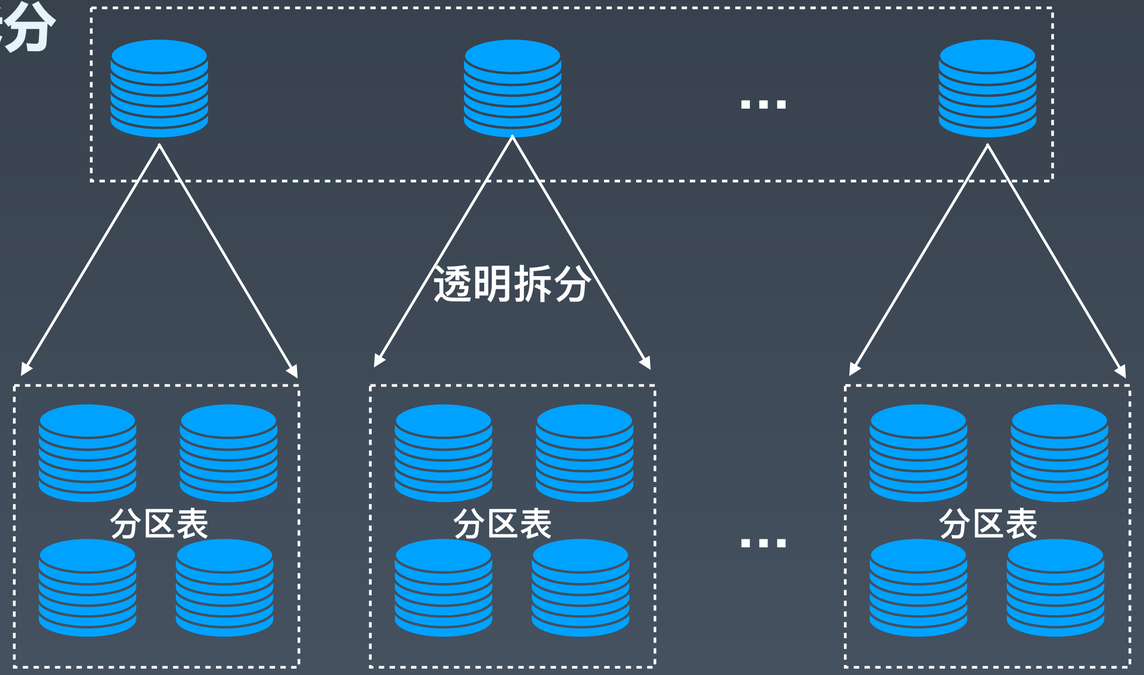
- 蚂蚁金服：支付宝核心链路100%支付量，网商银行全部流量，并已进军国际业务

- 在浙商银行、南京银行、苏州银行、广东农信、人保健康险等外部客户的互联网核心系统中，承担交易数据库的重要角色

浙商银行 | 南京银行 BANK OF NANJING | 苏州银行 BANK OF SUZHOU | 兴业银行 INDUSTRIAL BANK CO.,LTD.
广东农信 GDRC | PICC 中国人保健康 PICC HEALTH | 常熟农商银行 CHANGSHU RURAL COMMERCIAL BANK | 广发银行 CHINA GUANGFA BANK
中国建设银行 China Construction Bank | 招商证券 China Merchants Securities | 天津银行 BANK OF TIANJIN

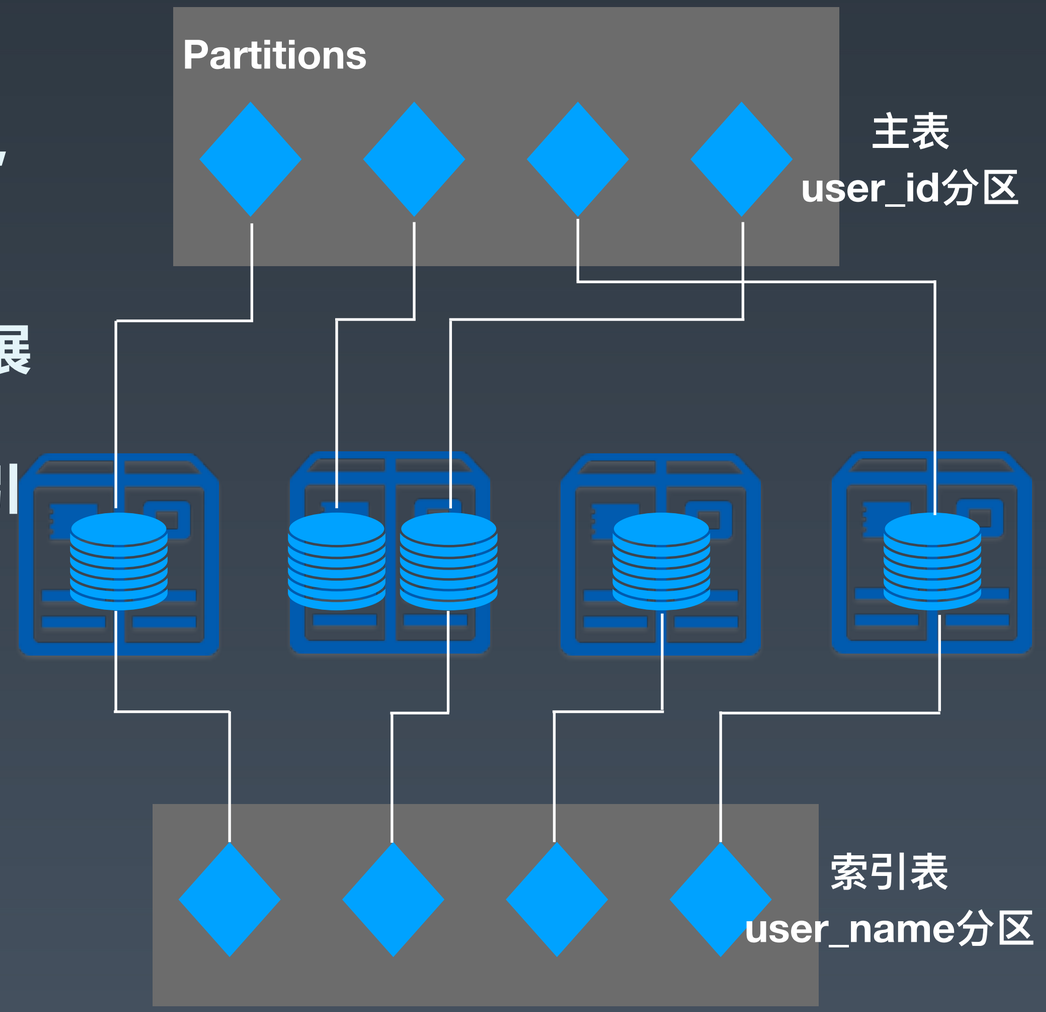
交易支付透明拆分

- 蚂蚁金服交易支付按照user_id拆分N份，需要扩容到M*N份
- 痛点：中间件+业务拆分，需要上百人年开发量，技术风险很高
- 解决方案：OceanBase分区表实现透明拆分



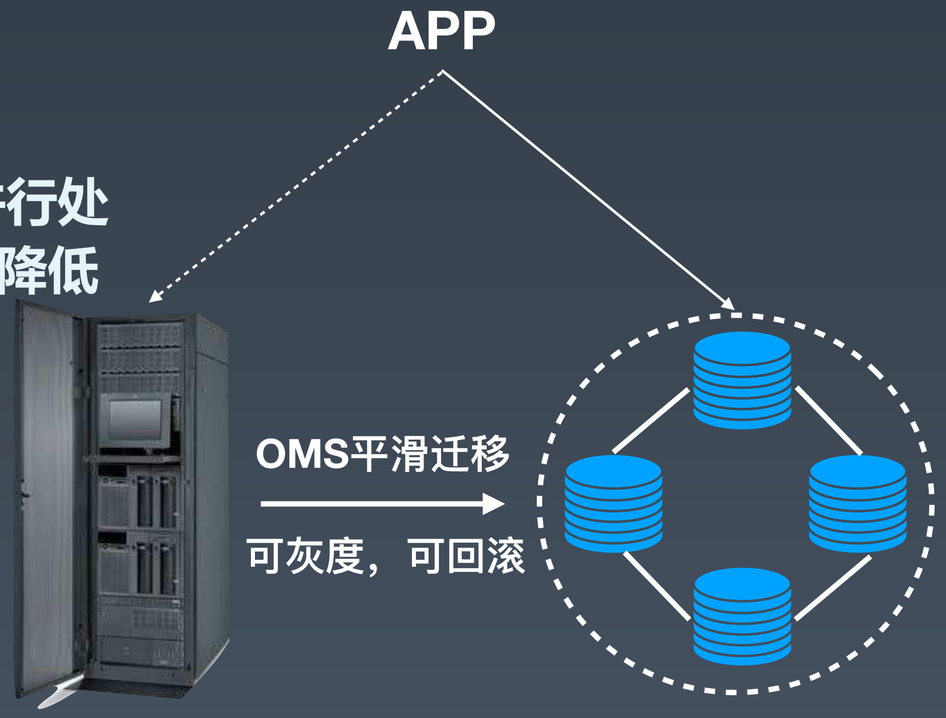
会员系统全局索引

- 蚂蚁金服会员系统，根据user_id，user_name，email查找用户信息
- 痛点：单机数据库，只能垂直扩展，无法水平扩展
- 解决方案：OceanBase分区表 + 强一致全局索引



结算系统小机下移

- 某金融机构有大量批处理场景，有多张大表关联的复杂计算，并且涉及到大量的数据更新
- 痛点：传统集中式数据库单点瓶颈，成本高
- 解决方案：透明可扩展的OceanBase，HTAP场景并行处理能力，处理时间缩短到现有系统的一半，TCO大幅降低



极客时间 | 企业服务

这是一封正经“福利”贴

想做团队的领跑者 需要迈过这些“槛”

成长型企业，易忽视人才体系化培养 企业转型加快，团队能力又跟不上	VS	从基础到进阶，超100+一线实战 技术专家带你系统化学习成长
团队成员技能水平不一， 难以一“敌”百人需求	VS	解决从小白到资深技术人所遇到 80%的问题
寻求外部培训，奈何价更高且 集中式学习	VS	多样、灵活的学习方式，包括 音频、图文 和视频
学习效果难以统计，产生不良循环	VS	获取员工学习报告，查看学习 进度，形成闭环

课程顾问「橘子」

回复「QCon」
免费获取
学习解决方案

 # 极客时间企业账号 # 解决技术人成长路上的学习问题

Geekbang. 极客邦科技 InfoQ

极客邦科技
会议推荐2019

5月

QCon 北京
全球软件开发大会
大会：5月6–8日
培训：5月9–10日

QCon 广州
全球软件开发大会
培训：5月25–26日
大会：5月27–28日

6月

GTLC GLOBAL TECH LEADERSHIP CONFERENCE 上海
技术领导力峰会
时间：6月14–15日

GMTC 北京
全球大前端技术大会
大会：6月20–21日
培训：6月22–23日

7月

ArchSummit 深圳
全球架构师峰会
大会：7月12–13日
培训：7月14–15日

10月

QCon 上海
全球软件开发大会
大会：10月17–19日
培训：10月20–21日

11月

GMTC 深圳
全球大前端技术大会
大会：11月8–9日
培训：11月10–11日

AiCon 北京
全球人工智能与机器学习大会
大会：11月21–22日
培训：11月23–24日

12月

ArchSummit 北京
全球架构师峰会
大会：12月6–7日
培训：12月8–9日

欢迎关注OceanBase公众号
了解更多OB最佳技术实践内容

QCon 10th

杨传辉 / 日照
rizhao.ych@alipay.com