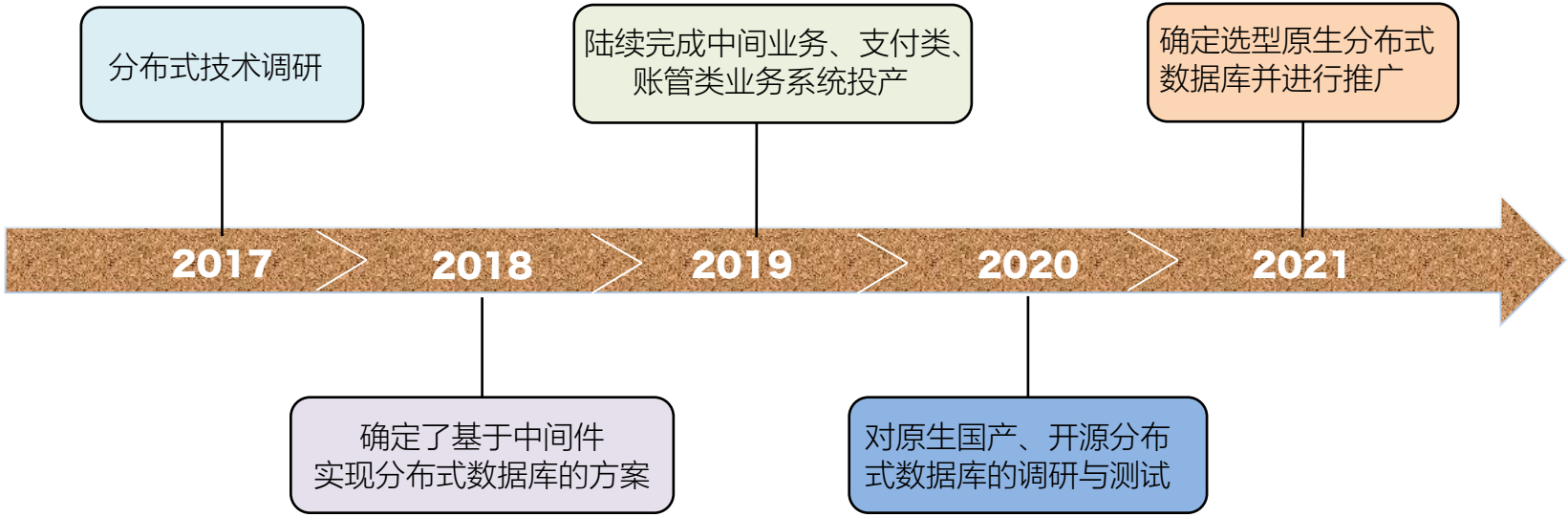
银行分布式数据库改造方案实践与探索

演讲人：王辉

目录

发展历程


分布式技术调研
2017
确定了基于中间件
实现分布式数据库的方案
2018
陆续完成中间业务、支付类、
账管类业务系统投产
2019
对原生国产、开源分布
式数据库的调研与测试
2020
确定选型原生分布式
数据库并进行推广
2021

分布式技术研究-分布式中间件

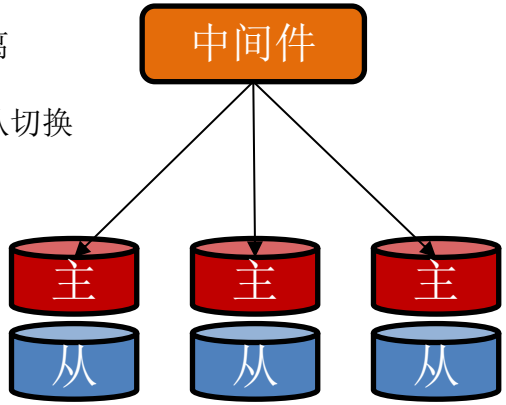
中间件

分布式技术研究-分布式中间件

① SQL解析、路由转发、聚合计算

② 分库分表、读写分离

③ 数据库高可用、主从切换



代表有：

MyCat、Atlas、

Cetus、ProxySql、

Sharding-JDBC、

TDDL、DBLE等

分布式技术研究-分布式中间件

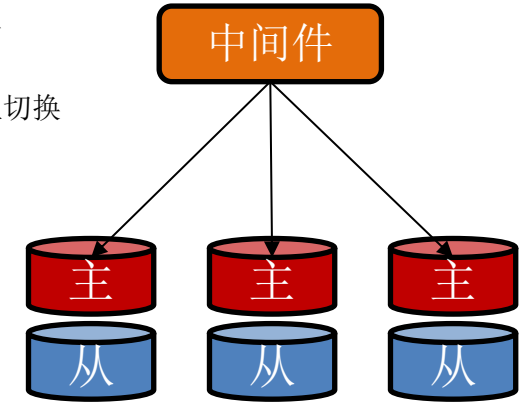
① SQL解析、路由转发、聚合计算

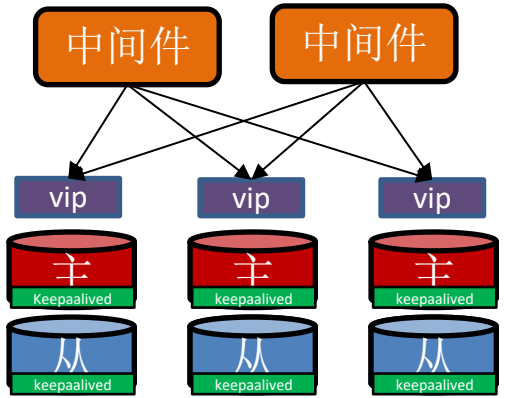
② 分库分表、读写分离

③ 数据库高可用、主从切换

代表有：

MyCat、Atlas、Cetus、ProxySql、 Sharding-JDBC、 TDDL、DBLE等





通过keepalive+mha的方式将数据库高可用下移到了数据层通过脚本实现，让中间件更专注于SQL解析与计算

分布式技术研究-分布式中间件

① SQL解析、路由转发、聚合计算

② 分库分表、读写分离

③ 数据库高可用、主从切换

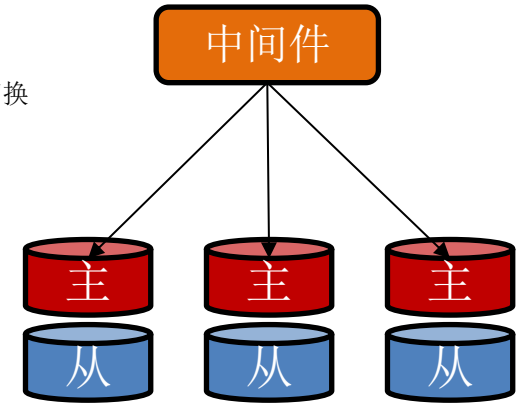
代表有：

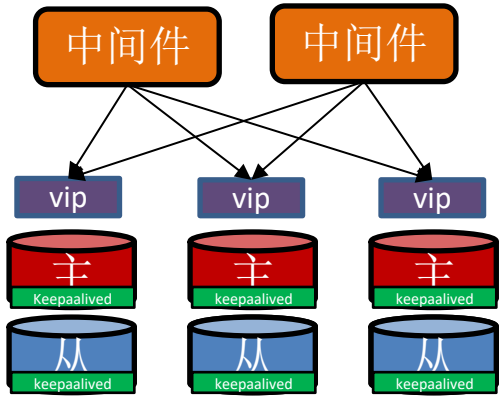
MyCat、Atlas、

Cetus、ProxySql、

Sharding-JDBC、

TDDL、DBLE等



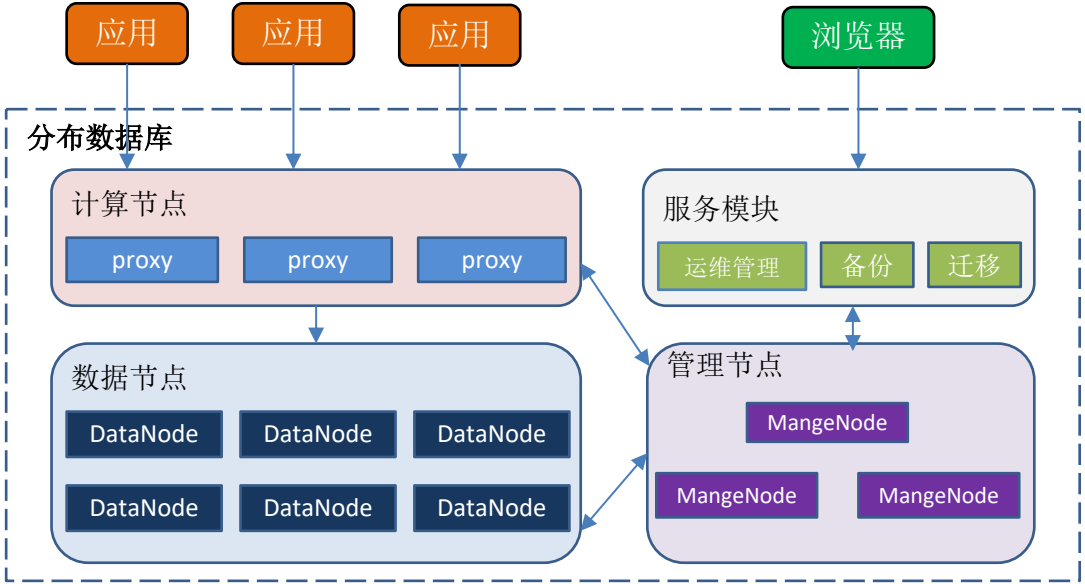


通过keepalive+mha的方式将数据库高可用下移到了数据层通过脚本实现，让中间件更专注于SQL解析与计算

1. 数据库副本一致性的问题
2. 数据库高可用可靠性问题
3. 中间件语法限制、性能、分布式事务能力等问题
4. 各组件统一运维管理的问题

分布式技术研究-架构体系

绝大部分分布式数据库架构设计由的计算、数据、管理节点组成



计算节点：

负责SQL解析、路由和汇聚、计算能力

数据节点：

数据多副本存储

管理节点：

管理和监控各节点运行情况、保存元数据及提供全局服务等功能

服务模块：

提供其他辅助功能便于运维管理

分布式技术研究-架构体系

1，优化了复制技术，延迟更低

2，优化了XA事务，性能更高

3，增加了自动高可用切换策略

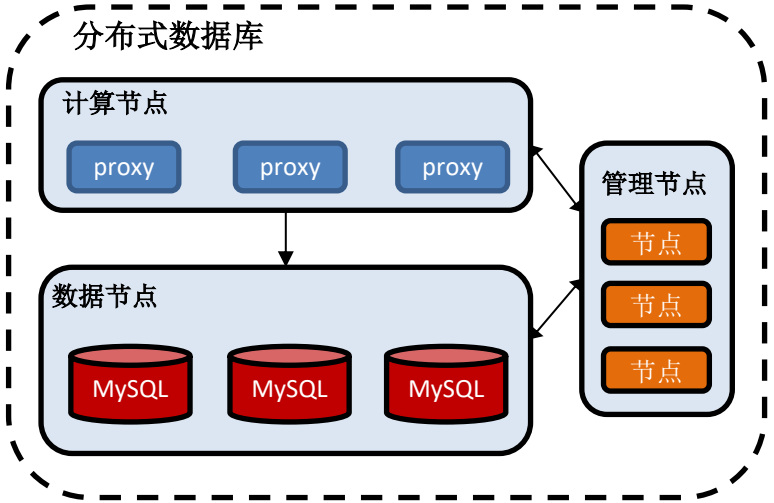
4，优化了Mysql半同步极端情况下数据丢失的情况，满足强一致要求

5，支持在线扩容

6，增加了GTM（全局事务管理器）、GSM（全局序列）及GTM（全局时钟管理器）

7，提供多租户技术

1



代表：TDSQL、GaiaDB 、GoldenDB、GreatDB、爱可生

分布式技术研究-架构体系

1，优化了XA事务，性能更高

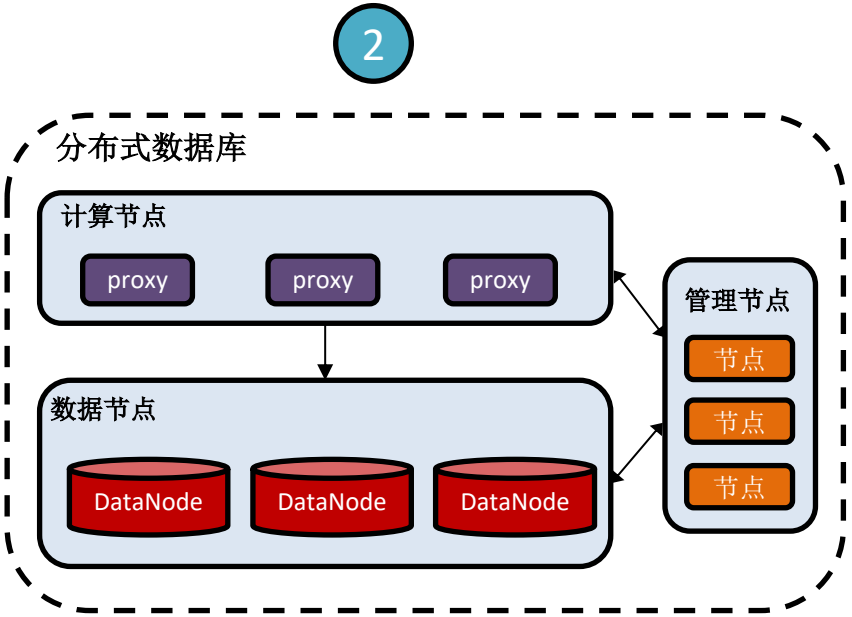
2，增加了自动高可用切换策略

3，数据同步基于Poxas、raft等共识协议，满足强一致要求

4，增加了GTM（全局事务管理器）、GSM（全局序列）及GTM（全局时钟管理器）

5，提供多租户技术

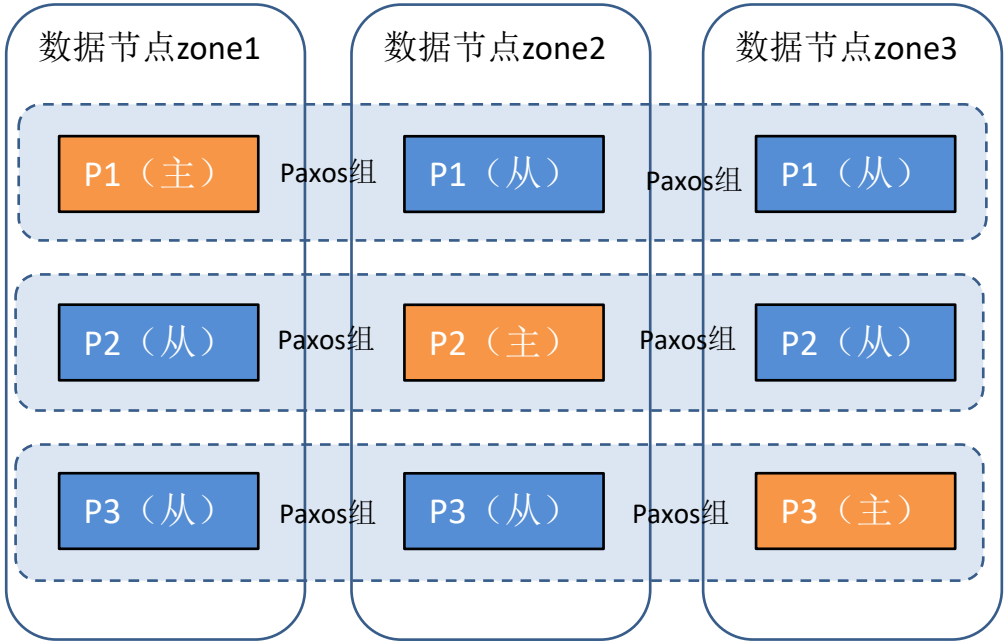
6，支持在线扩容与缩容



代表：OceanBase、巨杉、TiDB、华为

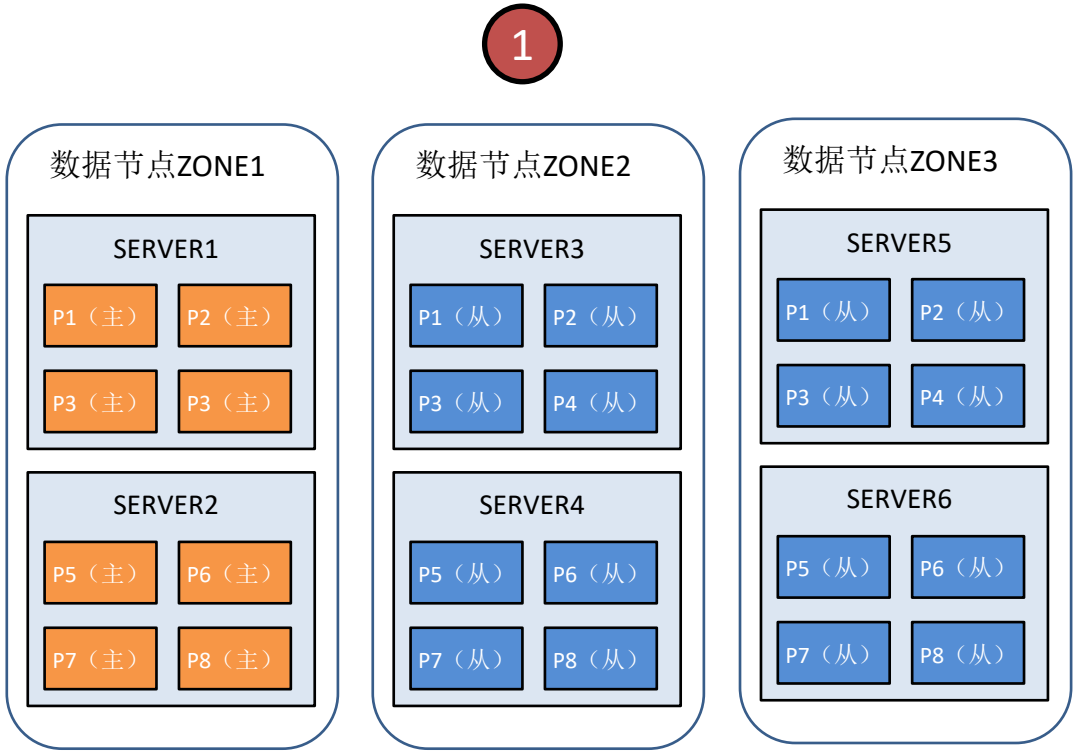
分布式技术研究-数据一致性

绝大部分分布式数据的数据节点都天然支持多副本，且基本都采用的paxos，raft等共识协议，确保副本数据的多数派强一致。



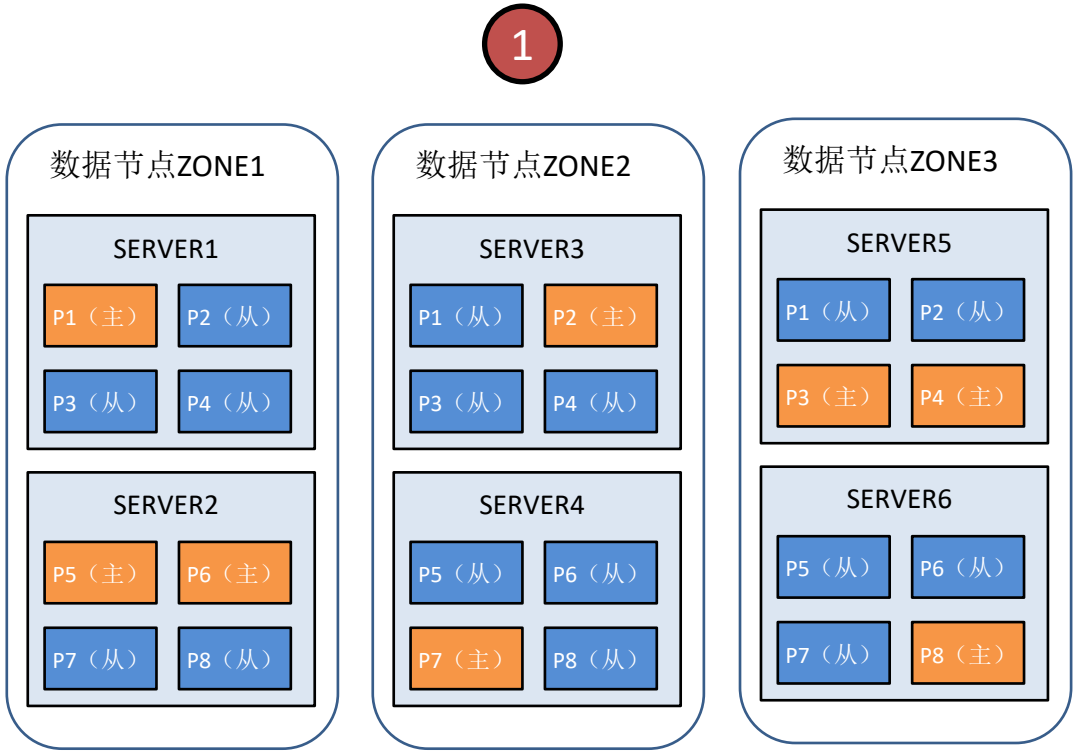
分布式技术研究-数据分布

- 逻辑拆分：指定数据表的分片键，采用 hash，range，list 等分片算法将数据拆分到各数据节点。
- 8个分片均匀分配到2个SERVER（数据库主机）中，每个SERVER4个分片，组成一个ZONE，每个ZONE都有完整的数据；一共3个ZONE，每个分片3个副本。
- 数据拆分后主节点可以放到一个区域避免网络交换（如右图），也可以全部可以均匀的分布到各区域。



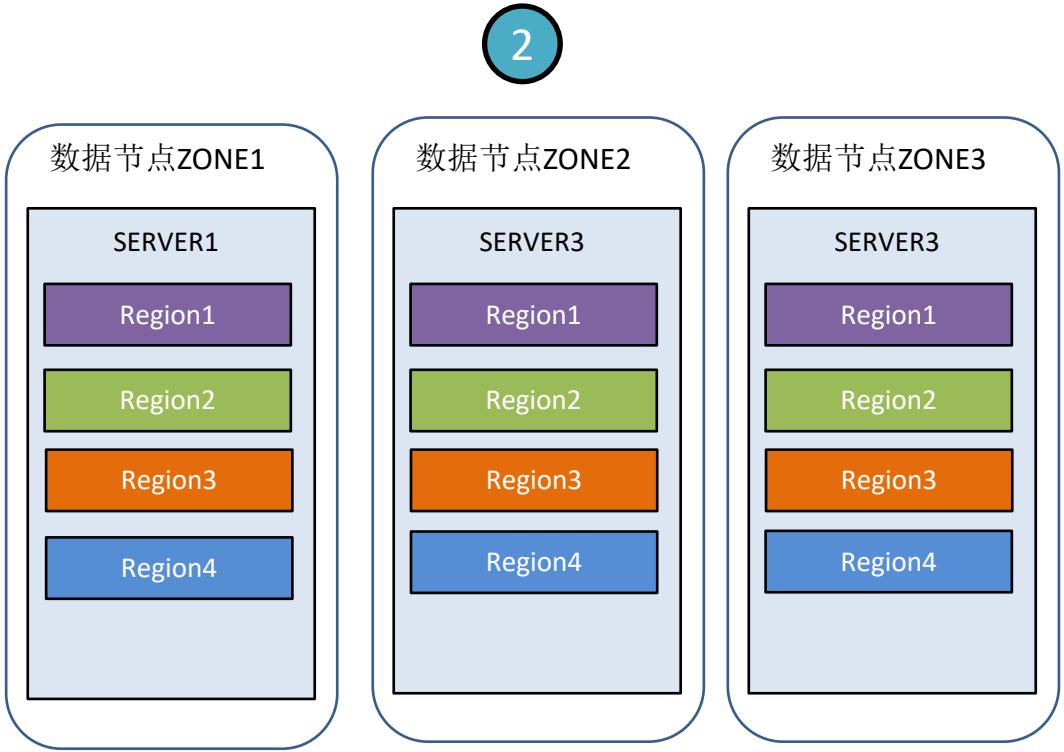
分布式技术研究-数据分布

- 逻辑拆分：指定数据表的分片键，采用 hash，range，list 等分片算法将数据拆分到各数据节点。
- 8个分片均匀分配到2个SERVER（数据库主机）中，每个SERVER4个分片，组成一个ZONE，每个ZONE都有完整的数据；一共3个ZONE，每个分片3个副本。
- 数据拆分后主节点可以放到一个区域避免网络交换（如右图），也可以全部可以均匀的分布到各区域。



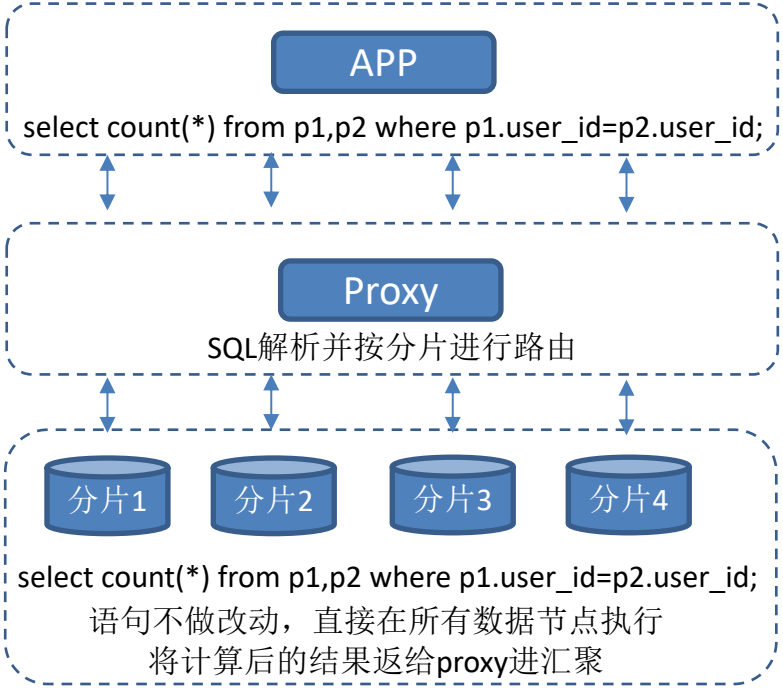
分布式技术研究-数据分布

物理拆分：最小存储单元是一个Region，与数据表结构无关，按照物理大小进行划分，达到一个存储单元后自动拆分。

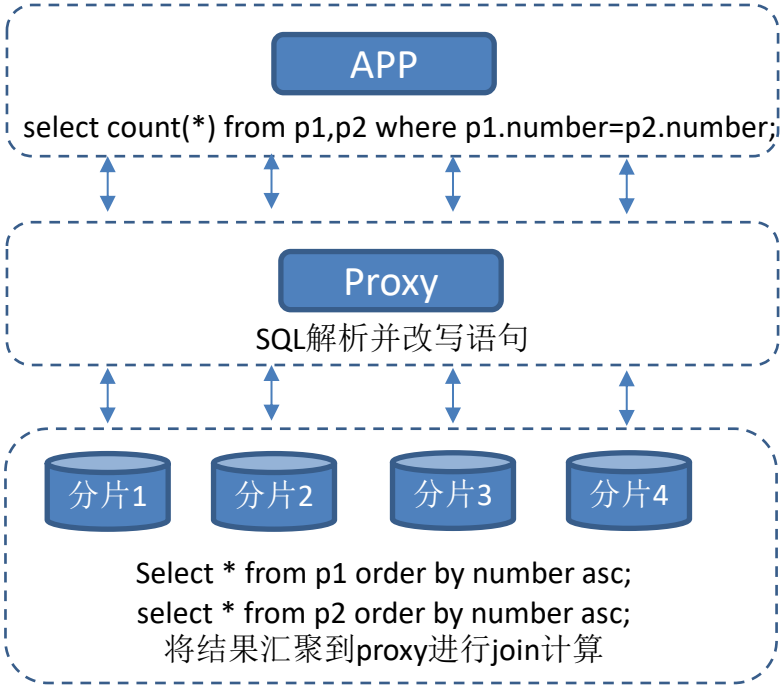


分布式技术研究-表关联

1 表关联采用分片键的处理

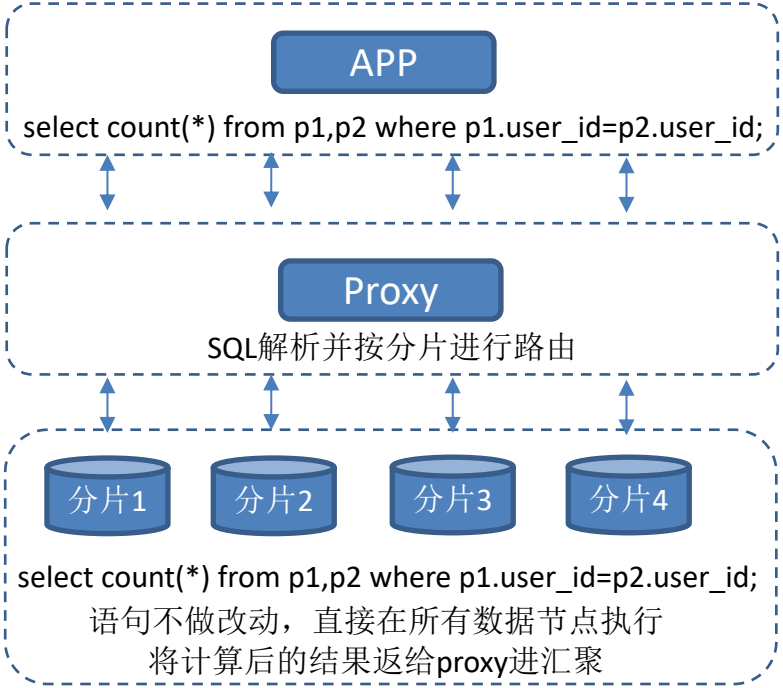


2 表关联采用非分片键的处理

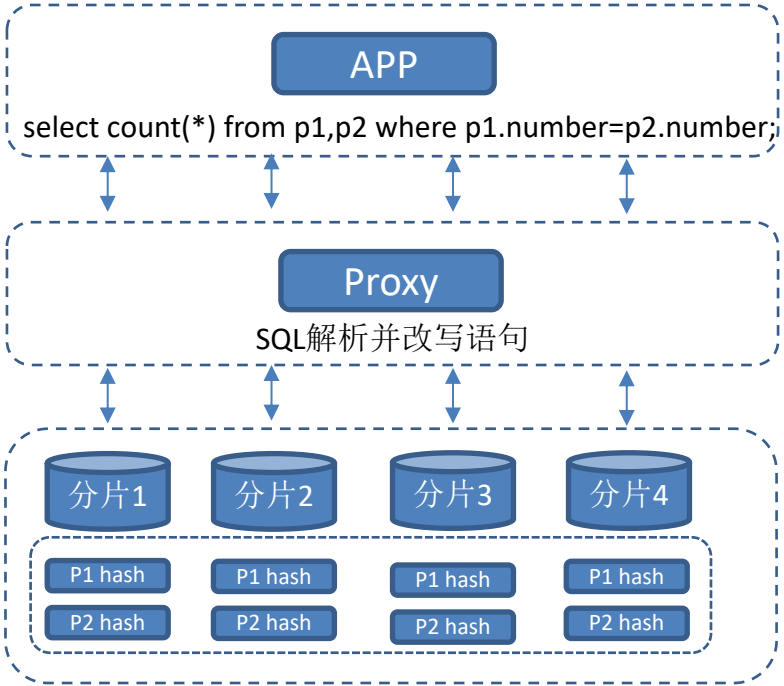


分布式技术研究-表关联

1 表关联采用分片键的处理



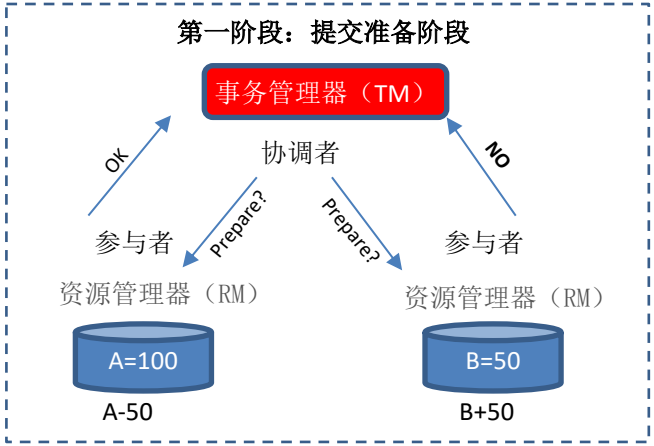
2 表关联采用非分片键的处理



分布式技术研究-分布式事务

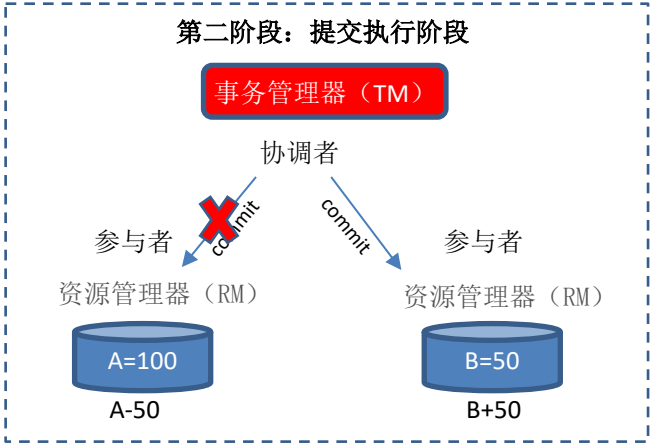
- 分布式事务处理能力，难点在ACID里面的A（原子性）和隔离性I（隔离性）
- 原子性：全部执行成功，全部不执行
- 集中式数据库由于不存在跨机操作的风险，如果服务故障，所有操作都不会完成，通过redo和undo机制，实现相对简单。
- 分布式环境下就会比较复杂：一个事务多条语句，在不同的节点（机器）完成，就会出现由有些主机异常，造成部分完成，部分未完成的情况，这样就影响事务的原则性。
- 而几乎所有的分布式数据库都采用2PC（两阶段提交）的方式来保证

分布式技术研究-分布式事务



单点故障

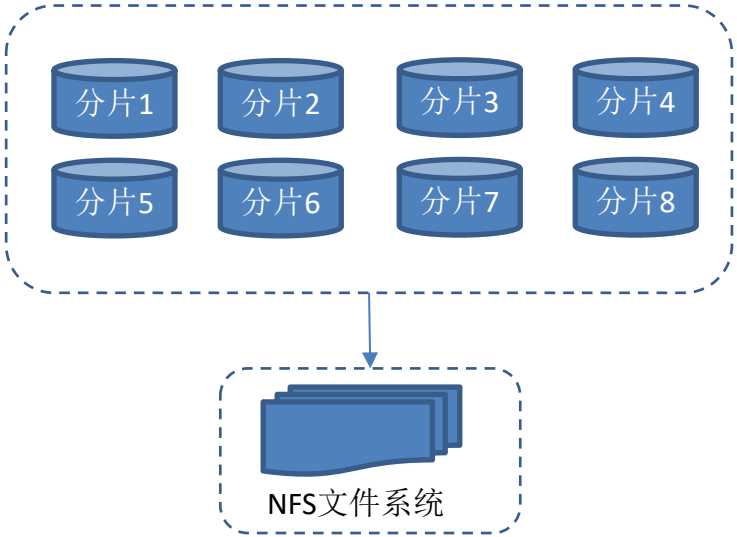
当参与者第一阶段执行完事务返回给协调者OK时、或者在第二阶段执行commit操作前协调者出现故障，那么参与者的事务会一致处于锁定状态，无法继续完成事务操作。



数据不一致

在二阶段提交的阶段二中，执行事务提交的时，当协调者向所有的参与者发送Commit请求后，发生局部网络异常或协调者在未全部发送完成commit请求前崩溃，导致只有部分参与者接收到了commit请求，于是整个分布式系统便出现数据不一致的现象。

分布式技术研究-一致性备份与恢复



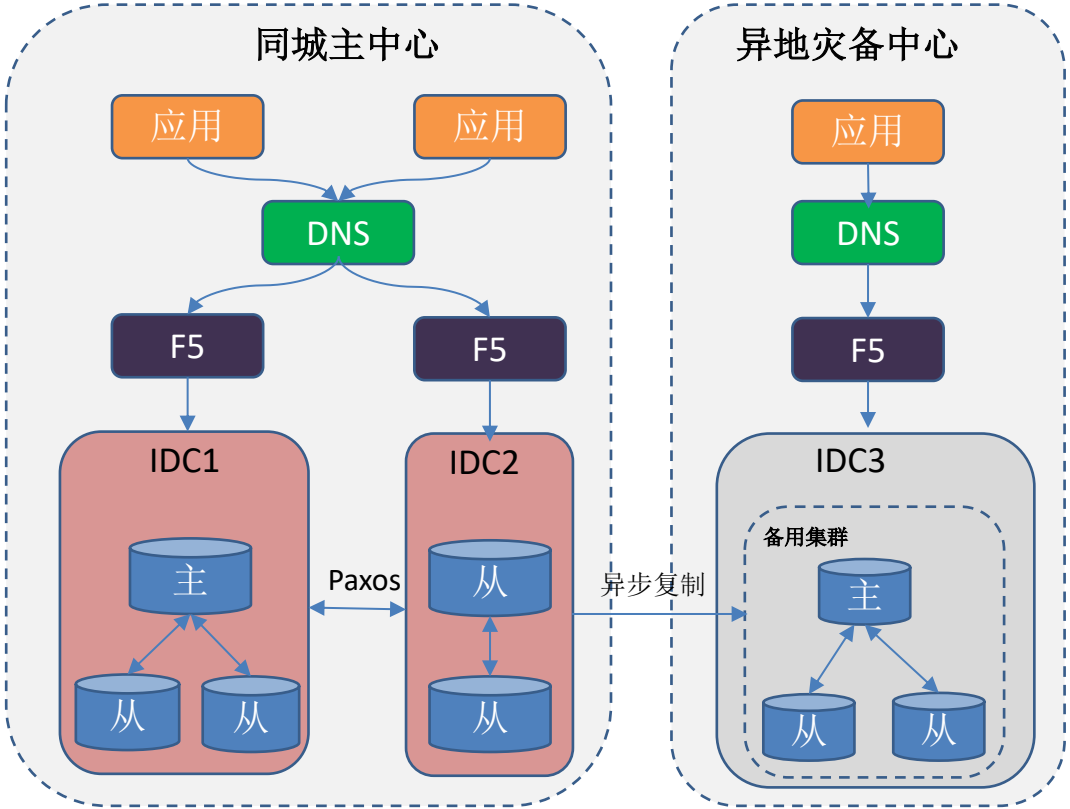
- 实现全局一致备份必须要有全局事务服务器 GTS（Global Transcation Service），通过快照的方式将所有节点的数据置于一点；同样基于GTS完成一致性恢复

- 一般为了减少主库影响，备份都在备库，另外备份一般存放到NFS文件系统，便于将所有分片的文件备份到一起，同时用于统一恢复

分布式技术研究-两地三中心架构

两地三中心同城5副本（32）基于Paxos强同步

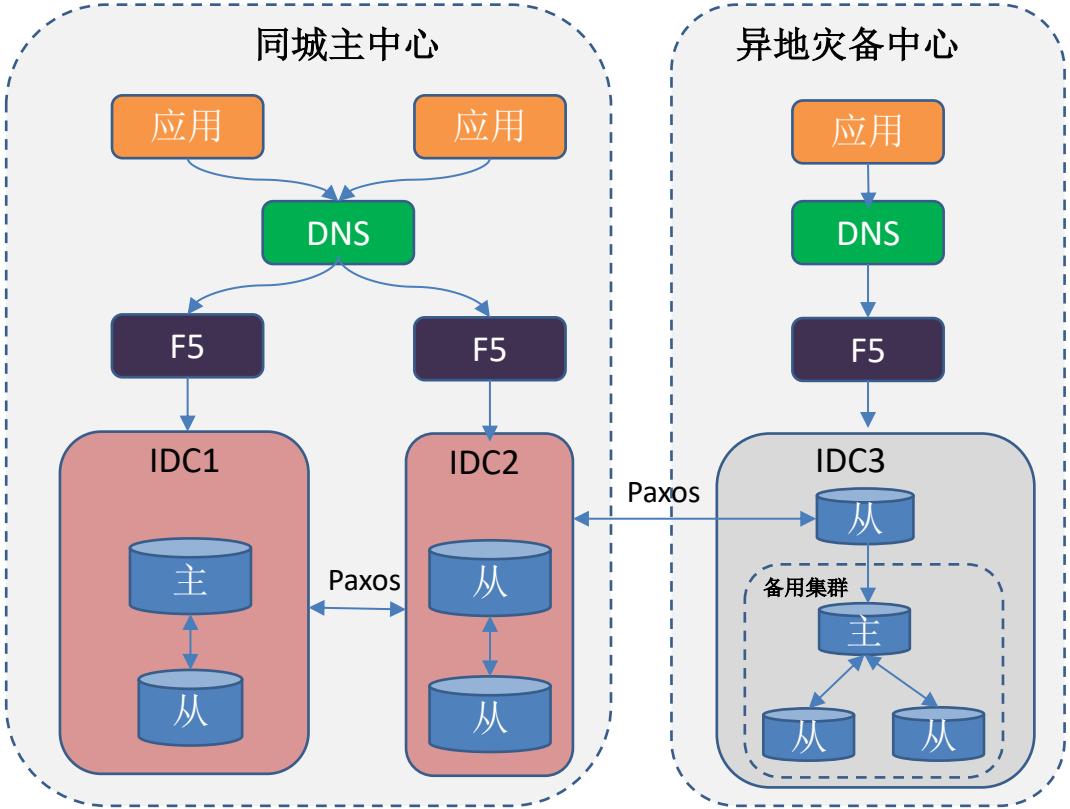
异地备用机群3副本异步复制方式（防止城市机房故障）



分布式技术研究-两地三中心架构

两地三中心同城5副本（221）基于Paxos强同步

异地备用机群3副本异步复制方式（防止城市机房故障）



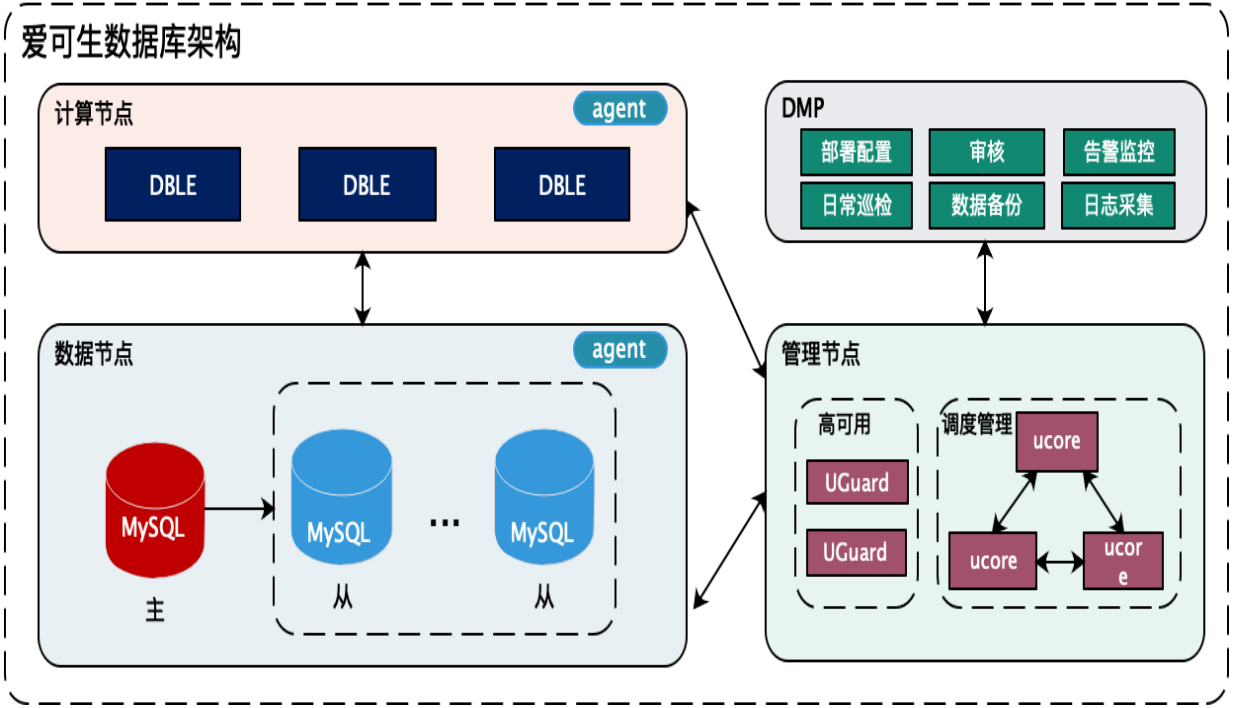
分布式数据库推广

计算节点（DBLE）：负责SQL解析、路由与聚合计算，数据分片；

数据节点：采用原生MySQL数据库，没有做任何的改动；

管理节点：负责存放元数据信息、数据库主从切换与数据补偿机制等

DMP（管理运维平台）：提供自动安装部署，告警监控，备份，巡检、节点管理等功能

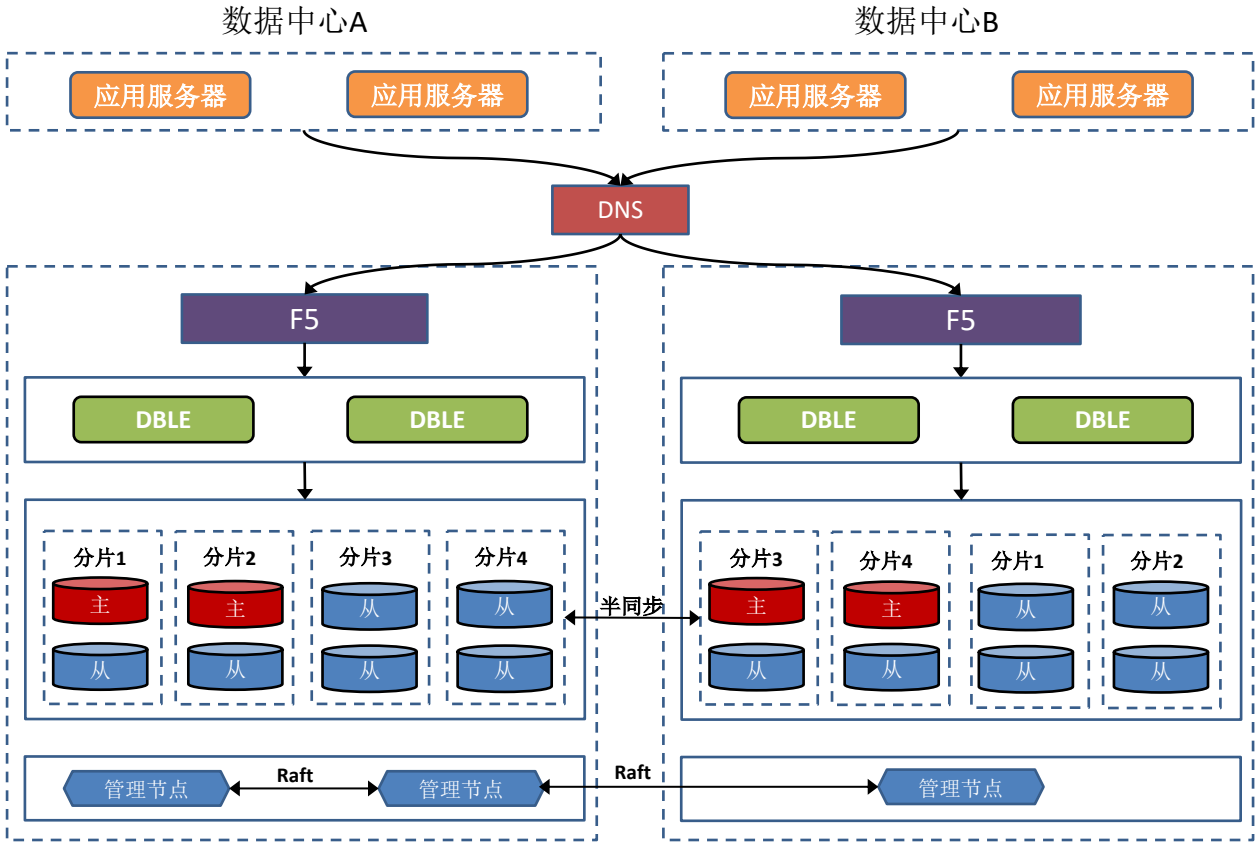


分布式数据库推广

场景 1：

同城双活：应用+数据库层 双活架构，OLTP，分库垂直拆分；中间业务系统，2018年上线；日均交易量100万笔，平均耗时10ms，tps500 ，

- 采用了DNS两层端口探活机制，负载均衡及中间件层联动切换。
- 中间件层无状态多活，可动态扩展。
- 无损复制技术、数据补偿技术，最大程度保障数据一致性。
- 高可用管理模块，故障自动切换。

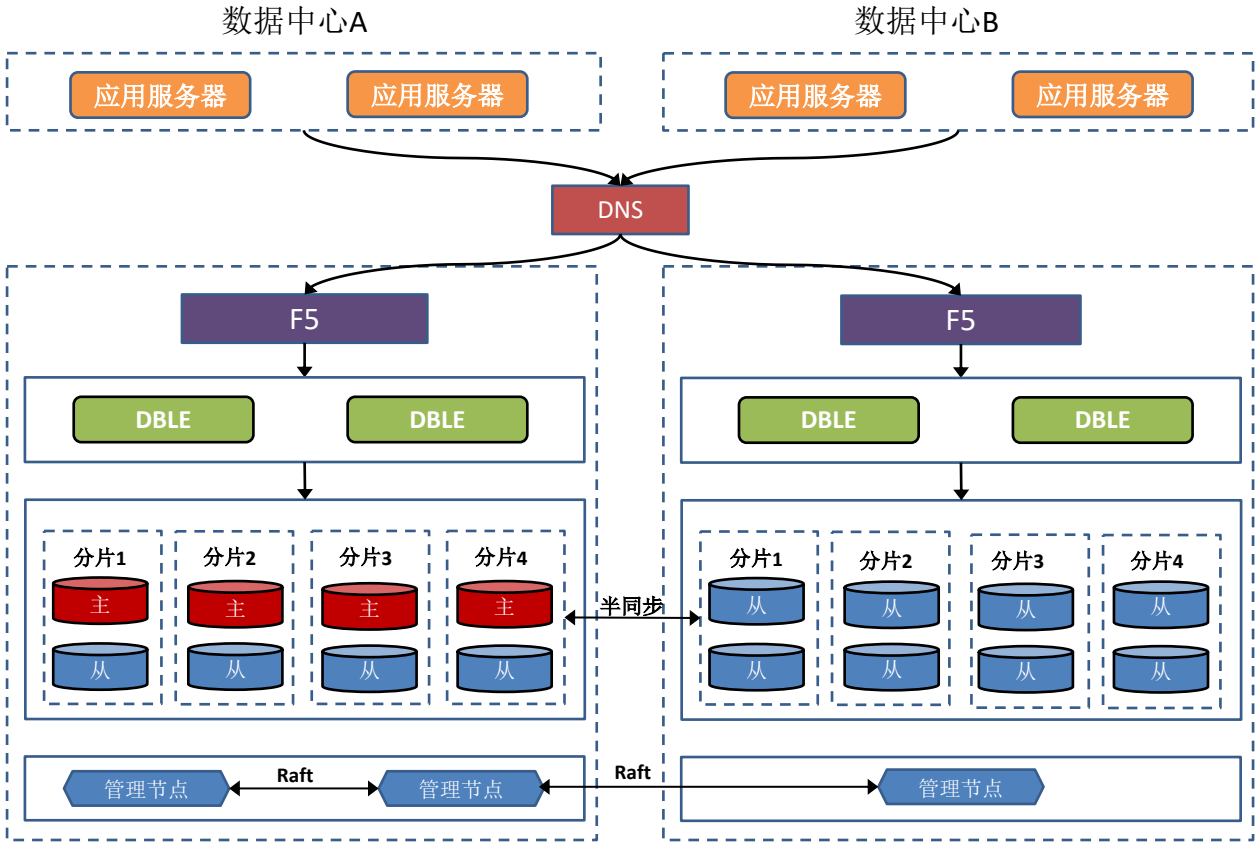


分布式数据库推广

场景 2：

同城双活：应用双活+数据库高可用架构，OLTP，水平拆分；支付类系统，2019年上线；日均交易量100万笔，平均耗时100ms，tps400，

- 采用了DNS两层端口探活机制，负载均衡及中间件层联动切换。
- 中间件层无状态多活，可动态扩展。
- 无损复制技术、数据补偿技术，最大程度保障数据一致性。
- 高可用管理模块，故障自动切换。

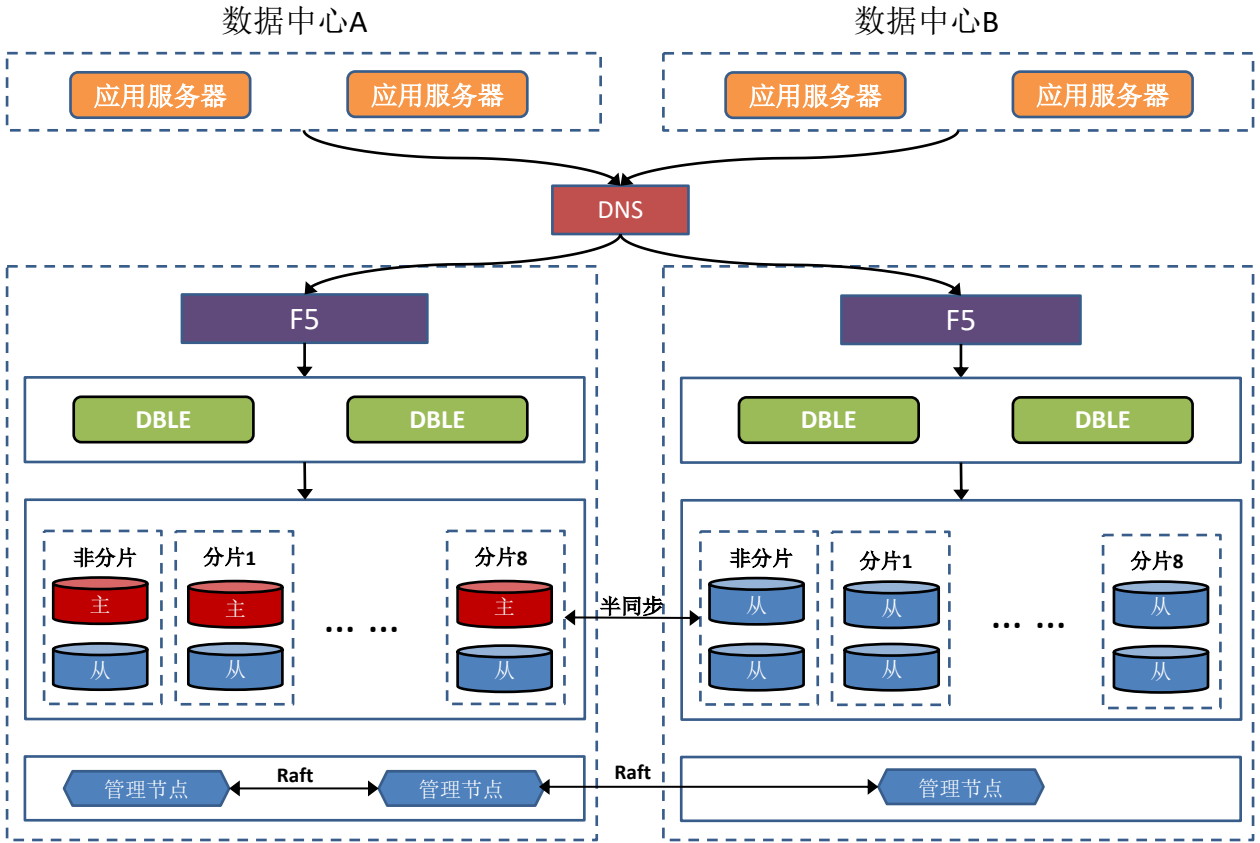


分布式数据库推广

场景 3：

同城双活-应用双活+数据库高可用架构，混合场景，水平拆分；账管类系统，2020年上线；平均耗时100ms，tps400 ，qps1000。

- 采用了DNS两层端口探活机制，负载均衡及中间件层联动切换。
- 中间件层无状态多活，可动态扩展。
- 无损复制技术、数据补偿技术，最大程度保障数据一致性。
- 高可用管理模块，故障自动切换。



应用场景

场景	耗时（秒）		对比
	Oracle 配置：物理机 32C+64G+SSD	DBLE+MySQL5.7 配置：虚拟机 8C16G+SSD（8分片）	
文本导入9千万数据	1517	582	较大提升
单表count	3.35	3.93	持平
单表点查询	<0.01	<0.01	持平
两表关联（带条件）	<0.01	<0.01	持平
两表关联（全表）	38	42.06	少量下降
三表关联（带条件）	<0.01	<0.01	持平
三表关联（全表）	72	72.94	持平

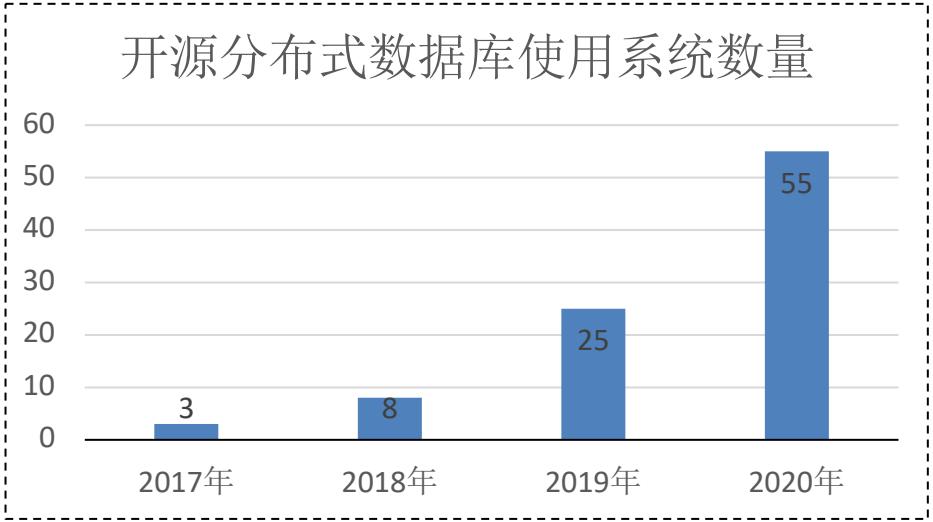
该项目因为涉及到AP场景，查询较为复杂，且DBLE本身会对分片场景有语法支持限制，这对开发前期代码逻辑质量及表结构设计要求较高。该方式对开发人员要求较高，需要相关查询都采用分片健，避免复杂SQL语句。虽然已经实现并上线，运行稳定，满足性能要求，但建议类似的场景采用HTAP数据库会降低开发成本，同时对复杂语句支持更好

推广成果

经过近几年的推广，我行在开源、分布式数据库方面得到了长足的发展，业务系统占有率由原来的2% 提升至 15% 以上

MySQL集中式+分布式

2017 年至今投产采用 MySQL 开源数据库的系统 46 个；采用爱可生MySQL 分布式数据库架构的系统有 10 几套，其中有 多个系统是从 Oracle 数据库迁移过来的，未来会逐步替换

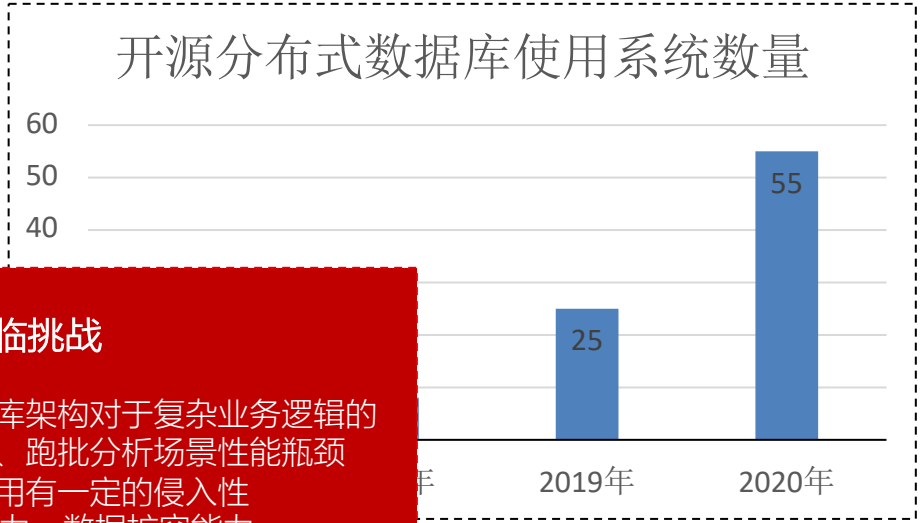


推广成果

经过近几年的推广，我行在开源、分布式数据库方面得到了长足的发展，业务系统占有率由原来的 2% 提升至 15% 以上

MySQL集中式+分布式

2017 年至今投产采用 MySQL 开源数据库的系统 46 个；采用爱可生MySQL 分布式数据库架构的系统有 10 几套，其中有
Oracle 数据库迁移过来的，



面临挑战

1. 基于现有分布式数据库架构对于复杂业务逻辑的局限性；高并发交易、跑批分析场景性能瓶颈
2. 语法限制较多，对应用有一定的侵入性
3. 分布式事务的处理能力、数据扩容能力

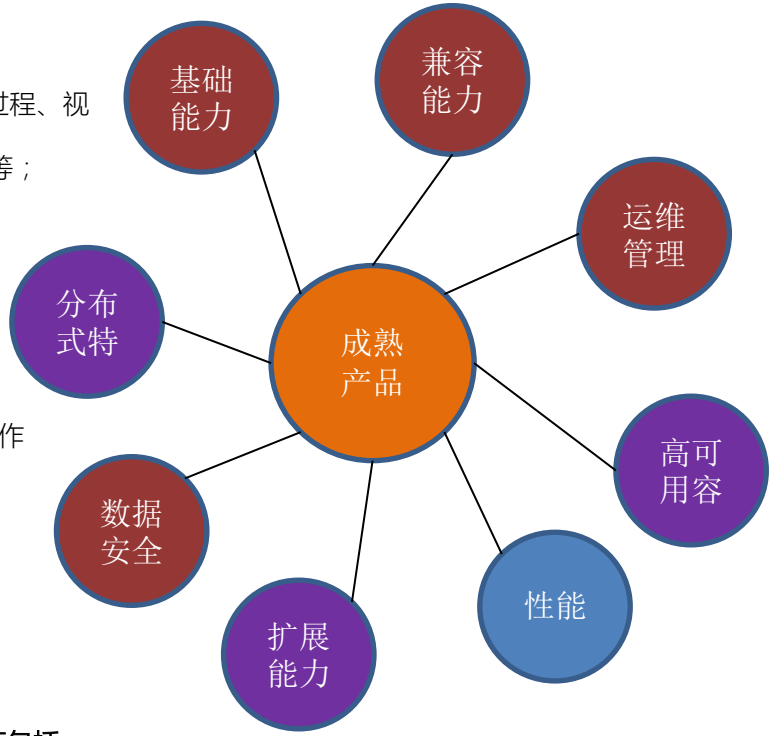
数据库测评指标

1. 基础能力
包括标准SQL与对象（函数、存储过程、视图等）、数据类型等；
多租户、字符集、数据分区、约束等；
查看元数据、执行计划等信息

2. 分布式特性
数据分布、读写分离、分布式事务

3. 数据安全能力
权限管理、口令安全、身份认证操作审计、数据加密、流量控制

4，扩展能力
计算负载均衡，数据分布均衡能力
数据节点在线扩缩容能力



5. 兼容能力
常用连接方式的兼容（JDBC等）
对Oracle、MySQL语法的兼容性
对X86，ARM等主流、国产软硬件的兼容性

6. 运维管理能力
包括安装部署与升级，配置管理，日常巡检，监控告警，数据备份恢复，集群节点管理，异构数据迁移等

7. 高可用容灾
针对各节点出现宕机，网络、磁盘故障，CPU、内存使用率过高等场景的高可用能力

8. 性能
业务数据多表关联、批量交易测试
Sysbench、TPCC模型测试

一个成熟的产品成还包括：

产品生态、商用年限、相关认证；技术支持体系，产品手册完善性；行业相关案例等指标

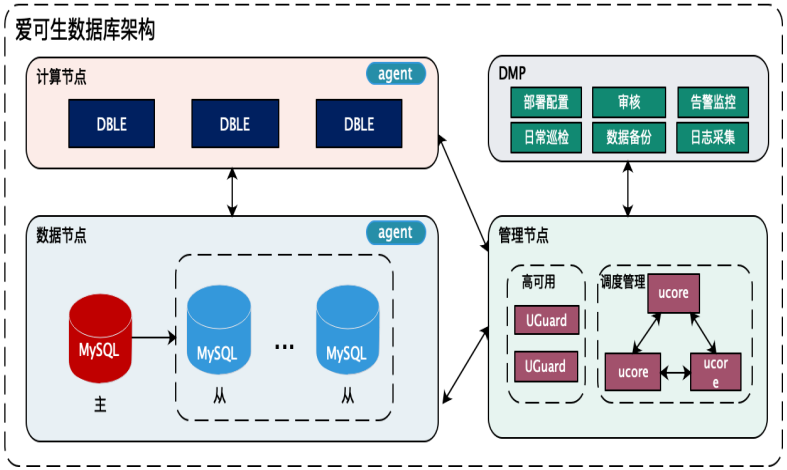
数据库测评指标

对运维友好，对开发透明，对架构可用

测评结果

测试项	结果
基础能力	50%不支持存储过程等特殊对象、语法和DBLINK
兼容性	50%不100%支持MySQL语法，所有都无法100%支持Oracle语法
分布式特性	50%不支持更新分片键，20%不支持读写分离
数据安全	40%不支持数据加密功能
运维管理	50%不支持一致性备份，运维管理功能不完善
扩展能力	50%不支持缩容
高可用	绝大部分支持较好
性能	大部分TPCC（400并发 1000仓库 6分片）TPMC>10w
商用年限	50%不满足5年时间

未来规划



应用场景

- 数据增长是可估算的，扩容频率较低
- 业务逻辑简单，拆分规则明确
- 无分布式事务或极少的分布式事务

应用场景

- 数据增长不可估算的，未来扩容不确定性
- 业务相对复杂，拆分规则复杂，较多计算与跑批操作
- 存在分布式事务

Gdevops

全球敏捷运维峰会

THANK YOU !