DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

云原生数据库技术内幕

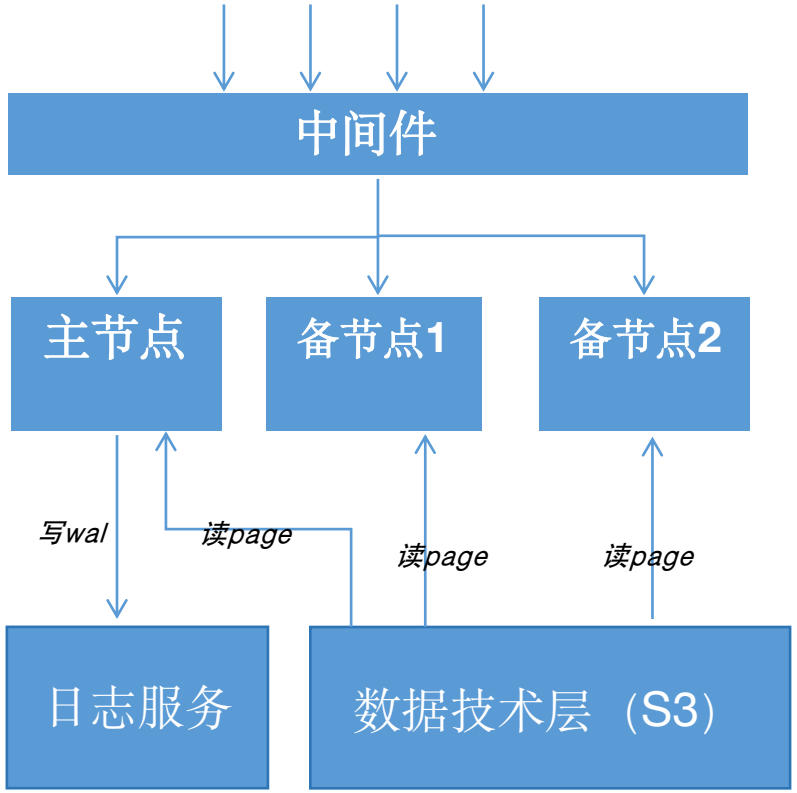
移动云 云原生数据库产品研发负责人 薛港

云原生数据库的设计原则

移动云He3DB是受Aurora论文启发，基于pg跟mysql自研的云原生数据库产品。总体设计遵守以下几个原则：

- **低破坏**：对PG跟MYSQL内核破坏尽可能低
- **高性能**：综合性能对齐RDS，没有短板，部分指标优于RDS
- **低成本**：RDS成本的70%（一主两备，100G以上数据量）
- **简单好用**：订购简单，无运维

云原生数据库的架构设计



计算节点：

支持一主多备，无状态。所有写转化为日志写（log is database），读请求从S3获取基础版本page，基于每个节点当前最新apply wal lsn,回放成不同版本的page

日志服务

存储wal 日志，通过一致性算法保证数据可靠性，只存储近期少量的wal日志，因此对容量要求不高，强调低delay,高吞吐能力。

数据技久层（S3）：

持久存储wal 以及Page多版本数据，实现共享存储

中间件：

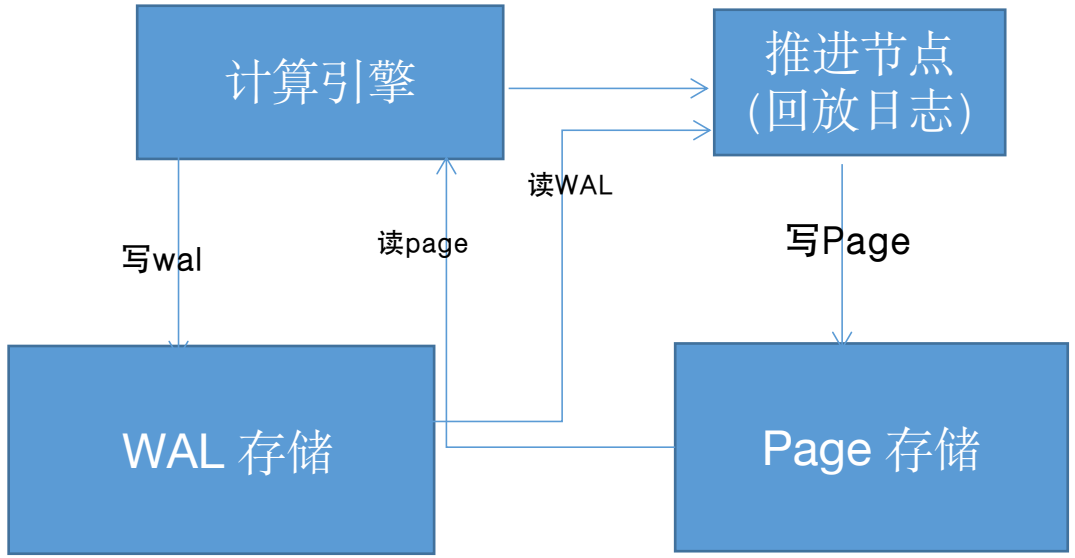
读写分离（保证读一致性），负载均衡，数据分片

云原生数据库的核心技术

log is database核心技术

设计目标：

- 减少写操作产生的数据量
- 提升写吐吞
- 基于WAL 日志异步回放wal生成Page数据



相关技术：

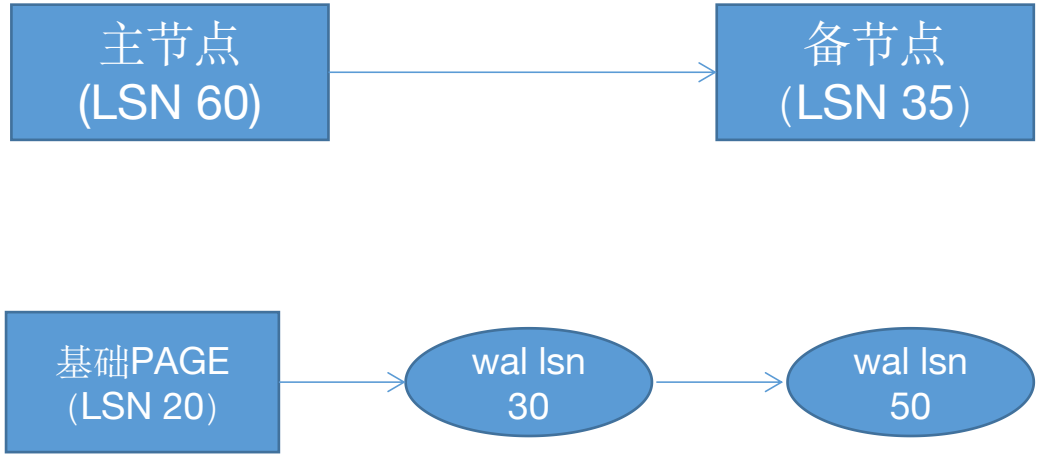
- 日志模式改造成KV 模式
- 并行提交，串行确认
- 所有的写转化为WAL 日志写
- 关闭Full Page Write,降低日志数据量
- 引入分布式KV作为WAL持久层
- 推进节点异步回放wal生成page数据

云原生数据库的核心技术

共享设计的核心技术

设计目标：

- 主备共享一份数据，并且主备能够读取不同版本的Page数据
- 主备之间wal delay绝对可控



相关技术：

- 存储层保存最慢备节点对应的基础数据Page ，以及对应的WAL 日志
- 计算引擎内存中缓存Page 的链表关系，提供多版本page读取功能
- 当读取特定版本Page 时,能够根据基础Page +WAL 日志回放到特定版本的Page

云原生数据库的核心技术

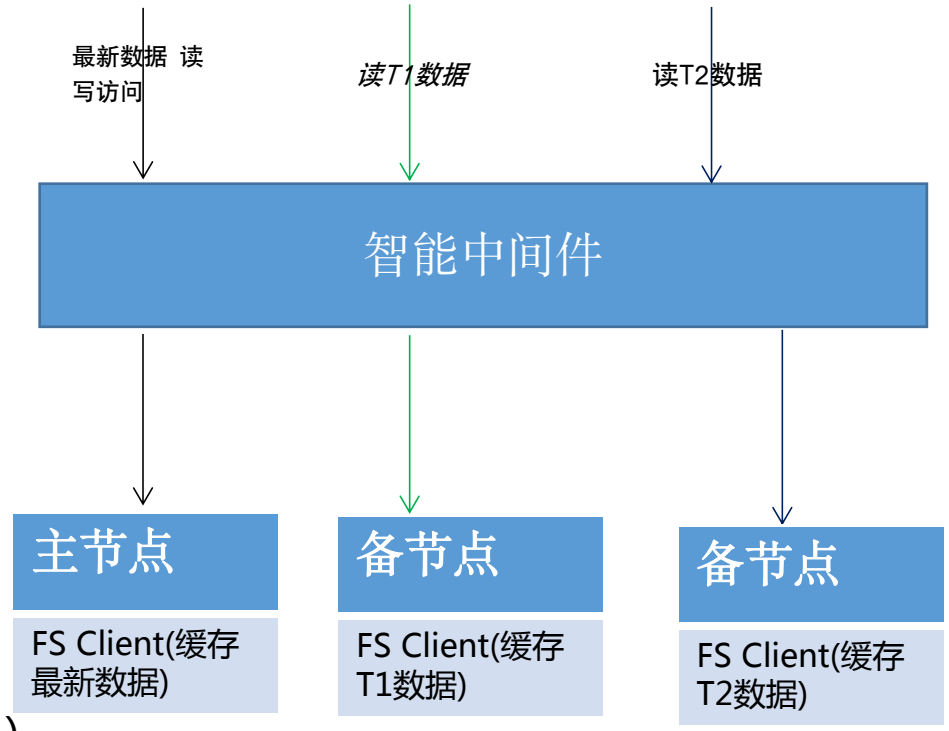
存储设计核心技术：

设计目标：

- 实现用户态类文件系统，降低对PG，Mysql存储模块的破坏性
- 读性能优化

相关技术：

- 冷热分层
- 智能中间件
- 数据负载分区（联邦内存池设计）
- 每个节点维护自身视角的文件系统，实现文件系统元数据管理（内存中）以及用户数据缓存在本地盘以及s3

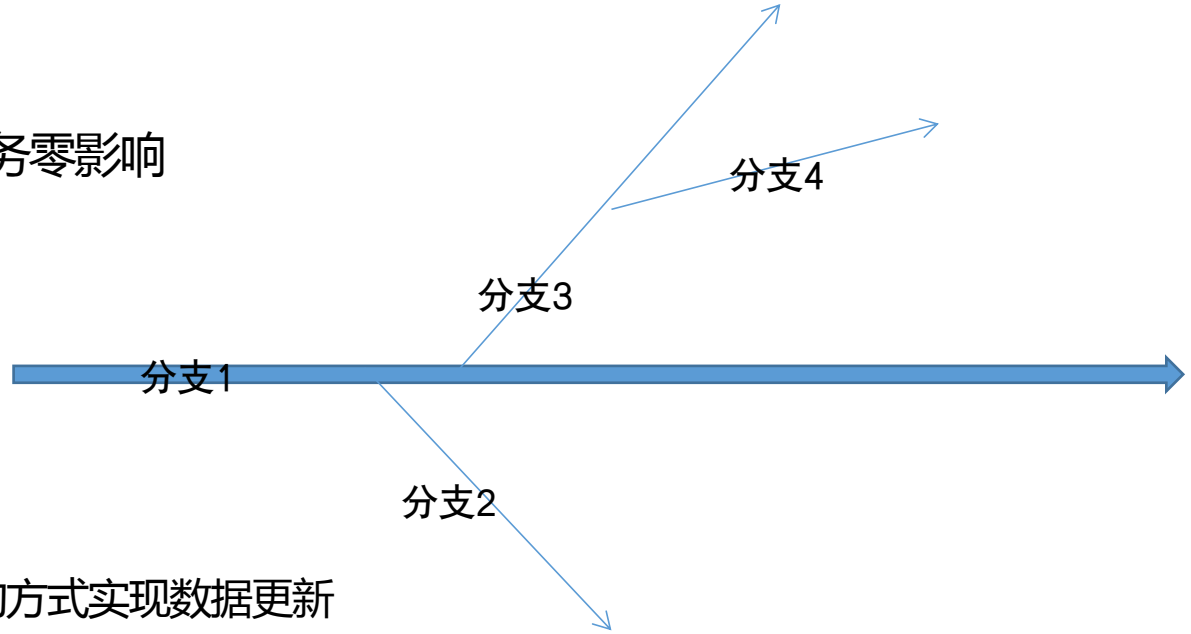


云原生数据库的核心技术

备份，恢复，clone 相关核心技术

设计目标：

- 不依赖于数据量大小，秒级完成备份，并且对业务零影响
- 秒级实现基于备份数据恢复实例
- 所有实例数据，备份数据， 统一纳管
- clone 实例跟原始实例实现数据共享



相关技术：

- 实现分布式存储，使用S3作为持久层，append only的方式实现数据更新操作
- 使用wal日志的lsn点作为快照点，通过copy on writer 的技术实现快照
- 数据加密以及压缩

云原生数据库对用户带来的改变

使用方式：提供serverless数据库服务，实现按量计费，按需扩缩容，零运维

备份：秒级完成备份，业务零影响，少量数据量增长开销

RTO时间：RTO上限绝对可控

备机管理：更快速加减备节点，读写分离价值放大（主备wal delay更低）

容量更大：s3作为持久层，支撑无限容量

性能更高：读写性能提升明显

总结

移动云云原生数据库的终极目标:在公有云上提供简单，便宜，好用的数据库服务。

简单：以serverless的方式提供服务，实现按需扩缩容，用户不用关心它的业务需要的资源

便宜: S3作为持久层，使用数据分层压缩，快照（copy on writer）等技术,实现所有资源按量计费

好用：对数据库而言，好用等价于：高性能，高容量，高可靠，高可用等能力。 He3DB 使用本地盘缓存设计， S3作为无限容量技久层，引入智能中间件，以及高性能KV存储层(wal) 实现好用目标

THANKS