TencentDB

CynosDB for PostgreSQL
一主多读架构

孙旭

腾讯云 高级工程师

TencentDB

SPEAKER

孙旭

腾讯云 高级工程师

10年数据库内核研发经验，熟悉PostgreSQL、Teradata数据库内核，熟悉数据库的查询优化、执行、事务并发以及存储等子系统；对分布式数据库有深入的研究和研发经验。

目前在腾讯云从事CynosDB数据库研发工作。

TencentDB

内容

- CynosDB for PostgreSQL架构
- CynosDB for PostgreSQL关键设计
- CynosDB for PostgreSQL一主多读架构

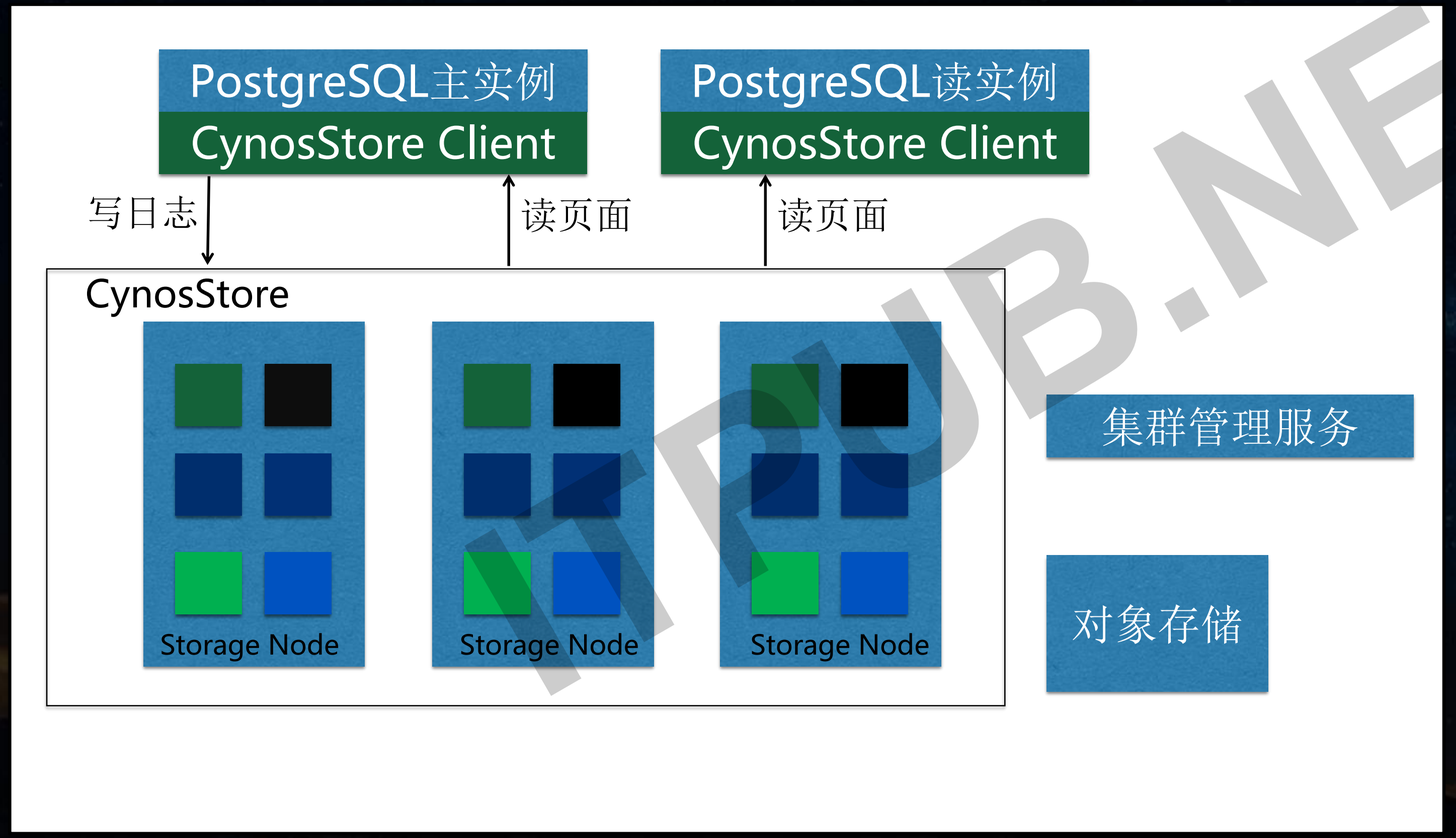
为什么需要CynosDB

- 传统数据库在云上面临的问题
  - 资源利用率低
  - 扩展能力不足
  - 资源规划难
  - 备份难

- CynosDB解决思路
  - 计算存储分离：计算资源弹性调度能力
  - 日志下沉以及异步回放：减少网络IO
  - 共享分布式存储：资源弹性扩展
  - 后台持续日志备份

CynosDB for PostgreSQL架构

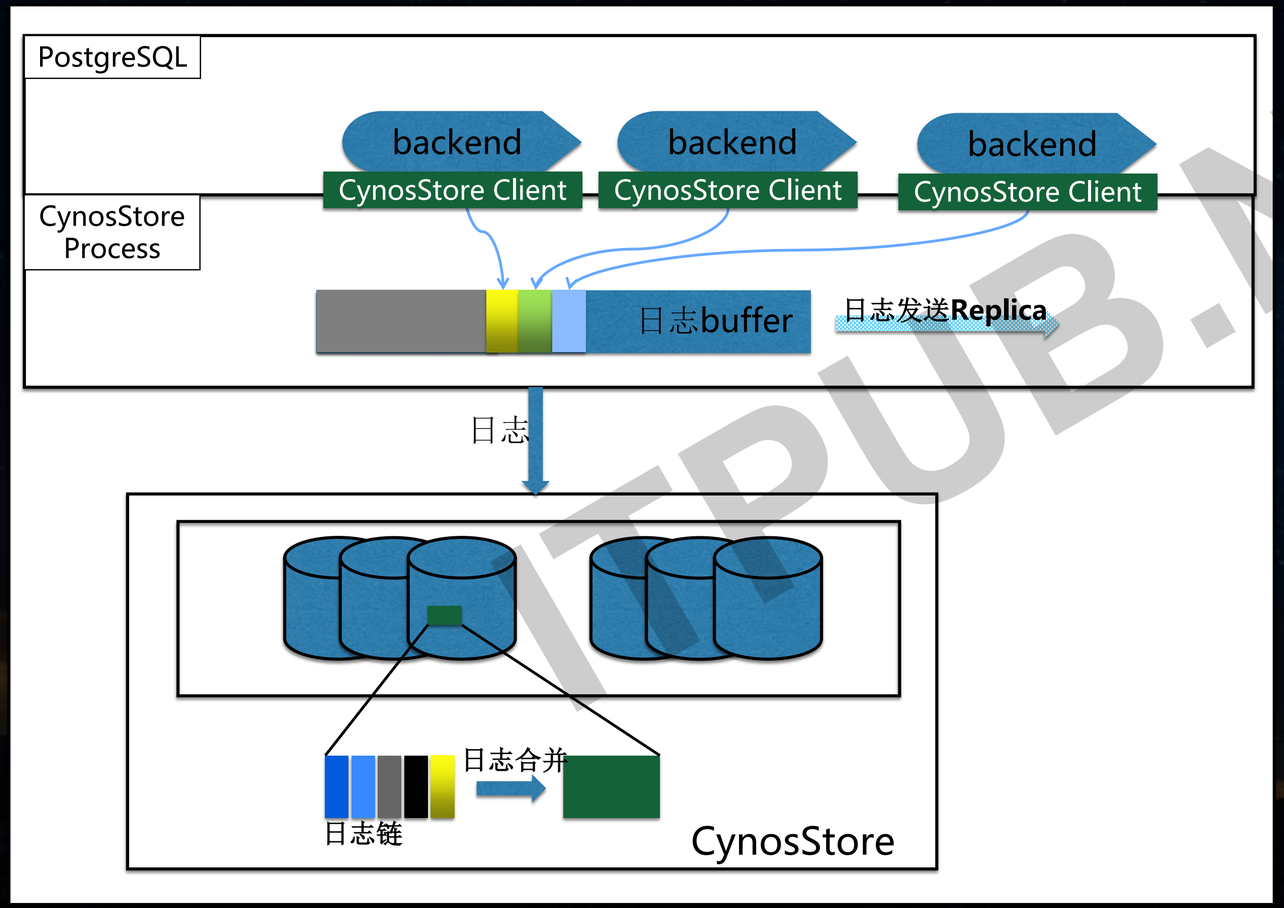
- CynosDB - 云原生数据库



- 核心架构设计
  - 日志下沉
  - 日志异步回放
  - 多版本读（同步）

CynosDB for PostgreSQL架构 - 关键设计

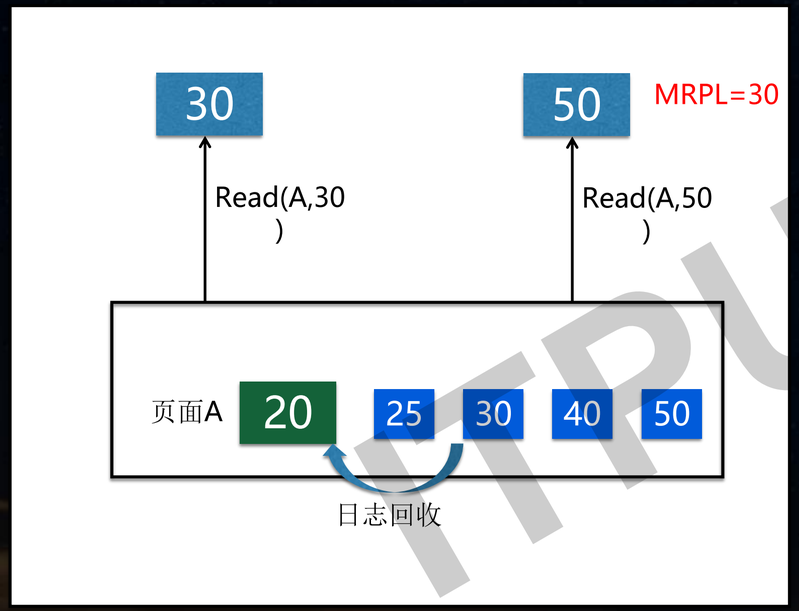
- 日志下沉、异步回放



- 数据原子修改（MTR）
  - MTR（Minimal Transaction Record）
  - CPL（Consistency Point LSN）
  - VDL（Volume Durable LSN）
- 日志异步写入
  - 日志写入buffer
- 日志并行插入

CynosDB for PostgreSQL架构 - 关键设计

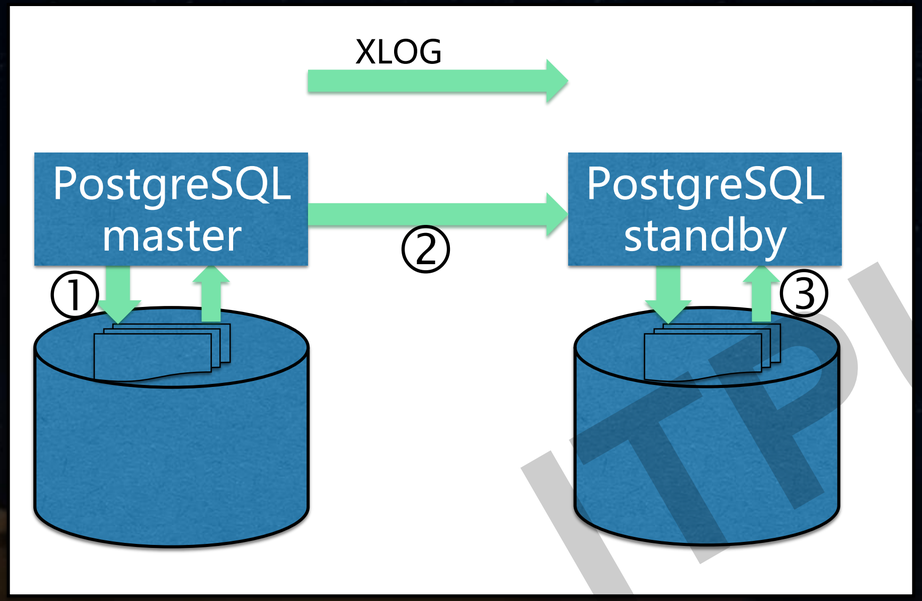
- 多版本读（同步）



- RPL（Read Point LSN）
  - 任何一个VDL都可以是一个RPL
- 日志回收
  - 最小读点（Minimal Read Point LSN）

一主多读

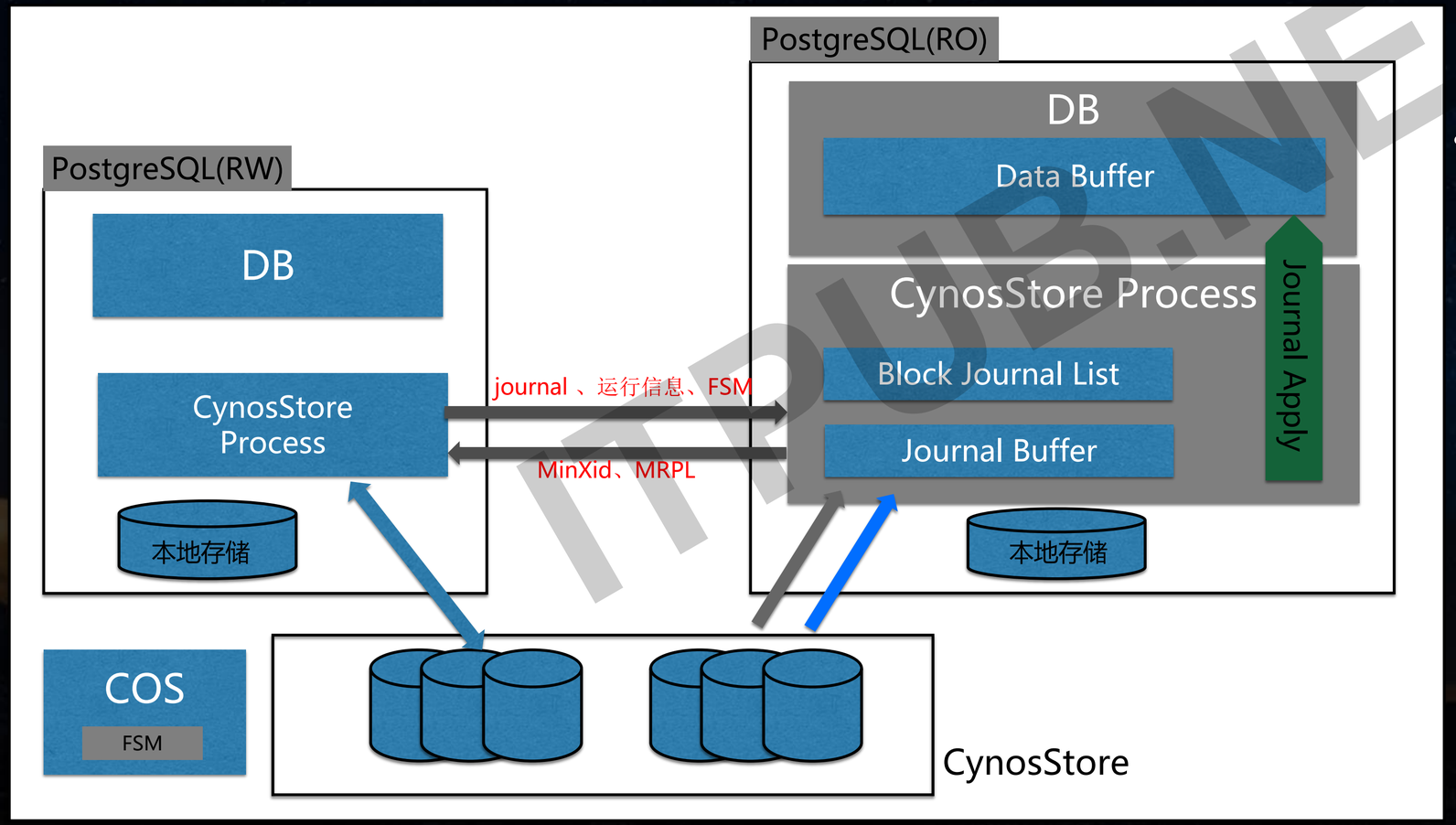
- 传统PostgreSQL主备模式仍有缺点



- 多读（Replica）优势
  - 横向扩展读能力
  - 提升数据库的可用性
- 传统主备模式的问题
  - 创建新备需要拷贝数据： 额外存储资源
  - 备机切换和启动需要恢复大量日志：慢
  - 收到日志需要写磁盘：慢
  - 解决BufferPinLock冲突

一主多读 - 架构

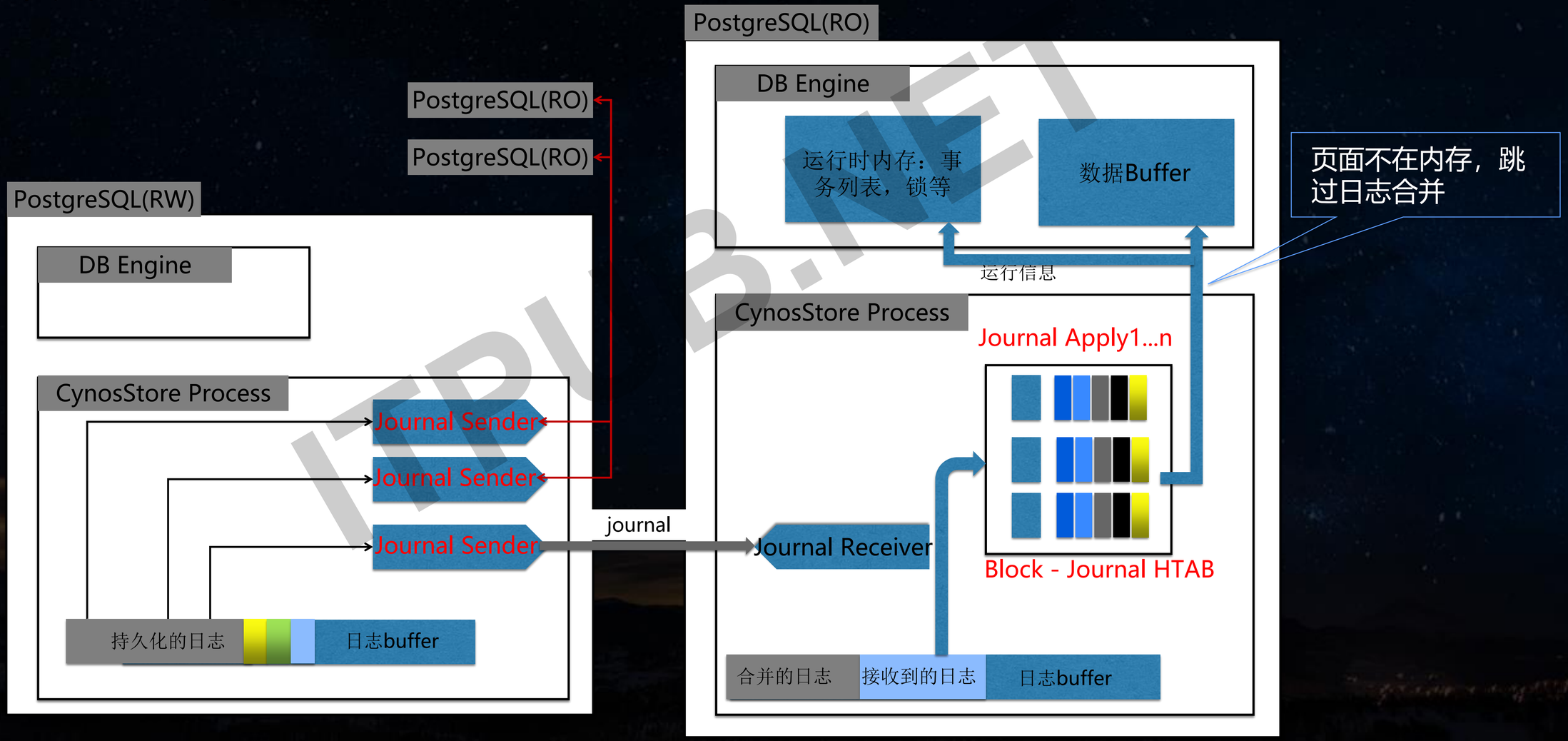
- 扩展系统读能力、快速主备切换



- 核心架构设计
  - Replica本地不存储数据
  - Repica并行恢复日志，不落盘
  - Repica多版本数据Buffer
  - 备份FSM到COS，备份配置文件到CynosStore

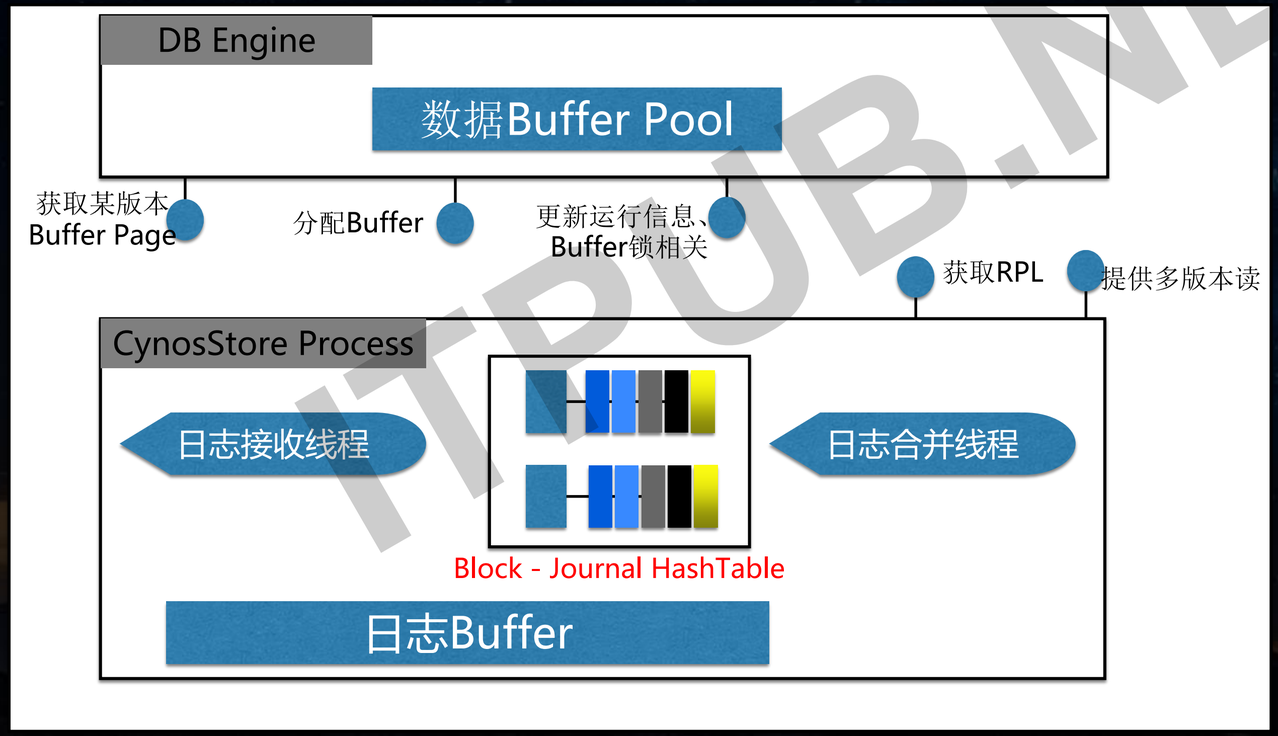
一主多读 - 架构

- 多个读节点、并行日志恢复



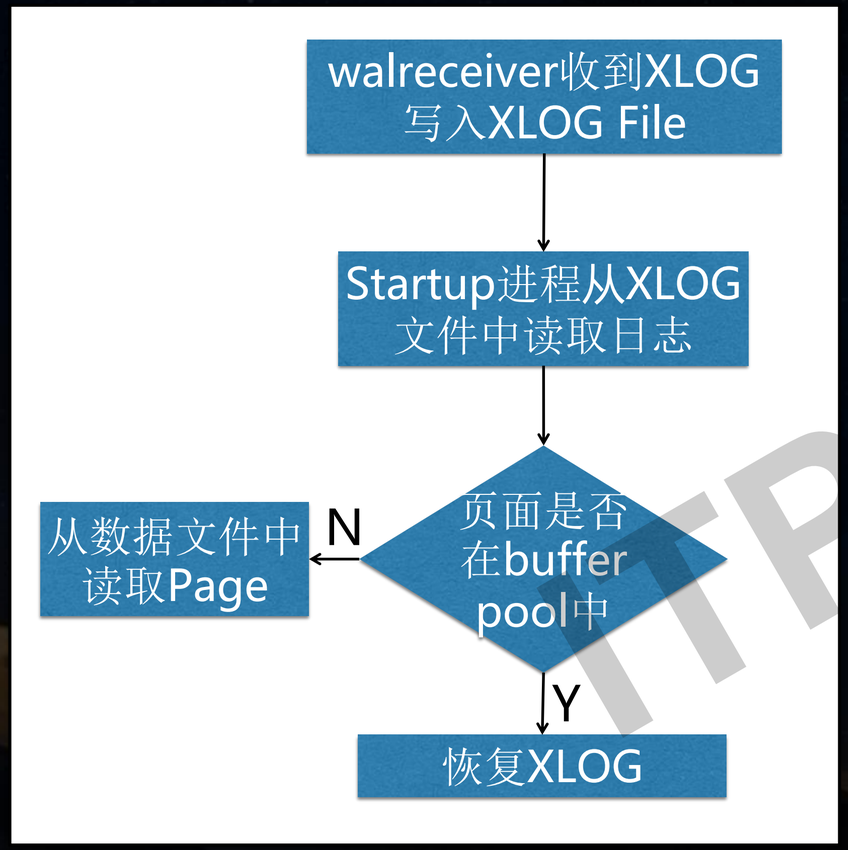
一主多读 - 架构

- CynosStore Process实现连接管理、日志管理
  - DB与CynosStore接口简单
  - DB无需关心日志管理、合并

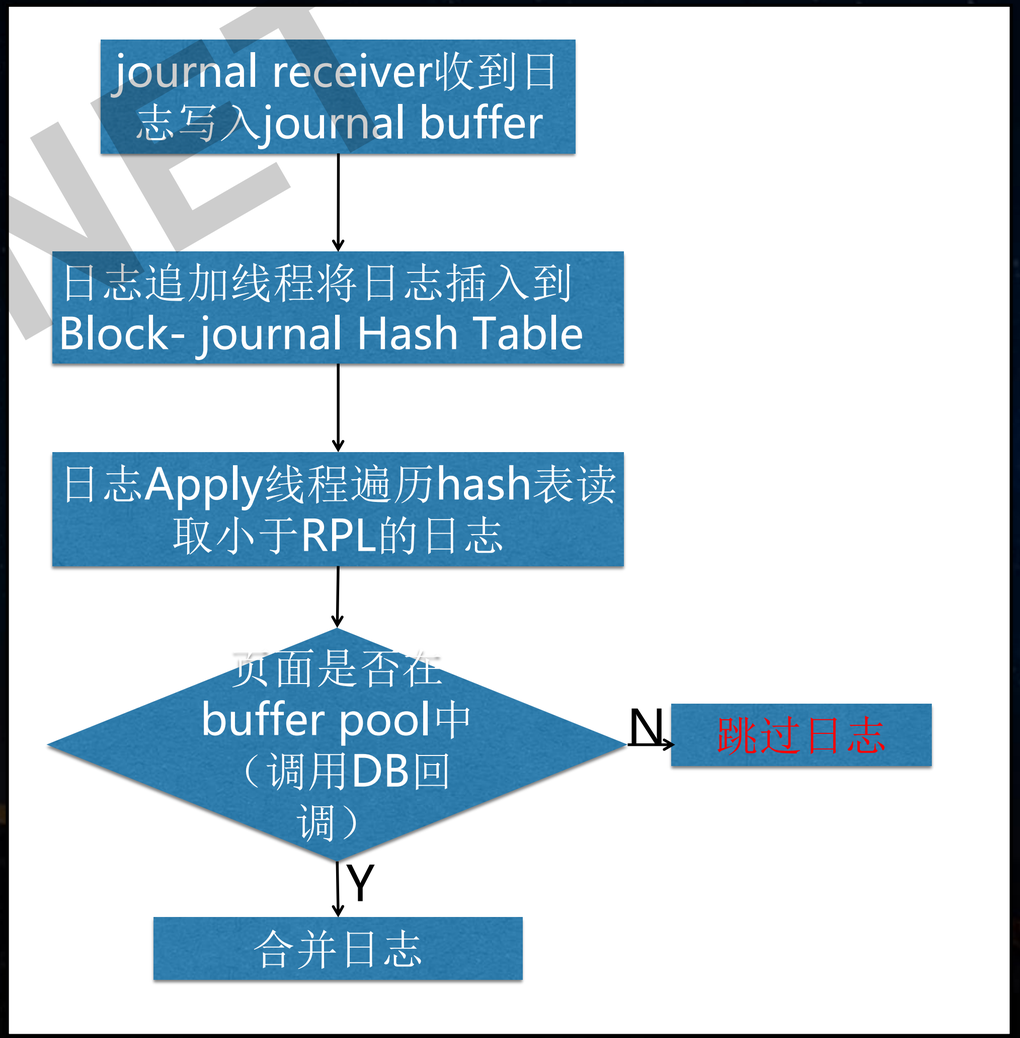


一主多读 - 架构

- 传统PostgreSQL恢复数据Buffer过程

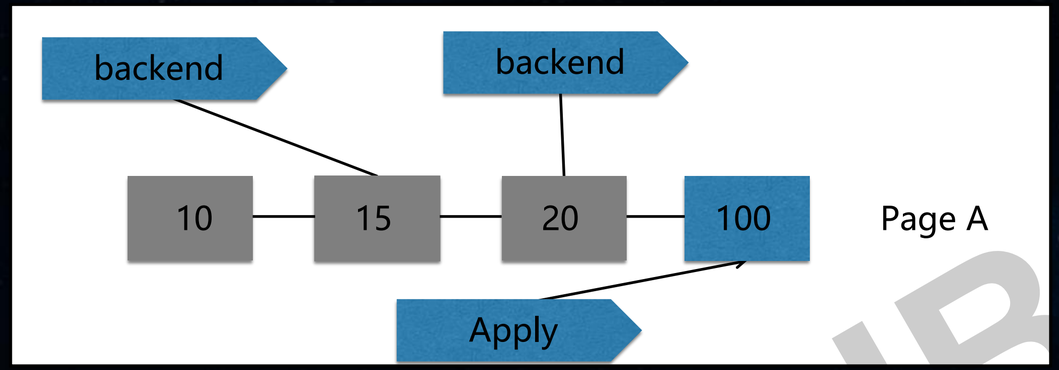


- CynosDB for PG

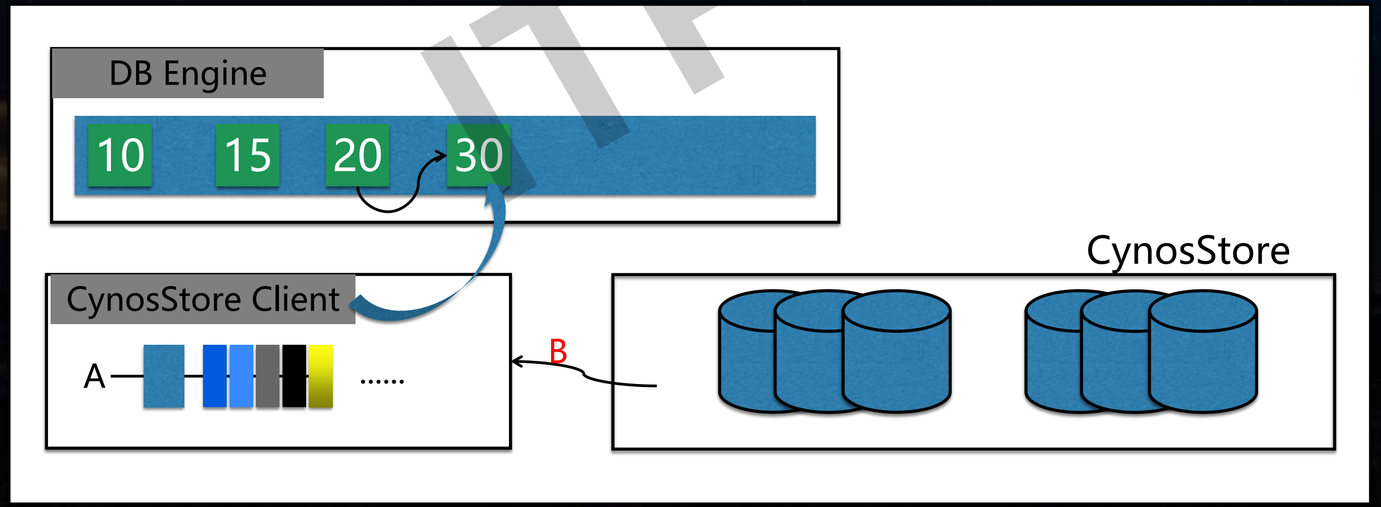


一主多读 - 读页面

- 多版本数据Buffer：事务访问旧版本页面，日志回放与读事务互相不阻塞

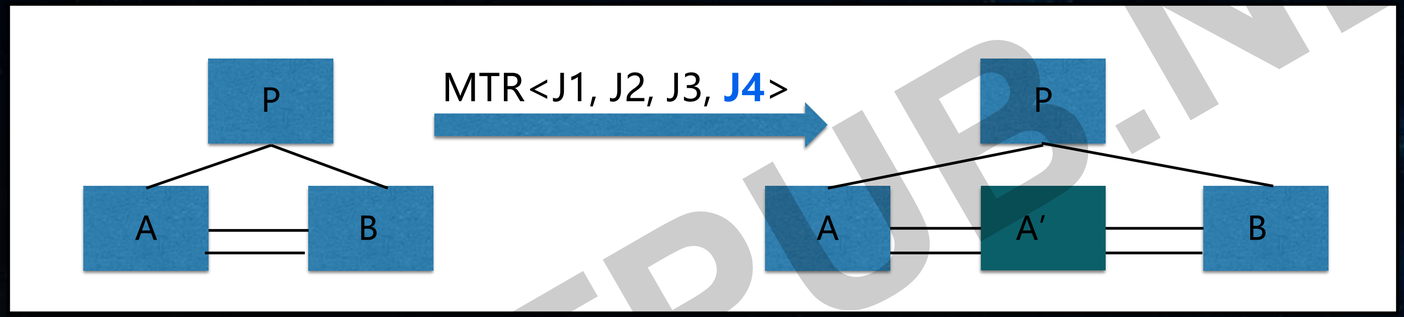


- 以较新版本为基础合并日志，快速返回请求页面



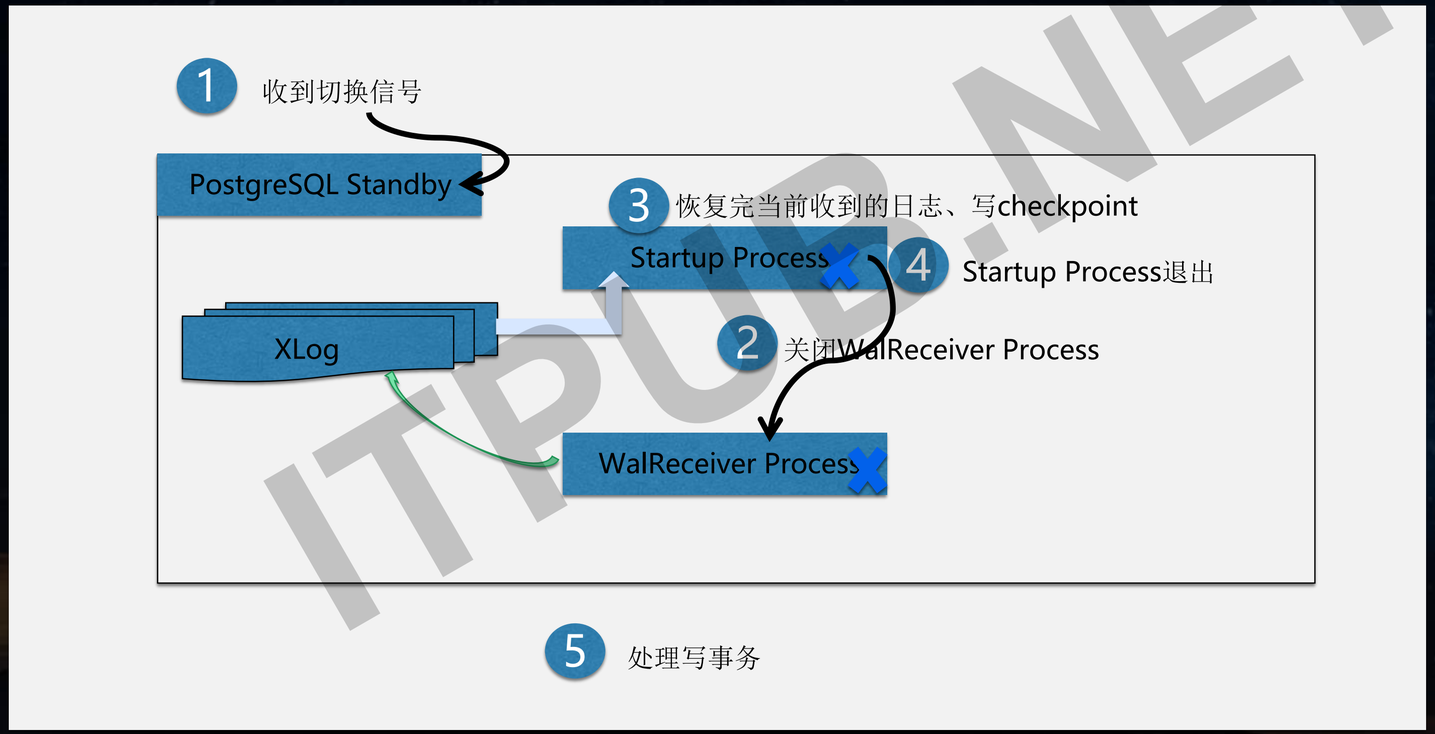
一主多读 - 读页面

- 原子性：Repilca按照MTR粒度进行访问页面，保证对象结构完整
  - 以索引页面分裂为例



一主多读 - 切换

- 传统PostgreSQL切换过程：



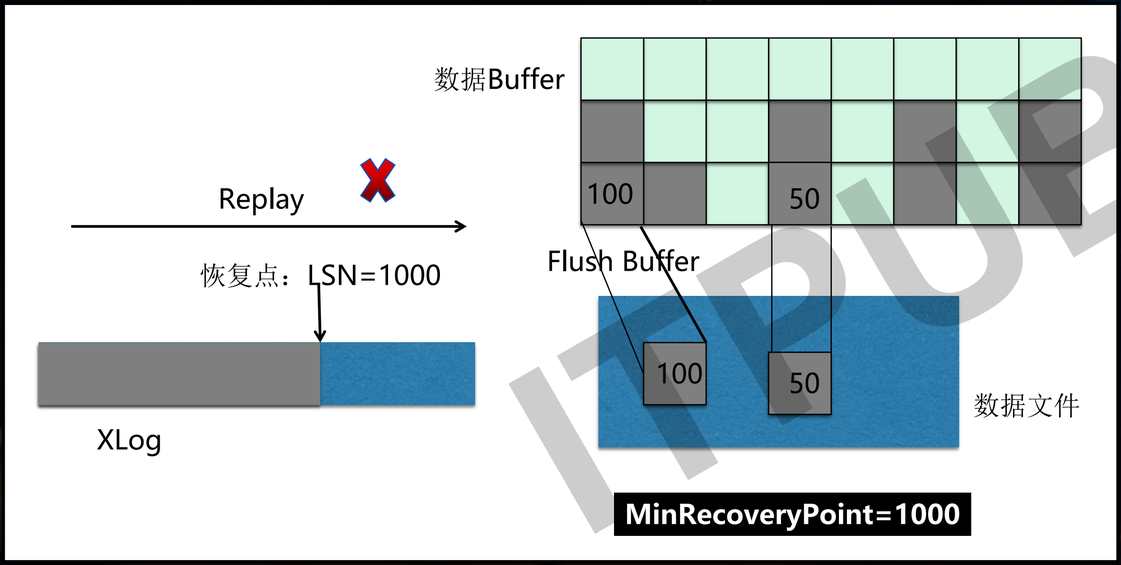
一主多读 - 切换

- CynosDB秒级切换

- HA组中的备实例参与主备切换
  - 备机切换不需要恢复大量日志
- 读实例仅仅参与读事务请求

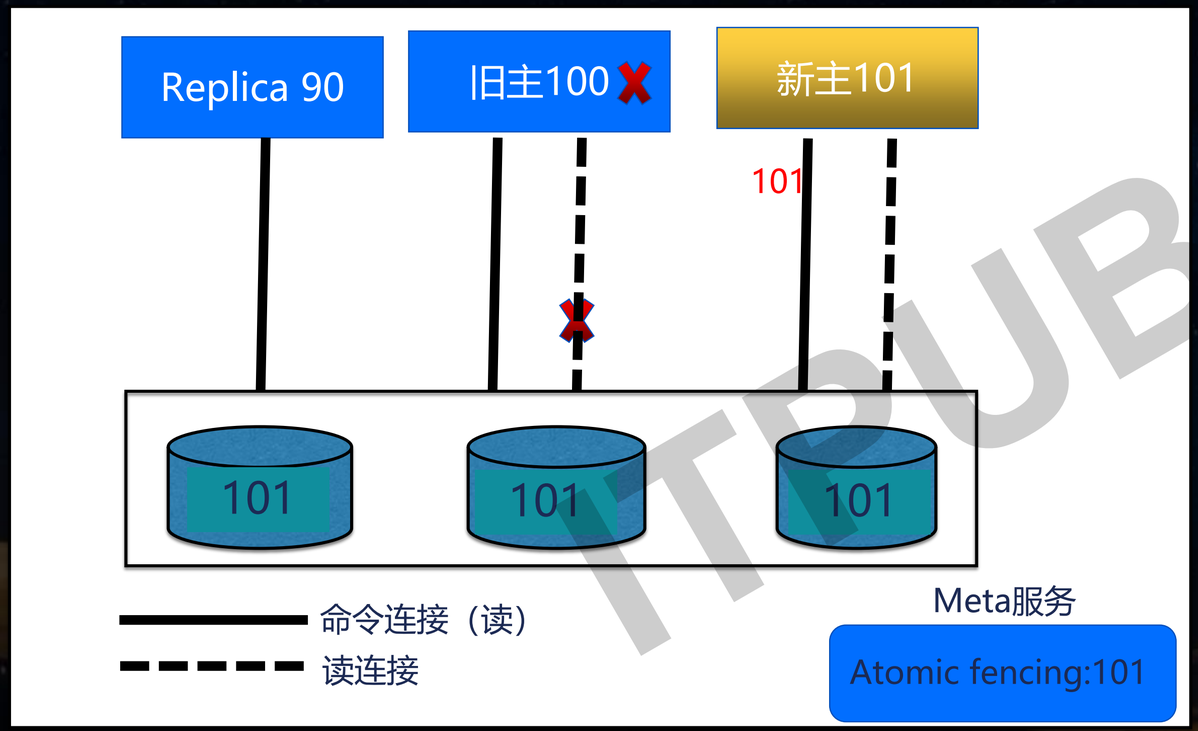
一主多读 - 启动

- 传统PostgreSQL备机的启动
  - 传统备机启动需要恢复到MinRecoveryPoint才能到一致状态



- CynosDB Replica启动需要一个持久化的RPL即可

一主多读 - 防止多写



- HA Fencing机制
  - 数据库启动，获取当前fencing 值
  - 使用获取的fencing值连接到存储
  - 存储将小于fencing的写连接断掉，读连接保持

TencentDB

客户案例

关注“**腾讯云数据库**”官方微信

体验**移动端一键管理数据库**

获取数据库技术干货和最新资讯

立享**10元腾讯云代金券**

TencentDB