Alibaba数据库运维最佳实践

张瑞 @ Alibaba运维部

Alibaba数据库发展历程

成长的烦恼

- 从PC服务器到小型机
- 集中式ORACLE数据库
- 可用性依赖高端硬件
- 性能无法线性扩展

解决方案

- **扩展性与高可用**
  - 分布式MySQL数据库集群
- **数据同步解决方案**
  - 基于日志解析的数据同步
- **提升数据库性能**
  - SSD高性能数据库集群

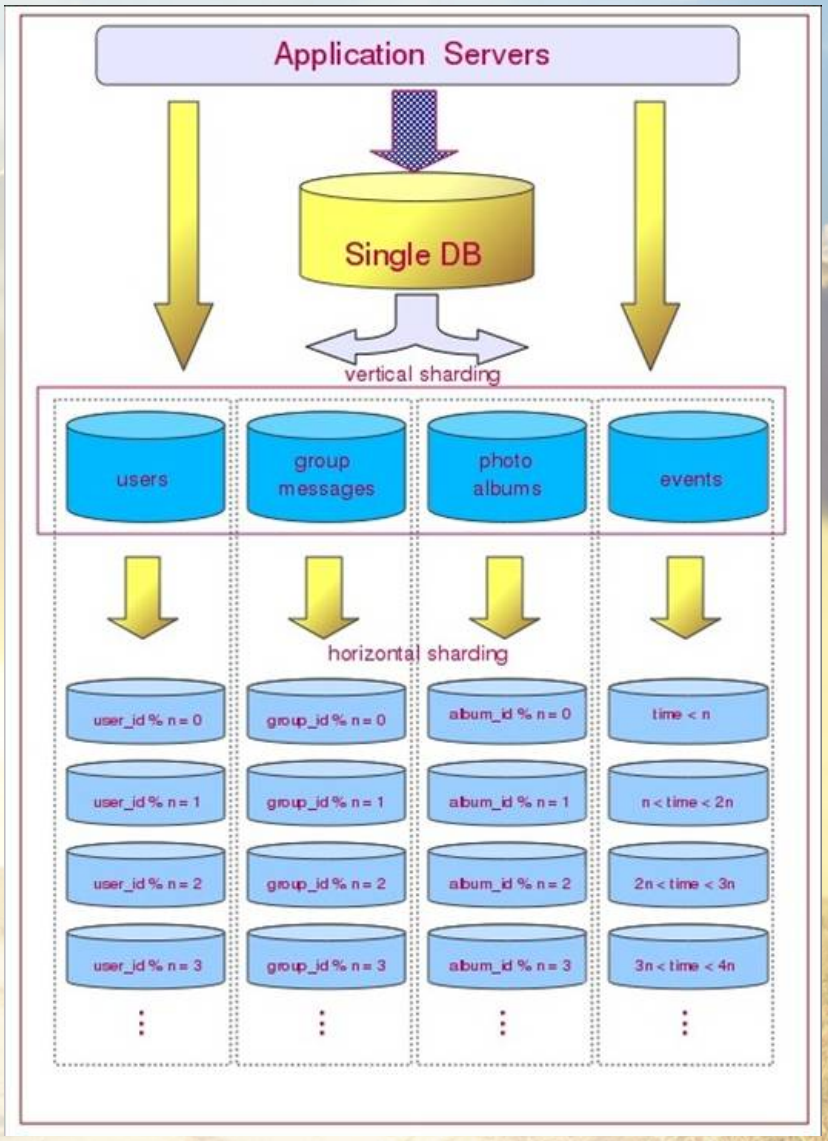

Alibaba分布式数据库

分布式数据库架构

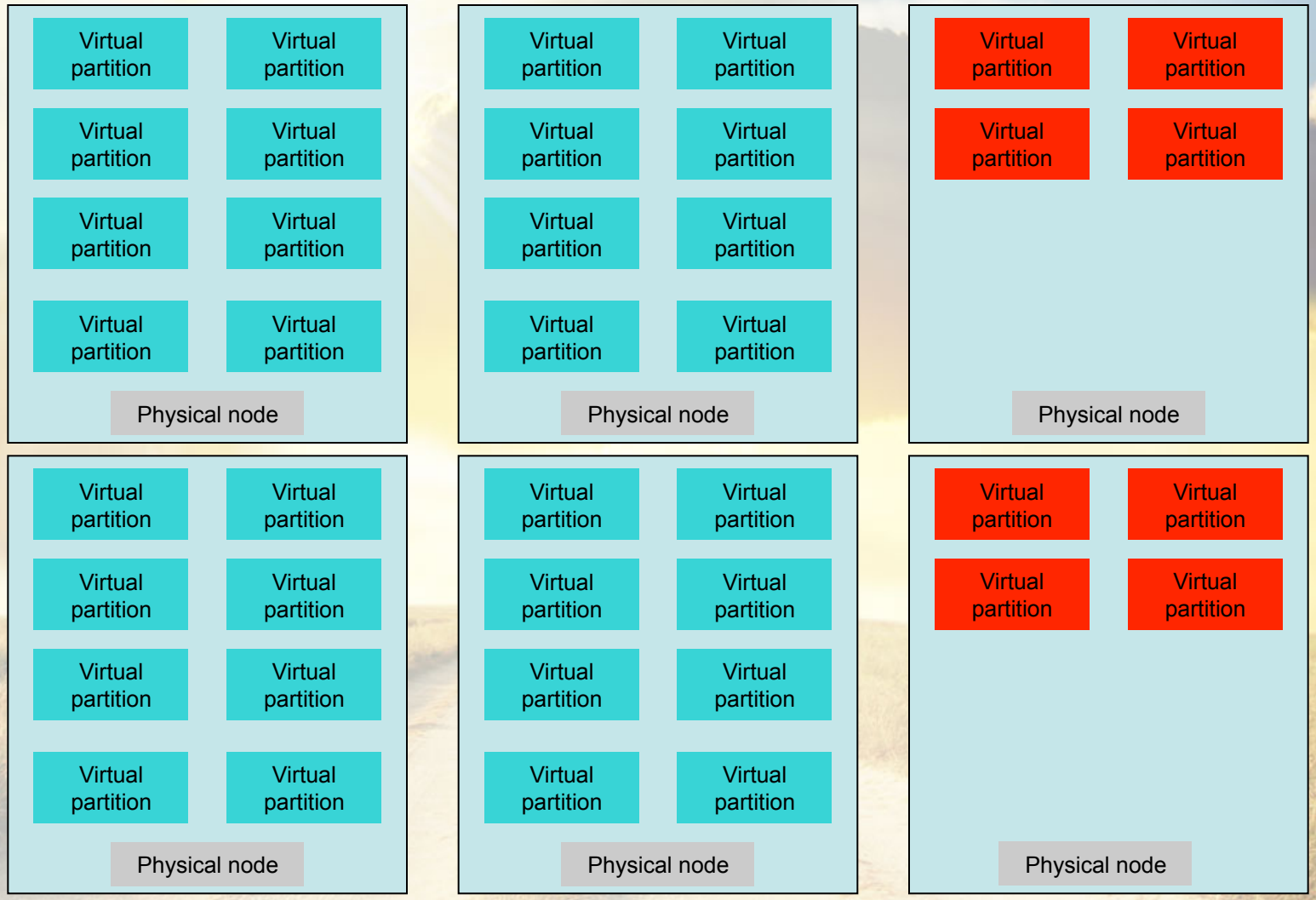
概述
- MySQL数据库
- 应用透明访问
- DB Proxy
- 功能分区
- 数据分片
- 高可用，可扩展
- 性能与运维

缺点分析
- 应用受限
- 关系型弱化
- 不支持事务

功能增强
- 跨节点Join
- 排序分页

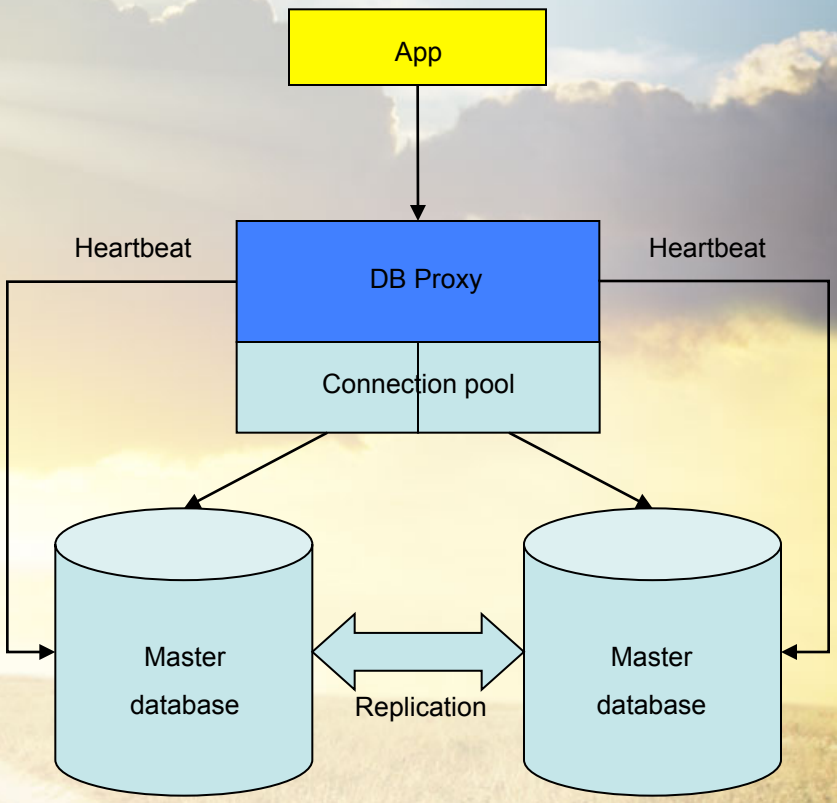


数据切分策略-Virtual Partition Hash



MySQL高可用方案

- 基于MySQL复制
- 应用透明访问
- 探测机制简单
- 切换快，影响小
- 配置灵活，节省资源
- DRBD+Heartbeat？



数据库高可用与硬件选型策略

ORACLE

- 集中式架构
- Shared-disk
- HA方案：商业软件
- 硬件选型策略
  - 小型机，4路高性能PC
  - 性能，扩展性，可用性
  - 外接存储设备

MySQL

- 分布式架构
- Shared-nothing
- HA方案：DB Proxy
- 硬件选型策略
  - 2路PC服务器
  - 处理能力与IO均衡
  - SAS+SSD混插

数据同步解决方案

数据同步方案分析

为什么需要数据同步？

- 多数据中心架构
- 系统之间的数据交互
- 跨平台数据库同步
- 数据库扩展性问题

现有解决方案

- 数据库触发器记录变化
- 系统之间数据交互
  - DBLink
  - 外部文件

商业产品分析

- ORACLE Dataguard，MySQL Replication
- Shareplex，Goldengate

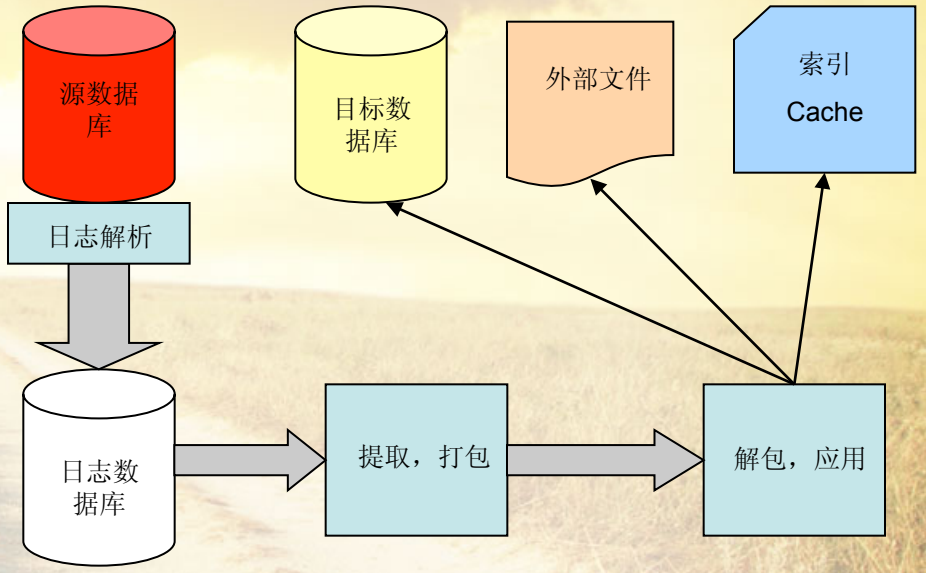
基于日志解析的数据同步方案

概述

- 替代触发器
- 实时解析，延迟小
- 对数据库性能影响小
- 数据库，文件，图片一致性
- ORACLE,MySQL统一解决方案
- 解析主键，实时抽取
- 支持多种目标端

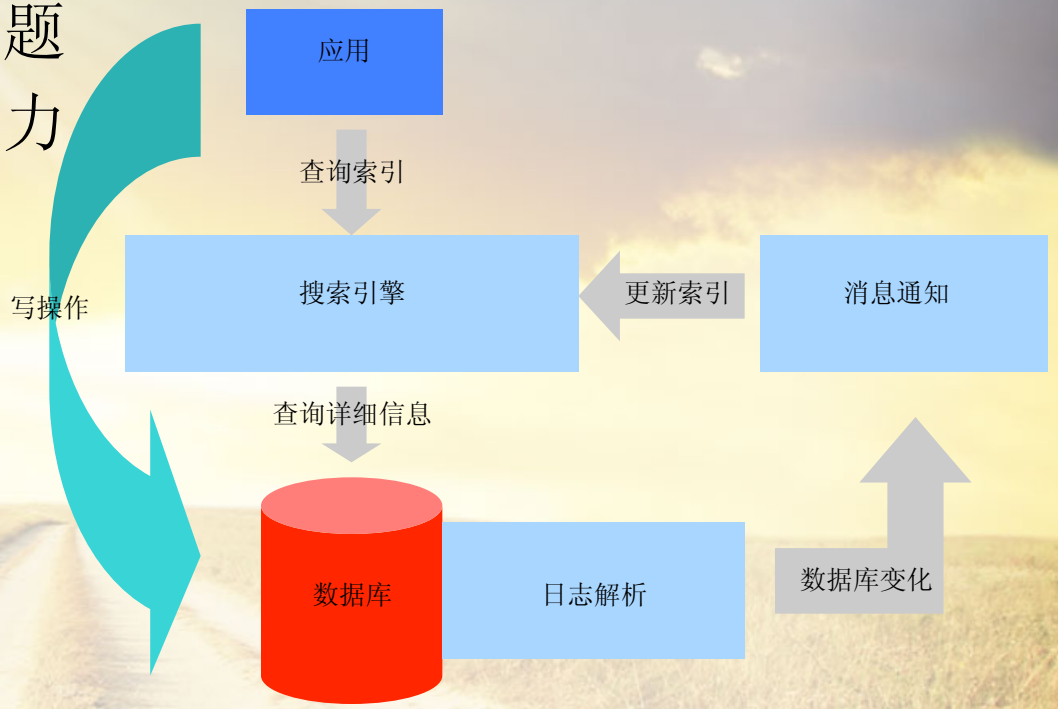
功能增强

- 解析字段
- 并行应用



搜索引擎实时更新

- 数据库日志实时解析
- 搜索引擎实时更新
- 解决搜索延迟的问题
- 提升数据库处理能力
- 读写分离架构




SSD高性能数据库

SSD性能特点分析

SSD vs Hard Disk

- SSD：
  - IOPS：8K read 10000+，8K write 2000+
  - 吞吐：1M read 200M，1M write 170M
- Hard Disk：
  - IOPS：8K read 150，8K write 150
  - 吞吐：1M read 170M，1M write 130M
- **SSD随机读非常好，随机写比较好，连续读写无优势**

SSD的性能特点

- 改写=读取+擦除+写入
- 随机写性能下降
- 损耗均衡算法
  - 均衡写磨损
  - 异步擦除
  - 提升写性能

数据库IO特点分析

数据库基于磁盘设计

- Sequential logging
- In-place update

日志文件顺序写

- 同步写入，响应延迟低
- 连续位置的随机写IO

数据文件随机写

- 异步写入
- 大量的随机写IO

瓶颈分析

- IOPS：
  - 小IO，数据文件随机读，随机写，日志文件写
- 吞吐量：
  - 大IO，数据文件连续读

基于SSD的数据库性能优化

为什么要使用SSD？
- CPU与IO性能不均衡
- 提升单机性能
- 减小集群规模
- RAM or SSD？

提升写性能
- 增加SSD保留空间

SSD-based database
- 减少擦除带来的影响
- IPL(In-page logging)
- 缓存写回机制

Flashcache方案
- 操作系统设备层实现
- 数据库存储引擎实现
- WB vs WT

SSD应用场景分析

SSD作为数据库主存储

- 依赖硬件层的损耗均衡算法
- 可靠性的问题
  - **硬件RAID** vs 软件RAID？
  - **RAID 5** vs RAID 10？
  - **SLC** vs MLC？
- 性价比较低

SSD存放日志文件

- 提升日志响应延迟
- HDD更合理 ？

SSD存放热点数据

- 提升随机查询性能
- 性能高，不灵活

SSD作为Flashcache

- 性价比高
- 复杂，有待实践考验

Alibaba使用SSD现状与发展

使用现状

- 2009年在数据库上使用SSD
- RAID 5 + BBU
- 基于SSD的MySQL集群
- 异构SAS磁盘备用集群
- 利用SSD存放热点数据
- 大幅度减小集群规模，降低成本

发展方向

- MLC替换SLC ?
- Intel SSD vs Fusion-IO
- Flashcache
  - MySQL替换SAS集群
- SSD-based database

Alibaba数据存储策略

ORACLE

- 强一致性，复杂查询
- 降低开发和管理成本
- ORACLE+数据同步+Cache

MySQL

- Sharding，去关系型
- 分布式，可扩展
- MySQL+Flashcache

KV store

- 场景优先
- 自主研发

Email: freezr@gmail.com
Blog: www.hellodba.net
Twitter: hellodba