ODF 2017 开源数据库论坛(北京)
OPEN-SOURCE DATABASE FORUM(BEIJING)

开源数据库正在改变世界

2017年8月24日-25日　北京－京仪大酒店

从零构建分布式数据库DDB

DDB团队负责人 马进

目录
CONTENTS

DDB简介

分库分表拆解

DDB架构变迁

总结

DDB简介

2006年开始， DDB为网易各大互联网产品提供透明分库分表服务

10年来不断完善，精益求精，是网易大体量互联网产品的立身之本

2006年
博客上线

简单SQL兼容
部分管理功能

2008年
V2.0发布

SQL兼容扩充
在线扩容功能
图形化管理工具

2010年
V3.0发布

分布式事务
在线修改表结构
SQL兼容扩充
管理功能完善
集群规模上千

2012年
V4.0发布

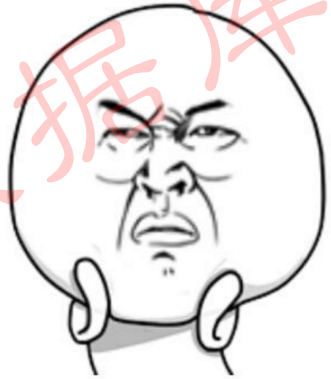
多语言支持
SQL统计功能
云计算DDB

2017 V5.0

架构简化
服务拆分
SQL兼容度进一步扩充

为什么分库分表？

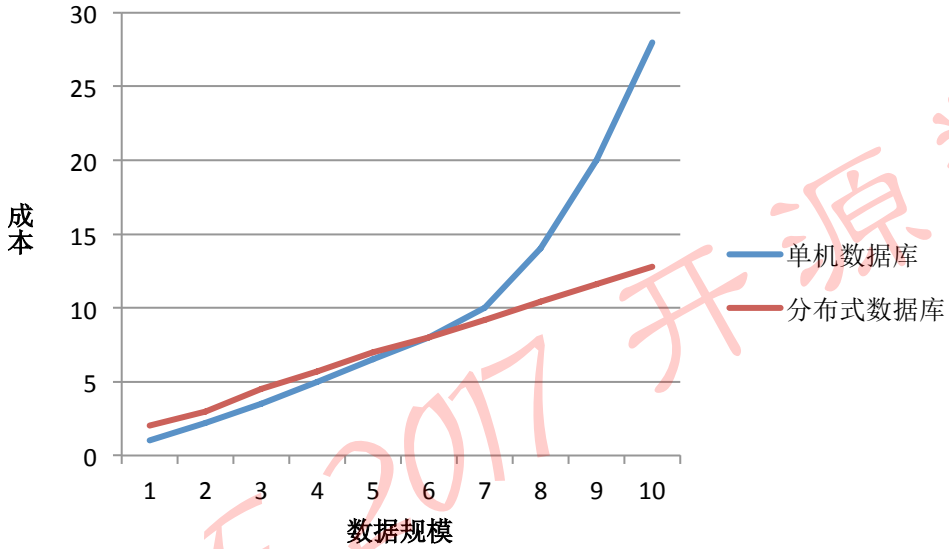
两个月后



- 目标上亿用户
- 要最好的性能
- 要最好的技术

- SQL兼容性不够
- Join性能不行啊
- 业务各种踩坑

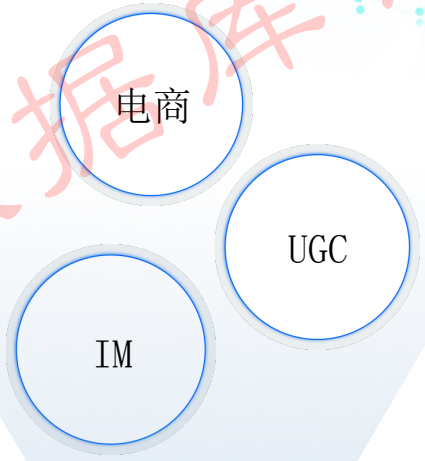
为什么分库分表？



数据库能力金字塔

DDB：一步到位的海量数据存储方案

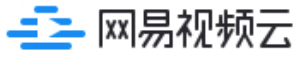
- PB级结构化数据存储
- 百万级qps
- 每日GB-TB数据增长
- 在线扩缩容
- 管理上千个数据节点
- 标准化的访问协议



DDB典型场景

DDB使用案例

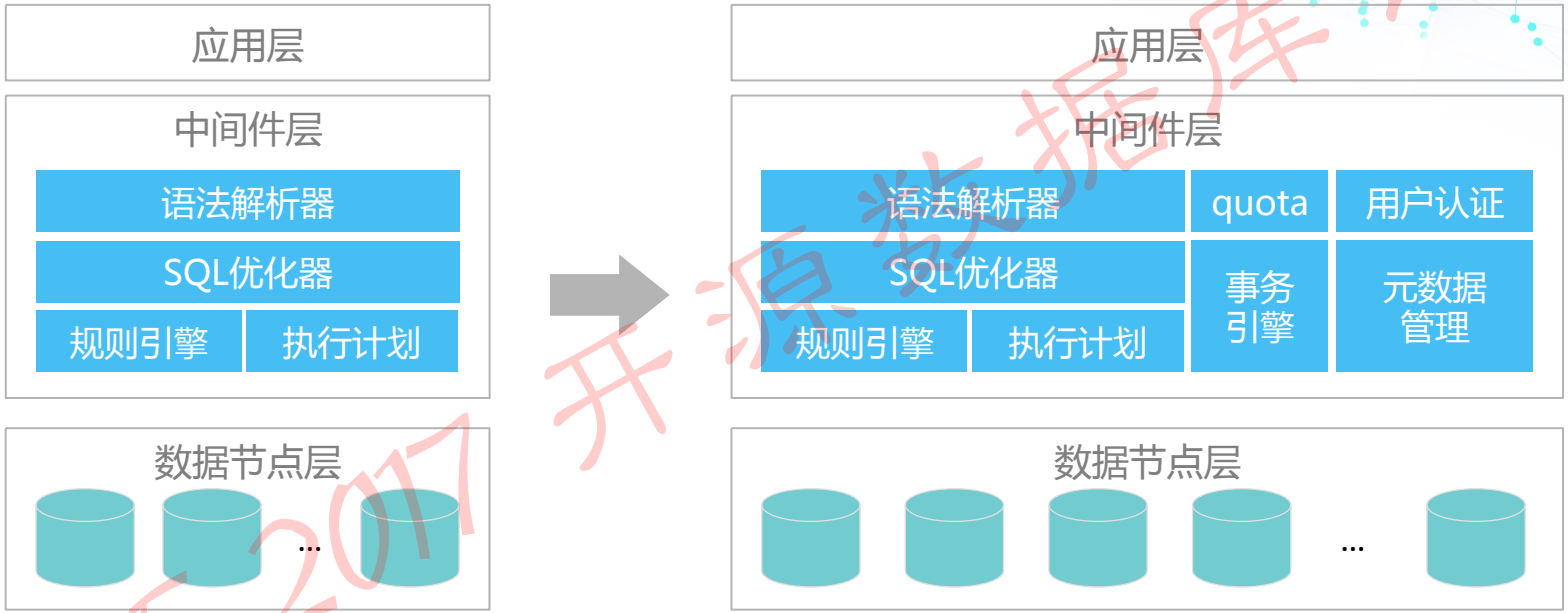
100+	10000+	1000000+
产品实践	数据节点	QPS



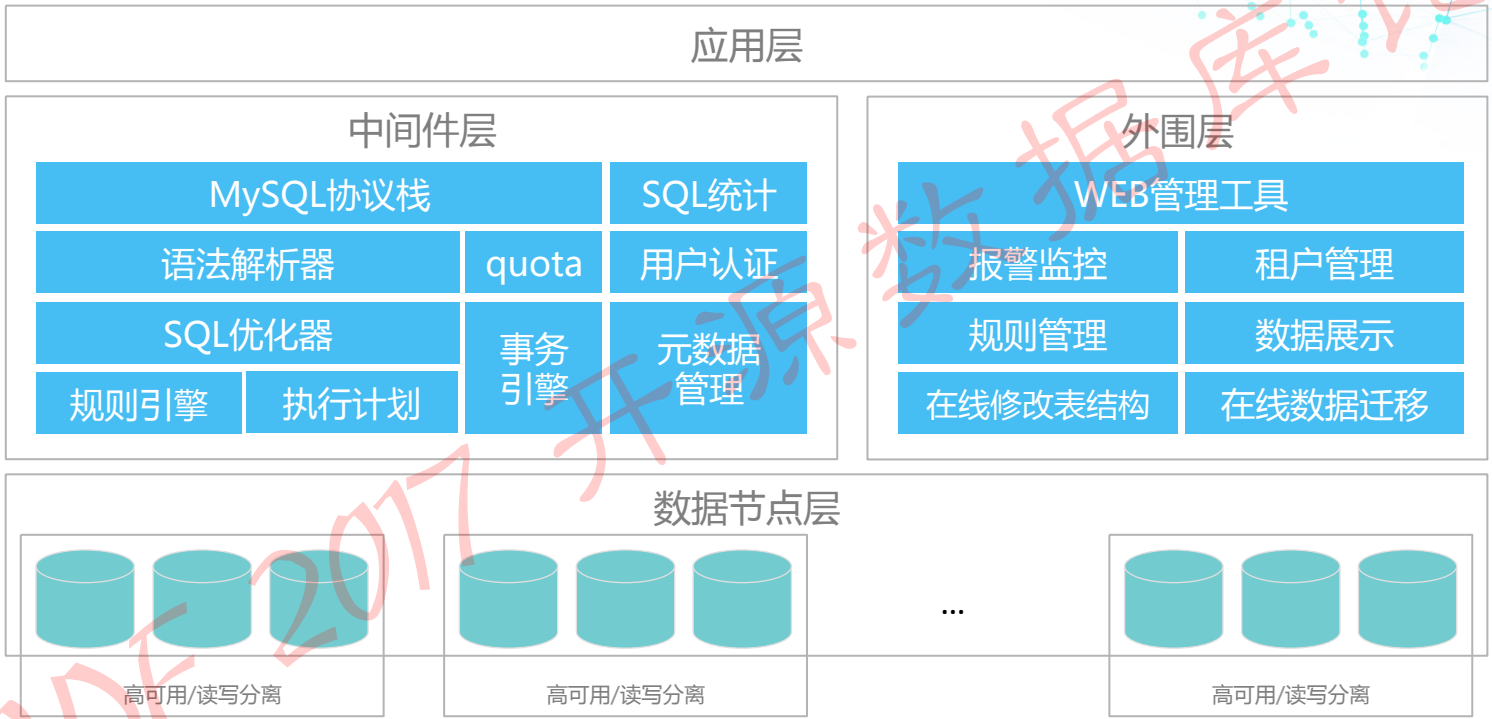
02

分库分表拆解

分库分表模块拆解



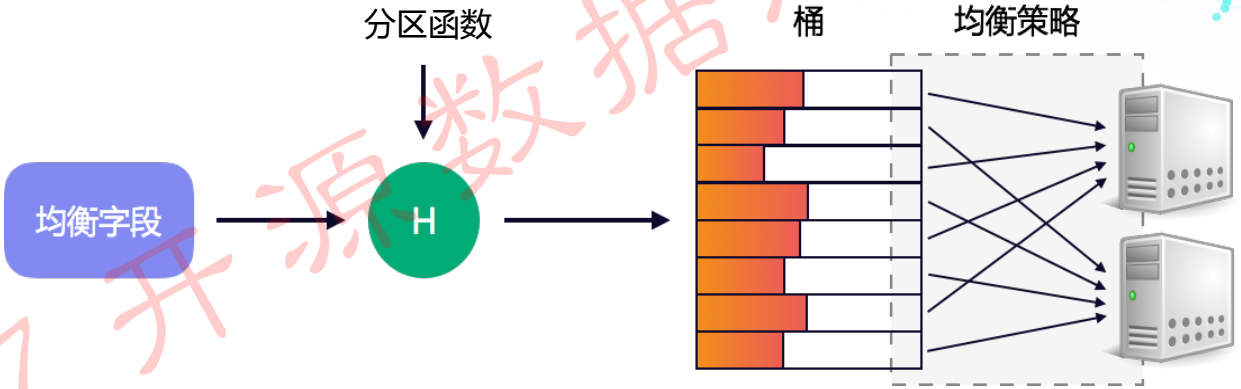
分库分表模块拆解



规则引擎

- 两级映射
  - 第一级：哈希函数
  - 第二级：均衡策略
  - 均衡性 + 单调性

- 自定义分区函数
  - range分区
  - list分区



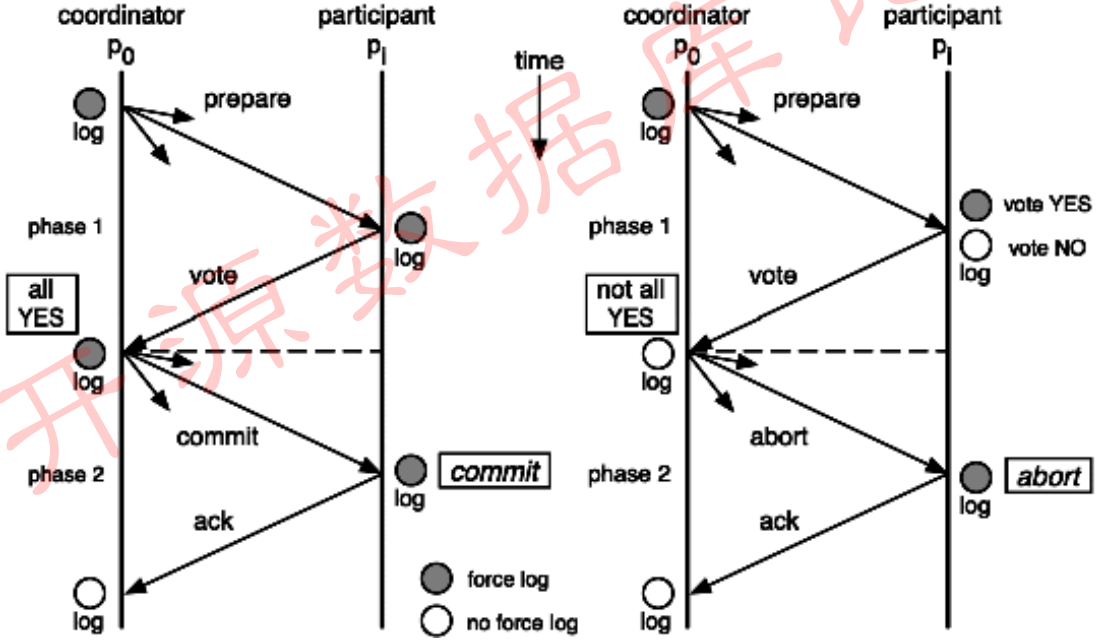
标准化——MySQL-like

- SQL兼容性90%
  - 支持所有简单SQL
  - 支持order by, group by, limit
  - 支持标量函数和大部分聚合函数
  - 支持部分特殊MySQL语法
  - 支持跨库join. etc
- 兼容MySQL通信协议
  - 支持任何语言MySQL客户端
  - 应用可使用任意ORM框架

```
mysql> explain select avg(age) from UserTest group by name limit 10,10;
+-----------------------------------------------------------------------------------------------+
| PLAN                                                                                          |
+-----------------------------------------------------------------------------------------------+
| LIMIT/OFFSET                                                                                  |
|   This plan will be dynamicly set disable/enable while running based on the underlying plan. |
|   /\                                                                                          |
|  /||\                                                                                         |
|   ||                                                                                          |
| AGGREGATE                                                                                     |
|   Do:                                                                                         |
|   /\                                                                                          |
|  /||\                                                                                         |
|   ||                                                                                          |
| PROJECT                                                                                       |
|   Project record to: SUM(age),COUNT(age),                                                     |
|   /\                                                                                          |
|  /||\                                                                                         |
|   ||                                                                                          |
| GROUP                                                                                         |
|   Group By: name,                                                                             |
|   /\                                                                                          |
|  /||\                                                                                         |
|   ||                                                                                          |
| MERGE-SELECT                                                                                  |
|   SQL: SELECT SUM(age), COUNT(age), name FROM UserTest GROUP BY name ORDER BY name ASC        |
|   Dest Node:                                                                                  |
|     dbn1[jdbc:mysql://10.120.146.129:3306/dbn1]                                               |
|     dbn2[jdbc:mysql://10.120.146.129:3306/dbn2]                                               |
|     dbn4[jdbc:mysql://10.120.146.130:3306/dbn4]                                               |
|     dbn3[jdbc:mysql://10.120.146.130:3306/dbn3]                                               |
|   Order by: name ASC,  with merge sort.                                                       |
+-----------------------------------------------------------------------------------------------+
```

分布式事务

- 2PC优化
  - 单节点事务一阶段提交
  - 省略第一步记日志
  - 日志组提交
- 2PC缺陷
  - 无法理论上保障事务ACID
  - XA事务提交效率差
- 2PC目的
  - 99.9%到99.99%的巨大提升
  - 很大应用分布式事务占比极低
  - 应用不想去操心分布式事务

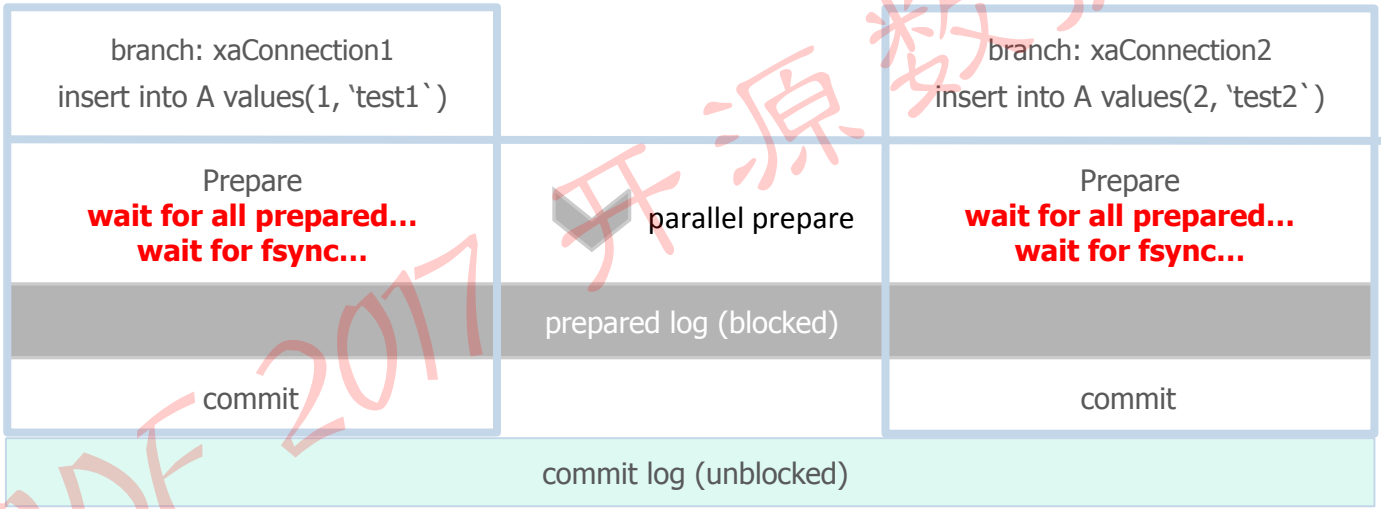


两阶段实现

- 自发判断是否需要两阶段

  insert into A values(1, 'test1'),(2, 'test2');

分区字段



在线扩缩容

- 初始化配置信息
- 初始化执行节点

- 启动增量数据拉取模块
- 增量变更数据缓存本地

- 分批迁移存量数据
- 根据源库压力调整速度

- 并发回放增量数据
- 幂等执行回放

- 设置源表或源库只读
- 瞬间切换（1-5s）

DDB功能特性——数据信道

数据分布

- 两级映射
- 自定义哈希函数

标准化

- SQL92 高兼容
- 全局自增ID
- 支持explain
- 数据导入导出
- 兼容MySQL通信协议

分布式事物

- 实现2PC协议
- 数据高一致
- 用户透明
- 自动识别

Hint功能

- 读写分离
- SQL路由

SQL统计

- SQL模式统计
- SQL频度统计
- 慢SQL统计
- 多维度QPS统计

DDB功能特性——管理信道

集群管理

- 配置管理
- 连接池管理
- 元数据管理和同步

表/策略管理

- 创建/删除/更新
- 在线修改表结构
- 支持show/desc等
- 兼容MySQL管理语法

用户管理

- 创建/删除/更新
- 支持常用授权操作
- 支持白名单操作
- 管理与访问权限分离

在线数据迁移

- 在线策略迁移
- 在线扩/缩容
- 更改均衡字段

高可用

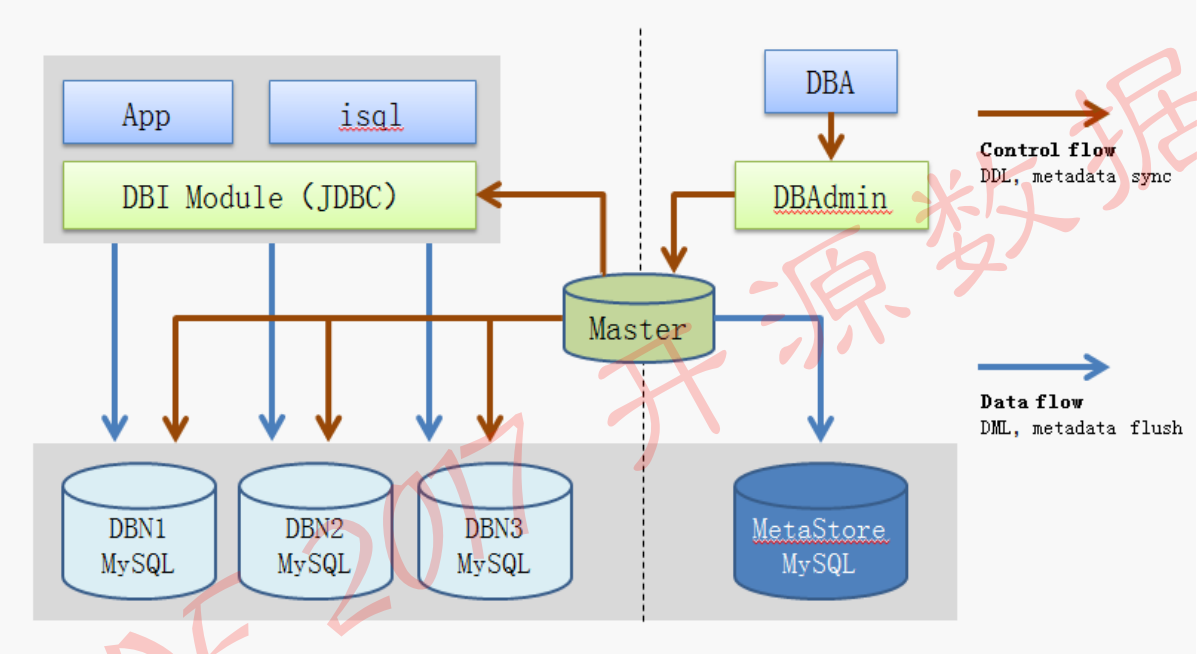
- 中间件节点高可用
- 数据节点高可用
- 数据节点手动切换

扩展功能

- 数据节点报警
- 中间件节点报警
- 悬挂事务报警
- 定时任务

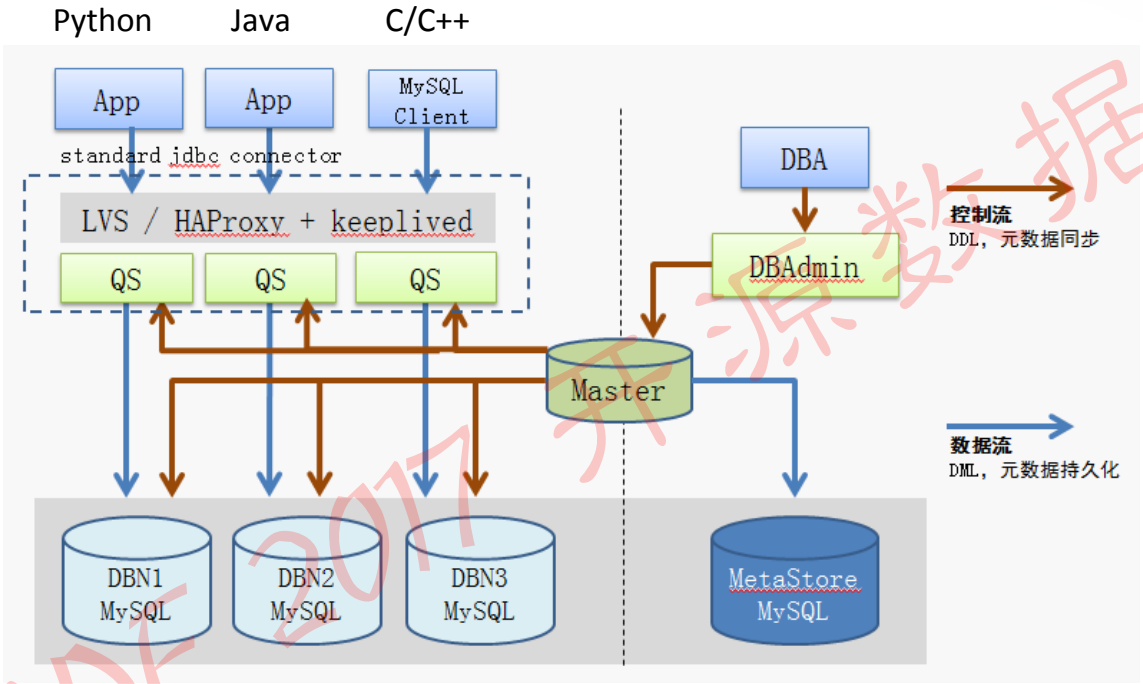
DDB架构变迁

DDB基础架构——DBI模式



- 版本难以管理
- 问题不好跟踪
- 连接无法收敛
- 侵入应用模块
- 不支持多语言

DDB基础架构——代理模式

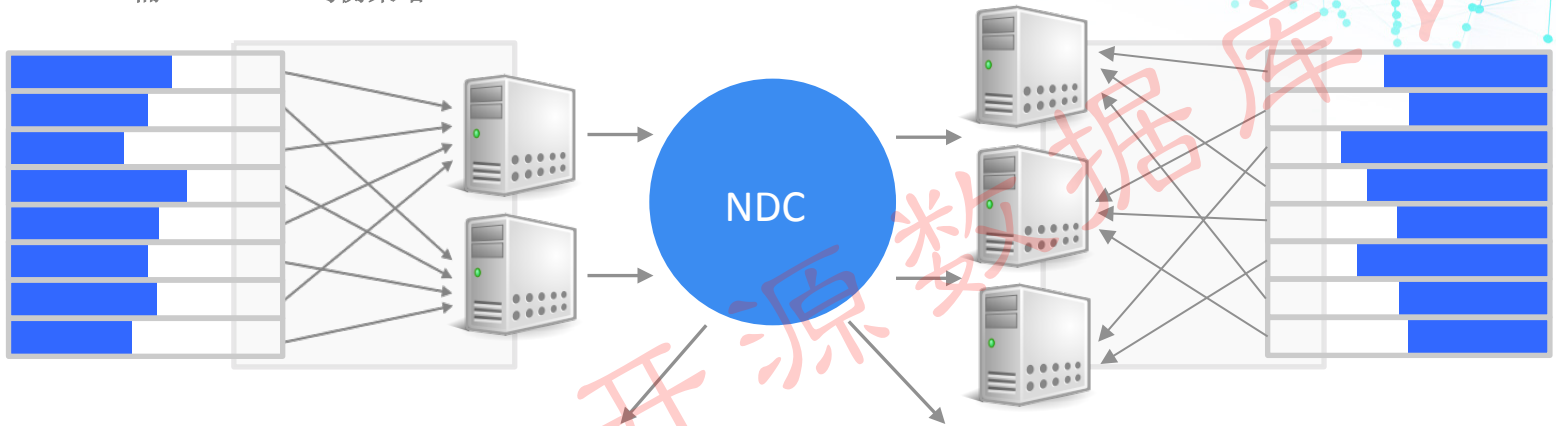


管理信道

数据信道

NDC——数据迁移一站式方案

桶　　均衡策略



Oracle　SQLServer　Postgres

Greenlum　HBase　Kudu

数据迁移

消息队列

风控　全文索引　流计算

数据订阅

04

总结

分库分表设计总结

- 设计哲学
  - 用不同的中间件解决不同问题
  - SQL兼容度不是越全越好（OLTP）
  - 规范很重要
- 架构设计
  - 保持系统精简，部分功能插件化
  - 中间件本身先平台化，再考虑适配PaaS
  - 考虑与其他中间件的集成问题（单元化架构）

Thanks

关注开源数据库论坛