OLTP和OLAP技术融合的探索实践

饿了么&百度外卖 梁福坤

目录

- 一、OLTP与OLAP 技术介绍
- 二、融合技术选型
- 三、Binlog+Kudu+impala最佳实践

目录

1.1 OLTP 背景介绍

联机事务处理OLTP（on-line transaction processing）

也称为面向交易的处理过程，其基本特征是前台接收的用户数据可以立即传送到计算中心进行处理，并在很短的时间内给出处理结果，是对用户操作快速响应的方式之一。[1]

关键词：

数据量少、面向应用、并行事务处理、分库分表、读写分离、Cache技术、B-Tree索引

实时性要求高、数据库作为载体、SQL交互

[1] OLTP概念引用自百度百科 https://baike.baidu.com/item/OLTP/5019563

1.2 OLAP 背景介绍

联机实时分析OLAP (OnlineAnalytical Processing,）

联机分析处理OLAP是一种软件技术，它使分析人员能够迅速、一致、交互地从各个方面观察信息，以达到深入理解数据的目的。它具有FASMI(Fast Analysis of Shared Multidimensional Information)，即共享多维信息的快速分析的特征。[2]

关键词：

数据海量、追加操作为主、数据分区、切片和切块、雪花模型

钻取、旋转、投影、数据仓库、MDX、实时性要求低

[2] OLAP概念引用自百度百科 https://baike.baidu.com/item/联机分析处理?fromtitle=OLAP

1.3 两者面向场景的分析-HTAP[3]

需求：

一份数据存储用于OLTP和OLAP处理。

1）数据实时可见 2）支持多维度低延迟查询交付 3）低成本

通用的解决方法：

数据Sharding：实例间share nothing，便于横向水平扩展

数据分区：满足数据线性扩展，通过引擎优化命中细节

分布式事务：两阶段提交

[3] https://en.wikipedia.org/wiki/Hybrid_transactional/analytical_processing_(HTAP)

存储模型

共享存储

- 典型的Shared Disk架构，从底层的存储层共享解决一致性问题，简单粗暴。
- 理论上无限扩展，基于Raft维持一致性
- 弱化OLAP功能，重点解决单点容量问题。
- **代表作：AWS Aurora、PolarDB**

Sharding

- 水平（Scale Out）/水平（Scale Up）切分、综合切分，把同一表数据分散到多个数据库或者多节点，增强并发能力，同时解决扩展能力。
- 多个表，可以跨越DB和服务器节点。
- Proxy层责任重，路由、优化、事务状态机、流式执行器。
- **代理**：Mysql方案的：Mycat、Baidu Dbproxy；**引擎**：Oracle Sharding、MongoDB；**DAO层**：Hibernate Shards、 Sharding-JDBC
- 切分策略根据业务键、时间。

Horizontal Partition

- 侧重单张表的水平切分，突破I/O瓶颈。
- 查询引擎负责任务计划、优化，不需要代理。
- 切分策略根据Hash、Range、List。
- 一般和多副本同时发挥作用
- 代表作：Kafka Partition、Kudu tablet、Greenplum segment。

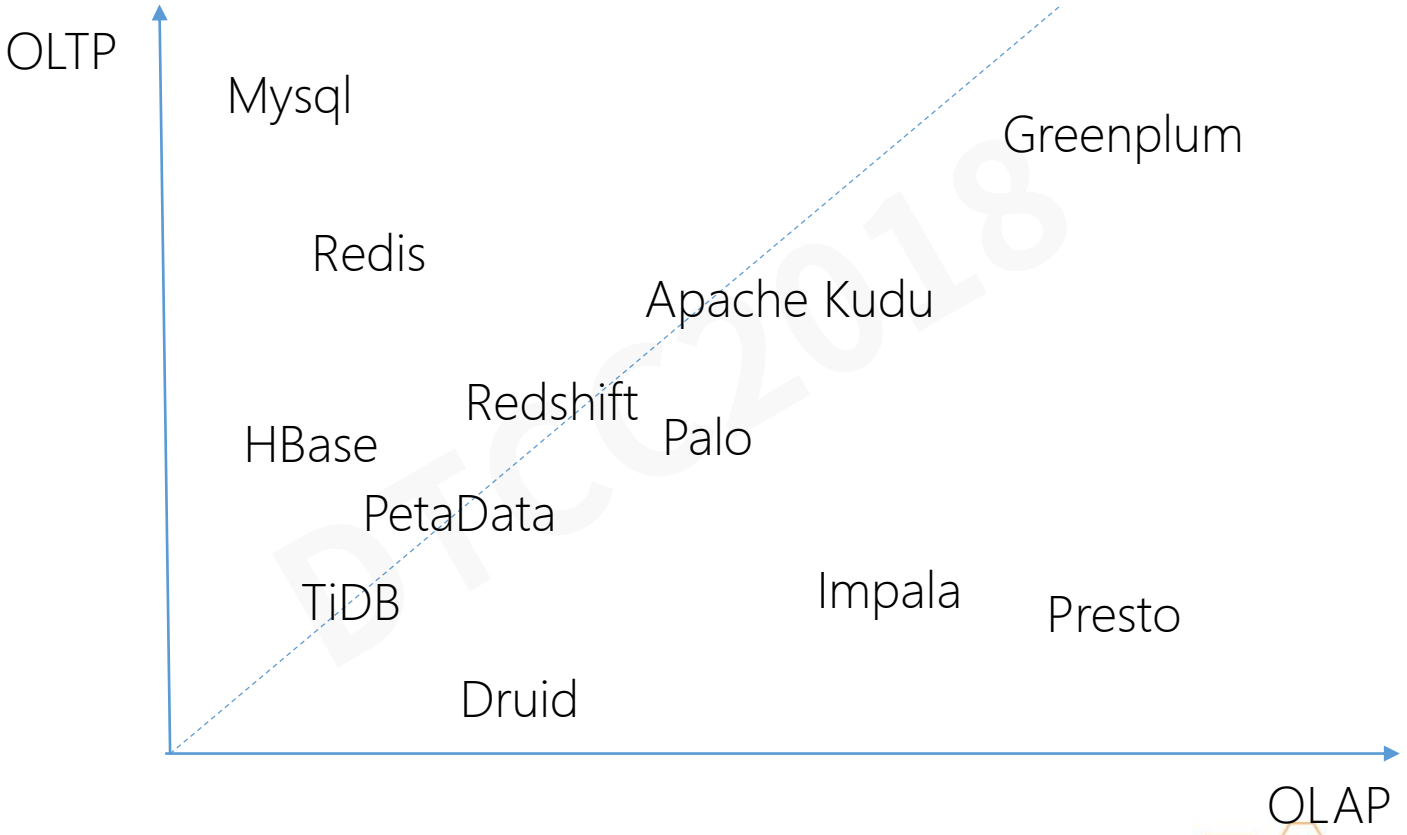
Sharding Nothing

- 在NewSQL、MPP模式下应用广泛，并行处理和扩展能力强。
- 节点独立，数据结果节点流转或者上层汇总。
- 底层存储多样，Kv解决方案多，类似Palo和TiKV底层存储的Rocksdb。
- 一般和多副本同时发挥作用、数据模型LSM Tree。
- 代表作：HybridDB for MySQL、TiDB、Teradata、DB2 DPF、GreenPlum

事务并发控制模型

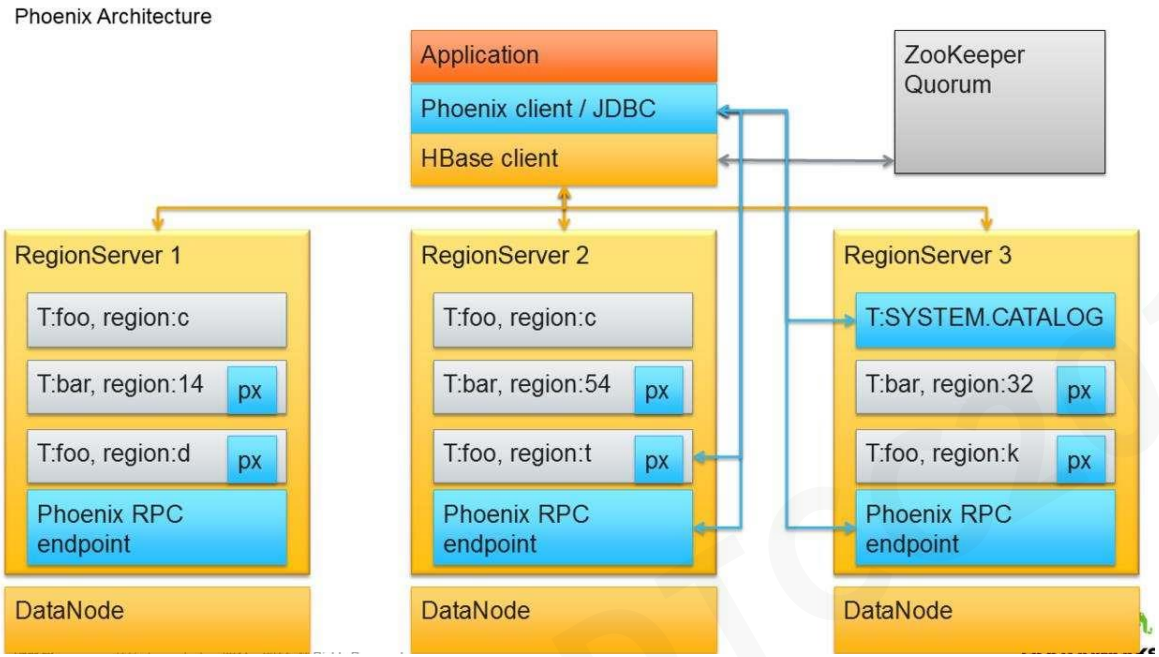
事务	特征	缺点	优点	应用代表
2PC/3PC	协调者、参与者 投票阶段+预提交+提交阶段	保守策略、同步堵塞、 单点故障、数据不一致	实现简单	Mysql、 Greenplum、 TiDB （Percolator）
Paxos	三角色对某个数据的值达成一致	推导过程复杂	保证安全和活性、允许日志空洞	阿里X-Paxos、 腾讯phxpaxos、 Zookeeper
MVCC	基于快照隔离机制进行并发控制，**解决读-写冲突**的无锁并发控制		写操作不用阻塞读操作的同时，避免了脏读和不可重复读	Mysql、Oracle、 Baidu TDB、 HBase
OCC	**解决写-写冲突**的无锁并发控制	假设竞争几率小	在资源冲突不激烈的场合，用乐观锁性能较好	DBMS

1.4 融合技术的应用场景需求



目录

2.1 SQL On Hbase：Apache Phoenix[4]

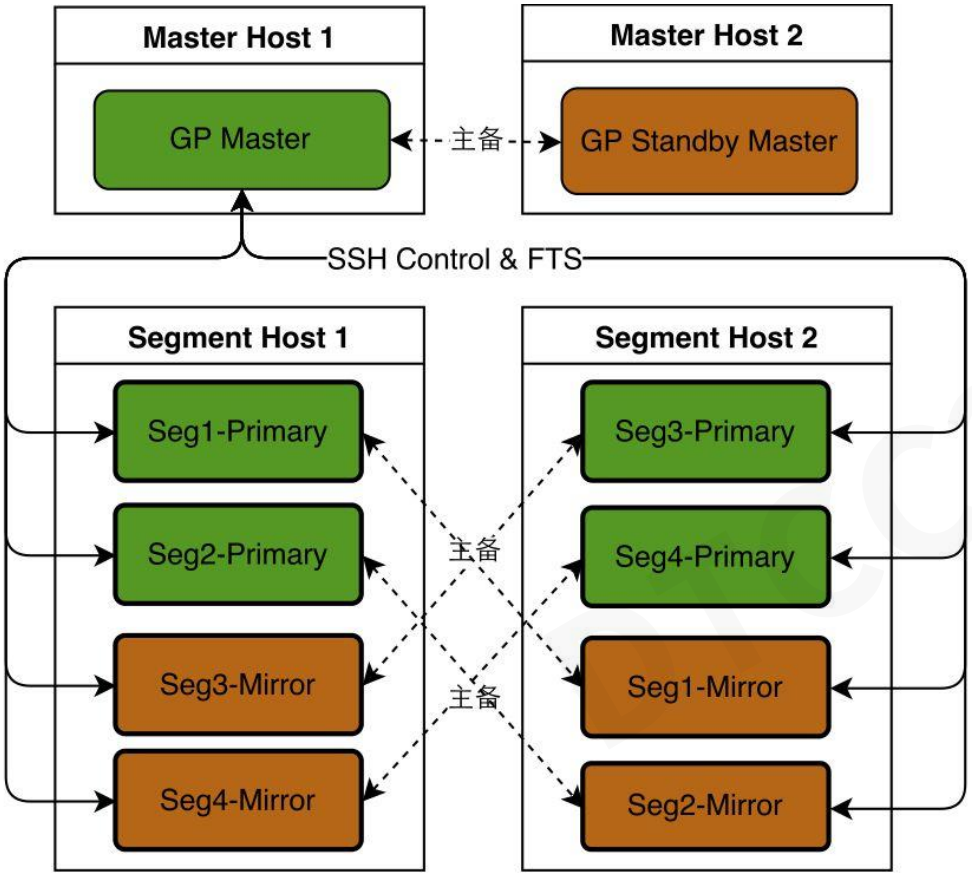


- 二级索引（四种）
- 统计信息收集
- SQL编译成Hbase Scans
- 基于Tephra支持全局事务
- 并行任务编排

- Phoneix重点强调低延迟的OLTP，基于Hbase提供分析能力。
- 擅长热数据的简单聚合分析能力
- 百度外卖用通过MQ对接数据DML的回放和数据暂存点

[4]图片来源https://www.cnblogs.com/linbingdong/p/5832112.html

2.2 Greenplum

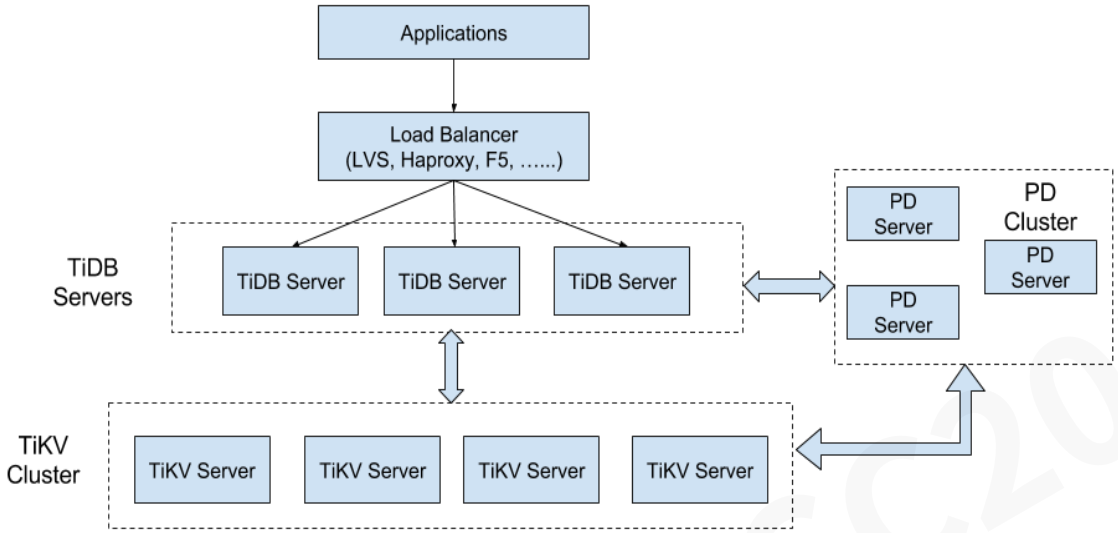


- 采用shared nothing架构（MPP）底层采用Postgresql
- 自有资源队列和优先级
- 运维管理工具丰富
- 索引方式丰富，命中索引场景速度较优

- Gpexpand可以实现动态扩容，但周期长
- 侧重OLAP能力
- Heap表容易实现膨胀
- 并发写入性能较Phoenix、TiDB差[6]

[5]图片来源https://m.aliyun.com/yunqi/articles/51176
[6]来源https://www.datanami.com/2018/02/22/hybrid-database-capturing-perishable-insights-yiguo/

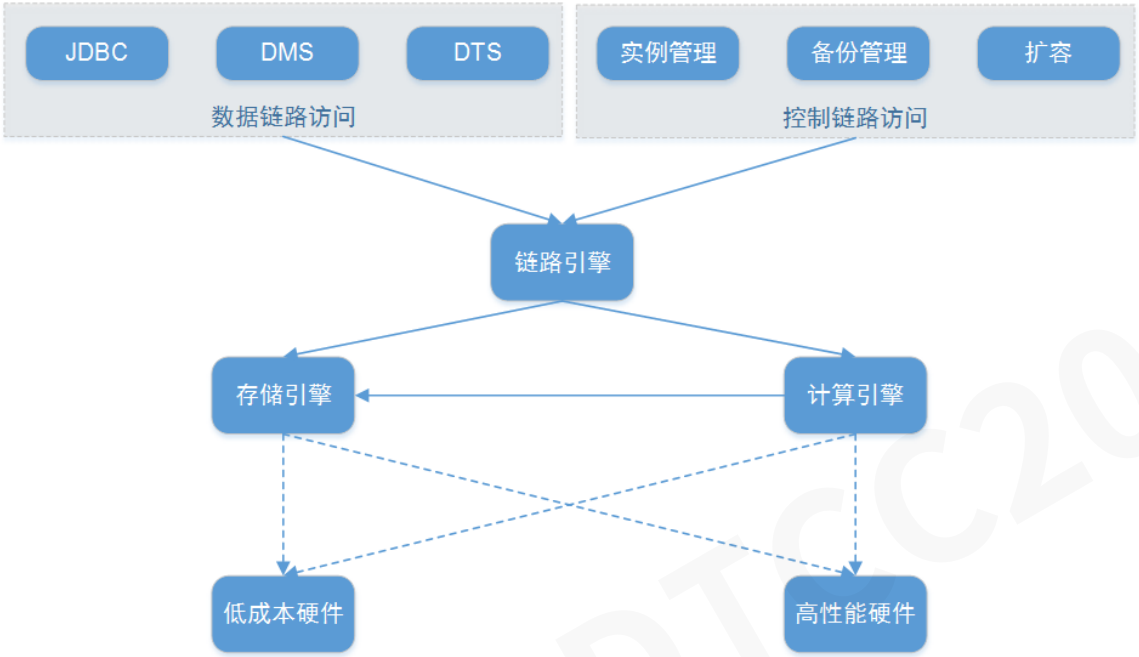
2.3 TiDB[7]



- 开源分布式HTAP数据库，目标是100% 的 OLTP 场景和 80% 的 OLAP 场景
- 兼容 MySQL
- 支持无限的水平扩展
- 具备强一致性和高可用性
- 运维工具和周边工具丰富
- TiKV以Region作为单元，对数据管理和复制
- 支持分区和索引
- 为云部署设计

[7]TiDB架构图来自https://pingcap.com/docs-cn/

2.4 HybridDB for MySQL[8]

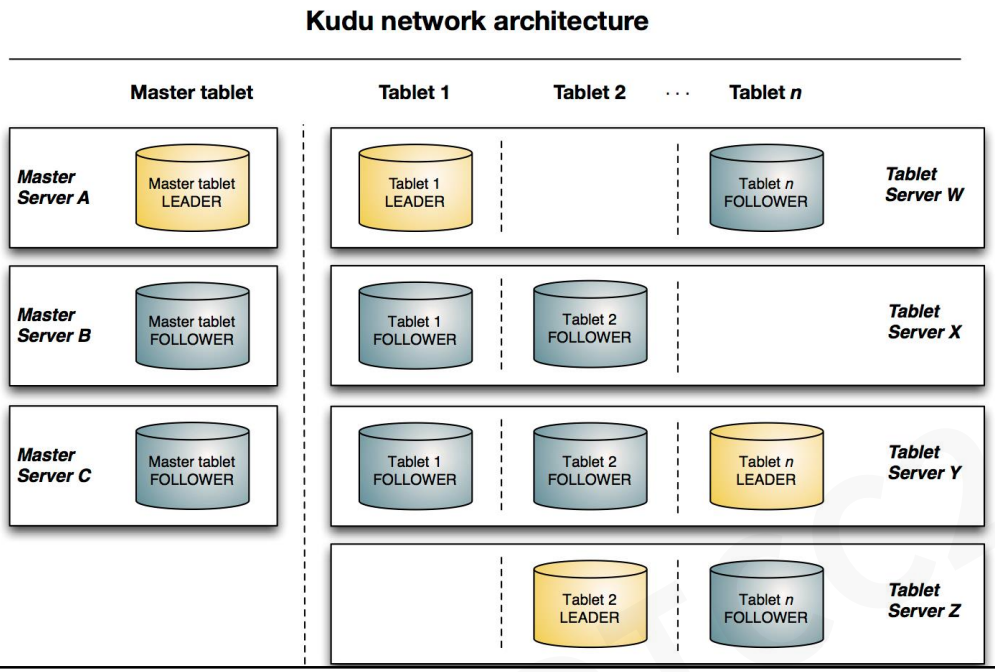


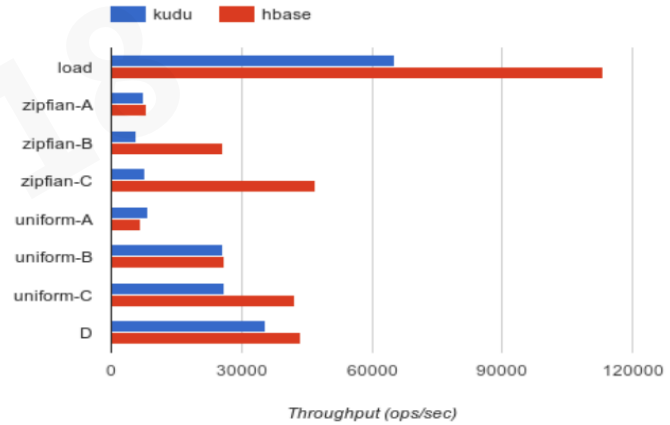
- 关系型 HTAP 类数据库，目标实时处理分析
- 分布式任务可以线性增长
- 兼容Mysql语法和函数
- 对Oracle常用分析函数的支持，100%完全兼容TPC-H和TPC-DS测试标准
- 支持分区内事务
- 以阿里云方式提供服务

[8]架构参照https://help.aliyun.com/product/26320.html?spm=a2c4g.11186623.3.1.qRhlHP

2.5 Impala+kudu[9]

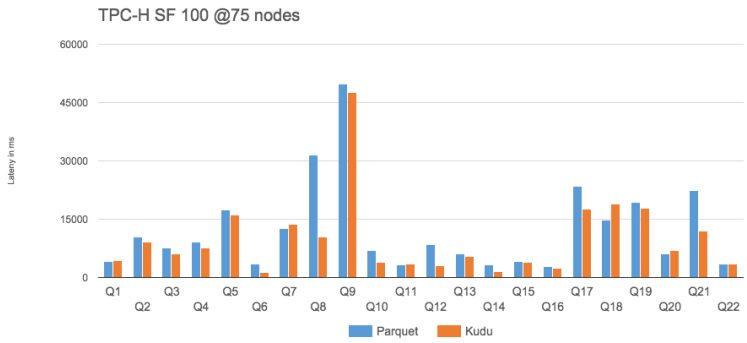
- Impala:基于MPP架构的即席查询引擎
- 内存shuffle，计算速度快
- 支持HDFS、KUDU、Hbase数据源
- 与Hive语法兼容性高
- Catalog和Statestore存在单点





- Kudu:融合OLTP型随机读写能力与OLAP型分析能力
- 开源的基于列式存储、与Hadoop生态结合好
- 强Schema，有限列数
- 顺序度和随机度综合性能强劲
- 有唯一主键约束，支持Upsert语法

Figure 1: Operation throughput of YCSB random-access workloads, comparing Kudu vs. HBase

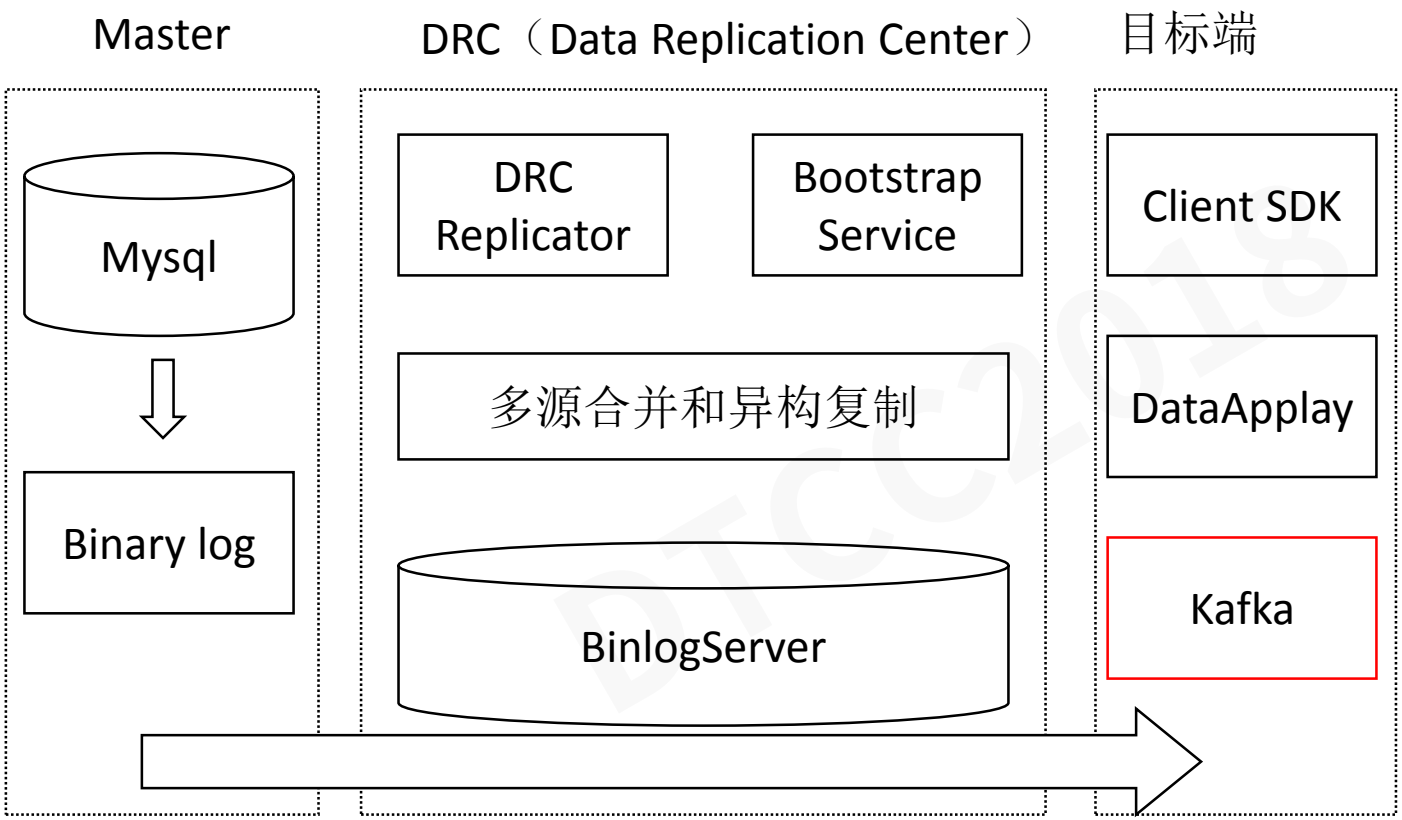


[9]参照https://kudu.apache.org/kudu.pdf http://kudu.apache.org/docs/

目录

- 一、OLTP与OLAP 技术介绍
- 二、融合技术选型
- 三、Binlog+Kudu+impala最佳实践

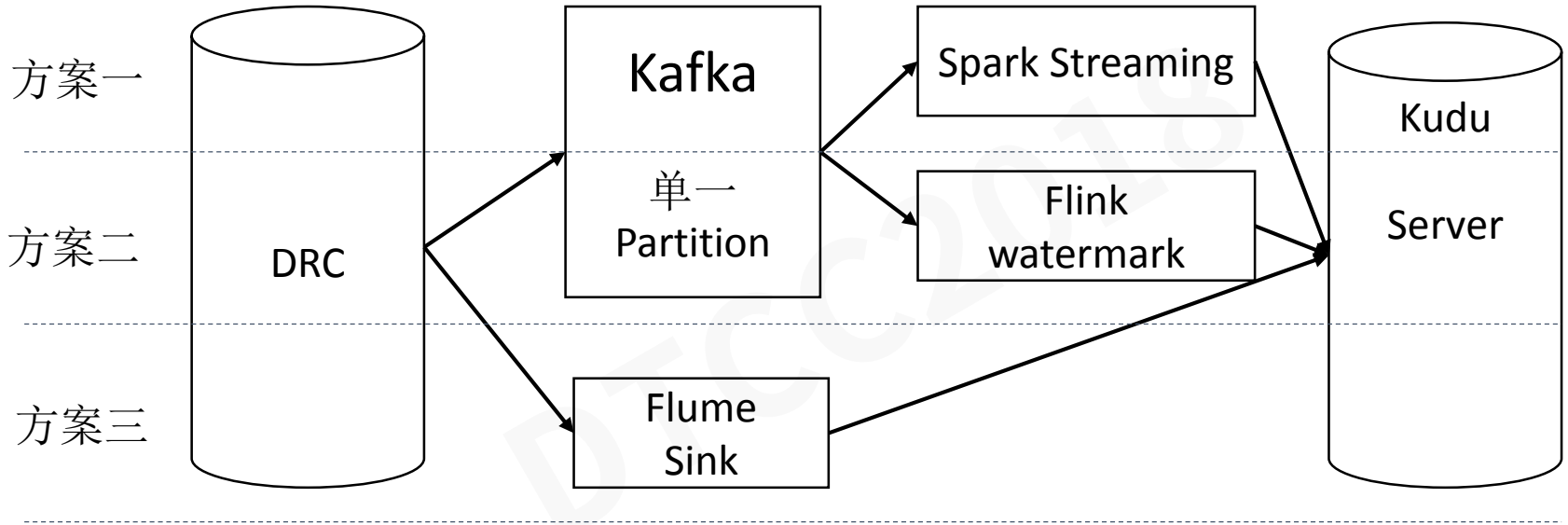
3.1 基于binlog的回放



DRC数据复制和调度中心

- 支持异构结构对接映射
- 支持基于时间点数据快速回放
- 支持多规则的过滤规则
- 整体服务高可用
- 不具备全局事务一致性，保证单个事务操作一致

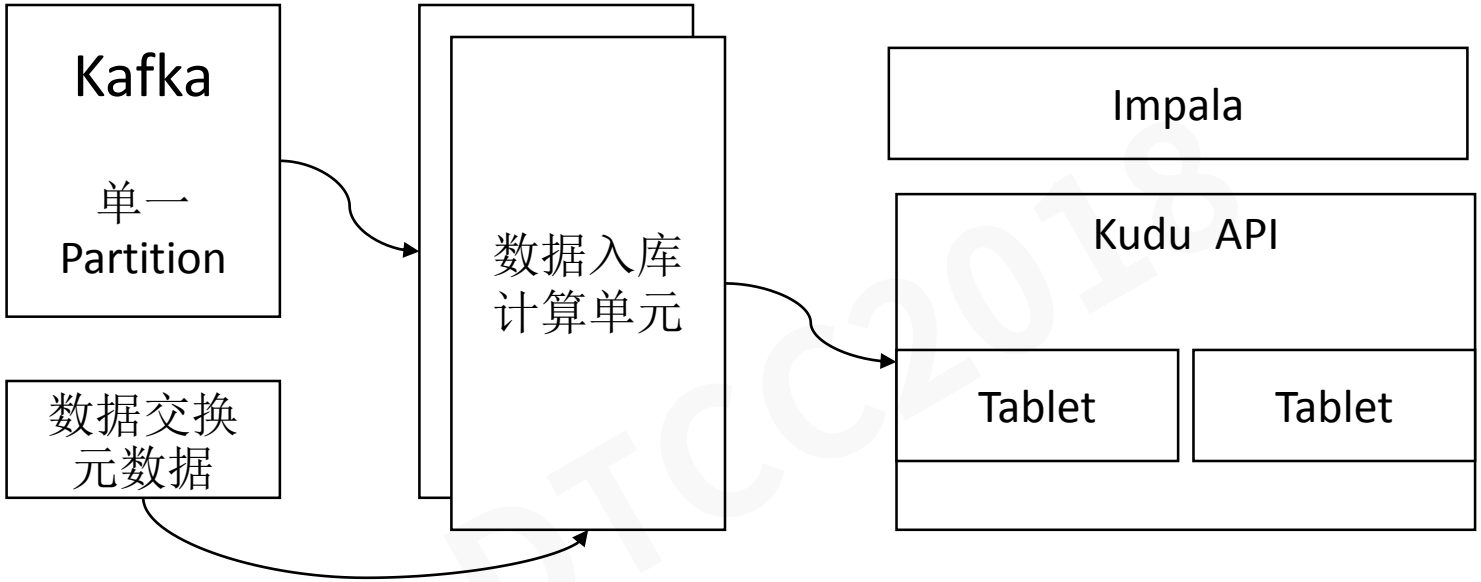
3.2 融合方案的设计架构



- 方案一：高并发支持，开发成本略低，实时性好，有乱序可能
- 方案二：高并发支持，在时间窗口内保证有序，实时性和有序做权衡
- 方案三：流程简单，处理效率低

- 支持轻量级ETL扇入
- 数据有序性
- 数据处理高效
- 精准insert、update、delete数据列

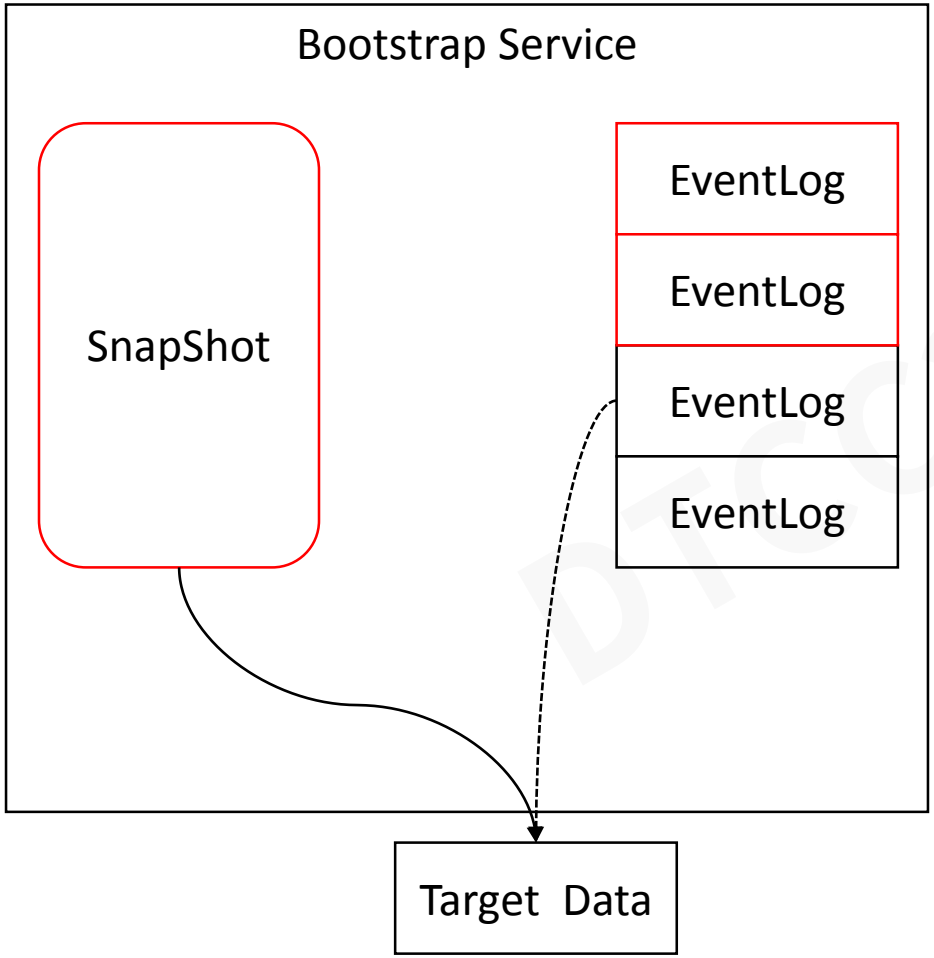
3.3 关键技术点：分库分表的融合



- **Kudu随机写压力分散到Tablet**
- **利用Hash Paritioning，实现随机写高性能**

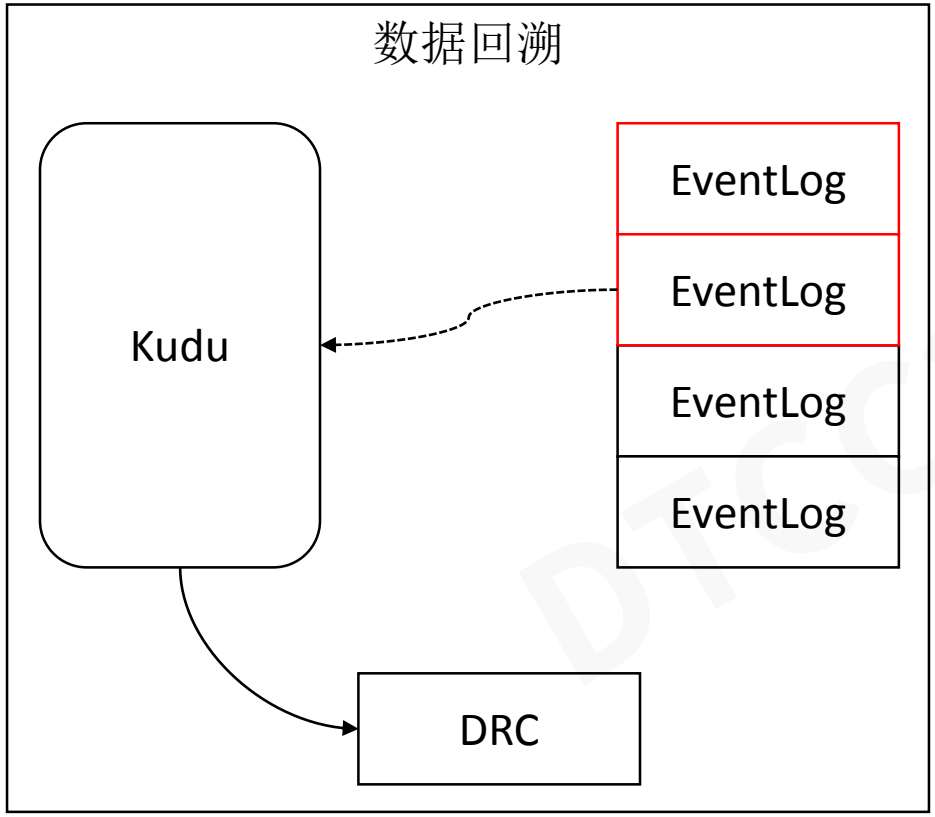
- **同一个Scan所需要的数据放在同一个tablet中，利用业务主键进行hash**
- **大范围检索需要Range切片（日期）**

3.4 关键技术点：数据冷启动



- 通过Slave系统复制一份快照，并且应用后续的Eventlog
- 对于批量数据方便装载
- 痛点：
- 数据快照占用存储空间，需要配合清理策略+压缩
- 定期快照 or 初始快照+DataEvent阶段

3.5 关键技术点：数据回溯

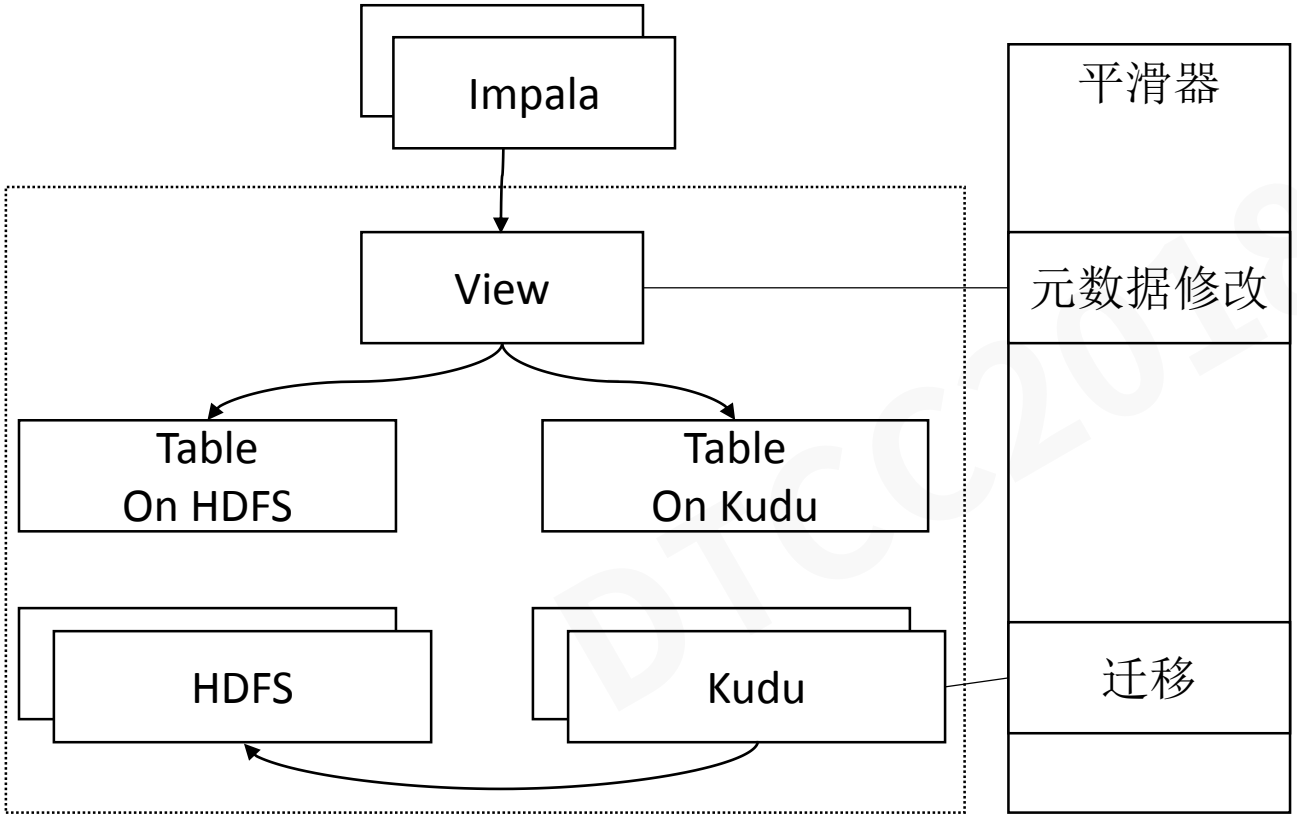


回溯用于历史数据的修正

- 通过客户端存储数据的复位点
- 数据回放不清理目标端数据

- 回溯周期较长，采取冷启动的方式快速对接。
- 需要：数据的增改删带全量变更前后数据。

3.6 关键技术点：热数据变冷归档0感知



- 基于视图对外交付
- 冷热交替数据通过调度定期变更
- 数据交替期间多存储
- 元数据变更与广播
- 迁移完Kudu清理

面临问题：

归档hdfs数据有变更

3.7 踩坑集合

- 没有主键
- 类型适配映射
- Binlog Row
- DDL变更
- 保证binlog回放顺序性
- 数据轻度ETL扇入

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigData
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168