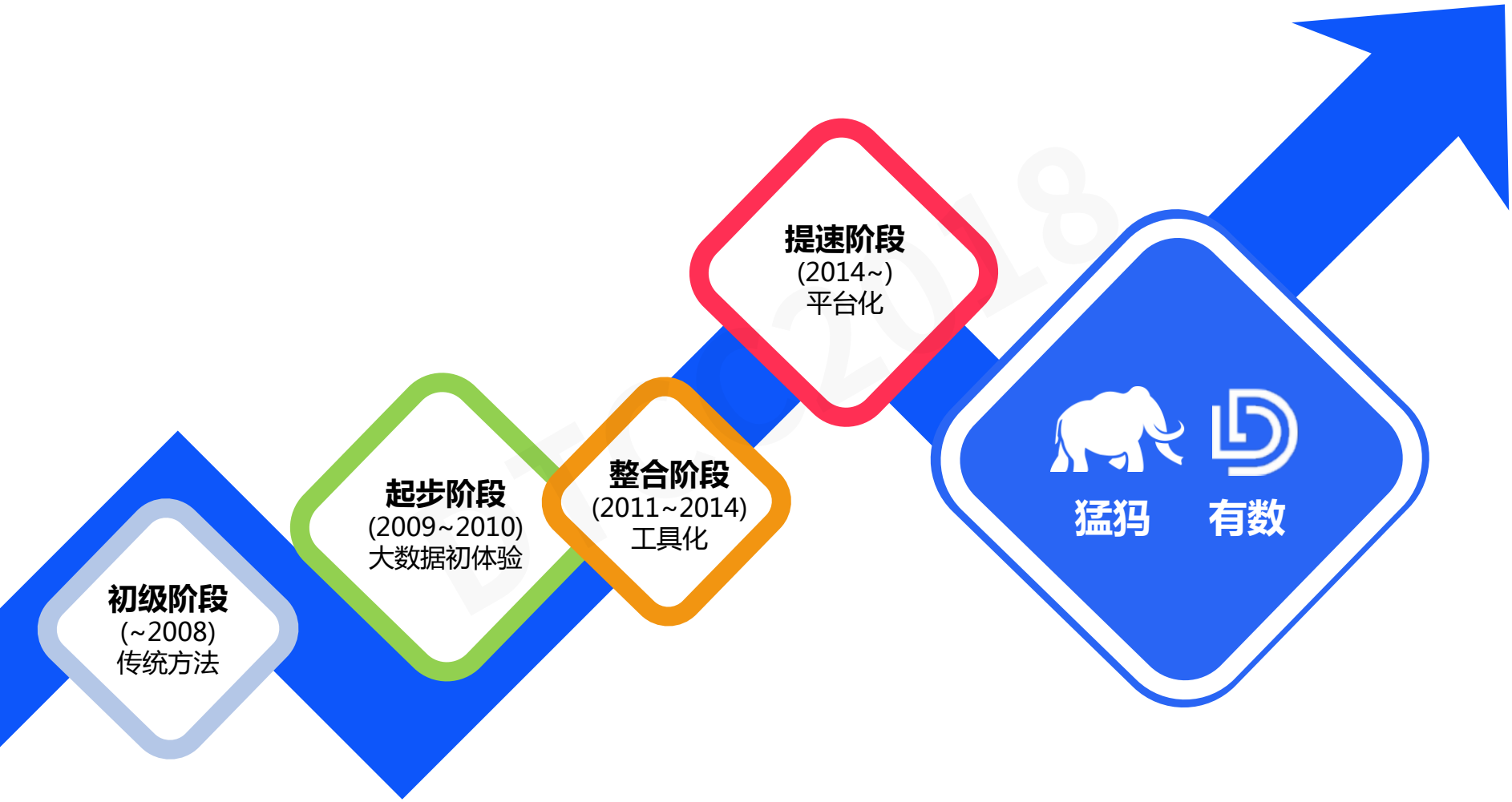
网易大数据平台实践

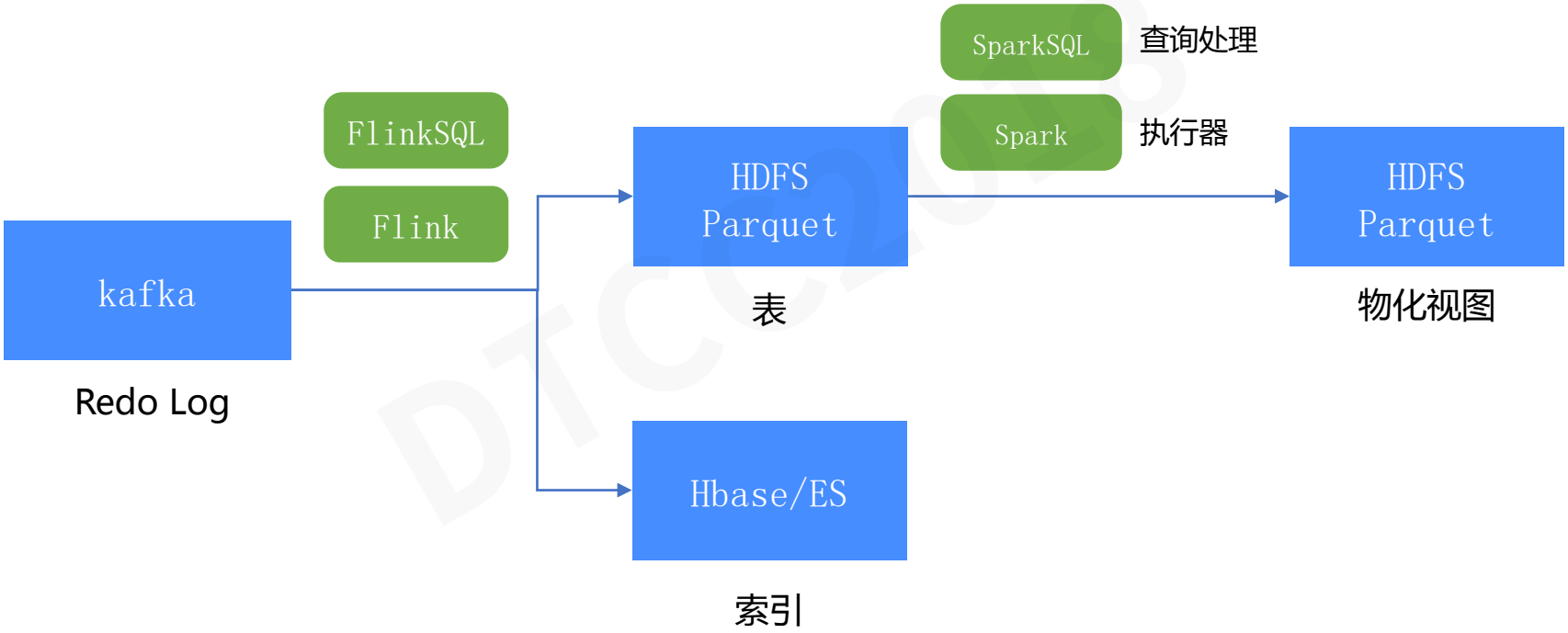
网易 余利华

PART 01 大数据平台概述

网易大数据发展历程


初级阶段
(~2008)
传统方法
起步阶段
(2009~2010)
大数据初体验
整合阶段
(2011~2014)
工具化
提速阶段
(2014~)
平台化
猛犸
有数

大数据系统为什么难用


SparkSQL
查询处理
FlinkSQL
Spark
执行器
HDFS
Parquet
HDFS
Parquet
Flink
kafka
表
物化视图
Redo Log
Hbase/ES
索引

平台的需求是什么

01 提供大数据基础能力

02 提升使用效率

03 提升管理效率

04 多租户和安全

大数据体系架构


大数据应用开发层
大数据开发套件（可视化IDE）
数据加工
数据集成
数据开发
任务运维
自助分析
数据管理
数据计算
离线计算
Hive
流式计算
Sloth
内存计算
Spark
资源管理
统一资源管理与调度
Yarn
数据存储
分布式文件系统
HDFS和Kudu
分布式数据库
HBase
数据集成
全量/非实时接入
Sqoop
实时/增量接入
NDC和DataStream
数据源
结构化数据
如RDBMS备库
半结构化数据
如JSON
非结构化数据
如音频文件
作业流开发
Azkaban
权限管理
Ranger
多租户管理
元数据管理
数据质量校验
DQC
秘钥管理
Kerberos
运维监控
Ambari

平台特色

统一元数据服务

- Hive, spark, impala,hbase元数据打通
- 数仓体系内，用户无需在不同的系统一之间做元数据同步
- 不同系统组件之间，数据全增量同步

流计算服务

- Sloth流计算服务化平台
- 通过增量计算的方式，来完成流计算任务
- 使用SQL作为开发方式，完全与离线SQL兼容，支持window/join/subquery/having等复杂SQL功能

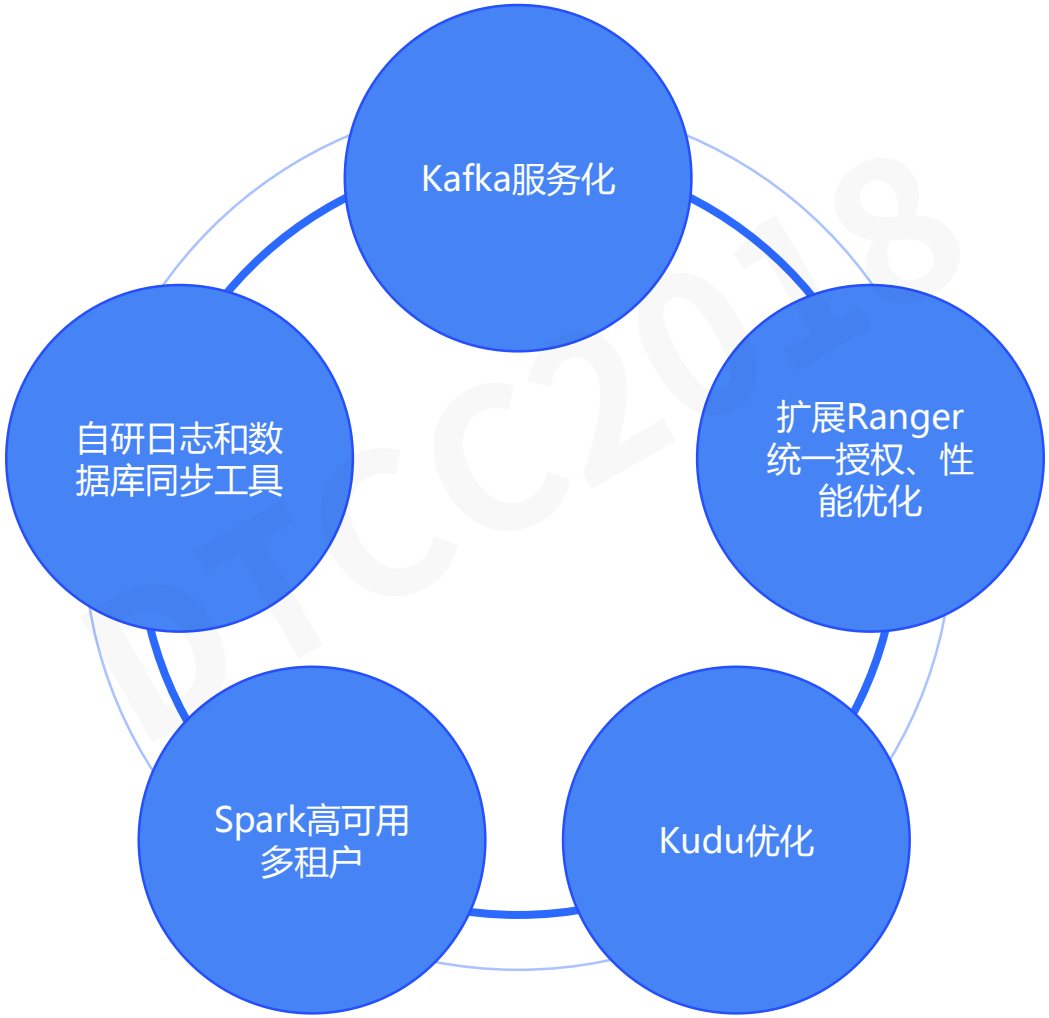
数据安全与权限

- HDFS/Hive/Impala/Spark等组件自动权限同步
- 支持到列级别的权限控制
- 基于角色访问控制，权限控制到个人
- 支持操作审计

一站式

- 一站式的数据平台
- 一站式的统一部署，监控，运维体系

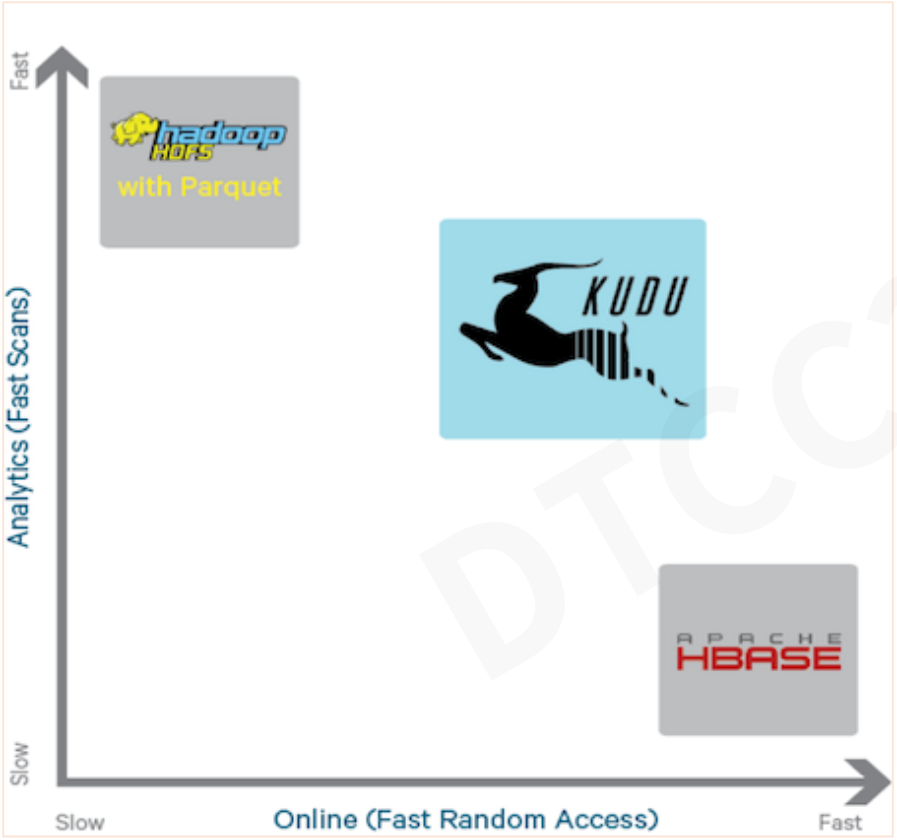
自研和开源相结合


Kafka服务化
扩展Ranger
统一授权、性
能优化
Kudu优化
Spark高可用
多租户
自研日志和数
据库同步工具

PART 02

Kudu：可更新存储

Kudu定位



HDFS：

批量数据写入能力，没有数据更改能力；在实时性要求较高的场景下，5~10min需要写入一个文件，造成小文件数量比较多，对NameNode压力较大；对大批量数据扫描比较又好，基本没有随机查询能力

HBase：

大批量数据写入能力；极高的随机数据读写能力；支持指定rowkey的update操作；扫描分析能力非常低下

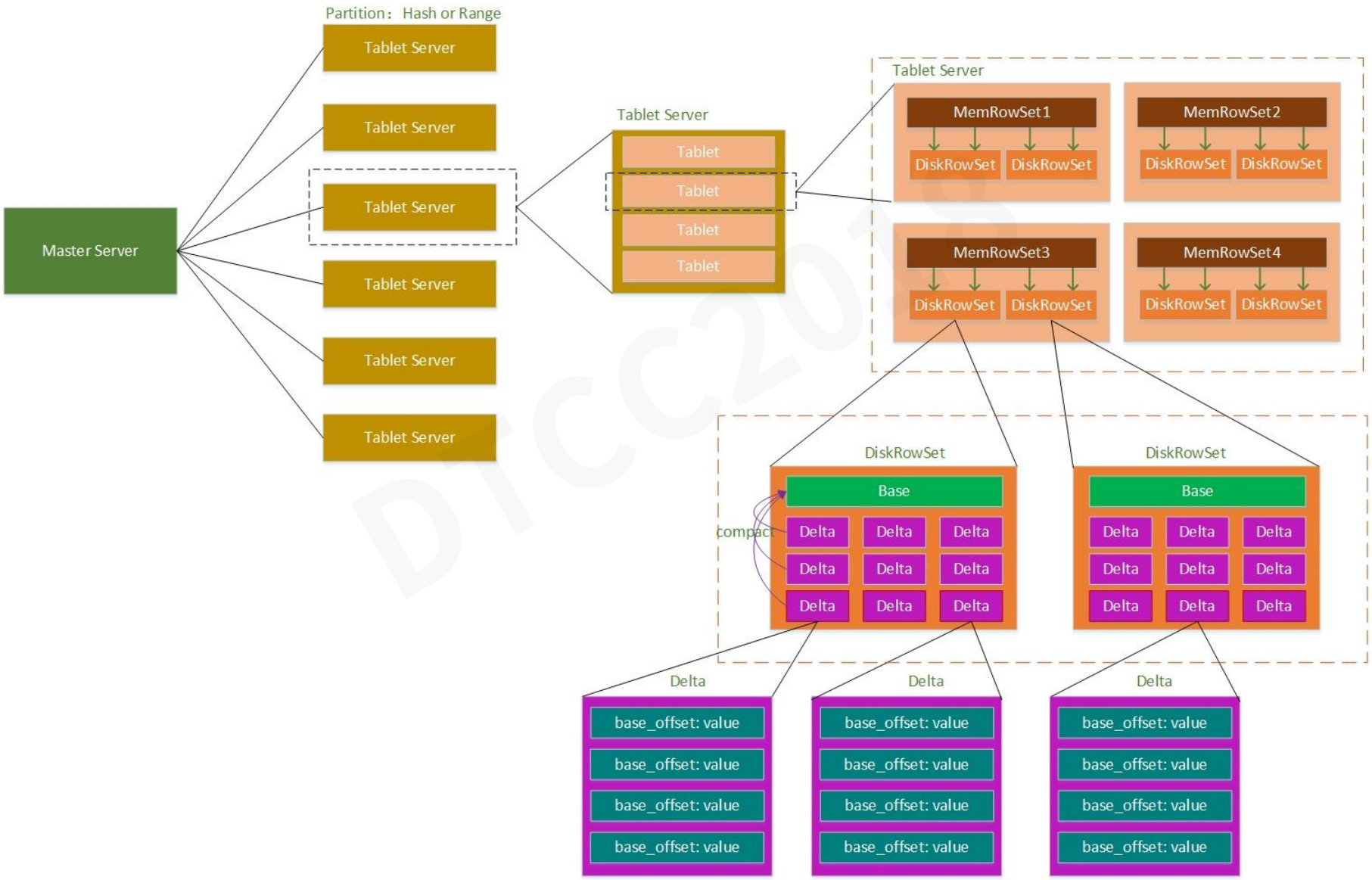
Kudu：

兼备HDFS大数据量写入与分析扫描能力，同时具备HBase的随机读写能力

与HBase对比

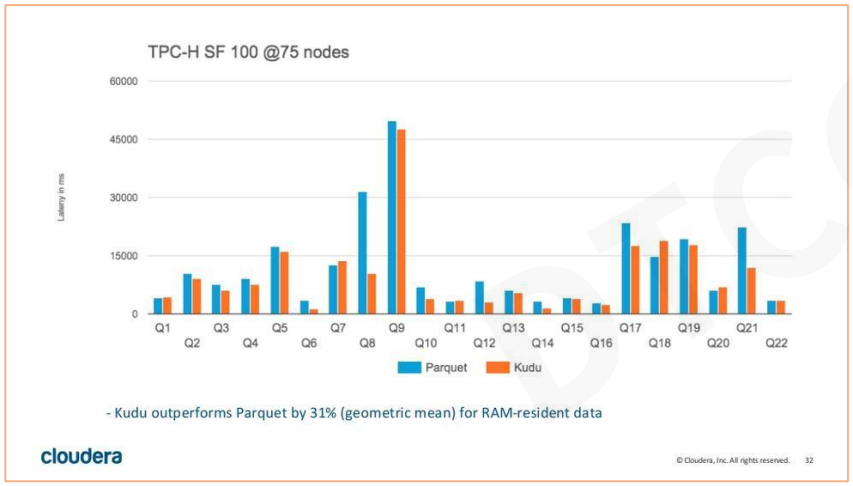
	HBase	Kudu
集群架构	Master-Slave结构	Master-Slave结构
选主方式	ZK选主	Raft内部自动选主
数据分布	Range方式分区	Range、HASH分区，支持组合分区
数据写入	HDFS(Pipleline)	Raft多副本
数据格式	ColumnFamily级别列存	RowGroup形式，同一个RG内部列存(类似Parquet)

Kudu原理


Partition： Hash or Range
Master Server
Tablet Server
Tablet Server
Tablet Server
Tablet Server
Tablet Server
Tablet Server
Tablet
Tablet
Tablet
Tablet
Tablet Server
MemRowSet1
MemRowSet2
MemRowSet3
MemRowSet4
DiskRowSet
DiskRowSet
DiskRowSet
DiskRowSet
DiskRowSet
DiskRowSet
DiskRowSet
DiskRowSet
compact
Base
Base
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
Delta
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value
base_offset: value

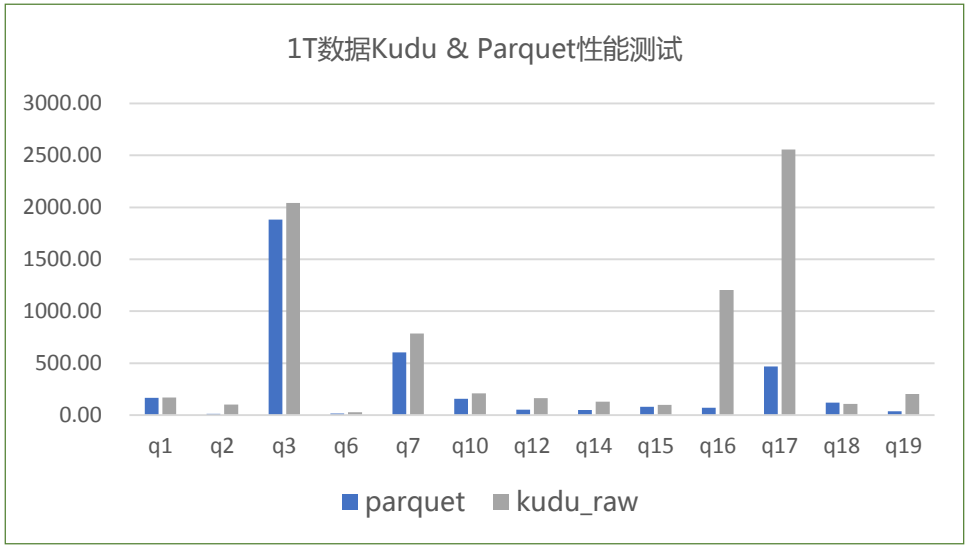
Kudu的缺陷

- Impala/Kudu与Impala/Parquet比有不小差距
- 没有Split & Merge功能



官方TPCH测试结果

结论：我们性能比Parquet就好那么一点点^_^



我们TPCH测试结果

大家都是搞技术的，还是诚实点好~_~

Kudu Runtime Filter

User表a(10万记录)

p289443643	1990-11-05	男	中国	广东	深圳
p297993524	1989-10-25	女	中国	江西	南昌
p302543202	1994-10-20	女	中国	广东	广州
p308578250	1990-12-02	女	中国	广东	广州
p347396619	1979-5-08	男	中国	广东	东莞
p358023170	1989-1-28	男	中国	辽宁	大连
p359123611	1993-4-27	女	中国	广东	阳江
p370138980	1996-9-30	男	中国	湖北	武汉
p402117135	1977-7-31	女	中国	北京	北京

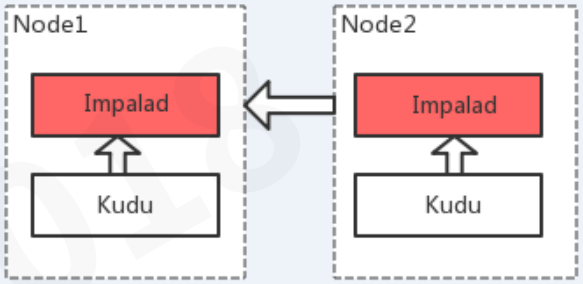
Event表b(10亿记录)

iPhone9,1	未知	750x1334	750	1334	中国移动	true	wifi	中文
iPhone9,2	未知	1242x2208	1242	2208	中国移动	true	wifi	中文
iPhone7,1	未知	1242x2208	1242	2208	中国移动	true	wifi	中文
iPhone8,1	未知	750x1334	750	1334	I WIND	true	wifi	中文
iPhone7,2	未知	750x1334	750	1334	Dialog	false	3G	English
iPad3,1	未知	640x960	640	960	未知	false	unreachable	中文
iPad3,1	未知	640x960	640	960	未知	false	unreachable	中文
iPad3,1	未知	640x960	640	960	未知	false	unreachable	中文

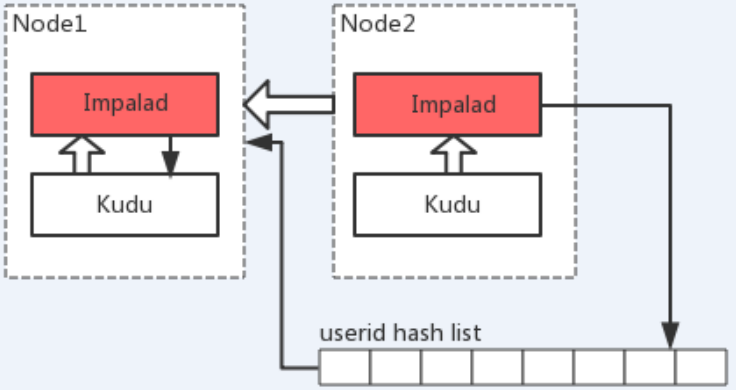
select xxx from user a, event b on a.userid = b.userid where xxx

通过runtime filter功能，小表的连接键被做成BF形式通过Impalad下发到Kudu节点，联合大表的连接键，在大表读取数据时参与数据的过滤，从而使得大表传递到Impalad层的数据大量减少，即在计算前减少参与计算的数据量，达到提升效率的结果

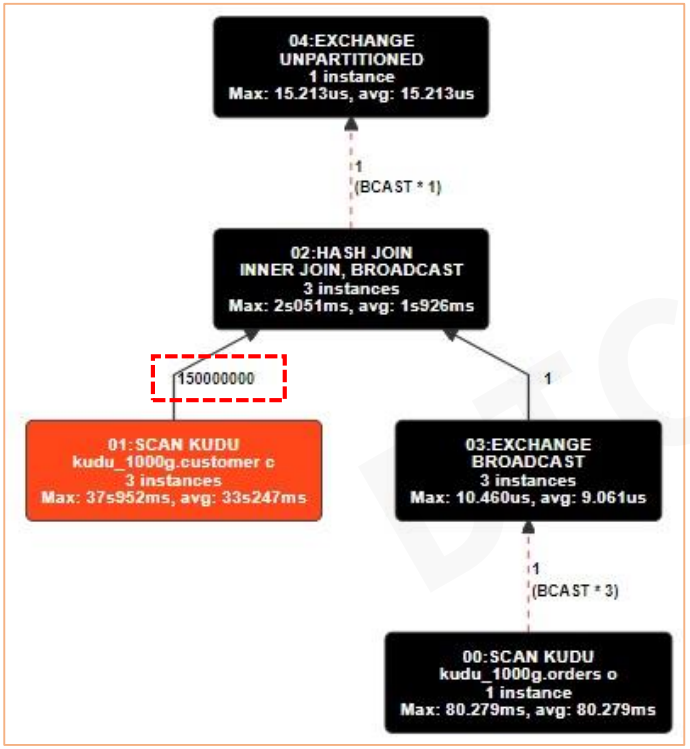
没有runtime filter

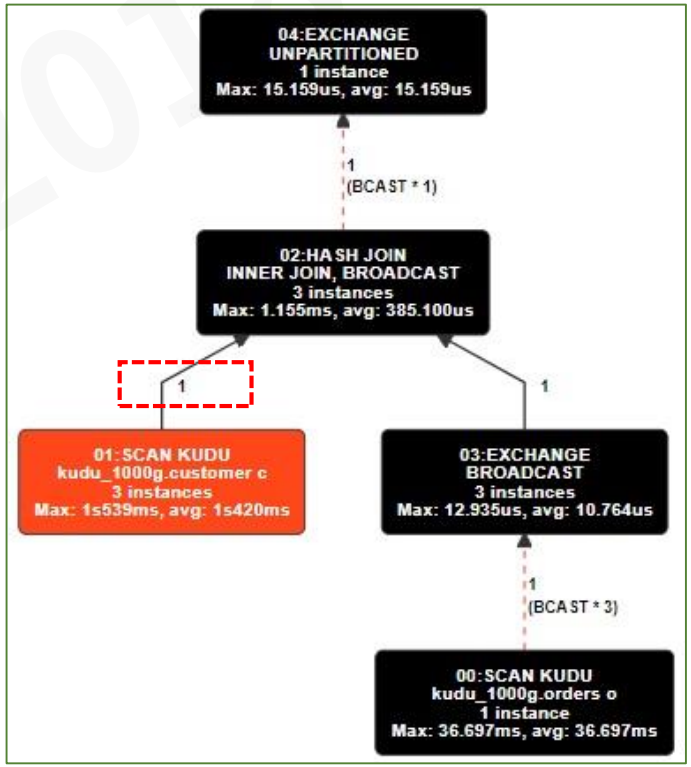


加入runtime filter功能

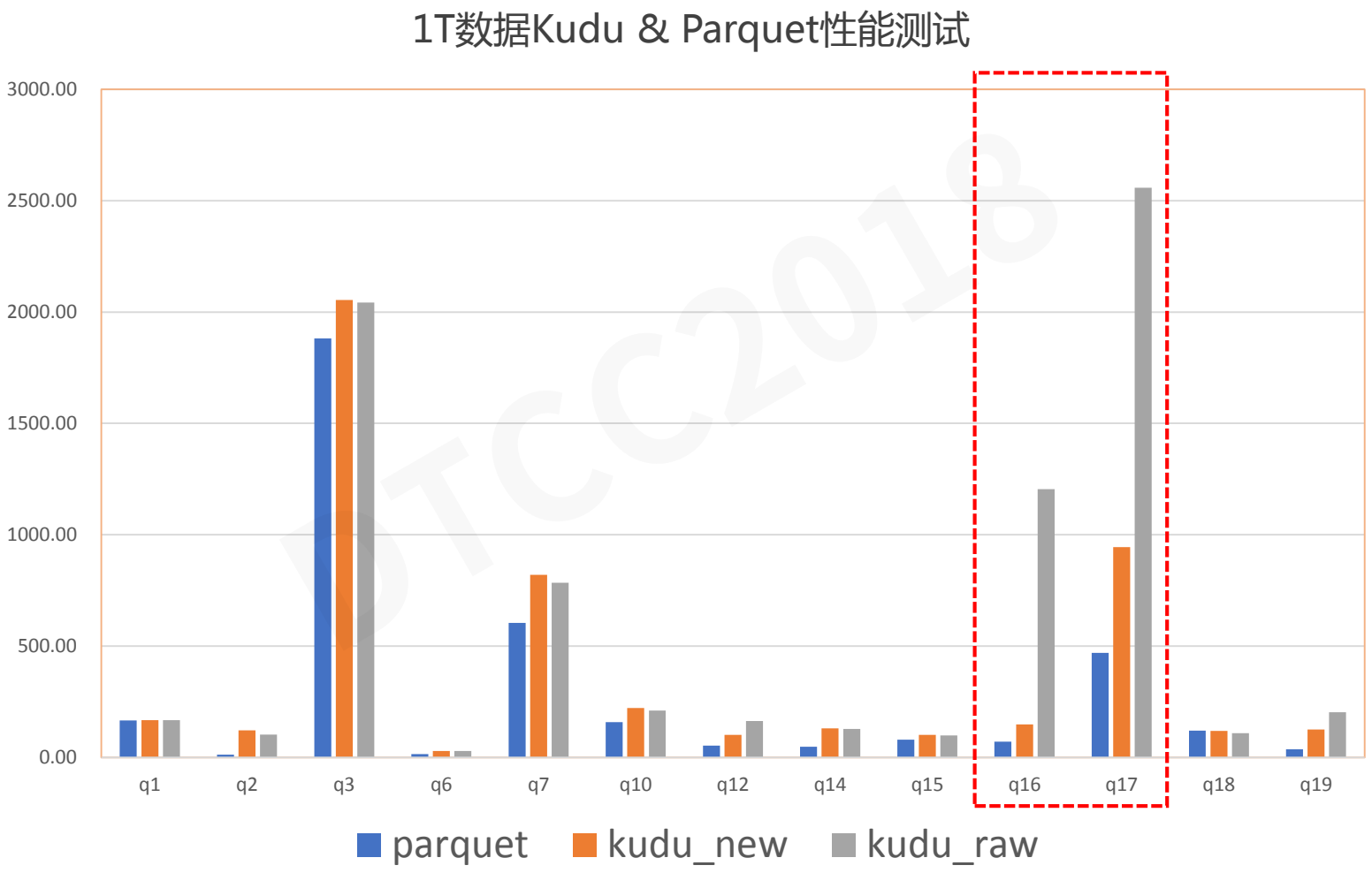


Kudu Runtime Filter



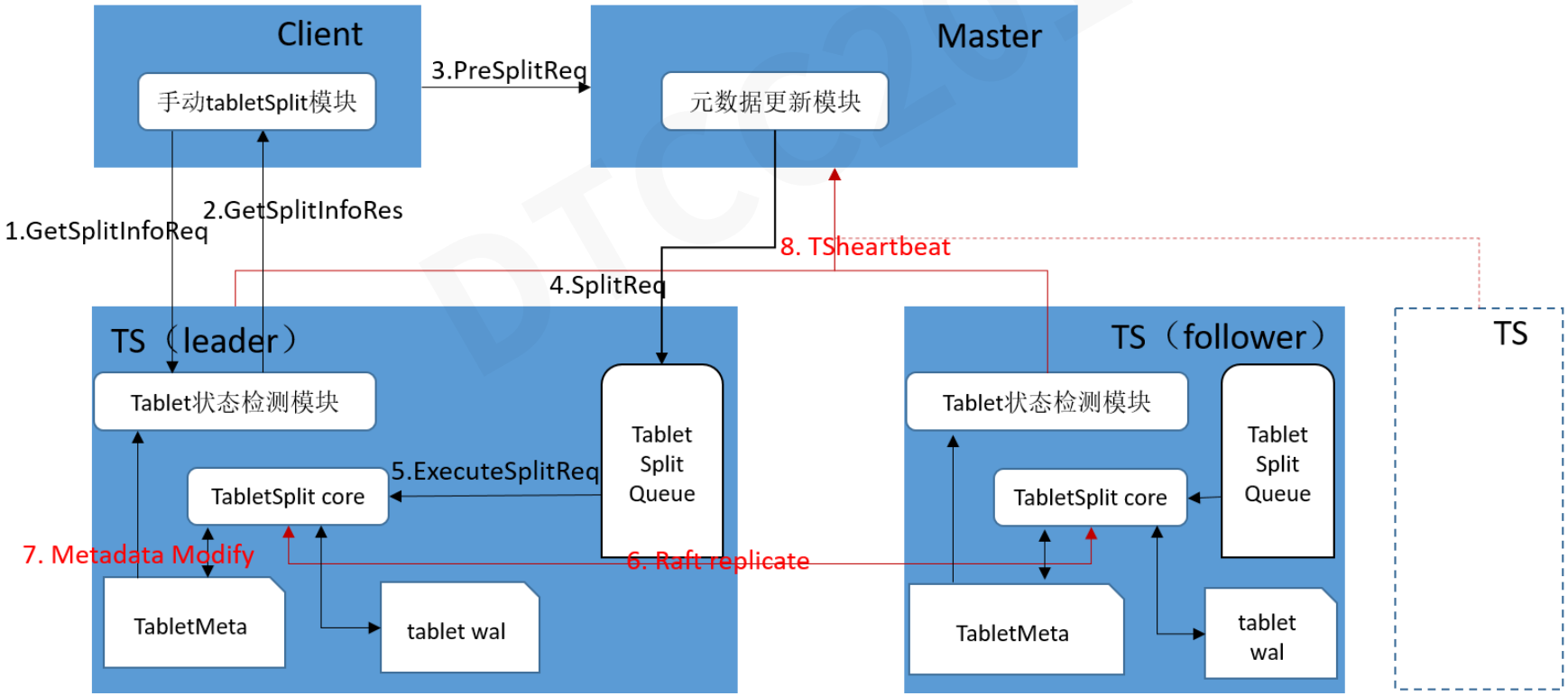


Kudu Runtime Filter



Kudu Tablet Split

- 支持Range分区分裂
- 仅修改元数据，在线完成分裂，compaction时再做物理分裂
- 主从协同



应用场景

01 秒级实时

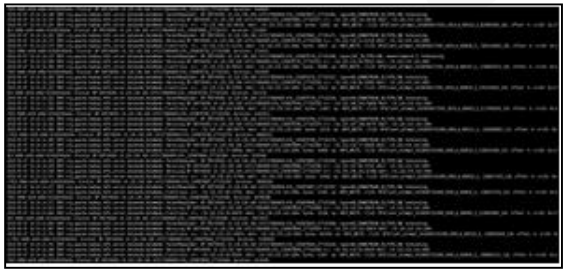
02 点查询和多维分析融合

03 实时维表

（一）秒级实时

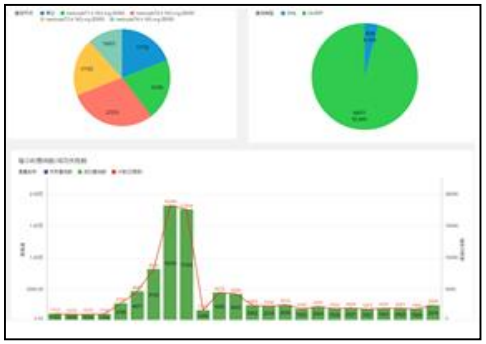
共享单车解决了出行最后一公里问题

Kudu解决分析数据最后半小时的实时性问题



实时写入

小批量写入



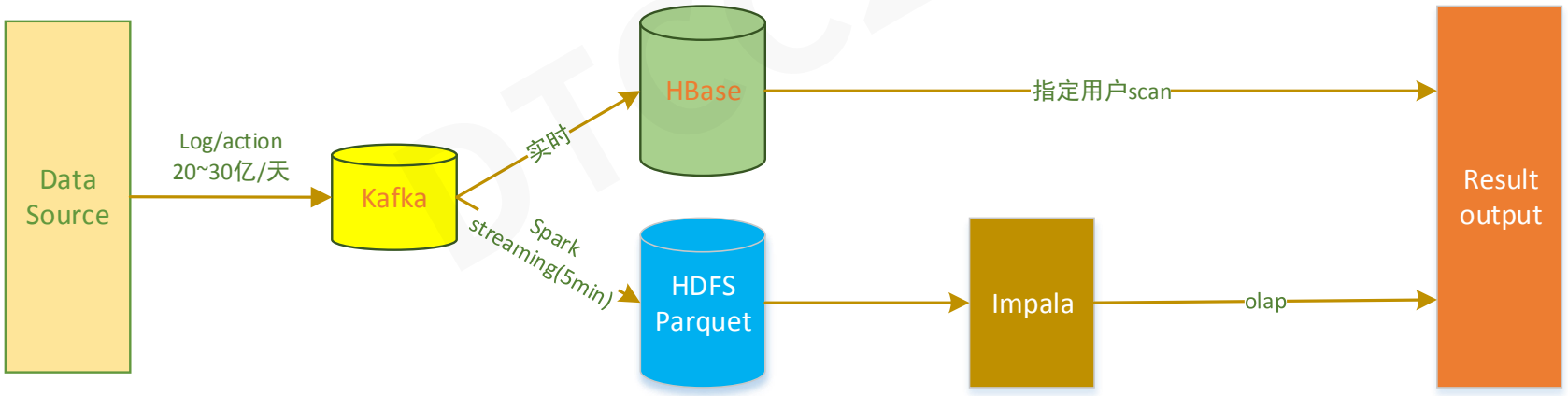
（二）点查和多维分析融合

游戏用户行为日志系统

游戏用户行为日志主要作用：

- 指定用户行为查询(给定用户id，查询某个时间段内的行为，可以进行反外挂等分析)
- 大批量用户行为分析(分析特定区域用户行为，比：如哪个区域玩家氪金较多？)

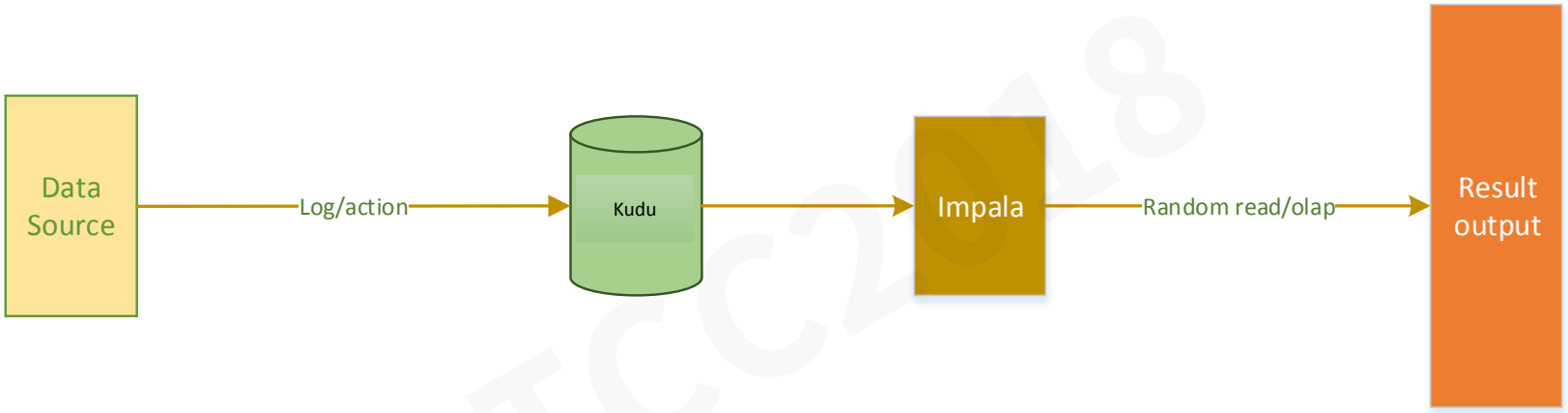
原先的架构



HBase： 指定用户id查询　　HDFS：批量用户行为分析

架构缺陷：两套系统，数据需要保存两份，6副本数据

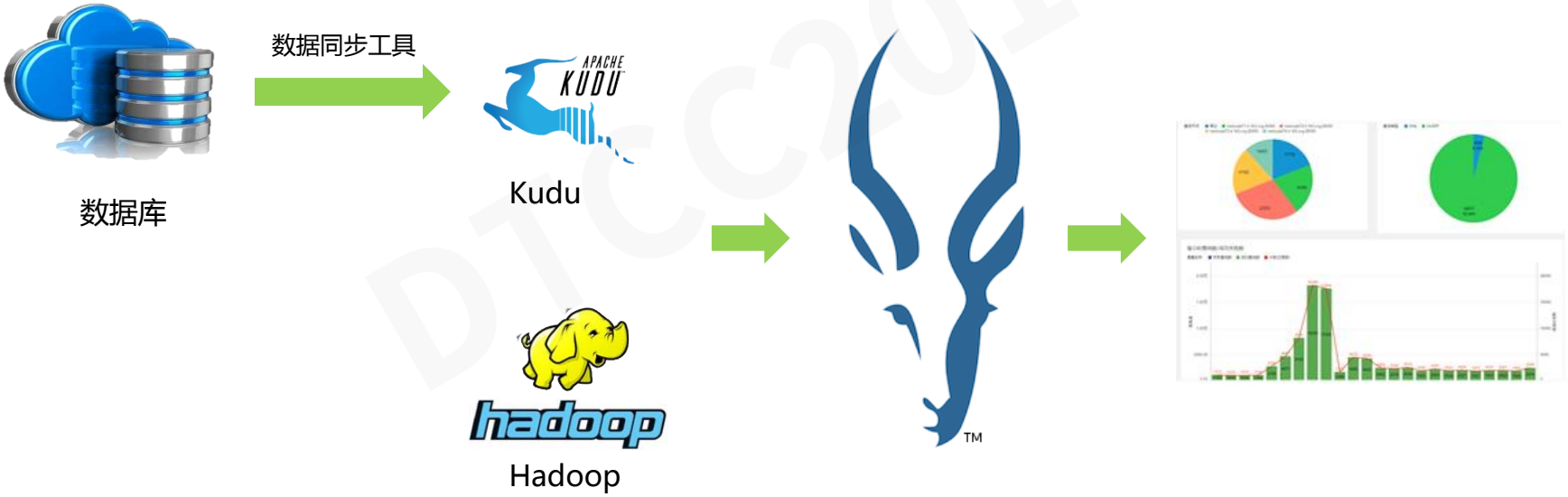
点查和多维分析融合



只需要一份数据，提供随机查询和数据分析

（三）实时维表

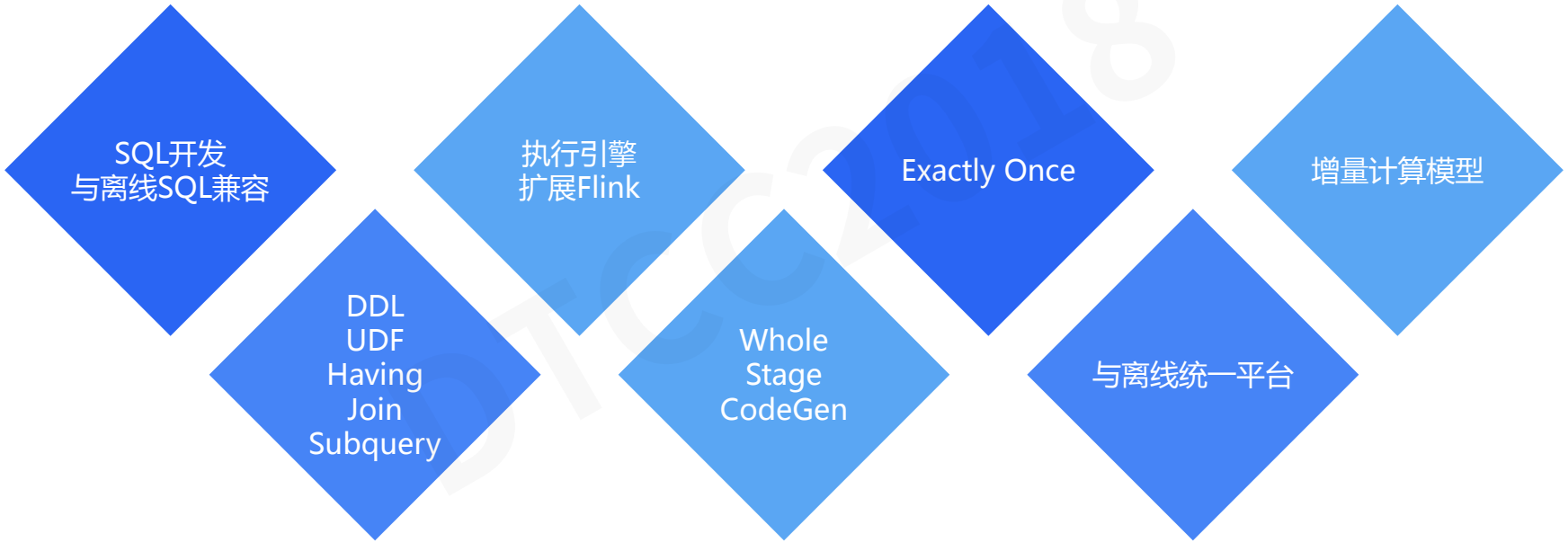
实时同步维度数据，联查分析



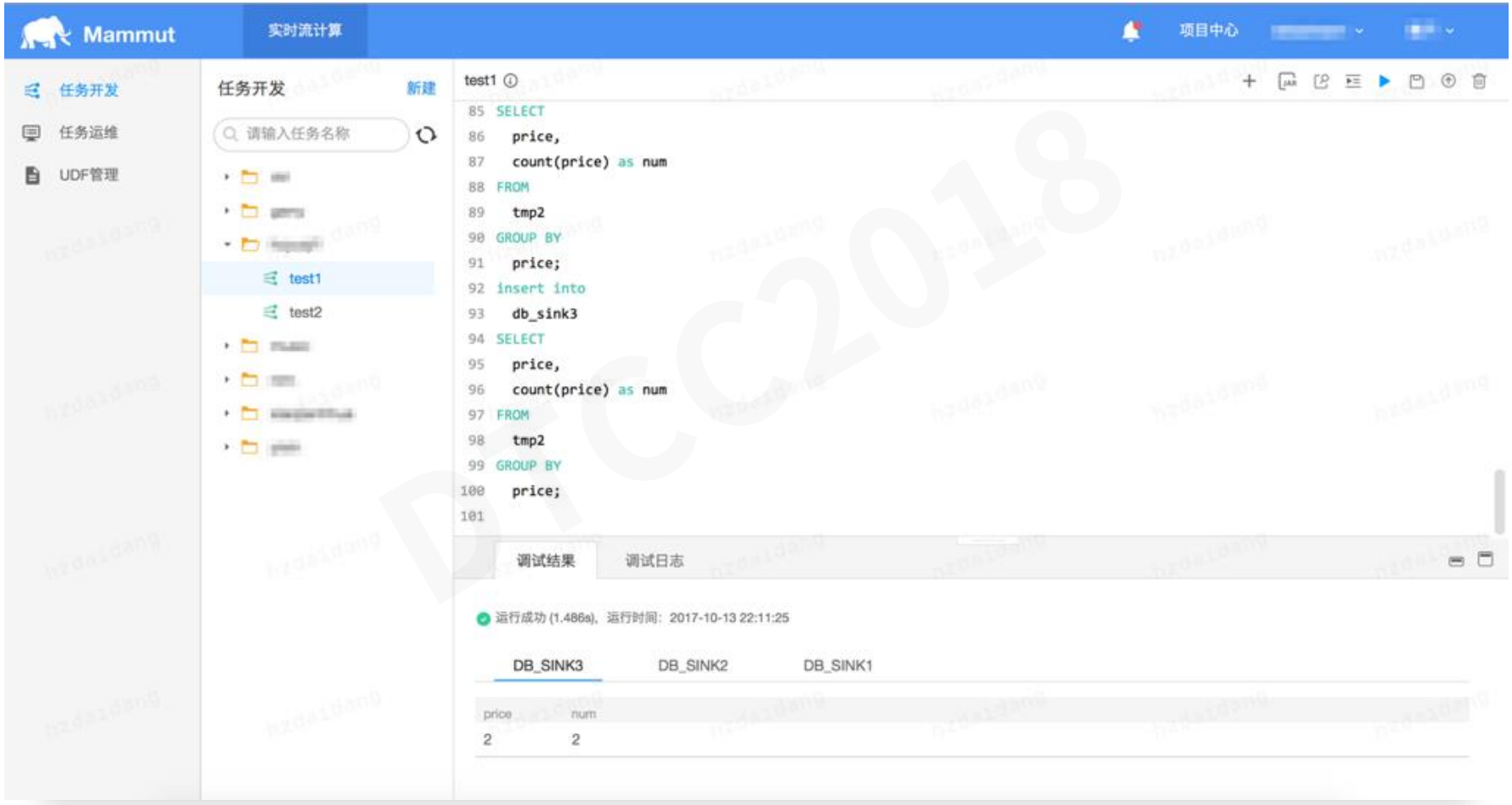
PART 03

Sloth：实时计算

Sloth特点


SQL开发
与离线SQL兼容
执行引擎
扩展Flink
Exactly Once
增量计算模型
DDL
UDF
Having
Join
Subquery
Whole
Stage
CodeGen
与离线统一平台

Sloth 开发平台


Mammut
实时流计算
项目中心
任务开发
任务运维
UDF管理
任务开发
新建
请输入任务名称
test1
test2
test1
SELECT
price,
count(price) as num
FROM
tmp2
GROUP BY
price;
insert into
db_sink3
SELECT
price,
count(price) as num
FROM
tmp2
GROUP BY
price;
调试结果
调试日志
运行成功 (1.486s)，运行时间：2017-10-13 22:11:25
DB_SINK3
DB_SINK2
DB_SINK1
price
num
2
2

Sloth 增量计算

案例： 所有商家按销售额做分类统计，销售额在[0,100]区间内的归为一类，[100,200]区间的的归为一类，以此类推，通过计算输出每个区间内的商家个数。

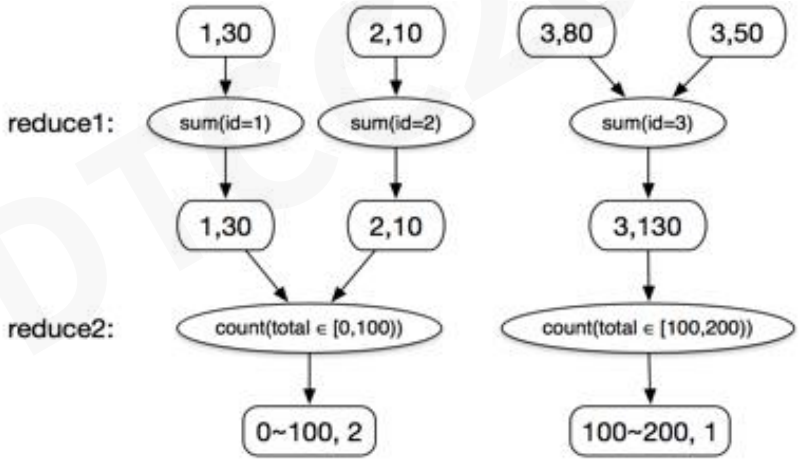
这个任务可以用SQL定义为：

```
-- stage1：计算每个商家的销售总额
INSERT INTO tmp
SELECT seller_id, sum(payment) as total FROM source GROUP BY seller_id;
-- stage2: 计算每个销售额区间内的商家个数
SELECT count(seller_id) as num, total/100 as range FROM tmp GROUP BY (total/100);
```

Sloth 增量计算

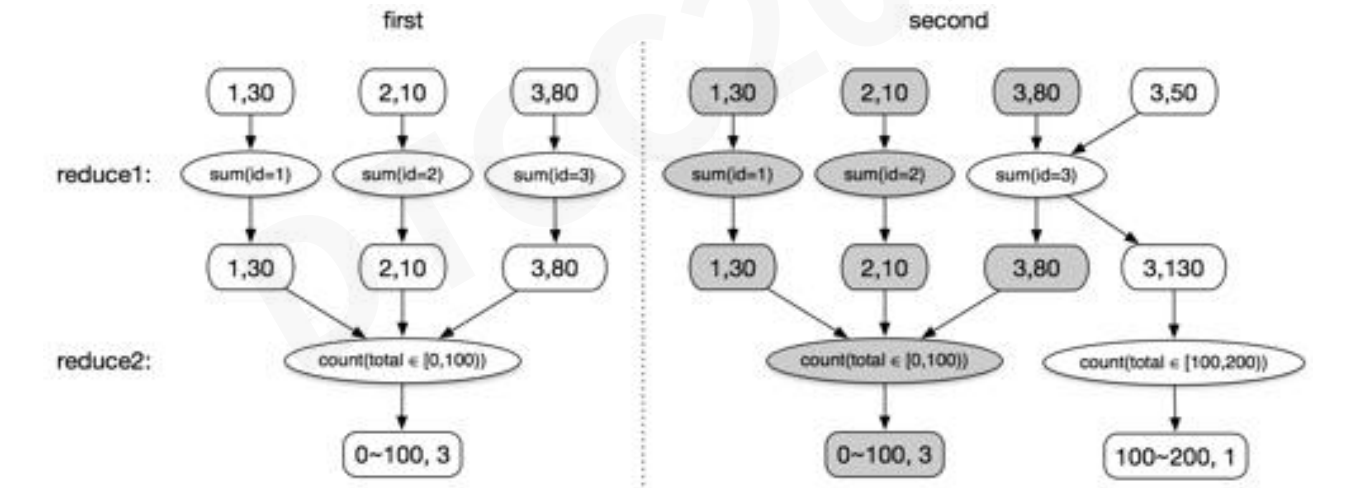
Sloth 增量计算

输入数据 | 离线计算 | 流式计算 | 增量计算



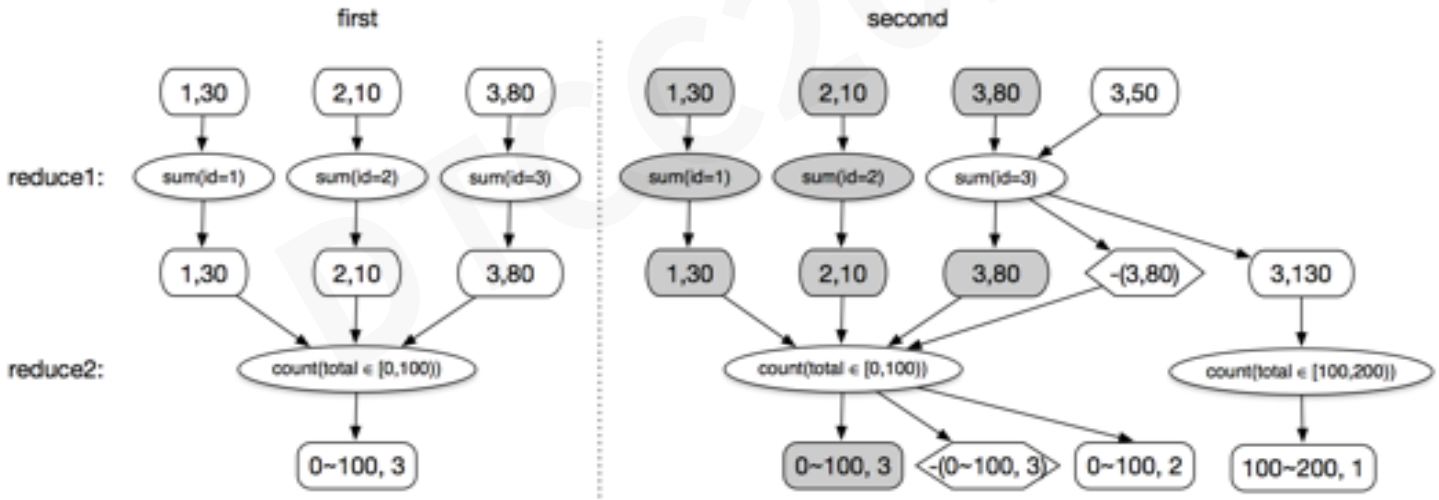
Sloth 增量计算

输入数据　离线计算　流式计算　增量计算

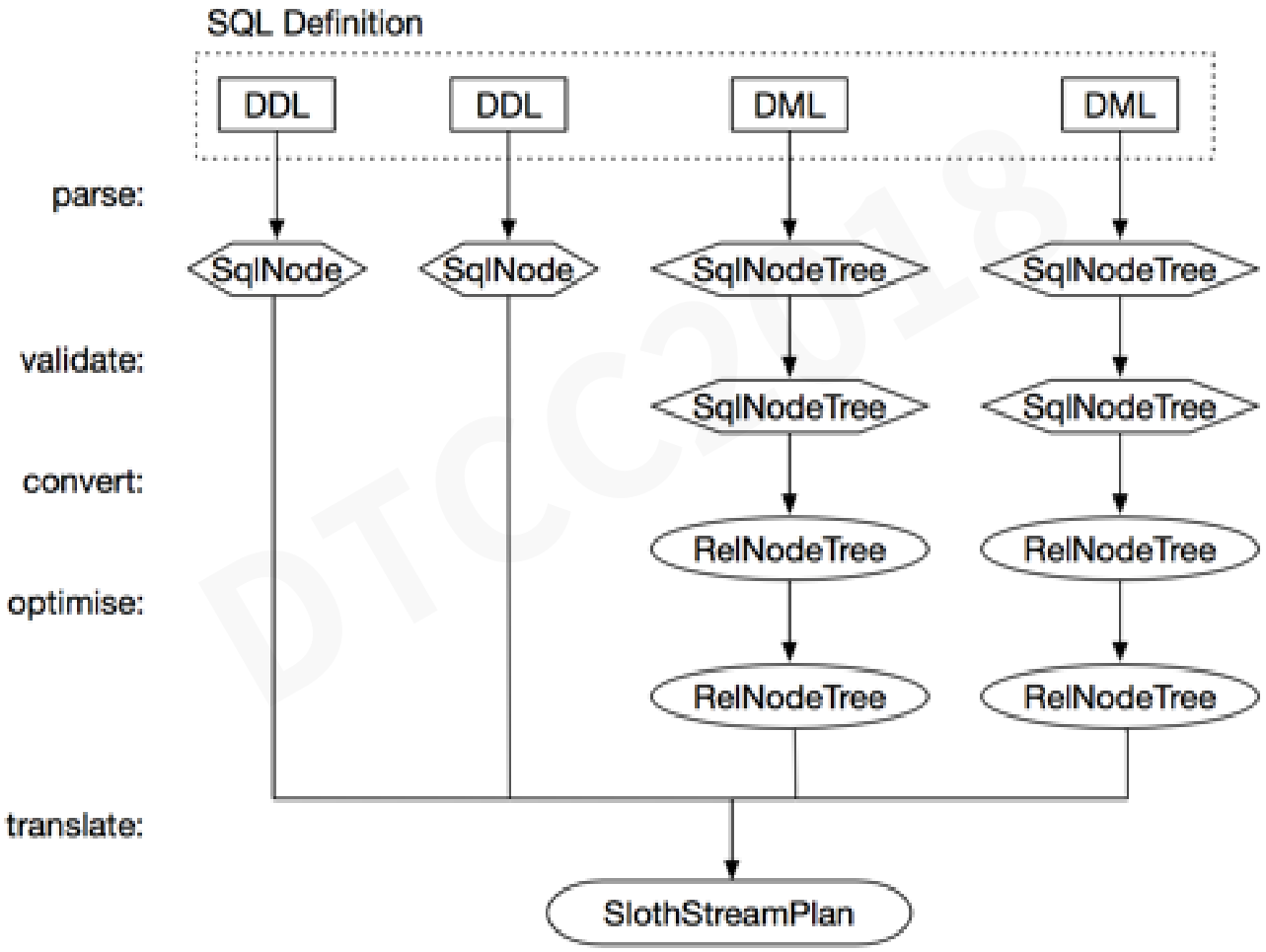


Sloth 增量计算

输入数据　离线计算　流式计算　**增量计算**



Sloth 执行计划生成



PART 04

Kyuubi：Spark 多租户

Kyuubi V.S. HiveSever2

相同点：

- xDBC方式，不依赖语言，方便与应用进行数据交互;
- 应用侧不依赖Hive/Hadoop libs, 部署简单，无客户端兼容性问题；
- MetaStore / HDFS 配置不可见，无数据及元数据泄露问题；
- Authentication & Authorization，解决数据安全问题；
- HA机制，动态扩容，解决应用侧的并发和负载均衡问题；

不同点：

- Spark SQL vs MapReduce

Kyuubi V.S. Spark Thrift Server

1. 多租户

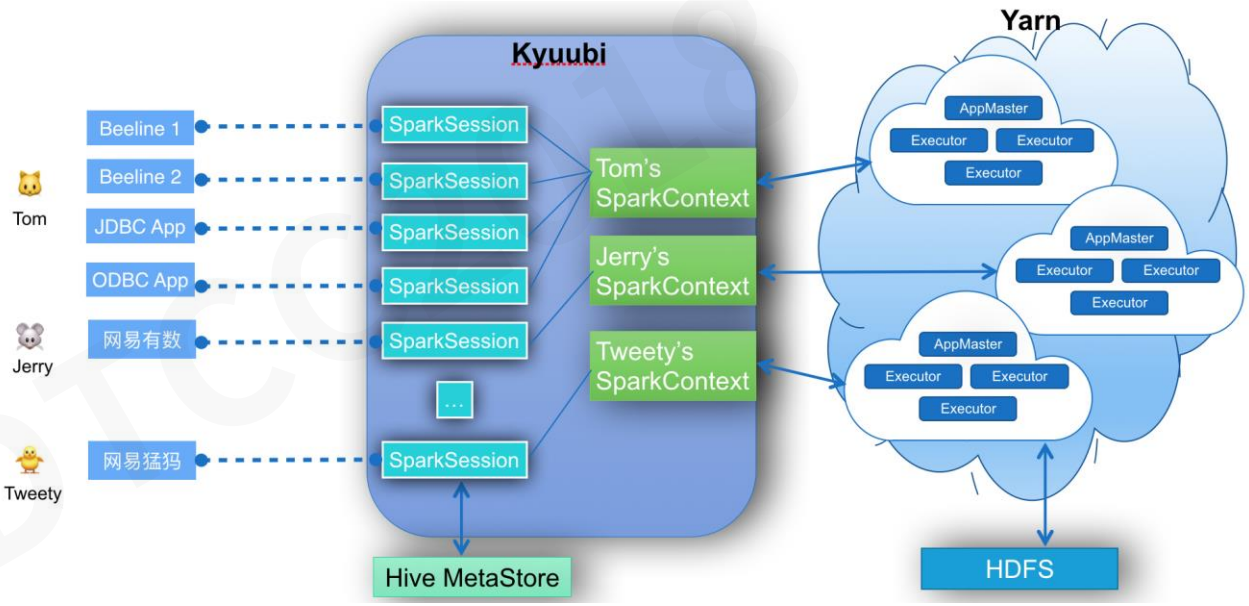
- SparkContext多实例
- Hadoop Impersonation

2. Spark SQL as a Service

- High Availability
- Operation log

3. Security

- Kerberos Support
- Row / Column Level Authorization



Kyuubi：统一接口

无缝衔接 易于集成

一套接口规范

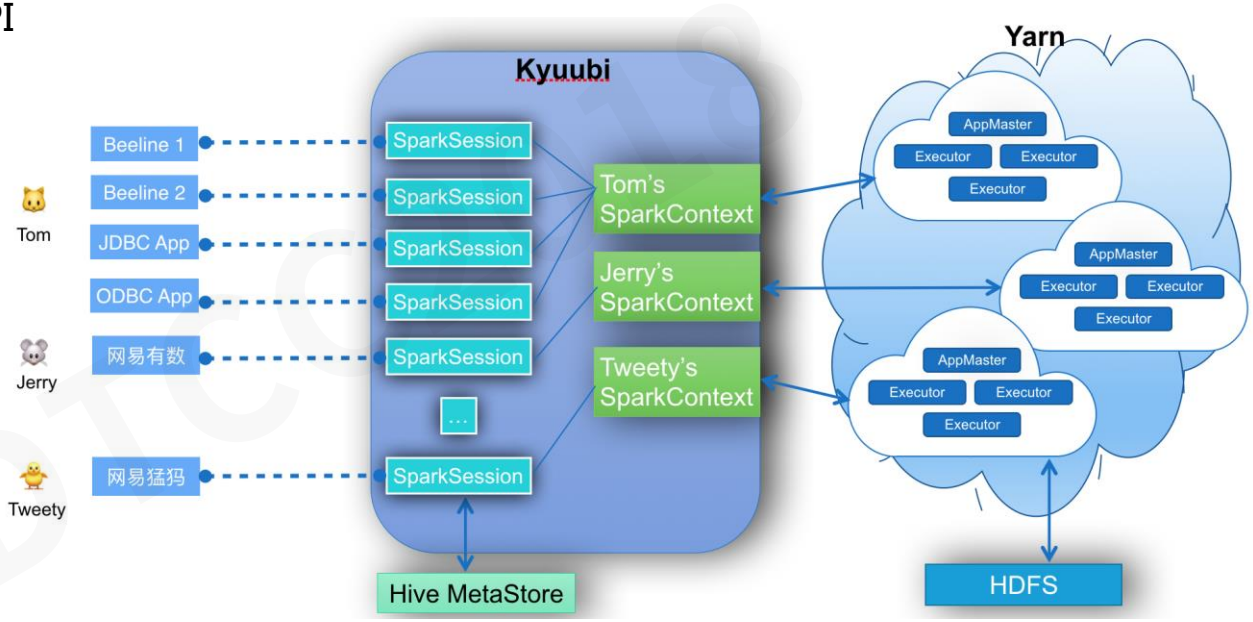
- HiveServer2 Thrift API

两种执行引擎

- Hive
- Spark SQL

多种连接方式

- Beeline
- JDBC
- ODBC
- 猛犸
- 有数
- ……



Kyuubi：弹性资源

资源高效 控制成本

Kyuubi动态资源申请

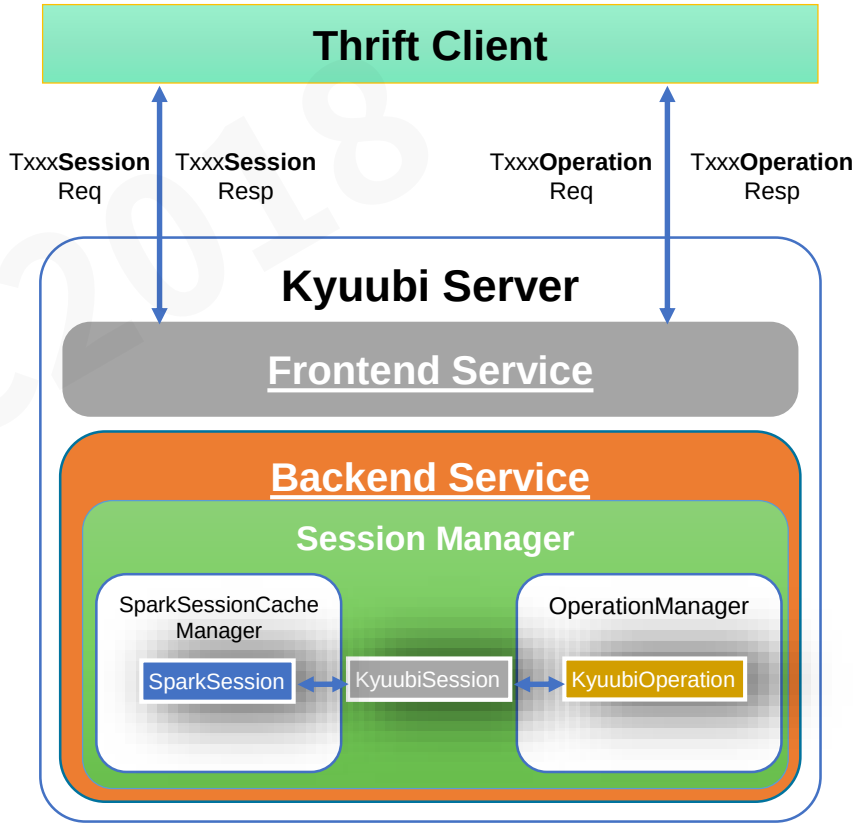
- 支持session级别配置：队列/核数/内存/等等
- 示例：
  jdbc:hive2://host:port/;hive.server2.proxy.user=tom#spark.yarn.queue=theque;spark.executor.instances=3;spark.executor.cores=3;spark.executor.memory=10g

Kyuubi动态缓存SparkContext

- 基于用户连接创建、注册
- 基于空闲策略缓存、回收

Spark动态资源分配特性

- spark.dynamicAllocation.enabled
- spark.dynamicAllocation.minExecutors
- spark.dynamicAllocation.maxExecutors
- .....



Kyuubi：安全特性

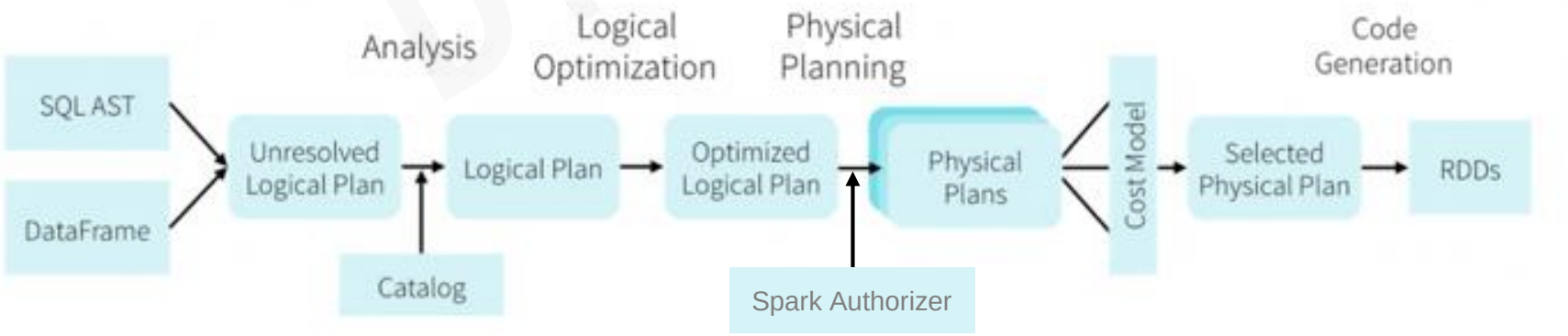
安全访问 权限隔离

安全认证

- Kerberos Support
- Hadoop Impersonation代理执行

权限控制

- 集成spark-authorizer GitHub: https://github.com/yaooqinn/spark-authorizer
- Row / Column Level Access Control
- 端到端数据安全隔离

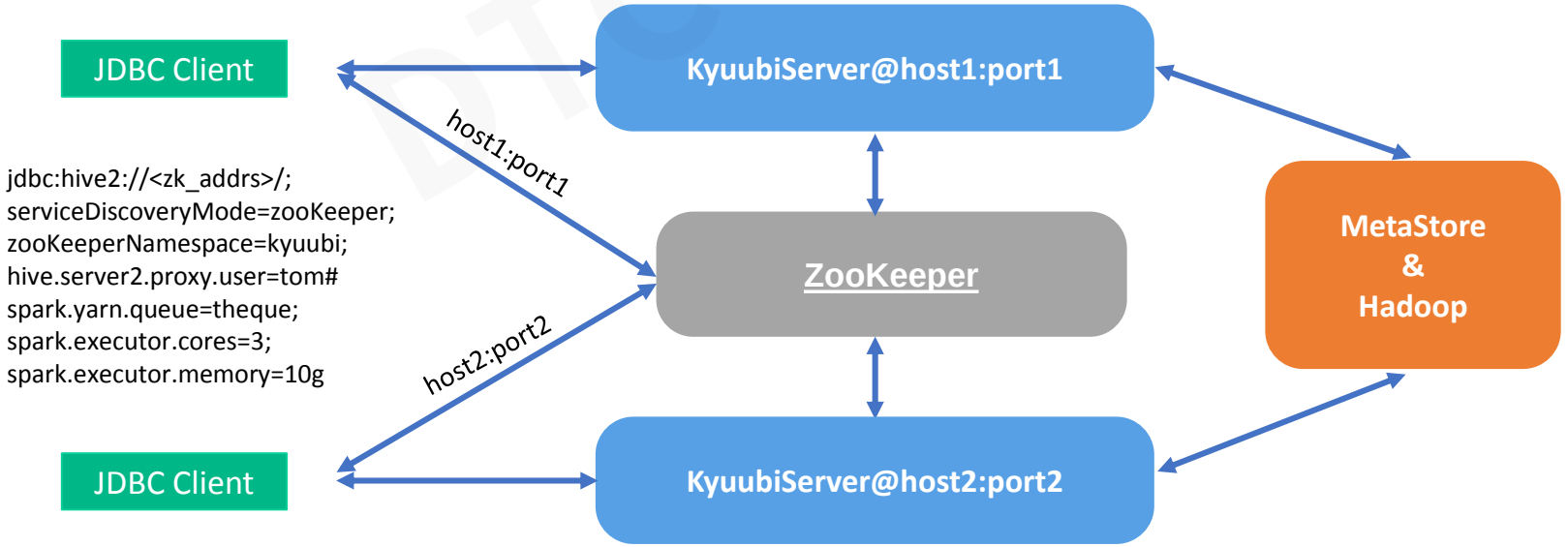


Kyuubi：服务可用性

水平扩展 负载均衡

基于ZooKeeper Namespace Discovery实现

- 并发
- 负载均衡
- 单点故障
- 服务可用性
- 兼容HiveServer2



总结

Kyuubi

一个以HiveServer2 Thrift API为接口协议，以Spark SQL的内置处理引擎的即席查询服务。它对于Spark自带的Thrift Server相比增加了适应于多租户场景下的特性

- 无缝衔接 易于集成
- 资源高效 控制成本
- 安全访问 权限隔离
- 水平扩展 负载均衡

Get Kyuubi

- GitHub: https://github.com/yaooqinn/kyuubi

PART 05

未来规划

未来规划

01 高性能查询引擎

02 离线和实时计算混部

03 新硬件加速GPU/FPGA

04 智能任务诊断和优化

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