UCLOUD

网易大数据
Netease Big Data

Ucan 下午茶
悟有所值
Ucan Afternoon Tea

基于Impala打造交互查询系统

网易杭州研究院

蒋鸿翔

CONTENTS

01 交互查询特点
02 选型因素
03 Impala 基本介绍
04 Impala 改进
05 Impala 使用场景
06 场景方案解析

Ucan
下午茶
悟有所值

PART 01

交互查询特点

网易大数据
Netease Big Data

交互查询特点

数据量基数庞大

关系模型相对较复杂

响应时间要求较高

数据基数庞大

根据数据量选用不同的存储

> 百亿级别，离线数仓，Hive，Spark

百万 ~ 百亿级别，分析型数仓，Impala，Presto，GreenPlum，Apache Drill

< 百万级别，普通数据库，MySQL，PostgreSQL

关系模型复杂

宽表型应用

ElasticSearch，Solr，ClickHouse，Druid

组合模型(星型 or 雪花型)

GreenPlum，Presto、Impala

响应时间要求高

应用场景

在线系统查询

(推荐系统、营销系统等)

BI报表

PART 02

选型因素

选型3大因素

本身项目熟悉度

大厂背书

性能和优缺点指标

选型因素

- 数据量
- 性能要求
- 架构复杂度
- 运维成本开销

选型因素

Impala

vs Presto

vs GreenPlum

- 与Hadoop/Hive兼容
- 成熟的社区环境
- 成熟的架构设计
- 查询性能
- 其他特性

悟有所值

PART 03

Impala 基本介绍

网易大数据
Netease Big Data

Impala 基本介绍

A MPP架构

B 执行节点无状态，去中心化

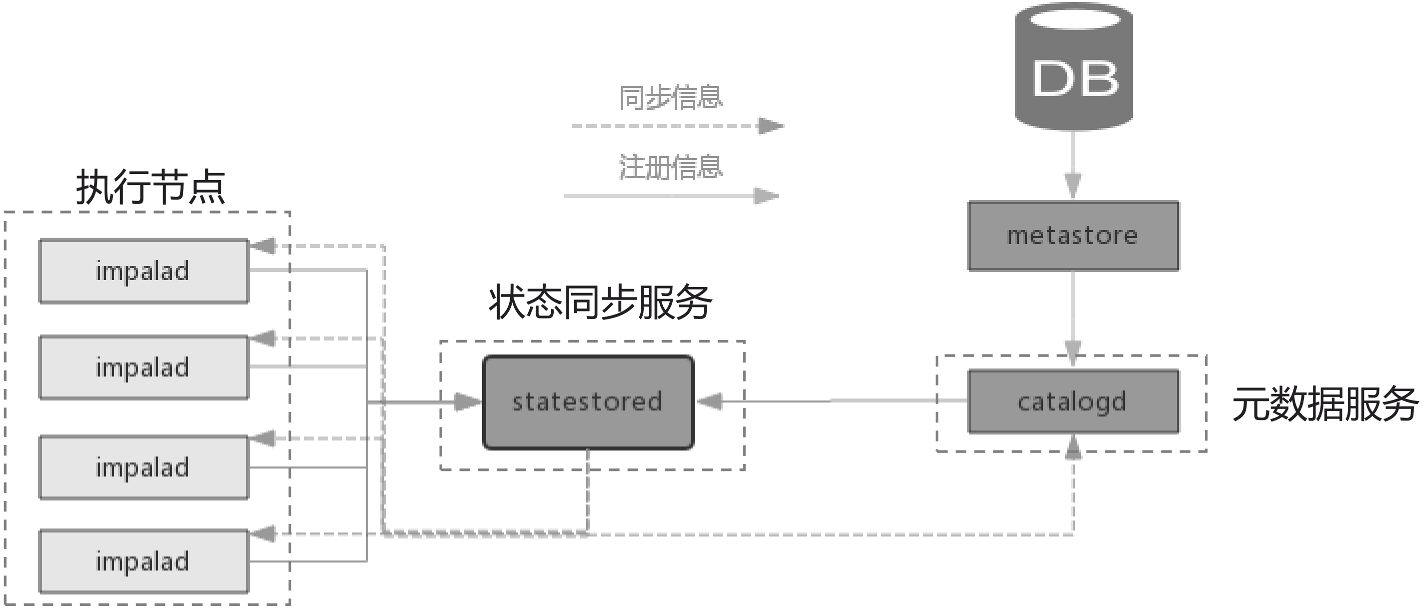
C 兼容Hive存储

D Apache顶级项目，成熟社区

E 多种数据格式兼容

F 高效的查询性能

Impala 基本介绍


同步信息
注册信息
DB
metastore
执行节点
impalad
impalad
impalad
impalad
状态同步服务
statestored
catalogd
元数据服务

Impala 基本介绍

Catalogd

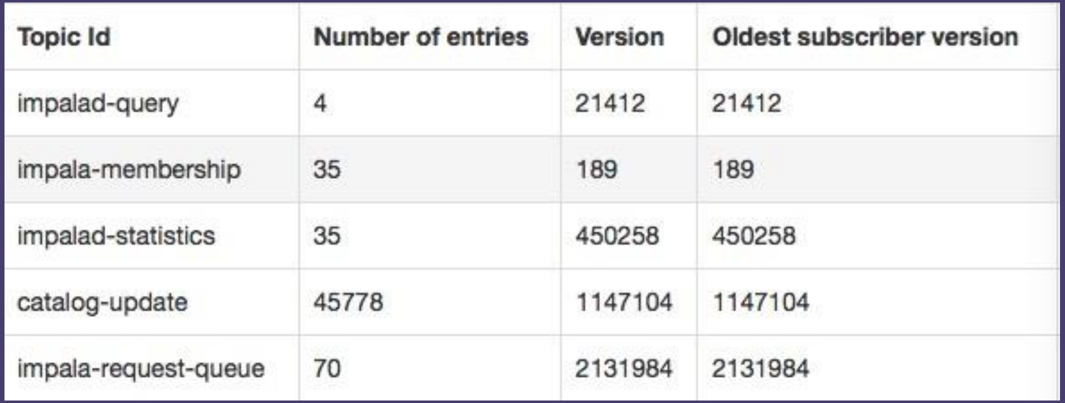
- 缓存表元数据
- 缓存HDFS数据块元数据
- DDL操作需要绑定catalogd

Statestored(sub/pub服务)

- Impalad节点状态信息
- 节点资源统计信息
- 元数据信息

Impalad

- SQL解析
- 执行计划生成
- 数据查询、聚合、结果返回
- 非DDL操作不需要catalogd参与



Topic Id	Number of entries	Version	Oldest subscriber version
impalad-query	4	21412	21412
impala-membership	35	189	189
impalad-statistics	35	450258	450258
catalog-update	45778	1147104	1147104
impala-request-queue	70	2131984	2131984

Statestored’s topics

Impala 基本介绍

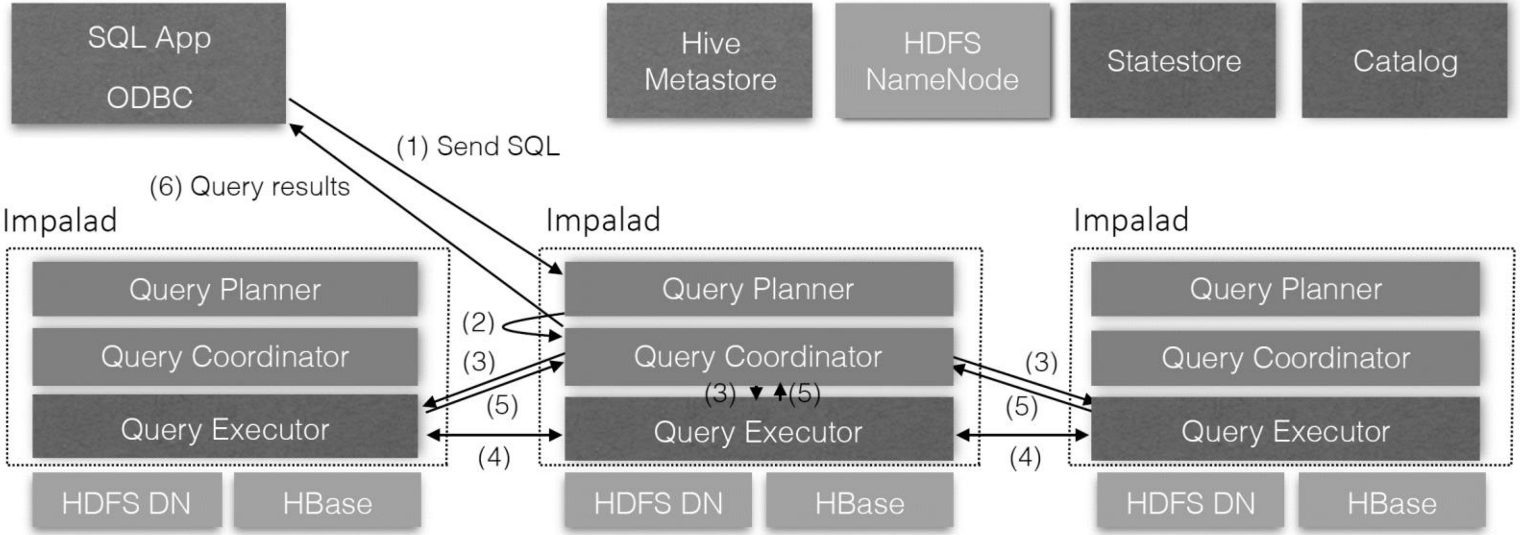

SQL App
ODBC
Hive Metastore
HDFS NameNode
Statestore
Catalog
(1) Send SQL
(6) Query results
Impalad
Impalad
Impalad
Query Planner
Query Coordinator
Query Executor
Query Planner
Query Coordinator
Query Executor
Query Planner
Query Coordinator
Query Executor
(2)
(3)
(5)
(4)
(3)
(5)
(3)
(5)
(4)
HDFS DN
HBase
HDFS DN
HBase
HDFS DN
HBase

Figure 1: Impala is a distributed query processing system for the Hadoop ecosystem. This figure also shows the flow during query processing.

Impala 基本介绍

高性能因素

- 元数据缓存
- MPP并行计算
- Codegen 技术，支持 LLVM，JIT
- 支持HDFS 本地读 (shortcircuit)
- 算子下推 (>,<,In List, runtime filter)

Impala 基本介绍

并行计算

并行连接（parallel join）

广播连接
broadcast join

被连接表被传输到所有节点进行连接

适用连接表较小

分区连接
partition join

hash partitioned

适用大表

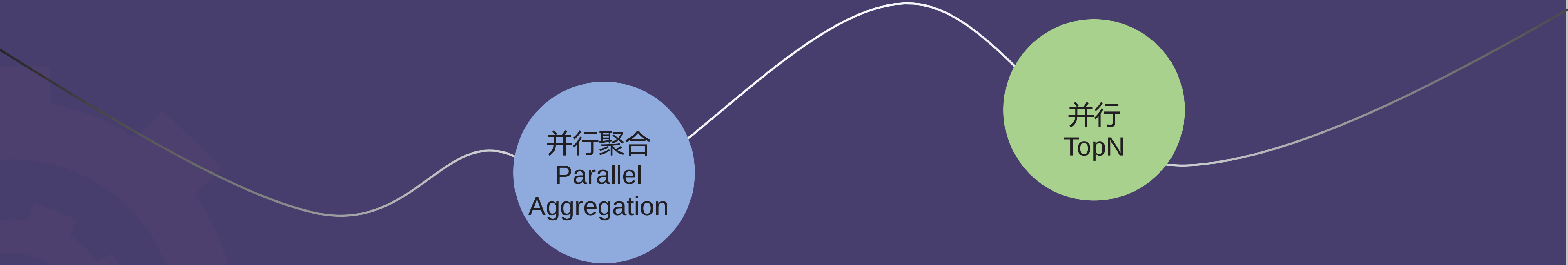
CBO

列统计（compute stats，只对hdfs有效）

数据传输代价

Impala 基本介绍

并行计算


并行聚合
Parallel
Aggregation
并行
TopN

数据生成后进行最初的TopN操作
在单节点计划上进行最后的TopN操作

数据生成时提前聚合

根据group by的column 进行聚合值的合并操作

Impala 基本介绍

存储支持

HDFS

Amazon S3

Azure Data Lake

EMC Isilon storage

HBase

Kudu

Alluxio

MySQL

Impala 基本介绍

授权 & 认证

认证

- ■ Kerberos
- ■ LDAP
- ■ Audit log

授权

- ■ Apache Sentry
  - 粒度：database，table，column
  - 权限：select， insert，all
  - 配置开启(authorization_policy_provider_class=org.apache.sentry.provider.file.LocalGroupResourceAuthorizationProvider)

Impala 基本介绍

资源管理

Admission Control

- fair-scheduler.xml（fair_scheduler_allocation_path）
- llama-site.xml（llama_site_path）

Impala节点通过statestored的impalad-statistics topic项同步资源信息，由于statestored通过心跳与impalad保持通信，这个资源信息实际上有些延迟；目前配置中，只有内存项有实际效果，vcore没有实现隔离，队列名配置如果与认证用户名相同，该用户提交的SQL自动分配到该队列

Impala 基本介绍

Web端

端 口

Impalad: 25000

catalogd：25020

statestored：25010

基本信息

集群节点

Catalog信息

内存信息

Query信息

……

Impala 基本介绍

Web端能做什么

节点内存消耗查看

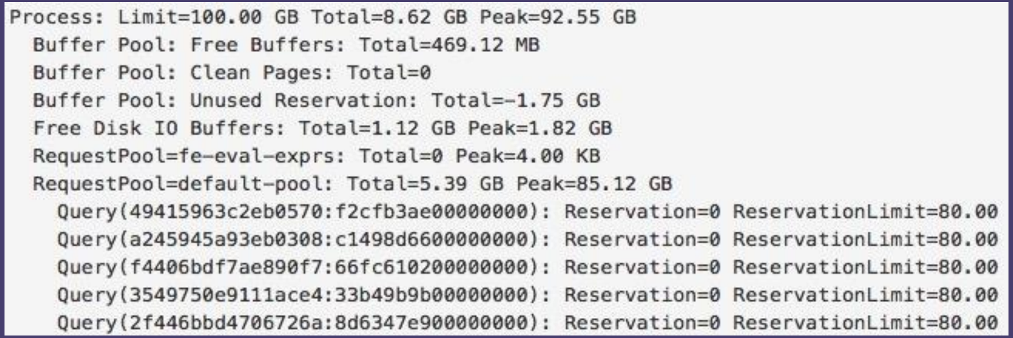
- 每个队列的内存消耗
- 每个查询在该节点内存消耗

该节点上查询分析

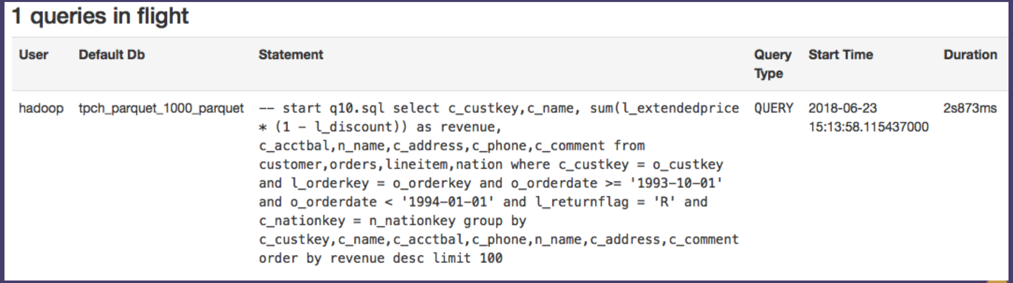
- 查询分析
- SQL诊断
- 异常查询终止

Metrics信息查看

……



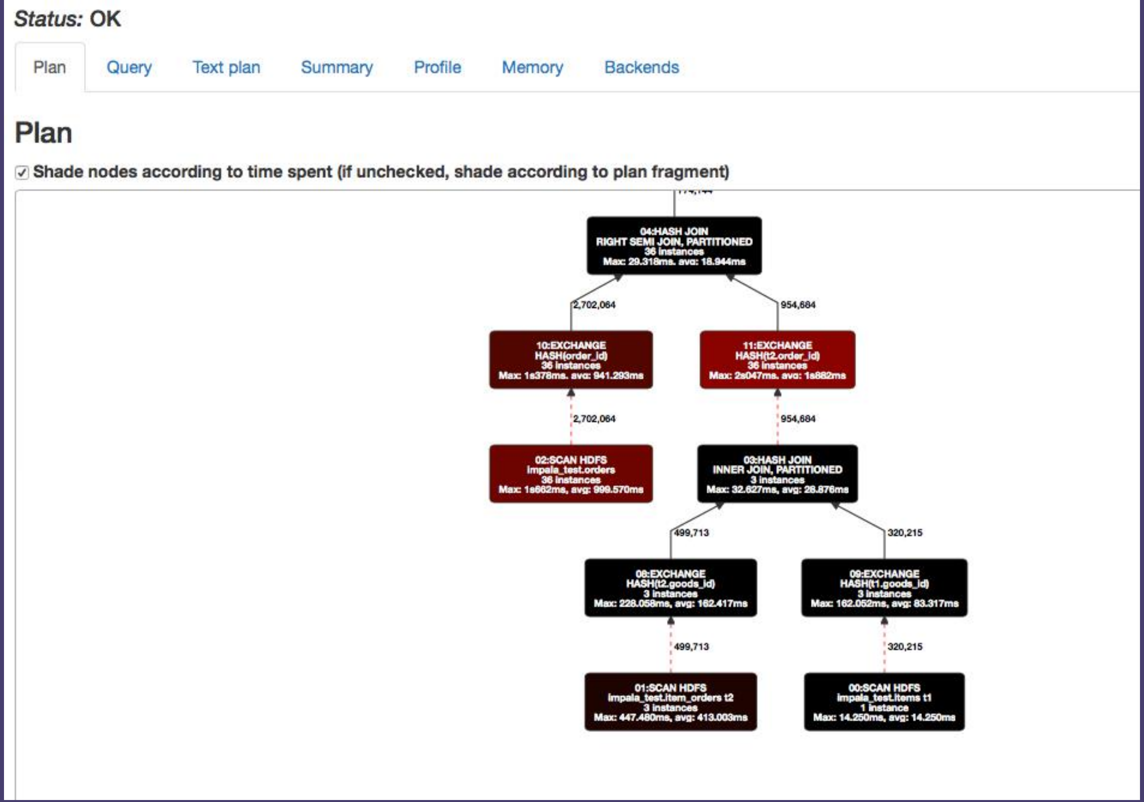
```
Process: Limit=100.00 GB Total=8.62 GB Peak=92.55 GB
  Buffer Pool: Free Buffers: Total=469.12 MB
  Buffer Pool: Clean Pages: Total=0
  Buffer Pool: Unused Reservation: Total=-1.75 GB
  Free Disk IO Buffers: Total=1.12 GB Peak=1.82 GB
  RequestPool=fe-eval-exprs: Total=0 Peak=4.00 KB
  RequestPool=default-pool: Total=5.39 GB Peak=85.12 GB
    Query(49415963c2eb0570:f2cfb3ae00000000): Reservation=0 ReservationLimit=80.00
    Query(a245945a93eb0308:c1498d6600000000): Reservation=0 ReservationLimit=80.00
    Query(f4406bdf7ae890f7:66fc610200000000): Reservation=0 ReservationLimit=80.00
    Query(3549750e9111ace4:33b49b9b00000000): Reservation=0 ReservationLimit=80.00
    Query(2f446bbd4706726a:8d6347e900000000): Reservation=0 ReservationLimit=80.00
```



1 queries in flight

User	Default Db	Statement	Query Type	Start Time	Duration
hadoop	tpch_parquet_1000_parquet	`-- start q10.sql select c_custkey,c_name, sum(l_extendedprice * (1 - l_discount)) as revenue, c_acctbal,n_name,c_address,c_phone,c_comment from customer,orders,lineitem,nation where c_custkey = o_custkey and l_orderkey = o_orderkey and o_orderdate >= '1993-10-01' and o_orderdate < '1994-01-01' and l_returnflag = 'R' and c_nationkey = n_nationkey group by c_custkey,c_name,c_acctbal,c_phone,n_name,c_address,c_comment order by revenue desc limit 100`	QUERY	2018-06-23 15:13:58.115437000	2s873ms

Impala 基本介绍



Status: OK

Plan Query Text plan Summary Profile Memory Backends

Plan

☑ Shade nodes according to time spent (if unchecked, shade according to plan fragment)

04:HASH JOIN
RIGHT SEMI JOIN, PARTITIONED
36 instances
Max: 29.318ms. avg: 18.944ms
2,702,064
954,684
10:EXCHANGE
HASH(order_id)
36 instances
Max: 1s378ms. avg: 941.293ms
11:EXCHANGE
HASH(t2.order_id)
36 instances
Max: 2s047ms. avg: 1s882ms
2,702,064
954,684
02:SCAN HDFS
impala_test.orders
36 instances
Max: 1s662ms, avg: 999.570ms
03:HASH JOIN
INNER JOIN, PARTITIONED
3 instances
Max: 32.627ms, avg: 28.876ms
499,713
320,215
08:EXCHANGE
HASH(t2.goods_id)
3 instances
Max: 228.058ms, avg: 162.417ms
09:EXCHANGE
HASH(t1.goods_id)
3 instances
Max: 162.052ms, avg: 83.317ms
499,713
320,215
01:SCAN HDFS
impala_test.item_orders t2
3 instances
Max: 447.480ms, avg: 413.003ms
00:SCAN HDFS
impala_test.items t1
1 instance
Max: 14.250ms, avg: 14.250ms

Impala 基本介绍

Impala缺陷

- Catalogd & statestored服务单点
- Web信息不持久化
- 资源隔离不精确
- 底层存储不能区分用户
- 负载均衡需要外部支持，如Haproxy
- 与Hive元数据同步都需要手动操作

PART 04

Impala 改进

网易大数据
Netease Big Data

- 基于ZK的loadbalance
- 管理服务器
- 细粒度权限 & 代理
- Json格式
- 兼容ranger权限管理
- 批量元数据刷新
- 元数据同步
- 元数据过滤
- 对接ElasticSearch查询

IMPROVE

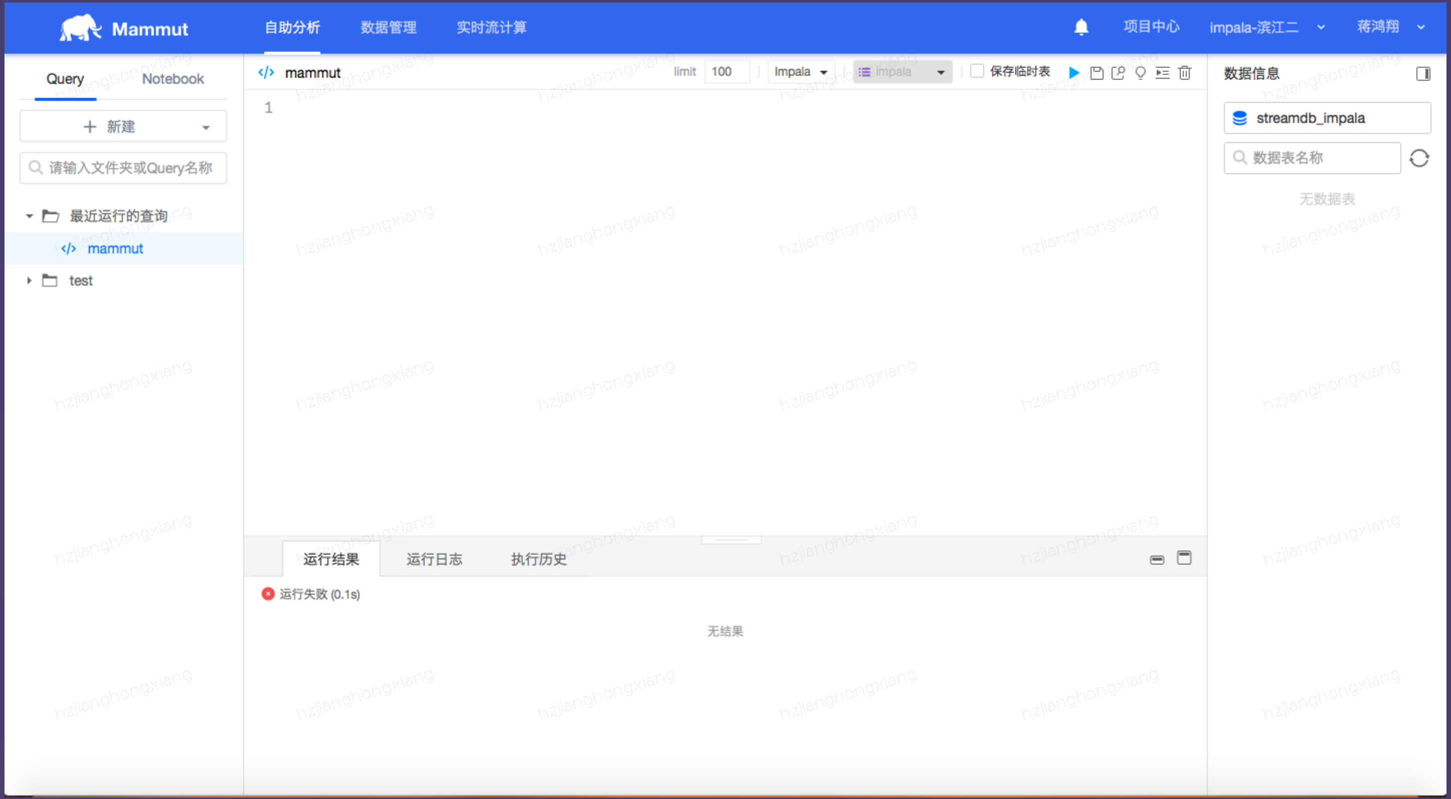
悟有所值
Ucan 下午茶
Ucan Afternoon Tea

PART 05

Impala 使用场景

网易大数据
Netease Big Data

Impala 使用场景

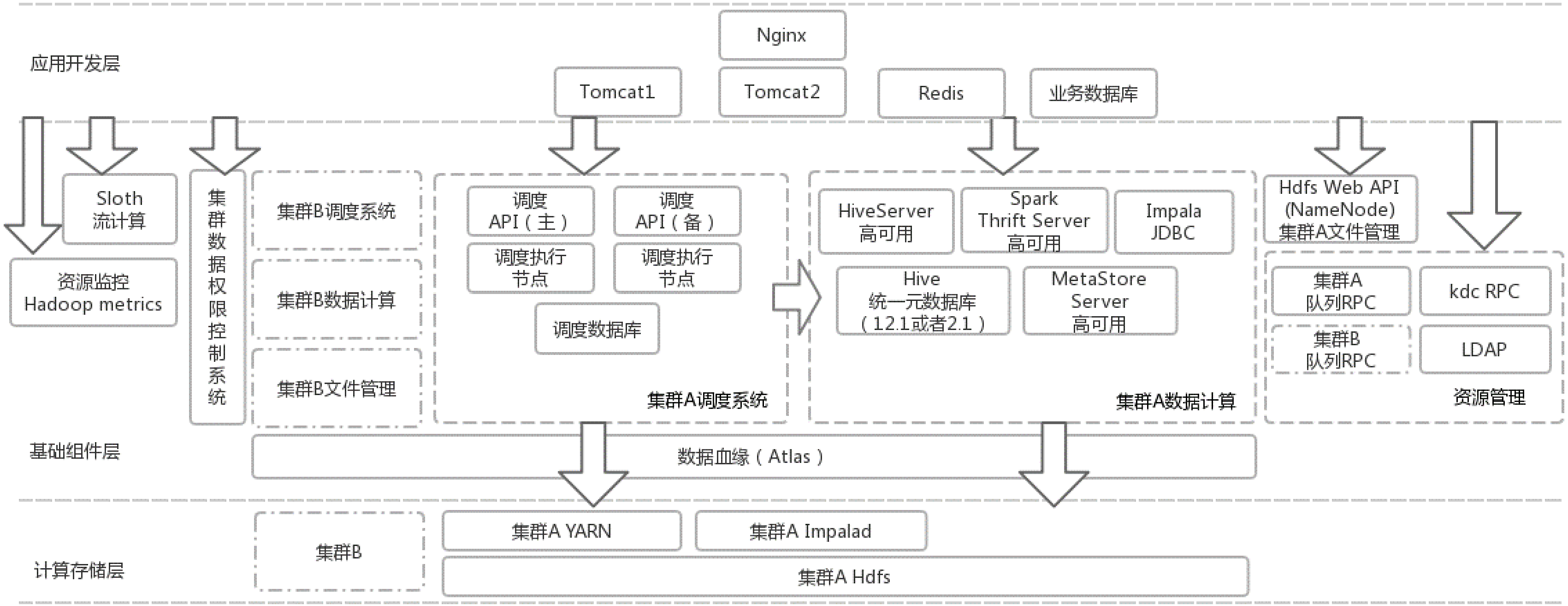

Mammut
自助分析
数据管理
实时流计算
项目中心
impala-滨江二
蒋鸿翔
Query
Notebook
+ 新建
请输入文件夹或Query名称
最近运行的查询
mammut
test
mammut
limit 100
Impala
impala
保存临时表
1
运行结果
运行日志
执行历史
数据信息
streamdb_impala
数据表名称
无数据表

运行失败 (0.1s)

无结果

hzjianghongxiang

Impala 使用场景



应用开发层

Nginx

Tomcat1

Tomcat2

Redis

业务数据库

Sloth
流计算

资源监控
Hadoop metrics

集群数据权限控制系统

集群B调度系统

集群B数据计算

集群B文件管理

调度
API（主）

调度
API（备）

调度执行
节点

调度执行
节点

调度数据库

集群A调度系统

HiveServer
高可用

Spark
Thrift Server
高可用

Impala
JDBC

Hive
统一元数据库
（12.1或者2.1）

MetaStore
Server
高可用

集群A数据计算

Hdfs Web API
(NameNode)
集群A文件管理

集群A
队列RPC

kdc RPC

集群B
队列RPC

LDAP

资源管理

基础组件层

数据血缘（Atlas）

计算存储层

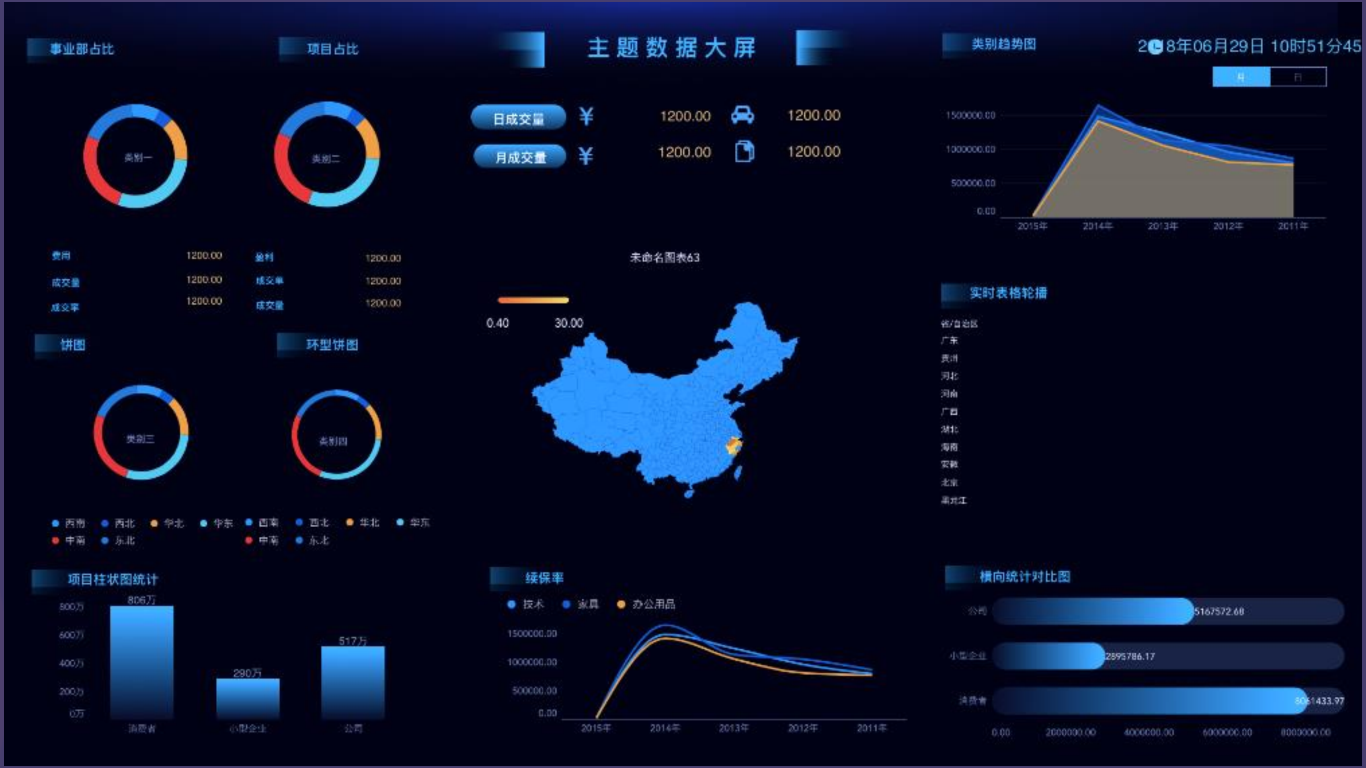
集群B

集群A YARN

集群A Impalad

集群A Hdfs

Impala 使用场景


主题数据大屏
事业部占比
项目占比
类别一
类别二
费用 1200.00
成交量 1200.00
成交率 1200.00
盈利 1200.00
成交单 1200.00
成交量 1200.00
日成交量 ¥ 1200.00 1200.00
月成交量 ¥ 1200.00 1200.00
饼图
环型饼图
类别三
类别四
西南
西北
华北
华东
中南
东北
西南
西北
华北
华东
中南
东北
未命名图表63
0.40
30.00
类别趋势图
2018年06月29日 10时51分45
年
月
1500000.00
1000000.00
500000.00
0.00
2015年
2014年
2013年
2012年
2011年
实时表格轮播
省/自治区
广东
贵州
河北
河南
广西
湖北
海南
安徽
北京
黑龙江
项目柱状图统计
800万
600万
400万
200万
0万
806万
290万
517万
消费者
小型企业
公司
续保率
技术
家具
办公用品
1500000.00
1000000.00
500000.00
0.00
2015年
2014年
2013年
2012年
2011年
横向统计对比图
公司
5167572.68
小型企业
2895786.17
消费者
8061433.97
0.00
2000000.00
4000000.00
6000000.00
8000000.00

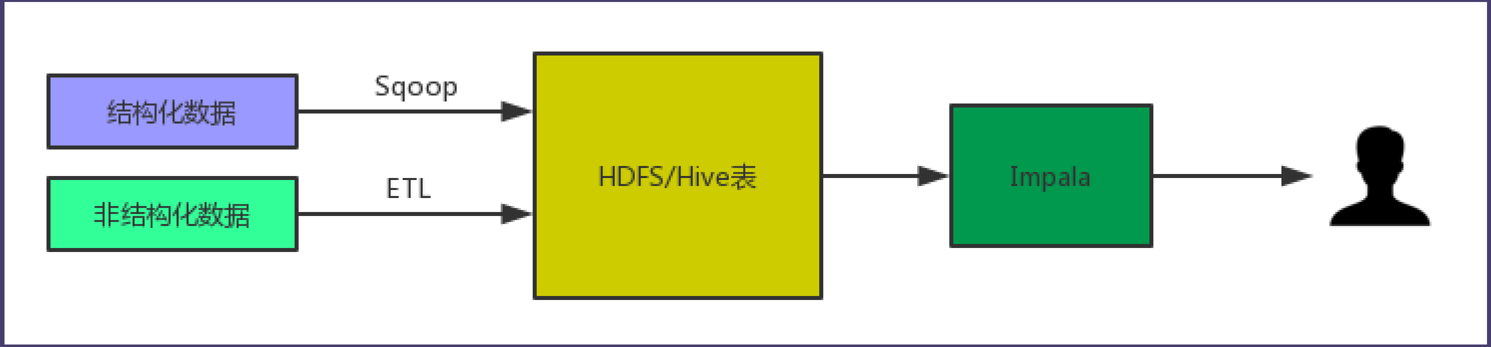
Ucan 下午茶 悟有所值 Ucan Afternoon Tea

PART 06

场景方案解析

网易大数据
Netease Big Data

场景方案解析

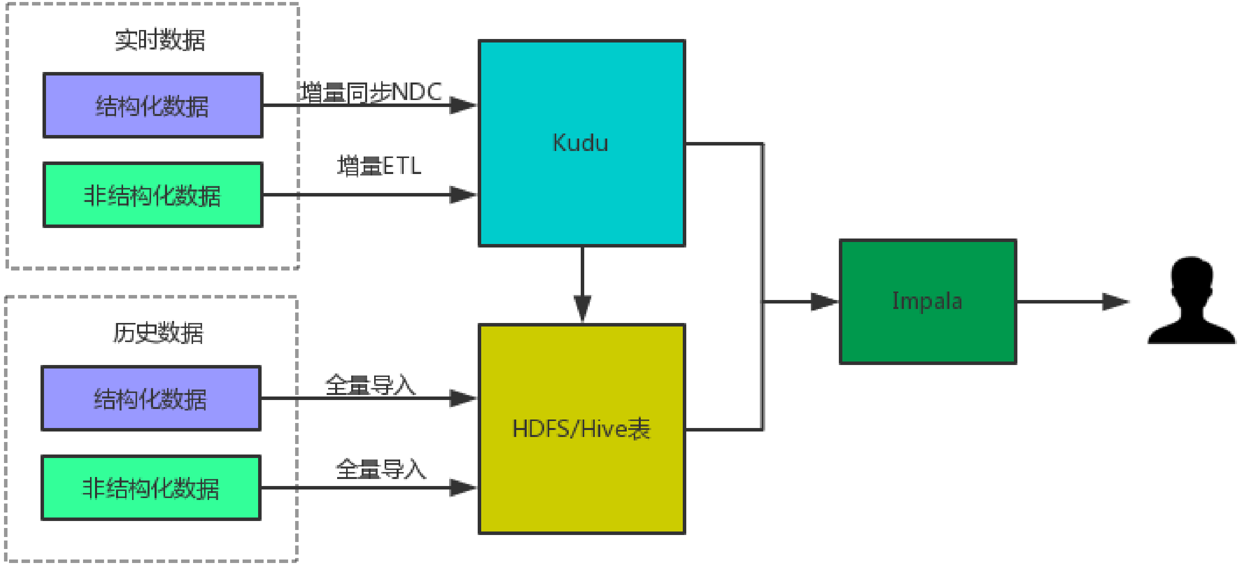

结构化数据
非结构化数据
Sqoop
ETL
HDFS/Hive表
Impala

需要考虑的问题：

1. 元数据同步问题，ETL写入数据Impala无感知，依赖元数据同步
2. 数据实时性问题，避免大量小文件导致NN不稳定，每次写文件的batch不能太小

场景方案解析

利用Kudu解决小文件问题


实时数据
结构化数据
非结构化数据
增量同步NDC
增量ETL
Kudu
历史数据
结构化数据
非结构化数据
全量导入
全量导入
HDFS/Hive表
Impala

UCLOUD

Ucan 下午茶
悟有所值
Ucan Afternoon Tea

关注“UCloud技术公告牌”，更多分享与交流