流式数据处理在百度数据工厂应用与实践

李俊卿

百度高级研发工程师
数据工厂流式数据处理负责人

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE
全球技术领导力峰会

Geekbang. 极客邦科技 | TGO 鲲鹏會

500+高端科技领导者与你一起探讨
技术、管理与商业那些事儿

2019年6月14–15日 | 上海圣诺亚皇冠假日酒店

扫码了解更多信

自我介绍

李俊卿，百度高级研发工程师，数据工厂流式数据处理负责人

经历百度大数据离线批处理从Hive到Spark1.x到Spark2.x技术方案的架构升级；

设计并研发百度数据工厂流式数据处理核心部分；

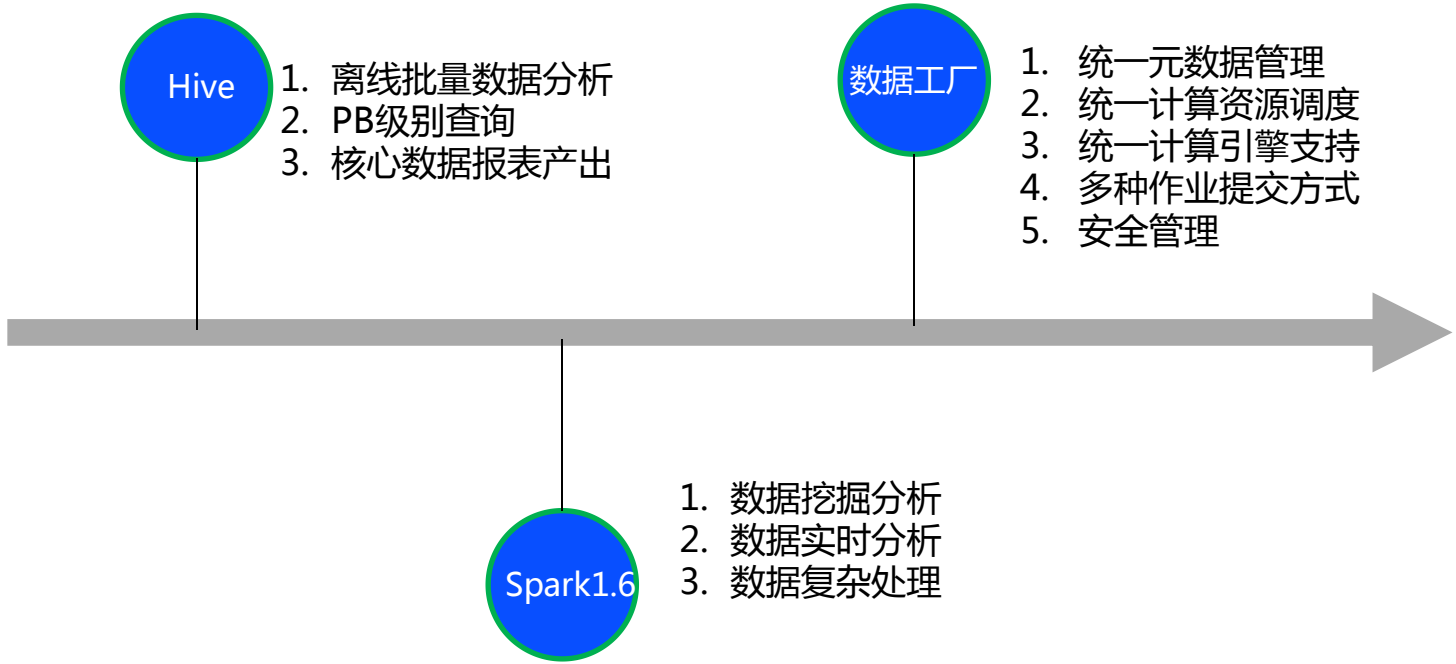
研发基于Spark的流批统一SQL引擎

CONTENT 目录

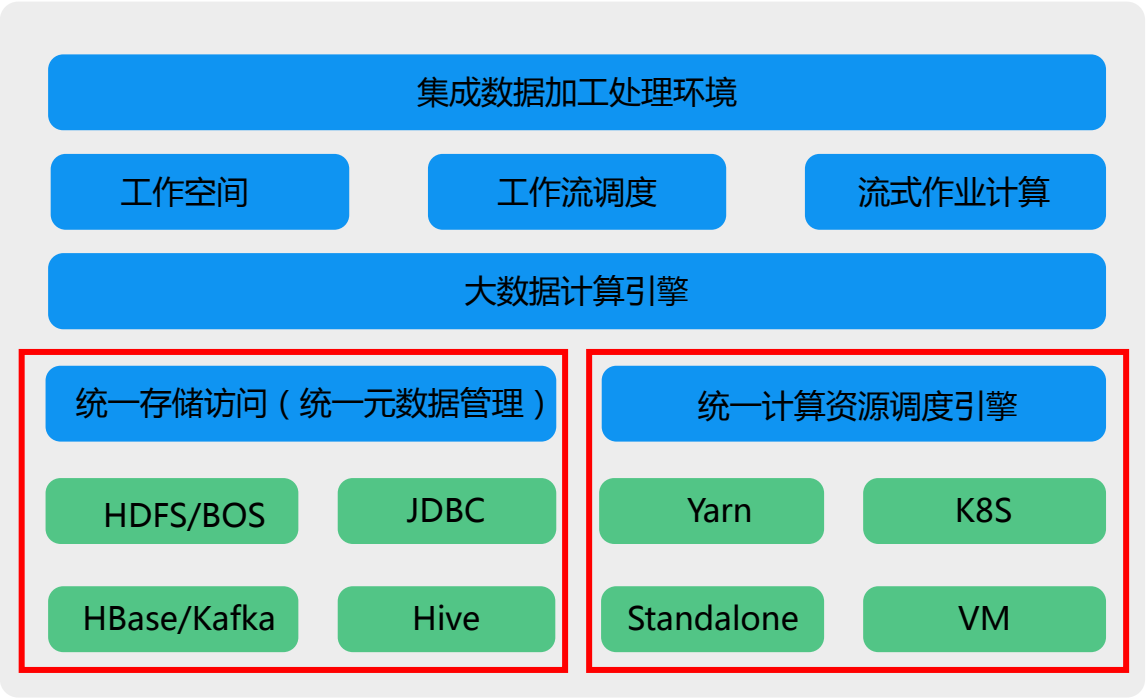
PART A

百度数据工厂介绍

百度数据工厂的发展史


Hive
1. 离线批量数据分析
2. PB级别查询
3. 核心数据报表产出
数据工厂
1. 统一元数据管理
2. 统一计算资源调度
3. 统一计算引擎支持
4. 多种作业提交方式
5. 安全管理
Spark1.6
1. 数据挖掘分析
2. 数据实时分析
3. 数据复杂处理

百度数据工厂整体介绍



百度内部

日例行任务数：10W+

日传输量：8PB+

应用业务数：200+

ToB

瑞声、大地等多家大型企业

PART B

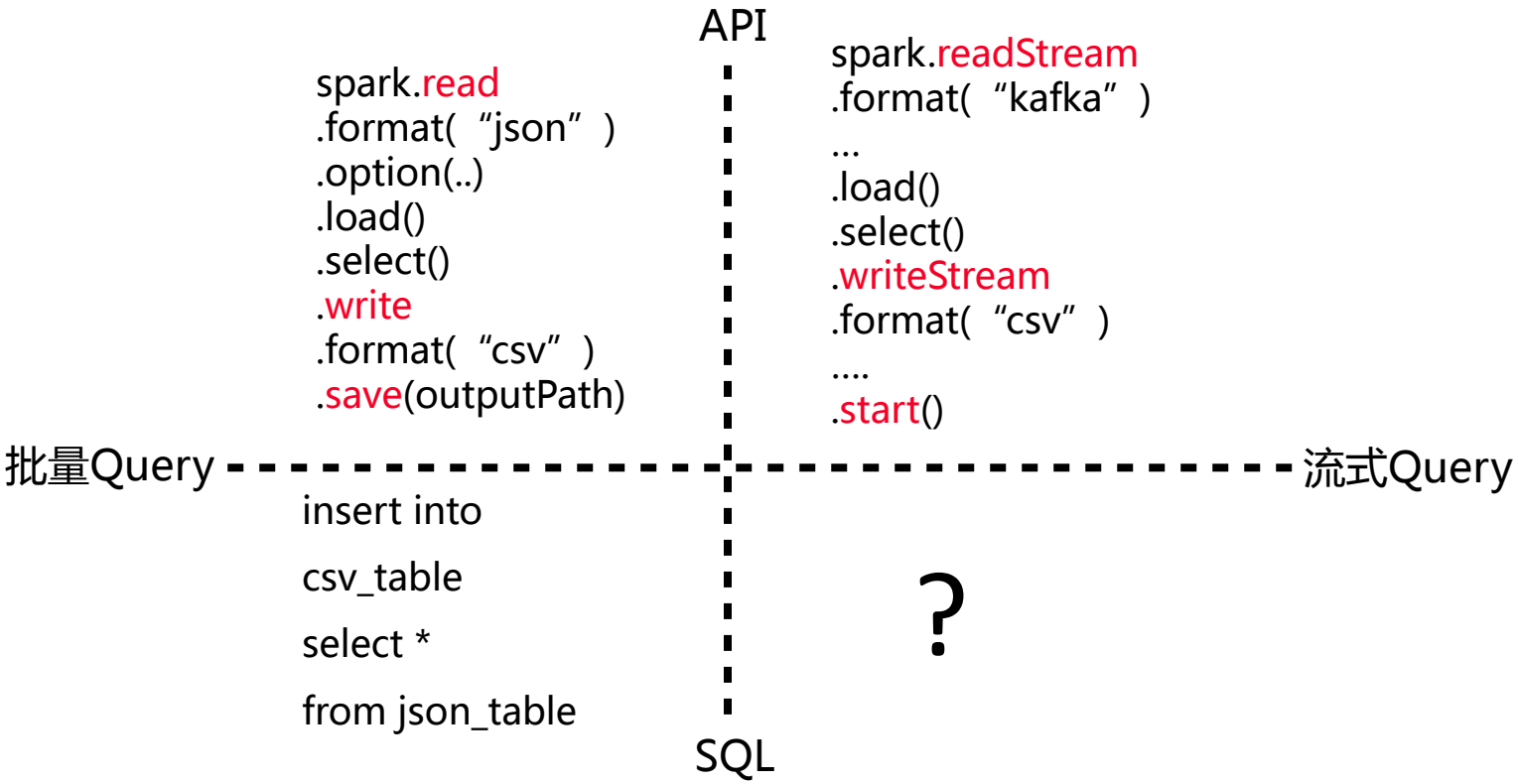
流式数据处理在百度数据工厂应用

Spark流式SQL问题
实时转离线问题
实时转大屏展示问题

PART B

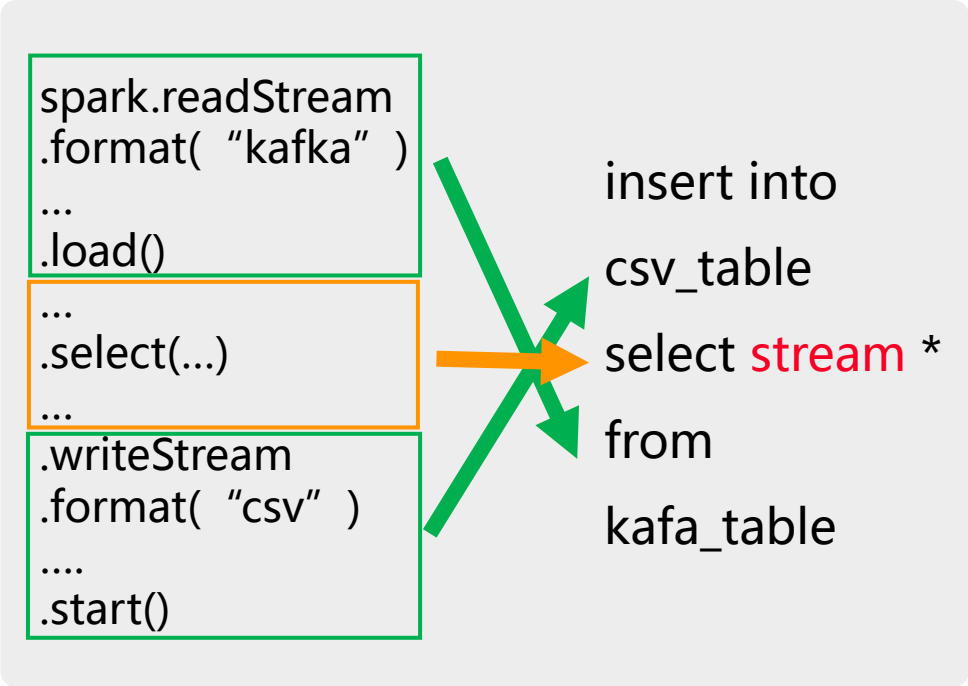
Spark流式SQL问题

B Spark未提供流式SQL



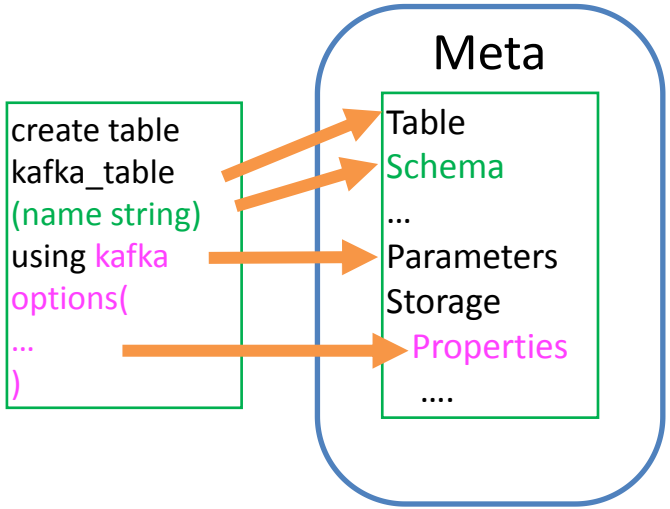
B 从流式API到流式SQL语法的定义

1. 数据源映射
    1. 读取数据源 -> 读取表
    2. 写入数据源 -> 写入表
2. 数据处理映射
    1. 数据处理 -> SELECT/JOIN/UNION等
3. 增加Stream关键字
    1. 区分批量SQL和流式SQL



B 流式SQL的统一元数据存储

以Properties存储数据源配置



保证表的通用性

1. 一张表对应于一个数据源
    1. 多个数据源会影响通用性
2. 只能定义通用配置
    1. 例如：watermark配置并不是通用配置，不能定义在表内
3. 允许流批2种方式读取
    1. 当带stream关键字时，表被翻译为流式读取；反之则翻译为批量读取

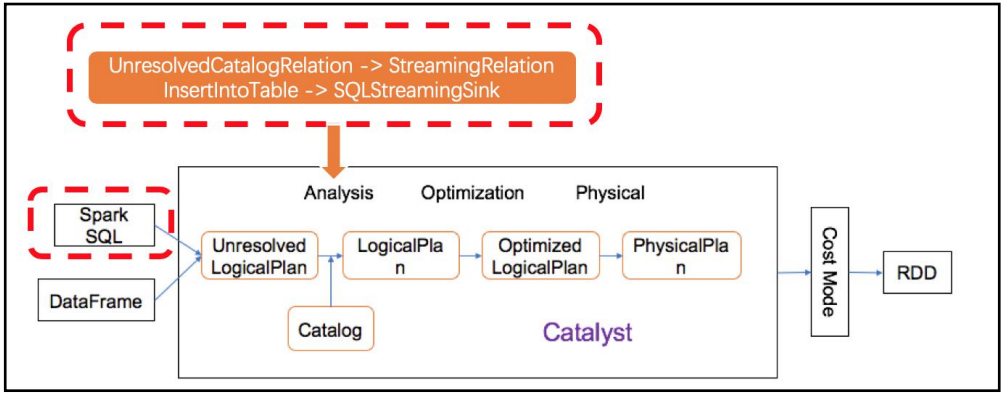
B 流式SQL语义层解析

1.更新语义解析规则

1. FindDataSource
    1. 完成流式表解析
    2. UnresolvedCatalogRelation -> StreamingRelation
    3. InsertIntoTable -> SQLStreamingSink

2.增加可执行类

1. SQLStreamingSink
    1. 负责触发流式任务执行
    2. 解析流式配置



insert into hive_table → SQLStreamingSink

select stream * → +- Project [*]

from kafa_table → +- StreamingRelation()

B Stream Join Batch语义解析问题

典型场景：

实时统计工厂工人生产速度，对Kafka(实时生产数据)和MySQL(员工数据)做join分析，从而获得每段时间内每个工人的产量。

按照目前的设计，在语义解析过程中，mysql_table会解析成StreamingRelation，无法完成正常的Streaming Join Batch。

```
insert into csv_table
select stream *
from kafka_table t1
join mysql_table t2
on t1.id = t2.p_id
```

```
SQLStreamingSink
+- Project [*]
   +- Join (t1.id,t2.p_id)
      +- Alias t1
         +- StreamingRelation
      +- Alias t2
         +- StreamingRelation
```

B 流式SQL的设计升级

新增流式表类型

只有流式表才能解析成StreamingRelation

流式表保留了流批查询的语义

创建流式表

在Option中添加Streaming标识

```
create table kafka_table(name string)
using kafka
options(
isStreaming=true,
...)
```

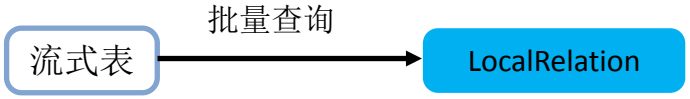
可选方案

1. 根据Source类型区分
   1. 例如：Kafka是流式Source， MySQL是批量Source
   2. pass原因：大部分Source即可以当流式也可以当批量
2. 建表时指定起止offset
   1. 例如：MySQL指定读取截止的数量等
   2. pass原因：这和我们对表的定义相违背，我们要做的是通用表，用户创建表后大家都可以用。

B 流式表的使用

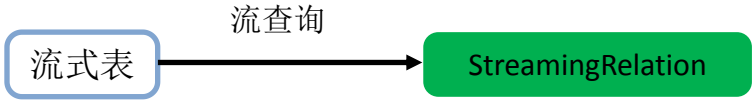
1. 批量查询流式表

1. 例如：select * from kafka_stream_table
2. 在此场景中，流式表会被当做普通表对待
3. 在语义解析层kafka_stream_table将被解析为KafkaRelation



2. 流式查询流式表

1. 例如：select stream * from kafka_stream_table
2. 在此场景中，流式表将转换为StreamingRelation



B 升级后的Stream Join Batch解析

升级：kafka_table创建为流式表，mysql_table创建为普通表

```
insert into csv_table
select stream *
from kafka_table t1
join mysql_table t2
on t1.id = t2.p_id
```

```
SQLStreamingSink
+- Join (t1.id,t2.p_id)
   +- Project [*]
      +- StreamingRelation
   +- Project[*]
      +- LocalRelation
```

B 流批统一SQL问题的总结

- **统一的表元数据存储**
- **流式SQL解析**
  - **升级FindDataSource**
    - insertIntoTable -> SQLStreamingSink
    - from Table -> StreamingRelation
  - **Streaming Join Batch处理**
    - 使用Stream/Normal Table区分表的使用

PART B

实时转离线问题

B 目前实时转离线存在的问题

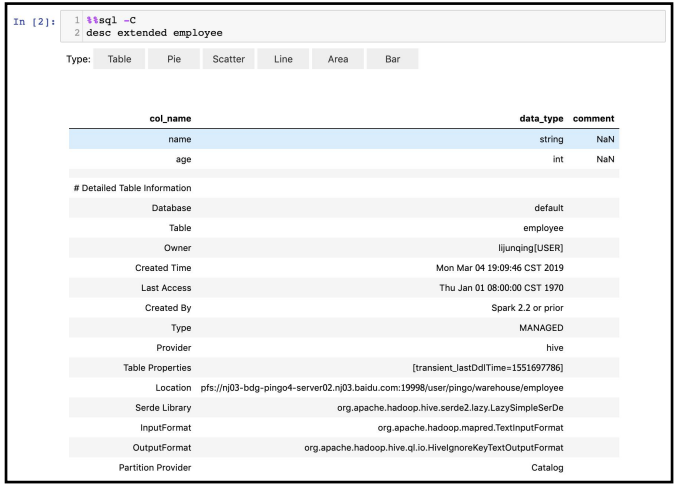
1. 输出信息缺乏管理
    1. 输出列信息？列分隔符？已输出分区有哪些？
2. 迁移升级代价较大
    1. 集群迁移(从测试集群迁移到正式集群)时，需要修改代码才能正常运行
3. 扩展文件格式繁杂
    1. 如果甲方要求输出sequenceFile，并指定了输出格式怎么办？

实时转离线样例Case

```
df.writeStream
.format(“csv”)
.partitionBy(“col1”, “col2”)
.option("path", “hdfs://hdfsPath“)
.start()
```

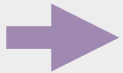
百度数据工厂实时转数仓方案

1. 输出信息管理规范
    1. 依托Hive元数据管理
2. 输出升级代价小
    1. 一般情况下修改Hive即可完成升级
3. 扩展文件格式简单
    1. 小批量写Hive，特殊格式添加Jar包即可



col_name	data_type	comment
name	string	NaN
age	int	NaN
# Detailed Table Information		
Database	default	
Table	employee	
Owner	lijunqing[USER]	
Created Time	Mon Mar 04 19:09:46 CST 2019	
Last Access	Thu Jan 01 08:00:00 CST 1970	
Created By	Spark 2.2 or prior	
Type	MANAGED	
Provider	hive	
Table Properties	[transient_lastDdlTime=1551697786]	
Location	pfs://nj03-bdg-pingo4-server02.nj03.baidu.com:19998/user/pingo/warehouse/employee	
Serde Library	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	
InputFormat	org.apache.hadoop.mapred.TextInputFormat	
OutputFormat	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat	
Partition Provider	Catalog	

```
df.writeStream
.format(“csv”)
.partitionBy(“col1”, “col2”)
.option("path", “hdfs://hdfsPath“)
.start()
```

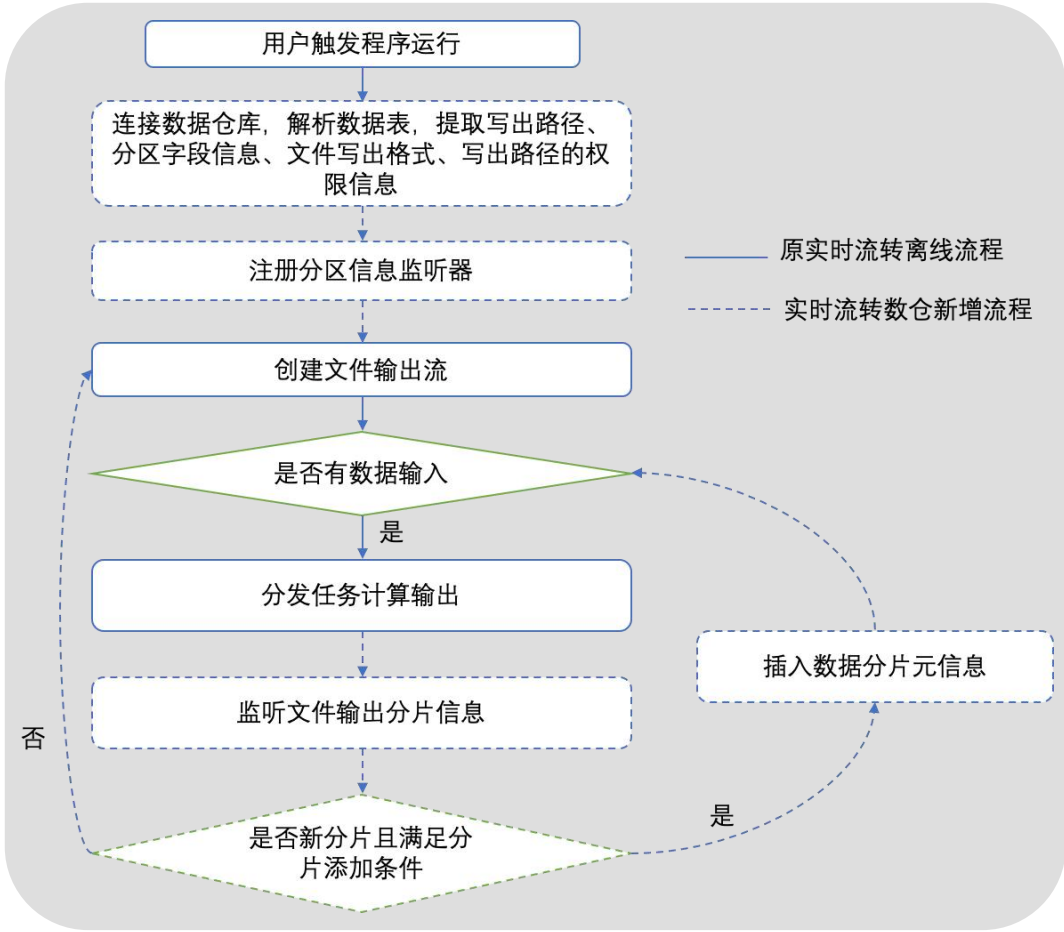


```
df.writeStream
.format("hive")
.option("table", "employee")
.option("database", "default")
.start()
```

B 实时转数仓的原理介绍

具体的解决方案：

1. 以实时转离线为基础
2. 在解析生成FileSink时
    1. 读取输出表，将配置信息、HiveDynamicPartitionSinker注入到FileSink中
3. 在小批量数据产出后
    1. 提取输出目录的分区信息，并将分区信息提供给HiveDynamicPartitionSinker，由HiveDynamicPartitionSinker添加Partition



PART B

实时转大屏展示问题

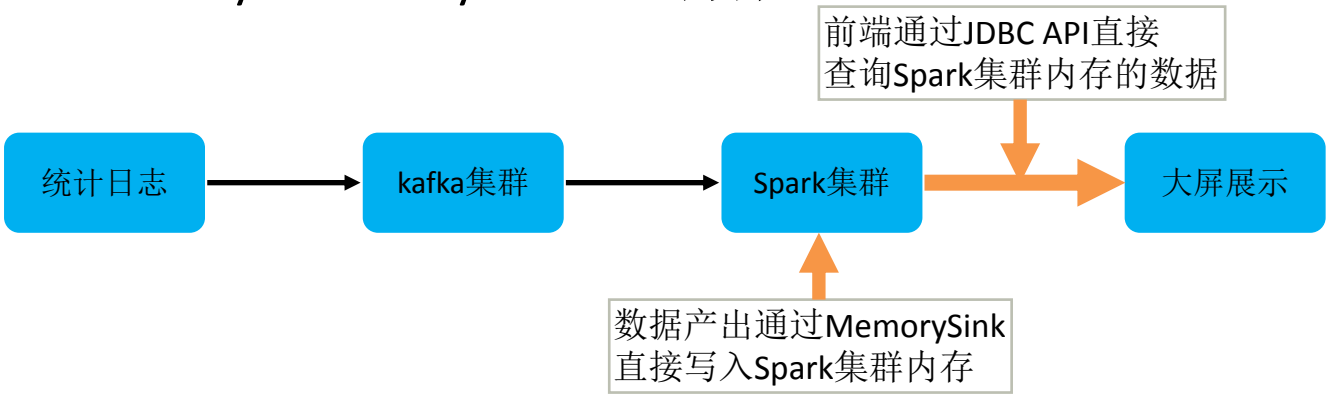
B 实时流输出大屏存在的问题

通用的实时流转OLAP/JDBC接大屏展示方案

1. 需额外部署其他组件
    1. 输出到OLAP，需要部署OLAP服务甚至需要kafka集群
2. 从计算产出到最终展示经多个系统，系统间延迟影响大
    1. 实时流写入OLAP的延迟
3. 产出数据中间落磁盘
    1. 一般OLAP是基于磁盘的，读写OLAP时会产生I/O操作

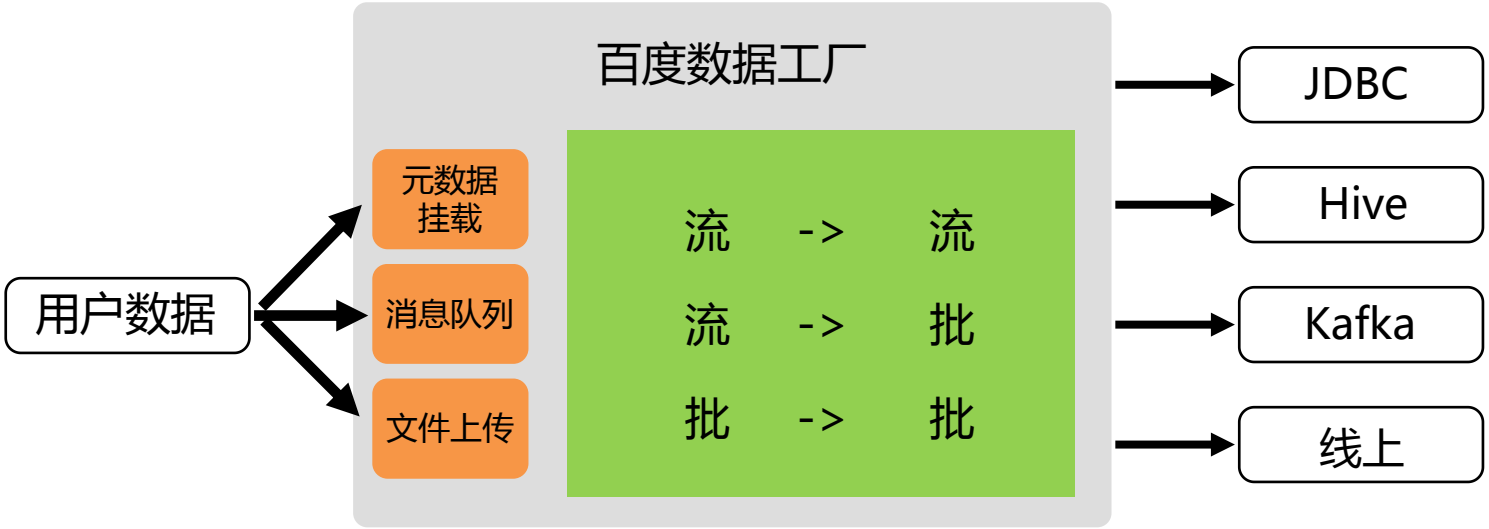
百度数据工厂实时流直接对接大屏方案

Spark MemorySink + Livy JDBC API方案



1. 通过Livy启动交互式分析任务
2. 数据产出到内存，无需落盘
3. 大屏通过JDBC接口直接获取内存数据

B 百度数据工厂的统一计算引擎



PART C

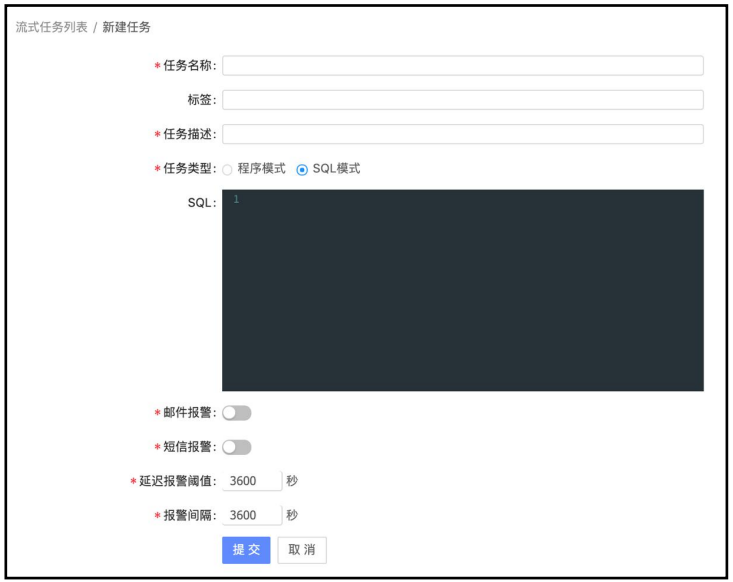
流式数据处理在百度数据工厂的实践

流式产品化页面

流式数据处理在广告物料分析的实践

百度数据工厂的统一计算引擎

Streaming SQL提交页面



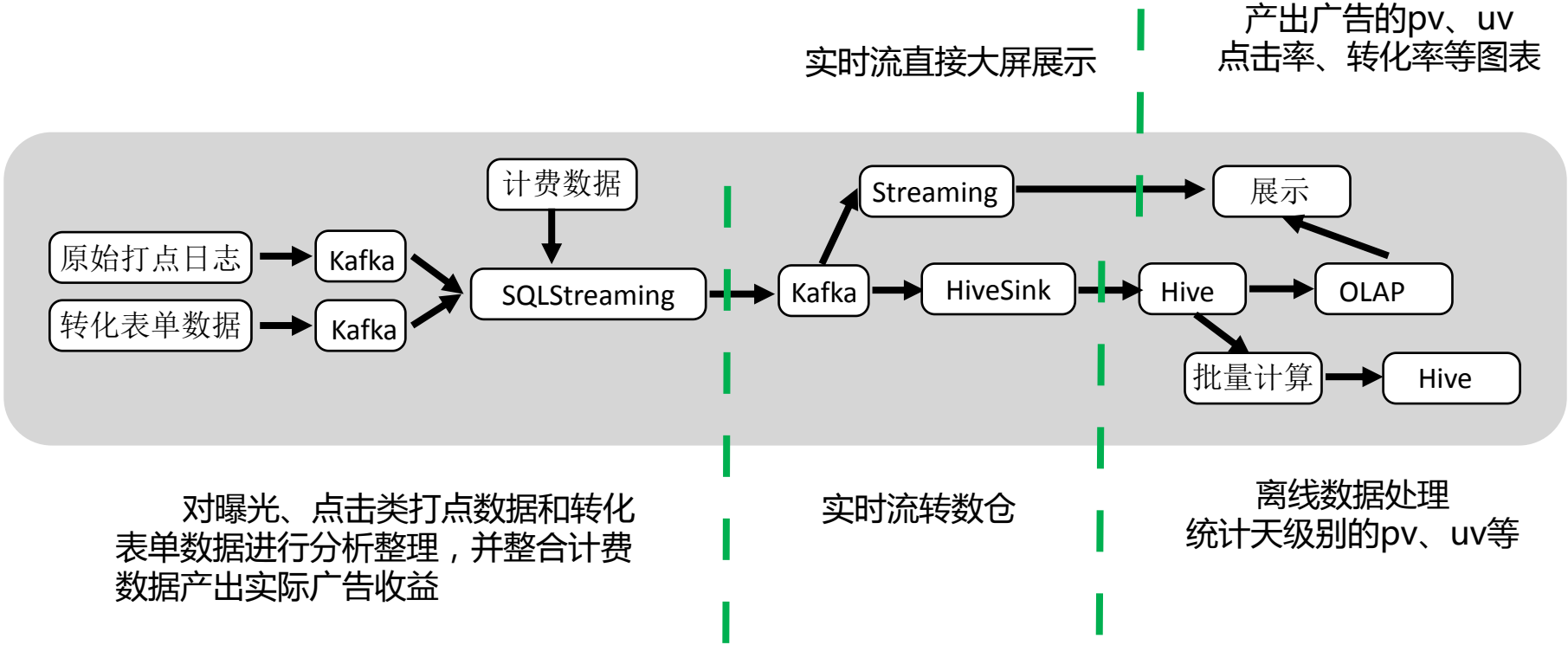
Streaming实时监控页面

案例：

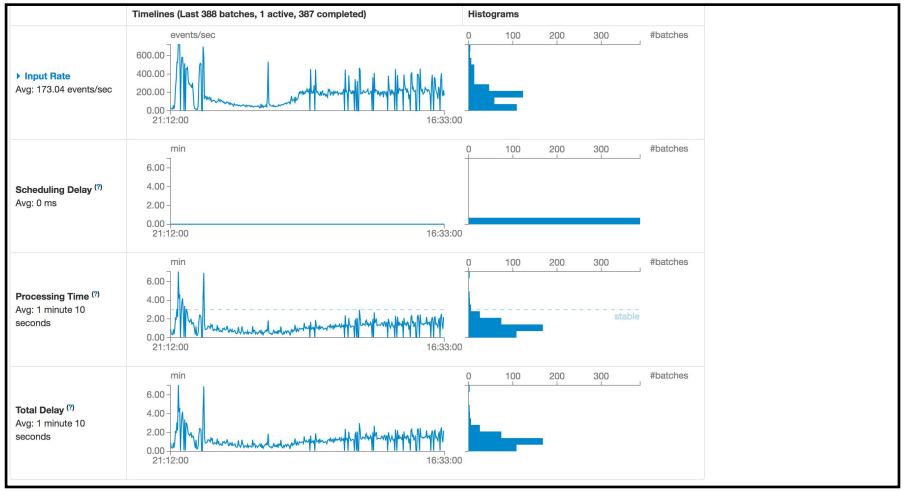
在实际的产品运维过程中，存在广告主投放广告的场景。广告主投放广告付钱是要看真实的点击率、曝光率和转化率的，而且很多广告主是根据曝光量、点击量、转化量来付钱的。

这种情况下，我们就需要专门针对广告物料进行分析，根据点击、曝光日志和转化数据生成广告的pv、uv、点击率和转化率，并根据计费数据生成广告收益。

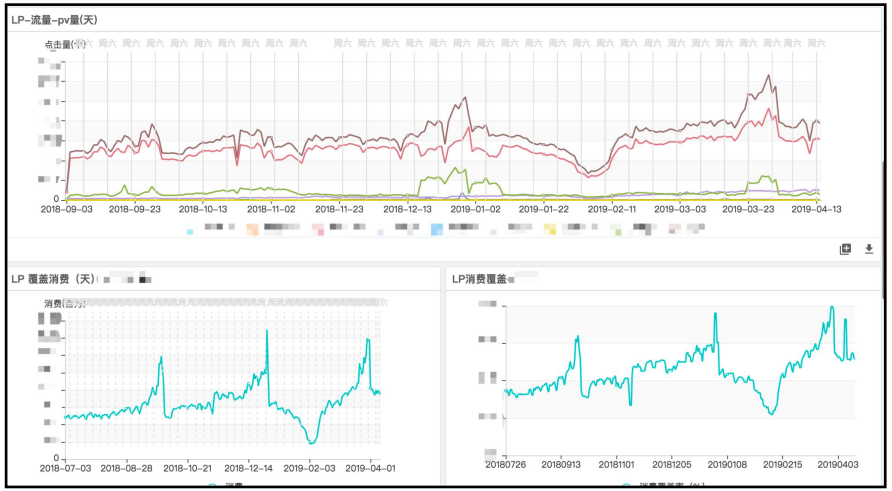
广告物料分析实践案例


实时流直接大屏展示
产出广告的pv、uv
点击率、转化率等图表
计费数据
原始打点日志
转化表单数据
Kafka
Kafka
SQLStreaming
Kafka
Streaming
展示
HiveSink
Hive
OLAP
批量计算
Hive
对曝光、点击类打点数据和转化表单数据进行分析整理，并整合计费数据产出实际广告收益
实时流转数仓
离线数据处理
统计天级别的pv、uv等

广告物料分析实践案例成果



实时流转数仓延迟图



广告物料报表实时展示

上: 每个投放渠道的pv统计

左下: 信息流渠道的消费情况

右下: 信息流渠道的消费占比

PART D

总结

总结

- **百度数据工厂介绍**
  - 大数据分析的一站式处理平台
  - https://cloud.baidu.com/product/pingo.html
- **流式数据处理在百度数据工厂的应用**
  - Spark统一SQL引擎的支持
  - 百度数据工厂实时转数仓方案
  - 百度数据工厂实时流直接对接大屏方案
- **流式数据处理在百度数据工厂的实践**
  - 介绍了流式产品以及典型使用案例

下一步

- **更强大的流式SQL引擎**
  - 基于Spark3.0 DataSource v2设计
- **更丰富的流式运维和监控**
  - 实时监控更多数据
- **更强大的流计算引擎**
  - Continue Processing

极客时间 | 企业服务

这是一封正经“福利”贴

想做团队的领跑者 需要迈过这些“槛”

成长型企业，易忽视人才体系化培养 企业转型加快，团队能力又跟不上	VS	从基础到进阶，超100+一线实战 技术专家带你系统化学习成长
团队成员技能水平不一， 难以一“敌”百人需求	VS	解决从小白到资深技术人所遇到 80%的问题
寻求外部培训，奈何价更高且 集中式学习	VS	多样、灵活的学习方式，包括 音频、图文 和视频
学习效果难以统计，产生不良循环	VS	获取员工学习报告，查看学习 进度，形成闭环

课程顾问「橘子」

回复「QCon」
免费获取
学习解决方案

 # 极客时间企业账号 # 解决技术人成长路上的学习问题

THANK YOU