Apache Flink

百亿级实时数据入湖实战

陈俊杰/腾讯数据湖研发高级工程师 2021-4-17

Apache Flink x Iceberg Meetup · 上海站

CONTENT

目录 >>

#1 腾讯数据湖介绍

更多平台对接

EMR
腾讯云弹性计算平台

DataHub
流批一体数据接入和治理工具

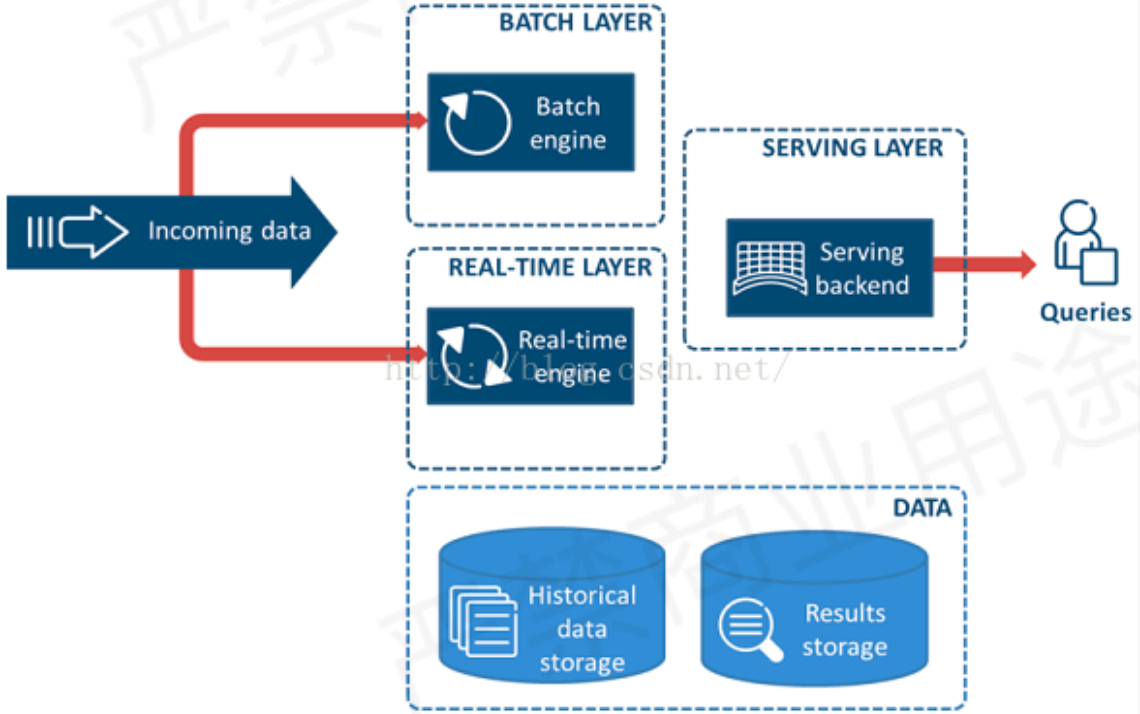
Oceanus
一站式流计算平台

TDBANK
内部数据接入平台

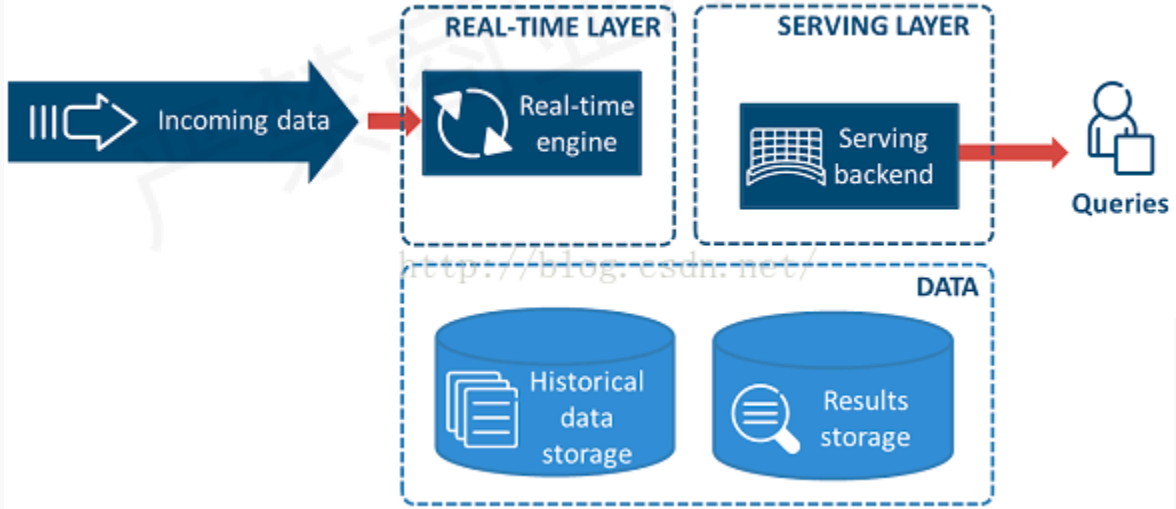
#2 百亿级数据落地实战

传统平台架构

Lambda架构



Kappa架构



Apache Flink

场景一：手Q安全数据入湖

小文件挑战

小文件爆炸

几天时间元数据和数据的小文件同时爆炸，集群压力巨大

延迟要求高

checkpoint 间隔短，commit间隔小，放大小文件问题

合并小文件又放大问题

为了解决小文件问题，开Action进行小文件合并，结果产生更多文件。

Flink Writer产生小文件

Flink写入没有shuffle,分发的数据无序，导致小文件多。

来不及删数据

删除快照，删孤儿文件，但是扫描文件太多，namenode压力巨大。。

解决方案

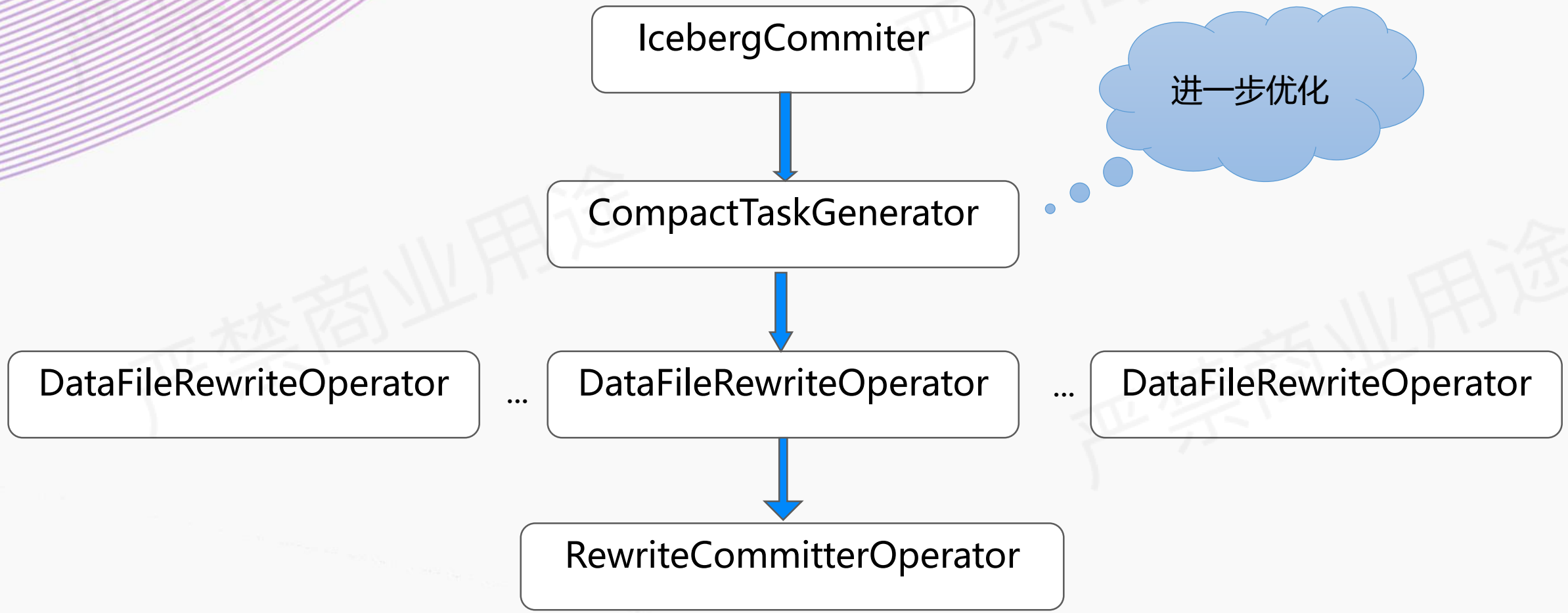
Flink同步合并

- 增加小文件合并Operators
- 增加Snapshot自动清理机制
  snapshot.retain-last.nums
  snapshot.retain-last.minutes

Spark异步合并

- 增加后台服务进行小文件合并和孤儿文件删除。
- 增加小文件过滤逻辑，逐步删除小文件。
- 增加按分区合并逻辑，避免一次生成太多删除文件导致任务OOM。

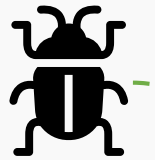
Flink同步合并



Fanout Writer的坑

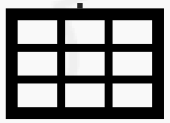
Flink写入无shuffle

为了低延迟写入，没有hash处理，数据均匀分布。



OOM

Fanout writer打开太多writer, OOM



多层分区

数据量大，用户除了定义时间分区，还会定义bucket分区

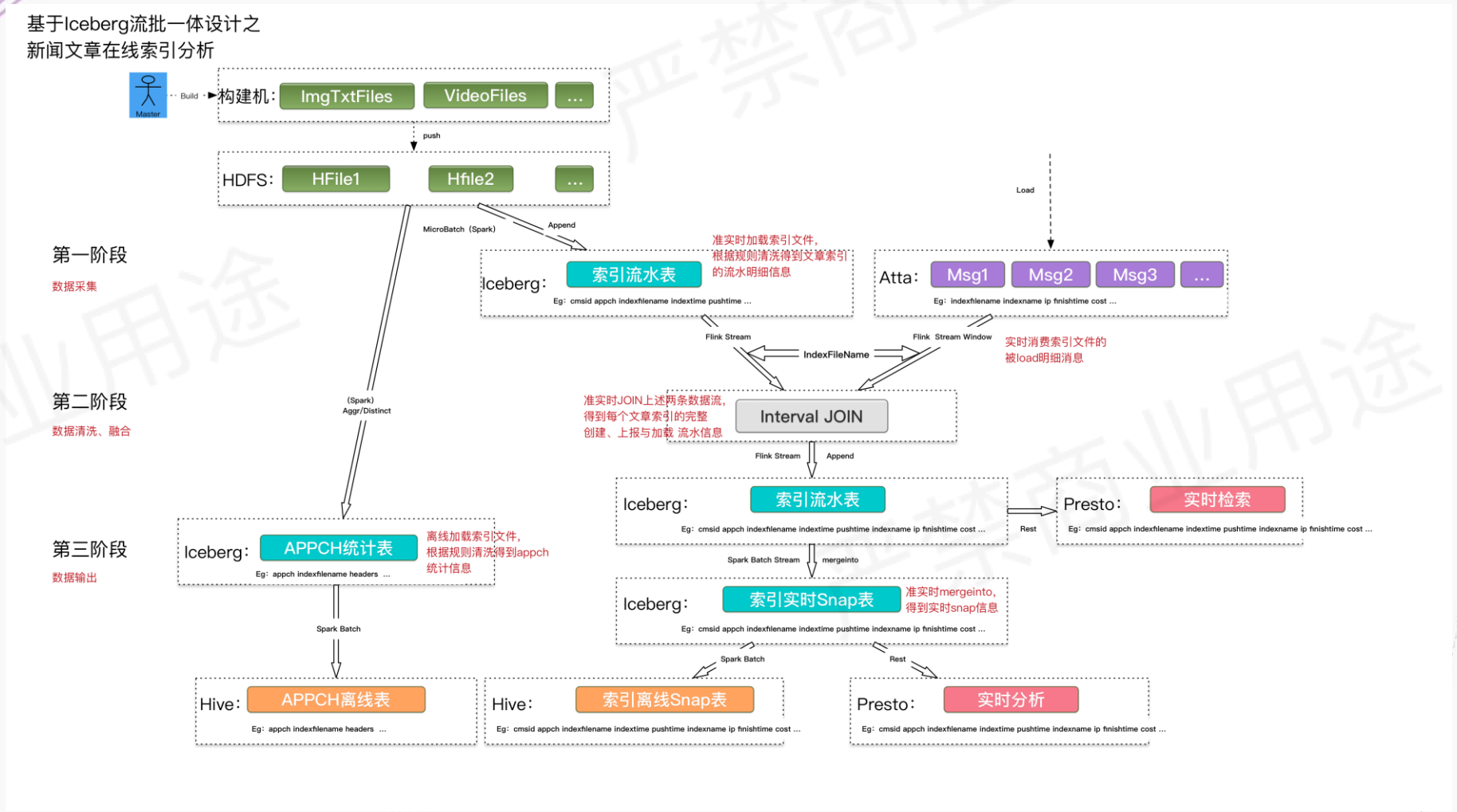
KeyBy 支持

根据用户定义的分区方式进行不同KeyBy策略

LRU Writer

场景二：新闻平台索引分析

- 准实时明细层
- 实时流式消费
- 流式MERGE INTO
- 多维分析
- 离线分析



场景特点

数量级

索引单表超千亿，单batch 2000万，日均千亿

数据源

全量、准实时增量、消息流

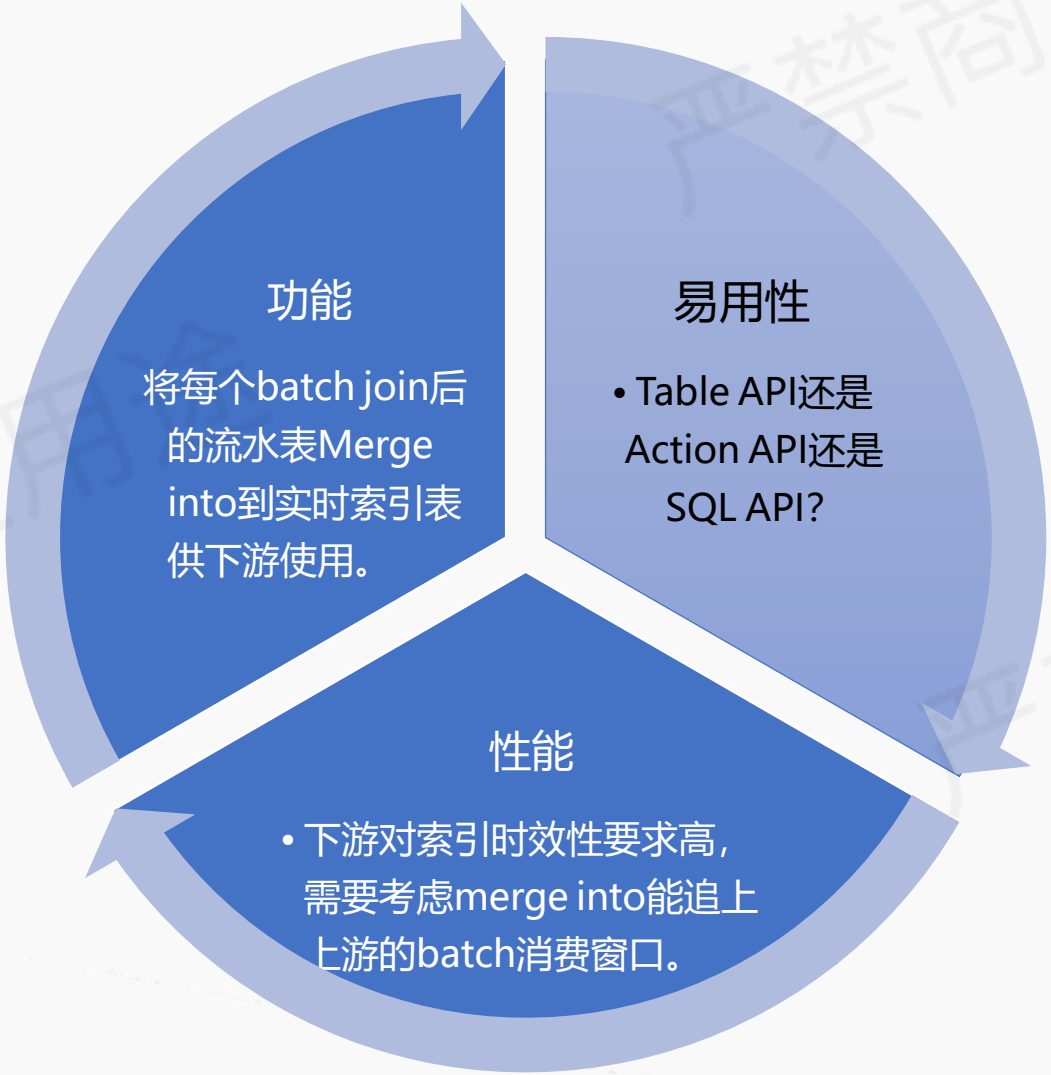
时延需求

端到端数据可见性分钟级

消费方式

流式消费，批加载，点查，行更新，多维分析

挑战：MERGE INTO



解决方案

#1

- 参考Delta Lake 设计 JoinRowProcessor.
- 利用Iceberg的WAP机制写临时快照

#2

- 可选cardinality-check
- 写入可选 sort

#3

- 支持 Dataframe API
- Spark 2.4支持SQL
- Spark 3.0使用社区版本

场景三：广告数据分析

数量级

日均千亿PB数据，单条2K

数据源

Spark Streaming 增量入湖

数据特点

标签不停增加，schema不停变换

使用方式

交互式查询分析

挑战

交互式查询

30 天数据基本集群撑爆

Schema嵌套复杂，平铺后近万列，一写就 OOM

解决方案

#1

- 默认每个Parquet page size 设置为1M。需要根据executor内存进行page size设置。

#2

- 提供Action进行生命周期管理。
- 文档！区分生命周期和数据生命周期。

#3

- column projection
- predicate push down

#3 未来规划

内核侧

更多的数据接入

- 增量入湖支持
- V2 Format支持
- Row Identity 支持

更快的查询

- 索引支持
- Alloxio加速层支持
- MOR 优化

更好的数据治理

- 数据治理Action
- SQL Extension支持
- 更好的元数据管理

平台侧

- 元数据清理服务化
- 数据治理服务化

- Spark 消费CDC入湖
- Flink 消费CDC入湖

- 写入数据指标
- 小文件监控和告警

#4 总结

可用性

通过多个业务线的实战，确认Iceberg经的起日均百亿，甚至千亿的考验。

易用性

使用门槛比较高，需要做更多的工作才能让用户使用起来。

场景支持

目前支持的入湖场景还没有Hudi多，增量读取这块也比较缺失，需要大家努力补齐。

Apache Flink

Thanks

Apache Flink x Iceberg Meetup · 北京站