FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

SmartNews 基于Flink的Iceberg实时数据湖实践

戢清雨 | 数据平台架构师，Apache Iceberg Contributor

FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

01 SmartNews数据湖介绍

02 基于Iceberg v1格式的数据湖实践

03 基于Flink实时更新的数据湖（Iceberg v2）解决方案

04 实时更新小文件问题的优化

05 总结与展望

SmartNews

01 SmartNews数据湖介绍

SmartNews

2012

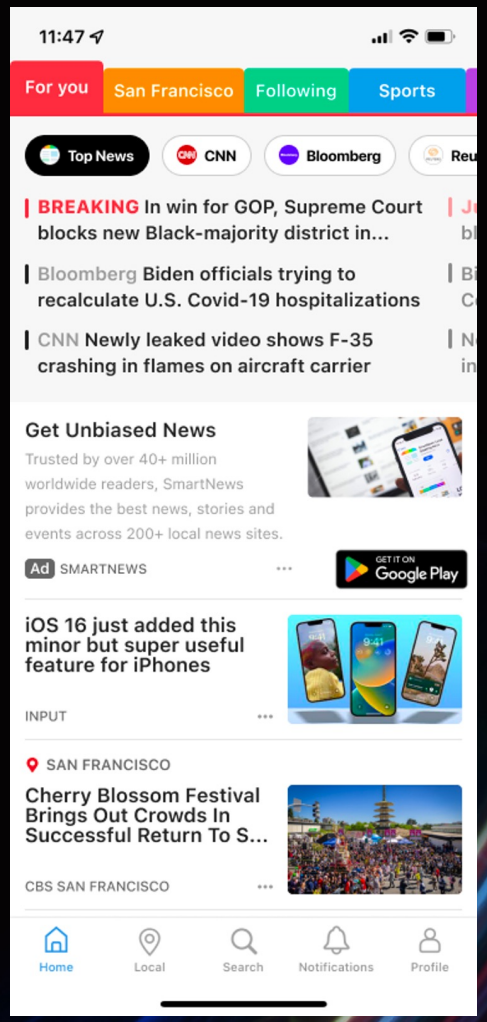
创立于东京

2014

纽约/旧金山/帕托 办公室成立

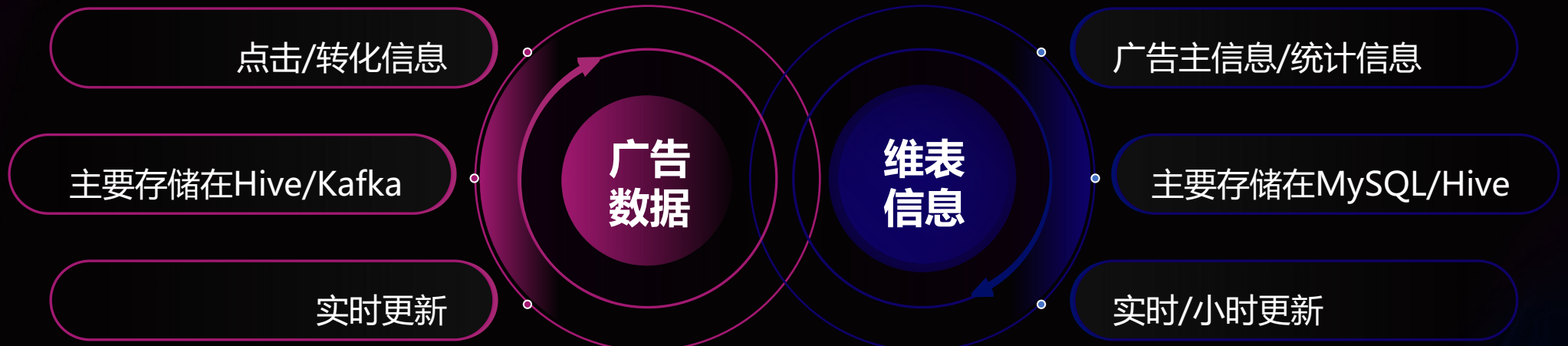
2019

上海/北京 办公室成立



SmartNews

广告数据湖



SmartNews

挑战

需要按广告主键去重

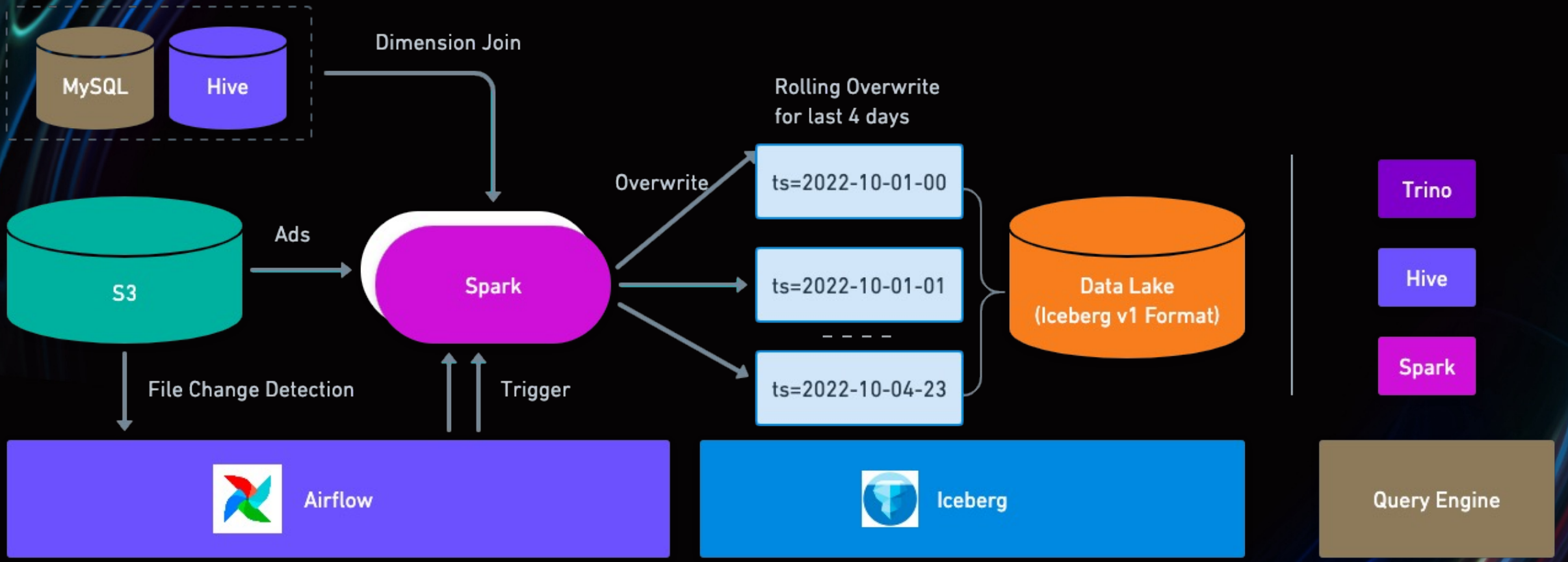
需要更新点击/转化时间戳字段

下游近实时读取

SmartNews

02 基于Iceberg v1格式的数据湖实践

SmartNews


MySQL
Hive
Dimension Join
Rolling Overwrite
for last 4 days
Overwrite
ts=2022-10-01-00
S3
Ads
Spark
ts=2022-10-01-01
Data Lake
(Iceberg v1 Format)
Trino
Hive
Spark
ts=2022-10-04-23
File Change Detection
Trigger
Airflow
Iceberg
Query Engine

SmartNews

解决挑战

在Spark作业中按照主键去重并且更新时间戳

通过Iceberg方案解决上下游同时读写问题

小时级别更新数据

SmartNews

方案不足

占用Infra资源太多

计算资源浪费 - 需要更新的行只占总体的1%左右

存储资源浪费 - 每次Overwrite都需要重写所有数据

并行提交到Iceberg的锁问题

SmartNews

03 基于Flink实时更新的数据湖(Iceberg v2)解决方案

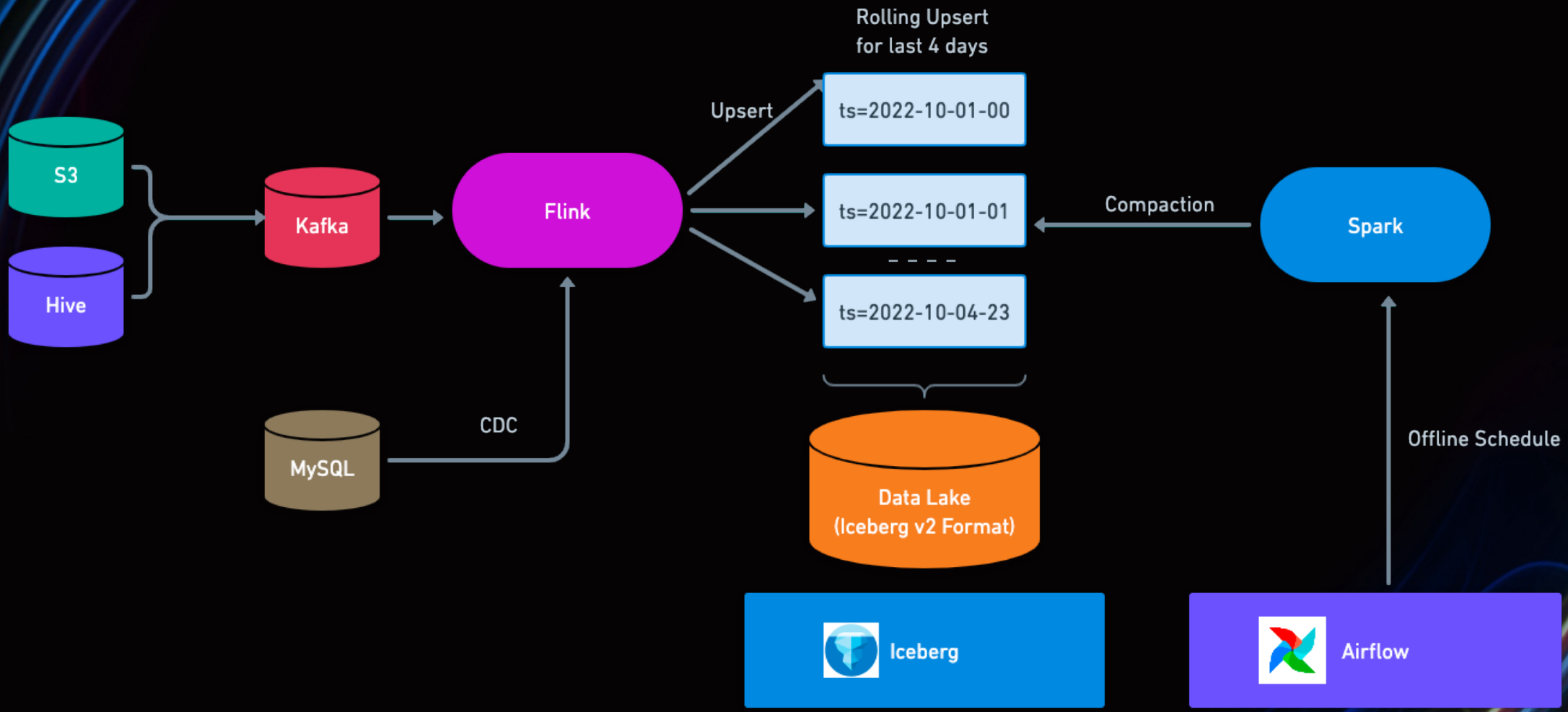
SmartNews

解决方案

Iceberg v2支持行级别更新

Flink实时写入 - Merge On Read

MySQL CDC流式解决dimension join


Rolling Upsert
for last 4 days
Upsert
S3
Hive
Kafka
Flink
ts=2022-10-01-00
ts=2022-10-01-01
ts=2022-10-04-23
Compaction
Spark
CDC
MySQL
Data Lake
(Iceberg v2 Format)
Offline Schedule
Iceberg
Airflow

SmartNews

解决方案对比

	Spark + Iceberg v1	Flink + Iceberg v2
写入方式	Overwrite	Upsert
输出文件数量	文件大小平均，数量可控	小文件数量巨大
计算方式	全部重新计算	Merge on Read，仅计算更新数据
实效性	小时级别	分钟级别

SmartNews

04 实时更新小文件问题的优化

SmartNews

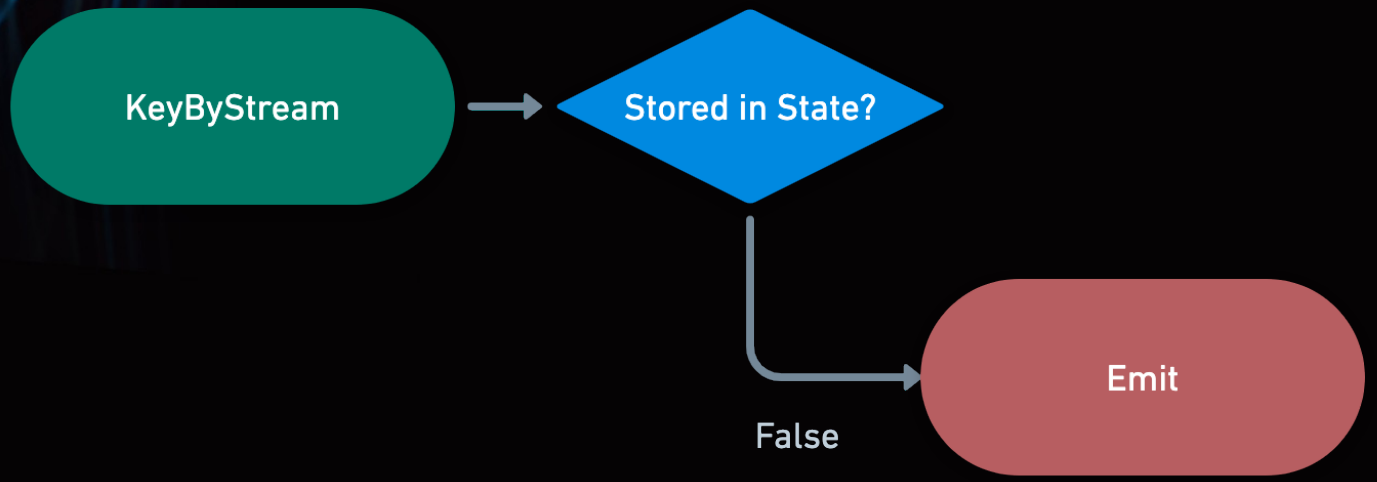
Iceberg Sink - Upsert Mode

每次插入数据会生成两条Record - Delete/Insert

存储空间浪费

下游Writer造成CPU压力

Flink Generated RowData

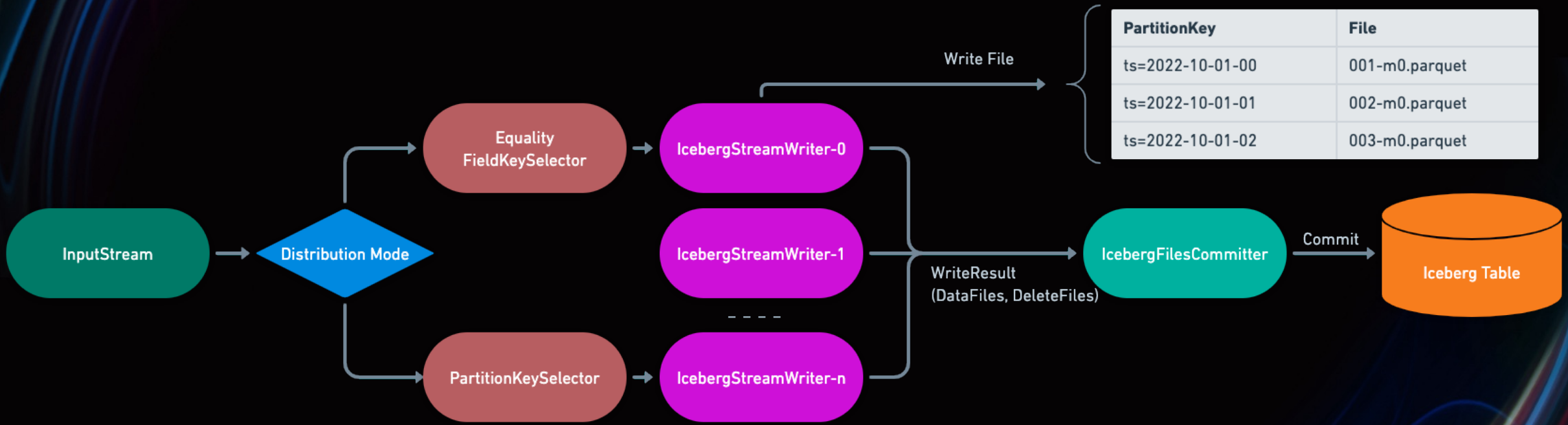


RowKind	Data
UPDATE_BEFORE	{Equaility Fields}
UPDATE_AFTER	{All Data Fields}

RowKind	Data
INSERT	{All Data Fields}

SmartNews

Iceberg Flink Sink



SmartNews

EqualityFieldKeySelector

按照Record主键shuffle到下游writer

在同一个partition路径下面会有多个writer同时写入

SmartNews

假设checkpoint的间隔为20分钟，使用10个writer去写文件

Partition	Record数量	每小时文件生成数量
ts=2022-10-01-23	xxx M	3(checkpoint interval) * 10(writer) * 3 files(data file/equality delete/position delete)
ts=2022-10-01-22	xx M	90
…	…	…
ts=2022-09-27-00	x K	90

SmartNews

PartitionKeySelector

按照Record的partition信息shuffle到下游writer

在同一个partition路径下面只会有1个writer写入

SmartNews

假设checkpoint的间隔为20分钟，使用10个writer去写文件

Partition	Record数量	每小时文件生成数量
ts=2022-10-01-23 BackPressure	xxx M	3 * 3（相同parition的record会被shuffle到同一个writer）
ts=2022-10-01-22 BackPressure	xx M	9
ts=2022-10-01-21	x M	9
…	…	…
ts=2022-09-27-00	x K	9

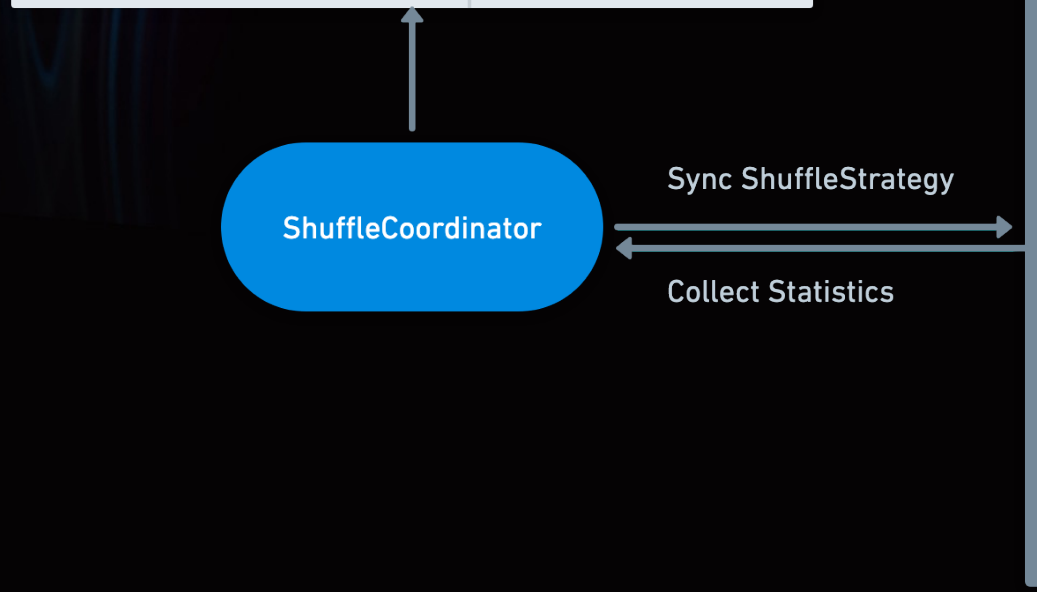
Dynamic Shuffle Operator

Partition	Record数量	Shuffle Strategy
ts=2022-10-01-23	xxx M	EqualityFieldKeySelector
ts=2022-10-01-22	xx M	EqualityFieldKeySelector
ts=2022-10-01-21	x M	PartitionKeySelector
…	…	…
ts=2022-09-27-00	x K	PartitionKeySelector

Dynamic Shuffle Operator

ShuffleStrategy

PartitionKey	Strategy
ts=2022-10-01-23	EqualityField
ts=2022-10-01-22	PartitionKey
ts=2022-10-01-01	PartitionKey



PartitionStatistics

PartitionKey	Statistics (Record Number)
ts=2022-10-01-23	xx M
ts=2022-10-01-22	x M
ts=2022-10-01-01	x K

SmartNews

DynamicShuffleKeySelector

按照当前最大PartitionKey来分配ShuffleStrategy

按照历史数据信息来动态分配ShuffleStrategy

确保所有subtask都使用相同的ShuffleStrategy

SmartNews

实验对比

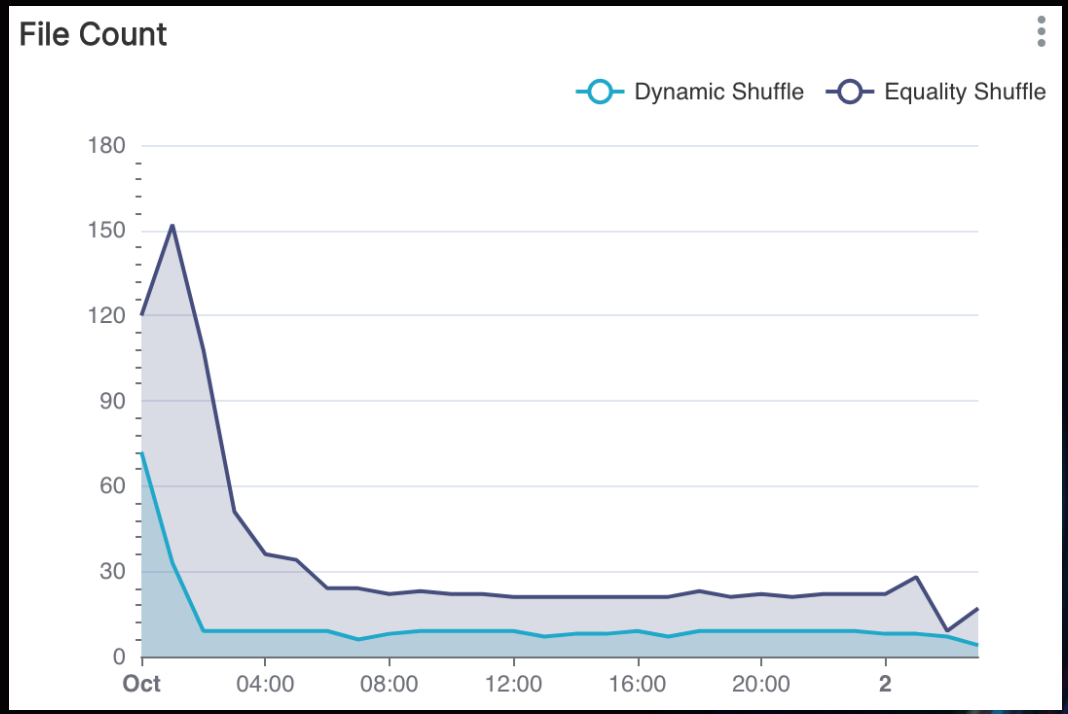
对比24小时以内每小时文件生成的数量以及平均大小

Flink并发设置为20

SmartNews

相同partition每小时新增文件数量

TS Offset (Hour)	No Shuffle	Dynamic Shuffle
+0	120	72
+1	152	33
+2	108	9
+3	51	9
+4	36	9
+5	34	9



SmartNews

相同partition每小时新增文件平均大小

TS Offset (Hour)	No Shuffle	Dynamic Shuffle
+0	34 MB	60 MB
+1	24 MB	40 MB
+2	1 MB	3 MB
+3	100 KB	600 KB
+4	60 KB	300 KB
+5	10 KB	50 KB
+6	20 KB	50 KB

05 总结与展望

SmartNews

总结与展望

相比较于Spark + Iceberg v1的方案，Flink实时写入的方案减少了50%的Infra成本

极大地避免了重复计算以及重复数据文件

实效性也从之前的小时级提高到了分钟级

DynamicShuffleOperator可以进一步按照写进文件的速率来分配ShuffleStrategy

FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

THANK YOU

谢 谢 观 看