

网易 ARCTIC

基于 APACHE ICEBERG 构建的实时湖仓一体系统

张永翔

网易数帆 资深大数据平台开发



目录 CONTENT

01

业务当前的挑战

Lambda 架构下流与批割裂带来的问题

02

网易Arctic

基于 iceberg 构建的湖仓一体系统

03

业务实践

Arctic 在网易内外的实践

04

未来规划



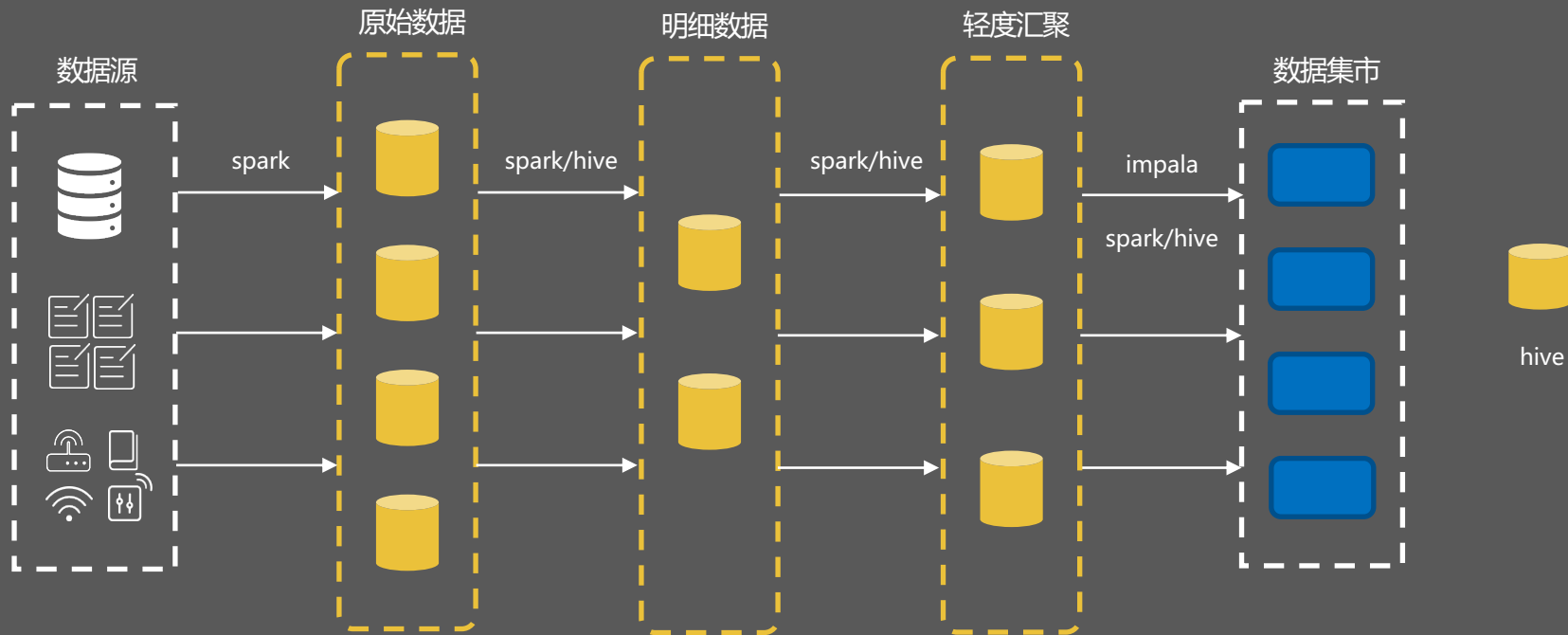
01

业务当前的挑战

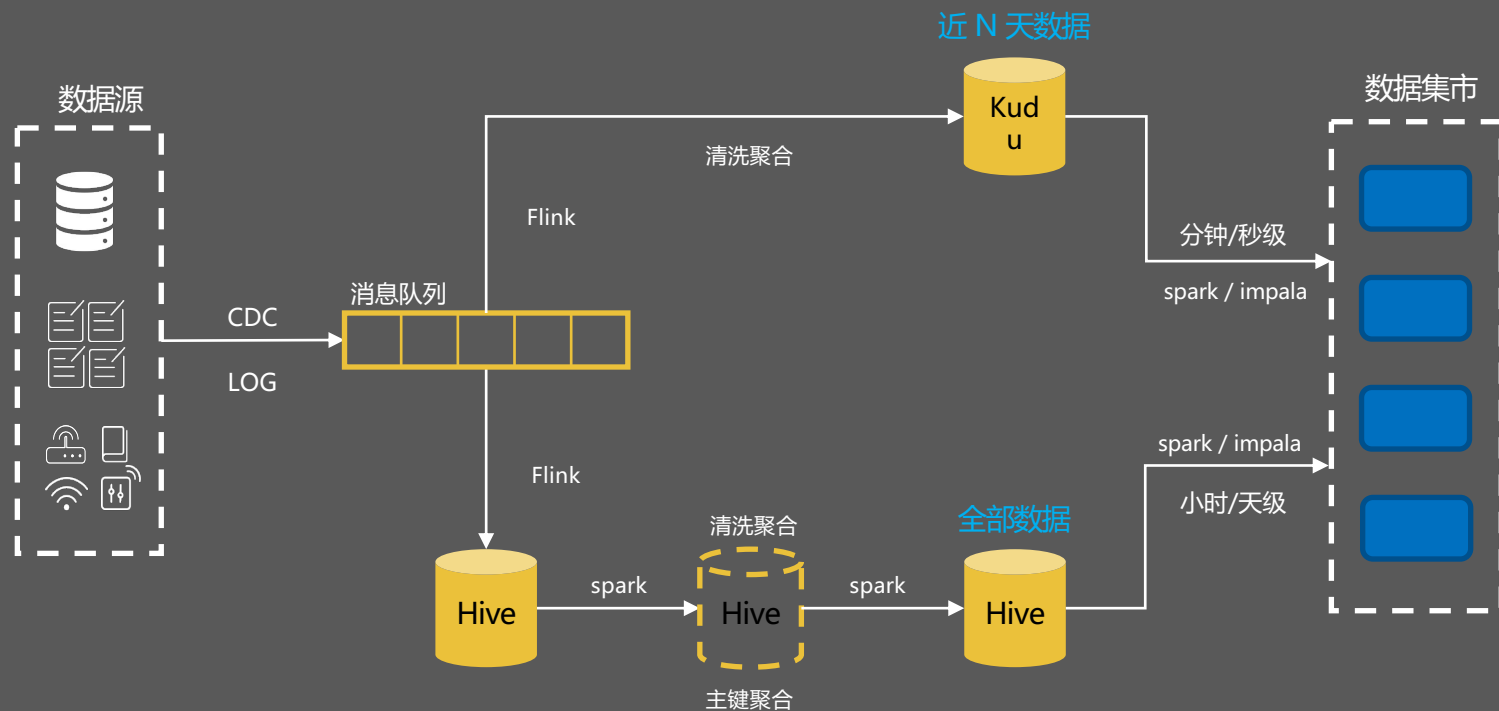
Lambda 架构下流与批割裂带来的问题



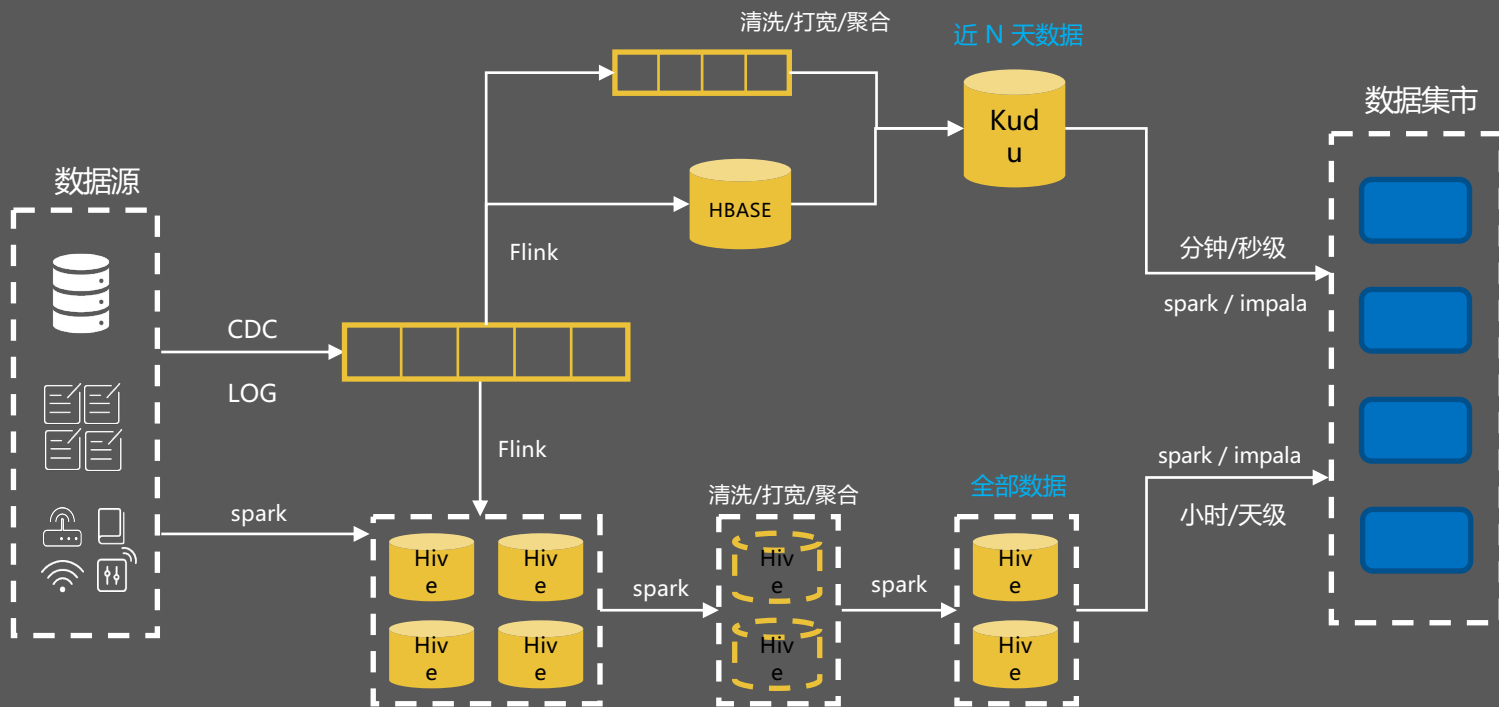
T+1 离线数据生产



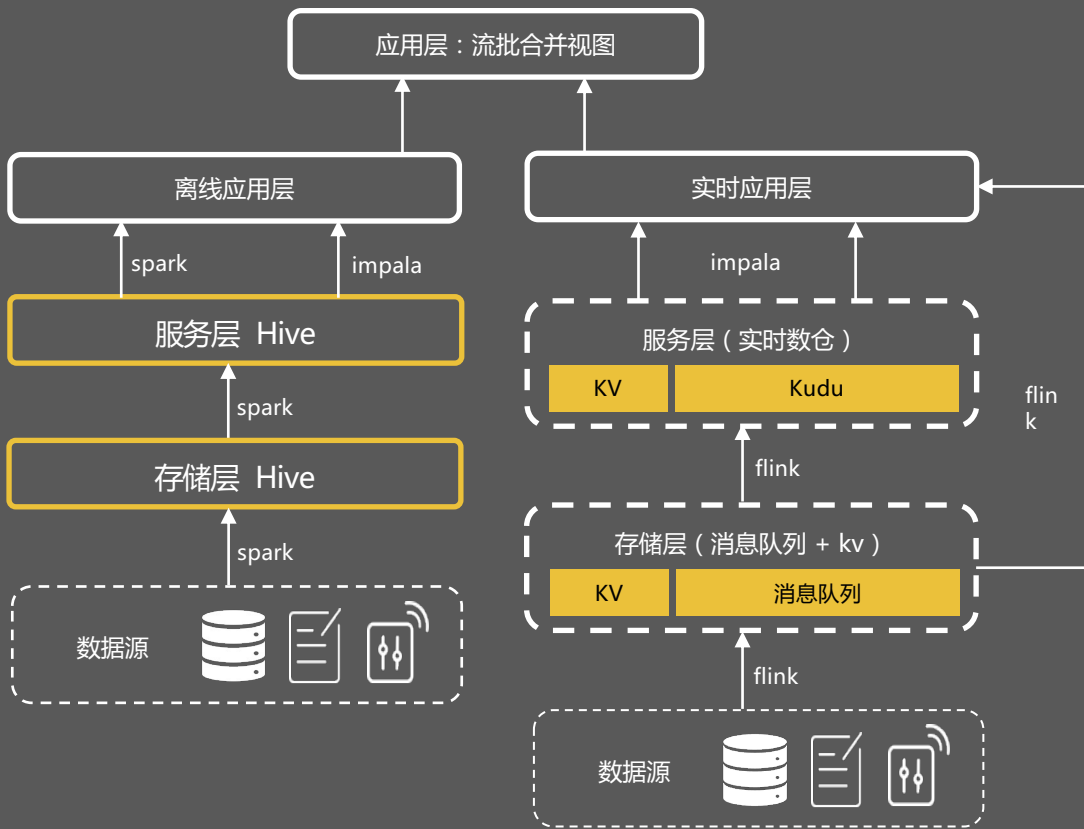
场景一：初步引入实时化



场景二：更加复杂的实时化



流批计算割裂下的 lambda 架构



- 数据孤岛 (Kudu 等)
 - 独立采购和部署
 - 冗余存储浪费成本
 - 难以数据复用和互通
- 研发体系割裂
 - 研发人效低
 - 研发规范不通用
 - 应用层视图合并复杂
- 指标和语义二义性



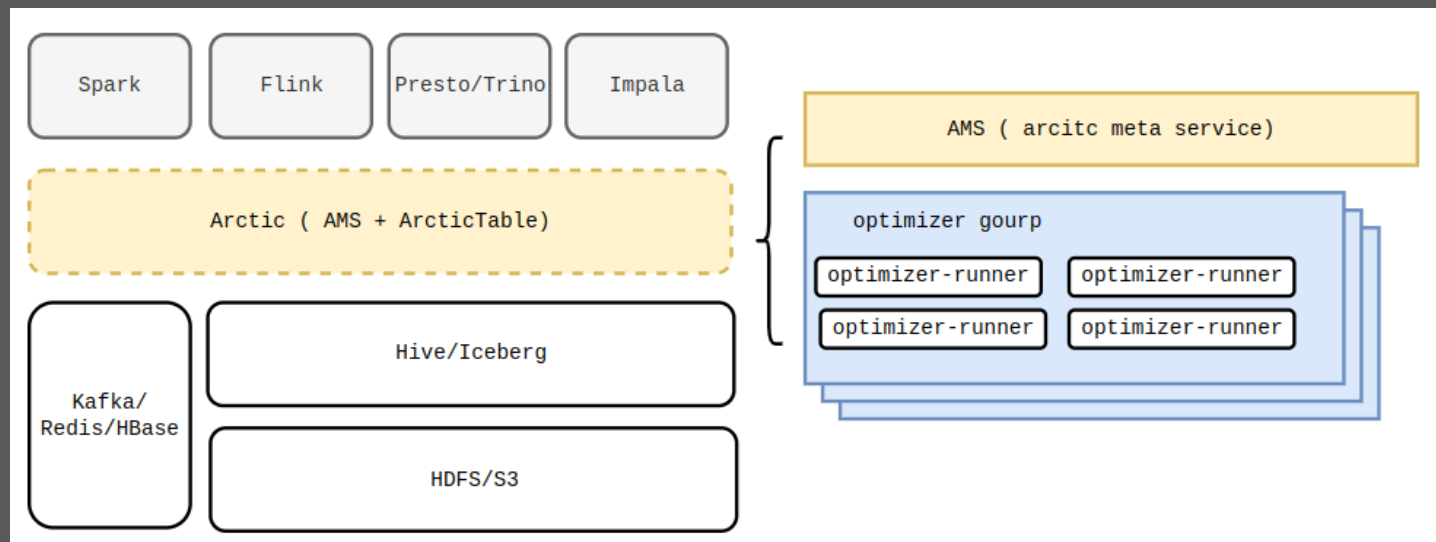
02

Arctic 功能特性

基于 Apache Iceberg 构建的湖仓一体系统



Arctic 的定位

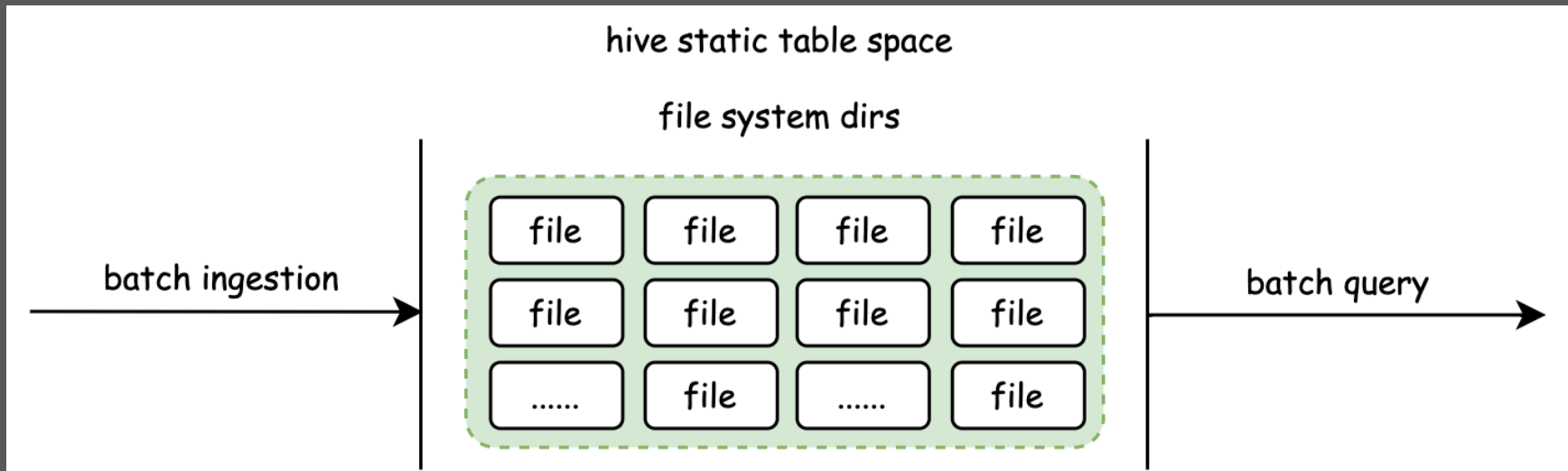


一句话概括：

定义在 **Hive/Iceberg** 表格式之上，计算引擎之下的 **TableService**，并提供表结构优化以及**Kafka**封装的实时湖仓系统

Arctic Table —— Optimize for Hive/Iceberg

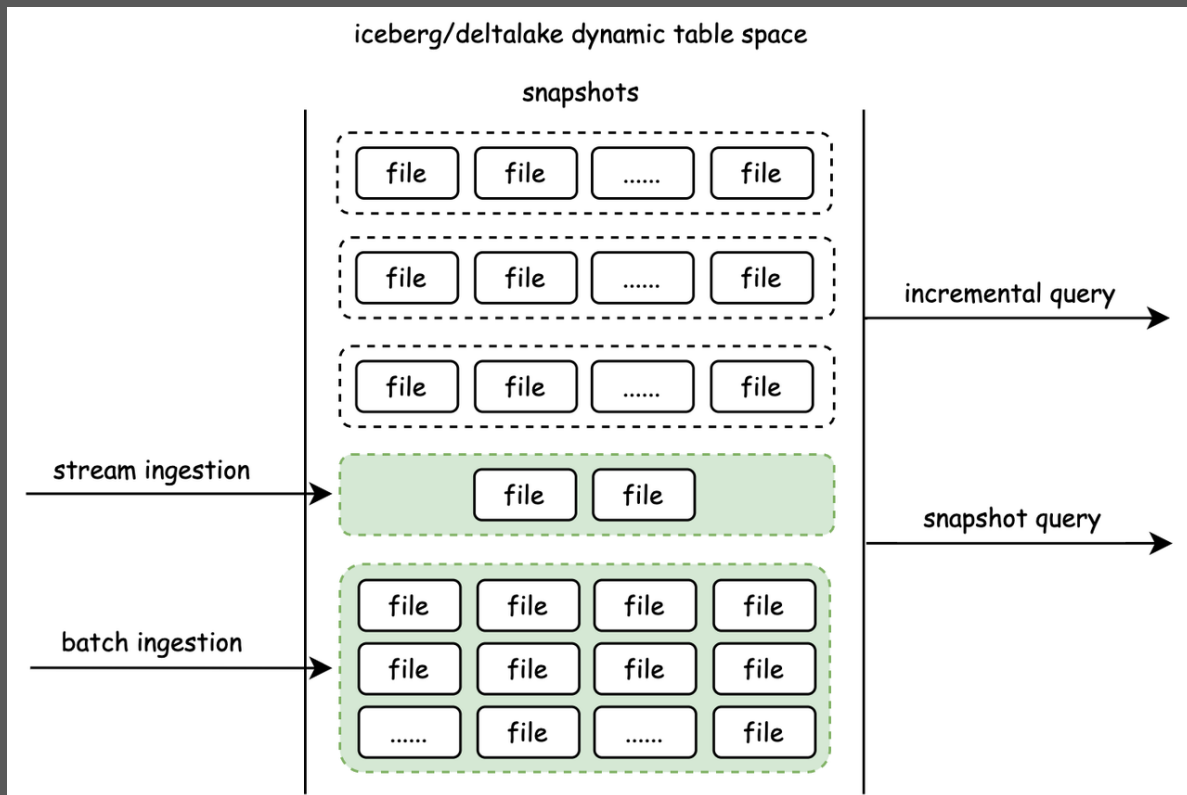
Hive 的摄取与计算场景



T+1/T+H 场景，每次进行全量的计算



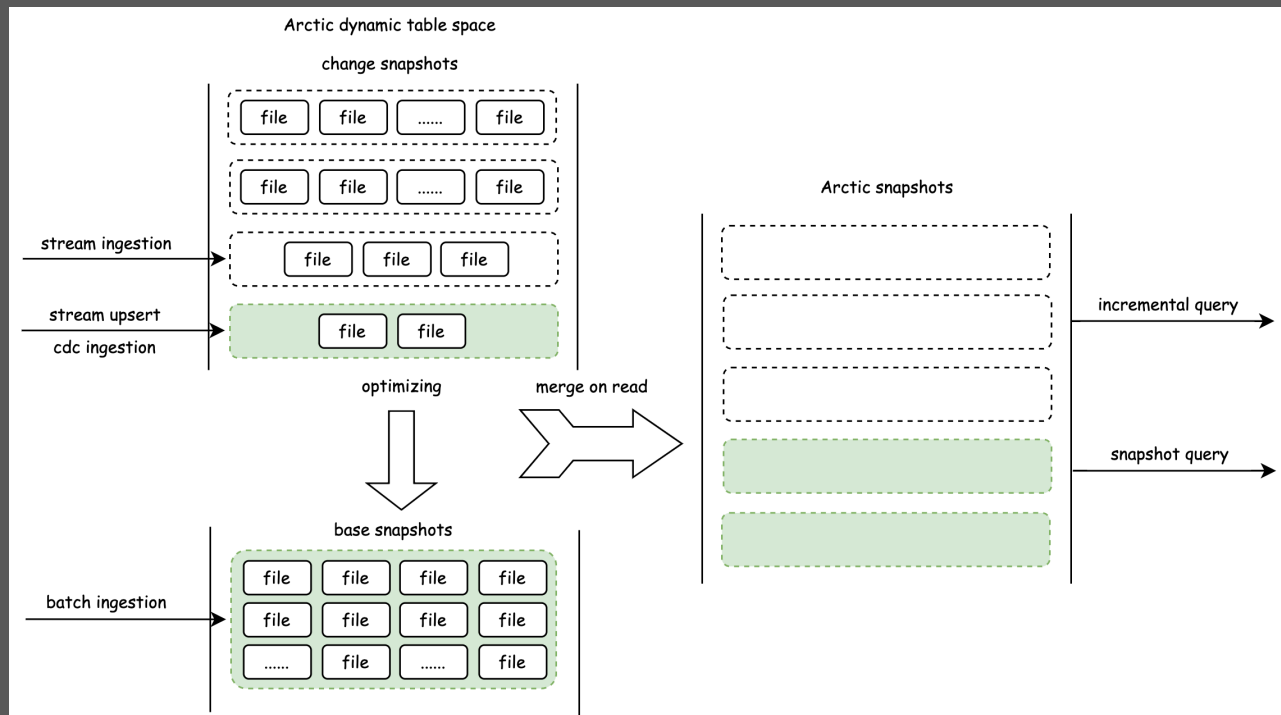
Arctic Table —— Optimize for Hive/Iceberg



Iceberg/Deltalake 的摄取与计算

抽象出 Snapshot 概念，通过快照隔离实现MVCC 和ACID，支持数据实时摄取

Arctic Table —— Optimize for Hive/Iceberg



Arctic 在 Iceberg 的基础上，将 Batch 和 Stream 写入的文件进行区分，分为 change store 和 base store.

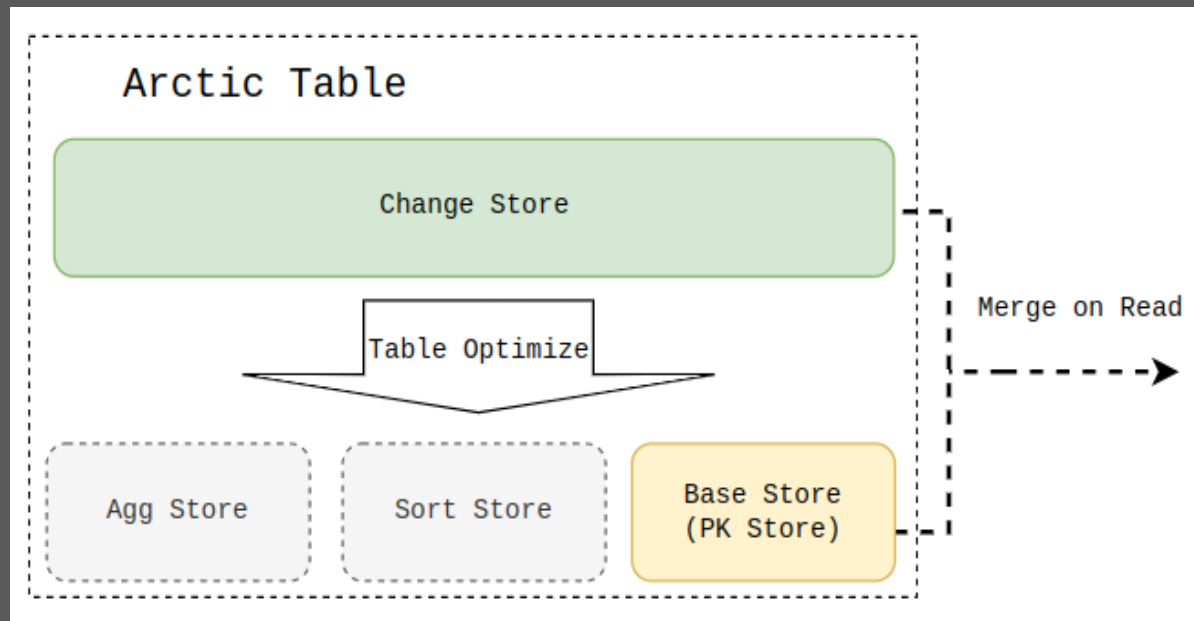
通过异步的 optimizing 对 stream 写入的文件进行合并，并提供了小文件治理、唯一键保证和 upsert 的能力

并通过 ArcticTable 封装的接口提供 merge on read, 实现准实时的读写能力

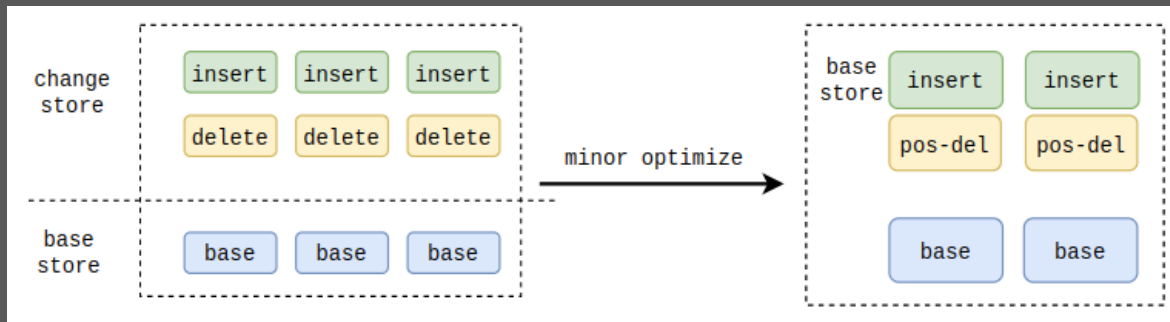


Arctic Table —— Primary Key Support

- 支持 Primary Key
- 支持 CDC ingestion
 - 实现 Upsert 语义
- 主键唯一性约束实现
 - Merge on read
 - optimize
- 未来扩展 SortKey / AggKey

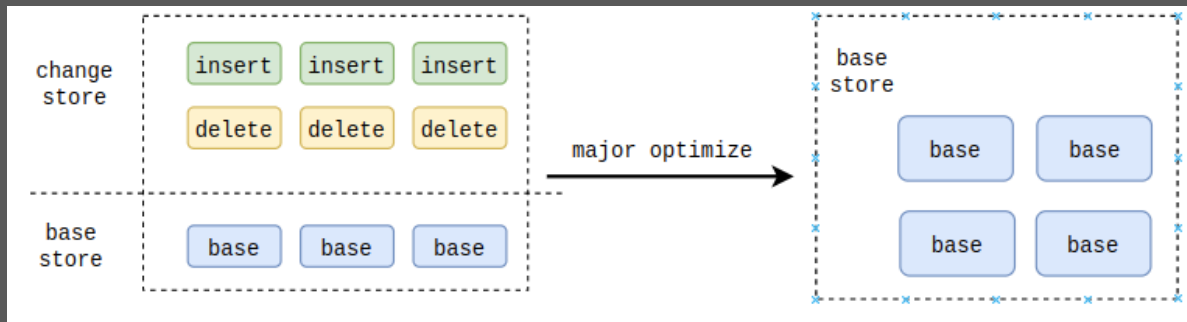


Arctic Table —— Auto Optimize



• Minor Optimize

- 执行周期短 (约10min 一次)
- 优化小文件数量
- eq-del 转换为 pos - del
- 只针对 change file



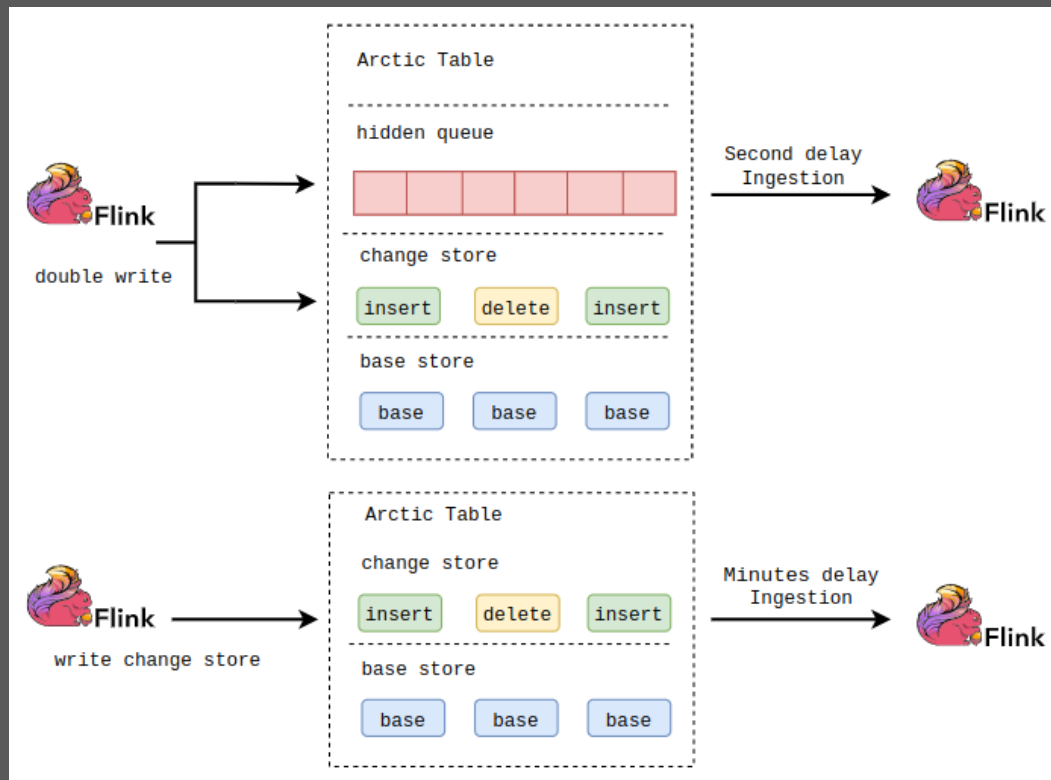
Major Optimize

- 执行周期长 (约1 day 一次)
- 合并 change file 到 base file
- 兼容 Hive 读

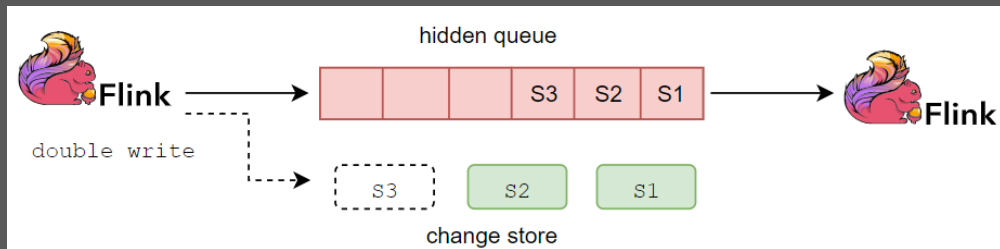


Arctic Table —— 流批一体表（支持CDC）

- 下游可订阅 Arctic 表变更
- 支持 Hidden Queue
 - 秒级延迟订阅
 - 通过消息回撤实现最终一致性
- arctic-flink-connector 封装双写和回撤实现细节
- 不开启 Hidden Queue
 - 分钟级延迟订阅

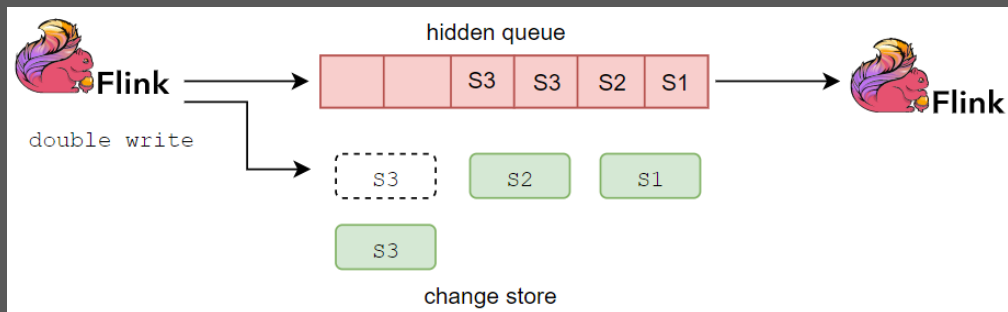


Arctic Table —— 流批一体表（双写一致性保证）

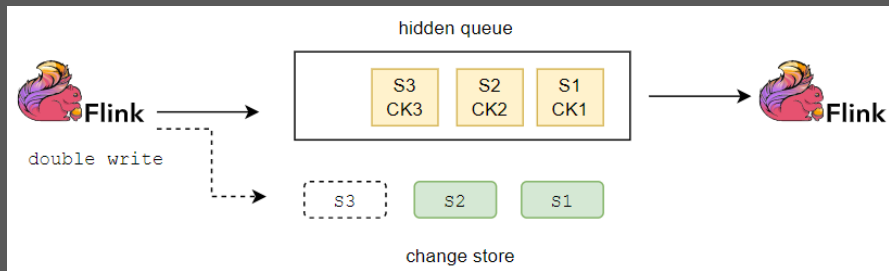


开启 hidden queue 后，上游消息先写 kafka 再写 change store，change store 在 checkpoint 时提交，此时如果上游任务failed会出现一致性问题。

当上游任务 failover 后，会导致部分数据重复写入 kafka，下游重复消费

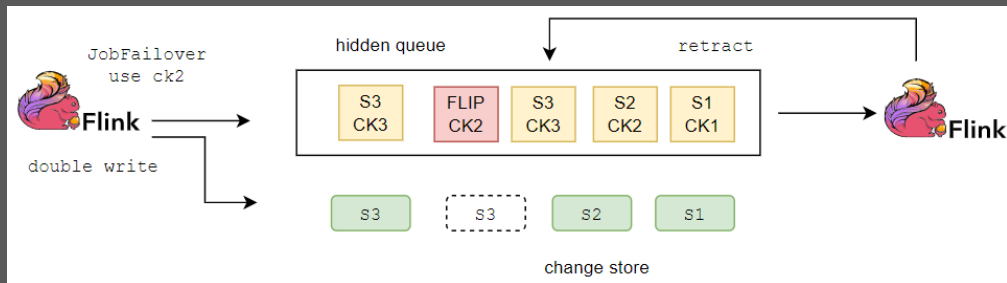


Arctic Table —— 流批一体表（Retract 实现双写一致性保证）



对消息进行封装，带上消息对应Writer 的 State 周期

下游算子记录每个CK 对应的offset 信息



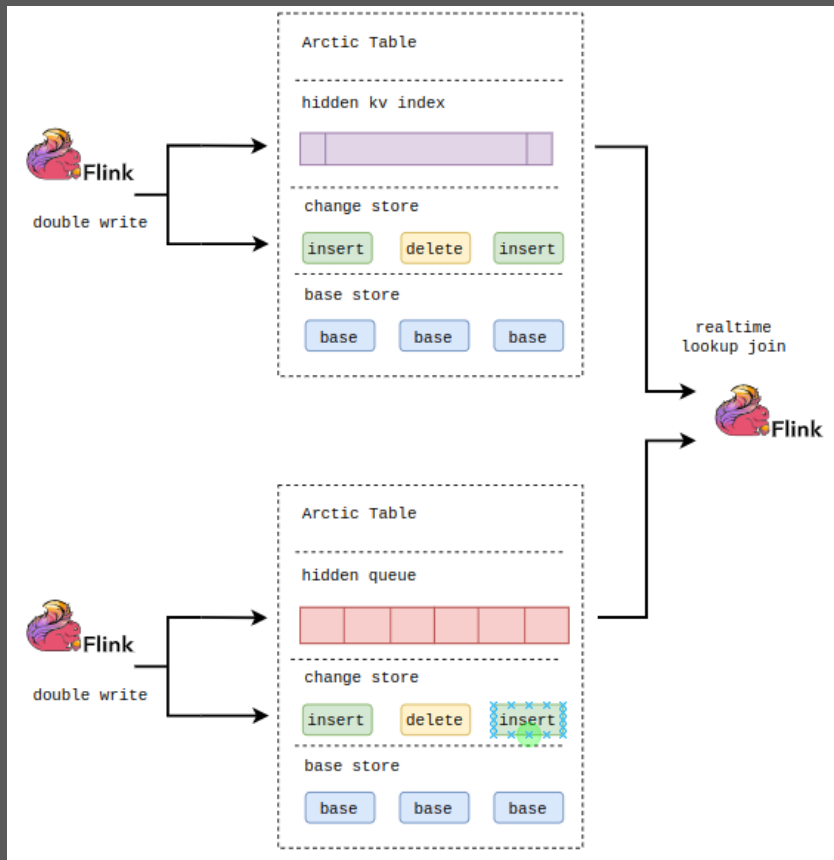
任务failover 后，先写入 Flip 消息，带上任务恢复的CP信息

下游任务收到 Flip 消息，从 Kafka 找到对应的消息并 retract

整个过程由 arctic-flink-connector 封装



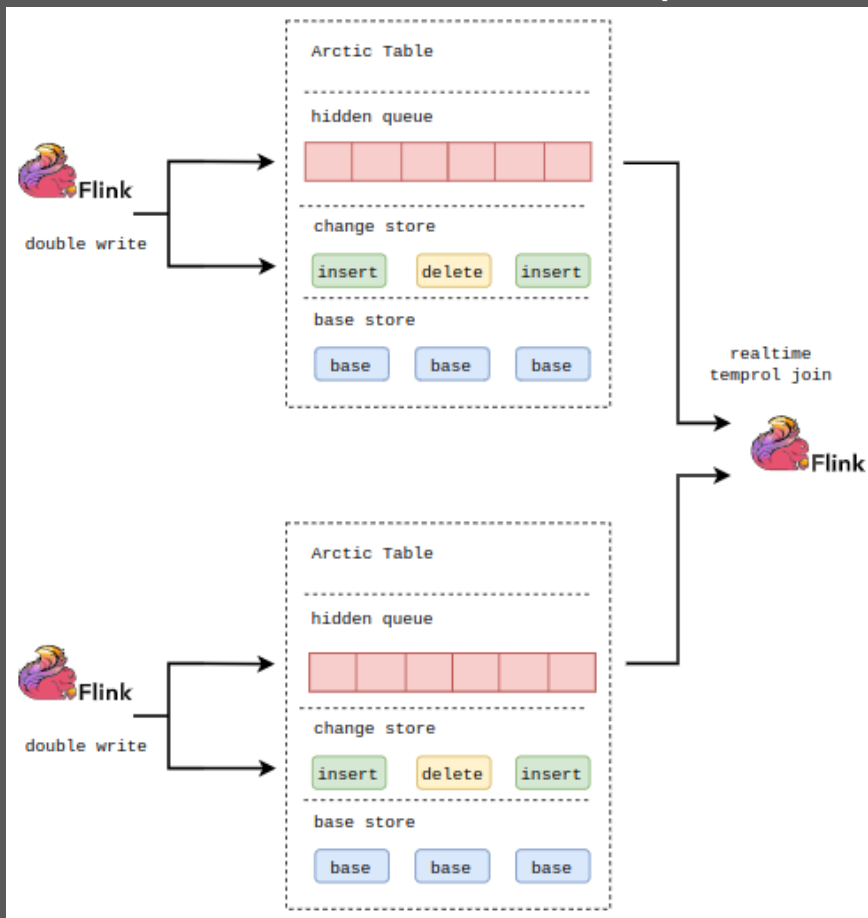
Arctic Table —— 流批一体表 (支持 Lookup join)



- 通过 Hidden kv index 支持 lookup join
- 同样不需要关心实现细节，Arctic Table 可以直接当维表用
- 未来会实现 Temporal Join 无需依赖外部 KV



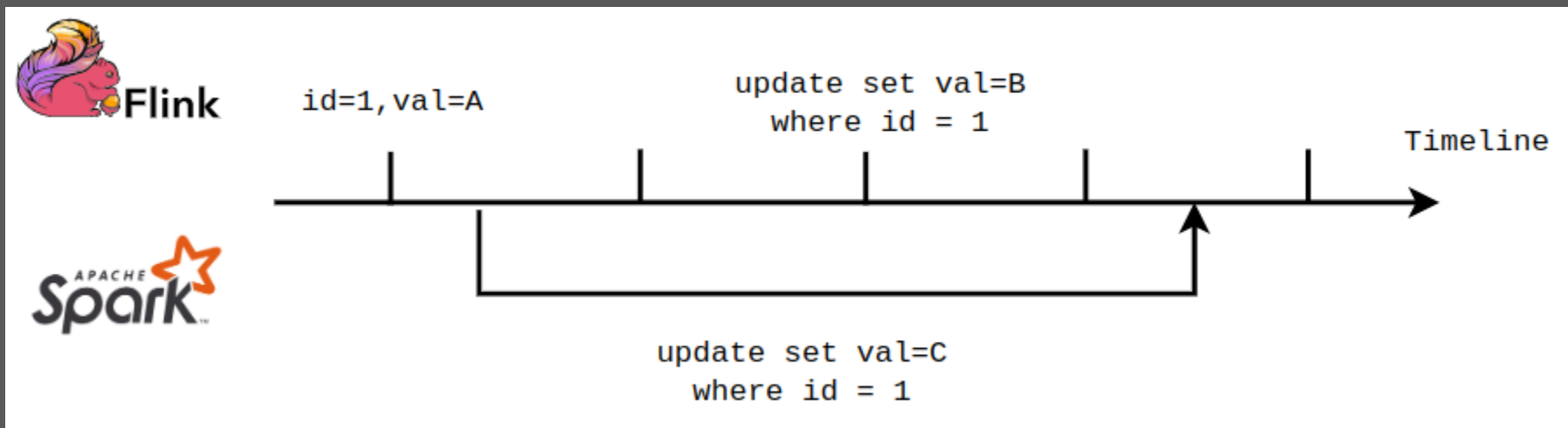
Arctic Table —— 流批一体表 (支持 Temporal join 规划中)



- 基于 Flink1.12 Temporal Table (时态表) 功能
- 不需要引入额外的 KV 组件
- 支持 event time join



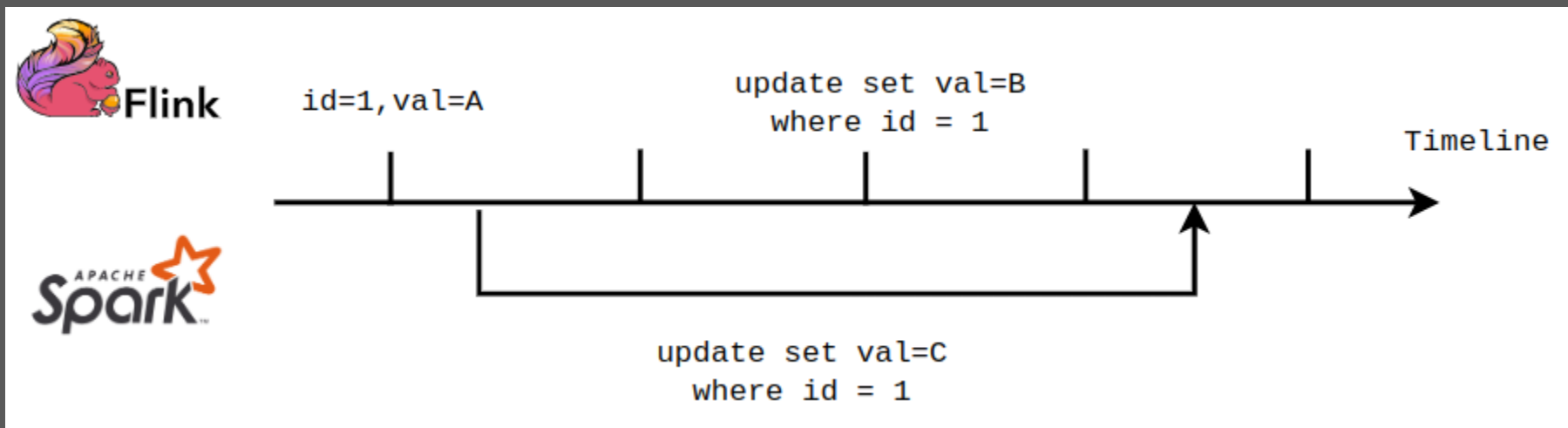
Arctic Table —— 并发写入与一致性保证



`id=1, val=?`

正确的值应该是什么

Arctic Table —— 并发写入与一致性保证



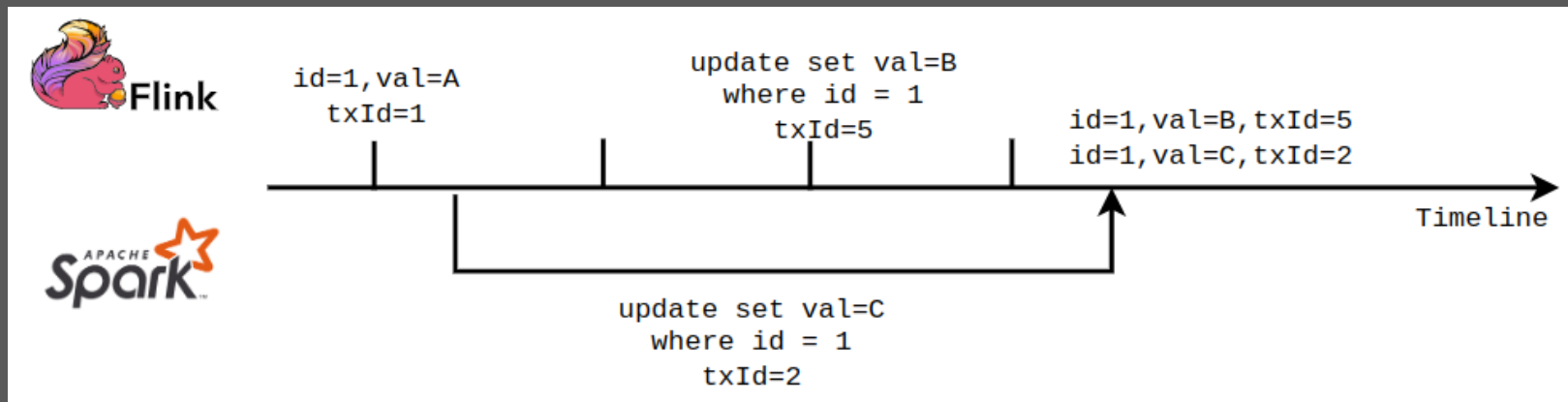
`id=1, val=?`

正确的值应该是什么

应用场景:

- 数据回补
(CDC 数据丢失, 批任务全量补数)
- GDPR/CPAA
(删除用户在大数据系统内的存量数据)

Arctic Table —— 并发写入与一致性保证



AMS 分配txId，标记记录写入先后顺序

Merge on read 时，确认可见 record 为哪一个

Minor/Major Optimize 时，确认应该保留哪一个 record



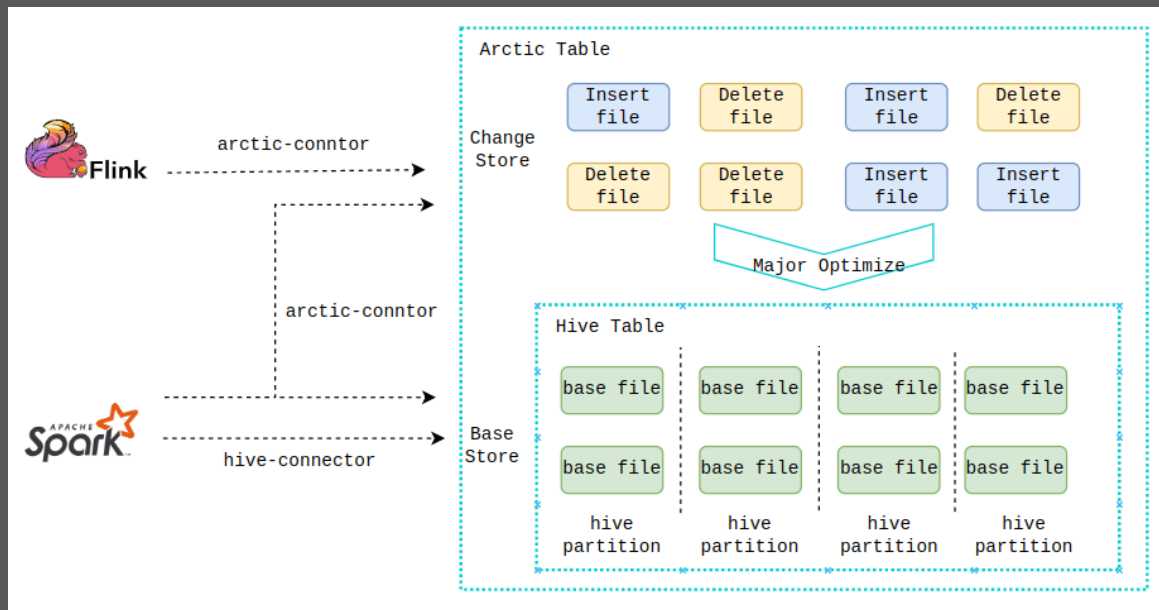
Arctic Table —— Hive 兼容

Hive兼容的原因

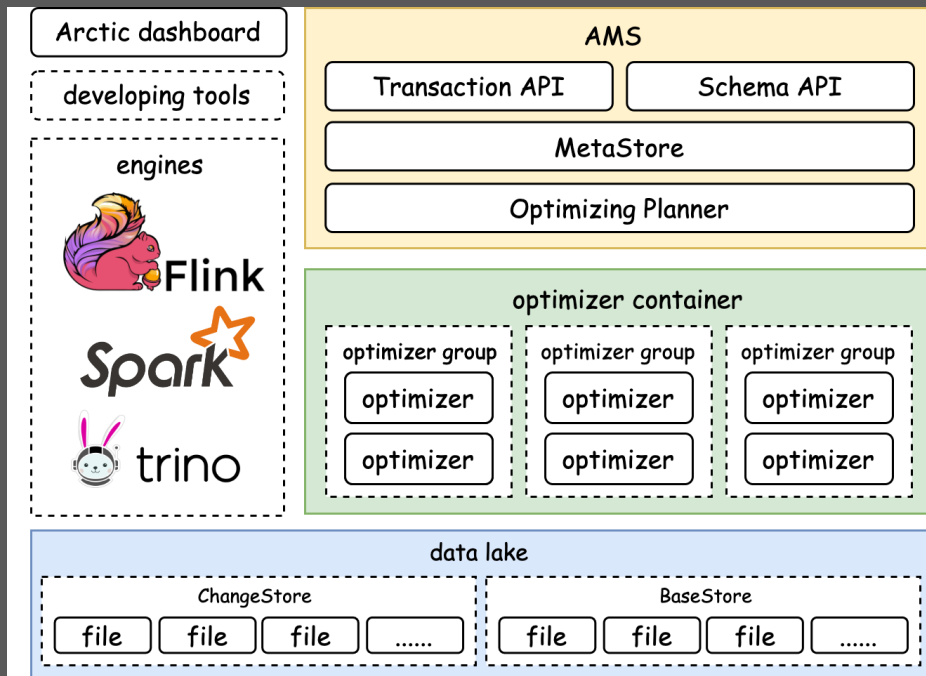
- 基于Hive的离线数据仓库已经非常成熟，有广泛应用
- 兼容离线数仓，方便离线数仓到实时数仓的升级

对Hive兼容能力

- 支持hive表原地升级为arctic表
- 支持将arctic base store 作为hive表读取
- 支持将arctic表作为hive表写入 (INSERT OVERWRITE)
- 通过 optimize 实现实时写入到hive表数据的同步
- 自动识别Hive 写入文件



Arctic Meta Service




AMS 被定义为新一代的HMS

- 负责Arctic Table Metadata 管理
- 分配事物ID
- 面向计算引擎的元数据服务
- 触发结构优化任务
- Optimizer 调度与资源管理
- 提供运维友好的 Dashboard

AMS Dashboard —— Transaction 管理

Dashboard 上展示某次Transaction提交的文件信息



dwd_rev_send_coin_di Create Time: 2022-02-23 10:25:23

TableSize: 16.40GB | File: 380 | Average File Size: 44.19MB

Details | Files | Transactions | Optimizes

All > Transaction ID 8433567282447972190

File	Partition	File Type	Size	Commit Time
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-03	BASE_FILE	19.51MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-04	BASE_FILE	2.36MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-03	BASE_FILE	19.51MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-04	BASE_FILE	2.35MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-03	BASE_FILE	19.50MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-03	BASE_FILE	19.50MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-04	BASE_FILE	2.36MB	2022-06-04 21:47:17
hdfs://hz-cluster11/user/da_music/hive/warehouse/tmp_music...	dt=2022-06-04	BASE_FILE	2.35MB	2022-06-04 21:47:17



AMS Dashboard —— Table Optimizer 管理

Dashboard 上展示某张表下的 Optimize 任务执行信息

<div><div></div><div>dwd_rev_send_coin_di</div><div>Create Time: 2022-02-23 10:25:23</div></div>					
TableSize: 16.40GB File: 380 Average File Size: 44.19MB					
<div>DetailsFilesTransactionsOptimizes</div>					
StartTime	Duration	Input		Output	
		Count	Size	Count	Size
2022-06-08 15:48:25	1 min 2 s	84	116.54MB	4	112.79MB
2022-06-08 15:38:25	59 s	84	115.08MB	4	111.33MB
2022-06-08 15:29:29	1 min 49 s	84	113.68MB	4	109.89MB
2022-06-08 15:20:52	3 min 15 s	68	111.55MB	4	108.43MB
2022-06-08 15:13:04	4 min 20 s	100	111.82MB	4	107.23MB
2022-06-08 14:58:31	1 min 14 s	84	109.14MB	4	105.47MB
2022-06-08 14:48:42	1 min 21 s	84	107.76MB	4	104.11MB
2022-06-08 14:39:26	1 min 57 s	84	106.44MB	4	102.74MB
2022-06-08 14:29:32	1 min 55 s	84	105.06MB	4	101.33MB
2022-06-08 14:19:54	2 min 33 s	84	103.68MB	4	99.94MB



对比总结

				Arctic
<ul style="list-style-type: none">不支持多写不支持秒级CDC不支持实时join	<ul style="list-style-type: none">不支持流式更新没有 merge on read对 Flink 不友好没有自动合并	<ul style="list-style-type: none">缺失文件合并功能不支持PrimaryKey多写场景一致性保障缺失	<ul style="list-style-type: none">存算不分离，无法利用HDFS 资源池数据孤岛，无法实现不同层数据串联写放大导致性能问题	<ul style="list-style-type: none">基于Iceberg，并兼容Iceberg所有功能Hive 兼容性好，业务升级阻力更低支持自动合并，动态调度合并任务支持 merge on read，提供分钟级延迟实时数仓提供流批一体功能，包括秒级实时订阅和实时join提供方便的运维管理平台
<ul style="list-style-type: none">对Hive 兼容支持不够友好缺失运维管理平台流批一体功能缺失（实时订阅，实时join）				

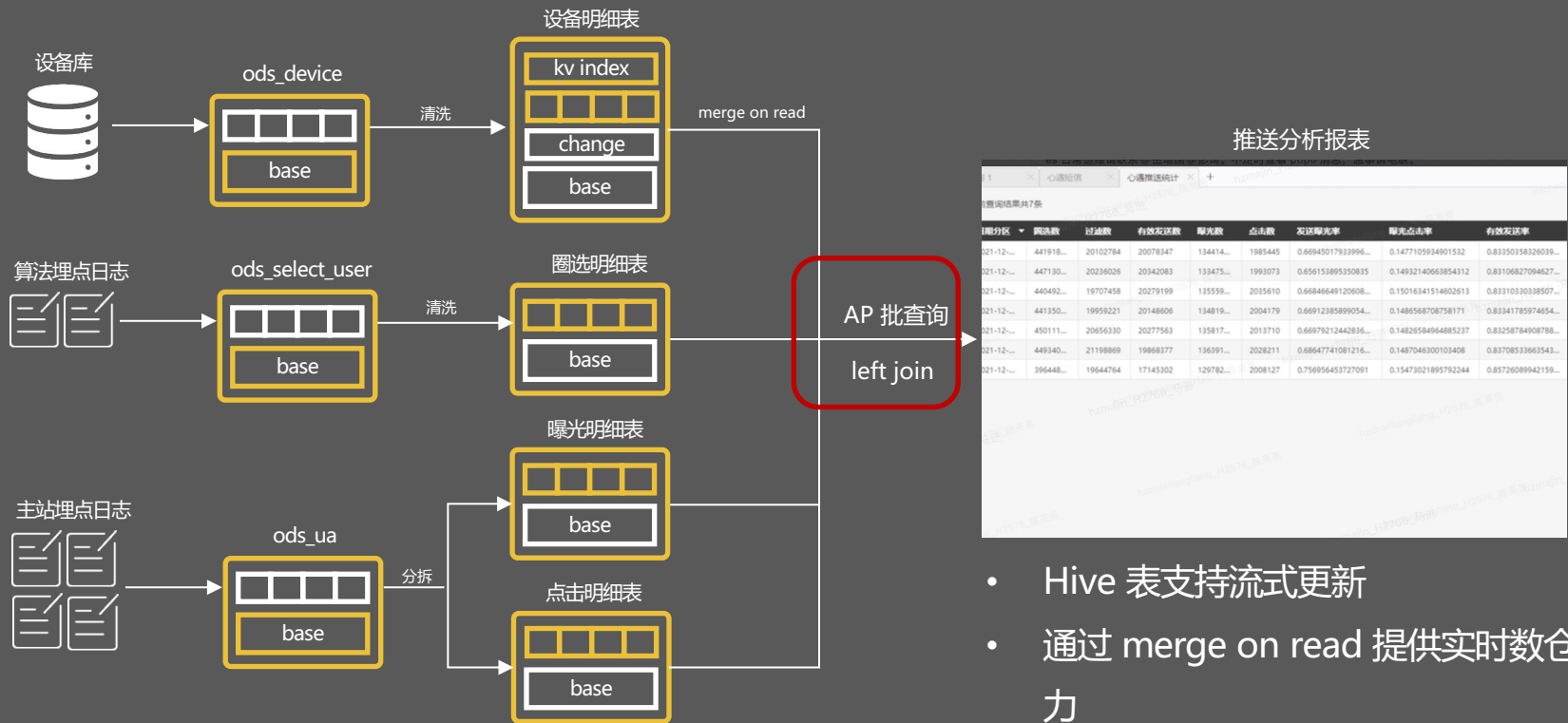


03

实践案例



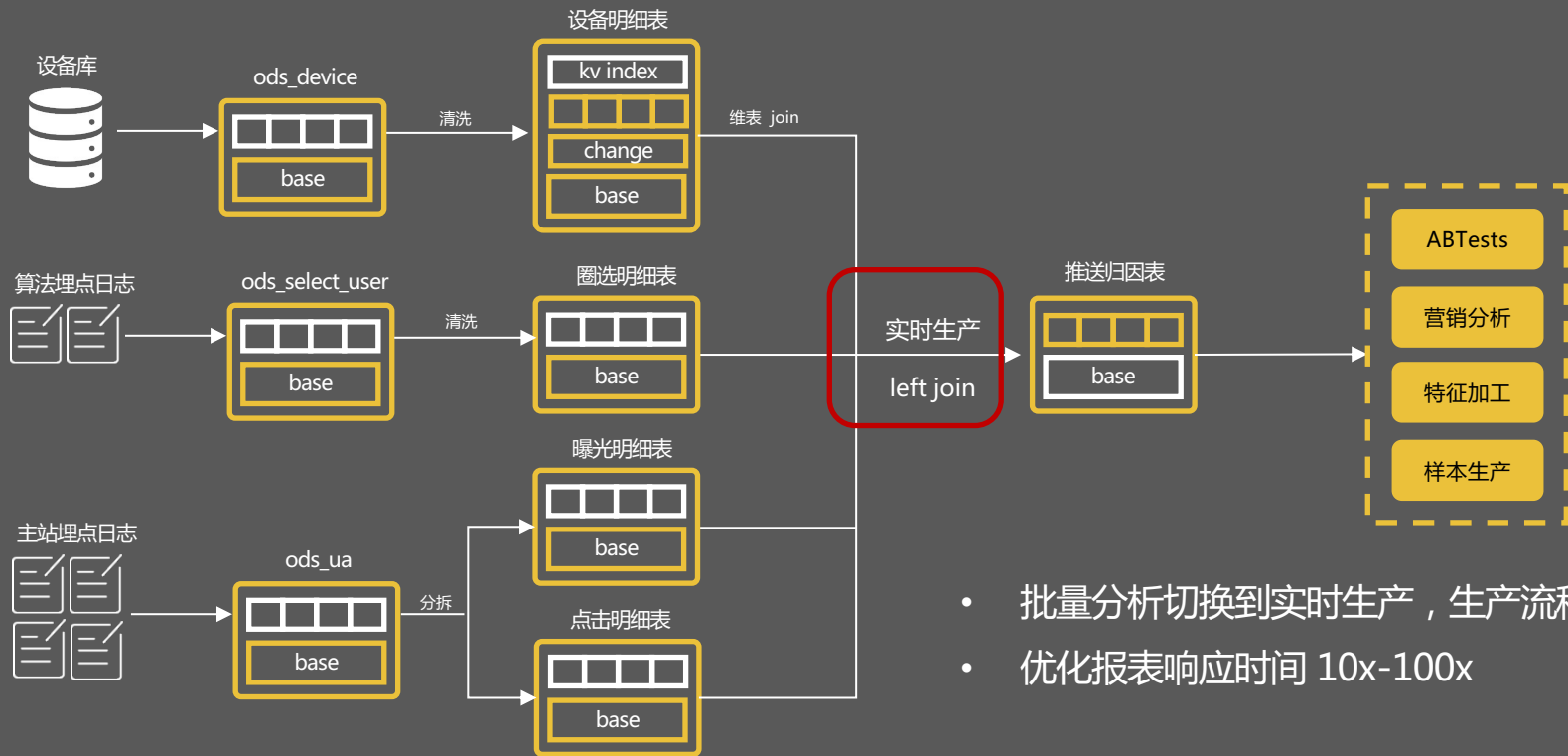
Arctic 应用举例 — 推送营销分析



- Hive 表支持流式更新
- 通过 merge on read 提供实时数仓能力



Arctic 应用举例 — 推送营销分析



- 批量分析切换到实时生产，生产流程复用
- 优化报表响应时间 10x-100x

04 未来规划



未来规划

- 更多流批一体场景
 - Rollup 聚合视图
 - Sort Key 支持, Z-ORDER 排序
 - 部分列的 Stream upsert
 - 支持 Temporal Join
- 更强的Dashboard
 - 任务血缘与数据血缘
 - SQL自助查询
- 安全体系完善
 - 支持开放式的权限插件, 支持对接Ranger
- 数据湖支持
 - S3/OSS



开源计划

预计6月底，敬请期待



网易数帆 |

DataFun.

非常感谢您的观看



网易数帆 |

DataFun.

