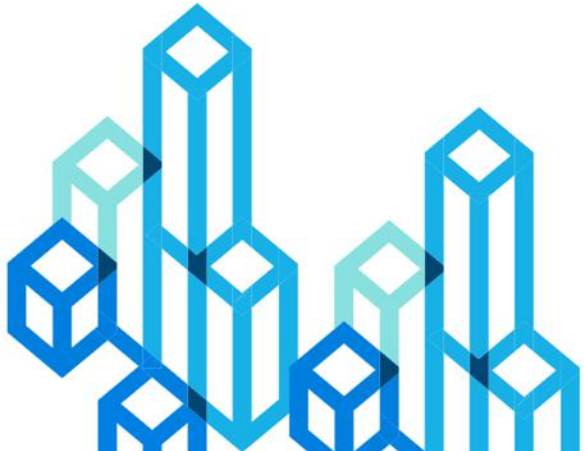
字节跳动基于 Iceberg 的海量特征存储实践

钱瀚 | 字节跳动



个人简介

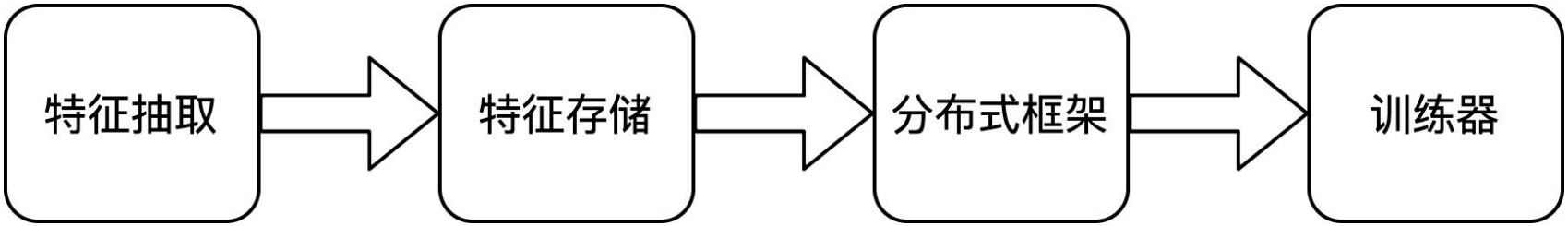
- 字节跳动基础架构资深研发
- 超过8年的研发经验
- 曾就职于搜狗、百度、京东
- 目前专注于大数据和AI基础架构方向

CONTENT

目录 >>

#1 背景

整体流程

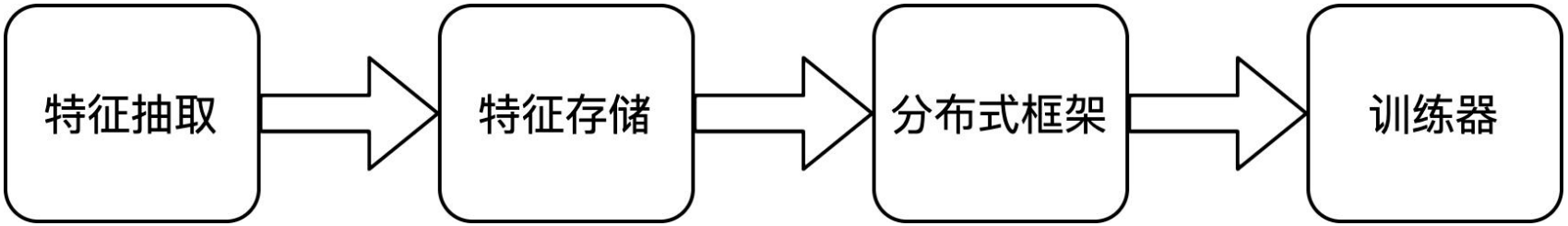


- 在线抽取特征
- 以行的格式存储在 HDFS
- 不存原始特征，只存抽取后的特征
- 数据的并发读取
- 断点恢复
- Format 解析
- 高速训练

规模

存储总量　　存储增量/天　　训练资源/天

痛点



- 特征抽取周期长
- 特征存储空间占用大
- 模型训练带宽大，数据读取有瓶颈

需求

- 存储原始特征
- 离线调研能力
- 支持特征回填
- 降低存储成本
- 降低训练成本
- 提升训练速度

#2 解決方案

选型

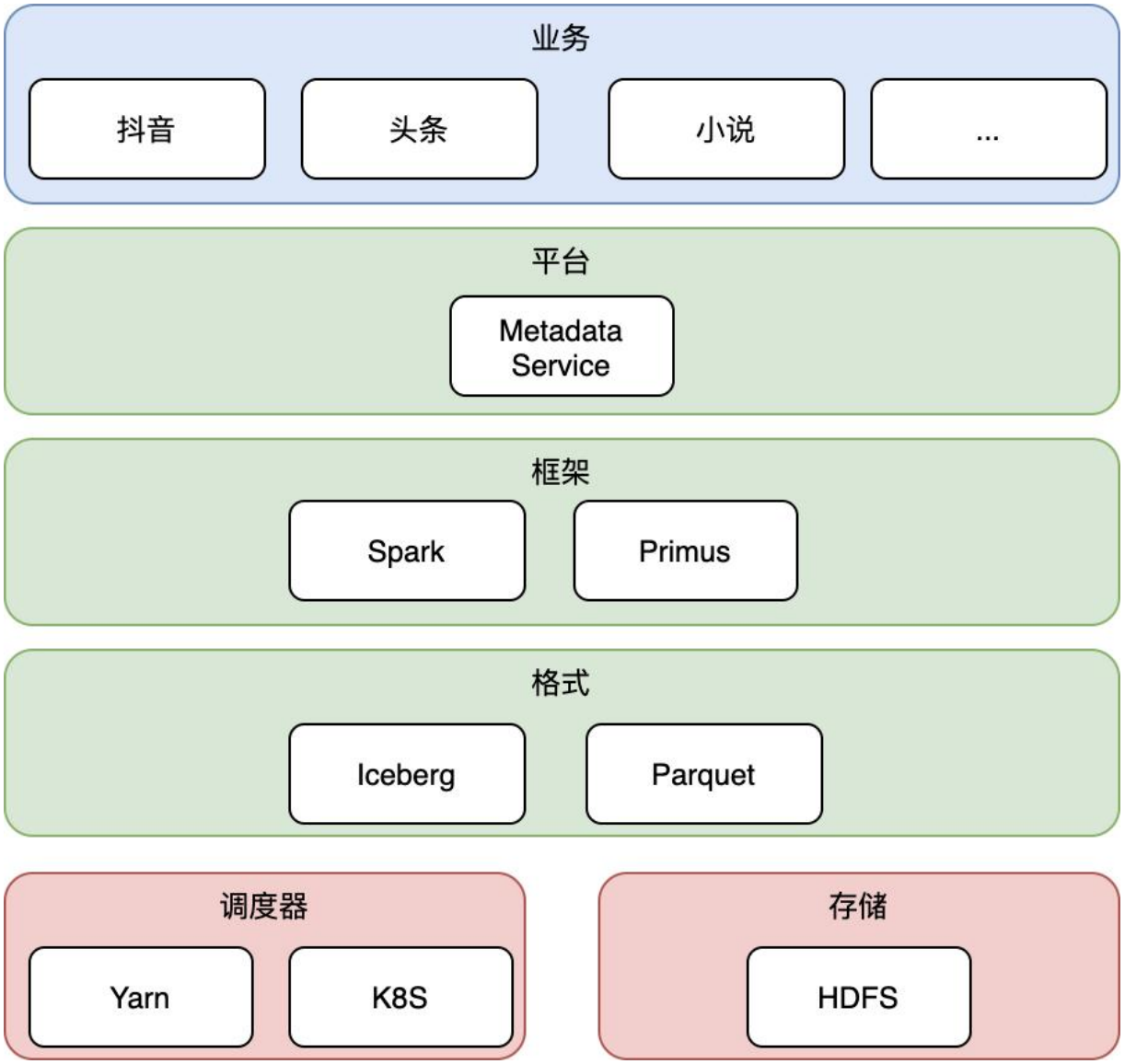
Parquet

1. 基于列存
2. 压缩率高
3. 支持选列

Iceberg

1. 模式演进
2. 特征回填
3. 并发读写

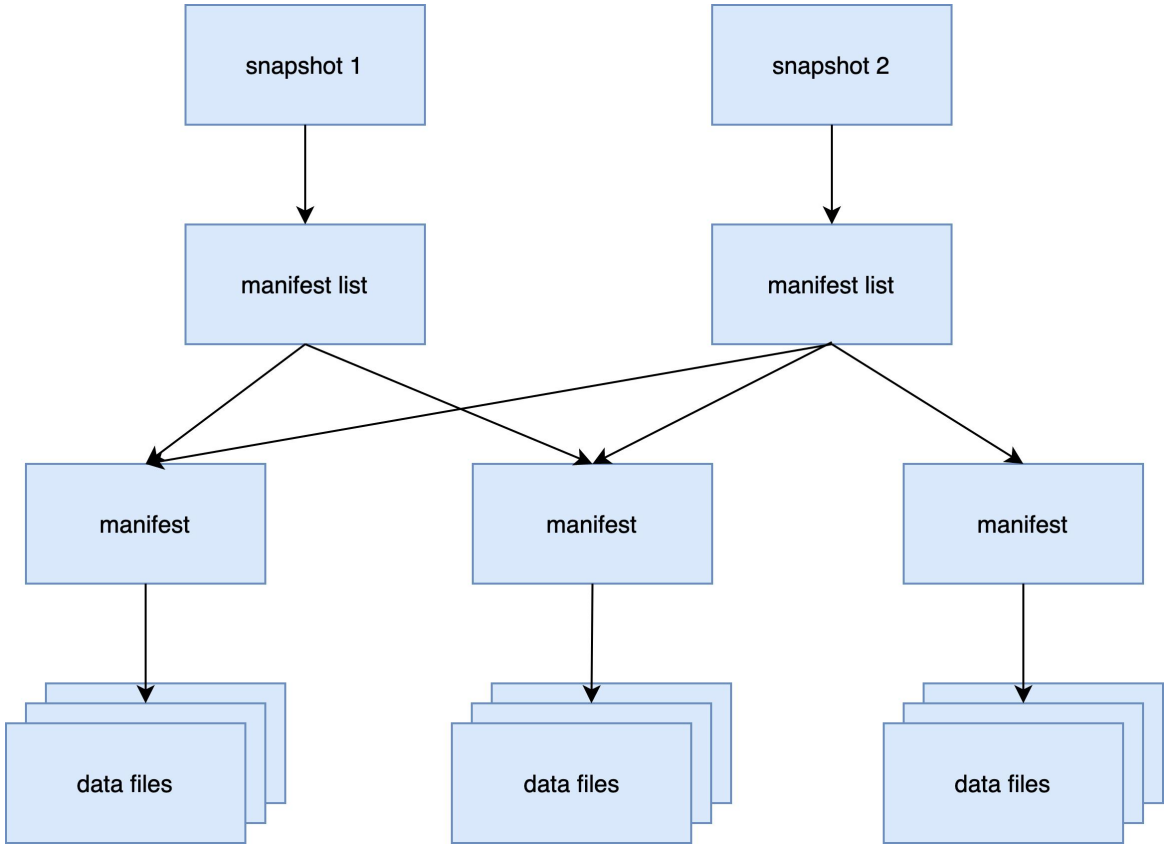
整体架构


业务
抖音
头条
小说
...
平台
Metadata Service
框架
Spark
Primus
格式
Iceberg
Parquet
调度器
Yarn
K8S
存储
HDFS

Iceberg 介绍

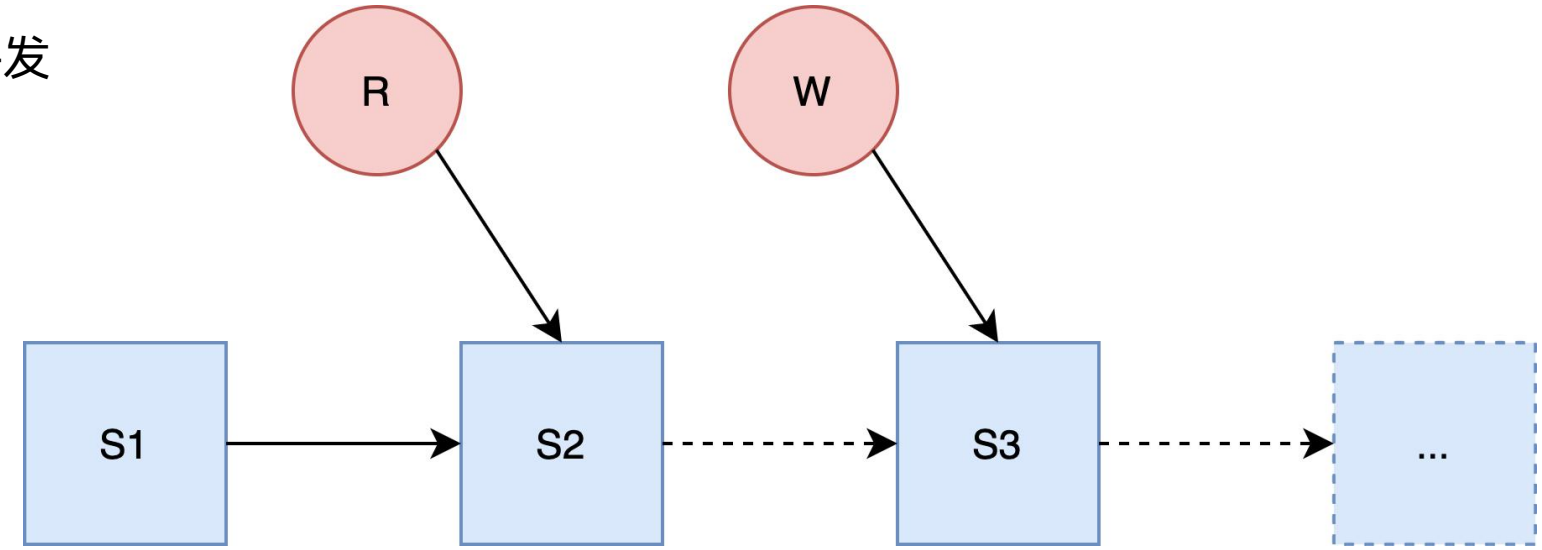
Apache Iceberg is an open table format for huge analytic datasets.

- 模式演进
- 隐藏分区&分区演进
- 支持事务
- MVCC
- 计算存储引擎解耦



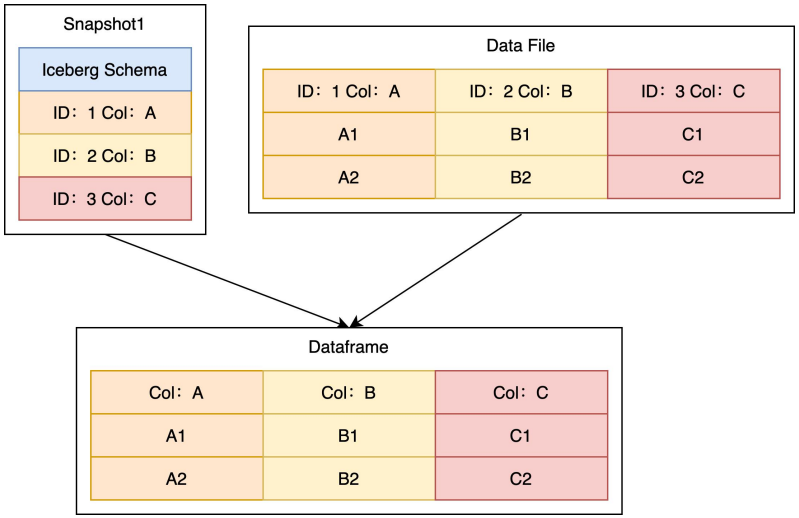
并发读写

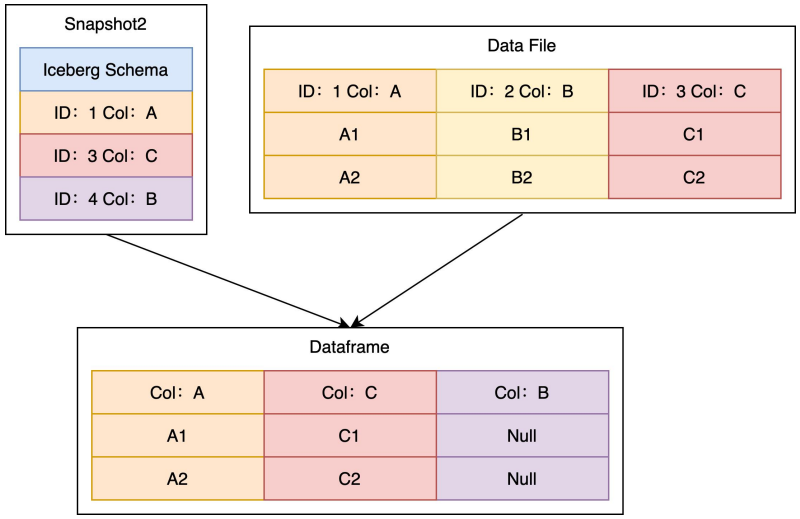
- 快照
- 乐观并发



模式演进

- 基于 id
- 框架解耦
- 完全的模式演进

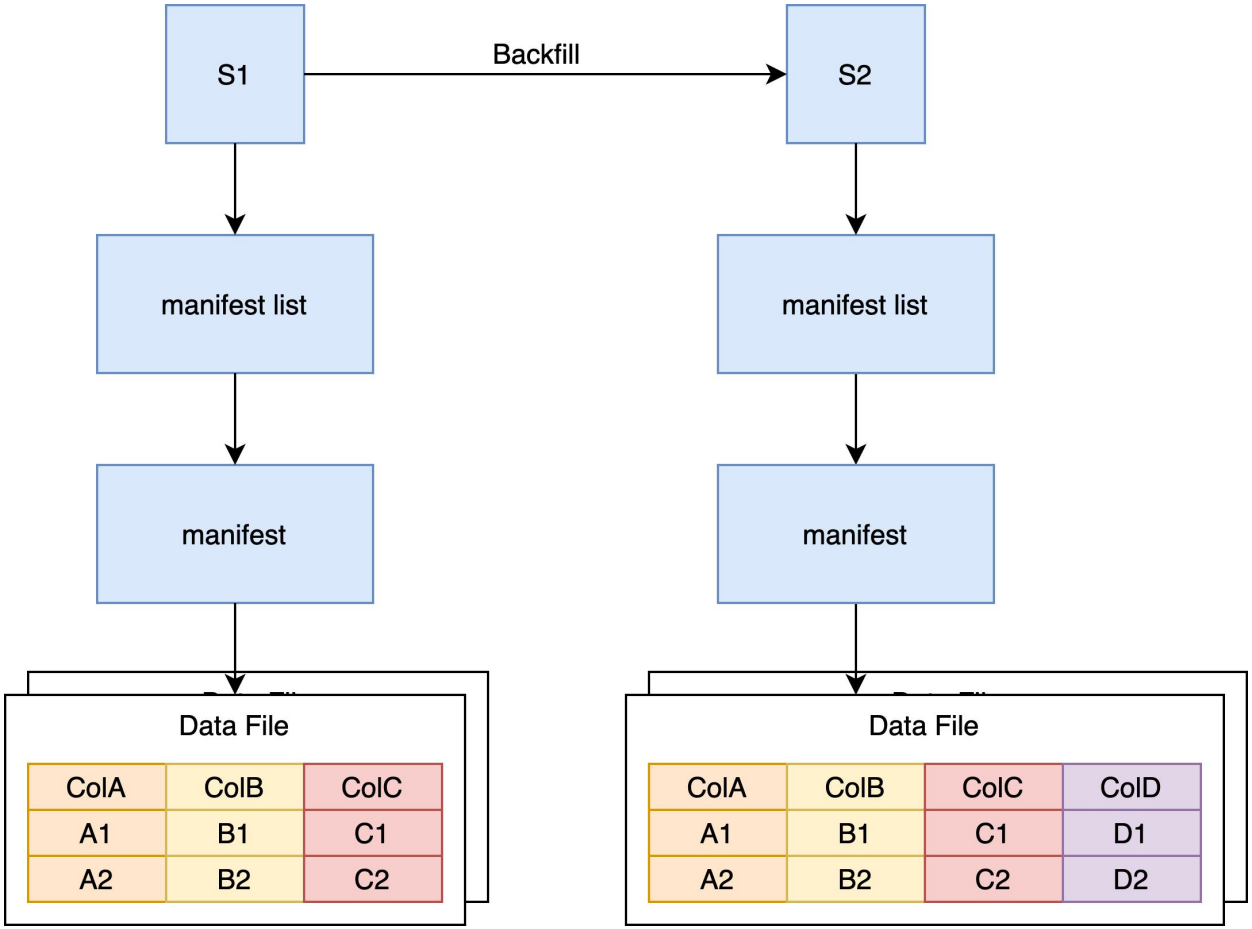




deleteColumn(colB)
addColumn(ColB)

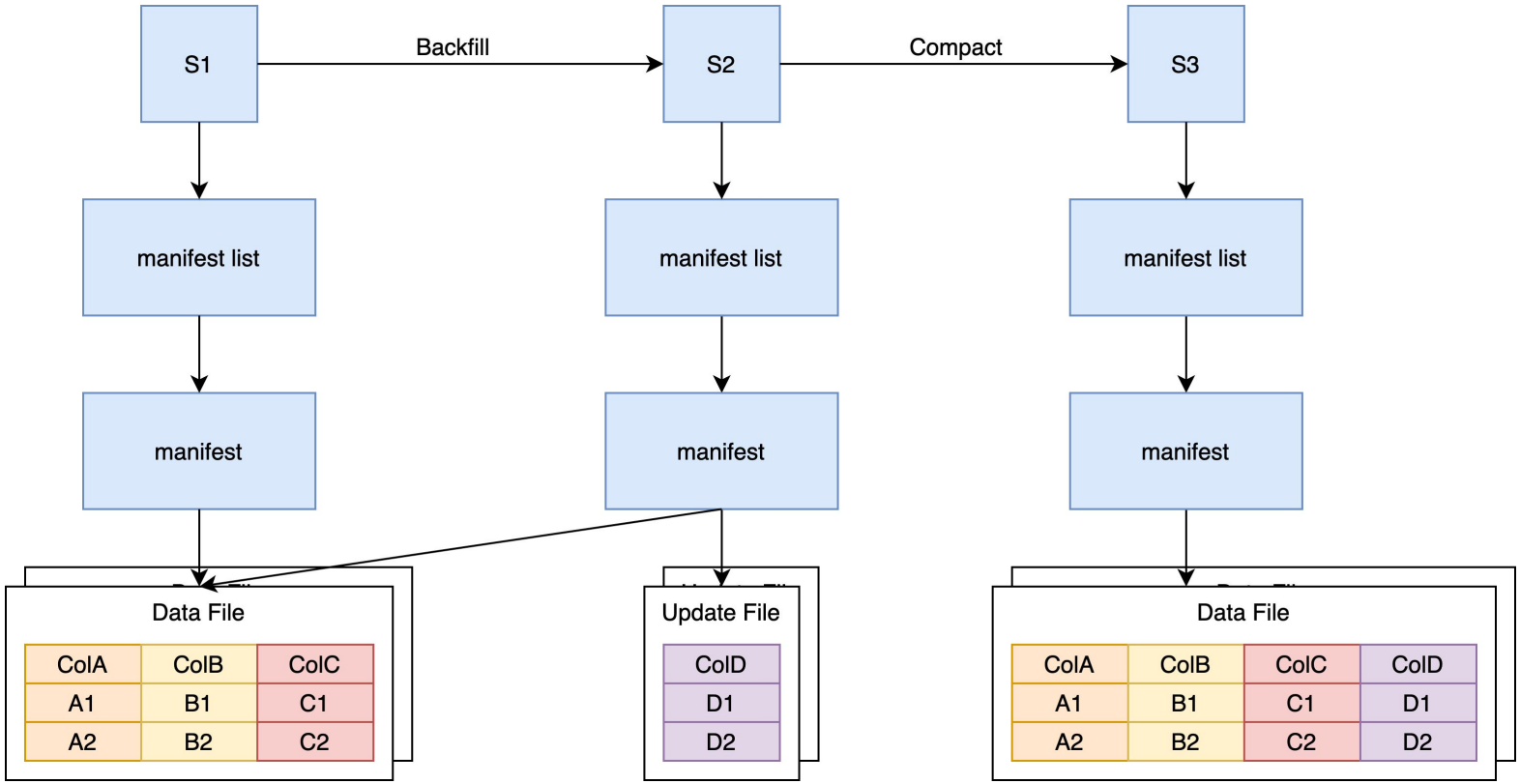
特征回填

- Copy On Write
  - Backfill -> Rewrite datafiles



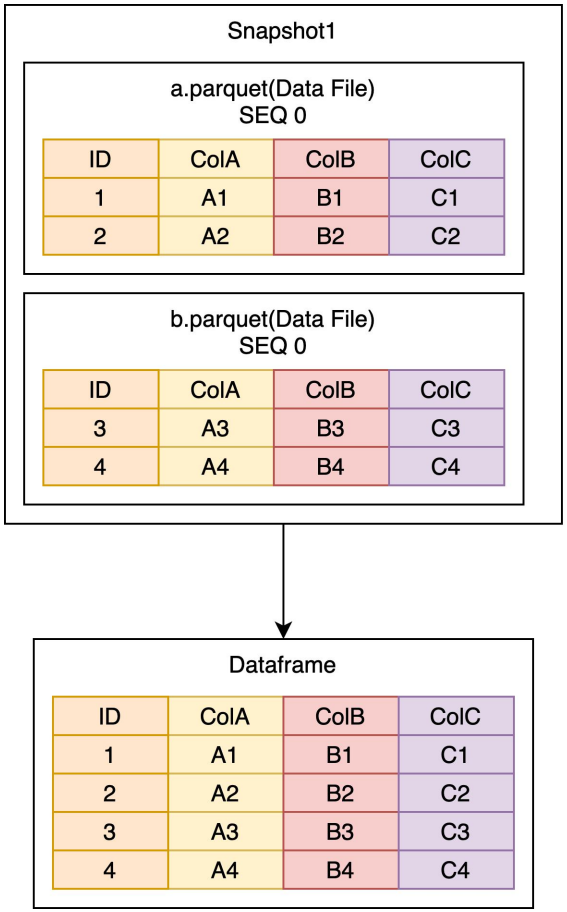
特征回填

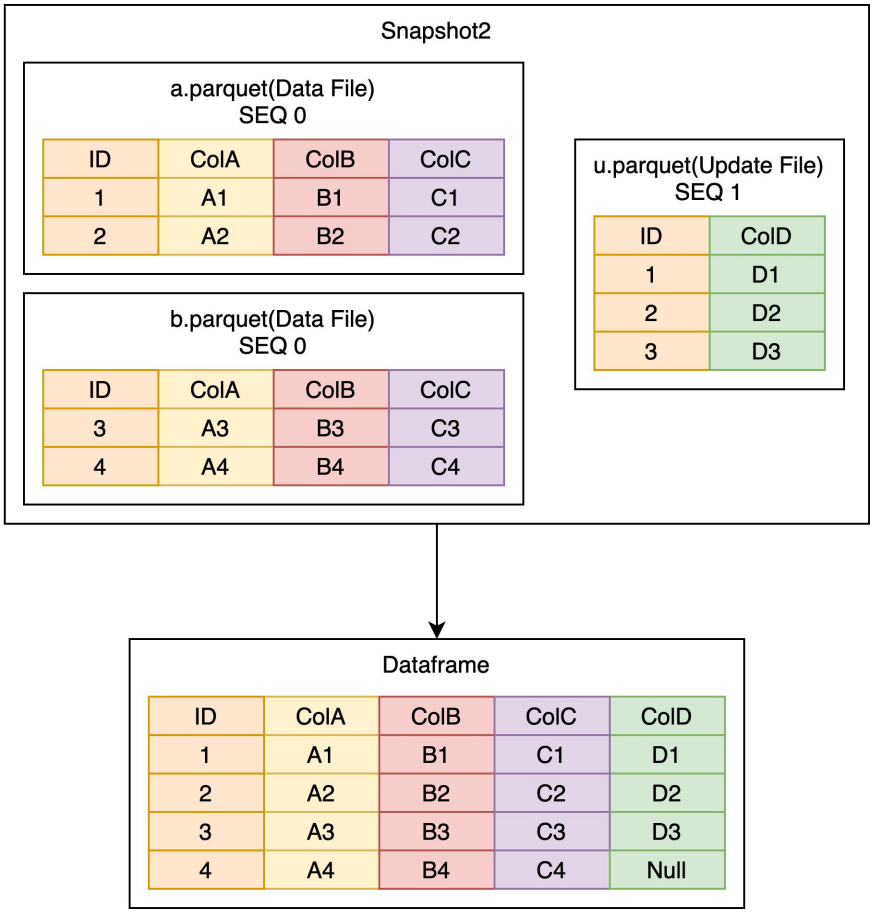
- Merge On Read
  - Backfill -> Append update files
  - Compact -> Rewrite data files and update files



特征回填

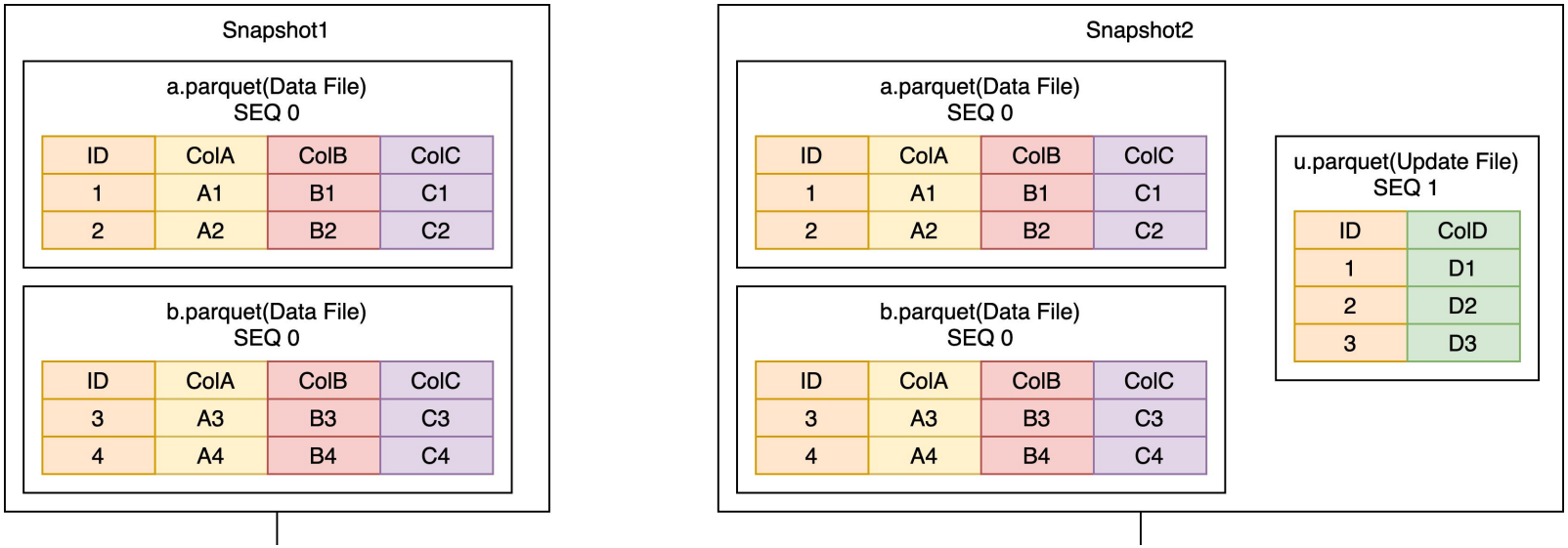
- Merge On Read





特征回填

- Merge On Read



限制

- datafile 和 updatefile 都需要有唯一主键，并按主键排序

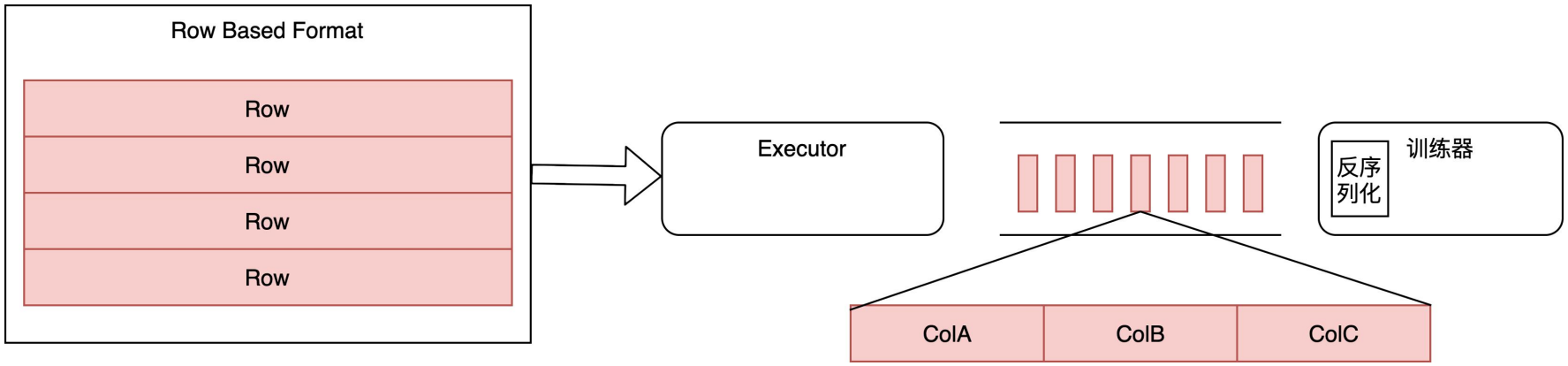
读取

- Projection 决定需要读哪些 datafile 和 updatefile
- 根据 Datafile 进行任务的拆分，根据 datafile 主键的 min - max 决定读哪个 updatefile
- 读取是一个datafile 和 updatefile 多路归并的过程
- SEQ 决定合并的顺序

特征回填

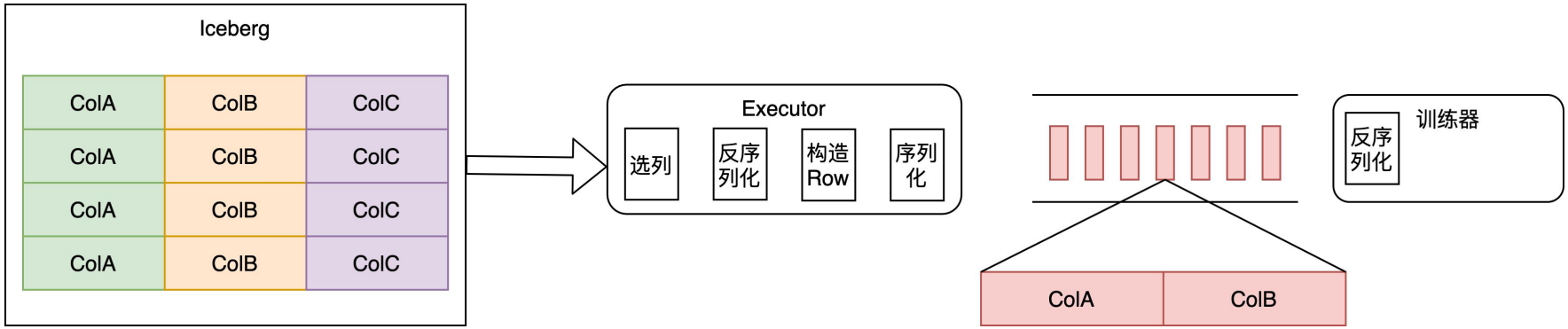
Copy on Write	Merge On Read
读写放大严重	没有读写放大
存储空间浪费	节省存储空间
读取逻辑简单	读取逻辑复杂
写入耗费更多资源	写入耗费更少资源
读取无需额外计算资源	绝大部分场景读取不需额外资源， 少部分场景需要额外资源

训练优化



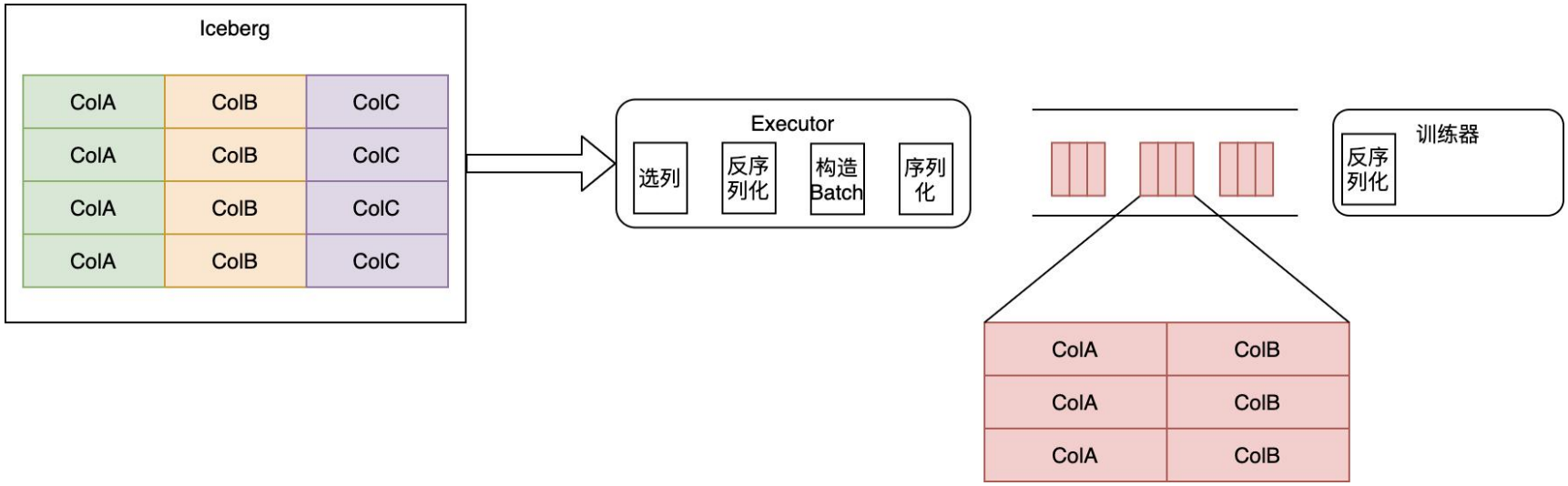
- 按行传输
- 框架透传数据给训练器

训练优化



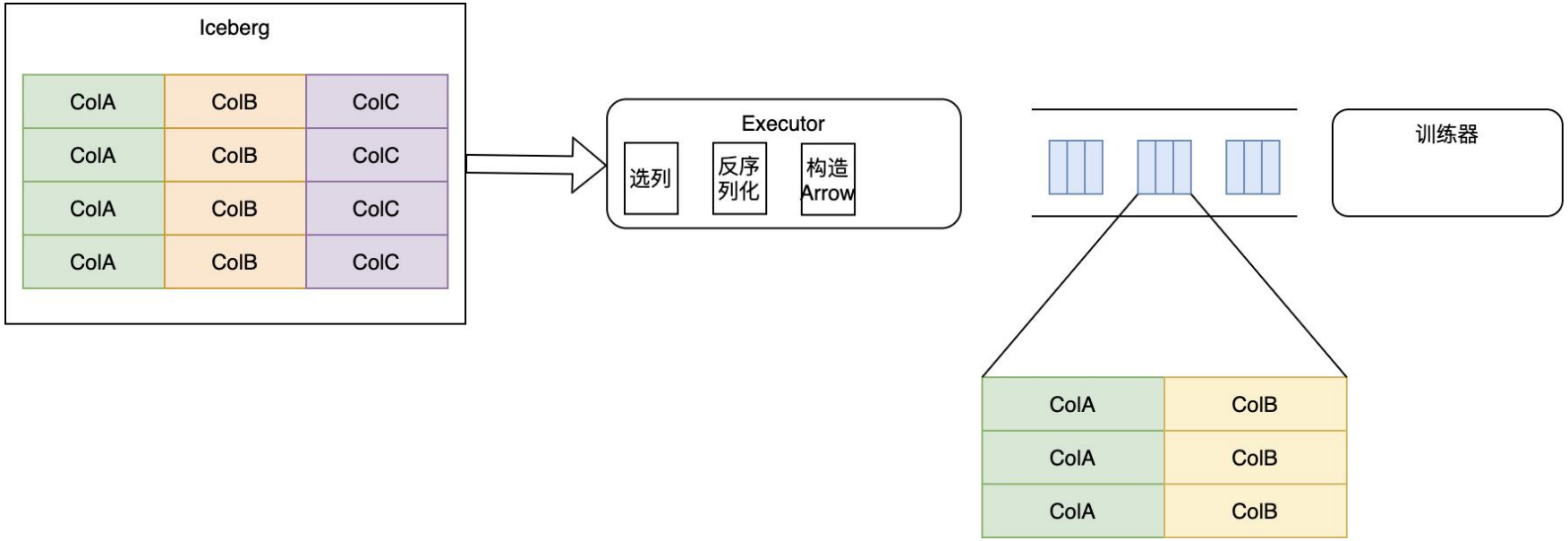
- 选列降低 IO
- 额外增加了序列化反序列化以及构造 Row 的开销，训练速度变慢，资源增加

训练优化



- 向量化读取提升训练速度，降低部分资源消耗

训练优化



- 基于 Arrow 的数据传输，降低序列化和反序列化开销，进一步提升训练速度，降低资源消耗

#3 收益

收益

#1

离线特征工程能力

#2

存储成本**降低40%**以上

#3

降低训练开销

- CPU **降低13%**
- 网络 IO **降低40%**

#4 未来规划

未来规划

支持 Upsert

物化视图

Data Skipping

基于 arrow 的
数据预处理

THANKS.

麦思博(msup)有限公司是一家面向技术型企业的培训咨询机构，携手2000余位中外客座导师，服务于技术团队的能力提升、软件工程效能和产品创新迭代，超过3000余家企业续约学习，是科技领域占有率第1的客座导师品牌，msup以整合全球领先经验实践为己任，为中国产业快速发展提供智库。

高可用架构公众号主要关注互联网架构及高可用、可扩展及高性能领域的知识传播。订阅用户覆盖主流互联网及软件领域系统架构技术从业人员。

高可用架构系列社群是一个社区组织，其精神是“分享+交流”，提倡社区的人人参与，同时从社区获得高质量的内容。