Apache Flink

Iceberg 和对象存储构建数据湖方案

孙伟/戴尔科技集团 高级软件研发经理 2021-4-17

Apache Flink x Iceberg Meetup · 上海站

CONTENT

目录 >>

#1 数据湖和Iceberg简介

数据湖生态

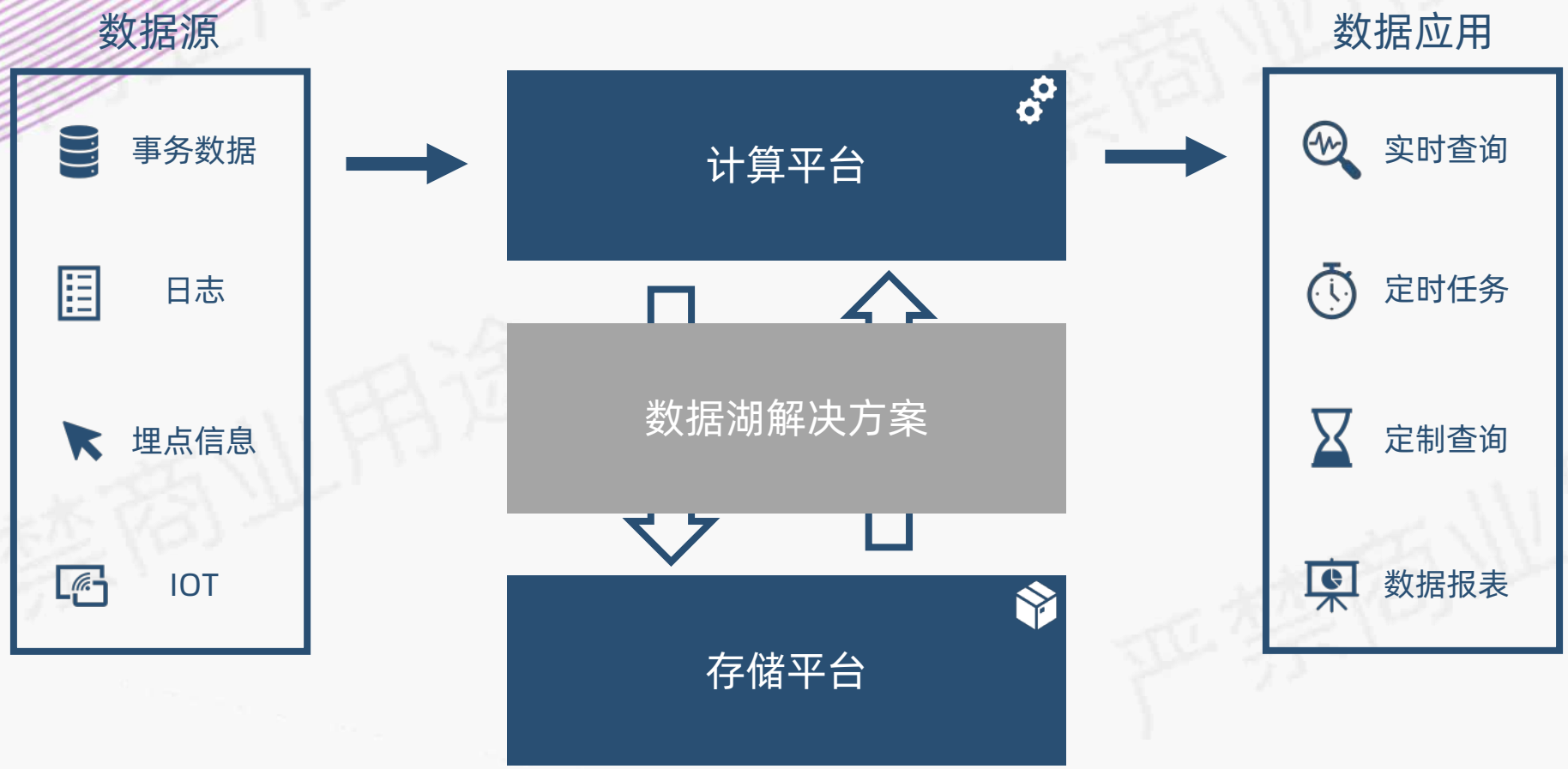
多计算引擎支持

高效统一的元数据管理

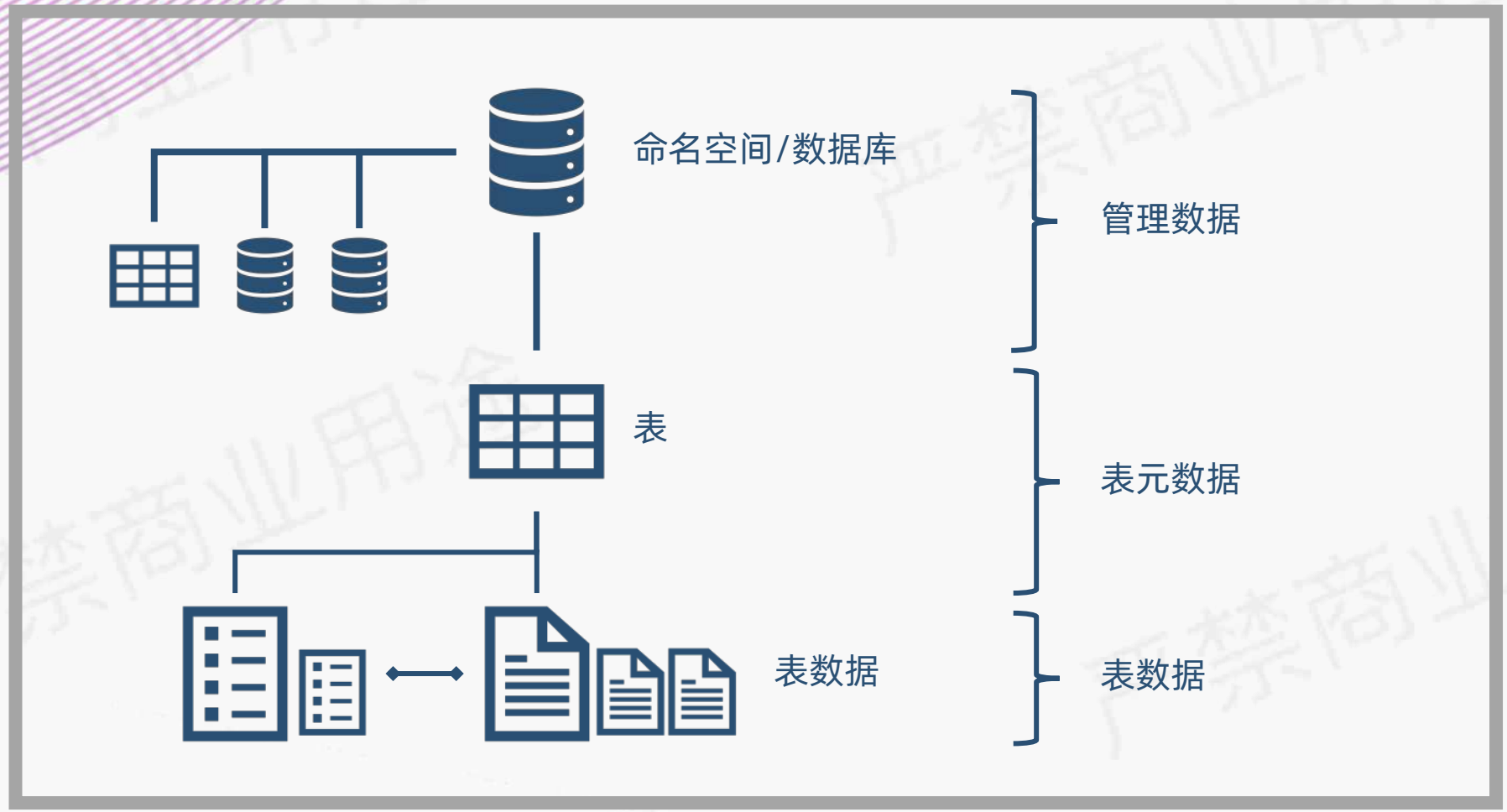
丰富的数据类型

海量数据统一存储

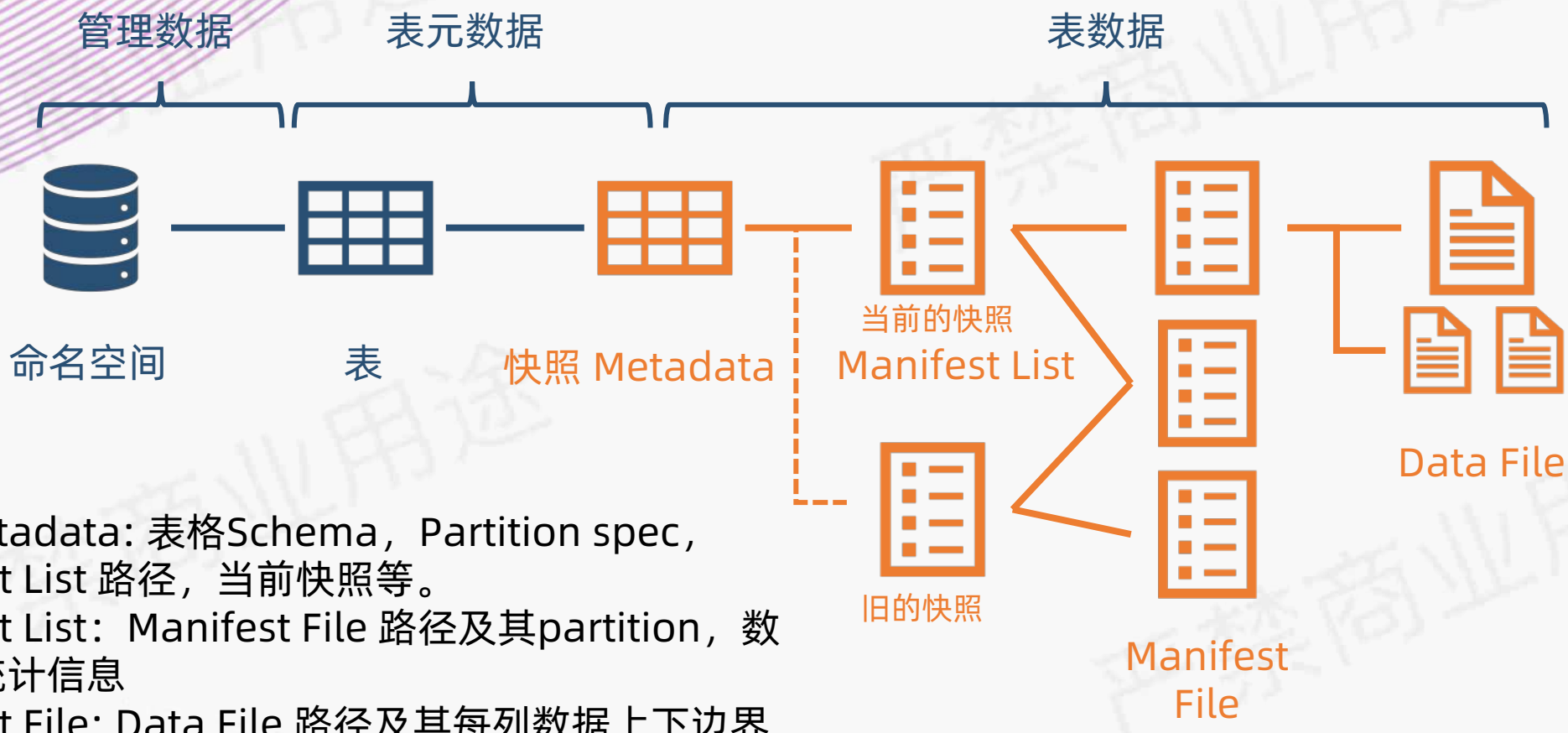
结构化数据在数据湖上的应用场景



结构化数据在数据湖上的典型解决方案

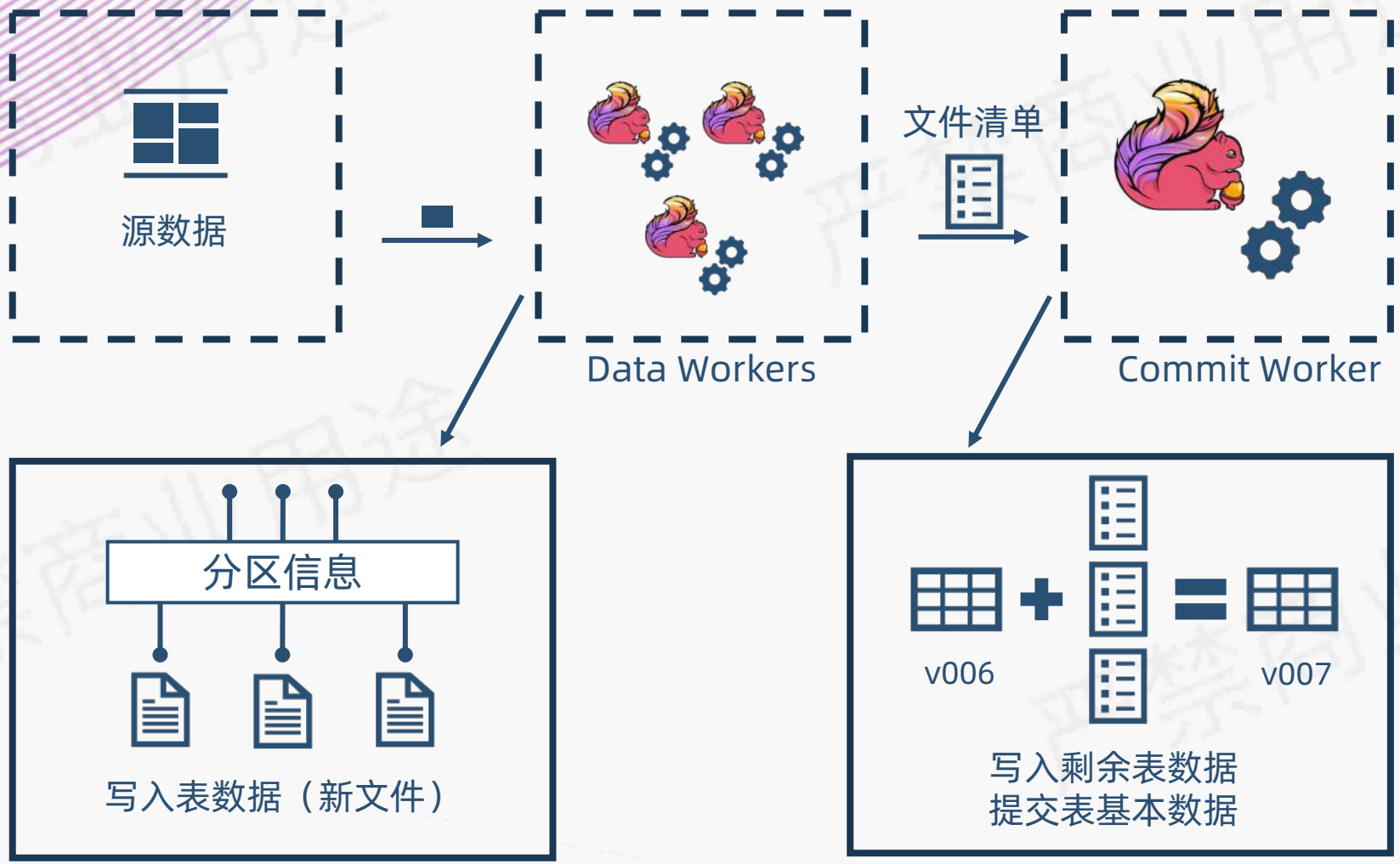


Iceberg 表数据组织架构

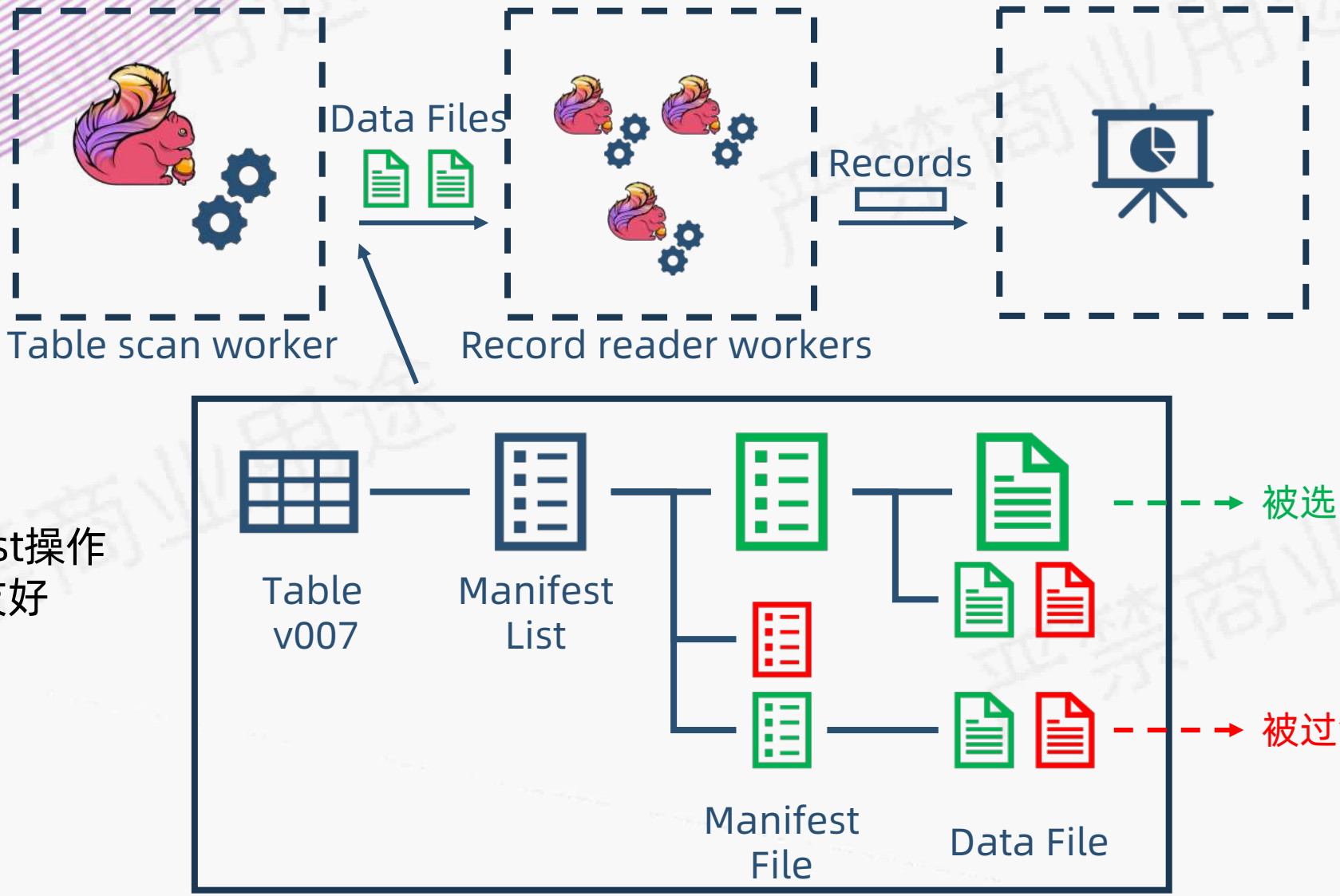


- 快照 Metadata: 表格Schema，Partition spec，Manifest List 路径，当前快照等。
- Manifest List：Manifest File 路径及其partition，数据文件统计信息
- Manifest File: Data File 路径及其每列数据上下边界
- Data File: 实际表内容数据，以Parquet, ORC, Avro 等格式组织

Iceberg 写入流程



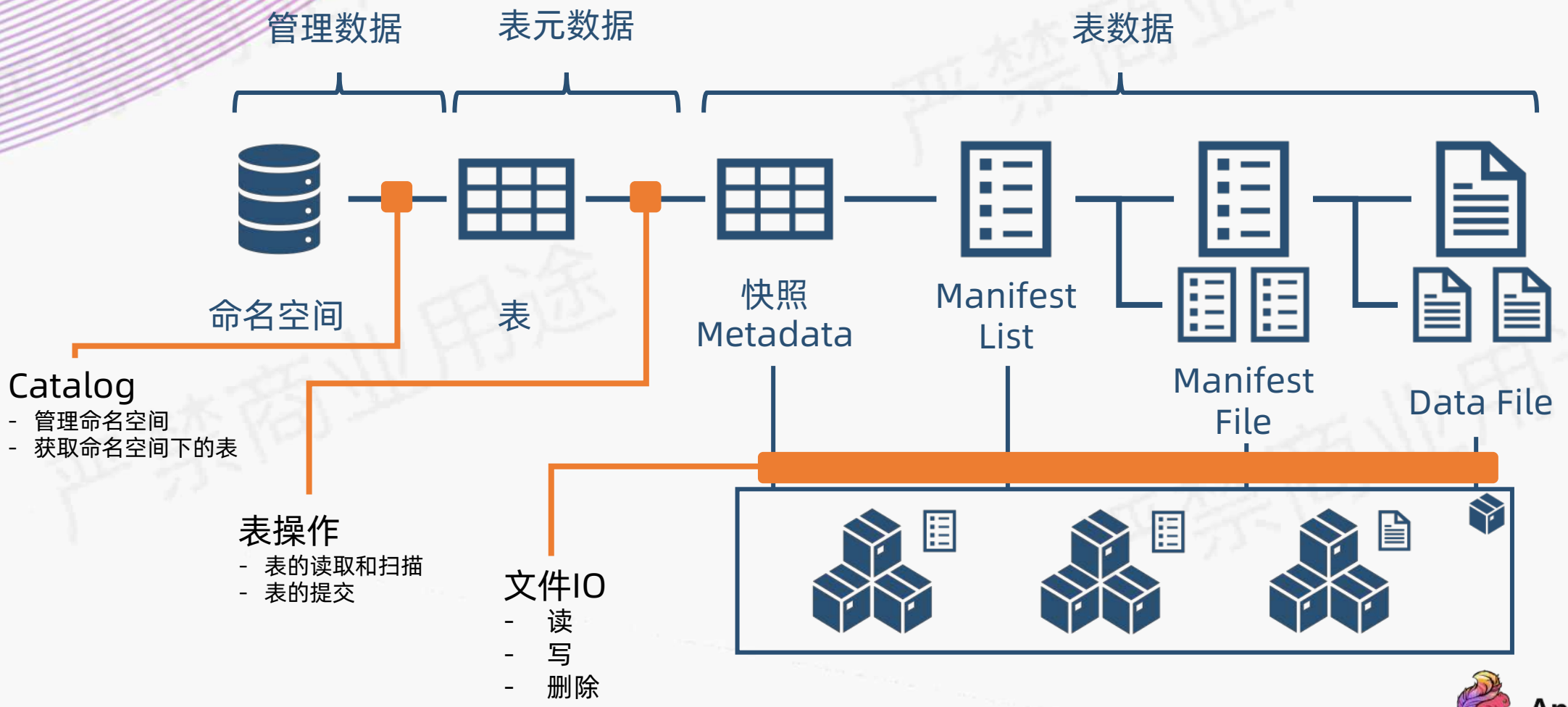
Iceberg 查询流程



- 无耗时的list操作
- 对象存储友好

Iceberg Catalog功能一览

良好的抽象来对接数据存储和元数据管理



#2 对象存储支撑Iceberg数据湖

当前Iceberg Catalog实现

Catalog实现	数据文件I/O	元数据管理
AWS	Amazon S3	Amazon Glue
Apache Hadoop	Apache HDFS	Apache HDFS
Apache Hive	Apache HDFS / S3A	Hive Metadata Store
Project Nessie	Apache HDFS / S3A	Project Nessie
对象存储（缺失）	S3 Compatible Storage	S3 Compatible Storage

对象存储与HDFS的比较

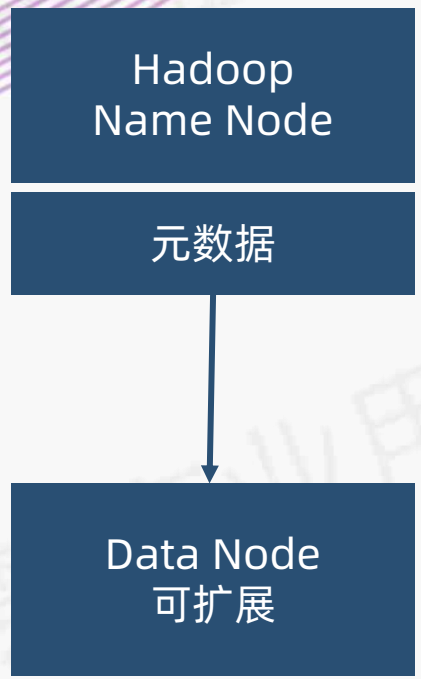
对象存储

- 集群扩展性
- 小文件友好
- 多站点部署
- 低存储开销

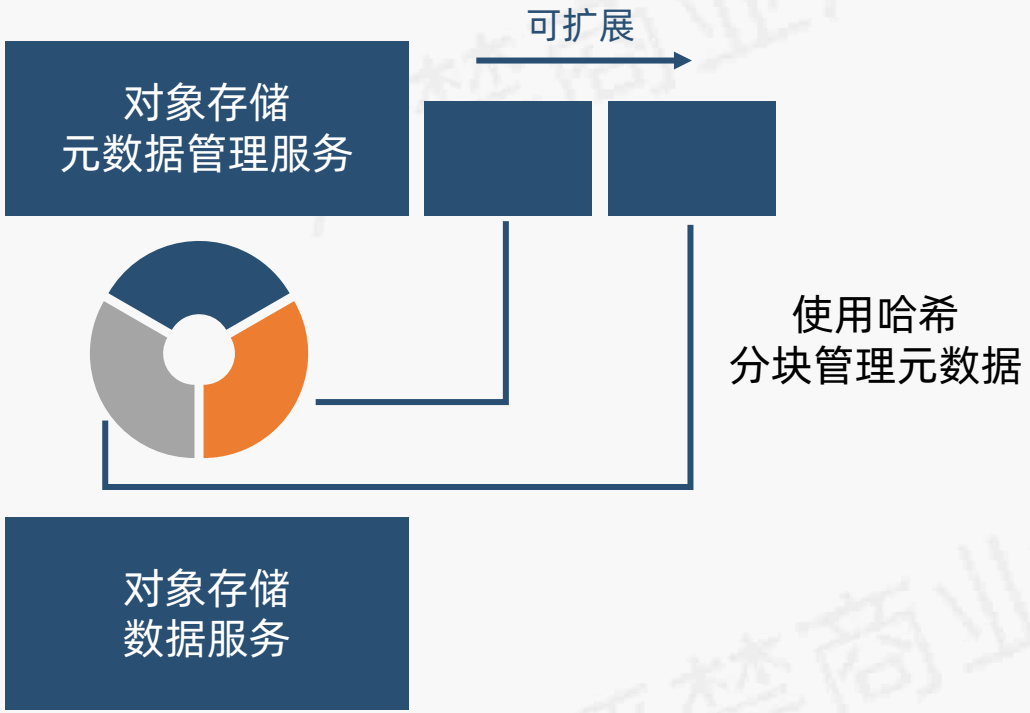
HDFS

- 追加上传
- 原子性rename

比较之：集群扩展性



- Name Node 单节点能力有限
- Name Node 无横向扩展能力



- 分布式管理元数据
- 极端情况下可重哈希（rehash）来横向扩展

比较之：小文件友好

HDFS

- 小文件存储受限于Name Node 内存等资源
- Archive等方法增加额外复杂性
- 小文件TPS 受限于Name Node处理能力

对象存储

- 分布式的元数据存储和管理
- 单节点海量小文件
- 多介质，分层加速

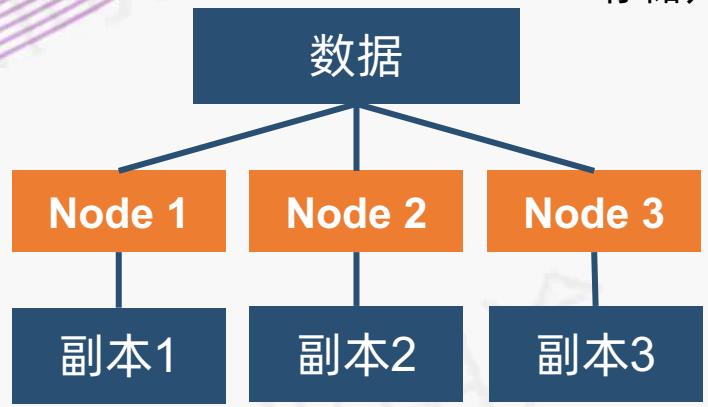
比较之：多站点部署

对象存储支持多站点部署

- 全局命名空间
- 支持丰富的规则配置

比较之：低存储开销

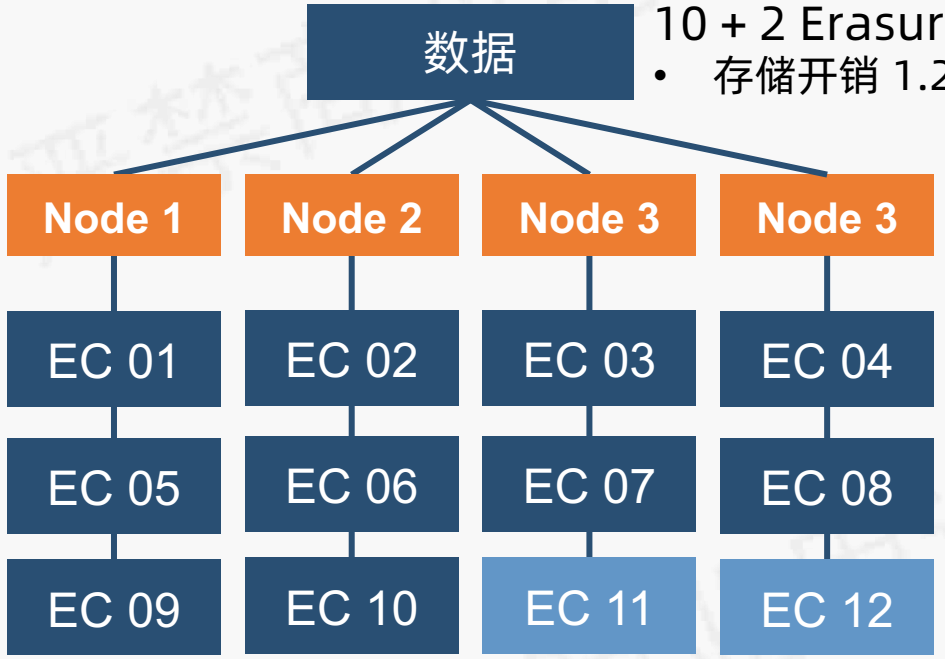
3副本
- 存储开销 3×



HDFS 默认3副本

HDFS Erasure Coding（EC）
- 基于文件做EC，小文件时EC开销大
- EC文件无法支持append, hflush, hsync

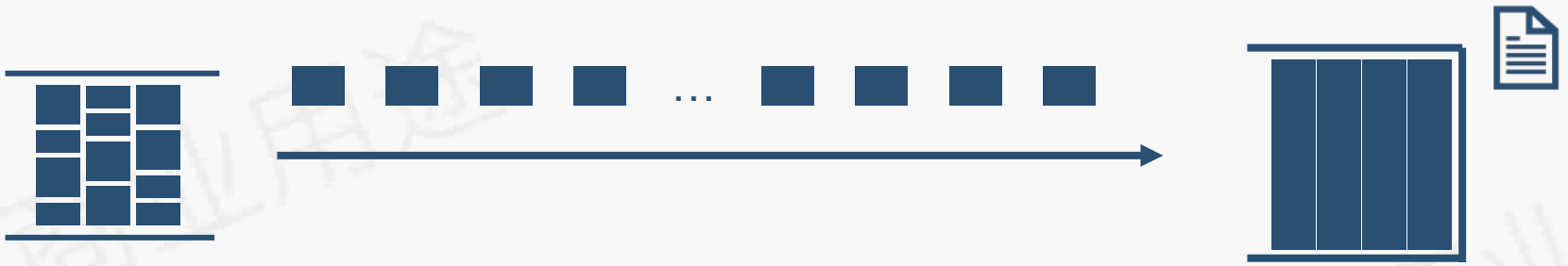
10 + 2 Erasure Coding
- 存储开销 1.2×



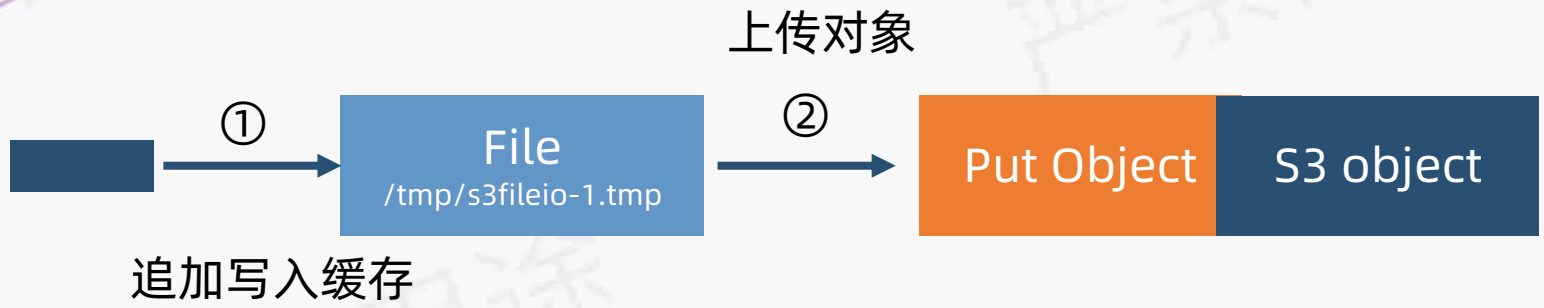
对象存储
- 原生支持EC
- 合并小文件，对块做EC

对象存储的挑战：数据的追加上传

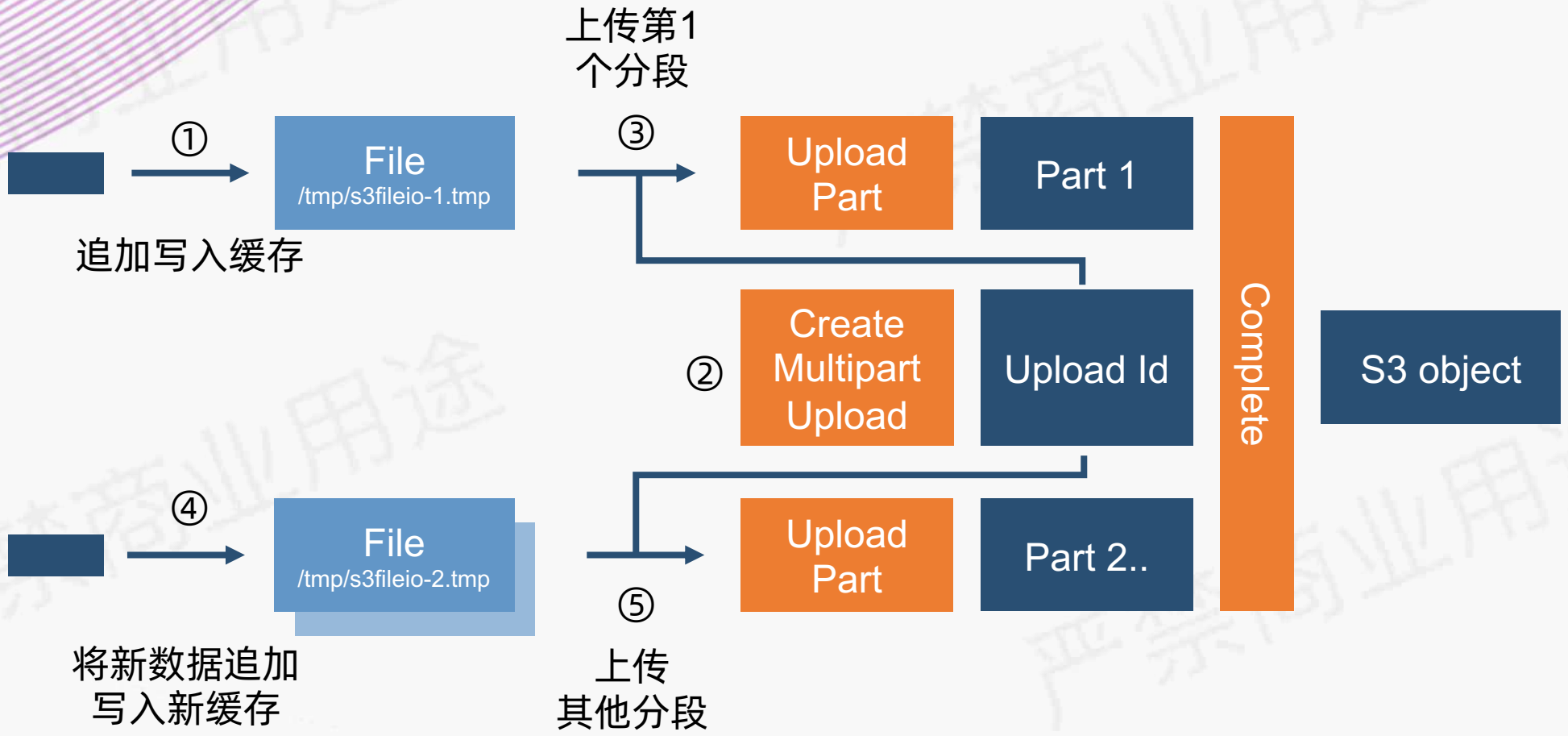
- 在 S3协议中，对象在上传时需要提供大小。



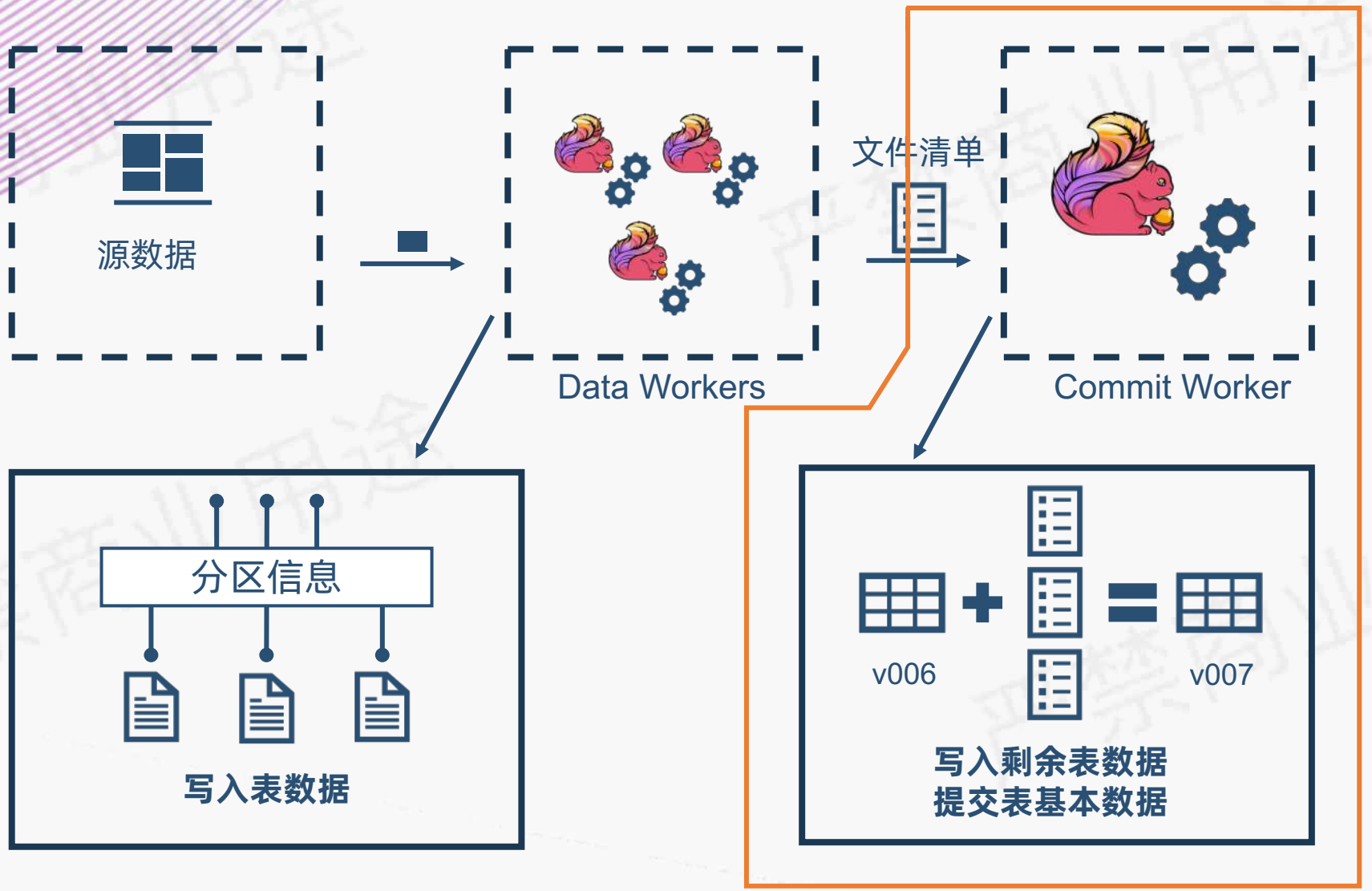
S3 Catalog数据追加上传 - 小文件缓存本地/内存



S3 Catalog数据追加上传 - MPU分段上传大文件

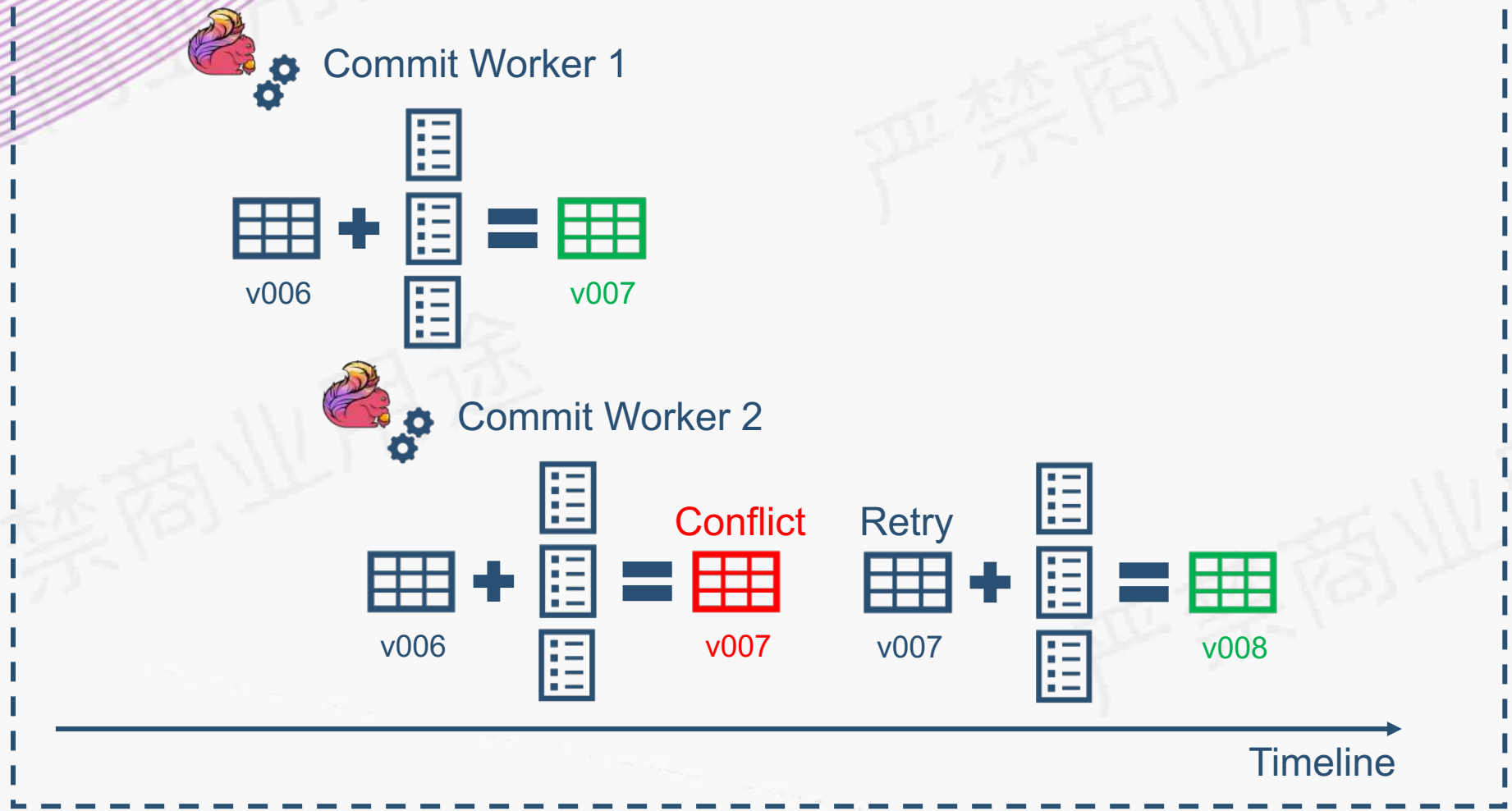


对象存储的挑战：原子提交

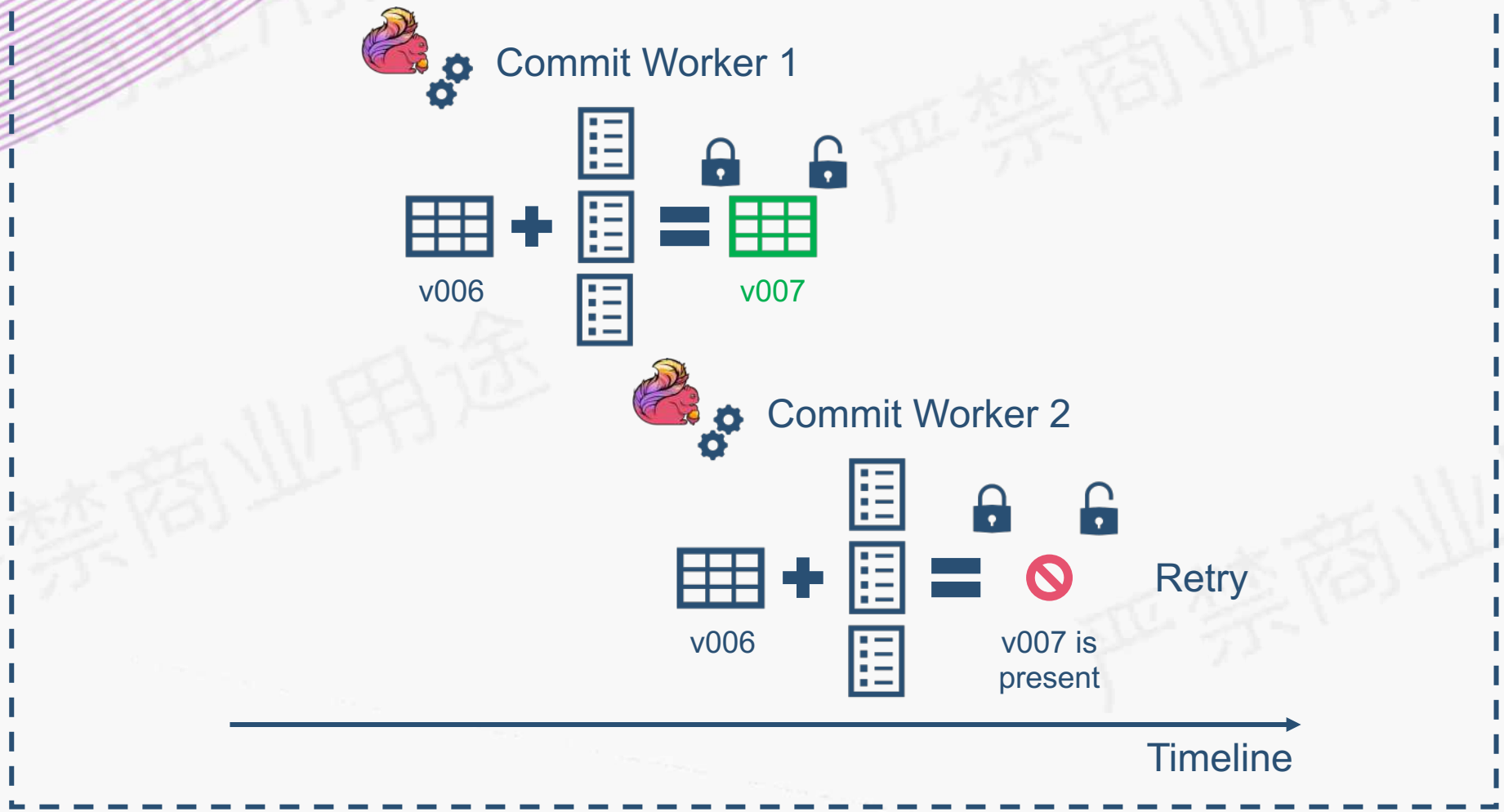


- 类似i=i+1

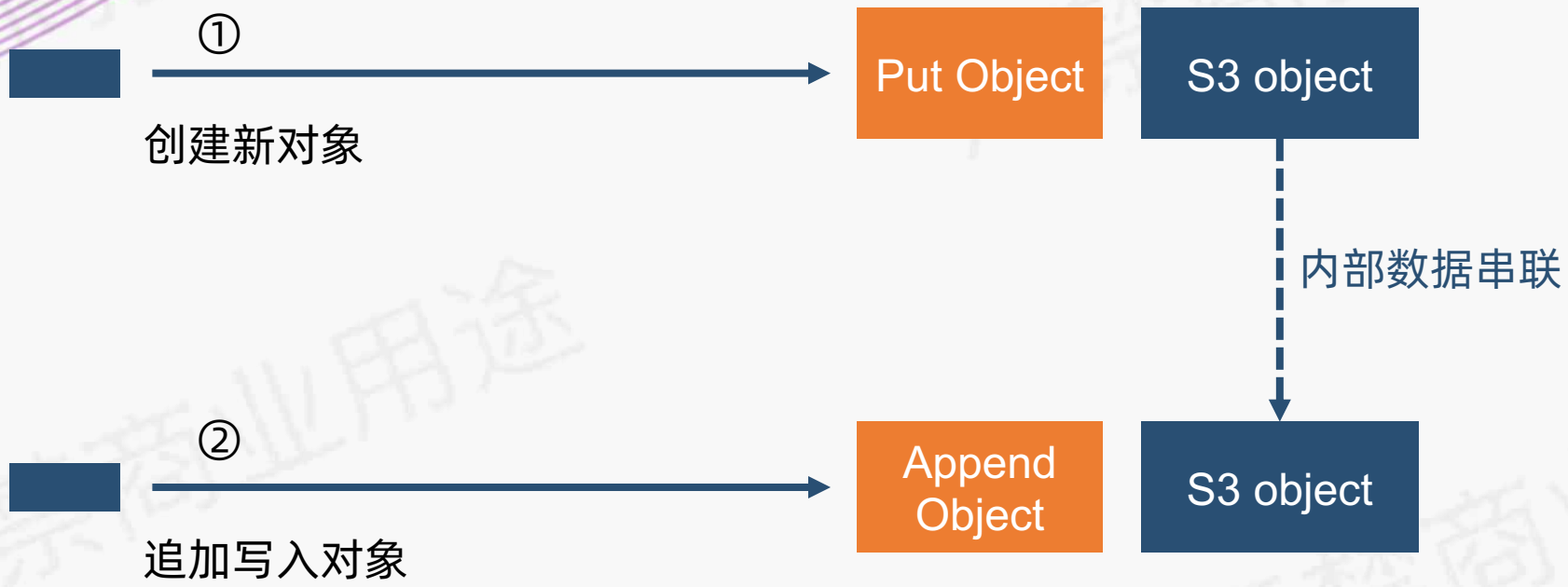
并发提交元信息的场景



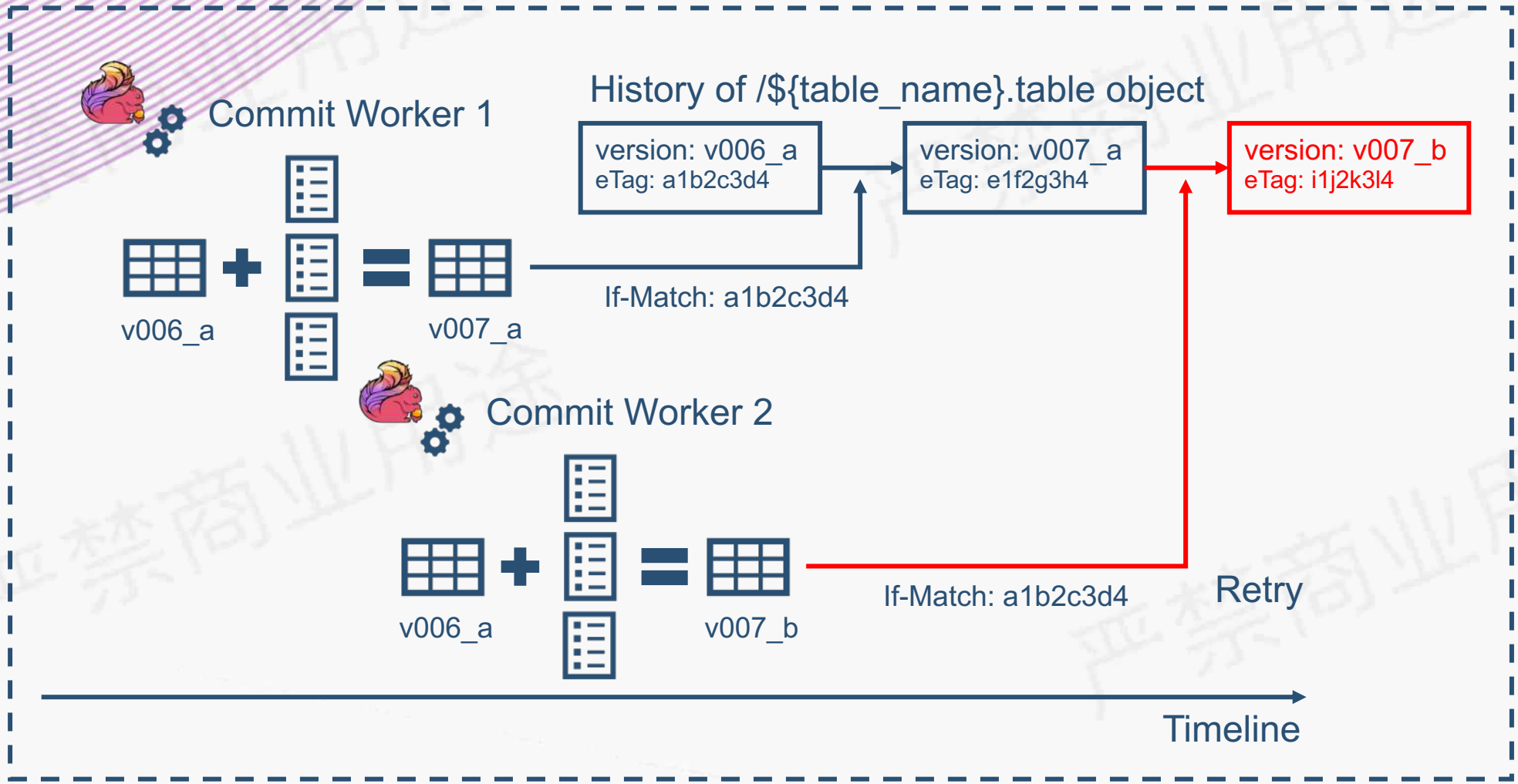
S3 Catalog: 使用锁机制处理并发提交



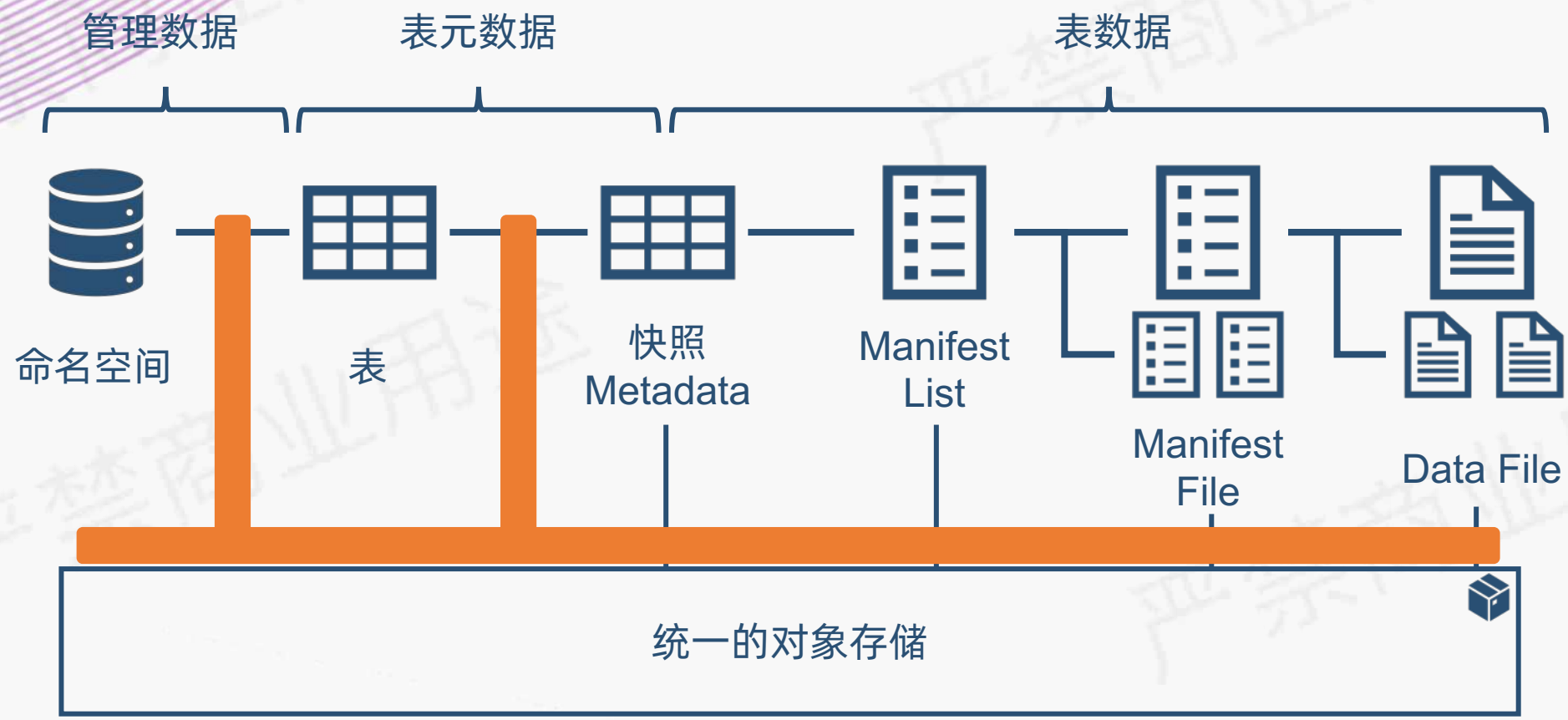
Dell EMC ECS的数据追加上传



Dell EMC ECS在并发提交下的解决方案

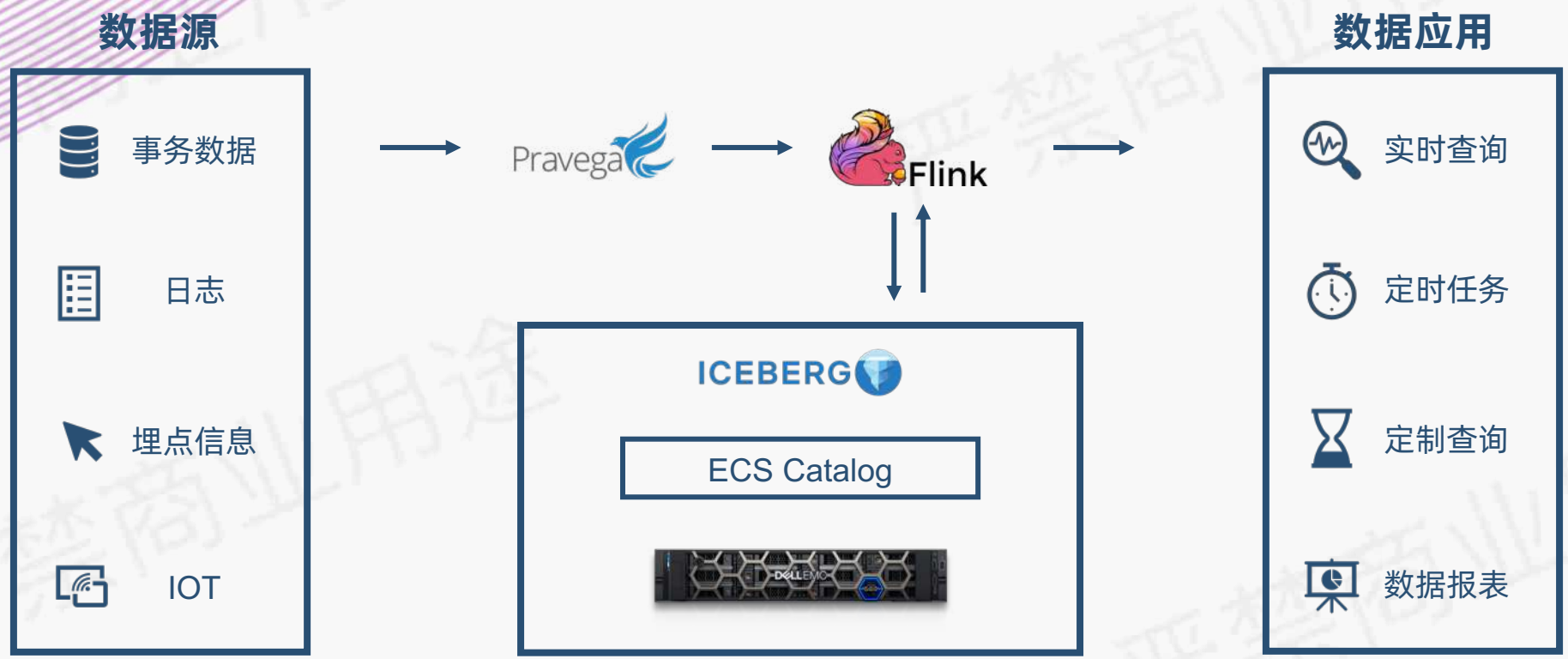


S3 Catalog - 统一存储的数据湖



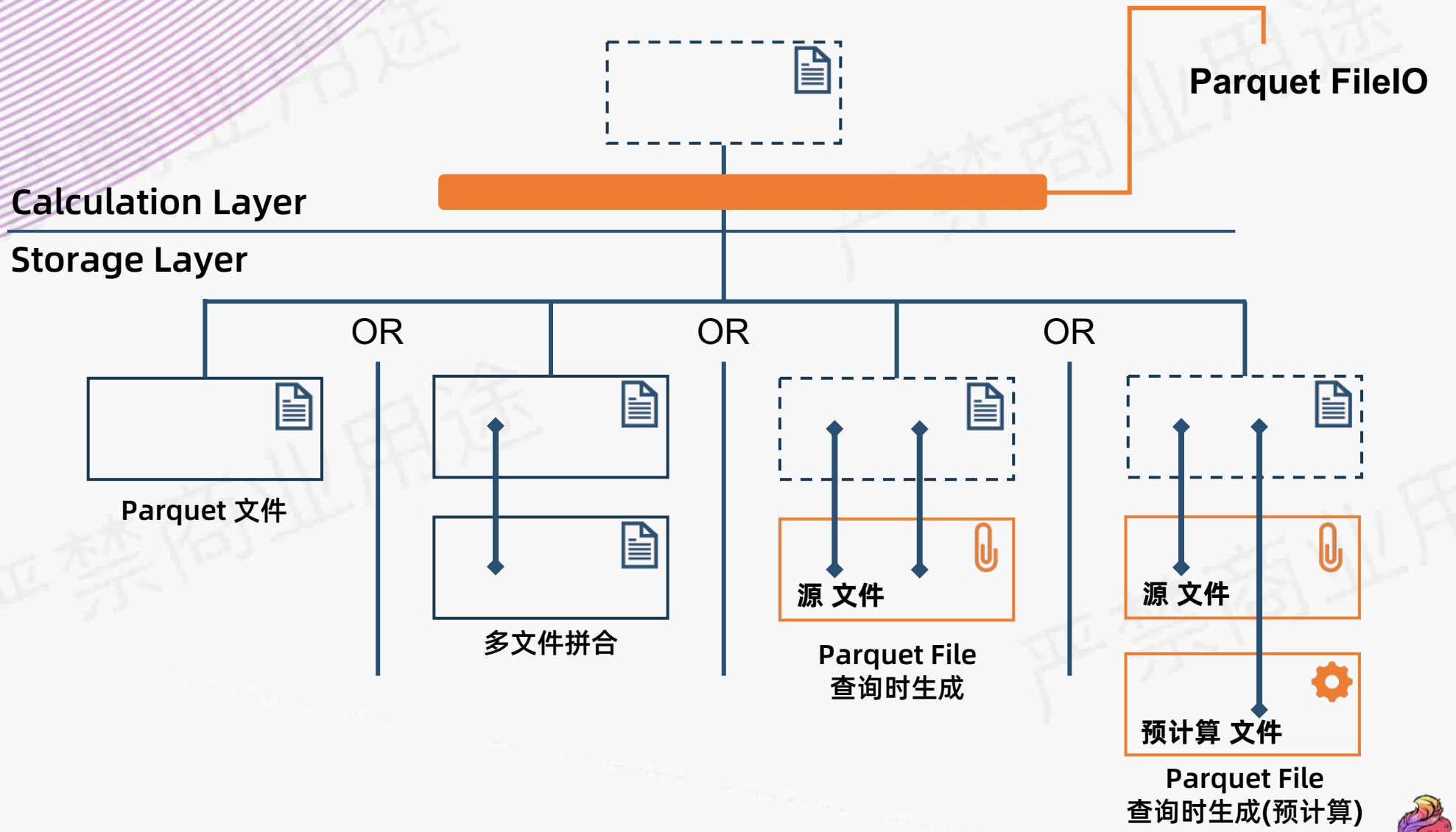
#3 方案演示

演示方案框架图



#4 存储优化的一些思考

Iceberg下多种结构化存储方案



Apache Flink

Thanks

Apache Flink x Iceberg Meetup · 上海站