

湖仓无界： 使用 Apache Doris 构建 Lakehouse

衣国垒

Apache Doris PMC

分享嘉宾



衣国垒

Apache Doris PMC

目录

01 为什么需要 Lakehouse

02 湖仓一体转型之痛

03 基于 Apache Doris 的湖仓一体解决方案

01

为什么需要 Lakehouse

数据一体化需求

实时与历史数据一体化

- 通过统一的存储架构来管理实时和历史数据，确保数据格式和存储方式的一致性，以便于数据的高效读取和处理。
- 实现数据的一致性管理，确保实时数据和历史数据在更新、查询时的一致性，避免数据冗余和冲突。

流批处理一体化

- 支持批处理和流处理的混合处理能力，能够同时处理实时数据流和历史数据批次，确保数据分析的及时性和全面性。
- 提供统一的数据访问接口和优化的查询引擎，支持流、批数据的快速检索和分析，提升数据访问效率。

功能一体化需求

统一元数据管理

- 实现不同数据源的统一元数据表示。
- 集中式的数据权限、质量、血缘管理。

系统开放性

- 通过 Parquet/ORC 等开放数据格式提供数据读写开放性。
- 通过对多计算引擎的支持，同一份数据支撑不同的分析负载。

半结构化数据

- 支持JSON等半结构数据得高效存储。
- 通过倒排索引等功能提供对半结构化数据的高效访问。

高性能与低成本

- 利用数据仓库的高性能查询引擎加速数据分析。
- 借助存算分离架构、对象存储等降低存储成本。

02

湖仓一体转型之痛

企业湖仓一体建设的痛点

数据统一之痛

- 各种数据源
- 数据表示不统一
- 数据权限不统一
- 数据查询不统一

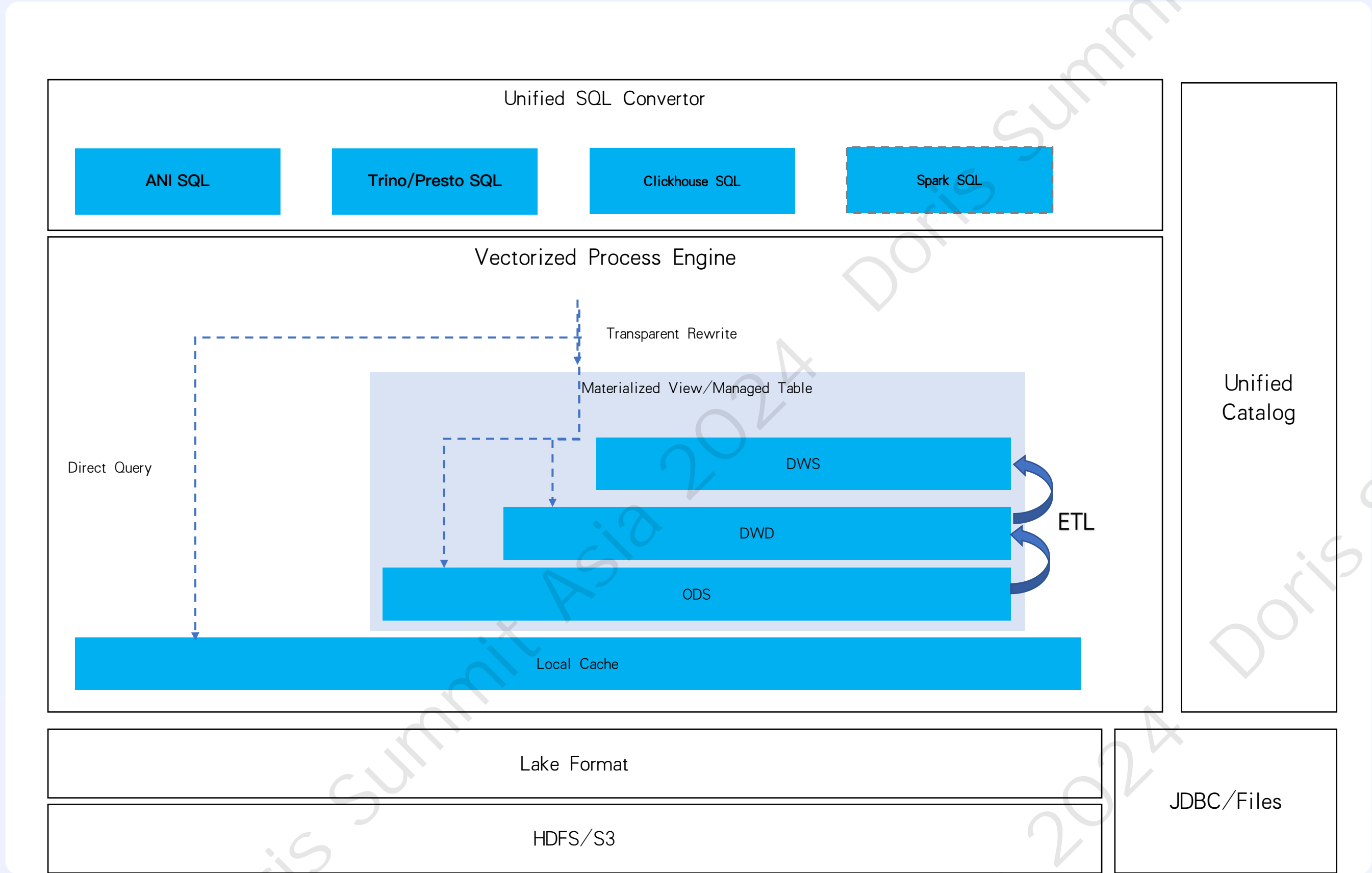
湖仓割裂之痛

- 湖和仓的数据流转成本高
- 冗余存储成本高
- 使用体验不一致

03

基于 Apache Doris 的湖仓一体方案

数据无界



丰富的数据源

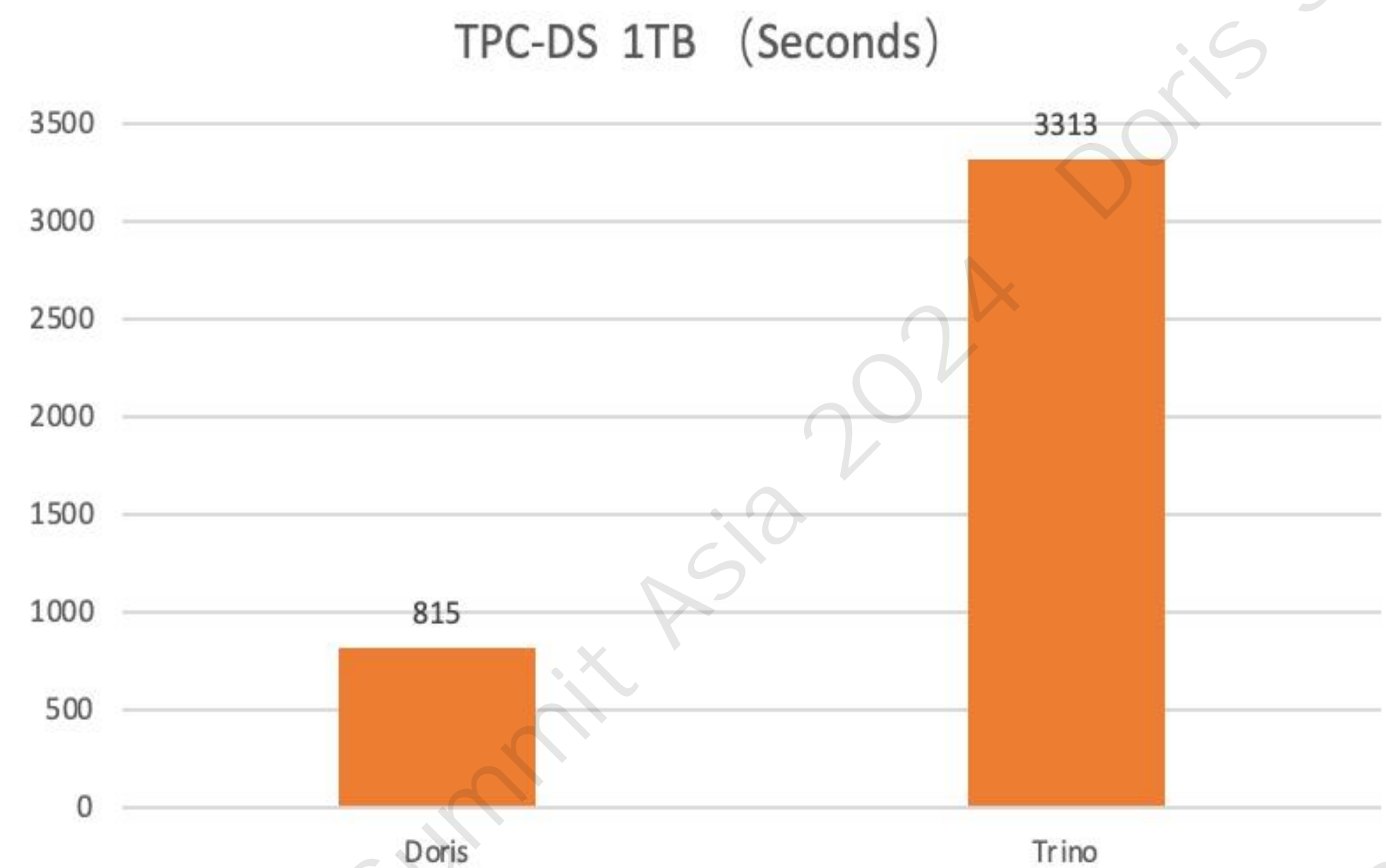
湖仓系统	数据库	文件格式	元数据服务
•Hive	•MySQL	•Parquet	•Hive Metastore
•Iceberg	•PostgreSQL	•ORC	•AWS Glue
•Hudi	•Oracle	•Text	•Unity Catalog
•Paimon	•SQL Server	•Json	•Aliyun DLF
•Delta Lake	•IBM Db2	•SequenceFile	•Iceberg Rest Catalog
•Kudu	•ClickHouse	•RCFile	•Filesystem
•Bigquery	•SAP HANA	•...	•...
•Max Compute	•OceanBase		
•Lakesoul	•Elasticsearch		
•...	•...		

SQL 方言兼容

- Presto/Trino
- Clickhouse
- Spark
- 兼容程度在实际客户现场验证中达到95%

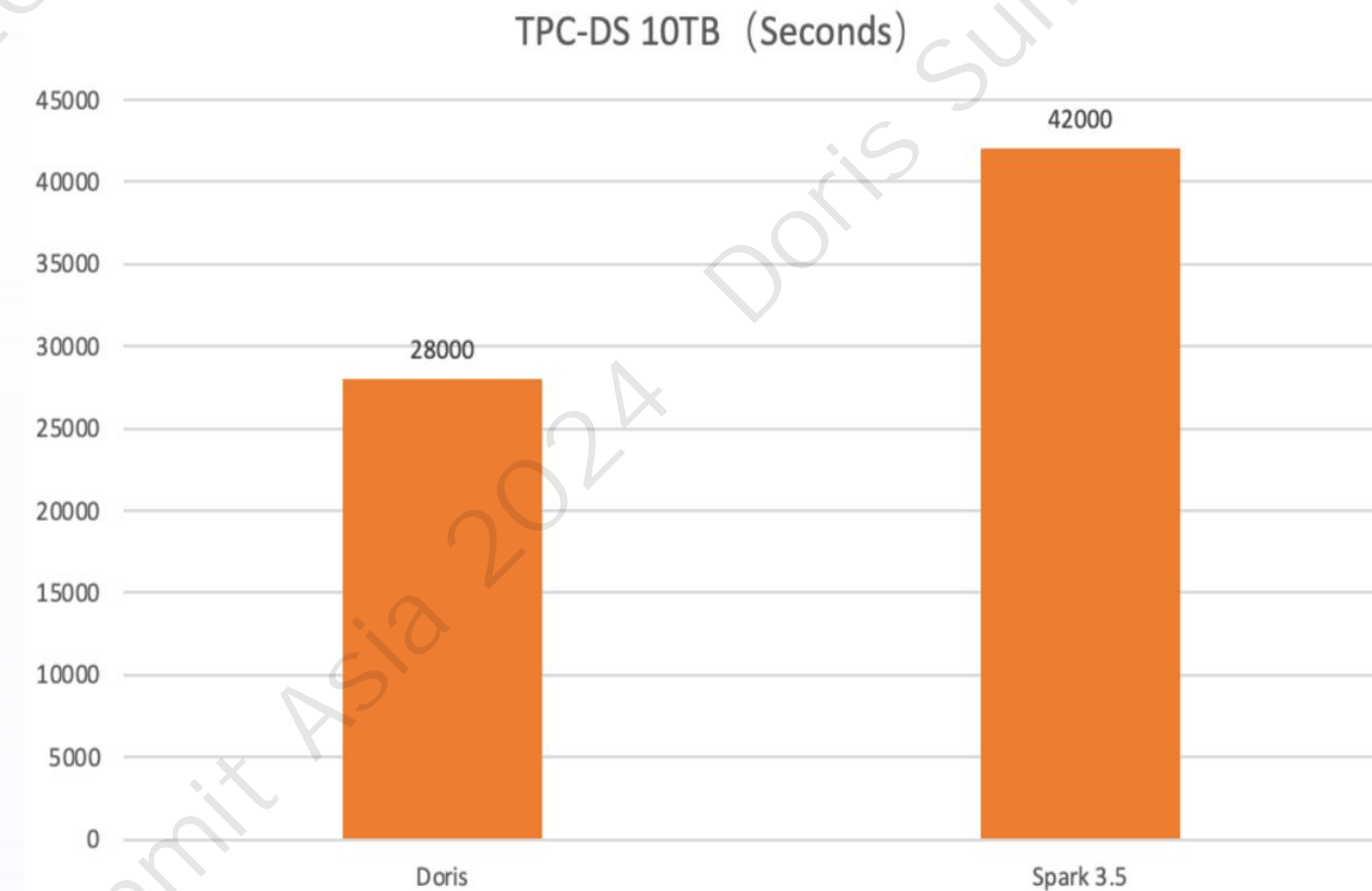
在线离线一体化

Adhoc 查询



Iceberg 外表, 5BE: 每台 16 core, 64GB

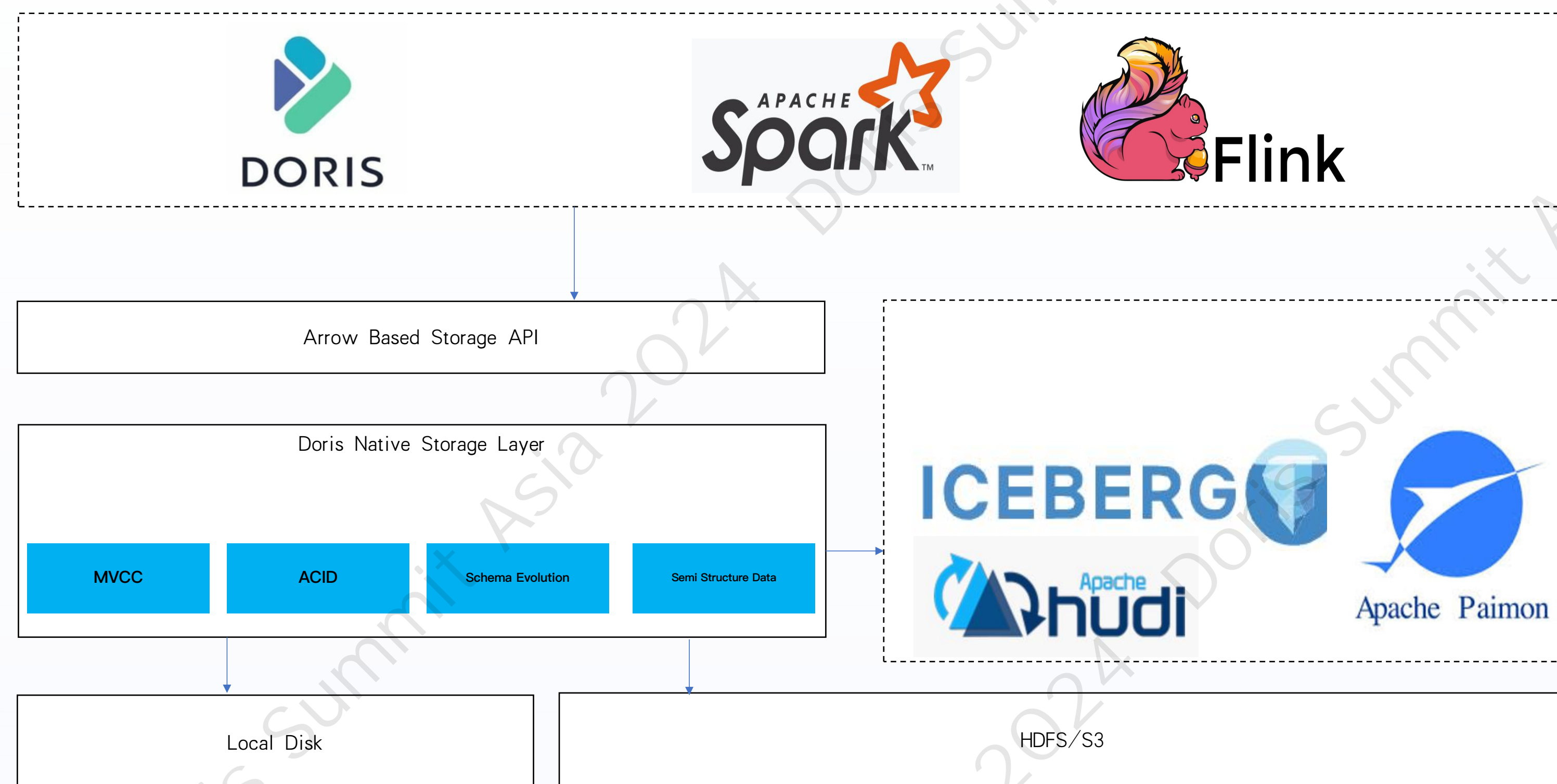
ETL



Iceberg 外表, 3BE: 每台 16 core, 64GB

湖仓融合

统一湖仓引擎



- 通过Storage API，支持Spark 和 Flink的高吞吐数据读取需求
- 冷热分离，存算分离降低成本
- 湖仓之间数据流动

Thanks for Watching!