

基于 Apache Doris 的高性能实时数仓架构设计与业务落地实践

邹婷

南京高泰信息科技有限公司, 江苏 南京 210046

DOI: 10.61369/TACS.2026050056

摘 要 : 数字化背景下, 企业对实时数据的分析与处理要求渐高。传统的离线数仓在数据处理时效性方面处于劣势, 并不能满足企业当下的需求, 实时数仓成为企业数据架构升级的重要方向。Apache Doris 作为分布式分析型数据库, 在低延迟查询、高吞吐写入与多源数据兼容方面具备显著优势, 适合用于搭建高性能实时数仓体系。基于此, 本文阐明实时数仓应用价值与优势, 结合 Apache Doris 技术特点设计实时数仓架构, 探讨其在电商、金融等行业的落地案例, 总结落地规划与优化思路, 期望能为各行业实时数据建设提供实践参考。

关 键 词 : Apache Doris; 实时数仓; 离线数仓; 架构设计; 业务落地

Architecture Design and Business Implementation Practice of High-Performance Real-Time Data Warehouse Based on Apache Doris

Zou Ting

Nanjing Gaotai Information Technology Co., Ltd., Nanjing, Jiangsu 210046

Abstract : In the context of digitalization, enterprises have increasingly stringent requirements for real-time data analysis and processing. Traditional offline data warehouses are at a disadvantage in terms of data processing timeliness and cannot meet current enterprise needs, making real-time data warehouses an important direction for upgrading enterprise data architectures. Apache Doris, as a distributed analytical database, offers significant advantages in low-latency querying, high-throughput writing, and compatibility with multi-source data, making it suitable for building a high-performance real-time data warehouse system. Based on this, this paper elucidates the application value and advantages of real-time data warehouses, designs a real-time data warehouse architecture incorporating the technical characteristics of Apache Doris, explores its implementation cases in industries such as e-commerce and finance, and summarizes implementation planning and optimization approaches. It aims to provide practical references for real-time data construction across various industries.

Keywords : Apache Doris; real-time data warehouse; offline data warehouse; architecture design; business implementation

引言

在实时数仓普及之前, 企业多依赖离线数仓处理数据, 这类数仓以固定周期批量加工数据, 数据从产生到可用存在较长延迟, 适用于历史复盘、固定报表等非即时性需求。实时数仓能够缩短数据处理周期, 让企业能在业务运行中快速获取数据, 拓展了数据应用场景, 成为企业数字化升级的关键。Apache Doris 凭借高并发、低延迟的核心优势, 成为企业搭建实时数仓的优选组件。本文围绕 Apache Doris, 梳理实时数仓架构设计思路, 结合行业案例阐述应用方式, 为企业实时数据体系建设提供参考。

一、实时数仓应用价值与优势

实时数仓是面向企业即时性数据处理需求搭建的数据分析体系, 主要用于对业务产生的流式数据进行快速接入、计算与查询。随着企业数字化运营的转型, 实时数仓从早期辅助性的数据应用工具, 逐步发展为支撑核心业务决策的关键基础设施, 在零

售、金融、物流等行业广泛应用^[1]。Apache Doris 作为成熟的分布式 MPP 分析型数据库, 凭借低延迟查询、高吞吐写入的特性, 成为当前企业搭建高性能实时数仓的主流技术选择。

实时数仓的应用, 可以使数据在极短的时间内完成加工并投入使用, 企业经营决策可以紧跟业务变化做出调整。企业内部分散在不同系统中的业务数据、行为数据得以高效整合, 数据之间

的关联价值也可以得到进一步发掘。流批一体的运行模式让数据处理流程更加简洁，企业在系统运维上的投入也随之减少，资源可以根据业务波动灵活调配，整体运行效率得到提升^[2]。传统离线数仓更适合长期历史数据的复盘分析，数据加工存在固定周期，在时效性和场景适配性上难以跟上当前业务的发展节奏，实时数仓则很好地弥补了这一不足，更能适配企业现代化的数据运营需求。

二、Apache Doris 的技术架构与原理分析

（一）Apache Doris 的整体架构设计

Apache Doris 的整体架构由前端节点和后端节点两部分构成，企业可以根据自身的业务规模和需求，选择存算一体或者存算分离的部署方式^[3]。

1. 前端节点（FE）

FE 是 Apache Doris 的控制中心，负责元数据管理、查询解析和优化、请求调度等工作^[4]。FE 节点有三种角色：

Leader 节点：负责元数据变更、集群管理和请求分发

Follower 节点：参与元数据投票，提供查询服务

Observer 节点：不参与投票，仅提供查询服务，用于扩展查询能力

FE 通过类 Raft 协议保障元数据的高可用和一致性。多个 FE 节点组成集群，避免单点故障。

2. 后端节点（BE）

BE 是数据存储和计算的核心节点，负责数据存储、副本管理和查询执行^[5]。BE 节点的核心功能包括：

数据分片存储，每个表被分为多个 Tablet，分布在不同 BE 节点

数据副本管理，默认3副本，保障数据可靠性

向量化执行引擎，利用 CPU SIMD 指令加速查询

MPP 并行计算，多个 BE 节点协同完成复杂查询

3. 存算分离架构

Apache Doris 3.0 版本引入存算分离架构，将系统分为三层：

元数据层：管理元数据和请求规划

计算层：由多个计算组组成，支持独立扩展

共享存储层：使用对象存储或分布式文件系统存储数据

存算分离架构让用户可以独立扩展存储和计算资源，提高资源利用率，降低总体拥有成本。

（二）核心技术特性

1. 多数据模型支持

多数据模型支持 Apache Doris 提供三种核心数据模型，适配不同业务场景：

明细模型（Duplicate Key）：适合日志、行为等原始数据存储，保留数据所有细节

聚合模型（Aggregate Key）：适合统计分析场景，自动聚合数据，提高查询效率

主键模型（Unique Key）：支持行级更新，适合需要实时更

新的业务场景

这些模型可以在同一集群中并存，满足企业多样化的数据存储需求。

2. LSM-Tree 存储引擎

Apache Doris 采用 LSM-Tree 存储引擎，实现秒级数据可见性。数据写入流程如下：

数据首先写入内存表（MemTable）

内存表达到阈值后冻结为不可变内存表

后台线程将不可变内存表刷盘为 Segment 文件

定期执行 Compaction 合并小文件，优化查询性能

这种机制兼顾了高写入性能和快速查询能力，适合实时数据场景。

3. 向量化执行引擎

Apache Doris 执行引擎基于 MPP 框架和 Pipeline 数据处理模型，支持向量化执行。向量化执行引擎可以一次处理一批数据，利用 CPU 缓存和 SIMD 指令，单节点扫描速度可达 2GB/s^[6]。

4. 湖仓一体能力

Apache Doris 支持直接访问外部数据湖（如 Hive、Iceberg、Hudi、Paimon）中的数据，这种能力让企业可以利用已有的数据湖资产，同时享受 Doris 的高性能查询体验^[7]。Apache Doris 3.0 版本还增加了数据湖写回功能，用户可以在 Doris 中完成多个数据源之间的数据分析、共享、处理和存储操作。

三、基于 Apache Doris 的实时数仓架构搭建思路

（一）数据接入层

数据接入层负责采集和传输各类数据，采用 Flink CDC 抓取业务库中的数据变更信息，

Kafka 则用来中转流式数据，让数据可以有序传递到后续环节。接入层支持多种数据格式，如 JSON、CSV、Parquet 等，不同来源、不同结构的数据都能顺利汇入数仓体系。

（二）数据存储层

数据存储层采用 Apache Doris 作为主存储，同时搭配数据湖存储原始数据，经加工处理后的数据统一放在 Doris 中，平台运行产生的完整历史数据则归档在数据湖内容^[8]。Apache Doris 自带自动分区与分桶的相关配置，数据在查询时可以更快被定位到，整体的查询体验也会更流畅。

（三）计算引擎层

计算引擎层由 Apache Doris MPP 引擎和 Flink 共同用承担计算任务。Flink 主要处理流式数据的加工计算，Apache Doris 负责复杂分析和批量数据处理。两者分工协作，形成流批一体的计算模式，数据不用重复处理，能够更高效流转到后续环节^[9]。Apache Doris 的向量化执行引擎搭配并行计算能力，面对复杂查询时，能快速给出结果，无需让用户长时间等待。

（四）数据服务层

数据服务层提供统一的数据访问接口，包括 JDBC/ODBC、REST API 等常用形式。通过 BI 工具如 Tableau、Power BI，用

户可以实时查询和分析数据^[10]。即便多个业务场景同时发起数据查询，系统也可以保持良好的运行状态，满足多场景并行使用的需求。

（五）应用层

应用层面向具体业务场景，如实时报表、业务监控、个性化推荐等。应用层根据业务需求调用数据服务层接口，获取实时数据结果。通过可视化展示，帮助用户快速理解数据，做出决策。

四、Apache Doris 实时数仓业务落地实践

（一）电商行业

电商行业具备数据流量庞大、业务链路零散的特征，行业对数据实时性分析要求严苛，数据处理门槛高。京东精准通广告平台每日产生大量广告曝光、用户点击以及交易转化行为数据，各类投放消耗与消费数据分散在不同业务模块中，进而加大数据整合与查询分析的工作难度。为改善该现状，京东广告团队引入 Apache Doris 搭建实时数据平台，面向广告商家提供投放数据检索、可视化报表统计等基础服务。平台结合广告业务实际使用场景，采用差异化的数据存储管理模式。一方面完整留存曝光、点击等原始明细数据，保障数据溯源能力；另一方面依据广告主常用分析维度，对投放消耗、推广计划等数据做聚合处理，实时同步预算管控、投放策略调整等业务信息。同时平台直连数据湖原始投放日志，简化数据流转链路，缩短数据加工延迟。由于广告主存在高频报表查看、多维数据分析的使用需求，技术团队针对性优化系统运行参数与集群调度逻辑，提升平台并发查询能力。在 618、双十一等电商大促节点，平台能够平稳承接查询流量峰值，方便商家实时掌握投放情况，并依托数据调整投放方案。当前平台日均可支撑千万级查询请求，即便处于大促高负载环境，系统依旧能够保持平稳运行。

（二）物流行业

纵腾集团需要处理大量物流数据，包括订单信息、运输状态和仓储数据。采用 Apache Doris 构建流批一体数仓后，实现了以

下优化：

数据模型选择：纵腾集团选用 Unique 模型模拟 ODS 层，每张表设置 70 个 Tablet，充分考虑 Compaction 性能及小文件场景。

写入性能优化后，最大写入时效达到 269 MB/S&680K ops/s，平均写入时效达到 70 MB/S&180K ops/s，提升了数据接入速度。

查询性能提升 Apache Doris 的查询压测性能优异，相比原来的方案，查询速度提升了 5 倍以上，可以提供更优质的物流业务实时查询服务。

（三）电信行业

天翼云基于 Apache Doris 和 Iceberg 构建湖仓一体平台，集群规模超过 3000 个节点，存储容量超过 15PB。

湖仓一体架构天翼云实现了“存储在湖、计算在仓”的高效协同模式，Apache Doris 作为计算引擎直接访问 Iceberg 数据湖中的数据。

大规模数据处理平台支持 PB 级数据的实时分析，可以满足电信行业的大规模数据处理需求。

多场景支持天翼云的湖仓一体方案满足了报表和 BI 分析、湖仓融合分析、日志存储分析、高并发实时分析等多种场景需求。

五、结语

Apache Doris 能够为企业搭建高性能实时数仓，更能适配当下企业即时性的数据应用需求，帮助企业快速落地实时数据处理体系。不同行业的业务特点存在区别，基于 Apache Doris 的实时数仓落地形式也会有所不同，企业可以结合自身数据源、业务需求和资源条件，定制适配的架构方案。随着企业数据化运营的不断深入，实时数仓的应用场景会持续拓展，Apache Doris 也会在更多行业的数据体系建设中发挥作用，推动业务的精细化运营与发展。

参考文献

[1] 丁肖, 陈辉, 耿兴隆, 等. 基于 Flink 的工业大数据实时数仓的设计与实现 [J]. 科技传播, 2025, 17(05): 24-27.
[2] 义天鹏. 银行湖仓一体大数据平台的探索与实践 [J]. 金融科技时代, 2024, 32(05): 64-71+75.
[3] 薛振文, 李彤, 牛帆. 大数据时代数据仓库技术的发展及应用 [J]. 中国传媒科技, 2023, (10): 154-15.
[4] 田名艳. 电信运营商大数据实时数仓系统的设计与实现 [D]. 北京工业大学, 2023.
[5] 张煜杨. 电商大数据处理平台的设计与实现 [D]. 华东师范大学, 2023.
[6] 贺晓松. 大数据背景下的数据仓库架构设计及实践研究 [J]. 中国新技术新产品, 2022, (19): 22-25.
[7] 孙志成, 许昆龙, 郑斌. 金融机构大数据技术应用创新实践 [J]. 信息技术与标准化, 2022, (06): 50-54+61.
[8] 李钊. 基于 Flink 的电商实时数仓系统的实现与应用 [D]. 大连交通大学, 2022.
[9] 吕家尧. 面向工业领域的实时数据仓库的设计与实现 [D]. 中国科学院大学 (中国科学院沈阳计算技术研究所), 2022.
[10] 卢山. 实时数仓架构探索与实践 [J]. 电子世界, 2022, (02): 96-97.