doi: 10.3969/j.issn.1672-6073.2024.01.009

基于“湖仓一体”技术的城轨大数据平台设计与升级改造实践

吴雁军，光志瑞，李明华，陈建华
（北京市地铁运营有限公司技术创新研究院分公司；地铁运营安全保障技术北京市重点实验室，北京 100082）

摘 要：为了探寻城市轨道交通行业大数据平台建设与升级改造的最优方案，本文以城轨大数据平台为研究对象，从城轨大数据平台发展历程出发，梳理城轨大数据平台发展的 3 个阶段，分析各阶段大数据平台所采用的技术与优缺点，重点总结当前阶段“湖仓一体”大数据技术所具备的湖仓一体、流批一体、OLTP+OLAP、多重负载等优点，研究了基于该技术的大数据平台架构升级改造设计要点，并将该技术在北京地铁数据中心的大数据平台升级改造中进行应用验证。结果表明：“湖仓一体”大数据平台技术兼具数据湖的低成本、数据仓库的高性能等优点，解决了原大数据平台在性能、容量与多用途支持上的不足，为城轨行业大数据平台建设与升级改造提供了新的解决思路。

关键词：城市轨道交通；大数据平台；升级改造；湖仓一体；流批一体；数据仓库；数据湖
中图分类号：U231　　**文献标志码：**A　　**文章编号：**1672-6073(2024)01-0054-09

Design and Upgrade Practice of Urban Rail Big Data Platform Based on “Data Lakehouse” Technology

WU Yanjun, GUANG Zhirui, LI Minghua, CHEN Jianhua
(Beijing Mass Transit Railway Operation Corp., Ltd., Technology Innovation Research Institute Branch, Beijing Key Laboratory of Subway Operation Safety Technology, Beijing 100082)

Abstract: To explore the optimal scheme for the construction and upgrading of the big data platform in the urban rail transit industry, this study takes the urban rail big data platform as the research object. Our study starts from the development process of the urban rail big data platform, sorts out the three stages of the development of the urban rail big data platform and analyzes the technology and advantages and disadvantages of the big data platform at each stage. Then it focuses on summarizing the advantages of “Data lake and Warehouse integration, stream processing and batch processing integration, OLTP+OLAP, multiple loads” and other advantages of the “Data Lakehouse” big data technology in the current stage, and studies the key points of the architecture upgrade and transformation design of the big data platform based on this technology. The technology was verified in the upgradation and transformation of the big data platform of the Beijing Metro Data Center. The application shows that the “Data lakehouse” big data platform technology combines the advantages of low cost of data lake and high performance of data warehouse, solves the shortcomings of the original big data platform in performance, capacity and multi-purpose support, and provides new solutions for the construction and upgradation of big data platforms in the urban rail industry.

Keywords: urban rail transit; big data platform; upgrading; data Lakehouse; stream processing and batch processing integration; data warehouse; data lake

中国城市轨道交通行业经历了近 20 年的快速发展。截至 2023 年 6 月 30 日，我国已有 57 座城市开通了城市轨道交通线路，运营总里程已达 10 566 km[1]，特别是北京、上海、广州、深圳等超大型城市的轨道交

收稿日期：2023-10-18　修回日期：2023-11-10
第一作者：吴雁军，男，本科，高级工程师，主要从事智慧交通、交通大数据、人工智能等方面研发工作，1532790289@qq.com
引用格式：吴雁军，光志瑞，李明华，等. 基于“湖仓一体”技术的城轨大数据平台设计与升级改造实践[J]. 都市快轨交通，2024，37(1)：54-62.
WU Yanjun, GUANG Zhirui, LI Minghua, et al. Design and upgrade practice of urban rail big data platform based on “data lakehouse” technology[J]. Urban rapid rail transit, 2024, 37(1): 54-62.

通已全面进入规模化和网络化运营阶段。城市轨道交通业务复杂，空间上跨越车站、线路、线网三种维度，专业上覆盖客流、行车、设备、供电、车辆、轨道等多种门类，关联信息化系统有综合监控系统(integrated supervision and control system，ISCS)、火灾报警系统(fire alarm system，FAS)、环境与设备监控系统(building automatic system，BAS)、监控和数据采集系统(supervisory control and data acquisition，SCADA)、自动列车监控系统(automatic train supervision，ATS)、自动售检票系统的清分中心(AFC clearing center，ACC)、多线路管理中心(multiple line center，MLC)、自动售检票系统(automatic fare collection，AFC)等数十个，每天会产生海量多源异构数据[2]。如何将这些数据有机地结合起来，来满足城市轨道交通快速发展产生的新需求，对城轨大数据平台建设提出了巨大的挑战。

一方面，城市轨道交通网络化快速发展，使网络化运营的调度指挥工作复杂程度大大增加，这对作为调度指挥系统“大脑”的数据中心平台提出了极高的要求。另一方面，轨道交通运营时间超过 10 年的城市已有 15 座，其配套的数据平台已经不能满足轨道交通安全提升、品质提升、效率提升、效益提升的需求，大多面临更新改造的问题。因此，如何升级改造既有数据中心平台或建设新的大数据平台，成为非常值得研究的课题。鉴于在城市轨道交通领域利用第三代大数据技术解决上述问题的研究成果并不多见，在参考轨道交通领域既有重要文献[3]的基础上，本文以城市轨道交通行业大数据平台的建设与升级改造为场景，研究并提出基于最新“湖仓一体”大数据技术的城轨大数据平台设计，助力城市轨道交通大数据体系的建设与完善。

1　城轨大数据平台技术发展阶段和现状

1.1　城轨大数据平台技术发展阶段

城轨大数据平台技术的发展与大数据技术自身的发展密不可分，到目前为止大致可分为 3 个阶段[4]，如图 1 所示。

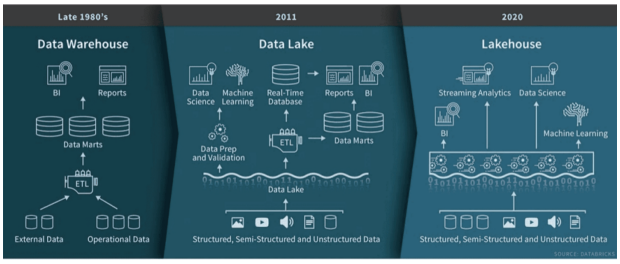

图 1　大数据技术发展历程

Figure 1　Big data technology development history

第 1 阶段，数据仓库(data warehouse)阶段。数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策和信息的全局共享，面向联机分析处理(on-line analytical processing，OLAP)，支持海量结构化数据的复杂分析。2011 年以前，城轨数据中心的核心任务是解决传统关系型数据库无法承载海量结构化数据复杂分析的难题，因此主流的城轨数据中心平台搭建主要采用数据仓库软硬件一体机技术。例如，北京市轨道交通指挥中心线网指挥中心(traffic control center，TCC)二期数据中心平台、广州市轨道交通工程线网指挥平台(comprehensive operation coordination center，COCC)一期数据中心平台、深圳市轨道交通网络运营控制中心(network operation control center，NOCC)一期数据中心平台均采用 Teradata Aster 数据仓库软硬件一体机，西安地铁线网(应急)指挥中心(network contingency and operating contrd center，NCC)一期采用 Oracle ExaData 数据仓库软硬件一体机。数据仓库拥有海量结构化数据的离线分析能力，但不支持非结构化数据分析，也不支持实时数据分析，导致数据中心平台功能单一，主要用于指标计算与报表输出，不能满足调度指挥系统实时监控需求。

第 2 阶段，数据湖(data lake)[5]阶段。随着城轨业务的发展，在指标分析的基础上实时监控和机器学习等新需求日增，这与数据仓库只支持结构化历史数据分析、不支持非结构数据、不支持实时分析的特性之间的矛盾日益突出。2011 年 12 月，Hadoop 1.0.0 的发布开启了基于数据湖技术搭建大数据平台的新篇章。以 Hadoop 为代表的数据湖技术构建的大数据平台可以存储任何形式(包括结构化、非结构化和半结构化)和任何格式(包括文本、音频、视频和图像等)的原始数据。这使得数据湖能搭建真正的大规模数据平台，数据存储成本也更为廉价。凭借该优点，Hadoop 在数年间逐步取代数据仓库软硬件一体机建设方案，成为城轨数据中心建设的主流技术。例如，青岛地铁线网运营管理与指挥中心(metro management & command center，MMCC)一期数据中心平台就采用 Hadoop 技术搭建，很好地解决了数据仓库不能处理非结构化数据、不支持流处理的问题。大数据平台成为统一的基础平台，既能支持离线指标计算，也能支持调度指挥系统所需的实时监控。但 Hadoop 本身并不支持结构化数据分析，使用 Hadoop 生态 HBase、Hive 等组件虽然可以实现类 SQL 语句查询，但不支持删除、修改操作，且执行效率不高、开发烦琐，不具备数据仓库操作管理的便利性。

第 3 阶段，湖仓一体(data lakehouse)阶段。随着云计算技术的兴起，依托云平台的新一代数据仓库——大规模并行处理数据库(massively parallel processing，MPP)逐渐发展成熟。MPP 能弥补 Hadoop 在结构化数据分析处理效率上的短板，因此，“Hadoop+MPP”混合架构理念被提出。例如，苏州市轨道交通线网指挥中心(network control center，NCC)一期数据中心平台利用 Hadoop 技术实现非结构化数据的分布式存储与实时计算，运用 MPP 实现结构化数据的快速分析，在一定程度上解决了数据湖技术存在的问题。但“Hadoop+MPP”的模式只是数据湖与数据仓库的简单组合，两套系统相对独立，同时开发与维护两套系统成本较高，且会造成数据冗余与逻辑不完全一致等问题，因此城轨数据中心亟需一套完美解决上述问题的方案。针对这一需求，2020 年“湖仓一体”概念被提出[4]。湖仓一体是一种通过一套数据湖仓组件实现数据湖和数据仓库两者优势的新范式，借助数据湖的低成本存储，实现与数据仓库中类似的数据结构和数据管理功能，这使得真正打通数据仓库和数据湖两套体系、构建一套有机的大数据技术生态体系成为可能。

1.2 湖仓一体大数据平台的技术特点

湖仓一体大数据平台实现原理如图 2 所示。数据湖仓由两部分组成：底层的数据湖作为存储层，用来存储包含结构化、半结构化和非结构化数据的任意种类数据；上层是由元数据、缓存、索引组成的数据管理层。数据管理层读取底层的数据，通过对数据的抽取、转化、加载(extract-transform-load，ETL)后，以接口的形式提供给上层应用访问。这些接口分为 3 大类：元数据接口、高性能 SQL 接口以及申明性数据帧 API 接口。其中，元数据接口实现数据湖仓对多源异构数据的统一管理，高性能 SQL 接口提供高性能 SQL 引擎实现对数据分析、报表报告的支持，申明性数据帧接口实现对数据科学与机器学习的支持。

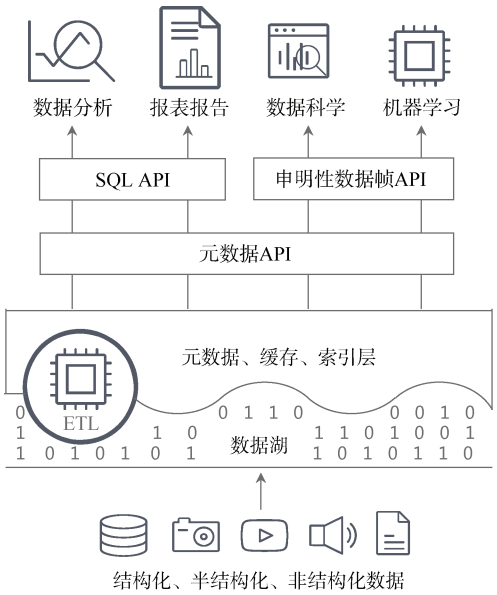

图 2 湖仓一体技术实现原理

Figure 2 Implementation principle of the Data Lakehouse technology

数据仓库、数据湖和湖仓一体 3 种技术，在数据格式、数据类型、数据访问、可靠性、应用场景支持等方面有明显区别[6]，具体如表 1 所示。

表 1 大数据平台构建技术对比

Table 1 Comparison of Big Data platform construction technologies

对比维度	数据仓库	数据湖	湖仓一体
数据格式	封闭的专有格式	开放格式	开放格式
存储数据类型	结构化数据为主，半结构化数据有限支持	所有类型：结构化、半结构化和非结构化数据	所有类型：结构化、半结构化和非结构化数据
数据访问	仅支持SQL访问	可直接访问到文件，并支持 SQL、R、Python 及其他语言	通过开放 API 可直接访问到文件，并支持 SQL、R、Python 及其他语言
可靠性	通过 ACID 事务处理提供高质量、可靠数据	低质量，易造成数据沼泽	通过 ACID 事务处理提供高质量、可靠数据
数据治理与安全	为数据表提供行/列级别的细粒度安全性和治理	安全性一般，因为需要将安全性应用到文件	为数据表提供行/列级别的细粒度安全性和治理
性能	高	低	高
扩展性	按比例扩展成本会成倍增加	可以低成本保存任何数量的数据而不考虑类型	可以低成本保存任何数量的数据而不考虑类型
用户场景支持	仅限于统计分析、报表报告	仅限于 AI 决策，统计分析需要使用额外组件	一个架构同时支持统计分析与 AI 决策

从技术形态上来说，数据仓库一般是独立的标准化系统，如 Teradata Aster 数仓软硬件一体机、华为 GaussDB for DWS。数据湖更像是一种架构指导，需要配合周边工具来实现业务需要。Hadoop 是最常用的数据湖技术，Hadoop 生态丰富的组件为其提供了高效的工具。湖仓一体也是一种架构，以数据湖为中心，把数据湖作为中央存储库，再围绕数据湖建立专用“数据服务环”，环上的服务包括了数据仓库、大数据处理、机器学习、交互式查询等一系列服务。目前，业界主

流湖仓一体开源技术有 Apache Iceberg、Apache Hudi 和 Delta Lake。国内厂商也纷纷推出了自己的湖仓一体解决方案，如华为云 FusionInsight、阿里云 MaxCompute 等。

综上所述，湖仓一体技术很好地避免了数据湖与数据仓库的局限性。基于湖仓一体技术搭建的城轨大数据平台有以下 4 大优势：

1) 存算分离。同时实现低成本与高性能。数据湖仓存储层通过分布式存储实现类似数据湖的大容量与低成本，计算层利用 Hudi 等数据湖仓组件实现类似数据仓库高性能分析能力，支持轨道交通海量数据低成本存储、高效率分析。

2) 流批一体。同时支持实时处理与离线处理。平台既能满足海量离线数据的批处理计算需求(如夜间运行批处理计算前一天所有运营指标)，也能满足实时数据的流处理计算需求(例如，持续处理当前进出站客流量，实时推算线网断面客流等实时指标)，更好地支持运营状态监控与分析。

3) OLAP+OLTP。同时支持联机分析处理(online analytical processing，OLAP)与联机事务处理(online transaction processing，OLTP)。平台既能支持从数据湖仓获取原始数据并在数据仓库通过 OLAP 进行复杂的分析挖掘，也能支持通过 OLTP 将处理结果反馈回数据湖仓，及时更新数据源，更好地支持复杂运营场景科学决策。

4) 多重负载。同时支持数据分析与机器学习。通过统一的数据管理，平台既能支持敏捷分析，也能支持人工智能，打通了数据分析与机器学习，使二者无缝集成，为 AI 大模型新技术在轨交领域的应用创造了条件，为智慧地铁建设提供坚实支撑。

2 架构设计

为了充分发挥“湖仓一体”技术的优势，指导城轨大数据平台的建设和升级改造，现结合城市轨道交通行业特点，从系统架构、部署架构与数据架构 3 个方面开展分析。

2.1 系统架构

依据湖仓一体架构“存算分离、流批一体”的机制，城轨大数据平台的总体架构可分为 6 层(如图 3 所示)，自下而上分别是数据源、采集层、存储层、计算层、服务层和应用层。

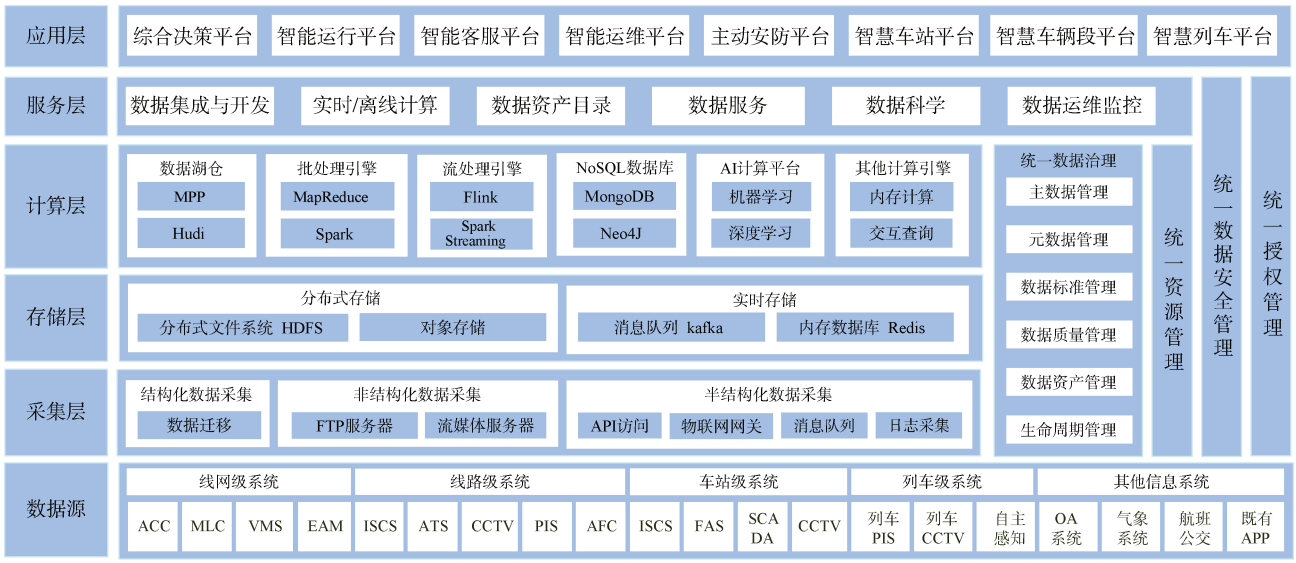

图 3　系统架构

Figure 3　System architecture

2.1.1 数据源

数据源，即大数据平台的数据来源系统，为大数据平台提供了海量要存储和应用的数据。数据源既包括 ACC、ISCS 等生产系统，也包括 OA 系统等管理系统，还包括气象信息系统、航班与公交等其他交通方式信息系统[7]。

2.1.2 采集层

采集层，负责从数据源采集各类数据，然后传递给存储层进行存储。数据采集时，根据存储结构类型不同分别采用不同的采集方式。对于结构化数据(如 OA 系统的员工表)，可通过数据迁移采集；对于非结构化数据(如 Excel 文件)，可通过 FTP 服务采集，CCTV

视频流可通过流媒体服务器采集；对于半结构化数据(如 Json 格式数据)，支持物联网的系统可以通过物联网网关系统采集，支持数据服务的系统可以通过访问其 API 接口采集，支持数据主动推送的系统可以通过消息队列接收数据。

2.1.3 存储层

存储层，也就是数据湖仓的存储部分，负责将采集层传递过来的数据进行存储，数据湖仓应是企业中全量数据的单一存储系统。数据按照时效性不同可分为实时数据、离线数据和近线数据。为了响应不同类似数据的存储和应用特征，湖仓一体数据平台中的存储层设计了两种存储模式，即实时存储和分布式存储。对于实时数据，为了保障其时效性和读取效率，通常将其保存在消息队列或内存数据库，如 Kafka 或 Redis；对于离线数据，为保证其安全性与并行性应采用分布式存储[8]，也可以通过 Hadoop 分布式文件系统(hadoop distributed system，HDFS)[9]或对象存储服务来存储；对于近线数据，可选内存数据库也可以选分布式存储，具体根据业务需求确定。

2.1.4 计算层

计算层，负责数据的处理、计算与管理。本层以 Apache Iceberg、Apache Hudi 和 Delta Lake 等数据湖仓组件为核心，以存储层的分布式存储作为支撑，利用流处理引擎与批处理引擎实现“流批一体”的数据处理机制，实现对在线分析(OLAP)与事务处理(on-line transaction processing，OLTP)的同时支持。流处理引擎用于以流处理的方式处理实时数据，常用的有 Hadoop 生态的 Flink 引擎和 Spark Streaming 引擎；批处理引擎用于以批处理的方式处理离线数据，常用的有 Hadoop 内置的 MapReduce 引擎和 Hadoop 生态的 Spark 引擎。数据仓库作为湖仓一体架构的重要组成部分也建在本层，目前最常用的数据仓库是 MPP，它同时支持行存储与列存储，支持事务处理机制，提供 PB 级结构化数据的高效在线分析能力，在结构化数据的处理方面具有明显优势，是湖仓一体架构的重要补充。此外，为了丰富数据库类型，本层还应部署 NoSQL 数据库，如文档型数据库 MongoDB、图数据库 Neo4J；为了支持 AI 计算，本层还应部署机器学习与深度学习平台；为了满足用户的多样化需求，还应部署内存计算引擎与交互式查询引擎等。

2.1.5 服务层

服务层，集成湖仓一体大数据平台对外提供服务。具体包括数据集成与开发、实时计算、离线计算、数据资产目录、数据服务、数据科学、数据运维监管等一系列服务。

2.1.6 应用层

应用层，即使用大数据平台的应用系统。大数据平台为应用系统提供数据支持与算法、算力支持，但严格意义上，应用层不划入大数据平台。在城市轨道交通领域，核心的应用软件有综合决策平台、智能运行平台、智能客服平台、智能运维平台和主动安防平台等智慧地铁中心级平台，以及智慧车站平台、智慧车辆段平台、智慧列车平台等智慧地铁现场级平台。

在这 6 层之外，还需要一些管理模块提供数据治理与数据安全等公共服务，确保大数据平台平稳、安全地运行。如图 3 系统架构所示，统一数据治理与统一资源管理模块跨越大数据平台采集层、存储层和计算层，提供大数据平台的数据管理服务；统一数据安全管理与统一授权管理模块跨越采集层和服务层，提供对数据安全与用户授权服务。

2.2 部署架构

城轨大数据平台应部署在云平台上[10]。云平台为大数据平台统一提供计算资源、存储资源和安全资源[11]。大数据平台按照所在云平台的网络空间可分为 3 个域：安全生产域、内部管理域和外部服务域。具体部署架构如图 4 所示。

安全生产域连接 AFC/MLC/ACC、ATS、ISCS、PSCADA、CCTV 等生产系统，获取实时数据用于支撑实时监控与生产大数据分析，需要部署 Hadoop 集群、MPP 集群、Redis 集群、kafka 集群和应用服务器。其中，Hadoop 集群需要部署 Hudi 数据湖仓、Spark 引擎、Flink 引擎等关键组件，以实现对“流批一体”数据处理能力和“多重负载”机器学习能力的支持。外部服务域主要用于发布面向互联网用户的服务和收发互联网数据的接口服务程序，考虑现有业务对计算资源的需求并不大，仅考虑部署应用服务器和关系数据库。内部管理域连接 OA 系统、EAM 系统等管理系统，并接收从安全生产域输入的生产数据，以及从外部服务域输入的互联网数据，进而实现三域数据的融合分析，因此需要部署安全生产域类似集群与组件。

2.3 数据架构

依据湖仓一体技术“存算分离、流批一体”的机制，按照城市轨道交通数据从采集、存储、处理到展示的过程，城轨大数据平台的数据架构可分为 4 层(如图 5 所示)，自下而上分别是数据源、存储层、计算层、展示层。

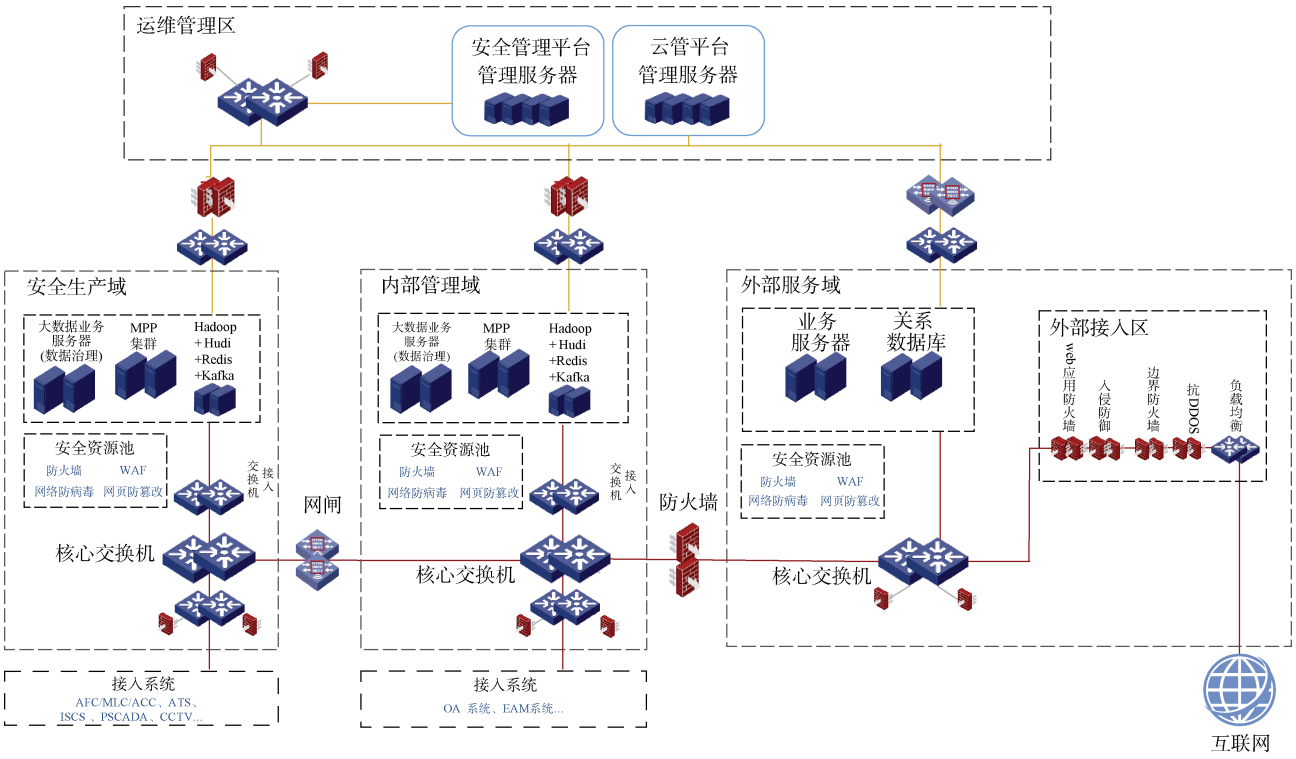

图 4　部署架构

Figure 4　Deployment architecture

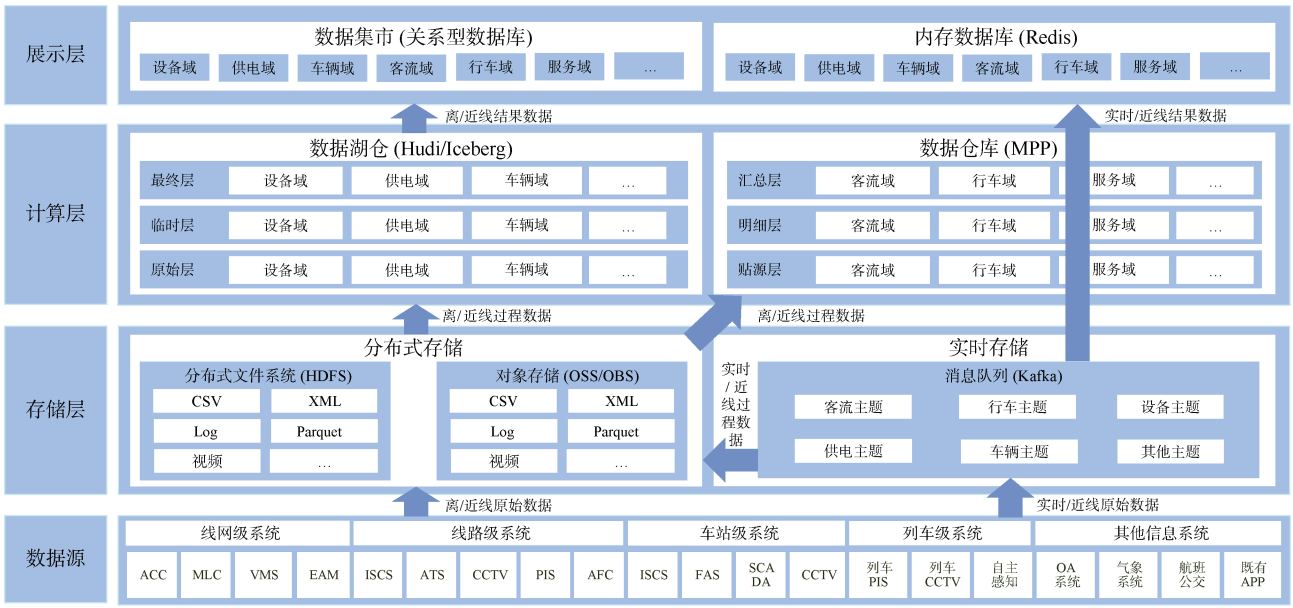

图 5　数据架构

Figure 5　Data architecture

2.3.1　数据源

数据源，即大数据平台的数据来源系统，与系统架构中的数据源一致。数据源包括各种生产系统、管理系统以及外部系统。这些系统为大数据平台提供数据，但严格意义上不属于大数据平台。

2.3.2　存储层

从数据源传递过来的数据在存储层进行存储，根据数据时效性不同分别采用不同的存储方式。实时数据以消息的形式按专业分别保存在 Kafka 消息队列的对应主题下；离线数据以 CSV 文件、日志文件、

Parquet 列式存储文件等形式保存在分布式文件系统(HDFS)或者对象存储系统中，作为数据湖仓的底层；而近线数据可以根据业务需要任意选择上述两种方式。实时数据也可以根据业务需求通过累积的方式转变为历史数据，然后作为离线数据进行存储。

2.3.3 计算层

数据在计算层进行处理、分析，形成分析结果，然后推送给展示层进行展示。对于离线数据，依托批处理引擎或流批一体引擎进行处理，在数据湖仓上层 Hudi 等组件或 MPP 中进行保存、分析。在数据湖仓中，从存储层读取的原始数据保存在原始层，分析过程的临时数据保存在临时层，分析结果保存在最终层。MPP 中的保存、分析与数据湖仓类似，只是 3 个内部层通常被称为贴源层、明细层、汇总层。对于实时数据，从消息队列 kafka 读取数据后，经流处理引擎或流批一体引擎处理，直接推给展示层进行展示。对于近线数据，同样既可以当作离线数据也可以作为实时数据来处理、分析。

2.3.4 展示层

展示层将计算层的分析结果提供给上层应用进行展示。对于分析结果中的离线数据，按照专业分别保存到数据集市的对应域中，数据集市通常选择关系型数据库。对于分析结果中的实时数据，按照专业分别保存到内存数据库的对应域中，内存数据库通常选择 Redis 键值库。对于分析结果中的近线数据，同样既可以当作离线数据也可以作为实时数据来保存、展示。

3 应用案例

2022 年 1 月，北京市地铁运营有限公司在既有全量数据仓库基础上，在打造的下一代企业级大数据平台中突破性地采用了湖仓一体架构设计，为智慧地铁建设更好地提供数据与算法支撑。如图 6 所示，Hadoop 平台的分布式文件系统(HDFS)具有提供海量多源异构数据的存储能力作为数据湖仓的存储层，Spark、Flink 等批处理和流处理引擎实现了对流处理与批处理的同时支持，Hudi 数据湖仓组件实现了在线分析(OLAP)与事务处理(OLTP)的同步支持，MPP 具有提供快速数据分析能力，既有全量数仓系统的接入确保了历史数据的延续性。此外，统一的数据管理系统实现了湖仓共享存储资源池，支持通过标准 SQL 访问跨域多源数据，支持数据科学与 AI 训练推理，减少了数据搬迁，实现了海量数据的快速价值挖掘。

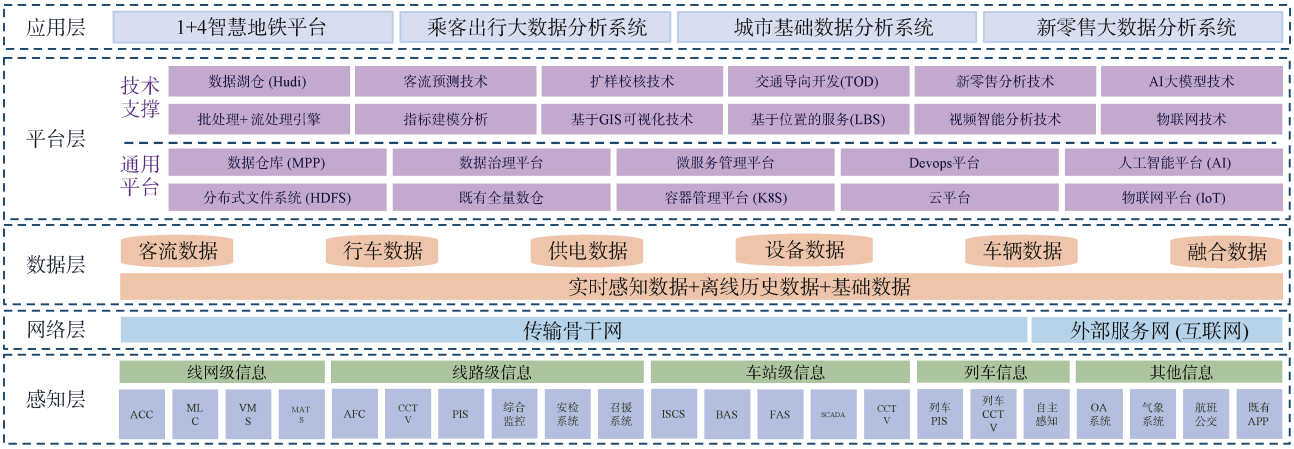

图 6 大数据平台总体架构

Figure 6 Big data platform overall architecture diagram

在系统部署方面，大数据平台基于云平台部署，依照业务需求与网络安全要求，在安全生产网和内部管理网均部署了数据湖仓集群(包含 Hadoop 节点、Kafka 节点、Redis 节点)、数据仓库集群(即 MPP 节点)和大数据业务集群(包含数据管理系统、各类大数据分析系统)，还在运维管理网部署了数据安全系统。在数据管理系统的协调下，打通了数据湖与数据仓库，构建了一套拥有完整的、有机的湖仓一体大数据技术生态体系。

北京地铁大数据平台(见图 7)作为新一代数据平台，与既有采用数据仓库技术的全量数据仓库系统相比，在 4 方面取得了突破性改善(见表 2)。

北京地铁智慧地铁大数据平台作为企业级大数据平台，接入了地铁运营所涉及的全专业数据，依托“湖仓一体”技术，不但具有了强大的运营指标融合计算能力，还具有了高效的运营数据实时分析能力，同时

通过对 AI 平台的数据支持，保障了人工智能模型优化能力，在提升地铁运营水平的同时支撑了智慧地铁建设诸多关键技术的科研攻关，为国家重点研发计划“超大城市轨道交通系统高效运输与安全服务关键技术”等项目的顺利实施奠定了坚实的基础，为北京地铁的数字化转型贡献了力量，也为湖仓一体大数据平台架构做了成功的验证，为城轨大数据平台建设与升级改造提供了新的解决方案。

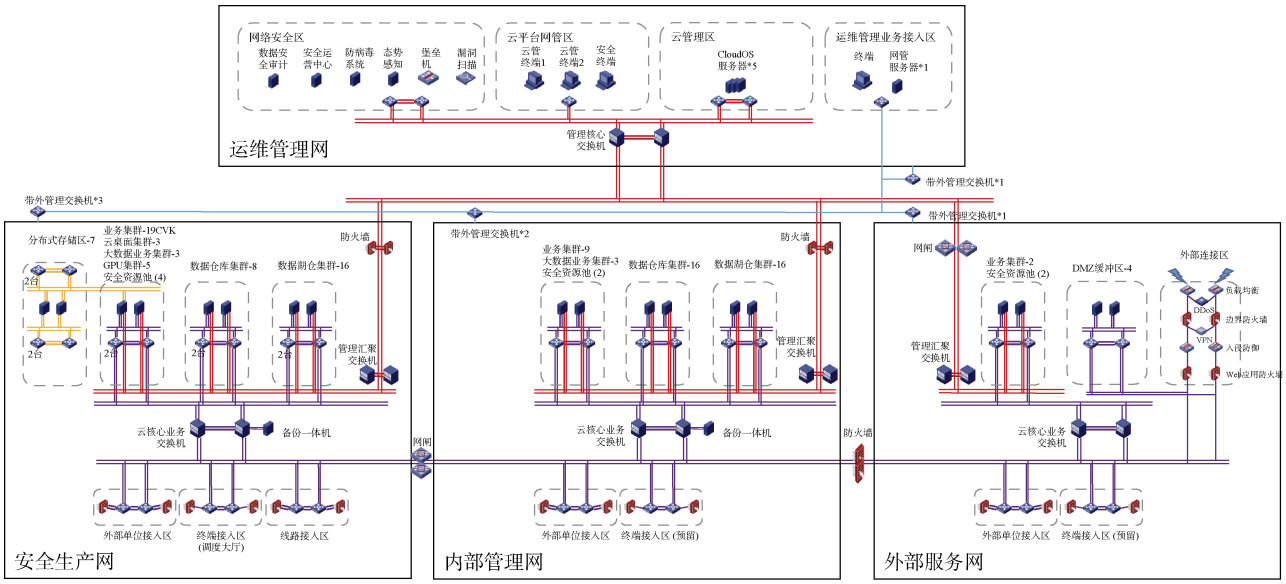

图 7　北京地铁大数据平台部署架构

Figure 7　Big data platform deployment architecture diagram

表 2　两代数据平台对比

Table 2　Comparison of the two generations of data platforms

对比维度	全量数据仓库 (基于数据仓库技术)	大数据平台 (基于湖仓一体技术)
成本	建设成本高昂，扩容成本按比例成倍增加	建设成本大幅降低，扩容成本低廉
数据范围	只支持结构化数据，不满足机器学习要求	支持所有数据类型，同时满足数据分析与 AI 技术需求
时效性	只支持离线数据批处理，不支持实时数据流处理	同时支持离线数据批处理与实时数据流处理
用途	只用于运营报表输出，不能支持调度系统等实时系统	可以作为智慧地铁所有在线分析、实时监控系统以及 AI 中台的统一数据支持平台

4　结束语

本文研究了基于“湖仓一体”技术的大数据平台升级改造设计要点，并在北京地铁数据中心的大数据平台升级改造中进行了应用验证，取得了良好效果。需要强调的是，首先，湖仓一体并不等同于数据湖+数据仓库。拥有数据湖和数据仓库的大数据平台必须具有统一的数据治理、统一的数据安全管理以及统一的资源管理等机制，将数据湖和数据仓库有机地融合起来才是真正的湖仓一体架构。其次，湖仓一体大数据平台架构在城市轨道交通领域的应用才刚刚起步，如何充分发挥数据湖仓组件的作用，使大数据平台摆脱对大规模并行处理数据库(MPP)的依赖还需进一步研究。总之，从发展趋势来看，“湖仓一体”技术必将在城市轨道交通企业数字化转型过程中发挥重要作用，值得继续研究和不断探索。

参考文献

[1] 2023 年上半年中国内地城轨交通线路概况[EB/OL]. 北京: 中国城市轨道交通协会. [2023-07-01]. https://mp.weixin.qq.com/s/xtN_so61W2QtqKi2jNynsA.

[2] 景亮, 方晖, 张森. 城市轨道交通信息化云平台及大数据平台建设[J]. 现代城市轨道交通, 2020(8): 129-134.
JING Liang, FANG Hui, ZHANG Sen. Construction of urban rail transit information Cloud and big data platform[J]. Modern urban transit, 2020(8): 129-134.

[3] 李中浩. 建设标准化的城市轨道交通云和大数据平台[J]. 城市轨道交通研究, 2021, 24(6): 12.
LI Zhonghao. Establish standardized urban rail transit cloud and big data platform[J]. Urban mass transit, 2021, 24(6): 12.

[4] MICHAEL Armbrust, ALI Ghodsi, REYNOLD Xin, et al.

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics[EB/OL]. [2020-12-22]. https://www.databricks.com/research/lakehouse-a-new-generation-of-open-platforms-that-unify-data-warehousing-and-advanced-analytics.

[5] 王健，徐炜，张宁，等. 南京地铁线网指挥中心大数据平台架构[J]. 都市快轨交通, 2021, 34(1): 138-143.
WANG Jian, XU Wei, ZHANG Ning, et al. The big data platform architecture of Nanjing metro network control center[J]. Urban rapid rail transit, 2021, 34(1): 138-143.

[6] 毛亮坚. 湖仓一体，构建企业数字化新基座[EB/OL]. [2021-11-24]. https://live.eyunbo.cn/live/74984?uin=1729.

[7] 徐炜，张宁，王健，等. 城轨线网指挥中心的大数据组织[J]. 铁路通信信号工程技术, 2020, 17(8): 62-66.
XU Wei, ZHANG Ning, WANG Jian, et al. Big data organization of network command center in urban rail transit[J]. Railway signalling & communication engineering, 2020, 17(8): 62-66.

[8] 胡彦. 城市轨道交通线网指挥中心大数据技术的应用[J]. 城市轨道交通研究, 2018, 21(增刊 2): 43-46.
HU Yan. Application of big data technology in urban rail transit network command center[J]. Urban mass transit, 2018, 21(S2): 43-46.

[9] 贾福宁. 轨道交通运营大数据：青岛地铁线网运营管理与指挥中心应用实践[M]. 北京：北京交通大学出版社, 2020.
JIA Funing. Big Data of Rail Transit Operation: Application Practice of Qingdao Metro Network Operation Management and Command Center[M]. Beijing: Beijing Jiaotong University Press, 2020.

[10] 城市轨道交通大数据平台技术规范：T/CAMET 11003—2020[S]. 北京：中国城市轨道交通协会, 2020.

[11] 吴昊，梁樑，张月坤，等. 超大线网标准城轨云及共享数据平台研究[J]. 都市快轨交通, 2022, 35(6): 69-74.
WU Hao, LIANG Liang, ZHANG Yuekun, et al. Standard urban rail cloud and shared data platform of a super-large line network[J]. Urban rapid rail transit, 2022, 35(6): 69-74.

（编辑：王艳菊）

（上接第 21 页）

[6] 李锦生，石晓冬，阳建强，等. 城市更新策略与实施工具[J]. 城市规划, 2022, 46(3): 22-28.
LI Jinsheng, SHI Xiaodong, YANG Jianqiang, et al. Urban renewal strategy and its implementation tools[J]. City planning review, 2022, 46(3): 22-28.

[7] 曹可心，邓羽. 可持续城市更新的时空演进路径及驱动机理研究进展与展望[J]. 地理科学进展, 2021, 40(11): 1942-1955.
CAO Kexin, DENG Yu. Spatio-temporal evolution path and driving mechanisms of sustainable urban renewal: progress and perspective[J]. Progress in geography, 2021, 40(11): 1942-1955.

[8] 毛羽. 城市更新规划中的体检评估创新与实践：以北京城市副中心老城区更新与双修为例[J]. 规划师, 2022, 38(2): 114-120.
MAO Yu. Innovation and practice of physical examination: old quarter renewal planning of Beijing sub-central district[J]. Planners, 2022, 38(2): 114-120.

[9] 杨元明. 既有铁路扩能改造若干问题探讨[J]. 铁道工程学报, 2010, 27(12): 1-4.
YANG Yuanming. Discussion on problems in upgrading of existing railway line[J]. Journal of railway engineering society, 2010, 27(12): 1-4.

[10] 杜连涛，许朝帅. 海外既有铁路升级改造可行性研究报告编制要点分析[J]. 中国铁路, 2018(3): 44-47.
DU Liantao, XU Chaoshuai. Key points for formulating feasibility study report of oversea existing railway line upgrading projects[J]. China railway, 2018(3): 44-47.

[11] 林柏梁，徐忠义，黄民，等. 路网发展规划模型[J]. 铁道学报, 2002, 24(2): 1-6.
LIN Boliang, XU Zhongyi, HUANG Min, et al. An optimization model to railroad network designing[J]. Journal of the China railway society, 2002, 24(2): 1-6.

[12] 《我国智慧城市建设若干关键问题研究》课题组. 走向智慧城市：我国智慧城市建设若干关键问题研究[M]. 北京：科学出版社, 2014.

[13] 张纯. 轨道交通与城市协同发展规划：理论、方法与评价[M]. 北京：中国建筑工业出版社, 2019.
ZHANG Chun. Coordinated development planning of rail transit and cities: theory, method and evaluation[M]. Beijing: China Architecture & Building Press, 2019.

[14] 建设部，国家发展改革委员会. 城市轨道交通工程项目建设标准：建标 104—2008[S]. 北京：中国计划出版社, 2008.

[15] 中华人民共和国住房和城乡建设部. 城市轨道交通工程项目规范：GB 55033—2022[S]. 北京：中国建筑工业出版社, 2022.

（编辑：王艳菊）