

基于湖仓一体技术的 “多湖一中台”工业大数据 平台研究与应用实践

蒋楠^{1,2}

1. 北京大学工学院, 北京 100871;

2. 中国石油化工集团有限公司, 北京 100728

摘要

数字经济时代数据量爆炸式增长, 数据平台在数据存储、处理、计算等方面遇到的挑战越来越大。提出基于湖仓一体技术的“多湖一中台”工业大数据平台架构, 旨在构建一个高效、灵活、可扩展的大数据管理平台, 以支持跨层级、跨组织、跨区域数据共享应用。通过3级数据湖和中台的统一部署, 构建“采、聚、理、用、保”五大核心功能, 形成多源异构数据采集、多模数据存储与数据处理、数据治理、数据服务、数据安全防护等能力。通过多湖协同联动和数据统一管控, 实现超大集团型企业数据的统筹管理、开放共享, 支持各业务领域数据应用场景落地, 促进数据要素价值释放。

关键词

数据湖; 湖仓一体; 数据中台; 工业大数据; 数据治理; 数据服务

中图分类号: TP311.13

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025069

Research and application practice of “Multi-Data Lake under One Middle-Platform” industrial big data platform based on Lakehouse technology

JIANG Nan^{1,2}

1. College of Engineering, Peking University, Beijing 100871, China

2. China Petroleum & Chemical Corporation, Beijing 100728, China

Abstract

With the explosive growth of data volume in the era of the digital economy, data platforms are facing increasing challenges in data storage, processing, and calculation. This paper proposes multiple Data Lakes under a unified Data Middle Platform, which is applied in industrial big data. The architecture, based on the integration of lake and warehouse technology, aims to build an efficient, flexible, and scalable data management platform to support cross-level, cross-organizational, and cross-regional data-sharing applications. With the centralized deployment of the three-level data lake and the middle platform deployed uniformly, the five core functions are constructed, forming the data

acquisition capabilities of multi-source heterogeneous data, multi-mode data storage, data processing, data governance, data service, and data security protection. Through multi-data-lakes, collaborative interaction, and unified data control ability, to realize the overall management and open sharing of data of super-large group companies, application scenarios in various business fields are supported, and the value of data elements is released.

Key words

Data Lake, Lakehouse, data middle platform, industrial big data, data governance, data service

0 引言

数据是国家的基础性战略资源，2020年4月被列为生产要素。随着数字经济的发展，企业数字化转型战略的持续推进，越来越多的新技术正在应用于传统工业领域，构建出企业的工业互联网系统。工业互联网可以将工业领域的生产设备、传感器和其他物理设备与计算机系统、网络通信等数字技术融合，实现数据信息的互通和智能化控制，不仅提升了传统企业的数字化水平，也带来了结构化、非结构化、半结构化数据的剧增。

石油化工行业是一个重资产行业，其企业数据具有数据量大、格式复杂、数据分散、类型模态多等诸多特点，大多数企业仍普遍采用关系型数据库或数据仓库来存储和处理数据，部分企业采用了数据湖技术来存储和处理数据，但普遍面临数据时效性不足、海量数据处理效率低下、无法有效利用非结构化数据、数据共享流通困难等挑战^[1]。

超大集团型企业迫切需要利用新一代数据技术研究适用于制造行业特点的工业大数据平台，以支撑总部和企业、企业和企业之间跨组织、跨层级、跨区域的数据高效采集处理和实时共享流通。本文提出基于湖仓一体技术“多湖一中台”工业大数据平台构建思路，将多级数据湖及数据

中台采用云边协同模式有机结合在一起，通过在企业边端部署的数据湖实现工业数据就近入湖、就近治理、就近计算，支撑时效性要求高的应用场景，通过总部云端部署的“一中台”实现集团统一数据目录、统一数据标准、统一数据分类、统一数据服务，有力支撑超大型集团企业之间、总部与企业之间跨组织、跨层级、跨区域的数据共享流通，促进数据要素价值释放。

1 研究现状

1.1 数据湖

基于工业大数据平台构建高兼容性数据湖是满足多模态数据存储及数据处理的关键。黄家凯等^[2]提出数据湖是一个存储所有类型数据的集中式存储库，可以存储结构化、半结构化和非结构化数据的原始格式；提供全局、统一的企业级数据概览视图，解决信息孤岛问题，为多源异构数据统一管理和数据隐藏价值挖掘提供可能。胡军军等^[3]提出数据统一存储、统一标准、近源采集、存储计算分离的电信运营数据湖建设方案，以解决数据质量不足、实时性差、缺乏灵活性、存储和应用相互牵制等问题。周大鸿等^[4]大数据平台的集中式数据存储库与计算引擎，允许对任意来源、任意规模、任意结构、任意传输速度的数据进行全量存储，并可跨平台、跨语言进

行分析和处理,提供数据的获取、存储、管理、分析、发布全周期全过程管理能力。王少杰^[5]提出通过数据湖的方法论打破数据孤岛,形成数据联邦,对企业多源异构数据进行统一管理。目前数据湖尚未有统一的定义,笔者认为数据湖是一套由数据存储与数据处理组成的解决方案。数据存储支持分布式存储,有足够的扩展性和可靠性,可以存储海量的任意类型的数据,包括结构化数据、半结构化数据和非结构化数据。数据处理则聚焦于如何把数据“搬到”湖里,包括定义数据源、制定数据同步策略、移动数据等,以支持对湖中的数据进行处理、分析、挖掘。

1.2 湖仓一体

湖仓一体概念由 Databricks 公司于 2020 年率先提出,并将其描述为“一个新的、开放的数据管理架构,将数据湖的灵活性、成本效益和规模与数据仓库的数据管理和事务管理结合起来”^[6]。陈氢等^[7]提出湖仓一体功能模块主要包括数据源、湖仓一体核心功能区和用户;流批一体数据流转过程包括批量数据过程和实时数据过程。李国华等^[8]提出总湖仓一体架构分为 3 层,与子湖仓一体架构相比没有采集摄取层,总湖仓一体的数据来源就是子湖仓一体的湖仓存储层级,数据来源安全且稳定,无采集存储层。总湖仓一体的数据湖可分别与子湖仓一体的数据湖和数据仓库之间进行数据传输。总湖仓一体与子湖仓一体组成分布式架构。目前湖仓一体还没有统一且成熟的定义,但业界普遍认为,湖仓一体架构同时吸收了数据仓库和数据湖的优势,应具有事务支持、数据模型化、开放性等多种工作负载支持、端到端实时等特性。

1.3 数据中台

朱一洲等^[9]提出了数据中台建设的特点,中台是一个整体解决方案,它既包含数据技术平台,也包含数据开发、数据模型、数据资产、中台治理和数据产品应用,以统一的数据平台实现企业数据汇聚、核心业务横纵连通、算法模型、数据管理能力的快速复用。数据中台是一套可持续“让企业数据用起来”的机制,是一套解决方案,不仅是一个技术平台^[10]。数据中台的本质在于数据的整合、共享和深度分析,为前台的决策赋能。ThoughtWork 首席咨询师提出“企业级的能力复用平台”,该定义强调了“企业级”和“能力复用”这两个关键词。笔者认为,“数据中台”不是简单的一套软件系统或者一个标准化产品,各类数据技术是构建数据中台的基础,通过提供收集、处理、存储、计算、分析和可视化呈现等各种数据能力,使数据成为可复用的资源,并与业务链条结合,真正转化为企业核心资产,发挥数据要素的价值^[11]。

综上,现有对数据湖、湖仓一体、数据中台的研究存在概念模糊和信息不完整等问题,数据湖与数据湖之间的协同机制尚无相关研究,未能提供涵盖全面的技术架构设计方法。当前,工业大数据平台主要是针对单一企业数据管理模式,不能满足集团型企业多数据来源、多管理层级、跨区域协同的数据共享模式^[12-14]。本文提出一种基于“湖仓一体”技术构建的“多湖一中台”工业大数据平台(以下简称平台),可解决超大集团型企业工业数据湖的高效搭建与快速融合,支撑企业经营、生产、服务等跨域数据的深度集成,结合数据治理体系的建立健全,实现集团数据资源“一个目录”,全面解决数据流通壁垒、数据服务能力重复建设等问题,满足集团

型企业工业大数据共享应用的需求，有效降低企业数据总体拥有成本，赋能业务运营决策。

2 “多湖一中台”工业大数据平台总体架构

2.1 总体架构

本文提出的“多湖一中台”工业大数据平台架构如图1所示，旨在通过3级数据湖的按需部署和数据中台的集中统一部署，实现数据治理能力、数据安全防护能力和数据服务输出能力的统一管理，构建一个高效、灵活、可扩展的集团级数据管理平台。

在多湖方面，平台涵盖总部数据湖、板块数据湖和企业数据湖3级数据湖。总部数据湖负责集中存储和管理全集团的核心数据，板块数据湖负责汇聚和管理各业务板块的数据，企业数据湖则针对企业级

数据进行存储和管理。

在中台方面，集中部署的“一中台”提供了统一的数据治理能力、数据安全防护能力和数据服务输出能力。通过统一的元数据管理、数据质量管理、数据安全管理等，实现全面数据治理。同时，数据中台还提供了强大的数据服务输出能力，能够将处理后的数据进行快速、高效的对外服务，对内满足总部、事业部和企业的数据共享需要，对外支持合作伙伴和客户的数据开放和运营。

2.2 功能架构

如图2所示，通过“采、聚、理、用、保”五大核心能力建设，该平台形成了多源异构数据采集、多模态数据存储与数据处理、数据治理、数据服务、数据安全防护等能力。通过多湖协同联动和数据统一管控，平台实现数据共享和开放。

数据湖包括数据采集、数据存储与数据处理模块，采用分布式存储和计算技术，

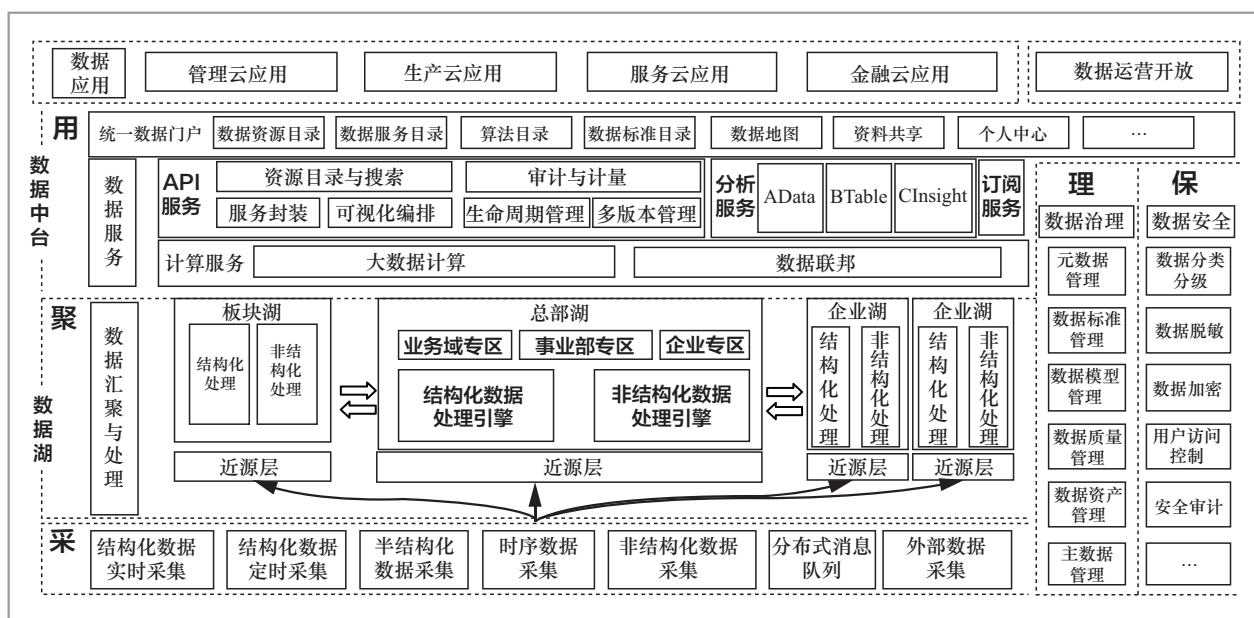


图1 “多湖一中台”工业大数据平台总体架构

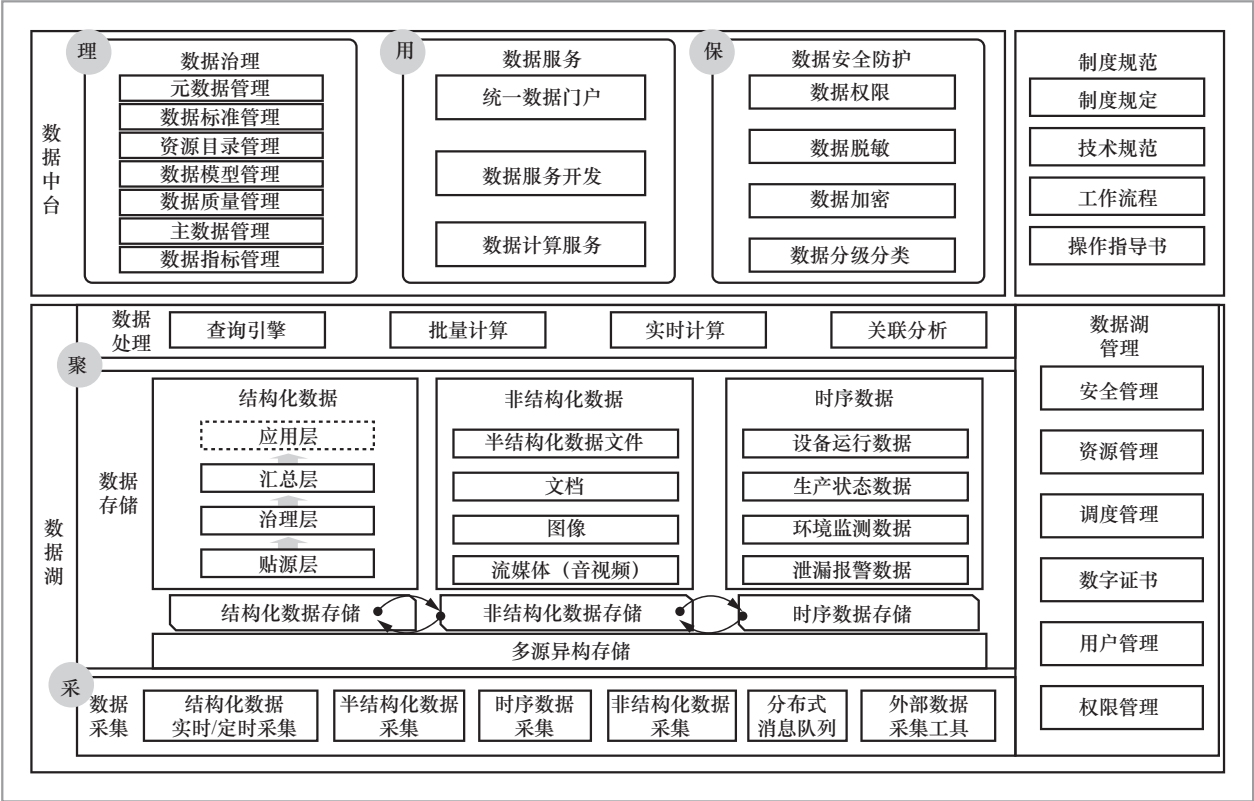


图2 “多湖一中台”工业大数据平台功能架构

高效地处理和存储海量数据，实现数据存储、备份、恢复、清洗等功能，从而保证数据的完整性、可靠性，提高数据处理效率和数据质量。在数据采集方面，支持从各种业务系统中获取原始数据，经过清洗和整合后，传输到数据湖中进行存储和管理。采集数据包括结构化数据、非结构化数据、时序数据等。在数据存储与数据处理方面，数据存储和数据处理是整个平台的核心，经过一系列的数据处理流程，将原始数据转换成有价值的资源。此外，数据湖支持多种数据处理任务，包括批处理、流处理、实时处理等，以满足不同的业务需求。

数据中台主要包括数据治理、数据服务、数据安全等模块。在数据治理方面，数据中台提供了统一的数据治理工具

技术组件，进行数据资产的统一管理和控制，包括元数据管理、数据标准管理、资源目录管理、数据模型管理、数据质量管理、主数据管理、数据指标管理等。在数据服务方面，数据中台通过提供统一的数据服务，主要包括统一数据门户、数据服务开发、数据计算服务等，并提供服务的管控、鉴权、计量等功能。数据服务方式包括查询、分析、挖掘等多种应用模式，以满足不同用户的需求。在数据安全方面，平台采用严格的数据安全保护措施，包括数据分类分级访问控制、加密通信、数据备份等数据全生命周期的安全防护与监测，确保数据的完整性和安全性；通过实施数据安全策略、访问控制、加密传输等手段，为企业提供一个安全可信的数据环境。

2.3 数据架构

如图 3 所示，平台采用“数据就近入湖、一次采集、多点应用”的策略，不仅可以在本地存储和处理数据，还能够实现数据跨湖流动和互联互通，从而更好地支持多湖协同联动和数据统一管控。

集团型企业总部、板块、子企业不同层级数据湖之间进行数据关联计算时，可根据不同应用场景分别采用外表关联（远程虚拟连接）、联邦查询、数据落地（ETL）3 种技术实现。数据服务就近部署，并统一注册到数据中台的服务网关中，实现统一管控。数据消费方在数据中台完成认证和鉴权后，通过路由的方式直接调用数据服务接口，使用数据服务。数据中台同时提供数据服务的开发能力，支持跨湖、跨分区数据服务的开发。生产企业时序数据就近进入本地的企业数据湖或企业采集节点的时序数据库，为本地应用提供近源、低时延数据服务；同时，总部数据湖按需采集时序数据或企业端计算结果数据，降低网络负载。

3 “多湖一中台”工业大数据平台系统实现

3.1 多湖一中台五大核心功能实现

3.1.1 数据采集

多湖一中台通过多种数据采集技术将各领域业务系统中的数据快速、安全地采集到就近的数据湖，包括使用数据库实时复制技术、数据库定时采集技术、非结构化数据采集技术、IoT 数据采集技术、网络爬虫采集技术、API 采集技术、日志采集技术、分布式消息队列技术等。

平台通过捕获变更数据（CDC）来实现数据跨库实时复制（如图 4 所示），适用于对时效性要求很高且对数据库中部分数据表进行选择性的数据同步的场景，其优势是数据高保真，可以捕获数据删除的行为。

平台通过查询 SQL 方式读取数据库表中的数据来实现数据定时采集，通常应用在对时效性要求不高的场合，其优势是易于实现，只要有数据库读写驱动程序，就可以通过编写程序来实现。

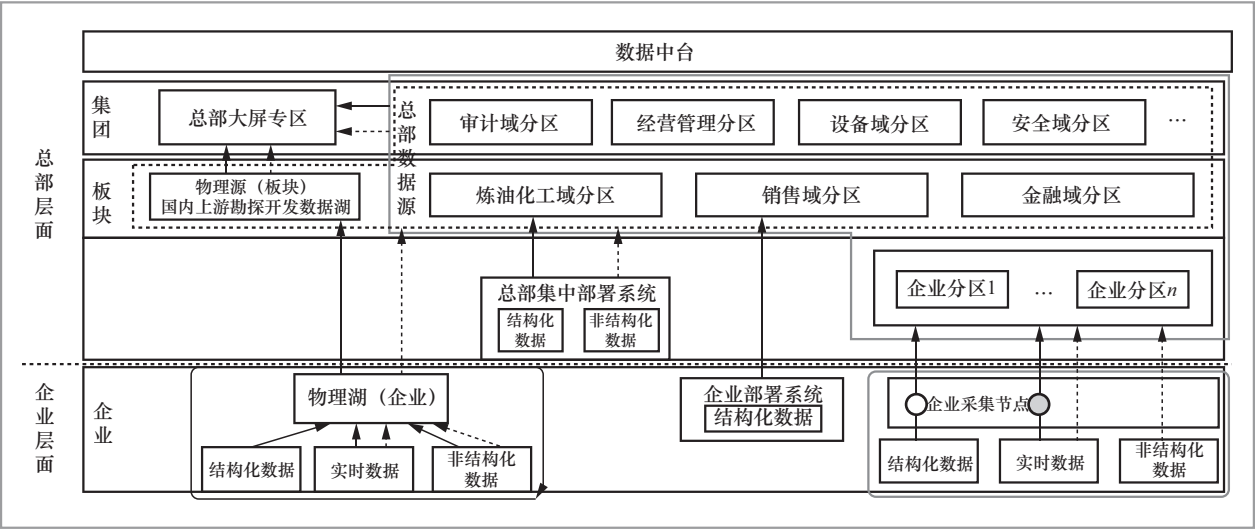


图3 “多湖一中台”工业大数据平台数据架构

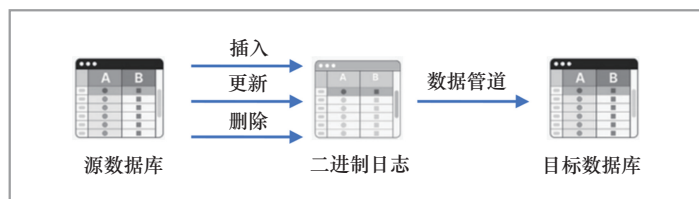


图4 数据实时复制技术

平台可监测指定目录中文件属性的变化，并采集新文件或修改的文件等非结构化数据，根据应用场景需要对文件进行解析处理，提取文件中有用信息。日志采集技术是一种特殊的非结构化文档采集技术，该技术可以将原始日志数据纳入采集范围，如工业生产行为、工业设备与系统状态等。通过分析日志文件将非结构化信息进行结构化，平台可以获取大量有价值的信息，这种方法形成的信息可普遍应用于系统监控、故障排查等领域。日志采集技术是工业大数据平台具体应用的主要支持性技术之一。

平台通过部署在各种设备和环境中的传感器，实时收集各种数据，如温度、湿度、压力、流量、光照等IoT数据。采集到的数据通过物联网网络进行传输并写入时序数据库，确保数据的实时性和准确性。随着物联网的普及应用，这种方式正在成为平台数据采集和应用的关键技术之一。

按照一定的规则，平台可自动抓取万维网信息的程序或者脚本。网络爬虫按照系统结构和实现技术，大致可以分为以下4种类型：通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫。实际应用中，4种爬虫技术通常结合使用。网络爬虫主要用于一些公开数据的自动获取，通常作为企业数据采集的辅助手段。

此外，数据提供方可以通过API方式提供数据，目前数据服务提供方大多通过这种方式来提供数据服务，比如天眼查、

企查查等。

分布式应用间交换信息使用分布式消息队列技术。消息队列可驻留在内存或磁盘上，直到它们被应用程序读走。它通过在生产者和消费者之间传递消息来实现系统的解耦、异步处理以及负载均衡等功能。消息队列可以避免系统因瞬时高峰而崩溃，能够帮助构建更加健壮、可伸缩和易于维护的分布式系统。

不同的数据采集技术分别有其适用的场景，比如时序数据可以用IoT数据采集技术来采集，存入时序数据库；半结构化数据可以通过非结构化数据采集技术进行采集和处理；对于外部数据采集，根据数据提供方的情况采用不同的技术，比如网络爬虫采集技术、API采集技术或数据文件采集技术等。

3.1.2 数据处理

数据处理作为工业要素连接的枢纽，致力于实现上中下游人财物数据的“统一存储、统一处理和统一调度”。多湖一中台采用“大数据Hadoop+分布式数据库+时序数据库+对象存储”等多模态存储技术路线，支持各种数据对象在HDFS、分布式数据库、时序数据库、OSS等组件进行存储，并通过MapReduce、Spark、Impala、Flink、Trino等计算引擎进行一系列的数据处理流程。

数据湖主要包括近源数据区、结构化加工区、非结构化加工区、数据主题区、数据实验室等功能区域。采用Iceberg、Hive、Kudu等Hadoop组件，结合DolphinSchedule调度技术，通过批量加工引擎MapReduce或计算引擎Impala/Trino，对结构化数据进行非实时加工处理；实时数据流处理方面，采用Flink流批一体技术实现准实时加工处理。在非结

构化数据处理方面，通过摘要提取、关键词提取等算法组件，对非结构化文档进行粗加工，为用户提供文档资料的搜索、在线浏览、全文检索和 API 服务共享等操作。基于结构化和非结构化加工区的处理结果，经过聚集、合并等操作，形成面向业务主题的数据集市模型。同时，平台可提供数据实验室功能，对数据主题区的数据进行数据预处理、模型搭建训练、模型性能评估、数据可视化等操作，形成特色算法模型，并向集团内企业进行算法模型的分享。

3.1.3 数据治理

平台通过制定数据治理体系，形成数据治理工作模式和实施方法，沉淀数据资产。在数据治理制度流程体系构建方面，编制了数据治理相关管理制度、流程、规范、模板及指导书，形成“1+6+N”制度体系，即以 1 项《数据资源管理规定》为基础，制定《数据架构管理细则》《主数据管理细则》《数据标准管理细则》《数据质量管理细则》《数据安全治理细则》《数据生命周期管理细则》6 项管理规范，以及一系列（N）工作指南及操作指导书等。在标准规范的建设过程中，结合数据角色的定义及数据分级原则，制定了数据资源发布、数据资源接入、数据资源申请、数据资源加工使用、数据资源维护、数据资源退役、数据服务评价等数据资源应用共享流程，明确了各数据角色在流程中的权责，制定了平台数据的所有权、持有权、管理权、加工使用权的确权规则，明确原始数据、二次加工数据的权属与责任，进一步促进数据共享流通。

多湖一中台的建设实现了数据治理组织、制度、流程、工具等各个方面的系统化落地，实现了数据接入、数据使用等流

程的线上申请和审批。其中，数据治理工具由自开发的数据资源管理工具、数据质量管理工具、指标管理工具和引进的元数据管理工具、数据标准管理工具、数据模型管理工具和主数据管理等工具组成。

3.1.4 数据服务

该平台打造了以“API 服务+数据分析服务+数据订阅服务”为代表的数应用支撑能力。

(1) API 服务

API 服务模块整体分为底层计算服务、API 服务网关、API 开发服务、统一数据门户服务 4 个部分，提供了 API 生命周期管理、API 安全控制、API 多版本管理和 API 的审计与计量功能，具备将数据资源封装为数据服务并提供安全、可追溯的共享使用能力，实现黑白名单、限流限次、流量统计、访问排名、API 活跃度等功能。

(2) 数据分析服务

平台配套的数据分析服务可为业务用户打造一站式数智化分析平台，实现企业数据、信息、知识、智慧的融会贯通和体系化沉淀，使业务用户切实成为数据使用主体，解决数据应用“最后一公里”的难题。该分析服务包括了 AData 全景可视化展示服务、BTable 在线报表服务，以及 CInsight 用户自助分析服务。

(3) 数据订阅服务

数据订阅服务允许用户或应用 App 按需获取多湖一中台里的数据资源。通过数据中台的数据资源目录，用户可以订阅所需的数据资源，经过数据持有者审批同意后，数据中台将根据订阅申请自动为用户推送增量数据。

(4) 跨湖联动

集团的数据资源分布在总部湖、板块

湖、企业湖中，对于跨域融合应用场景，需要多个数据湖参与关联计算，可以根据不同应用场景分别采用外表关联（远程虚拟连接）、联邦查询、数据落地（ETL）3种技术实现。外表关联技术用于需求明确、关联计算频繁的场合，其优势是数据不落地、及时性强。联邦查询技术用于即席查询、需求常变的场合，其优势是分布式任务、执行效率高，可以实现“移动计算，而不移动数据”的目的。数据落地技术用于数据量大且计算复杂的场合，其优势是数据一次性复制，方便数据分析，用户查询使用性能好。

3.1.5 平台安全

“多湖一中台”通过综合运用多种先进的数据安全技术手段，构建了一个安全、可信、高效的数据环境，如图5所示。

平台通过Kerberos、活动目录（active directory）和轻型目录访问协议（lightweight directory access protocol, LDAP）控制用户对数据和资源的访问，研发了大数据统一权限组件，实现了Hive、KUDU、HBase等大数据核心组件的统一权限管理，通过实施精细化的数据访问控制策略，遵循最小化权限原则，严格限制用户访问权限，进行细粒度的数据访问管理。

平台采用密钥管理系统实现密钥的集中管理，集成国产密码服务能力，根据不同的业务需求提供独享或共享的密码资源。针对用户密钥较多、并发量较小的密码应用场景，采用共享密码资源模式接入；针对用户密钥单一、并发量较大的密码应用场景，使用独享密码资源模式接入，也可根据不同的安全需求提供对称密钥加密、非对称密钥加密、数据信封加密、P1签名、P7签名等密码服务，为存储数据提供

保密性和完整性保护，并在数据采集、访问和共享过程中采用安全通信协议，确保数据传输过程安全。

平台可提供数据静态脱敏和数据动态脱敏功能。静态脱敏支持将生产环境的敏感数据脱敏后用于预生产、开发测试等非生产环境，动态脱敏支持SQL语句改写及HTTP应答改写，实现敏感数据访问时实时脱敏。平台实现了API服务接口文档、外发数据文件等与水印服务的功能集成，对共享文档进行保护，实现文档外发的可追溯；在API服务开发和运行环节，还增加了对update、delete等关键词的检查，避免出现数据非法操作或误操作。

平台在数据备份方面采取了本地备份与异地备份的双重保障措施，确保数据在任何情况下都能迅速恢复，保障数据的连续性和可用性。此外，平台内置了数据安全审计机制，实现了对主机、数据库等日志的集中收集、分析，并对日志数据进行6个月以上的安全存储，在实现用户数据操作行为追溯和审计的同时，有效满足了法律法规的相关要求。

在数据要素流通安全方面，平台基于敏感数据监测、数据防泄漏、日志审计等技术手段，建立“横向到边，纵向到底”的数据安全监控运营机制；通过可视化技术将静态资产和动态资产梳理出的信息以可视化的形式呈现，同时基于大数据技术框架和用户行为分析技术，统一发现全局数据泄露风险，构建数据分布情况的监测能力、数据流动监测能力以及用户恶意行为监控能力，实现一站式数据安全风险管理，指导数据安全持续运营工作。数据在流动中创造价值，同时在流动中也面临风险，只有直观地掌握数据流动的情况，才能有效地避免安全隐患。

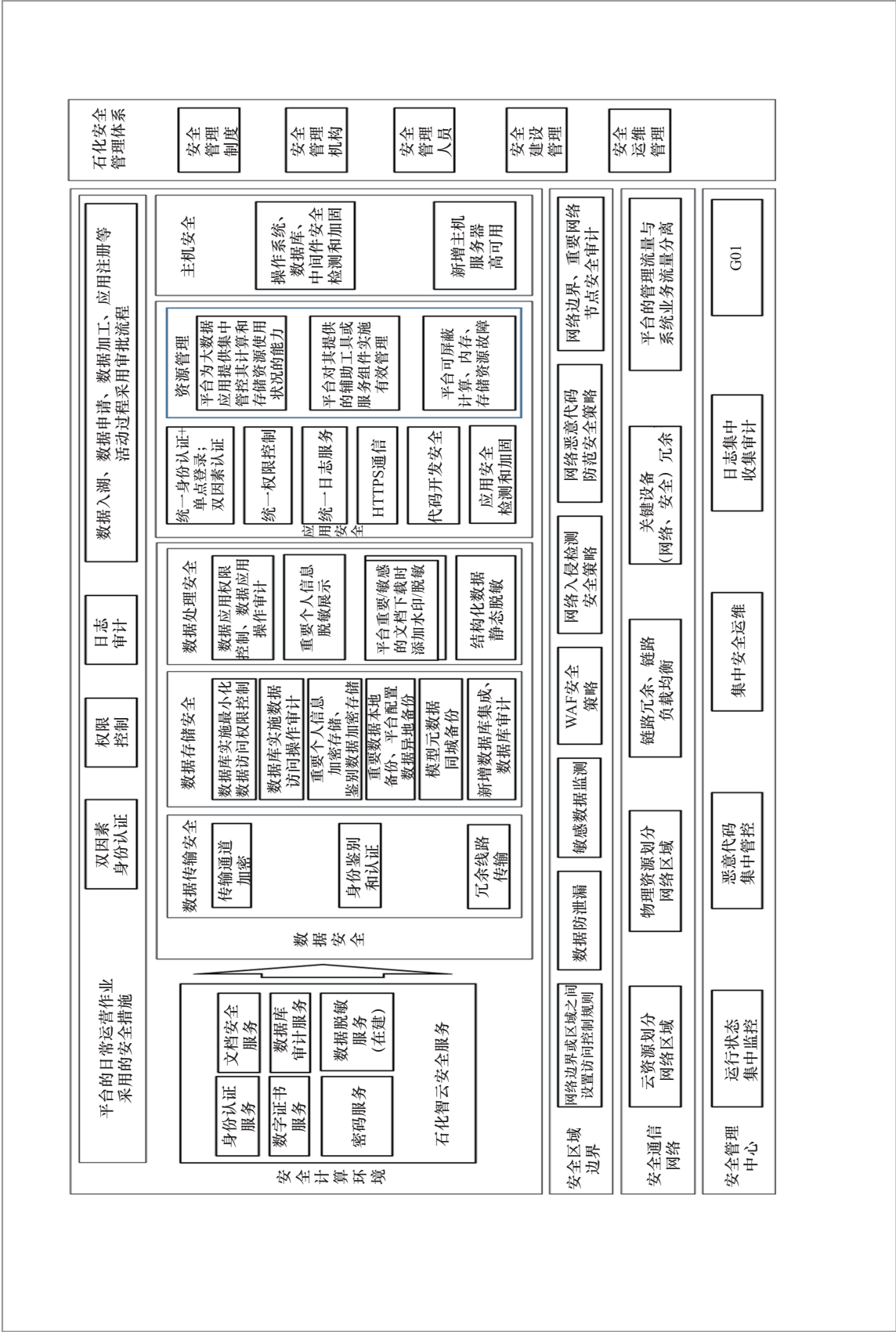


图 5 “多湖—中台”平台安全架构

3.2 “多湖一中台”落地成果

在数据采集方面，某集团“多湖一中台”工业大数据平台接入了19个业务域的500多个源系统，各类数据库表超过3.4万个，涵盖总部统建集中部署、总部统建分散部署、企业自建等多种部署方式。通过外部采购、外部获取等方式，该集团积累了宏观经济、工商企业、市场与行情、市场调研等5类1 200多个外部数据集。经过实际验证，数据集成工具PLT数据同步速度是SAP数据实时复制工具SLT 1.0的10倍以上（如图6所示）。

数据传输工具PLT和开源SeaTunnel对比实验表明，PLT数据同步时效性更好，同步速度更快（见表1）。

“多湖一中台”生产环境完成搭建后，对各项功能进行了测试。在性能测试方面，HDFS写、读能力分别可以达到3.01 Mbit/s、7.02 Mbit/s（见表2）；100 GB数据通过MapReduce的方式排序耗时184 s（如图7所示）；对1 000个文件的创建、读取、删除分别耗时约94 s、74 s、17 s（见表3）。

随着系统的持续使用、数据的持续增加，原有节点数量不能满足使用需求。基于湖仓一体架构的“多湖一中台”具有良好的可扩展性，集群按需进行了扩展，从最初的10节点逐步扩展至25节点。根据测试环境验证，Kudu资源从3节点增加到6节点时，高并发下的查询性能会有一定程度的提升（如图8所示）。

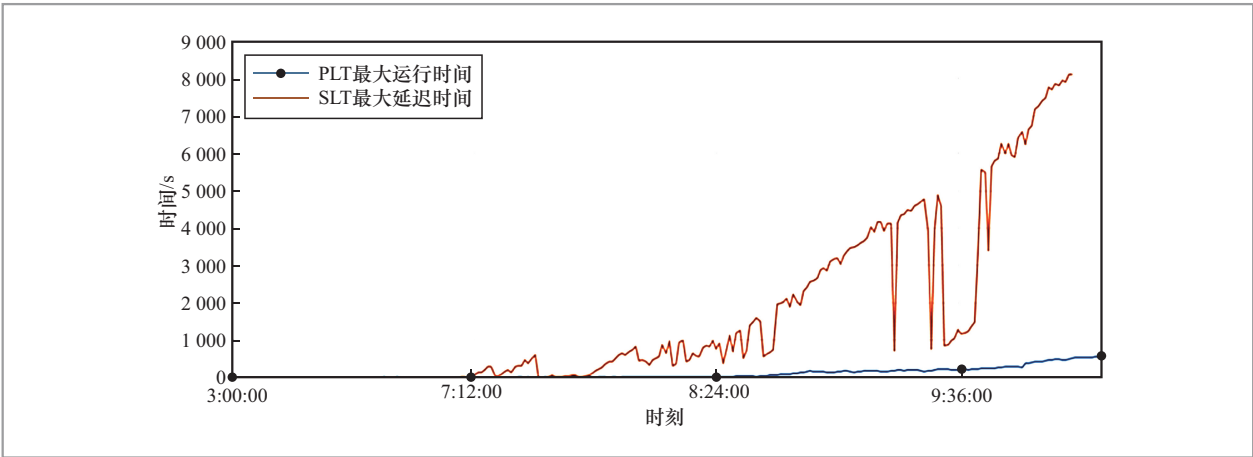


图6 数据传输工具PLT与SLT对比

表1 数据采集工具性能对比表

序号	名称	准确性	时效性	性能影响	同步速度
1	数据传输工具PLT	100%	+1 s	CPU:13%; 内存:65%; I/O:读1.2 Mbit/s,写5 Mbit/s	初始化/增量:2.030条/秒
2	数据集成平台SeaTunnel	100%	+2 s	CPU:10%; 内存:62%; I/O:读1 Mbit/s,写4.8 Mbit/s	初始化:960条/秒; 增量同步:2 040条/秒

表2 HDFS读写性能							
用例编号	用例说明	实际结果					
		文件数量/个	总处理数据 大小/MB	吞吐量/ (Mbit·s ⁻¹)	平均I/O速率/ (Mbit·s ⁻¹)	I/O速率标准 偏差	测试执行 时间/s
TC-01-01	HDFS文件 系统写性能	10	1 280	46.29	46.9	5.92	29.59
		10	10 240	95.08	95.26	4.16	38.46
		100	12 800	27.34	31.1	15.92	34.66
		100	102 400	44.63	44.9	3.65	52.67
		1 000	128 000	12.72	18.53	18.29	117.26
		1 000	1 024 000	3.01	3.45	1.49	1 474.43
TC-01-02	HDFS文件 系统读性能	10	1 280	239.88	445.81	422.91	27.44
		10	10 240	154.4	158.14	27.07	31.45
		100	12 800	181.89	421.71	435.35	28.68
		100	102 400	112.85	246.48	380.39	38.63
		1 000	128 000	89.48	238.53	257	73.24
		1 000	1 024 000	7.02	146.64	279.42	805.3

表3 1000个文件操作性能									
用例编号	用例说明	实际结果							
		任务执行 数/个	Reduce任务 执行数/个	块大小/ byte	文件数 量/个	复制 因子	原始数据: TPS总数/ms	原始数据:最长 Map时间/ms	原始数据: 异常
TC-02-01	批量创建						93 779	7 952	
TC-02-02	批量读取	12	6	1	1 000	3	74 015	6 287	12 000
TC-02-03	批量删除						17 374	1 500	

在治理方面，该平台实现了数据接入、数据使用等核心流程的在线化。截至目前，接入申请单1 078个，使用申请单1 590个，数据资源发布超过3万多个，形成了集团统一的数据资源目录，实现了数据资源的可视、可用、可管，解决了数据查找难、共享难等问题，推动了数据在集团内、外共享流通及数据要素价值释放。

在数据服务方面，该平台发布API服务8 500多个，服务用户3万余名，数据服务累计超过27.5亿次，支撑总部、事业部及企业三级数据应用共计220余个。每日API调用次数达到150万次左右，响应时间为1 s左右。

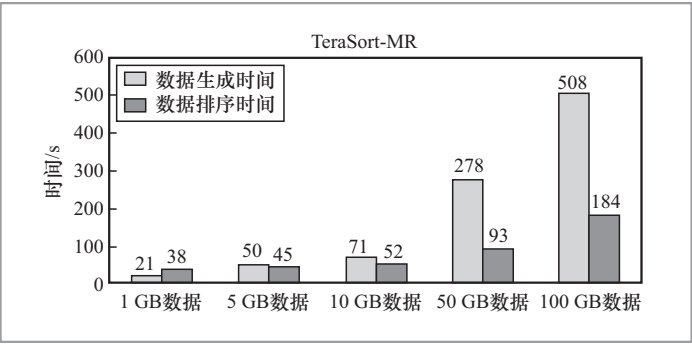


图7 MR排序性能

在数据安全方面，平台采用了数据脱敏、数据加密、权限管理、安全审计等方式确保数据安全。其中，数据脱敏能力与数据库有一定的关系，KUDU数据库脱敏

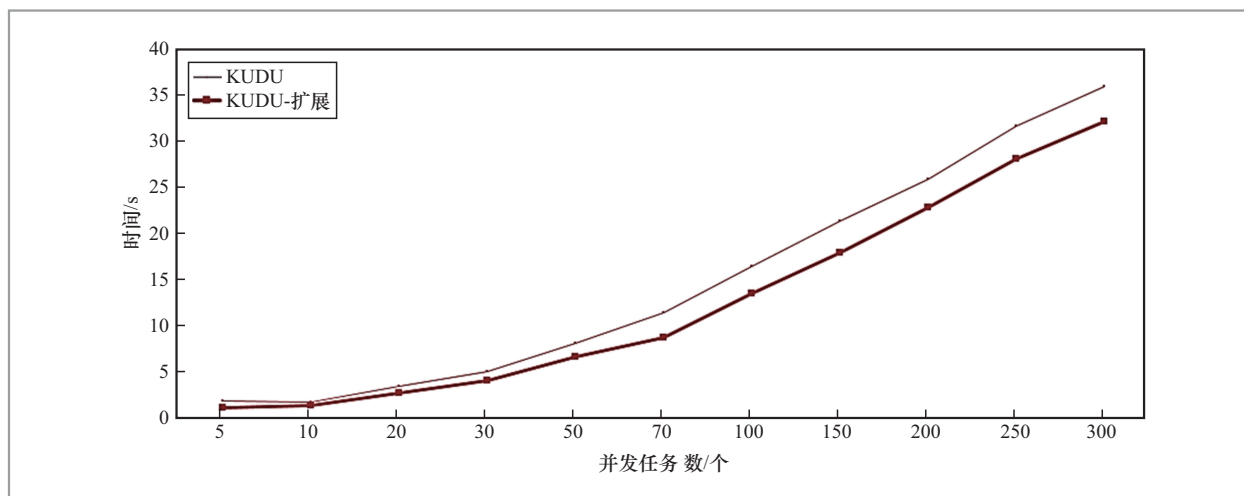


图8 KUDU扩展性测试

速度为可实现2 500条/秒，HANA 数据库脱敏速度可达到22 569条/秒。在数据流动监测方面，结合日志审计、态势感知、敏感数据监测等多种技术手段，平台日均处理安全威胁告警9 000个，为数据流通提供安全保障。

4 基于“多湖一中台”的大数据应用实践

基于该平台构建多业务场景主题数据模型库及数据治理核心能力，遵循统一数据治理方法，支撑集团型企业总部、板块各事业部和下属企业各类数据应用建设，营造多样繁荣的数据应用生态。

（1）产品数据分析

基于工业大数据平台，采用湖仓一体与实时数据湖改变了原有的数据处理模式，借助数据变更捕获（CDC）技术，将数据湖中的数据近实时处理，缩短了数据处理周期，提高了数据处理效率。以调度日报为例，之前采用批处理将多家炼厂生产产品的数量进行计算，一家炼厂生产的产品

只能第二天才能计算出来，板块汇总43家炼厂数据，需要3~4天。借助湖仓一体技术，集团能够实时查询各分公司生产情况，在大屏端实时展示全集团生产情况，更有利于产品生产的资源调配；同时总部数据分析人员借助数据湖的联邦计算能力，在数据湖中进行生产分析和参数优化，优化调配现场生产，提高了生产效率，从而使原料加工达到最优化。

（2）设备预警与故障诊断

构建设备功能图谱，在过往故障案例数据的基础上，提取故障特征，形成故障模式库，将离心机组振动数据和工艺数据进行处理，应用预警模型，推送预警信息，实现对设备的远程监测及优化。平台在瞬态数据记录密度、动态阈值智能预警、联锁保护耦合功能等多方面效果显著，且数据、算法自主可控。利用机理模型对关键设备进行建模，实现在线的资产绩效管理，在线评估塔板分离效率、换热器换热效率、压缩机性能、仪表测量偏差，确定处置措施，为提升装置运营效率、节能降耗提供数据基础。实现装置核心设备监测不仅针对设备本体，以及设备

故障预测和早期故障诊断，沉淀模型、案例库等，还为大数据分析提供数据基础，实现故障诊断和工艺耦合，技术、数据和算法融合，进一步提升设备状态的监测、预警及优化性能。

(3) 炼化业财融合

平台在炼化业财融合数据应用方面，建立了一套较为完善的炼化指标体系，实现了炼油化工、生产营运和财务管理三大业务域的跨域融合，为困扰炼化企业多年的装置效益问题提供灵活的分析手段，通过装置模型的定义实现装置的日效益分析、效益测算等分析功能，细化管理力度，改变效益分析模式，为企业实时了解经营效益情况，制定降本增效的生产策略提供了有力支撑。平台共集成三大业务域20余个业务核心系统，形成共性功能116个，沉淀指标328个，共性资源的复用率高达80%以上，减少三分之一的实施周期和项目成本。

5 结束语

“多湖一中台”工业大数据平台在某集团的研究和应用实践涵盖了数据采集、传输、存储、处理、开发和使用的全过程，覆盖了数据采集能力、数据整合处理能力、数据治理能力、数据服务能力、数据安全防护五大能力，实现了工业数据的采集入湖、多源数据整合、数据资源盘点与发布、数据服务的统一管理，有效解决了超大型集团企业之间、总部与企业之间跨组织、跨层级、跨区域的数据共享和高效流通的需求。

基于湖仓一体技术“多湖一中台”工业大数据平台的研究和应用有效实现了集团内部的数据共享和流通，未来需要结合

隐私计算、区块链确认等技术，有效保障数据安全，并实现数据要素的社会流通。

参考文献：

[1] 王建民. 工业大数据技术综述[J]. 大数据, 2017, 3(6): 3-14.
WANG J M. Survey on industrial big data [J]. Big data research, 2017, 3(6): 3-14.

[2] 黄家凯, 秦丽娟, 郑诗语, 等. 基于数据湖技术的地质大数据底座架构研究与应用[J]. 自然资源信息化, 2024(2): 43-49, 68.
HUANG J K, QIN L J, ZHENG S Y, et al. Research and application of geological big data base architecture based on data lake technology[J]. Natural Resources Informatization, 2024(2): 43-49, 68.

[3] 胡军军, 谢晓军, 石彦彬, 等. 电信运营商数据湖技术实施策略[J]. 电信科学, 2019, 35(2): 84-94.
HU J J, XIE X J, SHI Y B, et al. Implementation strategy of telecom operators' data lake technology[J]. Telecommunications Science, 2019, 35(2): 84-94.

[4] 周大鸿, 王红迁, 黄艺璠, 等. 医院数据湖系统设计与应用[J]. 中国数字医学, 2024, 19(4): 100-104.
ZHOU D H, WANG H Q, HUANG Y F, et al. Design and application of hospital data lake system[J]. China Digital Medicine, 2024, 19(4): 100-104.

[5] 王少杰. 基于数据湖的联邦数据管理系统的设计与实现[D]. 北京: 北京交通大学, 2020.
WANG S J. Design and implementation of federated data management system based on data lake[D]. Beijing: Beijing Jiaotong University, 2020.

[6] Databricks. Data Lakehouse[EB]. 2022.

[7] 陈氢, 宋仕伟. 数据治理视角下的湖仓一体架构研究[J]. 数字图书馆论坛, 2023, 19(4): 19-28.

- CHEN Q, SONG S W. Lakehouse architecture under perspective of data governance[J]. Digital Library Forum, 2023, 19(4): 19-28.
- [8] 李国华, 邹丹, 李海军, 等. 铁路数据分布式湖仓一体架构分析与设计[J]. 现代信息技术, 2024(1): 54-58.
- LI G H, ZOU D, LI H J, et al. Analysis and design of railway data distributed lake warehouse integrated architecture [J]. Modern Information Technology, 2024(1): 54-58.
- [9] 朱一洲, 吴莉珊, 谭英嘉. 基于数据中台的企业数据治理路径研究[J]. 信息技术与标准化, 2024(5): 75-79.
- ZHU Y Z, WU L S, TAN Y J. Research on enterprise data governance path based on data platform[J]. Information Technology & Standardization, 2024(5): 75-79.
- [10] 范欣, 吴先智, 王南南, 等. 商业银行数据中台实践[J]. 信息技术与标准化, 2024(5): 13-18.
- FAN X, WU X Z, WANG N N, et al. Practice of data hub in commercial banking[J]. Information Technology & Standardization, 2024(5): 13-18.
- [11] 杜啸争, 王笑非, 李娜, 等. 企业数据资产运营平台建设实践[J]. 大数据, 2024, 10(2): 17-31.
- DU X Z, WANG X F, LI N, et al. Construction of enterprise data asset operation platform[J]. Big Data Research, 2024, 10(2): 17-31.
- [12] 赵宏博, 刘伟, 李永杰, 等. 基于炼铁大数据智能互联平台推动传统工业转型升级[J]. 大数据, 2017, 3(6): 15-26.
- ZHAO H B, LIU W, LI Y J, et al. Transformation and upgrade of the traditional industry based on big data intelligent interconnection platform for iron-making[J]. Big Data Research, 2017, 3(6): 15-26.
- [13] 王冀彬, 杨海龙, 冯凯, 等. 面向大数据场景的系统性能优化实践[J]. 大数据, 2024, 10(4): 21-33.
- WANG J B, YANG H L, FENG K, et al. System performance optimization practice for big data scenarios[J]. Big Data Research, 2024, 10(4): 21-33.
- [14] 程伟, 马成, 凌捷. 大数据技术在数据安全治理中的应用[J]. 大数据, 2023, 9(6): 3-14.
- CHENG W, MA C, LING J. Application of big data technology in data security governance[J]. Big Data Research, 2023, 9(6): 3-14.

作者简介



蒋楠（1981-），男，博士，中国石油化工集团有限公司信息和数字化管理部副总经理、高级工程师，主要研究方向为人工智能、数据治理及数据集、基础设施及网络安全、工业互联网平台等。

收稿日期: 2024-10-28