耿恒高,彭苏萍,王占刚,等. 煤矿地学大数据智能湖仓系统建设与应用[J]. 矿业科学学报,2025,10(1):1-13. DOI:10.19606/j.cnki.jmst.2024940

GENG Henggao, PENG Suping, WANG Zhangang, et al. Construction and application of lakehouse system for coal mine big earth data[J]. Journal of Mining Science and Technology,2025,10(1):1-13. DOI:10.19606/j.cnki.jmst.2024940

煤矿地学大数据智能湖仓系统建设与应用

耿恒高[1,2],彭苏萍[1,2],王占刚[1,2],许娜[1,2],许献磊[1,2],杜文凤[1,2]

1. 中国矿业大学(北京) 地球科学与测绘工程学院,北京 100083;
2. 煤炭精细勘探与智能开发全国重点实验室,北京 100083

摘 要:为满足煤矿智能化发展过程中对海量、多源、异构数据高效管理与应用的需求,提出基于多种大数据技术的煤矿地学大数据智能湖仓系统。针对多源异构煤矿地学数据的高效索引需求,提出基于希尔伯特曲线与 GeoHash 编码的时空分类方法,将时间、空间、分类等多维属性进行统一编码,降低数据索引维度,提升数据检索效率。智能湖仓系统的构建不仅提高煤矿地学数据的整合与利用效率,还为煤矿企业的数字化转型与透明化发展提供重要实践经验,推动煤矿行业向智能化、透明化、高效化方向迈进。

关键词:煤矿智能化;地学大数据;大数据技术;智能湖仓

中图分类号:TD 311 **文献标志码**:A **文章编号**:2096-2193(2025)01-0001-13

Construction and application of lakehouse system for coal mine big earth data

GENG Henggao[1,2], PENG Suping[1,2], WANG Zhangang[1,2], XU Na[1,2], XU Xianlei[1], DU Wenfeng[1,2]

1. Couege of Geoscience and Surveying Engineering, China University of Mining and Technology-Beijing, Beijing 100083, China;
2. State Key Laboratory for Fine Exploration and Intelligent Development of Coal Resources, Beijing 100083, China

Abstract: In order to meet the needs of efficient management and application of massive, multi-sourc and heterogeneous data in the process of intelligent development of coal mines, this paper proposes an intelligent lakehouse system for coal mine big earth data based on multiple big data technologies. In response to the needs in efficient indexing of multi-source heterogeneous coal mine geological data, this paper proposes a spatiotemporal classification method based on Hilbert curves and GeoHash coding, which uniformly encodes multidimensional attributes such as time, space and classification, reduces the data index dimension, and improves retrieval efficiency. The construction of the intelligent lake house system not only optimizes the integration and utilization efficiency of coal mine geological data, but also provides important practical experience for the digital transformation and transparent development of coal mining enterprises, and promotes the coal mining industry to move towards intelligence, transparency, and high efficiency.

Key words: intelligent coal mine; big earth data; big data technology; lakehouse

收稿日期:2024-06-30 **修回日期**:2024-11-26

基金项目:国家重点研发计划(2023YFC3008904,2022YFF1303302);中央高校基本科研业务费专项资金(2023QNJS098)

作者简介:耿恒高(1993—),男,江苏灌云人,博士研究生,主要从事地球物理、计算机、大数据与地球科学交叉方面的研究工作。Tel:15351818127, E-mail:genghenggao@outlook.com

通信作者:彭苏萍(1959—),男,江西萍乡人,中国工程院院士,主要从事煤矿工程地质和工程物探方面的研究工作。E-mail:psp@cumtb.edu.cn

煤炭是我国能源结构的重要组成部分,煤炭工业的稳定健康发展对实现国家可持续发展战略具有不可替代的作用。随着全球能源转型和可持续发展战略的不断推进,煤炭行业面临着更高的发展需求和更严格的环保要求。近年来,随着科学技术的迅猛发展以及国家政策的强力推动,煤矿行业正逐步向智能化和透明化转型,以适应新时代发展需求。国家已经明确提出,将 5G 通信、人工智能、工业物联网、云计算、大数据、区块链、大模型等前沿技术与现代煤炭开发利用深度融合,推动煤矿生产过程的智能化运作,促进煤炭行业的高质量发展。

在这一背景下,煤矿智能化发展被赋予实现“矿井地质透明化”的核心任务,以推动煤炭资源的绿色开采和精准利用[1-3]。煤矿智能化作为煤炭工业高质量发展的重要阶段,其建设的关键之一在于深度整合和高效利用海量、多源、异构的煤矿地学大数据[4-6]。然而,当前煤矿企业对数据的管理和利用面临诸多挑战,包括数字化转型不充分、数据利用率低、业务之间缺少联动以及存在“数据孤岛”和“数据烟囱”现象,严重制约煤矿地学大数据的有效整合和挖掘分析。

煤矿地学大数据不仅是行业数字化转型的重要基础,也是驱动煤炭行业生产方式深刻变革的新质生产力。在煤矿智能化开采过程中,构建以地学数据要素为核心的煤矿数智体系已逐渐成为行业发展的重要方向。面对海量复杂的煤矿地学数据,亟须一种科学、系统且高效的存储管理方案,来实现煤矿地学数据的全生命周期管理与智能化应用,以满足多源异构数据的协同联动需求,打破“数据孤岛”问题,释放煤矿地学数据价值。

1 煤矿多源地学数据要素研究基础

随着煤矿行业智能化转型的加速,与先进技术的深度融合已成为行业发展的关键驱动力。特别是 5G 通信、人工智能、工业物联网、云计算、大数据、区块链、大模型、机器人以及智能装备等前沿技术的广泛应用,正深刻改变煤矿行业的生产和管理方式。这些技术不仅推动煤矿生产过程的智能化转型,而且促进煤矿多种探测技术的革新和智能化多系统平台的快速发展。

1.1 煤矿多场多源探测技术

煤矿智能化开采要求对煤矿地质环境的变化进行全方位、实时、精准的监测,并且需要综合利用多个时空尺度的变化规律和相互作用。这一过程不仅依赖于单一的技术手段,还需要多学科、跨领域的综合支撑,特别是多过程、长时间序列科学数据的支持[7-8]。煤矿探测技术作为支撑煤矿智能化开采的关键环节,涵盖多个领域和多种技术手段。当前,煤矿探测技术主要包括地面三维地震勘探技术[9-11]、地面电磁法勘探技术[12]、井下地球物理勘探技术[13]、随掘探测技术、随采探测技术[14]等多种探测技术。随着科学技术的进步,各类勘探传感器飞速发展,煤矿地学数据采集和传输技术取得显著突破,各种类型的勘探数据量呈现指数增长趋势,可达到 TB、PB 级规模,是典型的大数据[15]。图 1 为多种采集装置和仪器。

图 1 多种采集装置和仪器

Fig. 1 Various data acquisition devices and instruments

多种采集手段产生海量复杂的煤矿地学数据,包括地质调查、钻探、三维地震、物探、化探、抽水试验、采样测试、测量变形、位移和地表沉陷等数据,这些数据提供丰富的地质信息、地球物理属性、地下精细化结构等信息。煤矿地学大数据除了具有海量、多样、高速、价值密度低及真实性等常规大数据特点外[16-17],还具有多源异构、多时间、多标准、高维、高度复杂和不稳定性等独特的属性和空间位置特征,这些特点使得煤矿地学数据的存储、处理和分析面临前所未有的挑战[18-20]。图 2 为常规大数据的特点。

图 2　常规大数据的特点

Fig. 2　Characteristics of conventional big data

1.2　煤矿地学信息管理平台

面对煤矿地学数据的海量与复杂性,研究人员广泛开展煤矿数据管理系统的相关研究。毛善君等[21]基于 GIS 空间数据库,实现煤矿空间信息的集成化管理,并成功开发出具有完全自主知识产权的煤矿专用地理信息系统服务平台;康红普等[22]采用 MySQL 关系型数据库管理数据,研究中国煤矿井下地应力数据库的构建及其分布规律。疏礼春[23]基于数据中台架构,提出一种涵盖煤矿数据汇聚、数据开发、数据存储与数据资产管理的系统化方案;许娜等[24]基于 MongoDB 文档型数据库搭建集群,研究地震勘探数据管理系统;王霖等[25]深入探讨智能化煤矿数据仓库的建模方法,并构建智能煤矿数据仓库的分层架构,实现多业务系统数据的统一组织与管理,有效解决“数据孤岛”问题。针对传统煤矿数据中心存在各类数据离散存储、数据集成困难以及业务应用和数据分析难度大的问题,韩安[26]提出基于 Hadoop 分布式文件系统实现数据存储可靠性,同时采用 HBase 分布式数据库持久化存储历史数据,利用 Redis 实现实时数据的存储与最新数据的快速检索;廖志伟等[27]在大型智能矿山建设中,针对数据价值的挖掘,提出智能矿山数据标准体系与数据湖架构,并构建基于微服务平台的智能决策分析系统、智能经营管理系统、智能矿山生产执行系统及智能安全生产集成监测系统等。

当前我国煤矿智能化建设已经步入透明工作面 3.0 时代,多个煤矿信息化平台与系统逐步建设并投入使用。例如,国家能源、中煤能源、山东能源、晋能控股和陕煤集团等单位持续推进数字化与智能化建设,已建成并投入运营的一系列煤矿基础应用平台及子系统,涵盖掘进系统、开采系统、综合监控平台、瓦斯抽采效果评价系统、自动化信息平台等多个领域,形成包含近百个子系统的庞大信息化体系[5,28]。其他企业包括北京龙软、中煤科工西安研究院、山东蓝光及网格天地等相应开展煤矿相关应用平台建设,其中,中国地质调查局的地质云平台[29]、中煤科工的“煤智云”[30]均是行业中较具代表性的案例。

1.3　煤矿地学数据管理需求分析

1.3.1　当前煤矿地学数据管理模式的痛点

基于数据库管理系统和煤矿地学信息管理平台的研发与应用已取得显著进展,在一定程度上实现了某种类型的煤矿地学数据的数字化、标准化存储和流程化管理,显著提高了煤矿地学数据查询的便捷性。然而,当前煤矿地学数据的管理模式仍存在痛点[31],具体表现在以下 4 个方面。

(1) 数据存储标准缺乏,兼容性差。煤矿地学数据来源于多种勘探设备,而这些设备使用不同的操作系统和数据存储标准,导致数据存储格式不统一。缺乏统一的数据标准和元数据规范,造成数据格式之间的兼容性差异,进而使得多源异构数据的共享与协同分析变得困难,限制数据的整合与智能分析。

(2) 数据质量不高,数字化转型滞后。煤矿地学数据尚未建立完善的数据标准,数据的准确性、完整性和一致性难以保障。一方面,许多元数据缺乏必要的属性填充,部分煤矿地学数据缺失注释和精度等信息;另一方面,诸如图纸这一类型的历史数据由于数字化程度较低,面临老化问题,重要信息可能丢失。此外,仍有相当一部分煤矿地学数据依赖传统纸质记录,无法高效存档和检索,进一步影响数据的使用与分析。

（3）系统响应迟缓，数据需求响应能力不足。现有煤矿信息化平台多为早期建设，缺乏持续更新与维护，系统响应速度较慢，数据需求的响应时间过长。大部分基于传统关系型数据库的管理系统难以应对日益增长的数据量，扩展性和实时访问性能不足，无法满足大规模、动态地学数据的存储与快速访问需求。

（4）跨部门间协作困难，数据孤岛问题突出。不同部门在地学数据管理上缺乏统一的平台，数据在各系统之间相互隔离，甚至存在物理隔离现象。这类数据流通方式主要依赖磁盘拷贝，造成信息传递不畅，无法及时支持数据的动态处理与实时响应需求。

1.3.2 煤矿智能湖仓系统建设需求

针对煤矿地学数据管理模式中存在的上述痛点，传统的数据存储技术已无法满足当前煤矿智能化转型需求，亟须构建一种更加灵活、可扩展、智能化的数据管理系统。提出煤矿地学大数据智能湖仓系统，旨在通过构建统一的数据架构和先进的技术手段，解决现有数据管理中的主要问题，提升煤矿地学数据的整合与应用效率。系统建设需具备以下 4 个关键能力。

（1）煤矿多源地学大数据的汇聚集成能力。系统需要建立高效、规范的数据接入机制，支持多种数据来源的无缝接入，并利用数据治理技术进行自动化清洗、标准化和数据质量检测，确保数据的一致性和可用性。

（2）煤矿异构地学元数据的统一管理能力。系统需要打通元数据，实现煤矿地学数据全流程统一的元数据管理，确保各种现有数据库和数据仓库的元数据无缝打通及统一管理。

（3）煤矿多样信息巨系统的协同开发能力。系统支持批数据和流式数据多引擎计算，实现数据的统一开发、智能调度和数据治理，构建一站式、全托管、云原生与智能化的数据平台。

（4）煤矿海量地学大数据的存储分析能力。系统应如同数据湖一样具备可扩展性、经济高效性和灵活性，确保平台具有企业级高性能、稳定性和可靠性，支持存储、服务、计算的弹性伸缩，满足煤矿企业对系统可用性的高要求。

2 煤矿异构地学数据存储关键技术

海量集聚的煤矿地学数据作为最具时代特征的生产要素，为煤矿智能化发展带来新的机遇，并成为推动煤矿智能化发展的新质生产力。煤矿智能化发展需要新的数据存储管理思维。多年来，大数据技术的发展为建设煤矿地学大数据智能湖仓系统提供强有力的支撑。煤矿智能湖仓系统将严格按照煤矿智能化数据标准体系[32]，探索新的架构模式，以满足煤矿地学数据在存储、管理、分析应用和智能决策等方面的需求。该系统能够适应不同存储格式、不同数据类型与不同平台数据库的要求，通过统一的标准和规范，实现数据的高效管理和利用，推动煤矿智能化向更高水平发展。

2.1 煤矿智能湖仓技术选型

在存储管理海量煤矿地学大数据过程中，大数据技术组合应用显得尤为重要。传统的数据库及数据仓库因扩展性有限、数据存取效率低、性能不足，已无法满足当今煤矿地学数据的管理需求。智能湖仓系统将数据湖和数据仓库结合起来，有效实现数据的统一存储和管理，能够提供高效的数据存取和分析服务[33]。在应对海量异构的煤矿地学数据方面，智能湖仓系统具备水平扩展和高效存储的优势，支持高效地查询和分析结构化数据，确保数据的准确性和一致性。

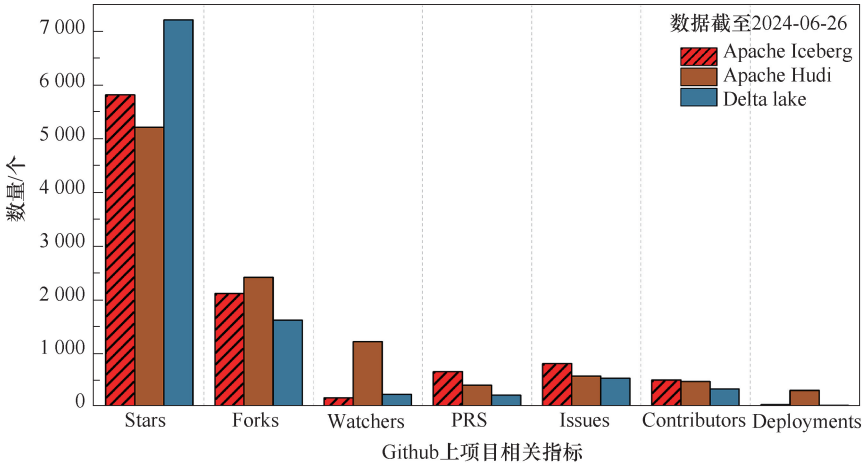


图 3 3 大框架 Github 统计信息

Fig. 3 Statistics of three frameworks on Github

当前可以建设煤矿地学大数据智能湖仓系统的主流框架有 Apache Iceberg、Apache Hudi 和 Delta Lake[34-35]。通过对比 3 大框架在 Github 社区的发展情况(图 3)、功能特性(表 1)、统计和分析 Stars、Forks、Watchers 等参数,结果表明,Apache Iceberg 的社区总体活跃度优于 Apache Hudi 和 Delta Lake,且具备与其类似的功能特性,基本满足煤矿地学大数据的存储和管理需求。

表 1　3 大框架功能特性

Table 1　Features of three frameworks

特性	Apache Iceberg	Apache Hudi	Delta Lake
数据模型	COW/MOR	COW/MOR	基于行的更新
文件格式	Parquet、Avro、ORC	Parquet	Parquet、ORC
事务支持	支持 ACID 事务	支持 ACID 事务	支持 ACID 事务
计算引擎	Spark、Flink、Trino	Spark	Spark
易用性	较高,SQL 查询支持	提供 API	使用 Delta 格式
底层存储	HDFS、S3、OSS 等	HDFS、S3、OSS 等	HDFS、S3、OSS 等
实时处理	支持	支持	部分支持
时间旅行	支持	支持	支持
社区生态	社区活跃,贡献者众多	社区活跃,应用广泛	以 Databricks 主导
适用场景	多样化数据湖,历史数据查询、大数据分析	实时数据处理、数据湖管理	强大的 ACID 事务、混合流处理和批处理

比较 3 大框架日志结构表在 Spark 上写时复制(Copy-on-Write, CoW)和读时合并(Merge-on-Read, MoR)等场景下的迭代执行时间,在规模因子为 100 的情况下,按照国际事务性管理委员会制定的标准规范(Transaction Processing Performance Council Decision Support Benchmark, TPC-DS)进行测试分析。图 4 为单用户只读和数据维护读写在 Spark 上迭代的执行时间比率。由图 4 可见,在单个用户只读和数据维护读写 2 种场景下,迭代过程中 Apache Iceberg 响应性能优于 Delta Lake,低于 Apache Hudi。

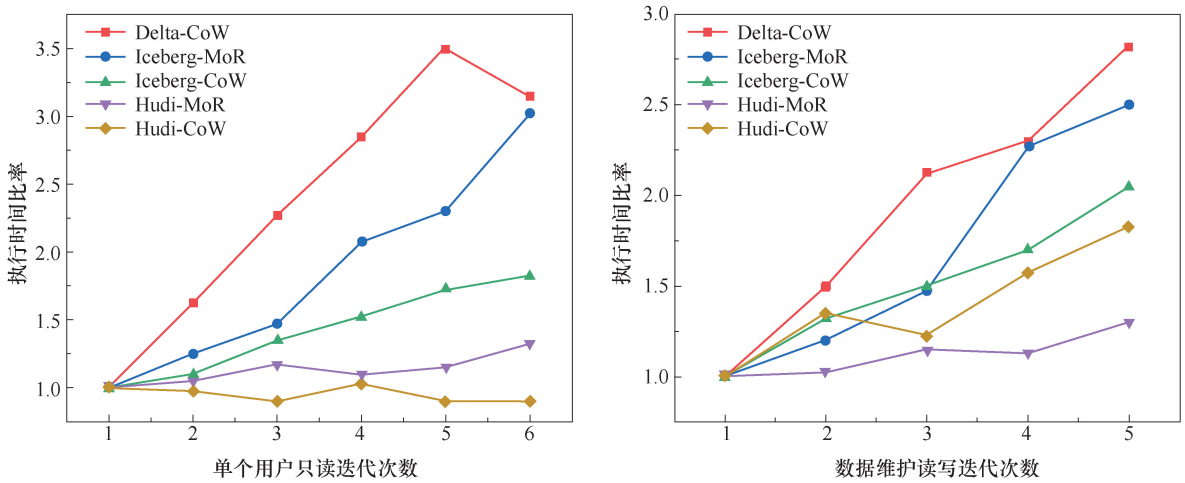


图 4　单用户和数据维护在 Spark 上迭代的执行时间

Fig. 4　Execution time of single user and data maintenance iterations on Spark

通过统计和分析 3 大框架的社区活跃、相关特性和性能,最终选用 Apache Iceberg 作为煤矿地学大数据智能湖仓系统的数据存储服务。Apache Iceberg 由 Netflix 开发,并在 Apache Software Foundation 的孵化器中维护,专为大规模数据集提供可靠、高效的存储和查询功能。其关键特性包括支持无中断的模式演化和动态分区策略,简化用户操作的隐藏分区,高可靠性的 ACID 事务,便于数据审计的时间旅行功能,高效的元数据处理。Apache Iceberg 具备良好的性能、可扩展性和灵活性,能

够满足煤矿智能化系统对数据存储、管理和分析的高标准要求。另外,Apache Iceberg 还兼容多种大数据处理引擎,如 Apache Spark、Apache Flink 和 Trino,确保能无缝集成到现有的大数据生态系统中。这些特性能够有效地克服传统数据存储方案在一致性、性能和易用性方面的挑战,是构建现代化、高效、可靠煤矿数据湖仓系统的理想选择。

2.2 煤矿智能湖仓总体架构

煤矿智能湖仓系统的设计目标是构建“多部门联动、全过程管控”的管理模式。通过这一系统,煤矿各部门能够实时共享数据、协调工作,提升煤矿生产的安全性和效率。其核心在于通过信息化手段实现数据的无缝连接和高效利用,从而为煤矿生产的各个环节提供及时、准确的数据支持,优化决策过程,减少潜在风险。因此,采用 Apache Spark、Apache Flink、Apache Zookeeper、Apache Kafka、HDFS 等多种大数据技术,在 Apache Iceberg 基础上构建一个集成化、智能化的煤矿地学大数据湖仓系统。图 5 为煤矿智能湖仓系统架构。其不仅构建一个新的统一数据存储层,还充分考虑与现有平台系统的兼容性。

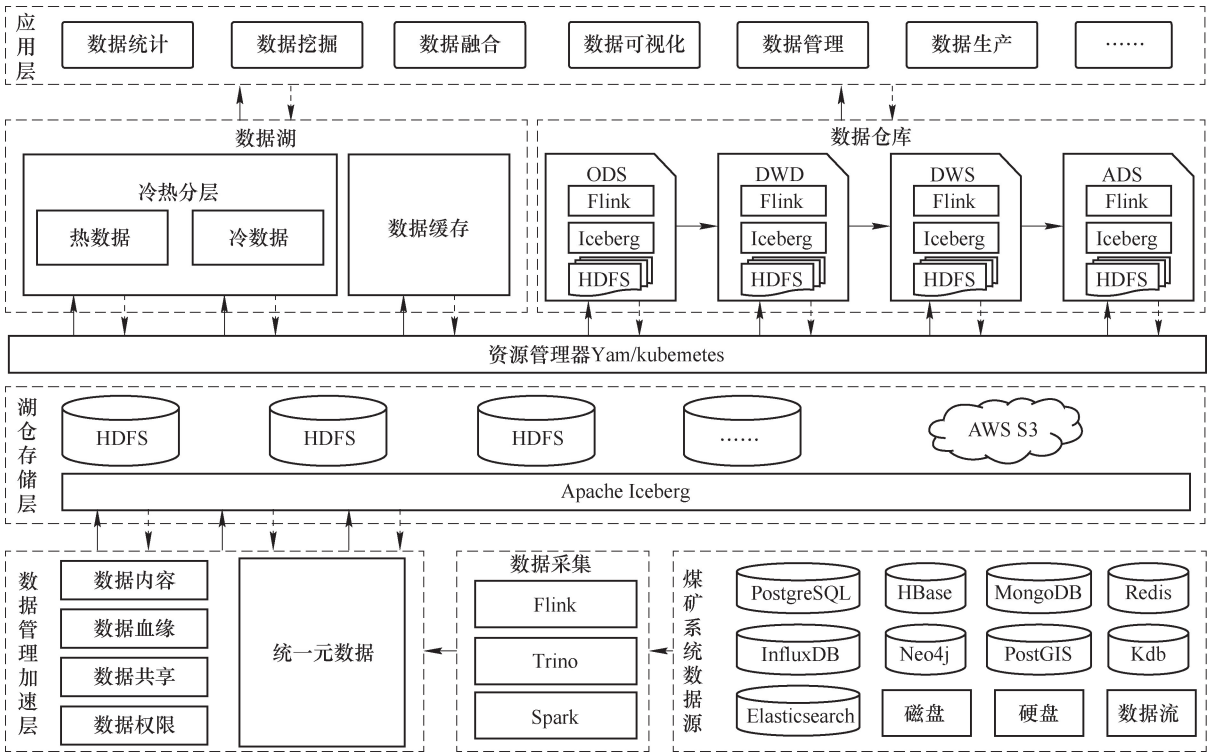


图 5 煤矿智能湖仓系统架构

Fig. 5 Architecture of coal mine lakehouse system

在数据管理加速层,煤矿地学大数据智能湖仓系统通过提取-转换-加载(Extract-Transform-Load, ETL)工具和变更数据捕获(Change Data Capture, CDC)技术处理,整合来自多源采集设备和现有数据库的数据,包括 PostgreSQL、MongoDB、Redis、Neo4j、InfluxDB 和 Elasticsearch 等多种数据库,确保不同来源的数据能够无缝集成。另外,系统提供数据采集、数据内容管理、数据血缘管理等关键组件,共同促进智能湖仓系统的数据管理加速服务。系统通过统一元数据管理和数据权限控制,优化数据存储结构,确保数据的一致性和安全性。

在湖仓存储层,智能湖仓系统采用分布式文件系统(Hadoop Distributed File System, HDFS)和亚马逊简易存储服务(Amazon Simple Storage Service, Amazon S3)作为数据存储源,构建私有云和公有云的混合架构,提高数据存储的灵活性和扩展性,确保数据的高可用性和安全性。

在应用层,智能湖仓系统为数据统计、数据挖掘、数据融合和数据可视化等提供强大的支持服务。基于 Apache Iceberg 的先进技术,通过这些组件的协同工作,智能湖仓系统能够为煤矿企业提供一个全面、高效和智能的数据管理解决方案,推动煤炭行业的数字化转型和高质量发展。

2.3 煤矿智能湖仓存储模型

煤矿地学大数据智能湖仓存储模型旨在应对

煤矿生产全周期涉及的多源异构地学数据所带来的挑战,通过制定规范的存储模型方案和智能化煤矿建设标准与技术规范体系,实现煤矿地学数据全周期共享和过程管理,满足地质测量、防突、生产、通风、瓦斯抽放、微震监测、综掘、综采等多个部门的数据需求。通过建立统一的煤矿地学数据标准和模型,系统能够有效解决数据来源广泛、种类繁多、结构不统一等问题,实现数据互联互通。智能湖仓系统支持多种常见的数据类型,涵盖文本、图像、音频、视频、传感器数据等,煤矿智能湖仓存储模型包括结构化、半结构化和非结构化煤矿地学数据存储模型。

2.3.1　结构化数据存储

煤矿地学数据常见的结构化数据格式包括 TXT、CSV 和 XLS/XLSX。这些格式在组织和存储煤矿数据时具有高度的结构性和可读性,是许多应用场景的标准选择。采用结构化数据存储格式能够整合和管理不同来源的数据,实现数据的标准化和统一化。以地质雷达数据的存储为例,地质雷达作为煤矿地质灾害隐患探测的重要手段,其数据主要包括雷达探测数据和解释数据。这些数据在不同阶段有不同的格式和结构,为便于数据的存储、查询和分析,通常会将其转换为结构化格式。图 6 为钻孔雷达探测数据 CSV 格式,对于这类数据采用 Iceberg 表结构进行规范化管理。

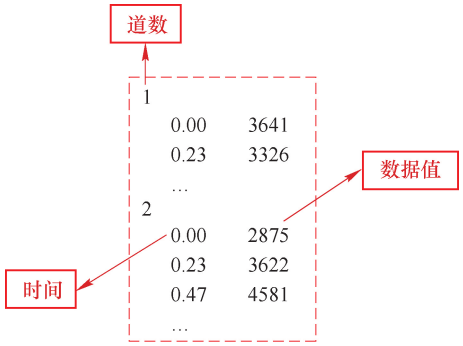
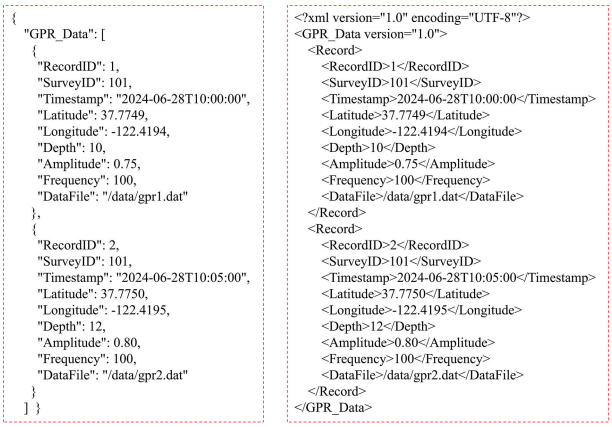


图 6　钻孔雷达探测数据 CSV 格式

Fig. 6　Borehole radar detection data in CSV format

2.3.2　半结构化数据存储

图 7 为雷达解释数据 JSON 和 XML 文档。在煤矿地学大数据管理系统中,日志文档、JSON 文档、XML 文档以及 HTML 等半结构化数据发挥重要作用。智能湖仓系统利用 Apache Iceberg 数据湖表结构管理半结构化文档数据,通过自定义规则实现数据的自动推断,并为这类数据提供与结构化数据相同的存储和检索功能。对于存储特定空间信息的数据,智能湖仓系统采用 GeoJSON 编码格式。GeoJSON 是一种用于表示地理特征的 JSON 格式,特别适合存储和交换包含地理坐标的信息。通过采用这些半结构化数据存储格式,煤矿地学大数据智能湖仓系统能够灵活管理和集成来自不同系统的地学数据,支持多样化的数据分析需求,提升数据处理效率和管理水平。



```
{
  "GPR_Data": [
    {
      "RecordID": 1,
      "SurveyID": 101,
      "Timestamp": "2024-06-28T10:00:00",
      "Latitude": 37.7749,
      "Longitude": -122.4194,
      "Depth": 10,
      "Amplitude": 0.75,
      "Frequency": 100,
      "DataFile": "/data/gpr1.dat"
    },
    {
      "RecordID": 2,
      "SurveyID": 101,
      "Timestamp": "2024-06-28T10:05:00",
      "Latitude": 37.7750,
      "Longitude": -122.4195,
      "Depth": 12,
      "Amplitude": 0.80,
      "Frequency": 100,
      "DataFile": "/data/gpr2.dat"
    }
  ]
}
```

(a) JSON文件结构

```
<?xml version="1.0" encoding="UTF-8"?>
<GPR_Data version="1.0">
  <Record>
    <RecordID>1</RecordID>
    <SurveyID>101</SurveyID>
    <Timestamp>2024-06-28T10:00:00</Timestamp>
    <Latitude>37.7749</Latitude>
    <Longitude>-122.4194</Longitude>
    <Depth>10</Depth>
    <Amplitude>0.75</Amplitude>
    <Frequency>100</Frequency>
    <DataFile>/data/gpr1.dat</DataFile>
  </Record>
  <Record>
    <RecordID>2</RecordID>
    <SurveyID>101</SurveyID>
    <Timestamp>2024-06-28T10:05:00</Timestamp>
    <Latitude>37.7750</Latitude>
    <Longitude>-122.4195</Longitude>
    <Depth>12</Depth>
    <Amplitude>0.80</Amplitude>
    <Frequency>100</Frequency>
    <DataFile>/data/gpr2.dat</DataFile>
  </Record>
</GPR_Data>
```

(b) XML文件结构

图 7　雷达解释数据 JSON 和 XML 文档

Fig. 7　Radar interpretation data in JSON and XML documents

2.3.3　非结构化数据存储

煤矿地学非结构化数据包括 DWG、DXF、SHP、LAS、SEG-Y 等,一般采用二进制形式存储管理。以地震勘探数据为例,在煤矿勘探过程中数据的存储格式十分复杂,常见的地震勘探数据类型分为:IBM 浮点型、IEEE 浮点型、16 位整型及 32 位整型。地震数据格式众多,野外数据采集仪器有专有的数据组织形式,数据格式也各不相同,其只针对软件本身,在其他软件中难以兼容管理。如地震勘探 SEG-Y 格式分为标准和非标准格式,区别在于是否含有 3 600 字节的文件头。标准 SEG-Y 格式主要由 3 部分组成。图 8 为地震勘探数据 SEG-Y 格式。

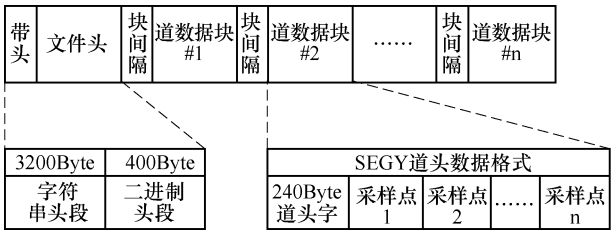


图 8　地震勘探数据 SEG-Y 格式

Fig. 8　Seismic exploration data in SEG-Y format

对于类似 SEG-Y 格式这种大规模的非结构化数据,煤矿地学大数据智能湖仓系统采用 Apache Iceberg 表结构进行管理,包括数据文件和元文件

两大部分。图 9 为 Apache Iceberg 数据存储格式。其中,数据文件(*.parquet 文件)存储非结构化数据,元文件管理这些数据文件的信息。元文件信息包括元数据描述信息文件、元数据文件和元数据文件列表。元数据描述信息文件(*.json 文件)描述表的结构、列类型和分区信息,以支持数据管理和查询优化;元数据文件(*.avro 文件),列出所有数据文件及其元数据,以便快速查找和访问,包括数据文件的状态、文件路径、分区信息、列级别的统计信息;元数据文件列表(snap-*.avro 文件)记录数据快照,以便进行版本管理和恢复,在查询时提供过滤以加快检索效率。这种方法将数据描述信息与 HDFS 中的数据源紧密结合,将大数据体分块切分并存储在数据集群中,通过 HDFS 进行统一管理,确保在读取数据时能够快速定位相关数据。

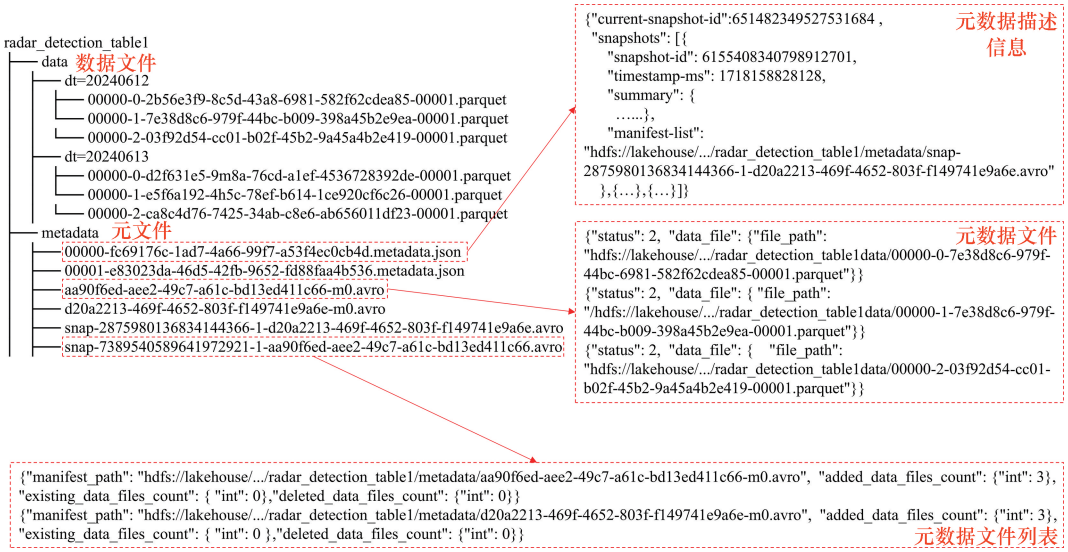


图 9 Apache Iceberg 数据存储格式

Fig. 9 Apache Iceberg data storage format

2.4 煤矿智能湖仓索引编码

海量的煤矿地学数据时空索引是智能湖仓系统数据管理核心问题。为提高数据检索效率,智能湖仓系统采用基于 Hilbert-Geohash 的时空分类编码,将时间、空间和分类进行统一编码。

2.4.1 空间编码

煤矿地学数据具有独特空间属性,在数据库中通常采用直角坐标系和球面坐标系进行存储管理。在查询空间信息时需要复合索引,检索效率低下。空间编码是一种快速检索策略,被广泛用于空间数据索引。智能湖仓系统基于 Hilbert-Geohash 编码,使用 Base64 字符映射将地理坐标系统转换为字符串。通过 Hilbert 曲线可以连续表达空间信息,将多维空间坐标信息进行降维,再利用 Geohash 编码,将空间坐标作为二进制表示,最后利用 Base64 进行编码。图 10 为 Hilbert-Geohash 空间编码。

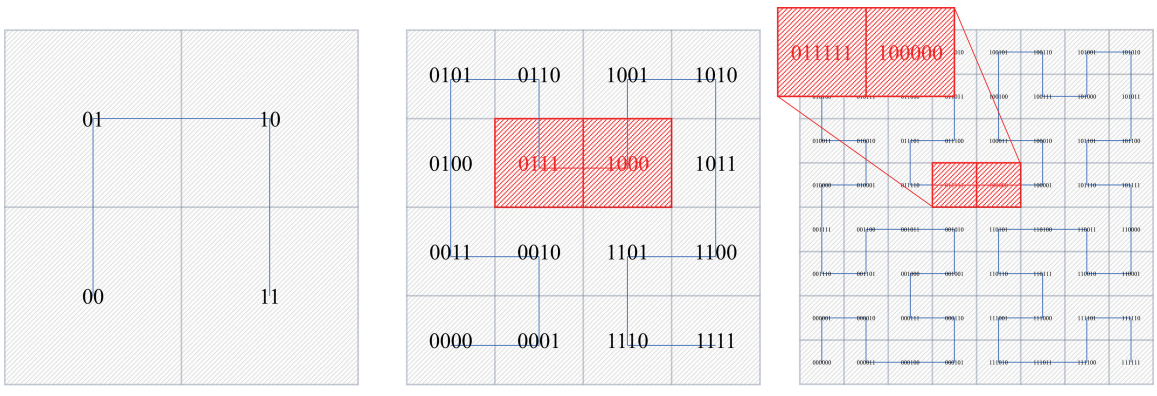


图 10 Hilbert-Geohash 空间编码

Fig. 10 Hilbert-Geohash spatial coding

2.4.2　时间编码

在煤矿地学大数据中,多种勘探数据的时间编码方式具有多样性,包括单尺度时间编码和多尺度时间编码,通常以字符串和时间戳的形式表达。字符串在进行运算时较为费时,例如“2023-12-31”进行加 1 d 运算,首先需要解码字符串,提取出年、月、日等信息;然后,按照历法规则加 1 d,得出结果为 2024-01-01。时间戳是以格林尼治时间 1970-01-01 的 00:00:00 起至当下的总秒数。受 32 位系统限制,时间戳的有效范围有限,且精度较低,只能精确到秒级。智能湖仓系统时间编码将时间进行切分,图 11 为时间编码。将时间按年、月、日、时、分、秒、毫秒转为二进制,其中年和毫秒转换为二进制后,将 12 位分别表示为 2 个 6 位二进制,采用和空间编码一致的 Base 64 进行重新编码。

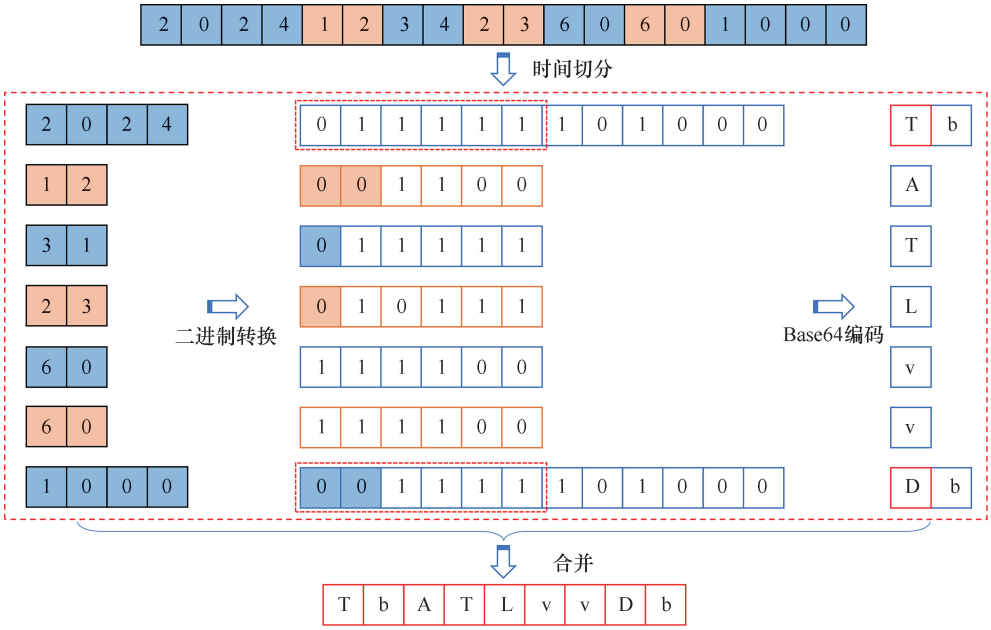


图 11　时间编码

Fig. 11　Temporal coding

2.4.3　分类编码

煤矿地学大数据来源多种勘探设备采集的数据,还有一些已经存储在数据库的数据。对于这些分散的数据,智能湖仓系统采用分类编码统一管理。分类编码采用“勘探类型+传感器类型+数据来源+部门”4 位组合作为分类,每一位由 Base 64 编码的 64 位字符构成,基本满足所有的煤矿地学数据分类,能唯一标识煤矿地学数据来源。

通过上述空间、时间和分类统一编码,煤矿地学大数据智能湖仓系统采用基于 Hilbert-Geohash 的时空分类编码作为索引键,在进行海量地学数据检索时,可以快速查询到某个空间范围和时空范围的所有类型数据。图 12 为时空分类索引编码。

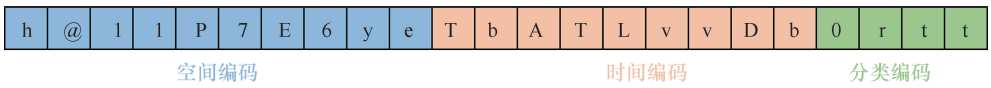


图 12　时空分类索引编码

Fig. 12　Spatiotemporal classification index encoding

3　煤矿智能湖仓系统技术应用案例

当前煤矿地质透明工作面存在如同“毛玻璃”的两张皮现象,透明化地质仍然存在“不透明”的问题[1]。为解决这一问题,研究团队在承担“十四五”国家重点研发计划中,探索性地开展煤矿地质灾害隐患透明化地图和自动更新平台建设工作。该平台主要实现复杂三维动态地质模型与多尺度网格化地质模型构建,精细化刻画工作面构造、岩性结构和地质灾害隐患区形态及内部参数空间分布。平台建设依赖于煤矿地质数据、物探、地理、生产和监测等数据汇聚和融合应用,因此需要构建一个能够满足多场多源异构数据存储的智能湖仓系统。

3.1 平台系统架构设计

图 13 为煤矿地质灾害隐患透明化地图和自动更新平台系统总体架构。该系统以煤矿地学大数据智能湖仓系统作为数据支撑平台,矿大技术服务节点作为技术支持,初步部署在煤炭精细勘探与智能开发全国重点实验室数据中心,集成先进的大数据技术和勘探数据汇聚系统,支持数据的自动更新,确保所有信息的准确性和时效性。

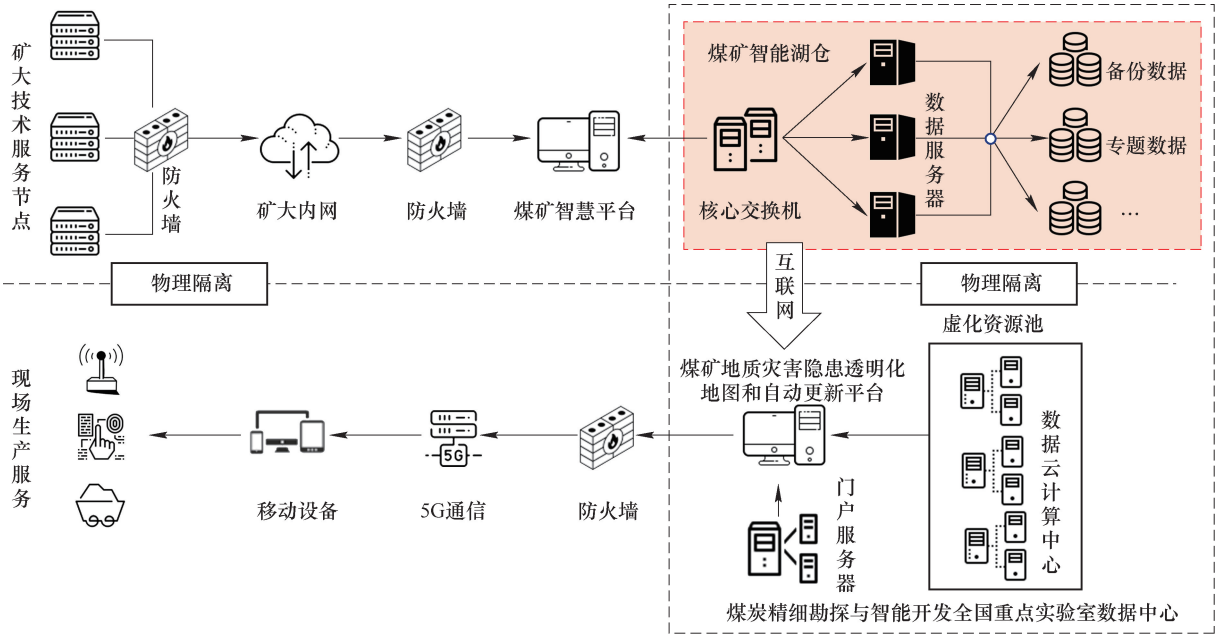


图 13 煤矿地质灾害隐患透明化地图和自动更新平台总体架构

Fig. 13 Overall architecture of transparent map and automatic update platform for coal mine geological hazard hidden dangers

智能湖仓系统负责存储和处理大量的煤矿地学数据,包含历史数据和采集实时数据。通过物理隔离和防火墙技术,系统设计多层安全防护,防止未授权访问和网络攻击,保障数据的安全性和可靠性。通过虚化资源池、现场生产服务、数据云计算中心等组件,共同构成一个高效、灵活的数据处理和分析环境。在通信方面,通过门户服务器和移动设备的支持,运用 5G 通信技术,使得煤矿地学数据访问和交互更加便捷。

3.2 智能湖仓系统搭建

智能湖仓系统采用前后端分离的 B/S 分层架构开发模式,选用 Vue. js 作为地质灾害隐患透明化地图和自动更新平台的前端开发框架,后端开发采用 Python 语言,结合 DRF(Django REST Framework)开发框架和 PyIceberg 模块开发系统接口。煤矿地学大数据智能湖仓系统集成多种大数据技术,包括 Apache Hadoop、HDFS、Apache Spark、Apache Flink、Apache Zookeeper、Apache Kafka、Apache Iceberg 等,每个节点分别承担不同的任务,利用多种大数据技术协同工作实现系统高效运行。表 2 为煤矿智能湖仓系统集群搭建信息。

表 2 煤矿智能湖仓系统集群搭建信息

Table 2 Cluster construction information of coal mine lakehouse system

节点	HDFS	YARN	Spark	Flink	Zookeeper/Kafka
Master: 192. 168. 55. 110	Name Node Secondary Name Node Data Node	Node Manager Resource Manager	Master	Job Manager Task Manager	Leader
Worker1: 192. 168. 55. 111	Data Node	Node Manager	Worker	Job Manager Task Manager	Follower
Worker2: 192. 168. 55. 112	Data Node	Node Manager	Worker	Job Manager Task Manager	Follower

在智能湖仓系统中,采用 Apache Hadoop 分布式文件系统 HDFS 提供高可用的存储基础设施,Map Reduce 计算框架提供分布式计算处理。通过集成 Apache Iceberg 支持大规模数据表的管理和高效查询,进一步提升智能湖仓系统的管理和查询能力。平台利用 Apache Spark 内存计算能力处理近实时数据,并将 Apache Flink 作为流式处理框架处理实时数据流,实现煤矿地学数据处理快速实时响应。系统部署 YARN 资源分配组件,使用核心组件 Resource Manager 和 Node Manager 管理集群资源,实现高效的资源管理和任务调度。另外,集成 Apache Zookeeper 和 Kafka 作为系统的协调和消息中间件,提供高可用的分布式协调和消息传递功能,确保系统的可靠性和扩展性。

3.3　平台核心功能实现

在煤矿地质灾害隐患透明化地图和自动更新平台中,采用 Vue. js 技术开发数据管理的 Web 页面实现功能页面,系统利用 PyIceberg 模块连接 Apache Iceberg,快速检索煤矿地学大数据。同时连接 PySpark 模块与 Apache Spark,实现对煤矿大数据的分布式处理和分析。对于实时数据处理,通过 PyFlink 模块连接 Apache Flink,利用 Flink Python API 实现流数据采集和处理。

在煤矿地学大数据智能湖仓系统的强大数据支撑下,煤矿地质灾害隐患透明化地图与自动更新平台成功实现地质模型的构建与实时更新功能。图 14 为煤矿地质灾害隐患透明化地图和自动更新平台。该平台能够展示煤矿三维地质模型,包括工业广场、钻孔、巷道和采掘工作面等多种精细化模型。平台地图精度小于 0. 2 m,地质信息反馈和自动更新时间控制在 10 min 以内,实时探测数据的响应速度控制在毫秒级。通过地质构造、岩性和物性参数的统一表达,该平台实现煤矿地学大数据的透明化融合,为煤矿智能化管理提供精准、高效的技术支撑。

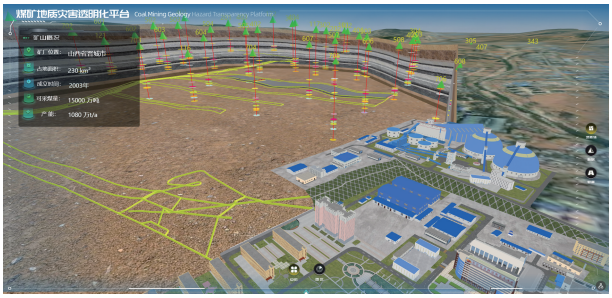


图 14　煤矿地质灾害隐患透明化地图和自动更新平台

Fig. 14　Transparent map and automatic update platform for coal mine geological hazard hidden dangers

4　结　论

在当今煤矿行业数智化转型的浪潮中,高效的数据管理与利用成为推动行业发展的关键要素。为解决传统的煤矿数据存储与管理方式难题,提高快速准确地检索海量数据以及充分挖掘数据要素的潜在价值,研究了煤矿地学大数据管理与应用模式,得到如下结论:

(1)基于多种大数据技术提出一种煤矿地学大数据智能湖仓系统。通过多种大数据技术的集成,实现煤矿结构化、半结构化和非结构化地学数据的统一存储与规范化管理。

(2)基于 Hilbert-Geohash 提出一种智能湖仓时空分类索引编码方法,将时空信息与数据分类进行统一编码,显著提高海量煤矿地学数据的检索效率。

(3)通过地质灾害隐患透明化地图和自动更新平台的应用验证煤矿地学大数据智能湖仓系统的可行性,为进一步提升煤矿数据要素的利用价值提供重要的理论与实践依据。

参考文献

[1]　彭苏萍. 我国煤矿安全高效开采地质保障系统研究现状及展望[J]. 煤炭学报,2020,45(7):2331-2345.
PENG Suping. Current status and prospects of research on geological assurance system for coal mine safe and high efficient mining[J]. Journal of China Coal Society,2020,45(7):2331-2345.

[2]　袁亮. 煤炭精准开采科学构想[J]. 煤炭学报,2017,42(1):1-7.
YUAN Liang. Scientific conception of precision coal mining[J]. Journal of China Coal Society,2017,42(1):1-7.

[3]　袁亮. 我国煤炭工业高质量发展面临的挑战与对策[J]. 中国煤炭,2020,46(1):6-12.
YUAN Liang. Challenges and countermeasures for high quality development of China's coal industry[J]. China Coal,2020,46(1):6-12.

[4]　王国法,刘峰,庞义辉,等. 煤矿智能化:煤炭工业高质量发展的核心技术支撑[J]. 煤炭学报,2019,44(2):349-357.
WANG Guofa,LIU Feng,PANG Yihui,et al. Coal mine intellectualization:The core technology of high quality development[J]. Journal of China Coal Society,2019,44(2):349-357.

[5] 王国法. 加快煤矿智能化建设 推进煤炭行业高质量发展[J]. 中国煤炭,2021,47(1):2-10.
WANG Guofa. Speeding up intelligent construction of coal mine and promoting high-quality development of coal industry[J]. China Coal,2021,47(1):2-10.

[6] 王国法,刘峰. 中国煤矿智能化发展报告-2022 年[M]. 北京:应急管理出版社,2022.
WANG Guofa, LIU Feng. China coal mine intelligent development report-2022[M]. Beijing: Emergency Management Press,2022.

[7] 王双明,孙强,谷超,等. 煤炭开发推动地学研究发展[J]. 中国煤炭,2024,50(1):2-8.
WANG Shuangming, SUN Qiang, GU Chao, et al. The development of geoscientific research promoted by coal exploitation[J]. China Coal,2024,50(1):2-8.

[8] PAL A,KUMAR P,SHAH F. Seismic data management for big data era[C]//. Day 3 Wed, November 13, 2019. November 11-14, 2019. Abu Dhabi, UAE. SPE, 2019: D032S194R001.

[9] 杜文凤,彭苏萍,师素珍. 基于三维地震勘探研究地裂缝空间展布特征[J]. 岩石力学与工程学报,2016,35(4):778-783.
DU Wenfeng, PENG Suping, SHI Suzhen. The spatial distribution characteristics of ground fissures based on 3D seismic exploration[J]. Chinese Journal of Rock Mechanics and Engineering,2016,35(4):778-783.

[10] 彭苏萍,赵惊涛,盛同杰,等. 煤田绕射地震勘探现状与进展[J]. 煤田地质与勘探,2023,51(1):1-20.
PENG Suping, ZHAO Jingtao, SHENG Tongjie, et al. Status and advance of seismic diffraction exploration in coalfield[J]. Coal Geology & Exploration, 2023, 51(1): 1-20.

[11] 彭苏萍. 滇东矿区喀斯特地貌高精度三维地震勘探技术研究[M]. 北京:科学出版社,2024.
PENG Suping. Research on high-resolution 3D seismic exploration technology under Karst landform conditions in the eastern Yunnan mining area[M]. Beijing: Science Press,2024.

[12] 程久龙,李飞,彭苏萍,等. 矿井巷道地球物理方法超前探测研究进展与展望[J]. 煤炭学报,2014,39(8):1742-1750.
CHENG Jiulong, LI Fei, PENG Suping, et al. Research progress and development direction on advanced detection in mine roadway working face using geophysical methods[J]. Journal of China Coal Society, 2014, 39(8): 1742-1750.

[13] 许献磊,马正,陈令洲. 煤矿地质灾害隐患透明化探测技术进展与思考[J]. 绿色矿山,2023,1(1):56-69.
XU Xianlei, MA Zheng, CHEN Lingzhou. Progress and thinking of transparent detection technology for hidden geological hazards in coal mines[J]. Journal of Green Mine,2023,1(1):56-69.

[14] 许献磊,彭苏萍,马正,等. 基于空气耦合雷达的矿井煤岩界面随采动态探测原理及关键技术[J]. 煤炭学报,2022,47(8):2961-2977.
XU Xianlei, PENG Suping, MA Zheng, et al. Principle and key technology of dynamic detection of coal-rock interface in coal mine based on air-coupled radar[J]. Journal of China Coal Society, 2022, 47(8): 2961-2977.

[15] 陈建平,李靖,谢帅,等. 中国地质大数据研究现状[J]. 地质学刊,2017,41(3):353-366.
CHEN Jianping, LI Jing, XIE Shuai, et al. China geological big data research status[J]. Journal of Geology, 2017,41(3): 353-366.

[16] L'HEUREUX A, GROLINGER K, ELYAMANY H F, et al. Machine learning with big data: challenges and approaches[J]. IEEE Access,2017,5: 7776-7797.

[17] ZHU Y Q, TAN Y J, LUO X, et al. Big data management for cloud-enabled geological information services [J]. Scientific Programming,2018,2018: 1327214.

[18] GUO H D. Big Earth data: a new frontier in Earth and information sciences[J]. Big Earth Data, 2017, 1(1/2): 4-20.

[19] MERRITT P, BI H X, DAVIS B, et al. Big Earth Data: a comprehensive analysis of visualization analytics issues[J]. Big Earth Data,2018,2(4): 321-350.

[20] 翟明国,杨树锋,陈宁华,等. 大数据时代:地质学的挑战与机遇[J]. 中国科学院院刊,2018,33(8):825-831.
ZHAI Mingguo, YANG Shufeng, CHEN Ninghua, et al. Big data epoch: challenges and opportunities for geology[J]. Bulletin of Chinese Academy of Sciences,2018, 33(8): 825-831.

[21] 毛善君,杨乃时,高彦清,等. 煤矿分布式协同"一张图"系统的设计和关键技术[J]. 煤炭学报,2018,43(1):280-286.
MAO Shanjun, YANG Naishi, GAO Yanqing, et al. Design and key technology research of coal mine distributed cooperative "one map" system[J]. Journal of China Coal Society,2018,43(1): 280-286.

[22] 康红普,伊丙鼎,高富强,等. 中国煤矿井下地应力数据库及地应力分布规律[J]. 煤炭学报,2019,44(1):23-33.
KANG Hongpu, YI Bingding, GAO Fuqiang, et al. Database and characteristics of underground in situ stress

distribution in Chinese coal mines[J]. Journal of China Coal Society,2019,44(1): 23-33.

[23] 疏礼春. 智能煤矿数据中台架构及关键技术研究[J]. 工矿自动化,2021,47(6): 40-44.
SHU Lichun. Research on the architecture and key technologies of intelligent coal mine data middle platform[J]. Industry and Mine Automation,2021,47(6): 40-44.

[24] 许娜,耿恒高,徐传鹏,等. 基于 MongoDB 的地震勘探数据管理系统的设计与实现[J]. 实验室研究与探索,2022,41(2): 251-260.
XU Na,GENG Henggao,XU Chuanpeng,et al. Design and implementation of seismic exploration data management system based on MongoDB[J]. Research and Exploration in Laboratory,2022,41(2): 251-260.

[25] 王霖,方乾,张晓霞,等. 智能化煤矿数据仓库建模方法[J]. 工矿自动化,2022,48(4): 5-13.
WANG Lin,FANG Qian,ZHANG Xiaoxia,et al. Intelligent coal mine data warehouse modeling method[J]. Journal of Mine Automation,2022,48(4): 5-13.

[26] 韩安. 基于 Hadoop 的煤矿数据中心架构设计[J]. 工矿自动化,2019,45(8): 60-64.
HAN An. Architecture design of coal mine data center based on Hadoop[J]. Industry and Mine Automation,2019,45(8): 60-64.

[27] 廖志伟,张建明,郭林峰,等. 煤矿智能化建设新型生产管控模式研究[J]. 煤炭经济研究,2024,44(4): 27-32.
LIAO Zhiwei,ZHANG Jianming,GUO Linfeng,et al. Research on new production control mode of intelligent construction in coal mine[J]. Coal Economic Research,2024,44(4): 27-32.

[28] 孙继平. 煤矿信息化与智能化要求与关键技术[J]. 煤炭科学技术,2014,42(9): 22-25,71.
SUN Jiping. Requirement and key technology on mine informationalization and intelligent technology[J]. Coal Science and Technology,2014,42(9): 22-25,71.

[29] CHEN J P,XIANG J,HU Q,et al. Quantitative geoscience and geological big data development: a review[J]. Acta Geologica Sinica-English Edition,2016,90(4): 1490-1515.

[30] 祁和刚,张建中,武光城,等. 构建“煤智云”大数据中心 引领煤炭产业数字化转型[J]. 数据,2022(5): 22-25.
QI Hegang,ZHANG Jianzhong,WU Guangcheng,et al. Building a “coal smart cloud” big data center and leading the digital transformation of coal industry[J]. Data,2022(5): 22-25.

[31] 王国法,任怀伟,赵国瑞,等. 煤矿智能化十大“痛点”解析及对策[J]. 工矿自动化,2021,47(6): 1-11.
WANG Guofa,REN Huaiwei,ZHAO Guorui,et al. Analysis and countermeasures of ten “pain points” of intelligent coal mine[J]. Industry and Mine Automation,2021,47(6): 1-11.

[32] 王国法. 煤矿智能化最新技术进展与问题探讨[J]. 煤炭科学技术,2022,50(1): 1-27.
WANG Guofa. New technological progress of coal mine intelligence and its problems[J]. Coal Science and Technology,2022,50(1): 1-27.

[33] ARMBRUST M,GHODSI A,XIN R,et al. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics[C]//. Proceedings of CIDR,2021.

[34] ARMBRUST M,DAS T,SUN L W,et al. Delta lake: high-performance ACID table storage over cloud object stores[J]. Proceedings of the VLDB Endowment,2020,13(12): 3411-3424.

[35] AIT ERRAMI S,HAJJI H,AIT EL KADI K,et al. Spatial big data architecture: From data warehouses and Data Lakes to the LakeHouse[J]. Journal of Parallel and Distributed Computing,2023,176: 70-79.

(责任编辑:陈贵仁)