doi: 10.12097/gbc.2022.11.004

地质调查业务管理数据集成辅助决策系统架构与关键技术

文敏[1,2,3,4]，月一[1,2,4]，张怀东[1,2,4]，王想红[1,2,4]，施艳[1,2,4]，刘荣梅[1,2,4]，孙涵睿[1,2,4]
WEN Min[1,2,3,4], YUE Yi[1,2,4], ZHANG Huaidong[1,2,4], WANG Xianghong[1,2,4], SHI Yan[1,2,4], LIU Rongmei[1,2,4], SUN Hanrui[1,2,4]

1. 中国地质调查局自然资源综合调查指挥中心, 北京 100055;
2. 中国地质调查局发展研究中心, 北京 100037;
3. 国家地理信息系统工程技术研究中心, 湖北 武汉 430078;
4. 自然资源部地质信息工程技术创新中心, 北京 100037
1. *Natural Resources Comprehensive Survey Command Center*, *China Geological Survey*, *Beijing* 100055, *China;*
2. *Development and Research Center*, *China Geological Survey*, *Beijing* 100037*;*
3. *National Engineering Research Center of Geographic Information System*, *Wuhan* 430078, *Hubei*, *China;*
4. *Technology Innovation Center for Geological Information*, *MNR*, *Beijing* 100037, *China*

摘要: 在地质调查工作过程中，各类信息系统产生了海量的业务管理数据，需要解决这些多来源、高动态、复杂异构数据的有效集成辅助决策问题，驱动管理决策现代化。基于大数据、GIS、数据挖掘等技术，构建了地质调查业务管理数据集成辅助决策的总体架构；通过多源异构数据自动化动态集成处理、基于 Hadoop 的“湖仓一体”混合式数据组织管理、数据挖掘融合地理智能的分析决策模型等关键技术方法，研发了国家地质调查业务管理大数据系统。系统已接入 24 个数据源，实现了自动化动态集成，完成了 1.5 亿余条、20 多万档异构数据的一体化组织管理，通过数据和分析服务有效辅助了管理决策。可有效解决大数据环境下数据集成辅助决策问题，提升国家地质调查工作管理决策的效率和水平。

关键词: 地质调查；业务管理；大数据；地理信息；数据集成治理；分析辅助决策
中图分类号: P628 **文献标志码:** A **文章编号:** 1671−2552(2024)07−1221−12

Wen M, Yue Y, Zhang H D, Wang X H, Shi Y, Liu R M, Sun H R. A framework and key technologies for national geological survey management data integration and analysis for decision support. *Geological Bulletin of China*, 2024, 43(7): 1221−1232

Abstract: Various information systems have been constructed for national geological survey organization, which have generated massive multi−source and heterogeneous management data. Effective integration and analysis of these data are in dire needs, for the collaborative and intelligent management of the national geological survey. This paper creates a framework based on big data, GIS and data mining technologies. Related key technologies are proposed, involving automatic and dynamic data integration, hybrid data management by Hadoop and "data−lake−warehouse" architecture, and decision support model for geological survey management. Based on the above, National Geological Survey Management Big Data System was constructed, which has integrated data from 24 different sources automatically and dynamically, more than 150 million records and 200 thousand documents have been organized in one, and has supported management decision by data or analysis services. It has been proved that can solve the problem of data integration for decision−making support, and has promoted the management efficiency of the national geological survey.

收稿日期: 2022−11−02; **修订日期:** 2023−02−16
资助项目: 中国地质调查局项目《地质调查业务管理与辅助决策系统建设》(编号: DD20160356A)、《地质调查业务管理信息支撑与服务》(编号: DD20190403A)、《地球科学数据集成与服务》(编号: DD20221785)
作者简介: 文敏(1986−)，男，硕士，正高级工程师，从事地学领域数据集成治理与辅助决策技术研究。E−mail: wenm@mail.cgs.gov.cn

Key words: geological survey; business management; big data; GIS; data integration; decision-making support

地质工作是保障国家能源资源安全、支撑国民经济发展的重要基础性工作。为更好地组织实施国家地质调查工作，中国地质调查局针对野外安全、部署规划、实施监督、成果服务、科技财务、人事装备等不同业务，先后研发了各类管理信息系统，有效支撑了工作的组织开展，但也面临系统不统一、业务不协同、数据不互通、辅助决策程度低等问题。各类系统的持续运行，产生了海量多源异构地质调查业务管理数据，既包括分布于各类数据库的结构化数据、非结构化技术管理文档、互联网数据等，也包括野外北斗、GPS、手机信令、工作部署实施等空间数据，具有鲜明的行业特点。实现这些数据的有效集成和分析辅助决策，驱动国家地质调查工作管理决策能力提升成为急需。

多源数据集成辅助决策是大数据研究的热点和难点，《Nature》《Science》等相继出版专刊，探讨大数据带来的机遇和挑战（Wang et al., 2014），研究认为数据能够促进高级决策模型的实施（Torre et al., 2022），有效的数据集成策略对有效辅助决策非常必要（Parimbelli et al., 2016; 洪之旭等, 2017）。在数据集成方面，从数据单一走向多元、从离线走向在线动态自动化，在数据仓库、联邦数据库（钟晓等,2001）、虚拟数据库（刘晴等,2020）等传统数据集成方法基础上，开展了集成模式与模型优化（Doan et al., 2002; Nottelmann et al., 2007; Das et al., 2008），提升集成效率（Do et al., 2002）与质量（Lehmberg et al., 2017），以及数据中台（赵伟伟等,2021）、自动集成（文敏等, 2011）等新的集成方法探索。在数据组织管理方面，传统数据管理系统难以应对新形势下的数据情况（Aissi et al., 2021），新一代数据仓库、数据湖（吴冲龙等, 2020）、分布式存储（王凯等, 2015; Halevy et al., 2016; 任晓霞等,2018; 刘文毅等,2019）等数据管理技术和架构成为研究热点。在辅助决策方面，生态（刘洪霞等,2018）、农业（韩家琪等, 2016）等领域已取得基于大数据的研究成果；商务智能、地理信息等技术交叉融合（徐佳沅等,2000; Angelaccio et al., 2012; Nasr et al., 2013），提供了新的思路。综上，大数据环境下的数据集成辅助决策技术，在互联网、金融等行业，面向通用数据类型，取得了创新进展。在地质领域，由于数据和业务的行业特性，针对多来源、复杂异构的地质调查业务管理数据，其自动化动态集成管理和分析辅助决策技术仍需要加强探索。

本文基于大数据、GIS、云计算、数据挖掘等技术，构建了数据自动化动态集成辅助决策的总体架构。通过多源异构地质调查业务管理数据自动化动态集成、基于 Hadoop 的“湖仓一体”异构数据组织管理、地理商务智能融合的辅助决策等关键技术方法，并应用于国家地质调查业务管理大数据系统研发，有效解决了多源异构地质调查业务管理数据的自动化动态接入集成、一体化组织管理和有效分析决策问题。

1 总体架构

1.1 总体技术架构

地质调查业务管理数据集成辅助决策总体技术流程，包括数据集成治理、统一组织管理、分析辅助决策 3 个步骤（图 1），据此形成总体技术架构（图 2）。

1.1.1 多源数据集成治理

数据中台在相关行业已经得到初步应用和实践，其中互联网等行业走在前沿（苏萌等,2019）。针对空间数据在内的多源异构数据源，将数据中台理念与 GIS、ETL 等技术融合，通过数据快速适配接入、统一开发和调度监控，实现地质调查业务管理数据从快速接入、集成治理、组织管理到数据服务一站式、自动化动态的全流程集成处理。

1.1.2 异构数据统一组织管理

在面向结构化数据的传统数据仓库（胡侃等,1998）、基于 Hadoop 体系的分布式存储基础上，针对同时包含空间数据、结构化、非结构化等类型的行业数据特点，采用“湖仓一体”（Saddad et al., 2020）架构下的混合式存储方案，保证最优存储；通过建立统一数据模型与主数据，实现异构数据的统一数据描述与有效组织；构建统一存取访问层，实现统一数据访问与组织管理。

1.1.3 数据分析辅助决策

采用 Apache Kylin 分布式分析引擎，对数据进行 OLAP 的多维数据立方体构建，为多维分析提供更加快速的预处理数据。利用多元回归、聚类等数据分析技术，融合空间分析方法，构建数据分析决策

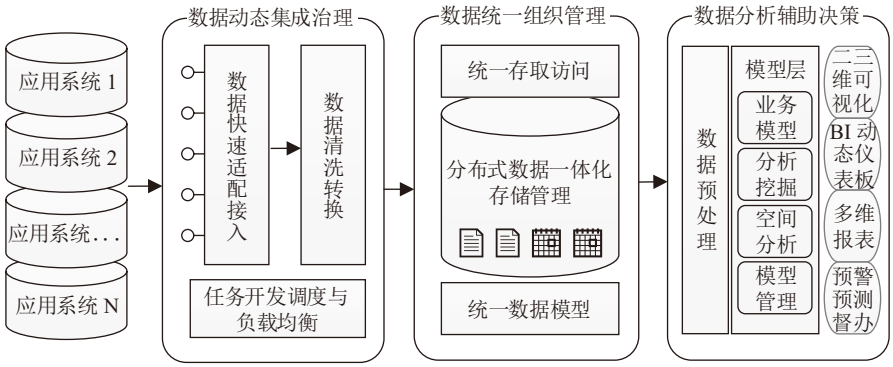

图 1　总体技术流程图

Fig. 1　Diagram of technical process

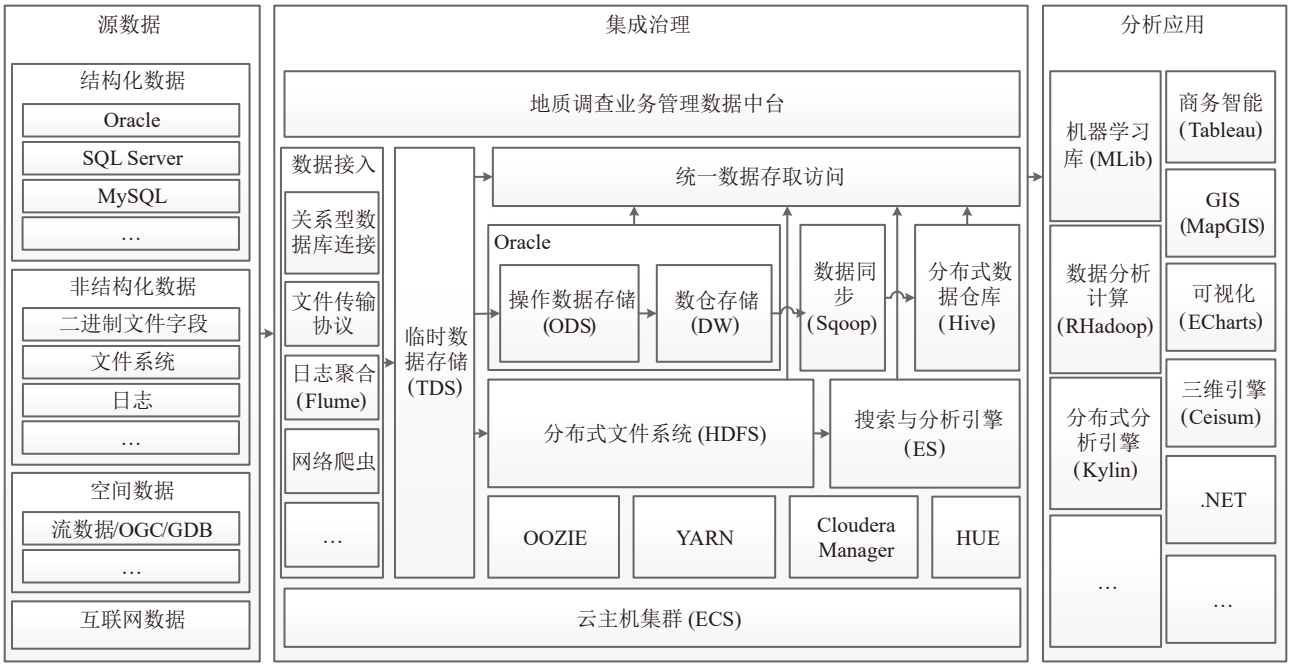

图 2　总体技术架构

Fig. 2　Overall technical architecture

模型, 实现地质调查业务运行的动态评估、诊断、预测。结合 GIS 与商务智能相关技术, 通过业务运行预警督办、二三维动态地图与仪表板、自定义多维报表等方法, 辅助管理决策。

1.2　总体应用架构

地质调查业务管理数据集成辅助决策总体应用架构, 包括云基础环境、数据源层、数据中台层、数据中心层、分析应用层 5 个部分(图 3)。

(1)云基础环境层, 基于中国地质调查局“地质云”平台, 使用 21 个高性能云虚拟节点。其中, 10 台为 Hadoop 平台节点, 2 台为数据存储节点, 5 台为数据中台服务节点, 1 台为分析引擎节点, 1 台为应用系统服务节点, 1 台为地图服务节点, 1 台为备用 ETL 节点(表 1)。

(2)数据源层, 主要包括物理分散的各类业务系统数据环境, 数据存储方式和类型各异, 数据随各系统运行动态产生。

(3)数据中台层, 提供对不同数据库、不同类型数据的接入、集成与处理的任务开发、调度与监控, 数据质量、资源管理和数据接口开发等。

(4)数据中心层, 负责异构数据的混合、分层存储, 并提供统一的组织、更新与检索等, 实现统一数据管理与操作。

(5)分析应用层面, 向各级各类用户, 划分基础

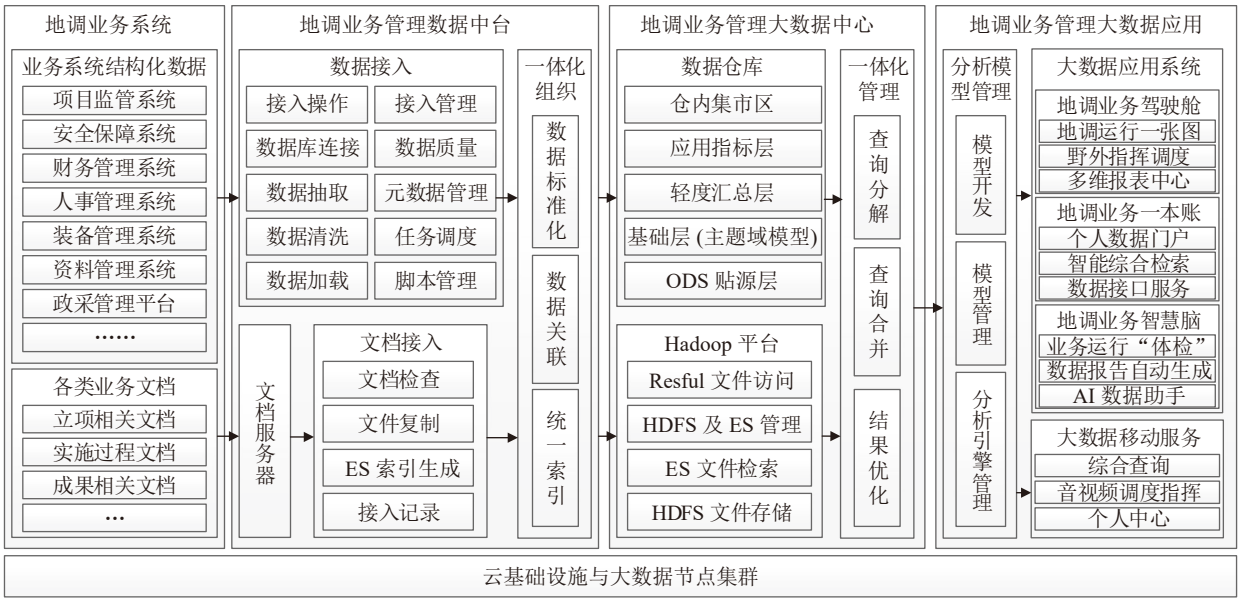

图 3 总体应用架构

Fig. 3 Overall application architecture

表 1 云基础环境层技术配置

Table 1 Technical configuration of Cloud layer

类型	节点	技术角色
Hadoop平台节点	BDP-01	HDFS NFS Gateway/Hive Metastore Server/Hue Server/Cloudera Management Service Activity Monitor/Cloudera Management Service Alert Publisher/Cloudera Management Service Event Server/Cloudera Management Service Host Monitor/Cloudera Management Service Monitor/Oozie Server/Sqoop 2 Server
	BDP-02	HDFS Failover Controller/HDFS HttpFS/HDFS JournalNode/HDFS NameNode/YARN (MR2 Included) ResourceManager
	BDP-03	HDFS Failover Controller/HDFS JournalNode/HDFS NameNode/Hue Server
	BDP-04	HDFS DataNode/HDFS JournalNode/YARN (MR2 Included) JobHistory Server/YARN (MR2 Included) NodeManager
	BDP-06	HDFS DataNode/Hive WebHCat Server/YARN (MR2 Included) NodeManager
	BDP-05	HDFS DataNode/Hive WebHCat Server
	BDP-07	HDFS DataNode/HiveServer2/YARN (MR2 Included) NodeManager
	BDP-08	ZooKeeper Server（Leader）/HDFS DataNode/HiveServer2/YARN (MR2 Included) NodeManager/Hive Gateway
	BDP-09	ZooKeeper Server（Follower）/HDFS DataNode/Hive Gateway/YARN (MR2 Included) NodeManager
	BDP-10	ZooKeeper Server（Follower）/HDFS DataNode/Hive Gateway/YARN (MR2 Included) NodeManager
数据节点	BDP-11	Oracle 12c (RDBMS Server-01)
	BDP-12	PostgreSQL,PostGIS(RDBMS Server-02)
备用ETL节点	BDP-13	Kettle (ETL Server)
数据中台节点	BDP-14	前端/网关（各模块前端/Gateway）
	BDP-15	用户中心/消息中心（UC-Server/SMS）
	BDP-16	开发套件/数据管理/机构管理（Dubhe Server/Megrez/Ent）
	BDP-17	开发套件/交换平台（Dubhe-Node/Datax/DSource）
	BDP-18	数据库（MySQL/Redis）
分析引擎节点	BDP-19	分析引擎服务（Tableau）
应用服务节点	BDP-20	应用服务（IIS/Apache）
地图服务节点	BDP-21	地图服务（MapGIS Server）

数据服务、多维分析可视化、分析决策辅助 3 个层次, 实现数据分析应用服务。

2　关键技术

2.1　多源数据自动化动态集成

基于第三方数据中台工具, 开展的数据集成研究(李文俊等,2020), 主要针对不包含空间数据的军队装备保障数据。针对地质调查业务管理数据特点, 通过数据的快速适配接入、自动化动态“清洗、转换、标准化、关联”集成处理, 实现地质调查业务管理数据的自动化动态集成(图 4)。

2.1.1　多源异构数据快速适配接入

针对多源数据的不同存储方式, 分别采用 ODBC、JDBC、FTP、Flume 等技术, 设计了“直读访问−主动接口−被动接口”的快速适配接入技术方法, 实现不同数据源快速适配接入, 降低集成复杂度, 提高集成灵活性。

(1)直读访问方式, 指针对关系型数据库, 使用 ODBC、JDBC 等连接协议, 面向 Oracle、SQL Server、MySQL 等各类主流关系型数据库, 封装实现快速配置接入, 利用只读访问方式, 直接读取源数据库。被动接口方式指数据提供方调用数据写入接口, 实现数据动态推送, 数据接收方校验数据并入库。主动接口方式指数据提供方提供数据访问接口, 通过该接口, 动态读取数据并解析入库。直读访问方式具有接入便捷、集成效率高等特点, 对于外围业务系统, 或无法提供数据库读取权限的系统, 采用主被动接口方式完成数据接入。

(2)地质调查业务管理数据中包含大量非结构数据, 如项目文档、地质报告、地质图件等, 涵盖 Word、Excel、PDF、JPEG 等不同格式, 包括文件系统、数据库 BLOB 大字段存储等不同存储方式, 利用 FTP 直接读取或接口同步的方式, 针对不同数据源情况, 实现接入集成。

(3)针对系统日志等半结构化数据, 采用 Hadoop 生态下的 Flume 组件, 实现半结构化数据的集成, 并同步存储至 HDFS 或 HBase; 针对地质调查项目招投标等互联网数据, 使用聚焦爬虫技术, 实现定向数据集成与处理。

对上述 3 项技术方法进行封装, 可以实现通过在线配置, 对指定数据源完成快速适配接入。

2.1.2　自动化动态“清洗、转换、标准化、关联”集成处理

采用数据标准化与清洗转换、基于主数据的动

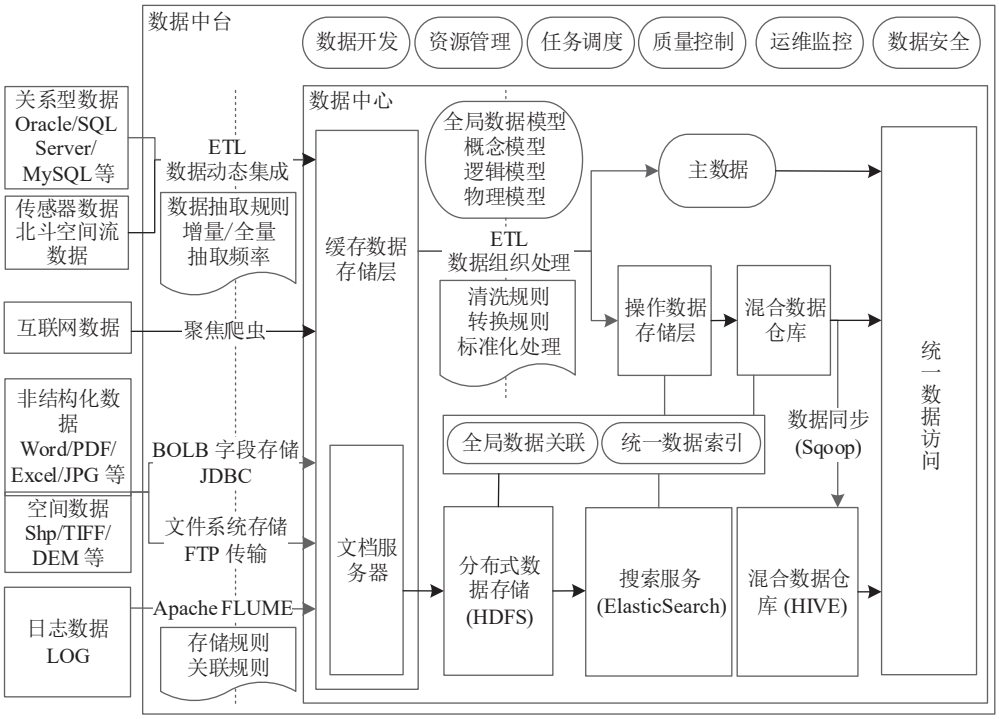

图 4　数据自动化动态集成治理与组织管理技术流程图

Fig. 4　Technology flowchart of data automation dynamic integration and governance

态关联、非结构化全文存储处理、数据脱敏在内的数据处理方法，将数据集成在统一的模型、基准和体系中，并利用负载均衡下的集成批处理任务和调度监控，实现数据自动化动态集成处理。

（1）有学者基于 ETL 技术，开展了 Shape 格式空间数据的坐标转换研究（刘文军等,2014）。针对地质调查业务管理中涉及的北斗、工区、野外手机信令等空间数据，采用经纬度和坐标清洗标准化、不同空间参考转换至 CGCS2000 统一坐标系、新旧图幅号转换、凸包计算等在内的技术方法，实现集成处理。

（2）针对结构化、半结构化数据，采用“平抽-转换-清洗-关联”的数据处理流程，通过定义主数据与构建统一数据模型，实现数据集成过程中的数据标准化，完成数据的有效组织关联。

（3）针对非结构化地质调查管理技术文档，本地持久化后，读取非结构化文件在数据库表中的元数据信息，建立与文件对象间的数据关联，同时将元数据与文件转换，用于全量数据统一检索与分析，实现跨文件类型的有效存储和关联组织。

（4）针对个人信息等敏感数据，采用对称加密 AES 算法，使用自定义对称密钥，完成加密解密脱敏过程。

基于上述技术方法，针对不同数据对象，确定全量、增量、条件等同步方式，数据模型映射关系和质量控制规则。基于中台理念，通过数据集成处理任务开发、基于 Nginx 负载均衡的任务调度和统一监控，可以实现从数据集成任务开发、动态调度到运行监控的自动化动态集成处理。

2.2 异构数据统一组织管理

2.2.1 地质调查业务管理统一数据模型

针对语义集成，基于本体技术开展了相关研究（Wen et al., 2012），对于地质调查业务管理数据，存在数出多源、口径不一、语义二义、无法关联等问题，基于元模型（姜楠等,2018）和数据仓库理论技术，通过多源融合、语义映射、多维关联，构建地质调查业务管理统一数据模型（图 5）。梳理各业务域不同来源数据的模型结构，进行模型集成，形成地质调查管理业务的统一数据实体、属性和关系；基于空间位置和主数据，关联建立异构数据的组织模型，为多来源异构数据形成统一的数据描述和业务抽象。在统一数据模型与多来源数据之间，建立映射关系和更新机制，通过包含概念、逻辑与物理模型的统一数据模型，以保证数据完整性、标准性、一致性。

2.2.2 异构数据“湖仓一体”混合式组织管理

通过构建总体的数据策略和架构，采用“湖仓一

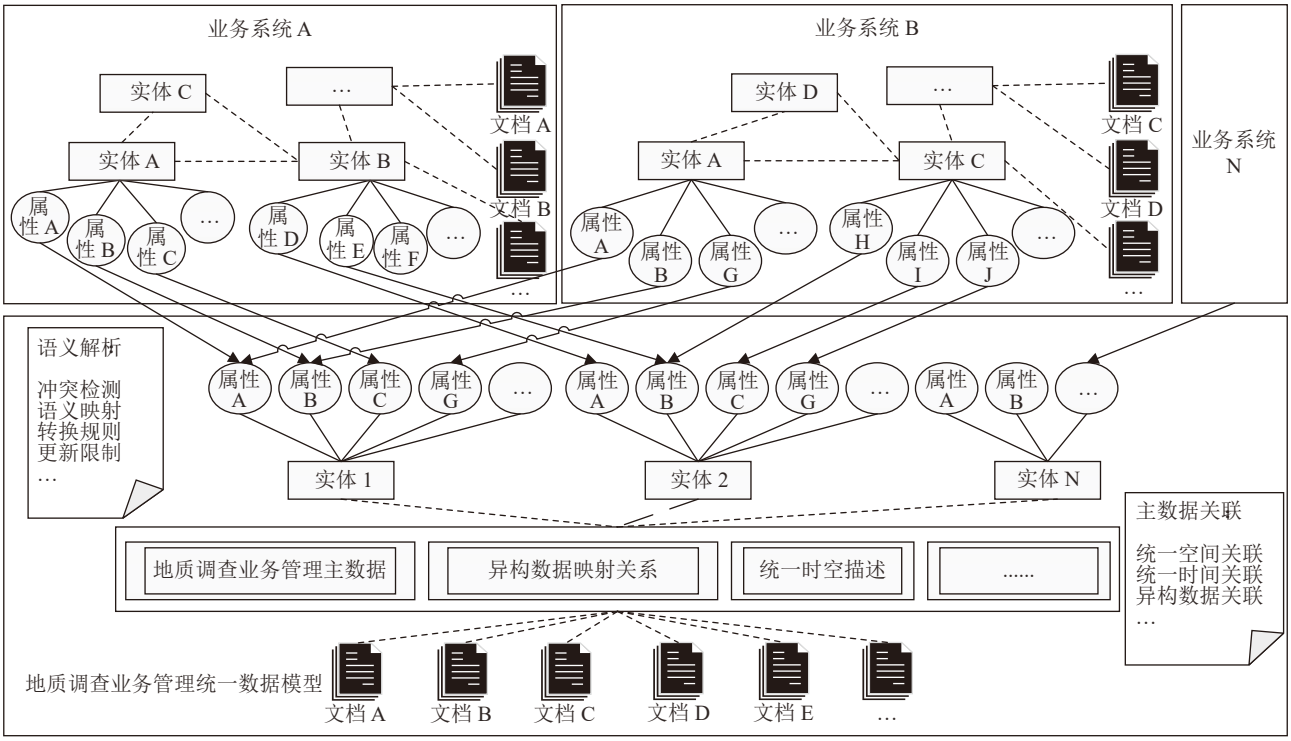

图 5 地质调查业务管理统一数据模型

Fig. 5 Universal data modal of geological surveying management

体”分布式混合存储, 实现地质调查业务数据的高效最优存储, 并利用支持空间/属性多条件的统一存取访问, 屏蔽数据结构与存储差异, 实现统一管理。

(1)为实现数据的有序集成管理, 采用横向覆盖多业务、纵向贯通多层级的总体数据策略, 构建了“源–缓冲–操作–仓库–集市”的 5 层数据架构。数据架构包括数据源层(DS)、临时数据存储层(TDS)、操作数据贴源层(ODS)、混合数仓层(DW, 包括汇总层与指标层)、数据集市层(DM)。其中, 临时数据存储层, 主要负责集成数据的本地临时存储缓冲; 操作数据贴源层, 主要存储经过轻度清洗转换后的数据; 混合数仓层, 对进一步处理进入统一数据模型的异构数据, 进行统一存储管理; 数据集市层, 主要是面向不同应用(张鸣之等,2013), 从混合数仓的全量数据中, 抽取转换形成的数据子集。

(2)“湖仓一体”架构, 是将传统数据仓库和 Hadoop、Spark 等大数据技术混合互补, 对结构化、非结构化等不同类型数据, 实现高可用、高可靠和可缩放的存储管理(Saddad et al., 2020)。在云环境中构建的 Hadoop 集群之上, 通过数据的关联组织与统一模型, 选用 Oracle、HDFS、Hive 等数据库, 配合 MySQL、PostGIS 等, 优化构建“湖仓一体”的混合式存储方案, 支持空间、结构化、非结构化在内的数据有效组织管理。针对结构化数据, TDS 层和 ODS 层主要由 Oracle 负责存储, 确保数据存取处理效率; 混合数仓层由 Hive 负责结构化数据存储、由 HDFS 负责存储非结构化等其他类型数据, 分布式存储确保数据安全和大数据量检索效率; 空间数据主要由 PostGIS 负责存储, 数据映射、关联和索引等数据组织信息, 由轻量级关系型数据库 MySQL 负责存储, 实现各类数据的最优存储。

(3)针对地质非结构化文档的内容检索, 已开展基于 Lucene 的探索应用(李超岭等,2015)。面向多类型异构数据, 采用分词技术、Lucene 搜索引擎和综合索引, 将结构化、半结构化等各类数据解析处理, 存储至 ElasticSearch, 并通过全局和空间索引、统一元数据、动态关联组织和检索分解合并, 构建统一存取访问层, 可以实现跨数据库、多类型数据的一体化管理。

2.3　地质调查业务管理决策模型方法

将地理信息与商务智能结合, 已经在公共卫生等领域辅助决策中, 取得研究进展(Sultan et al., 2013)。利用 GIS 结合 BI 的多维分析可视化、数据挖掘融合地理智能的分析决策模型(图 6), 辅助地质调查管理决策。

2.3.1　GIS 和 BI 融合分析可视化

采用多维数据预处理、二三维分析可视化、多维动态报表技术, 对地质调查业务的运行态势, 提供动态掌控和洞察, 服务各级管理决策。

(1)基于 OLAP 技术, 设定时间、单位、区域等数据维度和关键度量, 针对结构化数据, 开展数据立方体建模, 利用 Apache Kylin 进行数据预处理, 形成

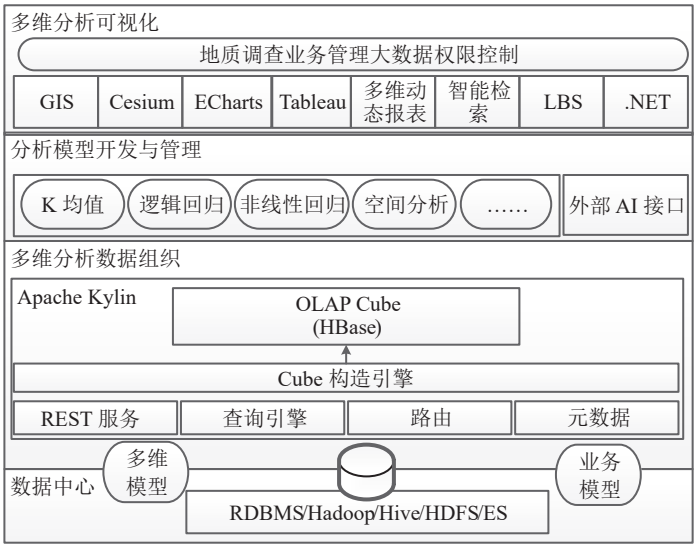

图 6　BI 与 GIS 融合的数据分析辅助决策技术框架

Fig. 6　Technical framework of data analysis for decision support

数据 Cube，提升数据多维汇总分析的效率。

（2）使用 Cesium 作为三维引擎、MapGIS 提供空间服务，整合可视化组件 ECharts、商务智能分析平台 Tableau，将二维和三维动态地图与仪表板结合，实现多元对象二维和三维可视化分析。按照用户角色，生成多级数据视图，实现数据的分级、动态、多维、时空的分析和可视化。

（3）采用角色权限约束下的灵活配置多维动态报表技术方案（图 7），通过多维数据报表、指标、维度等自由配置，报表动态挂载或下线，以及从功能级到菜单级、从菜单级到报表级的精细权限控制，实现多维动态报表的灵活授权和快速响应。

2.3.2 “描述–诊断–评估–预测”决策模型

在自然资源领域，有学者依托“一张图”数据，面向土地执法，开展了辅助决策系统研究（韩红太等，2019），其中辅助决策模型的研究值得进一步深入。针对地质业务管理特点，采用多元回归、决策树、K 均值等数据挖掘算法，融合空间分析技术，构建地质调查业务运行综合诊断分析模型，实现地质调查业务运行的动态评估预警。

模型分为描述、诊断、评估、预测 4 个维度，由基于空间分析的野外工作诊断、基于非线性回归的预算执行率预测、基于评分卡模型的项目承担单位综合动态评估、地质调查项目质量风险评估总体构成。从野外工作、经费执行、工作进度、工作质量、项目管理等维度，对国家地质调查工作的综合运行态势和风险，进行动态评估和预测预警。

（1）根据历年地质调查人员的野外北斗定位数据，计算全国不同区域的野外工作窗口期。结合每个项目的工区坐标与野外工作量数据，对项目野外定位数据进行交叉分析，结合年度累计野外工作时间、出野外时间、项目综合进度等数据，对项目野外工作开展情况进行动态评估（图 8）。

（2）基于历年全局预算执行流水数据，构建非线性回归预测模型，结合空间聚类算法，拟合出不同类型、不同地区地质调查工作预算执行规律（图 9）。对年底预算执行率进行预测、评价与预警。在指定执行率目标约束条件下，反向计算不同时间点的参考执行率，为划定全年不同时点考核线，科学推进预算执行，提供科学参考。

（3）承担单位综合评估采用评分卡模型，基于地质调查业务动态运行数据，选用人才队伍、工作经验、装备情况、历史任务完成情况、工作质量、诚信记录 6 个维度，共计 185 个关键指标，设置权重配比，通过计算和适度归一化，得出综合评估结果，对单位考核、能力评估等提供辅助。

（4）地质调查项目质量风险智能评估，采用相关性分析方法，结合主成分分析方法进行变量选取，采用 KNN 算法进行模型构建，利用 ROC 曲线、过拟合及欠拟合、交叉验证等方法进行模型评估，经分析后确定参数，得出项目运行质量风险 5 级分类结果。

3 实践与应用

3.1 地质调查业务管理大数据系统

根据总体研究成果，研发了国家地质调查业务管理大数据系统，包含地质调查业务管理数据中台、数据中心和应用系统 3 个部分。

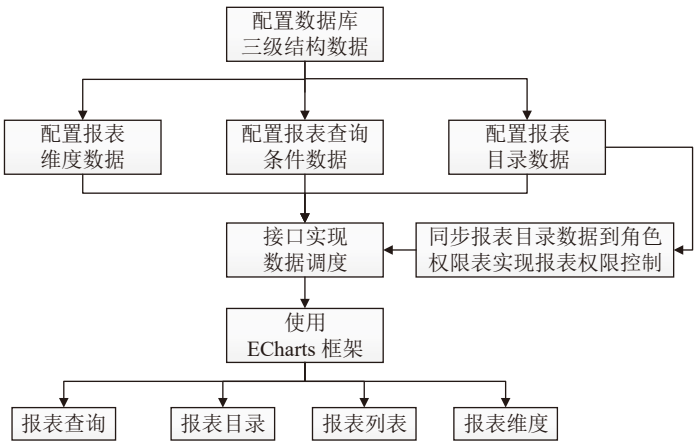

图 7 灵活配置多维动态报表技术路线

Fig. 7 Technology roadmap of multidimensional dynamic report

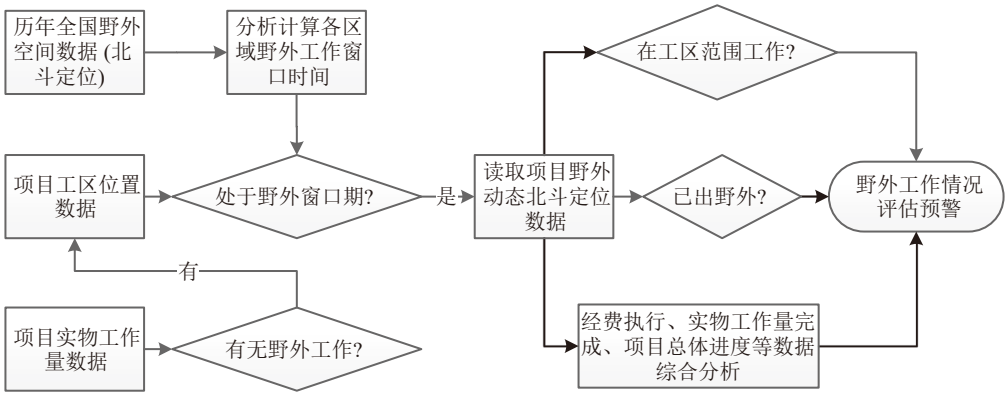

图 8　地质调查野外工作诊断评估模型流程图

Fig. 8　Flow chart of field work diagnostic evaluation model

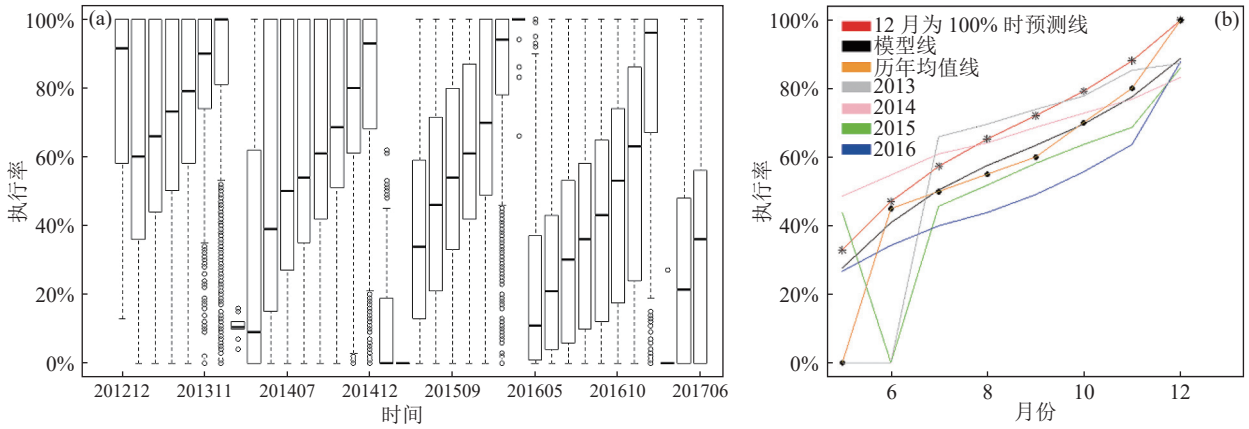

图 9　地质调查预算执行预测预警

Fig. 9　Geological survey budget execution prediction and early warning

a—预算执行历史数据分布箱图; b—预算执行预测拟合图

3.1.1　地质调查业务管理数据中台

研发了地质调查业务管理数据中台(图 10)。为多源异构地质调查业务管理数据的接入—集成—治理—服务, 提供了一站式全链路支撑平台。中台包括数据源管理、数据集成开发、集成处理任务调度、运维监控、数据质量、数据服务开发等模块, 且支持 Hadoop、Spark、Flink、Presto 等主流大数据计算引擎, 解决了地质调查不同业务管理系统、不同类型数据的快速接入和动态集成处理问题。

3.1.2　地质调查业务管理数据中心

建成了地质调查业务管理大数据中心(图 11), 实现了海量异构地质调查业务管理数据的统一组织管理。包括数据资源管理、统一检索和服务、统一元数据管理、统一数据备份与安全等功能, 解决了复杂异构的海量地质调查业务管理数据的一体化组织管理问题。

3.1.3　地质调查业务管理大数据应用系统

开发了地质调查业务管理大数据应用系统, 面向各级用户, 基于 Web 和移动端提供多端数据服务与决策支持(图 12)。系统包括地质调查业务管理一张图、综合态势、检索查询、报表中心、业务诊断、自助分析、智能助手等模块。面向全局用户, 实现了基于全量异构数据的统一数据检索; 面向各级各单位, 利用分层级、分单位的地质调查业务运行一张图, 为各单位实现业务运行态势一站掌控; 针对各级不同业务管理部门, 提供了权限控制的多维动态报表, 支撑跨部门、跨单位的业务协同; 通过地质调查业务运行综合诊断, 对工作综合运行情况, 进行动态分析和预警预测, 辅助各级管理决策。

3.2　应用效果

该系统已部署至中国地质调查局, 并经过了 5 年多的业务化运行。

图 10 地质调查业务管理数据中台

Fig. 10 Data middle platform of geological survey management

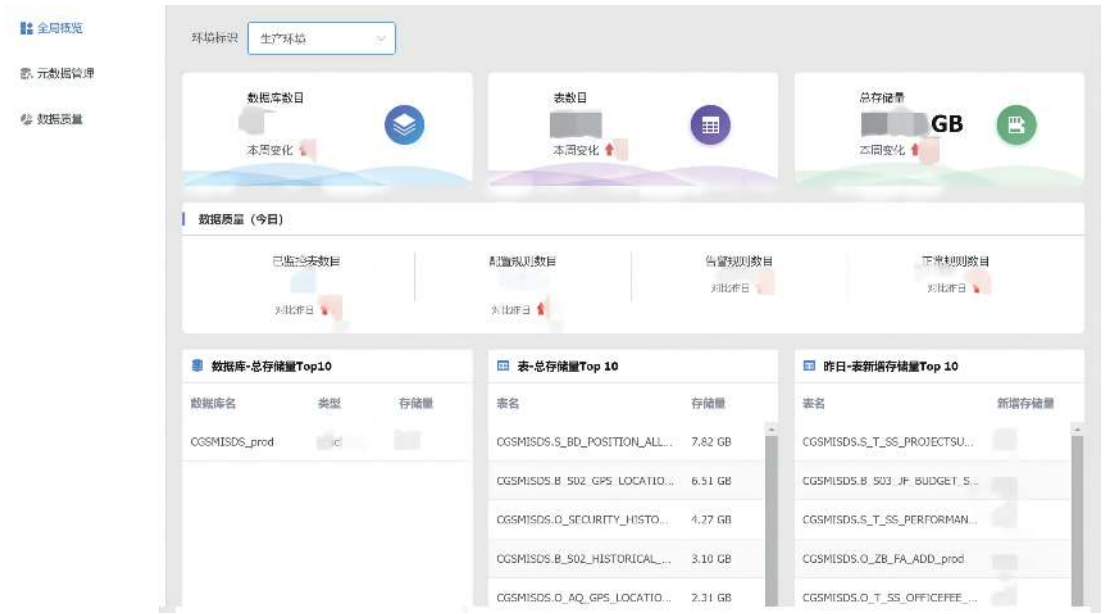

图 11 地质调查业务管理数据中心

Fig. 11 Data center of geological survey management

（1）地质调查业务管理数据中台：完成了中国地质调查局 24 个不同的系统数据源接入、284 个数据集成处理任务的在线开发和自动调度、7*24 小时稳定运行，实现了各类数据的自动化动态集成处理。

（2）地质调查业务管理数据中心：实现了 1999 年中国地质调查局建局以来的国家地质调查业务管理数据的有效组织管理与动态更新。目前结构化数据已超 1.5 亿条，各类非结构化文档已超 20 万档，并随着数据持续集成不断增加。

（3）地质调查业务管理大数据应用系统：面向地

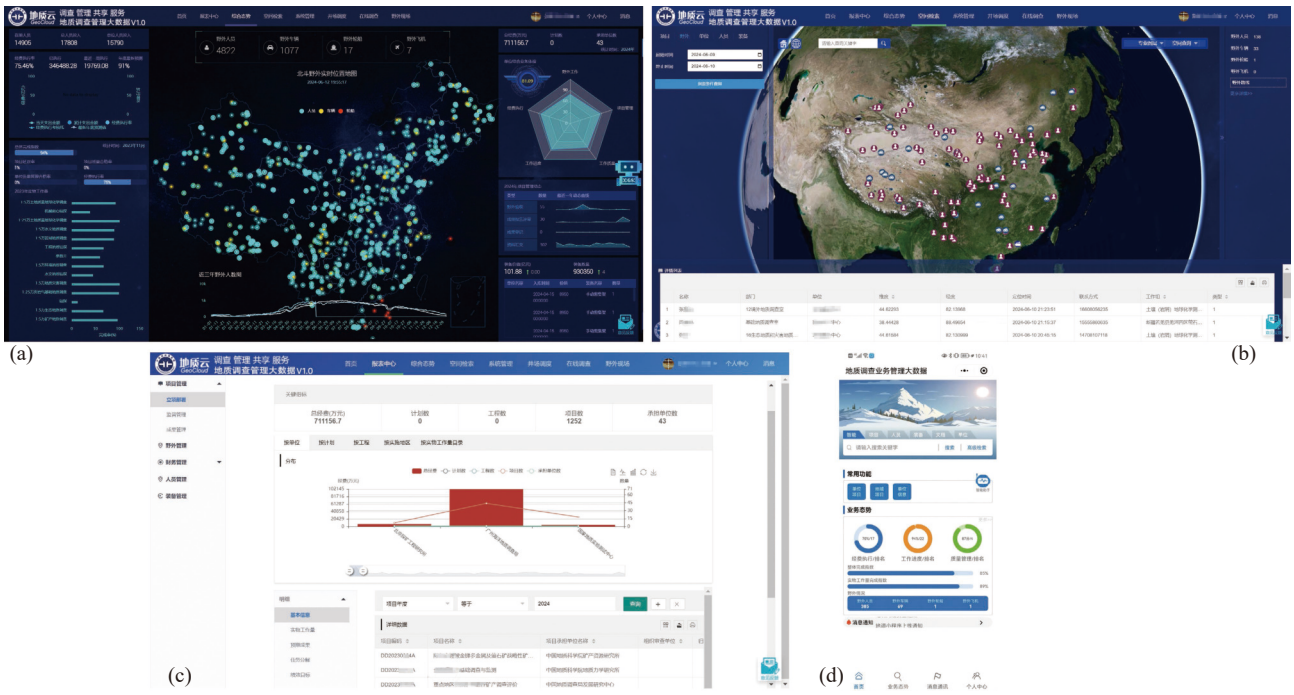

图 12　地质调查业务管理大数据应用系统

Fig. 12　Application system of geological survey management big data

a—地质调查业务运行一张图; b—国家地质调查业务管理综合检索; c—地质调查管理统一报表中心; d—地质调查业务管理数据移动端

质工作部署、实施监督、成果服务等各类业务, 通过数据有效驱动了流程优化、工作协同和决策服务, 已成为地质调查单位管理决策的工作平台。基于数据和分析决策模型, 连续 5 年发布《地质调查业务运行分析预警报告》5000 余份, 已成为国家地质调查工作指挥调度的支撑依据, 并形成了常态化工作机制。有效探索了国家地质调查跨层级、跨部门、跨业务的协同智能、精准靶向管理新模式, 服务了地质调查工作的组织实施。

4　结论与展望

(1)随着大数据技术发展, 多源异构数据集成走向在线动态化、趋于从数据接入到管理服务的全程化。将数据中台理念、ETL 与空间领域结合, 可以有效解决地学领域数据的自动化动态集成的问题。

(2)在关系型数据之外, 在地质领域还包括大量空间、物联网等数据, 面临海量非结构化等数据的共存管理问题。将分布式大数据结合传统数据库, 构建数据湖和数据仓库的混合式数据组织管理模式, 可以作为大数据环境下, 相关行业数据组织管理的一种有效方式。

(3)采用数据挖掘与地理智能的融合建模, 将空间分析和传统数据分析算法、商务智能和多维分析可视化有机结合, 对地质等空间属性较强的数据及应用领域, 是较有效的数据分析决策辅助方法。

在此基础上, 针对要素级时空数据集成治理的时空数据中台技术, 时序三维等空间数据和非结构化、结构化数据的统一组织互操作, 基于大数据和人工智能的时空分析辅助决策等方面, 值得进一步研究与深化。

参考文献

Aissi M, Benjelloun S, Loukili Y, et al. 2021. Data Lake Versus Data Warehouse Architecture: A Comparative Study[M]. New York: Springer.

Angelaccio M, Basili A, Buttarazzi B, et al. 2012. Using Geo-Business Intelligence to improve Quality of Life[C]//Satellite Telecommunications (ESTEL), 2012 IEEE First AESS European Conference on. IEEE.

Das Sarma A, Dong X, Halevy A. 2008. Bootstrapping pay-as-you-go data integration systems[C]//Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. ACM: 861-874.

Do H H, Rahm E. 2002. COMA: A system for flexible combination of schema matching approaches[C]//Proc. of the VLDB Endowment: 610-621.

Doan A H, Madhavan J, Domingos P, et al. 2002. Learning to map

between ontologies on the semantic Web[C]//Proc. of the 11th Int'l Conf. on World Wide Web. ACM: 662−673.

Halevy A, Korn F, Noy N, et al. 2016. Goods: Organizing google's datasets[C]//Proc. of the 2016 Int'l Conf. on Management of Data (SIGMOD '16). ACM: 795–806.

Lehmberg O, Bizer C. 2017. Stitching Web tables for improving matching quality[C]//Proc. of the VLDB Endowment, 10(11): 1502−1513.

Nasr M, Sultan T, Khedr A, et al. 2013. Dynamic AI−Geo Health Application based on BIGIS−DSS Approach[J]. IOSR Journal of Computer Engineering, 13. 36-42.

Nottelmann H, Straccia U. 2007. Information retrieval and machine learning for probabilistic schema matching[J]. Information Processing & Management, 43(3): 552−576.

Parimbelli E, Sacchi L, Bellazzi R. 2016. Decision support through data integration: Strategies to meet the big data challenge[J]. International Journal of Medical Research & Health Sciences, 12(1): 10−14.

Saddad E, El−Bastawissy A, Hoda M, et al. 2020, Lake Data Warehouse Architecture for Big Data Solutions[J]. International Journal of Advanced Computer Science and Applications, 11(8): 417−424.

Sultan T, Nasr M, Khedr A, et al. 2013. A Proposed Integrated Approach for BI and GIS in Health Sector to Support Decision Makers (BIGIS−DSS)[J]. International Journal of Advanced Computer Science and Applications, 4(1): 170−176.

Torre C, Guazzo G M, Ekani V, et al. 2022. The relationship between big data and decision making: A Systematic Literature Review[J]. Journal of Service Science and Management, 15: 89−107.

Wang Y Z, Jin X L, Cheng X Q. 2014. Network Big Data: Present and Future[J]. Chinese Journal of Computers, 36(6): 1125−1138.

Wen M, Tang X M, Shi S Y, et al. 2020. Semantic Integration for Multi−Source Geo−Data based on Ontology−A case integration of chart and map[C]//2010 3rd International Conference on Computer and Electrical Engineering (ICEE 2010), 7: 96−99.

韩红太, 焦利伟, 马林娜, 等. 2019. 自然资源管理辅助决策服务平台设计与实现[J]. 测绘科学, 44(6): 337−340.

韩家琪, 毛克彪, 夏浪, 等. 2016. 基于空间数据仓库的农业大数据研究[J]. 中国农业科技导报, 18(5): 17−24.

洪之旭, 陈浩, 程亮. 2017. 基于大数据的社会治理数据集成及决策分析方法[J]. 清华大学学报 (自然科学版), 57(3): 6.

胡侃, 夏绍玮. 1998. 基于大型数据仓库的数据采掘: 研究综述[J]. 软件学报, (1): 54−64.

姜楠, 文必龙, 林宗斌. 2018. 基于元模型驱动异构数据统一建模的研究[J]. 电脑知识与技术, 14(12): 4−5, 8.

李超岭, 李健强, 张宏春, 等. 2015. 智能地质调查大数据应用体系架构与关键技术[J]. 地质通报, 34(7): 1288−1299.

李文俊, 杨学强, 杜家兴. 2020. 基于数据中台的装备保障数据集成[J]. 系统工程与电子技术, 42(6): 1317−1323.

刘洪霞, 冯益明, 曹晓明, 等. 2018. 荒漠生态系统大数据资源平台建设与服务[J]. 干旱区资源与环境, 32(9): 126−131.

刘晴, 汤玮, 刘旭. 2020. 基于虚拟数据库技术的异地异构数据源整合[J]. 信息技术, 44(1): 130−133.

刘文军, 吴俐民, 方源敏. 2014. 基于 ETL 的多源异构空间数据集成技术研究[J]. 城市勘测, (2): 55−59.

刘文毅, 邓吉秋, 韩肖肖, 等. 2019. 大数据环境下地质资料的存储策略与文本化导入技术[J]. 地质学刊, 43(3): 367−371.

任晓霞, 喻孟良, 张鸣之, 等. 2018. 基于 Hadoop 分布式系统的地质环境大数据框架探讨[J]. 中国地质灾害与防治学报, 29(1): 130−134, 142.

苏萌, 贾喜顺, 杜晓梦, 等. 2019. 数据中台技术相关进展及发展趋势[J]. 数据与计算发展前沿, 1(1): 120−130.

王凯, 曹建成, 王乃生, 等. 2015. Hadoop 支持下的地理信息大数据处理技术初探[J]. 测绘通报, 2015(10): 114−117.

文敏, 唐新明, 史绍雨, 等. 2011. 针对海陆图融合的数字海图自动预处理及实现[J]. 地理空间信息, 9(1): 126−127, 135.

吴冲龙, 刘刚, 周琦, 等. 2020. 地质科学大数据统合应用的基本问题[J]. 地质科技通报, 39(4): 11.

徐佳沅, 文薪荐, 王思敏, 等, 彭晖儿. 2020. 一站式地理大数据智能化平台构建[J]. 测绘通报, (12): 132−137.

刘大杰, 陶本藻. 2000. 实用测量数据处理方法[M]. 北京: 测绘出版社: 79−81.

张鸣之, 喻孟良, 王勇, 等. 2013. 国家级地质环境数据仓库的设计与实现[J]. 地球科学 (中国地质大学学报), 38(6): 1347−1355.

赵伟伟, 王守东, 贾凉, 等. 2021. 地理信息中台在智慧城市中的应用——以南京市为例[J]. 工程勘察, 49(4): 57−61.

钟晓, 马少平, 张钹, 等. 2001. 数据挖掘综述[J]. 模式识别与人工智能, 14(1): 48−55.