技术方法

山东省地质资料集成与应用研究

吴红梅[1]，王志强[2]，吴友章[3]，毕亭亭[1]

(1.正元地理信息有限责任公司山东分公司，山东 济南　250101；2.青州市国土资源局，山东 青州 262500；3.山东中基地理信息科技有限公司，山东 济南　250101)

摘要：地质资料是极其宝贵的信息资源，也是开展地质工作的重要基础，利用大数据、云计算等技术，对地质成果资料进行集成与应用，有助于发挥地质成果资料的最大效能。通过全面系统收集山东省所有地质资料成果数据、成果图件、相关科研报告和地质论文，进行纸质资料扫描建库，利用 ETL 工具完成多源异构地质资料数据整合，完成山东省地质资料成果集成工作；研发“地质资料数据服务系统”，实现地质资料共享应用，系统结合云平台的设计思路，采用 SOA 架构，基于 Hadoop 和 Spark 进行搭建，支持数据库、文件等多源数据导入，提供地质资料空间数据查询检索、资料数据统计、空间分析、数据发布等功能，为不同专题的地质系统产品提供基础服务。地质资料数据服务系统为建立山东省“地质云”奠定基础。

关键词：地质资料；集成；应用；ETL；山东

中图分类号：P208　　**文献标识码**：B

引文格式：吴红梅，王志强，吴友章，等.山东省地质资料集成与应用研究[J].山东国土资源，2017，33(12)：70－74.

WU Hongmei，WANG Zhiqiang，WU Youzhang，etc.Study on Integration and Application of Geological Data in Shandong Province[J].Shandong Land and Resources，2017，33(12)：70－74.

地质资料具有广泛性、永久性和不可再生性[1-3]。地质成果资料是各类地质工作完成时形成的重要基础信息资源，可作为多次利用和再生产的重要基础，是地质勘探开发最宝贵的资源和财富[4-7]。国土资源部下发的国土资源信息化“十三五”规划(国土资厅函〔2017〕229 号)中第 24 项工作是：“构建全国地质信息协同服务体系”，要求：“开展地质信息服务聚合、数据资源描述与发现、大数据知识服务等理论技术研究；基本构建多层次、网格化地质信息协同服务体系；初步实现在地质大数据支撑平台框架下资源的统一汇集、互联互通[8]。”山东省绝大部分地质资料成果存储于山东省国土资源资料档案馆，但仍有部分地质资料成果存储于地勘单位中，并且各单位地质数据标准不统一，成果格式多样。山东省重视地质资料成果应用，山东省国土资源厅已完成多项地质资料成果应用系统，各局、院也开发了许多地质勘查、地质灾害、三维地质填图等应用系统，但这些系统多为试点项目或专项应用系统，没有大范围展开，也未相互建立联系并集成提供综合性的地学应用服务。山东省 TB 级的地质成果资料因为专业化的表达形式，使得地质成果资料只能服务于专业技术领域。有限的地质成果资料的信息服务对象和服务领域，不能很好的发挥地质成果资料的潜在价值。

1　研究目的与工作方法

随着社会经济的飞速发展，对地质资料的利用需求也随之大大增加[9-10]，利用大数据、云计算等技术，对地质成果资料进行分析挖掘，生成特色地质成果产品，更好地为国家、政府、民众提供服务，发挥地质成果资料的最大效能。

山东省地质资料集成应用工作方法见表 1。

2　系统设计

地质资料数据服务系统是地质资料大数据存

收稿日期：2017－07－17；**修订日期**：2017－09－12；**编辑**：王敏

作者简介：吴红梅(1978—)，女，黑龙江牡丹江人，高级工程师，从事于研发设计、项目管理等工作；E－mail：18653112787@163.com

储、计算平台，结合云平台的设计思路，采用 SOA 架构，实现基于数据服务的系统功能设计[14]。系统基于 Hadoop 和 Spark 进行搭建，支持数据库、文件等多源数据导入，可以进行分析代码和算法的导入，监控计算分析作业的执行，完成分析结果展现，实现完全平台化监控。地质资料数据服务系统共分为五个层次（图 1）。

表 1　地质资料集成应用工作方法

序号	工作方法	工　　作　　内　　容
1	山东省地质资料综合调查	全面系统收集山东省所有地质资料成果数据、地质成果图件、相关科研报告和地质论文[11]。对接整理“十一五”、“十二五”、“十三五”期间形成的地质资料成果，调查整理充分利用国土资源资料档案馆和地调局 2000 年以来的现有资料的数字化成果。针对 2000 年以前未进行扫描的纸质地质资料，遵循地质资料数字化工作流程、标准和相应规范，完成地质项目纸质资料数字化扫描建库工作
2	数据资源规划	调研分析省国土资源厅和各局院已建成的各地质、地勘、地灾等业务系统，遵循“一数一源”原则，理清行业对象定义及相互关系，业务与业务间的关系、业务与数据之间的关系的基础上，明确信息资源间的关联关系，保障数据的准确性与权威性[12]。以地理空间参考和数据库标准为统一基础，整合、集成矿山数据、土地数据、地质环境数据等成果数据，形成开放并可扩展的地质资料数据体系，以确保地质资料数据的唯一性[13]
3	地质资料集成整合	利用 ETL 工具完成多源异构地质资料数据整合。充分利用山东省现有地质工作成果，调查研究已建的各类地质数据库和应用系统，开发中间件，集成半结构化和非结构化数据源中信息，实现资源共享和利用。研究国家部委已建设完成的应用系统和数据标准，开发转化接口，实现已有成果的对接和利用
4	研发地质资料数据服务系统	为保障山东省地质资料成果发挥更大的效能，提供更快捷、智能的查询检索服务，研发地质资料数据服务系统。提供服务聚合、数据挖掘、数据建模和数据共享服务，建设玻璃地球和专题系统应用

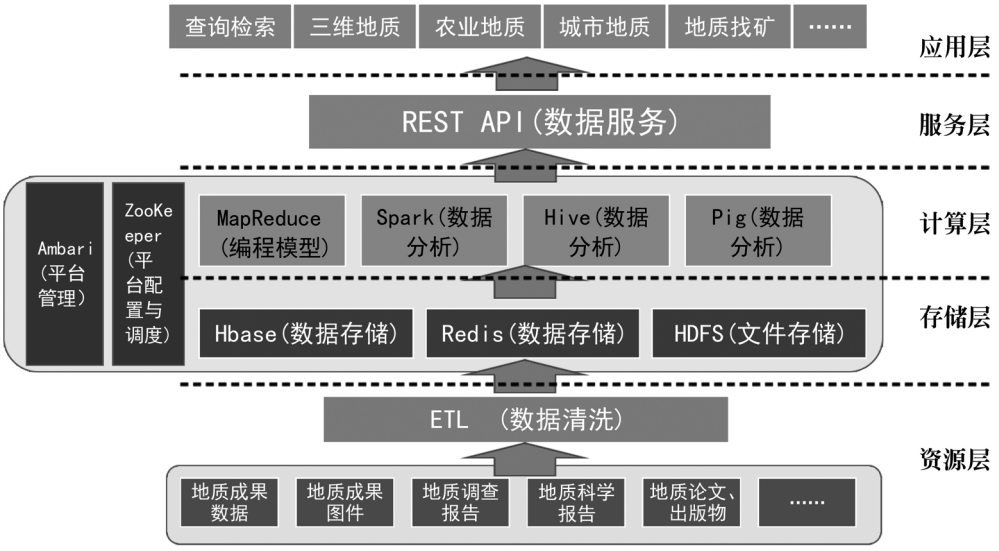

图 1　总体框架设计

从底层到顶层依次可分为资源层、存储层、计算层、服务层和应用层。资源层包含矿产地质、基础地质、农业地质、海洋地质、水文地质、工程地质、环境地质、地面沉降等地质成果数据和图件、相关科研报告和论文等；经 ETL（Extract－Transform－Load）工具进行数据抽取、转换和装载，形成标准化的地质资料成果数据，按照数据类型不同，分别存入 HDFS、Hbase 和 Redis，地质成果资料分别存储在不同结点，利用 Hadoop 本地计算的特性，减少数据的网络传输，充分发挥并行计算能力；数据计算分析和挖掘，采用 MapReduce、Spark、Hive 和 Pig，采用关键字生成算法，提供高效的地质资料成果检索查询；为方便数据共享和应用服务，提供 REST API 服务，方便应用系统调用。

3　关键技术

3.1　运用 ETL 工具完成多源异构地质资料数据集成整合

通过数据集成服务中的数据抽取引擎和数据计

算引擎等对获得的源数据进行转换和清洗，并对数据进行规范化，完成向地质资料数据中心的数据输送[15]。实现多源、异构数据源集成，并支持系统接口的动态扩展以及热部署，从而保证系统的灵活配置及扩展性需求。

数据集成服务的逻辑结构如图 2 所示：

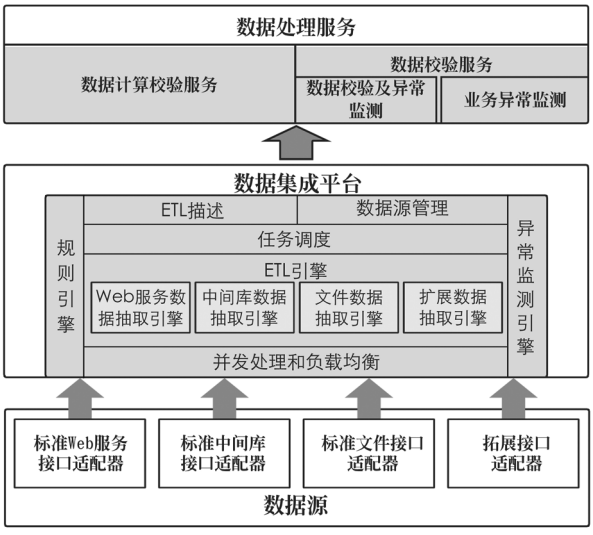

图 2　数据集成服务的逻辑结构

3.1.1　数据源管理

配置、管理和监测分布的数据源。描述每一个数据源的位置(IP)、接口适配器类型(文件方式、Web 服务方式、中间数据库方式等)、通信协议(FTP、HTTP 等)。

对于造成数据无法进行采集、数据采集不完整或不正确的数据采集方式的原因有：在数据源配置完成后，接口服务存在异常、通信异常、采集通道阻塞，数据源配置的方式不正确等情况。为解决上述问题，应实时监测数据源(图 3)。

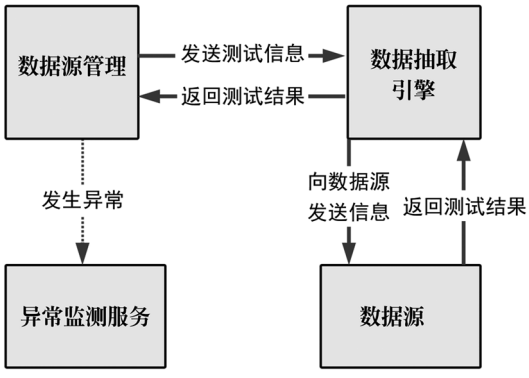

图 3　数据源监测

3.1.2　ETL 描述

ETL 是指数据抽取、数据转换以及数据加载[16]，提取、转换源系统中的数据成为一个标准的格式，并把数据加载到目标数据存储区的过程。对于 ETL 的过程描述采用可视化 ETL 配置工具，利用可视化的配置界面，通过拖拽操作及简单配置操作，完成各种数据源的配置操作及数据关系映射，可以以国际化标准的 XML 文件格式对 ETL 过程描述进行保存。

(1)资料档案 ETL 抽取描述

资料档案数据的特点是更新频率慢，抽取过程中需要进行代码转换、编码影射等过程。主要处理步骤为：

①去除重复记录，保证对象在系统中唯一性；

②编码映射：抽取过程中需要将数据源中的本地编码通过编码映射到地质资料数据中心中，保证各项数据的统一；

③字段选择：过滤符合要求的属性，去除数据源中非标准数据属性。

(2)负载均衡

数据源可能处于不同的网络区域，负载均衡能够有效的平衡各个接口服务器之间的负载压力，有效的提升数据采集效率。负载均衡使用基于轮转周期的动态反馈负载均衡算法，该算法结合了静态加权轮转算法的简单性、高效性和动态反馈机制的实时性等优点[17]。

3.2　采用关键字生成算法实现数据的动态分析

有效收集和管理地质资料结构化和非结构化数据，以此为基础进行统一的集成深度分析，通过结构化和非结构化数据之间的逻辑关联来获得更多的有效知识[18]。

在研究结构化和非结构化数据动态分析技术研究中，结构化和非结构化数据一体化管理和分析的关键技术之一便是采用关联查询与检索的技术，这种有机的统一查询分析处理，能够带给用户良好的数据分析体验。

在传统工作中，查询返回的结果是严格的精确的，查询对象限于结构化数据(使用 SQL)或半结构化 XML 数据(使用 XQuery)；与此相对，信息检索常用于对无结构化的文本或半结构化的 Web 网页数据的检索，检索结果是非精确的。为实现数据的动态分析，要扩展信息检索能力，能够对结构化和非

结构化数据进行统一的基于关键词的检索，将检索结果融合展现，在元数据支持下，能够支持基于概念的检索。

(1)基本框架

扩展信息检索能力，利用 SQL 检索出结构化信息，并从中自动提取出相应的关键词，然后利用这些关键词检索出非结构化信息，扩展了结构化查询能力，通过查询的方式，实现数据的动态分析(图 4)。

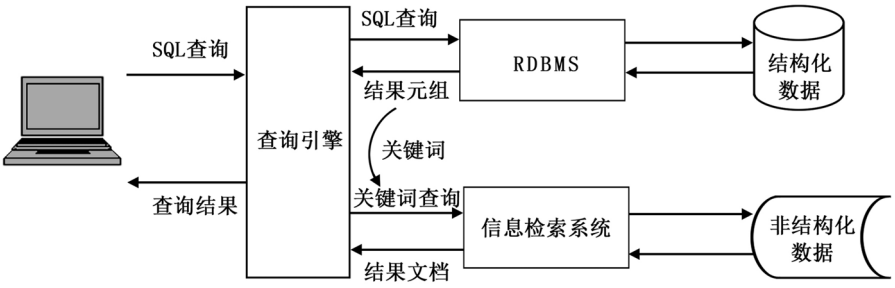

图 4　基本框架

(2)关键字生成算法

关键字生成算法是系统的核心部分，它有 2 个主要任务：一个是数据的范围，即从哪些表中选取关键词；二是逐步构造整个关键词表，即制定策略将新的数据加入进来。对于数据的范围，尽可能多地利用数据库中数据，因此不能仅限于 SQL 查询语句中指定的表的范围。可以利用数据库中表之间的主外键联系，将新的表添加进来，这样就大大扩展了可以选择的数据的范畴，所得到的关键词也会更加丰富，且和查询请求具有相关性。另一个是逐步构造关键词表。表之间的主外键联系比较复杂，一个表存在多个外键，和多个表相连。算法实现中利用了贪心算法的思想，在每一步添加新的表的时候，只选择一个与查询最相关的外键，并将该外键指向表包含进来。如此这样递归执行下去，直到没有新的表添加进来或者已经达到预先设定的最大递归深度。

4　功能及应用

地质资料数据服务系统提供地质资料空间数据查询检索、按时间轴浏览资料数据，资料数据统计、空间分析、数据发布等功能，为不同专题的地质系统产品提供基础服务。系统还提供云平台状态、文件管理、数据管理、平台服务等功能。

服务聚合——提供目录服务、地图服务、文档浏览、数据下载等在线数字资源服务聚合。

数据挖掘——建立地质成果图谱，关注地质成果资料的位置和时序变化，以三个空间维度和一个时间维度概念进行数据挖掘分析，描述不同时期地质空间的差异化形态，动态展现地质资料的空间形态结构和时空变化规律，赋予地质资料图形和谱系双重特性，“图”是指地质图、地质矿产勘查规划、地质项目范围等专题地图；“谱”是指按项目特性、时间序列所建立的资料成果体系。以地学知识、海量地质成果数据为驱动，坚持技术服务于业务的原则，通过数据计算、分析、挖掘，为地质找矿，城市地质、土地利用等提供科学数据支撑，从而满足政府、专业人员和社会公众的需求。

数据建模——建立地质项目卡片，纵向覆盖地质项目实施过程的各主要环节，横向覆盖到进度安排、人员投入、资金监管、照片记录、资料上报、预警提醒等内容，构建项目信息卡片。以卡片形式提供可视化、图形化、扁平化方式的地质资料快捷检索，根据用户检索热度和下载关注度模型进行卡片权重调整，能够主动推送信息到前端，方便用户使用。

建立玻璃地球应用，应用三维技术，搭建可视、多维、精确的山东地质三维模型；实现山东地质成果多维一体化存储、管理。

5　结论

(1)通过山东省地质资料集成应用，全面收集山东省分散存储的地质资料成果，梳理山东省地质资料成果和现存系统的关联关系，建立开放可扩展的山东省地质资料数据体系。

(2)运用大数据、云计算技术，搭建集约管理、资源共享、低耗拓展、应用创新于一体的“地质资料数据服务系统”，完成海量地质资料的快速浏览和便捷查询。

(3)“地质资料数据服务系统”为建立山东省“地

质云”奠定基础。

(4)为山东省各市、区地质灾害调查、城市地质、土壤污染防治、群测群防、搬迁避让、治理工程等提供地质信息支撑。

参考文献:

[1] 梁其华.对原始地质资料立卷归档与汇交问题的研究[J].中国国土资源经济,2016,(3):6-13.

[2] 王成锋,王丹辉,李乔乔,等.基于 GIS 的地质成果资料汇交管理系统研究及应用[J].山东国土资源,2017,33(7):82-85.

[3] 高延梅.现阶段原始地质资料立卷归档问题之我见[J].黑龙江国土资源,2005,(5):38-39.

[4] 袁宏,赖德军.可视化地质资料管理与共享平台研究[J].计算机与数字工程,2013,41(3):420-422.

[5] 党杰.广东省地质资料自动化管理系统建设探讨[J].山东国土资源,2012,28(7):66-68.

[6] 孙丽华,李树辉.浅议地质档案资料的专业化管理[J].吉林地质,2013,31(3):134-136.

[7] 张书波,张引,张斌,等.成果地质资料检索系统研究与实现[J].国土资源信息化,2016,(2):38-44.

[8] 叶恺,杨昊.浅析地质资料管理系统的建立与应用[J].中国化工贸易,2012,4(4):27-28.

[9] 李东风.浅析辽宁省地质资料管理信息服务系统建设[J].中国科技成果,2012,(18):34-37.

[10] 杨丽君.上海地质资料数据中心的构建与运行体系[J].上海国土资源,2012,33(3):79-84.

[11] 李军.信息资源规划在国土信息化中的重要作用[J].信息技术与信息化,2015,(3):142-143.

[12] 冯永玉.省级国土资源“一张图”数据中心建设探讨[J].山东国土资源,2014,30(11):67-70.

[13] 冀正强.基于 Web 数据分析的就业信息服务平台的设计实现[D].济南:山东大学,2013:18-44.

[14] 王芳,刘伟,吴红梅.建设项目压覆矿业权及矿产地应用研究与系统实现[J].山东国土资源,2016,32(1):66-70.

[15] 尤玉林,张宪民.一种可靠的数据仓库中 ETL 策略与架构设计[J].计算机工程与应用,2005,41(10):172-174.

[16] 许少华,夏智伟.基于轮转周期的动态反馈负载均衡算法[J].计算机技术与发展,2013,23(6):63-66.

[17] 邢胜南.基于 MDA 的数据分析过程研究[D].济南:山东大学,2010:31-44.

Study on Integration and Application of Geological Data in Shandong Province

WU Hongmei[1], WANG Zhiqiang[2], WU Youzhang[3], BI Tingting[1]

(1.Shandong Branch Corporation of Zhengyuan Geographical Information Limited Corporation, Shandong Jinan 250101, China; 2.Qingzhou Bureau of Land and Resources, Shandong Qingzhou 262500, China; 3. Shandong Zhongji Geographic Information Science and Technology Limited Corporation, Shandong Jinan 250101, China)

Abstract:Geological data is not only rare information resource, but also the basis for carrying out geological work. Based on big data and cloud computing technology, integration and application of geological data have been carried out. It can elaborate the best efficiency of geological information. Through systematic and comprehensive collection of all kinds of geological data, achievements, maps, related scientific research reports and geological papers of Shandong province, paper data scanning database construction has been carried out. Based on ETL tools, integration of multi—source heterogeneous geological data can be realized, and geological data and geological achievements can be collected. Study and development on the "service system of geological data" can realize the sharing of geological information. This systm is set up combining with the design idea of cloud platform, using SOA, and based on Hadoop and Spark. It can support multi—source data import of database and files, provide spatial data query, data statistics, spatial analysis, data release and other functions, and provide basic services for products of different special geological system. Geological data service system lays the foundation for the establishment of "geological cloud" in Shandong province.

Key words:Geological data; integration; application; ETL; Shandong province