智能地质调查大数据应用体系架构与关键技术

李超岭[1,2], 李健强[1], 张宏春[3], 龚爱华[3], 魏东琦[4]
LI Chaoling[1,2], LI Jianqiang[1], ZHANG Hongchun[3], GONG Aihua[3], WEI Dongqi[4],

1.中国地质调查局发展研究中心,北京 100037;
2.国土资源部地质信息技术重点实验室,北京100037;
3.武汉中地数码科技有限公司, 湖北 武汉 430073;
4.中国地质调查局西安地质调查中心,陕西 西安 710054
1. *Development & Research Center of China Geological Survey, Beijing* 100037, *China;*
2. *Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Beijing* 100037, *China;*
3. *Wuhan Zondy Cyber Science & Technology Co., Ltd., Wuhan* 430074, *Hubei, China:*
4. *Xi'an Center of Geological Survey, CGS, Xi'an* 710054, *Shaanxi, China*

摘要:地质调查数据主要由结构化和非结构化多样性的数据构成。由非结构化多样性数据文件组成的报告,由于技术原因,长期以来一直以传统的目录文件方式进行存储。这种存储方式导致数据的查询、统计、更新等操作不但低效,而且非常不利于检索、查询、挖掘等应用,使得数据服务能力极低。通过把Hadoop生态体系融入中国地质调查云平台架构,基于Hadoop HDFS和HBase存储架构,建立非结构化地质数据基础内容库存储组织模式,采用Lucene全文搜索引擎架和地质领域本体词库构建快速随机访问的索引文件机制,改变了多样化、碎片化的复杂地质调查非结构化数据的存储、阅读、搜索和应用模式,为智能地质调查提供精确、快速服务奠定基础。

关键词:智能地质调查;地质调查非结构化数据;分布式存储系统Hbase;Hadoop生态体系
中图分类号:P628 **文献标志码:**A **文章编号:**1671-2552(2015)07-1288-12

Li C L, Li J Q, Zhang H C, Gong A H, Wei D Q. Big data application architecture and key technologies of intelligent geological survey. *Geological Bulletin of China*, 2015, 34(7):1288-1299

Abstract: Geological survey data are mainly composed of structured and unstructured data. The composition report, consisting of Word, PDF, excel, graphics, pictures, video, PPT and other unstructured data files, is stored in files directory traditionally due to technical reasons. This traditional storage mode is not only inefficient in data retrieving, statistic analysis, updating and other operations but also not conducive to searching, querying and mining applications, thus resulting in extremely low data service capability. Through incorporating Hadoop ecosystem into cloud platform architecture of China Geological Survey, the authors established the geological data base of unstructured content library storage organization model based on HDFS and HBase storage architecture, and set up indexing mechanism for fast random access using full-text search engine 'Lucene' and geological domain main part lexicon to change the diversity of complex fragmentation of storage, reading, retrieval and application mode for geological survey unstructured data, and to provide accurate basis for fast service of smart geological survey.

Key words: intelligent geological survey; unstructured data of geological survey; Hbase distributed storage system; Hadoop ecological system

收稿日期:2014-04-15;**修订日期:**2015-02-15
资助项目:中国地质调查局项目(编号:12120115073201、1212011120436、1212011120449)
作者简介:李超岭(1957-),男,博士,研究员,从事智能地质调查技术和地质调查信息云计算与大数据技术相关理论、标准、方法与技术研究。E-mail: lichaoling@126.com, lcholing@mail.cgs.gov.cn

从20世纪80年代以来至今,在地质调查领域逐步建立了全国1∶250万、1∶100万、1∶25万、1∶50万、1∶20万地质图空间数据库,全国地质工作程度数据库、全国矿产地数据库、1∶20万自然重砂数据库、1∶20万水文地质数据库、1∶50万地质环境数据库、同位素测年数据库、区域地球化学数据库、区域重力数据库、航空磁测数据库、航天遥感数据库、地质图文资料数据库、地质文献资料数据库等一系列基础地质数据库,同时,完成了大量的1∶5万地质图数据库、航空影像数据库、成果资料、地质文献数据等数据建库工作,开展了全国钻孔数据库建库工作。这些数据建库的特点是,大多数据采用关系数据模型和GIS数据模型的方式来建立相关的数据库。

与结构化数据数量相比,非结构化数据在地质成果中所占比例要比结构化数据大得多,所涉及的内容更丰富,潜在的价值也是其他数据无法比拟的。以某幅1∶25万区域地质调查报告为例,多样化碎片化复杂地质调查非结构化数据占结构化数据的比例高达85%以上。这种由Word、PDF、Excel、PPT图形、图片、视频等非结构化多样性数据文件组成的报告,由于技术原因,对海量的非结构化多样性的数据构成的文件显得无能为力。如地质报告类,通常由不同格式和结构的正文类、附表类、附件类、附图类、多媒体类、审批类、其他类构成。一部报告可由上百个文件乃至上千个文件构成。这种具有碎片化、非结构化和多样性特点的数据,非常不利于计算机组织存储,导致一直以传统的目录文件方式进行存储。这种存储方式导致数据的查询、统计、更新等操作低效,而且不利于检索、阅读、挖掘等应用,使得这些内容丰富的数据服务能力不高,且应用率极低。

使海量、碎片化、非结构化与多样性的数据高效、快速存储,并利于数据挖掘、发现和应用,是大数据时代信息技术发展的热点。近年来,由云计算、社交计算和移动计算三大趋势推动的大数据技术[1]NoSQL技术[2]正在重塑业务流程、IT基础设施,以及人们对于数据、信息和知识的捕获与使用方式。数据源从关系数据库管理的结构化数据扩充到蕴含大量信息且数据巨大、无统一数据模型的非结构化数据;数据服务从常规分析转向深度分析;计算方式从集中式的串行计算向大规模的并行计算技术发展。这些诸多技术因素的变革及人们从数据中获得知识的客观需求,致使新一代信息技术不断出现,大数据技术就是典型的代表。

大数据相关技术的应用使人们拥有了获取、管理、分析、共享非结构化数据的能力。NoSQL技术的应运而生,给非结构化地质数据的同质化和面向数据挖掘的组织及应用带来机遇和变革。为高效、快速存储地质报告等相关非结构化碎片化和多样化的数据并利于数据挖掘、发现,笔者在中国地质调查网格平台架构中融入了Hadoop生态体系[3],基于Hadoop HDFS和HBase存储架构,提出了非结构化地质数据基础内容库存储组织模式,基于Lucene全文搜索引擎架和地质领域本体词库,构建地质知识内容库和快速随机访问的索引文件机制,改变了多样化碎片化复杂地质调查非结构化数据的存储、阅读、搜索、挖掘和应用模式,为智能地质调查提供精准地质知识内容和快速服务,以及以野外地质调查泛在服务为特征的智能地质调查模式奠定了基础。

1　智能地质调查大数据应用体系架构

1.1　智能地质调查定义与总体体系框架

智能地质调查以效率、精度、知识发现(数据挖掘)、服务、持续为坐标,以互联网、物联网、通讯网、云计算、大数据、智能技术(和设备)等新一代技术组合为框架,以地质调查、管理和服务智能化为研究内容。智能地质调查与数字地质调查技术[4]比较具有以下特点:①使野外数据采集更智能(智能技术与智能设备的引入)、感知化,如数据采集效率和精度更高,操作更简便,或有更多的知识模型提供支持;②智能地质调查更强调从领域或行业分割、相对封闭的信息化架构迈向复杂巨型系统的开放、整合、协同的智能地质调查信息化架构,发挥地调局信息化资源的整体效能,为地质调查提供泛在服务;③更注重通过泛在网络、移动技术实现无所不在的互联和随时、随地、随身的智能融合服务;④提供更多的数据处理和评价技术,使综合分析在新一代技术支持下更加智能,同时方便获取服务;⑤在智能地质调查框架下,通过动态数据获取、知识发现、数据挖掘等技术,实现天地空一体地质调查的智能管理和调度。智能地质调查体系框架如图1所示。

1.2 智能地质调查云(平台)架构

根据智能地质体系建设需求,本文在中国地质调查信息网格平台框架[5]的基础上,构建了智能地质调查云(网格)平台架构(图2)。通过MapGISK10的T-C-V云架构[6],升级与搭建了地质调查云服务仓库、构件生产中心、管理中心和应用生产中心,增加Hadoop大数据平台基本组件,对非结构化数据进行组织和数据挖掘。通过MapGIS K10基础云服务实现地质空间和非结构化数据的一体化管理和服务。在资源聚合层,强化对等式资源管理器的扩展,其资源总类重点增加了非结构化数据资源的管理。

1.3 基于智能地质调查云平台的大数据应用体系架构

基于大数据技术智能地质调查数据和发现,服务框架从底层到顶层依次可分为大数据资源层、汇聚层、数据挖掘与分析层及大数据应用层。在大数据资源层,为了能够快速地将成果文件存放到大数据存储介质中,采用Sqoop、Avro和Flume等主流的大数据存取工具来提高存储效率。在汇聚层,将原始地质成果资料文档存放到分布式文件系统HDFS/HDFS2中。为了对文件数据进行快速获取,将原始文档重新组织后存放在分布式实时访问数据库HBase中。其中,附图、附表、附件等文件均单独存放,主文件则按章节分开存储。同时对存储在HBase中的内容建立索引,存放到分布式缓存Memcached或Redis中。这样只需从内存中获取索引进行查找,可以极大地减少磁盘的I/O工作,便于下一步进行数据挖掘时快速检索定位文件。

在数据挖掘与分析层中,对地质成果资料文档中所蕴含的大数据进行分析处理之前,首先采用地质领域本体库和文本搜索框架Lucene对地质文档进行分词处理。地质领域本体库是专门为地学领域的科研、教学及语言比较研究而收集的文本集合。然后在Mahout框架中进行数据挖掘。利用Mahout可将机器学习中的多种算法有效地扩展到Hadoop集群平台上,通过与第二代Hadoop系统中的资源管理与计算调度框架结合,可实现海量数据的快速挖掘和并行处理,从中获取用户所需的地学信息知识并进行分析结果的可视化展示,为上层基于大数据环境支撑的地学信息综合应用提供决策支持。在大数据应用层,依托大数据处理技术可对用户提出的地学问题进行智能化分析处理,将其转化为地学问题求解任务,进而在大数据平台中将地学问题求解任务转化为并行任务执行,最后将执行结果返回给客户。

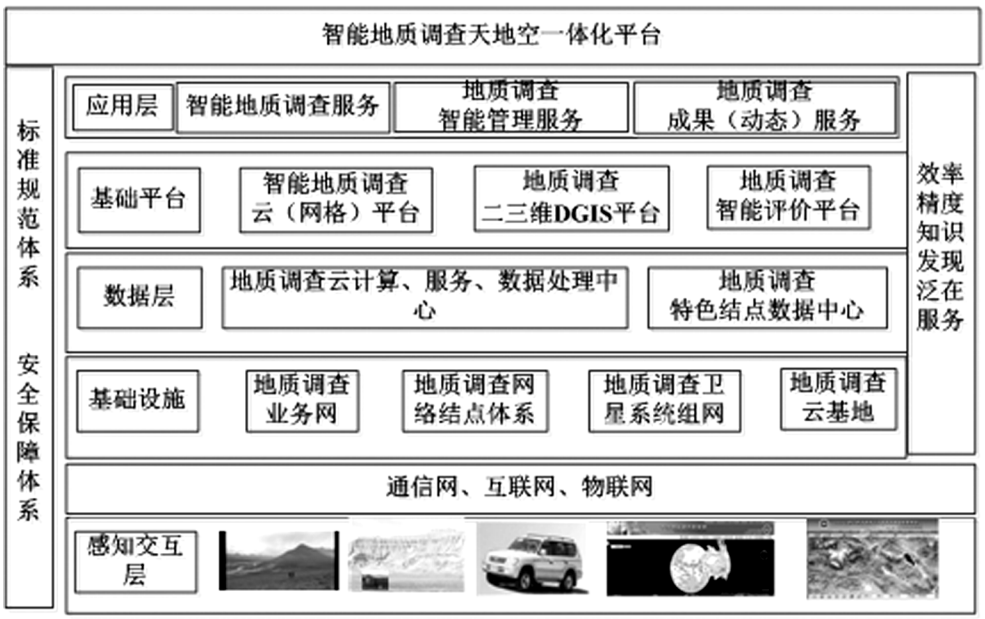

图1 智能地质调查云(网格)平台体系总体框架

Fig. 1 Overall framework of intelligent geological survey cloud (grid) platform system

图2　基于大数据技术的智能地质调查云(网格)平台架构

Fig. 2　Cloud (grid) platform architecture of intelligent geological survey based on big data tech

目前,该框架已在中国地质调查信息网格平台集成,并在天地空地质调查工作管理和安全保障服务中开展示范应用,为智能地质调查体系提供大数据技术支撑。

2　智能地质调查大数据应用体系架构关键技术

2.1　多样化碎片化复杂地质调查非结构化数据存储模型

2.1.1　数据特点

多样化碎片化复杂地质调查非结构化数据是地质调查成果报告最具特征的数据。以中华人民共和国地质图(H45C003004)日喀则市幅 1:25 万区域地质调查报告为例,一幅完整的 1:25 万区域地质调查报告合计有 308 份不同类型的文档,文件类型包括 Word、PDF、JPG、TIF、MapGIS、电子表、遥感影像、航卫片、显微照片等,数据内容涉及区域地质图编图说明书、地质图、矿产地质图、环境地质图、区域地质调查报告、区域地质调查报告专题报告——雅鲁藏布缝合带中段蛇绿岩成因、演化及工作方法研究、野外总图、实际材料图、编稿原图、野外记录本、野外手图、TM 遥感影像图、实测剖面记录表、地质资料质量检查卡、岩石光谱、岩石薄片、硅酸盐、稀土、热释光、粒度分析、孢粉、微体化石、岩矿绝对年龄测定、C^{14}测年、构造岩组和粒度、显微构造、电子探针、测年样品送样单、岩石薄片鉴定报告、化石鉴定报告、稀土分析报告、岩石光谱样分析报告、粒度分析报告、硅酸盐样分析报告、微体化石鉴定报告、岩石样分析报告、化石鉴定报告、同位素年龄测定报告、地质照片集、放射虫显微照片、区调野外验收决议书、原始资料检查记录、原始资料检查意见等。具体分类与数据特点见表 1。

2.1.2　基于 HBase 的多样化碎片化复杂地质调查非结构化数据存储模型

(1)基于 Hadoop HDFS 和 HBase 的存储架构

HBase 是一个面向列的非关系型分布式存储系统。它基于 Hadoop HDFS 文件存储系统,使用 MapRedue 处理海量数据,利用 Zookeeper 作为协同服务,使用简单的键值对映像关系为超大规模和高并发的海量数据实时响应系统提供很好的解决方案[7]。基于 Hadoop HDFS 和 HBase 的存储架构见图 4 。

HBase 是一个面向列存储的数据库,类似于 Google 的 BigTable。HBase 的物理数据存在 HDFS 文件系统集群,只有少量的元数据需要 Zookeeper 集群和 HDFS 集群来维护,因此,HBase 可以存储大表(百万行数据以上的表)。HBase 采用分布式 Master/Slave 方式,由 Master 进程负责管理 RegionServers 集群的负载均衡及资源分配,Zookeeper 负责集群元数据的维护并且监控集群的状态以防止单点故障。每个 RegionServers 会负责具体数据块的读写。

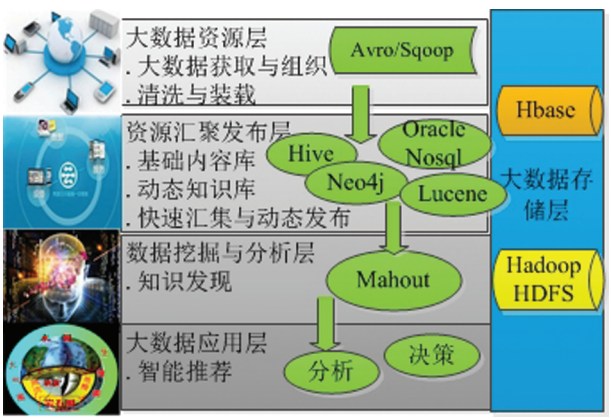

图3　基于大数据技术智能地质调查数据发现与服务总体框架

Fig. 3　Overall framework of intelligent geological survey for data discovery and services based on big data tech

表1 日喀则市幅1:25万区域地质调查报告内容分类

Table 1 Content classification of 250,000 regional geological survey report of Xigaze Sheet

文档名称	具 体 内 容	数据数量	数据类型	数据特点
成果报告	中华人民共和国区域地质调查报告1:25万日喀则市幅(H45C003004)中华人民共和国区域地质调查报告专题报告——雅鲁藏布缝合带中段蛇绿岩成因、演化及工作方法研究1:25万日喀则市幅H45C003004) 、关于建立日喀则国家地质公园的建议、成果总结	4份	Word文档	非结构化
主要成果图件	中华人民共和国地质图(H45C003004)日喀则市幅、中华人民共和国矿产地质图(H45C003004)日喀则市幅、中华人民共和国环境地质图(H45C003004)日喀则市幅	3幅	GIS图件	结构化
验收文件	地质调查项目成果报告认定书、地质调查项目成果报告初审意见书、地质调查项目成果报告评审意见书、原始资料初审竟见书、图幅中期评估意见、野外验收决议书、科技档案验收合格证	9份	Word文档	非结构化
野外记录本	野外路线记录、实测剖面记录	78本	XML或JPG或电子表或Word	非结构化
各种编图图件	野外手图、综合手图、野外总图、实际材料、编稿原图	31幅	GIS图件或JPG	非结构化
影像及解译图	遥感影像图、航卫片像解译图	11幅	GIS图件或JPG、TIF	非结构化
实测剖面	实测剖面记录表、实测剖面小结、实测剖面柱状图、实测地层剖面图	54份(幅)	GIS图件、Word、电子表	非结构化
样品	岩矿、化石、光谱、硅酸、稀土样送样单、微古送样单($1P_2$)、岩石光谱、岩石薄片、硅酸盐、稀土、热释光、粒度分析、孢粉、微体化石、岩矿绝对年龄测定、C^{14}测年、构造岩组和粒度、显微构造、电子探针、测年样品	35份	Word、电子表	结构化
质量检查	地质资料质量检查卡(地质路线、剖面)、原始资料检查记录、原始资料检查意见、质量检查工作小结	22册	Word、电子表	结构化
测定报告	岩石薄片鉴定报告、化石鉴定报告、稀土分析报告、岩石光谱样分析报告、粒度分析报告、硅酸盐样分析报告、微体化石鉴定报告、岩石样分析报告、化石鉴定报告、同位素年龄测定报告	21册	Word、电子表	非结构化、结构化
照片	放射虫显微照片、报告照片集、地质照片集	5集	JPG	非结构化
设计文件	总体工作设计、工作布置图、地质矿产草图、专题工作设计、编图说明书、计审查认定书	7份	Word、GIS	非结构化、结构化

Hadoop文件系统HDFS是能够运行在普通硬件之上的分布式文件系统,可实现大流量和大数据的随机读取。类似Google的CFS,其架构是典型的Master/Slave形式, Master 节点上启动一个进程Namenode(存放文件的元数据), 每个Slave节点上启动一个进程Datanode(存放具体的文件内容),Namenode与每个Datanode之间的通讯通过Heartbeat的方式进行通信。一个文件以多个block的形式存放在多个Datanode 节点上,每个block有多个副本,副本存放的具体位置按照hadoop放置的算法决定。HDFS的缺省block大小(64Mb)和副本数(3个)还可以重新设定。

(2)HBase存储模型

HBase利用HDFS作为底层存储系统,为其提供高可靠性、高性能、列存储、可伸缩、实时读写的数据库系统。因此,HBase具有通过大量容错分布式节点存储大量数据的能力。HBase使用列存式数据库存储海量数据,以表的形式存储数据,由行和列组成,其中列由很多的列族(rowfamily)组成。HBase数据模型以表的形式存储数据,它是一个稀疏多维度的排序的映射表,其特点是:①数据表大,一个表可以由百万级别或千万级别的列和行组成;②对象是列表,可以单独针对列(族) 的存储和检索;③稀疏性,对于空列(null),并不占用存储空间[8]。

HBase表行记录由3个基本类型RowKey、TimeStamp和Column构成。其中RowKey是唯一

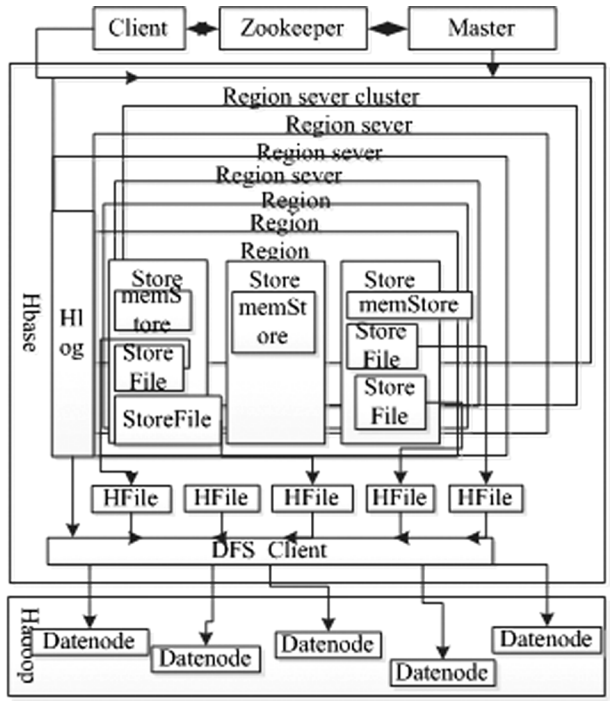

图4　基于Hadoop HDFS和 HBase的存储架构

Fig. 4　The storage structure based on Hadoop HDFS and HBase

标识,TimeStamp是每次数据操作的时间戳,每个更新都是一个新的版本,Column 列划分为若干个列族(row family),列名字的格式为:<family>:<label>,每一张表有一个family 集合,这个集合是固定不变的,但是label 值相对于每一行来说是可以改变的。初始Table对应一个Region,这个Region按照Family个数被切分为多个Store,每个Store包含一个或多个HFile,HFile是实际的存储文件。但HFile的长度超过配置参数时(256MB),其所在的Region一分为二,避免过多的分裂是提高HBase操作效率的关键。

(3)多样化碎片化复杂地质调查非结构化数据存储模型

如何充分利用HBase数据模型的特点,在不影响效率的情况下,对多样化碎片化复杂地质调查非结构化数据化"散"为"整",化"异构"为"同构",面向数据挖掘或形成"知识片段(知识库)"是非结构化数据集组织与分析方法的难点和重点。

通过对原始资料的访问处理,在不丢失原始资料信息量的原则上组织和描述数据的真实内容,使得计算分析更容易贴近数据描述的本质和发现数据之中蕴藏的知识。本文提出了基于HBase的多样化碎片化复杂地质调查非结构化数据存储模型:基础内容库存储模型和扩展动态内容库演化模型。

①基础内容库存储模型

基础内容库的目的就是化"散"为"整",化"异构"为"同构",除了为快速还原原样和方便阅读提供方便以外,另一个重要的目的就是为动态内容库演化模型奠定基础。基础内容库存储策略见图5。

地质调查数据原始资料存储在HDFS中,将数据的访问和存储分布在服务器集群之中,在多备份存储的同时还能将访问分布在各个服务器之上。HDFS 通过分布式计算的算法,将数据访问均摊到服务器阵列中每个服务器的多个数据拷贝之上,提供了极高的数据吞吐量。当整个系统容量需要扩充时,只需要增加DataNode的数量,系统会自动、实时地将新的服务器匹配到整体阵列中。之后,文件的分布算法会将数据块搬迁到新的NameNode之中,不需任何系统停机维护或人工干预。但是由于地质调查数据原始资料数量繁多且文件较小,HDFS存在着小文件问题,如图5所示,在HDFS之上扩充一套文件系统,将小文件打包成大文件,在减少namenode内存使用的同时,仍然允许对文件进行透明的访问。

基础内容库建立包含资料基础元数据的提取(资料名称、大小、空间范围、子标题等信息)、机械分页后的文本内容、图表内容、附件二进制(MapGIS或其他文件格式)内容等。HBase内容库包含多个逻辑属性组(列族)的任意属性值(列),如表2所示。

在具体部署时,要注意以下2个问题:①在

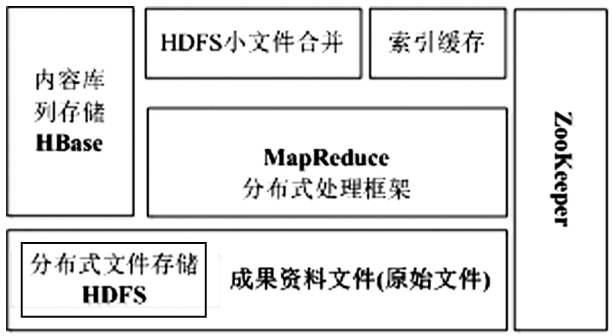

图5　基础内容库存储策略

Fig. 5　Basic content library storage strategy

HDFS中，每个Block对象约占150byte内存，如果有10 000 000个小文件则NameNode大约需要2G内存；在存储模型分析时，要考虑NameNode内存使用问题；②在应用HDFS文件系统时需要注意对大量小数据文件组成的集合处理效率较低的问题[9]。

②动态知识内容库存储模型

内容复杂、存在形式多样的大型非结构化数据蕴含的信息和知识并非以传统的关系性方式清晰表述，而是大多包含在非结构化的自然文字数据中。因此，基于基础内容库建立知识和特征内容库是进行有效表述的关键，同时也是进一步知识发现的基础。

建立知识内容库数据模型的目的在于重新构造属性或知识片段，以帮助数据挖掘和快速发现。在不丢失信息量的原则上尽量简单地组织和描述数据的真实内容，使得计算分析更容易贴近数据描述的本质和发现数据之中蕴藏的知识。模型力求对数据在语义清晰、组织高效的基础上，进行描述性和结构性建模，以便促进知识集成、共享和复用。

在传统的关系数据库领域，可将每个属性组赋予一个表，表结构与表关系决定了数据库架构。如果设计合理且数据规模适当，这样的系统是灵活的、可维护的。但是，当数据量超过一定规模且数据库架构需要调整时，传统设计就会遭遇致命的缺陷，对包含数百万或数亿行的关系数据库架构进行修改操作，或对架构的更改进行正确性和完整性检查是极其困难的。分布式数据库HBase可以避免以上缺陷，HBase列可以动态添加，在数据分析阶段，不需要对数据做特别的准备和运行时的处理工作。HBase建立在HDFS之上，具有高可靠性、高性能、列存储、可伸缩、实时读写的特性，适合存储非结构化和半结构化的松散数据。在内容库基础上可基于MapReduce实现即时查询及数据分析功能。

动态的创建知识内容库，是在基础内容库的基

表2 分多样化碎片化复杂地质调查非结构化数据存储模型

Table 2 Storage model of fragment and complex geological survey unstructured data

<table>
<tr><td colspan="9">表名[archive_table]</td></tr>
<tr><td rowspan="6">RowKey:
[档文件路径对应MD5码]</td><td colspan="8">列簇Column Family 1 [basic_info]</td></tr>
<tr><td rowspan="2">列</td><td>Key:
arch_id
value:
档案ID</td><td colspan="2">Key:create_time
value:
创建时间</td><td>Key:
authrity
value:
权限</td><td>Key:user
说明:所属
value:
用户</td><td>Key:size
value:
大小</td><td>Key:
name
valu:
名称</td><td>Key:pnt_
[FilePath]
Value:
提取文档空间位置坐标</td></tr>
<tr><td colspan="2">Key: title_ml_
[FilePath]
value:
ml提取文件名(注:ml指正规提交的归档文件的目录清单文件</td><td colspan="2">Key:title_djb_
[FilePath]
Value:
登记表提取文件名</td><td colspan="2">Key:title_src_[FilePath]
Value:
根据文档内容提取文件名</td><td colspan="2">Key:
title_cat_[FilePath]
Value:
自定义格式提取文件名</td></tr>
<tr><td colspan="8">列簇Column Family 2 [text_content]</td></tr>
<tr><td rowspan="2">列</td><td colspan="2" rowspan="2">文本页组[txt.]
(包括文档文本或空间数据注记形成的文本片段)</td><td colspan="6">图片组[img.]</td></tr>
<tr><td colspan="3">缩略图件(包括文档或空间数据全图范围生成的缩略图)</td><td colspan="3">原始图件</td></tr>
<tr><td colspan="2">Key:
txt.[FilePath].split.[Number]
value:
文本html片段,其中[Number]表示html片段序号</td><td colspan="3">Key:
img.[FilePath]/icon.[FileExt]
value:
包括图片文件及内嵌word文档里的缩略图</td><td colspan="3">Key:
img.[FilePath]/big.
[FileExt]
value:
包括图片文件及内嵌word文档里的原始图</td></tr>
</table>

注：[FilePath]—档案下某文件相对路径；[FileExt]—档案下某文件扩展名；[Number]—数字序列号；[Image-Name]—图片全名

础上,再次进行数据挖掘和学习,重新构造属性或知识片段,以帮助数据挖掘和快速发现。如地质报告,就是把一篇完整的报告,按照章、节或段落内容重新组织,形成知识片段,每个片段都会给出机器学习后的内容标签,帮助数据挖掘和快速发现。以中华人民共和国地质图(H45C003004)日喀则市幅 1∶25 万区域地质调查报告地层为例,在基础内容库的基础上可动态形成持久化地层知识库,至少有以下几个方面具有关联的知识内容库:①形成地层库,对本图幅地层特点进行综合描述;②通过地质领域本体,对不同地层进行语义一致性处理(标签);③形成剖面库,把原始剖面测量数据(含素描)与综合描述绑定;④形成地层对比图库;⑤形成与地层相关的照片、样品、样品鉴定报告库。

根据动态知识内容库存储模型的特点,本文确定了在不丢失原始资料信息量的原则上,组织和描述数据的真实内容,使得计算分析更容易贴近数据描述的本质和发现数据中蕴藏的知识。提出了包含多个逻辑属性组(列族)的知识内容库存储模型。存储模型基本与表 1 相同,不同点是文本的列族由页片段变为特定内容片段。

2.2　数据存储模型扩展:GIS 数据与内容库一体化

GIS 数据具有空间、时间和专题属性,地质调查成果资料中往往包含所属地区的地质图等空间数据。多样化碎片化复杂地质调查非结构化数据存储模型扩展,就是把 GIS 数据也作为内容库的一部分。以 MapGIS 数据为例,首先将 MapGIS67 格式的文档转换成 MapGIS10 文档格式,提取 MapGIS67 文档的地理范围、注记文本内容存储到内容库(HBase)中,注记文本内容的提取使得根据内容检索图件成为可能,区别于非矢量图件只能按文件名检索的方式,GIS 图件信息作为内容库的组成部分,与成果资料内容一起,支撑着地质大数据以后的数据挖掘;GIS 数据的可视化由 MapGIS10 IGSS 服务提供,空间数据处理流程能自动识别、转换、发布 GIS 数据,分布式文件系统(HDFS)存储原始的 MapGIS67 格式文件,作为 Web 服务器浏览地图服务的数据来源。空间数据与非结构化数据集组织与分析方法流程图见图 6。

2.3　数据发现与挖掘模式

快速搜索技术是内容搜索和知识发现领域研究的热点。在智能地质调查中,如何根据野外地质人员实时提出的问题,从基础内容库海量文档中快速返回相关知识片段,并按照用户兴趣相关度进行高效排序,同时根据聚类分析,找出相关推荐文档,及时指导野外工作,是未来智能地质调查重要应用模式之一。

在众多开源搜索引擎开发工具中,Lucene 是全文搜索引擎架构和开发工具包之一,以其优异的索引结构、高性能、可伸缩、跨平台、易使用性、开源等特性,被广泛地用来构建实用的全文搜索应用系统,或被集成于多类软件开发环境中。智能地质调查大数据应用体系架构的数据发现与推荐系统是基于 Lucene 全文搜索引擎架构开发的。使用 Lucene 时,选择合适的分析器非常关键,选择分析器的因素之一是待分析的语种,另一个因素是被分析的文本所属的行业领域。地质调查行业有特定的术语和缩略词,这时就需要创建自定义的分析解决方案,Lucene 的分析器支持自定义扩展。目前,开始接入中国地质领域本体库、地名词

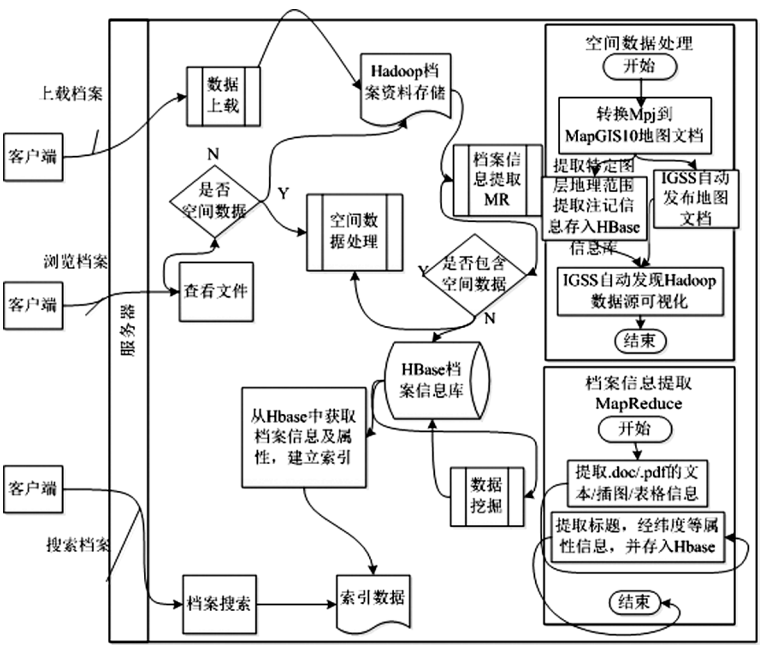

图 6　空间数据与非结构化数据集组织与分析方法流程

Fig. 6　Organization and analysis method flow of spatial data and unstructured data set

汇库(约80万个词或术语)的试验,取得了比较好的效果。

多样化碎片化复杂地质调查非结构化数据发现与挖掘模式包含三大内容:建立内容索引库、搜索和聚类推荐。

索引是采用一定的方法将基础内容库中的关键信息提取出来,并按照特定的数据结构和逻辑关系组织起来,形成一个支持快速随机访问的索引文件的过程(图7)。

Lucene 的分析指将域(Field)文本(在基础内容库指页,在知识库指知识片段)转换为索引表示单元(Term)的过程,分析器对分析操作进行了封装,分析器通过执行若干操作,将文本词汇单元化,这些操作包括提取单词、去除单元符号等,最终将文本流中的词汇单元与关联的域(Field)名对应形成了各个索引表示单元(Term)。分析期间提取的词汇单元被编入到索引项中,索引后的项就是可以搜索的项。

在Lucene内部,分析器的重点任务是将域的内容词汇单元化。HTML、Word、PDF、XML文档中包括了如作者、标题、章节标题等许多元数据,当索引一个完整文档时,这些元数据被分离,并索引成单独的域,分析器只分析特定的域并将域中的内容分解成词汇单元。因此为了对域进行分离,应预先解析文档,从而用独立的文本块表示域。

系统采用Lucene内置分析器StandardAnalyze将字母数字混合编列的词汇、e-mail地址和中文字符转化为语汇单元。Lucene支持自定义分析器的扩展,如处理关键词域、同义词、别名、其他表示相同意义的词等,都可以通过自定义算法来扩充完善分析器。为使分词更加符合地质专业内容库,笔者在Lucene开发了接口,并开展Paoding、mecab等分析器的试验。

搜索与推荐则是对索引的逆向操作。Lucene的搜索过程对搜索语句进行词法分析和语言处理,得到系列检索词,然后通过语法分析得到一个搜索树,再依据搜索树通过存取索引模块将相关索引调内存,得到与每个检索词相应的倒排文档索引链表。对文档链表众多文档ID进行布尔运算,并得到搜索结果文档集。对搜索结果文档集采用基于VSM的文档相关度计算模型和基于TF-IDF的权重评分机制等技术进行排序[10],在此基础上,通过聚类计算,将相关文档返回给用户。聚类算法MR流程见图8。算法将采用MapReduce分布式计算框架,先经过统计词频向量化[TF-IDF]预处理,得到文本的向量值,然后利用余弦距离公式和kmeans算法进行迭代计算,得到聚类的结果,然后再将结果信息保存在Hbase表中,便于搜索计算。最后对欧式距离或汉明距离等测试对比,将距离相近的项目归类到一起,给出聚类结果。

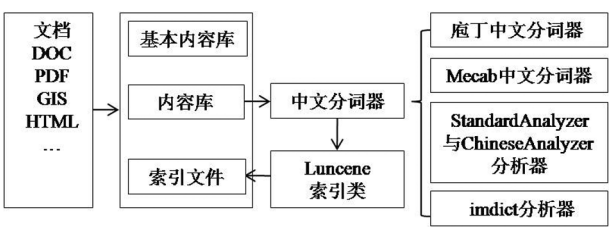

图7 支持快速随机访问的索引文件过程

Fig. 7 File indexing process for fast random access

2.4 数据集组织与分析方法流程

对地质调查非结构化文档大数据进行预处理,首先是将已有成果资料文档按照一定的数据结构进行序列化,存储到分布式文件系统中。然后对数据进行格式转换,将数据并行发布到分布式数据库中进行重新组织,并利用大数据缓存技术,根据资料成果的逻辑结构构建可高速访问的数据索引层。同时对存放在分布式文件系统中的数据,利用文本提取工具提取出来,建立全文索引,存放在分布式数据库中,供文本数据分析和挖掘。用户可以通过浏览器,提交领域问题,通过智能化分析处理,得出相应的解决方案。流程如图9所示。

2.5 数据结点部署模型

目前,中国地质调查信息网格在全国有25个结点,分布在大区中心、大学、地矿局、地调院等部门。根据各结点硬件资源状况,提出最小部署模型设计(图10)。

(1)主从节点部署模型:主节点作为管理节点,

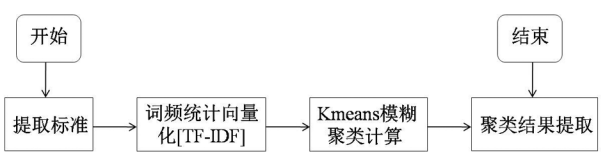

图8 聚类算法MR流程

Fig. 8 MR clustering process

从节点作为数据和计算节点。如果条件允许,主节点可以配置2台,工作模式为主备模式;从节点至少2台,工作模式为负载均衡。如果含大量GIS数据,应考虑增加GIS应用节点。

(2)HDFS 部署规划:主节点部署HDFS 的命名节点。如果条件允许,主节点配置2台,另一台的命名节点为备用。从节点部署HDFS 的数据节点。

(3)MapReduce 部署规划:主节点部署MapReduce的Jobtracker为主用。如果条件允许,主节点配置2台,另一台也部署MapReduce的Jobtracker为备用。从节点部署MapReduce的tasktracker。

(4) Hbase 部署规划:主节点部署 Hbase 的Hmaster,如果条件允许,主节点配置2台,另一台也部署 Hbase 的 Hmaster 为备用。从节点部署 Hbase 的RegionServer。

(5) ZooKeeper 部署规划:主从机分别部署Zookeeper。

3　智能地质调查大数据应用体系架构关键技术应用

为了验证本文复杂地质调查非结构化数据存储模型和数据发现与挖掘模型等关键技术,在省级信息资料中心,依据多样化碎片化复杂地质调查非结构化数据结点部署模型,部署了4台服务器,Hadoop平台为1.2版。传统存储档案文件约10237档、1766350个多样化和碎片化的数据文件,存储空间约3.5TB。根据本文提供的内容复杂的地质调查非结构化数据集组织与分析方法,对这些文件进行了试验,取得良好效果。部分部署技术指标参数见表3。

仍以中华人民共和国地质图(H45C003004)日喀则市幅 1:25万区域地质调查报告为例,数据映射到Hbase中存储的一条记录,内容见表4。

基于大数据技术的非结构化地质数据组织与集成架构研究,以Hadoop技术体系为基础,解决了非结构化地质数据的同质化和面向数据挖掘的组织问题,使多样化、碎片化的数据同质化和整体化,并为智能地质调查位置服务的快速数据挖掘和智能服务奠定了基础。

在多样化碎片化复杂的地质调查非结构化数据集组织与分析方法的基础上,笔者对省级资料馆开展了数据挖掘应用示范。首先基于大数据技术的非结构化地质数据组织与集成架构,从内容上把10237档、约176万个文件转变成一体化"同构同库"的数据体,为快速发现数据和多种数据挖掘方法试验奠定了基础。为验证数据挖掘的效果,针对新疆萨尔托海铬铁矿岩体特征问题,基于近10000档形成的基本内容库进行数据挖掘和发现,约3s左右获得新疆萨尔托海铬铁矿岩体特征约15个有效知识片段。图11是针对新疆萨尔托海铬铁矿岩体特征,通过基本内容库进行数据挖掘,从中找到相关内容片段目录的示例。通过数据挖掘,准确给出了新疆萨尔托海铬铁矿岩体特征的结果(图11)。

4　结　论

(1)智能地质调查大数据应用体系架构把智能

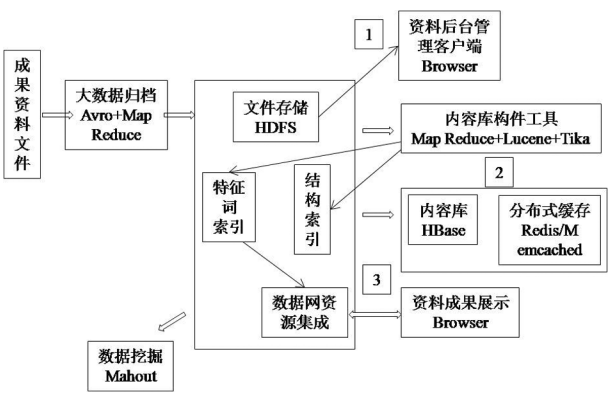

图9　内容复杂的地质调查非结构化数据集组织与分析方法流程

Fig. 9　Flow chart about organization and analysis method for complex content of unstructured geological survey data

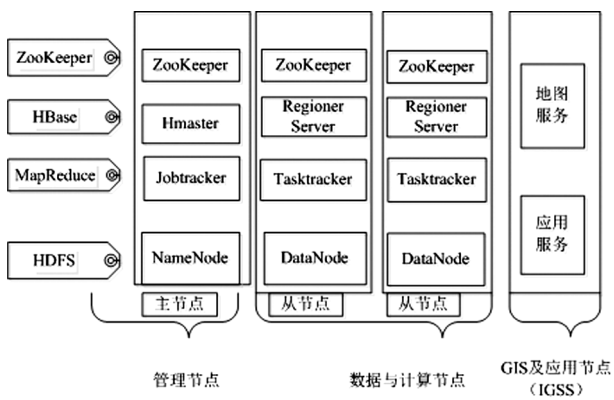

图10　多样化碎片化复杂地质调查非结构化数据结点部署模型

Fig. 10　Node deployment model of diversified fragmentation complex geological survey unstructured data

表3 部署技术指标参数

Table 3 Parameters and technical indicators

部署内容	时间	工作量	操作模式
系统部署	1d	4台服务器	交互
上载	39h	约1766350个文件	全自动化
基本内容库建立	1d	约500档数据资料	全自动化

地质调查云平台和Hadoop生态体系有效的融合一体，将极大地改变以“图”形式为主的服务模式，激活大量有价值却难以被利用的多样化碎片化复杂的地质调查非结构化数据，使信息更加多元化、丰富、和完整表达。

（2）多样化碎片化复杂的地质调查非结构化数据存储模型是未来地质调查数据组织的重要方式之一，将彻底改变中国地质调查领域一直采用回溯性方式整理原始数据的模式，不仅为国家节约大量的资金，同时为数据的应用从数据到知识高级应用奠定基础。

（3）基于地质领域本体的多分类内容扩展组织模型，是未来地质调查知识库或内容库建设的新模

表4 1∶25万区域地质调查报告文档集映射到Hbase中存储的一条记录内容

Table 4 1∶250000 regional geological survey report documentation set mapped to a record in Hbase

表 名	archive_table	
RowKey	MD5(“日喀则”)	
	Key	Value或说明
列族1 [basic_info]	arch_id	日喀则
	create_time	330584268037348
	authrity	777
	user	hadoop
	size	293854357
	name	中华人民共和国1:25万日喀则市幅区域地质调查报告(1-5章)
	title_ml_z01_0001.doc	中华人民共和国1:25万日喀则市幅区域地质调查报告(1-5章) [其他ml中的标题]
	title_src_z01_0001.doc	中华人民共和国区域地质调查报告比例尺:1:250000日喀则市幅(H45COO3OO4) [其他.doc/.pdf文档中标题]
列族2 [text_content]	Key	Value或说明
	txt.z01_0001.doc.split0	Z01_0001.doc中文本片段0 [Z01_0001.doc其他文本片段]
	txt.z01_0001.doc.split37	Z01_0001.doc中文本片段37 [其他doc/pdf中的片段]
	img.z01_0001.doc/image0.jpg/big.jpg	z01_0001.doc中图片0二进制数据 [z01_0001.doc中其他图片二进制数据]
	img.z01_0001.doc/image28.jpf/big.jpg	z01_0001.doc中图片28二进制数据
	img.T01_0001.JPG/icon.jpg	T01_0001.JPG图标图片数据
	img.T01_0001.JPG/big.jpg	T01_0001.JPG大图图片数据[Web快速可视化]
	img.t01_0001/t01_0001.JPG/icon.jpg	t01_0001/t01_0001.JPG图标图片数据
	img.t01_0001/t01_0001.JPG/big.jpg	t01_0001/t01_0001.JPG大图图片数据[Web快速可视化]
	MapgisPrj/z01_0001/ct01_0102/ct01_0001.mpj	对应Mapgis工程文件二进制数据
	MapgisJPG/z01_0001/ct01_0102/ct01_0001.mpj	自动生成工程文件范围缩略图

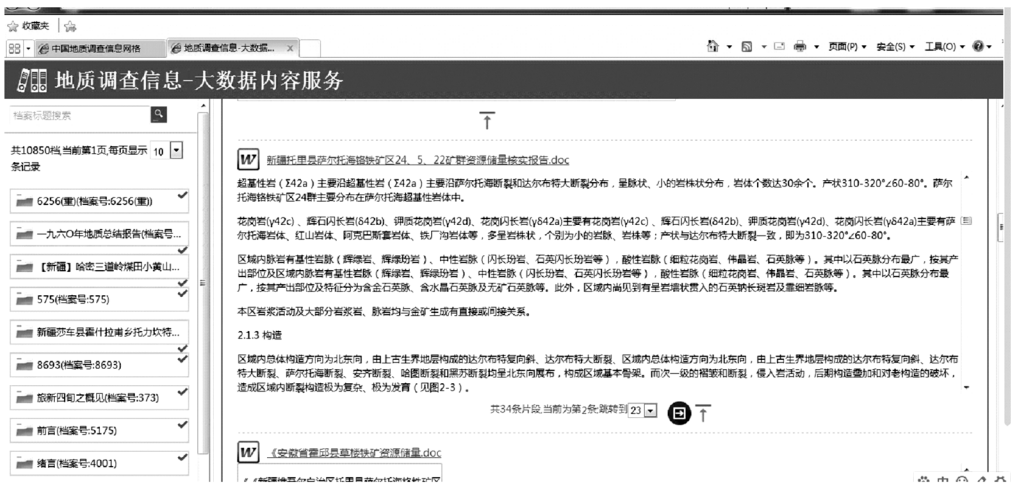

图11　基于基本内容库对区调报告相关内容的发现与推荐服务

Fig. 11　Content discovery and recommendation service for regional survey report based on basic contents library

式。在推进中国地质领域本体建设的基础上,可使多样化碎片化复杂的地质调查非结构化数据形成新的易于使用的数据源。同时将进一步推进地质数据挖掘技术的进步和适用化。

致谢:在多样化碎片化复杂的地质调查非结构化数据存储、数据发现与挖掘模式试验中,中国地质调查信息网格结点西藏地调院徐开锋研究员、新疆国土资源信息中心魏建新教授给予了大力支持,提供了很好的数据试验环境,在数据挖掘研究中给予了地质模型方面的指导,在此一并表示感谢。

参考文献

[1]涂新莉, 刘波, 林伟伟. 大数据研究综述[J]. 计算机应用研究, 2014, 31(6): 1612−1616.

[2]刘智慧, 张泉灵. 大数据技术研究综述[J]. 浙江大学学报(工学), 2104, 48(6): 957−972.

[3]陈吉荣, 乐嘉锦. 基于Hadoop生态系统的大数据解决方案综述[J]. 计算机工程与科学, 2013, 35(10): 25−35.

[4]李超岭, 杨东来, 李丰丹, 等. 中国数字地质调查系统的基本架构及其核心技术的实现[J]. 地质通报, 2008, 27(7): 923−944.

[5]李超岭, 吕霞, 李建强, 等. 中国地质调查信息网格——技术与方法[M]. 北京: 地质出版社, 2013: 99−124.

[6]吴信才, 徐世武, 万波, 等. 新一代的软件结构T−C−V结构[J]. 地球科学(中国地质大学学报), 2014, 39(2): 221−226.

[7]张智, 龚宇. 分布式存储系统HBase关键技术研究[J]. 现代计算机, 2014, 11: 33−37.

[8]王风, 杨璐璐. 基于Hadoop软件框架下海量数据集群处理的探究[J]. 长沙通信职业技术学院学报, 2013, 12(4): 58−62.

[9]陈吉荣,乐嘉锦. 基于hadoop生态系统的大数据解决方案[J]. 计算机工程与科学, 2013, 35(10): 25−35

[10]任树怀. LUCENE搜索算法剖析及优化研究[J]. 图书馆杂志, 2014, 33(12): 17−23.