地质大数据存储技术

李 婧[1], 陈建平[1], 王 翔[2]
LI Jing[1], CHEN Jianping[1], WANG Xiang[2]

1.中国地质大学(北京),北京 100083;
2.中国地质调查局发展研究中心,北京 100037
1. *China University of Geosciences, Beijing* 100083, *China;*
2. *Development and Research Center, China Geological Survey, Beijing* 100037, *China*

摘要:在大数据时代背景下,地质大数据的研究及大数据相关技术为实现地质工作的现代化发展和信息化提供了有效的支撑。而当代大数据发展趋势,就是海量数据的存储及越来越多的事物的数据存在形式。通过梳理大数据处理的关键技术,总结归纳了大数据背景下现有存储技术及数据库的类型。在地质大数据和地质云架构的基础上,讨论适合当前地质大数据的存储技术,地质数据具有多源、多元、异构、时空性、方向性、相关性、随机性、模糊性、非线性等特征。因此,对于存储与管理方式的选择应该是具体问题具体分析,建立多技术支持下的大数据架构,才能满足地质大数据的应用需求。

关键词:大数据; 地质大数据与地质云; 数据库类型; 大数据处理技术
中图分类号:P628^{+}.4 **文献标志码:**A **文章编号:**1671-2552(2015)08-1589-06

Li J, Chen J P, Wang X. A study of the storage technology of geological big data. *Geological Bulletin of China*, 2015, 34(8): 1589-1594

Abstract: Under the background of the age of big data, the study of geological data and the technology of big data provide effective support for the development and realization of the modernization of the geological work. The deep-seated reason of the contemporary trend of big data lies in mass data storage and data forms of more and more things. In this paper, through analyzing the key technology of data processing, the authors sum up the existing storage technologies under the background of big data. Based on geological data and geological cloud architecture, the authors discuss the geological data storage technology suitable for current geological work. Geological data are characterized by multiple sources, diversity, heterogeneity, space-time nature, directivity, correlation and randomness, fuzziness, and nonlinear nature. Therefore, the choice of the means of storage and management should be a specific issue deserving concrete analysis, and only the establishment of large data structure under multiple technologies can meet the demand of geological applications of big data.

Key words: big data; geological data; type of database; technology of big data

继云计算和物联网之后,大数据技术毫无疑问是再次掀起信息技术重大变革的前沿技术之一。就像维克托·迈尔-舍恩伯格在《大数据时代》开篇所说的那样"一场生活、工作与思维的大变革。大数据开启了一次重大的时代转型。大数据正在改变人们的生活及理解世界的方式,成为新发明和新服务的源泉"[1]。

大数据的影响力不仅限于商业范畴,还将深远地改变政府的运作方式、政治的性质,事实证明大数据也应用于农业、医疗、交通等各个领域。当代地球科学面临了很多重大问题,能源和矿产面临着严峻的合理利用及可持续发展问题,地质环境引起的自然灾害问题、地下水污染问题、生态系统平衡问题等。这些影响着国家且关联到世界的

收稿日期:2014-09-09;**修订日期:**2015-02-15
资助项目:国土资源部公益性行业科研专项项目(编号:201511079-02)
作者简介:李婧(1984-),女,在读博士生,从事地球探测与信息技术研究。E-mail: jessicalee1223@gmail.com

问题是人类活动的结果。对基本的地球系统产生了改变与破坏,并且打破了人类与地球的可持续发展平衡。与此同时,信息与网络技术的迅猛发展,不仅改变了人们的日常生活,也改变了科学家的思维模式。随着科学研究方式第四范式的诞生,即数据密集型的知识发现,地质学这种具有多源、多元、异构等复合型数据也被列入大数据的范畴,即地质大数据。

在大数据时代背景下,地质大数据及大数据相关技术为实现地质工作的现代化发展和信息化提供了有效的支撑。当代大数据发展趋势,是海量数据的存储及越来越多的事物的数据存在形式。

1 大数据的关键技术

当今社会经历了web2.0、物联网、云计算,已经迈进大数据时代。信息科技的进步,信息基础设施的持续完善,包括网络带宽的持续增加、存储设备性价比不断提升,为大数据的存储和传播准备了物质基础;云计算为大数据的集中管理和分布式访问提供了必要的场所和分享的渠道;物联网与移动中端持续不断地产生大量数据,数据类型丰富、内容鲜活,是大数据的重要来源[2]。

研究大数据的关键技术,最重要的是对大数据进行分析,通过分析才能获取更多智能、深入、有价值的信息。大数据的属性,包括数量、速度、多样性等都呈现了大数据不断增长的复杂性,所以大数据的分析方法在大数据领域显得尤为重要。

研究地质大数据,就是充分利用地质、矿产、地球物理、地球化学、遥感、地形、地貌、植被、建筑、水文、灾害等地表每一点上的数字化地质数据(结构化和非结构化数据),以大数据背景下的地质数据应用与服务为主线,以大数据技术和产业发展为指导,以挖掘地质信息资源为目的,在数据采集、资源整合、数据传输、信息提取、知识挖掘等相关技术研发与集成的基础上,建设地质大数据环境和地质大数据平台。

大数据的关键技术划分如表1所示。

2 地质大数据与地质云架构

长期以来,国土资源管理的专家们都致力于找到一个完整的解决方案,可以将数据转化为有用信息,从有用信息中集成为一个知识体系,并将其合理利用,完成产品发布与信息共享等服务,而应用和服务中又会产生新的数据,再从中提取有用的信息,从而构成一个循环,形成一条“数据链”。大数据正好提供了这样一种理念和技术方法。督促人们研究并创新各种数据分析方法,将各种数据类型标准化,建立以网络、计算资源为基础的实时信息,以及数据分析的信息管理和综合决策支撑平台。信息网络领域的发展使大数据存储技术取得突破,云计算、物联网、工业互联网等技术的兴起,使信息技术渗透方式、处理方法和应用模式发生变革,也使地质研究中多系统联合与结合成为可能,实现由“数字地质”向“智慧地质”的转变,对影响人类发展和社会文明进步的整个地球资源完成更合理的规划。

地球作为一个整体,无论是人类的行为、气候的变化、各种资源的开发与利用,还是自然灾害、环境污染及生态系统的循环,都是国土资源管理涵盖的内容。大数据的引入,可以将这些资源信息整合起来,提供统一调整整个地球信息资源问

表 1 大数据处理的关键技术

Table 1 The key technology of big data

大数据处理	关键技术
采集	数据采集:传感器技术等 数据存储:关系数据库、NoSQL、SQL等、分布式数据库、分布式缓存等
导入/预处理	数据管理技术:流数据管理、时空索引技术等 基础架构:云存储、分布式文件存储等
统计分析/挖掘	数据处理:NLP(自然语言处理) 统计分析:假设检验、显著性检验、差异分析、相关分析、T检验、方差分析、偏相关分析、距离分析、回归等 数据挖掘:分类(Classification)、估计(Estimation)、预测(Prediction)、相关性分组或关联规则(Affinity grouping or association rules)、聚类(Clustering)、描述和可视化、Description and Visualization)、复杂数据类型挖掘(Text、Web、图形图像、视频、音频等) 模型预测:预测模型、机器学习、建模仿真
结果呈现	可视化分析、标签云、关系图等

题的功能,对国土资源管理战略规划具有不容小觑的作用。

地质大数据与当前基于互联网、物联网的大数据分析不同。赵鹏大[3]指出,不同领域、不同问题的数据性质和类型各不相同,要收集好、处理好、分析好、解释好不同领域的数据,必须深刻了解数据的特性。地质学属于数据密集型科学,地质数据具有多源、多元、异构、时空性、方向性、相关性、随机性、模糊性、非线性等特征。 同时,地质数据也具有深地、深空、深海和深时的特点,空间和时间跨度大,数据获取难度大、成本高、局限性强,地质数据具有混合性和多总体性,地质体的变化性、观测的抽样性和事件结果的不确定性。地质研究的目标要求是定性、定量、定位、定向、定级、定度、定类、定型、定因、定果、定优劣、定概率,因此,从大数据中获取的有助于上述各"定"的关键数据或核心数据,具有专业性和一定的保密性,针对的是长期积累的大量的地质数据。地质灾害、地质环境等问题还有大量的实时数据。地质空间大数据的整合就是对数据的一致性处理,包括空间基准不一致、语义不一致和尺度不一致的地质空间大数据一致性处理,以及地质数据的一体化存储、管理。

3　地质大数据存储与管理的关键技术

大数据平台处理流程如图 1 所示。基本分为数据处理、预处理、存储、分析挖掘和结果展现 5 个步骤。

在大数据环境下实现地质数据的采集、存储、管理、共享、应用,其核心一是明确地质数据存储类型,以便选择相应的数据库存储模式。二是解决海量地质数据的分布式存储与并行计算。

3.1　地质大数据存储数据库

大数据时代需要云计算作为重要支撑,在云上的大数据拥有跨越大量节点、集群和层的众多潜在功能服务层。大数据平台融合需要一个虚拟化的架构,全面的云数据虚拟化基础设施应使用一种集成方法,以确保大数据的统一访问、建模、部署、优化和管理成为一种异构资源。

大数据概念中一个很基础但非常重要的问题就是如何在数以千计的服务器组成的集群中存储 PB 级及以上的海量数据。数据库的种类多种多样,不同的数据模型可以满足不同的应用需求(表 2)。明确这些数据库的应用场景,有助于根据应用需求选择数据存储方式。

目前现有的地质数据量非常庞大,其中,地质数据库总计 12 大类 150 个数据库集,数据量约 160TB;地质图件有 83978 幅,以全国性小比例尺图件为主;按全国地质资料馆馆藏成果地质资料统计,资料文献馆藏数量 120431 档。对于这些非结构化数据,文件系统是主流的存储选择,但是在存取、索引及元数据管理上不是最优。而结构化数据主要依靠关系型数据库,主要问题是结构变化时太复杂,当数据在 TB 级时处理速度也缓慢。NoSQL 数据库应时而生,一是能支持灵活的结构和非结构化数据,二是针对大数据体量可扩展性更好。同时文件系统也得到了发展,与对象存储相映生辉,提升效率的同时,也能更好地支持管理

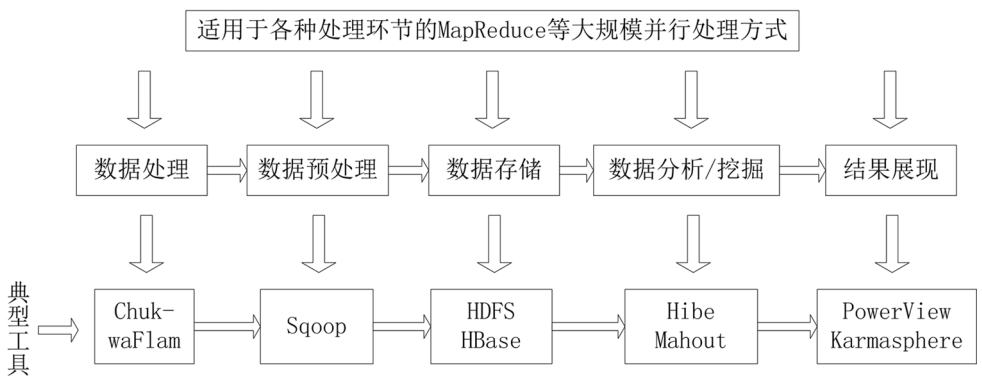

图 1　大数据平台处理流程(MapReduce 并行处理)

Fig. 1　The processing flow of the big data platform (MapReduce parallel processing)

表 2 数据模型分类[6]

Table 2 The Data models

数据库	数据模型	示例	优点
文档数据库	包含了key-value的文档集合	CouchDB, MongoDB	数据模型自然,编程友好,快速开发,web友好,CRUD
图数据库	节点和关系,也可处理键值对	AllegroGraph, InfoGrid, Neo4j	解决复杂的图问题
关系数据库	各种关系	VoltDB, Clustrix, MySQL	高性能、可扩展的OLTP,支持SQL,物化视图,支持事务,编程友好
对象数据库	对象	Objectivity, Gemstone	复杂对象模型,快速键值访问,键功能访问,以及图数据库的优点
Key-Value数据库	键值对	HBase, Hypertable, Cassandra	处理大量数据,应对极高写负载,高可用,支持跨数据中心, MapReduce
BigTable类型数据库	列簇,每一行在理论上都是不同的	HBase, Hypertable, Cassandra	处理大量数据,应对极高写负载,高可用,支持跨数据中心, MapReduce
数据结构服务	字典操作,lists, sets和字符串值	Redis	不同于以前的任何数据库
网格数据库	基于空间的架构	GigaSpaces, Coherence	适于事务处理的高性能和高扩展性

与分析[4]。因此需要了解关系型数据库和非关系型数据库的特点及应用场景,以便更好地选择地质大数据的存储方式。

关系型数据库是指采用关系模型来组织数据的数据库,关系模型由IBM研究员Codd在1970年首次提出[5]。简单来说,关系模型就是二维表格模型,而关系型数据库就是由二维表及其之间的联系组成的一个数据组织[6]。关系型数据库的特点是:数据存储以表格式存在,数据表可以彼此关联协作存储,方便提取数据,但在处理数据密集型应用方面显得力不从心,主要表现在灵活性差、扩展性差、性能差等方面;数据形式通常对应结构化数据,使用结构化查询语言(SQL)操作数据,但已经存入数据的表结构难以更改;存储规范性较高,把数据分隔成最小的逻辑表以避免重复,获得最精简的利用空间,使数据管理更清晰,但单个操作可能涉及多个关系表,具有复杂性;面对日益增长的数据量,支持纵向扩展,即提高处理能力,但成本较高;SQL数据库使用预定义优化方式(如列索引定义)帮助加速查询操作。

NoSQL一词最早出现于1998年,是Carlo Strozzi开发的一个轻量、开源、不提供SQL功能的关系型数据库[6]。大多数大数据环境下的云存储系统和NoSQL系统一般采纳BASE原则。①基本可用(Basically Available):在绝大多数时间内系统处于可用状态,允许偶尔的失败。②软状态或者柔性状态(Soft State):是指数据状态不要求在任意时刻都完全保持同步。③最终一致性(Eventual Consistency):软状态不要求数据时刻保持一致同步,但最终一致性要求在给定时间窗口内数据会达到一致状态。

NoSQL非关系型数据库特点:①通常存储在数据集中,例如文档、键值对或图结构;②基于动态结构,通常适用于非结构化数据,可以很容易地适应数据类型和结构的变化;③数据存储成一个整体,更易读写整块数据;④支持横向扩展,存储呈分布式,使用分布式节点集(集群)来提供高度弹性的扩展功能,让用户可以添加节点来动态处理负载;⑤以块为单元操作数据,使用非结构化查询语言(UnQL),采用简单精确的数据访问模式;⑥更适合大数据应用程序,无模式数据模型更适合于现在捕捉和处理数据种类和类型[7]。

非关系型数据库包括键值数据库、列式数据库、图数据库和文档数据库。

通过列举关系型数据库和非关系型数据库各自的优缺点,认为选择正确的架构取决于所构建应用的需求。传统SQL数据库依然非常强大,大数据是其可以支持的另一个领域。当SQL应用接近局

限性边缘时,NoSQL是可行的选择;涉及可扩展性、灵活性时,它是大数据的最佳选择。

中国工程院院士、中国互联网协会理事长邬贺铨[8]指出,大数据的主要挑战是实时性数据变化快。对于静态数据,可以将数据带进程序来处理,但对于动态数据,需要带程序进数据。大数据更大的挑战是品种多,特别是非结构化数据。对于结构化数据可以使用关系数据库技术来处理,对于非结构化数据则要用NoSQL来处理。针对结构化数据的虚拟存储平台采用动态分层技术,根据数据被调用的频率,自动将常用的数据搬到最高层。针对非结构化数据使用内容归档平台,把结构化和非结构化数据集成到一个单一的动态归档架构中,设计一套软件和元数据库规则,通过给数据加标签的方式,建立不同维度,从而实现模糊查询功能[9]。

目前国家地质资料主要包括文档、图件、数据库(图件、空间、属性数据库等)、图片、表格、视频、音频等结构化、半结构化、非结构化数据。分别保存在全国地质资料馆、国土资源实物地质资料中心、中国地质图书馆、各省(市、自治区)地质资料馆、各地质专业中心、中国地质科学院系统各地质资料室、中国地质调查局局属六大地质调查中心地质资料馆及各行业地质资料馆馆藏机构中。

这些地质数据主要针对地质科研机构、相关国有企业、政府部门,直接为其提供地质资料服务。在实现原始地质资料汇交条例后,地质资料种类、数量成倍增长。地质数据管理分散、条块分割严重,没有形成一个系统、完整的国家地质资料总体,存在"数据孤岛"现象。因此,明确地质大数据的应用需求,应用大数据的存储管理技术,建设国家地质资料大数据环境,是大数据时代带来的变革。

3.2　地质大数据分布式文件系统技术

由于地质资料数据量大,且其中的非结构化数据增长较快,这就需要构建一套地质大数据的分布式文件存储系统,完成海量地质资料稳定、高效的存放与读取。存储技术可以采用master/slave架构,将集群内的节点划分成2类,一类节点存放文件的元数据信息,维护文件系统的元数据信息,但不存储文件内容,负责管理文件系统和客户端对文件的访问;另一类节点放置文件信息。分布式文件系统可增加机器节点并具有可伸缩性,可以横向扩展,通过自动多数据备份存储提高数据的容错能力与可靠性[10]。

大数据技术是一整套技术体系,没有一种体系架构能完美解决所有大数据问题,需要根据实际的应用进行合体裁剪和扩展。因此,在构建地质大数据应用环境时应采用扩容性及兼容性好的方案,一个开放式体系结构的混合平台应该是较优的选择[11]。

4　讨　论

地质大数据的存储与管理技术研究,首先要明确地质大数据的应用需求,根据不同的需求来选择适合的数据模型和数据存储方式。例如,需要不同的访问方式和数据类型,可以选择文档数据库;需要大数据量的离线分析可以选择Hadoop;如果需要跨越多个数据中心,对于规模不断增长(真正的大数据场景),但是访问不频繁的数据,可以使用Bigatable类型的数据库,或其分布式的,能解决延迟问题、分区容错性问题的产品,因为它的数据存储在一个分布式文件系统上,很容易扩展;需要支持二级索引,想通过不同的键来检索,可以选用关系数据库和Cassandra,后者新增额对二级索引进行支持。SQL数据库可以通过ACID属性(原子性、一致性、隔离性、持久性)保证数据的完整性。而NoSQL数据库是基于节点的分布式系统,可以在一致性、可用性、分区容忍度中任意选择两项[12]。

地质大数据种类繁多、结构复杂,目前已有魏振华等人采用分层、分区、分类管理策略和要素扩展管理技术,创建对象–关系型数据库存储模型,来适应复杂数据的复杂查询,以及实现海量地质空间数据的对象——关系型一体化存储[13]。

对于地质资料大数据,为实现其更好的应用与服务功能,首先应构建一个互联互通的、连续的、动态的、完整的、反映国家版图内地质调查工作程度的地质资料逻辑总体。实现多节点的国家地质资料大数据环境,存储模式包括支持SQL语句查询的并行数据库与NoSQL等,地质数据的多元化需要根据不同的应用需求来选择多技术架构下的大数据存储[11]。单纯的数据并没有价值,必须有一套理念和机制对数据进行处理、对接,然后得出可以描述全局的"数据"[14-16]。

参考文献

[1](英)维克托·迈尔-舍恩伯格,(英)肯尼思·库克耶.大数据时代[M].杭州:浙江人民出版社,2013.

[2]赵国栋,易欢欢,糜万军,等.大数据时代的历史机遇[M]. 北京清华大学出版社,2013.

[3]赵鹏大.大数据时代的数字地质[C]//中国数学地质大会. 新疆,2013.

[4]吴甘沙. 大数据漫谈之三[EB/OL](2013-05-18)[2014-09-07]http://www.huxiu.com/article/14635/1.html. 2013.

[5]李新安.数据库技术发展前景展望[J].山东电力高等专科学校学报, 2005, 2: 40-43

[6]张俊林. 大数据日知录:架构与算法[M]. 北京:电子工业出版社,2014.

[7]数控小V. 大数据应用程序最佳选择:是SQL还是NoSQL[EB/OL](2014-03-20)[2014-09-07]http://www.36dsj.com/archives/6868. 2014.

[8]邬贺铨. 大数据时代的机遇和挑战[J]. 信息安全与通信保密,2013, 11: 16-19.

[9]邬贺铨. 我们需要国家大数据战略[EB/OL](2013-12-16)[2014-09-07]http://www.cnii.com.cn/informatization/2013-12/16/content_1271495.htm. 2013.

[10]朱月霞,侯建光. 基于大数据的地质数据存储与管理研究[C]//江苏省测绘地理信息学会2014年学术年会论文集,2014.

[11]Wu B,Kshemkalyani A D. Objective optimal algorithms for long-term Web perfecting[J]. Computers, IEEE Transactions on, 2006, 55(1): 2-17.

[12]网络大数据论坛. 数据库各个派系的起源、应用场景和选择指南[EB/OL](2015-02-09)(2015-07-27)http://www.raincent.com/content-85-3842-1.html. 2015.

[13]魏振华,刘志锋,李金萍,等. 基于要素扩展管理的海量地质空间数据存储模型的设计与实现[J]. 计算机应用与软件,2014,31(7):36-39.

[14]Holland B. 全面梳理SQL和NoSQL数据库的技术差别[C]//TechTarget中国,2014.

[15]徐民先. 多技术架构大数据存储应用[J]. 中国公共安全(综合版)2014, 9: 135-137.

[16]周侠. 单纯的数据并没有价值"大数据"未来市场巨大[EB/OL](2013-06-13)[2014-09-07]http://www.bj.xinhuanet.com/hbpd/jrpd/jrpd/2013-06/13/c_116132705.htm. 2013.