DOI:10. 16031/j. cnki. issn. 1003-8035. 2018. 01. 20

基于 Hadoop 分布式系统的地质环境大数据框架探讨

任晓霞,喻孟良,张鸣之,陈一超,韩明伟,曾青石
(中国地质环境监测院,北京 100081)

摘要:在分析目前地质环境数据现状基础上,结合大数据特点,分析了地质环境大数据特征,并结合目前地质环境应用情况,设计了地质环境大数据框架。该框架包括对地质环境大数据的数据进行清洗与转换、分布式数据存储管理、数据挖掘、文本检索、大数据可视化等功能。该框架对于后续实现地质环境大数据分析应用具有指导意义。

关键词:大数据;地质环境数据;Hadoop 生态系统;云计算
中图分类号:P628 **文献标识码**:A **文章编号**:1003-8035(2018)01-0130-05

Geological environment big data framework based on Hadoop

REN Xiaoxia, YU Mengliang, ZHANG Mingzhi, CHEN Yichao, HAN Mingwei, ZENG Qingshi
(*China Institute of Geological Environment Monitoring*, *Beijing* 100081, *China*)

Abstract: The characteristics of geological environment big data were analyzed based on current geological environment data and geological environment big data framework was put forward based on Hadoop ecological system architecture, which supported data cleaning and conversion, distributed data storage management, data mining, text searching, data visualization and other functions for geological environment big data. The framework has guiding significance for the application of geological environment big data in future.

Keywords: big data; geological environment data; Hadoop ecological system; cloud computing

0 引言

地质环境数据包括地质灾害、地下水、矿山地质环境、地质遗迹、水土地质环境等业务的调查与监测数据,可为国家重大战略、资源合理开发、环境保护和生态文明建设等提供有力数据支撑。近年来,国土资源部和中国地质调查局等在国土资源信息化方面的工作不断深入,“数字国土工程”、“金土工程”、“地质调查项目”、“国家地下水监测工程”等的实施,积累了海量地质环境数据资料[1-2]。如已经完成的 1∶10 万县市地质灾害调查数据、全国 1∶20 万区域水文地质调查数据和正在进行的国家地下水监测数据、1∶5万地质灾害详查数据、1∶5万水文地质调查数据以及 1∶5万环境地质调查数据等。这些地质环境数据具有涉及领域多、数据格式多样、数据量大、数据更新快等特点,数据本身主要包含结构化和非结构化数据[3]。

但同时,随着业务管理和新技术的不断发展,原有的地质环境数据管理和应用模式面临新的需求,主要表现在:

(1)多源异构数据的集成管理需求。各类地质环境数据的生产来源不同,相应的数据格式多样。如何满足多源、异构数据的统一高效管理成为亟需解决的问题。

(2)地质环境数据高效率存储管理需求。随着业务的不断发展,传统的单服务器存储已经不能满足快速增长的业务需求。如何将多源、量大、应用复杂的数据进行高效存储管理迫在眉睫。

(3)海量数据的数据挖掘与展示需求。地质环境数据通过不断的积累,产生了海量数据,如何从海量数据中进行挖掘并提取有价值的数据以及进行可视化展

收稿日期:2017-05-08;**修订日期**:2017-06-08
基金项目:中国地质调查局项目(地质大数据支撑平台建设(中国地质环境监测院))(DD20179373)
第一作者:任晓霞(1981-),女,山东安丘人,硕士,高级工程师,主要从事水工环信息化工作。E-mail:renxx@ mail. cigem. gov. cn

示也是亟需解决的问题。

(4)数据快速识别与组装要求。根据用户要求进行用户数据定制,对数据进行组装与分发,满足地质环境多专题数据的个性化定制要求。

(5)地质环境信息服务的新需求。地质环境数据种类繁多、数据产生量飞速增长、应用复杂等特点,给地质环境信息服务提出新挑战。同时,经济社会发展对地质环境信息服务提出了全方位需求[4]。

近年来,随着虚拟化、云计算等信息技术的飞速发展,全球数据量飞速增长,人类已经进入了大数据时代。大数据技术吸引了企业、政府、学术界等高度重视。Google 公司设计开发了 GFS(Google file system)分布式文件系统和 BigTable 非关系数据库[5]。2012 年 3 月美国政府公布的"大数据研究和发展倡议"[6]使"大数据研发计划"成为国家层面的指导文件。2005 年,Apache 受 Google GFS 的启发,提出了 Hadoop 大数据框架,并在各个行业得到广泛应用[4]。马友忠等[7]提出了云数据管理索引技术,李超岭等[8]提出了智能地质调查体系。但是,目前所形成的大数据技术主要应用于互联网文本搜索、商品推荐、Hadoop 算法改进、地质调查应用体系等领域,在地质环境领域大数据技术应用相对较少。

随着新技术的不断发展,如何借鉴大数据技术,加快地质环境数据的集成管理与信息挖掘,以满足社会各界对地质环境信息日益增长的需求是接受大数据时代对地质环境信息服务提出的机遇与挑战。为此,本文基于大数据技术,分析了地质环境大数据特点,讨论了地质环境数据集成大数据框架和关键问题,为今后大数据技术在地质环境领域应用提供参考。

1　地质环境大数据特征分析

大数据(Big Data)是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产,具有 4V 特点[9-10]。

(1)数据量大(Volume)。第一个特征是数据量大,包括采集、存储和计算的量都非常大。

(2)类型繁多(Variety)。第二个特征是种类和来源多样化。大数据包括结构化、半结构化和非结构化数据,具体表现为网络日志、音频、视频、图片、地理位置信息等等。

(3)价值密度低(Value)。第三个特征是数据价值密度相对较低,或者说是浪里淘沙却又弥足珍贵。随着互联网以及物联网的广泛应用,信息感知无处不在,信息海量,但价值密度较低。如何结合业务逻辑并通过强大的机器算法来挖掘数据价值,是大数据时代最需要解决的问题。

(4)速度快时效高(Velocity)。第四个特征数据增长速度快,处理速度也快,时效性要求高。比如搜索引擎要求几分钟前的新闻能够被用户查询到,个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。

地质环境数据主要产生于基础地质、水文地质、环境地质、工程地质、地质灾害等相应的调查、监测以及科研过程中,包括地质灾害、地下水、矿山地质环境、地质遗迹、水土地质环境等业务的各类数据资源。地质环境主要数据情况见表 1。

由表 1 可知,地质环境数据类型各异、数据格式不同,总体可分为结构化、半结构化和非结构化。随着地质环境信息化的不断深入,地质环境数据的数据量飞速增长,数据量大,数据种类繁多,除了传统的 MapGIS 和 ArcGIS 矢量数据、关系型数据库、空间数据库、地质报告、图件、表格外,也出现了大量图片和视频等格式的数据。数据本身价值大但提取难度大,监测数据时效高更新速度快。地质环境数据这些特点符合大数据的 4V 特点,是时空大数据[9]。

地质环境大数据的上述特征,决定了其存储管理、挖掘处理和服务应用方式的特殊性和挑战性。为充分发挥地质环境大数据的作用,解决当前地质环境面临的困难和问题,应充分利用大数据技术,采用大数据的存储管理体系架构对数据进行存储管理和分析处理,分析地质环境哪些方面的应用或者算法适合改造为大数据环境下的算法,以提高系统的效率和可靠性,进一步提升用户体验度和满意度。

2　地质环境大数据总体框架

2.1　设计目标

针对地质环境大数据特点和面临的需求挑战,地质环境大数据框架应具能达到以下目标。

(1)支持多源异构数据的数据集成处理

地质环境数据来源于不同生产源,大部分数据通过调查与监测获取。其中通过地质调查手段获得的数据主要包括地质灾害县市调查数据、地质灾害详查数据、水文地质调查数据、矿山地质环境摸底调查数据、环境地质调查数据(地质遗迹、矿山地质环境、地面沉降等数据)等,其格式一般为 MapGIS 和 ArcGIS 数据格式。通过监测获得的数据主要包括地下水数据、地

表 1　地质环境主要数据情况列表

Table 1　Lists of geological environmental main data

数据集名称	数据类型	数据格式	主要内容	更新频率
1:10 万县市地质灾害调查数据	属性数据、图件、报告	MS Access、MapGIS、MS Word、图片等格式	崩塌、滑坡等调查表以及地质灾害防治区划图、报告等	调查频率决定
1:5 万县市地质灾害详查数据	属性数据、图件、报告	MS Access、MapGIS、MS Word、图片等格式	崩塌、滑坡等调查表以及地质灾害防治区划图、报告等	调查频率决定
国家级地下水动态调查数据	属性数据	MS Access	水位、水温、水质	动态
地下水动态监测数据	属性数据	SQL Server、MS Access	水位、水温、水质	动态
全国矿山地质环境摸底调查数据	属性数据	SQL Server	矿山基本信息表	调查频率决定
地质遗迹调查数据	属性数据、图件、报告	MS Access、MapGIS、MS word 、图片等	地质遗迹点等调查表以及图件报告等	调查频率决定
1:5 万水文地质调查数据	矢量数据	MapGIS	地下水类型等空间专题数据	调查频率决定
1:5 万环境地质调查数据	矢量数据	ArcGIS、MS Access、图片	矿山地质环境、地面沉降等调查数据	调查频率决定
1:20 万区域水文地质空间数据库	矢量数据	MapGIS	地下水类型等空间专题数据以及基础地理数据	调查频率决定
1:50 万分省环境地质空间数据库	矢量数据	ArcGIS	环境地质分区等空间专题数据以及基础地理数据	调查频率决定
30 m DEM 数据	影像数据	TIF	图像	一次性
基础地质	矢量数据	MapGIS、ArcGIS	岩石、地层等	一次性
基础地理	矢量数据	MapGIS、ArcGIS	行政区、水、路等	一次性
其他报告类	报告、图片	MS Word 、PDF、JPG、Excel 等	生产的报告文档类等数据	一次性

质灾害示范区监测数据、地面沉降监测数据等。多源、异构、动态高速增长的数据采集与处理,需利用目前主流大数据处理技术,完成对数据的采集与处理。对于传统数据库结构数据,可利用 Sqoop 开源工具完成从传统数据库中的地质环境数据到 Hadoop(如 HBase、Hive 和 HDFS)的数据传递。Sqoop 其架构参见图 1。对于非结构化或者半结构化数据,可利用 Avro 开源工具完成到 Hadoop 的存储和交换。Avro 是一个基于二进制数据传输高性能的中间件,是数据序列化的系统,适合远程或者本地大规模数据传输。

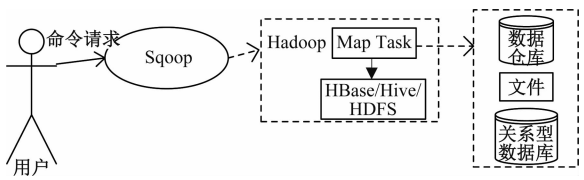

图 1　Sqoop 架构示意图

Fig. 1　Schematic diagram of sqoop architecture

(2)可扩展的动态存储管理方式

大数据体量大、速度快、种类多等特征带来了存储管理上的质变。相对于静态、有限的数据集,地质环境大数据存储管理系统需要具有可扩展性,以处理动态增长数据的存储、更新和查询等问题。云计算技术通过互联网按需进行动态部署(provision)、配置(configuration)、重新配置(reconfigure)以及取消服务(deprovision),能够提供动态资源池、虚拟化和高可用性的下一代计算平台[11]。Hadoop 是大数据处理中常用的软件框架,它实现了 MapReduce 编程模型,能够对大量数据进行分布式处理,将应用程序分割成许多小的工作单元,并把这些单元放到任何集群节点上执行,是具有高可靠性和良好扩展性的分布式系统[12]。HDFS(Hadoop Distributed File System)分布式文件系统、Hadoop MapReduce 分布式计算模型和 HBase 分布式数据库是 Hadoop 的三大核心技术。

为此,可充分利用云计算技术和 Hadoop 软件框架,搭建云计算 Hadoop 平台,为地质环境行业提供“私有云”或者“公有云”服务。

(3)海量数据的数据挖掘与分析

大数据必然要依靠数据挖掘,从地质环境数据库或者数据仓库中挖掘出隐含的、非显见的知识和规律,以满足地质灾害防治、环境保护等领域的应用。数据挖掘可采用 Hive 工具完成数据分析。Hive 是基于 Hadoop 的一个数据仓库工具,可以将结构化的数据文件映射为一张数据库表,存储为 Hadoop 兼容的文件系统(如 Amazon S3,HDFS),并提供简单的 SQL 查询功能,可以将 SQL 语句转换为 MapReduce 任务运行。

对于空间数据,利用 ArcGIS 的 GIS Tools for Hadoop 完成对地质环境空间大数据的 Hadoop 的 GIS 应用等。ArcGIS 提供了一套关于 JAVA 的 Geometry API,通过这些 API,可对存储在 Hadoop 的 HDFS 中数据进行处理。

(4)传统算法的并行处理算法改造

充分利用 Hadoop 的高效性、高可靠性和高容错性等优点,研究哪些传统算法适合在 Hadoop 系统上运行,将其改造为基于 Hadoop 平台的并行处理算法。比如,影像数据金字塔处理可改进为并行处理算法,查询某地区地质灾害受灾情况排名或者地下水水位下降情况排名等均可使用并行处理算法。

(5)高效的地质环境信息服务

传统系统的服务方式可作为大数据平台下的服务方式,如数据服务、地图服务等。对于数据快速识别与组装需要考虑,以满足数据定制等要求。同时,需考虑数据的全文检索,如使用 Lucene 或者 Nutch 等开源搜索引擎包,建立数据的全文检索,以从海量数据中快速定位到数据。通过数据挖掘分析得到的结果,可利用大数据的可视化工具 Hue 等工具来完成展示。或者将挖掘分析结果与传统方式相结合,如 ArcGIS 或者 MapGIS 或者已有地图平台等进行地图展示、地图浏览查询、三维可视化等。

2.2　总体框架

根据设计目标,地质环境大数据框架主要采用目前主流大数据技术,包括数据清洗工具、数据序列化工具、分布式数据库、分布式数据仓库、大数据文本搜索框架等技术。利用上述多种技术,对地质环境数据、资料文档、图件等数据进行存储、组织,经过一系列的数据清洗、数据集成、数据存储、分析和挖掘,实现对海量数据的大数据采集、存储、分析、管理与服务应用。同时,对于传统地质环境信息服务方式进行保留,进行并服务。其框架设计见图 2。

(1)基础设施层

基于云计算平台的基础设施层可以更方便快捷地管理计算、存储等资源。基础设施层主要用于存放和处理地质环境大数据的物理设施,包括主机、存储、网络设备等。在云计算环境下,使用云一体机提供基础设施资源池,并且基础设施池的计算资源和存储资源可以动态伸缩地提供给地质环境内部业务人员和科研人员使用,以实现资源的整合,大大提高资源利用率。

(2)源数据层

主要包含地质环境的结构化与非结构化数据,如地质灾害、地下水、矿山地质环境等业务的调查与监测数据以及报告、文档、图件等数据。

(3)大数据资源层

海量地质环境数据从原始采集阶段到形成地质环境大数据,需要进行清洗、集成等综合处理,包括对传统数据库的清洗与装载、非结构数据的序列化与装载传输等。清洗后的数据在进行海量数据存储时,将地质环境数据存放到分布式文件系统 HDFS/HBase/Hive/Druid 中。大数据资源层主要负责对数据文件进行并行提取、加载、转换以及存储。

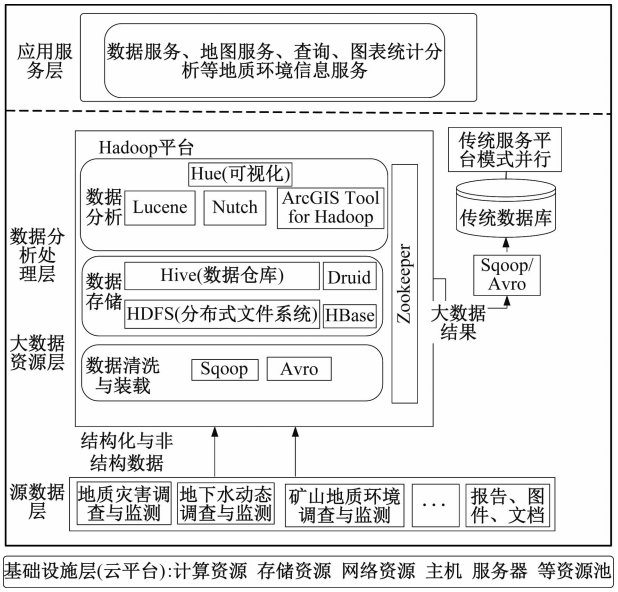

图 2　基于 Hadoop 的地质环境大数据框架示意图

Fig. 2　Schematic diagram of geological environment big data framework based on Hadoop

(4)数据分析处理层

数据分析处理层对地质环境数据建立列索引,进行数据识别、全文检索、GIS 空间分析等操作,然后再利用可视化工具将分析挖掘的结果进行显示。数据识别可借助 HBase 列索引也可借助全文检索引擎 Lucene 或者 Nutch。GIS 空间分析可使用 ArcGIS Tool for Hadoop 工具包进行分析处理。数据识别、全文检索等技术能从海量数据中提取蕴含的地质环境信息知识,可视化工具如 Hue 将结果进行展示分析。

(5)应用服务层

通过大数据分析技术获取的结构可进行可视化、地图服务、统计分析结果展示、数据定制等服务。对于获取的大数据结果也可结合传统地质环境信息服务模式,利用 Sqoop、Avro 等工具将大数据分析结果导出与传统服务平台进行对接,从而丰富服务方式。

3　应用场景展现

结合地质环境实际数据情况(表 1),考虑目前地质环境各方面的需求和实际应用情况,利用地质环境

大数据框架,可有如下几个方面应用。

3.1 查询统计分析类应用场景

场景 1 描述:

对某地下水水位监测点一定时间段内的水位数据情况查询统计,也可对某地区内一定时间段内的水位数据情况查询统计,从而获取特定地区特定时间段内水位变化特征。

大数据处理流程:

地下水动态调查和动态监测数据,目前格式均为结构化数据。可利用 Sqoop 工具对原始数据进行清洗装载到 Hadoop 的 Hive 中或者 HBase 中,然后进行数据查询统计。

其他应用场景跟场景 1 类似的,均可采用以上处理过程,例如地质灾害县市调查属性数据(MS Access 结构化数据)涉及到的查询统计分析类。

3.2 基于地理位置查询数据类应用场景

场景 2 描述:

地质灾害数据可根据位置信息或者地理位置名称或者区域查询发生的地质灾害数据。

大数据处理流程:

已有地质灾害县市调查数据格式均为结构化数据。可利用 Sqoop 工具对原始数据清洗装载到 Hadoop 的 Hive 中或者 HBase 中,基于地理位置名称或者经纬度进行数据查询或者数据定位。

若利用已有报告类数据,如地质灾害通报数据,则需要利用 Avro 工具对原始数据进行序列化后提取有效信息存入 HBase 中,利用 Lucene 建立全文索引,基于文本进行数据查询。

3.3 并行处理类应用场景

场景 3 描述:

影像数据的金字塔并行处理。

大数据处理流程:

影像数据格式为非结构化数据。可利用 Avro 工具对原始小的数据进行合并序列化后,存储到 HDFS 中,再利用 MapReduce 改进金字塔处理算法以便进行并行处理。

场景 4 描述:

大数据分析验证采集数据正确性或者各项数据指标关联性或者预测数据分布趋势。

4 结语

地质环境数据资料为国家的宝贵财富,符合地质环境大数据特点。本文以地质环境数据为例,提出了地质环境大数据设计目标,并基于当前主流大数据技术,设计了地质环境大数据框架,该框架为后续地质环境大数据平台建设实施提供了技术参考和思路。今后,通过建立地质环境大数据平台,地质环境大数据应用主要表现以下几个方面。

(1)地图瓦片等数据并行处理。地图瓦片、遥感影像等数据预处理可改造为大数据并行处理算法,以充分减少处理时间,提高数据发布效率。

(2)多区域、多维度数据综合分析。数据识别使用列索引,将数据作多种数据标签,从而加快数据跨区域、跨时间等分析,如进行多维数据的关联分析等。

(3)全文检索的高效数据定位。利用全文检索的大数据工具,提高数据查询效率,做到数据快速定位。

参考文献:

[1] 刘晓慧, 吴信才, 罗显刚. 面向对象的地质灾害数据模型与时空过程表达[J]. 武汉大学学报(信息科学版), 2013,38(8): 958 – 962.
LIU Xiaohui, WU Xincai, LUO Xiangang. Object-oriented geological disaster data model and spatio-temporal process expression [J]. Geomatics and Information Science of Wuhan University, 2013, 38(8): 958 – 962.

[2] 任晓霞, 曾青石, 喻孟良, 等. 地质环境数据交换与共享思路探讨[J]. 国土资源信息化, 2015, (4):17 – 22.
REN Xiaoxia, ZENG Qingshi, YU Mengliang, et al. Discussion of geological environment data exchange and sharing schema idea[J]. Land and Resources Informatization, 2015, (4): 17 – 22.

[3] 喻孟良, 任晓霞, 曾青石, 等. 地质环境数据集成方法探讨及实例应用[J]. 中国地质灾害与防治学报, 2016, 27(4):103 – 108.
YU Mengliang, REN Xiaoxia, ZENG Qingshi, et al. Discussion and application of data integration method for geological environment data [J]. The Chinese Journal of Geological Hazard and Control, 2016, 27(4):103 – 108.

[4] 郑啸, 李景超, 王翔, 等. 大数据背景下的国家地质信息服务系统建设[J]. 地质通报, 2015, 34(7):1316 – 1322.
ZHENG Xiao, LI Jingchao, WANG Xiang, et al. Construction of the national geological information service system in the age of big data[J]. Geological Bulletin of China, 2015, 34(7):1316 – 1322.

(下转第 142 页)