

大数据理念在海洋环境观测数据共享中的应用研究^{*}

宋坤

(国家海洋技术中心 天津 300112)

摘要:随着海洋观测技术、科学调查手段、信息技术等的发展,海洋观测数据量日益庞大、观测要素的种类日益繁多,而我国海洋观测平台的条块管理模式造成了观测数据的分布式和区域化,使得资料共享成为近海研究的大障碍,影响了对数据价值的挖掘。文章介绍了近年来蓬勃发展的大数据的概念及其核心价值,并分析了当前海洋环境观测数据共享的需求。在此基础上,提出了一种基于大数据理念的海洋环境观测数据共享平台框架及相应的共享技术方案,并重点探讨了分布式的数据存储结构,以及从数据级和应用级两个层次进行信息共享的机制,数据更新和维护策略等。

关键词:系统工程;大数据;数据共享

中图分类号:X145 **文献标志码:**A **文章编号:**1005—9857(2015)06—0043—0003

1 引言

海洋是关系国家领土、海洋权益和军事对抗的战略空间和主题,其具有重要的政治、经济和军事意义。海洋环境是指以人类生存与发展为中心,相对其存在并产生直接或间接影响的海洋自然和非自然全部要素的整体。海洋环境数据来源于人们对海洋环境的了解、观测和掌握,是人类进行海洋开发、海洋环境保护、军事保障和海洋管理的主要决策依据。

海洋环境数据覆盖了水文、气象、化学、生物、地质、地球物理等多个学科,主要包括大气(气温、风、雨、云、雾等)、水文(海水温度、盐度、深度、潮汐、海浪、海流等)以及海底地形、地貌、地质、重力、磁力、海底扩张等各种数据信息^[1]。由于数据采集的分布性、多源性和多样性等,海洋环境观测数据具有类型复杂、结构多样等显著特征。有效地组织管理数据信息,实现观测数据共享,对海洋环境科学研究和应用具有重要的意义和价值。

大数据(Big Data)是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合。而海洋环境观测数据的多样

性、多源性、复杂性与大数据的主要特征数据量巨大(Volume)、数据类型多(Variety)、数据流动快(Velocity)和数据潜在价值(Value)相一致。在海洋环境观测数据共享应用中引入大数据的理念不止是为了更好地组织管理庞大的数据信息,其主要意义还在于更大地挖掘其潜在的应用价值。

2 大数据理念

当气象学家在做气象地图分析、物理学家在建立大物理仿真模型等过程中,由于数据量巨大而不能再用传统的计算技术来实现时,大数据的概念在这些科学研究领域首先被提出来。随着互联网和电子商务的快速发展,大数据的理念和技术开始被实际应用。大数据的新技术和新架构也在面临解决数据量大、数据种类多、数据流动速度快、数据不完备甚至不可理解等问题中得以不断开发出来。

大数据应用真正要实现的是“用数据说话”,而不再是直觉或者经验。其主要包括以下 3 个方面:一是发现过去没有发现的数据潜在价值;二是通过不同数据集的整合创造新的数据价值;三是把在一个领域已经发挥过价值的数据再次

^{*} 基金项目:赤潮监测与预警系统集成与示范(2007AA092004)。

应用在新的领域创造出新的价值^[2]。

3 海洋环境观测数据共享需求

(1)随着国家对海洋科学研究重视程度的提高,国家公益性专项、“863”计划等的支持创建了许多海洋工程项目,对海洋生物、化学、物理等各领域进行了探究,同时也产生了许多的数据资料。但由于我国海洋观测平台的条块管理模式使得观测数据具有分布式、区域化的特征,很多数据只在某一个领域或某一个课题中得以利用,并未能充分挖掘数据的价值。考虑到数据资源的分布式特征需要引入大数据的理念,建立数据共享机制,实现数据的重复利用,挖掘数据的潜在价值。

(2)由于海洋环境观测数据获取的手段不同、表述的方式不同,使得数据的类型多样。按照数据结构分为结构化数据即存储在数据库中、可以用二维表结构来逻辑表达实现的数据;半结构化数据即格式较为规范的,可以通过某种方式解析得到每项的数据;无结构的非结构化数据是指非纯文本类数据,没有标准格式,无法直接解析出来相应的值,如图像、视频等。基于海洋环境观测数据的异构性特征,需要借鉴大数据的数据存储理念对异构数据进行整合、对其存储进行重新规划和部署。

(3)在传感器技术、海洋观测技术、通信技术的飞速发展形势下,海洋环境观测数据的获取频率在不断提高,如何快速地完成数据的更新、分析并提供相应的应用及产品,是实现数据共享需要解决的重要问题。在数据存储和计算中,可以采用大数据的分布式存储+计算平台的模式,以满足观测数据的快速流动性。

4 海洋环境观测数据共享平台框架设计

依据海洋环境观测数据共享的需求,结合近年来蓬勃发展的大数据的概念和技术,本研究提出了一种基于大数据理念的海洋环境观测数据共享平台框架,如图 1 所示。

该框架主要由数据源、数据操作、数据存储、数据访问、数据分析、应用服务以及相关数据共享管理制度和安全机制等组成。各层相对独立,层间的耦合性比较小。

数据源处于框架的最底层主要包括需要共

享的基础数据、各类历史资料、各项目和区域获得的实时数据和延时数据,以及获取和制作的信息产品等,作为整个平台运行的基础。

数据操作是对异构的多源数据进行抽取、排序/过滤、转换等处理,以实现数据的重新整合、部署,完成数据的规范化存储。

数据访问是各数据集与数据分析之间的桥梁,实现数据的提取以及分析信息的加载。

数据分析是根据应用的需要对存储的数据进行检索、统计分析、信息产品制作、数据同化融合等分析处理,为应用服务提供后台并行计算。数据的操作、存储、访问及分析都要以元数据管理为依托。

应用服务是平台最终能提供给用户的功能以及为用户交互的接口,包括数据查询、信息可视化、信息共享及数据评价与反馈等。

数据管理、安全和备份恢复是帮助进行数据治理、实现数据保护。

数据共享管理制度是指与海洋环境观测数据共享相关的各种法律、法规和文件。

整个框架以多源数据为基础,以提供各类应用服务为目标,以数据管理、安全和共享管理制度为运行保障,最终实现海洋环境观测数据的共享。

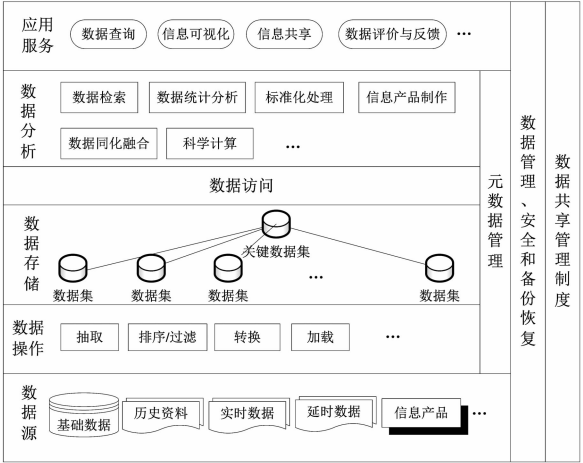


图 1 海洋环境观测数据共享平台框架

4.1 分布式数据存储

为了有效地避免数据重复建设和信息孤岛,以及海洋环境观测数据异构性的特征,需要基于基础数据和元数据对多源异构海洋环境观测数据信息进行整合,并对信息资源的存储进行重新

部署^[3]。根据数据涉及学科的不同、安全等级的不同和应用需求的不同,将海洋观测数据分为多个数据集,并通过关键数据集来维护其间的关系。关键数据集主要包括元数据、基础数据等。结合观测数据应用的情况,从管理和应用效率的角度出发,采用数据集中存储与分散存储相结合的方式,共享频率高公开数据存放在本地数据库中。对于某项目或地区的非公开数据则存储在数据生产者的数据库中,其数据源的生产、维护更新和所有权归数据生产者。该平台只在元数据的描述中提供该类数据的数据标志信息、内容信息及发布信息等。

4.2 信息共享机制

根据共享内容的粒度差异,信息共享需求分为数据级共享和应用级共享^[4]。数据级共享指共享的内容为数据,是粒度较低的共享,主要是纯数据库级的数据提取和发送。在该共享平台中体现为以元数据为核心建立信息目录,通过数据检索提供数据查询、信息共享等服务。应用级共享指从应用层面依据用户需求提供经数据统计分析、同化融合和科学计算等处理分析后的信息产品及资料,虽然展示的是信息产品,但实质提供的是在 Web Services 技术规范下的与逻辑相关的通用服务共享功能,是粒度较高的共享。

4.3 数据更新和维护策略

海洋环境观测数据的流动速度很快,数据的更新和维护关系到共享平台的运行效果。在海洋信息交换领域数据源非常广泛,新数据源的添加也是不可避免的,所以该平台对数据的更新和维护采用动态管理的方式,对数据的加载、检索、分析等,使用数据配置的关联方式和管理方式,以适应海洋数据源的变动,提高平台对数据的支持能力。

5 结束语

所谓信息共享指信息的双向流动,即任何单位或个人履行了把所掌握的全部或部分数据贡献出来作为社会发展之用的义务,就有权利获得其他人提供的数据或信息,只有形成信息的双向流动才能促进海洋事业的发展^[5]。本研究以海洋观测数据共享的需求为依托,引入了大数据的理念,提出了一种海洋环境观测数据共享平台的框架,在理论与技术层面上为海洋观测数据的共享提供了一种模型,以实现数据的共享及其潜在价值的挖掘。该平台的搭建和运行还需要海洋信息管理制度的完善和海洋信息共享相关法律法规的建立,以保障数据生产者的利益和敏感数据的安全。

参考文献

[1] 李琳,郝建新,汤晓安,等. 基于 GML 的海洋空间数据共享研究[J]. 现代电子技术,2009(6):147—153.
[2] 赵刚. 大数据技术与应用实践指南[M]. 北京:电子工业出版社,2013:1—21.
[3] 丁火平,陈建平,余剑平,等. 基于 SOA 架构的数字城市信息共享方法研究[J]. 计算机工程与设计,2009,30(20):4632—4625.
[4] 钱大君,吴建平,余柏菴,等. 上海市资源与环境数据共享平台的实现[J]. 计算机工程,2008,34(10):283—285.
[5] 常虹,于华明,鲍献文,等. 我国海洋数据信息共享现状及立法建议[J]. 海洋开发与管理,2008,25(1):134—138.