大数据背景下海洋数据管理的挑战与对策

黄冬梅 赵丹枫 魏立斐 杜艳玲 王振华
(上海海洋大学信息学院 上海 201306)

摘 要 空天地底海洋立体观测技术的飞速发展催生了呈指数级增长的多精度、多频度、大覆盖、多模态的海洋数据。然而,目前的研究主要集中于通用大数据,针对海洋数据的专门研究仍处于起步阶段。海洋数据的时空性、多源多类性、海量性、敏感性等特性为其管理带来了新的挑战。阐述了海洋数据的特点和海洋数据管理的基本架构,探讨了海洋数据在数据存储、数据质量、数据安全等环节面临的挑战,分析了相应的对策,为海洋科学与工程技术的研究提供了重要的参考与依据。

关键词 海洋,大数据,数据存储,数据质量,数据安全
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.003

Managing Marine Data as Big Data: Uprising Challenges and Tentative Solutions

HUANG Dong-mei ZHAO Dan-feng WEI Li-fei DU Yan-ling WANG Zhen-hua
(College of Information Technology, Shanghai Ocean University, Shanghai 201306, China)

Abstract Big data have been continually drawing extensive interests in both academia and industry. Currently, the scale of marine data is increasing consecutively and exponentially with the rapid development of ocean observation technologies and data acquisition methodologies. Until recently, most of the solutions focus on the generic big data, while an extensive study over marine data is still left undiscussed, since the uniqueness of marine data brings new challenges for its management. As a result, this article first outlined the characteristics of marine data as well as the fundamental architecture of marine data management. Secondly, this paper also analyzed the problems of data storage, data quality and data security as well as the corresponding tentative solutions, which will provide significant evidence and references for the future study over ocean science and engineering technology.

Keywords Marine, Big data, Data storage, Data quality, Data security

1 引言

信息技术尤其是云计算、物联网、信息获取技术、社交网络等新兴服务的快速发展,促使了各行业数据量的急剧增长,行业大数据已经成为目前研究的热点。在海洋领域,国际上已经开展了诸多观测计划(包括 Argo、海王星、OOI、GOOS、IOOS 等多个观测计划)并发射了多颗海洋观测卫星,采用多种不同的数据获取技术同时对海洋数据进行采集,导致海洋数据量急剧增加。如:Argo 计划目前在全球共投放浮标 10231 个,对海水温度、盐度、酸度和密度、二氧化碳等多个海洋要素数据进行实时采集,其中一个 Argo 数据中心在过去一年里就处理了 657 个活跃浮标观测的 21954 条剖面数据[1,2];NASA 发射的海洋观测卫星"水瓶座"对全球大洋的海洋环流、温度、成分以及海平面高度等数据信息进行扫描,扫描周期为 7 天,"水瓶座"2 个月内采集的数据量相当于调查船和浮标 125 年测量的历史总记录;截止到 2012 年底,NOAA 管理的年数据量高达 30 PB,每日能从卫星、船只、飞机、浮标以及其它传感器收集超过 35 亿份观测资料。全方位、多手段的海洋观测现状使得海洋数据具有量大、增长速度快、类型多样化以及蕴含价值大等特点,海洋数据呈现指数级增长。

海洋数据蕴含着巨大的价值,为人类更深入地感知、认识和控制物理世界提供了前所未有的丰富信息。例如:通过对 Argo 数据进行分析,可以发现地球正在寻找一个全球水文循环的强化[3];通过对声学遥感数据进行分析,能够得到海洋中的生物群落和物种分布,为保证海洋生态平衡提供了强大的科学参考[4];通过对"海王星"计划获取的地震活动、断层活动、洋中脊岩浆活动观测数据的分析,能够对海底地震和海啸进行预警预报[5,6]。由此可见,海洋数据能够实现生态、气候、灾害等多领域的预测、预警以及辅助决策。发挥海洋数据蕴含的巨大价值是海洋数据管理的目标。因此,海洋数据存储、分析、安全和质量控制等管理环节的研究对生态系统、人

到稿日期:2015-04-26 返修日期:2015-08-27 本文受国家 973 项目(2012CB316200),国家自然科学基金项目(61272098),上海高校青年教师培养资助计划(ZZHY14024)资助。

黄冬梅(1964—),女,硕士,教授,主要研究方向为数据库技术、海洋大数据管理、智能决策,E-mail:dmhuang@shou.edu.cn;**赵丹枫**(1982—),女,博士,讲师,主要研究方向为业务流程管理、DAS 模型;**魏立斐**(1982—),男,博士,讲师,主要研究方向为密码学、网络安全;**杜艳玲**(1987—),女,博士生,主要研究方向为云存储、海洋大数据管理;**王振华**(1982—),女,博士,副教授,主要研究方向为数据质量管理。

类社会、科学研究等各个方面都具有重要的战略与现实意义。

目前的研究工作主要集中在解决大数据管理的通用问题上，然而，海洋数据多源、超高维、海量、实时、多类、敏感以及空间性的特征给其管理带来了诸多挑战。在质量控制环节存在数据质量水平不一、质量问题多样化、检测方案不固定等问题；在数据存储环节存在存储系统可扩展性低、时效性不高的问题；在数据分析环节存在速度和实时性的问题；由于数据安全涉及海洋数据管理的各个环节，因此，在数据安全环节存在存储安全、访问安全、计算安全、共享安全和监管安全的问题。若上述问题得不到很好的解决，海洋数据的价值将得不到有效发挥。本文通过对海洋数据管理流程进行梳理，讨论了海洋数据管理各个环节中面临的挑战，综述了适用于海洋数据的对策，展望了海洋数据管理的发展方向。

本文第 2 节从海洋数据的来源分析了其特征；第 3 节探讨了海洋数据存储及分析面临的挑战，并综述了与其对应的对策；第 4 节描述了海洋数据质量控制问题并介绍了相关对策；第 5 节研究海洋数据面临的问题及关键解决方案；第 6 节给出了海洋大场景应用实例；最后对本文进行了总结与展望。

2 海洋数据管理

2.1 海洋数据

通过卫星、航空遥感、海洋站、调查船、浮标等手段获取的服务于海洋相关领域的一类大数据称为海洋数据。通过对海洋数据的分析，能够为生态平衡、人类文明提供有力的科学参考。

海洋数据具有显著的特征[7]，具体如下所述。

(1)多源多类性。海洋数据的来源广泛，包括卫星、航空遥感、海洋站、调查船、浮标以及海底观测系统等。各数据源采用不同的数据获取技术对海洋进行观测，然而，不同的数据获取技术在技术指标、数据格式、参数以及使用区域上存在巨大的差异，造成了海洋数据呈现多类性。

(2)时空性。海洋数据具有强时空性，每一次数据观测都对应具体的时间与空间位置信息，海洋数据的应用价值一定是在具体的时间与空间位置下才具有。

(3)超高维。海洋科学涉及多个学科，包括海洋物理、海洋化学、海洋生物、海洋环境和海洋经济等。每一个数据对象除具有时间、空间位置信息之外，还包含多个维度信息，如：海水温度、盐度、酸度、密度和流速等。

(4)海量性。海洋数据在具有超高维特征的同时，各海洋观测计划覆盖全球几乎所有大洋，利用多种观测技术进行周期的、实时的数据采集，导致海洋数据呈现指数级增长。

(5)敏感性。海洋数据中包含大量机密敏感数据，如长周期的海洋气象、水文、潮位数据，海洋渔业和油气矿产资源数据，大比例尺的海岛暗礁、近海岸线数据，灾害预警与评估等，需要采取数据安全保障措施。

2.2 海洋数据管理架构

海洋数据来源广泛，应用需求和数据类型不尽相同。通过对海洋数据的分析，梳理出海洋数据管理过程，主要包括数据采集、数据处理、数据存储、数据分析、数据应用以及贯穿整个流程的质量控制和数据安全保障等多个环节。海洋数据管理的基本架构如图 1 所示。

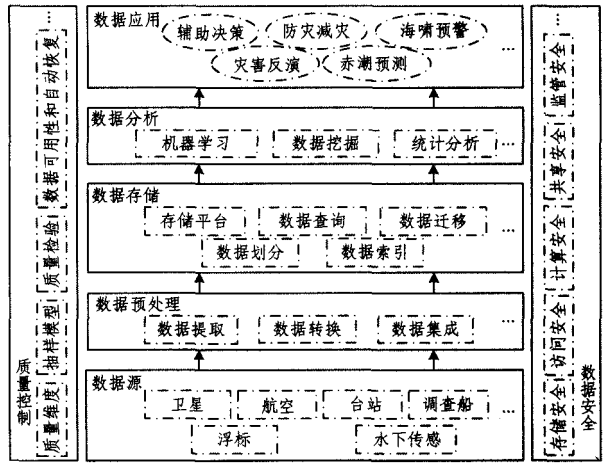

图 1 海洋数据管理的基本架构

海洋数据的来源包括卫星、航空遥感、海洋观测站、调查船、浮标和海底观测系统等。由于海洋数据结构和种类极为繁杂，因此需要对不同数据源获取的数据进行预处理，在数据存储环节，包括了存储平台、数据划分、建立索引、数据查询和数据迁移等。海洋数据的分析包括机器学习、数据挖掘、统计分析等，对海洋数据的分析可以为海啸预警、赤潮预测、辅助决策、防灾减灾和灾害反演提供精确、可靠的科学依据。质量控制和数据安全是整个体系架构的保障。

3 海洋数据的存储与分析

数据存储和分析是大数据管理的主要组成部分，对于海洋数据特殊的行业应用需求，通用的大数据存储和分析技术往往不能完全适用。本节分析了海洋数据存储和分析所面临的挑战，并从云存储、数据划分、数据索引、查询处理、数据分析等方面讨论了现有方法中适用于海洋数据的对策。

3.1 海洋数据存储与分析的挑战

海洋数据具有海量、多模态、多维多类等特征，对存储的可扩展性以及数据模型有较高要求。合理的数据存储模式为大规模海洋信息决策分析提供了重要技术支撑和保障，准确的数据分析是支撑海洋数据应用的关键。本小节结合海洋数据的特征，探讨了目前海洋数据分析与存储的技术瓶颈。

3.1.1 海洋数据存储的挑战

数据存储是数据高效应用的有力保障。传统数据存储方案存在存储系统可扩展性差、数据模型相对固定等问题。然而海洋数据多源、超高维、海量、实时、多类以及敏感的特征对数据存储提出了新的挑战，主要包括：

(1)存储空间可扩展性要求。海洋数据海量性、实时性的特征使得数据存储系统在硬件架构和文件系统上的要求远远高于传统技术，要求数据存储空间具有高扩展性，随着实时观测数据的采集，数据存储空间应具有强大的弹性。

(2)存储系统存储模型多样性要求。海洋数据的多源性导致了海洋数据模态千差万别，包括结构化属性数据(*.MDB，*.dbf，*.bak，*.dmp 等)、空间数据(*.shp，*.adf，*.tif，*.jpg，DEM 等)、非结构化数据(*.doc，*.xls，*.pdf，*.txt，*.xml 等)。数据格式的多样性对数据库的一致性(Consistency)、可用性(Availability)和分区容错性都提出了更高的要求。

3.1.2 海洋数据的分析挑战

数据分析的目的是从多、广、杂的数据中发现规律，提取

新的知识，是挖掘海洋数据价值的关键。传统数据分析对象多是结构化、单一对象的小数据集，分析挖掘更侧重于根据先验知识预先人工建立模型，然后依据既定模型进行分析。多源异构的海洋数据存在着数量庞大、格式不一、质量不高等问题，相应的数据分析需要建立在一定的质量控制和抽样模型的基础上。此外，传统的分析技术如数据挖掘、机器学习、统计分析等要应用于海洋数据，则需要作出调整，面临的挑战主要包括：

(1)速度要求。海洋数据量大、类型庞杂，数据的分布特点不确定，需要根据处理的数据类型和分析目标，采用适当的算法模型快速处理数据，对于机器硬件以及算法都具有一定的挑战。

(2)实时性要求。海洋数据的应用常常具有实时性的特点，例如在雪龙号极地极端环境下工作，需要对天气、海冰、海底、船自身等各类实时信息进行综合分析。大量实时数据的处理和分析需要消耗大量的计算资源，传统的单机或并行计算技术很难保障，需要与云计算相结合，因此对算法的实时性和可扩展性提出了考验。

3.2 海洋数据存储及分析的对策

结合海洋数据的特征，本小节从云存储、数据划分、索引与查询、数据迁移、数据分析这几方面讨论目前可应用于海洋数据的存储系统与分析模型。

云存储在大数据存储中被广泛应用。目前主要的云存储平台有 Google Store、Amazon 的 S3、Microsoft 的 Azure 以及 IBM 的“蓝云”等。为了使云存储能较好地适用于敏感的空间海洋数据，需根据海洋数据安全密级的不同对其进行数据划分，建立适当的索引结构，以提升超高维海洋数据的查询效率。随着实时观测数据的持续采集以及数据存储系统数据量的不断累积，需要在考虑海洋数据特性的基础上对其进行动态迁移，实现存储系统资源的最优化利用。

如何进行数据的划分是影响拓展性、负载平衡以及系统性能的关键问题，它影响着数据访问速度以及数据利用效率。现有数据划分方法主要从以下几个角度出发。根据数据敏感和密保级别的不同，数据划分的方法有：通过计算数据敏感度实现动态划分[8]、采取物理隔绝[9]及用户访问设限[10]等密文方式分流数据到相应节点；另一方面，也有一些基于统计分析理论的数据划分方法，主要有：采取聚类方式对数据进行划分[11]、根据数据分布自适应选择划分边界[12]或采用抽样[13]等方法来减轻海量数据的处理压力，避免数据倾斜，实现数据稳定及动态分布。

索引是影响整个数据库系统效率的关键，是提高数据库系统执行效率的一种有效工具。海洋数据主要采用分布式的云存储方式，在此基础上的研究包括哈希结构索引[14]、树状索引[15]、以时间为主体的复合索引[8]、基于并行处理技术的优化索引[16]、随数据迁移而动态调节的索引[17]。

大数据查询技术指的是对相关数据进行检索，方便用户快速有效地找到需要及感兴趣的数据。为了提高云存储平台的查询效率，需减轻计算压力并提高传输速度，研究查询优化技术。目前的研究主要从算法实现角度出发，如共享历史查询结果作为中间结果[18]，根据数据特征自适应抽取样本[19]及根据度量标准提取代表元组[20]等；此外，相关研究也有从硬件角度出发，采取任务调度的方式[21,22]，实现高效并行处理。

存储平台上数据的动态迁移保证了存储资源的优化利用。两种传统的数据迁移方法分别是基于存储空间的高低水位法[22]和基于数据访问频率的 Cache 替换迁移算法[23]。随着存储技术的发展，出现了多种数据存储模式。在分级存储中，根据迁移模型实现数据自动迁移[24]；在多级存储中，采用 CuteMig 迁移方法实现数据的迁移[25]；在混合云存储中，通过对数据敏感度和迁移函数的计算对数据进行动态迁移[7]。

大数据分析模型的基本要求是实时性。MapReduce 适用于批处理[26]。Pregel[27]主要用于图的计算。Dremel[28]能够实现极短时间内的海量数据分析，支持云端大数据分析平台 BigQuery[29]。分析工具 PowerDrill 采用列存储，且使用压缩技术将尽可能多的数据装载进内存[30]。Dryad[31]与 Cascading 模型均支持有向无环图(Directed Acycline Graph，DAG)类型的应用。

4 海洋数据质量控制理论研究

海洋数据的质量是海洋地理信息科学生存和发展的基础，一套立足于全周期、针对海洋数据的质量控制方案是实现海洋时空大数据精确有效分析及应用的重要保证。本节主要分析了具有海量、多源等特性的海洋时空大数据的质量控制体系所面临的挑战，并从海洋时空大数据的采集抽样、质量检验、数据可用、自动检验与修复等方面综述了贯穿整个数据生命周期的质量控制理论模型和方法。

4.1 海洋数据质量控制挑战

由于海洋数据具有多源、多类、多维、多尺度等特性，导致原本应用于传统工业产品的质量管理手段不能完全适用于对数据批量和内部关联有明显需求的海洋时空大数据产品。因此基于海洋数据特性，面向海洋数据的整个采集、处理、再生过程的质量控制理论研究，是海洋数据发展亟待解决的关键问题之一。结合海洋数据的特征，其质量控制理论的研究面临的挑战主要包括：

(1)如何制定适用于海洋时空大数据的质量检验方案。面对来源多样性、形式多样化以及具有空间相关性的海洋数据，如何基于其特点从海量数据中选择“适量的样本数据”，并根据数据的应用精度要求给出“合理的质量判定”，是海洋数据质量控制的首要问题。

(2)如何平衡海洋数据质量需求和信息冗余之间的关系。海洋数据的空间相关性使得质量检验中样本点的选择不同于传统的抽样方法。其数据间距离的远近制约了样本点之间的信息冗余度。充分考虑海洋数据质量的相关性，使得其在检验费用一致的信息情况达到最大，是保障海洋数据质量控制方案有效实施的关键问题。

(3)如何界定和利用弱可用数据。海洋数据根据其质量检验结果可以分为可用数据和弱可用数据。由于海洋数据的获取途径不同，很多海洋数据具有不可逆性。研究海洋数据产品的可用性，对弱可用数据进行必要的数据清洗和自动修复，也是完善数据管理机制的关键问题之一。

4.2 海洋数据质量控制的对策

针对上述问题，需要贯穿整个数据生命周期的数据质量控制机制，包括对数据质量维度上的约束、抽样方案的选择、质量检验标准的制定以及对弱可用数据的处理和自动修复方案等。

(1)海洋数据的质量维度与标准。数据质量本质上可以理解为数据对应用的适用性[32]，并可以细分为数据质量相关的多种维度。数据质量的概念一般使用一致性、完整性、时效

性、可用性以及可信性来描述[33]，可以将上百个对数据质量有影响的因素归结为内在因素、应用因素、数据表述和数据存取4个大类。对于空间数据，现有研究提出了5个重要的数据质量评价要素，包括空间精度、主题精度、逻辑一致性、完整性和谱系[34]。针对不同的数据质量维度，相关学者和机构也先后制定了相应的数据质量标准[35]来对空间数据质量进行约束和界定。

(2)空间数据的抽样模型。抽样方法是处理海量信息的一种有效方法，通过选取小量样本代表总体，以较小的精度牺牲换取较大的效率提升，具有效率高、费用低等优点。在空间抽样调查方面，在样本不独立的情况下，Bootstrap算法通过引入二次抽样思想[36]，大大提高了分布式计算环境下大数据质量评估的效率。在空间数据的抽样方面，基于分层抽样思想的"Sandwich"抽样模型[37]通过考虑空间对象的自相关性，解决了空间异质性带来的问题。

(3)空间数据的质量检验模型。现有的针对空间数据的质量检验模型的研究主要集中于统计学理论和模糊集理论两方面。基于统计学理论的质量检验模型研究主要包括标准型、挑选型、调整型3种，现有研究主要体现在标准型质量检验模型[38]和调整型质量检验模型[39]以及两种模型结合的方法研究[40]；模糊集理论方面主要有基于检验批一阶、二阶的抽样检验模型，解决了模型中参数不确定的问题[41]。

(4)数据可用性理论。一个空间数据集合的可用性包括数据的一致性、精确性、完整性、时效性和实体同一性[42]。在数据的一致性方面的研究主要包括基于语义规则描述的研究[43]和基于统计学描述方法的研究[44]；在数据完整性方面的最经典的研究是基于条件表的不完整数据表述系统[45]；在数据精确性方面的研究较少；在数据时效性方面的研究主要是解决数据时效性的判定问题以及数据时效性的自动发现和修复问题[46]；在数据的实体同一性方面的研究主要是实体同一性错误检测的问题。

(5)数据的自动检测与修复。数据错误的自动检测研究主要包括一致性和对实体同一性两方面。一致性错误的研究集中于对自动检测算法[47]、分布式数据库检测方法的设计和探索。实体同一性方面主要是以最大化识别精度[48]和最大化识别效率为目标的研究[49]。在数据的自动修复方面，解决数据不一致问题主要采用传统的函数依赖所发现的数据不一致问题的研究[50]。在解决数据实体同一性问题上采用的是数据融合技术[51]。

5 海洋数据安全

海洋数据应用涉及到国家行业的开展，因此，海洋数据的收集、处理、发布都存在着安全控制。安全控制的主要方式有物理隔绝和访问控制等。目前国内外针对数据安全的研究具有很强的针对性，本节将从数据存储安全、访问安全、计算安全、共享安全、监管安全等安全领域出发，描述现有的安全挑战及相应的解决方案。

5.1 海洋数据安全的挑战

海洋数据的安全与传统的通信模式下的数据安全与隐私保护有显著不同，海洋数据呈现出典型的结构型特征，包括"一对多"(一个用户存储，多个用户访问)、"多对一"或"多对多"等数据安全与隐私保护模式。从数据的业务流程上看，对海洋数据的管理可以分为数据存储服务、数据访问服务、数据计算服务、数据共享服务和数据监管服务；从而，海洋数据的安全问题可以简单概括为"存得住、易共享、可计算、查得到、能监管"。海洋数据安全与隐私保护的需求问题集中体现在上述5个环节上，分别面向海洋数据的存储安全、访问安全、计算安全、共享安全和监管安全。

(1)从斯诺登事件人们已经意识到：如果大数据未能妥善存储，会对用户的隐私造成极大的危害。海洋数据的存储手段往往依赖于服务器/节点的存储安全或节点本身的可信性，无法抵抗节点管理者或敌手对数据的窃取或篡改。如果对数据不加甄别而直接利用，实时的数据也会欺骗用户；尤其是伪造或刻意制造的数据，往往会导致错误的结论。

(2)大数据的访问控制是实现数据受控共享的有效手段。海洋数据可能用于不同的场景访问中。因此，海洋数据被多个不同用户、不同角色、不同密级的人访问，其访问控制需求也十分突出。传统的访问控制技术主要依赖于对数据库的访问控制，一旦数据库管理者或者服务提供商出现了恶意行为，数据的访问控制将难以确保安全，从而对用户隐私和机密数据造成侵害。

(3)计算分析是海洋数据的一个重要应用。由于提供计算服务的大数据服务商不能被完全信任或者计算服务往往通过外包的方式进行，如何能够在实现数据隐私与机密性的前提下，依然能够实现数据的有效计算与分析，是海洋数据的重要需求。同时，它能够克服目前已有同态算法效率低下的缺陷，提高计算与分析的效率，保证数据的适用性。

(4)在海洋数据共享与分发过称中，由于用户/节点的密钥可能被有意或无意泄露，导致数据被泄露或被非法窃取，无法实现云环境下数据共享和分发机制的健康运行。由于现代密码技术往往仅依赖于密钥的安全性，如果无法对泄露者的密钥进行追踪和撤消，数据的安全体系可能会整体瓦解。

(5)对海洋数据监管是保证海洋数据安全的又一重要手段。在数据存储、计算、共享与分发的过程中，恶意的用户可能会插入伪造数据，无意的用户可能会插入错误数据，如果缺少有效的监管监控(拦截与删除违法信息，减少和降低冗余开销，检验存储内容完整性，验证计算结果的正确性等手段)，都有可能导致数据利用环节出现问题。

综上所述，只有确保了这5个环节中海洋数据的机密性、完整性、认证性、可用性、可控性等安全属性，才能达到完整的海洋数据安全与隐私保护目标。

5.2 海洋数据安全的对策

海洋数据的安全与传统的点对点通信模式下的数据安全与隐私保护有显著不同，导致了传统的数据安全与隐私保护方法在海洋数据环境中的应用已受到严重制约。因此，面向海洋数据，需要从数据存储安全、访问安全、计算安全、共享安全、监管安全5个方面研究与开发数据安全与隐私保护方案。

(1)海洋数据的存储安全。在海洋数据存储中，现有的存储安全依赖于服务器/节点的安全或节点本身的可信性。为了改变这种现状，需要研究基于密文的数据存储[52]来抵抗节点管理者或敌手对数据的窃取或篡改，并且将数据访问的管理权交予多个不同的管理者[53]，以减小因为单一管理者的恶意行为而带来的数据损失。同时，在密文存储结构中对数据进行完整性检验[54]和数据存储证明[55]，因此需要支持密文存储的数据隐私保护技术[56]来实现存储安全。

(2)海洋数据的访问安全。在海洋数据访问中，海洋数据

被多个不同用户、不同角色、不同密级的人访问，传统对明文的访问控制技术主要依赖于对数据库的访问控制，难以对非可信的大数据平台实施基于密文的访问控制。采用了基于密文的数据存储技术后，需要支持密文检索[57]、支持细粒度访问[58]、支持“与、或、非”逻辑功能的灵活丰富访问[59]和基于密文数据的索引[60]、搜索[61,62]等的数据隐私保护来实现访问安全。

(3)海洋数据的计算安全。在海洋数据计算分析中，由于提供计算服务的大数据服务商不能被完全信任或者计算服务往往通过外包的方式进行，计算分析功能所需要的输入/输出均应该以密文形式进行传递；需要研究在密文的基础上实现密文的直接计算[63]，而不是将密文进行解密后再计算。在海洋数据计算分析过程中，需要支持密文的线性方程组求解[64]、数据分析与挖掘、图像处理[65]、全同态加/解密[66-68]等数据隐私保护来实现计算安全。

(4)海洋数据的共享安全。海洋数据共享依赖于用户的密钥，确保云环境下基于密文的数据共享和分发机制[69]的健康运行，必然需要支持数据泄露时可追踪技术[70-72]、访问权限撤销技术[73]等数据隐私保护技术来实现共享安全。同时，在面对海量数据时，需要支持密文数据的批量共享与分发[74]，研究海洋数据隐私方案的优化[75]和高效实现[76]，以提高数据批量处理能力。

(5)海洋数据的监管安全。在海洋数据监管中，为了保证数据的有用性，数据存储、计算与共享的过程中需要有效的监管监控[77]：拦截与删除违法信息[78]，减少和降低冗余开销，存储内容完整性检验[79]，计算结果正确性验证[65]和敏感信息提炼挖掘等。在监管监控时，还需要对用于个人隐私保护和大数据的监管监控进行协调处理[80]。因此，需要有效的监控与监管手段来实现监管安全。

6 海洋数据应用实例

由海洋数据管理基本框架可知，海洋要素、场景、灾害过程等的可视化需要兼顾海洋数据的敏感性、时空相关性、海量性等特征。本节通过实际项目对海洋数据大场景应用实例进行简要介绍。该项目主要对灾害发生海域的地形进行精细化建模，利用高精度的DEM数据进行通视分析、视域分析和扩散分析，进而实时生成动态最优撤离路径，辅助人员撤离。

该项目涉及灾害发生海域的精细化地形数据、高精度DEM数据、实时遥感监测数据等各类环境数据，数据量达到PB级，并以每天数十GB的速度增长。针对该数据的多源性、敏感性、海量性等特点，设计了混合云存储架构，按照数据的时空信息进行划分存储，并进行动态迁移。在此基础上，对数据的分析、计算和再现并行化，保证了应用的时效性。其中，可视化效果如图2—图5所示。

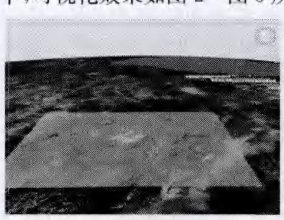

图2 三维地形

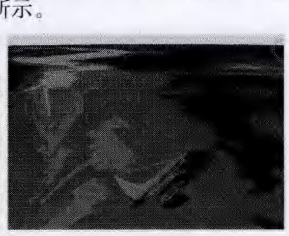

图3 视域分析

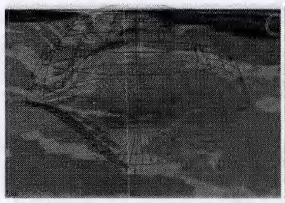

图4 扩散分析

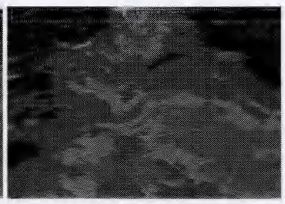

图5 最优路径

结束语 海洋领域的需求具有极其明显的行业特征，海洋数据的合理利用对生态系统和人类社会都具有重大的意义。现有大数据的研究工作多集中于互联网数据，针对海洋数据管理的研究刚刚起步。已有少数研究工作涉及到海洋数据的质量控制和存储方法，但基于海洋数据特征的数据分析方法、可视化方法的研究还很少见，目前仍采用大数据通用方法来解决。针对海洋数据多源、多类、多模态、时空性的特点，如何发挥海洋数据的价值，尚需要我们从数据管理各个环节进行深入、系统的科学研究。

参考文献

[1] Argo 简讯[EB/OL]. http://www.argo.org.cn

[2] Argo data center in China[OL]. http://www.argo.org.cn

[3] Durack P J, Wijffels S E, Matear R J. Ocean Salinities Reveal Strong Global Water Cycle Intensification During 1950 to 2000 [J]. Science, 2012, 336(6080): 455-458

[4] Brown C J, Smith S J, Lawton P, et al. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques[J]. Estuarine Coastal and Shelf Science, 2011, 92(3): 502-520

[5] Rogers G C, Meldrum R, Baldwin R, et al. The NEPTUNE Canada Seismograph Network[J]. Seismological Research Letters, 2010, 81(2): 369-379

[6] Rabinovich A B, Thomson R E, Fine I V. The 2010 Chilean Tsunami off the west coast of Canada and the northwest coast of the United States[J]. Pure and Applied Geophysics, 2013, 170(9): 1529-1565

[7] Huang Dong-mei, Du Yan-ling, He Qi. Migration Algorithm for Big Marine Data in Hybrid Cloud Storage[J]. Journal of Computer Research and Development, 2014, 51(1): 199-205(in Chinese)
黄冬梅，杜艳玲，贺琪. 混合云存储中海洋大数据迁移算法的研究[J]. 计算机研究与发展，2014，51(1)：199-205

[8] Huang Dong-mei, Sun Le, Zhao Dan-feng, et al. An efficient hybrid index structure for temporal marine data[C]// Proceeding of Conference on Web-Age Information Management. 2014: 187-199

[9] Zhou Xiang-min, Wang Guo-ren. Key Dimension Based High-Dimensional Data Partition Strategy[J]. Journal of Software, 2004, 15(9): 1361-1374(in Chinese)
周项敏，王国仁. 基于关键维的高维空间划分策略[J]. 软件学报，2004，15(9)：1361-1374

[10] Ren Ping, Liu Wu, Sun Dong-hong. Partition-based data cube storage and parallel queries for cloud computing[C]// Proceedings of the Ninth International Conference on Natural Computation (ICNC). 2013: 1183-1187

[11] Zhao Dan-feng, Jin Shun-fu, Liu Guo-hua, et al. A cryptograph index technology based on query probability in DAS model[J]. Journal of Yanshan University, 2008, 32(6): 77-482

[12] Han Lei, Sun Xu-zhan, Wu Zhi-chuan, et al. Optimization Study on Sample Based Partition on MapReduce[J]. Journal of Computer Research and Development, 2013, 50(6): 77-84 (in Chinese)
韩蕾,孙徐湛,吴志川,等. MapReduce 上基于抽样的数据划分最优化研究[J]. 计算机研究与发展,2013,50(6):77-84

[13] Xu Yu-jie, Zou Peng, Qu Wen-yu. et al. Sampling-Based Partitioning in MapReduce for Skewed Data[C]// Proceeding of the 7th Conference on China Grid. 2012: 1-8

[14] Shi Sui-xiang, Lei Bo. Theory and Practice on China Digital Ocean[M]. Beijing: Ocean Press, 2011

[15] Fox A, Eichelberger C, Hughes J, et al. Spatio-temporal indexing in non-relational distributed databases[C]// Proceeding of IEEE Conference on Big Data. 2013: 291-299

[16] Zhong Yun-qin, Fang Jin-yun, Zhao Xiao-fang. VegaIndexer: A Distributed composite index scheme for big spatio-temporal sensor data on cloud[C]// Proceedings of the IEEE Conference on Geoscience and Remote Sensing Symposium (IGARSS). 2013: 1713-1716

[17] Su Chen, Beng C O, Tan K L, et al. ST^{2}B-tree: a self-tunable spatio-temporal b$^+$-tree index for moving objects[C]// Proceeding of Conference on ACM Special Interest Group Conference on Management of Data(SIGMOD). 2008: 29-42

[18] Kaufmann M, Manjili A A, Vagenas P, et al. Timeline index: a unified data structure for processing queries on temporal data in SAP HANA[C]// Proceeding of Conference on ACM Special Interest Group Conference on Management Of Data(SIGMOD). 2013: 1173-1184

[19] Hu Xiao-cheng, Miao Qiao, Tao Yu-fei. Independent Range Sampling[C]// Proceedings of the 33rd ACM Special Interest Group Conference on Management of Data. 2014: 246-255

[20] Zhang Jin, Chen Guo-qing, Tang Xiao-hui. Extracting Representative Information to Enhance Flexible Data Queries[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(6): 928-941

[21] Thomas S, Kevin L. MapReduce Optimization Using Regulated Dynamic Prioritization[C]// Proceedings of the SIGMETRICS. 2009: 299-310

[22] Gharibi W, Mousa A. Query optimization based on time scheduling approach[C]// IEEE East-West Design & Test Symposium. 2013: 1-7

[23] Gibson T, Smith C H M, Miller E. An improved long-term file usage prediction algorithm[C]// Annual International Conference on Computer Measurement & Performance. 2002: 639-648

[24] Jeong J, Dubois M. Cost-sensitive cache replacement algorithms [C]// Proceedings of the 9th International Symposium on High-Performance Computer Architecture (HPCA-9). 2003, 327

[25] Reed B, Darrell D. Analysis of caching algotithms for distributed file systems[C]// Proceedings of the ACM SIGOPS Operating Systems Review. 1996: 12-21

[26] He Ding-shan, Zhang Xian-bo, Grider G, et al. Coordinating parallel hierarchical storage management in object-based cluster file system[OL]. http://wiki. lustre. org/images/f/fc/MSST-2006-paper. pdf

[27] Ao L, Yu D, Shu J, et al. A tiered storage system for massive data: TH-TS[J]. Journal of Computer Research and Development, 2011, 48(6): 1089-1100

[28] Li Feng, Beng C O, Ozsu M T. Distributed Data Management Using MapReduce[J]. ACM Computing Surveys (CSUR), 2014, 46(3): 1-41

[29] Malewicz G, Austern M H, Bik A J C, et al. Pregel: A system for large-scale graph processing[C]// Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. 2010: 135-146

[30] Valiant L G. A bridging model for parallel computation[J]. Communicationg of ACM, 1990, 33(8): 103-111

[31] Hall A, Bachmann O, Büssow R, et al. Processing a trillion cells per mouse click[J]. Proceedings of the VLDB Endowment, 2012, 5(11): 1436-1446

[32] Shank G, Wang R Y, Mostapha Z. IP-map: Representing the manufacture of an information product[C]// Proceedings of the Information Quality Conference. 2000

[33] Yair W, Wang R Y. Anchoring data quality dimensions in ontological foundations[J]. Communications of the ACM, 1994, 39(11): 86-95

[34] Zargar A, Devillers R. An operation-based communication of spatial data quality[C]// Proceedings of the International Conference on Advanced Geographic Information Systems & Web Services. 2009: 140-145

[35] ISO 2859. 0. Sampling Procedures for Inspection by Attributes-Part 0[J]. Introduction to the ISO 2859 attribute sampling system, International Organization for Standardization, 1995: 56-63

[36] Kleiner A, Talwalkar A, Sarkar P, et al. A scalable bootstrap for massive data[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76: 795-816

[37] Li Lian-fa, Wang Jin-feng. Spatial Sampling model of Geographic Data[J]. Progress in Natural Science, 2002, 12(5): 99-102 (in Chinese)
李连发,王劲峰. 地理数据空间抽样模型[J]. 自然科学进展,2002,12(5):99-102

[38] Akhavan S T, Nezhad M S. Designing an optimum acceptance sampling plan using bayesian inferences and a stochastic dynamic programming approach[J]. Transaction E-Industrial Engineering, 2009, 16(1): 19-25

[39] Jamkhaneh E B, Gildeh B S. AOQ and ATI for Double Sampling Plan with Using Fuzzy Binomial Distribution[C]// Proceedings of International Conference on IEEE Intelligent Computing and Cognitive Informatics (ICICCI). 2010: 45-49

[40] Duarte B P M. An optimization-based approach for designing attribute acceptance sampling plans[J]. International Journal of Quality & Reliability Management, 2008, 25(8): 824-841

[41] Wang Zhen-hua, Zhou Xue-nan, Huang Dong-mei. Sampling Model for Quality Inspection of Uncertain Ocean Data[J]. Computer Science, 2015, 42(2): 182-184 (in Chinese)
王振华,周雪楠,黄冬梅. 不确定海洋数据的质量抽样检验模型研究[J]. 计算机科学,2015,42(2):182-184

[42] Li Jian-zhong, Liu Xian-min. An Important Aspect of Big Data: Data Usability[J]. Journal of Computer Research and Development, 2013, 50(6): 1147-1162 (in Chinese)
李建中,刘显敏. 大数据的一个重要方面:数据可用性[J]. 计算机研究与发展,2013,50(6):1147-1162

[43] Fan Wen-fei, Geerts F, Li J, et al. Discovering conditional functional dependencies[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 683-698

[44] Golab L, Korn F, Srivastava D. Efficient and Effective Analysis

of Data Quality using Pattern Tableaux[J]. IEEE on Data Engineering,2011,34(3):26-33

[45] Grahne G. The Problem of Incomplete Information in Relational Databases[M]. Berlin:Springer,1991

[46] Fan Wen-fei, F Geerts, Jia Xi-bei. Conditional functional dependencies for capturing data inconsistencies[J]. ACM Transactions on Database Systems (TODS),2008,33(2):1-48

[47] Chaves L,Buchmann W F,Bohm E K. Finding misplaces items in retail by clustering RFID data[C]// Proceedings of the 13th International Conference on Extending Database Technology. 2012:501-512

[48] Fan Wen-fei,Geerts F,Shuai M,et al. Detecting in consistencies in distributed data[C]// Proceedings of IEEE ICDE. 2010:64-75

[49] Fan Wen-fei, Geerts F, Wijsen J. Determining the currency of data[J]. ACM Transactions on Database Systems (TODS), 2012,37(4):1-46

[50] Whang S E,Menestrina D,Koutrika G,et al. Entiry resolution with interative blocking[C]// Proceedings of the 35th SIGMOD Conference on Managemnt of Data. 2009

[51] Chomicki J,Marcinkowski J. Minimal-change intergrity maintenance using tuple deletions[J]. Information and Computation, 2005,197(1):90-121

[52] Lin Huang,Cao Zhen-fu,Liang Xiao-hui,et al. Secure Threshold Multi Authority Attribute based Encryption without a Central Authority[M]// Progress in Cryptology—INDOCRYPT 2008. 2008:426-436

[53] Liu Zhen,Cao Zhen-fu,Huang Qiong,et al. Fully Secure Multi-Authority Ciphertext-Policy Attribute-Based Encryption without Random Oracles[M]// Computer Security—ESOR2CS 2011. 2011:278-297

[54] Yang Kan,Jia Xiao-hua. Data Storage Auditing Service in Cloud Computing:Challenges, Methods and Opportunities[J]. World Wide Web (WWW),2012,15(4):409-428

[55] Wang Cong,Wang Qian,Ren Kui,et al. Privacy-preserving public auditing for data storage security in cloud computing[C]// 29th IEEE Conference on Computer Communications (INFOCOM'10). 2010:1-9

[56] Li Ming,Yu Shu-cheng,Ren Kui,et al. Toward privacy-assured and searchable cloud data storage services[J]. IEEE Network, 2013,27(4):56-62

[57] Cao Ning,Wang Cong,Li Ming,et al. Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data[C]// IEEE Transactions on Parallel and Distributed Systems. 2014:829-837

[58] Liang X, Cao Z, Lin H, et al. Provably Secure and Efficient Bounded Ciphertext Policy Attribute Based Encryption[C]// ASIACCS. 2009:343-352

[59] Yang Kan,Jia Xiao-hua,Ren Kui,et al. Enabling Efficient Access Control with Dynamic Policy Updating for Big Data in the Cloud[C]// INFOCOM 2014. 2014:2013-2021

[60] Wang Hui, Lakshmanan L. Efficient secure query evaluation over encrypted XML databases[C]// VLDB Endowment,2006: 127-138

[61] Cao Ning,Wang Cong,Li Ming,et al. Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data[C]// Proceedings of IEEE INFOCOM. 2011:829-837

[62] Wang Cong,Ren Kui,Yu S,et al. Achieving Usable and Privacyassured Similarity Search over Outsourced Cloud Data[C]// Proceedings of IEEE INFOCOM. 2012:25-30

[63] Shen E,Shi E,Waters B. Predicate privacy in encryption systems [C]// TCC. 2009:457-473

[64] Wang Cong,Ren Kui,et al. Harnessing the Cloud for Securely Outsourcing Large-scale Systems of Linear Equations[C]// IEEE Transactions on Parallel and Distributed Systems. 2013: 1172-1181

[65] Wang Cong,Xu Zhen,Zhang Bin-sheng,et al. OIRS:Outsourced Image Recovery Service from Compressive Sensing with Privacy Assurance[C]// NDSS. 2013

[66] Smart N P,Vercauteren F. Fully homomorphic encryption with relatively small key and ciphertext sizes[C]// Proc of the Public Key Cryptography. 2010:420-443

[67] Brakerski Z,Gentry C G,et al. Fully Homomorphic encryption without bootstrapping[C]// Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 2012:309-325

[68] Cheon J H,Coron J S,Kim J,et al. Batch fully homomorphic encryption over the integers[C]// EUROCRYPT. 2013:315-335

[69] Li Ming,Yu Shu-cheng,Ren Kui,et al. Scalable and Secure Sharing of Personal Health Records in Cloud Computing using Attribute-based Encryption[C]// IEEE Transactions on Parallel and Distributed Systems. 2013:131-143

[70] Liu Zhen,Cao Zhen-fu,Wong D S. White-box Traceable Ciphertext-Policy Attribute-Based Encryption Supporting Any Monotone Access Structures[J]. IEEE Transactions on Information Forensics and Security,2013,8(1):76-88

[71] Liu Zhen, Cao Zhen-fu, Wong D S. Blackbox Traceable CPA-BE:How to Catch People Leaking Their Keys by Selling Decryption Devices on eBay[C]// ACM CCS. 2013:4-8

[72] Ning Jian-ting,Cao Zhen-fu,Dong Xiao-lei,et al. Large Universe Ciphertext-Policy Attribute-Based Encryption with White-Box Traceability[C]// European Symposium on Research in Computer Security. 2014:55-72

[73] Yang Kan, Jia Xiao-hua, Ren Kui, et al. DAC-MACS: Effective Data Access Control for Multi-Authority Cloud Storage Systems [C]// IEEE Transactions on Information Forensics and Security. 2013:1790-1801

[74] Cao Zhen-fu. New Directions of Modern Cryptography [M]. Florida:CRC Press,2012

[75] Yang Kan,Jia Xiao-hua,Ren Kui. Attribute-based Fine-Grained Access Control with Efficient Revocation in Cloud Storage Systems[C]// ASIACCS. 2013:523-528

[76] Wei Li-fei,Zhu Hao-jin,Cao Zhen-fu,et al. Security and Privacy for Storage and Computation in Cloud Computing[J]. Information Sciences,2014,258:371-386

[77] Wang Cong, Ren Kui, Yu Shu-cheng, et al. Achieving Usable and Privacy-assured Similarity Search over Outsourced Cloud Data[C]// IEEE INFOCOM. 2012:25-30

[78] Feng Deng-guo,Zhang Min,Li Hao. Big Data Security and Privacy Protection[J]. Chinese Journal of Computers,2014,37(1): 1-13(in Chinese)
冯登国,张敏,李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014,37(1):1-13

[79] Wang Qian,Ren Kui,Meng Xiao-qiao. When Cloud Meets eBay: Towards Effective Pricing for Cloud Computing[C]// IEEE INFOCOM. 2012:936-944

[80] 曹珍富,大数据时代,如何提升政府治理能力[N]. 光明日报, 2014(11)