计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20160847
54(2):235-247,2017

科学大数据管理:概念、技术与系统

黎建辉[1] 沈志宏[1] 孟小峰[2]
[1](中国科学院计算机网络信息中心 北京 100190)
[2](中国人民大学信息学院 北京 100872)
(lijh@cnic.cn)

Scientific Big Data Management: Concepts, Technologies and System

Li Jianhui[1], Shen Zhihong[1], and Meng Xiaofeng[2]
[1](*Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190)
[2](*School of Information, Remin University of China, Beijing* 100872)

Abstract In recent years, as more and more large-scale scientific facilities have been built and significant scientific experiments have been carried out, scientific research has entered an unprecedented big data era. Scientific research in big data era is a process of big science, big demand, big data, big computing, and big discovery. It is of important significance to develop a full life cycle data management system for scientific big data. In this paper, we first introduce the background of the development of scientific big data management system. Then we specify the concepts and three key characteristics of scientific big data. After an review of scientific data resource development projects and scientific data management systems, a framework is proposed aiming at the full life cycle management of scientific big data. Further, we introduce the key technologies of the management framework including data fusion, real-time analysis, long termstorage, cloud service, and data opening and sharing. Finally, we summarize the research progress in this field, and look into the application prospects of scientific big data management system.

Key words scientific data; big data; data pipeline; full life cycle of data

摘 要 近年来,随着越来越多的大科学装置的建设和重大科学实验的开展,科学研究进入到一个前所未有的大数据时代.大数据时代科学研究是一个大科学、大需求、大数据、大计算、大发现的过程,研发一个支持科学大数据全生命周期的数据管理系统具有重要的意义.分析了研发科学大数据管理系统的背景,阐述了科学大数据的概念和三大特征,通过对科学数据资源发展和科学数据管理系统的研究进展进行综述分析,提出了满足科学数据管理全生命周期的科学大数据管理框架,并从数据融合、数据实时分析、长期存储、云服务体系以及数据开放共享机制5个方面分析了科学大数据管理系统中的关键技术.最后,结合科学研究领域展望了科学大数据管理系统的应用前景.

关键词 科学数据;大数据;数据流水线;数据全生命周期

中图法分类号 TP391

收稿日期:2016-11-15;**修回日期**:2017-01-14

基金项目:国家重点研发计划项目(2016YFB1000600)

This work was supported by the National Key Research Program of China (2016YFB1000600).

大规模巡天望远镜、大型粒子加速器、高通量基因测序仪等源源不断产生巨量科学数据,使得全球科技创新进入一个前所未有的科学大数据时代.科学大数据已成为科学发现的新型战略资源,一个国家的科学研究水平将直接取决于其在科学大数据的优势以及将数据转换为知识的能力.

面向大规模的科学数据管理,以及科学大数据应用,往往需要突破当今所有数据管理系统的极限,才能实现高效的科学知识发现,这也成为当下科学界和数据管理领域携手攻坚的"难题".概括起来,科学大数据管理面临的主要问题和挑战包括:1)超大规模关系数据管理.如天文领域多个数据中心千亿乃至万亿行天文星表数据的管理.2)多源数据关联和知识发现.如全球开放生物资源、文献、序列和疾病等万种数据源100亿级关联数据的知识发现,需6步以上关联挖掘.3)实时的高效数据处理.如引力波科学发现中,16 MHz采样频率10 000信道数据需要近似零延迟数据处理.

1 科学大数据概念与特性

1.1 科学大数据概念

科学数据是科研活动的输入、输出和资产.但究竟"什么是科学数据?",如何给"科学数据"一个确切的定义?迄今为止,还在困扰着学术界.Greenberg在其最近出版的著作《大数据,小数据,没数据》[1]中,列举了学术界对数据各种不同的认识和理解,"在自然科学、社会科学和人文科学领域,学者们创造、使用、分析和解释数据,但往往不知道这些数据的真正含义."

科学数据是对所研究的客观对象的某些现象的描述.这种描述,一般是指在领域或学科知识指导下,对客观对象进行科学抽象和概念化后,就其中的某些现象进行系统地、有目的地观测、调查、实验所形成的实体.因此,数据不是客观事物,数据不是带有自身特征的自然对象,数据只是对学术研究的客观对象中某些可观测到的现象的描述.这些描述会因人而异、因地而异和因时而异.把一些事物概念化为数据,本身就是一种学术研究活动.

科学数据是以科学证据形式存在的事实,它至少应该包括科学观测与监测的数据、实验数据、计算与模型模拟输出的数据、对情景或现象的描述数据、对行为的观测或定性描述数据、用于管理或者商业目的统计数据等,以及描述数据的元数据.它们通常是科研活动的输入,是证实、证伪科学发现、科学观点的事实与证据,或者是论证推理的基础.

科学数据从历史上非自动化的"手工采集"的方式,逐渐地过度到自动化的"机器采集".非自动"手工采集"的数据,其产生的速度较慢,数据量与复杂度不高,但数据的价值密度高.而通过大型仪器设备、大科学装置、大规模传感器网络等自动化采集的数字化数据,其产生的速度快,数据量和复杂性高,存在着不确定性和噪声.对这些数据进行存储、分析和应用需要新技术与更强的基础设施环境支持.科学大数据主要是指这种通过"机器"自动化快速采集、规模化存储与分析处理、具有较高维度和复杂关联的数据及其衍生产品.

随着越来越多的诸如500 m口径球面射电望远镜(five-hundred-meter aperture spherical radio telescope, FAST)、中国散裂中子源(China spallation neutron source, CSNS)等大科学装置的建设和重大科学实验的开展,以及无所不在的科学传感器和传感器网络广泛应用于天空、陆地和海洋,对自然环境进行全方位的探测、监测,源源不断产生的科学数据将科学研究快速推进到一个前所未有的大数据时代.科学大数据将改变人类几个世纪以来科学研究主要在于理解相对简单、未耦合或弱耦合系统这一局面,增强我们详细表征和描述复杂性的能力,以及分析高度耦合复杂系统的动态行为的能力,催生如希格斯粒子和引力波等重大科学发现.可以这样比喻,科学大数据为科学发现提供了一种新型的"望远镜"和"显微镜",在宏观上大大扩展了我们对复杂系统整体性进行研究的能力,在微观上,让我们的视线可以深入到复杂系统内部细微的行为和动态变化.

1.2 科学大数据的特征

相较于其他类型的大数据,科学大数据除了具有明显的"4V"特征[2-4]之外,还具有多层次逐级演化、全生命周期以及流水线处理和应用等特征.

1.2.1 多层次演化特征

科学大数据具有多层次逐级演化的显著特征.如图1所示,由大型仪器设备、大科学装置和计算模拟等产生的海量原始数据,经过校对、刻度、特征提取等处理形成具有科学意义的实例对象数据,并与相关的数据关联融合,形成知识网络.典型例子如美国航空航天局(NASA)地球观测系统(earth observing system, EOS)[5]卫星获取的数据按照其不断加工和演化过程,区分为0级、1A级、1B级、2级、3级、4级6个不同的级别.根据科学应用和目标的不同,

科学家可以直接使用精加工的 4 级数据,也可以使用 1A 级,甚至 0 级数据.

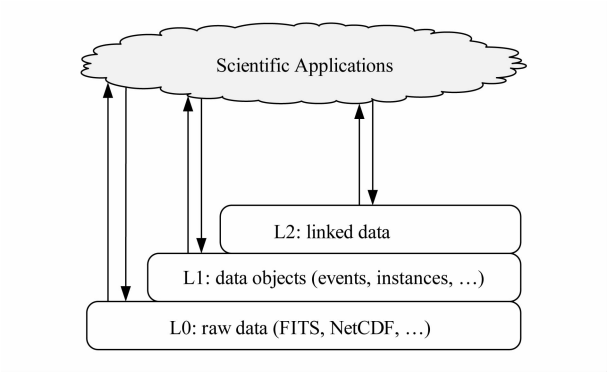

Fig. 1 Characteristics of multi-level progressive evolution of scientific big data

图 1 科学大数据具有多层次逐级演化的显著特性

1.2.2 全生命周期特征

科学大数据具有明显的涉及“采集与实时分析—存储与处理—发布与共享—再分析与重用—归档与长期保存”全过程的全生命周期特征.其中,采集与实时分析阶段主要完成科学实验装置、仪器设备、观测台站等数据的采集,并实现数据的实时筛选、处理和分析;存储与处理阶段主要完成对采集/筛选的数据的持久化存储,同时通过批量分析任务,完成初步的科学分析和科学发现;发布与共享阶段主要按照特定的主题,对科学数据进行组织管理,形成系列的数据集产品,通过 Web 等方式对科研界发布,提供数据共享与交换服务;再分析与重用阶段主要支持用户对发布的数据集进行二次整合分析,实现进一步的科学发现;归档与长期保存阶段主要完成历史数据的归档,通过采用持久的存储设备,实现海量历史数据的长期保存.整个流程如图 2 所示.

在如上不同阶段中,对科学数据的操作方式具有不同的特征,如表 1 所示.

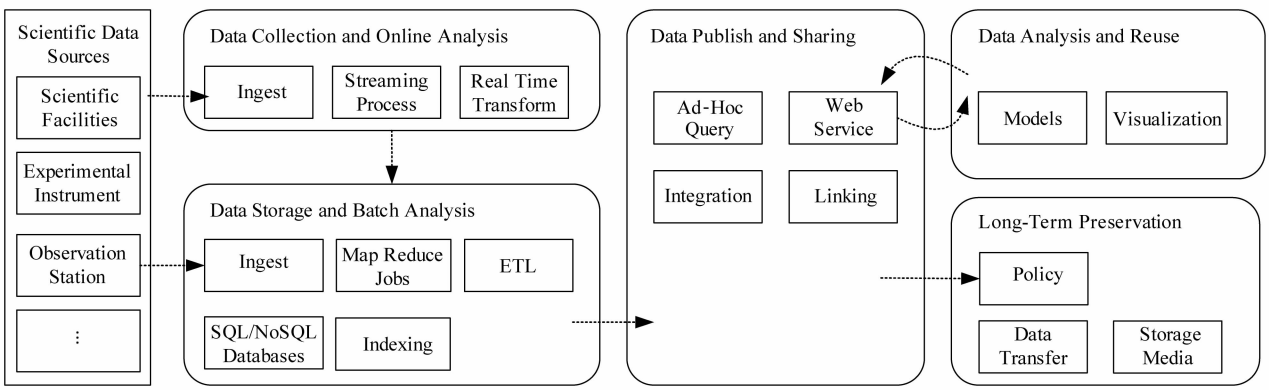

Fig. 2 Full life cycle of scientific big data

图 2 科学大数据全生命周期

Table 1 Stages of Full Life Cycle Management of Scientific Big Data

表 1 科学大数据全生命周期的不同过程

Stages	Data Operation	Data Analysis	Details
Collection & Real Time Analysis	Fast Insert & Online Analysis	RTAP	Collection, Stream Processing, Real Time Extracting and Transforming
Storing & Processing	Offline Analysis(SCAN) of Large Scale Data	OLAP	Database, Index, Collection, Batch Analysis, ETL
Publish & Sharing	Fast Query	OLTP	Linking, Integration, Online Query, Web Service
Reanalysis & Reuse	Online Visual Analysis	OLTP OLAP	Online Analysis Model, Visualization
Archiving & Long Term Storage	Reliable Storage	Analysis is not involved	Archiving strategy, media, data copy and migration

1.2.3 流水线处理特征

科学大数据具有“流水线处理和应用”的特征.以 GWAC(The ground-based wide-angle camera array)为例,GWAC 是中法合作伽玛暴探测天文卫星 SVOM 的关键地面设备,一个 GWAC 相机每 15 s 产生一个大小为 32 MB 的天区图,图像的点源提取

和接下来的光变曲线处理流程应该在一帧的15 s内快速处理完.这个实时处理约束是由于很多短时标的光变,例如微引力透镜事件,需要通过对光变曲线数据实时分析才能得以发现.这个过程就是一个典型的数据流水线,包括天区图采集、图像处理、点源提取、交叉证认、光变曲线处理等步骤[6],如图3所示.为了满足特定的科学目标,科学数据流水线一般对数据处理的精度或者对数据处理的速度等方面会有明确而苛刻的要求,从而为预期的科学目标或者科学发现提供保证.

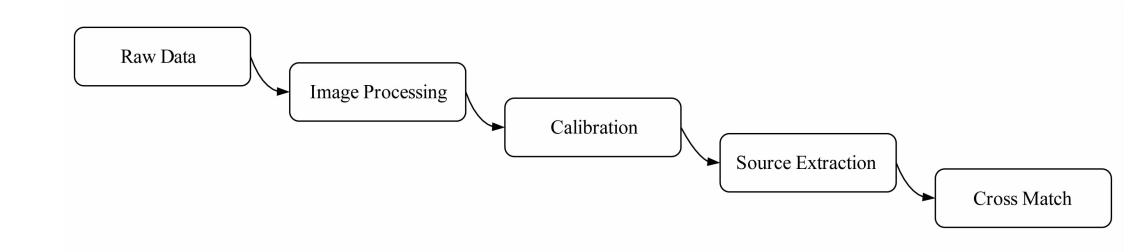

Fig. 3 Collecting and analyzing pipeline of astronomic data

图3 天文数据采集分析流水线

科学数据流水线具有如下特点:

1) 一条流水线通常会涉及到科学数据采集、存储、分析等不同环节.如:从GWAC望远镜获取到天区图,就是一个海量数据采集步骤,图像处理则是一个数据分析的过程.因此,除了需要提供数据分析的支持,还需要考虑到数据的采集等管理功能的支持.

2) 一条流水线会涉及到多元的大数据管理与处理系统.为了达到高效的科学发现目标,往往需要组合不同的数据管理系统,如高吞吐的消息队列系统、高效交互式查询的SQL数据库系统、高可靠的HDFS管理系统等.同时,根据任务的不同特征,也会组合用到不同时效性要求的计算框架,如流计算框架、实时计算框架、离线计算框架等.

3) 完整的科学发现过程往往需要多个流水线并行执行,因此需要考虑CPU/GPU、内存、存储等资源的共享和分配问题.

2 发展现状

科学大数据已成为科学发现的新型战略资源,为了抢占科技竞争的至高点,世界各国已纷纷把科学大数据纳入国家战略,并开始重点部署.美国国立卫生研究院2013年启动了"从大数据到知识"(BD2K)计划[7],总投资达到6.56亿美元.欧盟"地平线2020"计划[8]将科学大数据关键技术和基础设施列为了重点支持领域.欧盟宣布,将投资65亿欧元用于建设"欧洲开放科学云"(Europe Open Science Cloud)[9],重点支持大数据驱动的科学发现.在我国发布的《促进大数据发展行动纲要》中,首次将科学大数据上升到国家战略层面,明确提出"发展科学大数据"的战略目标.中国科学院在"十三五"信息化发展规划中,也明确提出将实施科学大数据工程,全面提升大数据驱动的科技创新能力.

2.1 国际科学大数据资源发展趋势

大规模巡天望远镜、大型粒子加速器、高通量基因测序仪等大科学装置,使得科学大数据呈几何级数增长态势.在天文学领域,人类正在设计和制造各种大型巡天望远镜,试图实现对宇宙多波段、多时域等数字化全覆盖,实现其"虚拟天文台"的伟大构想.如斯隆数字巡天(SDSS)、"泛星计划"(Pan-STARRS)、大型巡天望远镜LSST(Large Synoptic Survey Telescope)等[10].LSST[11]将每3天完成对南半球的天空巡天1次,每15 s记录3幅10亿像素图像(每幅图像包含百万个天体),每晚需对30 TB原始数据准实时的分析,同时对转瞬即逝的千万级突发天体事件,需在60 s之内完成数据分析、插入和分发,并向全世界发出预警[12-13].

在生命科学领域,第二代测序技术使得基因组数据发生了爆炸式的增长.相比于2000年,2010年的基因组数据产量增大了8个数量级.仅华大基因这一个基因组研究机构每天就产生约15 TB数据[14].世界著名的三大基因序列数据库GenBank,EMBL,DDBJ收录了70 000多种生物的核苷酸序列[15-16],其数据量以指数形式增长,核酸碱基数目大概每14个月就翻一倍.再以脑科学为例,用电子显微镜重建大脑中的突触网络,1 mm^3 大脑的图像数据就超过了1 PB[17].

在高能物理领域,位于欧洲核子研究组织CERN的大型强子对撞器LHC每年将产生有15 PB左右的原始数据,利用原始数据进行事例重建以及物理分析所产生的数据规模更大.以其中的ATLAS实验[18]为例,仅2011年产生的总数据就达40 PB.

在对地观测领域,刚刚退役的Landsat 5[19]卫星保持在每天67 GB的观测数据获取量,而2012年

发射的 ZY3 卫星,每天的观测数据获取量可以达到 10 TB 以上,类似能力的传感器现已大量部署在卫星、飞机等飞行平台上,未来 10 年全球部署的对地观测平台的数据获取能力将超过 10 PB/d.

各个不同的领域都在讲述着一个类似的故事,那就是爆炸式增长的数据.这种增长超出了我们创造机器和软件工具的速度,甚至超出了我们的想象.

2.2 我国科学大数据资源现状

我国从 20 世纪 80 年代就持续进行数据资源的积累.1982 年,中国科学院正式提出科学数据库及其应用系统建设项目.经过 30 余年的持续发展,截止十二五"科技数据资源整合与共享工程"项目验收[20],该项目系统地整合了 58 家单位的 1 340 个科学数据库,数据下载量累计达 175 TB.国家科技基础条件平台持续资助了林业科学数据平台、地球系统科学数据共享平台、人口与健康科学数据共享平台、农业科学数据共享中心、地震科学数据共享中心、气象科学数据共享中心等.

以中国科学院为例,中国科学院在生命与健康领域、地球与空间领域、基础与前沿领域积累了丰富的数据资源.其中生物多样性与生物资源数据比较完善,已建成 3 类资源体系:生物多样性与生物资源、组学、医药与健康.在地球与空间领域已建成的数据资源体系包括:固体地球、陆地表层和空间天文等.其中陆地表层又分为地形地貌、气象、水文、生态、自然资源、海洋等内容.此外,在基础与前沿领域已建成的数据资源体系包括物理、化学、能源、材料、脑科学、信息科学等.各领域积累的数据资源如表 2 所示:

Table 2 Typical Scientific Data Resources

表 2 部分典型科学数据资源

Resource Categories		Data Contents	Data Size
Life & Health Area	Biodiversity & Biological Resources	Animal, plant, microorganisms, virus, etc.	2 085 TB
	omics	Genomics, transcriptome, proteome, metabolome, etc.	23 632 TB
Earth and Space Areas	Solid Earth	geomagnetic field data, station observation data, satellite observation data, GPS data, GNSS observation data, superconducting gravimeter data, earthquake observation data, etc.	133 TB
	Meteorological Data	Observation data, reanalysis data, simulated and predicted data, institute internal data, etc.	130 TB, 45 000 Records
	Spatial Data	Spatial scientific satellite data, meridian project data, international meridian data, double star data, space radiation environment, international SPIDR mirror data, WDC exchange data, earth observation data, space environment detection data, spatial astronomical data, spatial scientific experimental data, etc.	145 TB
	Astronomical Data	Image data, radio data, spectrum data, radio spectrometer, numerical simulation data, solar radio spectrometer, measured data, monitoring data, meteorological data, satellite telemetry data, earthquake emergency spatial data, remote sensing data, etc.	502 TB
Foundation & Frontier Areas	Chemical Data	Analytical chemistry data, spectrum data, applied chemistry data, natural products and medicine chemistry data, environmental chemistry data, organic chemistry data, physical chemistry data, etc.	50 GB, 17.38 million Records
	Energy Data	Chromatogram data, component screening data, standard natural product chemistry data, alga fatty acid data, coal fundamental property data, coal-derived oil & chemistry data, coal gasification development & engineering data, carbon fiber research engineering data, pollutant & C02 emission data, green chemical process data, etc.	900 000 Records
	Material Data	Ancient Chinese data, inorganic non-metallic materials data, phase portrait data, superalloy data, titanium alloy data, aluminum alloy data, magnesium alloy data, copper alloy data, precision tube, welding materials data, material corrosion & invalidation data, nanophase reference materials/certified reference material data, formally published nanophase reference data, project information of nanophase major research plan, nanophase material data, etc.	360 000 Records
	Brain Science Data	Brain image data, large-scale brain knowledge base, multimode schizophrenia image data, etc.	5 TB
	Information Science Data	Multimode biometrics(iris, fingerprint, human face, palm print, gait, etc.), tumor data about national image omics, multilingual parallel corpora, speciation analysis corpora, word alignment corpora, language corpora, handwritten digit recognition data, etc.	10 TB

2.3 科学大数据管理系统

针对科学数据,不同科研机构相继研发了科学数据管理系统,包括 SRB[21-22],iRODS[23-25],SciDB[26-35],Hama[36-39],SkyServer[40-41]等.美国圣地亚哥超算中心(San Diego Supercomputer Center, SDSC)为了解决复杂海量科学数据的方便、高效、透明、统一的数据管理和访问,研发了融合资源保存代理(storage resource broker, SRB)系统,在数据网格、数字图书馆、永久保存和实时数据系统中得到了较好的应用,并继而推出了开源分布式数据管理系统 iRODS (Integrated Rule-Oriented Data System).结合科学研究所产生的数据特征,Stonebraker 等人在列存储的基础上,研发了一套开源的数据管理系统 SciDB. SciDB 不同于传统的关系数据库管理系统,它是一个数据管理和分析软件系统,侧重于科学数据的分析操作,设计目标是与 R,MATLAB 以及 IDL 等科学分析软件结合来分析管理科学数据. Hama 作为 Hadoop 项目的大规模计算子项目,利用 Hadoop 强大的分布式存储与处理性能,针对部分科学问题的计算提供基于整体同步并行计算(bulk synchronous parallel, BSP)模型及 graph 模型的计算框架.针对 SDSS 的数据,Gray 主导研发了 SkyServer 天文数据管理系统,实现 TB 量级天文数据的管理与探索.由于现有数据库管理在处理像 LSST 这样 100~200 PB 量级的数据时依然显得力不从心,因此 LSST 启动研发了可管理百亿级天文对象的数据库 Qserv[42],借助多数据中心、大规模分布式并行数据库等技术,实现更加强大的数据管理、访问和共享的能力.

这些科学数据管理系统在功能、原理以及特色上的差异如表 3 所示,可以看出,目前的科学数据管理系统仅关注科学数据全生命周期的某个环节,还缺乏面向大数据的、涉及全生命周期的、与分析系统紧密集成的科学数据管理系统.

表 3 现有科学数据管理系统对比分析

Table 3 Comparative Analysis of Existing Scientific Data Management Systems

Systems	Functions	Principles	Characteristics
iRODS	The unified access interface of multiple hybrid storage such as file system, archive system, and database	Rule based data grid middleware	Completing a series of complex and linked tasks via rules
SciDB	Vector/matrix based management and analysis	Array data model	Imperfect in application, but valuable to the research and development of the scientific data management and analysis
Hama	Large-scale scientific computing, especially the matrix and graph computing	BSP & Graph based Model	Focusing on matrix and graph computing
SkyServer	Manage TB level astronomical data	Columnar Database, MonetDB	With flexible SQL query and direct visual interface, but incapable to support the LSST
Qserv	Ten billion level astronomical objects	Based on ZONE	Relying on technologies such as multiple data center, large-scale distributed parallel database, etc.

我国在科学数据管理技术与平台软件方面也有一些工作在展开,典型的如中国科学院通过信息化专项项目在“十二五”期间率先建成了“科学数据云”,形成了 52 PB 云存储和上万个虚拟机的云计算环境,研发部署了科学数据管理软件 TeamDR、数据发布与集成软件 VisualDB/VDBCloud[43-44]、数据服务注册系统 RSR、可视化服务平台 DVIZ[45]等 20 余项软件工具.

面对源源不断快速产生的大量数据文件以及从中分析生成的千亿级科学对象的管理,我们还面临着一系列的挑战,包括 EB 级文件和万亿行关系数据的高效率、低成本、一体化存储和管理,科学大数据快速索引,以支持大规模、交互式的查询和处理;海量多源、多学科数据的自动关联与融合;瞬时产生的海量数据实时或准实时的高效分析;以流水线的方式实现海量数据资源与科学模型的快速融合与并行处理等.

3 全域科学大数据管理系统框架

科学大数据数据管理的目的是最大限度提高科学发现的速度和能力,因此管理必须与科学发现的过程有机融合,要实现科学数据的采集、存储、分析处理、发布与关联融合、归档等全域管理,支持数据按需快速流动,支持各种类型的科学数据流水线的动态集成与调度.此外,要充分考虑到科学数据类型

多样性,应用需求多样性和计算框架的多样性,能以开放架构实现系统的按需扩展和动态演进.

为此,本文提出全域科学大数据管理框架,具体如图 4 所示:

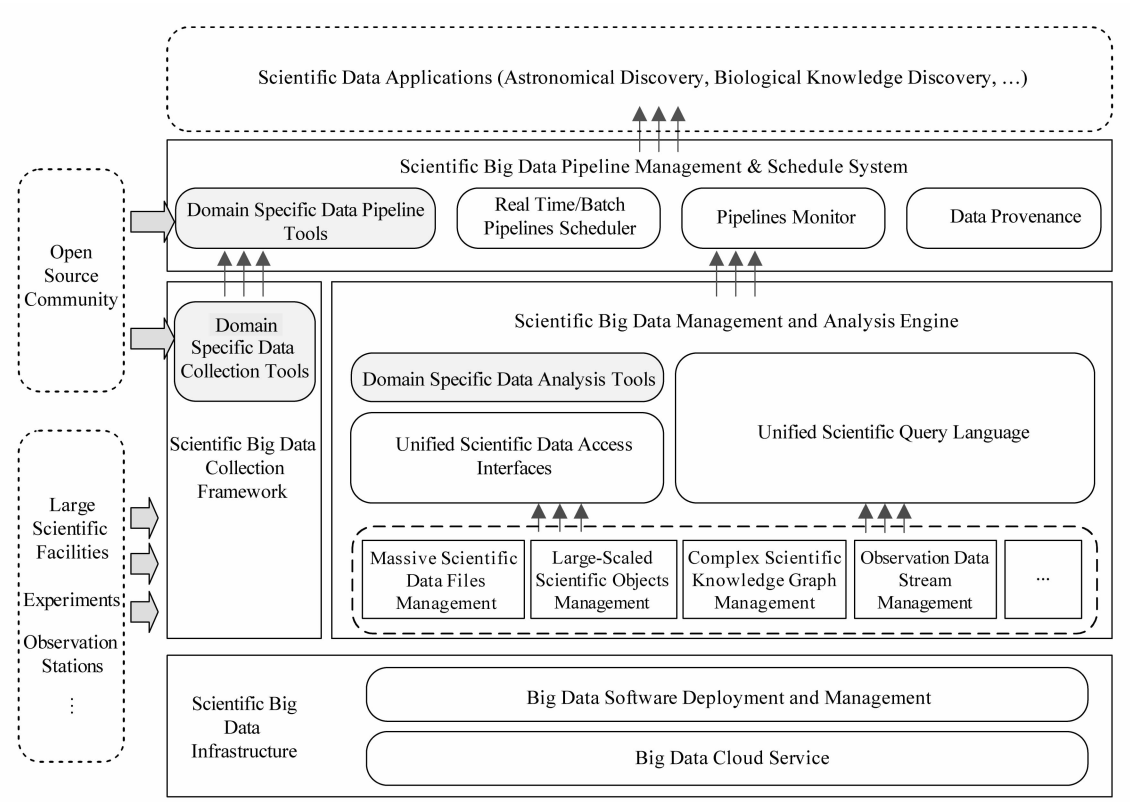

Fig. 4 Scientific big data management system

图 4 科学大数据管理系统

主要组成部件包括科学大数据基础平台、科学大数据统一采集与汇聚框架、科学大数据管理与分析引擎、科学大数据流水线管理与调度系统,以及科学大数据应用环境.

科学大数据基础平台旨在构建大数据存储与计算的云服务平台,对存储和计算资源进行管理及优化,提供基础的大数据存储和并行计算能力.同时配置大数据软件部署与管理工具,实现 Impala[46],HBase[47],Solr[48],TITAN[49],Cassandra[50]等大数据集群的按需部署与配置化管理,实现集群的横向扩展,并通过提供运行监控界面,实现资源状态可视化和及时告警.

科学大数据统一采集与汇聚框架是一个可扩展的、高容错的、高吞吐量的科学大数据采集框架,实现科学大装置、实验观测、台站网络等各类科学数据的统一接入,同时提供包括 morphine 转换、正则转换、模板转换等灵活的数据转换能力.针对各领域科学数据的采集的不同需求,提供个性化的学科领域大数据采集软件,如天文巡天图像数据采集、实验观测数据采集、台站网络观测数据采集等.

科学大数据管理与分析引擎旨在支持海量分布式科学数据文件的索引和管理、万亿级事例数据的在线查询与提取、高吞吐的观测/实验流数据的在线分析与管理,以及大规模关联图的管理与分析计算.该引擎通过统一的查询语言,实现对多元数据管理模型的统一访问和查询,包括对关系型数据、图数据、Key/Value 数据、列数据,以及文件系统的查询.同时通过统一的编程式查询分析一体化操作语言,实现对科学数据的大批量写入与分析,通过函数式编程语言的特性,支持用户在操作语句中自定义数据的转换和分析算法.

科学大数据流水线管理与调度系统通过对数据的采集、存储、查询和分析过程的封装,形成科学大数据流水线的软件表达模型.通过流水线管理模块,实现各领域数据流水线的统一集成管理.同时,基于大数据计算环境,实现数据流水线任务的转换和运行调度,支持数据流水线任务的启停、再放与回溯.针对各领域科学数据的分析处理的不同特征,集成个性化的学科领域大数据流水线处理软件,如天体交叉证认流水线、生物信息关联发现流水线、高能物理事件抽取流水线等.

4 关键技术

针对科学大数据的管理需求与特点,我们可将其涉及到的关键技术归纳为:科学大数据的融合、实时分析、长期存储、云服务技术、开放共享机制等.

4.1 科学大数据融合

大数据时代人们面临的最根本挑战是从数据中凝练可领悟的知识[51-52].大数据融合的概念[53]是指聚合数据间、信息间、知识片断间多维度、多粒度的关联关系实现更多层面的知识交互,已广泛应用于各个领域.比如商业领域中 IBM Watson[54]利用大数据融合的关键技术辅助认知商业发展;生命科学利用 Bio2RDF[55],Neurocommons[56]等知识图谱做问答和决策等.

大数据融合不同于传统数据库领域的数据集成技术[57-59],也不同于传统人工智能与认知科学中的知识融合技术[60-61].数据融合需要用动态的方式统一不同的数据源,将离散的数据转化为统一的知识资源.知识融合是将数据融合阶段获得的笼统的知识转化为可领悟知识,面向需求提供知识服务.它需要挖掘隐含知识,寻找潜在知识关联,进而实现知识的深层次理解,以便更好地解释数据.

以微生物领域为例,比如后基因组时代的系统生物学把生物系统内不同性质的构成要素以及系统内各个不同层次整合在一起进行研究[62].那么首先势必要将基因、mRNA、蛋白质、生物小分子,以及从基因到细胞、到组织、再到每个水平的有机体等不同来源的数据进行融合.这个过程分 4 步完成:

1) 需要从不同数据源(如 Taxonomy,Genbank,Gene,UniProt,PDB,KEGG,Pfam,GO 等)抽取相关的实体和关系,或者从现存知识库(如 Neurocommons,Bio2RDF)中直接转化数据,这一过程中,随着数据体量、种类、来源等动态变化,需要对构建的知识库进行动态更新;

2) 识别出相同实体,并进行实体链接,比如识别出 Bt 蛋白与苏云金杆菌蛋白是同一个蛋白,并且它们与知识库中的实体 Bt 蛋白进行链接;

3) 在进行实体关联时可能会存在歧义、冲突的情况,比如 BT 既可以表示苏云金杆菌,也可以表示蚂蚁磁力链接搜索引擎,这就需要冲突解决技术消除歧义;

4) Bt 蛋白属于晶体蛋白,如果我们为 Bt 蛋白构建了本体——晶体蛋白,那么也可以加速融合的效率,比如中国科学院微生物研究所构建 Speices taxonomy, Protein (uniprot), Gene, Pathway (Kegg), Genome, Enzyme Reaction Data(Kegg)六个本体用于促进生物大数据的融合.

经过上述数据融合,我们仅仅使碎片化的数据相联系、将分散的数据相集中,形成表层知识,即微生物知识资源;但是为了更好地探究生物数据之间繁杂的逻辑关系和特征,就要使隐性知识显性化,使表层知识上升为普适机理.这个过程分 4 步完成:

1) 根据数据的分布规律归纳出数据的结构规则进而抽象出数据之间的关联模式来表示知识,即要对微生物知识进行抽象与建模,比如把"苏云金杆菌是产生 Bt 蛋白质的土壤细菌"这一知识用 RDF 三元组〈苏云金杆菌,产生,Bt 蛋白质〉和〈苏云金杆菌,属于,土壤细菌〉表示或者用低维向量的形式表示.

2) 通过关系推演技术显性化隐性知识,比如中科院微生物研究所融合了 36 个不同的数据源约 830 万个数据,从约 4 000 万个显示关联关系中推演得到约 1.4 亿个隐式关联关系.

3) 除了隐性知识,还有更重要的深度知识,包括高阶多元关系和隐含语义关系,比如鱼类中的掠食者在食物富集时运动轨迹呈布朗运动,微生物菌群共生体系中可能存在基因共振现象,而单个培养的微生物中没有共振现象[63].这种知识一般需要通过领域理论,运用数学、物理等工具,进行理论建模、解析、逻辑演绎、公式推演和证明获得,如采用统计分析和深度学习的方法.

4) 人的智力能透过现象看到本质,只有发现大数据所呈现出的普遍现象背后的普适原理才能对客观世界产生更大的影响.比如,社会网络中社群的消失现象,他们背后的普适原理是生物进化论[64];增长和择优机制在复杂网络自组织演化中具有普遍性,它们使网络在宏观上具有幂律度分布的普适现象[65].这就搭建起了庞大复杂的人类社会与渺小精细的微生物群落之间的关联.

从上述案例我们也可以看出,微生物大数据融合的数据融合用于"喂饱"人类对微生物知识的需求,而知识融合"反哺"生态系统的和谐发展.二者相互协调启发才能最大限度地提升微生物大数据的价值.

4.2 科学大数据实时分析

科学领域已进入一个信息丰富的大数据时代,数据量正以 TB 级甚至 PB 级的速度增长.科学大数据的分析正在从传统的批量处理向实时分析快速发展.

以天文领域 GWAC 全天短时标观测系统为例,整个天区由 40 个 GWAC 相机阵同时监控,一个 GWAC 相机每 15 s 产生一个大约 32 MB 的天区图,通过点源提取该天区图将生成 1.7×106 条星表记录.每副图片的点源提取和星表记录与模板表的交叉证认时间之和需小于 15 s 的延迟,这是一个典型的实时分析的应用场景.

天文大数据具有产生速度快、数据量大、周期时间长等特点,需要设计可快速入库的缓存机制或消息队列,提高数据的存储能力和消息队列的吞吐率.并采用分布式多级缓存机制或可扩展的消息队列实现科学数据的快速存储和传输.

为满足高速数据采集下的实时分析,一般分为针对批量外存数据的大规模并行处理(massively parallel processing, MPP)技术和基于流式内存数据的数据流查询处理技术.为便于快速查询和实时分析内外存数据,可设计同时进行批量处理和流式处理的查询适配器,通过统一的查询接口实现不同数据类型的全量查询结果.

此外,随着数据量的累积和维度的增加,以及查询和分析复杂度的不断增长,实时返回用户查询结果越来越成为科学大数据系统的一个重要挑战.目前,学术界和工业界的一个研究重点就是如何在科学大数据系统中支持交互式的数据查询.这里的交互性体现在处理用户查询过程中系统及时不断地提供反馈,这样使得用户能够快速地做出反应和根据反馈结果更改或优化下一步的查询条件,以找到最相关和最有意义的查询结果.因此,交互性查询分析也是实时分析的一个重要研究方向.

4.3 科学大数据长期存储

现代科学大数据需要花费成百上千万美元产生数据,通常会积累几年到十几年的数据,这些数据该如何有效地保存和利用一致是科学数据面临的重大问题.大数据时代数据产生的速度更快,产生的量更大,如何长期存储这些数据并提供高效的处理,或者说如何决定保存哪些数据淘汰哪些数据成为了当务之急.

以 GWAC 为例,根据天文数据的独特要求,为了满足对短期数据的快速实时查询以及对数据的长期存储,设计使用了正三角和倒三角模型对数据进行处理分析(如图 5 所示).在数据的底层存储中,通过使用 HDFS 对数据按照文件的方式进行存储.将每一个星的数据保存成一个文件.单个星的文件随着时间的积累不断增加,而文件总数却始终保持在百万级.而 HDFS 面对海量小文件时的处理应对能力较弱,因此我们使用三角模型对数据进行处理.随着时间的增加,将海量小文件逐步合并,越久远的数据合并率越高,而近期的数据则保持不变,不进行合并.同时,随着文件的合并,文件大小也会有所变化,当久远的数据合并后,单一文件大小会不断增大.通过这样的方法,在文件个数和文件大小之间寻找平衡以满足对数据的有效管理.

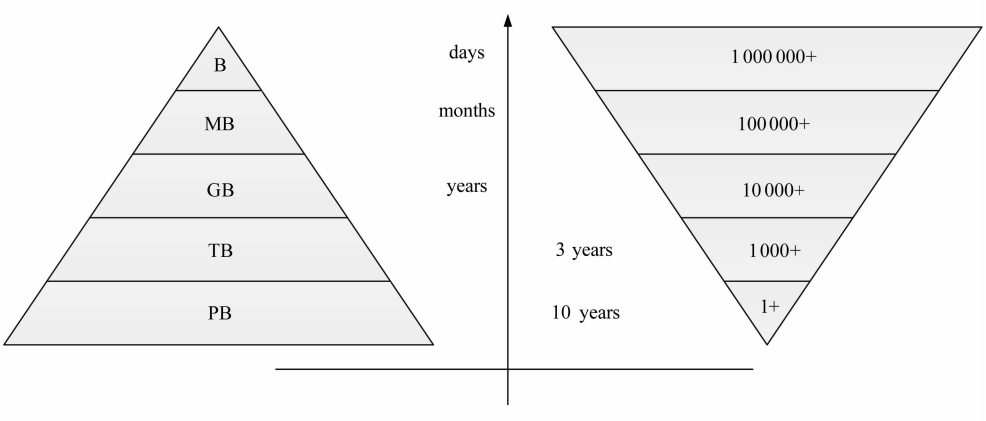

Fig. 5 Counts and sizes of long term stored scientific data

图 5 科学大数据长期存储文件数目与大小

总之,长期存储系统的目标可以归为 3 个:1)设计一个简单一致的解决方案,计算与存储资源混合在同一节点上,使其具备独立运行能力;2)完成可扩张的和轻便的设计,以便能够将所有设计布置到位于全球任何地方的合作单位;3)集安全性和适应性于一体,对于磁盘或结点丢失应具有健壮性,所有后备成员具备完全独立性.

4.4 科学大数据云服务技术

随着云服务提供给大数据管理和分析的质量得到不断提高,云服务的多样性也在稳步增长.科学大

数据的管理与分析正好可以借助云服务的进步来更好地为科学研究提供助力,将计算资源和数据资源合理高效地整合到云端,更好地为科技工作者提供服务和帮助.

科学大数据云存储服务不同于普通云存储,其主要是面向大数据分析的超大规模存储库,一般要求能存储非结构化和结构化数据且能提高分析性能的大吞吐量.由于受到传统分析体系结构(例如架构的预定义)的限制,需要事先定义数据模式.为应对这一挑战,引入数据湖概念,将它作为存储在单一位置收集的各种类型数据的企业级存储库.出于科学探索分析目的,可在定义架构之前,所有类型的数据都可以存储在数据湖中.因此当面对某种分析时,动态的创建数据模式是未来的主要挑战.

科学领域的数据分析往往需要深度定制自身分析流程.科研人员需要调用基本 API 编程,但不同系统的 API 差异很大(如 Spark 和 Hadoop 的编程接口差异),导致程序移植性差.因此云分析服务面临的基本挑战是解耦底层数据分析系统和分析 API 直接的联系,从而实现同样的分析程序可在不同的大数据系统之间轻松移植,从而减轻科研人员的工作压力.

4.5 科学大数据开放共享

数据只有在不断的使用中才能产生价值,而且,数据资源天然具有可重复使用的特征.开放科学数据可确保科学研究结论的真实性和可重现性,可确保公共财政投入获取的公共资源,能最大限度地产生价值,可支持数据跨领域、跨学科的融合和重复使用,从而加快科学发现的进程.世界经济合作与发展组织 OECD 提出了科学数据开放的基本原则[66],Force11 明确了有效开放的“FAIR”(findable, accessible, interoperable, reusable)标准[67-68],国际科学联合会 ICSU 发布了“大数据时代开放数据公约”,明确了在数据开放过程中各利益相关方的责任.

科学大数据开放共享一个基本的共识是,研究项目及其相关数据收集完成时,公共财政支持产生的数据应可公开访问及最大限度地再利用.如生命科学领域的基因序列数据库 GenBank,通过和国际著名学术期刊合作,强制要求学术论文作者在提交论文时,必须先将数据提交到 GenBank 数据库中,为全人类积累了一个庞大的基因序列数据库.再如,Sloan 数字巡天项目 SDSS,已经先后向全世界发布了 13 版的巡天数据.

概括起来,科学大数据开放共享方式主要包括:

1) 通过国际合作项目或合作网络驱动的开放共享,典型的如 GEOSS[69],GBIF[70], WDCM[71]等.这种方式要求所有参与者按照大家共同认可的规则开放数据和使用数据.

2) 通过学术期刊驱动的开放共享,典型的如基因序列数据库 EMBL/Genbank/DDBJ.

3) 通过公共存储库和公共服务驱动的开放共享,典型的如 SDSS,Dryad[72],Fig share[73]等.这种方式通过建立一个领域内或者跨领域的公共数据库或公共数据存储平台,以服务的方式来汇聚和开放数据资源.

4) 数据出版和引用机制,典型的如 Nature 旗下的 Scientific Data[74]、ESSD[75]、《中国科学数据》[76]等,通过数据论文的发表和引用来激励科研人员开放数据,并提高数据的可理解性和可重用性.此外,以数据交易的形式提供服务的数据集市机制,也开始有一些尝试,但是其是否适合于科学数据,还有待进一步观察.

在科学数据的开放共享中,不同的学科、不同的数据、不同的组织乃至国家,其采用的机制、模式等可能均不同,不能一概而论,也不存在“One size fits all”的解决方案.但在任何一种机制的设计中,必须首先明确参与数据开放共享的各相关方的利益和诉求,要通过建立有效的激励机制、利益分配机制和评估评价机制等来有序推进,而且其中数据权属的问题、隐私问题、安全问题,也不可忽视.

5 总结和展望

大数据时代科学研究是一个大科学、大需求、大数据、大计算、大发现的过程.数据密集型科学发现已经成为继实验科学、理论推演、计算机仿真这 3 种科研范式相辅相成的科学研究第四范式.先进的科学大数据管理和处理可以为各学科领域的新发现提供坚实的技术基础,能够加速具有国际影响力的科技成果的产出过程,具有重要的科学价值.

然而,为了更好地促进科学研究,科学大数据的管理还存在着较大的技术挑战,包括 EB 级文件和千亿行关系数据的高效率、低成本、一体化存储和管理,科学大数据快速索引,以支持大规模、交互式的查询和处理;海量多源、多学科数据的自动关联与融合;瞬时产生的海量数据实时或准实时的高效分析;以流水线的方式实现海量数据资源与科学模型的快速融合与并行处理等.

为此,我们需突破科学大数据管理与分析的关键问题,研发一体化全流程科学大数据管理系统,成为大数据时代重大科技创新活动必要的“使能利器”,也成为广大科研人员“军械库”中的“杀手锏”,帮助他们从大数据中高效、快速地发现新知识,取得新的突破.

参 考 文 献

[1] Greenberg J. Big data, little data, no data: Scholarship in the networked world [J]. Leonardo, 2016, 49(1): 91-92

[2] Barwick H. The “four Vs” of Big Data, Implementing Information Infrastructure Symposium [EB/OL]. North Sydney NSW: IDG Communications Pty Ltd. (2012-10-02) [2016-10-10]. http://www.computerworld.com.au/article/396198/iiis_four_vs_big_data/

[3] IBM. What is big data? [EB/OL]. Armonk, NY: IBM Corporation. (2012-10-02) [2016-10-12]. http://www-01.ibm.com/software/data/bigdata/

[4] Wikimedia. Big data [EB/OL]. 2016 [2016-10-02]. http://en.wikipedia.org/wiki/Big_data

[5] Kaufman Y J, Justice C, Flynn L, et al. Monitoring global fires from EOS-MODIS [J]. Journal of Geo-Physical Research, 1998, 103(D24): 32215-32238

[6] Wan Meng, Wu Chao, Wang Jing, et al. Column store for GWAC: A high-cadence, high-density, large-scale astronomical light curve pipeline and distributed shared-nothing database [J]. Publications of the Astronomical Society of the Pacific, 2016, 128(969): 114501-114516

[7] Bourne P E, Bonazzi V, Dunn M, et al. The NIH big data to knowledge (BD2K) initiative [J]. Journal of the American Medical Informatics Association, 2015, 22(6): 1114-1114

[8] Chen Guangren, Zhu Yu, Su Qing. Science programs lead to the future [J]. Science Technology Review, 2014, 32(31): 15-28 (in Chinese)
(陈广仁, 朱宇, 苏青. 引领未来的科学计划[J]. 科技导报, 2014, 32(31): 15-28)

[9] Jones B. Towards the open European science cloud [C] // Digital Era Forum. Zenodo, 2015: 1-21

[10] Lupton R, Gunn J E, Ivezić Z, et al. The SDSS imaging pipelines [C] //Astronomical Data Analysis Software and Systems X. Boston, MA: Astronomical Society of the Pacific, 2001: 269-278

[11] LSST. LSST Public Website Sitemap [OL]. Tucson, AZ. LSST Corporation. [2016-10-02]. http://www.lsst.org/lsst/science/scientist_transient

[12] Ivezic Z, Tyson J A, Abel B, et al. LSST: From science drivers to reference design and anticipated data products [J]. American Astronomical Society, 2008, 41: 366

[13] Becla J, Szalay A, Gray J. Designing a multi-petabyte database for LSST [C] //Proc of SPIE Astronomical Telescopes+ Instrumentatio. Bellingham: WASPIE Publications, 2006: 62700R-62700R

[14] Mao Daowei, Su Xia. The initial progress of the model reforming, the characteristics of the cultivating talents-Students fromthe Beijing Genomics Institute (BGI) frequently publish works in Science and Nature [J]. Guangdong Science & Technology, 2010, 19(11): 15-18 (in Chinese)
(毛道伟, 孙侠. 模式改革初显成效人才培养渐成特色——华工-华大基因组科学创新班学生《Science》、《Nature》频亮相引关注[J]. 广东科技, 2010, 19(11): 15-18)

[15] Brooksbank C, Cameron G, Thornton J. The European Bioinformatics Institute's data resources: Towards systems biology [J]. Nucleic Acids Research, 2005, 33(Suppl 1): 46-53

[16] Rao Dongmei. NCBI data base and its resource access [J]. Science & Technology Vision, 2013 (7): 53-54 (in Chinese)
(饶冬梅. NCBI 数据库及其资源的获取[J]. 科技视界, 2013 (7): 53-54)

[17] Li Guojie. The recognition of big data [J]. Big Data, 2015, 1(1): 1-9 (in Chinese)
(李国杰. 对大数据的再认识[J]. 大数据, 2015, 1(1): 1-9)

[18] Andreeva J, Campana S, Fanzago F, et al. High-energy physics on the grid: The ATLAS and CMS experience [J]. Journal of Grid Computing, 2008, 6(1): 3-13

[19] Chen J, Wang W, Li Z Y, et al. Landsat 5 satellite overview [J]. Remote Sensing Information, 2007, 43(3): 85-89

[20] The results summary of the information special project “integration and share of data resources” in Chinese Academy of Science. Science and Technology Daily [N]. Beijing: Science and Technology Daily Press, 2016-04-05 (in Chinese)
(中科院“十二五”信息化专项科技数据资源整合与共享工程成果概述. 科技日报[N]. 北京: 科技日报社, 2016-04-05)

[21] Moore R, Chen S Y, Schroeder W, et al. Production storage resource broker data grids [C] //Proc of IEEE Int Conf on E-Science & Grid Computing. Los Alamitos, CA: IEEE Computer Society, 2006: 147

[22] Manandhar A, Dam K K V, Berrisford P, et al. Deploying a distributed data storage system for grid applications on the National Grid Service using federated SRB [C] //Proc of the UK e-Science All Hands Meeting. Edinburgh. UK: National e-Science Centre, 2004

[23] Hedges M, Hasan A, Blanke T. Management and preservation of research data with iRODS. [C] //Proc of the 16th ACM Conf on Information and Knowledge Management, Workshop on Cyberinfrastructure: Information Management in Escience (CIMS 2007, CIKM 2007). New York: ACM, 2007: 17-22

[24] Conway M, Moore R, Rajasekar A, et al. Demonstration of policy-guided data preservation using iRODS [C] //Proc of IEEE Int Symp on Policies for Distributed Systems and Networks. Los Alamitos, CA: IEEE Computer Society, 2011: 173-174

[25] Antunes G, Barateiro J. Securing the iRODS metadata catalog for digital preservation [M] //Research and Advanced Technology for Digital Libraries. Berlin: Springer, 2009: 412-415

[26] Cudre-Mauroux P, Kimura H, Lim K T, et al. A demonstration of SciDB: A science-oriented DBMS [J]. VLDB, 2009, 2(2): 1534-1537

[27] Stonebraker M. SciDB: An open-source DBMS for scientific data [J]. ERCIM News, 2012, 89: 13

[28] Stonebraker M, Becla J, Dewitt D J, et al. Requirements for science data bases and SciDB [C] //Proc of the Conf of CIDR. New York: ACM, 2009: 173-184

[29] Hammami R, Zouhir A, Naghmouchi K, et al. SciDBMaker: New software for computer-aided design of specialized biological databases [J]. BMC Bioinformatics, 2008, 9(1): 1-6

[30] Stonebraker M, Brown P, Becla J, et al. SciDB: A database management system for applications with complex analytics [J]. Computing in Science & Engineering, 2013, 15(3): 54-62

[31] Cudré-Mauroux P, Kimura H, Lim K T, et al. A demonstration of SciDB: A science-oriented DBMS [J]. VLDB Endowment, 2009, 2(2): 1534-1537

[32] Stonebraker M, Duggan J, Battle L, et al. SciDB DBMS research at MIT [J]. IEEE Data Engineering Bulletin, 2013, 36(4): 21-30

[33] Paul G Brown. Overview of SciDB: Large scale array storage, processing and analysis [C] //Proc of Conf of SIGMOD. New York: ACM, 2010: 963-968

[34] Stonebraker M, Brown P, Poliakov A, et al. The architecture of SciDB [C] // Proc of Scientific and Statistical Data Management Conf. Berlin: Springer, 2011: 1-16

[35] Becla J, Lim K T. Report from the SciDB workshop [J]. Data Science Journal, 2008, 7: 88-95

[36] Seo S, Yoon E J, Kim J, et al. HAMA: An efficient matrix computation with the MapReduce framework [C] //Proc of Cloud Computing Technology and Science (CloudCom). Piscataway, NJ: IEEE, 2010: 721-726

[37] Luo S, Liu L, Wang H, et al. Implementation of a parallel graph partition algorithm to speed up BSP computing [C] // Proc of Fuzzy Systems and Knowledge Discovery (FSKD). Piscataway, NJ: IEEE, 2014: 740-744

[38] Suchanek F M, Weikum G. Knowledge bases in the age of big data analytics [J]. VLDB Endowment, 2014, 7(13): 1713-1714

[39] Suchanek F, Weikum G. Knowledge harvesting in the big-data era [C] //Proc of ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 933-938

[40] Szalay A S, Gray J, Thakar A R, et al. The SDSS SkyServer, public access to the sloan digital sky server data [C] //Proc of SIGMOD. New York: ACM, 2002: 570-581

[41] Raddick M J, Szalay A S, Gray J N, et al. Two years of SkyServer: Education and outreach with sloan digital sky survey data [J]. Bulletin of the American Astronomical Society, 2003, 35(3): 718

[42] Wang D L, Monkewitz S M, Lim K T, et al. Qserv: A distributed shared-nothing database for the LSST catalog [C] //Proc of High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2011: 1-11

[43] Shen Z, Li J, Li C, et al. VisualDB: Managing and publishing scientific data on the Web [C] //Proc of Int Conf on Cyber-Enabled Distributed Computing and Knowledge Discovery, Cyberc. Piscataway, NJ: IEEE, 2011: 399-404

[44] Huo D M, Li S, Xu C. Service system of the South China Sea science data products based on VisualDB [J]. Journal of Tropical Oceanography, 2012, 31(2): 118-122

[45] Du Yi, Guo Danhuai, Chen Xi, et al. Model-driven visualization generation system [J]. Journal of Software, 2016, 27(5): 1199-1211 (in Chinese)
(杜一, 郭旦怀, 陈昕, 等. 一种模型驱动的可视化生成系统 [J]. 软件学报, 2016, 27(5): 1199-1211)

[46] Taft D K. Cloudera Impala 1.0 Brings SQL to Hadoop for Real-Time Queries [EB/OL]. Foster City, CA: Eweek, (2013-05-12) [2016-10-10]. http://www.eweek.com/database/cloudera-impala-1.0-brings-sql-to-hadoop-for-real-time-queries

[47] Vora M N. Hadoop-HBase for large-scale data [C] //Proc of Int Conf on Computer Science and Network Technology. Piscataway, NJ: IEEE, 2011: 601-605

[48] Abdelouarit K A, Sbihi B, Aknin N. Solr, lucene and Hadoop: Towards a complete solution to improve research in big data environment (Case of the UAE) [C] //Proc of the Mediterranean Congress of Telecommunications. Los Alamitos, CA: IEEE Computer Society, 2016: 363-367

[49] Jouili S, Vansteenberghe V. An empirical comparison of graph databases [C] //Proc of Int Conf on Social Computing. Piscataway, NJ: IEEE, 2013: 708-715

[50] Lakshman A, Malik P. Cassandra: A decentralized structured storage system [J]. AcmSigops Operating Systems Review, 2010, 44(2): 35-40

[51] Suchanek F M, Weikum G. Knowledge bases in the age of big data analytics [J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1713-1714

[52] Suchanek F, Weikum G. Knowledge harvesting in the big-data era [C] //Proc of the 2013 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 933-938

[53] Meng Xiaofeng, Du Zhijuan. Research on the big data fusion: Issues and challenges [J]. Journal of Computer Research and Development, 2016, 53(2): 231-246 (in Chinese)
(孟小峰, 杜治娟. 大数据融合研究: 问题与挑战 [J]. 计算机研究与发展, 2016, 53(2): 231-246)

[54] IBM. Shop hardware, software and services from IBM and our partners [OL]. IBM Watson. 2016 [2016-10-13]. http://www-31.ibm.com/ibm/cn/cognitive/outthink/

[55] Belleau F, Nolin M A, Tourigny N, et al. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems [J]. Journal of Biomedical Informatics, 2008, 41(5): 706-716

[56] Lenzerini M. Data integration: A theoretical perspective [C] //Proc of the 21st ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM, 2002: 233-246

[57] Meng Xiaofeng, Liu Wei, Jiang Fangjiao, et al. Web Data Management Principle and Technology [M]. Beijing: Tsinghua University Press, 2014 (in Chinese)
(孟小峰, 刘伟, 姜芳艽, 等. Web 数据管理:概念与技术 [M]. 北京: 清华大学出版社, 2014)

[58] Dong X L, Srivastava D. Big data integration [C] //Proc of Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2013: 1245-1248

[59] Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion [C] // Proc of SIGKDD. New York: ACM, 2014: 601-610

[60] Jan M. Linked data integration [D]. Progue: Charles University in Prague, 2013

[61] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information (abstract) [C] //Proc of PODS. New York: ACM, 1998: 188

[62] Zieglgänsberger W, Toile T R. The pharmacology of pain signalling [J]. Current Opinion in Neurobiology, 1993, 3 (4): 611-618

[63] Chen Y, Kim J K, Hirning A J, et al. Emergent genetic oscillations in a synthetic microbial consortium [J]. Science, 2015, 349(6251): 986-989

[64] Givan M, Newman M E J. Community structure in social and biological networks [C] //Proc of the National Academy of Sciences of the United States of America. Los Gatos, CA: HighWire Press, 2001: 7821-7826

[65] Barabási A L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286(5439): 509-512

[66] Pilat D, Fukasaku Y. OECD principles and guidelines for access to research data from public funding [J]. Data Science Journal, 2007, 6: OD4-OD11

[67] Wilkinson M D, Dumontier M, Aalbersberg I J J, et al. The FAIR guiding principles for scientific data management and stewardship [J]. Scientific Data, 2016, 3: 1-9

[68] Force11. Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0 [EB/OL]. [2016-09-10]. https://www.force11.org/fairprinciples

[69] Wikimedia. GEO. GEOSS: The Global Earth Observation System of Systems. [EB/OL]. [2016-09-10]. http://www.earthobser vations.org/geoss.shtml

[70] GBIF (Global biodiversity information facility). Free and Open Access to Biodiversity Data | GBIF.org [EB/OL]. [2016-09-10]. http://www.gbif.org/

[71] Sun Q, Li L, Wu L, et al. Web resources for microbial data [J]. Genomics Proteomics Bioinformatics, 2015, 42(1): 69-72

[72] Dryad. Submission integration [EB/OL]. [2016-09-10]. http://datadryad.org/

[73] Hahnel M. Exclusive: Figshare a new open data project that wants to change the future of scholarly publishing [EB/OL]. 2012 [2016-09-10]. https://core.ac.uk/download/pdf/16380431.pdf

[74] Nature. Scientific Data [EB/OL]. [2016-09-10]. http://www.nature.com/sdata/

[75] ESSD. Earth System Science Data [EB/OL]. [2016-09-10]. http://www.earth-system-science-data.net/

[76] CSData. Chinese Science Data [EB/OL]. [2016-09-10]. http://www.csdata.org/ (in Chinese)
(CSData. 中国科学数据(中英文网络版) [EB/OL]. [2016-09-10]. http://www.csdata.org/)

Li Jianhui, born in 1973. PhD, professor. His main research interests include open data policy and practice, large scale distributed data integration and data cloud service, big data management, big data computing and analysis for science discovery.

Shen Zhihong, born in 1977. PhD, professor. His main research interests include scientific data management and integration, linked data and big data management.

Meng Xiaofeng, born in 1964. PhD, professor at Renmin University of China. CCF fellow. His main research interests include data fusion and knowledge fusion, big data management for new hardware, big data real time and interactive analysis, and big data privacy management.