大规模分布式科学数据管理与服务技术架构及系统

刘峰，陈昕，黎建辉，刘昂，韩芳
中国科学院计算机网络信息中心，北京 100190

摘要
随着信息化进程的发展，大规模分布式多源异构科学数据的管理和应用问题凸显，如何有效地实现分布式数据的管理、整合、服务，成为推动科研与应用的共性需求和必要手段。深入分析核心需求和关键问题，对服务体系进行了系统化的顶层设计，提出一套面向大规模分布式科学数据管理与服务的技术架构，从自治管理、整合管理、集成服务3个层级对服务体系进行了组织和规划，并建设了完备的服务平台和软件体系，为科学数据的管理与服务提供了从管理到应用的一体化解决方案。

关键词
科学数据；科学数据管理；资源服务系统；分布式服务；服务框架；技术架构

中图分类号：G311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016062

Large scale distributed scientific data management and service technology framework and system

LIU Feng, CHEN Xin, LI Jianhui, LIU Ang, HAN Fang
Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China

Abstract
With the development of the information process, problems in the management and application of large-scale distributed With three decades of experience in related fields, after studying key issues and core requirements, the systematic top-level design was conducted. The technical architecture for large-scale distributed scientific data in management and services was proposed, and the service system was organized and planned from three levels: autonomy management, integration management and integration services, also a complete service platform and software system were constructed to provide integration solutions for scientific data management and services from management to application.

Key words
scientific data, scientific data management, resource service system, distributed service, service framework, technology framework

1 引言

科学数据是人类社会科技活动所产生的基本数据、资料以及按照不同需求而系统加工的数据产品和相关信息[1]。作为现代科学可持续发展的重要资源，科学数据与科技创新密不可分。科学数据不断积累和发展，逐渐呈现出规模巨大、分布广泛、结构多样的特点，这为科学数据的管理与共享服务带来了巨大的挑战。

近年来，许多政府机构和科研组织从不同层面开展了大量科学数据管理与服务的研究与实践工作[2]。一些发达国家和国际组织建立了国家级科学数据中心群和数据共享服务网络，如国际科学联合会理事会的国际科学技术数据委员会（Committee on Data for Science and Technology，CODATA）①、世界数据中心/系统（World Data Center/System，WDC/WDS）②、地球观测组织（Group on Earth Observations，GEO）③、国际研究数据联盟（Research Data Alliance，RDA）④和全球生物多样性信息网络（Global Biodiversity Information Facility，GBIF）⑤等，美国建立的分布式最活跃数据档案中心群（distributed active archive centers，DAAC）⑥、全球变化主目录（global change master directory，GCMD）⑦、美国国立卫生研究院数据共享库⑧、地球观测数据网（data observation network for earth，DataOne）⑨，欧洲空间信息基础设施⑩等。我国从20世纪80年代起，从多个层面推动了科学数据的管理与共享，启动了科学数据工程、国家科技基础条件平台（National Science and Technology Infrastructure，NSTI）⑪等。

然而，大规模分布式科学数据的管理与服务尚缺乏完整的理论体系和解决方案，这为科学数据更为广泛高效的开放共享带来了障碍。对此，本文提出了一套面向大规模分布式科学数据的管理与服务技术架构，从技术视角对服务体系的整体框架、技术架构和系统设计进行了完整的描述和分析。该体系架构已在多个应用项目中使用，取得了良好的服务效果并具有广泛的适用性。

2 体系框架设计

2.1 体系框架活动分析

大规模分布式科学数据资源管理与服务体系是指面向大规模分布式科学数据，以提供敏捷安全的数据资源共享服务为目标所形成的与之相关的一系列概念、政策、目标、方法、规范、系统等，可实现科学数据的有效标引、发现、共享、服务，推动科学数据的有效管理和敏捷交付。

服务体系的核心目标在于对科学数据资源进行有效的管理与敏捷共享。具体包括如下内容。

（1）使全部的基础性与公共性数据依据可发现、可访问、可理解、可评估、可使用、可治理的原则以服务的形式发布，使其成为科研活动和经济发展的大数据资源与公共资产。其核心目标是实现科学数据的快速有效共享，使数据存得下、取得到、易分享、安全性高。

（2）形成专业化的数据服务，使数据服务与科研活动有机融合、协调发展，创造数据服务可持续发展的新生态。

（3）实现应用驱动的数据服务按需敏捷集成，以充分挖掘数据的潜在应用价值，有力支撑大数据环境下的科研工作。

大规模分布式科学数据管理与服务基

① http://www.codata.org/
② http://www.icsu-wds.org/
③ http://www.earthobservataions.org/
④ http://www.rd-alliance.org/
⑤ http://www.gbif.org/
⑥ http://earthdata.nasa.gov/about/daacs/
⑦ http://gcmd.gsfc.nasa.gov/
⑧ http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
⑨ http://www.dataone.org/
⑩ http://inspire.ec.europa.eu/
⑪ http://www.escience.gov.cn/default.jsp

本需求的概要描述如图1所示。数据生产者在服务系统的协助下，实现数据资源的管理及发布；服务体系对分布的资源进行整合与集成管理，并通过门户服务等形式提交给用户实现数据资源的交付，主要涉及数据管理、服务管理、用户服务等核心环节。

2.2 体系框架设计难点与重点

科学数据的“大规模”主要体现在数据量大、分布广泛、结构多样3个方面，而科学数据的服务又要求快速有效，这对服务体系的设计与实现提出了挑战。

（1）如何快速整合资源

科学数据基本掌握在各科研单位手中，其分布极为分散，这种分布式的存储形式对快速有效地组织和整合形成了障碍。

科学数据的类型和存储形式多种多样，结构化、半结构、非结构化同时存在，例如气象数据、地学数据等都有其独特的数据结构和存储方式，如何将这些多源异构的数据管理和集成起来并提供统一的服务，是服务体系需要解决的重要问题之一。

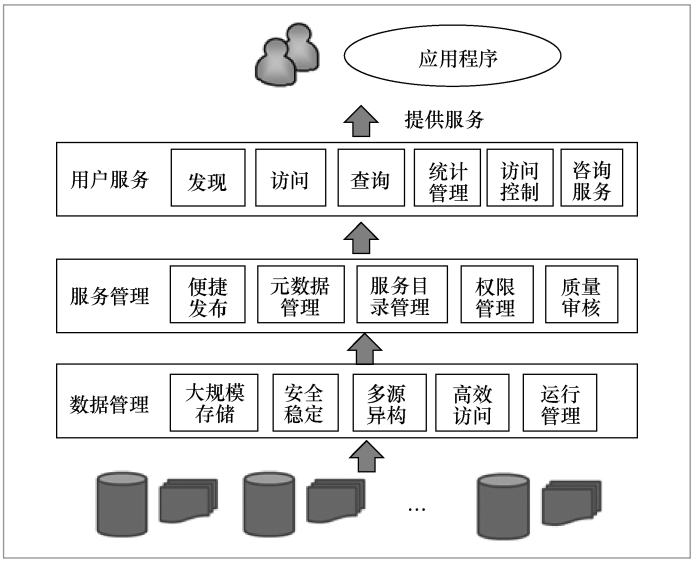


图1 大规模分布式科学数据管理与服务需求概览

同时，科学数据往往作为重要资源掌握在各单位手中，如何获取并促进开放也是推动科学数据开放共享的重要课题，虽然这不是技术层面的问题、不是本文研究的范围，但如何更好地提供服务，使科研人员从分享中获得益处，从而推进共享，也是本文思考的问题之一。

（2）如何提供高质量的数据服务

服务体系的最终目标是为科研人员提供高质量的数据服务，这对数据服务的组织形式、交付方式都提出了较高的要求，因此，好的服务体系设计需要对服务模式、交互方式等有深入的研究和分析。

（3）如何形成可持续的管理与发展

持续化的管理是服务体系长期运行和有效服务的关键环节，与政策等密切相关，但对技术架构也提出了要求，通过技术和服务的设计促进服务体系的有效管理、形成激励等也是本文致力研究的问题之一。

2.3 体系框架分层设计

针对大规模数据资源分散存储与统一服务的总体需求，结合上述体系框架设计的难点与问题的分析，在整个体系框架设计中要求必须采用分层设计的模式，以满足不同层次管理与服务的需求。整体框架分层结构如图2所示。

服务体系框架共分3层，自底向上分别是自治管理层、整合管理层、集成服务层。其中，自治管理层重点实现分布式数据资源自治管理与服务，完成数据资源的本地化集成注册、服务封装及发布管理。整合管理层重点实现数据资源与服务的集中注册、审核与发布管理，进而形成统一的资源服务目录，同时实现对数据资源与服务的监控、统计和评估管理，为分布式数据资

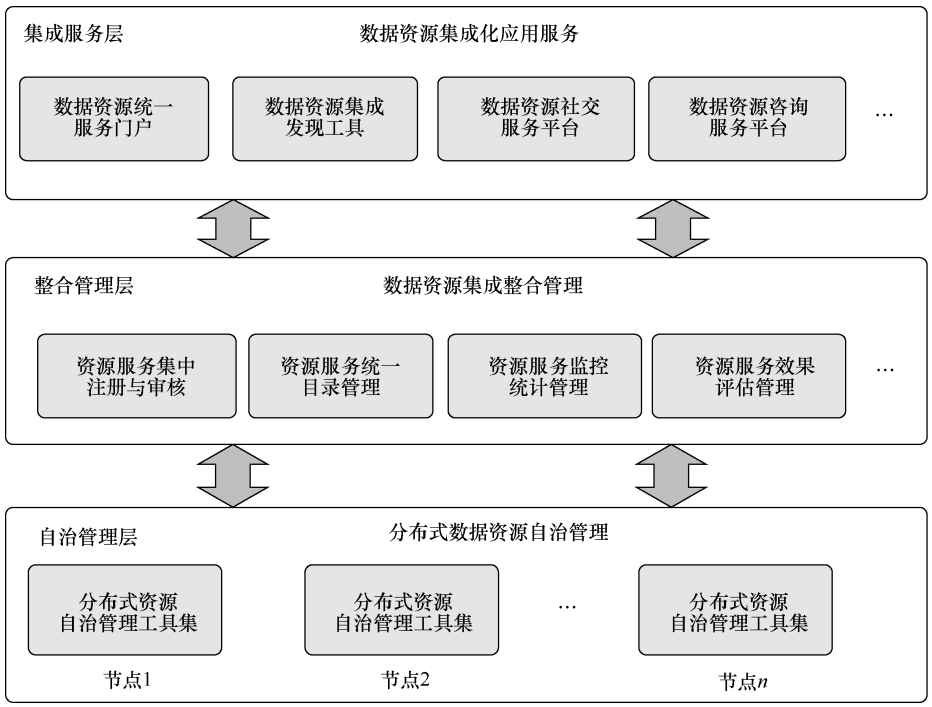


图 2 大规模分布式科学数据管理与服务体系分层框架

源与服务的稳定、优质服务提供支撑和保证。集成服务层是整个体系的对外服务门户，该层重点实现数据资源的目录、发现、访问、获取等公共服务，同时面向最终用户实现以数据资源为中心的集成、交流、共享、咨询方面的服务系统。

3 技术架构设计

3.1 分层技术体系设计

根据第2节体系框架设计，将系统技术架构进行分层设计，如图3所示，共分为基础支撑层、数据管理层、数据服务层、集成服务层。其中，基础支撑层为其他3个层提供公共支撑。

(1) 基础支撑层

基础支撑层主要为整个体系提供公共支撑服务。其中，消息队列技术主要用于系统之间消息传递和任务分发；权限控制技术采用单点登录授权与权限验证相结合的方式，其中登录与授权采用OAuth 2.0协议[3]来实现开放式授权；数据加密模块主要用在重要信息存储和传输，其中在数据传输上采用安全超文本传输协议(hyper

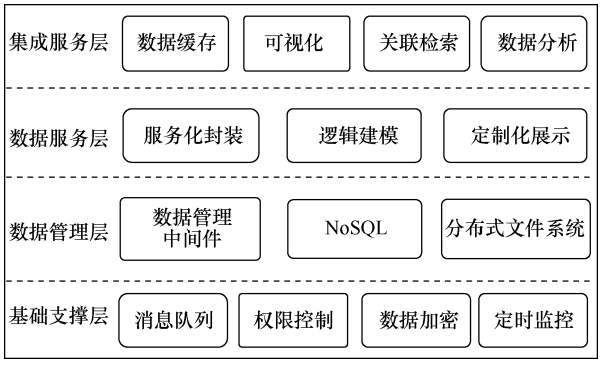


图 3 分布式科学数据管理与服务技术架构分层框架

text transfer protocol over secure socket layer, HTTPS)[4], 保证传输过程安全, 在重要信息存储方面, 采用对称加密和非对称加密相结合的方式; 定时监控主要使用线程池与定时任务技术以及日志分析算法和相关可视化技术, 对各个系统进行运行情况监控、访问情况分析等。

(2)数据管理层

数据管理层主要提供底层数据管理服务。数据管理中间件技术主要用于整合不同类型关系数据库的访问, 提供不同关系数据库的统一访问接口, 提供不同类型数据库的集成功能; NoSQL[5]主要用于存储非结构化数据, 在效率和可用性方面强于关系型数据库; 分布式文件系统主要处理大型文件存储和处理, 如日志文件等。

(3)数据服务层

数据服务层主要提供数据服务化封装、数据逻辑模型建立以及个性化服务页面定制等功能。数据服务化封装技术对数据的内容进行解释包装, 在逻辑模型之上建立服务接口, 采用表述性状态传递(representational state transfer, REST)风格, 以一种数据交换格式(javascript object notation, JSON)作为数据传输对象; 逻辑建模技术用于将同源或异源的数据集成起来, 并对其进行描述和解释; 定制化展示采用前端JS(javascript)技术, 用户可以定制服务页面内容和布局。

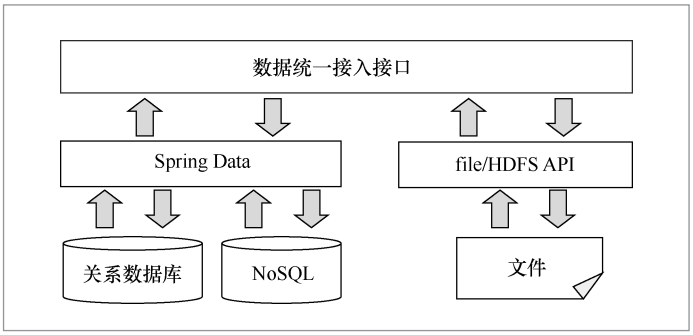


图 4 数据统一接入接口关系

(4)集成服务层

集成服务层主要将分布式数据集成处理, 统一对外提供服务。缓存技术能够提高系统响应效率, 减轻数据库压力; 可视化技术为用户提供所见即所得的数据服务, 提供基本的数据分析和展示功能; 关联检索技术可以使用户搜索结果更加准确, 同时根据用户搜索行为进行定制化推荐; 数据分析技术利用常用的大数据分析技术[6], 如Spark、R等, 为用户提供大数据分析服务。

3.2 重点接口与协议设计

(1)基础数据统一接入接口

为了解决科学数据以各种形式存储带来的不便, 体系框架提供基础数据统一接入接口, 对已存在的数据, 不论是以关系数据库、NoSQL数据库还是文件形式存储, 都可以接入系统中进行统一管理。

如**图4**所示, 数据统一接入接口通过Spring Data与关系数据库和NoSQL数据库进行交互, 通过file/HDFS[7] 应用程序编程接口(application programming interface, API)与本地文件或分布式文件系统进行交互。接口中统一定义常用的数据操作: 数据查询、数据预览、数据修改、数据删除、表结构信息查询、文件读取、文件属性查询、文件写入、文件上传等。

(2)服务统一接入接口

体系框架的统一接入接口服务于数据管理层。数据管理层发布的数据服务可以通过服务统一接入接口进行集成发布, 同时接口也支持第三方数据服务接入。为了实现接口统一接入, 系统设计了一套元数据管理体系, 描述数据服务基本信息。接口主要功能分两种: 基本信息接口和内容信息接口。基本信息接口主要用于查询数据服务相关信息, 如基本描述、数据内

容结构、返回数据格式等；内容信息接口主要用于查询数据内容。接口以超文本传输协议（hyper text transfer protocol，HTTP）为传输协议，采用REST风格设计，支持get、post、put请求格式，采用轻量级的JSON作为数据交换格式。

（3）开放服务接口

体系框架在集成服务层提供了统一开放服务接口，便于第三方应用调用服务或编程使用。为了使服务被程序或第三方应用理解，框架设计提供了两种服务接口：元数据服务接口和数据服务接口。接口采用HTTP、REST风格设计，主要以JSON作为数据交换格式。元数据服务接口包括元数据的获取功能、程序或第三方应用，可以调用元数据获取接口读取服务的元数据信息，接口提供JSON和XML两种形式的返回数据，程序可以根据元数据内容解析服务数据。数据服务接口包括数据查询接口和数据获取接口；数据查询接口可供程序通过关键字查询相关服务，获取服务ID、服务名称、服务简介等信息；数据获取接口可供程序通过服务ID和元数据中描述的参数信息，调用相应服务并获取数据，数据以JSON格式返回，对于文件型数据，接口会暂存文件到可访问空间中，返回文件的访问链接地址，程序可以直接访问链接地址获取文件。

3.3 关键技术的设计实现

3.3.1 逻辑模型映射实现

为了将科学数据转换成用户可以理解的数据，需要对科学原始数据进行抽象，建立逻辑模型，增加数据描述。通过逻辑模型映射模块可以将科学原始数据描述成用户可理解的数据或服务。为了使数据可理解，需要为数据增加额外的属性：语义属性、结构属性。

逻辑模型={语义属性，结构属性，原始数据}。

语义属性：主要描述数据的含义和用途、在上下文中的含义、与其他数据的关系等。

结构属性：主要描述数据的结构信息、数据的权限、特征属性等，如文件的类型、权限、关系数据库的列信息等。

为了实现逻辑建模，系统对常用的数据格式进行封装，对外提供丰富的可理解的数据类型，如文件、字典、日期、链接、图片、音频、视频、化学结构式、HTML等，并可扩展。逻辑模型映射模块负责将基本数据类型根据数据服务的定义，包装成用户可理解的数据格式。

逻辑模型映射的实现过程对用户是透明的，用户在数据发布时，根据所发布的数据填写部分描述信息，这部分信息被逻辑模型映射模块收集，与其他自动抽取的结构属性信息共同建立起相应数据的逻辑模型。

3.3.2 定制化数据服务发布

为了方便科学数据快速发布并提供服务，体系框架服务平台提供了快速定制化数据服务发布功能。定制化数据服务发布模块为用户提供透明简便的数据发布功能，实现简单配置、快速发布、所见即所得的发布功能。模块采用HTML和JavaScript的前端技术，实现数据发布功能。同时采用组件化技术，使用React框架将页面拆分成一个个组件，由用户进行配置。用户的配置信息会保存在配置文件中，可以重复使用，同时在系统启动时加载到缓存中，提高数据服务页面响应效率。采用React框架进行页面组件化，可以使页面上的逻辑业务模块之间的耦合度降低，使页面能够模块化、可拼装，同时完全采用前端技术，能够降低内存使用率，提高页面响应效率。

4 服务系统设计

4.1 分层系统体系设计

针对大规模分布式资源服务的需求，中国科学院计算机网络信息中心大数据部依托科学数据工程项目，围绕分布式资源服务体系框架设计，在“十五”到“十二五”期间重点建设了一批数据资源管理与服务系统。这些系统分别从资源服务体系框架的数据自治管理、数据集成整合管理和数据集成服务3个不同层次进行建设研发。整个系统建设的整体层次框架如图5所示。

数据资源自治管理层：重点面向各分布端的数据资源管理员，实现分布式数据资源的自主管理，主要系统包括面向科研团队的数据管理工具TeamDR（课题数据宝）、面向数据自主管理与发布的工具VisualDB和基于规则的数据校验工具iCheck。

数据资源集成整合管理层：重点面向数据资源及服务的集中监控与管理人员，包括数据资源整合管理和数据资源服务管理两个方面的功能。其中，数据资源整合管理包括：科学数据资源与服务注册系统RSR、数据资源在线映射与集成中间件SDM、分布式数据收割系统DDHS；数据资源服务管理包括：数据网络资源量在线统计系统Resstat、数据网络服务与访问监控统计系统Msis和科学数据服务效果评测系统Sees。

数据集成服务层：重点面向广大公众用户和科研用户，实现数据资源的集成化服务。主要系统包括：科学数据云门户CSDB、科学数据共享社区DataPub、科学数据搜索引擎Voovle和数据参考咨询与服务系统DRS。

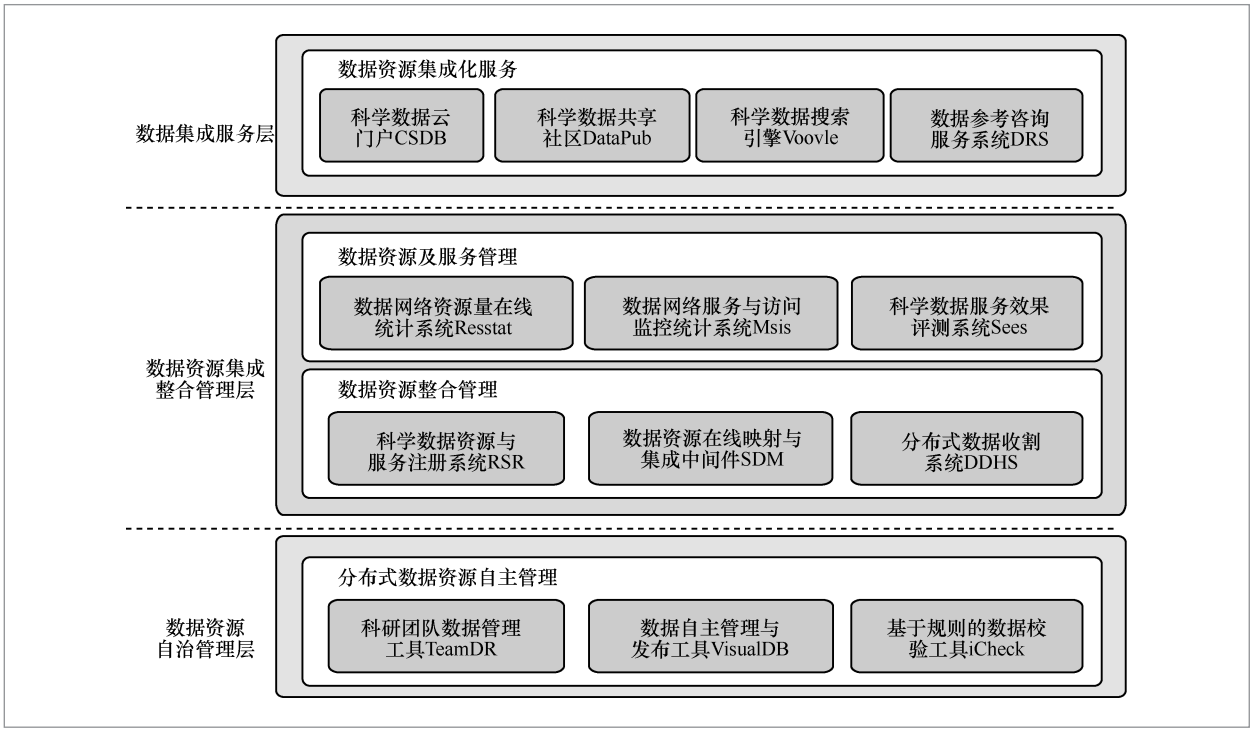


图 5 分布式科学数据管理系统分层框架

4.2 分系统功能实现

下面对系统分层框架中的部分关键系统进行重点介绍。

4.2.1 数据资源自治管理层

（1）科研团队数据管理工具TeamDR

该系统中文名为课题数据宝。系统定位于面向课题组等科研团队打造专属日常科研数据存储、组织、协作与共享的管理服务云平台和本地管理工具，是一套课题组数据管理与共享的解决方案，是一个稳定且可持续积累的课题组数据资源库。

（2）数据自主管理与发布工具VisualDB

该系统面向数据资源管理者集成异构数据源的需求，提供可视化管理、发布云平台服务与本地管理工具。它是一个帮助数据管理者管理和发布关系型数据库和文件系统的工具；一个帮助应用研发人员快速开发面向数据应用的研发框架；一套帮助数据应用低成本集成异构数据源的解决方案。

4.2.2 数据资源集成整合管理层

（1）科学数据资源与服务注册系统RSR

该系统重点实现科学数据各类资源的集中汇交、注册、审核管理。系统覆盖的资源类型包括：数据库元数据、公共服务接口、服务案例、科研论文、服务公告、软件与专利著作权、手册素材、项目文档材料。

（2）数据网络服务与访问监控统计系统Msis

该系统目标是面向站点信息管理、监测、访问统计和分析报告的需求，建成基于B/S架构的站点监测及访问统计管理的浏览平台，为各站点进一步提高服务水平提供支撑和保障。

（3）科学数据服务效果评测系统Sees

该系统通过对数据服务效果相关的定量与定性指标的采集，将评估指标体系固化在软件工具中，有效地实现对各科学数据服务系统服务效果的监控与评估。

4.2.3 数据集成服务层

（1）科学数据云门户CSDB

中国科学院数据云门户在整合“十二五”资源和服务的建设成果基础上，重点实现基础设施、平台及应用各层次科学数据云服务相关系统网站和接口的服务集成，对服务案例和检索接口进行整合；对云服务的系统进行服务状况的监测、服务状况可视化的展示；对云服务的数据库进行元数据、论文、服务API的集成发现。

（2）数据参考咨询服务系统DRS

该系统是科学数据的参考咨询服务平台，其建设目标是为用户提供一个在访问和使用数据资源遇到问题时可方便地寻求和获得帮助的平台，该系统将用户、服务专员和有关知识紧密联系起来。

5 典型项目应用

大规模分布式科学数据管理与服务体系设计完成后，在中国科学院科学数据库项目、科技部基础科学数据共享网项目和国家生态系统研究网络等项目中得到了广泛的应用和验证，极大地提高了这些重大项目的建设服务成效。下面选取典型项目应用进行介绍。

5.1 中国科学院科学数据库

5.1.1 项目背景

从“十一五”的科学数据库建库项目到“十二五”期间的科技数据资源整合与

共享工程，中国科学院计算机网络信息中心逐渐完善了一整套构建科学数据库并提供数据共享服务的服务体系，形成支持科研活动与科技创新的数据云，并从基础设施（IaaS）、数据资源（DaaS）、应用平台（PaaS）三大类服务的角度整合各类资源和服务，形成科技数据云环境。

5.1.2 系统体系框架

“中国科学院科学数据库”的分层支撑系统体系主要是由3个层次构成：基础设施服务层、数据资源服务层、应用平台服务层。

在基础设施服务层上，为各个研究所提供必需的硬件设备，配备构建数据库的基本设施，支持MySQL、Oracle等多种数据库的建设，支持文件型等多种类型资源的管理，提供多种管理工具和服务。

在数据资源服务层上，基于软件发布工具及各类资源与服务注册系统，可将建库单位的个体专业库进行整合，通过数据映射和集成，形成主题相关的专题数据库。同时也提供集中的数据资源管理，利用科学数据搜索引擎进行数据目录和资源的搜索，通过资源服务监控平台获取资源与服务的基本信息。

在应用平台服务层上，基于数据协同服务，能够提供多种数据资源的应用服务，如各数据库会最终汇聚到数据云门户站点，提供统一的服务和接口，提供数据访问标准。

5.1.3 系统建设成效

科技数据云环境的应用服务基本实现了对分布式科研数据的统一管理、发现与共享，发展到目前“十二五”期间，共有58家建库单位、1 340个数据库，中国科学院数据云整合了从资源学科领域到植物学科领域等多领域数据库资源，提供的共享数据量增加到655 TB，年均在线访问超过千万人次。“十二五”期间，累计为131项科研项目提供了数据支持和服务，数据云存储环境运行服务总容量达52 PB，云存储规模达8 PB，共拥有物理服务器约300台，虚拟机5 000多台的计算服务能力。

5.2 国家生态系统观测研究网络

5.2.1 项目背景

国家生态系统观测研究网络（CNERN）是跨部门、跨行业、跨地区的科技基础条件平台，它将各主管部门的野外观测研究基地资源、观测设备资源、数据资源以及观测人力资源等进行整合和规范化，构建国家层次的生态系统观测与研究的野外基地平台、数据资源共享平台、生态学研究的科学家合作与人才培养基地。CNERN包含53个野外观测站和一个综合研究中心，野外观测站中36个隶属于中国科学院，是中国生态系统研究网络的成员站；15个属于其他部门野外站；两个属于其他部门的子网。

5.2.2 平台系统框架

国家生态系统观测研究网络云平台（生态云平台）属于典型的分布式数据资源管理与服务体系架构，该云平台框架主要包括3个主要的部分：支撑子系统、业务系统和门户系统。

其中统一认证系统完成整个云平台用户身份的集中统一认证功能；数据汇交和集成采编两个重要的业务系统完成行政、实物等七大基础数据及采编信息的集中汇聚；考核评估业务系统实现数据资源与服务的质量效果评估；资源服务门户和综合信息门户两个重要的门户系统，主要完成信息资源的发布和数据资源在线对外服务；台站门户系统完成台站数据资源的管理和对外服务以及信息发布功能。

5.2.3 平台建设成效

国家生态系统观测研究网络云平台建设应用云计算和大数据技术，建成“运行管理”与“开放服务”相结合的统一高效的CNERN云服务环境。同时以平台建设为契机，提高CNERN的资源信息化程度，强化CNERN的资源质量，提升CNERN的服务能力，提升CNERN平台在基础条件平台中的作用，为平台规范化运行和深化服务提供技术支撑系统。建成了安全的国家生态网络私有云平台；实现了云用户的安全验证；建立了数据上报规范体系，实现了元数据的自动收割，实现了一次汇交、多处共用，保障了平台数据质量和长效更新；建立、健全了生态资源的集成并创新引导了服务考核评估的建设。

6 结束语

本文提出了一套面向大规模分布式科学数据管理与服务的技术架构与系统实现方案，充分考虑了科学数据管理与服务的核心需求，贯穿了数据服务全生命周期的各个环节，兼顾了实用性和可扩展性，其服务效果已在实际项目中得到了实践与验证。

当前，随着科学大数据的到来，科学数据的管理与服务又迎来了新的机遇和挑战。科学数据资源的管理、服务、共享得到了空前的重视，但对数据的规模、敏捷集成和交付方面又提出了新的要求。服务体系仍需要在大数据服务模式与服务的敏捷集成等方面继续探索与完善。

此外，必须说明的是，分布式科学数据管理共享是一项长期的系统工程，涉及政策、规范、系统等各个方面，需要全社会的联合推动和长期努力。只有建立了完善的共享政策、标准规范体系和管理体系，才能真正实现科学数据的潜在价值，使科学数据资源的积累与共享达到基本满足科技创新和国家发展的需求，提高国家科技创新能力和竞争力。

参考文献:

[1] 黄鼎成，郭增艳. 科学数据共享管理研究[M]. 北京：中国科学技术出版社，2002.
HUANG D C, GUO Z Y. Study on the management of scientific datasharing[M]. Beijing: Science and Technology of China Press, 2002.

[2] 诸云强，孙九林，廖顺宝，等. 地球系统科学数据共享研究与实践[J]. 地球信息科学学报，2010, 12(1): 1-8.
ZHU Y Q, SUN J L, LIAO S B, et al. Earth system scientific data sharing research and practice[J]. Journal of GEO-Information Science, 2010, 12(1): 1-8.

[3] 时子庆，刘金兰，谭晓华. 基于 OAuth2.0 的认证授权技术[J]. 计算机系统应用，2012, 21(3): 260-264.
SHI Z Q, LIU J L, TAN X H. Authentication and authorization technique based on OAuth2.0[J]. Computer Systems and Applications, 2012, 21(3): 260-264.

[4] 沈涛，马红光，薛文通. 网络数据加密算法研究及其应用[J]. 计算机工程与应用，2002, 38(19): 156-158.
SHEN T, MA H G, XUE W T. Research and application on network data encryption[J]. Computer Engineering and Applications, 2002, 38(19): 156-158.

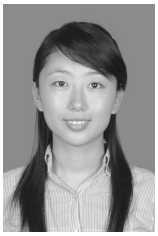
[5] 申德荣，于戈，王习特，等. 支持大数据管理的NoSQL系统研究综述[J]. 软件学报，2013(8): 1786-1803.
SHEN D R, YU G, WANG X T, et al. Survey on NoSQL for management of big data[J]. Journal of Software, 2013(8): 1786-1803.

[6] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014(9): 1889-1908.
CHENG X Q, JIN X L, WANG Y Z, et al. Survey on big data system and analytic technology[J]. Journal of Software, 2014(9): 1889-1908.

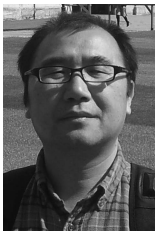
[7] GHEMAWAT S, GOBIOFF H, LEUNG S T. The Google File System[C]//19th ACM Symposium on Operating Systems Principles, October 19-22, 2003, Lake George, USA. [S.l.:s.n.], 2003: 29-43.

作者简介

刘峰(1974-),男,中国科学院计算机网络信息中心高级工程师,主要研究方向为科学数据管理与服务体系构建。

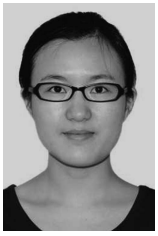


陈昕(1982-),女,博士,中国科学院计算机网络信息中心研究员,主要研究方向为数据可视分析、科学数据管理与服务。



黎建辉(1973-),男,博士,中国科学院计算机网络信息中心研究员、博士生导师,大数据技术与应用发展部主任,CODATA中国委员会秘书长,主要研究方向为大数据管理、大数据分析与处理。

刘昂(1990-),男,中国科学院计算机网络信息中心工程师,主要研究方向为科学大数据技术与服务。



韩芳(1987-),女,中国科学院计算机网络信息中心工程师,主要研究方向为自然语言处理。

收稿日期: 2016-10-08

基金项目: 国家“十二五”科技支撑计划资助项目(No. 2013BAD15B02);国家自然科学基金资助项目(No.91224006);中国科学院“十二五”信息化基金资助项目(No.XXH12504)

Foundation Items: National “Twelfth Five-Year” Plan for Science & Technology Support(No. 2013BAD15B02), The National Natural Science Foundation of China(No.91224006), Special Project of Informatization of Chinese Academy of Sciences in “the Twelfth Five-Year Plan” (No.XXH12504)