·管理纵横·

国家自然科学基金大数据知识管理服务平台总体方案及关键技术研究

姚　畅[1]　王晓帆[2]　杜　一[3]　张兆田[4]　李建军[4]　郝艳妮[1*]

（1. 国家自然科学基金委员会 信息中心，北京 100085；
2. 西安理工大学 计算机科学与工程学院，西安 710048；
3. 中国科学院计算机网络信息中心 大数据技术与应用发展部，北京 100190；
4. 国家自然科学基金委员会 信息科学部，北京 100085）

[摘　要] 科学基金数据资源是我国基础研究的重要战略资源，对我国基础科学研究具有巨大的应用潜力和价值。本文阐述了科学基金大数据的基本概念和特征，分析了科学基金大数据的应用现状及需求，设计了国家自然科学基金大数据知识管理服务平台的总体架构，剖析了平台选型的关键技术，对促进大数据技术在科学数据开放共享、知识服务等方面的应用研究具有一定的指导意义。

[关键词] 国家自然科学基金；大数据；知识管理服务；知识图谱；多维分析

随着云计算、物联网等新一代信息技术的发展及与经济社会各领域的深度融合，引发了数据量的爆炸式增长，大数据在此背景下应运而生，数据资源已经成为国家重要的战略资源和核心创新要素。2015 年 8 月 31 日，国务院印发《促进大数据发展行动纲要》（以下简称《纲要》），系统部署大数据发展工作[1]。《纲要》明确指出，推动大数据发展和应用，适应国家创新驱动发展战略，实施大数据创新行动计划，发展万众创新大数据；推动大数据发展与科研创新有机结合，形成大数据驱动型的科研创新模式，发展科学大数据；推动由国家公共财政支持的公益性科研活动获取和产生的科学数据逐步开放共享，利用大数据、云计算等技术，对各领域知识进行大规模整合，搭建层次清晰、覆盖全面、内容准确的知识资源库群，建立国家知识服务平台与知识资源服务中心，为科技创新提供精准、高水平的知识服务。

国家自然科学基金委员会（以下简称基金委）作为国家科研资助体系的重要组成部分，经过多年信息化工作建设，已经建成科学基金网络信息系统和信息开放共享环境[2, 3]。科学基金网络信息系统主要支撑基金委项目管理全过程中各项业务工作的开展，包括从项目申请阶段开始的项目在线申请、项目接收、项目受理，到同行评审、会议评审，再到项目审批、项目发布、资助计划书，以及项目在研阶段的项目进展、中期报告、中期报告评审、项目变更管理，最后到项目结题阶段的结题报告、成果研究报告、工作报告、结题验收评审、成果发布等，基本覆盖了项目管理过程中的各个环节。同时系统还包含项目经费管理、依托单位管理、评审专家管理等多项功能。截至目前，系统已积累有海量的结构化和非结构化数据，且各类数据仍逐年快速增长。

由于数据以结构化和非结构化的结构形式存储，数据扩展性较弱，项目数据关联分析和多维展现存在困难。此外，目前科学基金数据的组织、采集、存储、分析等模式与方法已不能满足更高层次的项目管理需求。近年来云计算、大数据、知识图谱等新兴技术的发展成熟，为基于知识网络的科学基金大数据处理与分析提供了新的方案。

收稿日期：2018-10-30；修回日期：2018-11-29

* 通信作者，E-mail：haoyn@nsfc.gov.cn

1 科学基金大数据的概念及特征

大数据(Big Data),是指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的海量、高增长率和多样化的信息资产。通常,大数据是以多元的形式,从多个来源搜集而来的庞大数据组,一般满足“4V”特点,即 Volume(数据量大)、Velocity(高速)、Variety(数据类型多样)、Value(价值密度较低)[4-7]。

科学基金大数据是大数据在科学研究领域的缩影,是指由科学基金项目全过程管理、经费管理、依托单位管理、评审专家管理、电子文件管理等各业务领域的结构化和非结构化数据所汇集而成的数据集合。

科学基金大数据的典型特征是数据量大、数据类型多、数据增长快、数据价值大。

(1) 数据量大。近年来,随着信息化建设的逐步深入,信息系统已经覆盖人员管理、项目申请、项目评审、项目执行、项目结题、成果提交、经费拨付、单位信息管理、专家库管理、共享服务、开放获取等各个领域,各子系统都积累了海量的数据,包括 1 200 多万人次的申请数据、47 万的获资助项目数据、33.5 万的结题项目数据、300 万项目成果数据、4 000 多个单位数据等。数据总量已达到上百 TB 量级,且各类数据仍逐年快速增长。

(2) 数据类型多。科学基金数据主要包括结构化和非结构化数据。结构化数据主要包括:项目基本信息、人员基本信息、成果基本信息、单位信息、统计汇总数据等。非结构化数据主要包括:项目申请书全文 PDF 数据、经费预算说明文件、补充说明文件(作为项目附件提交)、成果全文数据、技术支持服务数据和办公文件等。

(3) 数据增长快。随着我国从事自然科学研究的队伍不断扩大,项目申请数量逐年快速增长,这也导致项目全过程管理的各环节数据同步快速增长。此外,我国科研人员每天产生的研究成果,包括期刊论文、会议论文、专著、专利、获奖等,都将作为个人研究成果或者项目研究成果而进入信息系统。

(4) 数据价值大。科学基金数据资源是我国基础研究的重要战略资源,对我国基础科学研究具有巨大的应用潜力和价值。历年申请和资助项目数据对我国自然科学领域各学科的发展规划具有重要意义,资助项目数据对学科热点分析和发现具有重要参考价值,申请项目数据、专家个人信息、专家承担项目信息以及成果信息等的关联分析对于项目辅助指派推荐评审专家等具有重要价值。

2 科学基金大数据应用现状及需求

2.1 应用现状

当前,科学基金网络信息系统存储和管理了海量的科学基金数据。由于长期运行积累的数据包括结构化、非结构化等多种类型,传统的基于关系数据库的数据架构面临大规模数据的挑战。数据之间缺少知识性富关联,系统的知识本体、知识模型、知识服务不能有效进行区分,知识不能通过关联进行自动发现,使得表现度、可视化和友好性缺乏充分的展现。

由于数据分散在不同的业务系统,各子系统各自为政,独立建设,基础数据多头维护,数据标准化程度不高,数据质量有待提高,数据资源有待统一整合、管理和监控。数据异构特征明显,在各子系统的输入、输出接口多样化,形成的业务孤岛、信息孤岛有待打破。

此外,数据的利用仍停留在初级阶段,数据服务以查询统计为主,深层次的数据分析、数据挖掘较少,描述项目、人员、成果、单位等的数据模型缺乏自解释能力,基于知识图谱的科研项目热点分析、评审推荐、关联分析等更高层次的项目管理需求难以满足和实现。

2.2 建设需求

(1) 数据组织。随着信息化的深入以及数据规模、种类、来源的不断增加,当前基于关系数据库的组织无法满足逐渐增长的业务需求,在数据组织上需要建立对项目、人员、成果、单位进行唯一标识的 NSFCID 体系,理清当前项目、人员、成果、单位等各要素之间的关系,建立关系数据模型,理清当前针对项目、人员、成果、单位的多维分析需求,建立多维分析数据模型。

(2) 成果集成。基金委通过软课题等方式探索建设了共享服务网、专家 Profile 系统、基础研究知识库等应用子系统,但是各子系统彼此独立建设,形成了信息孤岛。因此,本平台需要建立系统、算法的集成机制,使得各子系统与算法能够快速集成到本项目所建设的大数据中心来,将现有成果有效地集成到大数据平台中。同时,为进一步完善基金委的成果数据数量和质量,需要实现第三方数据的采集与集成。

(3) 科学基金大数据支撑平台建设。当前各个

子系统直接从科学基金网络信息系统导入数据存在稳定性隐患和性能瓶颈等问题，需要建立统一的存储环境，用于存储采集与集成的数据，且要求具有良好的伸缩性，适应数据的动态变化；需要建立数据流水线机制，设计通用的数据采集、清洗、存储机制，支持数据的快速采集与存入；需要建立大数据服务和节点管理环境，建成可以快速添加大数据服务、灵活管理节点的大数据环境，以支持新的大数据服务和应用。

（4）高效数据服务。在基金大数据支撑平台及相关的数据整合业务需求基础上，需要为第三方应用系统提供数据服务，包括查询、多维分析、网络服务等；需要为基金委内部的业务系统提供数据服务，包括丰富的查询服务、多维统计分析服务等。此外，平台需要建立服务授权机制，支持新的应用系统的接入。

（5）完善的安全机制。数据开放共享意味着面临更大的安全威胁，为保证平台和数据的安全性，需要从大数据基础平台安全、应用子系统安全和数据安全等多方面统筹考虑，构建完善的安全保障体系，包括数据的不同安全等级存储和处理、系统访问的权限设置和管理、数据的加密存储和大数据环境保持高在线率，通过分片、数据冗余等方法实现单服务节点损坏时仍能保持服务等。

3　平台总体方案

3.1　建设目标

国家自然科学基金大数据知识管理服务平台采用图数据库、数据立方体等新型大数据技术构建，一体化汇聚和存储项目、人员、成果、单位等科研实体以及实体关联，实现自然科学基金业务生产环境数据到数据中心的一次性抽取和多次使用。在不需要与业务生产环境建立关联的情况下，大数据知识管理服务平台基于图数据库关联分析的优势实现交互式查询、人员关系发现、科研合作网络分析、科研影响力分析、科研社区发现等关联分析需求；基于数据立方体统计分析的优势，实现历史基金数据的快速多维统计分析，最终形成基金委大数据知识库，提供独立于业务生产环境的安全高效数据分析平台，支持未来数据分析业务的动态扩展和知识库数据的大规模增长。

3.2　总体架构

国家自然科学基金大数据知识管理服务平台包括五个层次，分别是设施层、数据层、分析层、应用层和用户层（图 1）。

（1）设施层。设施层包括基金委目前各系统运行所使用的硬件设施及云平台环境。在该层次上，运行当前基金委各个系统及相关的业务数据库，国家自然科学基金大数据知识管理服务平台利用当前已有的硬件设施及云平台，进行数据的整合和平台的构建。

（2）数据层。数据层在设施层之上，首先构建多维数据仓库及科研项目关系网络体系，通过对业务的理解，基于现有基金委运行的各个子系统，完成开放数据与业务数据的抽取、重新组织，并通过唯一标识、网络关联及结构化等方法完成对数据的预处理。

（3）分析层。分析层在数据层之上，在完成对数据组织与管理的基础上，实现如维度统计、Top-N 统计等多维统计分析与关联检索、关键节点发现、聚类分析、社区发现、PageRank 分析等网络分析功能。

（4）应用层。应用层在分析层完成基本的多维分析及网络分析的基础上，面向基金委的相关知识管理需求，实现交互式查询控制台、回避关系挖掘、科研合作网络、知识关联查询、科研影响力分析及成果多维统计等应用功能。

（5）用户层。用户层利用应用层完成的相关功能，实现基金知识管理服务门户、国家科技成果信息系统及掌上 NSFC-KBMS APP 三个应用系统供用户使用。

在各个层次的构建过程中，设计整体系统的数据组织与服务接口，以及平台的安全体系建设。其中，平台安全体系建设尤为重要，一方面要保证各类用户能够充分分析、挖掘、共享和使用科技资源，保证系统的灵活性和稳定性，以及用户的体验度，另一方面也要防止科技资源被不恰当地使用、泄漏甚至是破坏，要保证平台和数据的可靠性、权威性和安全性。平台安全体系建设遵循信息安全技术信息系统安全等级保护基本要求第三级标准，从硬件、软件、网络、服务器、数据库等多个方面设计系统的安全保障。针对平台应用安全方面，一方面规范系统用户管理，对系统用户、角色、权限进行划分和制定，制定系统用户和权限管理基本原则，对访问系统的用户进行分类和角色确认，另一方面对用户的创建、注册、口令设置及注销等进行管理。针对平台数据安全方面，对数据按照类型和级别进行管理，制定相应的数据安全策略，同时采用数据存取和访问控制、数据加密存储、数据监控、数据审计以及数据备份与容

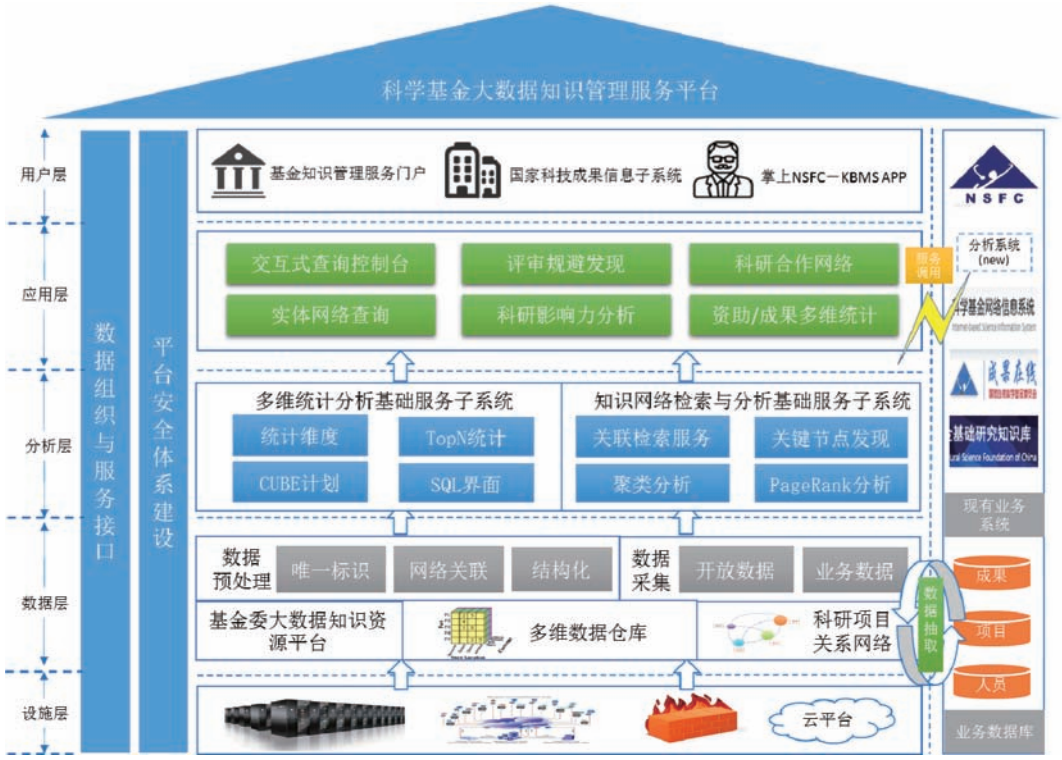

图 1 总体功能框架

灾等安全管理技术和管理措施对平台数据进行安全管理，使数据资源在使用过程中能够满足业务需求的同时又能获得安全保护。

3.3 功能模块

通过对需求的进一步分析，将整体的功能进一步整理，国家自然科学基金大数据知识管理服务平台的建设可以划分成两大部分。第一部分是应用子系统，包括知识管理服务门户、科技成果信息子系统和掌上 NSFC-KBMS APP。第二部分是国家自然科学基金知识管理服务平台的大数据部分，包括基金大数据采集、加工与汇聚，基金大数据支撑平台，基金大数据服务三个主要功能模块(图 2)。每个功能模块均由基于对业务系统及大数据建设的需求理解所定义的与业务相关的功能性需求以及通用需求组成。与基金委业务相关的功能，例如，基金大数据采集中对业务数据的采集，基金大数据支撑平台中网络关系数据环境搭建，基金大数据服务中的多维分析服务。通用需求是为了能够支持新的应用系统对新的大数据服务、节点、数据采集、应用服务的需求而建立，包括基金大数据采集模块中的流水线采集配置模块，以及大数据支撑平台中的基金大数据基础环境管理模块。

(1) 基金大数据支撑平台。基金大数据支撑平台包括四类子功能，分别是基金大数据存储环境、多维数据环境、网络关系数据环境以及大数据基础环境管理。其中，基金大数据存储环境为基金大数据构建高效的存储环境，包括大规模实体-关系数据存储环境、大规模文档数据存储环境两个主要功能模块；多维数据环境在存储环境基础上，构建和管理多维数据，包括多维分析立方体构建及多维分析模型管理；网络关系数据环境在存储环境基础上，构建和管理网络环境关系数据，包括网络关系数据分析模型构建及网络分析模型管理。大数据基础环境管理实现对整个基金大数据底层环境的管理，包括大数据基础环境服务部署与配置、大数据基础环境服务管理、大数据基础环境节点部署与配置及知识资源中心权限管理四个功能模块。

(2) 基金大数据采集、加工与汇聚。基金大数据采集、加工与汇聚为数据存储、管理提供基础的大数据环境，包括三类子功能，分别是基金大数据采集、基金大数据加工及基金大数据流水线管理。其中基金大数据采集包括业务数据采集和开放文献数据采集两个主要功能模块；基金大数据加工包括热词数据加工、NSFCID 数据加工及网络关系加工三个主要功能模块；基金大数据流水线管理包括流水线采集配置、流水线计算控制等主要功能模块。

(3) 基金大数据服务。基金大数据服务包括基本查询服务、多维分析服务和网络查询分析服务。

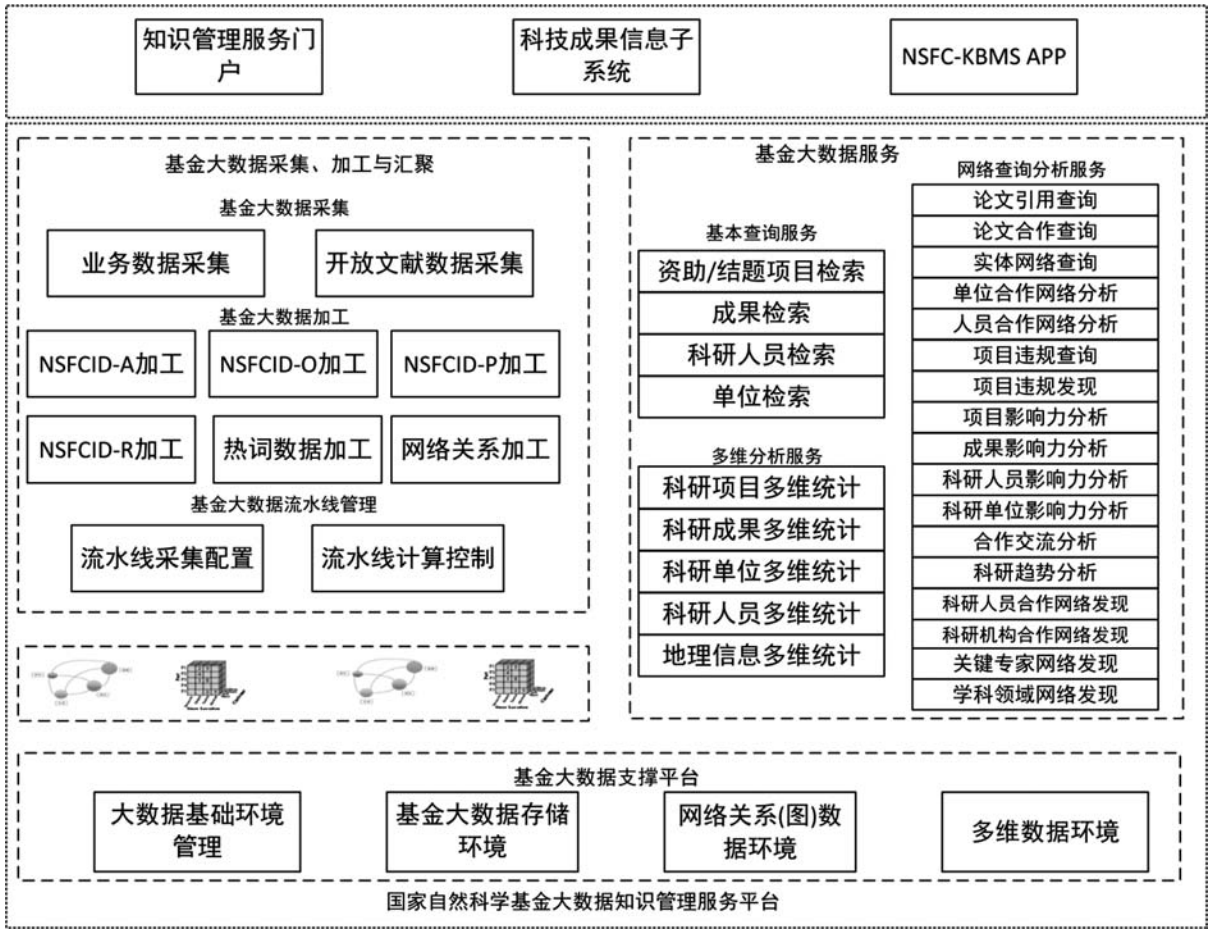

图 2　大数据平台功能模块图

科学基金数据主要包含项目、人员、成果、单位等数据，除共有的时间、空间等维度外，项目数据包含项目类型、项目金额、项目起止时间、项目申请数据等维度；人员数据包含人员的学历、年龄、研究领域、工作年限、承担项目数量、发表论文数量等维度；成果数据包含成果的类型、成果级别、成果影响力等维度；单位数据包含单位所在的行政区域、单位类型、单位规模等维度。经过数据的汇聚、清洗及组织后，需要构建多维统计分析系统来实现不同维度的数据分析，为上层应用子系统提供分析服务。项目、人员、成果、单位等数据，除具有多维的特征外，还具有典型的网络特点。不同类的数据之间存在明显的关联特性，例如，一个项目包含一个或多个参与人员，每个人员属于一个或多个单位；一名人员参与一个或多个项目，产出多种成果，并且可能供职于不同的单位；一个成果由一名或多名人员合作完成，由一个或多个项目资助完成；一家单位承担多个不同的项目、拥有多名人员、产出多个成果。除各类数据之间的关联特性外，同一类数据之间也存在着关联关系。在同一类数据内部及不同类数据之间，存在大量的网络特性，利用经典的网络分析算法，在知识库管理子系统提供的数据处理能力的基础上，实现知识网络关联查询、知识网络图分析及知识网络图分析挖掘算法库，为上层应用子系统提供分析服务及算法支持。

4　关键技术选型及优势分析

系统设计过程中，优先采用成熟先进的大数据技术，具体设计方案为：采用 HDFS 文件系统及 Hadoop 存储环境，进行基础大数据环境的搭建，在 Hadoop、Spark 环境基础上搭建 Hive 及 HBase 等大数据管理环境，然后在 Hive 及 HBase 基础上搭建相关的图数据库环境 Titan 及多维数据仓库环境 Kylin。在面向用户使用的子系统实现上，采用成熟的基于微服务架构的 SpringBoot 框架实现 Web 应用程序开发与集成，通过负载均衡、消息队列方式实现系统高性能服务，使用 Ajax 异步调用、Bootstrap 模板完成 Web 前端界面开发。在数据采集、存储、分析、计算领域，数据与服务的跨平台扩展与应用将是考量系统可移植性的重要标准。该

平台将基于先进的 J2EE 技术架构，构建多层软件体系，并通过面向服务方式，使用 WebService 和 OSGi 等技术集成并整合平台内不同应用系统，实现不同功能的模块化开发，从而保证服务级别的分布性与可移植性，为基金大数据知识管理与服务平台提供先进、安全、可靠、兼容性强、易扩展的平台级解决方案。

4.1 大规模分布式图数据库 Titan

网络关系数据环境采用 Titan 数据库进行构建。Titan 是一个分布式的图数据库，支持横向扩展，可容纳数千亿个顶点和边。Titan 支持事务，并且可以支撑上千并发用户和计算复杂图形遍历，支持 HBase、Cassandra 等分布式存储。因此，Titan 支持较大的数据规模，并且面向基金委未来的数据发展趋势，具有较好的水平可扩展性。

4.2 大规模数据仓库平台 Hive

平台需要构建大规模数据仓库来存储流水线采集的业务数据、开放文献数据、关键词数据及 NSFCID、网络关系数据，采用 Hive 实现大规模数据表的存储。

Hive 是建立在 Hadoop 上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载(ETL)，这是一种可以存储、查询和分析 Hadoop 中大规模数据的机制。Hive 定义了简单的类 SQL 查询语言，称为 HQL，允许熟悉 SQL 的用户查询数据。同时，这个语言也允许熟悉 MapReduce 的开发者开发自定义的 mapper 和 reducer 来处理内建的 mapper 和 reducer 无法完成的复杂分析工作。

Hive 的底层存储使用 HDFS。HDFS 是一个高度容错的系统，拥有故障检测和自动快速恢复的特性。典型的 HDFS 适合大规模数据集应用，默认提供三副本机制，保证数据安全可靠使用。因此通过采用 Hive，数据表存储环境可支持 10 亿条以上记录，由于 HDFS 具有可以快速添加存储节点的特性，存储能力可得到快速提升。

4.3 数据统计分析引擎 Apache Kylin

针对项目、人员、成果、单位在时空维度上的自由统计分析，为了得到良好的统计性能，多维数据环境采用 Kylin 进行构建。

Kylin 是一个 MOLAP 系统，它的工作原理就是对数据模型做 Cube 预计算，并利用计算的结果加速查询。由于 Kylin 的查询过程不会扫描原始记录，而是通过预计算预先完成表的关联、聚合等复杂运算，并利用预计算的结果来执行查询，因此相比非预计算的查询技术，其速度一般要快一到两个数量级，并且这点在超大数据集上优势更明显。该模式保证了平台具有较大的高维度数据统计能力，并面向不断增长的基金委数据具备较强的水平扩展能力。

4.4 大数据流水线软件 Apache Nifi

平台的数据采集和加工是一个常态化、耗时的过程，这个过程需要采用高效的管理和调度工具来实现自动化、高可靠运行。本平台采用流水线软件 Apache Nifi 来实现复杂多样的流程管理。

大数据流水线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。

Apache Nifi 是一个易于使用、功能强大而且可靠的数据处理和分发系统，可以实现基金委大数据流水线的可视配置和监控。在性能方面，具有低延迟和高吞吐量特性，满足基金委大数据流水线的需求。在扩展性方面，用户可根据自己的需求开发特定处理器组件，根据不同需求完成不同流水线的配置。在安全性方面，支持 SSL，SSH，HTTPS 加密内容和可插拔的基于角色的验证/授权。

4.5 分布式大数据处理引擎 Spark

为了实现知识大数据的分布式计算，提高数据的处理和分析性能，平台采用 Apache Spark 系统来实现大规模的批量计算和流式计算。Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，换句话说，Spark 启用了内存分布数据集，除了能够提供交互式查询外，还可以优化迭代工作负载。

Spark 是在 Scala 语言中实现的，它将 Scala 用作其应用程序框架。与 Hadoop 不同，Spark 和 Scala 能够紧密集成，其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集，具有运行速度快、易用性好、通用性强和随处运行等特点，这四项特征保证了平台具备较好的计算性能、可移植性和水平扩展性。

4.6 大数据基础环境管理软件 PackOne

为了应付大规模的数据管理、处理和分析，需要采取并维护 Hadoop、Spark、Titan 等大数据工具，这

种多元复杂的大数据基础环境，需要强大的管理工具。平台研发了 PackOne 实现对大数据基础环境的管理。PackOne 是快速部署、管理、监控的大数据管理平台，可以帮助用户快速搭建大数据环境，降低大数据技术的使用门槛，同时对各类产品的版本进行严格兼容性测试和调优，降低选型的时间。

PackOne 可以呈现指定服务的总览、图表和告警。通过 Web 可以进行大数据产品的安装和管理，执行基本操作，可以进行开始和停止服务，添加节点，更新服务配置。

5　结束语

大数据技术发展变化日新月异，大数据应用已经深入各行各业。本文采用最新的图数据库、大数据多维分析、知识图谱等大数据技术构建国家自然科学基金大数据知识管理服务平台，实现了一体化采集、清洗、存储多源海量数据和基于图数据模型的多元数据融合，建立了包括项目、人员、单位、成果的科学基金大数据知识图谱和基于大数据立方体的多维数据大仓库，构建并实现了基于关系网络的深层知识分析挖掘和可视化展示。国家自然科学基金大数据知识管理服务平台的建设对充分释放我国科学基金大数据所产生的红利，推动科学数据开放共享，促进从传统的数据服务向知识服务、从服务用户向服务社会迈进具有重要的现实意义，对推动大数据发展与科研创新有机结合，形成大数据驱动型的科研创新模式，为科技创新提供精准、高水平的知识服务，为提升国家科技创新能力提供强大助力具有重要意义。

参　考　文　献

[1] 中华人民共和国国务院. 国发(2015)50 号 促进大数据发展行动纲要. 北京:中华人民共和国国务院. 2015. 8.

[2] 李建军，卿来云. 国家自然科学基金委员会“十三五”期间信息化建设展望. 中国科学基金，2017，31(2)：170—175.

[3] 李东，马建，姚畅，等. 信息化助力国家自然科学基金实现精准管理与开放共享. 中国科研信息化蓝皮书，2017：239—249.

[4] 孟小峰. 大数据管理概论. 北京：机械工业出版社，2017.

[5] 邓奕军,李燮慧. 基于大数据知识服务体系的数字图书馆构建. 广西民族大学学报(自然科学版)，2017，23(3)：76—80.

[6] 史天运，刘军，李平，等. 铁路大数据平台总体方案及关键技术研究. 计算机应用，2016，25(9)：1—6.

[7] 王丽，王苹，沈俊辉. 基于 Hadoop 的中医药大数据平台基础架构的设计与研究. 中国医药导报，2018，15(6)：158—162.

Overall scheme and key technologies: knowledge management and service platform based on NSFC big data

Yao Chang[1]　Wang Xiaofan[2]　Du Yi[3]　Zhang Zhaotian[1]　Li Jianjun[4]　Hao Yanni[1]

(1. *Information Center, National Natural Science Foundation of China, Beijing* 100085;
2. *School of Computer Science and Engineering, Xi'an University of Technology, Xi'an* 710048;
3. *Department of Big Data Technology and Application Development, Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190;
4. *Department of Information Science, National Natural Science Foundation of China, Beijing* 100085)

Abstract　Scientific fund data resources are important strategic resources and have great utilization potentiality for basic research in China. This paper expounds the basic concept and characteristics of scientific fund big data, analyzes the current situation and demand of scientific fund big data application, designs the overall architecture of NSFC big data knowledge management and service platform, analyzes the key technologies, and provides some guidance to promote the application of big data technology in scientific data open sharing and knowledge service.

Key words　National Natural Science Foundation of China (NSFC); big data; knowledge management and service; knowledge graph; multidimensional analysis