面向特定领域大数据平台架构及标准化研究

林旺群，高晨旭，陶克，邓波，包阳
北京系统工程研究所，北京 100101

摘要
分析了与国家民生和安全密切相关的特定领域的大数据发展现状，介绍了大数据标准研究的相关工作。研究了特定领域大数据的特点和现实需求，给出了适合这些领域大数据建设的参考架构，结合大数据主流技术和基础平台，详细分析了特定领域大数据参考架构的不同部件。考虑大数据未来技术发展，梳理了面向特定领域大数据标准体系发展脉络，提出了相应的大数据参考标准体系，为面向特定领域大数据相关体系架构和标准化建设提供技术参考。

关键词
特定领域大数据；参考架构；标准体系

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017041

Research on big data platform architecture and standardization for specific fields

LIN Wangqun, GAO Chenxu, TAO Ke, DENG Bo, BAO Yang
Beijing Institute of System Engineering, Beijing 100101, China

Abstract
The research development and related works of big data were analyzed. According to the characteristics of big data and the particularly requirements of the specific fields mentioned above, the reference architecture was presented. The different components of the inference architecture in detail combing the mainstream technology and basic platform were analyzed. The milestones of big data standardization were studied and the standards system of big data was presented. All of these contribute the technology development of reference architecture and standardization of big data in future, and provide technical reference for architecture-oriented and standardization-oriented construction of large data-related fields.

Key words
big data in specific fields, reference architecture, standards system

1 引言

随着移动互联网、物联网和云计算等技术的不断进步和用户数量的急剧增加，与国家民生和安全密切相关的银行、电子商务、航空航天、网络安全、应急反恐、军事对抗等特定领域（以下简称“特定领域”）产生的数据规模急剧增长。这些数据具有类型复杂、结构多样、体量巨大、价值密度低、挖掘潜力巨大等特点，大数据时代已经来临[1-3]。与其他领域大数据相比，上述特定领域应用大数据有独特的要求。一是安全性要求更高，比如在与网络安全相关的大数据系统中，需要考虑应对信息对抗条件下的各类安全威胁，对数据和系统的保护和授权必须严格遵守相关权限管理要求[4]。二是可靠性要求更高，从硬件到软件，从数据到应用，在大数据系统的各个层次都要提供各类容错抗毁措施。比如航空航天领域的在轨通信卫星7×24 h不间断地向地面发送图像影片资料，同时地面基站需要实时向太空卫星发送各类指令数据，这就要求相关大数据系统具有高稳定性。三是时效性要求更高，对各类动态信息的收集、融合和分析速度往往要求达到实时或近实时。比如在银行领域，在大型节假日等高峰期，相关大数据系统每秒需要处理的事务数已经超过十万条。四是多样性要求更高，尤其是各类传感器产生的动态流数据占据重要比例。比如在应急反恐和军事对抗领域的相关大数据处理平台中，需要全面分析互联网、战场物联网传送过来的文本、图片、语音、视频等各类数据。五是数据质量要求更高。比如在银行信用卡消费相关业务中，需要对实时处理的各类数据进行真伪校验和欺诈识别。在事关国计民生和国家安全的特定领域，在复杂对抗条件下数据质量攸关成败，对其甄别、清洗、融合和分析处理要求更为严格[5]。

大数据强调尊重事实理念，将精准计算与概率评估两种客观要求有机统一，突出用数据来评估、用数据来管理、用数据来决策、用数据来创新[6]。大数据相关技术是对传统数据处理技术进行继承和创新，通过有效解决大数据的高效存储、快速计算、多维关联、深度分析、趋势预测和辅助决策等问题，挖掘大数据隐藏的客观规律，预测事物及行为发展变化的趋势，支持做出正确的判断与决策。大数据已经成为国家竞争的战略前沿、企业创新发展的重要来源。

面向特定领域的大数据标准化工作是实现相关大数据系统互连、互通、互操作的重要前提，是实现大数据系统中重要信息跨部门、跨领域、跨平台安全共享的坚实基础。加强特定领域大数据平台架构研究，构建特定领域大数据相关标准体系，对加速这些领域大数据的快速发展具有重要意义。

2 相关工作

2.1 主要国家相关战略

美国政府的网络与信息技术研发（NITRD）计划早在2011年就成立了大数据高级指导小组，负责大数据项目的遴选、启动和协调。2012年3月美国政府发布《大数据研究和发展倡议》，将大数据提升为国家战略。2013年4月美国国会科学、空间和技术委员会下的研究专委和技术专委专门就“下一代计算与大数据分析”举行联合听证会，邀请政府部门（自然科学基金委员会）、IT企业（IBM）和教育部门（北卡

罗来纳州立大学）的专家就大数据的影响及国家应做的工作进行深入讨论和佐证。欧盟2011年12月推出《开放数据战略》，以开放数据为核心，制定了迎接大数据时代的相关策略。受此影响，各界从领域需求出发，积极推进大数据应用。欧盟通过“开放数据平台”向公众提供包括地理、大气、国际贸易、农业等信息在内的5 815个数据集，大力推进了欧洲数字化进程。日本政府2013年发布《创建最尖端IT国家宣言》，阐述2013—2020年期间以发展开放公共数据和大数据为核心的日本新IT国家战略。日本产业界也十分强调利用大数据的相关理念方法解决工程应用中的实际问题。此外，英国、法国、澳大利亚等国家也相继制定了相关大数据国家发展战略。

我国政府2014年《政府工作报告》将大数据列为我国政府推动国家科技创新和引领经济结构优化升级、赶超国际先进水平、引领国家未来产业发展的六大战略性创新产业之一。2015年6月，国务院办公厅印发《关于运用大数据加强对市场主体服务和监管的若干意见》指出“加强政府信息标准化建设和分类管理”。同年8月，国务院以国发〔2015〕50号印发《促进大数据发展行动纲要》，明确指出要建立标准规范体系，推进大数据产业标准体系建设，加快建立大数据市场交易标准体系。

2.2 特定领域研究情况

在银行、电子商务等核心金融领域，大数据建设十分关注建立各应用之间的数据共享交互标准，消除应用之间的数据壁垒和信息孤岛，统一业务和技术的数据字典，构建完整、全面、统一的企业数据信息视图。在银行系统，大数据相关标准建设正经历由传统数据标准规范向大数据背景下的数据标准发展。ISO/TC68金融服务技术委员会是国际标准化组织下设的从事金融标准化工作的机构。该机构近年来紧密跟踪大数据背景下金融标准化研究工作，先后在金融业务、业务数据安全等领域提出了多项国际标准。在国内，2004年中国建设银行提出银行业“数据标准化”概念；2008年中国光大银行成为国内第一家实施数据标准化的股份制银行，并将成果运用到数据仓库建设中；2014年中国人民银行相继发布《银行数据标准定义规范》《统计数据与元数据交换（SDMX）》等标准规范，为大数据背景下统一银行数据属性描述，促进银行间数据共享等方面发挥积极意义。在电子商务领域，大数据相关建设如火如荼。2016年电子商务大数据专题组在电子商务大数据采集规范、仓库建设模型和应用指标体系等方面已经向相关部门提交了标准规范。同时，在用户画像规范、管理系统规范、营销分析规范等方面正在展开深入研究。

在航空航天领域，大数据相关系统建设为相关问题的研究和解决提供了新的手段。美国国家航天局约翰逊航天中心和企业合作开发的可扩展大数据处理分析平台每秒能处理700 TB数据，通过该系统实现了对太空探测设备回传数据的实时处理，提高了外太空异常信号捕捉与探测能力。荷兰、瑞士、英国和其他17个国家与IBM公司合作，开发了一个名为DOME的超级计算系统，该系统每天可处理1 EB的天文数据，通过该系统加速了太空科学问题的研究。

在网络安全、应急反恐等领域，美国十分重视大数据相关系统建设。美国国防部近年年均投资超过2.5亿美元进行大数据研究，相继开展了数据到决策、网络内部威胁、视频与图像检索分析工具（VIRAT）、洞察（Insight）、多尺度异常

检测（ADAMS）、XDATA等20多项军事大数据项目。美国国家安全局（National Security Agency，NSA）“棱镜计划”的主要技术基础是大规模图数据管理技术，其中基于Hadoop的Accumulo系统已成为NSA安全战略的关键，该系统可在大约140 h内完成对4.4万亿（2^{42}）个节点和70万亿（2^{46}）条边的图数据的分析计算，可对参与计划公司（包括微软、雅虎、谷歌、苹果等公司）的全世界用户进行全面监听。英国情报机构的“时代计划”通过采用大数据技术中高可扩展的体系架构，能够在线收集、存储所有流经英国服务器的所有互联网数据，并进行情报分析。

在军事对抗领域，美国海军为了实现在全球范围内对数据资源进行统一调配和优化使用，相继启动了海军大数据、远征作战海军战术云等项目，以提高其远征作战能力。美国海军甚至在探讨利用大数据技术颠覆传统的潜艇作战模式。美国陆军早在2009年就启动了有关分布式通用地面系统——陆军分系统（DCGS-A）的大数据支撑建设，其利用私有云平台为驻阿富汗美军提供最新战场情报信息。DCGS-A借助有效的大数据交互标准实现系统间的相互访问，替代了以往一系列烟囱式的情报数据收集与分析系统，实现了系统级别的互联、互通和互操作。美国军方和相关国防项目承包商积极参与国际大数据标准化组织以及产业联盟的相关标准化工作，比如美国著名的国防承包商Level 3通信控股公司下属的数据分析部门，向美国国家标准与技术研究所大数据工作组提交了国防军事领域中3个大数据典型用例，其中涉及地理信息数据分析与可视化技术（基于云平台的大规模地理空间数据分析和可视化）、非结构化数据的持续实时分析与应用（基于广域大幅面图像与全动态视频的目标识别和跟踪）、情报数据处理与分析技术。

2.3 大数据标准建设

在大数据标准研究方面，美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）于2013年6月成立了大数据公共工作组（NBD-PWG），将达成共识的定义、术语、安全参考体系结构和技术路线图形成标准，并提出数据分析技术应满足的互操作性、移植性、可用性和扩展性需求，以解决如何安全有效地支持大数据应用技术基础设施[7]。国际标准化组织/国际电工委员会第一联合技术委员会（ISO/IEC JTC1）于2013年11月全会上成立负责大数据国际标准化的大数据研究组（ISO/IEC JTC1 SG2），负责调研各个标准化组织在大数据领域的关键技术、参考模型以及用例等标准基础，评估分析当前大数据标准的具体需求，提出大数据标准的优先顺序。国际电信联盟远程通信标准化组织ITU-T将物联网、云计算与大数据相关技术结合，在高吞吐、低延迟、安全、灵活和规模化的网络基础设施以及垂直行业平台的互操作等领域开展了相关标准化研究工作[8]。

国内学术界和产业界也意识到了大数据标准体系研究工作的重要性，做了很多重要工作。全国信息化标准委员会近年相继在元数据、数据库、数据建模、数据交换与管理等领域推动相关标准的研制与应用，为提升跨行业领域数据管理能力提供标准化支持[9]。中国电子技术标准化研究院于2015年12月份推出《大数据标准化白皮书V2.0》[10]，这为推进我国大数据体系架构研究和标准化工作提供了重要技术基础。

3 面向特定领域大数据参考架构

在特定领域面临的任务种类多样，需要处理的数据类型繁多，比如在面向实时处理的电子商务在线交易、银行系统在线支付欺诈检测等任务时，需要在线分析大量结构化和非结构化数据；在面向网络安全密码破译、海量文本检索等任务时，需要离线分析大量结构化、半结构化数据。此外，在面向航天侦察、国防安全等任务时，相应的大数据系统往往还需要同时满足在线计算和离线分析的要求。

由于面向特定领域大数据系统在安全性、可靠性、时效性、多样性以及数据质量等方面的特殊要求，传统面向单一任务类型的数据处理平台和框架无法有效应对特定领域的大数据处理需求。比如，基于传统MapReduce架构的大数据处理平台对于需要迭代离线计算的文本处理和数值计算具备优势，但是可能无法应对数据流条件下的大规模图数据处理；同样基于Spark架构的大数据分析平台在大规模流式数据处理上具备较大优势，但是对于密码破译和气象预报等巨大计算量类型的大数据业务可能无法有效应对。因而，面向特定领域的大数据平台参考架构既需要考虑特定领域的特殊性，又需要考虑这类任务的通用性。

综合考虑特定不同任务特点、不同用户需求和不同数据差异，提出如**图1**所示的面向特定领域大数据参考架构模型。参考架构由数据源层、大数据管理层、大数据计算层、大数据应用层、大数据相关通信协议和标准、大数据安全维护和管理等部件组成，其中大数据相关通信协议和标准、大数据安全维护和管理两大部件贯穿大数据参考架构始终。大数据管理层和大数据计算层由底层各类物理存储资源、计算资源和网络资源等通过虚拟化和分布式技术形成的虚拟资源提供支撑，构成大数据体系平台。自底向上，面向特定领域大数据参考架构体现了“数据→信息→知识→决策”的转化过程，实现大数据到大价值的转变。

(1)数据源层

主要负责数据的供给和数据清洗等。面向特定领域大数据来源广泛，类型多样，体量巨大。数据源层数据来源包括面向特定领域专门业务数据、系统运行数据、外部采集数据、环境感知数据和信息系统迁移改造过程中留下的历史遗留数据等。对于面向特定领域的大数据数据源，从数据类型维度分析，这些数据包括结构化数据、半结构化数据和非结构化数据；从数据时间维度分析，这些数据包括离线数据、近似实时数据和实时数据。这些数据体量巨大，比如航空航天领域每分钟实获卫星下传数据量可能超过1 GB，网络安全领域每天需要实时处理的数据量可能超过1 TB。各类数据由各类传感器传输到代理服务器，然后由各类代理服务器对这些数据进行数据去噪和清洗，并且实现数据的抽取、转换和加载(ETL)过程。

(2)大数据管理层

负责对特定领域大数据的存储、组织和管理。由于不同领域面临的任务不同，因而对数据格式、存储方法、读写方式、存储周期等要求差异较大，面向特定领域的大数据参考架构在大数据管理层设计了针对不同任务所需要的大数据组织和管理方法，包括采用分布式关系数据库、分布式文件系统、NoSQL数据库等。目前主流的分布式关系数据库包括Oracle RAC、MySQL cluster、MemSQL等。分布式数

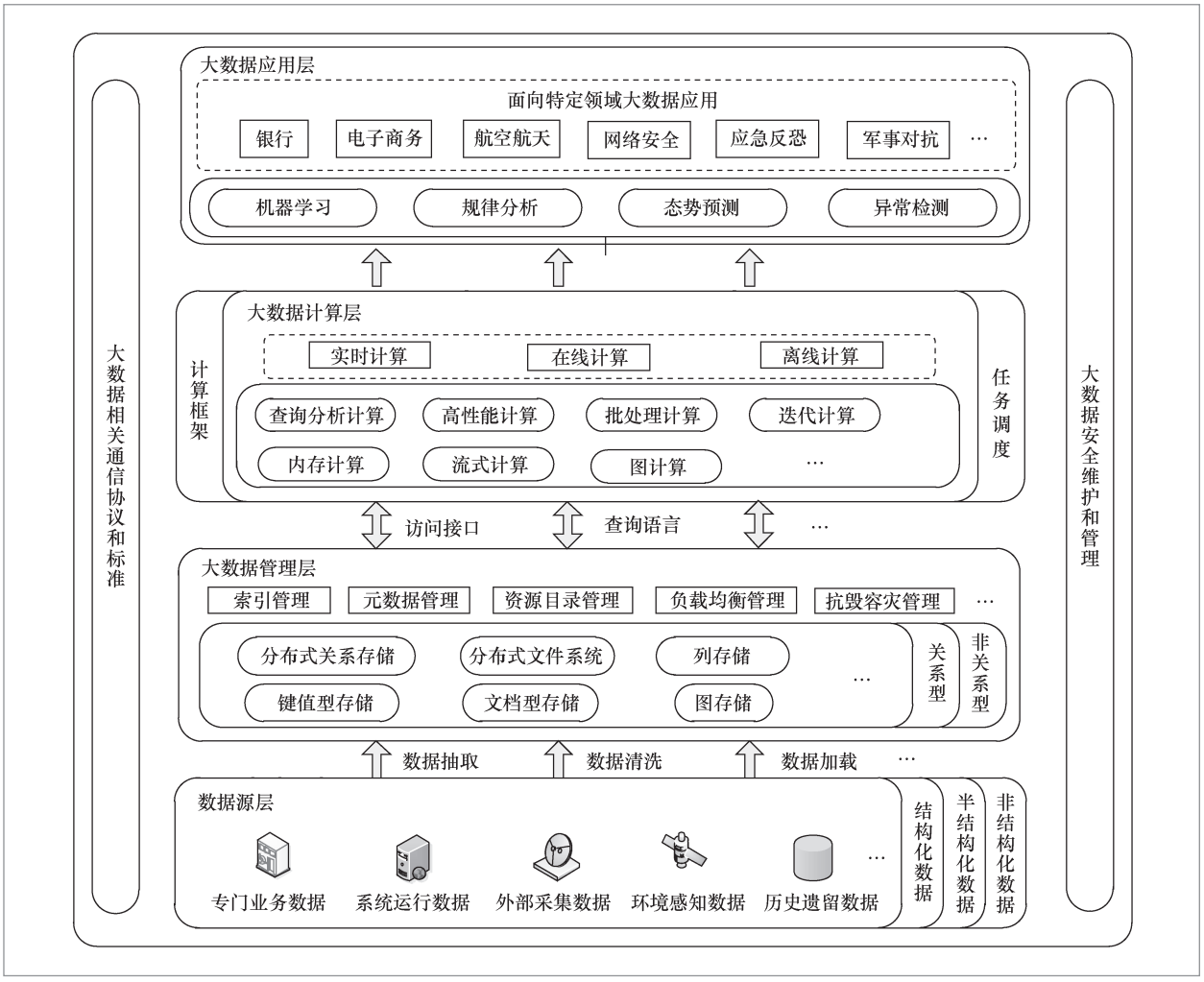

图1 面向特定领域大数据参考架构

据库实现高扩展的方法主要有两个，即数据分片和读写分离。对于不同厂商的分布式关系数据库，其集群解决方案的技术架构可能存在巨大差异，比如Oracle RAC采用共享存储（shared-disk）架构，整个RAC集群建立在一个共享存储设备之上，节点之间采用高速网络互连，而MySQL cluster则采用无共享（shared-noting）方式，整个集群由管理节点、处理节点和存储节点组成，并且处理节点和存储节点可以线性增加，从而实现高效的横向扩展能力。对于银行、电子商务的大多数业务数据可以用关系模型描述的结构化数据，考虑采用分布式关系数据库系统进行组织管理。常见的分布式文件系统包括Ceph、HDFS、Lustre、GridFS等。由于Hadoop生态系统的日益壮大，以HDFS为代表的分布式文件系统被大数据系统广泛采用。HDFS数据存储时采用数据块方式按照一定策略多数据节点冗余部署，能够有效满足“一次写入、多次读出”要求，适合航空航天侦察等领域的视频、图像等海量数据的存储管理。

NoSQL数据库大多具有无需预先定义数据模式和表结构、无共享架构、异步复制、最终一致性和软事务等特点。列存

储数据库代表性系统包括Tokyo、Redis、Voldemort等，键值存储数据库代表性系统包括 BigTable、Dynamo、HBase、Gemfire、Cassandra等，文档型数据库系统的代表包括MongoDB和Couchbase等，图数据库的代表是Neo4j和Graph等。对于不同领域的任务需求，可以根据不同类型的NoSQL数据库特点，选取适合的大数据组织管理方式。比如，列存储数据库以列族式存储实现高可扩展，适合军事对抗环境下由各类传感器传送回的水文气象和地理环境大数据管理；键值型数据库通过散列表的方式实现Key-Value，具有查找速度快的特点，适合于电子商务广告业务分析中需要海量文本检索等场景；文档型数据库弱化数据结构要求，无需像关系型数据库一样需要预先定义表结构，适合于网络安全内容分析中原始数据管理；图数据库采用图结构的顶点和边方式存储，适合于电子商务推荐系统中商品交易网络关系分析和应急反恐中敏感人物社会关系分析等场景；对于面向大数据的分布式数据管理方案，大多都提供了成熟的索引管理、元数据管理、资源目录管理、负载均衡管理和抗毁容灾管理等功能，为上层的大数据计算层提供坚实的数据基础。在同一个大数据系统中，完全可以按照实际任务需求，对不同来源和不同类型的数据，分别采用不同类型的数据组织管理方法，实现对特定领域大数据系统的灵活部署和合理开发利用。

（3）大数据计算层

提供大数据运算所需要的计算框架和任务调度等功能，负责对特定领域大数据的计算、分析和处理等。根据大数据处理多样性的需求和不同的特征维度，大数据计算模式可以大致分为查询分析计算、高性能计算、批处理计算、流式计算、内存计算、迭代计算和图计算等。根据不同领域任务要求和数据特点，可以灵活采用上述一种或多种计算模式提供实时计算、在线计算或离线计算。查询分析计算模式适合于存储于数据仓库的大数据处理，这类计算一般对计算实时性要求不高，但要求能够保证在数据体量极大时仍有效提供多维数据查询分析能力，类似系统包括Hadoop下的Hive、Facebook开发的Cassandra等。高性能计算采用高性能互联方式将众多处理器联合起来进行科学运算，这类计算大多配套专门的操作系统和软件环境，适合通过科学发现第三范式即数值模拟的方法处理网络安全密码破译、航空航天飞行器空洞试验等典型任务。以我国“天河”“太湖之光”等为代表的系列高性能计算机代表了当前高性能计算的最高水平。

大数据批处理的代表性计算模式是MapReduce，这是一种适合于网络内容安全分析、航空航天水文气象环境模拟等离线计算的处理模式。MapReduce采用“分而治之”的思想，将大量重复的数据记录处理过程总结成map和reduce两个抽象的操作，诸多系统层细节都在计算框架中自动完成。MapReduce计算模式在批处理计算方面优势明显，但是在电子商务和银行等领域的交易数据在线计算方面存在诸多不足。迭代计算模式在MapReduce基础上通过优化数据存储位置、持久化map和reduce任务、引入可缓存的map和reduce机制等手段，有效实现了迭代计算应用需求[11]。目前，最为典型的迭代计算系统是UC Berkeley AMPLab的Spark，其采用了基于分布式内存的弹性数据集模型实现快速的迭代计算。内存计算通过虚拟化和高效数据管理方法，在体系结构层面提高数据的读写效率，由于内存计算模式能够大幅减少磁盘I/O，因而在计算速度上远高于普通的MapReduce批处理。内存计算适合于实时性要求较高的大数据支撑

的军事对抗领域指挥决策和实时情报处理等场景。流式计算是一种高实时性的计算模式，该模式需要对时间窗口内的新数据实时计算处理，从而避免数据堆积和丢失。在航空航天和军事对抗等特殊领域，相关数据系统往往需要同时具备处理高流量的流式数据和大量积累的历史数据的能力，因而在提供批处理数据模式的同时，系统还需要具备高实时性的流式计算能力。目前最具代表性的流式系统包括Twitter公司的Storm和加州大学伯克利分校的Spark Steaming等。图计算是采用以顶点、边和属性为主要元素进行组织存储的一种计算模式，这种计算模式特别适合于应急反恐敏感人物分析和电子商务推荐系统等重要领域中对各类社交网络、Web链接关系分析等应用场景，代表系统包括Facebook公司的Pregel和微软公司的Trinity等。

(4)大数据应用层

构建在大数据存储架构和计算架构之上，为了满足特定领域需要而开发的面向专门任务的大数据应用系统集合。大数据应用层提供各种挖掘模型和工具，并以可视化的方式展现给最终用户。常见的大数据挖掘和分析任务包括机器学习、规律分析、态势预测、异常检测等。为了满足大数据平台多用户的特点，系统采用虚拟化方法引入多租户模式，提供各类数据的访问控制方式。面向特定领域的大数据应用层包括涉及国家民生和安全密切相关的银行、电子商务、航空航天、网络安全、应急反恐、军事对抗等特定应用。此外，还可以根据特定领域任务的需要，以安全可控的方式向第三方提供开放数据平台等功能。

(5)大数据相关通信协议和标准

负责消息传输、数据管理和功能接口交互等的相关规则和约定。大数据相关通信协议和标准通常以协议栈和标准集合的形式定义数据处理和信息交互时数据单元应使用的格式、信息单元应包含的信息与语义、连接方式、信息发送和接收的时序等。通信协议和标准均具有层次性特点，每个层次完成一部分功能，各个层次相互配合共同完成相关功能。目前大数据相关协议和标准在充分继承传统数据处理的相关方法上不断创新发展。比如在数据管理层的分布式文件系统、NoSQL键值存储，在大数据计算层的MapReduce计算框架等，因为被大多具有影响力的开源系统所使用而成为事实上的协议和标准。此外，一些国际标准化组织和机构积极制定相关协议，NIST联合微软、Oracle、IBM以及加利福尼亚大学圣地亚哥分校等机构和组织联合编写了《大数据互操作框架》等大数据指导性文件，规范大数据相关通信协议和标准；云安全联盟(CSA)自2012年以来相继发布了《大数据安全与隐私十大挑战》《大数据安全和隐私手册》等报告；国际电信联盟标准化部门(ITU-T)2015年12月批准了代号为ITU-T Y.3600的首个大数据国际电联标准“大数据—基于云计算的要求及能力”，该标准描述了如何利用云计算相关技术对大数据进行处理，并提供大数据服务。大数据相关通信协议和标准有效保证了大数据架构内部各部件之间以及大数据系统之间互联、互通、互操作，具有重要意义。

(6)大数据安全维护和管理

作为面向特定领域大数据参考架构的重要部分有着十分重要的意义。由于面向特定领域的相关大数据系统涉及国家民生和安全，容易成为犯罪分子和敌对势力的攻击目标。大数据系统的攻防双方在目标上具有不对称性特点，因而在安全性上，面向特定领域大数据具有更高和更特殊的要求。面向特定领域大数据安全既包括传

统数据平台的物理安全、系统安全、网络安全等，又包括大数据特有的数据安全、隐私防护等。由于面向特定领域大数据系统属于弱电工程，耐压值很低，所以在物理安全上除了应该考虑传统物理打击外，还需要考虑地震、水灾、火灾、电源故障等环境事故，操作失误或错误等人为事故以及数据的外部电磁干扰和线路截获等。主机和数据中心级别的冗余通常是处理物理安全的有效手段。面向特定领域的大数据参考架构中对数据存储资源和计算资源均采用分布式冗余部署，具备较高的物理安全防护级别。系统安全主要表现在大数据运行平台使用的核心软硬件是否能够自主可控。同时应避免系统关键软硬件的崩溃、损坏和供应链的断裂导致的对系统存储、处理和传输的破坏和损失，保证大数据系统信息处理和传输的安全[12,13]。面向特定领域大数据系统核心软硬件产品包括操作系统、数据管理平台、数据计算平台、处理器和关键板卡等，在同等条件下应优先采用国内自主品牌，对于开源软件应具备吸收引进后的二次开发能力。网络安全体现在大数据系统内部网络和开放数据共享平台上的系统信息的安全，包括用户口令鉴别、用户存取权限控制、安全审计、计算机病毒防治等。数据安全主要体现在数据的分级访问控制、数据加密存储和传输、数据完整性和数据真实性等。此外，面向特定领域大数据隐私保护也非常重要，比如在军事对抗领域，很多关键信息都涉及大数据隐私保护问题，一旦出现隐私泄露将可能造成严重的后果。

对上述面向特定领域大数据参考架构中不同部件进行角色模型分析，数据源扮演大数据提供者角色，大数据管理层和大数据计算层扮演大数据运行框架提供者角色，大数据应用层扮演大数据消费者角色，大数据安全和维护管理扮演大数据安全维护者角色，大数据相关通信协议、标准以及大数据系统维护者扮演大数据协调运维者角色。面向特定领域大数据参考架构各个部件密不可分，角色互补，形成一个有机统一总体。

4 面向特定领域大数据标准体系

大数据标准之间存在着相互依存、相互衔接、相互补充、相互制约的内在联系，通过建立大数据标准体系，可以形成科学的有机整体[10]。同时，大数据标准体系是大数据标准化研究发展的基石，对标准化研制工作具有重要的指导意义。

由于特定领域大数据在安全性、可靠性、时效性、多样性和数据质量等方面有着更高的要求，所以在相关标准的制定过程中应区别于传统领域，重点关注其业务特点，突出特定领域的特殊性，增强数据质量、数据安全等相关标准在体系中的地位，有针对性地进行特定领域大数据标准体系的建设工作。

本文针对特定应用领域的大数据的内在需求，结合国内外大数据标准化情况、大数据技术发展现状以及大数据未来发展的趋势，充分考虑特定领域的业务特点和特殊要求，提出面向特定领域大数据标准体系，见**表1**。

特定领域大数据标准体系主要由6个部分的标准组成，分别为：基础标准、技术标准、平台标准、应用标准、安全标准、质量标准。

（1）基础标准

基础标准是大数据标准体系的根基，是解决跨平台异构数据互操作问题的前提，主要用于规范大数据基础层面标准。基础标准主要包括术语、大数据参考模型、大数据架构标准。

表1 面向特定领域大数据标准体系

一级分类	二级分类	标准名称
基础	术语	
	大数据参考模型	
	大数据架构标准	
	元数据	
技术	处理生命周期技术	数据收集
		数据预处理
		数据存储与管理
		数据处理与分析
		数据访问与接口
		数据可视化
	分布式技术	分布式存储
		分布式计算
	互操作技术	不同数据模型间互操作
		不同平台架构间互操作
	大数据集描述	
平台	基础平台	平台基础设施
		预处理类平台
	管理平台	数据库平台
	计算平台	分布式计算平台
		智能分析平台
	系统平台	
应用	大数据资产管理	
	数据共享模式	
	特定应用领域	银行金融大数据
		电子商务大数据
		航空航天大数据
		网络安全大数据
		应急反恐大数据
		军事对抗大数据
安全	物理安全	环境安全
		设备安全
	网络安全	传输安全
		入侵防护
	系统安全	存储安全
		计算安全
		权限管理
	内容安全	隐私保护
		数据加密
质量	质量模型	
	数据质量测试	
	产品和平台测试	测试场景
		测试指标
		测试工具

目前国际上比较有影响力的基础类标准是NIST发布的《大数据互操作框架》[14]，该标准对基础层面的大数据术语定义、参考模型和大数据架构等都进行了详细阐述，并提出了大数据标准体系发展路线图。近些年，中国电子技术标准化研究院针对我国大数据技术和标准发展现状，相继发布了两个版本的《大数据白皮书》[10]，

补充了国内基础标准研究工作，为特定领域大数据基础标准相关研究提供了重要依据。

（2）技术标准

该部分标准主要针对大数据相关技术进行规范，包括大数据集描述、处理生命周期技术、分布式技术和互操作技术4类标准。其中，处理生命周期技术主要针对数据收集、数据预处理、数据存储与管理、数据处理与分析、数据访问与接口、数据可视化等方面进行规范。同时，面向特定领域大数据实际应用需求，还提出了分布式技术和互操作技术标准，以规范解决数据规模大和异构性问题的方法。其中，分布式技术包括分布式存储和分布式计算等规范，互操作技术包括不同数据模型间互操作和不同平台架构间互操作等技术规范。

对于特定领域，需要对相关技术进行更加严格的规范，以保证其对时效性、多样性和数据质量的强烈需求。同时，该部分也是大数据领域研究的核心内容，在学术界得到普遍关注，催生出诸多不同技术流派。在大数据处理生命周期技术标准领域，国外多家知名标准化研究机构在数据收集、存储、处理、可视化等方面发布了诸多技术标准，例如ISO/IEC JTC1/SC32工作组推出的数据管理标准《信息技术 数据管理参考模型》，美国存储网络工业协会（SNIA）颁布的大数据云存储标准《信息技术 云数据存储和管理》。同时，随着特定领域数据量的日益激增，分布式技术和互操作技术已经是大数据解决方案的关键环节，也一直受到国外标准研究机构的关注，如ISO/IEC JTC1发布的《信息技术 开放分布式处理》和NIST发布的《大数据互操作框架》[14]。

（3）平台标准

平台标准主要针对大数据相关技术平台进行规范，包括基础平台、管理平台、计算平台、系统平台。其中基础平台包括平台基础设施、预处理类平台等；管理平台包括数据库平台等；计算平台包括分布式计算平台、智能分析平台等。

由于日益膨胀的互联网数据对大数据平台的迫切需求，相关技术方案在互联网行业有着广泛的研究基础，例如谷歌公司的Hadoop计算平台和HDFS存储平台、开源数据库平台HBase、UC Berkeley AMP lab开源的SPARK平台、适用于图数据计算的Giraph平台等。大数据平台的解决方案种类繁多，新技术新架构层出不穷，很难形成统一的标准规范，该领域的标准化研究相对薄弱。同时，特定领域大数据应用对高安全、高可靠、强实时的需求迫切，对成熟平台技术的依赖程度较高，急需一整套健全的平台标准规范。

（4）应用标准

应用标准主要针对特定领域大数据相关技术应用与服务产品进行规范，主要包括大数据资产管理、数据共享模式、特定应用领域等。其中特定应用领域包括银行金融大数据、电子商务大数据、航空航天大数据、网络安全大数据、应急反恐大数据、军事对抗大数据等。

相比于传统应用领域，特定领域大数据在安全性、可靠性和实时性等方面有着更高的需求，急需规范大数据资产的管理和共享模式以及特定领域的应用标准。中国电子技术标准化研究院在开放数据集、数据交易平台、数据开放共享平台、领域应用数据等方面拟开展标准化研究工作，尤其是领域应用大数据标准，现已成立工业大数据专题组，非常重视大数据在各行各业的应用标准化研制，对特定应用领域制定大数据应用标准具有重要的指导作用。

（5）安全标准

数据安全是信息安全的一个分支，

NISTIR 7298技术报告描述数据安全为旨在防止对数据的未授权（事故或有意）篡改、破坏和泄露[12]。大数据背景下的安全标准主要由物理安全、网络安全、系统安全、内容安全等标准组成。其中物理安全标准主要指环境安全和设备安全等标准规范；网络安全标准主要指大数据传输安全、入侵防护等标准规范；系统安全标准包括存储安全、计算安全和权限管理等标准规范；内容安全标准包含隐私保护、数据加密等标准规范。

对于特定领域大数据应用，数据安全问题或隐私泄露可能引发灾难性后果，导致这些领域不计成本地追求系统的高安全和高可靠性，并且积极投身于数据安全标准化的制定工作中。例如在航空领域，中国民用航空局在《民用航空空中交通管理管理信息系统技术规范》中提到了数据安全问题；在金融领域，中国人民银行发布的《银行卡联网联合技术规范》中，提到了数据的安全传输部分，这些都为大数据背景下的安全标准研究提供了参考。

同时在国际上，安全标准的研究也得到了普遍关注。ISO TC68/SC2工作组专门为金融领域制定了标准性文件《银行领域—安全和其他金融服务—金融系统安全架构》，规范了金融数据安全的加密方法、密钥管理、API、协议等。JTC1/SC27工作组在《信息技术—安全技术—隐私框架》中定义了数据隐私架构的基本元素组成和相关基本概念解释，并指出了数据隐私的11个基本原则。鉴于特定领域大数据对高安全、高可靠性的特殊需求，安全标准对规范大数据的安全管理和使用具有重要意义。

（6）质量标准

针对特定领域对高质量数据和高可靠应用的特殊需求，本文将质量标准作为面向特定领域大数据标准体系的重要组成部分之一。质量标准主要针对大数据的数据质量和产品质量进行规范，包括质量模型、数据质量测试、产品和平台测试等标准。其中数据质量测试以及产品和平台测试又包括测试场景、测试指标、测试工具等方面标准规范。

国内外大数据标准化研究机构在质量标准方面关注较少，一般某些特定领域的研究机构在该部分标准中成果较多，根据实际应用中的质量问题和测试方法，发布了质量标准规范文件。例如在地理信息系统领域，ISO/TC 211于2003年发布了地理信息数据质量评估过程标准《地理信息—质量评估过程》，并于2013年对其进行修订，发行新版标准《地理信息数据质量》。我国国家测绘局在2010年也发布了地理信息数据测试标准文件《基础地理信息数据库测试规程》，规范了地理信息数据库的测评，整体提升了地理信息数据质量。金融、军事、航空航天等特定领域对于质量模型和质量测评等相关标准的制订，可参照国内外先进标准成果，研制大数据质量标准规范，完善特定领域大数据标准体系。

5 结束语

与国家民生和安全密切相关的银行、电子商务、航空航天、网络安全、应急反恐、军事对抗等特定领域大数据是国家安全的重要战略资源。研究面向特定领域的大数据标准是实现相关大数据系统互连、互通、互操作的重要前提，是实现大数据系统中重要信息跨部门、跨领域、跨平台安全共享的坚实基础。本文系统分析了特定领域国内大数据研究现状，结合不同领域任务大数据特点和需求差异，提出了面向特定领域大数据参考架构，并对架构的各个功能部件从技术基础、角色模型和主

流产品等角度进行系统分析。进一步，结合大数据架构模型和特定领域大数据发展需求，提出了面向特定领域大数据标准体系框架，并对相关标准进行有效分析。面向特定领域大数据参考架构和标准体系框架将随用户需求的不断拓展和信息技术的不断进步而持续演化，为面向特定领域大数据相关技术发展提供重要技术支撑。

参考文献:

[1] MANYIKA J, CHUI M, BROWN B, et al. Big data: the next frontier for innovation, comptetition, and productivity[J]. Analytics, 2011.

[2] LOHR S. The age of big data[J]. The New York Times, 2012, 16(4): 10-15.

[3] 李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊，2012，27(6)：5-15.
LI G J, CHENG X Q. Big data research: a major strategic field for future technology and economic development: research status and scientific thinking of big data [J]. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 5-15.

[4] 王元卓，靳小龙，程学旗. 网络大数据：现状与展望[J]. 计算机学报，2013，36(6)：1125-1138.
WANG Y Z, JIN X L, CHENG X Q. Network big data: present and future[J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.

[5] TOLLE K M, TANSLEY D, HEY A J G. The fourth paradigm: data-intensive scientific discovery [point of view][J]. Proceedings of the IEEE, 2012, 99(8): 1334-1337.

[6] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(1)：146-169.
MENG X F, CI X. Big data management: concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[7] FITZPATRICK G J, WOLLMAN D A. NIST interoperability framework and action plans[C]//2011 IEEE Power & Energy Society General Meeting, July 28-29, 2010, Providence, RI, USA. New Jersey: IEEE Press, 2010: 1-4.

[8] 饶玮，蒋静，周爱华，等. 面向全球能源互联网的电力大数据基础体系架构和标准体系研究[J]. 电力信息与通信技术，2016(4)：1-8.
RAO W, JIANG J, ZHOU A H, et al. Research on power large data infrastructure and standard system for global energy internet[J]. Electric Power Information and Communication Technology, 2016(4): 1-8.

[9] 韩晶，王健全. 大数据标准化现状及展望[J]. 信息通信技术，2014(6)：38-42.
HAN J, WANG J Q. Status and prospect of large data standardization[J]. Information and Communications Technolog, 2014(6): 38-42.

[10] 中国电子技术标准化研究院. 大数据标准化白皮书V2.0[R]. 北京：工业和信息化部，2015.
China Electronics Standardization Institute. Big data standardization white paper V2.0[R]. Beijing: Ministry of Industry and Information Technology, 2015.

[11] 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，2014，25(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, 25(9): 1889-1908.

[12] 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)：246-258.
FENG D G, ZHANG M, LI H. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258.

[13] 叶润国，胡影，韩晓露，等. 大数据安全标准化研究进展[J]. 信息安全研究，2016，2(5)：404-411.
YE R G, HU Y, HAN X L, et al. Research progress on standardization of large data security[J]. Information security research, 2016, 2(5): 404-411.

[14] NIST Big Data Working Group. NIST big data interoperability framework [R]. US: National Institute of Standards and Technology, 2015.

作者简介

林旺群（1983-），男，博士，北京系统工程研究所助理研究员，主要研究方向为大数据、分布式计算、机器学习。

高晨旭（1990-），男，北京系统工程研究所研究实习员，主要研究方向为大规模图数据管理、大数据挖掘与分布式计算、大数据体系架构等领域的科学研究和工程开发，发表论文多篇，申请专利5项，目前已有3项获得授权。

陶克（1988-），男，博士，北京系统工程研究所助理研究员，主要研究方向为网络信息系统、大数据挖掘、信息检索、用户建模与个性化等，在WWW、IJCAI、UMAP、IJDL等国际会议与国际期刊上发表论文10余篇。

邓波（1973-），男，博士，北京系统工程研究所研究员，长期从事软件体系结构、分布式计算与数据处理以及软件质量保证等方向的技术研究工作，是所在单位计算机软件总体技术的学术带头人，目前是中国计算机学会高级会员、大数据专家委会员委员。

包阳（1978-），男，北京系统工程研究所副研究员，主要研究方向为大数据、机器学习、软件测试。

收稿日期： 2017-02-22