DOI: 10.13382/j.jemi.2015.04.001

大数据:内涵、技术体系与展望*

彭 宇 庞景月 刘大同 彭喜元

(哈尔滨工业大学电气工程及自动化学院自动化测试与控制系 哈尔滨 150080)

摘 要: 随着复杂系统测试、试验和状态监测产生的数据呈级数增长,大数据逐渐成为各种工业领域的研究热点。为此,从大数据定义、产生、特性的角度阐述大数据的内涵,重点强调大数据产生三要素的彻底变革,并按照大数据处理流程——产生、存储、预处理、分析及挖掘、呈现,归纳得出大数据处理的通用技术体系,分析了技术体系中各环节技术的发展现状。最后,从数据科学、工业 4.0 以及信息物理系统的角度,阐述大数据发展的趋势,并分析了大数据发展的挑战。

关键词: 大数据; 工业 4.0; 信息物理系统; Hadoop; 云存储

中图分类号: TP806 **文献标识码:** A **国家标准学科分类代码:** 510.4030

Big data: connotation, technical framework and its development

Peng Yu Pang Jingyue Liu Datong Peng Xiyuan

(Department of Automatic Test and Control, Harbin Institute of Technology, Harbin 150080, China)

Abstract: The large amount of data increases significantly in the tests, experiments and condition monitoring of complex system, the big data becomes the research hotspot in various industrial fields. Thus, connotation of big data is concluded from its definition, production and characteristics, etc. And this work also focuses on the three generation elements to show the inevitability and specialty of the era of big data. In accordance with the big data processing procedures involving its production, storage, pretreatment, analysis, mining and presentation, this paper summarizes the universal technology framework for the big data processing. Moreover, the development status of technologies involved in the framework is also analyzed in detail. Finally, from the view of data science, industrial 4.0 and cyber-physical system, the trend and challenges are further explained.

Keywords: big data; industrial 4.0; cyber-physical system; Hadoop; cloud storage

1 引 言

近些年,由于计算机、物联网等信息化技术以及传感技术的发展,使得现代生活中出现了“一切皆可数据化”的思维[1],数据的产生方式由“人机”、“机物”的二元世界向着融合社会资源、信息系统以及物理资源的三元世界转变[2-3],数据规模呈膨胀式发展。例如,互联网领域中[4],谷歌搜索引擎的每秒使用用户量达到 200 万,Twitter 每天的推特量已经超过了 3.4 亿;科研领域中,仅某大型强子对撞机在一年内积累的新数据量就达到 15 PB 左右[5];电子商务领域中,作为世界连锁性企业沃尔玛,其每小时可处理的客户交易可超过 100 万笔,相应为数据库注入超过 2.5 PB 的数据;航空航天领域中,仅一架双引擎波音 737 在横贯大陆飞行的过程中,传感器网络便会产生近 240 TB 的数据。综合各个领域,目前积累的数据量已经从 TB 级上升至 PB、EB 甚至已经达到 ZB 级别,其数据规模已经远远超出了现有计算机所能够处理的量级,而且全球的数据量正以每 18

收稿日期: 2015-01 Received Date: 2015-01

* 基金项目:国家自然科学基金(61301205)、高校博士点基金(20112302120027)、部委预研重点基金(9140A17050114HT0-1054)资助项目

个月翻一倍的速度呈膨胀式增长[1]。对此全球著名的管理咨询公司 McKinsey 首先提出了“大数据时代”的到来[6]，其认为数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。

“大数据”这一词语并不是近几年才出现，其最早是由美国著名未来学家 Alvin Toffer 在《第三次浪潮》一书中提出，其将大数据赞颂为“第三次浪潮的华彩乐章”[7]；而 2000 年 Diebold 所撰写的论文[8]是大数据第一次出现在学术期刊。但“大数据”并不等同于“大规模数据”，Viktor Mayer – Schönberger 和 Kenneth Cukier 在《“大数据”时代》[1]中提出：大数据应具有 4V 特性，即 Volume（数据量大）、Velocity（数据处理速度快）、Variety（数据具有多样性）和 Value（数据价值密度低）。

大数据时代的到来颠覆了工业界、学术界对传统数据的认知，同时也引起了数据获取、存储、分析、挖掘以及可视化等技术的变革[9-12]。例如，在大数据背景下，新型数据库的开发，大规模存储设备的研制，云存储服务方案的提出等，大数据相关技术的更新换代为大数据价值的快速、有效挖掘提供了技术基础。

与此同时，大数据以及其相关技术的发展也将成为改变目前人类生产以及生活方式的重要基础。如美国国家科学基金会（National Science Foundation，NSF）科学家 Helen Gill 在 2006 年提出了信息物理系统（Cyber-Physical System，CPS）的概念[13]，CPS 将实现感知基础上的人、机、物的深层融合，而基于大数据的分析是 CPS 系统得以智能运行的关键；2013 年举办的“Hannover Messe 2013”上，由产官学专家组成的德国“工业 4.0 工作组”发表了最终报告——《保障德国制造业的未来：关于实施“工业 4.0”战略的建议》[14]，宣告以物联网和制造业服务化为特征的第四次工业革命的到来，其强调第四次工业革命将以大数据分析以及 CPS 为基础，最终实现“智能工厂”与“智能生产”，解脱对人的依赖性。以上无论是 CPS 还是工业 4.0 的提出，其本质都是制造业基于数据分析的转型，基础都是大数据的分析。由此可见，大数据的发展以及有效运用将引起人类生产以及生活方式的巨大变革，只有在大数据背景下迅速抓住机遇，并将其转化为决策信息，才能在未来市场以及科技的竞争中取得胜利。

虽然目前大数据在商业领域已经得到广泛关注，相关概念愈炒愈热，但是对于有效的大数据处理技术体系认识不足，且并未清晰阐述大数据在工业领域、科学领域的发展趋势。因此，本文在具体介绍大数据内涵的基础上归纳总结了大数据处理（包含大数据采集、存储以及挖掘等）的技术体系，并从数据科学、工业 4.0 以及 CPS 的角度，对大数据的发展进行了展望。

2 大数据内涵

2.1 大数据定义

大数据自提出至今得到广泛关注，其并无统一的定义，由于大数据是相对概念，因此目前的定义都是对大数据的定性描述，并未明确定量指标。维基百科中指出，大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间限制的数据集[15]；全球著名的管理咨询公司 McKinsey 则将数据规模超出传统数据库管理软件的获取、存储、管理以及分析能力的数据集称为大数据[6]；研究机构 Gartner 将大数据归纳为需要新处理模式才能增强决策力、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[16]；徐宗本院士则在第 462 次香山科学会议上的报告中，将大数据定义为“不能够集中存储、并且难以在可接受时间内分析处理，其中个体或部分数据呈现低价值性而数据整体呈现高价值的海量复杂数据集[17]”。

虽然以上关于大数据定义的定义方式、角度以及侧重点不同，但是所传递的信息基本一致，即大数据归根结底是一种数据集，其特性是通过与传统的数据管理以及处理技术对比来突显，并且在不同需求下，其要求的时间处理范围具有差异性，最重要的一点是大数据的价值并非数据本身，而是由大数据所反映的“大决策”、“大知识”、“大问题”等。

2.2 大数据产生

“大数据”并不是一个空的概念，其出现对应了数据产生方式的变革。如果从事件发生的三要素来看，需要具备时间、地点以及人物要求，事件才能完整。但是对于“大数据”而言，其产生方式已经分别在这三要素上突破了限制，即传统数据产生方式的变革导致了具有 4V 特性的“大数据”的出现。为此，本小节将从事件发生三要素的独特视角，清晰、全面地分析大数据产生的特点以及变化。

大数据的产生如图 1 所示。

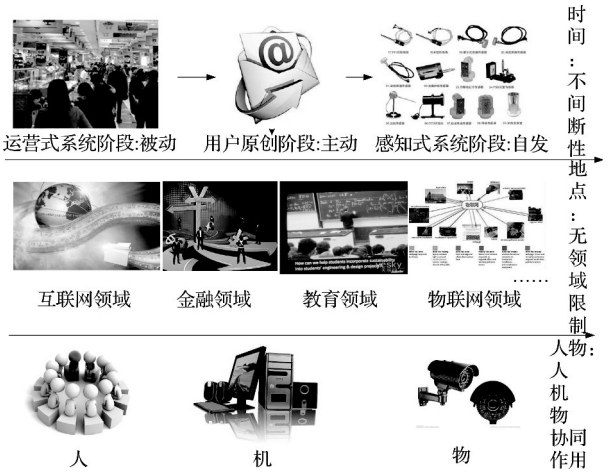

图 1　大数据的产生

Fig. 1　The production of big data

1) 时间: 不间断性

传统数据是伴随着一定的运营活动而产生的，并在产生后存储至数据库，如超市只有用户发生购买行为之后才会产生交易信息，该阶段数据的产生是被动的，具有这种数据产生方式的阶段被称为“运营式系统阶段”; 随着互联网技术的发展，以智能移动终端以及社交平台为媒介，大量通话以及聊天记录的产生标志着“用户原创阶段”的到来，该阶段数据产生呈现主动性; 而后，云计算、物联网以及传感技术的发展，使得数据以一定的速率源源不断的产生，该阶段的数据呈现自发性，该阶段被称为“感知式系统阶段”。通过以上分析可知，数据产生方式经历了被动、主动以及自发式的历程[18]，其已经脱离了对活动的依赖性，突破了传统时间的限制，具备了持续不间断产生的特性。

2) 地点: 无领域限制

大数据已经出现在各种领域，包括互联网[19]、金融[20]、医疗[21-22]、教育[23]、科研[24]、航空航天[25]以及物联网[26]等。例如，互联网领域的网络点击流、网络日志、电子邮件以及交易记录，金融领域的股票交易、用户消费记录以及账户信息，物联网领域中大量分布的传感器感知的环境信息、设备信息，科研领域中仿真实验数据、实验报告、论文等，这些都是构成大数据的重要组成部分。但产生大数据的领域并不局限于此，其甚至已经分布在了我们能够想象到的生产生活的各种领域。例如，学生的考试成绩、学号信息，购物清单以及手机短信、通话记录等，其都会形成数据并保存。由此可见，领域的扩展已经为“大数据”的形成提供了重要基础。

3) 人物: 人、机、物协同作用

众所周知，人物是传统事件发生的重要因素，而对于数据的产生，其主体已经从传统的“人”的概念扩展到“人”、“机”、“物”以及三者的融合。首先，“人”指的是人类的活动，包括人的日常消费，使用移动互联网、移动设备终端等; 其次，很重要的一部分数据来源于“机”，即信息系统本身，计算机信息系统产生的各类数据，其以文件、多媒体等形式存在，包括计算机虚拟镜像、内容拷贝以及数据备份等; 另外，“大数据”同样也来源于“物”，即我们所处的物理世界，其涉及到各种具有采集功能的设备，如摄像头、医疗设备、传感器等。而且，随着云计算、物联网等信息技术的发展，“人”、“机”及“物”的规模逐渐扩大，相互之间的作用越来越明显，数据的产生方式也已经由“人机”或“机物”的二元世界向着融合社会资源、信息系统以及物理资源的三元世界转变[2-3]。

通过以上分析可知，数据产生的三要素已经发生了历史性的变革，人、机、物协同作用下，不间断、无领域限制的数据产生方式已经突破了传统数据的概念，其必然导致数据性质的变革，这也就衍生出了新的概念——“大数据”。

2.3　大数据特性

在“大数据”的定义中，已经包含了大数据的特性，即数据量大、处理速度要求快、价值密度低等，目前对于大数据的特性认可度较高的是 3V 特性: 即数据的规模性(Volume) 、高速性(Velocity) 以及数据结构多样性(Variety) ，而在此基础上已经有不同的公司以及研究机构对其进行了扩展，大数据特性描述的演化如表 1 所示。

根据上表可以看出，随着时间的演化，业界对于大数据的认识也更深入、全面。除以上对大数据特性的通用性描述之外，不同应用领域的大数据的具体特性也存在差异性。如互联网领域需要实时处理和分析用户购买行为，以便及时制定推送方案，返回推荐结果来迎合和激发用户的消费行为，精度以及可靠性要求较高; 医疗领域需要根据用户病例以及影像等信息判断病人的病情，由于其与人们的健康息息相关，所以其精度以及可靠性要求非常高。

表 1 大数据特性描述的演化情况

Table 1 The description evolutions of big data characteristics

特点	提出时间	作者或者机构	内涵
规模性(Volume)	2001	DougLaney[27](Gartner Meta Group research)	数据量大
高速性(Velocity)			数据分析和处理速度快
多样性(Variety)			数据类型多样
价值性(Value)[28]	2012	IDC	价值稀疏性
真实性(Veracity)[29]	2012	IBM	数据反应客观事实
易变性(Variability)[30]	2012	Brian Hopkins 和 Boris Evelson(Forrester)	大数据具有多层结构

表 2 列举了不同领域大数据的具体特点[18]以及应用案例。

表 2 不同领域大数据的具体特点

Table 2 The specific features of big data in different areas

领域	用户数目	响应时间	数据规模	可靠性要求	精度要求	应用案例
科学技算	小	慢	TB	一般	非常高	大型强子对撞机数据分析
金融	大	非常快	GB	非常高	非常高	信用卡营销
医疗领域	大	快	EB	非常高	非常高	病历、影像分析
物联网	大	快	TB	高	高	迈阿密戴德县的智慧城
互联网	非常大	快	PB	高	高	网络点击流入侵检测
社交网络	非常大	快	PB	高	高	Facebook、QQ 等结构挖掘
移动设备	非常大	快	TB	高	高	可穿戴设备数据分析
多媒体	非常大	快	PB	高	一般	史上首部大数据制作的电视剧《纸牌屋》

由表 2 可以看出，不同应用领域的数据规模、用户数目以及精度要求等均存在较大差异。例如，互联网领域与人的正常活动息息相关，其数据量达 PB 级别，用户数目非常大，而且以用户实时性请求为主。与此不同，在科研领域中，其用户数目相对较少，产生的数据量级别在 TB 级。因此，对大数据后续的分析以及处理必须因地制宜，才能实现大数据价值的最大化。

3 大数据技术体系分析

大数据出现颠覆了传统数据处理的一系列技术，如大数据获取方式的改变导致数据规模迅速膨胀，相对于传统的数据库系统，其索引、查询以及存储都面临着严峻的考验，而且怎样快速地完成大数据的分析也是传统数据分析方法无法解决的。为此针对规模大、速度快、数据多样、价值密度低的大数据，本文将大数据处理技术体系总结如图 2 所示。

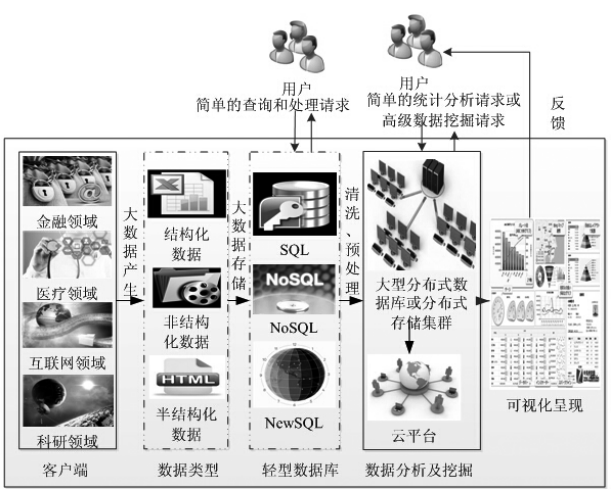

图 2 大数据处理技术体系

Fig. 2 The technical framework for big data processing

如图 2 所示，大数据处理技术体系主要涉及大数据的采集技术、存储技术、分析及挖掘技术、可视化呈现技术 4 个部分。

1) 大数据的采集：来自于不同领域的大数据，其特点、数据量以及用户数目不同，按照结构特点，可划分为 3 种类型：结构化数据、半结构化数据以及非结构化数据。大数据采集的挑战是并发数高、流式数据速度快。

2) 大数据的存储：改进的轻型数据库可用于完成大数据的存储并响应用户的简单查询与处理请求；而当数据量超过轻型数据库的存储能力时，

则需要借助于大型分布式数据库或存储集群平台，且随着互联网技术和云计算技术的发展，建立在分布式存储基础上的云存储已经成为大数据存储的主要趋势。大数据存储的主要挑战是数据异构、结构多样、规模大。

3) 大数据的分析及挖掘: 大数据的分析涉及简单的统计分析以及分类汇总，其挑战在于导入数据量大，查询请求多; 而大数据挖掘涉及数据的分类、聚类、频繁项挖掘等，其算法复杂，计算量大。

4) 大数据可视化: 大数据的挖掘及分析结果将在显示终端以友好、形象、易于理解的形式呈现以供专业人士分析结果的准确性或为用户提供决策信息支持[31]。大数据呈现的挑战在于数据维度高、呈现需求多样化。

大数据处理环节中各技术功能的相互配合使用可为大数据价值的有效实现提供技术基础。

3.1　大数据获取

不同领域对应的数据采集方法以及工具也不同，如互联网领域中，用于日志采集的大数据获取工具，Hadoop 的 Chukwa[32]、Cloudera 的 Flume、Facebook 的 Scribe、LinkedIn 的 Kafka 等，用于网络数据采集的网络爬虫或网站公开 API 等方式; 物联网领域中，用于数据感知的 MEMS 传感器、光纤传感器、无线传感器等。数据产生以及采集方式的发展为大数据的获得提供了重要基础。

获取的大数据按照结构的不同，可分为结构化数据、非结构化数据以及半结构化数据，其特点如表 3 所示。

表 3　不同数据类型的特点分析

Table 3　The characteristics of different data types

数据类型	举例	特点
结构化数据 (structured)	二维表	先有结构后有数据、行数据
半结构化数据[33] (semi-structured)	HTML 文档、XML 文档[34]、SGML 文档	先有数据后有模式、无规则性结构
非结构化数据 (unstructured)	图形、文本、声音、视频	模式具有多样性

其中结构化数据可用二维表结构来逻辑表达实现，一般采用数据记录存储，而非结构化数据一般采用文件系统存储。据统计，目前大数据的构成中非结构化数据与半结构化数据占据主体地位，且非结构化数据以及半结构化数据规模呈膨胀式增长。而由于半结构化数据以及非结构化数据的模式多样，并无强制性的结构要求，为大数据的存储、分析、呈现带来巨大挑战[35]。

3.2　大数据存储

3.2.1　轻型数据库

对应于大数据获取环节，当数据量在轻型数据库存储能力范围内，且仅为响应用户简单的查询或者处理请求的情况下可将数据存储至轻型数据库内。图 2 中对应的大数据存储的轻型数据库[36]包括关系型数据库 SQL、非关系型数据库 NoSQL 以及新型数据库 NewSQL，通过轻型数据库可响应简单的大数据查询以及处理需求，与此相关的大数据轻型数据库总结如表 4 所示。

表 4　用于大数据存储的轻型数据库

Table 4　The lightweight databases for big data storage

分类	举例		
	现属公司	数据库名称	主要特点
SQL	EMC	Greenplum[37]	关系型数据库集群
	HP	Vertica[38]	分布式 MPP 列式数据库具有数据库内分析功能
	Teradata	Aster Data[39]	结合 SQL 与 MapReduce
NoSQL	Google	HBase[40]	分布式、面向列、开源
	10gen	MongoDB[41]	操作简单、完全免费、源码公开、随时下载
	Facebook	Cassandra[42]	分布式网络服务、高扩展
	VMware	Redis	超高性能的键值数据库
NewSQL	Google	Spanner[43]	可扩展、全球分布式
	Google	Megastore[44]	融合 NoSQL 的可扩展性和传统的关系型数据库
	Google	F1[45]	动态扩展、并行 SQL 执行引擎

关系型数据库 SQL 是把所有的数据都通过行和列的二元表现形式表示出来，其具有非常好的通用性和非常高的性能，但是 SQL 并不适

宜于以下情况: 大量数据的查询、简单查询需要快速返回结果、非结构化数据的应用等,所以用于大数据存储的关系型数据库需要做出不同的改进才能满足大数据的存储以及查询要求,如表 4 所示的现所属 EMC 公司的 Greenplum,其并不是简单的关系型数据库,而是属于关系型数据库集群,且采取了 MPP 并行处理架构,查询速度快,数据装载速度快,批量 DML 处理快; Vertica 是具有 MPP 架构的分布式列式存储关系型数据库,其属于高效能、低成本的海量数据实时分析数据库; 而 Teradata 公司开发的 Aster Data,其提供两种分析框架,SQL 与 MapReduce,并具有近似线性的扩展能力。

NoSQL(NoSQL = Not Only SQL) [46] ,意即“不仅仅是 SQL”,相对于 SQL,NoSQL 具有非常高的读写性能、灵活的数据模型以及高可用性,NoSQL 为非关系型数据库,主要分为键值(Key-Value) 存储数据库、列存储数据库、文档存储数据库、图形(Graph) 数据库。上表中 HBase 与 Cassandra 属于列存储数据库,MongoDB 属于文档型数据库,Redis 属于键值(Key-Value) 存储数据库。

NewSQL 一词是由 451 Group 的分析师 Matthew Aslett 在文献 [47] 中提出,其是对各种新的可扩展、高性能数据库的简称,这类数据库不仅具有 NoSQL 对海量数据的存储管理能力,还保持了传统数据库支持 ACID 和 SQL 等特性,表 4 中的 Google 推出的 Spanner、Megastore 以及 F1 等均可归为 NewSQL 类型。

3.2.2 大数据存储平台

当用户提出大数据分析以及复杂的挖掘请求或数据量已经远超过轻型数据库的存储能力时,应将大数据导入大型分布式存储数据库或者分布式存储集群。目前典型的大数据存储平台包括 Infobrignt、Hadoop (Pig 和 Hive [48]) 、YunTable、HANA [49] 以及 Exadata [50] 等,以上数据库中除 Hadoop 外均可满足大数据的在线分析请求。

而随着宽带网络技术、WEB2.0 技术、应用存储、集群技术、存储虚拟化技术的发展,云环境下的大数据存储将成为未来数据存储的发展趋势。云存储并不是存储,而是一种服务,其将数据放在云上以供使用者在不同的时间、地点、通过任何可联网的设备对数据进行获取。目前很多公司推出的网盘便是云存储的应用实例,其一经推出便得到了大家的广泛青睐,包括迅雷快传、115 网盘、163 网盘、腾讯微云、新浪微盘、360 云盘、百度云等,虽然各个网盘的上传、下载速度以及容量等具有差异性,但网盘的推出以及流行反映了云存储的良好发展趋势。现在很多公司也相继推出了云存储平台,如 Amazon S3 [51] 、Microsoft 的 Azure [52] 等,云存储平台的出现为企业以及研究机构带来了便利,其可利用云存储平台开发自己的云存储系统,但是对应于云存储,成本以及安全性、隐私性的问题也是未来需要突破的重点。

3.3 大数据查询及处理需求

由于大数据所属领域不同,其查询及处理需求的分类不同。例如,互联网行业按照其业务需求,可以将大数据处理技术分为在线、近线以及离线 [18] ,其中在线模式下数据的处理时间一般限定在毫秒甚至是微秒范围内,而离线模式下数据的处理时间可延长至以天为单位,近线模式的数据处理时间则位于二者之间,即可在分钟级以及小时级之间; 而按照处理需求划分,大数据的处理需求可面向于海量数据的分布式处理、非结构化数据处理以及实时数据处理。按照上述划分方式,总结其核心技术如表 5 所示。

表 5 大数据处理对应的核心技术

Table 5 The key techniques for big data processing

划分标准	处理模式	内涵	核心技术
按照处理时间	在线	处理时间在秒级甚至毫秒级	流式处理技术
	近线	处理时间在分钟甚至小时级	批量数据处理
	离线	处理时间以天为基本单位	批量数据处理
按照处理需求	海量数据分布处理	批处理	Hadoop 生态系统
	非结构化数据处理	特殊数据处理	文本处理、多媒体处理、图处理技术
	实时数据处理	流处理	流式处理技术

目前典型的批量数据处理系统包括 2003 年 Google 研发的 Google 文件系统 GFS [53] 以及 2004 年的 MapReduce [54] 编程模型,以及在此基础上,2006 年 Nutch 项目子项目之一的 Hadoop 实现的两个强有力的开源产品: HDFS 和 MapReduce,目前

Hadoop 已经成为了典型的大数据处理架构。而对应的实时处理需求衍生出的典型流式数据处理系统包括 Twitter 推出的可用于实时处理新数据和更新数据库的 Storm 系统[55]、2013 年 Linkedin 开发的自己的流式数据处理框架 Samza[56]、Berkeley 提出的基于内存计算的可扩展的开源集群计算系统 Spark[57] 以及 Google 研发的交互式数据分析系统 Dremel[58]。而针对于非结构化数据处理，文献[6]中作者综述了典型的图数据处理系统，涉及 GraphLab[59]、Giraph[60]、Neo4j[61]、HyperGraphDB[62]、InfiniteGraph、Cassovary[63]、Trinity 以及 Grappa 等。

3.4　大数据的计算平台

最早的计算资源是只能由专业人员使用的大型机，之后发展成个人电脑走进千家万户，现为了满足海量数据运算的需要，这些小型的服务器又通过网络搭建集群提供更强大的计算资源，且为了方便管理、部署及提高资源使用率，虚拟化技术应运而生。最终所有的 IT 资源都会迁移到“云”中。其计算资源演变如图 3 所示。

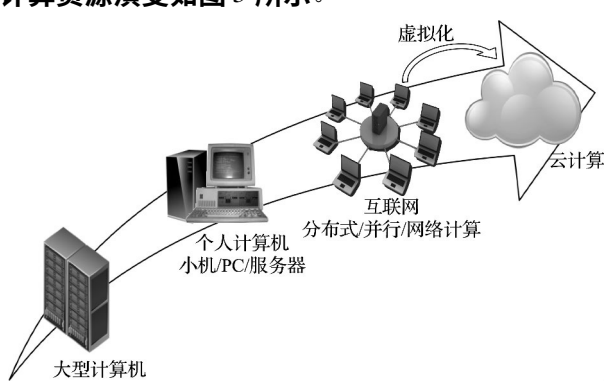

图 3　数据计算技术演化

Fig. 3　The evolutions of data computing

云计算在未来将成为重要的计算模式[64]，其将根据需求实现资源的有效分配以及应用。云计算平台涉及谷歌推出的 Google Compute Engine、微软推出的 Azure、亚马逊推出的 AWS 等。

3.5　大数据挖掘技术

对于大数据的挖掘请求，包括面向于文本的挖掘、机器学习等，挖掘算法的复杂度高、数据的计算量大，针对于大数据的规模大、速度快以及类型多样的特点，将大数据挖掘算法的研究方向总结如下:

1) 有效的大数据预处理技术

大数据的规模大、处理速度快以及流式查询处理的需求使得在对大数据进行分析以及挖掘时，必须提高数据预处理能力，以提升响应效率。目前针对于流式大数据的约简技术，包括 2 种方式[65]，一是基于数据的技术，其通过生成整个流式数据的概要或者选择其中的部分子集来实现约简，包括采样(Sampling)、卸载技术(Load-shedding)、梗概(Sketching)、数据概要结构(Synopsis Data Structures)、集成(Aggregation)，其中 Sampling、Load-shedding 以及 Sketching 通过一定规则选取整个流式数据的子集来代替原始数据从而减少数据存储量、而 Synopsis Data Structures 以及 Aggregation 方法则通过概括整个数据流的方式实现约简; 另一种约简方式是基于任务的技术，包括近似算法(Approximation algorithms)、滑动窗口技术(Sliding Window)以及输出粒度(Algorithm Output Granularity)的方法，其主要是从空间上减少整个数据流的计算规模，这种对原始数据进行压缩表达的思想更是在信号重建及还原领域得到充分体现，如文献[66]将压缩感知理论用于宽带 SAR 信号侦察，其基于信号的稀疏性，利用较少的压缩采样数据获得了较高的信号估计精度。

2) 非向量数据挖掘

以前数据挖掘多假设数据为向量数据，而大数据其结构具有多样性，包含了半结构化以及非结构化数据，所以大数据算法应提高非向量数据挖掘能力。对于非结构化数据挖掘算法研究，涉及频繁项挖掘、分类以及聚类等。例如，文献[67]提出了 XRules 算法，其为面向半结构化数据的基于规则的分类方法，通过挖掘 XML 数据中的频繁结构来建立分类规则，以发现文件中隐含的重要信息; Xproj[68] 算法则通过将数据中特殊频繁子结构出现的频度定义为类间的相似性，将相似性定量化，从而实现 XML 文档的聚类; POTMiner[69] 通过半序树的并行挖掘实现 XML 文档的结构信息表达。但是由于非结构化数据以及半结构化数据的结构具有不确定性，其价值的挖掘仍然面临巨大挑战，包括结构化信息的表达，类间相似性函数的构建、相似性函数的使用以及聚类中间结果的表达等。

3) 分布式大数据挖掘算法

早期的数据挖掘研究集中于单任务计算算法的性能提升,而随着现今数据规模的增长以及类型复杂度的提升,尤其是数据源的异构性以及分布式存储的方式,使得大数据的挖掘算法应具有分布式数据挖掘能力。如 TPFP-tree 和 BTP tree 算法通过并行计算实现了电网系统中数据的频繁项挖掘,其均采用了数据库分而治之的思想; CARM 算法虽没有直接对数据库进行划分,但是其将数据分布于云环境中的各个节点; ARMH[70] 算法采用了基于 Hadoop 分布式框架下不同云服务的可用资源实现大规模数据的频繁项挖掘,其可用于有效的处理增量数据库。文献[71]基于 Hadoop MapReduce 框架实现了并行的 RIPPER (Repeated Incremental Pruning for Error Reduction) 算法,该算法利用每个节点处理部分数据,然后将不同节点的结果集成为一个分类器。由此可知,以上分布式数据挖掘的实现必须有效的结合大数据的相关技术,如 Hadoop Mapreduce 框架以及云服务等,才能更有效地解决分布式数据挖掘问题。

4) 可扩展的大数据挖掘算法

大数据的高速性以及规模的不断增长,使得大数据挖掘算法应具有可扩展性,即在数据规模扩大的情况下,大数据挖掘算法仍能在有效的时间内快速响应挖掘请求。如文献[72]通过不同的并行策略以及云服务增强了 PIC 算法的可扩展性,实现了大数据的聚类; 文献[73]提出了基于 MapReduce 模型和云计算的序列模式挖掘算法(SPAMC) ,将树构建的子任务并行的分配于独立的 Mappers,并且并行的计算支持度,从而减少了大数据的挖掘时间。

4 大数据发展趋势及挑战

大数据的出现以及其相关技术在近几年的迅速突破使得大数据在改变人类生产生活方式中逐渐承担重要角色,美国政府甚至将其称为“未来的石油”,可见大数据的重要性。目前大数据的膨胀式发展已经改变了人类的思维方式,“一切皆可数据化”的思维已经出现,并且必然会在以后的科学研究中占据主导地位。同时大数据在人类生产方式上的应用将会加速工业 4.0 的到来,而大数据在人类生活方式上的应用也会助阵 CPS 系统价值的展现。为此,本部分将从改变思维方式、改变生产方式以及生活方式的三个角度阐述大数据的发展趋势,并分析其发展所面临的挑战。

4.1 大数据发展趋势

1) 改变思维方式

在 2007 年,图灵奖的获得者 Jim Gary 提出了科学的第四范式—— “数据密集型科学”,之前的三种科学范式分别为实验科学、理论科学以及计算科学,第四范式的提出标志着数据对于科学研究的重要性的提升,其实质是科学研究将从以计算为中心向以数据为中心转变,即数据思维的到来。

“数据密集型科学”一经提出就得到了领域内研究学者的广泛关注,如微软在 2009 年 10 月发布了《e-science 科学研究的第四种范式》,其对 Jim Gary 的观点进行了应用扩展,首次全面描述了快速兴起的数据密集型科学的研究,并将一个完整的科学研究分为四个部分,分别是数据收集、数据整理、数据分析以及数据可视化,其强调大量收集的数据需要有效分析才能实现其价值,e-science 提供了一种新的科学思维,即各种工具的使用都应用于解决科学研究中海量数据问题,由此可见大数据的发展已改变了科学研究的思维方式,为了促进大数据的认知以及发展,微软研究院已经于 2012 年 10 月 23 日发布《第四范式: 数据密集型的科学发现》中文版。

大数据的发展不仅改变了科学思维,也必然会引起企业以及政府、个人的思维方式的变革,维克托·尔耶·舍恩伯格在《大数据时代: 生活、工作与思维的大变革》中指出对于大数据时代,应放弃对因果关心的渴求,而更关注相关关系,正如其在福布斯·静安南京路论坛上的演讲所述 “在大数据时代,人们每天醒来,要想的事情就是这么多大数据可以用来做什么,其价值可以体现在哪些方面,而且是否可以找到一个别人从未涉及的事情使得思路以及想法成为重要的资产”。由此可见,大数据时代必然会引起思维的转变,而且思维的转变越快,越能在如今竞争激烈的社会中抢占先机。

2) 改变生存方式

在 21 世纪,信息技术突飞猛进的今天,物联网、嵌入式技术、传感技术等的发展,为人类更全面地感知客观存在的物理世界提供了基础; 而互联

网、云计算等信息技术的发展更是改变了人类通信与管理信息的方式。随着技术的发展以及工具的更新换代,人类也提出了更高的生存需求,美国国家科学基金委员会在 2006 年提出了 CPS 的概念; 2007 年,不同机构及研究学者对其进行了定义,包括 LEE,Baheli,Sastry 以及 Krogh 等,强调计算元素以及物理元素,实体与虚拟网络的关系,并注重通信、计算以及控制能力,尽管不同定义的描述不同[74],但是都明确了 CPS 的内涵: Cyber 与 Physical 的深度融合后形成的智能系统,CPS 的含义如图 4 所示。

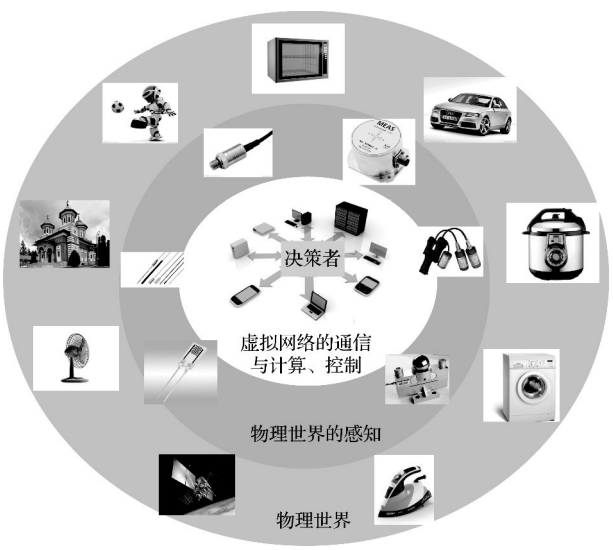

图 4　信息物理系统运行

Fig. 4　the operation for CPS

图 4 中,最外面一层是物理实体,其代表我们生活的物理世界; 中间一层为感知层,包括了传感器等具有采集功能的设备; 第三层为计算机等具有计算功能的设备,其负责实现对采集数据的分析以及可视化呈现; 最里面一层为决策层(具有决策能力的人或者其他事物) ,其通过感知以及分析结果做出决策,并作用于物理实体。CPS 的运行图体现了在感知基础上,“人”、“机”、“物”的深层融合。CPS 系统的有效工作将改变人类的生存方式,如其可应用于无人机、自主导航的汽车等以实现物理实体的自主工作,医疗领域中可应用于自动手术,物联网领域中可实现生活中的智能家居以及智慧城市等。上述 CPS 的成功实现,最重要的基础就是系统中收集的大量数据的有效分析以及处理,其是决策支持的重要来源。即如果没有大数据的积累以及分析,那么 CPS 系统也就无从谈起。由此可见,大数据的产生以及有效分析是 CPS 的重要资源和基础,结合其他技术的发展,将为改变人类生存方式提供重要动力。

3) 改变生产方式

目前已经先后经历了三次工业革命,包括 1760 ~ 1840 年因为蒸汽动力的发明产生了生产制造的机械化,开创了“蒸汽时代”; 1840 ~ 1950 年因为电的发明开创了“电气时代”,使得生产得以批量化; 1950 年至今,电子技术和计算机等信息技术的发展开创了“信息时代”,使得产品更为丰富,功能性更强; 而随着科技的进一步发展,科技的进步也必定引起生产方式的变革。为此,德国提出了“工业 4.0”,即第四次工业革命,以智能制造为主导实现生产制造人机一体化,“工业 4.0”的提出预示着革命性的生产方式的诞生,而实现“工业 4.0”的基础就是大数据的分析以及 CPS 的推广,其标志着生产制造业必须转向以数据分析为中心。由此可见,大数据的发展将在生产方式改变中起到关键作用。四次工业革命演化如图 5 所示。

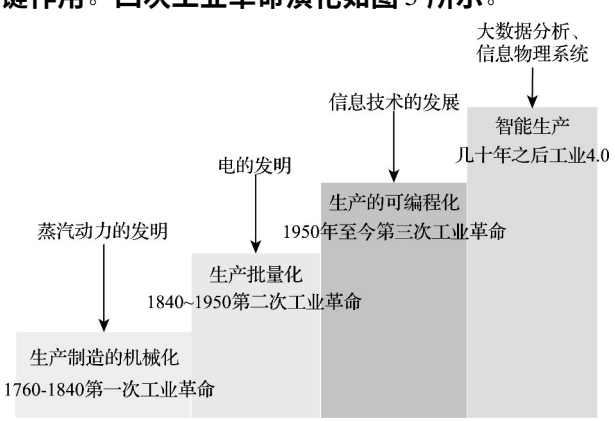

图 5　四次工业革命演化

Fig. 5　The four times evolutions for industrial revolution

工业 4.0 计划是在 2013 年举办的 Hannover 工业博览会中,由德国“工业 4.0 工作组”在《保障德国制造业的未来: 关于实施“工业 4.0”战略的建议》中正式提出,其强调了以物联网和制造业服务化的第四次工业革命,虽然第四次工业革命是否到来还存在很大争议,但是目前很多国家已经投入了大量资金以及精力来推进“工业 4.0”的进程。成功的典范是特斯拉以及西门子,特斯拉将自己的核

心定位于大型可移动的智能终端，通过互联网将汽车设计为包含软件、硬件以及内容和服务的体验工具，将互联网思维引入制造业；而西门子的电子车间更是将“工业 4.0”付诸实践的典型代表，其建立了一个紧密结合的技术网络，通过技术整合形成更智能、高效的整体，使生产线的可靠性达到 99%，追溯性达到 100%。

“工业 4.0”将要达到的目标是通过物联网系统实现智能工厂，即每件产品、零部件都会包含大量的信息，包括何时生产、可以用多久、是否需要替换等，通过非人为干预的智能方式实现自主处理。由此可见，大数据将在改变生产方式中扮演重要角色，由大数据到决策的实现将加速工业 4.0 时代的到来。

4.2 大数据发展的挑战

大数据其规模大、速度快以及结构多样的特点，为传统数据的分析、存储以及管理技术带来的挑战不言而喻。对大数据处理流程的挑战总结如表 6 所示。

表 6 大数据发展的挑战

Table 6 The challenges for the big data development

研究主题	挑战
大数据预处理及集成	广泛的异构性、时空特性、数据质量
大数据分析	先有数据后有模式、动态增长、先验知识的缺乏、实时性
大数据硬件处理平台	硬件异构性、新硬件
性能测试基准	系统复杂性高、案例多样性、数据规模庞大、系统的快速演变
隐私保护	隐性数据的暴露、数据公开与保护、数据动态性
大数据管理的易用性	可视化、人机交互、数据起源技术、海量元数据的高效管理
大数据的能耗	低功耗、新能源

大数据的规模大，其质量影响算法的效率以及精度，大数据预处理作为数据分析的第一步，至关重要。而大数据来源的多样性，使得数据具有广泛的异构性、时空特性等，其为大数据预处理及集成带来严峻的考验；大数据规模动态增长使得大数据的模式获取困难，加之先验知识的缺乏，如何在规定的时间内返回有价值的分析结果也是研究学者设计算法时不得不考虑的问题；这种需求也给大数据的计算系统提出了挑战，高性能计算面临着访存墙、通信墙、可靠性墙、能耗墙的问题[75]，如信息系统正从“数据围着处理器转”转变为“处理能力围着数据转”，在这种情况下为了满足持续的数据存取要求，系统结构设计的出发点要从重视单任务的完成时间转变为提高系统吞吐率和并发处理能力，且必须转变为以数据为中心的计算系统的基本思路，从根本上消除不必要的数据流动，即使是必要的数据搬运也应由“大象搬木头”（少量强核处理复杂任务）转变为“蚂蚁搬大米”（大量弱核处理简单任务），即以数据为中心的系统结构要消除不必要的数据存放、通信和计算。但是与之相对应的系统并未完全实现这一思路，所以这也是大数据计算系统未来所必须解决的问题。同时，作为大数据处理的支撑技术，包括隐私保护、硬件平台以及大数据管理、能耗等也有很多难题需要突破。大数据发展的挑战也为大数据的发展指明了方向，需要大数据相关工作者突破领域限制，共同努力。

5 结 论

大数据作为现在以及未来的重要资源，已经出现在生产生活的多个领域，引起了各部门的广泛关注，并将成为未来市场竞争以及科技创新的重要争夺资源，但是其价值的体现需要突破传统数据分析以及处理的限制，重视数据之间的相关关系，在满足精度要求的情况下快速响应分析需求。

本文在大数据内涵的基础上，对于大数据处理中各个环节涉及的关键技术进行了归纳与分析，并从数据科学、工业 4.0 以及信息物理系统的角度，展望了大数据的发展对于改变人类思维方式、生产方式以及生活方式的重要作用。研究学者应在“大数据”大热的趋势下，冷静分析，按照自身定位以及需求，定义科学问题，以切入切实可行的研究方向，并通过把握大数据处理流程中的关键技术，建立持续的研究体系，以把握大数据这一发展机遇，充分利用大数据创造大价值。

参考文献

[1] MAYER-SCHONBERGER V, CUKIER K. Big data: a revolution that will transform how we live, work, and

think [M]. John Murray Publishers Ltd, 2013: 174-180.

[2] 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 [J]. 中国科学院院刊, 2012, 27(6): 647-657.

[3] 武延军. 大数据时代已经来临——人机物融合的大数据时代 [J]. 高科技与产业化, 2015(5): 46-49.

[4] 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.

[5] 覃雄派, 王会举, 杜小勇, 等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生 [J]. 软件学报, 2012, 23(1): 32-45.
QIN X P, WANG H J, DU X Y, et al. Big data analysis——Competition and symbiosis of RDBMS and MapReduce [J]. Journal of Software, 2012, 23(1): 32-45.

[6] MANYIKA J, CHUI M, BROWN B, et al. Big Data: The next frontier for innovation, competition and productivity [R]. Mckinsey Global Institute, 2011.

[7] TOFFER A. The third wave [M]. New York: Bantam Books, 1981.

[8] DIEBOLD F X. "Big Data" Dynamic factor models for macroeconomic measurement and forecasting [M]. Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society, Cambridge: Cambridge University Press, 2003: 115-122.

[9] HU H, WEN Y G, CHUA T S, LI X L. Toward scalable systems for big data analytics: a technology tutorial [J]. IEEE Access, 2014(2): 652-687.

[10] KATAL A, WAZID M, GOUDAR R H. Big data: issues, challenges, tools and good Practices [C]. 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, INDIA, 2013: 404-409.

[11] LIU Z Y, YANG P, ZHANG L X. A sketch of big data technologies [C]. 2013 Seventh International Conference on Internet Computing for Engineering and Science (ICICSE 2013), Shanghai, China, 2013: 26-29.

[12] 刘吉臻, 刘继伟, 曾德良, 等. 大数据多尺度状态检测方法在磨损检测的应用 [J]. 仪器仪表学报, 2013, 34(1): 180-186.
LIU J ZH, LIU J W, ZENG D L, et al. Application of multi-scale state detection method based on big data in wear detection [J]. Chinese Journal of Scientific Instrument, 2013, 34(1): 180-186.

[13] EDWARD A L, SANJIT A S. Introduction to embedded systems, a cycber-physical systems approach [EB/OL]. http://LeeSeshia.org.

[14] KAGERMANN H, WAHLSTER W, HELBIG J. Securing the future of German manufacturing industry recommendations for implementing the strategic initiative INDUSTRIE 4.0 [R]. Germany: Federal Ministry of education and research, Final report of the Industrial 4.0 working group, April 2013.

[15] BigData [EB/OL]. [2012-10-02]. http://en.wikipedia.org/wiki/Big_data.

[16] Gartner. IT glossary-big data [EB/OL]. http://www.gartner.com/it-glossary/big-data.

[17] 徐宗本, 张维, 刘雷, 等. "数据科学与大数据的科学原理及发展前景"——香山科学会议第 462 次学术讨论会专家发言摘登 [J]. 科技促进发展, 2014, 10(1): 66-75.

[18] 孟小峰, 慈祥. 数据管理: 概念、技术与挑战 [J]. 计算机研究与发展, 2013, 50(1): 146-149.
MENG X F, CI X. Big data management: concepts, techniques and challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-149.

[19] 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 ZH, JIN X L, CHENG X Q. Network Big Data: Present and Future [J]. Chinese Journal of Computers, 2013, 36(6): 1125-1138.

[20] ALYA A N, ALLAN T, SERGIO D C. Big data analysis of stock twits to predict sentiments in the stock market [J]. Lecture Notes in Computer Science, 2014, 8777: 13-24.

[21] 周光华, 辛英, 张雅洁, 等. 医疗卫生领域大数据应用探讨 [J]. 中国卫生信息管理杂志, 2013, 10(4): 296-304.
ZHOU G H, XIN Y, ZHANG Y J, et al. Study on big data's applications [J]. Medical and Health Field, 2013, 10(4): 296-304.

[22] 孙磊, 胡学龙, 张晓斌, 等. 生物医学大数据处理的云计算解决方案 [J]. 电子测量与仪器学报, 2014, 28(11): 1190-1197.
SUN L, HU X L, ZHANG X B, et al. Cloud computing solutions for processing biomedical data [J]. Journal of Electronic Measurement and Instrumentation, 2014, 28(11): 1190-1197.

[23] AGHABOZORG I, MAHROEIAN H, DUTT A, et al. An approachable analytical study on big educational data mining [C]. Computational Science and Its Applica-

tions-ICCSA 2014 Lecture Notes in Computer Science, 2014, 8583: 721-737.

[24] GUO H D, WANG L Z, CHEN F, LIANG D. Scientific big data and digital earth [J]. Chinese Science Bulletin, 2014, 59(35): 5066-5073.

[25] ARMES T, REFERN M. Using big data and predictive machine learning in aerospace test environments [C]. IEEE AUTOTESTCON, 2013: 16-19.

[26] WANG H Z, LIN G. W, WANG J Q, et al. Management of big data in the internet of things in agriculture based on cloud computing [C]. Applied Mechanics and Materials, 3rd International Conference on Manufacturing Engineering and Process, ICMEP, 2014: 1438-1444.

[27] LANEY D. 3D Data Management: Controlling Data Volume, Velocity, and Variety [R]. Meta Group, 2001. http://blogs. gartner. com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety. pdf

[28] BARWICK H. The "four Vs" of big data [EB/OL]. (2011-08-05) [2012-10-02]. http://www. Computerworld. com. au/article/396198/iii3_four_vs_big_data/

[29] IBM. What is big data? [EB/OL]. [2012-10-02]. http://www-01. ibm. com/software/data/bigdata/what-is-big-data. html

[30] BRIAN H, BORIS E. Expand your digital horizon with big data [EB/OL], 2012.

[31] **任永功，于戈. 数据可视化技术的研究与进展** [J]. **计算机科学**, 2004, 31(12): 92-96.
REN Y G, YU G. Research and Development of the Data Visualization Techniques [J]. Computer Science, 2004, 31(12): 92-96.

[32] ARI R, RANDY K. Chukwa: a scalable log collector [EB/OL]. [2010-1-15] https://www. usenix. org/legacy/event/lisa10/tech/full_papers/Rabkin. pdf

[33] **王静，孟小峰. 半结构化数据的模式研究综述** [J]. **计算机科学**, 2011, 28(2): 6-11.
WANG J, MENG X F. Schema of Semistructured Data: A Survey [J]. COMPUTER SCIENCE, 2011, 28(2): 6-11.

[34] WILDE E, GLUSHKO R J. XML fever [J]. Communications of the ACM, 2008, 51(7): 40-46.

[35] HUANG S, CAI L, LIU Z, HU Y. Non-structure data storage technology-An discussion [C]. 2012 IEEE/ACIS 11th International Conference on Computer and Information Science, Shanghai; China, 2012: 482-487.

[36] **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]. 2013, 24(8): 1786-1803.
SHEN D R, YU G, WANG X T, et al. Survey on NoSQL for Management of Big Data [J]. 2013, 24(8): 1786-1803.

[37] WAAS F M. Beyond conventional data warehousing-Massively parallel data processing with greenplum database [C]. 2nd International Workshop on Business Intelligence for the Real-Time Enterprise, Auckland, New Zealand, Lecture Notes in Business Information Processing, 2009, 27: 89-96.

[38] BEAR C, LAMB A, TRAN N. The vertica database: SQL RDBMS for managing big data [C]. Proceedings of the 2012 Workshop on Management of Big Data Systems, Co-located with ICAC12, San Jose, CA, United States, 2012: 37-38.

[39] SIMMEN D, SCHNAITTER K, DAVIS J, et al. Large-scale graph analytics in Aster 6: Bringing context to big data discovery [J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1405-1416.

[40] VORA M N. Hadoop-HBase for large-scale data [C]. Proceedings of 2011 International Conference on Computer Science and Network Technology, ICCSNT 2011, Harbin, China, 2011(1): 601-605.

[41] DEDE E, GOVINDARAJU M, GUNTER D, et al. Performance evaluation of a MongoDB and Hadoop platform for scientific data analysis [C]. Proceedings of the 4th ACM Workshop on Scientific Cloud Computing, New York, NY, United States, 2013: 13-20.

[42] LAKSHMAN A, MALIK P. Cassandra-A decentralized structured storage system [J]. Operating Systems Review (ACM), 2010, 44(2): 35-40.

[43] CORBETT J C, DEAN J, EPSTEIN M, et al. Spanner: Google's globally distributed database [J]. ACM Transactions on Computer Systems, 2013, 31(3): 8.

[44] BAKER J, BOND C, CORBETT J C, et al. Megastore: Providing scalable, highly available storage for interactive services [C]. 5th Biennial Conference on Innovative Data Systems Research, Conference Proceedings, Asilomar, CA, United States, 2011: 223-224.

[45] RAE I, ROLLINS E , SHUTE J, SODHI S, VINGRALEK R. Online, Asynchronous schema change in F1 [C]. The 39th International Conference on Very Large Data Bases, Riva del Garda, Trento, Italy, Proceedings of the VLDB Endowment, 2013, 6(11): 1045-1056.

[46] MICHAEL S. SQL databases v. NoSQL databases [J]. Communications of the ACM, 2010, 53(4): 10-11.

[47] MATTHEW A. How Will The Database Incumbents Re-

spond To NoSQL And NewSQL [R]. 451 Group, [2011-04-04] https://cs.brown.edu/courses/cs227/archives/2012/papers/newsql/aslett-newsql.pdf

[48] THUSOO A, SARMA J S, JAIN N, et al. Hive-A petabyte scale data warehouse using hadoop [C]. International Conference on Data Engineering, Long Beach, CA; United States, 2010: 996-1005.

[49] FÄRBER F, CHA S K, PRIMSCH J, et al. SAP HANA database-Data management for modern business applications [J]. SIGMOD Record, 2011, 40(4): 45-51.

[50] KUNCHITHAPADAM K, ZHANG W, GANESH A, MUKHERJEE N. Oracle database filesystem [C]. Proceedings of the ACM SIGMOD International Conference on Management of Data, Athens, Greece, 2011: 1149-1160.

[51] BRANTNER M, FLORESCU D, GRAF D, et al. Building a database on S3 [C]. Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 2008: 251-263.

[52] CALDER B, WANG J, OGUS A, et al. Windows Azure storage: a highly available cloud storage service with strong consistency [C]. Proceedings of the 23rd ACM Symposium on Operating Systems Principles, Cascais, Portugal, 2011: 143-157.

[53] GHEMAWAT S, GOBIOFF H, LEUNG S-T. The Google file system [J]. ACM, 2003, 37(5): 29-43.

[54] DEAN J, GHEMAWAT S. MapReduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.

[55] TOSHNIWAL A, TANEJA S, SHUKLA A, et al. Storm @ Twitter [C]. 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, United States, 2014: 147-156.

[56] **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 ZH, et al. Survey on Big Data System and Analytic technology [J]. Journal of Software, 2014, 25(9): 1889-1908.

[57] ZAHARIA M, CHOWDHURY M, FRANKLIN M, et al. Spark: Cluster computing with working sets [C]. HotCloud10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, 2010: 1-10.

[58] MALEWICZ G, AUSTERN M H, BIK A J, et al. Pregel: A system for large-scale graph processing [C]. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data, 2010: 135-146.

[59] LOW Y, GONZALEZ J, KYROLA A, et al. GraphLab: A new framework for parallel machine learning [C]. 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, United States, 2010: 340-349.

[60] YANG S, SPIELMAN N D, JACKSON J C, RUBIN B S. Large-scale neural modeling in MapReduce and Giraph [C]. 2014 IEEE International Conference on Electro/Information Technology, Milwaukee, WI, United States, 2014: 556-561.

[61] WEBBER J. A programmatic introduction to Neo4J [C]. 2012 3rd ACM Conference on Systems, Programming, and Applications, Tucson, AZ, United States, 2012: 207.

[62] IORDANOV B. Hypergraphdb: a generalized graph database [C]. In Proceedings of the 2010 international conference on Web-age information management, WAIM'10, Berlin, Heidelberg, 2010: 25-36.

[63] GUPTA P, GOEL A, LIN J, SHARMA A, et al. WTF: The Who to Follow service at Twitter [C]. 22nd International Conference on World Wide Web, WWW 2013, Rio de Janeiro, Brazil, 2013: 505-514.

[64] BUYYA R, YEO C S, VENUGOPAL S, BROBERG J, BRANDIC I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility [J]. Future Generation Computer Systems, 2009, 25(6): 599-616.

[65] GABER M M, ZASLAVSKY A, KRISHNASWAMY S. ACM SIGMOD Record, Mining Data Streams: A Review [J]. 2005, 34(2): 18-26.

[66] **王康, 叶伟, 劳国超, 等. 一种基于压缩感知的宽带信号侦察方法** [J]. **国外电子测量技术**, 2014, 33(4): 40-43.
WANG K, YE W, LAO G CH, et al. Reconnaissance method to wideband SAR signals based on compressed sensing [J]. Foreign Electronic Measurement Technology, 2014, 33(4): 40-43.

[67] ZAKI M J, AGGARWAL C C. XRules: An effective algorithm for structural classification of XML data [J]. Machine Learning, 2006, 62(1-2): 137-170.

[68] AGGARWAL C C, TA N, WANG J, et al. Xproj: a framework for projected structural clustering of XML documents [C]. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, California, 2007: 46-55.

[69] JIMENEZ A, BERZAL F, CUBERO J C. POTMiner: Mining ordered, unordered, and partially-ordered trees

[J]. Knowledge and Information Systems, 2010(23): 199-224.

[70] NATARAJAN S, SEHAR S. A Noval Algorithm for Distributed Data Mining in HDFS [C]. Advanced Computing (ICoAC), 2013 Fifth International Conference, Chennai, India, 2013: 93-99.

[71] GUGNANI S, KHANOLKAR D, BIHANY T, KHADILKAR N. Rule Based Classification on a Multi Node Scalable Hadoop Cluster [C]. Internet and Distributed Computing Systems. 7th International Conference, IDCS 2014, Calabria, Italy, 2014: 22-24.

[72] YAN W Z, BRAHMAKSHATRIYA U, XUE Y, et al. p-PIC: Parallel power iteration clustering for big data [C]. Journal Of Parallel And Distributed Computing, 2013, 73(3): 352-359.

[73] CHEN C C, TSENG C Y, CHEN M S. Highly scalable sequential pattern mining based on MapReduce model on the cloud [C]. 2013 IEEE International Congress on Big Data, 2013: 310-317.

[74] 王中杰，谢璐璐. 信息物理融合系统研究综述 [J]. 自动化学报，2011，37(10)：1157-1166.
WANG ZH J, XIE L L. Cyber-physical Systems: A Survey [J]. ACTA AUTOMATICA SINICA, 2011, 37(10): 1157-1166.

[75] 王之元. 并行计算可扩展性分析与优化 [D]. 湖南：国防科技大学，2011：54-60.
WANG Z Y. Analysis and optimization of scalability for parallel computing-energy, reliability and performance [D]. Changsha, Hunan, P. R. China: National University of Defense Technology, 2011: 54-60.

作者简介

彭宇(通讯作者)，工学博士、教授、博士生导师，1973 年出生，哈尔滨工业大学电气工程及自动化学院副院长、自动化测试与控制研究所副所长，主要研究方向为虚拟仪器和自动测试、故障预测与健康管理、可重构计算等。

E-mail: pengyu@hit.edu.cn

Peng Yu(Corresponding author) was born in June 1973. He is now a professor and Ph. D candidate supervisor in the Major of Instrumentation Science and Technology, Harbin Institute of Technology (HIT). He is the vice Dean of School of Electrical Engineering and Automation and the vice Director of Automatic Test and Control Institute in HIT. His main research fields include virtual instruments and automatic test technologies, prognostics and system health management, and reconfigurable computing, etc.

庞景月，工学在读博士，1988 年出生，2011 年于重庆理工大学获得学士学位，2013 年于哈尔滨工业大学获得硕士学位，现在哈尔滨工业大学攻读博士学位。主要研究方向为故障预测与健康管理、数据流异常检测与挖掘。

E-mail: jypang@hit.edu.cn

Pang Jingyue, Ph. D. candidate, was born in December 1988. She received the B. Sc. from Chongqing University of Technology in 2011, and received M. Sc. from Harbin Institute Technology in 2013. Now she is a PhD candidate in Harbin Institute of Technology, her research interest is prognostics and health management, abnormal detection and mining for data stream.

刘大同，工学博士、副教授、硕士生导师，1982 年出生，哈尔滨工业大学电气工程及自动化学院自动化测试与控制系，主要研究方向为自动测试技术、智能测试信息处理、故障预测和健康管理。

E-mail: liudatong@hit.edu.cn

Liu Datong was born in September 1982. He is currently an associate professor and graduates advisor with the Department of Automatic Test and Control, School of Electrical Engineering and Automation, Harbin Institute of Technology (HIT). His research interests include automatic test technologies, intelligent test data processing, and prognostics and health management.

彭喜元，工学博士、教授、博士生导师，1961 年 12 月生。现任哈尔滨工业大学电气工程及自动化学院院长，兼任自动化测试与控制研究所所长。主要研究方向是自动测试理论、技术及系统，先进故障诊断技术及应用。

E-mail: pxy@hit.edu.cn

Peng Xiyuan was born in December 1961. He is now a professor and Ph. D candidate supervisor in the Major of Instrumentation Science and Technology, Harbin Institute of Technology(HIT). He is also the Dean of School of Electrical Engineering and Automation and the Director of Automatic Test and Control Institute in HIT. His main research fields include automatic test and advanced fault diagnosis.