大数据研究综述*

涂新莉[1]，刘 波[1]，林伟伟[2†]
(1. 华南师范大学 计算机学院，广州 510631；2. 华南理工大学 计算机科学与工程学院，广州 510640)

摘 要： 主要从大数据的概念着手，对比分析了国内外大数据研究和应用现状，重点分析比较当前大数据主流处理工具的优缺点，并深入归纳总结了基于数据存储的大数据处理技术、基于数据挖掘的大数据处理技术、基于查询的大数据处理技术的优缺点和适用场景。最后，在前面比较和分析的基础上给出了大数据研究和发展的方向，为大数据的研究提供有益参考。

关键词： 大数据；数据处理；数据挖掘
中图分类号： TP311 **文献标志码：** A **文章编号：** 1001-3695(2014)06-1612-05
doi:10.3969/j.issn.1001-3695.2014.06.003

Survey of big data

TU Xin-li[1], LIU Bo[1], LIN Wei-wei[2†]
(1. *School of Computer, South China Normal University, Guangzhou* 510631, *China*; 2. *School of Computer, Science & Engineering, South China University of Technology, Guangzhou* 510640, *China*)

Abstract: This paper mainly compared and analyzed the research and application status of big data home and abroad, from the concept of big data, laying special stress on analyzing and comparing current main processing tools of big data, focusing on advantages and disadvantages, summarizing the advantages, disadvantages and applicable scenario of big data processing techniques, which were based on data storage, data mining and queries. Finally, it presented important directions of research and development of big data for future, providing reference for big data research.

Key words: big data; data processing; data mining

提及大数据几乎是无人不晓，但大数据这个概念并不是近几年才有的，早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。在 20 世纪 80 年代我国已经有一些专家学者谈到了海量数据的加工和管理，但是由于计算机技术和网络技术的限制大数据未能引起足够的重视，它蕴藏的巨大信息资源也暂时隐藏了起来。随着云计算技术的发展，互联网的应用越来越广泛，以微博和博客为代表的新型社交网络的出现和快速发展，以及以智能手机、平板电脑为代表的新型移动设备的出现，计算机应用产生的数据量呈现了爆炸性增长的趋势。2012 年末出版的《大数据时代》的作者英国牛津大学网络学院互联网研究所治理与监管专业教授维克托·尔耶·舍恩伯格在书的引言中说，大数据正在改变人们的生活以及理解世界的方式，而更多的改变正蓄势待发。美国总统奥巴马的成功竞选及连任的背后都有大数据挖掘的支撑，美国政府认为，大数据是“未来的新石油”，并将对大数据的研究上升为国家意志，这对未来的科技与经济发展必将带来深远影响[1]。如今，大数据已成为一项业务上优先考虑的工作任务，因为它能够对全球整合经济时代的商务产生深远的影响。大数据的应用范围如此广泛，与大数据相关的很多问题都引起了专家和学者的重视。大数据最基本的问题——大数据的定义目前还没有一个统一的定论，但大数据作为一种基础性资源需要被处理才能显现其潜在的价值，那么如何更好地处理大数据这种基础性资源就显得特别重要，因为这些问题都关系到大数据核心价值的体现。为此，本文从大数据若干个版本的概念出发，调查分析了大数据的研究和应用现状，重点分析了当前主流的大数据处理工具和技术，最后预测了大数据未来的几个研究和发展方向。

1 大数据的概念

大数据本身就是一个很抽象的概念，提及大数据很多人也只能从数据量上去感知大数据的规模，如：百度每天大约要处理几十 PB 的数据；Facebook 每天生成 300 TB 以上的日志数据；据著名咨询公司 IDC 的统计，2011 年全球被创建和复制的数据总量为 1.8 ZB(10^{21})，但仅仅是数据量并不能区分大数据与传统的海量数据的区别。在 2008 年《Science》杂志出版的专刊中，大数据被定义为“代表着人类认知过程的进步，数据集的规模是无法在可容忍的时间内用目前的技术、方法和理论去获取、管理、处理的数据”[2]。比较有影响力的 Gartner 公司也给出了大数据的定义[3]，大数据是高容量、高生成速率、种类繁多的信息价值，同时需要新的处理形式去确保判断的作出、洞察力的发现和处理的优化。这种定义不仅是数据规模大，更重要的是如何从这些动态快速生成的数据流或数据块中

收稿日期：2013-10-14；修回日期：2013-11-28 基金项目：国家自然科学基金资助项目(61070015)；广东省自然科学基金资助项目(S2011010001754，S2012030006242)；广东省科技计划资助项目(1311050300017，1311020500039)；中央高校基本科研业务费专项资金资助项目(2013ZZ0044)

作者简介：涂新莉(1988-)，女，硕士研究生，主要研究方向为分布计算系统；刘波(1968-)，男，教授，博士，主要研究方向为分布计算系统；林伟伟(1980-)，男(通信作者)，副教授，博士，主要研究方向为云计算、大数据、分布式系统(linww@scut.edu.cn)。

获取有用的具有时效性价值的信息，但是这些数据类型众多，结构化、半结构化、非结构化的数据对已有的数据处理模式带来了巨大的挑战，其中也体现了大数据在 3V 基础上发展的 4V 定义。4V 定义即 volume，variety，velocity，value，关于第 4 个 V 的说法并不统一，国际数据公司(International Data Corporation，IDC) 认为大数据还应当具有价值性(value) [4]，大数据的价值往往呈现出稀疏性的特点; 而 IBM 认为大数据必然具有真实性(veracity) [5]，这样有利于建立一种信任机制，有利于领导者的决策。百度百科对大数据的定义是: 大数据(big data)，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。大数据的科学家 Rauser 提到一个简单的定义: 大数据就是超过了任何一个计算机处理能力的庞大数据量。

2　大数据的研究与应用现状

虽然大数据的概念没有一个统一的定论，但这对于大数据的研究而言并不是最重要的，如何使用大数据才是关键。研究大数据其实也就是为了更好地应用大数据，所以国内外对大数据的研究与应用都相当重视。事实上，大数据的研究与应用已经在互联网、商业智能、咨询与服务以及医疗服务、零售业、金融业、通信等行业显现，并产生了巨大的社会价值和产业空间。来自麦肯锡 2012 年大数据报告中的一组数据显示，大数据产业为美国医疗系统带来每年 3000 亿美元的收益; 为欧洲公共管理部门带来 2500 亿欧元的收益; 为零售业增加 60% 的净利润; 为制造业减少 50% 的产品研发等成本。而 Canner 认为，2015 年超过 85% 的财富 500 强企业将在大数据竞争中失去优势[6]。据市场调研机构 IDC 预测，大数据技术与服务市场将从 2010 年的 32 亿美元攀升到 2015 年的 169 亿美元，实现 40% 的年增长率(IT 与通信产业增长率的 7 倍) [7]。从上面的统计数据很容易看出大数据的应用之广，价值之大。

国外的大数据研究工作主要集中在如何进行大数据存储、处理、分析以及管理的技术及软件应用上。在学术界，《Nature》早在 2008 年就推出了“big data”专刊，从互联网技术、超级计算、生物医学等方面来专门探讨对大数据的研究。2012 年 3 月，美国公布了旨在提高和改进人们从海量信息数据中获取信息能力的“大数据研发计划”[1]。2012 年 4 月欧洲信息学与数学研究协会会刊《ERCIM News》出版专刊“big data”[8]，讨论了大数据时代的数据管理、数据密集型研究的创新技术等问题。2012 年 7 月，日本推出“新 ICT 战略研究计划”，其中重点关注大数据应用，将大数据定位为战略领域之一。在具体的实际应用方面，大数据也显现出了它的价值所在。文献 [9] 中，谷歌公司通过对人们在网上检索的词条与疾病中心的数据进行分析处理，有效及时地判断出了流感的传播来源，为公共卫生机构提供了有价值的信息，这是来自 2009 年《Science》杂志上发表的一篇论文。乔布斯通过大数据辅助癌症治疗，丹麦癌症协会通过大数据研究手机是否致癌等。美国最大的西奈山医疗中心(Mount Sinai Meddical Center) 使用来自大数据创业公司 Ayasdi 的技术分析大肠杆菌的全部基因序列，包括超过 100 万个 DNA 变体，来了解为什么菌株会对抗生素产生抗药性。Ayasdi 的技术使用了一种全新的数学研究方法——拓扑数据分析(topological data analysis) 来了解数据的特征。医疗行业的大数据不仅量大，而且繁杂，其中蕴涵的信息价值也是丰富且多样。英特尔全球医疗解决方案架构师吴闻新等人也预测了医疗行业数据的增长之快，特别是影像数据和 EMR 电子病历数据[10]。英特尔协助用友医疗进行了合理的架构分析和指导，对于基于大数据分析的解决方案进行了深入的探索和研究，并且制定了基于英特尔大数据解决方案的区域卫生数据中心建设目标: 文档快速检索，存储模式满足数据模式的更新，透明化扩展容量和性能。美国俄亥俄州运输部(ODOT) 利用 INRIX 的云计算分析处理大数据来了解和处理恶劣天气的道路状况[11]，减少了冬季连环撞车发生的概率，方便了人们的出行。在能源行业，SaaS 型软件公司 Opower 使用数据分析提供消费用电的能效[3]。2012 年 11 月 6 日，美国总统奥巴马成功击败对手罗姆尼再次赢得美国总统，奥巴马总统获胜的秘密——通过大数据系统进行数据挖掘，用科学的方法指定策略，它帮助奥巴马在获取有效选民、投放广告、募集资金等方面起到了很大的作用。

与国外相比，国内大数据的研究和应用还处在起步阶段。2012 年 5 月，香山科学会议组织了以“大数据科学与工程——一门新兴的交叉学科”为主题的会议，深入讨论了大数据的理论与工程数据研究、应用方向，指出目前最重视的都是大数据分析算法和大数据系统效率，通过研究大数据的关系网络整体而全面地研究大数据。同年 6 月，中国计算机学会青年计算机科技论坛(CCF YOCSEF) 举办了“大数据时代，智谋未来”学术报告会，就大数据时代的数据挖掘、体系架构理论、大数据安全、大数据平台开发与大数据现实案例进行了全面的讨论。随着大数据时代的到来，油田勘探开发过程中也产生了规模巨大、类型多样的数据。文献 [12] 在计算机集群上构建油田勘探开发一体化数据管理模型和数据访问基础架构，从而解决油田实际应用中所面临的大数据问题，即交叉复用、信息可见、信息传承。应用文献 [12] 中构建的数据模型及其接口，专业分析软件可以很容易地获得本研究区域齐、全、准的勘探开发信息，从而进行分析，部署勘探开发生产任务。以部署探井为例，分析软件可以利用“大数据”接口非常方便地获得探井区域的地震剖面、测井曲线、层位、断层等信息。文献 [7，13] 分别从商务管理、大城市亟待解决的交通问题进行相关的研究和实验，应用实例表明，在营销策略的制定、智能化的交通管理方面都得益于大数据的分析。

如果在国内能够搭建一个大数据共享平台，经过预处理，抽取和集成的数据可通过相关的平台交换和共享，让大数据处理更便捷、更快速、更贴近用户、更容易去实现或者去操作，那么也就实现了数据的流通，数据才会更加有生命力，使用价值也会增值。对大数据的处理和应用，其核心还是需要从业务层面进行科学规划。

3　大数据的处理工具与技术

从大数据比较有影响力的概念和大数据的研究现状来看，推动大数据发展的核心力量之一就是大数据的分析处理工具和技术。因为传统的数据分析处理技术已经无法满足大数据的需求，大数据的出现也必然伴随着新的处理工具和新技术的出现。

3.1　大数据的处理工具

大数据处理技术的不断更新也促使了大数据处理工具的

出现。在大数据的处理平台中，大家最熟悉的莫过于 Apache 的 Hadoop 的块处理平台，Hadoop 主要是基于 MapReduce[14] 编程框架和 HDFS[15]。HPCC（high perform-ance computing cluster）[16] 系统也是一种开源的分布式密集数据处理平台，主要有以下组件：a）Thor（HPCC data refinery cluster）主要是作为一个能够并行处理跨节点的分布式文件系统进行工作，主要负责大量数据的接收、传输、连接和检索工作，对数据进行整合；b）Roxie（HPCC data delivery engine）提供了大量的高性能的多用户在线查询功能；c）ECL（enterprise control language）是一种适合处理大数据的功能强大的编程语言；d）ECL IDE 主要是与 ECL 配合工作的，用来编码、调试、监控 ECL 的程序；e）ESP（enterprise services platform）提供了一个易用的访问 ECL 查询接口，一般支持 SOAP、XML、HTTP 和 REST 等。Hadapt[17] 是一种高性能的自适应分析平台，其分层结构如图 1[17] 所示。

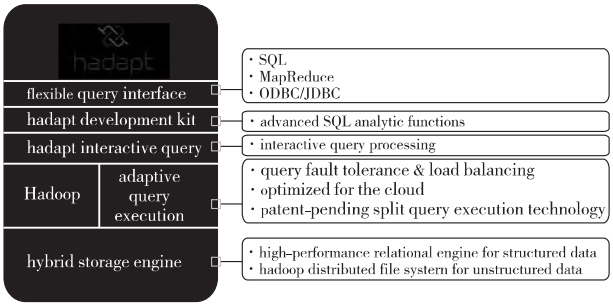

图 1 Hadapt 的结构

针对上面提到的三种处理工具的优缺点在表 1 进行了具体的比较和分析。

表 1 Hadoop、Hadapt 和 HPCC 的比较

处理工具	核心组件	优点	缺点
Hadoop	MapReduce 编程框架和 HDFS 分布式文件系统	出现得比较早，使用得比较成熟，应用范围广，开源；采用大规模廉价服务器集群，数据的存储和服务也分为 HDFS 和 HBase 两个层次，从而最大限度地利用机器资源；低成本、高可扩展性、容错性以及无须构建预定义模式，精于数据处理和分析以及原始数据的存储，索引编制、模式识别、推荐引擎建立和分析中使用较多	追求高吞吐率的同时也带来了批量处理的延迟，且对 MapReduce 的数据处理有很强的依赖性，大公司使用开源项目要考虑技术支持、技术保密性
Hadapt	hybrid storage engine，Hadoop，HDK	Hadapt 结合了 Hadoop 和关系数据库管理软件的优点，既可以在私有云上运行又可以在公共云上运行；每个节点的结构化数据存储在 RDBMS 中，非结构化数据存储在 HDFS 中，这样 Hadapt 可以在 Hadoop 层和关系数据库层之间自动划分查询执行任务；HDK 允许分析者创建高级的 SQL 分析功能统一结构和灵活的模式性能可减少分析用例的复杂性	Hadapt 通常采用的方法是扩充型连接件连接起两个不同的系统，这样做的结果是带来了一定的延迟，因而这种方法显得很孤立
HPCC	Thor 集群、Roxie 集群以及 ECL、ECL IDE、ESP 等组件	HPCC 是一种分布式密集数据处理平台，可充分满足数据密集型计算需求，提供了大数据流管理服务，组件间相对独立，处理数据高速并行，以数据为中心的语言；可靠性高，扩展性好，亚马逊已经部署 HPCC 在其云计算平台上	HPCC 未能在开源社区让更多的大型企业和开发者看到它处理大数据的优势，开发生态环境有待改善

3.2 大数据的处理技术

从大数据的处理过程来看，大数据处理的关键技术包括：大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现和应用（大数据检索、大数据可视化、大数据应用、大数据安全等）。

3.2.1 基于数据存储的大数据处理技术

在大数据处理技术方面，Google 起步比较早，自行开发了 GFS，随着发展的需要又不断出现了第二代 GFS——Colossus、BigTable[18] 和 Megastore[18]。在 BigTable 和 Megastore 的基础上诞生了 Spanner[18]，其功能主要是源于一个用 GPS 和原子钟实现的时间 API，这个 API 能将数据中心之间的时间同步精确到 10 ms 以内。基于 Spanner 服务器，2012 年 6 月，Google 研究院就推出被称为 F1（fault tolerant distributed RDBMS）[19] 的新型数据库。微软自行开发的分布式计算平台 Cosmos[20]，能够存储和分析大规模数据集，其宗旨是能够在成千上万台服务器集群上运行。Cosmos 这个平台主要包括 Cosmos 存储系统、Cosmos 执行环境和一种高级脚本语言 SCOPE（structured computations optimized for parallel execution）。

作为社交网络的代表，FaceBook 也在变革着自己原来的存储技术。Facebook 推出了海量小文件的文件处理系统 Haystack[21]，同时 Facebook 还结合自己的应用场景提出了实时的 Hadoop 系统[23]。为了改善 MapReduce 的易用性，Facebook 提出了基于 Hadoop 的大型数据仓库 Hive[23]，它的目标就是简化 Hadoop 上的数据聚集、Ad hoc 查询和大数据的分析等操作。表 2 是对这些相关的存储技术优缺点及应用场景的比较。

3.2.2 基于数据挖掘的大数据处理技术

数据的分析离不开数据的挖掘，大数据也不例外。表 3 详细比较了几种重要的数据挖掘技术的优缺点和适用场景。

表 2 数据存储技术的比较

处理技术	优点	缺点	适用场景	开发机构
BigTable	提供了容错能力和持续的数据库，能自动负载均衡，可扩展	不能提供强一致性和事务级别的需求	大规模数据存储	Google
Megastore	有类似 RDBMS 的数据模型，支持同步复制，利用 Paxos 协议保证实体群组内的数据具有 ACID 语义的强一致性	吞吐量小，不能适应应用要求	大规模数据存储	Google
Spanner	“临时多版本”的数据库取代了 BigTable 的版本化 key-value 存储，时间 API，这个 API 能将数据中心之间的时间同步精确到 10 ms 以内，并且可扩展，全球分布，支持外部一致的事务	在复杂的 SQL 查询上并不能保证所有的节点都能高效地执行，数据中心间数据的传输时间的延迟较长	超大规模的数据存储	Google
F1	融合了 BigTable 的高扩展性和 SQL 数据库的可用性和功能性，底层由 Spanner 支撑，可以同时提供强一致性和弱一致性；高可用性；事务提交延迟为 50 ~ 100 ms，读延迟为 5 ~ 10 ms，高吞吐率	并行查询执行、故障恢复、隔离、优化、迁移应用时要求不宕机等方面还面临很多挑战	大规模的数据存储	Google
Cosmos	数据的多次复制保证了它强大的数据容错能力，拥有自己的硬件和软件机制，可保证系统的可靠运行；通过增加集群中服务器的数量来实现存储容量和计算量的增加；数据并行分布减少了总的运行时间，处理 PB 级的数据比传统的方法花费少	元数据处理性能较差，这是由其元数据分布策略决定的；提供管理工具，使用起来不够方便，系统的可扩展性还有待提高	大规模数据的存储	微软
Haystack	副本技术实现容错，简化元数据结构，以追写的方式存储图片，效率高，图片对应有 index 文件，重启时易创建；通过在主存中执行所有元数据的查询来减少磁盘的操作，提高整个系统的吞吐量，双节点热备的 Avatar-Node，提高了结点的可用性	系统的可扩展性有待提高	图片的存储，大规模的社交网络中共享图片的请求	Facebook

表 3　数据挖掘技术的比较

处理技术	优点	缺点	适用场景
WEKA [24]	开源，基于 Java，集合了大量能承担数据挖掘任务的机器学习算法，可通过简单的 API 进行扩展，可通过 GUI 自动整合一些新的学习算法，WEKA 3.6 版本 [25] 支持导入的 PMML 模型	WEKA 提供的分类功能还不够灵活，只能定长度和定频率地分类	大数据挖掘和分析、大数据的预处理，单机
RapidMiner [26～28]	较先进，接口易用，拥有自己的数据库，直观的 GUI，不受限制的运行平台，可实现多种数据资源的访问，如 Excel、Access、Oracle、IBM DB2，结合 Hadoop 出现的 Radoop 隐藏数据分析的复杂性	对并行性支持的能力有限，也缺乏在多机上长时间运行的能力	大数据的分析处理，单机
PMML [29,30]	基于 XML，定义了表达数据挖掘模型的格式，拥有自己的数据字典和挖掘模型，无须独自开发数据挖掘模块，灵活，可共享；结合算术和逻辑操作在复杂的数据预处理过程中构建的模型能够预测现实世界中不断增长的数据，能够加速分析模型的调用，从而减少大数据处理的时间延迟，让大数据的实时分析成为可能	建模前需要进行大量的数据转换，模型并没有真正地实现与数据的分离，也没有被所有的数据挖掘软件商采纳	医疗和疾病预测中数据的统计、分析，模型的预测
Mahout [31]	基于 Hadoop 的分布式数据挖掘开源项目，具有平台独立性；基于 Java 的可扩展推荐引擎，在 Apache 的 Hadoop Map/Reduce 下高效实现了基于内容和基于用户的推荐算法，解决当前大数据条件下推荐时延问题 [32]	有的算法不是很高效，集群算法和数据类型有时会影响结果的输出	在大数据的预处理过程中可对不必要的数据进行过滤
Dryad [33]	通过集群处理大规模的数据挖掘，对无环图类型数据流和对 TCP 管道的利用，避免了昂贵的磁盘写以及数据边共享内存，在性能上有很大的改善；此模型与 MapReduce 有很多相似之处，但不必按照 map/distribute/ sort/ reduce 的顺序去操作，与 MPI [34]、PVM [35]、GPUs [36] 计算有密切联系	处理数据的可靠性不够高	大规模的数据挖掘，特别是针对有向无环图的数据流
Pregel [37]	是一种可扩展的图处理计算模型，顶点并行计算，可扩展性好，可随着图规模的增加进行分块处理；解决了传统的单机图处理算法限制处理问题的规模、现存的并行图处理系统的容错等问题，PageRank 等算法都已在其上实现	计算模型与 MapReduce 有很多相似之处但更复杂	挖掘图数据，比如在线社交网络的社交图谱

表 4　查询技术的比较

查询技术	优点	缺点	查询性能	适用场景
Dremel [38]	列存储，结合了 Web 搜索的多级查询树(图 2 [38]) 和并行的 DBMS 技术，聚集查询；嵌套的层次数据模型避免大量的连接操作，节省查询时间，速度和规模兼顾	在代数规范形式、连接、可扩展机制方面需深入研究	执行同样的操作，MapReduce 需要的时间是分钟级，而 Dremel 需要的时间是秒级	Web 数据级别的交互式数据分析
PowerDrill [39]	列存储，基于双层字典的压缩存储的技术，普通的数据模型，对数据分区进行了组合，分析时可以跳过很多不需要的分区	PowerDrill 数据 load，增加数据(估计) 不太方便	速度快于 Dremel	处理核心数据，分析少量的大数据集
Impala [40]	基于 Hadoop，采用并行数据库的思想，省去了启动任务的开销以及一些查询上的优化和高效的 C++ 语言	批处理任务的查询优势不明显	比基于 MapReduce 的 Hive SQL 查询速度提升 3～90 倍，在 SQL 功能上要优于 Dremel	处理输出数据适中或比较小的查询
Caffeine [41]	构建在 Spanner 之上，采用 Percolator 更新索引 [42]，提高了网络索引的时效性，与之前的系统相比新系统可提供“50% 新生”的搜索结果	对 PR 值很高的网站没什么优势	速度快，相同时间内检索出的相关结果多	快速检索
Percolator [43] Nectar [44] DryadInc [45]	基于增量计算，避免了从零开始重新计算，通过增加对资源的利用来减少检索的延迟；Percolator 引入了对事务的处理，Nectar 共享子计算，DryadInc 主要依靠 IDE 和 MER 两种算法 [33] 实现	DryadInc 实际应用受很多现实因素的制约	对旧数据的重复计算上性能提升较多	大规模数据的冗余计算

3.2.3　基于查询的大数据处理技术

在大数据的处理过程中，数据分析是关键，数据分析主要是依赖于数据分析工具。表 4 对几种比较典型的数据查询技术进行了深入分析。

除表 4 所列举的技术外，文献 [46] 提出一种新奇的大数据分析方法——危险理论(danger theory)，这种危险理论是来源于生物免疫系统，但又不同于传统的人工免疫系统。在关键特征和属性的选择上引入危险理论，主要是被用做数据过滤策略，可提高数据分析的效率。在危险理论中关注的是潜在危险，捕捉危险信号，用数值微分法判断危险信号。这个处理模型与之前的编程处理模型相比具有自学习能力和智能性，它在数据的预处理阶段有明显的效果，更适合于快速的数据过滤。

Dremel 的系统结构及多级查询图如图 2 所示。

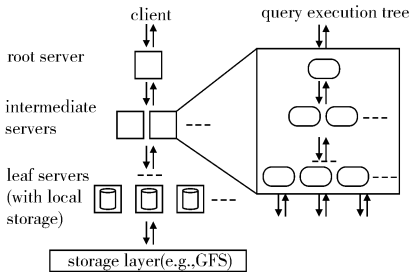

图 2　Dremel 的系统结构及多级查询图

4　大数据的研究与发展方向

尽管大数据的时代已经到来，各界也发现了大数据的巨大价值，但是大数据的研究还处在初始阶段。随着研究的不断深入，大数据所面临的问题也越来越多，如何让大数据朝着有利于全社会的方向发展就需要全面地研究大数据，以下是几种可能的大数据未来的研究与发展方向。

1) 关系数据库和非关系数据库的融合

众所周知，关系数据库系统在数据分析中占据着主要地位，但是随着后来半结构化和非结构化数据的大量涌现，关系数据库系统就无所适从了。而类似于 MapReduce 的大数据处理工具在容错性、可扩展性、数据的移动性上明显优于关系数据库系统，但在处理数据的实时性能上，MapReduce 与 RDBMS 相比还有一定的差距。关系数据库和非关系数据库各有所长，如果在以后的大数据的研究处理过程中，能将关系数据库系统和分布式并行处理系统进行有效的结合，而不是将二者明显地区分开来，那么大数据的分析效率将在很大程度上得到提高。

2) 数据的不确定性与数据质量

大数据，顾名思义是数据量非常大，如何从这些庞大的数据量中提取到尽可能多的有用信息就涉及到数据质量的问题。在网络环境下，不确定性的数据广泛存在，并且表现形式多样，这样大数据在演化的过程中也伴随着不确定性。文献 [47] 提到了网络大数据的不确定性，其实大数据的不确定性不仅仅适用于网络大数据，对一般大数据而言也存在这种不确定性。大数据的不确定性要求人们在处理数据时也要应对这种不确定性，包括数据的收集、存储、建模、分析都需要新的方法来应对。

这样也给学习者和研究者带来了很大的挑战，数据质量就很难得到保证，况且大数据的研究领域尚浅，本身就有很多亟待解决的问题。面对不断快速产生的数据，在数据分析的过程中很难保证有效的数据不丢失，而这种有效的数据才是大数据的价值所在，也是数据质量的体现。所以需要研究出一种新的计算模式，一种高效的计算模型和方法，这样数据的质量和数据的时效性才能有所保证。文献[48]中几位从事大数据研究的专家也强调了数据质量的重要性，中国工程院院士、西安交通大学教授汪应洛认为，在大数据产业发展中，数据质量也是一大障碍，不容忽视，他说“数据质量是大数据产业这座大厦的基础，如果数据质量不高，基础不牢靠，大数据产业就可能岌岌可危，甚至根本无从发展。”所以处理好大数据的不确定性、提高数据质量是大数据研究中的重中之重。

3) 跨领域的数据处理方法的可移植性

大数据自身的特点决定了大数据处理方法的多样性、灵活性和广泛性。而今几乎每个领域都有涉及到大数据，在分析处理大数据的建模过程中除了要考虑大数据的特点外还可以结合其他领域的一些原理模型，如文献[46]提出的用来源于生物免疫系统的计算模型去处理大数据中的关键属性的选择。还有统计学中的统计分析模型，特别是对原始数据的统计和计量，音频、视频、照片等重要信息。广泛吸纳其他研究领域的原理模型，然后进行有效的结合，从而提高大数据处理的效率，这可能会成为以后大数据分析处理的重要方法。

4) 大数据的预测性作用日益凸显

提及大数据，它的作用自然是不言而喻，也有不少专家进行了总结，大数据有变革价值的力量、大数据有变革经济的潜力、大数据有变革组织的潜能。但是从很多大数据的应用案例分析不难发现，无论是大数据的研究者还是普通人，大数据给人们带来的最直接的利益就是对未来的预见。气象部门可以根据气象数据预测未来的天气变化；经销商可根据商品的销量分析客户的喜好从而制定未来的采购计划及时调整经营模式，增加利润；通信部门通过对大数据的分析实时了解市场行情，从而作出合理决策。由已知推测未知，通过大数据可以提高对未知预测的可靠性和精准性，这对整个人类来说都是一种进步。

5 结束语

大数据已经涉及到生活的各个领域，对于大数据的研究涉及的领域也很广。与人们直接利益相关的大数据的能耗、安全、隐私保护等都受到了很多企业和个人的关注，还有更多未知的领域也不例外。本文主要是在对大数据处理工具和处理技术对比分析的基础上给出了大数据未来几个可能的研究和发展方向：关系数据库和非关系数据库的融合、数据的不确定性和数据质量、跨领域的数据处理方法的可移植性、大数据的预测性作用日益凸显。大数据的发展尚在起步阶段，需要人们不断开拓的空间很大，如何高效地处理大数据、合理地利用大数据仍需要不断地探索发现。

参考文献：

[1] 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)：647-657.

[2] GRAHAM-ROWE D，GOLDSTON D，DOCTOROW C，*et al*. Big data：science in the petabyte era[J]. Nature，2008，455(7209)：8-9.

[3] JI Chang-qing，LI Yu，QIU Wen-ming，*et al*. Big data processing in cloud computing environments[C]//Proc of the 12th International Symposium on Pervasive Systems，Algorithms and Networks. 2012：17-23.

[4] 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/.

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

[6] 顾芳，刘旭峰，左超. 大数据背景下运营商移动互联网发展策略研究[J]. 邮电设计技术，2012(8)：21-24.

[7] GANTZ J，REINSEL D. 2011 digital universe study：extracting value from chaos[EB/OL]. (2011-07). http://www.b-eye-network.com/blogs/devlin/archives/2011/071.

[8] Big data[J]. ERCIM News，2012，89.

[9] 冯海超. 透视美国大数据爆发全景[N]. 互联网周刊，2013-01-14.

[10] 邬雪艳，孙永杰. 云计算和大数据助力医疗协同[N]. 通信世界，2013-04-17.

[11] 陈美. 大数据在公共交通中的应用[J]. 图书与情报，2012(6)：22-28.

[12] 李伟，赵春宇. 油田勘探开发“大数据”管理及应用[J]. 信息技术，2013(4)：196-198.

[13] 冯芷艳，郭迅华，曾大军，等. 大数据背景下商务管理研究若干前沿课题[J]. 管理科学学报，2013，16(1)：1-9.

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

[15] HDFS architecture guide[EB/OL]. [2013-08-04]. http://hadoop.apache.org/docs/rl.2.1/hdfs_design.html.

[16] SAGIROGLU S，SINANC D. Big data：a review[C]//Proc of International Conference on Collaboration Technologies and Systems. 2013：42-47.

[17] Hadapt[EB/OL]. http://hadapt.com/product/.

[18] CORBETT J C，DEAN J，EPSTEIN M，*et al*. Spanner：Google’s globally distributed database[C]//Proc of the 10th USENIX Conference on Operation Systems Design and Implementation. Berkeley：USENIX Association，2012：251-264.

[19] SHUTE J，OANCEA M，ELLNER S. F1：the fault-tolerant distributed RDBMS supporting Google’s Ad business[C]//Proc of ACM SIGMOD. 2012.

[20] CHAIKEN R，JENKINS B，LARSON P，*et al*. SCOPE：easy and efficient parallel processing of massive data sets[J]. Proceedings of the VLDB Endowment，2008，1(2)：1265-1276.

[21] BEAVER D，KUMAR S，LI H C，*et al*. Finding a needle in Haystack：Facebook’s photo storage[C]//Proc of the 9th USENIX Conference on Operation Systems Design and Implementation. Berkeley：USENIX Association，2010：47-60.

[22] BORTHAKUR D，SARMA J S，GRAY J，*et al*. Apache Hadoop goes realtime at Facebook[C]//Proc of ACM SIGMOD Conference on Management of Data. New York：ACM Press，2011：1071-1080.

[23] 王珊，王会举，覃雄派，等. 架构大数据：挑战、现状与展望[J]. 计算机学报，2011，34(10)：1741-1752.

[24] 郑世明，苗壮，宋自林，等. WEKA 环境下基于模糊理论的聚类算法[J]. 解放军理工大学学报：自然科学版，2012，13(1)：22-26.

[25] HALL M，FRANK E，HOLMES G，*et al*. The WEKA data mining software：an update[J]. ACM SIGKDD Explorations，2009，11(1)：10-18.

[26] ARIMOND A，KOFLER C，SHAFAIT F. Distributed pattern recognition in RapidMiner[C]//Proc of RapidMiner Community Meeting. 2010：1-6.

[27] ARIMOND A. A Distributed system for pattern recognition and machine learning[D]. Kaiserslautern：DFKI，2010.

[28] PREKOPCSÁK Z，MAKRAI G，HENK T，*et al*. Radoop：analyzing big data with RapidMiner and Hadoop[C]//RapidMiner Community Meeting and Conference. 2011.

[29] 焦雷. 基于 PMML 数据挖掘应用研究[J]. 电子设计工程，2012，20(8)：20-23.

(下转第 1623 页)

[53] 施文，刘志学，杨威. 零部件循环取货越库物流系统仿真优化[J]. 计算机集成制造系统，2012，18(12)：2765-2776.
[54] WAN Xiao-tao，PEKNY J F，REKLAITIS G V. Simulation-based optimization with surrogate models：application to supply chain management[J]. Computers & Chemical Engineering，2005，29(6)：1317-1328.
[55] SANCHEZ S M，SANCHEZ P J，RAMBERG J S，*et al*. Effective engineering design through simulation[J]. International Trans on Operational Research，1996，3(2)：169-185.
[56] FEYZIOGLU O，PIERREVAL H，DEFLANDRE D. A simulation-based optimization approach to size manufacturing systems[J]. International Journal of Production Research，2005，43(2)：247-266.
[57] RAMAEKERS K，JANSSENS G K，LANDEGHEM H V. Toward logistics systems parameter optimization through the use of response surfaces[J]. Quarterly Journal of the Belgian，French and Italian Operations Research Societies，2006，4(4)：331-342.
[58] KUMAR S，NOTTESTAD D A. Capacity design：an application using discrete-event simulation and designed experiments[J]. IIE Transactions，2006，38(9)：729-736.
[59] EKREN B Y，HERAGU S S，KRISHNAMURTHY A，*et al*. Simulation based experimental design to identify factors affecting performance of AVS/RS[J]. Computers & Industrial Engineering，2010，58(1)：175-185.
[60] BARTON R R，MECKESHEIMER M. Chapter 18 metamodel-based simulation optimization[M]//SHANE G H，BARRY L N. Handbooks in Operations Research and Management Science. Boston：Elsevier，2006：535-574.
[61] MYERS R H，KHURI A I，CARTER W H. Response surface methodology：1966—1988[J]. Technometrics，1989，31(2)：137-157.
[62] KLEIJNEN J P C，Van BEERS W C M. Application-driven sequential designs for simulation experiments：Kriging metamodelling[J]. Journal of the Operational Research Society，2004，55(8)：876-883.
[63] KLEIJNEN J P C. Kriging metamodeling in simulation：a review[J]. European Journal of Operational Research，2009，192(3)：707-716.
[64] LOPHAVEN S N，NIELSEN H B，SØNDERGAARD J. DACE：a MATLAB Kriging toolbox (version 2.0)[M]. Lyngby：Technical University of Denmark，2002.
[65] KLEIJNEN J P C，SARGENT R G. A methodology for fitting and validating metamodels in simulation[J]. European Journal of Operational Research，2000，120(1)：14-29.
[66] SARGENT R G. Verification and validation of simulation models[C]//Proc of Winter Simulation Conference. 2009：162-176.
[67] TAGUCHI G. System of experimental designs[M]. New York：UNIPUB/Krauss International Publications，1987.
[68] NAIR V N，ABRAHAM B，MACKAY J，*et al*. Taguchi's parameter design：a panel discussion[J]. Technometrics，1992，34(2)：127-161.
[69] SANCHEZ S M. Design of experiments：robust design：seeking the best of all possible worlds[C]//Proc of the 32nd Conference on Winter Simulation. San Diego：Society for Computer Simulation International，2000：69-76.
[70] DELLINO G，KLEIJNEN J P C，MELONI C. Robust optimization in simulation：Taguchi and response surface methodology[J]. International Journal of Production Economics，2010，125(1)：52-59.
[71] DELLINO G，KLEIJNEN J P C，MELONI C. Robust optimization in simulation：Taguchi and Krige combined[J]. INFORMS Journal on Computing，2011，24(3)：471-484.
[72] FOWLER J W，ROSE O. Grand challenges in modeling and simulation of complex manufacturing systems[J]. Simulation，2004，80(9)：469-476.

(上接第 1616 页)

[30] PMML：accelerating the time to value for predictive analytics in the big data era[R]. [S. l.]：Sybase，2012.
[31] OWEN S，ANIL R，DUNNING T，*et al*. Mahout in action[M]. [S. l.]：Manning Publications，2011：3-10.
[32] 朱倩，钱立. 基于 Mahout 的推荐系统的分析与设计[J]. 科技通报，2013，29(6)：35-36.
[33] ISARD M，BUDIU M，YU Yuan，*et al*. Dryad：distributed data-parallel programs from sequential building blocks[C]//Proc of the 2nd ACM SIGOPS/Euro Sys Conference on Computery System. New York：ACM Press，2007：59-72.
[34] Open MPI[EB/OL]. http://www.open-mpi.org/.
[35] SUNDERAM V S. PVM：a framework for parallel distributed computing concurrency[J]. Concurrency，1990，2(4)：315-339.
[36] TARDITI D，PURI S，OGLESBY J. Accelerator：using data-parallelism to program GPUs for general-purpose uses[C]//Proc of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems. New York：ACM Press，2006：325-335.
[37] MALEWICZ G，AUSTERN M H，BIK A J C，*et al*. Pregel：a system for large-scale graph processing[C]//Proc of ACM SIGMOD Conference on Management of Data. New York：ACM Press，2010：135-146.
[38] MELNIK S，GUBAREV A，LONG Jing-jing，*et al*. Dremel：interactive analysis of Web-scale datasets[J]. Proceedings of the VLDB Endowment，2010，3(1-2)：330-339.
[39] HALL A，BACHMANN O，BUSSOW R，*et al*. Processing a trillion cells per mouse click[J]. Proceedings of the VLDB Endowment，2012，5(11)：1436-1446.
[40] 耿益锋，陈冠诚. Impala：新一代开源大数据分析引擎[J]. 程序员，2013(8)：95-97.
[41] IYER S C，UTTS M. Help test some next generation infrastructure[EB/OL]. (2009-08-10). http://googlewebmastecentral. blogspot.com/2009/08/help-test-some-next-generation. html.
[42] 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.
[43] PENG D，DABEK F. Large-scale incremental processing using distributed transactions notifications[C]//Proc of the 9th USENIX Conference on Operating Systems Design and Implementation. Berkeley：USENIX Association，2010：1-15.
[44] GUNDA P K，RAVINDRANATH L，THEKKATH C A，*et al*. Nectar：automatic management of data and computation in data centers[C]//Proc of the 9th USENIX Conference on Operating Systems Design and Implementation. Berkeley：USENIX Association，2010：75-88.
[45] POPA L，BUDIU M，YU Yuan，*et al*. DryadInc：reusing work in large scale computations[C]//Proc of Conference on Hot Topics in Cloud Computing. Berkeley：USENIX Association，2009：1-14.
[46] LU Lin，LIANG Yi-wen，YANG He，*et al*. Danger theory：a new approach in big data analysis[C]//Proc of International Conference on Automatic Control and Artificial Intelligence. 2012：739-742.
[47] 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)：1126-1136.
[48] 赵海娟. 掘金大数据 亟待国家战略支持[N]. 中国经济时报，2013-01-22(A02).
[49] GÜNNEMANN S，KREMER H，MUSIOL R，*et al*. A subspace clustering extension for the KNIME data mining framework[C]//Proc of the 12th IEEE International Conference on Data Mining. Washington DC：IEEE Computer Society，2012：886-889.