中国科学: 信息科学 2015 年 第 45 卷 第 1 期: 1–44
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

大数据系统综述

李学龙[①*], 龚海刚[②]

① 中国科学院西安光学精密机械研究所光学影像分析与学习中心 (OPTIMAL), 西安 710119
② 电子科技大学计算机科学与工程学院, 成都 611731
* 通信作者. E-mail: xuelong_li@opt.ac.cn

收稿日期: 2014–09–30; 接受日期: 2014–11–21
国家自然科学基金 (批准号: 61125106) 资助项目

摘要 随着科学、技术和工程的迅猛发展, 近 20 年来, 许多领域 (如光学观测、光学监控、健康医护、传感器、用户数据、互联网和金融公司以及供应链系统) 都产生了海量的数据 (更恰当的描述或许是 "无限" 的数据, 例如, 在光学观测和监控等应用中, 数据都是源源不断而来的, 形成了 "数据灾难"), 大数据的概念也随之再次引起重视. 与传统的数据相比, 除了大容量等表象特点, 大数据还具有其他独特的特点, 例如大数据通常是无结构的, 并且需要得到实时分析, 因此大数据的发展需要全新的体系架构, 用于处理大规模数据的获取、传输、存储和分析. 本文对大数据分析平台进行了尽可能详尽的文献调研, 首先介绍了大数据的基本定义和大数据面临的一些挑战; 然后提出了大数据系统框架, 将大数据系统分解为数据生成、数据获取、数据存储和数据分析等 4 个模块, 这 4 个模块也构成了大数据价值链; 随后讨论了学术界和工业界中和大数据相关的方法和机制; 最后介绍了典型的大数据系统基准和大数据的一些科学问题. 本文意图为非专业读者提供大数据的全景知识, 也为高级读者定制自己的大数据解决方案提供辅助思想, 希望能够对大数据相关的科技和工程人员起到一些参考作用.

关键词 大数据 数据获取 数据存储 数据处理 数据分析

1 引言

近年来, "大数据" 已广为人知, 并被认为是信息时代的新 "石油", 这主要基于两点共识. 首先, 在过去 20 年间, 数据产生速度越来越快. 据国际数据公司 IDC 报道 [1], 2011 年产生和复制的数据量超过 1.8 Z 字节, 是过去 5 年数据增长的 9 倍, 并将以每两年翻倍的速度增长. 其次, 大数据中隐藏着巨大的机会和价值, 将给许多领域带来变革性的发展. 因此, 大数据研究领域吸引了产业界、政府和学术界的广泛关注. 例如, 产业界报告 [2] 和公共媒体 (*Economists* [3]1), *New York Times* [4], 美国国家公共广播电台 [5,6]) 中充斥了大数据的相关信息; 政府部门设立重大项目加速大数据的发展 [7]; *Nature*2)和*Science*3) 等期刊也发表了大数据挑战相关的论点. 毫无疑问, 大数据时代已经到来.

1) Economist T. Drowning in numbers – digital data will flood the planet and help us understand it better. http://www.economist.com/blogs/dailychart/2011/11/bigdata-0/.
2) Big Data. 2008. http://www.nature.com/news/specials/bigdata/index.html.
3) Special Online Collection: Dealing with Big Data. 2011. http://www.sciencemag.org/site/special/data/.

引用格式: 李学龙, 龚海刚. 大数据系统综述. 中国科学: 信息科学, 2015, 45: 1–44, doi: 10.1360/N112014-00290

大数据的数据集大小以难以想象的速度增长, 给数据处理带来了极大的挑战. 首先, 信息技术的发展使得数据的生成和消费变得更容易. 例如, 每分钟有 72 小时长度的视频被上传到 Youtube 服务器[4]. 大数据的这种大容量特性使得数据难以可伸缩地从分布式的地点收集并集成. 第二, 数据采集后, 如何以最小的硬件和软件代价存储和管理这些海量异构数据是非常具有挑战性的问题. 第三, 由于大数据的异构性、规模性、实时性、复杂性和隐私性等特点, 大数据分析必须在不同层次 (建模、可视化、预测和优化) 高效地挖掘数据以提高决策效率. 这些挑战迫切地需要对整个数据管理系统的各个层次 (从体系架构到具体机制) 进行变革. 但是如果能有效地管理大数据, 就能够给许多领域, 如科学和环境建模、健康医护和能源保护带来巨大的变革. 国际策略咨询公司 McKinsey 的研究报告 [2] 表明, 全球个人位置信息的潜在价值达到 7000 亿, 并且能降低产品开发和集成成本的一半以上.

然而, 传统的数据管理和分析系统是基于关系型数据库管理系统 (RDBMS) 的. 这些系统在处理结构化数据时性能突出, 但是对半结构化或无结构化数据的处理却无法提供有力的支持. 此外, RDBMS 可以通过增加昂贵的硬件向上扩展 (scale up), 但是无法通过并行增加硬件实现向外扩展 (scale out). 显然, 传统的 RDBMS 无法处理如今大数据的规模和异构性. 为了解决这些挑战, 学术界和产业界从不同角度提出了大数据系统的解决方案. 而云计算平台可以作为大数据系统的基础设施层以满足特定的基础设施需求, 例如成本效率、灵活性以及向上或向下扩展的能力.

分布式文件系统 [8] 和 NoSQL 数据库 [9] 适用于数据持久存储和模式自由 (scheme free) 的海量数据管理. MapReduce [10] 编程框架在处理组聚合 (group-aggregation) 任务, 如网站排名方面, 获得了极大的成功. Hadoop [11] 则集成了数据存储、数据处理、系统管理和其他模块, 提供了强大的系统级解决方案, 成为大数据处理的主流. 基于这些革新性的技术和平台, 可以构建多样的大数据应用.

本文对大数据领域进行系统性的介绍, 为理解大数据平台、开发大数据应用以及从事大数据的研究提供指导. 本文引入了大数据系统的通用框架, 该框架将大数据平台分为数据生成、数据获取、数据存储和数据分析 4 个处理阶段, 并对每一个阶段的当前研究进展进行了调研, 提出了架构设计的工程级观点, 对大数据的不同分析实例进行了探讨. 此外, 本文还比较了大数据系统的评价基准, 并归纳了大数据存在的科学问题和研究方向.

2 大数据国内外现状

大数据的快速发展, 使之成为信息时代的一大新兴产业, 并引起了国内外政府、学术界和产业界的高度关注.

2.1 国外研究现状

早在 2009 年, 联合国就启动了 “全球脉动计划”, 拟通过大数据推动落后地区的发展, 而 2012 年 1 月的世界经济论坛年会也把 “大数据, 大影响” 作为重要议题之一. 在美国, 2009 年至今, Data.gov (美国政府数据库) 全面开放了 40 万政府原始数据集, 大数据已成为美国国家创新战略、国家安全战略以及国家信息网络安全战略的交叉领域和核心领域. 2012 年 3 月, 美国政府提出 “大数据研究和发展倡议”, 发起全球开放政府数据运动, 并投资 2 亿美元促进大数据核心技术研究和应用, 涉及 NSF, DARPA 等 6 个政府部门和机构, 把大数据放在重要的战略位置. 英国政府也将大数据作为重点发展的科技领域, 在发展 8 类高新技术的 6 亿英镑投资中, 大数据的注资占三成. 2014 年 7 月, 欧盟委员

4) Youtube Statistics. http://www.youtube.com/yt/press/statistics.html.

会也呼吁各成员国积极发展大数据, 迎接 “大数据” 时代, 并将采取具体措施发展大数据业务. 例如建立大数据领域的公私合作关系; 依托 “地平线 2020” 科研规划, 创建开放式数据孵化器; 成立多个超级计算中心; 在成员国创建数据处理设施网络.

在学术界, 美国麻省理工大学 (MIT) 计算机科学与人工智能实验室 (CSAIL) 建立了大数据科学技术中心 (ISTC). ISTC 主要致力于加速科学与医药发明、企业与行业计算, 并着重推动在新的数据密集型应用领域的最终用户体验的设计创新. 大数据 ISTC 由 MIT 作为中心学校, 研究专家们来自 MIT、加州大学圣巴巴拉分校、波特兰州立大学、布朗大学、华盛顿大学和斯坦福大学等 6 所大学. 通过明确和资助领域带头人、提供合作研究中心的方式, 目标是发掘共享、存储和操作大数据的解决方案, 涉及 Intel, Microsoft, EMC 等多家国际产业巨头. 同时, 英国牛津大学成立了首个综合运用大数据的医药卫生科研中心, 该中心的成立有望给英国医学研究和医疗服务带来革命性变化, 它将促进医疗数据分析方面的新进展, 帮助科学家更好地理解人类疾病及其治疗方法. 该中心通过搜集、存储和分析大量医疗信息, 确定新药物的研发方向, 减少药物开发成本, 同时为发现新的治疗手段提供线索. 而以英国为首的欧洲核子中心 (CERN) 也在匈牙利科学院魏格纳物理学研究中心建设了一座超宽带数据中心, 该中心将成为连接 CERN 且具有欧洲最大传输能力的数据处理中心.

在产业界, 国外许多著名企业和组织都将大数据作为主要业务, 例如 IBM, Microsoft, EMC, DELL, HP 等国际知名厂商都提出了各自的大数据解决方案或应用. IBM 宣布了收购 Star Analytics (星分析公司) 软件产品组合的消息. 除了 Star Analytics, 在 IBM 最新的收购计划中, Splunk 和 NetApp 是最热门的收购目标. 据不完全统计, 从 2005 年起, IBM 花费超过 160 亿美元收购了 35 家与大数据分析相关的公司. 此外, IBM 还和全球千所高校达成协议, 就大数据的联合研究、教学、行业应用案例开发等方面开展全面的合作.

无疑, 欧美等国家对大数据的探索和发展已走在世界前列, 各国政府已将大数据发展提升至战略高度, 大力促进大数据产业的发展.

2.2 国内研究现状

我国政府、学术界和产业界也早已经开始高度重视大数据的研究和应用的工作, 并纷纷启动了相应的研究计划. 挂一漏万, 鉴于我们的了解面所限, 本文仅能够简要介绍其中的一些.

在政府层面, 科技部 “十二五” 部署了关于物联网、云计算的相关专项. 2012 年, 中国科学院院长白春礼院士呼吁中国应制定国家大数据战略. 同年 3 月, 科技部发布的《“十二五” 国家科技计划信息技术领域 2013 年度备选项目征集指南》中的 “先进计算” 板块已明确提出 “面向大数据的先进存储结构及关键技术”, 国家 “973 计划”、“863 计划”、国家自然科学基金等也分别设立了针对大数据的研究计划和专项. 目前已立项 “973 计划” 项目 2 项, “973 计划” 青年项目 2 项, 国家自然科学基金重点项目 2 项. 地方政府也对大数据战略高度重视, 2013 年上海市提出了《上海推进大数据研究与发展三年行动计划》, 重庆市提出了《重庆市人民政府关于印发重庆市大数据行动计划的通知》, 2014 年广东省成立大数据管理局负责研究拟订并组织实施大数据战略、规划和政策措施, 引导和推动大数据研究和应用工作. 贵州、河南和承德等省市也都推出了各自的大数据发展规划.

在学术研究层面, 国内许多高等院校和研究所开始成立大数据的研究机构. 与此同时, 国内有关大数据的学术组织和活动也纷纷成立和开展. 2012 年中国计算机学会和中国通信学会都成立了大数据专家委员会, 教育部也在人民大学成立 “萨师煊大数据分析与管理国际研究中心”. 近年来开展了许多学术活动, 主要包括: CCF 大数据学术会议、中国大数据技术创新与创业大赛、大数据分析与管理国际研讨会、大数据科学与工程国际学术研讨会、中国大数据技术大会和中国国际大数据大会等.

在产业层面, 国内不少知名企业或组织也成立了大数据产品团队和实验室, 力争在大数据产业竞争中占据领先地位.

3 大数据基础

本节首先介绍了大数据的一些主流定义, 随后介绍大数据的发展历史, 并讨论两种大数据处理方式: 流处理和批处理.

3.1 大数据定义

随着大数据的流行, 大数据的定义呈现多样化的趋势, 达成共识非常困难. 本质上, 大数据不仅意味着数据的大容量, 还体现了一些区别于 "海量数据" 和 "非常大的数据" 的特点. 实际上, 不少文献对大数据进行了定义, 其中三种定义较为重要.

• 属性定义 (Attributive definition): 国际数据中心 IDC 是研究大数据及其影响的先驱, 在 2011 年的报告中定义了大数据 [1]: "大数据技术描述了一个技术和体系的新时代, 被设计于从大规模多样化的数据中通过高速捕获、发现和分析技术提取数据的价值". 这个定义刻画了大数据的 4 个显著特点, 即容量 (volume)、多样性 (variety)、速度 (velocity) 和价值 (value), 而 "4Vs" 定义的使用也较为广泛. 类似的定义也出现在 2001 年 IT 分析公司 META 集团 (现在已被 Gartner 并购) 分析师 Doug Laney 的研究报告中 [2], 他注意到数据的增长是三维的, 即容量、多样性和速度的增长. 尽管 "3Vs" 定义没有完整描述大数据, Gartner 和多数产业界巨头如 IBM [12] 和 Microsoft [13] 的研究者们仍继续使用 "3Vs" 模型描述大数据 [14].

• 比较定义 (Comparative definition): 2011 年, McKinsey 公司的研究报告中 [2] 将大数据定义为 "超过了典型数据库软件工具捕获、存储、管理和分析数据能力的数据集". 这种定义是一种主观定义, 没有描述与大数据相关的任何度量机制, 但是在定义中包含了一种演化的观点 (从时间和跨领域的角度), 说明了什么样的数据集才能被认为是大数据.

• 体系定义 (Architectural definition): 美国国家标准和技术研究院 NIST 则认为 [15] "大数据是指数据的容量、数据的获取速度或者数据的表示限制了使用传统关系方法对数据的分析处理能力, 需要使用水平扩展的机制以提高处理效率". 此外, 大数据可进一步细分为大数据科学 (big data science) 和大数据框架 (big data frameworks). 大数据科学是涵盖大数据获取、调节和评估技术的研究; 大数据框架则是在计算单元集群间解决大数据问题的分布式处理和分析的软件库及算法. 一个或多个大数据框架的实例化即为大数据基础设施.

此外, 还有不少产业界和学术界对大数据定义的讨论 [16][5)].

然而对于大数据定义, 要达成共识非常困难. 一种逻辑上的选择是接受所有的大数据定义, 其中每种定义反映了大数据的特定方面. 本文采取这种方式理解大数据科学和工程的共同问题和相关机制. 前面提到的大数据定义给出了一系列工具, 用于比较大数据和传统的数据分析, 比较结果如表 1 所示. 首先, 数据集的容量是区分大数据和传统数据的关键因素. 例如, Facebook 报道 2012 年每天有 27 亿用户登录并发表评论 [17]. 其次, 大数据有三种形式: 结构化、半结构化和无结构化. 传统的数据通常是结构化的, 易于标注和存储. 而现在 Facebook, Twitter, YouTube 以及其他用户产生的绝大多数数据都是非结构化的. 第三, 大数据的速度意味着数据集的分析处理速率要匹配数据的产生速率. 对于

5) Grobelnik M. Big Data Tutorial. http://videolectures.net/eswc2012_grobelnik_big_data.

表 1 大数据和传统数据比较

Table 1 Comparison between big data and traditional data

	Traditional data	Big data
Volume	GB	Constantly updated (TB or PB currently)
Generated rate	Per hour, day, ...	More rapid
Structure	Structured	Semi-structured or un-structured
Data source	Centralized	Fully distributed
Data integration	Easy	Difficult
Data store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real-time

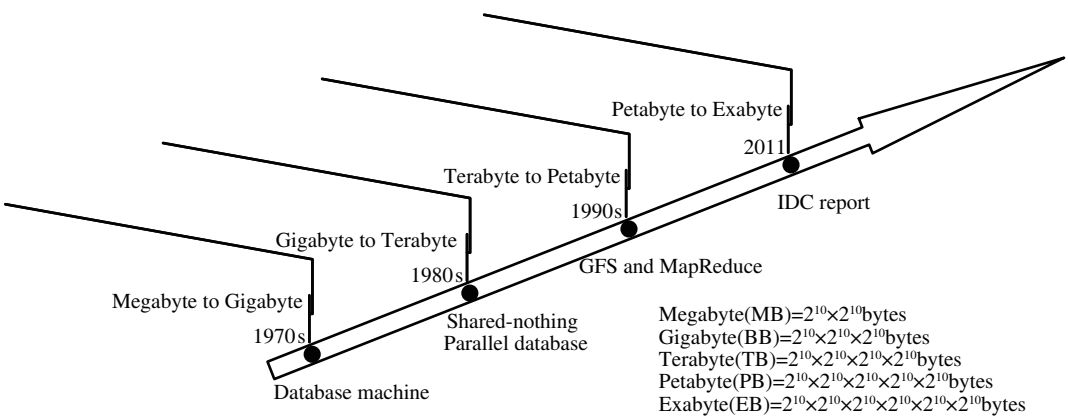

图 1 大数据主要历史里程碑

Figure 1 Milestones of big data history

时间敏感的应用, 例如欺诈检测和 RFID 数据管理, 大数据以流的形式进入企业, 需要尽可能快地处理数据并最大化其价值. 最后, 利用大量数据挖掘方法分析大数据集, 可以从低价值密度的巨量数据中提取重要的价值.

3.2 大数据的历史

以往对大数据的演化过程通常从单方面的观点描述, 例如从年代 [18] 或技术里程碑 [19] 等方面. 本文对大数据的演化过程则根据数据大小来刻画, 大数据的发展历史和有效存储管理日益增大的数据集的能力紧密联系在一起. 每一次处理能力的提高都伴随着新数据库技术的发展, 如图 1 所示. 因此, 大数据的历史可以大致分为以下几个阶段.

• Megabyte 到 Gigabyte: 20 世纪 70 年代到 80 年代, 历史上的商业数据从 Megabyte 达到 Gigabyte 的量级, 从而引入最早的 "大数据" 挑战. 当时的迫切需求是存储数据并运行关系型数据查询以完成商业数据的分析和报告. 数据库计算机 (database machine) 随之产生, 它集成了硬件和软件解决问题, 其思想是通过硬件和软件的集成, 以较小的代价获得较好的处理性能. 一段时间后, 专用硬件的数据库计算机难以跟上通用计算机的发展. 因此, 后来的数据库系统是软件系统, 对硬件几乎没有什么限制, 可以运行在通用计算机上.

• Gigabyte 到 Terabyte: 20 世纪 80 年代末期, 数字技术的盛行导致数据容量从 Gigabyte 达到

Terabyte 级别, 这超出了单个计算机系统的存储和处理能力. 数据并行化技术被提出, 用于扩展存储能力和提高处理性能, 其思想是分配数据和相关任务 (如构建索引和评估查询) 到独立的硬件上运行. 在此基础上, 提出了几种基于底层硬件架构的并行数据库, 包括内存共享数据库、磁盘共享数据库和无共享 (share nothing) 数据库. 其中, 构建在互连集群基础上的无共享数据库取得了较大的成功. 集群由多个计算机构成, 每个计算机有各自的 cpu、内存和磁盘 [20]. 在过去几年, 也出现了无共享数据库类型的产品, 包括 Teradata[6], Netazza[7], AsterData[8], Greenplum[9] 和 Vertica[10]. 这些系统产品使用关系型数据模型和说明性关系查询语言, 并成为使用分治法并行化数据存储的先驱.

• Terabyte 到 Petabyte: 20 世纪 90 年代末期, web 1.0 的迅猛发展将世界带入了互联网时代, 随之带来的是巨量的达到 Petabyte 级别的半结构化和无结构的网页数据. 这需要对迅速增长的网页内容进行索引和查询. 然而, 尽管并行数据库能够较好地处理结构化数据, 但是对于处理无结构的数据几乎没有提供任何支持. 此外, 并行数据库系统的处理能力也不超过几个 Teragbytes. 为了应对 web 规模的数据管理和分析挑战, Google 提出了 GFS 文件系统 [21] 和 MapReduce 编程模型 [10]. GFS 和 MapReduce 能够自动实现数据的并行化, 并将大规模计算应用分布在大量商用服务器集群中. 运行 GFS 和 MapReduce 的系统能够向上和向外扩展, 因此能处理无限的数据. 2000 年代中期, 用户自主创造内容 (user generated contents, UGC)、多种多样的传感器和其他泛在的数据源产生了大量的混合结构数据, 这要求在计算架构和大规模数据处理机制上实现范式转变 (paradigm shift). 模式自由、快速可靠、高度可扩展的 NoSQL 数据库技术开始出现并被用来处理这些数据. 2007 年 1 月, 数据库软件的先驱者 JimGray 将这种转变称为 “第 4 范式” [22]. 他认为处理这种范式的唯一方法就是开发新一代的计算工具用于管理、可视化和分析数据.

• Petabyte 到 Exabyte: 根据现有的发展趋势, 大公司存储和分析的数据毫无疑问将在不久后从 Petabyte 级别达到 Exabyte 级别. 然而, 现有的技术只能处理 Petabyte 级别的数据, 目前仍没有革命性的新技术能够处理更大的数据集. 2011 年 7 月, EMC 发布了名为 “Extracting Value from Chaos” 的研究报告 [1], 讨论了大数据的思想和潜在价值. 该报告点燃了产业界和学术界对大数据研究的热情, 随后几年几乎所有重要的产业界公司, 如 EMC, Oracle, Microsoft, Google, Amazon 和 Facebook, 都开始启动各自的大数据项目. 2012 年 3 月, 美国政府宣布投资 2 亿美元推动大数据研究计划, 并涉及 DAPRA、国家健康研究所 NIH、国家自然科学基金 NSF [7] 等美国国家机构.

3.3 大数据处理方式: 流式处理和批处理

大数据分析是在强大的支撑平台上运行分析算法发现隐藏在大数据中潜在价值的过程, 例如隐藏的模式 (pattern) 和未知的相关性. 根据处理时间的需求, 大数据的分析处理可以分为两类.

• 流式处理: 流式处理假设数据的潜在价值是数据的新鲜度 (freshness) [23], 因此流式处理方式应尽可能快地处理数据并得到结果. 在这种方式下, 数据以流的方式到达. 在数据连续到达的过程中, 由于流携带了大量数据, 只有小部分的流数据被保存在有限的内存中. 流处理理论和技术已研究多年, 代表性的开源系统包括 Storm, S4 [24] 和 Kafka [25]. 流处理方式用于在线应用, 通常工作在秒或毫秒级别.

6) http://www.teradata.com/.
7) http://www-01.ibm.com/software/data/netezza/.
8) http://www.asterdata.com/.
9) http://www.greenplum.com/.
10) http://www.vertica.com/.

表 2 批处理和流处理比较

Table 2 Comparison between batch processing and stream processing

	Stream processing	Batch processing
Input	Stream of new data or updates	Data chunks
Data size	Infinite or unknown in advance	Known and finite
Storage	Not store or store non-trial portion in memory	Store
Hardware	Typical single limited amount of memory	Multiple CPUs and memory
Processing	A single or few pass(es) over data	Multiple rounds
Time	A few seconds or even milliseconds	Much longer
Applications	Web mining, sensor networks, traffic monitoring	Widely adopted in almost every domain

• 批处理: 在批处理方式中, 数据首先被存储, 随后被分析. MapReduce 是非常重要的批处理模型. MapReduce 的核心思想是, 数据首先被分为若干小数据块 chunks, 随后这些数据块被并行处理并以分布的方式产生中间结果, 最后这些中间结果被合并产生最终结果. MapReduce 分配与数据存储位置距离较近的计算资源, 以避免数据传输的通信开销. 由于简单高效, MapReduce 被广泛应用于生物信息、web 挖掘和机器学习中.

两种处理方式的区别如表 2 所示. 通常情况下, 流处理适用于数据以流的方式产生且数据需要得到快速处理获得大致结果. 因此流处理的应用相对较少, 大部分应用都采用批处理方式. 一些研究也试图集成两种处理方式的优点.

大数据平台可以选择不同的处理方式, 但是两种处理方式的不同将给相关的平台带来体系结构上的不同. 例如, 基于批处理的平台通常能够实现复杂的数据存储和管理, 而基于流处理的平台则不能. 在实际应用中, 可以根据数据特性和应用需求订制大数据平台. 本文将主要针对基于批处理的大数据平台进行探讨.

4 大数据系统架构

本节主要介绍大数据价值链, 大数据价值链由 4 个阶段构成: 数据生成、数据获取、数据存储和数据分析.

4.1 大数据系统: 价值链观点

大数据系统是一个复杂的、提供数据生命周期 (从数据的产生到消亡) 的不同阶段数据处理功能的系统. 同时, 对于不同的应用, 大数据系统通常也涉及多个不同的阶段 [26, 27]. 本文采用产业界广为接受的系统工程方法, 将典型的大数据系统分解为 4 个连续的阶段, 包括数据生成、数据获取、数据存储和数据分析, 如图 2 中水平轴所示.

数据生成阶段关心的是数据如何产生. 此时 “大数据” 意味着从多样的纵向或分布式数据源 (传感器、视频、点击流和其他数字源) 产生的大量的、多样的和复杂的数据集. 通常, 这些数据集和领域相关的不同级别的价值联系在一起 [2]. 本文将集中在商业、互联网和科学研究这三个重要的领域, 因为这些领域的数据价值相对容易理解. 但是, 在收集、处理和分析这些数据集时存在巨大的技术挑战, 需要利用信息通信技术 (ICT) 领域的最新研究技术提出新的解决方案.

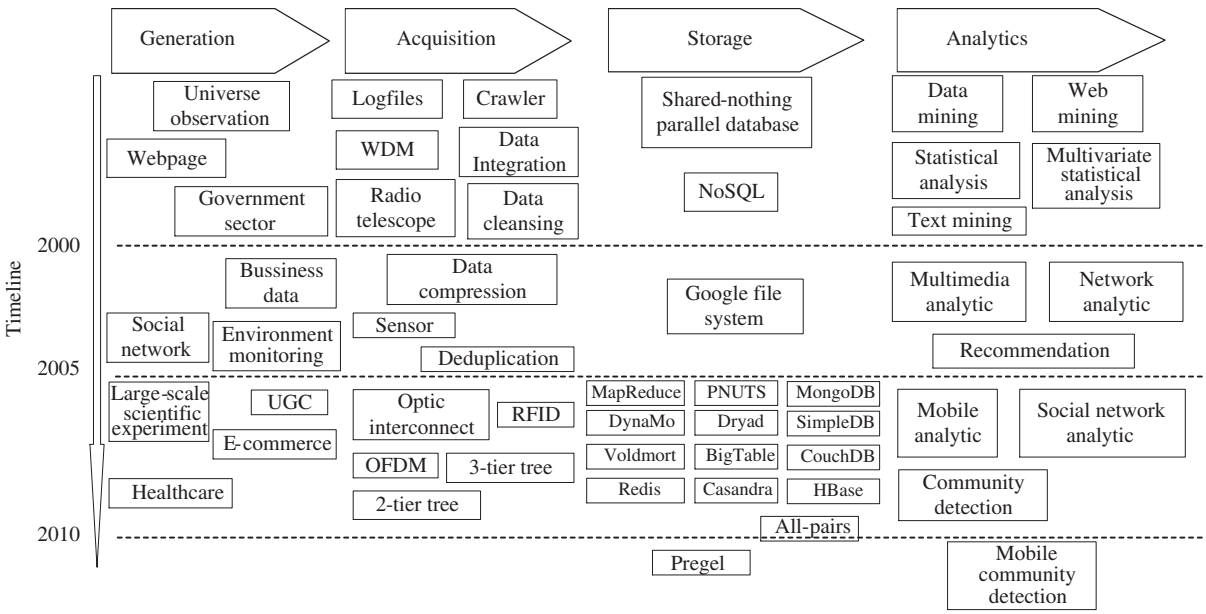

图 2　大数据价值链及其技术地图

Figure 2　Big data value chain and technology map

数据获取则是指获取信息的过程, 可分为数据采集、数据传输和数据预处理. 首先, 由于数据来自不同的数据源, 如包含格式文本、图像和视频的网站数据, 数据采集是指从特定数据生产环境获得原始数据的专用数据采集技术. 其次, 数据采集完成后, 需要高速的数据传输机制将数据传输到合适的存储系统, 供不同类型的分析应用使用. 再次, 数据集可能存在一些无意义的数据, 将增加数据存储空间并影响后续的数据分析. 例如, 从监控环境的传感器中获得的数据集通常存在冗余, 可以使用数据压缩技术减少数据传输量. 因此, 必须对数据进行预处理, 以实现数据的高效存储和挖掘.

数据存储解决的是大规模数据的持久存储和管理. 数据存储系统可以分为两部分: 硬件基础设施和数据管理软件. 硬件基础设施由共享的 ICT 资源池组成, 资源池根据不同应用的即时需求, 以弹性的方式组织而成. 硬件基础设施应能够向上和向外扩展, 并能进行动态重配置以适应不同类型的应用环境. 数据管理软件则部署在硬件基础设施之上用于维护大规模数据集. 此外, 为了分析存储的数据及其数据交互, 存储系统应提供功能接口、快速查询和其他编程模型.

数据分析利用分析方法或工具对数据进行检查、变换和建模并从中提取价值. 许多应用领域利用领域相关的数据分析方法获得预期的结果. 尽管不同的领域具有不同的需求和数据特性, 它们可以使用一些相似的底层技术. 当前的数据分析技术的研究可以分为 6 个重要方向: 结构化数据分析、文本数据分析、多媒体数据分析、web 数据分析、网络数据分析和移动数据分析.

大数据的研究涉及许多学科技术, 图 2 显示了大数据技术地图, 图中将大数据价值链不同阶段和相应的开源或专有技术联系在一起. 图 2 反映了大数据的发展趋势. 在数据生成阶段, 大数据的结构逐渐复杂, 从结构化或无结构的数据到不同类型的混合数据. 在数据获取阶段, 数据采集、数据预处理和数据传输的研究则出现在不同的时期. 而数据存储的相关研究则大部分始于 2005 年. 数据分析的基本方法形成于 2000 年前, 随后的研究则使用这些方法解决领域相关的问题. 从该图中, 可以在不同阶段选择合适的技术和方法定制大数据系统.

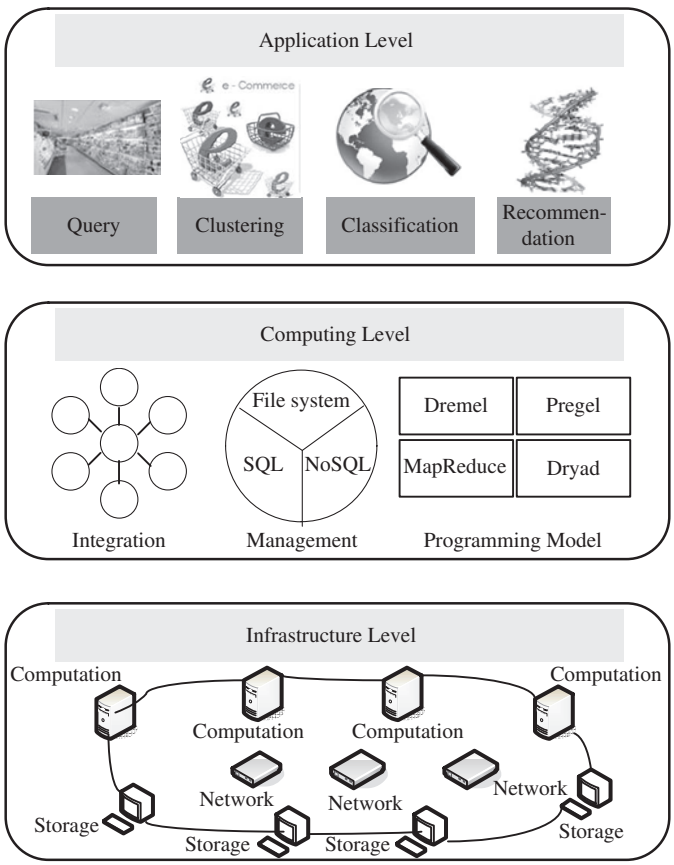

图 3 大数据系统的层次架构

Figure 3 Layered architecture of big data system

4.2 大数据系统: 层次观点

另一方面, 从层次观点, 可以将大数据系统分解为 3 层: 基础设施层、计算层和应用层, 如图 3 所示. 这种层次观点仅提供概念上的层次以强调大数据系统的复杂性.

• 基础设施层: 由 ICT 资源池构成, 可利用虚拟技术组织为云计算基础设施. 这些资源通过特定的服务级别协定 (service-level agreement, SLA) 以细粒度的方式提供给上层子系统, 资源的分配需要满足大数据需求, 同时通过最大化系统利用率、能量感知和操作简化等方式实现资源使用的有效性.

• 计算层: 将多种数据工具封装于运行在原始 ICT 硬件资源之上的中间件中, 典型的工具包括数据集成、数据管理和编程模型等. 数据集成是指从独立的数据源中获取数据, 并通过必要的预处理技术将数据集合成为统一形式. 数据管理是指提供数据的持久存储和高效管理的机制和工具, 例如分布式的文件系统和 SQL, NoSQL 数据存储. 编程模型实现应用逻辑抽象并为数据分析应用提供便利. MapReduce [10], Dryad [28], Pregel [29] 和 Dremel [30] 是几个典型的编程模型.

• 应用层: 利用编程模型提供的接口实现不同的数据分析功能, 包括查询、统计分析、数据的聚类和分类等, 同时通过组合基本分析方法开发不同的领域相关应用. McKinsey 公司提出了 5 个潜在的大数据应用领域: 医疗康护、公众部门管理、零售、全球制造和个人位置信息.

4.3 大数据系统面临的挑战

设计和实现一个大数据系统不是一个简单的任务, 如同大数据定义描述的, 大数据超出了现有硬件和软件平台的处理能力. 新的硬件和软件平台反过来要求新的基础设施和编程模型解决大数据带来的挑战. 最近的研究工作 [26,31,32] 讨论了影响大数据应用的潜在障碍. 本文将大数据面临的挑战分为 3 类: 数据采集和管理、数据分析和系统观点. 近年来, 本文作者参加了一些学术界和工业界的相关座谈和讨论, 所以文中可能也包含了领域内一些同行们的见解.

数据采集和管理处理异构复杂的海量数据, 其面临的部分挑战包括:

• 数据表示. 许多数据集在类型、结构、语义、组织、粒度和可访问性等方面是异构的. 合适的数据表示方法能够反映数据的结构、层次和多样性, 并且需要设计一个集成技术实现跨数据集的有效操作.

• 冗余缩减 (Redundancy reduction) 和数据压缩. 通常在原始数据集中存在大量的冗余数据. 不损毁数据价值的冗余缩减和数据压缩是减少系统整体开销的有效方法.

• 数据生存周期管理: 普适的感知和计算以难以想象的速率和规模产生数据, 远超现有存储技术的发展. 一个迫切的挑战是现有的存储系统难以容纳海量数据. 而数据的潜在价值和数据新鲜度有关, 因此应该设置和隐藏价值相联系的数据重要性原则, 以决定哪部分数据需要存档, 哪部分数据可以丢弃.

• 数据隐私和安全. 随着在线服务和移动手机的增长, 与访问控制、个人信息分析相关的隐私和安全问题日益得到关注. 了解需要提供什么样的系统级别隐私保护机制至关重要.

大数据分析技术的发展为数据解释、建模、预测和模拟带来了重大的影响. 然而, 海量数据、异构数据结构和多样化的应用也带来了许多挑战.

• 近似分析: 随着数据集的增长和实时处理需求的提出, 对整个数据集的分析越来越难. 一个潜在的解决方案是给出近似结果, 例如使用近似查询. 近似的含义有两个方面: 结果的准确度和从输出中删除的数据组.

• 连接社交媒体: 社交媒体具有独特的性质, 如巨量性、统计冗余性和用户反馈的可用性. 不同的提取技术已成功用于标识从社交媒体到具体产品名称和位置等参照物. 通过连接领域间的数据和社交媒体, 应用能够获得更高的精确性.

• 深度分析: 大数据的一个令人兴奋的研究动机是期望获得新的领悟. 诸如机器学习等复杂的分析技术对发现新的知识非常必要, 而有效地使用这些分析工具包需要了解概率和统计. 安全和隐私机制的核心是强制的访问控制和安全通信, 多粒度访问控制, 隐私感知的数据挖掘和分析, 以及安全存储和管理.

最后, 大规模并行处理系统通常面临几个共同的问题, 而大数据的出现则放大了这些问题.

• 能量管理: 大规模计算系统的能量消耗从经济和环境的观点吸引了较大的关注. 随着数据量和分析需求的增长, 数据传输、存储和处理无疑将消耗更多的能量. 因此, 在大数据系统中必须提供系统级的能量控制和管理机制, 同时提供可扩展性和可访问性.

• 可扩展性: 大数据系统应该能够支持现在以及将来产生的巨大的数据集. 大数据系统中的所有组件都能扩展以解决复杂数据集的日益增长.

• 协作性: 大数据分析是一个交叉学科研究领域, 需要来自不同专业领域的专家协作挖掘数据中隐藏的价值. 因此需要建立一个综合的大数据基础设施, 允许不同领域的科学家和工程师访问多样的数据, 并应用各自的专业知识, 协作完成分析任务.

5 阶段 I: 数据生成

本节将介绍大数据源的两个方面: 大数据源的历史趋势和三种典型的数据源.

5.1 数据源

大数据生成的发展趋势可由数据产生速率来描述. 随着技术的发展, 数据产生速率也不断增长. 事实上, IBM 认为现在世界上 90% 的数据是近两年产生的[11)]. 数据爆炸的原因被广为争论. Cisco 认为数据的增长来自于视频、互联网和摄像头 [33]. 由于数据实际上是能被计算机可读的信息抽象, 信息通信技术 (ICT) 是使得信息可读并且产生或捕获数据的主要驱动力. 因此本节首先从 ICT 技术的发展开始, 以历史的观点解释数据爆炸的发展趋势.

数据生成的模式可分为 3 个顺序的阶段.

• 阶段 1 始于 20 世纪 90 年代. 随着数字技术和数据库系统的广泛使用, 许多企业组织的管理系统存储了大量的数据, 如银行交易事务、购物中心记录和政府部门归档等. 这些数据集是结构化的, 并能通过基于数据库的存储管理系统进行分析.

• 阶段 2 则始于 web 系统的日益流行. 以搜索引擎和电子商务为代表的 web 1.0 系统在 20 世纪 90 年代末期产生了大量的半结构化和无结构的数据, 包括网页数据和事务日志等. 而自 2000 年初期以来, 许多 web 2.0 应用从在线社交网络 (如论坛、博客、社交网站和社交媒体网站等) 中产生了大量的用户创造内容.

• 阶段 3 因移动设备 (如智能手机、平板电脑、传感器和基于传感器的互联网设备) 的普及而引发. 在不久的将来, 以移动为中心的网络将产生高度移动、位置感知、以个人为中心和上下文相关的数据.

可以发现, 数据生成模式是从阶段 1 的被动记录到阶段 2 的数据主动生成, 再到阶段 3 的自动生成.

除了用数据产生速率描述, 大数据源还与数据产生领域相关. 本文主要对商业、网络和科学研究这三个领域进行大数据相关技术的调研. 首先, 大数据和商业活动联系紧密, 许多大数据工具已经被开发并广泛使用; 其次, 大部分的数据是由互联网、移动网络和物联网产生的. 再次, 科学研究会产生大量的数据, 高效的数据分析将帮助科学家们发现基本原理, 促进科学发展. 这三个领域在对大数据的处理方面具有不同的技术需求.

(1) 商业数据

过去几十年中, 信息技术和数字数据的使用对商业领域的繁荣发展起到了重要的推动作用. 全球所有公司商业数据量每 1.2 年会翻番. 互联网上的商业事务, 包括 B2B 和 B2C 事务, 每天有 4500 亿条 [34]. 日益增长的商业数据需要使用高效的实时分析工具挖掘其价值. 例如, Amazon 每天要处理几百万的后端操作和来自第三方销售超过 50 万的查询请求. 沃尔玛每小时要处理上百万的客户事务, 这些事务被导入数据库, 约有超过 2.5 PB 的数据量 [3]. Akamai 每天则需分析 7500 万事件, 以更好地实现广告定位[12)].

(2) 网络数据

网络 (互联网、移动网络和物联网) 已经和人们的生活紧密联系在一起. 网络应用如搜索、社交网络服务 SNS、网站和点击流是典型的大数据源. 这些数据源高速产生数据, 需要先进的处理技术. 例如, 搜索引擎 Google 在 2008 年每天要处理 20 PB 的数据 [10]; 社交网络应用 Facebook 则每天需存

11) IBM. What is big data. http://www-01.ibm.com/software/data/bigdata/.
12) Kelly J. Taming Big Data. http://wikibon.org/blog/taming-big-data/.

表 3 典型大数据源

Table 3 Typical big data sources

Data source	Application	Data scale	Type	Response time	Number of users	Accuracy
Walmart	Retail	PB	Structured	Very fast	Large	Very high
Amazon	e-commerce	PB	Semi-structured	Very fast	Large	Very high
Google search	Internet	PB	Semi-structured	Fast	Very large	High
Facebook	Social network	PB	Structured, unstructured	Fast	Very large	High
AT&T	Mobile network	TB	Structured	Fast	Very large	High
Health care	Internet of Things	TB	Structured, unstructured	Fast	Large	High
SDSS	Scientific research	TB	Unstructured	Slow	Small	Very high

储、访问和分析超过 30 PB 的用户创造数据; Twitter 每月会处理超过 3200 亿的搜索[13)]. 在移动网络领域, 2010 年有 40 亿人持有手机, 其中约 12% 的手机是智能手机. 而在物联网领域, 有超过 3000 万的联网传感器工作在运输、汽车、工业、公用事业和零售部门并产生数据. 这些传感器每年仍将以超过 30% 的速率增长.

(3) 科学研究数据

越来越多的科学应用正产生海量的数据集, 若干学科的发展极度依赖于对这些海量数据的分析, 这些学科主要包括:

光学观测和监控. 在光学遥感和对地观测领域、基于光学等设备的视频监控领域等, 往往需要获取连续大量的数据. 这些几乎造成管理和处理灾难的数据有一定的周期性, 而用户关心的又往往是其中的差异和异常的部分. 考虑到这类数据的分析和学习过程往往又同获取这些数据时的装置和参数密切相关, 再加上视觉信息对人类的重要性以及用户同系统的必要交互, 对光学观测和监控数据的管理和处理已经提高到重要日程.

计算生物学. 美国国家生物信息中心 NCBI 维护了 GenBank 的核苷酸序列数据库, 该数据库大小每 10 个月翻倍. 2009 年 8 月, 数据库中存储了来自 15 万多有机生物体的超过 2500 亿条核苷酸碱基 [35].

天文学. 从 1998 年到 2008 年, 最大的天文目录 SDSS 从天文望远镜中获取了 25 Terabytes 数据. 随着天文望远镜分辨率的提高, 每晚产生的数据量将在 2014 年超过 20 Terabytes[14)].

高能物理. 欧洲粒子物理实验室中大型强子对撞机实验, 在 2008 年初起以 2 PB/s 的速率产生数据, 每年将存储约 10 PB 经过处理的数据[15)].

这些领域不但要产生海量的数据, 还需要分布在世界各地的科学家们协作分析数据 [36, 37]. 表 3 列举了这三个领域中具有代表性的大数据源及其应用属性和数据分析的需求. 可以看出, 大部分的数据源产生 PB 级别的无结构数据, 并且需要得到快速准确的分析.

5.2 数据属性

普适感知和计算产生前所未有的复杂的异构数据, 这些数据集在规模、时间维度、数据类型的多样性等方面有着不同的特性. 例如, 移动数据和位置、运动、距离、通信、多媒体和声音环境等相关 [38]. NIST 提出了大数据的 5 种属性 [15].

13) Wikibon. A Comprehensive List of Big Data Statistics. http://wikibon.org/blog/big-data-statistics/.

14) http://www.sdss.org/.

15) http://atlasexperiment.org/.

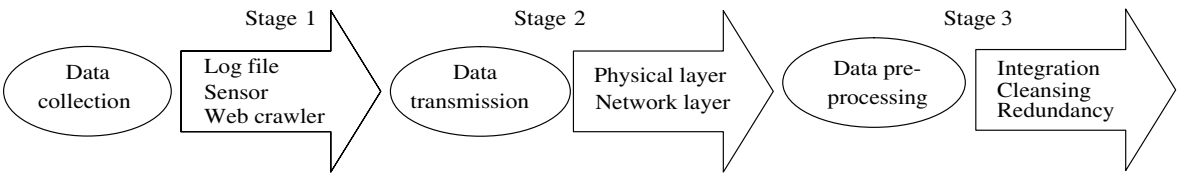

图 4 数据获取的 3 个步骤

Figure 4 Three steps of data acquisition

- 容量: 数据集的大小.
- 速度: 数据生成速率和实时需求.
- 多样性: 结构化、半结构化和无结构的数据形式.
- 水平扩展性: 合并多数据集的能力.
- 相关限制: 包含特定的数据形式和查询. 数据的特定形式包括时间数据和空间数据; 查询则可以是递归或其他方式.

通常, 科学研究领域的数据源在 5 种属性中具有最小的属性值; 商业领域的数据源则具有较高的水平扩展性和相关限制的需求; 而网络领域的数据源具有较高的容量、速度和多样性特征.

6 阶段 II: 数据获取

在大数据价值链中, 数据获取阶段的任务是以数字形式将信息聚合, 以待存储和分析处理. 数据获取过程可分为三个步骤: 数据采集、数据传输和数据预处理, 如图 4 所示. 数据传输和数据预处理没有严格的次序, 预处理可以在数据传输之前或之后.

6.1 数据采集

数据采集是指从真实世界对象中获得原始数据的过程. 不准确的数据采集将影响后续的数据处理并最终得到无效的结果. 数据采集方法的选择不但要依赖于数据源的物理性质, 还要考虑数据分析的目标. 随后将介绍 3 种常用的数据采集方法: 传感器、日志文件和 web 爬虫.

(1) 传感器

传感器常用于测量物理环境变量并将其转化为可读的数字信号以待处理. 传感器包括声音、振动、化学、电流、天气、压力、温度和距离等类型. 通过有线或无线网络, 信息被传送到数据采集点.

有线传感器网络通过网线收集传感器的信息, 这种方式适用于传感器易于部署和管理的场景. 例如视频监控系统通常使用非屏蔽双绞线连接摄像头, 摄像头部署在公众场合监控人们的行为, 如偷盗和其他犯罪行为 [39]. 而这仅仅是光学监控领域一个很小的应用示例, 在更广义的光学信息获取和处理系统中 (例如对地观测、深空探测等), 情况往往更复杂.

另一方面, 无线传感器网络利用无线网络作为信息传输的载体, 适合于没有能量或通信的基础设施的场合. 近年来, 无线传感器网络得到了广泛的研究, 并应用在多种场合, 如环境 [40,41]、水质监控 [42]、土木工程 [43,44]、野生动物监控 [45] 等. WSNs 通常由大量微小传感器节点构成, 微小传感器由电池供电, 被部署在应用制定的地点收集感知数据. 当节点部署完成后, 基站将发布网络配置/管理或收集命令, 来自不同节点的感知数据将被汇集并转发到基站以待处理 [46].

表 4 三种数据采集方法的比较

Table 4 Comparison among three data collection methods

Method	Mode	Data structure	Data scale	Complexity	Applications
Sensor	Pull	Structured or unstructured	Median	Sophisticated	Video surveillance, Inventory management
Log file	Push	Structured or semi-structured	Small	Easy	Web log, click stream
Web crawler	Pull	Mixture	Large	Median	Search, social networks analysis

基于传感器的数据采集系统被认为是一个信息物理系统 (cyber-physical system) [47]. 实际上, 在科学实验中许多用于收集实验数据的专用仪器 (如磁分光计、射电望远镜等)16), 可以看作特殊的传感器. 从这个角度, 实验数据采集系统同样是一个信息物理系统.

(2) 日志文件

日志是广泛使用的数据采集方法之一, 由数据源系统产生, 以特殊的文件格式记录系统的活动. 几乎所有在数字设备上运行的应用使用日志文件非常有用, 例如 web 服务器通常要在访问日志文件中记录网站用户的点击、键盘输入、访问行为以及其他属性 [48]. 有三种类型的 web 服务器日志文件格式用于捕获用户在网站上的活动: 通用日志文件格式 (NCSA)、扩展日志文件格式 (W3C) 和 IIS 日志文件格式 (Microsoft). 所有日志文件格式都是 ASCII 文本格式. 数据库也可以用来替代文本文件存储日志信息, 以提高海量日志仓库的查询效率 [49, 50]. 其他基于日志文件的数据采集包括金融应用的股票记帐和网络监控的性能测量及流量管理.

和物理传感器相比, 日志文件可以看作是 "软件传感器", 许多用户实现的数据采集软件属于这类 [38].

(3) Web 爬虫

爬虫 [51] 是指为搜索引擎下载并存储网页的程序. 爬虫顺序地访问初始队列中的一组 URLs, 并为所有 URLs 分配一个优先级. 爬虫从队列中获得具有一定优先级的 URL, 下载该网页, 随后解析网页中包含的所有 URLs 并添加这些新的 URLs 到队列中. 这个过程一直重复, 直到爬虫程序停止为止. Web 爬虫是网站应用如搜索引擎和 web 缓存的主要数据采集方式. 数据采集过程由选择策略、重访策略、礼貌策略以及并行策略决定 [52]. 选择策略决定哪个网页将被访问; 重访策略决定何时检查网页是否更新; 礼貌策略防止过度访问网站; 并行策略则用于协调分布的爬虫程序. 传统的 web 爬虫应用已较为成熟, 提出了不少有效的方案. 随着更丰富更先进的 web 应用的出现, 一些新的爬虫机制已被用于爬取富互联网应用的数据 [53].

除了上述方法, 还有许多和领域相关的数据采集方法和系统. 例如, 政府部门收集并存储指纹和签名等人体生物信息, 用于身份认证或追踪罪犯 [54]. 根据数据采集方式的不同, 数据采集方法又可以大致分为以下两类:

- 基于拉 (pull-based) 的方法, 数据由集中式或分布式的代理主动收集.
- 基于推 (push-based) 的方法, 数据由源或第三方推向数据汇聚点.

表 4 对上述三种数据采集方法进行了比较, 日志文件是最简单的数据采集方法, 但是只能收集相对一小部分结构化数据; web 爬虫是最灵活的数据采集方法, 可以获得巨量的结构复杂的数据.

6.2 数据传输

原始数据采集后必须将其传送到数据存储基础设施如数据中心等待进一步处理. 数据传输过程可

16) Wikipedia. Scientific Instrument. http://en.wikipedia.org/wiki/Scientific_instrument/.

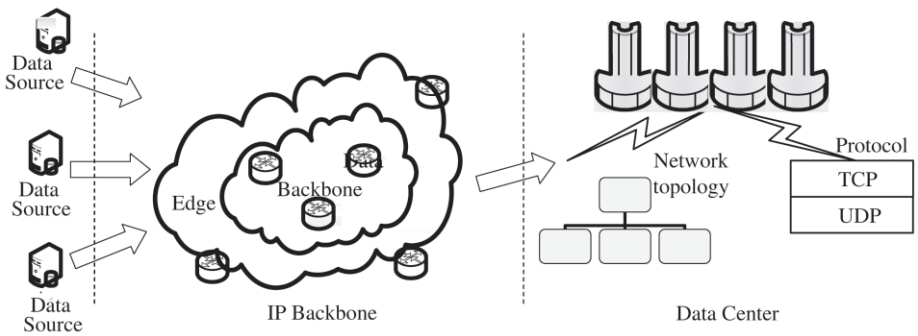

图 5 大数据传输过程

Figure 5 Big data transmission procedure

以分为两个阶段, IP 骨干网传输和数据中心传输, 如图 5 所示.

(1) IP 骨干网传输

IP 骨干网提供高容量主干线路将大数据从数据源传递到数据中心. 传输速率和容量取决于物理媒体和链路管理方法.

• 物理媒体: 通常由许多光缆合并在一起增加容量, 并需要存在多条路径已确保路径失效时能进行重路由.

• 链路管理: 决定信号如何在物理媒体上传输. 过去 20 年间 IP over WDM 技术得到了深入地研究 [55, 56]. 波分复用技术 (WDM) 是在单根光纤上复用多个不同波长的光载波信号. 为了解决电信号带宽的瓶颈问题, 正交频分复用 OFDM 被认为是未来的高速光传输技术的候选者. OFDM 允许单个子载波的频谱重叠, 能够构建具有更灵活的数据率、资源有效使用的光网络 [57, 58].

目前 IP 骨干网部署了每信道 40 Gbps 带宽的光传输系统, 100 Gbps 的接口也已经商用. 不久的将来 Tbps 级别的传输也将实现 [59].

由于在现有的互联网骨干网中增强网络协议功能较为困难, 必须遵循现有的互联网协议传输大数据. 然而, 对于区域或私有 IP 骨干网, 对于特定的应用, 一些专用的传输方法能够获得更好的性能 [60].

(2) 数据中心传输

数据传递到数据中心后, 将在数据中心内部进行存储位置的调整和其他处理, 这个过程称为数据中心传输, 涉及到数据中心体系架构和传输协议.

• 数据中心体系架构: 数据中心由多个装备了若干服务器的机架构成, 服务器通过数据中心内部网络连接. 许多数据中心基于权威的 2 层 [61] 或 3 层 [62] fat-tree 结构的商用交换机构建. 一些其他的拓扑也用于构建更加高效的数据中心网络 [63~66]. 由于电子交换机的固有缺陷, 在增加通信带宽的同时减少能量消耗非常困难. 数据中心网络中的光互联技术能够提供高吞吐量、低延迟和减少能量消耗, 被认为是有前途的解决方案. 目前, 光技术在数据中心仅用于点对点链路, 这些链路基于低成本的多模光纤并连接交换机, 带宽只能达到 10 Gbps [67]. 数据中心的光互联 (交换机以光的形式交换) [68] 能够提供 Tbps 级别的带宽, 并能提高能量效率. 许多光互联机制 [65, 69~73] 已被应用于数据中心网络. 一些方案建议增加光电路升级现有数据中心网络, 而另一些方案则认为需完全替换现有交换机.

• 传输协议: TCP 和 UDP 是数据传输最重要的两种协议, 但是它们的性能在传输大量的数据时并不令人满意. 许多研究致力于提高这两种协议的性能. 一些增强 TCP 功能的方法目标是提高链路吞

吐率并对长短不一的混合 TCP 流提供可预测的小延迟. 例如, DCTCP [74] 利用显示拥塞通知对端主机提供多比特反馈; Vamanan 等 [75] 提出了用于数据中心网络的 deadline 感知的 TCP 协议, 用于分配带宽, 确保在软实时限制下完成网络传输. UDP 协议适用于传输大量数据, 但是缺乏拥塞控制. 因此高带宽的 UDP 应用必须自己实现拥塞控制机制, 这是一个困难的任务并会导致风险. Kholer 等 [76] 在类 UDP 的基础协议上设计添加了一个可拥塞控制的不可靠传输协议, 该协议类似于 TCP 但是没有可靠传输和累积确认机制.

6.3 数据预处理

由于数据源的多样性, 数据集由于干扰、冗余和一致性因素的影响具有不同的质量. 从需求的角度, 一些数据分析工具和应用对数据质量有着严格的要求. 因此在大数据系统中需要数据预处理技术提高数据的质量. 本节讨论三种主要的数据预处理技术 [77~79].

(1) 数据集成 (Data integration)

数据集成技术在逻辑上和物理上把来自不同数据源的数据进行集中, 为用户提供一个统一的视图 [80]. 数据集成在传统的数据库研究中是一个成熟的研究领域 [81], 如数据仓库 (data warehouse) 和数据联合 (data federation) 方法. 数据仓库又称为 ETL, 由 3 个步骤构成: 提取、变换和装载.

- 提取: 连接源系统并选择和收集必要的数据用于随后的分析处理.
- 变换: 通过一系列的规则将提取的数据转换为标准格式.
- 装载: 将提取并变换后的数据导入目标存储基础设施.

数据联合则创建一个虚拟的数据库, 从分离的数据源查询并合并数据. 虚拟数据库并不包含数据本身, 而是存储了真实数据及其存储位置的信息或元数据.

然而, 这两种方法并不能满足流式和搜索应用对高性能的需求, 因此这些应用的数据高度动态, 并且需要实时处理. 一般地, 数据集成技术最好能与流处理引擎 [23] 或搜索引擎集成在一起 [82].

(2) 数据清洗 (Data cleansing)

数据清洗是指在数据集中发现不准确、不完整或不合理数据, 并对这些数据进行修补或移除以提高数据质量的过程. ~~一个通用的数据清洗框架由 5 个步骤构成: 定义错误类型, 搜索并标识错误实例, 改正错误, 文档记录错误实例和错误类型, 修改数据录入程序以减少未来的错误.~~

此外, 格式检查、完整性检查、合理性检查和极限检查也在数据清洗过程中完成. 数据清洗对保持数据的一致和更新起着重要的作用, 因此被用于如银行、保险、零售、电信和交通的多个行业.

在电子商务领域, 尽管大多数数据通过电子方式收集, 但仍存在数据质量问题. 影响数据质量的因素包括软件错误、定制错误和系统配置错误等. Kohavi 等 [83] 讨论了通过检测爬虫和定期执行客户和帐户的重复数据删除 (de-duping), 对电子商务数据进行清洗.

在 RFID 领域, 文献 [84] 研究了对 RFID 数据的清洗. RFID 技术用于许多应用, 如库存检查和目标跟踪等. 然而原始的 RFID 数据质量较低并包含许多由于物理设备的限制和不同类型环境噪声导致的异常信息. Zhao 等在 [85] 中提出了一个概率模型解决移动环境中的数据丢失问题. Khoussainova 等在 [86] 中设计了一个能根据应用定义的全局完整性约束自动修正输入数据错误的系统.

文献 [87] 则实现了一个框架 BIO-AJAX, 用于对生物数据进行标准化. 在该框架的辅助下, 生物数据中的错误和副本可以消除, 数据挖掘技术能够更高效地运行.

数据清洗对随后的数据分析非常重要, 因为它能提高数据分析的准确性. 但是数据清洗依赖复杂的关系模型, 会带来额外的计算和延迟开销, 必须在数据清洗模型的复杂性和分析结果的准确性之间进行平衡.

(3) 冗余消除 (Redundancy elimination)

数据冗余是指数据的重复或过剩, 这是许多数据集的常见问题. 数据冗余无疑会增加传输开销, 浪费存储空间, 导致数据不一致, 降低可靠性. 因此许多研究提出了数据冗余减少机制, 例如冗余检测 [88] 和数据压缩 [89]. 这些方法能够用于不同的数据集和应用环境, 提升性能, 但同时也带来一定风险. 例如, 数据压缩方法在进行数据压缩和解压缩时带来了额外的计算负担, 因此需要在冗余减少带来的好处和增加的负担之间进行折中.

由广泛部署的摄像头收集的图像和视频数据存在大量的数据冗余. 在视频监控数据中, 大量的图像和视频数据存在着时间、空间和统计上的冗余 [90]. 视频压缩技术被用于减少视频数据的冗余, 许多重要的标准 (如 MPEG-2, MPEG-4, H.263, H.264/AVC) 已被应用以减少存储和传输的负担 [91]. Tsai 等在 [92] 中研究了通过视频传感器网络进行智能视频监控的视频压缩技术. 通过发现场景中背景和前景目标相联系的情境冗余, 他们提出了一种新的冗余减少方法.

对于普遍的数据传输和存储, 数据去重 (data deduplication) 技术 [93] 是专用的数据压缩技术, 用于消除重复数据的副本. 在存储去重过程中, 一个唯一的数据块或数据段将分配一个标识并存储, 该标识会加入一个标识列表. 当去重过程继续时, 一个标识已存在于标识列表中的新数据块将被认为是冗余的块. 该数据块将被一个指向已存储数据块指针的引用替代. 通过这种方式, 任何给定的数据块只有一个实例存在. 去重技术能够显著地减少存储空间, 对大数据存储系统具有非常重要的作用.

除了前面提到的数据预处理方法, 还有一些对特定数据对象进行预处理的技术, 如特征提取技术, 在多媒体搜索 [94] 和 DNS 分析 [95,96] 中起着重要的作用. 这些数据对象通常具有高维特征矢量. 数据变形技术 [97] 则通常用于处理分布式数据源产生的异构数据, 对处理商业数据非常有用. Gunter 在文献 [98] 中提出了 MapLan, 对瑞士国家银行的调查信息进行影射和变形. Wang 等在 [99] 中提出了一种在分布式存储系统中异构感知的数据重生成机制, 在异构链路上传递最少的数据以保持数据的完整性.

然而, 没有一个统一的数据预处理过程和单一的技术能够用于多样化的数据集, 必须考虑数据集的特性、需要解决的问题、性能需求和其他因素选择合适的数据预处理方案.

7 阶段 III: 数据存储

大数据系统中的数据存储子系统将收集的信息以适当的格式存放以待分析和价值提取. 为了实现这个目标, 数据存储子系统应该具有如下两个特征:

- 存储基础设施应能持久和可靠地容纳信息;
- 存储子系统应提供可伸缩的访问接口供用户查询和分析巨量数据.

从功能上, 数据存储子系统可以分为硬件基础设施和数据管理软件.

7.1 存储基础设施

硬件基础设施实现信息的物理存储, 可以从不同的角度理解存储基础设施. 首先, 存储设备可以根据存储技术分类. 典型的存储技术有如下几种.

- 随机存取存储器 (Random access memory, RAM): 是计算机数据的一种存储形式, 在断电时将丢失存储信息. 现代 RAM 包括静态 RAM、动态 RAM 和相变 RAM.

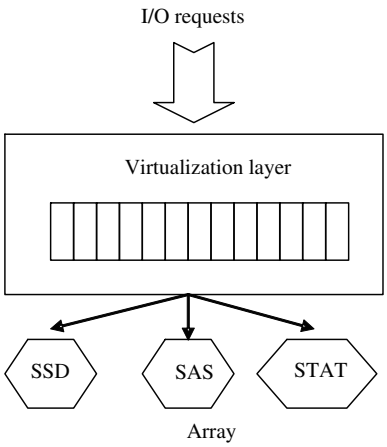

图 6　基于 SSD 的多层存储系统

Figure 6　SSD-based multi-layer storage system

• 磁盘和磁盘阵列: 磁盘 (如硬盘驱动器 HDD) 是现代存储系统的主要部件. HDD 由一个或多个快速旋转的碟片构成, 通过移动驱动臂上的磁头, 从碟片表面完成数据的读写. 与 RAM 不同, 断电后硬盘仍能保留数据信息, 并且具有更低的单位容量成本, 但是硬盘的读写速度比 RAM 读写要慢得多. 由于单个高容量磁盘的成本较高, 因此磁盘阵列将大量磁盘整合以获取高容量、高吞吐率和高可用性.

• 存储级存储器: 是指非机械式存储媒体, 如闪存. 闪存通常用于构建固态驱动器 SSD, SSD 没有类似 HDD 的机械部件, 运行安静, 并且具有更小的访问时间和延迟. 但是 SSD 的单位存储成本要高于 HDD.

这些存储设备具有不同的性能指标, 可以用来构建可扩展的、高性能的大数据存储子系统. 文献 [100] 更为详细地讨论了存储设备的发展. 文献 [101, 102] 则结合 HDD 和 SSD 的优点提出了一种混合层次存储系统, 该存储系统包括一个大容量的硬盘和一个 SSD 缓存, 经常访问的数据存放在 SSD 缓存中, 从而提高存取性能. 图 6 显示了一个典型的基于 SSD 的多层存储系统, 该系统由三个部件构成, 即 I/O 请求队列、虚拟化层和阵列 [103]. 目前, IBM, EMC, 3PAR 等公司的基于 SSD 的商用多层存储系统已能获得较好的性能.

其次, 可以从网络体系的观点理解存储基础设施 [104], 存储子系统可以通过不同的方式组织构建.

• 直接附加存储 (Direct attached storage, DAS): 存储设备通过主机总线直接连接到计算机, 设备和计算机之间没有存储网络. DAS 是对已有服务器存储的最简单的扩展.

• 网络附件存储 (Network attached storage, NAS): NAS 是文件级别的存储技术, 包含许多硬盘驱动器, 这些硬盘驱动器组织为逻辑的冗余的存储容器. 和 SAN 相比, NAS 可以同时提供存储和文件系统, 并能作为一个文件服务器.

• 存储区域网络 (Storage area network, SAN): SAN 通过专用的存储网络在一组计算机中提供文件块级别的数据存储. SAN 能够合并多个存储设备, 例如磁盘和磁盘阵列, 使得它们能够通过计算机直接访问, 就好象它们直接连接在计算机上一样.

这三种存储技术的存储网络体系架构如图 7 所示, SAN 具有最复杂的网络架构, 并依赖于特定的存储网络设备.

最后, 尽管已有的存储系统架构得到了深入的研究, 但是却无法直接应用于大数据系统中. 为了适应大数据系统的 “4Vs” 特性, 存储基础设施应该能够向上和向外扩展, 以动态配置适应不同的应用.

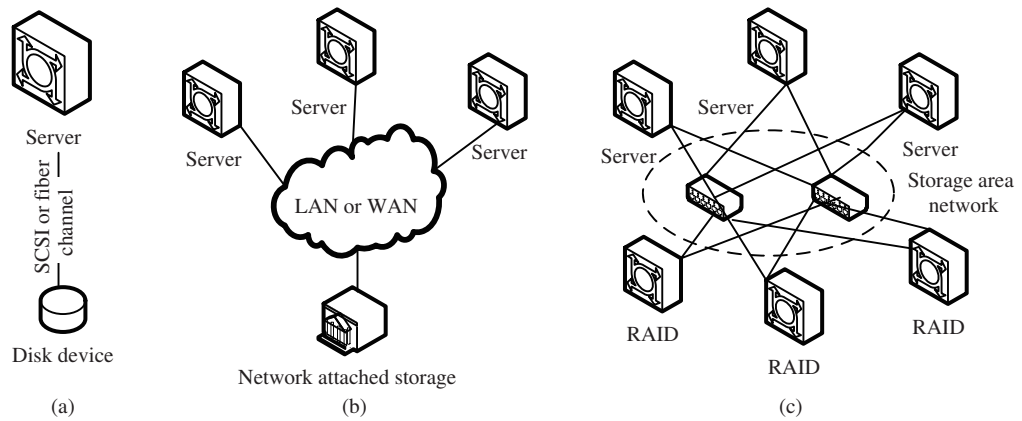

图 7 存储系统网络架构

Figure 7 Network architecture of storage system. (a) DAS; (b) NAS (file oriented); (c) SAN(block oriented)

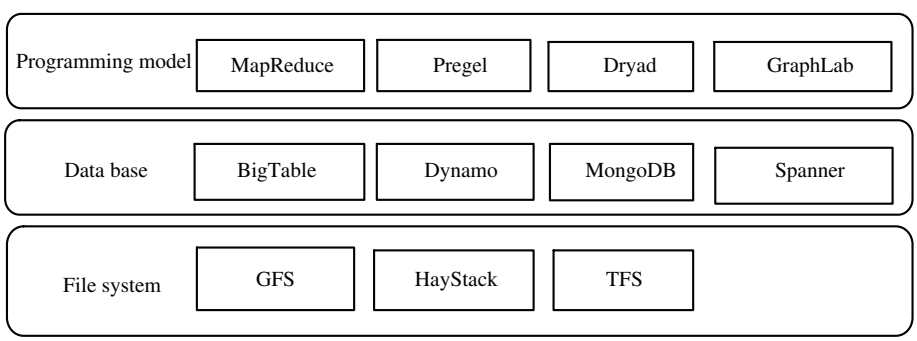

图 8 数据管理技术

Figure 8 Data management technology

一个解决这些需求的技术是云计算领域提出的存储虚拟化 [105]. 存储虚拟化将多个网络存储设备合并为单个存储设备. 目前可以在 SAN 和 NAS 架构上实现存储虚拟化 [106]. 基于 SAN 的存储虚拟化在可扩展性、可靠性和安全方面能够比基于 NAS 的存储虚拟化具有更高的性能. 但是 SAN 需要专用的存储基础设施, 从而带来较高的成本.

7.2 数据管理框架

数据管理框架解决的是如何以适当的方式组织信息以待有效地处理. 在大数据出现之前, 数据管理框架就得到了较为广泛的研究. 本文从层次的角度将数据管理框架划分为 3 层: 文件系统、数据库技术和编程模型, 如图 8 所示.

(1) 文件系统

文件系统是大数据系统的基础, 因此得到了产业界和学术界的广泛关注. Google 为大型分布式数据密集型应用设计和实现了一个可扩展的分布式文件系统 GFS [21]. GFS 运行在廉价的商用服务器上, 为大量用户提供容错和高性能服务. GFS 适用于大文件存储和读操作远多于写操作的应用. 但是 GFS 具有单点失效和处理小文件效率低下的缺点, Colossus [107] 改进了 GFS 并克服了这些缺点. 此

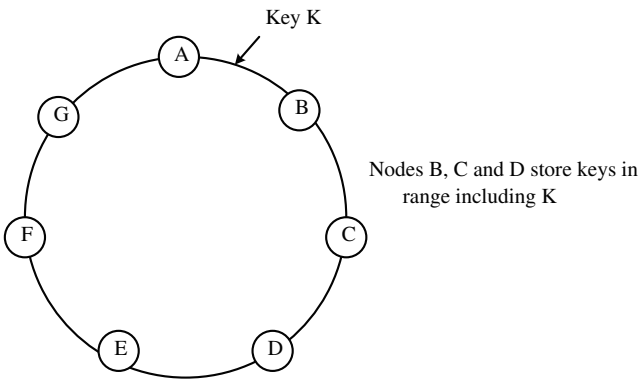

图 9 Dynamo 环中键的分割和复制

Figure 9 Partition and replication of keys in Dynamo ring

外, 其他的企业和研究者们开发了各自的文件存储解决方案以适应不同的大数据存储需求. HDFS[17)]和 Kosmosfs[18)]是 GFS 的开源产物; Microsoft 开发了 Cosmos 支持其搜索和广告业务 [108]; Facebook 实现了 Haystack 存储海量的小照片 [109]; 淘宝则设计了两种类似的小文件分布式文件系统: TFS[19)] 和 FastFS[20)]. 由于分布式文件系统已较为成熟, 因此本文将重点介绍数据管理框架的其余两层.

(2) 数据库技术

数据库技术已经经历了 30 多年的发展, 不同的数据库系统被设计用于不同规模的数据集和应用. 传统的关系数据库系统难以解决大数据带来的多样性和规模的需求. 由于具有模式自由、易于复制、提供简单 API、最终一致性和支持海量数据的特性, NoSQL 数据库逐渐成为处理大数据的标准. 随后将根据数据模型的不同, 讨论三种主流的 NoSQL 数据库: 键值 (key-value) 存储数据库、列式存储数据库和文档存储数据库.

(a) 键值存储数据库

键值存储是一种简单的数据存储模型, 数据以键值对的形式储存, 键是唯一的. 近年出现的键值存储数据库受到 Amazon 公司的 Dynamo 影响特别大 [110]. 在 Dynamo 中, 数据被分割存储在不同的服务器集群中, 并复制为多个副本. 可扩展性和持久性 (durability) 依赖于以下两个关键机制.

• 分割和复制: Dynamo 的分割机制基于一致性哈希技术 [111], 将负载分散在存储主机上. 哈希函数的输出范围被看作是一个固定的循环空间或 “环”. 系统中的每个节点将随机分配该空间中的一个值, 表示它在环中的位置. 通过哈希标识数据项的键, 可以获得该数据项在环中对应的节点. Dynamo 系统中每条数据项存储在协调节点和 N−1 个后继节点上, 其中 N 是实例化的配置参数. 如图 9 所示, 节点 B 是键 k 的协调节点, 数据存储在节点 B 同时复制在节点 C 和 D 上. 此外, 节点 D 将存储在 (A, B], (B, C] 和 (C, D] 范围内的键.

• 对象版本管理: 由于每条唯一的数据项存在多个副本, Dynamo 允许以异步的方式更新副本并提供最终一致性. 每次更新被认为是数据的一个新的不可改变的版本. 一个对象的多个版本可以在系统中共存.

17) Hadoop Distributed File System. http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html.
18) https://code.google.com/p/kosmosfs/.
19) Taobao File System. http://code.taobao.org/p/tfs/src/.
20) Fast Distributed File System. https://code.google.com/p/fastdfs/.

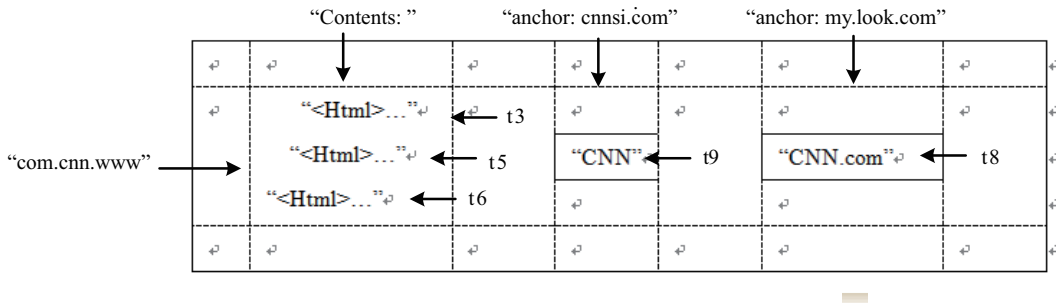

图 10 BigTable 数据模型

Figure 10 BigTable data model

其他键值存储包括 Voldemort[21)], Redis[22)], Tokyo Cabinet[23)], Tokyo Tyrant[24)], Memcached[25)], MemchacheDB[26)], Riak[27)]以及 Scalaris[28)]. Voldemort, Riak, Tokyo Cabinet 和 Memcached 可以将数据存储在 RAM 或带附件的磁盘中; 其他的则存储在 RAM 并使用磁盘作为备份, 或者无需备份而使用复制和恢复机制.

(b) 列式存储数据库

列式存储数据库以列存储架构进行存储和处理数据, 主要适合于批量数据处理和实时查询. 下面介绍典型的列式存储系统.

• Bigtable [112]: 是 Google 公司设计的一种列式存储系统. Bigtable 基本的数据结构是一个稀疏的、分布式的、持久化存储的多维度排序映射 (map), 映射由行键、列键和时间戳构成. 行按字典序排序并且被划分为片 (tablet), 片是负载均衡单元. 列根据键的前缀成组, 称为列族 (column family), 是访问控制的基本单元. 时间戳则是版本区分的依据. 图 10 给出了一个在单个表中存储大量的网页的示例, 其中 URL 作为行键, 网页的不同部分作为列名. 网页的多个版本内容存储在单个列中. Bigtable 的实现包括三个组件: 主服务器、tablet 服务器和客户端库. Master 负责将 tablet 分配到 tablet 服务器, 检测 tablet 服务器的添加和过期, 平衡 tablet 服务器负载, GFS 文件的垃圾回收. 另外, 它还会处理 schema 的变化, 比如表和列族的创建. 每个 tablet 服务器管理一系列的片, 处理对 tablet 的读取以及将大的 tablet 进行分割. 客户端库则提供应用与 Bigtable 实例交互. Bigtable 依赖 Google 基础设施的许多技术, 如 GFS、集群管理系统、SSTable 文件格式和 Chubby [113].

• Cassandra [114]: 由 Facebook 开发并于 2008 年开源, 结合了 Dynamo 的分布式系统技术和 Bigtable 的数据模型. Cassandra 中的表是一个分布式多维结构, 包括行、列族、列和超级列. 此外, Cassandra 的分割和复制机制也和 Dynamo 的类似, 用于确保最终一致性.

• Bigtable 改进: 由于 Bigtable 不是开源的, 因此开源项目 HBase[29)] 和 Hypertable[30)]进行合并, 同时吸收了 Bigtable 的思想, 实现了类似的系统.

21) http://www.project-voldemort.com/voldemort/.
22) http://redis.io/.
23) http://fallabs.com/tokyocabinet/.
24) http://fallabs.com/tokyotyrant/.
25) http://memcached.org.
26) http://memcachedb.org.
27) http://basho.com/riak/.
28) http://code.google.com/p/scalaris/.
29) http://hbase.apache.org.
30) http://hypertable.org.

列式存储数据库大部分是基于 Bigtable 的模式, 只是在一致性机制和一些特性上有差异. 例如 Cassandra 主要关注弱一致性, 而 HBase 和 HyperTable 则关注强一致性.

(c) 文档数据库

文档数据库能够支持比键值存储复杂得多的数据结构. MongoDB[31], SimpleDB 和 CouchDB 是主要的文档数据库, 它们的数据模型和 JSON 对象类似 [115]. 不同文档存储系统的区别在于数据复制和一致性机制方面.

• 复制和分片 (Sharding): MongoDB 的复制机制使用主节点的日志文件实现, 日志文件保存了所有数据库中执行的高级操作. 复制过程中, 从节点向主节点请求自它们上一次同步之后所有的写操作, 并在它们的本地数据库中执行日志中的操作. MongoDB 通过自动分片将数据分散到成千上万的节点, 自动实现负载平衡和失效回复, 从而支持水平缩放. SimpleDB 将所有的数据复制到不同数据中心的不同服务器上以确保安全和提高性能. CouchDB 没有采用分片机制, 而是通过复制实现系统的扩展, 任一 CouchDB 数据库可以和其他实例同步, 因此可以构建任意类型的复制拓扑.

• 一致性: MongoDB 和 SimpleDB 都没有版本一致性控制和事务管理机制, 但是它们都提供最终一致性. CouchDB 的一致性则取决于是使用 master-master 配置还是 master-slave 配置. 前者能提供最终一致性, 而后者只能提供强一致性.

(d) 其他 NoSQL 和混合数据库

除了前面提到的数据存储系统, 还有许多其他项目支持不同的数据存储系统, 如 Neo4j[32], DEX[33], PNUTS [116], 以及 Dache [117].

由于关系型数据库和 NoSQL 数据库有着各自的优缺点, 结合两者的优势以获取较高的性能是一个较好的选择. 基于这种思想, Google 近来开发了如下几种集成了 NoSQL 和 SQL 数据库优点的数据库系统.

• Megastore [118] 将 NoSQL 数据库数据存储的可伸缩性和关系型数据库的便利结合在一起, 能够获得强一致性和高可用性. 其思想是首先将数据分区, 每个分区独立的复制, 在分区内提供完整 ACID 语义, 但是在分区间仅保证有限一致性. Megastore 的数据模型介于 RDBMS 的三元组和 NoSQL 的行 – 列存储之间, 其底层的数据存储依赖于 Bigtable.

• Spanner [119] 是第一个将数据分布到全球规模的系统, 并且在外部支持一致的分布式事务. 不同于 Bigtable 中版本控制的键值存储模型, Spanner 演化为时间上的多维数据库. 数据存储在半关系表中并创建版本, 每个版本根据提交的时间自动生成. 旧版本数据根据可配置的垃圾回收政策处理, 应用可以读取具有旧时间戳的数据. 某一粒度数据的复制可以由应用控制. 此外, 数据在服务器甚至数据中心上可以重新分片以均衡负载或应对失效. Spanner 显著的特点是外部一致读写和在某一时间戳的全度跨数据库一致读取.

• F1 [120] 是 Google 公司提出的用于广告业务的存储系统, 建立在 Spanner 的基础上. F1 实现了丰富的关系型数据库的特点, 包括严格遵从的 schema, 强力的并行 SQL 查询引擎, 通用事务、变更与通知的追踪和索引. 其存储被动态分区, 数据中心间的一致性复制能够处理数据中心崩溃引起的数据丢失.

(e) NoSQL 数据库比较

尽管有很多不同类型的数据库, 但没有一种数据库适用于任何场景, 不同的数据库在多个性能之

31) http://www.mongodb.org.
32) http://www.neo4j.org.
33) http://www.sparsity-technologies.com/dex.php.

表 5 NoSQL 存储系统设计

Table 5 Design of NoSQL storage system

Data model	Name	Producer	Data storage	Concurrency control	CAP option	Consistency
Key-value	Dynamo	Amazon	Plug-in	MVCC	AP	Eventually consistent
	Voldemort	LinkedId	RAM	MVCC	AP	Eventually consistent
	Redis	Salvatore Sanfilippo	RAM	Locks	AP	Eventually consistent
Column	BigTable	Google	GFS	Locks+stamps	CP	Eventually consistent
	Cassandra	Facebook	Disk	MVCC	AP	Eventually consistent
	HBase	Apache	HDFS	Locks	CP	Eventually consistent
	HyperTable	HyperTable	Plug-in	Locks	AP	Eventually consistent
Document	SimpleDB	Amazon	S3	None	AP	Eventually consistent
	MongoDB	10gen	Disk	Locks	AP	Eventually consistent
	CouchDB	Couchbase	Disk	MVCC	AP	Eventually consistent
Row	PNUTS	Yahoo	Disk	MVCC	AP	Timeline consistent

间进行不同的折中. Cooper 等 [121] 考虑了基于云的数据管理系统中的性能折中问题, 包括读和写、延迟和持续、同步和异步复制以及数据分区. 文献 [122~124] 等则讨论了一些其他的指标.

表 5 比较了一些系统的显著特点.

• 数据模型: 本文主要讨论键值存储、列式存储和文档存储等数据模型, PNUTS 的数据模型则是行式存储.

• 数据存储: 一些系统采用 RAM 快照或磁盘复制, 另一些系统使用带 RAM 缓存的磁盘存储, 还有些系统则采用可插入的后端, 允许接入不同的数据存储媒体, 或者需要一个标准的底层文件系统.

• 并发控制: 主要有 locks, MVCC 和无并发控制等三种方式. Locks 机制只允许一个用户在某一时刻读或者修改数据条目. MVCC 机制确保数据库的读一致性视图, 但是可能会导致同一时刻多个用户修改同一数据条目出现的多版本冲突问题. 另一些系统没有提供原子性, 允许不同用户并行修改同一数据对象的不同部分, 但是对版本没有进行控制.

• CAP 选项: CAP 定理 [125, 126] 表明共享数据系统最多只能选择一致性、可用性和分区容错性这三种性质中的两种. 为了解决部分失效, 基于云的数据库将数据广泛复制, 必须考虑数据的一致性和可用性. 因此在一致性和可用性间存在性能上的折中. 弱一致性模型的不同实现形式 [127] 已被实现以获得可以接受的系统可用性.

• 一致性: 根据 CAP 定理, 在满足可用性和分区容错性的同时难以获得严格的一致性, 通常采用两种弱一致性原则即最终一致性和 timeline 一致性. 最终一致性是指所有的更新将在系统中传播, 在很长一段时间内副本是一致的. Timeline 一致性则是指指定记录的所有副本按相同的次序更新记录 [116].

通常情况下, 在大数据应用中维持 ACID 原则非常困难. 数据管理工具的选择取决于前面提到的多种因素, 例如数据模型和数据源相关, 数据存储设备影响访问速率等. 大数据系统需要在成本、一致性和可用性间寻求平衡.

(3) 编程模型

尽管 NoSQL 数据库具有很多关系型数据库不具备的优点, 但是没有插入操作的声明性表述, 对查询和分析的支持也不够. 编程模型则对实现应用逻辑和辅助数据分析应用至关重要. 但是, 使用传

统的并行模型如 OpenMP [128] 和 MPI [129] 在大数据环境下实现并行编程非常困难. 许多并行编程模型已被提出应用于领域相关的应用. 这些模型有效地提高了 NoSQL 数据库的性能, 缩小了 NoSQL 和关系型数据库性能的差距, 因此 NoSQL 数据库逐渐成为海量数据处理的核心技术. 目前主要有三种编程模型: 通用处理模型、图处理模型以及流处理模型.

• 通用处理模型: 这种类型的模型用于解决一般的应用问题, 被用于 MapReduce [10] 和 Dryad [28] 中. 其中, MapReduce 是一个简单但功能强大的编程模型, 能将大规模的计算任务分配到大的商用 PC 集群中并行运行. 它的计算模型由用户定义的 Map 和 Reduce 两部分组成. MapReduce 将所有具有相同中间键 I 的中间结果聚合, 并且将其传递到相应的 Reduce 函数. Reduce 函数收到中间键 I, 并将和该键关联的一系列值进行合并, 产生更小集合的值. 简化的 MapReduce 只提供两个不透明的函数, 而无需一些最常用的操作 (如映射和过滤). 在 MapReduce 框架上添加 SQL 的特点是可以让 SQL 程序员快速高效地使用 MapReduce. 一些高级语言如 Google 的 Sawzall [130], Yahoo 的 Pig Latin [131], Facebook 的 Hive [132] 和 Microsoft 的 Scope [108] 已被开发用于提高程序员的效率. Dryad 是一个粗粒度的并行应用的通用分布式执行引擎. 一个 Dryad 作业是一个有向无环图, 图中顶点是程序, 边是数据信道. Dryad 在图中顶点所对应的一组计算机上运行作业, 并通过文件、TCP 管道和共享内存 FIFO 等数据信道通信. 运行时, 逻辑计算图自动映射到物理资源. MapReduce 可以看作是 Dryad 的特殊情况, 即图中只有两个阶段: 映射阶段和 Reduce 阶段. Yang 等在文献 [133] 中对使用 MapReduce 模型处理大数据进行了详尽的研究.

• 图处理模型: 社交网络分析和 RDF 等能够表示为实体间的相互联系, 因此可以用图模型来描述. 和流类型 (flow-type) 的模型相比, 图处理的迭代是固有的, 相同的数据集将不断被重访, 如 Google 的 Pregel [29], Graphlab [134] 和 X-stream [135]. Pregel 是用于大规模图计算 (如 web 图和社交网络分析) 的模型, 计算任务表示为一个有向图. 图中的顶点和一个用户定义的可修改的值相关, 有向边则和源顶点相关, 每条边有一个可变化的值和目标顶点的标识. 图完成初始化后, 程序迭代运行, 每一次迭代称为一个 superstep, 由同步点分离直到程序结束. 在每一个 superstep 中, 顶点以并行的方式执行给定算法逻辑的用户定义函数. 顶点能够修改它自身或者边输出的状态, 接收来自上一 superstep 的数据, 发送消息到其他顶点以及改变图的拓扑. 边没有与之联系的计算. 顶点可以通过投票终止其运行. 当所有的节点都未激活, 并且没有任何消息需要传递时, 程序将终止. Pregel 程序的结果是顶点的输出值集合, 通常和有向图输入是同构的. GraphLab 是一种面向机器学习算法的图处理模型, 包含三个组件: 数据图、更新函数和同步操作. 数据图是一个管理用户定义数据的容器, 包括模型参数、算法状态和统计数据. 更新函数是一个无状态的过程, 用于修改顶点范围内的数据, 调度将来在另一个顶点运行的更新函数. GraphLab 和 Pregel 的主要区别在于它们的同步模型. Pregel 在每一次迭代后有一个屏障 barrier, 所有节点在一次迭代后会完成全局的同步, 而 GraphLab 则是完全异步的. GraphLab 提供了三种一致性模型, 即完全、边和顶点一致性, 以允许不同级别的并行化. 此外, 相对于 Pregel 和 Graphlab 的以顶点为中心的计算模式, X-stream 是一种以边为中心的计算模式. X-Stream 提供了两个 API 函数用于表示图计算, Edge-centric scatter 和 Edge-centric gather. Edge-centric scatter 以一条边作为输入, 根据边的源点进行计算, 是否需要把更新后的值发送到边的目的顶点, 从而更新目的顶点的值. Edge-centric gather 以更新作为其输入, 重新计算目的顶点的数据域. 整个计算过程可组织成循环, 每一个循环步由 scatter 和 gather 组成: 首先 scatter 迭代所有边, 然后 gather 迭代 scatter 所产生的更新. 因此, X-stream 的 scatter 和 gather 是同步进行的. 当应用程序符合终止条件时, 循环终止.

• 流处理模型: S4 [24] 和 Storm 是两个运行在 JVM 上的分布式流处理平台. S4 实现了 actor 编程模型, 每个数据流中 keyed tupple 被看作是一个事件并被以某种偏好路由到处理部件 (processing

表 6 编程模型的特点

Table 6 Features of programming models

	MapReduce	Dryad	Pregel	GraphLab	Storm	S4
Application	General purpose parallel execution engine	General purpose parallel execution engine	Large scale graph processing	Large scale machine learning and data mining	Distributed stream processing	Distributed stream processing
Programming model	Map and Reduce	Directed acyclic graph	Directed graph	Directed graph	Directed acyclic graph	Directed acyclic graph
Parallelism	Concurrent execution within map and reduce phases	Concurrent execution of vertices during a stage	Concurrent execution over vertices within a superstep	Concurrent execution of non-overlaping scopes, defined by consistency model	Worker processes and executors	Worker processes and executors
Data handling	Distributed file system	Various storage media	Distributed file system	Memory or disk	Memory	Memory
Architecture	Master-slaves	Master-slaves	Master-slaves	Master-slaves	Master-slaves	Decentralized and symmetric
Fault tolerance	Node-level fault tolerance	Node-level fault tolerance	Checkpointing	Checkpointing	Partial fault tolerance	Partial fault tolerance

elements, PEs). PEs 形成一个有向无环图, 并且处理事件和发布结果. 处理节点 (processing nodes, PNs) 是 PEs 的逻辑主机并能监听事件, 将事件传递到处理单元容器 PEN 中, PEN 则以适当的顺序调用处理部件. Storm 和 S4 有着许多相同的特点. Storm 作业同样由有向无环图表示, 它和 S4 的主要区别在于架构: S4 是分布式对称架构, 而 Storm 是类似于 MapReduce 的主从架构.

表 6 比较了上述几种编程模型的特点. 首先, 尽管实时处理越来越重要, 批处理仍然是重要的数据处理方式. 其次, 大多数系统采取图作为编程模型是因为图能够表示更复杂的任务. 再次, 所有的系统都支持并发执行以加速处理速度. 第四, 流处理模型使用内存作为数据存储媒体以获得高访问和处理速率, 而批处理模型使用文件系统或磁盘存储海量数据以支持多次访问. 第五, 这些系统的架构通常是主 – 从式的, 但是 S4 则是分布式的架构. 最后, 不同的系统具有不同的容错策略. 对于 Storm 和 S4, 当节点失效时, 失效节点的任务将转移到备用节点运行; Pregel 和 GraphLab 则是用检查点用于容错; MapReduce 和 Dryad 则仅指出节点级别的容错.

此外, 一些其他的研究工作关注特定任务的编程模型, 如插入两个数据集 [136], 迭代计算 [137, 138], 容错内存计算 [139], 增量计算 [140~143] 和数据依赖的流控制决策 [144]. 特别地, 在内存计算模型中, Spark 是由 UC Berkeley 开发的一个的分布式计算框架 [139]. 它的特点是处理任务的中间结果能够保存在内存中, 从而避免传统数据重用 (如数据拷贝、硬盘 I/O、序列化) 的运行开销, 因此 Spark 能很好地使用于迭代算法与数据挖掘的应用之中.

8 阶段 IV: 大数据分析

大数据价值链最后也是最重要的阶段就是数据分析和处理, 其目标是提取数据中隐藏的数据, 提供有意义的建议以及辅助决策制定. 本节将首先讨论数据分析目标和分类, 然后介绍不同数据源的应用演化及其 6 个相关应用方向, 最后介绍数据分析中起着重要作用的几个常用方法.

8.1 数据分析目的和分类

数据分析处理来自对某一兴趣现象的观察、测量或者实验的信息. 数据分析目的是从和主题相关的数据中提取尽可能多的信息. 主要目标包括:

- 推测或解释数据并确定如何使用数据;
- 检查数据是否合法;
- 给决策制定合理建议;
- 诊断或推断错误原因;
- 预测未来将要发生的事情.

由于统计数据的多样性, 数据分析的方法大不相同. 可以将数据根据下述标准分为几类: 根据观察和测量得到的定性或定量数据, 根据参数数量得到的一元或多元数据. 此外, 有些工作对领域相关的算法进行了总结. Manimom 等在 [145] 中对数据挖掘算法进行了分类, 将其分为描述性 (descriptive)、预测性和验证性 (veryfying). Bhatt 等则在 [146] 中将多媒体分析方法划分为特征提取、变形、表示和统计数据挖掘. 然而并没有对大数据处理方法进行分类的工作. Blackett 等[34)]根据数据分析深度将数据分析分为三个层次: 描述性 (descriptive) 分析, 预测性分析和规则性 (prescriptive) 分析.

- 描述性分析: 基于历史数据描述发生了什么. 例如, 利用回归技术从数据集中发现简单的趋势, 可视化技术用于更有意义地表示数据, 数据建模则以更有效的方式收集、存储和删减数据. 描述性分析通常应用在商业智能和可见性系统.
- 预测性分析: 用于预测未来的概率和趋势. 例如, 预测性模型使用线性和对数回归等统计技术发现数据趋势, 预测未来的输出结果, 并使用数据挖掘技术提取数据模式 (pattern) 给出预见.
- 规则性分析: 解决决策制定和提高分析效率. 例如, 仿真用于分析复杂系统以了解系统行为并发现问题, 而优化技术则在给定约束条件下给出最优解决方案 [147].

8.2 应用演化

数据驱动的应用在过去几十年里已经出现. 例如, 20 世纪 90 年代在商业领域出现的商业智能, 21 世纪初期出现的基于数据挖掘的 web 搜索引擎. 接下来将介绍在不同时期典型大数据领域中具有高影响力的大数据分析应用的发展.

(1) 商业应用演化. 早期的商业数据是结构化的数据, 由企业或公司收集并存储在关系数据库管理系统中. 这些系统应用的数据分析技术通常是直观简单的. Gartner [148] 总结了商业智能应用的常用方法, 包括报表 (reporting)、仪表盘 (dashboard)、即时查询 (ad hoc query)、基于搜索的商业智能、在线事务处理、交互可视化、计分卡、预测模型和数据挖掘. 21 世纪初期, 互联网和 web 使得企业将其业务上线, 并能和客户直接联系. 大量的产品和客户信息如点击流数据日志和用户行为可以通过 web 收集. 通过使用不同的文本和 web 挖掘技术, 可以完成产品放置优化, 客户事务分析, 产品推荐和

34) Blackett G. Analytics Network - O.R. & Analytics. http://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx.

市场结构分析. 据报道[35], 2011 年移动手机和平板电脑的数量首次超过了笔记本和 PC 机. 移动手机和物联网构建了具有位置感知、个人为中心和上下文感知的革新性应用.

(2) 网络应用演化. 早期的网络提供电子邮件和网站服务, 因此文本分析、数据挖掘和网页分析技术被用于挖掘邮件内容、创建搜索引擎. 网络数据占据了全球数据的绝大部分, 包含文本、图像、视频、照片和交互式内容等多种类型的数据. 随后, 用于半结构化和无结构数据的分析技术得到了发展. 例如, 图像分析技术可以从照片中提取有意义的信息, 多媒体分析技术可以使商业或军事领域的视频监控系统自动化. 2004 年后, 诸如论坛、博客、社交网站、多媒体分享站点等在线社交媒体的出现使得用户能够产生、上传和共享丰富的用户自主创造内容. 从这些不同人们发布社交媒体内容中可以挖掘每天的热门事件和社会政治观点等, 从而提供及时的反馈和意见.

(3) 科学应用演化. 科学研究的许多领域中高生产量的传感器和仪器将产生大量的数据, 如天文学、海洋学、基因学和环境研究等学科领域. 美国 NSF 宣布对 BIGDATA 项目进行立项, 促进数据分享和分析[36]. 有些科学研究学科以前已开发出对海量数据的分析平台, 并取得了有效地成果. 例如在生物学科, iPlant[37]利用信息基础设施, 物理计算资源和支持互操作的分析软件等, 向致力于丰富植物科学知识的研究者、教育者和学生提供数据服务. iPlant 数据集是多样性的数据, 包含权威的和供参考的数据、实验数据、仿真建模数据、观察数据和其他处理后的数据.

基于以上的分析, 可以将数据分析的研究分为 6 个方向: 结构化数据分析、文本分析、web 数据分析、多媒体数据分析、社交网络数据分析和移动数据分析. 结构化数据分析是指传统的数据分析. Web 数据、多媒体数据、社交网络数据和移动数据, 从数据形态上可能包括结构化数据的某些数据类型 (如文本), 但是在特定的应用领域里面, 具有新的分析要求和特性. 所以需要从分析方法的角度对其分别分析, 这将在下节进行详细讨论.

8.3 常用分析方法

尽管目标和应用领域不同, 一些常用的分析方法几乎对所有的数据处理都有用, 下面将讨论三种类型的常用数据分析方法.

- 数据可视化: 与信息绘图学和信息可视化相关. 数据可视化的目标是以图形方式清晰有效地展示信息[38]. 一般来说, 图表和地图可以帮助人们快速理解信息. 但是, 当数据量增大到大数据的级别, 传统的电子表格等技术已无法处理海量数据. 大数据的可视化已成为一个活跃的研究领域, 因为它能够辅助算法设计和软件开发. Friedman[39]和 Frits [149] 分别从信息表示和计算机科学领域对数据可视化进行了探讨. Tabusvis [150] 则是一个轻型的可视化系统, 提供对多维数据的灵活、可定制的数据可视化.
- 统计分析: 基于统计理论, 是应用数学的一个分支. 在统计理论中, 随机性和不确定性由概率理论建模. 统计分析技术可以分为描述性统计和推断性统计. 描述性统计技术对数据集进行摘要 (summarization) 或描述, 而推断性统计则能够对过程进行推断. 更多的多元统计分析包括回归、因子分析、聚类和判别分析 [151].

35) Economist T. Beyond the PC, http://www.economist.com/node/21531109.

36) Foundation N S. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA). http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm.

37) http://www.iplantcollaborative.org/about/.

38) Friedman V. Data visualization and infographics. http://www.smashingmagazine.com/2008/01/14/monday-inspiration-data-visualization-and-infographics/.

39) Friedman V. Data Visualization: Modern Approaches. http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/.

• 数据挖掘: 是发现大数据集中数据模式的计算过程. 许多数据挖掘算法已经在人工智能、机器学习、模式识别、统计和数据库领域得到了应用. 2006 年 ICDM 国际会议上总结了影响力最高的 10 种数据挖掘算法 [152], 包括 C4.5, k-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, 朴素贝叶斯和 CART, 覆盖了分类、聚类、回归和统计学习等方向. 此外, 一些其他的先进技术如神经网络和基因算法也被用于不同应用的数据挖据. 有时候, 几乎可以认为很多方法间的界线逐渐淡化 (当然, 还存在), 例如数据挖掘、机器学习、模式识别、甚至视觉信息处理、媒体信息处理等等. 此处以 “数据挖掘” 作为一个通称.

9 大数据分析分类

表 7 从数据生命周期的角度, 从数据源、数据特性等方面总结比较了主要的数据分析方法, 包括结构化数据分析、文本分析、web 数据分析、多媒体数据分析、社交网络数据分析和移动数据分析.

9.1 结构化数据分析

在科学研究和商业领域产生了大量的结构化数据, 这些结构化数据可以利用成熟的 RDBMS、数据仓库、OLAP 和 BPM [32] 等技术管理, 而采用的数据分析技术则是前面介绍的数据挖掘和统计分析技术. 近来深度学习 (deep learning) 逐渐成为一个主流的研究热点 [193, 194]. 许多当前的机器学习算法依赖于用户设计的数据表达和输入特征, 这对不同的应用来说是一个复杂的任务. 而深度学习则集成了表达学习 (representation learning), 学习多个级别的复杂性/抽象表达 [195]. 此外, 许多算法已成功用于一些最近的应用. 例如, 统计机器学习, 基于精确的数据模型和强大的算法, 被应用在异常检测 [196] 和能量控制 [197]. 利用数据特征, 时空挖掘技术能够提取模型中的知识结构, 以及高速数据流与传感器数据中的模式 (pattern) [198], 文献 [199] 对大规模图像的模式挖掘进行了研究. 由于电子商务、电子政务和医疗健康应用对隐私的需求, 隐私保护数据挖掘也被广为研究 [200]. 随着事件数据, 过程发现和一致性检查技术的发展, 过程挖掘也逐渐成为一个新的研究方向, 即通过事件数据分析过程 [201].

9.2 文本分析

文本数据是信息储存的最常见形式, 包括电子邮件、文档、网页和社交媒体内容, 因此文本分析比结构化数据具有更高的商业潜力. 文本分析又称为文本挖掘, 是指从无结构的文本中提取有用信息或知识的过程. 文本挖掘是一个跨学科的领域, 涉及信息检索、机器学习、统计、计算语言和数据挖掘. 大部分的文本挖掘系统建立在文本表达和自然语言处理 (NLP) 的基础上.

文档表示和查询处理是开发矢量空间模型、布尔检索模型和概率检索模型 [153] 的基础, 这些模型又是搜索引擎的基础.

NLP 技术能够增加文本的可用信息, 允许计算机分析、理解甚至产生文本. 词汇识别、语义释疑、词性标注和概率上下文无关文法 [154] 等是常用的方法. 基于这些方法提出了一些文本分析技术, 如信息提取、主题建模、摘要 (summarization)、分类、聚类、问答系统和观点挖掘. 信息提取技术是指从文本中自动提取具有特定类型的结构化数据. 命名实体识别 (named-entity recognition, NER) 是信息提取的子任务, 其目标是从文本中识别原子实体并将其归类到人、地点和组织等类别中. NER 最近被应用于一些新的分析应用 [155] 和生物医学 [156] 中. 主题模型则建立在文档包含多个主题的情况. 主题是一个基于概率分布的词语, 主题模型对文档而言是一个通用的模型, 许多主题模型被用于分析文

表 7 大数据分析方法分类

Table 7 Taxonomy of big data analytics methods

Analysis domains	Sources	Characteristics	Approaches
Structured data analysis	Customer transactions	Structured records	Data mining [152]
	Scientific data	Less volume and real time	Statistical analysis [151]
Text analysis	Logs	Unstructured	Document presentation [153]
	Email	Rich textual	NLP [154]
	Corporate documents	Context	Information extraction [155, 156]
	Government regulations	Semantic	Topic model [157]
	Text content of webpages	Language dependent	Summarization [158]
	Feedback and comments		Categorization [159]
			Clustering [160]
			Question answering [161]
			Opinion mining [162]
Web analytics	Various webpages	Text and hyperlink	Web content mining [163]
		Symbolic	Web structure mining [164∼166]
		Metadata	Web usage mining [167]
Multimedia analytics	Corporation and user	Image, audio, video	Summarization [168], Annotation [169]
	Generated multimedia	Massive	Indexing and retrieval [170]
	Surveillance	Redundancy	Recommendation [171, 172]
	Health and patient media	Semantic gap	Event detection [173]
Social network analytics	Bibliometric	Rich content	Link prediction [174∼176]
	Sociology network	Social relationship	Community detection [177, 178]
	Social networks	Noisy and redundancy	Network evolution [179∼182]
		Fast evolution	Influence analysis [183, 184]
			Key words search [185]
			Classification [186], Clustering [187]
			Transfer learning [188, 189]
Mobile analytics	Mobile apps	Location based	Monitoring [190∼192]
	Sensors	Person specific	Location based mining
	RFID	Fragmented information	

档内容和词语含义 [157]. 文献 [202] 引入一个新的主题模型, 即主题超图, 用于描述长文档的主体结构. 文本摘要技术从单个或多个输入的文本文档中产生一个缩减的摘要, 分为提取式 (extractive) 摘要和概括式 (abstractive) 摘要 [158]. 提取式摘要从原始文档中选择重要的语句或段落并将它们连接在一起, 而概括式摘要则需理解原文并基于语言学方法以较少的语句复述. Morrison 等 [203] 提出一种演化网络, 用于多元数据摘要. 文本分类技术则用于识别文档主题, 并将之归类到预先定义的主题或主题集合中, 基于图表示和图挖掘的文本分类在近年来得到了关注 [159]. 文本聚类技术用于将类似的文档聚合, 和文本分类不同的是, 文本聚类不是根据预先定义的主题将文档归类. 文本聚类中, 文档可以表现出多个子主题. 一些数据挖掘中的聚类技术可以用于计算文档的相似度. 有研究证实了结构化的关系信息能够用于增加 Wikipedia 的聚类效率 [160]. Deng 等 [204] 提出了一个统一的时空数据聚类分析框

架, 基于时空统计方法和智能计算技术设计了一种新的时空聚类算法. 问答系统主要设计用于如何为给定问题找到最佳答案, 涉及问题分析、源检索、答案提取和答案表示等技术 [161]. 问答系统可以用在教育、网站、健康和答辩等场合. 观点挖掘类似于情感分析, 是指提取、分类、理解和评估在新闻、评论和其他用户自主创造内容中观点的计算技术, 它能够为了解公众或客户对社会事件、政治动向、公司策略、市场营销活动和产品偏好看法提供机会 [162].

9.3 Web 数据分析

过于十几年间网页数据爆炸式的增长, 使得网页数据分析也成为活跃的领域. Web 数据分析的目标是从 web 文档和服务中自动检索、提取和评估信息以发现知识, 涉及数据库、信息检索、NLP 和文本挖掘, 可分为 web 内容挖掘、web 结构挖掘和 web 用法挖掘 (web usage mining) [205].

Web 内容挖掘是从网站内容中获取有用的信息或知识. web 内容包含文本、图像、音频、视频、符号、元数据和超链接等不同类型的数据. 而关于图像、音频和视频的数据挖掘被归入多媒体数据分析, 将在随后讨论. 由于大部分的 web 数据是无结构的文本数据, 因此许多研究都关注文本和超文本的数据挖掘. 如前所述, 文本挖据已经比较成熟, 而超文本的挖掘需要分析包含超链接的半结构化 HTML 网页. 有监督学习 (Supervised learning) 或分类在超文本分析中起到重要的作用, 例如电子邮件管理、新闻组管理和维护 web 目录等 [163]. Web 内容挖掘通常采用两种方法: 信息检索和数据库. 信息检索方法主要是辅助用户发现信息或完成信息的过滤; 数据库方法则是在 web 上对数据建模并将其集成, 这样能处理比基于关键词搜索更为复杂的查询.

Web 结构挖掘是指发现基于 web 链接结构的模型. 链接结构表示站点内或站点之间链接的关系图, 模型反映了不同站点之间的相似度和关系, 并能用于对网站分类. Page Rank [164], CLEVER [165] 和 Focused Crawling [166] 利用此模型发现网页. Focused Crawling 的目的是根据预先定义的主题有选择地寻找相关网站, 它并不收集或索引所有可访问的 web 文档, 而是通过分析 crawler 的爬行边界, 发现和爬行最相关的一些链接, 避免 web 中不相关的区域, 从而节约硬件和网络资源.

Web 用法挖掘则是对 web 会话或行为产生的次要数据进行分析. 与 web 内容挖掘和结构挖掘不同的是, web 用法挖掘不是对 web 上的真实数据进行分析. Web 用法数据包括 web 服务器的访问日志, 代理服务器日志, 浏览器日志, 用户信息、注册数据, 用户会话或事务, cookies, 用户查询、书签数据, 鼠标点击及滚动数据, 以及用户与 web 交互所产生的其他数据. 随着 web 服务和 web 2.0 系统的日益成熟和普及, web 用法数据将更加多样化. Web 用法挖掘在个性化空间、电子商务、web 隐私和安全等方面将起到重要的作用. 例如, 协作推荐系统可以根据用户偏好的相同或相异实现电子商务的个性化 [167].

9.4 多媒体数据分析

多媒体数据分析是指从多媒体数据中提取有趣的知识, 理解多媒体数据中包含的语义信息. 由于多媒体数据在很多领域比文本数据或简单的结构化数据包含更丰富的信息, 提取信息需要解决多媒体数据中的语义分歧. 多媒体分析研究覆盖范围较广, 包括多媒体摘要、多媒体标注、多媒体索引和检索、多媒体推荐和多媒体事件检测.

音频摘要可以简单地从原始数据中提取突出的词语或语句, 合成为新的数据表达; 视频摘要则将视频中最重要或最具代表性的序列进行动态或静态的合成. 静态视频摘要使用连续的一系列关键帧或上下文敏感的关键帧表示原视频, 这些方法比较简单, 并已被用于 Yahoo, Alta Vista 和 Google, 但是

它们的回放体验较差. 动态视频摘要技术则使用一系列的视频片段表示原始视频, 并利用底层视频特征进行平滑以使得最终的摘要显得更自然 [168].

多媒体标注是指给图像和视频分配一些标签, 可以在语法或语义级别上描述它们的内容. 在标签的帮助下, 很容易实现多媒体内容的管理、摘要和检索. 由于人工标注非常耗时并且工作量大, 没有人工干预的自动多媒体标注得到了极大的关注. 多媒体自动标注的主要困难是语义分歧, 即底层特征和标注之间的差异. 尽管取得了一些重要的进展, 目前的自动标注方法性能并不能令人满意. 一些研究开始同时利用人和计算机对多媒体进行标注 [169].

多媒体索引和检索处理的是多媒体信息的描述、存储和组织, 并帮助人们快速方便地发现多媒体资源 [170]. 一个通用的视频检索框架包括 4 个步骤: 结构分析, 特征提取, 数据挖掘、分类和标注, 以及查询和检索. 结构分析是通过镜头边界检测、关键帧提取和场景分割等技术, 将视频分解为大量具有语义内容的结构化元素. 结构分析完成后, 第 2 步是提取关键帧、对象、文本和运动的特征以待后续挖掘 [206~208], 这是视频索引和检索的基础. 根据提取的特征, 数据挖掘、分类和标注的目标就是发现视频内容的模式, 将视频分配到预先定义的类别, 并生成视频索引. Shao 等 [209] 提出一种基于内容的视频检索方法, 通过时间和空间定位从数据库中有效地检索相关行为的视频. 在大规模图像检索方面, Chen 等 [210] 提出一种基于图哈希的方法 (spectral embedded hashing). Song 等 [211] 提出一种基于哈希方法的近似相似多媒体检索, 通过机器学习方法有效地学习一组哈希函数来给数据产生哈希码. 此外, Dong 等 [212] 利用 shearlets 和线性回归, 对图像进行纹理分类与检索, 其平均分类正确率比现有技术要高.

多媒体推荐的目的是根据用户的偏好推荐特定的多媒体内容, 已被证明是一个能提供高质量个性化内容的有效方法. 现有的推荐系统大部分是基于内容和基于协作过滤的机制. 基于内容的方法识别用户兴趣的共同特征, 并且给用户推荐具有相似特征的多媒体内容. 这些方法依赖于内容相似测量机制, 容易受有限内容分析的影响. 基于协作过滤的方法将具有共同兴趣的人们组成组, 根据组中其他成员的行为推荐多媒体内容 [171]. 混合方法则利用基于内容和基于协作过滤两种方法的优点提高推荐质量 [172].

多媒体事件检测是在事件库视频片段中检测事件是否发生的技术 [213]. 视频事件检测的研究才刚刚起步, 已有的大部分研究都集中在体育或新闻事件, 以及重复模式事件 (如监控视频中的跑步) 或不常见的事件. Ma 等在 [173] 中提出了一种新的即时多媒体事件检测算法, 以应付训练正例不足的场景.

9.5 社交网络数据分析

随着在线社交网络的兴起, 网络分析从早期的文献计量学分析 [214] 和社会学网络分析 [215] 到 21 世纪的社交网络分析. 社交网络包含大量的联系和内容数据, 其中联系数据通常用一个图拓扑表示实体间的联系; 内容数据则包含文本、图像和其他多媒体数据. 显然, 社交网络数据的丰富性给数据分析带来了前所未有的挑战和机会. 从以数据为中心的角度, 社交网络的研究方向主要有两个: 基于联系的结构分析和基于内容的分析 [216].

基于联系的结构分析关注链接预测、社区发现、社交网络演化和社交影响分析等方向. 社交网络可以看成一个图, 图中顶点表示人, 边表示对应的人之间存在特定的关联. 由于社交网络是动态的, 新的节点和边会随着时间的推移而加入图中. 链接预测对未来两个节点关联的可能性进行预测. 链接预测技术主要有基于特征的分类、概率方法和线性代数方法. 基于特征的分类方法选择节点对的一组特征, 利用当前的链接信息训练二进制分类器预测未来的链接 [174]; 概率方法对社交网络节点的链接概率进行建模 [175]; 线性代数方法通过降维相似矩阵计算节点的相似度 [176]. 社区是指一个子图结构, 其

中的顶点具有更高的边密度, 但是子图之间的顶点具有较低的密度. 用于检测社区 [177] 的方法中, 大部分都是基于拓扑的, 并且依赖于某个反映社区结构思想的目标函数. Du 等 [178] 利用真实世界中社区存在重叠的特性, 提出了大规模社交网络中的社区发现算法. Pelechrinis 等也通过基于位置的社交网络挖掘以用户为中心的网络结构. 社交网络演化研究则试图寻找网络演化的规律, 并推导演化模型. 部分经验研究 [179~181] 发现, 距离偏好、地理限制和其他一些因素对社交网络演化有着重要的影响. 一些通用的模型 [182] 也被提出用于辅助网络和系统设计. 当社交网络中个体行为受其他人感染时即产生社交影响, 社交影响的强度取决于多种因素, 包括人与人之间的关系、网络距离、时间效应和网络及个体特性等. 定量和定性测量个体施加给他人的影响, 会给市场营销、广告和推荐等应用带来极大的好处 [184].

随着 web 2.0 技术的发展, 用户自主创造内容在社交网络中取得了爆炸性的增长. 社交媒体是指这些用户自主创造的内容, 包括博客、微博、图片和视频分享、社交图书营销、社交网络站点和社交新闻等. 社交媒体数据包括文本、多媒体、位置和评论等信息. 几乎所有的对结构化数据分析、文本分析和多媒体分析的研究主题都能转移到社交媒体分析中. 但是社交媒体分析面临着前所未有的挑战. 首先, 社交媒体数据每天不断增长, 应该在一个合理的时间限制范围对数据进行分析; 其次社交媒体数据包含许多干扰数据, 例如博客空间存在大量垃圾博客; 再次社交网络是动态、不断变化、迅速更新的. 简单来说, 社交媒体和社交网络联系紧密, 社交媒体数据的分析无疑也受到社交网络动态变化的影响. 社交媒体分析即社交网络环境下的文本分析和多媒体分析. 社交媒体分析的研究处于起步阶段. 社交网络的文本分析应用包括关键词搜索、分类、聚类和异构网络中的迁移学习. 关键词搜索利用了内容和链接行为 [185]; 分类则假设网络中有些节点具有标签, 这些被标记的节点则可以用来对其他节点分类 [186]; 聚类则确定具有相似内容的节点集合 [187]. 由于社交网络中不同类型的对象之间存在大量链接的信息, 如标记、图像和视频等, 异构网络的迁移学习用于不同链接的信息知识迁移 [188]. 在社交网络中, 多媒体数据集是结构化的并且具有语义本体、社交互动、社区媒体、地理地图和多媒体内容等丰富的信息. 文献 [217] 讨论了地域社交多媒体信息挖掘的应用, 包括移动位置检索、地标识别、场景重构、景点推荐等. 社交网络的结构化多媒体又称为多媒体信息网络. 多媒体信息网络的链接结构是逻辑上的结构, 对网络非常重要. 多媒体信息网络中有四种逻辑链接结构: 语义本体、社区媒体、个人相册和地理位置 [216]. 基于逻辑链接结构, 可以提高检索系统 [218]、推荐系统 [219]、协作标记 [220] 和其他应用 [221,222] 的性能.

9.6 移动数据分析

随着移动计算 [223~225] 的迅速发展, 更多的移动终端 (移动手机、传感器和 RFID) 和应用逐渐在全世界普及. 2012 年末移动数据流量每月达到 885 PB [226]. 巨量的数据对移动分析提出了需求, 但是移动数据分析面临着移动数据特性带来的挑战, 如移动感知、活动敏感性、噪声和冗余. 目前移动数据分析的研究远未成熟, 下面介绍一些具有代表性的移动数据分析应用.

RFID 能够在一定范围内读取一个和标签 (tag) 相联系的唯一产品标识码 [227], 标签能够用于标识、定位、追踪和监控物理对象, 在库存管理和物流领域得到了广泛的应用. 然而, RFID 数据给数据分析带来了许多挑战: (i) RFID 数据本质上是充斥着干扰数据和冗余数据的; (ii) RFID 数据是时间相关的、流式的、容量大并且需要即时处理. 通过挖掘 RFID 数据的语义 (如位置、聚集和时间信息), 可以推断一些原子事件追踪目标和监控系统状态 [190].

无线传感器、移动技术和流处理技术的发展促进了体域传感器网络的部署, 用于实时监控个体健康状态 [191,192]. 医疗健康数据来自具有不同特性的异构传感器, 如多样化属性、时空联系和生理特征

等特性, 并存在隐私和安全问题.

从上述讨论可以发现, 大部分的移动数据分析技术既是描述性分析, 也是预测性分析.

9.7 光学观测和监控数据分析

光学观测和监控数据的分析不仅仅存在于数据的本身, 同数据获取端的设置和参数等都密切相关, 是一个非常复杂的系统工程. 本文暂不深入进行讨论.

10 大数据系统基准 (benchmark)

10.1 面临的挑战

事务处理性能委员会 (Transaction Processing Performance Council) 制定的系列基准极大地促进了传统关系型数据库的发展和商业化. 随着大数据系统研究的逐步成熟, 学术界和产业界试图创建新的类似 TPC 的基准, 对大数据系统的性能进行比较和评估. 然而到目前为止, 还没有一个可用的标准基准. 大数据系统的独特性质对新基准的提出带来了如下的挑战.

• 系统复杂性: 大数据系统通常由多个模块或组件组成, 这些模块有着不同的功能并耦合在一起. 对整个系统建模和为所有模块提供一个统一的框架并不容易.

• 应用多样性: 一个好的基准应该反映大数据系统的典型特性, 例如应用访问模式和性能需求等. 由于大数据系统的多样性, 使得提取显著特征非常复杂.

• 数据规模: 在传统的 TPC 基准中, 测试集通常比真实的客户数据集大得多, 因此测试结果能精确的反映真实性能. 然而, 大数据的数据量巨大并且不断增长, 必须考虑一种有效的方式测试具有小数据集的产品.

• 系统演化: 大数据增长率不断增加, 大数据系统必须不断演化, 以适应日益变化的需求, 因此大数据基准也要迅速变化.

10.2 研究现状

大数据基准的研究也刚刚起步, 可以分为组件级别 (component-level) 的基准和系统级别的基准. 组件级别的基准也称为微基准 (micro benchmark), 用于评价独立组件的性能; 系统级基准提供端到端系统测试框架. 在大数据相关组件中, 数据存储已发展成熟并可以准确地建模. 因此许多微基准被提出用于评价数据存储组件, 主要可以分为三类.

• TPC 基准: TPC 系列基准[40)]用于评价关系型数据库的事务性工作负荷. TCP-DS [228] 是 TPC 最近颁布的支持决策制定的基准, 它事实上已涉及大数据系统的一些方面. 具体来说, TCP-DS 能够产生最多 100 Terabytes 的结构化数据, 并且通过初始化数据库, 能在单用户和多用户模型下执行 SQL 查询.

• NoSQL 基准: NoSQL 数据库能够高效地处理半结构化和无结构数据, 这对大数据集中占较大比例的无结构数据非常适用. Yahoo 开发了它的云服务基准 —— YCSB [121], 用于评价 NoSQL 数据库. YCSB 由产生工作负载的客户和一个标准负载包构成, 负载包覆盖了部分性能空间, 如大量读操作负载、大量写操作负载和扫描负载. 这三种负载可针对 Cassandra, HBase, PNUTs 和简单的共享 MySQL

40) http://www.tpc.org/information/benchmarks.asp.

等 4 种数据存储系统运行. 其他一些研究 [229, 230] 扩展了 YCSB 框架, 集成了一些高级特征, 例如预分割、大容量加载和服务器方过滤等.

• Hadoop 基准: Hadoop 已逐渐成为大数据分析的主流框架, 一些研究者试图构建类似 TPC 的 MapReduce 基准. GridMix[41)]和 PigMix[42)]是 Apache 的 Hadoop 项目中内置的两个测试框架, 可以评估 Hadoop 集群和 Pig 查询的性能. Palvod 等 [231] 定义了由任务集合构成的基准, 将 Hadoop 和其他两种并行 RDBMS 系统进行了性能比较, 测试结果表明了性能上的 tradesoff, 并认为未来的系统应该同时考虑这两种类型的体系架构. GraySort[43)]是一个已被广泛使用的大规模排序基准, 这些基准可以看成是许多类型和大小作业的复杂迭加. 通过对 Facebook 和 Yahoo! 中 MapReduce 追踪信息的比较和分析, Chen 等 [232] 开发了一个开源的统计工作负载注入器 (SWIM), SWIM 套件包括三个关键组件: 真实 MapReduce 工作负载仓库, 生成代表性工作负载的负载合成工具, 和执行历史工作负载的负载重放工具. SWIM 套件能够获得基于现实工作负载的性能评估, 并能发现系统资源瓶颈. 随后他们在文献 [233] 中对工作负载进行了更复杂的分析. PDMiner [234] 则是一个基于大规模数据处理平台 Hadoop 的并行分布式数据挖掘工具平台. 在 PDMiner 中开发实现了各种并行数据挖掘算法, 比如数据预处理、关联规则分析以及分类、聚类等算法.

Ghazal 等 [235] 基于生产零售模型第一次提出了一个端到端的大数据基准 —— BigBench, 由两个主要部件构成: 数据生成器和工作负载查询规范. 数据生成器可以产生结构化、半结构化和无结构数据这三种类型的原始数据; 查询规范则根据 McKinsey 报告中生产零售商的典型特征, 定义了查询类型、数据处理语言和分析算法的类型. BigBench 覆盖了大数据系统的 “3Vs” 特性.

11 大数据科学问题

大数据系统面临的许多挑战需要通过后续的研究解决 [236~238]. 在整个大数据生命周期中, 从大数据平台和处理模型到应用场景等各方面, 都存在一些值得研究的方向.

• 大数据基础平台: 尽管 Hadoop 已成为大数据分析的主流框架, 但是和发展了 40 余年的 RDBMS 系统相比, 大数据平台还远未成熟. 首先, Hadoop 需要集成实时的数据采集和传输机制, 提供非批处理方式的快速处理机制. 其次, Hadoop 提供了一个简化的用户编程接口, 隐藏了复杂后台执行的细节, 这种简化在一定程度会降低处理性能. 应该设计类似于 DBMS 系统的更先进的接口, 从多个角度优化 Hadoop 性能. 再次, 大规模 Hadoop 集群由成千上万甚至几十万台服务器构成, 要消耗大量的能量. Hadoop 能否大范围部署取决于其能量效率. 此外, 基础平台的研究还包括海量数据分布式存储管理, 实时索引查询, 大数据平台功耗, 以及海量数据实时采集、传输和处理等问题. Hu 等 [239] 提出了一个基于 SDN 的大数据平台, 用于社交 TV 数据分析.

• 处理模式: 现有的批处理模式难以适应海量数据实时处理的需求, 需要设计新的实时处理模式. 在传统的批处理模式中, 数据首先被存储, 随后扫描整个数据集并进行处理得到分析结果, 时间极大地浪费在数据传输、存储和重复扫描上. 新的实时处理模式可以减少这种浪费. 例如, 现场 (in-situ) 分析可以避免因数据传输到集中存储基础设施所带来的开销, 从而提高实时性能. 大数据系统是个系统问题, 在处理模式上需要考虑多方面因素. 一个任务的解决不仅仅是算法的问题, 与传输和存储等各方面也有关系. 仅从计算复杂度来进行分析并不足够, 因为理论上计算复杂度低的算法, 实际在机器上

41) http://hadoop.apache.org/docs/stable/gridmix.html.
42) https://cwiki.apache.org/PIG/pigmix.html.
43) http://sortbenchmark.org.

运行也不一定快. 此外, 由于大数据低价值密度的特点, 可以采取降维或基于采样的数据分析减少处理的数据量. 具体而言, 处理模式研究涉及大数据可视化计算分析、大数据处理复杂性问题、并行化深度机器学习和数据挖掘算法、异构数据融合、基于海量数据低价值密度采样问题和高维海量数据降维问题.

• 大数据应用: 大数据的研究刚刚起步, 典型大数据应用的研究能够给商业带来利润, 提高政府部门效率, 并且促进人类科学的发展. 主要的应用场景有: 图数据并行计算模型和框架, 社会网络分析、排名和推荐, web 信息挖掘和检索, 媒体分析检索和自然语言处理.

• 大数据隐私: 隐私也是大数据领域的重要问题. 用户的信息可能会被遭到暴露, 比如企业的营销策略、个人的消费习惯等. 特别是在电子商务、电子政务和医疗健康领域, 隐私保护显得尤其重要, 需要增强访问控制. 此外, 还需要在增强访问控制和数据处理的便利性之间达到一个平衡 [240].

• "无限" 数据: 随着云计算、物网联、移动终端、可穿戴设备等技术的发展, 我们已经进入了大数据的时代. 然而, 产生的数据量也随之日益增长. 目前的大数据, 在不久的将来还只会是小数据. 因此, 对于未来的大数据最确切的描述, 或许会是 "无限" 数据. 相应地, 数据的增量和学习方法会是一个重要的问题. 例如, 当前用 10 亿个样本训练了一个分类器, 效果很好, 但未来样本数增加到 15 亿的时候 (之前的 10 亿样本已经不能完全表达数据的特征), 就会面临一个问题, 是利用 15 亿个样本重新训练一个分类器, 还是利用新增加的 5 亿个样本来修正原来用 10 亿个样本训练得到的分类器呢? 如果重新训练分类器, 这将会造成过大的时间和空间开销, 并且可扩展性差. 以往, 为了避免重复学习历史样本和减少后继的训练时间, 我们可以采用增量学习的方法, 即利用历史学习的结果和新增加的样本来修正之前的分类器. 但面对不断演化的 "无限" 大数据, 是否需要研究新型的增量学习方法, 从而动态自适应地进行预测并确保模型的准确性, 或许将会是大数据未来发展需要解决的重要问题.

12 结论

本文介绍了大数据的基本概念, 强调了覆盖数据生命周期的大数据价值链. 大数据价值链由数据生成、数据获取、数据存储和数据分析 4 个阶段构成. 此外, 从系统角度, 本文介绍并讨论了不同阶段的一些方法和机制. 在数据生成阶段, 给出了一些数据源并讨论了数据属性; 在数据获取阶段, 讨论了典型的数据采集方法, 数据传输机制和数据预处理技术; 在数据存储阶段, 分析了 NoSQL 数据存储, 讨论了一些代表性的计算模型; 在数据分析阶段, 根据数据特性介绍了不同的数据分析方法, 随后详细讨论了不同的大数据分析研究方向. 最后本文对大数据系统基准进行了介绍, 并对大数据的一些科学问题进行了探讨和总结.

参考文献

1 Gantz J, Reinsel D. Extracting value from chaos. IDC iView, 2011: 1–12
2 Manyika J, Chui M, Brown B, et al. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011
3 Cukier K. Data, data everywhere. Economist, 2010, 394: 3–16
4 Lohr S. The age of big data. New York Times, 2012, 11
5 Noguchi Y. Following digital breadcrumbs to big data gold. National Public Radio, 2011
6 Noguchi Y. The search for analysts to make sense of big data. National Public Radio, 2011
7 White House. Fact Sheet: Big Data Across the Federal Government. Office of Science and Technology Policy, 2012
8 Howard J H, Kazar M L, Menees S G, et al. Scale and performance in a distributed file system. ACM Trans Comput Syst, 1988, 6: 51–81

9 Cattell R. Scalable SQL and NoSQL data stores. SIGMOD Rec, 2011, 39: 12–27
10 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM, 2008, 51: 107–113
11 White T. Hadoop: the definitive guide. O'Reilly Media, Inc., 2012
12 Zikopoulos P, Eaton C. Understanding big data: analytics for enterprise class hadoop and streaming data. New York: McGraw-Hill Osborne Media, 2011
13 Meijer E. The world according to LINQ. Commun ACM, 2011, 54: 45–51
14 Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Gartner, 2001
15 Cooper M, Mell P. Tackling Big Data. NIST, 2012
16 Team O R. Big Data Now: Current Perspectives from O'Reilly Radar. O'Reilly Media, 2011
17 Marche S. Is Facebook making us lonely. Atlantic, 2012, 309: 60–69
18 Borkar V R, Carey M J, Li C. Big data platforms: what's next? XRDS: Crossroads, The ACM Magazine for Students, 2012, 19: 44–49
19 Borkar V, Carey M J, Li C. Inside Big Data management: ogres, onions, or parfaits? In: Proceedings of the 15th International Conference on Extending Database Technology, Berlin, 2012. 3–14
20 Dewitt D J, Gray J. Parallel database systems: the future of high performance database systems. Commun ACM, 1992, 35: 85–98
21 Ghemawat S, Gobioff H, Leung S T. The Google file system. In: Proceedings of the nineteenth ACM symposium on Operating systems principles, New York, NY, USA, 2003. 29–43
22 Hey A J, Tansley S, Tolle K M, et al. The fourth paradigm: data-intensive scientific discovery. 2009
23 Tatbul N. Streaming data integration: Challenges and opportunities. In: Proceedings of the 26th International Conference on Data Engineering Workshops, California, 2010. 155–158
24 Neumeyer L, Robbins B, Nair A, et al. S4: distributed stream computing platform. In: Proceedings of IEEE International Conference on Data Mining Workshops, Sydney, 2010. 170–177
25 Goodhope K, Koshy J, Kreps J, et al. Building LinkedIn's real-time activity data pipeline. Data Engineering, 2012, 35: 33–45
26 Agrawal D, Bernstein P, Bertino E, et al. Challenges and opportunities with big data — a community white paper developed by leading researchers across the United States. Computing Research Association, 2012
27 Fisher D, DeLine R, Czerwinski M, et al. Interactions with big data analytics. Interactions, 2012, 19: 50–59
28 Isard M, Budiu M, Yu Y, et al. Dryad: distributed data-parallel programs from sequential building blocks. In: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems, New York, 2007. 59–72
29 Malewicz G, Austern M H, Bik A J, et al. Pregel: a system for large-scale graph processing. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Indianapolis, 2010. 135–146
30 Melnik S, Gubarev A, Long J J, et al. Dremel: interactive analysis of web-scale datasets. Proc VLDB Endowment, 2010, 3: 330–339
31 Labrinidis A, Jagadish H V. Challenges and opportunities with big data. Proc VLDB Endowment, 2012, 5: 2032–2033
32 Chaudhuri S, Dayal U, Narasayya V. An overview of business intelligence technology. ACM Commun, 2011, 54: 88–98
33 Evans D, Hutley R. The Explosion of Data. White Paper, 2010
34 Gantz J, Reinsel D. The digital universe decade-are you ready. White Paper, IDC, 2010
35 Bryant R E. Data-intensive scalable computing for scientific applications. Comput Sci Eng, 2011, 13: 25–33
36 Wang X Q. Semantically-aware Data Discovery and Placement in Collaborative Computing Environments. Dissertation for Ph.D. Degree. Taiyuan: Taiyuan University of Technology, 2012
37 Middleton S E, Sabeur Z A, Löwe P, et al. Multi-disciplinary approaches to intelligently sharing large-volumes of real-time sensor data during natural disasters. Data Sci J, 2013, 12: WDS109–WDS113
38 Laurila J K, Gatica-Perez D, Aad I, et al. The mobile data challenge: big data for mobile computing research. In: Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing, Newcastle, 2012. 1–8
39 Chandramohan V, Christensen K. A first look at wired sensor networks for video surveillance systems. In: Proceedings of the 27th Annual IEEE Conference on Local Computer Networks, Tampa, 2002. 728–729
40 Selavo L, Wood A, Cao Q, et al. Luster: wireless sensor network for environmental research. In: Proceedings of the 5th International Conference on Embedded Networked Sensor Systems, Sydney, 2007. 103–116
41 Barrenetxea G, Ingelrest F, Schaefer G, et al. Sensorscope: out-of-the-box environmental monitoring. In: Proceedings of International Conference on Information Processing in Sensor Networks, St. Louis, 2008. 332–343
42 Kim Y, Schmid T, Charbiwala Z M, et al. NAWMS: nonintrusive autonomous water monitoring system. In: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, New York, 2008. 309–322
43 Kim S, Pakzad S, Culler D, et al. Health monitoring of civil infrastructures using wireless sensor networks. In:

Proceedings of the 6th International Conference on Information Processing in Sensor Networks, Cambridge, 2007. 254–263
44 Ceriotti M, Mottola L, Picco G P, et al. Monitoring heritage buildings with wireless sensor networks: the torre tquila deployment. In: Proceedings of the 2009 International Conference on Information Processing in Sensor Networks, San Francisco, 2009. 277–288
45 Tolle G, Polastre J, Szewczyk R, et al. A macroscope in the redwoods. In: Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems, San Diego, 2005. 51–63
46 Wang F, Liu J C. Networked wireless sensor data collection: issues, challenges, and approaches. IEEE Commun Surv Tutor, 2011, 13: 673–687
47 Shi J H, Wan J F, Yan H H, et al. A survey of cyber-physical systems. In: Proceedings of International Conference on Wireless Communications and Signal Processing, Nanjing, 2011. 1–6
48 Wahab M H A, Mohd M N H, Hanafi H F, et al. Data pre-processing on web server logs for generalized association rules mining algorithm. World Academy Sci Eng Technol, 2008, 48: 970
49 Nanopoulos A, Manolopoulos Y, Zakrzewicz M, et al. Indexing web access-logs for pattern queries. In: Proceedings of the 4th International Workshop on Web Information and Data Management, Hong Kong, 2002. 63–68
50 Joshi K P, Joshi A, Yesha Y. On using a warehouse to analyze web logs. Distributed Parallel Databases, 2003, 13: 161–180
51 Cho J, Garcia-molina H. Parallel crawlers. In: Proceedings of the 11th International Conference on World Wide Web, Honolulu, 2002. 124–135
52 Castillo C. Effective web crawling. In: Proceedings of ACM SIGIR Forum, New York, 2005. 39: 55–56
53 Choudhary S, Dincturk M E, Mirtaheri S M, et al. Crawling rich internet applications: the state of the art. In: Proceedings of CASCON, Tronto, 2012. 146–160
54 Jain A K, Bolle R, Pankanti S. Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers, 1999
55 Ghani N, Dixit S, Wang T S. On IP-over-WDM integration. IEEE Commun Mag, 2000, 38: 72–84
56 Manchester J, Anderson J, Doshi B, et al. IP over Sonet. IEEE Commun Mag, 1998, 36: 136–142
57 Armstrong J. OFDM for optical communications. J Lightwave Technol, 2009, 27: 189–204
58 Shieh W. OFDM for flexible high-speed optical networks. J Lightwave Technol, 2011, 29: 1560–1577
59 Jinno M, Takara H, Kozicki B. Dynamic optical mesh networks: Drivers, challenges and solutions for the future. In: Proceedings of the 35th European Conference on Optical Communication, Vienna, 2009. 1–4
60 Goutelle M, Gu Y, He E, et al. A survey of transport protocols other than standard TCP. Global Grid Forum, 2004
61 Hoelzle U, Barroso L A. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. 1st ed. Morgan and Claypool Publishers, 2009
62 Cisco. Cisco data center interconnect design and deployment guide. Cisco, 2009
63 Greenberg A, Hamilton J R, Jain N, et al. VL2: a scalable and flexible data center network. In: Proceedings of the ACM SIGCOMM Conference on Data Communication, Barcelona, 2009. 51–62
64 Guo C, Lu G, Li D, et al. BCube: a high performance, server-centric network architecture for modular data centers. SIGCOMM Comput Commun Rev, 2009, 39: 63–74
65 Farrington N, Porter G, Radhakrishnan S, et al. Helios: a hybrid electrical/optical switch architecture for modular data centers. In: Proceedings of the ACM SIGCOMM Conference, New Delhi, 2010. 339–350
66 Abu-Libdeh H, Costa P, Rowstron A, et al. Symbiotic routing in future data centers. ACM SIGCOMM Comput Commun Rev, 2010, 40: 51–62
67 Lam C, Liu H, Koley B, et al. Fiber optic communication technologies: What's needed for datacenter network operations. IEEE Commun Mag, 2010, 48: 32–39
68 Kachris C, Tomkos I. The rise of optical interconnects in data centre networks. In: Proceedings of the 14th International Conference on Transparent Optical Networks, Coventry, 2012. 1–4
69 Wang G, Andersen D G, Kaminsky M, et al. c-Through: part-time optics in data centers. SIGCOMM Comput Commun Rev, 2010, 41: 327–338
70 Ye X, Yin Y, Yoo S B, et al. DOS: A scalable optical switch for datacenters. In: Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems, San Diego, 2010. 24
71 Singla A, Singh A, Ramachandran K, et al. Proteus: a topology malleable data center network. In: Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, Monterey, 2010. 1–6
72 Liboiron-Ladouceur O, Cerutti I, Raponi P G, et al. Energy-efficient design of a scalable optical multiplane interconnection architecture. IEEE J Selected Topics Quantum Electron, 2011, 17: 377–383
73 Kodi A K, Louri A. Energy-efficient and bandwidth-reconfigurable photonic networks for high-performance computing (HPC) systems. IEEE J Selected Topics Quantum Electron, 2011, 17: 384–395

74 Alizadeh M, Greenberg A, Maltz D A, et al. Data center tcp (dctcp). ACM SIGCOMM Comput Commun Rev, 2010, 40: 63–74
75 Vamanan B, Hasan J, Vijaykumar T. Deadline-aware datacenter tcp (d2tcp). ACM SIGCOMM Comput Commun Rev, 2012, 42: 115–126
76 Kohler E, Handley M, Floyd S. Designing DCCP: Congestion control without reliability. ACM SIGCOMM Comput Commun Rev, 2006, 36: 27–38
77 Müller H, Freytag J C. Problems, Methods, and Challenges in Comprehensive Data Cleansing. Professoren des Inst. Für Informatik, 2005
78 Noy N F. Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record, 2004, 33: 65–70
79 Han J W, Kamber M, Pei J. Data mining: concepts and techniques. Morgan Kaufmann, 2006
80 Lenzerini M. Data integration: A theoretical perspective. In: Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Madison, 2002. 233–246
81 Silberschatz A, Korth H F, Sudarshan S. Database System Concepts. New York: McGraw-Hill Hightstown, 1997
82 Cafarella M J, Halevy A, Khoussainova N. Data integration for the relational web. Proc VLDB Endowment, 2009, 2: 1090–1101
83 Kohavi R, Mason L, Parekh R, et al. Lessons and challenges from mining retail e-commerce data. Mach Learn, 2004, 57: 83–113
84 Chen H Q, Ku W S, Wang H X, et al. Leveraging spatio-temporal redundancy for RFID data cleansing. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, Indianapolis, 2010. 51–62
85 Zhao Z, Ng W. A model-based approach for RFID data stream cleansing. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, 2012. 862–871
86 Khoussainova N, Balazinska M, Suciu D. Probabilistic event extraction from rfid data. In: Proceedings of the 24th International Conference on Data Engineering, Cancún, 2008. 1480–1482
87 Herbert K G, Wang J T. Biological data cleaning: a case study. Int J Inf Quality, 2007, 1: 60–82
88 Zhang Y, Callan J, Minka T. Novelty and redundancy detection in adaptive filtering. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, 2002. 81–88
89 Salomon D. Data Compression. Berlin: Springer-Verlag, 2004
90 Dufaux F, Ebrahimi T. Video surveillance using JPEG 2000. Proc SPIE, 2004, 5588: 268–275
91 Symes P D. Digital Video Compression. New York: McGraw-Hill/TAB Electronics, 2004
92 Tsai T H, Lin C Y. Exploring Contextual Redundancy in Improving Object-Based Video Coding for Video Sensor Networks Surveillance. IEEE Trans Multimedia, 2012, 14: 669–682
93 Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, 2002. 269–278
94 Huang Z, Shen H T, Liu J J, et al. Effective data co-reduction for multimedia similarity search. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Athens, 2011. 1021–1032
95 Kamath U, Compton J, Islamaj-Dogan R, et al. An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice site prediction. IEEE/ACM Trans Comput Biol Bioinform, 2012, 9: 1387–1398
96 Leung K, Lee K, Wang J, et al. Data mining on DNA sequences of hepatitis B virus. IEEE/ACM Trans Comput Biol Bioinform, 2011, 8: 428–440
97 Bleiholder J, Naumann F. Data fusion. ACM Comput Surv, 2009, 41: 1–41
98 Günter M. Introducing MapLan to map banking survey data into a time series database. In: Proceedings of the 15th International Conference on Extending Database Technology, Berlin, 2012. 528–533
99 Wang Y, Wei D S, Yin X R, et al. Heterogeneity-aware data regeneration in distributed storage systems. In: Proceedings of IEEE International Conference on Computer Communications, Toronto, 2014. 1878–1886
100 Goda K, Kitsuregawa M. The History of Storage Systems. Proc IEEE, 2012, 100: 1433–1440
101 Strunk J D. Hybrid aggregates: combining SSDs and HDDs in a single storage pool. ACM SIGOPS Operating Syst Rev, 2012, 46: 50–56
102 Soundararajan G, Prabhakaran V, Balakrishnan M, et al. Extending SSD lifetimes with disk-based write caches. In: Proceedings of the 8th USENIX Conference on File and Storage Technologies, San Jose, 2010. 101–114
103 Guerra J, Pucha H, Glider J S, et al. Cost Effective Storage using Extent Based Dynamic Tiering. In: Proceedings of USENIX Conference on File and Storage Technologies, San Jose, 2011. 273–286
104 Troppens U, Erkens R, Mueller-Friedt W, et al. Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, iSCSI, Infiniband and FCoE. New York: John Wiley & Sons, 2011
105 Mell P, Grance T. The NIST definition of cloud computing. NIST Special Publication 800-145, 2011

106 Clark T. Storage Virtualization: Technologies for Simplifying Data Storage and Management. Boston: Addison-Wesley Professional, 2005
107 McKusick M K, Quinlan S. GFS: Evolution on fast-forward. ACM Queue, 2009, 7: 10–20
108 Chaiken R, Jenkins B, Larson P, et al. SCOPE: easy and efficient parallel processing of massive data sets. Proc VLDB Endowment, 2008, 1: 1265–1276
109 Beaver D, Kumar S, Li H C, et al. Finding a needle in Haystack: Facebook's photo storage. In: Proceedings of 9th USENIX Symposium on Operating Systems Design and Implementation, Vancouver, 2010
110 DeCandia G, Hastorun D, Jampani M, et al. Dynamo: Amazon's highly available key-value store. SIGOPS Oper Syst Rev, 2007, 41: 205–220
111 Karger D, Lehman E, Leighton T, et al. Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the World Wide Web. In: Proceedings of the 29th Annual ACM Symposium on Theory of Computing, El Paso, 1997. 654–663
112 Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data. ACM Trans Comput Syst, 2008, 26: 4:1–4:26
113 Burrows M. The Chubby lock service for loosely-coupled distributed systems. In: Proceedings of the 7th Symposium on Operating Systems Design and Implementation, Seattle, 2006. 335–350
114 Lakshman A, Malik P. Cassandra: structured storage system on a P2P network. In: Proceedings of the 28th ACM Symposium on Principles of Distributed Computing, Calgary, 2009. 5
115 Crochford D. The application/json Media Type for JavaScript Object Notation (JSON), RFC 4627, 2006
116 Cooper B F, Ramakrishnan R, Srivastava U, et al. PNUTS: Yahoo!'s hosted data serving platform. Proc VLDB Endowment, 2008, 1: 1277–1288
117 Zhao Y X, Wu J. Dache: A data aware caching for big-data applications using the MapReduce framework. In: Proceedings of IEEE International Conference on Computer Communications, Turin, 2013. 35–39
118 Baker J, Bond C, Corbett J, et al. Megastore: Providing scalable, highly available storage for interactive services. In: Proceedings of Conference on Innovative Data Systems Research, Asilomar, 2011. 223–234
119 Corbett J C, Dean J, Epstein M, et al. Spanner: Google's globally-distributed database. In: Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation, Hollywood, 2013. 251–264
120 Shute J, Oancea M, Ellner S, et al. F1: the fault-tolerant distributed RDBMS supporting google's ad business. In: Proceedings of the 2012 International Conference on Management of Data, Scottsdale, 2012. 777–778
121 Cooper B F, Silberstein A, Tam E, et al. Benchmarking cloud serving systems with YCSB. In: Proceedings of the 1st ACM Symposium on Cloud Computing, Indianapolis, 2010. 143–154
122 Kraska T, Hentschel M, Alonso G, et al. Consistency rationing in the cloud: pay only when it matters. Proc VLDB Endowment, 2009, 2: 253–264
123 Keeton K, Morrey C B III, Soules C A, et al. LazyBase: freshness vs. performance in information management. SIGOPS Oper Syst Rev, 2010, 44: 15–19
124 Florescu D, Kossmann D. Rethinking cost and performance of database systems. SIGMOD Rec, 2009, 38: 43–48
125 Brewer E A. Towards robust distributed systems (abstract). In: Proceedings of the 19th Annual ACM Symposium on Principles of Distributed Computing, Portland, 2000. 7
126 Gilbert S, Lynch N. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. SIGACT News, 2002, 33: 51–59
127 Tanenbaum A S, Steen M v. Distributed Systems: Principles and Paradigms. 2nd ed. Upper Saddle River: Prentice-Hall, Inc., 2006
128 Dagum L, Menon R. OpenMP: an industry standard API for shared-memory programming. IEEE Comput Sci Eng, 1998, 5: 46–55
129 Walker D W, Dongarra J J. MPI: a standard message passing interface. Supercomputer, 1996, 12: 56–68
130 Pike R, Dorward S, Griesemer R, et al. Interpreting the data: Parallel analysis with Sawzall. Sci Program, 2005, 13: 277–298
131 Gates A F, Natkovich O, Chopra S, et al. Building a high-level dataflow system on top of Map-Reduce: the Pig experience. Proc VLDB Endowment, 2009, 2: 1414–1425
132 Thusoo A, Sarma J S, Jain N, et al. Hive: a warehousing solution over a map-reduce framework. Proc VLDB Endowment, 2009, 2: 1626–1629
133 Yang W C, Chen H, Qu Q Y. Research of a MapReduce Model to Process the Traffic Big Data. Appl Mech Mater, 2014, 548: 1853–1856
134 Low Y, Bickson D, Gonzalez J, et al. Distributed GraphLab: A framework for machine learning and data mining in the cloud. Proc VLDB Endowment, 2012, 5: 716–727
135 Roy A, Mihailovic I, Zwaenepoel W. X-stream: edge-centric graph processing using streaming partitions. In: Pro-

ceedings of the 24th ACM Symposium on Operating Systems Principles, Farmington, 2013. 472–488
136 Moretti C, Bulosan J, Thain D, et al. All-pairs: An abstraction for data-intensive cloud computing. In: Proceedings of IEEE International Symposium on Parallel and Distributed Processing, Miami, 2008. 1–11
137 Bu Y, Howe B, Balazinska M, et al. HaLoop: efficient iterative data processing on large clusters. Proc VLDB Endowment, 2010, 3: 285–296
138 Ekanayake J, Li H, Zhang B, et al. Twister: a runtime for iterative MapReduce. In: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, Chicago, 2010. 810–818
139 Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, Berkeley, 2012. 2
140 Bhatotia P, Wieder A, Rodrigues R, et al. Incoop: MapReduce for incremental computations. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, Cascais, 2011. 1–14
141 Peng D, Dabek F. Large-scale incremental processing using distributed transactions and notifications. In: Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation, Vancouver, 2010. 1–15
142 Yan C R, Yang X, Yu Z, et al. IncMR: Incremental data processing based on MapReduce. In: Proceedings of the 5th International Conference on Cloud Computing, Honolulu, 2012. 534–541
143 Olston C, Chiou G, Chitnis L, et al. Nova: continuous pig/hadoop workflows. In: Proceedings of the International Conference on Management of Data, Athens, 2011. 1081–1090
144 Murray D G, Schwarzkopf M, Smowton C, et al. CIEL: a universal execution engine for distributed data-flow computing. In: Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation, Boston, 2011. 9
145 Eschenfelder A H. Data Mining and Knowledge Discovery Handbook. Berlin: Springer-Verlag, 1980
146 Bhatt C A, Kankanhalli M S. Multimedia data mining: state of the art and challenges. Multimed Tools Appl, 2011, 51: 35–76
147 Slavakis K, Giannakis G, Mateos G. Modeling and optimization for big data analytics: (statistical) learning tools for our era of data deluge. IEEE Signal Process Mag, 2014, 31: 18–31
148 Hagerty J, Sallam R L, Richardson J. Magic Quadrant for Business Intelligence Platforms. Gartner Research G00225500, 2012
149 Post F H, Nielson G M, Bonneau G P. Data visualization: the state of the art. Berlin: Springer, 2003
150 Nguyen Q V, Qian Y, Huang M L, et al. TabuVis: a tool for visual analytics multidimensional datasets. Sci China Inf Sci, 2013, 56: 052105
151 Anderson T W. An Introduction to Multivariate Statistical Analysis. 3rd ed. New York: John Wiley & Sons, 2003
152 Wu X, Kumar V, Ross-Quinlan J, et al. Top 10 algorithms in data mining. Knowl Inf Syst, 2007, 14: 1–37
153 Salton G. Automatic text processing. Science, 1970, 168: 335–343
154 Manning C D, Schütze H. Foundations of statistical natural language processing. Cambridge: MIT Press, 1999
155 Ritter A, Clark S, Mausam, et al. Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, 2011. 1524–1534
156 Li Y P, Hu X H, Lin H F, et al. A Framework for Semisupervised Feature Generation and Its Applications in Biomedical Literature Mining. IEEE/ACM Trans Comput Biol Bioinform, 2011, 8: 294–307
157 Blei D M. Probabilistic topic models. ACM Commun, 2012, 55: 77–84
158 Balinsky H, Balinsky A, Simske S J. Automatic text summarization and small-world networks. In: Proceedings of the 11th ACM Symposium on Document Engineering, Mountain View, 2011. 175–184
159 Mishra M, Huan J, Bleik S, et al. Biomedical text categorization with concept graph representations using a controlled vocabulary. In: Proceedings of the 11th International Workshop on Data Mining in Bioinform, Beijing, 2012. 26–32
160 Hu J, Fang L J, Cao Y, et al. Enhancing text clustering by leveraging Wikipedia semantics. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 2008. 179–186
161 Maybury M T. New Directions in Question Answering. Menlo Park: AAAI press, 2004
162 Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inf Retr, 2008, 2: 1–135
163 Chakrabarti S. Data mining for hypertext: A tutorial survey. ACM SIGKDD Explorations, 2000, 1: 1–11
164 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the 7th International Conference on World Wide Web, Brisbane, 1998. 107–117
165 Konopnicki D, Shmueli O. W3QS: A query system for the World Wide Web. In: Proceedings of the 21st International Conference on Very Large Data Bases, San Francisco, 1995. 54–65
166 Chakrabarti S, van den Berg M, Dom B. Focused crawling: a new approach to topic-specific Web resource discovery. Comput Netw, 1999, 31: 1623–1640

167 Xu B, Bu J J, Chen C, et al. An exploration of improving collaborative recommender systems via user-item subgroups. In: Proceedings of the 21st International Conference on World Wide Web, Lyon, 2012. 21–30

168 Ding D, Metze F, Rawat S, et al. Beyond audio and video retrieval: towards multimedia summarization. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, 2012. 2

169 Wang M, Ni B B, Hua X S, et al. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. ACM Comput Surv, 2012, 44: 25

170 Hu W M, Xie N H, Li L, et al. A survey on visual content-based video indexing and retrieval. IEEE Trans Syst Man Cybern Part C-Appl Rev, 2011, 41: 797–819

171 Park Y J, Chang K N. Individual and group behavior-based customer profile model for personalized product recommendation. Expert Syst Appl, 2009, 36: 1932–1939

172 de Campos L M, Fernández-Luna J M, Huete J F, et al. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. Int J Approx Reasoning, 2010, 51: 785–799

173 Ma Z G, Yang Y, Cai Y, et al. Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In: Proceedings of the 20th ACM International Conference on Multimedia, Nara, 2012. 469–478

174 Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 2011. 1046–1054

175 Ninagawa A, Eguchi K. Link prediction using probabilistic group models of network structure. In: Proceedings of the ACM Symposium on Applied Computing, Sierre, 2010. 1115–1116

176 Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorizations. ACM Trans Knowl Discov Data, 2011, 5: 10

177 Leskovec J, Lang K J, Mahoney M. Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web, 2010. 631–640

178 Du N, Wu B, Pei X, et al. Community detection in large-scale social networks. In: Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis, San Jose, 2007. 16–25

179 Garg S, Gupta T, Carlsson N, et al. Evolution of an online social aggregation network: an empirical study. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, Chicago, 2009. 315–321

180 Allamanis M, Scellato S, Mascolo C. Evolution of a location-based online social network: analysis and models. In: Proceedings of the ACM Conference on Internet Measurement Conference, Boston, 2012. 145–158

181 Gong N Z, Xu W C, Huang L, et al. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In: Proceedings of the ACM Conference on Internet Measurement Conference, Boston, 2012. 131–144

182 Zheleva E, Sharara H, Getoor L. Co-evolution of social and affiliation networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009. 1007–1016

183 Tang J, Sun J M, Wang C, et al. Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009. 807–816

184 Li Y H, Chen W, Wang Y J, et al. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, 2013. 657–666

185 Lappas T, Liu K, Terzi E. Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009. 467–476

186 Zhang T, Popescul A, Dom B. Linear prediction models with graph regularization for web-page categorization. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, 2006. 821–826

187 Zhou Y, Cheng H, Yu J X. Graph clustering based on structural/attribute similarities. Proc VLDB Endowment, 2009, 2: 718–729

188 Dai W Y, Chen Y Q, Xue G R, et al. Translated learning: Transfer learning across different feature spaces. In: Proceedings of the Advances in Neural Information Processing Systems, Vancouver, 2008. 353–360

189 Shao L, Zhu F, Li X L. Transfer learning for visual categorization: A survey. IEEE Trans Neural Netw Learn Syst, 2014, DOI: 10.1109/TNNLS.2014.2330900

190 Wu E, Diao Y L, Rizvi S. High-performance complex event processing over streams. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, Chicago, 2006. 407–418

191 Garg M K, Kim D J, Turaga D S, et al. Multimodal analysis of body sensor network data streams for real-time healthcare. In: Proceedings of the International Conference on Multimedia Information Retrieval, Philadelphia, 2010. 469–478

192 Park Y, Ghosh J. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium,

Miami, 2012. 445–454
193 Chen X W, Lin X T. Big data deep learning: challenges and perspectives. IEEE Access J, 2014, 2: 514–525
194 Shao L, Wu D, Li X L. Learning deep and wide: a spectral method for learning deep networks. IEEE Trans Neural Netw Learn Syst, 2014, 25: 2303–2308
195 Hinton G E. Learning multiple layers of representation. Trends Cogn Sci, 2007, 11: 428–434
196 Baah G K, Gray A, Harrold M J. On-line anomaly detection of deployed software: a statistical machine learning approach. In: Proceedings of the 3rd International Workshop on Software Quality Assurance, Portland, 2006. 70–77
197 Moeng M, Melhem R. Applying statistical machine learning to multicore voltage & frequency scaling. In: Proceedings of the 7th ACM International Conference on Computing Frontiers, Bertinoro, 2010. 277–286
198 Gaber M M, Zaslavsky A, Krishnaswamy S. Mining data streams: a review. SIGMOD Rec, 2005, 34: 18–26
199 Lei Z. Large-scale web image search and pattern mining. Sci Sin Inform, 2013, 43: 1641–1653 [张磊. 大规模互联网图像检索与模式挖掘. 中国科学: 信息科学, 2013, 43: 1641–1653]
200 Verykios V S, Bertino E, Fovino I N, et al. State-of-the-art in privacy preserving data mining. SIGMOD Rec, 2004, 33: 50–57
201 van der Aalst W. Process mining: overview and opportunities. ACM Trans Manag Inf Syst, 2012, 3: 7
202 Wang G Z, Wen C K, Yan B H, et al. Topic hypergraph: hierarchical visualization of thematic structures in long documents. Sci China Inf Sci, 2013, 56: 052111
203 Morrison D A. Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis. Wiley Interdiscip Rev Data Mining Knowl Discov, 2014, 4: 296–312
204 Deng M, Liu Q, Wang J, et al. A general method of spatio-temporal clustering analysis. Sci China Inf Sci, 2013, 56: 102315
205 Pal S K, Talwar V, Mitra P. Web mining in soft computing framework: relevance, state of the art and future directions. IEEE Trans Neural Netw, 2002, 13: 1163–1177
206 Li X L, Lin S, Yan S C, et al. Discriminant locally linear embedding with high-order tensor data. IEEE Trans Syst Man Cybern Part B-Cybern, 2008, 38: 342–352
207 Li X L, Pang Y W. Deterministic column-based matrix decomposition. IEEE Trans Knowl Data Eng, 2010, 22: 145–149
208 Li X L, Pang Y W, Yuan Y. L1-norm-based 2DPCA. IEEE Trans Syst Man Cybern Part B-Cybern, 2010, 40: 1170–1175
209 Shao L, Jones S, Li X L. Efficient Search and Localization of Human Actions in Video Databases. IEEE Trans Circuit Syst Video Technol, 2014, 24: 504–512
210 Chen L, Xu D, Tsang I W-H, et al. Spectral embedded hashing for scalable image retrieval. IEEE Trans Cybern, 2014, 44: 1180–1190
211 Song J K, Yang Y, Li X L, et al. Robust hashing with local models for approximate similarity search. IEEE Trans Cybern, 2014, 44: 1225–1236
212 Dong Y S, Tao D C, Li X L, et al. Texture classification and retrieval using shearlets and linear regression. IEEE Trans Cybern, 2014, DOI: 10.1109/TCYB.2014.2326059
213 Jiang Y G, Zeng X H, Ye G N, et al. Columbia-UCF TRECVID2010 multimedia event detection: combining multiple modalities, contextual concepts, and temporal matching. In: Proceedings of NIST TRECVid Workshop, 2010. 2:6
214 Hirsch J E. An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA, 2005, 102: 16569–16572
215 Watts D J. Six Degrees: The Science of A Connected Age. New York: WW Norton, 2004
216 Aggarwal C C. An Introduction to Social Network Data Analytics. Berlin: Springer, 2011
217 Ji R R, Gao Y, Liu W, et al. When location meets social multimedia: a survey on vision-based recognition and mining for geo-social multimedia analytics. ACM Trans Intell Syst Technol, 2014, 5
218 Rabbath M, Sandhaus P, Boll S. Multimedia retrieval in social networks for photo book creation. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Trento, 2011. 72
219 Shridhar S, Lakhanpuria M, Charak A, et al. SNAIR: a framework for personalised recommendations based on social network analysis. In: Proceedings of the 5th International Workshop on Location-based Social Networks, Redondo Beach, 2012. 55–61
220 Maniu S, Cautis B. Taagle: efficient, personalized search in collaborative tagging networks. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Scottsdale, 2012. 661–664
221 Hu H, Huang J, Zhao H, et al. Social TV analytics: a novel paradigm to transform TV watching experience. In: Proceedings of the 5th ACM Multimedia Systems Conference, Singapore, 2014. 172–175
222 Hu H, Wen Y G, Luan H, et al. Towards multi-screen social TV with geo-aware social sense. IEEE Multimedia, 2014, 21: 10–19

223 Zhang H Z, Zhang Z Y, Dai H Y. Gossip-based information spreading in mobile networks. IEEE Trans Wirel Commun, 2013, 12: 5918–5928
224 Zhang H Z, Zhang Z Y, Dai H Y. Mobile conductance and gossip-based information spreading in mobile networks. In: Proceedings of IEEE International Symposium on Information Theory, 2013. 824–828
225 Zhang H Z, Huang Y F, Zhang Z Y, et al. Mobile conductance in sparse networks and mobility-connectivity tradeoff. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), 2014
226 Cisco. Cisco visual networking index: global mobile data traffic forecast update, 2012-2017. Cisco, 2013
227 Han J, Lee J G, Gonzalez H, et al. Mining massive RFID, trajectory, and traffic data sets. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 2
228 Nambiar R O, Poess M. The making of TPC-DS. In: Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, 2006. 1049–1058
229 Patil S, Polte M, Ren K, et al. Ycsb++: benchmarking and performance debugging advanced features in scalable table stores. In: Proceedings of the 2nd ACM Symposium on Cloud Computing, Cascais, 2011. 14
230 Rabl T, Gómez-Villamor S, Sadoghi M, et al. Solving big data challenges for enterprise application performance management. Proc VLDB Endowment, 2012, 5: 1724–1735
231 Pavlo A, Paulson E, Rasin A, et al. A comparison of approaches to large-scale data analysis. In: Proceedings of the 35th SIGMOD International Conference on Management of Data, Providence, 2009. 165–178
232 Chen Y P, Ganapathi A, Griffith R, et al. The case for evaluating MapReduce performance using workload suites. In: Proceedings of IEEE 19th International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems, Singapore, 2011. 390–399
233 Chen Y P, Alspaugh S, Katz R. Interactive analytical processing in big data systems: a cross-industry study of MapReduce workloads. Proc VLDB Endowment, 2012, 5: 1802–1813
234 He Q, Zhuang F, Zeng L, et al. PDMiner: cloud computing based parallel and distributed data mining toolkit platform. Sci Sin Inform, 2014, 44: 871–885 [何清, 庄福振, 曾立, 等. PDMiner: 基于云计算的并行分布式数据挖掘工具平台. 中国科学: 信息科学, 2014, 44: 871–885]
235 Ghazal A, Rabl T, Hu M, et al. Bigbench: Towards an industry standard benchmark for big data analytics. In: Proceedings of the International Conference on Management of Data, New York, 2013. 1197–1208
236 Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools and good practices. In: Proceedings of the 6th International Conference on Contemporary Computing, Noida, 2013. 404–409
237 Sagiroglu S, Sinanc D. Big data: a review. In: Proceedings of International Conference on Collaboration Technologies and Systems, San Diego, 2013. 42–47
238 Hu H, Wen Y G, Chua T-S, et al. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access J, 2014, 2: 652–687
239 Hu H, Wen Y G, Gao Y, et al. Towards SDN-enabled big-data platform for social TV analytics. IEEE Network, in press
240 Lu R X, Zhu H, Liu X M, et al. Toward efficient and privacy-preserving computing in big data era. IEEE Network, 2014, 28: 46–50

A survey on big data systems

LI XueLong[1*] & GONG HaiGang[2]

1 *Center for OPTical IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an* 710119*, China*;
2 *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu* 611731*, China*
*E-mail: xuelong_li@opt.ac.cn

Abstract With the development of the latest technologies, a large amount of data is generated from various domains (such as optical observation and control, healthcare, sensors, user-generated data, Internet and financial companies, supply chain systems, etc.) during the last two decades. (A more appropriate description could be "infinite" data, e.g., in the application of optical observation and control, data are continuously generated, creating a data disaster.) The term of big data is coined to capture the profound meaning of this emerging trend.

Compared with traditional data, big data exhibits some unique characteristics besides the sheer volume, such as commonly un-structured data and more real-time analysis requirements. The development of big data calls for new system architectures for data storage and large-scale data processing mechanisms. In this paper, we present a literature survey of big data analytics. Firstly, the definition of big data and big data challenges are presented. Secondly, a systematic framework to decompose big data system into four sequential modules, namely data generation, data acquisition, data storage, and data analytics, which form the value chain for big data, is proposed. A detailed survey of numerous approaches and mechanisms related to each module, from research and industry communities is discussed. Finally, some evaluation benchmarks and potential scientific problems in big data systems are outlined.

Keywords big data, data acquisition, data storage, data processing, data analytics

LI XueLong is currently a full professor at Xi'an Institute of Optics and Precision Mechanics of Chinese Academy of Sciences, and he is also the director of Center for OPTical IMagery Analysis and Learning (OPTIMAL). He is a fellow of the OSA, SPIE, IEEE, and IAPR.

GONG HaiGang is currently an associate professor at University of Electronic Science and Technology of China, and he is also an associate director of Southwest Technical Committee on Networks and Information system.