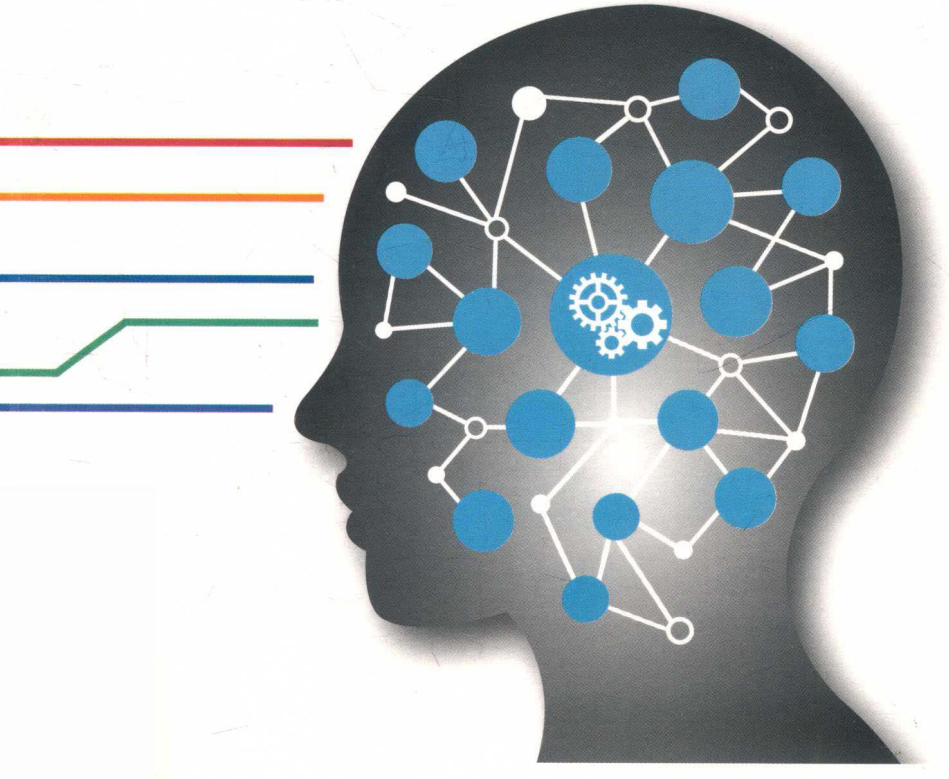
大数据巨量分析与机器学习的整合与开发

◎ 韦鹏程 冉 维 段 昂 著



电子科技大学出版社

需要全本请访问 www.ertongbook.com

大数据巨量分析与机器学习的整合与开发

◎ 韦鹏程　冉　维　段　昂　著

电子科技大学出版社

图书在版编目(CIP)数据

大数据巨量分析与机器学习的整合与开发/韦鹏程，冉维，段昂著. -- 成都：电子科技大学出版社，2017.5
ISBN 978-7-5647-4512-7

Ⅰ.①大… Ⅱ.①韦…②冉…③段… Ⅲ.①数据处理②机器学习 Ⅳ.①TP274②TP181

中国版本图书馆CIP数据核字（2017）第101564号

大数据巨量分析与机器学习的整合与开发

韦鹏程 冉 维 段 昂 著

策划编辑 李述娜

责任编辑 熊晶晶

出版发行 电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼 邮编 610051

主 页 www.uestcp.com.cn

服务电话 028-83203399

邮购电话 028-83201495

印 刷 北京一鑫印务有限责任公司

成品尺寸 170mm × 240mm

印 张 20.5

字 数 358千字

版 次 2017年5月第一版

印 次 2017年5月第一次印刷

书 号 ISBN 978-7-5647-4512-7

定 价 68.00元

安徽理工大学 藏書 图書馆

版权所有，侵权必究

前　言

以深度学习为代表的机器学习是当前最接近人类大脑的智能学习方法和认知过程，充分借鉴了人脑的多分层结构、神经元的连接交互、分布式稀疏存储和表征、信息的逐层分析处理机制，自适应、自学习的强大并行信息处理能力，在语音、图像识别等方面取得了突破性进展，在诸多应用领域取得巨大商业成功。随着产业界数据量的爆炸式增长，大数据概念受到越来越多的关注。由于大数据的海量、复杂多样、变化快的特性，对于大数据环境下的应用问题，传统的在小数据上的机器学习算法很多已不再适用。因此，研究大数据环境下地机器学习算法成为学术界和产业界共同关注的话题。

本书一共9章，从概念到应用，由浅入深，全面深入地探析了大数据巨量分析与机器学习的理论及应用。

其中，第1~3章是理论部分，第1章主要介绍了大数据的发展历程、国内外现状、大数据的概念、系统架构、分类、基准、面临的挑战和科学问题；第2章主要介绍了机器学习发展简史、发展现状、策略与方法、经典算法；第3章主要介绍了大数据机器学习系统的研究背景、现状、技术特征及主要研究问题、相关技术、总体架构等。

第4~8章是应用部分，其中第4章从互联网、商业、农业、医疗卫生、城市规划及其他领域的应用做了系统介绍，并选取当前主流的四个应用场景：标签系统、数据自助营销平台、基于Mahout的个性化推荐系统、图计算与社会网络，介绍如何实现数据驱动，让数据“自动”流转于各个环节。

最后，第9章从哲学的角度探索了机器学习的未来发展方向，包括机器学习的前沿科学基础、可能实现途径分析、算法及其知识发现功能，并对机器学习的前景做了展望。

本专著由重庆市第二师范学院韦鹏程教授、冉维和段昂三位教师完成，并得到重庆市交互式电子教育工程技术研究中心和重庆市第二师范学院交互式电子产品协同创新中心支持，在此表示感谢！

目　录

第 1 章　大数据发展综述 1

1.1　大数据的发展历程 1
1.2　大数据国内外现状 5
1.3　大数据的概念 7
1.4　大数据系统架构 13
1.5　大数据分析分类 34
1.6　大数据基准 41
1.7　大数据面临的挑战 42
1.8　大数据科学问题 51

第 2 章　机器学习研究进展 54

2.1　机器学习发展简史 55
2.2　机器学习发展现状 58
2.3　机器学习策略与方法 60
2.4　机器学习的经典算法 63

第 3 章　大数据机器学习系统 70

3.1　大数据机器学习系统研究背景 70
3.2　大数据机器学习研究现状 73
3.3　大数据机器学习系统的技术特征及主要研究问题 77
3.4　大数据机器学习相关技术 81
3.5　大数据机器学习平台总体架构 101

第 4 章　大数据巨量分析与机器学习的应用领域 107

4.1　互联网领域 108
4.2　商业领域 116

需要全本请访问 www.ertongbook.com

4.3 工业领域 122
4.4 农业信息化建设领域 127
4.5 医疗行业 133
4.6 城市规划与建筑工程 141
4.7 其他研究领域 144

第 5 章 标 签 系 统 149

5.1 认识标签系统 149
5.2 标签系统的设计 151
5.3 标签系统的实现 154

第 6 章 数据自助营销平台 163

6.1 数据自助营销平台的价值所在 163
6.2 数据自助营销平台的实现原则 168
6.3 数据自助营销平台的场景实例 175

第 7 章 基于 Mahout 的个性化推荐系统 182

7.1 Mahout 的推荐引擎 182
7.2 规模与效率 190
7.3 实现一个推荐系统 199

第 8 章 图计算与社会网络 206

8.1 社会网络和属性图 206
8.2 Spark Graph X 与 Neo4j 208
8.3 使用 Spark Graph X 和 Neo4j 处理社会网络 211

第 9 章 机器学习的哲学探索 225

9.1 机器学习哲学前沿科学基础 226
9.2 机器学习的可能实现途径分析 248
9.3 机器学习算法及其知识发现功能 272
9.4 机器学习展望 299

参考文献 301

第 1 章　大数据发展综述

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，2011 年产生和复制的数据量超过 1.8 ZB 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（Economists，经济学家）和公共媒体（New York Times，美国国家公共广播电台）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；Nature 和 Science 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。

1.1　大数据的发展历程

1.1.1　国际发展历程

大数据的历史最早可以追溯到 18 世纪 80 年代，1887—1890 年美国统计学家赫尔曼·霍尔瑞斯为了统计 1890 年的人口普查数据，发明了一台电动器来读取卡片上的洞数，该设备让美国用一年时间就完成了原本耗时 8 年的人口普查活动，由此在全球范围内引发了数据处理的新纪元。

1944 年，卫斯理大学图书馆员弗莱蒙特·雷德对大数据时代的到来进行了预见。他出版了《学者与研究型图书馆的未来》一书，在书中他估计美国高校图书馆的规模每 16 年就翻一番。

1961 年德里克·普赖斯出版了《巴比伦以来的科学》，在这本书中，普赖斯通过观察科学期刊和论文的增长规律来研究科学知识的增长。他得出以下结论：新期刊的数量以指数方式增长而不是以线性方式增长，每 15 年翻一番，每 50 年以 10 为指数倍进行增长。普赖斯将其称之为“指数增长规律”。

1980 年 4 月，I.A. 特詹姆斯兰德在第四届美国电气和电子工程师协会（IEEE）“大规模存储系统专题研讨会”上做了一个报告，题为《我们该何去何从？》。在报告中，他指出所有数据正在被无选择地保存以避免错失有价值的信息。

1981 年，匈牙利中央统计办公室开始实施了一项调查国家信息产业的研究项目，包括以比特为单位计量信息量。这项研究一直持续至今。

1986 年 7 月，哈尔·B. 贝克尔在《数据通信》上发表了《用户真的能够以今天或者明天的速度吸收数据吗？》一文，预计数据记录密度将大幅增长。

1993 年，匈牙利中央统计办公室首席科学家伊斯特万·迪恩斯编制了一本国家信息账户的标准体系手册。

1997 年 10 月，迈克尔·考克斯和大卫·埃尔斯沃思在第八届美国电气和电子工程师协会（IEEE）关于可视化的会议论文集中发表了《为外存模型可视化而应用控制程序请求页面调度》的文章。这是在美国计算机学会的数字图书馆中大数据发展历程综述第一篇使用“大数据”这一术语的文章。

1999 年 8 月，史蒂夫·布赖森、大卫·肯怀特、迈克尔·考克斯、大卫·埃尔斯沃思以及罗伯特·海门斯在《美国计算机协会通讯》上发表了《千兆字节数据集的实时性可视化探索》一文。这是《美国计算机协会通讯》上第一篇使用“大数据”这一术语的文章。

2001 年，美国一家在信息技术研究领域具有权威地位的咨询公司 Gartner 首次开发了大数据模型。

2001 年 2 月，梅塔集团分析师道格·莱尼发布了一份研究报告，题为《3D 数据管理：控制数据容量、处理速度及数据种类》。十年后，3D 作为定义大数据的三个维度而被广泛接受。

2005 年，Hadoop 项目诞生。Hadoop 是由多个软件产品组成的一个生态系统，这些软件产品共同实现全面功能和灵活的大数据分析。

2007 年，著名图灵奖获得者 Jim Gray 提出，“数据密集型科学发现”（Data-Intensive Scientific Discovery）将成为科学研究的第四范式。

2008 年年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟（Computing Community Consortium）发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不再局限于数据处理的机器，此组织可以说是最早提出大数据概念的机构。

2008 年，在 Google 成立 10 周年之际，著名的《自然》杂志出版了一期专刊，专门讨论未来的大数据处理相关的一系列技术问题和挑战，其中就提出了“Big Data”的概念。

大约从 2009 年开始，“大数据”逐渐成为互联网信息技术行业的流行词汇。

2009 年，印度政府建立了用于身份识别管理的生物识别数据库，联合国全球脉冲项目已研究了对如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病暴发之类的问题。

2009 年年中，美国政府通过启动 Data.gov 网站的方式进一步开放了数据的大门，这个网站向公众提供各种各样的政府数据，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。库克尔在报告中提到："世界上有着无法想象的巨量数字信息，并以极快的速度增长。科学家和计算机工程师已经为这个现象创造了一个新词汇：'大数据'。"库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2010 年 12 月，美国总统办公室下属的科学技术顾问委员会（PCAST）和信息技术顾问委员会（PITAC）向奥巴马和国会提交了一份《规划数字化未来》的战略报告，把大数据收集和使用的工作提升到体现国家意志的战略高度。

2011 年 2 月，IBM 的沃森超级计算机每秒可扫描并分析 4 TB（约 2 亿页文字量）的数据量，并在美国著名智力竞赛电视节目"Jeopardy"（危险边缘）上击败两名人类选手而夺冠。

后来纽约时报认为这一刻为一个"大数据计算的胜利"。2011 年 5 月，全球知名咨询公司麦肯锡的全球研究院（MGI）发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，这项研究估计 2010 年所有的公司存储了 7.4 EB 新产生的数据，消费者存储了 6.8 EB 新数据。大数据开始备受关注，这也是专业机构第一次全方面地介绍和展望大数据。

2012 年 1 月，瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响（Big Data，Big Impact）》宣称，数据已经成为一种新的经济资产类别。

2012 年，美国总统选举中，那些精于数字计算的数据挖掘团队把传统的投票放在一边不用，而是利用"大数据"来规划这次选举将如何进行。如利用房产记录、选举记录甚至是期刊的订阅注册等来预测人们对候选人的看法、这些看法是否能被改变，以及为此要采取怎样的措施等。

2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。

2012 年 3 月 22 日，美国政府宣布投资 2 亿美元于大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭，在次日的电话会议中，政府对数据的定义“未来的新石油”，大数据技术领域的竞争，事关国家安全和未来。

2012 年 4 月，美国软件公司 Splunk 于 19 日在纳斯达克成功上市，成为第一家上市的大数据处理公司。Splunk 成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快大数据布局。

2012 年 7 月，联合国在纽约发布了一本关于大数据政务的白皮书《大数据促发展：挑战与机遇》，全球大数据的研究和发展进入了前所未有的高潮。这本白皮书总结了各国政府如何利用大数据响应社会需求，指导经济运行，更好地为人民服务，并建议成员国建立“脉搏实验室”（Pulse Labs），挖掘大数据的潜在价值。

2014 年 4 月，世界经济论坛以“大数据的回报与风险”为主题发布了《全球信息技术报告（第 13 版）》。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要，接下来将对数据保密和网络管制等议题展开积极讨论。

2014 年 5 月，美国白宫发布了 2014 年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。

由于大数据技术的特点和重要性，目前国内外已经出现了“数据科学”的概念，即数据处理技术将成为一个与计算科学并列的新的科学领域。

1.1.2 国内发展状况

为了紧跟全球大数据技术发展的浪潮，我国政府、学术界和工业界对大数据也予以了高度的关注。

2011 年 12 月，工信部发布的物联网十二五规划上，信息处理技术作为 4 项关键技术创新工程之一被提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。

2012 年 7 月，为挖掘大数据的价值，阿里巴巴集团在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。

随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。阿里巴巴也是最早提出通过

数据进行企业数据化运营的企业。

为了推动我国大数据技术的研究发展，2012 年中国计算机学会（CCF）发起组织了 CCF 大数据专家委员会，CCF 专家委员会还特别成立了一个“大数据技术发展战略报告”撰写组，并已撰写发布了《2013 年中国大数据技术与产业发展白皮书》。

2013 年 4 月 14 日和 21 日，央视著名“对话”节目邀请了《大数据时代——生活、工作与思维的大变革》作者维克托·迈尔·舍恩伯格，以及美国大数据存储技术公司 LSI 总裁阿比分别做客“对话”节目，做了两期大数据专题谈话节目“谁在引爆大数据”“谁在掘金大数据”，国家央视媒体对大数据的关注和宣传体现了大数据技术已经成为国家和社会普遍关注的焦点。

国内的学术界和工业界也都迅速行动，广泛开展大数据技术的研究和开发。

2013 年以来，国家自然科学基金、973 计划、核高基、863 等重大研究计划都已经把大数据研究列为重大的研究课题。

清华信息学院、国家实验室也成立了数据科学学院，并于 2014 年 12 月 22 日举办了“大数据论坛——数据科学与技术”，对大数据发展战略和各大数据专项进行了探讨。

1.2　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。

1.2.1　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，欧盟委员会也呼吁各成员国

需要全本请访问 www.ertongbook.com

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，美国麻省理工学院（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 将 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 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，找到发掘共享、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 Net App 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。

1.2.2 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。

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。

1.3　大数据的概念

1.3.1　大数据的定义

大数据（big data），又称海量资料，是由数量巨大、结构复杂、类型众多的数据构成的数据集合，其所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为能帮助政府机构和企业进行管理、决策的资讯。

有关大数据的定义有多种，其中一个狭义的定义为：大数据是指不能装载进计算机内存储的数据。尽管这是一个非正式的定义，但易理解，因为每台电脑都有一个大到不能装载进内存的数据集。李国杰等对大数据的定义为：一般意义上，大数据是指无法在可容忍的时间内用传统 IT 技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合。

此外，不少文献对大数据进行了定义，其中三种定义较为重要。

1. 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："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，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 和 Microsoft 的研究者们仍继续使用"3Vs"模型描述大数据。

2. 比较定义（Comparative definition）

2011 年，Mc Kinsey 公司的研究报告将大数据定义为"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。

3. 体系定义（Architectural definition）

美国国家标准和技术研究院 NIST 则认为"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。然而对于大数据定义，要达成共识非常困难。

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 1-1 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。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、You Tube 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。

表 1-1　大数据和传统数据比较

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

1.3.2　大数据处理方式

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。

1. 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），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 和 Kafka。流处理方式用于在线应用，通常工作在秒或毫秒级别。

2. 批处理

在批处理方式中，数据首先被存储，随后被分析。Map Reduce 是非常重要的批处理模型。Map Reduce 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。Map Reduce 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，Map Reduce 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 1-2 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。

表 1-2　批处理和流处理比较

	Stream processing	Batch processing
Input	Stream of new data or updates	Data chunks
Data size	Infinite or unknow in advance	Know and finite
Storage	Not store or store non-trial portion in memory	Store
Hardware	Typice single limited smount 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	Wed mining，sensor networke，traffic monitoring	Widely adopted in almost every domain

1.3.3　大数据系统观点

1. 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。按照产业界广为接受的系统工程方法，典型的大数据系统可分解为 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 1-2 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。

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 1-1 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 1-1 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。

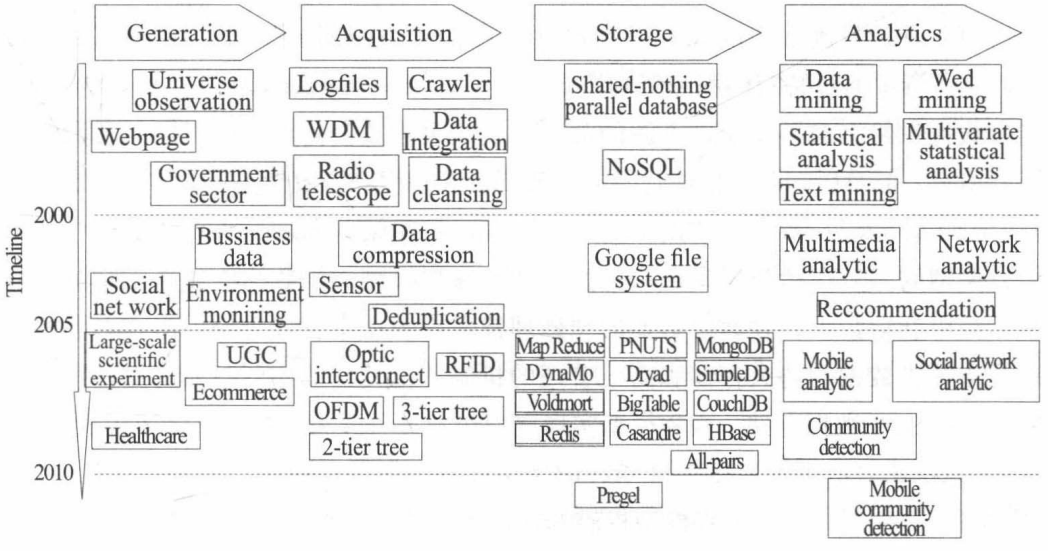


图 1-1　大数据价值链及其技术地图

需要全本请访问 www.ertongbook.com