BIG DATA

大数据概论

主　编　崔奇明
副主编　喻洪辉　金绍玉　赵永彬
　　梁　凯　张紫荣　叶　丹

東北大學出版社
Northeastern University Press

大数据概论

主　编：崔奇明

副主编：喻洪辉　金绍玉　赵永彬　梁　凯
张紫荣　叶　丹

编　委：（按姓氏笔划）：

于　鑫　王　玲　王乃玉　王诺笛
方　洁　许　斌　刘　晖　孙道军
陈国栋　张晓冬　张　荻　张兆贵
李伟明　周雨田　胡家瑛　高秀平
高传杰　常　健　黄甦龙　崔舒婷

东北大学出版社

·沈　阳·

图书在版编目（CIP）数据

大数据概论 / 崔奇明主编. — 沈阳：东北大学出版社，2016.9
ISBN 978-7-5517-1416-7

Ⅰ. ①大… Ⅱ. ①崔… Ⅲ. ①数据处理—概论 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 234801 号

出 版 者：东北大学出版社
地址：沈阳市和平区文化路三号巷 11 号
邮编：110819
电话：024－83687331（市场部） 83680267（社务部）
传真：024－83680180（市场部） 83687332（社务部）
E-mail：neuph@neupress.com
http：//www.neupress.com
印 刷 者：沈阳航空发动机研究所印刷厂
发 行 者：东北大学出版社
幅面尺寸：178mm×233mm
印 张：15
字 数：254 千字
出版时间：2016 年 9 月第 1 版
印刷时间：2016 年 9 月第 1 次印刷
责任编辑：李 佳 责任校对：汪彤彤
封面设计：刘江旸 责任出版：唐敏志

ISBN 978-7-5517-1416-7 定 价：48.00 元

前言

近年来，大数据的浪潮冲击着世界上的每个角落。在大数据的驱动下，人类社会在加速发展。大数据已经成为企业界、科技界和政府等关注的热点。2012 年 3 月，美国奥巴马政府宣布投资 2 亿美元启动“大数据研究和发展计划”，这是继 1993 年美国宣布“信息高速公路”计划后的又一次重大科技发展部署。在美国宣布大数据计划后，世界其他国家以及各大商业公司也对大数据给予了极大关注。大数据涉及的范围非常广泛，它也将持续地影响人们生活的各个方面。

2015 年 9 月 5 日，《国务院关于印发促进大数据发展行动纲要的通知》（国发〔2015〕50 号）正式发布，在全社会引起广泛影响。《促进大数据发展的行动纲要》是到目前为止我国促进大数据发展的第一份权威性、系统性文件，从国家大数据发展战略全局的高度，提出了我国大数据发展的顶层设计，是指导我国未来大数据发展的纲领性文件。

中国计算机学会（the China Computer Federation，CCF）作为国内计算机及相关领域的专业学术团体及时开展了相应的工作。自 2012 年成立以来，CCF 大数据专家委员会已经连续两年发布了《中国大数据技术与产业发展报告》。该报告汇聚大数据专家委员会一百多位专家的知识和智慧，为业界梳理大数据应用现状及发展趋势，为政府制定产业发展政策提供建议；同时，探讨大数据研究面临的科学问题和技术挑战，为科研机构和科研人员提供参考指南。

全球实施互联网，万众应用大数据。在上述背景下，编者学习、整理、分类、总结了国内外一些学者关于大数据理论及技术的研究成果、论述及部分公司在大数据应用方面的实践，并结合编者在大数据应用方面最基本的实践，旨在从整体上对大数据给出描述，为大数据知识的学习及应

用者提供参考，借以为推动大数据知识在我国的传播与应用贡献自己的绵薄之力。

本书由国网辽宁省电力有限公司鞍山供电公司崔奇明、金绍玉、梁凯、张紫荣、叶丹、高秀平、王玲、高传杰、张晓冬、许斌、黄甦龙、王诺笛、刘晖、孙道军，国网辽宁省电力有限公司信息通信分公司喻洪辉、赵永彬，国网技术学院王乃玉、陈国栋、胡家瑛，山东电力高等专科学校张兆贵，国网辽宁省电力有限公司丹东供电公司周雨田、于鑫，留美学生崔舒婷，国网四川省电力公司信息通信公司常健，国网陕西省电力公司华阴市供电分公司方洁，国网湖北省电力公司技术培训中心张荻，国网浙江宁波市鄞州区供电公司李伟明编写。其中第1、2、3章由崔奇明、喻洪辉、赵永彬、周雨田、王乃玉、于鑫、胡家瑛、孙道军、张兆贵、方洁编写，第4、5、6、7、8章由崔奇明、金绍玉、梁凯、张紫荣、叶丹、高秀平、王玲、高传杰、张晓冬、许斌、黄甦龙、王诺笛、刘晖、崔舒婷编写，第9章由喻洪辉、赵永彬编写，第10章由崔奇明、许斌编写，第11章由崔奇明、常健、陈国栋、张荻、李伟明编写，全书由崔奇明负责修改定稿。

由于编者水平有限，书中难免存在错误或漏洞，敬请读者给予批评指正。本书在编写过程中，得到了许多领导、专家、同事的帮助，在此一并表示感谢。

编　者

2016年5月

目录

第 1 章　概　述 …… 1

1.1　大数据基本概念 …… 6
1.2　大数据基本特征 …… 9
1.3　大数据基本技术 …… 10
1.4　大数据基本处理 …… 11
1.5　大数据基本分析 …… 12
1.6　大数据应用简介 …… 13

第 2 章　机遇与挑战及思考 …… 17

2.1　大数据带来的新机遇和新挑战 …… 17
2.2　大数据与机器学习 …… 21
2.3　大数据处理模式与处理器体系结构 …… 26
2.4　计算与存储融合体系结构及内存计算 …… 33
2.5　大数据与数据管理及数据密集型计算 …… 36
2.6　大数据与可视分析 …… 38
2.7　大数据时代软件技术及大数据系统引擎技术面临的挑战 …… 40

第 3 章　大数据存储与云计算及物联网 …… 44

3.1　大数据存储渊源与大数据技术发展新目标 …… 44
3.2　大数据与数据流及算法 …… 48
3.3　大数据与云端存储 …… 51
3.4　大数据分析的可伸缩云与数据密集型计算的应用 …… 52
3.5　大数据与物联网：海 - 云计算模型 …… 55

第 4 章　大数据与软件工程及软件技术 …… 57

4.1　大数据与软件工程新思维 …… 57
4.2　移动应用大数据与软件工程 …… 58
4.3　大数据与软件工程教育 …… 59
4.4　大数据与软件工程关键技术 …… 60
4.5　基于大数据的软件设计模型 …… 64

第 5 章　大数据时代的研究与应用 …… 67

5.1　大数据与典型学科领域 …… 67
5.2　大数据与生物多样性信息学及生物医学 …… 71
5.3　大数据与用户理解及情感计算 …… 73
5.4　大数据与文化及计算机专业教育 …… 78
5.5　大数据与金融及实时竞价 …… 79
5.6　大数据与复杂工业系统 …… 81
5.7　大数据与电信业务 …… 82
5.8　新生儿重症监护室中的大数据 …… 87
5.9　淘宝自主研发的分布式数据库 OceanBase 简介 …… 88
5.10　携程大数据平台简介 …… 89
5.11　精硕科技——Hadoop 在广告监测技术上的实践 …… 91
5.12　武汉绿网——大数据在网络服务中的应用及网络演进 …… 92

第 6 章　大数据与数据治理、挖掘、预测及分析 …… 95

6.1　大数据：公司治理观点的价值、风险和成本 …… 96
6.2　转换大数据为集体意识 …… 98
6.3　大数据可能的意想不到的结果 …… 99
6.4　使用大数据方法的业务过程分析 …… 100
6.5　谷歌流感趋势预测的启示：大数据分析中的陷阱及小数据 …… 101
6.6　大数据与预测分析 …… 102
6.7　大数据与数据挖掘 …… 103
6.8　大数据与数据分析及认知计算 …… 106
6.9　淘宝真实数据的统计分析 …… 109

第 7 章　大数据与数据科学及数据驱动 …… 112

7.1　我们为什么需要数据科学 …… 112
7.2　大数据研究的科学价值 …… 113
7.3　数据科学家与领域专家 …… 114
7.4　对于数据驱动创新公共政策的考虑 …… 116

第 8 章　大数据与政府 …… 118

8.1　政府与企业中大数据的应用比较 …… 118
8.2　大数据和政府转型 …… 119
8.3　城市计算与大数据 …… 121
8.4　从数字脚印到城市计算 …… 122
8.5　大数据下的灾难行为分析和城市应急管理 …… 123

第 9 章　中国电力大数据 …… 125

9.1　《中国电力大数据发展白皮书》简介 …… 125
9.2　电网大数据研究及应用简介 …… 126
9.3　国家电网公司信息通信新技术推动智能电网和“一强三优”现代公司创新发展行动计划简介 …… 127

第 10 章　中国计算机学会大数据专家委员会关于大数据热点问题及发展趋势分析简介 …… 129

10.1　2012 年 CCF 给出的大数据热点问题 …… 129
10.2　2013—2015 年大数据十大趋势预测对比 …… 130
10.3　2016 年大数据发展趋势预测 …… 132

第 11 章　虚拟机上 Hadoop 等的基本安装实践 …… 135

11.1　Hadoop 简介 …… 135
11.2　安装 VM ware-workstation 10.0 及 RedHat Linux 6.2 …… 139
11.3　伪分布式模式下安装 Hadoop …… 145
11.4　完全分布式模式下安装 Hadoop …… 159
11.5　安装 HBase …… 187

11.6 安装 Hive 及 MySQL …… 191
11.7 Windows 平台下安装 Cygwin 及 Hadoop …… 203
11.8 Hadoop 进程启动过程初步分析 …… 211
11.9 安装及启动过程常见问题处理方法 …… 214
11.10 主要文件配置内容示例 …… 217

参考文献 …… 222

第1章 概 述

近年来，大数据的浪潮冲击着世界上的每个角落。在大数据的驱动下，人类社会在加速发展。大数据已经成为企业界、科技界和政府等关注的热点。2012年3月22日，奥巴马宣布美国政府投资2亿美元启动“大数据研究和发展计划”（Big Data Research and Development Initiative），这是继1993年美国宣布“信息高速公路”计划后的又一次重大科技发展部署[1]。美国政府的此项计划旨在提高和改进人们从海量和复杂的数据中获取知识的能力，进而加速美国在科学与工程领域发明的步伐，增强国家安全，共同提高收集、储存、管理、分析和共享海量数据所需核心技术的先进性，扩大大数据技术开发和应用所需人才的供给。对于大数据，美国政府认为它是“未来的新石油”，并将“大数据研究”上升为国家意志。美国政府的大数据计划及发展部署对未来的科技与经济发展必将带来深远的影响。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制也将成为国家间和企业间新的争夺焦点[2]。

在美国宣布大数据计划后，世界其他国家及各大商业公司也对大数据给予了极大关注。文献［2］认为，目前推动大数据研究的动力主要是大型互联网公司、政府机构及创业企业。巨大的经济利益驱使大企业不断扩大数据处理规模。例如，自2005年以来，IBM投资160亿美元进行了30次与大数据有关的收购，促使其业绩稳定高速增长。

2015年9月5日，《国务院关于印发促进大数据发展行动纲要的通知》（国发〔2015〕50号）正式发布，在全社会引起广泛影响[3]。《促进大数据发展行动纲要》是到目前为止我国促进大数据发展的第一份权威性、系统性文件，从国家大数据发展战略全局的高度，提出了我国大数据

发展的顶层设计，是指导我国未来大数据发展的纲领性文件。

IBM 执行总裁罗睿兰认为："数据将成为一切行业当中决定胜负的根本因素，最终数据将成为人类至关重要的自然资源"[4]。IBM 积极提出了 IBM 的"大数据平台"架构，如图 1－1 所示。该平台的四大核心能力包括 Hadoop 系统、流计算（Stream Computing）、数据仓库（Data Warehouse）和信息整合与治理（Information Integration and Governance）。

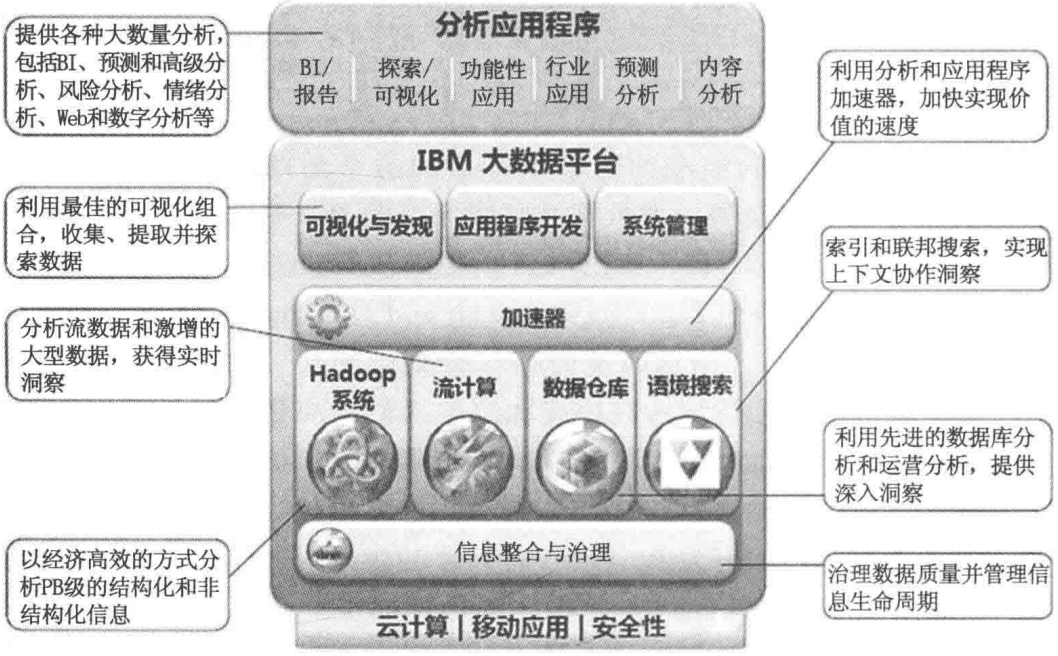

图 1－1　IBM 的"大数据平台"架构（来源：文献［4］）

大数据是什么，如何系统地认知大数据，许多专家、学者从不同的角度进行了探讨。文献［4］认为，应从三个层面来展开：① 理论。理论是认知的必经途径，也是被广泛认同和传播的基线。应从大数据的特征定义理解行业对大数据的整体描绘和定性，从对大数据价值的探讨来深入解析大数据的珍贵所在，从对大数据的现在和未来去洞悉大数据的发展趋势，从大数据隐私这个特别而重要的视角审视人和数据之间的长久博弈。② 技术。技术是大数据价值体现的手段和前进的基石。应从云计算、分布式处理技术、存储技术和感知技术的发展来了解大数据从采集、处理、存储到形成结果的整个过程。③ 实践。实践是大数据的最终价值体现。应从互联网的大数据、政府的大数据、企业的大数据和个人的大数据四个方面来描绘大数据已经展现的美好景象及即将实现的蓝图，如图 1－2 所示。

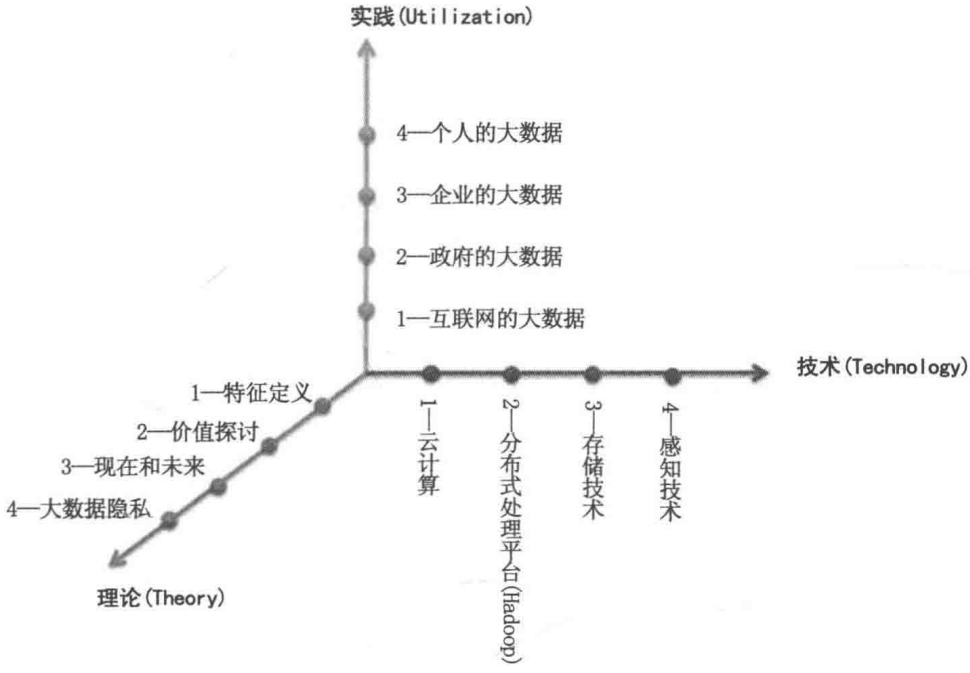

图 1－2　大数据认知的三个层面（来源：文献［4］）

关于大数据技术的发展，从 Gartner 公司发布的 2012 年技术超周期图（见图 1－3）来看，Gartner 公司将各类大数据技术的技术成熟度划分为 5 个阶段：技术触发期、过热期、幻灭期、复苏期和生产力成熟期[5]。这 5 个阶段分别采取如下定义[6]。

① 技术触发期（Technology Trigger）：产生之初，被业界和媒体广泛关注，曝光率直线上升。

② 过热期（Peak of Inflated Expectations）：概念炒作达到顶峰，媒体对新技术产生了大量不切实际的期望。

③ 幻灭期（Trough of Disillusionment）：泡沫破灭，媒体态度骤变，从原先的新技术支持者变为声讨者，概念逐渐被人淡忘。

④ 复苏期（Slope of Enlightenment）：虽然该技术已经很少被曝光，但并未在业内完全消失，不少企业在慢慢推动该技术走向成熟，技术本身的优势和局限性已经被业内人士逐渐了解。

⑤ 生产力成熟期（Plateau of Productivity）：已经达到成熟的新技术找到了市场定位，虽然不像先期媒体期望的那样具有颠覆性，但却实实在在地改变着人们的生活。

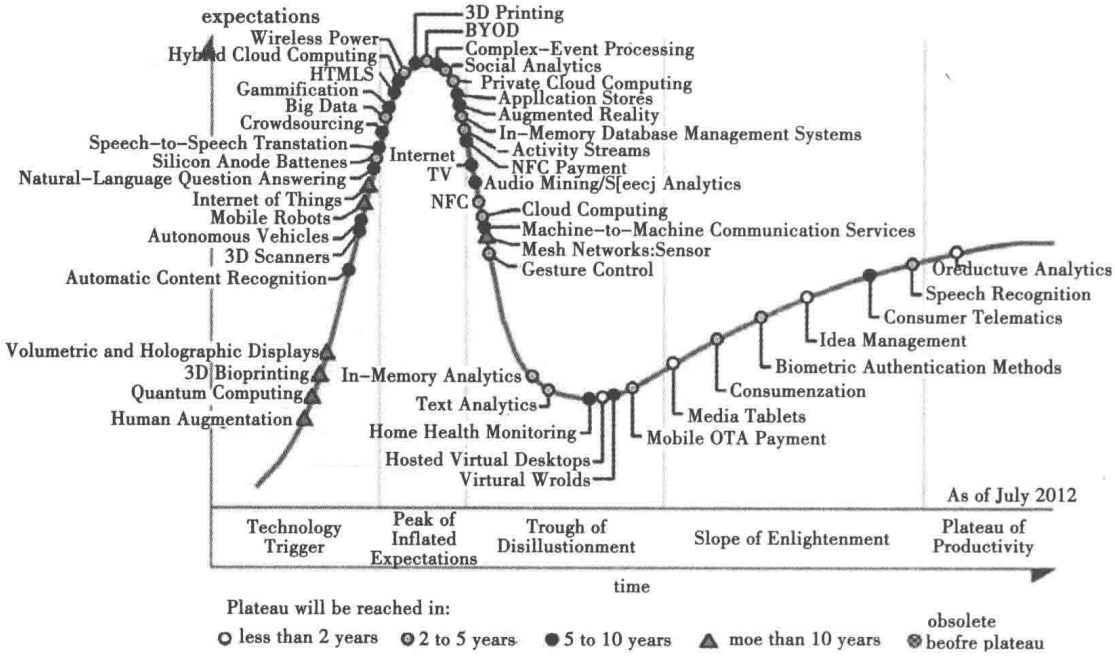

图 1-3　2012 年技术超周期图（来源：中国计算机学会通讯 2012 年第 9 期）

按照 Gartner 公司的分析，大数据技术目前正处在技术的诱发期和过热期，进入主流应用还需要时间。如何抓住这一时机，是摆在学术界、工业界及政府面前的机会与挑战。文献［7］认为，大数据技术的发展趋势包括：① 各个行业对数据的收集能力越来越强，而且数据获取朝着越来越自动化、多元化的方向发展。云计算使大数据存储和管理成为可能。② 智能技术的发展（包括机器学习和人工智能）使数据产生价值。其核心技术是大规模计算和构建更复杂的模型，并使其可以在不同层面取代人工操作。③ 行业创新可以让大数据在更大范围内得以应用。大数据技术的发展将使社会更加智能化。

中国计算机学会（CCF）作为国内计算机及相关领域的专业学术团体及时开展了相应的大量工作。2012 年 6 月 9 日，中国计算机学会常务理事会决定成立大数据专家委员会（CCF Task Force on Big Data，CCF TFBD）。2012 年 11 月 30 日，在中国计算机学会主办的“2012 Hadoop 与大数据技术大会”上，CCF 大数据专家委员会发布了《大数据热点问题和 2013 年发展趋势分析》的报告。在报告中，大数据专家委员会对大数据热点问题与发展趋势进行了预测分析。2013 年 12 月，由中国计算机学会主办，CCF 大数据专家委员会承办，中国人民大学、中科院计算所、南京大学共

同协办的第一届 CCF 大数据学术会议在北京成功召开，会议期间，CCF 大数据专家委员会发布了《2013 中国大数据技术与产业发展白皮书》。2014 年 12 月，由中国计算机学会主办，CCF 大数据专家委员会承办，中科院计算所与 CSDN 共同协办，以推进大数据科研、应用与产业发展为主旨的 2014 中国大数据技术大会（Big Data Technology Conference 2014，BDTC 2014）暨第二届 CCF 大数据学术会议在北京盛大开幕。大会现场正式发布了《中国大数据技术与产业发展白皮书（2014）》和《2015 大数据十大发展趋势预测》报告。第三届 CCF 大数据学术会议（CCF Big Data 2015）于 2015 年 10 月在合肥举行。几次会议上，多位业内专家在大数据典型应用现状、大数据技术体系现状、大数据 IT 产业链与生态环境、大数据人才资源、大数据发展趋势与建议等方面进行了深入思考，为我国未来大数据产业的发展奠定了坚实的理论基础，为我国企业制定大数据战略规划提供了极具价值的参考建议。

文献［4］总结了大数据的现在表现并展望了其未来，指出大数据在当下的表现主要包括以下几方面。

① 大数据帮助政府实现市场经济调控、公共卫生安全防范、灾难预警和社会舆论监督。

② 大数据帮助城市预防犯罪，实现智慧交通，提升应急能力。

③ 大数据帮助医疗机构建立患者的疾病风险跟踪机制，帮助医药企业提升药品的临床使用效果，帮助艾滋病研究机构为患者提供定制的药物。

④ 大数据帮助航空公司节省运营成本，帮助电信企业实现售后服务质量提升，帮助保险企业识别欺诈骗保行为，帮助快递公司监测分析运输车辆的故障险情以提前预警维修，帮助电力公司有效识别并预警即将发生故障的设备。

⑤ 大数据帮助电商公司向用户推荐商品和服务，帮助旅游网站为旅游者提供心仪的旅游路线，帮助二手市场的买卖双方找到最合适的交易目标，帮助用户找到最合适的商品购买时期、商家和最优惠的价格。

⑥ 大数据帮助企业提升营销的针对性，降低物流和库存的成本，减少投资的风险，以及帮助企业提升广告投放的精准度。

⑦ 大数据帮助娱乐行业预测歌手、歌曲、电影、电视剧的受欢迎程度，并为投资者分析评估拍一部电影需要投入多少钱才最合适。

⑧ 大数据帮助社交网站提供更准确的好友推荐，为用户提供更精准的

企业招聘信息，向用户推荐可能喜欢的游戏以及适合购买的商品。

……

文献［4］认为，未来大数据的身影应该无处不在。在互联网、云计算、移动互联网发展的同时，当物联网发展到达一定规模时，借助条形码、二维码、RFID 等就能够唯一标识产品。而传感器、可穿戴设备、智能感知、视频采集、增强现实等技术可实现实时的信息采集和分析，这些感知或采集的数据能够支撑智慧城市、智慧交通、智慧能源、智慧医疗和智慧环保的理念需要。

1.1 大数据基本概念

目前，大数据没有一个统一的定义。麦肯锡全球研究院（McKinsey Global Institute，MGI）在《大数据：下一个创新、竞争和生产力的前沿》报告中给出大数据的一个描述是："大数据是指无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合"。维基百科中关于大数据的定义为："大数据是指利用常用软件工具来获取、管理和处理数据所耗时间超过可容忍时间的数据集"。文献［8］给出了一些大数据的定义或描述。例如，亚马逊网络服务（AWS）、大数据科学家 John Rauser 提到一个简单的定义："大数据就是任何超过了一台计算机处理能力的庞大数据量"；研发小组对大数据的定义："大数据是最大的宣传技术、是最时髦的技术，当这种现象出现时，定义就变得很混乱"；Kelly 说："大数据是可能不包含所有的信息，但我觉得大部分是正确的。对大数据的一部分认知在于，它是如此之大，分析它需要多个工作负载，这是 AWS 的定义。当你的技术达到极限时，也就是数据的极限"。

人类正处在一个高度数字化的社会中，那么，全世界的数据究竟有多少？据 IDC（International Data Corporation）公司统计，2011 年，全球被创建和被复制的数据总量为 1.8ZB（10^{21}），其中，75% 来自个人（主要是图片、视频和音乐），远远超过人类有史以来所有印刷材料的数据总量 200PB（10^{15}）。文献［2］给出了较具体的数据总量：谷歌公司通过大规模集群和 MapReduce 软件，每个月处理的数据量超过 400PB；百度每天大

约要处理几十 PB 数据；Facebook 注册用户超过 10 亿，每月上传的照片超过 10 亿张，每天生成 300TB 以上的日志数据；淘宝网会员超过 3.7 亿，在线商品超过 8.8 亿，每天交易数千万笔，产生约 20TB 数据；雅虎的总存储容量超过 100PB；传感网和物联网的蓬勃发展是大数据的又一推动力，各个城市的视频监控每时每刻都在采集巨量的流媒体数据；工业设备的监控也是大数据的重要来源。数据以更快的速度产生与消费[9]，一个简单说明如图 1 -4。上述的数据总量是个什么概念？大家可以先看看存储容量单位的定义[10]，如表 1 -1 所示。对于 1.8ZB 的大数据，文献［10］给出了一个描述：假如用 1TB 2.5 寸硬盘分别保存 1.8ZB 数据，这些硬盘叠加起来的高度超过 1.7 万千米，接近地球周长的一半。据 IDC 公司预测，到 2020 年，用户的数据量将达到 35ZB，全球数据量预测如图 1 -5 及图 1 -6[11]所示。

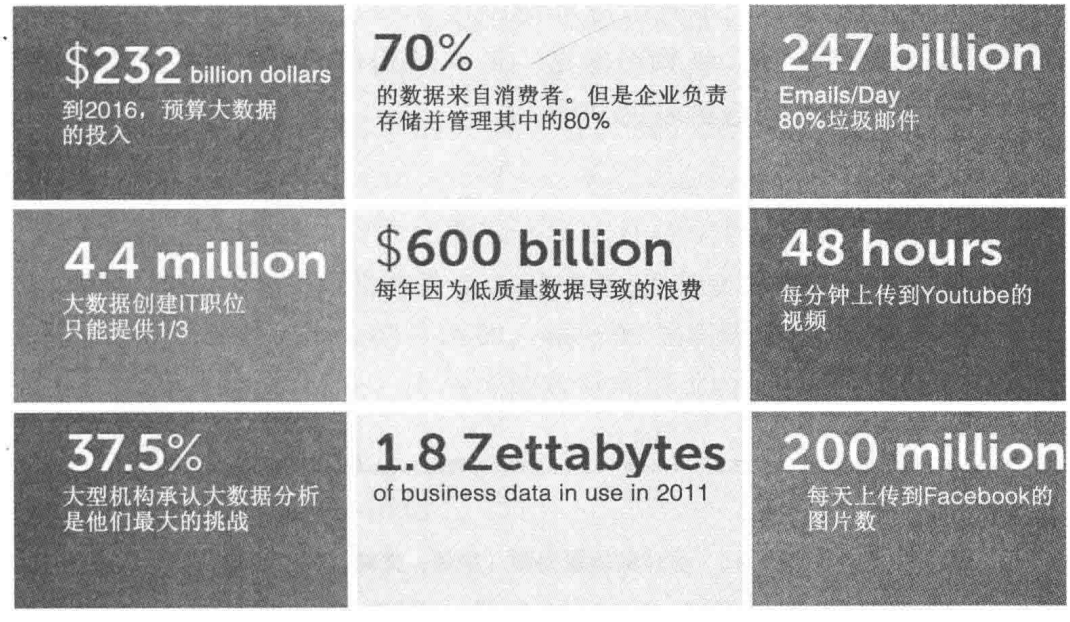

图 1 -4　数据产生与消费（来源：文献［9］）

表 1 -1　存储容量单位的定义（来源：中国计算机学会通讯 2012 年第 6 期）

单位	定义	字节数（2 进制）	字节数（10 进制）
Kilobyte（千）	1024 - Byte	2^{10}	10^3
Megabyte（兆）	1024 - Kilobyte	2^{20}	10^6

续表 1-1

单位	定义	字节数（2 进制）	字节数（10 进制）
Gigabyte（吉）	1024 - Megabyte	2^{30}	10^{9}
Terabyte（太）	1024 - Gigabyte	2^{40}	10^{12}
Petabyte（拍）	1024 - Terabyte	2^{50}	10^{15}
Exabyte（艾）	1024 - Petabyte	2^{60}	10^{18}
Zettabyte（泽）	1024 - Exabyte	2^{70}	10^{21}
Yottabyte（尧）	1024 - Zetabyte	2^{80}	10^{24}

一个很自然的问题就是，这些数据是从哪里来的？文献［10］给出了说明，这些数据主要来自如下几方面。

（1）视频。视频图像需要以每秒几十帧的速度持续记录运动着的物体，即使经过压缩，1 个小时的标准清晰度视频也在 GB 量级，高清晰度视频的数据量就更大了。视频图像是一种非结构化的数据，它与其他各种非结构化数据贡献了总数据量的 90% 以上。

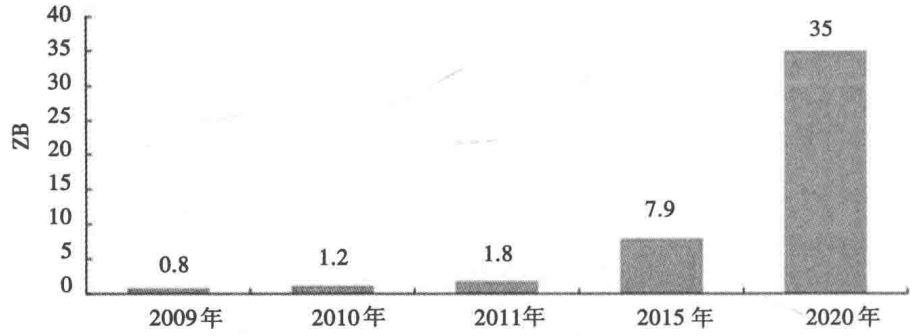

图 1-5　全球数据量预测（来源：文献［6］）

（2）图片和照片。截至 2011 年 9 月，用户向 Facebook 上传了超过 1400 亿张照片。如果平均每张照片大小为 1MB，这些照片的总数据量就是 140PB。图片和照片也是非结构化数据。

（3）日志。网站日志记录了用户对网站的访问，电信日志记录了用户拨打和接听电话的信息。以电信日志为例，假设为 5 亿用户，每个用户每天呼出呼入一共 10 次，每条日志为 400Byte，并且要保存 5 年，那么总数据量是：5 亿 ×10 ×365 ×400Byte ×5 =3.65PB。日志也是非结构化数据。