

大数据 技术与应用

任友理◎著



西北工业大学出版社

大数据技术与应用

任友理 著

西北工业大学出版社
西安

【内容简介】 本书针对计算机、信息管理和其他相关专业学生的发展需求,系统、全面地阐述了大数据技术与应用的基本知识和技能,详细介绍了大数据概念及影响、Hadoop 基础、HDFS 基本应用、分布式数据库 HBase、NoSQL 数据库、云数据库、MapReduce 应用开发、大数据分析、大数据应用等内容,具有极强的系统性、可读性和实用性。

本书是为高等学校相关专业开设大数据相关课程而全新设计编写的主教材,具有丰富实践特色,也可供有一定实践经验的软件开发人员和管理人员参考或作为继续教育的教材使用。

图书在版编目(CIP)数据

大数据技术与应用/任友理著. —西安:西北工业大学出版社,2018.9

ISBN 978-7-5612-6268-9

I. ①大… II. ①任… III. ①数据处理—高等学校—教材 IV. ①TP274

中国版本图书馆 CIP 数据核字(2018)第 213854 号

策划编辑:季 强

责任编辑:张 潼 郭 睿

出版发行:西北工业大学出版社

通信地址:西安市友谊西路 127 号

邮编:710072

电 话:(029)88493844 88491757

网 址:www.nwpup.com

印刷者:西安真色彩设计印务有限公司

开 本:710mm×1000mm 1/16

印 张:16

字 数:282 千字

版 次:2019 年 5 月第 1 版

2019 年 5 月第 1 次印刷

定 价:65.00 元

前 言

由于互联网和信息技术的快速发展,大数据(Big Data)越来越引起人们的关注,已经引发自互联网、云计算之后 IT 行业的又一大颠覆性的技术革命。面对信息的激流,多元化数据的涌现,大数据已经为个人生活、企业经营,甚至国家与社会的发展都带来了机遇和挑战,成为 IT 信息产业中最具潜力的蓝海。人们用大数据来描述和定义信息爆炸时代产生的海量数据,并命名与之相关的技术发展与创新。云计算主要为数据资产提供了保管、访问的场所和渠道,而数据才是真正有价值的资产。企业内部的经营信息、互联网世界中的商品物流信息,以及互联网世界中的人与人交互信息、位置信息等,其数量将远远超越现有企业 IT 架构和基础设施的承载能力,实时性要求也将大大超越现有的计算能力。如何盘活这些数据资产,使其为国家治理、企业决策乃至个人生活服务,是大数据的核心议题,也是云计算内在的灵魂和必然的发展方向。

大数据技术与应用是一门理论性和实践性都很强的课程。在长期的教学实践中,笔者体会到,坚持“因材施教”的重要原则,把实践环节与理论教学相融合,用实践教学促进理论知识的学习,是有效改善教学效果和提高教学水平的重要方法之一。本书的主要特色是理论联系实际,结合一系列了解和熟悉大数据技术与应用的学习和实践活动,把大数据的相关概念、基础知识和技术技巧融入实践当中,使学生保持浓厚的学习热情,加深对大数据技术的认识、理解和掌握。

笔者在写作过程中参考了大量文献著作,在此一并向原作者表示感谢。由于水平有限,书中不妥之处,恳请业界同仁和广大读者批评指正!

编 者

2018 年 6 月

目 录

第 1 章 大数据概述	1
1.1 大数据时代	1
1.2 大数据概念	8
1.3 大数据的影响	11
1.4 大数据的应用	17
1.5 大数据关键技术	18
1.6 大数据计算模式	19
1.7 大数据产业	21
1.8 大数据与云计算、物联网	22
第 2 章 Hadoop 基础	33
2.1 Hadoop 概述	33
2.2 Hadoop 原理	34
2.3 Hadoop 的安装与配置	39
2.4 Hadoop 生态系统简介	56
第 3 章 HDFS 基本应用	58
3.1 实战命令行接口	58
3.2 实战 Java 接口	61
3.3 数据流	65
第 4 章 分布式数据库 HBase	70
4.1 HBase 简介	70
4.2 HBase 接口	71
4.3 安装 HBase 集群	71
4.4 HBase Shell	76
4.5 HBase API	78
4.6 HBase 综合实例	81

第 5 章 NoSQL 数据库	87
5.1 NoSQL 简介	87
5.2 NoSQL 兴起的原因	88
5.3 NoSQL 与关系数据库的比较	91
5.4 NoSQL 的四大类型	93
5.5 NoSQL 的三大基石	98
5.6 从 NoSQL 到 NewSQL 数据库	103
第 6 章 云数据库	106
6.1 云数据库概述	106
6.2 云数据库产品	113
6.3 云数据库系统架构	117
6.4 云数据库实践	123
第 7 章 MapReduce 应用开发	130
7.1 配置 Hadoop MapReduce 开发环境	130
7.2 编写和运行第一个 MapReduce 程序前的准备	134
7.3 MapReduce 应用案例	144
第 8 章 大数据分析	156
8.1 数据分析的演变	156
8.2 大数据分析平台	160
8.3 大数据与数据挖掘	166
8.4 数据挖掘的高级分析方法	170
8.5 数据挖掘项目的生命周期	176
8.6 大数据可视化	180
第 9 章 大数据在互联网领域的应用	188
9.1 推荐系统概述	188
9.2 协同过滤	192
9.3 协同过滤实践	197
第 10 章 大数据在生物医学领域的应用	202
10.1 流行病预测	202
10.2 智慧医疗	206
10.3 生物信息学	207
10.4 案例：基于大数据的综合健康服务平台	209

第 11 章 大数据的其他应用	215
11.1 大数据在物流领域的应用	215
11.2 大数据在城市管理中的应用	220
11.3 大数据在金融行业的应用	224
11.4 大数据在汽车行业的应用	227
11.5 大数据在零售行业的应用	228
11.6 大数据在餐饮行业中的应用	231
11.7 大数据在电信行业的应用	234
11.8 大数据在能源行业的应用	235
11.9 大数据在体育和娱乐领域的应用	236
11.10 大数据在安全领域的应用	240
11.11 大数据在政治领域中的应用	242
11.12 大数据在日常生活中的应用	243
参考文献	247

► 第 1 章

大数据概述

大数据时代悄然来临，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人民生活的方方面面。在全球范围内，世界各国政府均高度重视大数据技术的研究和产业发展，纷纷把大数据上升为国家战略加以重点推进。企业和学术机构纷纷加大技术、资金和人员投入力度，加强对大数据关键技术的研发与应用，以期在“第三次信息化浪潮”中占得先机、引领市场。大数据已经不是“镜中花、水中月”，它的影响力和作用力正迅速触及社会的每个角落，所到之处，或是颠覆，或是提升，都让人们深切感受到了大数据实实在在的威力。

对于一个国家而言，能否紧紧抓住大数据发展机遇，快速形成核心技术和应用参与新一轮的全球化竞争，将直接决定未来若干年世界范围内各国科技力量博弈的格局。大数据专业人才的培养是新一轮科技较量的基础，高等院校承担着大数据人才培养的重任，因此，各高等院校非常重视大数据课程的开设，大数据课程已经成为计算机科学与技术专业的重要核心课程。

1.1 大数据时代

第三次信息化浪潮涌动，大数据时代全面开启。人类社会信息科技的发展为大数据时代的到来提供了技术支撑，而数据产生方式的变革是促进大数据时代到来至关重要的因素。

1.1.1 第三次信息化浪潮

根据 IBM 前首席执行官郭士纳的观点，IT 领域每隔 15 年就会迎来一次重大变革（见表 1-1）。1980 年前后，个人计算机（PC）开始普及，使得计算机走入企业和千家万户，大大提高了社会生产力，也使人类迎来了第一次信息化浪潮，Intel、IBM、苹果、微软、联想等企业是这个时期的标志。随后，在 1995 年前后，人类开始全面进入互联网时代，互联网的普及把世界变成“地球村”，每个人都可以自由徜徉于信息的海洋，由此，人类迎来了第二次信息化浪潮，这个时期也缔造了雅虎、谷歌、阿里巴巴、百度等互联网巨头。时隔 15 年，在 2010 年前后，云计算、大数据、物联网的快速发展，拉开了第三次信息化浪潮的大幕，大数据时代已经到来，也必将涌现出一批新的市场标杆企业。

表 1-1 三次信息化浪潮

信息化浪潮	发生时间	标 志	解决的问题	代表企业
第一次浪潮	1980 年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1985	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010 年前后	物联网、云计算和大数据	信息爆炸	亚马逊、谷歌、IBM、VMWare、pALAMTIR、HORTONWORKS、LOUDER、阿里云等

1.1.2 信息科技为大数据时代提供技术支撑

信息科技需要解决信息存储、信息传输和信息处理 3 个核心问题，人类社会在信息科技领域的不断进步，为大数据时代的到来提供了技术支撑。

1. 存储设备容量不断增加

数据被存储在磁盘、磁带、光盘、闪存等各种类型的存储介质中，随着科学技术的不断进步，存储设备的制造工艺不断升级，容量大幅增加，速度不断提升，价格却在不断下降（见图 1-1）。

早期的存储设备容量小、价格高、体积大，例如，IBM 在 1956 年生产的一个早期的商业硬盘，容量只有 5MB，不仅价格昂贵，而且体积有一个冰箱那么大（见图 1-2）。相反，今天容量为 1TB 的硬盘，大小只有 3.5 英寸（约 8.89cm），读写速度达到 200MB/s，价格仅为 400 元左右廉价、高性能的硬盘存储设备，不仅提供了海量的存储空间，同时大大降低了数据存储成本。

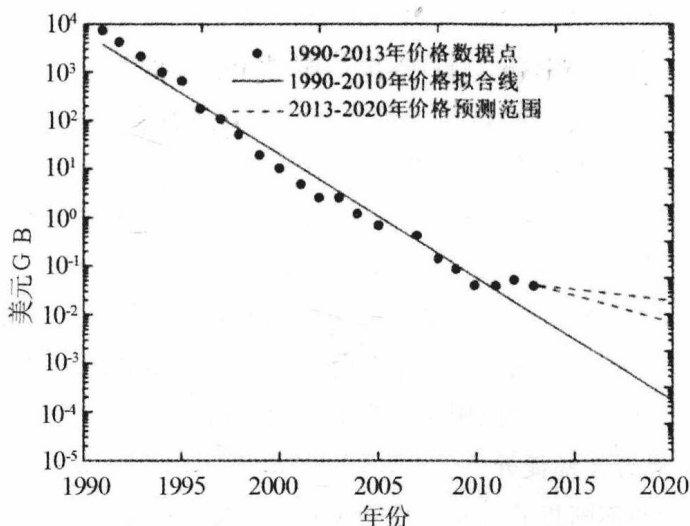


图 1-1 存储设备的价格随时间变化的情况



图 1-2 IBM 在 1956 年生产的一个早期的商业硬盘

与此同时，以闪存为代表的新型存储介质也开始得到大规模的普及和应用。闪存是一种新兴的半导体存储器，从 1989 年诞生第一款闪存产品开始，闪存技术不断获得新的突破，并逐渐在计算机存储产品市场中确立了自己的重要地位。闪存是一种非易失性存储器，即使发生断电也不会丢失数据；因此，可以作为永久性存储设备，它具有体积小、质量轻、能耗低、抗振性好等优良特性。

闪存芯片可以被封装制作成 SD 卡、U 盘和固态硬盘等各种存储产品，SD 卡和 U 盘主要用于个人数据存储，固态硬盘则越来越多地应用于企业级数据存储。一个 32GB 的 SD 卡，体积只有 $24\text{mm} \times 32\text{mm} \times 2.1\text{mm}$ ，质量只有 0.5g。以前 7 200 r/min 的硬盘，一秒钟读写次数只有 100 IOPS (Input/Output Operations Per Second)，传输

速率只有 50MB/s，而现在基于闪存的固态硬盘，每秒钟读写次数有几万甚至更高的 IOPS，访问延迟只有几十微秒，允许以更快的速度读写数据。

总体而言，数据量和存储设备容量二者之间是相辅相成、互相促进的。一方面，随着数据的不断产生，需要存储的数据量不断增加，对存储设备的容量提出了更高的要求，促使存储设备生产商制造更大容量的产品满足市场需求；另一方面，更大容量的存储设备进一步加快了数据量增长的速度，在存储设备价格高企的年代，由于考虑到成本问题，一些不必要或当前不能明显体现价值的数据往往会被丢弃。但是，随着单位存储空间价格的不断降低，人们开始倾向于把更多的数据保存起来，以期在未来某个时刻可以用更先进的数据分析工具从中挖掘价值。

2. CPU 处理能力大幅提升

CPU 处理速度的不断提升也是促使数据量不断增加的重要因素。性能不断提升的 CPU，大大提高了处理数据的能力，使得人们可以更快地处理不断累积的海量数据。从 20 世纪 80 年代至今，CPU 的制造工艺不断提升，晶体管数量不断增加（见图 1-3），运行频率不断提高，核心（Core）数量逐渐增多，而同等价格所能获得的 CPU 处理能力也呈几何级数上升。在 30 多年里，CPU 的处理速度已经从 10 MHz 提高到 3.6 GHz，在 2013 年之前的很长一段时期，CPU 处理速度的增加一直遵循“摩尔定律”，性能每隔 18 个月提高一倍，价格下降一半。

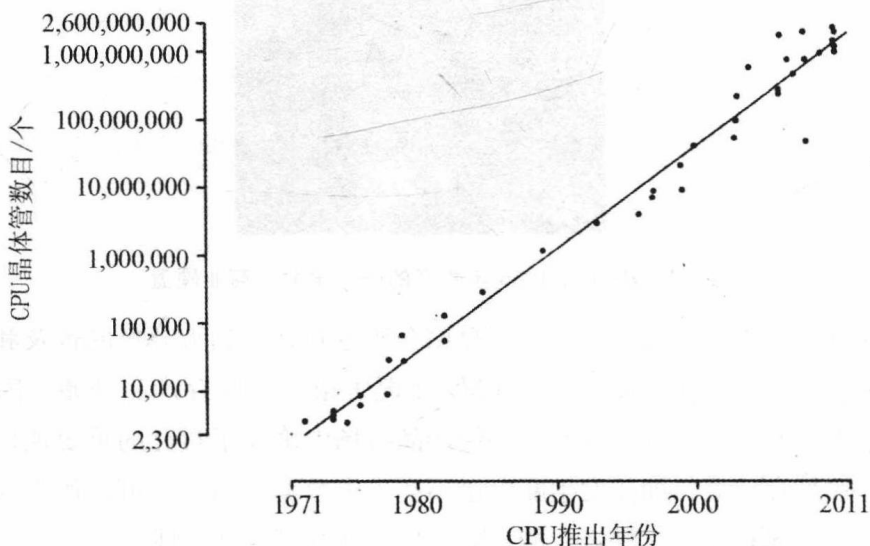


图 1-3 CPU 晶体管数目随时间变化的情况

3. 网络带宽不断增加

1977 年，世界上第一条光纤通信系统在美国芝加哥市投入商用，该光纤数据传输速率为 45 Mb/s，从此，人类社会的信息传输速度不断被刷新。进入 21 世纪，世

界各国更是纷纷加大宽带网络建设力度，不断扩大网络覆盖范围和传输速度。以我国为例，截至 2012 年 6 月，92.6% 的固定宽带用户接入速率达到或超过 2 Mb/s，国际互联网出口带宽达到 1.48 Tb/s，是 2005 年的 11.4 倍。与此同时，移动通信宽带网络迅速发展，3G 网络基本普及，4G 网络覆盖范围不断加大，各种终端设备可以随时随地传输数据。大数据时代，信息传输不再遭遇网络发展初期的瓶颈和制约。

1.1.3 数据产生方式的变革促成大数据时代的来临

数据是人们通过观察、实验或计算得出的结果。数据和信息是两个不同的概念。信息是较为宏观的概念，它由数据的有序排列组合而成，传达给读者某个概念方法等；而数据则是构成信息的基本单位，离散的数据没有任何实用价值。

数据有很多种，比如数字、文字、图像和声音等。随着人类社会信息化进程的加快，在日常生产和生活中每天都会产生大量的数据，比如商业网站、政务系统、零售系统、办公系统、自动化生产系统等，每时每刻都在不断产生数据。数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素，从创新到所有决策，数据推动着企业的发展，并使得各级组织的运营更为高效，可以这样说，数据将成为每个企业获取核心竞争力的关键要素。数据资源已经和物质资源、人力资源一样成为国家的重要战略资源，影响着国家和社会的安全、稳定与发展，因此，数据也被称为“未来的石油”。

数据产生方式的变革，是促成大数据时代来临的重要因素。总体而言，人类社会的数据产生方式大致经历了 3 个阶段：运营式系统阶段、用户原创内容阶段和感知式系统阶段（见表 1-2）。

表 1-2 数据产生方式的变革

阶段	内容
运营式系统阶段	数据库的出现使数据管理的复杂度大大降低，数据往往伴随着一定的运营活动而产生并记录在数据库中，数据的产生方式是被动的
用户原创内容阶段	数据爆发产生于 Web2.0 时代，而 Web2.0 的最重要标志就是用户原创内容智能手机等移动设备加速内容产生数据产生方式是主动的
感知式系统阶段	感知式系统的广泛使用人类社会数据量第三次大的飞跃最终导致了大数据的产生

1. 运营式系统阶段

人类社会最早大规模管理和使用数据，是从数据库的诞生开始的。大型零售超市销售系统、银行交易系统、股市交易系统、医院医疗系统、企业客户管理系统等

大量运营式系统，都是建立在数据库基础之上的，数据库中保存了大量结构化的企业关键信息，用来满足企业各种业务需求。在这个阶段，数据的产生方式是被动的，只有当实际的企业业务发生时，才会产生新的记录并存入数据库。比如，对于股市交易系统而言，只有当发生一笔股票交易时，才会有相关记录生成。

2. 用户原创内容阶段

互联网的出现，使得数据传播更加快捷，不需要借助于磁盘、磁带等物理存储介质传播数据，网页的出现进一步加速了大量网络内容的产生，从而使得人类社会数据量开始呈现“井喷式”增长。但是，互联网真正的数据爆发产生于以“用户原创内容”为特征的 Web 2.0 时代。Web 1.0 时代主要以门户网站为代表，强调内容的组织与提供，大量上网用户本身并不参与内容的产生。而 Web 2.0 技术以 Wiki、博客、微博、微信等自服务模式为主，强调自服务，大量上网用户本身就是内容的生成者，尤其是随着移动互联网和智能手机终端的普及，人们更是可以随时随地使用手机发微博、传照片，数据量开始急剧增加。

3. 感知式系统阶段

物联网的发展最终导致了人类社会数据量的第三次跃升。物联网中包含大量传感器，如温度传感器、湿度传感器、压力传感器、位移传感器、光电传感器等，此外，视频监控摄像头也是物联网的重要组成部分。物联网中的这些设备，每时每刻都在自动产生大量数据，与 Web 2.0 时代的人工数据产生方式相比，物联网中的自动数据产生方式，将在短时间内生成更密集、更大量的数据，使得人类社会迅速步入“大数据时代”。

1.1.4 大数据的发展历程

大数据的发展历程总体上可以划分为 3 个重要阶段：萌芽期、成熟期和大规模应用期（见表 1-3）。

表 1-3 大数据发展的 3 个阶段

阶 段	时 间	内 容
第一阶段：萌芽期	20 世纪 90 年代至 21 世纪初	随着数据挖掘数据库技术的逐步成熟，一批商业智能工具呈知识管理技术开始以被应用，如数据仓库、专家系统、知识管理系统等
第二阶段：成熟期	21 世纪前 10 年	Web2.0 应用迅猛发展，非结构化数据 大量产生，传统处理方法验证以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统 两磊核心 技术，哇哇大哭的 GFS 和 MapReduce 等大数据技术 受到追捧，Hadoop 平台开始大行其道
第三阶段：大规模应用期	2010 年以后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化和程度大幅提高

这里简要回顾大数据的发展历程。

*1980年,著名未来学家阿尔文·托夫勒在《第三次浪潮》一书中,将大数据热情地赞颂为“第三次浪潮的华彩乐章”。

*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第一篇使用“大数据”这一术语的文章。

*1999年10月,在美国电气和电子工程师协会(IEEE)关于可视化的年会上,设置了名为“自动化或者交互:什么更适合大数据?”的专题讨论小组,探讨大数据问题。

*2001年2月,梅塔集团分析师道格·莱尼发布题为《3D数据管理:控制数据容量、处理速度及数据种类》的研究报告。10年后,“3V”(Volume、Variety和Velocity)作为定义大数据的三个维度而被广泛接受。

*2005年9月,蒂姆·奥莱利发表了《什么是Web 2.0》一文,并在文中指出“数据将是下一项技术核心”。

*2008年,《自然》杂志推出大数据专刊;计算社区联盟(Computing Community Consortium)发表了报告《大数据计算:在商业、科学和社会领域的革命性突破》,阐述了大数据技术及其面临的一些挑战。

*2010年2月,肯尼斯·库克尔在《经济学人》上发表了一份关于管理信息的特别报告《数据:无所不在的数据》。

*2011年2月,《科学》杂志推出专刊《处理数据》,讨论了科学研究中的大数据问题。

*2011年,维克托·迈尔·舍恩伯格出版著作《大数据时代:生活、工作与思维的大变革》,引起轰动。

*2011年5月,麦肯锡全球研究院发布《大数据:下一个具有创新力、竞争力与生产力的前沿领域》,提出“大数据”时代到来。

*2012年3月,美国奥巴马政府发布了《大数据研究和发展倡议》,正式启动“大数据发展计划”,大数据上升为美国国家发展战略,被视为美国政府继信息高速公路计划之后在信息科学领域的又一重大举措。

*2013年12月,中国计算机学会发布《中国大数据技术与产业发展白皮书》,系统总结了大数据的核心科学与技术问题,推动了我国大数据学科的建设与发展,并为政府部门提供了战略性的意见与建议。

※2014年5月，美国政府发布2014年全球“大数据”白皮书《大数据：抓住机遇、守护价值》，报告鼓励使用数据来推动社会进步。

※2015年8月，国务院印发《促进大数据发展行动纲要》，全面推进我国大数据发展和应用，加快建设数据强国。

※2016年5月，在“2016大数据产业峰会”上工信部透露，我国将制定出台大数据产业“十三五”发展规划，有力推进我国大数据技术创新和产业发展。

1.2 大数据概念

随着大数据时代的到来，“大数据”已经成为互联网信息技术行业的流行词汇。关于“什么是大数据”这个问题，大家比较认可关于大数据的“4V”说法。大数据的4个“V”，或者说是大数据的4个特点，包含4个层面：数据量大（Volume）、数据类型繁多（Variety）、处理速度快（Velocity）和价值密度低（Value）。

1.2.1 数据量大

人类进入信息社会以后，数据以自然方式增长，其产生不以人的意志为转移。从1986年开始到2010年的20多年时间里，全球数据的数量增长了100倍，今后的数据量增长速度将更快，人们正生活在一个“数据爆炸”的时代。今天，世界上只有25%的设备是联网的，大约80%的上网设备是计算机和手机，而在不久的将来，将有更多的用户成为网民，汽车、电视、家用电器、生产机器等各种设备也将接入互联网。随着Web 2.0和移动互联网的快速发展，人们已经可以随时随地、随心所欲发布包括博客、微博、微信等在内的各种信息。以后，随着物联网的推广和普及，各种传感器和摄像头将遍布人们工作和生活的各个角落，这些设备每时每刻都在自动产生大量数据。

综上所述，人类社会正经历第二次“数据爆炸”（如果把印刷在纸上的文字和图形也看作数据的话，那么人类历史上第一次“数据爆炸”发生在造纸术和印刷术发明的时期）。各种数据产生速度之快，产生数量之大，已经远远超出人类可以控制的范围，“数据爆炸”成为大数据时代的鲜明特征。根据著名咨询机构IDC（Internet Data Center）做出的估测，人类社会产生的数据一直都在以每年50%的速度增长，也就是说，每两年就增加一倍，这被称为“大数据摩尔定律”。这意味着，

人类在最近两年产生的数据量相当于之前产生的全部数据量之和。预计到 2020 年，全球将总共拥有 35ZB（见表 1-4）的数据量，与 2010 年相比，数据量将增长到近 30 倍。

表 1-4 数据存储单位之间的换算关系

单 位	换算关系
B (Byte, 字节)	1B = 8b
KB (Kiobyte, 千字节)	1KB = 1024B
MB (Megabyte, 兆字节)	1MB = 1 024KB
GB (Gigabyte, 吉字节)	1GB = 1 024MB
TB (Trillionbyte, 太字节)	1TB = 1 024GB
PB (prtabyte, 拍字节)	1PB = 1 024TB
EB (Exabyte, 艾字节)	1EB = 1 024PB
ZB (Zettabyte, 泽字节)	1ZB = 1 024EB

1.2.2 数据类型繁多

大数据的数据来源众多，科学研究、企业应用和 Web 应用等都在源源不断地生成新的数据。生物大数据、交通大数据、医疗大数据、电信大数据、电力大数据、金融大数据等都呈现出“井喷式”增长，所涉及的数量十分巨大，已经从 TB 级别跃升到 PB 级别。

大数据的数据类型丰富，包括结构化数据和非结构化数据，其中，前者占 10% 左右，主要是指存储在关系数据库中的数据；后者占 90% 左右，种类繁多，主要包括邮件、音频、视频、微信、微博、位置信息、链接信息、手机呼叫信息、网络日志等。

如此类型繁多的异构数据，对数据处理和分析技术提出了新的挑战，也带来了新的机遇。传统数据主要存储在关系数据库中，但是，在类似 Web 2.0 等应用领域中，越来越多的数据开始被存储在非关系型数据库（Not Only SQL, NoSQL）中，这就必然要求在集成的过程中进行数据转换，而这种转换的过程是非常复杂和难以管理的。传统的联机分析处理（On-Line Analytical Processing, OLAP）和商务智能工具大都面向结构化数据，而在大数据时代，用户友好的、支持非结构化数据分析的商业软件也将迎来广阔的市场空间。

1.2.3 处理速度快

大数据时代的数据产生速度非常迅速。在 Web 2.0 应用领域,在 1min 内,新浪可以产生 2 万条微博, Twitter 可以产生 10 万条推文, 苹果可以下载 4.7 万次应用, 淘宝可以卖出 6 万件商品, 人人网可以发生 30 万次访问, 百度可以产生 90 万次搜索查询, Facebook 可以产生 600 万次浏览量。大名鼎鼎的大型强子对撞机 (LHC), 大约每秒产生 6 亿次的碰撞, 每秒生成约 700 MB 的数据, 有成千上万台计算机分析这些碰撞。

大数据时代的很多应用都需要基于快速生成的数据给出实时分析结果, 用于指导生产和生活实践。因此, 数据处理和分析的速度通常要达到秒级响应, 这一点和传统的数据挖掘技术有着本质的不同, 后者通常不要求给出实时分析结果。

为了实现快速分析海量数据的目的, 新兴的大数据分析技术通常采用集群处理和独特的内部设计。以谷歌公司的 Dremel 为例, 它是一种可扩展的、交互式的实时查询系统, 用于只读嵌套数据的分析, 通过结合多级树状执行过程和列式数据结构, 它能做到几秒内完成对万亿张表的聚合查询, 系统可以扩展到成千上万的 CPU 上, 满足谷歌上百万用户操作 PB 级数据的需求, 并且可以在 2~3s 内完成 PB 级别数据的查询。

1.2.4 价值密度低

大数据虽然看起来很美, 但是价值密度却远远低于传统关系数据库中已经有的那些数据。在大数据时代, 很多有价值的信息都是分散在海量数据中的。以小区监控视频为例, 如果没有意外事件发生, 连续不断产生的数据都是没有任何价值的, 当发生偷盗等意外情况时, 也只有记录了事件过程的那一小段视频是有价值的。但是, 为了能够获得发生偷盗等意外情况时的那一段宝贵的视频, 人们不得不投入大量资金购买监控设备、网络设备、存储设备, 耗费大量的电能和存储空间, 来保存摄像头连续不断传来的监控数据。

如果这个实例还不够典型的话, 那么可以想象另一个更大的场景。假设一个电子商务网站希望通过微博数据进行有针对性的营销, 为了实现这个目的, 就必须构建一个能存储和分析新浪微博数据的大数据平台, 使之能够根据用户微博内容进行有针对性的商品需求趋势预测。愿景很美好, 但是现实代价很大, 可能需要耗费几百万元构建整个大数据团队和平台, 而最终带来的企业销售利润增加额可能会比投入低许多, 从这点来说, 大数据的价值密度是较低的。