全国高校标准教材《云计算》姊妹篇，剖析大数据核心技术和实战应用

大数据 BIG DATA

刘　鹏 ◎ 主编　　张　燕　张重生　张志立 ◎ 副主编

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

全国高校标准教材《云计算》姊妹篇，剖析大数据核心技术和实战应用

大数据

刘　鹏　主　编

张　燕　张重生　张志立　副主编

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书是国内绝大多数高校采用的知名教材《云计算》（1～3 版）的姊妹篇，是中国大数据专家委员会刘鹏教授联合国内多位专家历时两年的心血之作。大数据领域一直缺乏一本权威教材，希望本书能够填补空白。本书系统地介绍了大数据的理论知识和实战应用，包括大数据采集与预处理、数据挖掘算法与工具、深度学习以及大数据可视化等，并深度剖析了大数据在互联网、商业和典型行业的应用。刘鹏教授创办的网站中国大数据（thebigdata.cn）、中国云计算（chinacloud.cn）和微信公众号刘鹏看未来（lpoutlook）将免费提供本书配套 PPT 和其他资料。本书配套的大数据实验体系已经在郑州大学等高校成功应用。

“让学习变得轻松”是本书的初衷。本书适合作为相关专业本科和研究生教材。高职高专学校也可以选用部分内容开展教学。本书也很适合作为大数据研发人员的自学书籍。

图书在版编目（CIP）数据

大数据 / 刘鹏主编. —北京：电子工业出版社，2017.1
ISBN 978-7-121-30430-9

Ⅰ. ①大… Ⅱ. ①刘… 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（2016）第 284502 号

策划编辑：董亚峰
责任编辑：董亚峰　特约编辑：刘广钦
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×1092　1/16　印张：22.5　字数：360 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
定　　价：58.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254694。

编 写 组

主　编： 刘　鹏

副主编： 张　燕　张重生　张志立

编　委： 翟洪军　潘永东　邵奇峰　张　鑫　武郑浩
桂文明　张晓民　郭东恩　吴　刚　叶　崧
谢党恩　陈会忠　于继明　周　端　张佩云
曹　骝　吴　伟　胡　勇　杨震宇　沈大为
蒋永和　顾才东　车战斌　秦恩泉　高秀斌
王建林　袁　科　闫永航　刘畅畅　郝　昱
张　愿　赵冬冬　赵晓东　戎新堃　贾文周
汪洲权　马　慧　陈常伟

基金支持

2015年度江苏高校优秀科技创新团队“大数据智能挖掘信息技术研究”
金陵科技学院高层次人才科研启动基金资助，项目编号：40610186
国家自然科学基金（61472005）资助
江苏省高校软件工程品牌专业建设项目系列教材

前 言

在未来 5～10 年，我国大数据市场规模年均增速将超过 30%。未来 5 年，国内大数据人才缺口将突破 150 万。在 BAT 发布的招聘职位中，目前大数据岗位占比已经超过 60%。现在业界有一种观点：即使把全国所有计算机专业都做成大数据专业，仍然无法满足国内对大数据人才的需求量。

在快速膨胀的需求与国家扶植政策的推动下，全国高校、高职、高专院校纷纷启动大数据人才培养计划。然而，大数据专业建设却面临重重困难。首先，大数据是个新生事物，懂大数据的老师少之又少，院校缺“人”；其次，尚未形成完善的大数据人才培养和课程体系，院校缺“机制”；再次，大数据实验需要为每位学生提供集群计算机，院校缺“机器”；最后，院校不拥有海量数据，开展大数据教学科研工作缺“原材料”。

其实，在 2000 年网格计算兴起时和 2008 年云计算兴起时，我国科技工作者都曾遇到过类似的挑战问题，我有幸参与了这些问题的解决过程。

为了解决网格计算挑战问题，我在清华大学读博期间，于 2001 年创办了中国网格信息中转站（chinagrid.net）网站，每天花好几个小时收集和分享有价值的资料给学术界。于 2002 年与人合作出版了《网格计算》教材。并多次筹办和主持全国性的网格计算学术会议。

为了解决云计算挑战问题，我于 2008 年创办了中国云计算（chinacloud.cn）网站，于 2010 年出版了《云计算（第一版）》、2011 年出版了《云计算（第二版）》、2015 年出版了《云计算（第三版）》，每一版都花费大量成本制作并免费分享对应的几十个教学 PPT。这些 PPT 的下载总量达到了几百万次之多。早在 2010 年，我就在南京组织了全国高校云计算师资培训班，培养了国内第一批云计算老师。并通过与华为、中兴、360 等知名企业合作，输出云计算技术，培养云计算研发人才。为社区做贡献，收获是沉甸甸的：我获得了大家的好评与认可，担任了一些全国性专家委员会的专家，《云计算》教材成为国内高校的首选教材，中国云计算网站成为国内排名第一的云计算网站。

近几年，我用类似的办法来解决我们所面临的大数据挑战问题。为了解决大数据技术资料缺乏和存在交流障碍的问题，我于 2013 年创办了中国大数据（thebigdata.cn）网站，投入大量的人力每天维护，该网站已经在各大搜索引擎排名“大数据”关键词第一

名；为了解决大数据师资匮乏的问题，我面向全国院校，陆续举办多期大数据教师培训班。最近在南京举办的全国高校/高职/中职大数据免费培训班，报名的老师已有 400 多位；为了解决缺乏权威大数据教材的问题，我所负责的南京大数据研究院，联合金陵科技学院、河南大学、中原工学院、南阳理工学院、云创大数据、许昌学院、安徽师范大学、才云科技、中国地震局、南京公安研究院等多家单位，历时两年，编著了《大数据》教材和《大数据库》教材。并计划为高职和中职院校专门编写大数据专业系列教材。我们将在中国大数据（thebigdata.cn）、中国云计算（chinacloud.cn）和刘鹏看未来（lpoutlook）微信公众号等陆续免费提供配套 PPT 和其他资料；为了解决大数据实验难以开展的问题，我带领云创大数据（www.cstor.cn）的科研人员，研发成功 BDRack 大数据实验一体机，它打破虚拟化技术的性能瓶颈，可以为每一位参加实验的人员虚拟出 Hadoop 集群、Spark 集群、MongoDB 集群、Storm 集群等，自带实验所需数据，并准备了详细的实验手册、PPT 和视频，可以开展大数据管理、大数据挖掘等各类实验，并可进行精确营销、信用分析等多种实战演练。目前该平台已经在郑州大学等高校成功应用。我们还开放了免费的物联网大数据托管平台万物云（wanwuyun.com）和环境大数据免费分享平台环境云（envicloud.cn）

在此，特别感谢我的硕士导师谢希仁教授和博士导师李三立院士。谢希仁教授出版的《计算机网络》已经更新到第 6 版，与时俱进且日臻完美，时时提醒学生要以这样的标准来写书。李三立院士是留苏博士，为我国计算机事业做出了杰出贡献，曾任国家攀登计划项目首席科学家。他的严谨治学带出了一大批杰出的学生。

本书是集体智慧的结晶，在此谨向付出辛勤劳动的各位作者致敬！书中难免会有不当之处，请读者不吝赐教。我的邮箱：gloud@126.com，微信公众号：刘鹏看未来（lpoutlook）。

刘鹏　教授
于南京大数据研究院
2016 年 12 月 24 日

目 录

第 1 章　大数据概念与应用

大数据的出现开启了大规模生产、分享和应用数据的时代，能让我们通过对海量数据进行分析，以一种前所未有的方式获得全新的产品、服务或独到的见解，最终形成变革之力，实现重大的时代转型。这就好比当我们感受浩瀚无垠的宇宙时，用望远镜只能看到宇宙的冰山一角，但更广阔的区域都在表面之后，等待着进一步的探索。云计算正是大数据探索过程中的动力源泉，通过对大数据进行检索、分析、挖掘、研判，可以使得决策更为精准，释放出数据背后隐藏的价值。大数据正在改变我们的生活及理解世界的方式，正在成为新发明和新服务的源泉，而更多的改变正蓄势待发……

1.1　大数据之“大”

英特尔创始人戈登·摩尔（Gordon Moore）在 1965 年提出了著名的“摩尔定律”，即当价格不变时，集成电路上可容纳的晶体管数目，约每隔 18 个月便会增加一倍，性能也将提升一倍。1998 年图灵奖获得者杰姆·格雷（Jim Gray）提出著名的“新摩尔定律”，即人类有史以来的数据总量，每过 18 个月就会翻一番[1]。

从图 1-1 中可以看出，2004 年，全球数据总量是 30EB[2]（1EB=10^{18}B=1024PB）；2005 年达到了 50EB，2006 年达到了 161EB；到 2015 年，达到了惊人的 7900EB；到 2020 年，预计将达到 35000EB。

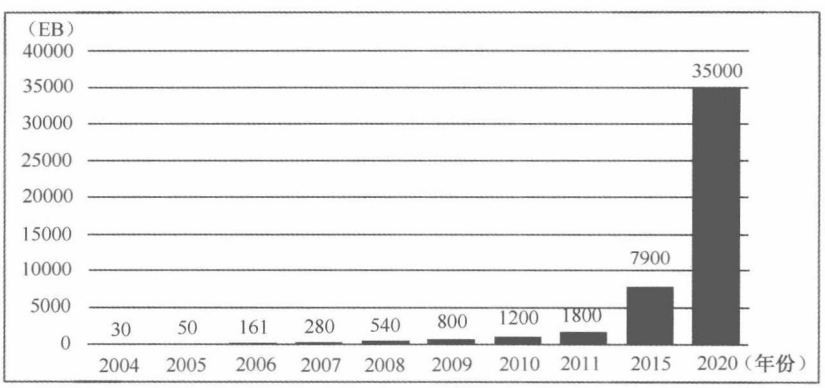

图 1-1　全球数据总量[1]

大数据到底有多大？下面列举出一组互联网数据展示给大家。

（1）互联网每天产生的全部内容可以刻满 6.4 亿张 DVD。

（2）Google 每天需要处理 24PB 的数据。

（3）网民每天在 Facebook 上要花费 234 亿分钟，被移动互联网使用者发送和接收的数据高达 44PB。

（4）全球每秒发送 290 万封电子邮件，一分钟读一篇的话，足够一个人昼夜不停地读 5.5 年。

（5）每天会有 2.88 万个小时的视频上传到 YouTube，足够一个人昼夜不停地观看 3.3 年。

（6）Twitter 上每天发布 5000 万条消息，假设 10 秒浏览一条消息，足够一个人昼夜不停地浏览 16 年。

为什么会产生如此海量的数据？主要有 3 个因素：一是大人群产生的海量数据，全球已经有大约 30 亿人接入了互联网，在 Web 2.0 时代，每个人不仅是信息的接受者，也是信息的产生者，每个人都成为数据源，几乎每个人都在用智能终端拍照、拍视频、发微博、发微信等。二是大量传感器产生的海量数据，目前全球有 30 亿～50 亿个传感器，到 2020 年将达到 1000 亿个之多，这些传感器 24 小时不停地产生数据，这就导致了信息的爆炸。三是科学研究和各行各业越来越依赖大数据手段来开展工作，例如，欧洲粒子物理研究所的大型强子对撞机每年需要处理的数据是 100PB，且年增长 27PB；又如，石油部门用地震勘探的方法来探测地质构造、寻找石油，需要用大量传感器来采集地震波形数据；高铁的运行要保障安全，需要在铁轨周边大量部署传感器，从而感知异物、滑坡、水淹、变形、地震等异常。

也就是说，随着人类活动的进一步扩展，数据规模会急剧膨胀，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的各行业累积的数据量越来越大，数据类型也越来越多、越来越复杂，已经超越了传统数据管理系统、处理模式的能力范围，于是“大数据”这样一个在含义上趋近于“无穷大”的概念才会应运而生[3]。

那么，何为大数据？大数据又称巨量数据，指的是无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。

大数据已经渗透到每一个行业和业务职能领域，并成为重要的生产因素。目前工业界普遍认为大数据具有 5V+1C 的特征：大量（Volume）、多样（Variety）、高速（Velocity）、价值（Value）、准确性（Veracity）和复杂（Complexity）[3]。

（1）大量（Volume）：存储的数据量巨大，PB 级别是常态，因而，对其分析的计算量也大。

（2）多样（Variety）：数据的来源及格式多样，数据格式除了传统的格式化数据外，还包括半结构化或非结构化数据，如用户上传的音频和视频内容。随着人类活动的进一步拓宽，数据的来源更加多样。

（3）高速（Velocity）：数据增长速度快，同时要求对数据的处理速度也要快，以便能够从数据中及时地提取知识，发现价值。

（4）价值（Value）：需要对大量的数据处理，挖掘其潜在的价值，因而，大数据对我们提出的明确要求是设计一种在成本可接受的条件下，通过快速采集、发现和分析从大量、多种类别的数据中提取价值的体系架构。

（5）准确性（Veracity）：即处理的结果要保证一定的准确性。

（6）复杂（Complexity）：对数据的处理和分析的难度大。

从大数据的特征可以看出 3 个层次的内容。①海量的数据。数据获取和用户使用需求呈指数级增长，数量极其庞大。②数据复杂度高。其非结构化特征非常明显，传统的数据处理方式无法来处理。③处理时效与分析得到的结果的可用性。数据海量加之结构复杂，对分析处理的技术要求相当高，数据的及时处理难度相当大；同时，从大数据中提取出来的规律或结果必须是真实的、有价值的、可用的。可见，大数据问题涉及从存储、转换、传输直到分析的每一个层面，运用传统的数据处理工具和技术无法满足实时处理大数据的需求。

1.2　大数据的来源

最早提出“大数据”时代已经到来的机构是全球知名咨询公司麦肯锡[4]。根据麦肯锡全球研究所的分析，利用大数据能在各行各业产生显著的社会效益。美国健康护理利用大数据每年产出 3000 多亿美元，年劳动生产率提高 0.7%；欧洲公共管理每年价值 2500 多亿欧元，年劳动生产率提高 0.5%；全球个人定位数据服务提供商收益 1000 多亿美元，为终端用户提供高达 7000 多亿美元的价值；美国零售业净收益可增长 6%，年劳动生产率提高 1%；制造业可节省 50%的产品开发和装配成本，营运资本下降 7%。可见，大数据无所不在，已经对人们的工作、生活和学习产生了深远的影响，并将持续发展。

大数据的来源这个问题其实很简单，其无非就是通过各种数据采集器、数据库、开源的数据发布、GPS 信息、网络痕迹（如购物、搜索历史等）、传感器收集的、用户保存的、上传的等结构化或者非结构化的数据，非常广泛。正是由于这种广泛性，人们才可以从产生数据的主体、数据来源的行业、数据存储的形式三个方面来对大数据的来源进行分类，并借此窥视大数据的来源途径。

1．按产生数据的主体划分

（1）少量企业应用产生的数据：如关系型数据库中的数据和数据仓库中的数据等。

（2）大量人产生的数据：如推特、微博、通信软件、移动通信数据、电子商务在线交易日志数据、企业应用的相关评论数据等。

（3）巨量机器产生的数据：如应用服务器日志、各类传感器数据、图像和视频监控数据、二维码和条形码（条码）扫描数据等。

2．按数据来源的行业划分

（1）以 BAT 为代表的互联网公司：百度公司数据总量超过了千 PB 级别，数据涵盖了中文网页、百度推广、百度日志、UGC 等多个部分，并以 70%以上的搜索市场份额坐拥庞大的搜索数据。阿里巴巴公司保存的数据量超过了百 PB 级别，拥有 90%以上的电商数据，数据涵盖了点击网页数据、用户浏览数据、交易数据、购物数据等。腾讯公司总存储数据量经压缩处理以后仍然超过了百 PB 级别，数据量月增加达到 10%，包括大量社交、游戏等领域积累的文本、音频、视频和关系类数据。

（2）电信、金融、保险、电力、石化系统：电信行业数据包括用户上网记录、通话、信息、地理位置数据等，运营商拥有的数据量将近百 PB 级别，年度用户数据增长

超过 10%。金融与保险包括开户信息数据、银行网点数据、在线交易数据、自身运营的数据等，金融系统每年产生的数据超过数十 PB，保险系统的数据量也超过了 PB 级别。电力与石化方面，仅国家电网采集获得的数据总量就达到了数十 PB，石油化工领域每年产生和保存下来的数据量也将近百 PB 级别。

（3）公共安全、医疗、交通领域：一个中、大型城市，一个月的交通卡口记录数可以达到 3 亿条；整个医疗卫生行业一年能够保存下来的数据就可达到数百 PB 级别；航班往返一次产生的数据就达到 TB 级别；列车、水陆路运输产生的各种视频、文本类数据，每年保存下来的也达到数十 PB。

（4）气象、地理、政务等领域：中国气象局保存的数据将近 10PB，每年约增数百 TB；各种地图和地理位置信息每年约数十 PB；政务数据则涵盖了旅游、教育、交通、医疗等多个门类，且多为结构化数据。

（5）制造业和其他传统行业：制造业的大数据类型以产品设计数据、企业生产环节的业务数据和生产监控数据为主。其中产品设计数据以文件为主，非结构化，共享要求较高，保存时间较长；企业生产环节的业务数据主要是数据库结构化数据，而生产监控数据则数据量非常大。在其他传统行业，虽然线下商业销售、农林牧渔业、线下餐饮、食品、科研、物流运输等行业数据量剧增，但是数据量还处于积累期，整体体量都不算大，多则达到 PB 级别，少则数十 TB 或数百 TB 级别。

3. 按数据存储的形式划分

大数据不仅仅体现在数据量大，还体现在数据类型多。如此海量的数据中，仅有20%左右属于结构化的数据，80%的数据属于广泛存在于社交网络、物联网、电子商务等领域的非结构化数据。

结构化数据简单来说就是数据库，如企业 ERP、财务系统、医疗 HIS 数据库、教育一卡通、政府行政审批、其他核心数据库等数据。

非结构化数据包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频、视频信息等数据。

大数据的价值不在于存储数据本身，而在于如何挖掘数据，只有具备足够的数据源才可以挖掘出数据背后的价值，因此，获取大数据是非常重要的基础。就数据获取而言，大型互联网企业由于自身用户规模庞大，可以把自身用户产生的交易、社交、搜索等数据充分挖掘，拥有稳定安全的数据资源。对于其他大数据公司和大数据研究机构而言，目前获取大数据的方法有如下 4 种：

（1）系统日志采集。可以使用海量数据采集工具，用于系统日志采集，如 Hadoop 的 Chukwa、Cloudera 的 Flume、Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足大数据的日志数据采集和传输需求。

（2）互联网数据采集。通过网络爬虫或网站公开 API 等方式从网站上获取数据信息，该方法可以数据从网页中抽取出来，将其存储为统一的本地数据文件，它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。除了网站中包含的内容之外，还可以使用 DPI 或 DFI 等带宽管理技术实现对网络流量的采集。

（3）APP 移动端数据采集。APP 是获取用户移动端数据的一种有效方法，APP 中的