广东省专业技术人员继续教育公需科目教材

Dashuju Qianyan Jishu Yu Yingyong

大数据

前沿技术与应用

何克晶　阳义南◎编著

非外借

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS

广东省专业技术人员继续教育公需科目教材

大数据

前沿技术与应用

何克晶　阳义南　编著

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS
·广州·

图书在版编目（CIP）数据

大数据前沿技术与应用/何克晶，阳义南编著．—广州：华南理工大学出版社，2017.3

ISBN 978-7-5623-5238-9

Ⅰ.①大… Ⅱ.①何… ②阳… Ⅲ.①数据管理 Ⅳ.①F279.23

中国版本图书馆 CIP 数据核字（2017）第 069056 号

大数据前沿技术与应用

何克晶 阳义南 编著

出 版 人：卢家明

出版发行：华南理工大学出版社

（广州五山华南理工大学 17 号楼，邮编 510640）

http://www.scutpress.com.cn E-mail:scutc13@scut.edu.cn

营销部电话：020-87113487 87111048（传真）

策划编辑：卢家明 袁 泽

责任编辑：刘 锋 袁 泽

印 刷 者：佛山市浩文彩色印刷有限公司

开 本：787mm×960mm 1/16 **印张**：6.25 **字数**：107 千

版 次：2017 年 3 月第 1 版 2017 年 3 月第 1 次印刷

定 价：28.00 元

版权所有 盗版必究 印装差错 负责调换

前　言

PREFACE

当前，正处于以数据为驱动的时代，大数据发展浪潮持续汹涌，它所涉及的行业十分广泛，从金融、教育、医疗等传统行业到电子商务、移动社交、智能设备等新兴产业，都有大数据的影子。大数据相关技术更是国家科技创新和“互联网+”产业发展的基础。而大数据相关技术及应用跨越多个领域，尤其对非计算机专业背景的人来说，理解起来有一定难度。因此，本书旨在通过通俗易懂的语言介绍大数据领域的前沿技术与应用，帮助读者了解未来信息时代的发展脉络。

大数据技术横跨多个技术领域，从数据存储、数据计算到云存储、并行计算，再到数据挖掘。考虑到本书篇幅有限，且侧重于大数据的入门介绍和科普性质，本书着重介绍大数据中典型的相关技术。在大数据的应用层面，本书主要结合国外、国内及广东省的大数据发展情况进行阐述。

本书共分4章，内容安排如下：

第1章是走进大数据时代。主要从整体上介绍大数据涉及的技术范畴和应用领域，在了解大数据发展现状的基础上介绍大数据的概念、特征及技术架构，然后介绍大数据在不同领域的应用，最后是探讨大数据发展的机遇与挑战。

第2章是大数据关键技术。主要介绍大数据存储、大数据处理与计算以及流式大数据处理的技术，涉及的具体技术包括NoSQL、NewSQL、分布式计算、MapReduce、Hadoop、Spark和Storm等框架或平台。本章专业性比较强，理解起来可能存在一定难度。

第3章是大数据挖掘与分析。主要介绍大数据挖掘与分析的应用领域，并从机器学习与数据挖掘的角度介绍大数据挖掘与分析的理论与方法，最后简要介绍相关技术工具。本章内容是实现大数据价值最大化的关键部分，理论性和实践性较强，且跨多个学科。

第4章是大数据的应用。主要介绍促进大数据应用的战略意义、我国发展大数据的优势及发展方向，然后介绍大数据在国内外的经典应用案例，最后是探讨大数据在应用中存在的问题以及未来的趋势。

本书根据广东省人力资源和社会保障厅对专业技术人员继续教育培训的要求而编写，作为一本技术普及与行业应用相结合的读物，在写作方式上力求深入浅出，使未具备此专业知识的读者能阅读此书后对大数据技术的前沿动态和行业应用情况有初步的了解与认识。尽管本书在编写中投入了大量的资源与精力，付出了诸多艰苦努力，但囿于时间仓促，书中难免存在疏漏之处，恳请读者和专家批评指正。

编　者

2017年3月于广州

试读结束：需要全本请在线购买：　www.ertongbook.com

目　录

CONTENTS

第 1 章

走进大数据时代

早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据比作“第三次浪潮的华彩乐章”。现如今，数据已经呈爆炸式增长，足以引发全世界范围内的一次技术变革，“第三次浪潮”——大数据时代已经来到。

1.1 大数据发展现状

1.1.1 大数据国外发展现状

大数据是新资源、新技术和新理念的混合体。“大数据”这个概念在 20 世纪 80 年代就出现了，到了 2008 年，“大数据”这个词已经不再陌生，并有了广泛传播之势。移动互联网和云计算的出现，让人们逐渐认识到大数据的重大意义，国际顶级学术刊物相继出版大数据方面的专刊，讨论大数据的特征、技术与应用。2011 年，麦肯锡全球研究所发布名为《大数据：创新、竞争和生产力的下一个前沿》的报告，提出了大数据概念，认为数据已经成为经济社会发展的重要推动力。2012 年，大数据成为 IT 界的热门话题。2013 年 3 月，美国政府宣布推出“大数据研究和发展计划”（big data research and development initiative），有人将其比喻为美国的下一个“信息高速公路”计划。

英国政府紧随美国之后，推出一系列支持大数据发展的举措，首先便是给予研发资金支持。2013 年 1 月，英国政府向航天、医药等 8 类高新技术领域注资 6 亿英镑研发，其中大数据技术获得 1.89 亿英镑的资金，是获得资金最多的领域。为了在医疗领域更好地应用大数据，2013 年 5 月，英国政府和李嘉诚基

金会联合投资设立全球首个综合运用大数据技术的医药卫生科研机构，将透过高通量生物数据，与业界共同界定药物标靶，处理目前在新药开发过程中关键的瓶颈，之后还将聚集遗传学、流行病学、临床、化学和计算机科学等领域的顶尖人才，集中分析庞大的医疗数据。

法国政府为促进大数据领域的发展，将以培养新兴企业、软件制造商、工程师、信息系统设计师等为目标，开展一系列的投资计划。法国政府在其发布的《数字化路线图》中表示，将大力支持包括大数据在内的战略性高新技术。法国软件编辑联盟曾号召政府部门和私人企业共同合作，投入3亿欧元资金用于推动大数据领域的发展。

日本政府认为，要提升日本的国际竞争力，大数据应用不可或缺。日本在新一轮IT振兴计划中把发展大数据作为国家战略的重要内容，新的ICT（information communications technology，信息、通信和技术）战略重点关注大数据应用技术。日本总务省于2012年7月推出了新的综合战略“活力ICT日本”，重点关注大数据应用，并将其作为2013年六个主要任务之一，聚焦大数据应用所需的社会化媒体等智能技术开发，以及在新医疗技术开发、缓解交通拥堵等公共领域的应用。

1.1.2 我国促进大数据发展的公共政策

我国高度重视大数据的应用和发展。自2014年3月“大数据”首次出现在《政府工作报告》中以来，国务院常务会议一年内6次提及大数据运用。在2015年6月17日的国务院常务会议上，李克强总理再次强调“我们正在推进简政放权，放管结合、优化服务，而大数据手段的运用十分重要”。2015年7月1日，国务院办公厅印发了《关于运用大数据加强对市场主体服务和监管的若干意见》。表1-1是2012年以来国内关于大数据行业的相关政策汇总。

表1-1 国内重要大数据相关政策行动

时 间	部门/地方	政策行动名称
2012年7月	国务院	“十二五”国家战略性新兴产业发展规划
2013年7月	重庆	重庆市大数据行动计划
2013年7月	上海	上海推进大数据研究与发展三年行动计划（2013—2015）
2013年8月	国务院	关于信息消费扩大内需的若干意见

续表1-1

时 间	部门/地方	政策行动名称
2014年2月	贵州	关于加快大数据产业发展应用若干政策的意见
2015年3月	国务院	制定“互联网+”行动计划
2015年4月	发改委	创新投资管理方式建立协同监管机制的若干意见
2015年5月	工信部	将编制实施软件和大数据“十三五”规划
2015年6月	国家信息中心	联合深圳大学建立大数据研究院
2015年6月	工信部	加快推进云计算和大数据标准体系建设
2015年7月	国务院	关于运用大数据加强对市场主体服务和监管的若干意见

2015年8月31日，国务院正式印发了《促进大数据发展行动纲要》，系统部署大数据应用的发展工作。2016年，继国家发改委印发了《关于组织实施促进大数据发展重大工程的通知》后，环保部、国务院办公厅、国土资源部、国家林业局、煤工委、交通运输部、农业部均推出大数据发展意见和方案。大数据政策从全面、总体规划逐渐朝各大产业、各细分领域延伸，大数据产业发展也逐步从理论研究走向实际应用之路。

2016年1月15日，贵州省通过了《贵州省大数据发展应用促进条例》，这是中国首部大数据地方法规，该条例将大数据产业纳入法治轨道，以立法推动大数据产业蓬勃发展。条例的出台不仅是贵州作为大数据综合试验区迈出的坚实一步，对大数据产业的健康发展具有很大的促进作用，更为重要的是，条例填补了中国大数据立法的空白。

截至2016年9月，工业和信息化部透露，全国已有30多个省市专门出台了大数据相关的政策文件，十余个地方专门设置了大数据的管理部门，统筹推进大数据发展，呈现出京津冀、长三角、珠三角、中西部、东北部全面开花的格局。2016年10月8日，国家发展改革委、工业和信息化部、中央网信办发函批复，在京津冀、珠江三角洲、上海市、河南省、重庆市、沈阳市、内蒙古七个区域推进国家大数据综合试验区建设，这是继贵州之后第二批获批建设的国家级大数据综合试验区。此次批复是贯彻落实国务院《促进大数据发展行动纲要》的重要举措，将在大数据制度创新、公共数据开放共享、大数据创新应用、大数据产业聚集、大数据要素流通、数据中心整合利用、大数据国际交流合作等方面进行试验探索，推动我国大数据创新发展。

大数据产业“十三五”发展规划在征求专家意见并集中讨论和修改之后，

试读结束：需要全本请在线购买： www.ertongbook.com

于2017年1月17日正式发布《大数据产业发展规划（2016—2020年）》（以下简称“规划”）。《规划》作为引领DT（data technology，数据处理技术）时代的指导性文件，涉及内容包括推动大数据在工业研发、制造、产业链全流程各环节的应用，支持服务业利用大数据建立品牌、精准营销和定制服务等。总之，随着一系列大数据产业政策的出台，我国大数据产业发展有了依据和指导，并得以实现规范化。在政策支持下，大数据产业实现创新和应用也会更加积极，我国开展产业结构调整和升级也就有了更科学的依据。另外，大数据产业相关的标准制定、推广与国际合作等方面继续完善，也会进一步促进大数据产业的可持续发展，拓宽大数据的应用领域。

1.1.3 广东省的大数据发展政策

广东省委、省政府高度重视大数据工作，明确提出要发展成为大数据应用先行区。2012年底提出实施大数据战略以来，广东省大数据发展取得了良好的成效。2013年4月，成立广东省实施大数据战略专家委员会，10位国内外知名大数据专家担任委员。2014年2月，广东省政府批准在省经济和信息化委设立省大数据管理局。2014年9月1日起施行的《广东省信息化促进条例》，大数据发展是其中一项重要内容。在《广东省电子政务发展规划（2014—2020年）》中，也明确提出要加快政务大数据发展。

2016年4月22日，《广东省促进大数据发展行动计划（2016—2020年）》正式印发。该行动计划明确提出要加快大数据基础设施建设，要利用大数据提高社会治理能力，利用大数据促转型升级，打造新经济增长点，“推进东莞、佛山、惠州、中山等地创建数据资源应用试点，促进区域数据资源的汇聚应用”。2016年10月，珠江三角洲国家大数据综合试验区作为全国首批确定的跨区域类综合试验区正式启动建设，其目标是到2020年，力争将广东打造成为全国数据应用先导区、大数据创业创新集聚区，抢占数据产业发展高地，建成具有国际竞争力的国家大数据综合试验区。

广东省地方各市也相继出台了大数据应用政策。2016年8月，东莞市出台了《东莞大数据行动计划（2016—2020年）》，不久之后又出台了《东莞大数据发展实施方案》。2016年12月，中山市出台了《中山市促进大数据发展行动计划（2016—2020年）》。2017年1月，广州市也出台了《广州市人民政府办公厅

关于促进大数据发展的实施意见》。此外，为贯彻落实大数据发展行动计划，充分发挥应用示范效应，广东省大数据管理局还组织开展了工业大数据应用示范项目推荐工作。

1.2 认识大数据

1.2.1 大数据的概念

“大数据”一词由英文“Big Data”翻译而来。麦肯锡全球研究所报告《大数据：创新、竞争和生产力的下一个前沿》对“大数据”的定义如下（James，2011）：大数据是指大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据群。这个定义有意地带有主观性，对于“究竟多大才算是大数据”，其标准是可以调整的，即我们不以超过多少 TB（$1TB = 2^{10}GB$）为大数据的标准，我们假设随着时间的推移和技术的进步，大数据的“量”仍会增加。应注意到，该定义可以因部门的不同而有所差异，也取决于什么类型的软件工具是通用的，以及某个特定行业的数据集通常的大小。

大数据研究机构 Gartner 指出，大数据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。大数据的目标不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”，大数据是为解决巨量复杂数据而生的。巨量复杂数据有两个核心点，一个是巨量，一个是复杂。“巨量”意味着数据量大，要实时处理的数据越来越多，一旦在处理巨量数据上耗费的时间超出了可承受的范围，将意味着企业的策略落后于市场。“复杂”意味着数据是多元的，不再是过去的结构化数据了，必须针对多元数据重新构建一套有效的理论或分析模型，甚至分析行为所依托的软硬件都必须进行革新。

1.2.2 大数据的特征

一般认为，大数据主要具有以下四个方面的典型特征：volume（大量）、variety（多样）、value（价值）、velocity（高速），这四个典型特征通常称为大

数据的“4V”特征。

1. 数据体量巨大（volume）

大数据的特征首先就体现为数据体量大。如今存储的数据数量正在急剧增长，我们身边的所有数据，包括财务数据、医疗数据、监控数据等，都快将人类“淹没”在数据的“海洋”中。随着计算机深入到人类生活的各个领域，数据基数在不断增大，数据的存储单位已经从过去的 GB 级升级到 TB 级，再到 PB 级甚至 EB 级。要知道，每一个单位都是前面一个单位的 2^{10} 倍！

2. 数据类型多（variety）

广泛的数据来源，决定了大数据形式的多样性。以往的数据尽管数量庞大，但通常是事先定义好的结构化数据。结构化数据是将事物向便于计算机存储、处理的方向抽象后的结果，结构化数据在抽象的过程中，忽略了一些在特定的应用下可以不考虑的细节。相对于以往的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这一类数据的大小、内容、格式、用途可能完全不一样，对数据的处理能力提出了更高的要求。无论是企业还是人们日常生活中接触到的数据，绝大部分都是非结构化的。而半结构化数据，就是介于完全结构化数据和完全非结构化数据之间的数据，HTML 文档就属于半结构化数据，它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

3. 价值高，但价值密度低（value）

价值密度的高低与数据总量的大小成反比。大数据为了获取事物的全部细节，不对事物进行抽象、归纳等处理，直接采用原始的数据，保留了数据的原貌。因此相对于特定的应用，大数据关注的非结构化数据的价值密度偏低。以视频为例，一部 1 小时的视频，在连续不间断的监控中，有用数据可能仅有一两秒。如何通过强大的算法更迅速地完成数据的价值“提纯”，成为目前大数据背景下亟待解决的难题。大数据最大的价值在于通过从大量不相关的各种类型的数据中，挖掘出对未来趋势与模式预测分析有价值的数据，发现新规律和新知识。

4. 处理速度快（velocity）

数据的增长速度和处理速度是大数据高速性的重要体现。根据 IDC（Internet data center，互联网数据中心）的报告，预计到 2020 年，全球数据使

用量将达到35.2ZB。在如此海量的数据面前，处理数据的效率显得格外重要。企业不仅需要了解如何快速获取数据，还必须知道如何快速处理、分析，并返回结果给用户，以满足他们的实时需求。新数据不断涌现，快速增长的数据量要求数据处理的速度也要相应的提升，才能让大量的数据得到有效利用。此外，一些数据是在互联网中不断流动，且随着时间推移而迅速衰减的，如果数据尚未得到及时有效的处理，就失去了价值，大量的数据就没有意义。对不断增长的海量数据进行实时处理，是大数据与传统数据处理技术的关键差别之一。

1.3 大数据技术架构

大数据技术包含各类基础设施支持，底层计算资源支撑着上层的大数据处理。底层主要是数据采集、数据存储阶段，上层则是大数据的计算、处理、挖掘与分析和数据可视化等阶段。

1.3.1 基础设施支持

大数据处理需要拥有大规模物理资源的云数据中心和具备高效的调度管理功能的云计算平台的支撑。云计算管理平台能为大型数据中心及企业提供灵活高效的部署、运行和管理环境，通过虚拟化技术支持异构的底层硬件及操作系统，为应用提供安全、高性能、高可扩展性、高可靠和高伸缩性的云资源管理解决方案，降低应用系统开发、部署、运行和维护的成本，提高资源使用效率。

云计算平台可分为3类：以数据存储为主的存储型云平台、以数据处理为主的计算型云平台以及计算和数据存储处理兼顾的综合云计算平台。目前在国内外已经存在较多的云计算平台，开源的有OpenStack、CloudStack、Hadoop等。商业化的云计算平台国外有Google公司的AppEngine、微软公司的Azure、Amazon公司的EC2等，国内也有阿里云、百度云和腾讯云等。

1.3.2 数据采集

足够的数据量是企业大数据战略建设的基础，因此数据采集是大数据价值挖掘中的重要一环。数据的采集有基于物联网传感器的采集，也有基于网络信息的数据采集。比如在智能交通中，数据的采集有基于GPS的定位信息采集、

基于交通摄像头的视频采集、基于交通卡口的图像采集、基于路口的线圈信号采集等。而在互联网上的数据采集是对各类网络媒介的，如搜索引擎、新闻网站、论坛、微博、博客、电商网站等的各种页面信息和用户访问信息进行采集，采集的内容包括文本信息、网页链接、访问日志、日期和图片等。之后我们需要把采集到的各类数据进行清洗、过滤、去重等各项预处理并分类归纳存储。

在数据量呈爆炸式增长的今天，数据的种类丰富多样，也有越来越多的数据需要放到分布式平台上进行存储和计算。数据采集过程中的 ETL（extract，transform，load，提取、转换和加载）工具将分布的、异构数据源中的不同种类和结构的数据抽取到临时中间层进行清洗、转换、分类、集成，最后加载到对应的数据存储系统，如数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。在分布式系统中，经常需要采集各个节点的日志，然后进行分析。企业每天都会产生大量的日志数据，对这些日志数据的处理也需要特定的日志系统。因为与传统的数据相比，大数据的体量巨大，产生速度非常快，对数据的预处理也需要实时快速，所以在 ETL 的架构和工具选择上，也许要采用分布式内存数据、实时流处理系统等技术。根据实际生活环境中应用环境和需求的不同，目前已经产生了一些高效的数据采集工具，包括 Flume、Scribe、Chukwa 和 Kafka 等。

1.3.3 数据存储

大数据中的数据存储是实现大数据系统架构中的一个重要组成部分。大数据存储专注于解决海量数据的存储问题，它既可以给大数据技术提供专业的存储解决方案，又可以独立发布存储服务。云存储将存储作为服务，它将分别位于网络中不同位置的大量类型各异的存储设备通过集群应用、网络技术和分布式文件系统等集合起来协同工作，通过应用软件进行业务管理，并通过统一的应用接口对外提供数据存储和业务访问功能。云存储系统具有良好的可扩展性、容错性，以及内部实现对用户透明等特性，这一切都离不开分布式文件系统的支撑。现有的云存储分布式文件系统包括 GFS 和 HDFS 等。此外，目前存在的数据库存储方案有 SQL、NoSQL 和 NewSQL。SQL 是目前为止企业应用中最为成功的数据存储方案，仍有相当大一部分的企业把 SQL 数据库作为数据存储方案。NoSQL 和 NewSQL 则是为了解决 SQL 的不足而产生的。

1.3.4 数据计算

面向大数据处理的数据查询、统计、分析、数据挖掘、深度学习等计算需求，促生了大数据计算的不同计算模式，整体上我们把大数据计算分为离线批处理计算和实时计算两种。

其中，离线批处理计算模式最典型的应该是 Google 提出的 MapReduce 编程模型。MapReduce 的核心思想就是将大数据并行处理问题分而治之，即将一个大数据通过一定的数据划分方法，分成多个较小的具有同样计算过程的数据块，数据块之间不存在依赖关系，将每一个数据块分给不同的节点去处理，最后再将处理的结果进行汇总。

实时计算一个重要的需求就是能够实时响应计算结果，主要有以下两种应用场景：一种是数据源是实时的、不间断的，同时要求用户请求的响应时间也是实时的；另一种是数据量大，无法进行预算，但要求对用户请求实时响应的。实时计算在流数据不断变化的运动过程中实时地进行分析，捕捉到可能对用户有用的信息，并把结果发送出去。整个过程中，数据分析处理系统是主动的，而用户却处于被动接收的状态。数据的实时计算框架需要能够适应流式数据的处理，可以进行不间断的查询，同时要求系统稳定可靠，具有较强的可扩展性和可维护性，目前较为主流的实时流计算框架包括 Storm 和 Spark Streaming 等。

1.3.5 数据可视化

数据可视化是将数据以不同形式展现在不同系统中。计算结果需要以简单、直观的方式展现出来，才能最终被用户理解和使用，形成有效的统计、分析、预测及决策，应用到生产实践和企业运营中。想要通过纯文本或纯表格的形式理解大数据信息是非常困难的，相比之下，数据可视化却能够将数据网络的趋势和固有模式展现得更为清晰。可视化会为用户提供一个总的概览，再通过缩放和筛选，为人们提供其所需的更深入的细节信息。可视化的过程在帮助人们利用大数据获取较为完整的信息时起到了关键性作用。可视化分析是一种通过交互式可视化界面，来辅助用户对大规模复杂数据集进行分析推理的技术。可视化分析的运行过程可以看作是“数据—知识—数据”的循环过程，中间经过两条主线：可视化技术和自动化分析模型。

试读结束：需要全本请在线购买： www.ertongbook.com