普通高等教育“十三五”规划教材

Free Courseware

大数据技术及其应用

吕林涛 等 编著

科学出版社

普通高等教育“十三五”规划教材

大数据技术及其应用

吕林涛 等 编著

冯博琴 主审

科学出版社

北京

内 容 简 介

本书分为上篇（基础篇）、中篇（编程篇）和下篇（应用篇）三篇，共 13 章。书中主要内容包括大数据技术概述、大数据处理平台 Hadoop、分布式文件系统 HDFS、分布式计算框架 MapReduce、内存型计算框架 Spark、分布式数据库 HBase、数据仓库 Hive、Pig 语言、Python 语言、分布式数据收集系统 Chukwa、分布式协调服务 ZooKeeper、大规模微博传播分析案例和图书推荐案例等。

本书将理论与科研实践相结合，注重大数据技术的系统性、实用性和先进性，配有大量的应用案例，不仅能够帮助读者提高大数据技术的应用与研究水平，而且能够提高读者的综合应用创新能力。

本书可作为高等院校计算机科学与技术、物联网工程、数据科学与大数据技术等专业，或新工科相关专业本科生、研究生的教材，也可供从事大数据技术应用与开发，以及大数据系统运营与维护的科研、工程技术人员参考使用。

图书在版编目（CIP）数据

大数据技术及其应用/吕林涛等编著. —北京：科学出版社，2019.5

ISBN 978-7-03-056143-5

I. ①大… II. ①吕… 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（2017）第 326443 号

责任编辑：孙露露 常晓敏 / 责任校对：王万红

责任印制：吕春珉 / 封面设计：耕者设计工作室

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

三河市骏杰印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2019 年 5 月第 一 版 开本：787×1092 1/16

2019 年 5 月第一次印刷 印张：16 3/4

字数：379 000

定价：42.00 元

（如有印装质量问题，我社负责调换〈骏杰〉）

销售部电话 010-62136230 编辑部电话 010-62135120-2010

前　言

进入云计算和大数据时代以来，大数据技术在“互联网+”“一带一路”以及新经济、新产业、新业态中发挥着越来越重要的作用。当前，大数据技术已成为高等院校计算机科学与技术、物联网工程、数据科学与大数据技术等专业，或新工科相关专业学生以及科技人员必不可少的基本技能。为适应当前高等院校服务于国家社会经济的发展需求，以及培养具有一定行业深度、特色鲜明的创新应用型人才的需求，作者科研团队在大数据领域研究成果的基础上编著了本书。

针对读者对大数据技术的不同需求，以及高等院校开设本课程的课时不同等情况，本书分为上、中、下三篇，共 13 章，相关学校可根据授课课时选取不同授课内容。上篇（基础篇）主要讲述大数据技术基础，中篇（编程篇）主要讲述大数据应用开发编程技术，下篇（应用篇）介绍两个大数据技术应用案例。

本书的特点是，注重大数据技术的系统性、实用性和先进性；坚持理论与实践相结合、基本原理与创新能力相结合；强调基本原理，概念准确，论述严谨，内容新颖，既介绍大数据基本技术，也介绍大数据最新技术，力求反映大数据技术的最新发展成果。为进一步拓展和提升读者的大数据技术的应用水平和创新能力，各章均附有习题和参考文献。另外，为便于教师教学和学生学习，本书配有电子课件等教学资源，可到科学出版社网站（www.abook.cn）下载或发邮件至 360603935@qq.com 索取。

本书由吕林涛完成大纲制定及统稿，由国家级教学名师冯博琴主审，由黄文准、郭建新、乌伟审校。本书第 1、2 章由马亚红撰写，第 3、4 章由徐鲁辉和姚全珠撰写，第 5～8、10、11 章由吕林涛、吕晖撰写，第 9 章由张玉成和黄世奇撰写，第 12 章由乌伟和孔韦韦撰写，第 13 章由傅海明和胡文斌撰写。

本书在编著过程中参考了许多相关文献，在此对相关作者一并表示感谢。

由于时间仓促，加之作者水平有限，本书难免存在疏漏之处，殷切希望广大读者批评指正。

目　录

上篇　基　础　篇

中篇 编 程 篇

上篇

基础篇

大数据是为了适应新经济、新产业、新业态和“互联网+”应用等需求出现的一种新技术。本篇以Hadoop框架为研究对象，主要介绍大数据技术发展概况、大数据处理平台 Hadoop、分布式文件系统 HDFS、分布式计算框架 MapReduce、内存型计算框架 Spark、分布式数据库 HBase 和数据仓库 Hive。

通过本篇的学习，读者可以全面了解和掌握 Hadoop 系统架构及大数据分析技术，从而为后续两篇内容的学习奠定基础。

第1章 大数据技术概述

随着互联网技术的蓬勃发展，大数据（big data）已经渗透到每个人的日常生活之中。传统的数据挖掘和处理技术已经无法满足大数据的处理要求。大数据技术是信息技术领域又一次颠覆性的技术变革，其核心在于为客户从数据中挖掘出蕴藏的价值。

本章主要介绍大数据的发展历史、基本特征，大数据处理框架，以及大数据技术的主要应用领域。

1.1 大数据的发展历史

Hadoop 项目诞生于 2005 年，其最初只是 Yahoo 公司用来解决网页搜索问题的一个项目，后来因其技术的高效性，被阿帕奇软件基金会（Apache Software Foundation）引入并成为开源应用。Hadoop 本身不是一个产品，而是由多个软件产品组成的一个生态系统。从技术上看，Hadoop 关键服务主要包括：采用 Hadoop 分布式文件系统（HDFS）的可靠数据存储服务；MapReduce 技术的高性能并行数据处理服务。这两项服务为实现结构化和复杂数据的快速、可靠分析奠定了基础。

2008 年年末，美国“计算社区联盟”（Computing Community Consortium）发表了一份具有影响力的白皮书——《大数据计算：在商务、科学和社会领域创造革命性突破》。它使人们的思维不仅仅局限于数据处理的机器，提出大数据的新用途和新见解。

2009 年，印度政府建立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研究如何利用手机和社交网站的数据源来分析预测从失业率到疾病暴发之类的问题。美国政府通过启动 Data.gov 网站的方式进一步开放数据大门，并向公众提供各种各样的政府数据。欧洲一些领先的研究型图书馆和科技信息研究机构建立伙伴关系，致力于在互联网上改善获取科学数据的简易性。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表长达 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。从经济界到科学界，从政府部门到艺术领域，很多方面都已经感受到这种巨量信息的影响，科学家和计算机工程师就此现象提出了“大数据”。

2011 年 2 月，IBM 的沃森超级计算机每秒扫描并分析 4TB（约 2 亿页的文字量）的数据，并在美国著名智力竞赛电视节目《危险边缘》（*Jeopardy*）上击败两名人类选手而夺冠。后来，《纽约时报》认为这一成果应归功于“大数据计算”的胜利。

2011 年 5 月，全球知名咨询公司麦肯锡全球研究院（McKinsey Global Institute，MGI）发布一份报告——《大数据：创新、竞争和生产力的下一个新领域》，由此大数据开始备受

人们关注。报告中指出，大数据已经渗透到当今每一个行业和业务的职能领域，并成为重要的生产因素。人们对海量数据的挖掘和运用，预示着新一波生产率的增长和消费者盈余浪潮的到来。报告还提到，“大数据”源于数据生产和收集的能力和速度的大幅提升，由于越来越多的人、设备和传感器通过数字网络连接起来，获取、传送、分享和访问数据的能力也发生了彻底的变革。

2011 年 12 月，工业和信息化部发布的《物联网“十二五”发展规划》中，信息处理技术作为四项关键技术的创新工程之一被提出来，其中包括海量数据存储、数据挖掘、图像视频智能分析，这些都是大数据的重要组成部分。

2012 年 1 月，瑞士达沃斯召开的世界经济论坛，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 3 月，美国政府在白宫网站发布《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；美国政府宣布投资 2 亿美元以推动大数据技术发展，是大数据技术从商业行为上升到国家科技战略的分水岭。大数据技术领域的竞争，事关国家的安全和未来。美国政府认识到国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及对数据的解释、运用能力；国家数字主权体现为对数据的占有和控制，数字主权将是继边防、海防、空防之后另一个大国博弈的空间。

2012 年 4 月，美国软件公司 Splunk 在纳斯达克成功上市，成为第一家上市的大数据处理公司。鉴于美国经济持续低迷、股市持续震荡的大背景，Splunk 首日股价暴涨一倍多。Splunk 是首家大数据监测和分析服务的软件提供商，其成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快大数据战略布局。

2012 年 7 月，联合国在纽约发布一份关于大数据政务的白皮书，总结各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求。其主要包括个人由于对价格的关注和更好服务的渴望，提供数据和众包信息，并对隐私和退出权提出需求；公共部门出于改善服务、提升效益的目的，提供诸如统计数据、设备信息、健康指标及税务和消费信息等，对隐私和退出权提出需求；私人部门出于提升客户认知和预测趋势的目的，提供汇总数据、消费和使用信息，并对敏感数据所有权和商业模式更加关注。白皮书还指出，人们如今可以使用丰富的数据资源，包括旧数据和新数据，对社会人口进行前所未有的实时分析。联合国还以爱尔兰和美国的社交网络活跃度增长作为失业率上升的早期征兆为例，说明政府如果能合理分析所掌握的数据资源，将能“与数俱进”，快速应变。

2012 年 7 月，为挖掘大数据的价值，阿里巴巴集团全面推进“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，阿里巴巴集团董事局主席马云提出，从 2013 年 1 月 1 日起，集团将转型，重塑平台、金融和数据三大业务。阿里巴巴集团希望通过资源共享和挖掘海量数据来创造价值。此举是国内企业把大数据提升到企业管理高度的一个重大里程碑。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。

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会变得越来越重要。全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到

大数据在推动经济发展、改善公共服务、增进人民福祉乃至保障国家安全方面的重大意义。

2014 年 5 月，美国发布《大数据：把握机遇，守护价值》白皮书，再次重申要把握大数据可为经济社会发展带来创新动力的重大机遇，同时也要高度警惕大数据应用所带来的隐私、公平等问题，以积极、务实的态度深刻剖析可能面临的治理挑战。

2017 年，中国大数据产业生态大会发布了《2017 中国大数据产业发展白皮书》，其中指出，与 2016 年相比，中国大数据产业最大的变化在于生态系统的完善。2016 年，我国大数据产业逐步形成了以京津冀、长三角、珠三角、中西部以及东北地区为集聚发展区的发展格局，产业生态日渐成熟，大数据产业增长迅速且产业规模持续放大。基础支撑层作为整个大数据产业链的核心环节，预计 2017 年的规模约为 2246 亿元，增长 68.2%；融合应用层作为大数据产业未来发展的着力点，预计 2017 年规模约为 16 998 亿元，增长率为 30.7%；数据服务层围绕各类大数据应用需求提供辅助性服务，预计 2017 年规模约为 326 亿元，增长率达到 60.6%。

2018 年 10 月，中国国际大数据大会聚焦大数据产业高质量发展，围绕“大数据与实体经济深度融合”，从生态完善、技术突破、融合应用、环境优化等维度进行了讨论，并且把大数据安全作为一个重要的研究领域。

1.2　大数据的基本特征

大数据，指无法在一定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利用它获得更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。IBM 提出大数据具有以下 5V 特征。

（1）海量的数据规模（volume）

第一个特征是数据量大，包括采集、存储和计算的量都非常大。大数据的起始计量单位至少是拍字节（PB）、艾字节（EB）或泽字节（ZB）等[①]。

（2）快速的数据流转和动态的数据体系（velocity）

数据增长速度快，处理速度也快，时效性要求高。比如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。

（3）多样的数据类型（variety）

种类和来源多样化，包括结构化、半结构化和非结构化数据，具体表现为网络日志、音频、视频、图片、地理位置信息等，多类型的数据对数据处理能力提出更高的要求。

（4）低价值密度（value）

随着互联网以及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的机器算法来挖掘数据价值，是大数据时代最需要解决的问题。

（5）真实性（veracity）

真实性表现为数据的准确性和可信赖度，即数据的质量。

① 1PB=1024TB，1EB=1024PB，1ZB=1024EB。