教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目
数据科学与大数据技术专业系列规划教材
华为信息与网络技术学院指定教材

Hadoop
集群程序设计与开发

王宏志 李春静 ◉ 编著

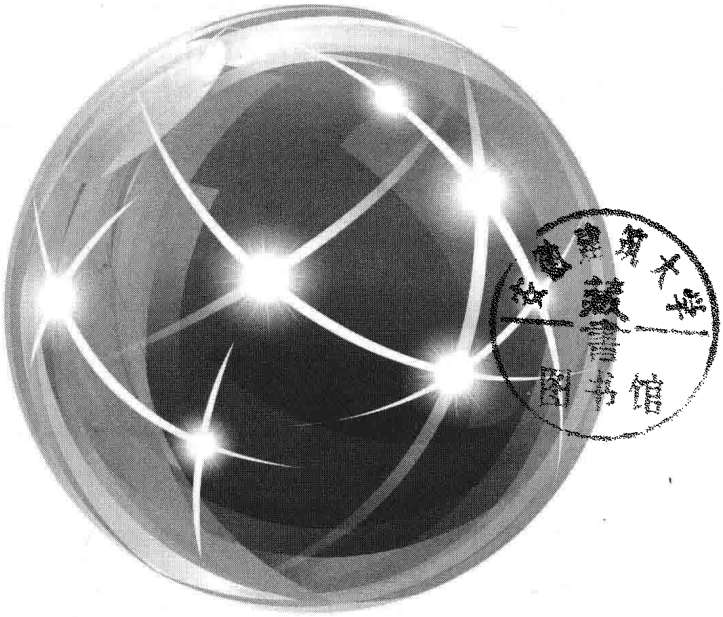
系统、完整的数据科学与大数据技术专业解决方案
名校名师打造大数据领域精品力作
全面讲授Hadoop生态与系统开发
系统原理与开发实战相结合

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目
数据科学与大数据技术专业系列规划教材
华为信息与网络技术学院指定教材

Hadoop
集群程序设计与开发

王宏志 李春静 ◉ 编著



人民邮电出版社
北京

图书在版编目（CIP）数据

Hadoop集群程序设计与开发 / 王宏志，李春静编著
. -- 北京 : 人民邮电出版社，2018.8
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-48304-1

Ⅰ. ①H… Ⅱ. ①王… ②李… Ⅲ. ①数据处理软件－程序设计－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第111464号

内 容 提 要

本书系统地介绍了基于 Hadoop 的大数据处理和系统开发相关技术，包括初识 Hadoop、Hadoop 基础知识、Hadoop 开发环境配置与搭建、Hadoop 分布式文件系统、Hadoop 的 I/O 操作、MapReduce 编程基础、MapReduce 高级编程、初识 HBase、初识 Hive。通过本书的学习，读者可以较全面地了解 Hadoop 的原理、配置和系统开发的相关知识，并且可以从 Hadoop 的角度学习分布式系统和 MapReduce 算法设计的相关知识。

本书可作为大数据技术相关专业本科生、研究生的教材，也可作为大数据技术的培训用书，还可作为大数据技术相关工作人员的参考用书。

◆ 编　　著　王宏志　李春静
策划编辑　戴思俊
责任编辑　李　召
责任印制　马振武

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京市艺辉印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：21.25　　2018 年 8 月第 1 版
字数：554 千字　　2018 年 8 月北京第 1 次印刷

定价：59.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目
数据科学与大数据技术专业系列规划教材

编　委　会

丛书序一 PREFACE

毫无疑问，我们正处在一个新时代。新一轮科技革命和产业变革正在加速推进，技术创新日益成为重塑经济发展模式和促进经济增长的重要驱动力量，而“大数据”无疑是第一核心推动力。

当前，发展大数据已经成为国家战略，大数据在引领经济社会发展中的新引擎作用更加突显。大数据重塑了传统产业的结构和形态，催生了众多的新产业、新业态、新模式，推动了共享经济的蓬勃发展，也给我们的衣食住行带来根本改变。同时，大数据是带动国家竞争力整体跃升和跨越式发展的巨大推动力，已成为全球科技和产业竞争的重要制高点。可以大胆预测，未来，大数据将会进一步激起全球科技和产业发展浪潮，进一步渗透到我们国计民生的各个领域，其发展扩张势不可挡。可以说，我们处在一个“大数据”时代。

大数据不仅仅是单一的技术发展领域和战略新兴产业，它还涉及科技、社会、伦理等诸多方面。发展大数据是一个复杂的系统工程，需要科技界、教育界和产业界等社会各界的广泛参与和通力合作，需要我们以更加开放的心态，以进步发展的理念，积极主动适应大数据时代所带来的深刻变革。总体而言，从全面协调可持续健康发展的角度，推动大数据发展需要注重以下五个方面的辩证统一和统筹兼顾。

一是要注重“长与短结合”。所谓“长”就是要目标长远，要注重制定大数据发展的顶层设计和中长期发展规划，明确发展方向和总体目标；所谓“短”就是要着眼当前，注重短期收益，从实处着手，快速起效，并形成效益反哺的良性循环。

二是要注重“快与慢结合”。所谓“快”就是要注重发挥新一代信息技术产业爆炸性增长的特点，发展大数据要时不我待，以实际应用需求为牵引加快推进，力争快速占领大数据技术和产业制高点；所谓“慢”就是防止急功近利，欲速而不达，要注重夯实大数据发展的基础，着重积累发展大数据基础理论与核心共性关键技术，培养行业领域发展中的大数据思维，潜心培育大数据专业人才。

三是要注重“高与低结合”。所谓“高”就是要打造大数据创新发展高地，要结合国家重大战略需求和国民经济主战场核心需求，部署高端大数据公共服务平台，组织开展国家级大数据重大示范工程，提升国民经济重点领域和标志性行业的大数据技术水平和应用能力；所谓“低”就是要坚持“润物细无声”，推进大数据在各行各业和民生领域的广泛应用，推进大数据发展的广度和深度。

四是要注重“内与外结合”。所谓“内”就是要向内深度挖掘和深入研究大数据作为一门学科领域的深刻技术内涵，构建和完善大数据发展的完整理论体系和技术支撑体系；所谓“外”就是要加强开放创新，由于大数据涉及众多学科领域和产业行业门类，也涉及国家、社会、个人等诸多问题，因此，需要推动国际国内科技界、产业界的深入合作和各级政府广泛参与，共同研究制定标准规范，推动大数据与人工智能、云计算、物联网、网络安全等信息技术领域的协同发展，促进数据科学与计算机科学、基础科学和各种应用科学的深度融合。

五是要注重“开与闭结合”。所谓“开”就是要坚持开放共享，要鼓励打破现有体制机制障碍，推动政府建立完善开放共享的大数据平台，加强科研机构、企业间技术交流和合作，推动大数据资源高效利用，打破数据壁垒，普惠数据服务，缩小数据鸿沟，破除数据孤岛；所谓“闭”就是要形成价值链生态闭环，充分发挥大数据发展中技术驱动与需求牵引的双引擎作用，积极运用市场机制，形成技术创新链、产业发展链和资金服务链协同发展的态势，构建大数据产业良性发展的闭环生态圈。

总之，推动大数据的创新发展，已经成为了新时代的新诉求。刚刚闭幕的党的十九大更是明确提出要推动大数据、人工智能等信息技术产业与实体经济深度融合，培育新增长点，为建设网络强国、数字中国、智慧社会形成新动能。这一指导思想为我们未来发展大数据技术和产业指明了前进方向，提供了根本遵循。

习近平总书记多次强调“人才是创新的根基”“创新驱动实质上是人才驱动”。绘制大数据发展的宏伟蓝图迫切需要创新人才培养体制机制的支撑。因此，需要把高端人才队伍建设作为大数据技术和产业发展的重中之重，需要进一步完善大数据教育体系，加强人才储备和梯队建设，将以大数据为代表的新兴产业发展对人才的创新性、实践性需求渗透融入人才培养各个环节，加快形成我国大数据人才高地。

国家有关部门“与时俱进，因时施策”。近期，国务院办公厅正式印发《关于深化产教融合的若干意见》，推进人才和人力资源供给侧结构性改革，以适应创新驱动发展战略的新形势、新任务、新要求。教育部高等学校计算机类专业教学指导委员会、华为公司和人民邮电出版社组织编写的《教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目——数据科学与大数据技术专业系列规划教材》的出版发行，就是落实国务院文件精神，深化教育供给

侧结构性改革的积极探索和实践。它是国内第一套成专业课程体系规划的数据科学与大数据技术专业系列教材，作者均来自国内一流高校，且具有丰富的大数据教学、科研、实践经验。它的出版发行，对完善大数据人才培养体系，加强人才储备和梯队建设，推进贯通大数据理论、方法、技术、产品与应用等的复合型人才培养，完善大数据领域学科布局，推动大数据领域学科建设具有重要意义。同时，本次产教融合的成功经验，对其他学科领域的人才培养也具有重要的参考价值。

我们有理由相信，在国家战略指引下，在社会各界的广泛参与和推动下，我国的大数据技术和产业发展一定会有光明的未来。

是为序。

中国科学院院士　郑志明

2018 年 4 月 16 日

丛书序二 PREFACE

在500年前的大航海时代，哥伦布发现了新大陆，麦哲伦实现了环球航行，全球各大洲从此连接了起来，人类文明的进程得以推进。今天，在云计算、大数据、物联网、人工智能等新技术推动下，人类开启了智能时代。

面对这个以“万物感知、万物互联、万物智能”为特征的智能时代，“数字化转型”已是企业寻求突破和创新的必由之路，数字化带来的海量数据成为企业乃至整个社会最重要的核心资产。大数据已上升为国家战略，成为推动经济社会发展的新引擎，如何获取、存储、分析、应用这些大数据将是这个时代最热门的话题。

国家大数据战略和企业数字化转型成功的关键是培养多层次的大数据人才，然而，根据计世资讯的研究，2018年中国大数据领域的人才缺口将超过150万人，人才短缺已成为制约产业发展的突出问题。

2018年初，华为公司提出新的愿景与使命，即“把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界”，它承载了华为公司的历史使命和社会责任。华为企业BG将长期坚持“平台+生态”战略，协同生态伙伴，共同为行业客户打造云计算、大数据、物联网和传统ICT技术高度融合的数字化转型平台。

人才生态建设是支撑“平台+生态”战略的核心基石，是保持产业链活力和持续增长的根本，华为以ICT产业长期积累的技术、知识、经验和成功实践为基础，持续投入，构建ICT人才生态良性发展的使能平台，打造全球有影响力的ICT人才认证标准。面对未来人才的挑战，华为坚持与全球广大院校、伙伴加强合作，打造引领未来的ICT人才生态，助力行业数字化转型。

一套好的教材是人才培养的基础，也是教学质量的重要保障。本套教材的出版，是华为在大数据人才培养领域的重要举措，是华为集合产业与教育界的高端智力，全力奉献的结晶和成果。在此，让我对本套教材的各位作者表示由衷的感谢！此外，我们还要特别感谢教育部高等学校计算机类专业教学指导委员会副主任、北京大学陈钟教授以及秘书长、北京航空航天大学马殿富教授，没有你们的努力和推动，本套教材无法成型！

同学们、朋友们，翻过这篇序言，开启学习旅程，祝愿在大数据的海洋里，尽情展示你们的才华，实现你们的梦想！

华为公司董事、企业BG总裁　阎力大

2018年5月

前言 FOREWORD

本书的缘起与成书过程

Hadoop 是最早得到广泛使用的大数据计算平台之一，也是目前生态系统最完整、参与开发人员最多的大数据计算平台之一。尽管 Hadoop 有诸多的竞争者，但其具有的开发配置便利、可扩展性好、生态完整等优点，使其得到了大数据开发人员的普遍认同，成为许多大数据应用系统的基础软件平台。而 Hadoop 本身也在不断演化中，新近发布的 Hadoop 3.1 开始支持 GPU 和 FPGA。

鉴于此，笔者认为，尽管大数据计算平台很多，但 Hadoop 是其中非常重要的一种，也是学习门槛相对比较低的一种，读者学习之后可以快速地掌握大数据处理与系统设计。笔者撰写本书的出发点是提供一本适用于学习 Hadoop 平台安装、设置与系统开发的教材。

在撰写本书的过程中，有专家提出，本书的内容有些偏重于应用，不适合作为学历教育的教材。因此，笔者进一步增加了分布式系统的介绍、对 MapReduce 程序设计方法的介绍及对一些改进 Hadoop 的技术的介绍，期望读者可以以 Hadoop 为范例，学习分布式系统的相关知识，并且可以通过本书初步学习 MapReduce 程序设计，以使本书的深度适应学历教育，这就是本书现在的版本。

本书的内容

本书对基于 Hadoop 大数据处理与开发进行了系统的介绍，同时还对 Hadoop 系统原理、Hadoop 开发环境配置与搭建、Hadoop 系统开发相关知识、MapReduce 程序设计、HBase 和 Hive 配置与开发等相关知识进行了介绍。考虑到读者的多样性，书中对于不同的内容采取了不同的介绍方式。

在原理部分，主要突出了 Hadoop 的原理介绍，并且介绍了一些延伸 Hadoop 技术和网络与分布式系统的背景知识。通过这一部分内容的学习，读者可以较为深入地了解 Hadoop 及其相关组件（HDFS）的运行原理。从理论学习的角度来说，Hadoop 是一个典型的分布式文件管理与计算系统，对其进行深入剖析，读者会加深对分布式系统诸多概念的认知；从实践的角度来说，了解 Hadoop 的原理，对系统的配置、运维、调优及高效程序的设计，都非常有帮助。

系统开发环境配置与搭建部分则更加面向实战，这一部分通过实例讲解了系统安装、部署、环境搭建、配置、应用程序部署等一系列过程，帮助读者搭建 Hadoop 开发环境。

由于 Hadoop 的核心是处理“数据”，因而 Hadoop 系统开发相关知识部分着重介绍了 HDFS 和 I/O 操作，使读者能较为深入地了解这两部分的相关技术。

基于 Hadoop 的系统开发需要 MapReduce 程序设计，本书介绍了 MapReduce 程序设计和算法设计。读者通过学习，可以掌握利用 MapReduce 编程模式解决计算问题的方法，从而能够根据需求为基于 Hadoop 的大数据计算系统设计有效的程序。

在 Hadoop 的大数据系统中，数据管理扮演着重要角色。本书介绍了基于 Hadoop 的数据管理系统，即 HBase 和 Hive 相关知识。读者通过学习，可以掌握 HBase 和 Hive 的配置、使用及相关程序的设计方法。

本书试图以 Hadoop 为主线，兼顾理论与实战，较全面地介绍可操作的大数据平台配置与系统开发的相关知识，和大数据算法、大数据分析、大数据系统等图书具有互补性，可以相互参考。

本书的适用对象

本书适合作为本科生和研究生“Hadoop 系统程序设计”“大数据系统开发”等课程的教材，也可以作为“高级语言程序设计”“分布式系统”“数据库系统”等课程的补充教材或课外读物。同时，本书也可供大数据领域从业人员参考。

致使用本书的教师

本书涉及多方面内容，对于教学而言，本书适用于多门课程，除了直接用于“Hadoop 系统程序设计”“大数据系统开发”等课程之外，还可以作为“分布式系统”“数据库系统”“高级语言程序设计”等课程的补充教材。教师可以根据这些课程的具体内容补充学习内容。

不同层次的教学可以从本书选择不同的内容。偏重系统原理的教学，可以着重讲授本书的第 2 章，而偏重应用的教学，则可以略讲这一部分；偏重程序设计的教学，可以着重讲授第 6～7 章，而偏重系统运维的教学，则可以略讲这一部分；偏重原理和程序设计的教学，可以把第 3～4 章留给学生自学而不需要详细讲解，而偏重应用的教学，则需要详细讲解这两章。

致使用本书的学生

希望本书为学生提供比较全面的基于 Hadoop 的大数据处理与系统开发的相关知识，本书不同部分需要的背景知识不尽相同。例如，第 2 章对 Hadoop 原理的介绍，需要一部分分布式系统和操作系统的背景知识；第 4～7 章的学习，需要一些 Java 语言的相关知识；如果读者学习过数据库系统相关知识，则比较容易学习第 8～9 章的内容。

为了帮助读者理解本书内容，对一些读者可能不容易理解的概念，本书以“学习提示”的形式进行了介绍；同时也对一些分布式系统的知识进行了简要介绍。

致使用本书的专业技术人员

本书可以作为一本 Hadoop 大数据处理和系统开发的参考书，供专业技术人员参考。各部分内容针对的人群有所不同，可以单独查阅涉及的主题。对于相关的知识（包括库函数和语法），本书尽量提供比较全面的列表，供专业人员查阅之用。同时对系统的安装和配置，提供了尽可能详细的步骤，供读者在安装和配置系统时参考。

致谢

首先感谢本书的共同作者李春静老师，本书的大部分内容来源于李春静老师多年的教学实践，这使本书更加贴近实战。

感谢哈尔滨工业大学的李建中教授、高宏教授及国际大数据计算研究中心的诸位同事，对本书内容、表述给予的指导和建议，以及在专业上的帮助。

在本书的撰写过程中，哈尔滨工业大学的张梦、孟凡山等同学在资料翻译、搜集、整理、文本校对、作图等多个方面提供了帮助和支持，在此表示感谢。

非常感谢我的爱人黎玲利副教授，感谢她一直以来对我的支持，以及她在大数据相关课程授课的过程中对本书提出的一系列有益的建议。感谢我的母亲和岳母，在本书写作期间，她们料理家务，照顾我的宝宝“壮壮”茁壮成长，使我有时间从事本书的写作。

本书的写作得到了“教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目”的资助，感谢华为公司的张志峰、刘洁在本书成书过程中提供的帮助。

还要感谢在哈尔滨工业大学选修我讲授的“大数据管理与分析”课程的同学，他们给我的意见和建议对本书的写作大有裨益。

由于 Hadoop 一直处于演化之中，本书涉及的内容也比较广泛，限于笔者的水平，本书在内容安排、表述等方面存在着各种不当之处，敬请读者在阅读本书的过程中，不吝提出宝贵建议，以期共同改进本书。读者的任何意见和建议请发邮件至 wangzh@hit.edu.cn。此外，读者如果想了解更多关于数据科学与大数据技术方向的科学研究、专业建设、人才培养及教学资源等信息，可关注作者公众号“大数据与数据科学家”。

最后，笔者关于大数据管理和分析方面的研究和本书的写作，还得到了国家自然科学基金项目（编号：U1509216，61472099）、国家重点研发计划项目（编号：2016YFB1000703）、国家科技支撑计划项目（编号：2015BAH10F01）、黑龙江省留学回国人员基金（编号：LC2016026）和微软-教育部语言语音重点实验室的经费资助，在此一并表示感谢。

王宏志

2018 年 4 月于哈尔滨

目 录 CONTENTS

01 第1章 初识Hadoop

【内容概述】

本章通过与目前较热门的云计算、大数据技术做对比的方式，从实际应用的角度来介绍 Hadoop。主要包括云计算介绍、大数据介绍及 Hadoop 相关项目人才要求三大块内容。

【知识要点】

- 了解 Hadoop 产生过程、应用场景
- 理解云计算、大数据概念及 Hadoop 与它们的关系
- 了解 Hadoop 学习过程及目前 Hadoop 人才需求情况