普通高等教育“十三五”规划教材
国家新闻出版改革发展项目库入库项目
数据科学与大数据技术专业教材丛书

大数据技术基础

BIG DATA TECHNOLOGY FOUNDATION

鄂海红　宋美娜　欧中洪◎编著

北京邮电大学出版社
www.buptpress.com

普通高等教育“十三五”规划教材
国家新闻出版改革发展项目库入库项目
数据科学与大数据技术专业教材丛书

大数据技术基础

鄂海红　宋美娜　欧中洪　编著

北京邮电大学出版社
www.buptpress.com

内容简介

本书围绕大数据技术基础，重点介绍了大数据存储系统（分布式文件系统和 NoSQL 数据库）、大数据处理框架（Hadoop 的 MapReduce、Spark 及实时处理框架 Storm 和 Flink）、大数据仓库技术（Hive、Druid 等）、大数据多维分析（Kylin）、大数据可视化技术和大数据综合应用等，以及当今主流的大数据平台构建技术和开源组件实践知识，可以指导读者全面、系统地掌握大数据各层的实现方案，开展各领域的大数据实践。本书可作为计算机学科相关专业，特别是数据科学与大数据技术专业的教材。

图书在版编目(CIP)数据

大数据技术基础 / 鄂海红，宋美娜，欧中洪编著. -- 北京：北京邮电大学出版社，2019.9
ISBN 978-7-5635-5878-0

Ⅰ. ①大… Ⅱ. ①鄂… ②宋… ③欧… 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（2019）第 204848 号

书　　名：大数据技术基础
作　　者：鄂海红　宋美娜　欧中洪
责任编辑：孙宏颖
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：保定市中画美凯印刷有限公司
开　　本：787 mm×1 092 mm　1/16
印　　张：15.5
字　　数：401 千字
版　　次：2019 年 9 月第 1 版　2019 年 9 月第 1 次印刷

ISBN 978-7-5635-5878-0　　定价：48.00 元

大数据顾问委员会

宋俊德　王国胤　张云勇　郑　宇

段云峰　田世明　娄　瑜　孙少陵

王　柏

大数据专业教材编委会

总主编： 吴　斌

编　委： 宋美娜　欧中洪　鄂海红　双　锴

于艳华　周文安　林荣恒　李静林

袁燕妮　李　劼　皮人杰

总策划： 姚　顺

秘书长： 刘纳新

序　言

党的十八届五中全会明确提出实施国家大数据战略，至此大数据技术成为塑造国家竞争力的战略制高点之一。掌握和运用大数据技术的能力成为一个国家竞争力的重要体现。国内许多行业如互联网、电信、金融和交通等开始实际部署大数据平台并付诸实践，这带动了软件、硬件及服务市场的快速发展。

大数据正在成为产业发展的重要推动力，大数据相关产业的高速发展带来了大数据人才严重短缺的问题，大数据人才的培养成为当前急迫的任务。近年来大数据专业建设在全国各大高校如火如荼地开展，设立该专业的学校数量也在快速增长。截止到 2019 年 4 月，教育部累计批准 486 所高校设立“数据科学与大数据技术”专业，其中，2016 年 3 所高校获批，2017 年 32 所高校获批，2018 年 248 所高校获批，2019 年 203 所高校获批。如何更好地建设大数据专业和培养产业迫切需求的高水平专业人才，成为高校人才培养工作的重要挑战。

自 2018 年起，北京邮电大学出版社联合北京邮电大学计算机学院、网络技术研究院的多位知名教授、副教授及任课教师，共同开启“数据科学与大数据技术专业教材丛书”的出版工作。这套丛书包括《大数据技术基础》《大数据技术基础实验》《R 语言编程与数据科学》《网络科学与计算》《计算机视觉》《NoSQL 数据库技术》《流数据分析技术》《数据可视化》《机器学习》《分布式计算与云计算》《数据仓库与数据挖掘》《Python 语言程序设计》等。这些教材的出版凝炼了众多大数据领域教学、科研专家的心得体会，为大数据创新型人才的培养奠定了基础。

《大数据技术基础》是“数据科学与大数据技术”专业重要的基础教材之一，主要讲授大数据知识体系中理论与工程实践结合的技术基础。该书涵盖大数据采集、存储、处理、分析、可视化及应用等一整套全流程所需的基础理论知识。为了使读者能够快速地掌握大数据工程实践的知识，书中还介绍了多种开源大数据实践工具组件的技术架构和使用方法。可以说，该书所设计的内容一方面体现了对学生理论知识培养的重视，另一方面强调了计算机专业背景下数据科学的系统观，注重学生实际应用能力的培养。

该书的作者一直在大数据领域从事一线的教学和科研工作，这些工作基础为大数据专业人才的培养和大数据专业教材的出版提供着有力的支撑。该书作为北京邮电大学计算机学院“数据科学与大数据技术”专业的第一批正式出版教材，我很期待在以后的教学和科研实践中该书能够得到不断升华，也恳请全国同行在使用该书的同时予以批评指正，让我们一起为中国的大数据事业添砖加瓦。

国家“万人计划”领军人才
北京邮电大学计算机学院执行院长
博士生导生、教授
苏　森

前　言

本书一共分为 9 章。

第 1 章为大数据概述。本章首先介绍了大数据的发展历程、大数据的定义与特征、大数据与传统数据的区别；然后介绍了大数据平台应具备的能力和大数据平台架构；最后介绍了 Hadoop生态开源组件和大数据技术的应用领域。

第 2 章为大数据存储技术。本章主要介绍主流的分布式存储系统，包括相关概念、体系结构、存储机制和操作方法，主要涵盖了分布式文件系统 HDFS 以及 4 种 NoSQL 数据库。

第 3 章、第 4 章、第 5 章为大数据处理技术。第 3 章介绍了 Hadoop 的 MapReduce 并行计算框架，第 4 章介绍了 Spark 内存计算框架，第 5 章介绍了实时计算框架。

第 6 章为大数据仓库技术。本章介绍了分布式数据仓库和数据查询技术，主要包括 3 个组件：Hive 分布式数据仓库、Druid 时序数据仓储和 Drill 分布式实时查询。

第 7 章为大数据多维分析技术。本章的主要内容包括大数据多维分析技术演进的需求和背景、开源 Kylin 的基本概念与原理、技术架构和实战操作方法。

第 8 章为大数据可视化技术。本章详细介绍了数据可视化的定义及其分类、可视化流程，以及时空数据可视化、层次和网络数据可视化、文本和文档可视化的概念，并对商业智能中的数据可视化及其应用进行了介绍；同时讲解了常见的数据可视化的实现技术和方法。

第 9 章为大数据应用案例。本章选择了某电影大数据平台案例，结合某电影大数据平台的技术体系架构，对大数据应用的构建流程进行了介绍，可以帮助读者整体性地理解和掌握本书知识内容的实践方法。

本书可以作为数据科学与大数据技术专业的本科高年级专业课教材，也可以作为研究生相关课程的参考材料。同时本书还配套了《大数据技术基础实验》，用于指导读者学习具体的实践课程知识，以使读者掌握实际大数据平台和大数据应用系统的研发能力。

本书的编写得到了北京邮电大学 PCN&CAD 中心、教育部信息网络工程研究中心和北

京邮电大学计算机学院数据科学与服务中心教师与研究生的支持，他们分别是宋美娜、欧中洪、宋俊德、毕秋波、韩鹏昊、田川、孔慧慧、赵淑晨、吴金盛、温宇飞、万仁山、谭泽华、陈小康、韦帅丽、朱永波，在此一并表示感谢。

感谢国家重点研发计划项目“大数据征信及智能评估技术”和“基于大数据的科技咨询技术与服务平台研发”、国家科技条件平台计划项目“国家人类遗传资源共享服务平台北京创新中心建设”的支持。

作者作为在计算机领域从事科研和教学的教师，由于在专业知识的深度和广度上的局限性使得本书存在不足之处，欢迎广大读者反馈对本书的意见和建议，我们将随着“大数据技术基础”专业课程的建设，不断地改进本书的质量。

鄂海红
于北京

目　录

第 1 章

大数据概述

本章思维导图

随着数据的爆炸式增长和计算机技术的迅速发展，大数据技术迎来了前所未有的发展，它使人们的生活发生新变化的同时，也给人们带来了许多挑战，包括如何存储、查询、计算这些海量数据等，因此构建一个统一的大数据平台显得尤为重要。目前业界普遍认为大数据平台应具有数据源、数据采集、存储、处理、分析、可视化及其应用这 6 个层次。Hadoop 作为一个开源的大数据平台，目前已成为大数据领域的技术标准，它具有高可靠性、高扩展性、高效性和高容错性等优点，这些优点使它能应对大数据领域的大部分问题。

本章首先介绍了大数据的发展历程、大数据的定义与特征、大数据与传统数据的区别，使读者对大数据概念有个整体的了解；然后介绍了大数据平台应具备的能力和大数据平台架构，使读者对大数据平台的架构有大体的轮廓；接着介绍了 Hadoop 生态系统，使读者能够认识其基础的组件；最后介绍了大数据应用，使读者能了解目前现实生活中大数据应用的例子。本章思维导图如图 1-0 所示。

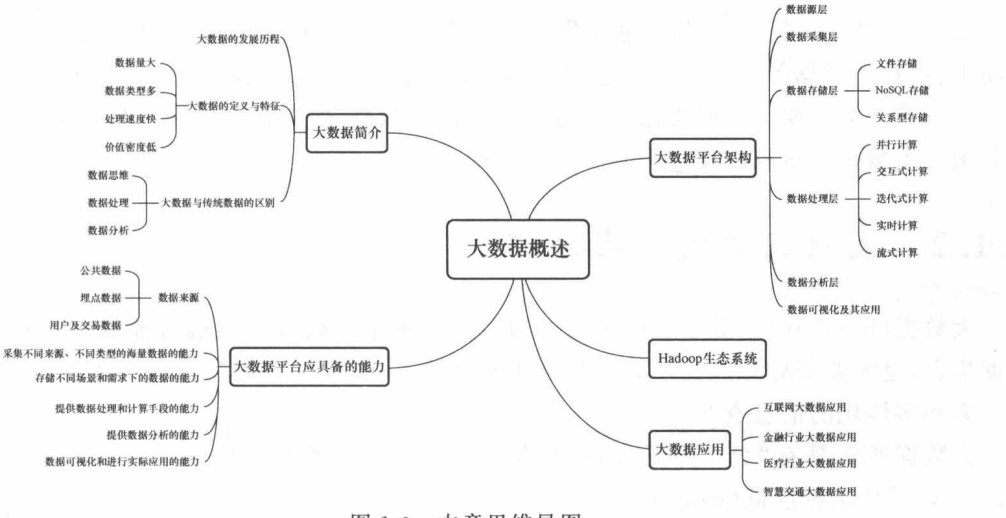

图 1-0　本章思维导图

1.1 大数据简介

21世纪以来，随着计算机技术，尤其是互联网和移动技术的发展，使得数据规模呈爆炸性增长，因此“大数据”概念应运而生。大数据是继云计算、物联网之后信息技术产业领域的又一重大技术革新，它使人们的生活发生了新的变化。本节首先帮助读者更好地认识和了解大数据的发展历程、大数据的定义与特征以及大数据与传统数据的区别等[1]。

1.1.1 大数据的发展历程

2005年Hadoop项目诞生。Hadoop是由多个软件产品组成的一个生态系统，这些软件产品共同实现全面功能和灵活的大数据分析[2]。

2008年9月，*Nature*推出*Big Data*专刊[3]，并邀请一些研究人员和企业家预测大数据所带来的革新。同年，计算社区联盟发表了报告“Big-data computing: creating revolutionary breakthroughs in commerce, science, and society”[4]，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术以及大数据在商业、科研和社会领域所面临的一些挑战。

2011年2月，*Science*推出*Dealing with Data*专刊[5]，该专刊围绕着科学研究中大数据的问题展开讨论。麦肯锡公司在同年5月份发布了“Big data: the next frontier for innovation, competition, and productivity”[6]，对大数据的影响、关键技术和应用领域等进行了详细的介绍。

拓展阅读

工业和信息化部——《大数据产业发展规划（2016—2020年）》

2012年3月，美国政府在白宫网站发布了“Big data research and development initiative”，这一举动标志着大数据已经成为重要的时代特征[7]。同年7月，联合国在纽约发布了一本关于大数据政务的白皮书*Big Data for Development: Opportunities & Challenges*[8]，标志着全球大数据的研究和发展进入了前所未有的高潮阶段。

2014年，“大数据”一词首次写入我国《政府工作报告》，报告中指出，要设立新兴产业创业创新平台，在大数据等方面赶超先进，引领未来产业发展[2]。2017年12月，习近平主席在中共中央政治局第二次会议时提出“实施国家大数据战略 加快建设数字中国”的目标，这代表着我国对大数据的重视程度上升到了一个新的高度[9]。

1.1.2 大数据的定义与特征

大数据(big data)是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具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产[10]。

大数据通常具有“4V”特征，即数据量大(volume)、数据类型多(variety)、处理速度快(velocity)和价值密度低(value)[7]。

① 数据体量庞大。采集、存储和计算的量都非常大。数据时代刚刚来临的时候，一般的数