大数据创新人才培养系列

大数据技术原理与应用

第2版

概念、存储、处理、分析与应用

PRINCIPLES AND APPLICATIONS OF BIG DATA TECHNOLOGY (2ND)

◎ 林子雨 编著

搭建起通向“**大数据知识空间**”的桥梁和纽带

构建知识**体系**、阐明基本**原理**、引导初级**实践**、介绍相关**应用**

为读者在大数据领域“**深耕细作**”奠定基础、指明方向

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

大数据创新人才培养系列

大数据技术原理与应用 第2版

概念、存储、处理、分析与应用

PRINCIPLES AND APPLICATIONS OF BIG DATA TECHNOLOGY (2ND)

◎ 林子雨 编著

人民邮电出版社

北京

图书在版编目（CIP）数据

大数据技术原理与应用 ：概念、存储、处理、分析与应用 / 林子雨编著. -- 2版. -- 北京 ：人民邮电出版社，2017.1
（大数据创新人才培养系列）
ISBN 978-7-115-44330-4

Ⅰ. ①大… Ⅱ. ①林… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第313809号

内 容 提 要

本书系统介绍了大数据的相关知识，分为大数据基础篇、大数据存储与管理篇、大数据处理与分析篇、大数据应用篇。全书共 15 章，内容包含大数据的基本概念、大数据处理架构 Hadoop、分布式文件系统 HDFS、分布式数据库 HBase、NoSQL 数据库、云数据库、MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学领域和其他行业的应用。本书在 Hadoop、HDFS、HBase、MapReduce 和 Spark 等重要章节安排了入门级的实践操作，以便读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考。

◆ 编　　著　林子雨
责任编辑　吴　婷
责任印制　沈　蓉　彭志环

◆ 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
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：18.75　　2017 年 1 月第 2 版
字数：487 千字　　2017 年 1 月河北第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前　言（第2版）

《大数据技术原理与应用》第1版于2015年8月出版，虽然距今仅有一年左右的时间，但是在过去一年里，大数据技术发展迅猛，诸如Spark等新技术迅速崛起，开始改变Hadoop一枝独秀的市场格局。因此，我们及时对第1版内容进行了补充和修订，以适应大数据技术的快速发展，保持本书的先进性和实用性。

本书依然沿用第1版的篇章设计，共分四大部分，包括大数据基础篇、大数据存储与管理篇、大数据处理与分析篇和大数据应用篇。在大数据基础篇中，第1章介绍大数据的基本概念和应用领域，并阐述大数据、云计算和物联网的相互关系；第2章介绍大数据处理架构Hadoop，并补充介绍了Hadoop版本演化。在大数据存储与管理篇中，第3章介绍了分布式文件系统HDFS，在编程实践部分根据最新版本的API进行了修订；第4章介绍了分布式数据库HBase，在编程实践部分根据最新版本的API进行了修订；第5章介绍了NoSQL数据库；第6章介绍了云数据库。在大数据处理与分析篇中，首先在第7章介绍了分布式并行编程模型MapReduce，然后在新增的第8章中对Hadoop进行了再探讨，介绍了Hadoop的发展演化和一些新特性，并在新增的第9章中介绍了当前比较热门的、基于内存的分布式计算框架Spark，在第10章和第11章分别介绍了两种典型的大数据分析技术——流计算和图计算，最后在第12章简单介绍了可视化技术。在大数据应用篇中，用3章（第13章～第15章）内容介绍了大数据在互联网、生物医学领域和其他行业的典型应用。

本书第1版于2015年8月出版后，厦门大学数据库实验室建设了与本书配套的“中国高校大数据课程公共服务平台”（http://dblab.xmu.edu.cn/post/bigdata-teaching- platform/），为教师教学和学生学习大数据课程提供PPT讲义、学习指南、备课指南、上机习题、实验指南、技术资料、授课视频等全方位、一站式免费服务，并提供面向全国高校的大数据实验平台建设方案和大数据课程师资培训服务。

本书是厦门大学计算机科学系大数据课程的配套教材，根据近几年的教学实践，建议安排32学时理论课，16个教学周，每周2学时。每章的具体学时分配如下：第1、3、4、5、6、8、10、11、12、13章每章安排2学时；第2、7、9章每章安排4学时；第14、15章这两章内容由学生自学完成。已经建设大数据教学实验室的高校，可以增加16学时上机实践课，分成4次上机，每次连续4节课，“中国高校大数据课程公共服务平台”的“教师服务站”为本书提供了配套的上机实验指南。

本书第1版出版后，笔者收到了大量的读者来信，对本书提出了许多宝贵的改进意见和建议，这里表示衷心的感谢。同时，笔者举办了多期全国高校大数据课程教师培训交流班和全国高校大数据教学论坛，开展了全国高校大数据公开课巡讲计划与辅助国内高校开设大数据课程公益项目，建立了大数据课程教师交流群，与全国高校大数据课程教师进行了广泛的接触、沟通和交流，更好地了解了当前国内高校

大数据课程教学发展需求和前进方向，这也为本书第2版撰写奠定了很好的基础。这里向参与交流的全国高校大数据课程教师表示衷心的感谢！

本书由林子雨执笔。在撰写第2版过程中，厦门大学计算机科学系硕士研究生蔡珉星、李雨倩、谢荣东、罗道文、邓少军、阮榕城、薛倩、魏亮、曾冠华等做了大量辅助性工作，在此，向他们的辛勤工作表示衷心的感谢。

大数据技术发展日新月异，在今后的工作中，笔者以及厦门大学数据库实验室会持续跟踪大数据技术发展趋势，把大数据最新技术和本书相关补充资料及时发布到“中国高校大数据课程公共服务平台”，方便本书读者通过网络及时免费获取相关信息。由于笔者能力有限，书中难免存在不足之处，望广大读者不吝赐教。

林子雨

厦门大学计算机科学系数据库实验室

2016年9月

前　言
（第1版）

大数据作为继云计算、物联网之后IT行业又一颠覆性的技术，备受人们关注。大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业，都融入了大数据的印迹，大数据对人类的社会生产和生活必将产生重大而深远的影响。

大数据时代的到来，迫切需要高校及时建立大数据技术课程体系，为社会培养和输送一大批具备大数据专业素养的高级人才，满足社会对大数据人才日益旺盛的需求。本书定位为大数据技术入门教材，为读者搭建起通向“大数据知识空间”的桥梁和纽带。本书将系统梳理、总结大数据相关技术，介绍大数据技术的基本原理和大数据的主要应用，帮助读者形成对大数据知识体系及其应用领域的轮廓性认识，为读者在大数据领域“深耕细作”奠定基础、指明方向。在本书的基础上，感兴趣的读者可以通过其他诸如《Hadoop权威指南》等工具书，继续深入学习和实践大数据相关技术。

本书紧紧围绕“构建知识体系，阐明基本原理，引导初级实践，了解相关应用”的指导思想，对大数据知识体系进行系统梳理，做到“有序组织、去粗取精、由浅入深、渐次展开”。本书共分四大部分，包括大数据基础篇、大数据存储篇、大数据处理与分析篇和大数据应用篇。在大数据基础篇中，第1章介绍大数据的基本概念

和应用领域，并阐述大数据、云计算和物联网的相互关系；第2章介绍大数据处理架构Hadoop，由于Hadoop已经成为应用最广泛的大数据技术，因此，本书的大数据相关技术主要围绕Hadoop展开，包括Hadoop MapReduce、HDFS和HBase，本章是第3、4、7章的基础。在大数据存储篇中，用4章（第3～6章）的内容介绍了大数据存储相关技术的概念与原理，包括分布式文件系统（HDFS）、分布式数据库（HBase）、NoSQL数据库和云数据库。在大数据处理与分析篇，首先在第7章介绍了大数据处理和分析的核心技术——分布式并行编程模型MapReduce，然后，在第8章和第9章分别介绍了大数据时代两种新兴的数据分析技术——流计算和图计算，最后在第10章简单介绍了可视化技术。在大数据应用篇，用3章（第11章～第13章）内容介绍了大数据在互联网、生物医学领域和其他行业的典型应用。

本书面向高校计算机和信息管理等相关专业的学生，可以作为专业必修课或选修课的教材。在教学过程中，建议安排32学时，16个教学周，每周2学时。每章的具体学时分配如下：第1、2、5、6、8、10、11章每章安排2学时；第3、4、9章每章安排4学时；第7章安排6学时；第12、13章这两章内容由学生自学完成。

本书由林子雨执笔。在撰写过程中，厦门大学计算机科学系硕士研究生刘颖杰、叶林宝、蔡珉星、李雨倩、谢荣东、罗道文以及本科生黄梓铭、李粲等做了大量辅助性工作，在此，向这些同学的辛勤工作表示衷心的感谢。

本书官方网站是http://dblab.xmu.edu.cn/post/bigdata，提供教学PPT和相关资料的下载，并接受错误反馈和发布教材勘误信息。厦门大学数据库实验室为本教材配套建设了国内高校首个大数据课程公共服务平台（http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/），为教师教学和学生学习大数据课程提供讲义PPT、学习指南、备课指南、上机习题、技术资料、授课视频等全方位、一站式免费服务。

本书在撰写过程中，参考了大量国内外的教材、专著、论文和资料，对大数据知识进行了系统梳理，有选择地把一些重要知识纳入本书。本书也是笔者多年在数据科学领域从事教学、科研、产业方面工作的系统总结。由于笔者能力有限，本书难免存在不足之处，望广大读者不吝赐教。

林子雨
厦门大学计算机科学系数据库实验室
2015年3月

作者介绍

林子雨（1978－），男，博士，厦门大学计算机科学系助理教授，厦门大学云计算与大数据研究中心创始成员，海峡云计算与大数据应用研究中心副主任。于2001年获得福州大学水利水电专业学士学位，2005年获得厦门大学计算机专业硕士学位，2009年获得北京大学计算机专业博士学位。中国高校首个“数字教师”提出者和建设者（http://www.cs.xmu.edu.cn/linziyu），2009年至今，“数字教师”大平台累计向网络免费发布超过100万字高价值的教学和科研资料，累计网络访问量超过100万次。

主要研究方向为数据库、数据仓库、数据挖掘、大数据和云计算，发表期刊和会议学术论文多篇，并作为课题组负责人承担了国家自然科学基金和福建省自然科学基金项目。曾作为志愿者翻译了Google Spanner、BigTable和《Architecture of a Database System》等大量英文学术资料，与广大网友分享，深受欢迎。2013年在厦门大学开设大数据课程，并因在教学领域的突出贡献和受到学生的认可，成为2013年度厦门大学教学类奖教金获得者。

主讲课程：《大数据处理技术》。

个人主页：http://www.cs.xmu.edu.cn/linziyu。

E-mail: ziyulin@xmu.edu.cn。

数据库实验室网站：http://dblab.xmu.edu.cn。

建设了中国高校大数据课程公共服务平台（http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/），为教师教学和学生学习大数据课程提供包括教学大纲、讲义PPT、学习指南、备课指南、实验指南、上机习题、授课视频、技术资料等全方位、一站式免费服务，平台年访问量超过50万次，同时提供面向高校的大数据实验平台建设方案和大数据课程师资培训服务。

中国高校大数据课程
公 共 服 务 平 台

扫一扫访问平台主页

目 录

第一篇 大数据基础

第二篇　大数据存储与管理

第三篇 大数据处理与分析

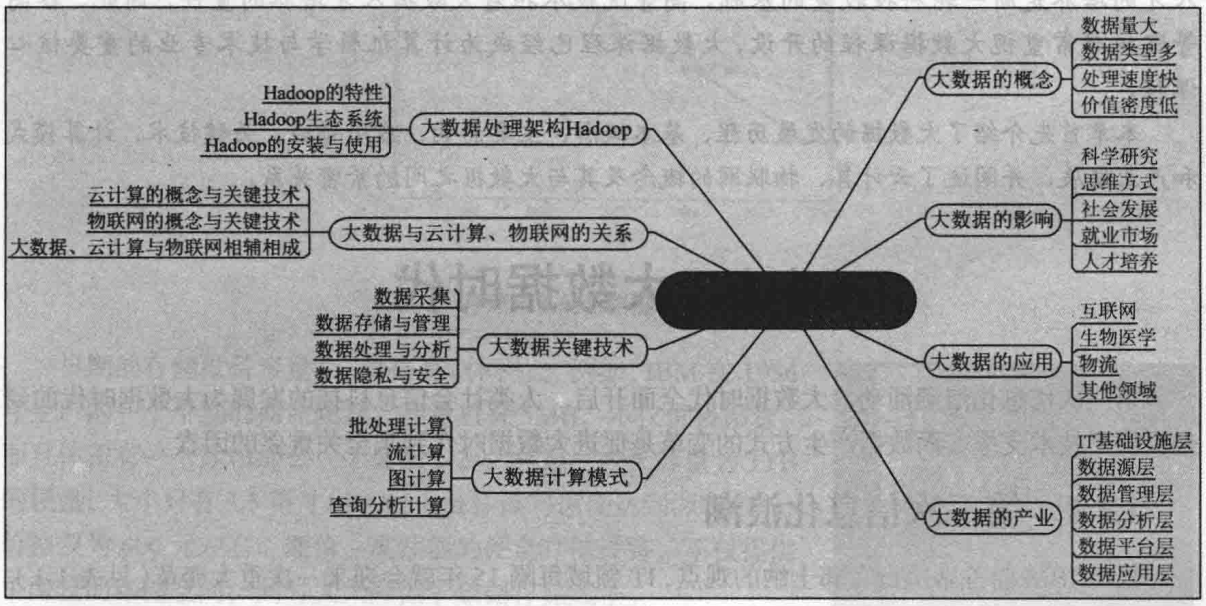
第一篇
大数据基础

本篇内容

本篇介绍大数据（Big Data）的基本概念、影响和应用领域，并阐述大数据、云计算和物联网的相互关系，还介绍了大数据处理架构 Hadoop。由于 Hadoop 已经成为应用最广泛的大数据技术，因此本书的大数据相关技术主要围绕 Hadoop 展开，包括 Hadoop、MapReduce、HDFS 和 HBase。本篇内容是学习后续内容的基础。

本篇包括 2 章。第 1 章介绍大数据的概念和应用，分析了大数据、云计算和物联网的相互关系；第 2 章介绍大数据处理架构 Hadoop。

知识地图



重点与难点

重点为理解大数据的概念，大数据对科学研究、思维方式和社会发展的影响，以及大数据处理架构 Hadoop。难点为掌握 Hadoop 的安装与使用方法。

第1章 大数据概述

大数据时代悄然来临，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人民生活的方方面面。全球范围内，世界各国政府均高度重视大数据技术的研究和产业发展，纷纷把大数据上升为国家战略加以重点推进。企业和学术机构纷纷加大技术、资金和人员投入力度，加强对大数据关键技术的研发与应用，以期在“第三次信息化浪潮”中占得先机、引领市场。大数据已经不是“镜中花、水中月”，它的影响力和作用力正迅速触及社会的每个角落，所到之处，或是颠覆，或是提升，都让人们深切感受到了大数据实实在在的威力。

对于一个国家而言，能否紧紧抓住大数据发展机遇，快速形成核心技术和应用参与新一轮的全球化竞争，将直接决定未来若干年世界范围内各国科技力量博弈的格局。大数据专业人才的培养是新一轮科技较量的基础，高等院校承担着大数据人才培养的重任，因此，各高等院校非常重视大数据课程的开设，大数据课程已经成为计算机科学与技术专业的重要核心课程。

本章首先介绍了大数据的发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展，并阐述了云计算、物联网的概念及其与大数据之间的紧密关系。

1.1 大数据时代

第三次信息化浪潮涌动，大数据时代全面开启。人类社会信息科技的发展为大数据时代的到来提供了技术支撑，而数据产生方式的变革是促进大数据时代到来至关重要的因素。

1.1.1 第三次信息化浪潮

根据IBM前首席执行官郭士纳的观点，IT领域每隔15年就会迎来一次重大变革（见表1-1）。1980年前后，个人计算机（PC）开始普及，使得计算机走入企业和千家万户，大大提高了社会生产力，也使人类迎来了第一次信息化浪潮，Intel、IBM、苹果、微软、联想等企业是这个时期的标志。随后，在1995年前后，人类开始全面进入互联网时代，互联网的普及把世界变成“地球村”，每个人都可以自由徜徉于信息的海洋，由此，人类迎来了第二次信息化浪潮，这个时期也缔造了雅虎、谷歌、阿里巴巴、百度等互联网巨头。时隔15年，在2010年前后，云计算、大数据、物联网的快速发展，拉开了第三次信息化浪潮的大幕，大数据时代已经到来，也必将涌现出一批新的市场标杆企业。

表 1-1　　三次信息化浪潮

信息化浪潮	发生时间	标志	解决的问题	代表企业
第一次浪潮	1980 年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995 年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010 年前后	物联网、云计算和大数据	信息爆炸	亚马逊、谷歌、IBM、VMWare、Palantir、Hortonworks、Cloudera、阿里云等

1.1.2　信息科技为大数据时代提供技术支撑

信息科技需要解决信息存储、信息传输和信息处理 3 个核心问题，人类社会在信息科技领域的不断进步，为大数据时代的到来提供了技术支撑。

1. 存储设备容量不断增加

数据被存储在磁盘、磁带、光盘、闪存等各种类型的存储介质中，随着科学技术的不断进步，存储设备的制造工艺不断升级，容量大幅增加，速度不断提升，价格却在不断下降（见图 1-1）。

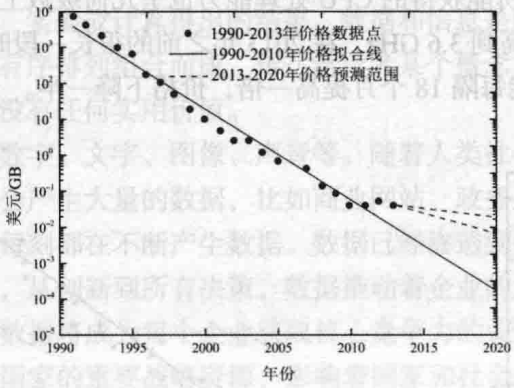


图 1-1　存储设备的价格随时间变化的情况

早期的存储设备容量小、价格高、体积大，例如，IBM 在 1956 年生产的一个早期的商业硬盘，容量只有 5MB，不仅价格昂贵，而且体积有一个冰箱那么大（见图 1-2）。相反，今天容量为 1TB 的硬盘，大小只有 3.5 英寸（约 8.89cm），读写速度达到 200MB/s，价格仅为 400 元左右。廉价、高性能的硬盘存储设备，不仅提供了海量的存储空间，同时大大降低了数据存储成本。

图 1-2　IBM 在 1956 年生产的一个早期的商业硬盘

与此同时，以闪存为代表的新型存储介质也开始得到大规模的普及和应用。闪存是一种新兴的半导体存储器，从 1989 年诞生第一款闪存产品开始，闪存技术不断获得新的突破，并逐渐在计算机存储产品市场中确立了自己的重要地位。闪存是一种非易失性存储器，即使发生断电也不会丢失数据；因此，可以作为永久性存储设备，它具有体积小、质量轻、能耗低、抗振性好等优良特性。