BIG DATA

大数据

基础编程、实验和案例教程

林子雨　编著

- **步步引导，循序渐进**　详尽的安装指南为顺利搭建大数据实验环境铺平道路
- **深入浅出，去粗取精**　丰富的代码实例帮助快速掌握大数据基础编程方法
- **精心设计，巧妙融合**　五套大数据实验题目促进理论与编程知识的消化和吸收
- **结合理论，联系实际**　大数据课程综合实验案例精彩呈现大数据分析全流程

清华大学出版社

大数据

基础编程、实验和案例教程

林子雨　编著

清华大学出版社
北京

内 容 简 介

本书以大数据分析全流程为主线，介绍了数据采集、数据存储与管理、数据处理与分析、数据可视化等环节典型软件的安装、使用和基础编程方法。本书内容涵盖操作系统（Linux 和 Windows）、开发工具（Eclipse）以及大数据相关技术、软件（Sqoop、Kafka、Flume、Hadoop、HDFS、MapReduce、HBase、Hive、Spark、MySQL、MongoDB、Redis、R、Easel. ly、D3、魔镜、ECharts、Tableau）等。同时，本书还提供了丰富的课程实验和综合案例，以及大量免费的在线教学资源，可以较好地满足高等院校大数据教学实际需求。

本书是《大数据技术原理与应用——概念、存储、处理、分析与应用》的“姊妹篇”，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辅助教材，用于指导大数据编程实践；也可供相关技术人员参考。

图书在版编目（CIP）数据

大数据基础编程、实验和案例教程/林子雨编著. —北京：清华大学出版社，2017
ISBN 978-7-302-47209-4

Ⅰ. ①大…　Ⅱ. ①林…　Ⅲ. ①数据处理—教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 125787 号

责任编辑：白立军
封面设计：杨玉兰
责任校对：梁　毅
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www. tup. com. cn，http://www. wqbook. com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup. tsinghua. edu. cn
质量反馈：010-62772015，zhiliang@tup. tsinghua. edu. cn
课件下载：http://www. tup. com. cn，010-62795954

印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：24　　**字　　数**：568 千字
版　　次：2017 年 8 月第 1 版　　**印　　次**：2017 年 8 月第1 次印刷
印　　数：1～2000
定　　价：59.00 元

产品编号：074870-01

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,厦门大学数据库实验室负责人,中国计算机学会数据库专委会委员,中国计算机学会信息系统专委会委员;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学生的认可,成为2013年度和2017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/),成为全国高校大数据教学知名品牌。平台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100万次;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

前　言

大数据带来了信息技术的巨大变革，并深刻影响着社会生产和人民生活的方方面面。大数据专业人才的培养是世界各国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，需要及时建立大数据课程体系，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。

高质量的教材是推进高校大数据课程体系建设的关键支撑。2013 年 12 月，笔者根据自己主讲厦门大学计算机系研究生大数据课程的教学实践，编写了电子书《大数据技术基础》，通过网络免费发布，获得了较好的反响。此后两年多的时间里，笔者继续对大数据技术知识体系进行深入学习和系统梳理，并结合教学实践和大量调研，编著出版了《大数据技术原理与应用》教材，该书第 1 版于 2015 年 8 月出版发行，第 2 版于 2017 年 2 月出版发行。《大数据技术原理与应用》一书侧重于介绍大数据技术的实现原理，编程实践内容较少，该教材定位为入门级大数据教材，以“构建知识体系、阐明基本原理、开展初级实践、了解相关应用”为原则，旨在为读者搭建起通向大数据知识空间的桥梁和纽带，为读者在大数据领域深耕细作奠定基础、指明方向。教材系统论述了大数据的基本概念、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、分布式并行编程模型 MapReduce、大数据处理架构 Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。

《大数据技术原理与应用》一书出版以后，获得了读者较高的认可，目前已经成为国内多所高校的大数据课程教材。与此同时，笔者在最近两年通过各种形式助力全国高校加快推进大数据课程建设，包括建设全国高校大数据课程公共服务平台、开展全国高校大数据公开课巡讲计划、组织全国高校大数据教学论坛、举办全国高校大数据课程教师培训交流班等。通过这些活动，笔者与全国高校广大大数据课程教师有了更深的接触和交流，也收集到了广大一线教师的核心教学需求。很多高校教师在高度肯定《大数据技术原理与应用》教材的同时，也提出了很多中肯的改进意见和建议，其中，有很多教师指出，应该加强大数据实践环节的训练，提供实验指导和综合案例。

为了更好地满足高校教学实际需求，笔者带领厦门大学数据库实验团队，

开展了大量的探索和实践，并对实践材料进行系统整理，在此基础上编写了本教程。本教程侧重于介绍大数据软件的安装、使用和基础编程方法，并提供大量实验和案例。由于大数据软件都是开源软件，安装过程一般比较复杂，也很耗费时间。为了尽量减少读者搭建大数据实验环境时的障碍，笔者在本书中详细写出了各种大数据软件的详细安装过程，可以确保读者顺利完成大数据实验环境搭建。

本书共13章，详细介绍系统和软件的安装、使用以及基础编程方法。第1章介绍大数据的关键技术和代表性软件，帮助读者形成对大数据技术及其代表性软件的总体性认识。第2章介绍Linux系统的安装和使用方法，为后面其他章节的学习奠定基础。第3章介绍分布式计算框架Hadoop的安装和使用方法。第4章介绍分布式文件系统HDFS的基础编程方法。第5章介绍分布式数据库HBase的安装和基础编程方法。第6章介绍典型NoSQL数据库的安装和使用方法，包括键值数据库Redis和文档数据库MongoDB。第7章介绍如何编写基本的MapReduce程序。第8章介绍基于Hadoop的数据仓库Hive的安装和使用方法。第9章介绍基于内存的分布式计算框架Spark的安装和基础编程方法。第10章介绍5种典型的可视化工具的安装和使用方法，包括Easel.ly、D3、魔镜、ECharts、Tableau等。第11章介绍数据采集工具的安装和使用方法，包括Flume、Kafka和Sqoop。第12章介绍一个大数据课程综合实验案例，即网站用户购物行为分析。第13章通过5个实验让读者加深对知识的理解。

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辅助教材。本书是《大数据技术原理与应用》的“姊妹篇”，可以作为《大数据技术原理与应用》的辅助配套教程，两本书组合使用，可以达到更好的学习效果。此外，本书也可以和市场上现有的其他大数据教材配套使用，作为教学辅助用书。

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，向这些同学的辛勤工作表示衷心的感谢。

本书的官方网站是http://dblab.xmu.edu.cn/post/bigdatapractice/，免费提供了全部配套资源的在线浏览和下载，并接受错误反馈和发布勘误信息。同时，在学习大数据课程的过程中，欢迎读者访问厦门大学数据库实验室建设的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/)，该平台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，参考了大量网络资料，对大数据技术及其典型软件进行了系统梳理，有选择地把一些重要知识纳入本书。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

2017年2月于厦门大学计算机科学系数据库实验室

目　录

第1章

大数据技术概述

大数据的时代已经到来，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，都融入了大数据之中，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。

本章首先介绍大数据关键技术和各类典型的大数据软件，帮助读者形成对大数据技术及其代表性软件的总体性认识；然后，给出本书的整体内容安排，帮助读者快速找到相关技术所对应的章节；最后，详细给出与本书配套的在线资源，帮助读者更好、更深入地学习理解相关大数据技术知识。

1.1 大数据时代

人类全面进入信息化社会以后，数据以自然方式增长，其产生不以人的意志为转移。从1986年开始到2016年的30年时间里，全球数据的数量增长了100多倍，今后的数据量增长速度将更快，我们正生活在一个“数据爆炸”的大数据时代。今天，世界上只有大约25%的设备是联网的，大约80%的上网设备是计算机和手机，而在不远的将来，随着物联网的发展和大规模普及，汽车、电视、家用电器、生产机器等各种设备也将联入互联网，各种传感器和摄像头将遍布人们工作和生活的各个角落，这些设备每时每刻都在自动产生大量数据。可以说，人类社会正经历第二次数据爆炸（如果把印刷在纸上的文字和图片也看作数据，那么，人类历史上第一次数据爆炸发生在造纸术和印刷术发明的时期）。各种数据产生速度之快，产生数量之大，已经远远超出传统技术可以处理的范围，“数据爆炸”成为大数据时代的鲜明特征。

在数据爆炸的今天，人类一方面对知识充满渴求，另一方面为数据的复杂特征所困惑。数据爆炸对科学研究提出了更高的要求，需要人类设计出更加灵活高效的数据存储、处理和分析工具，来应对大数据时代的挑战。由此，必将带来云计算、数据仓库、数据挖掘等技术和应用的提升或者根本性变革。在存储效率（存储技术）领域，需要实现低成本的大规模分布式存储；在网络效率（网络技术）方面，需要实现及时响应的用户体验；在数据中心方面，需要开发更加绿色节能的新一代数据中心，在有效面对大数据处理需求的同时，实现最大化资源利用率、最小化系统能耗的目标。面对数据爆炸的大数据时代，我们人类不再从容！

1.2 大数据关键技术

大数据的基本处理流程，主要包括数据采集、存储管理、处理分析、结果呈现等环节。因此，从数据分析全流程的角度，大数据技术主要包括数据采集与预处理、数据存储和管理、数据处理与分析、数据可视化、数据安全和隐私保护等几个层面的内容（具体如表 1-1 所示）。其中，数据可视化有时也被视为数据分析的一种，即可视化分析，因此，数据可视化也可被归入“数据处理与分析”这一大类。

表 1-1 大数据技术的不同层面及其功能

技术层面	功能
数据采集与预处理	利用 ETL（Extraction-Transformation-Loading）工具将分布的、异构数据源中的数据，如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到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；也可以利用日志采集工具（如 Flume、Kafka 等）把实时采集的数据作为流计算系统的输入，进行实时处理分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL 数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析
数据可视化	对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据安全和隐私保护	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

需要指出的是，大数据技术是许多技术的集合体，这些技术也并非全部都是新生事物，诸如关系数据库、数据仓库、数据采集、ETL、OLAP（On-Line Analytical Processing）、数据挖掘、数据隐私和安全、数据可视化等已经发展多年的技术，在大数据时代得到不断补充、完善、提高后又有了新的升华，也可以视为大数据技术的一个组成部分。对于这些技术，除了数据可视化技术以外，我们将不再介绍，本书重点阐述近些年新发展起来的大数据核心技术及其代表性软件使用方法，包括分布式并行编程、分布式文件系统、分布式数据库、NoSQL 数据库、日志采集工具等。

此外，大数据技术及其代表性软件种类繁多，不同的技术都有其适用和不适用的场景。总体而言，不同的企业应用场景，都对应着不同的大数据计算模式，根据不同的大数据计算模式，可以选择相应的大数据计算产品，具体如表 1-2 所示。

表 1-2 大数据计算模式及其代表产品

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark 等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等

续表

大数据计算模式	解 决 问 题	代 表 产 品
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb 等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala 等

批处理计算主要解决针对大规模数据的批量处理，也是人们日常数据分析工作中非常常见的一类数据处理需求。例如，爬虫程序把大量网页抓取过来存储到数据库中以后，可以使用 MapReduce 对这些网页数据进行批量处理，生成索引，加快搜索引擎的查询速度。

流计算主要是实时处理来自不同数据源的、连续到达的流数据，经过实时分析处理，给出有价值的分析结果。例如，用户在访问淘宝网等电子商务网站时，用户在网页中的每次点击的相关信息（比如选取了什么商品）都会像水流一样实时传播到大数据分析平台，平台采用流计算技术对这些数据进行实时处理分析，构建用户“画像”，为其推荐可能感兴趣的其他相关商品。

在大数据时代，许多大数据都是以大规模图或网络的形式呈现，如社交网络、传染病传播途径、交通事故对路网的影响等，此外，许多非图结构的大数据，也常常会被转换为图模型后再进行处理分析。图计算软件是专门针对图结构数据开发的，在处理大规模图结构数据时可以获得很好的性能。

查询分析计算也是一种在企业中常见的应用场景，主要是面向大规模数据的存储管理和查询分析，用户一般只需要输入查询语句（如 SQL），就可以快速得到相关的查询结果。

流计算软件（Storm、S4 等）和图计算软件（Pregel、Hama 等）学习门槛稍高，一般适合作为高级教程内容，本书作为入门级教程，没有涉及流计算和图计算的内容。感兴趣的读者，可以访问本书官网，学习《大数据技术原理与应用》在线视频的内容，了解图计算和流计算的技术原理和相关软件的使用方法。

1.3　大数据软件

本书涉及的大数据软件涵盖数据采集、数据存储与管理、数据处理与分析、数据可视化等环节，每个环节所采用的相关软件如表 1-3 所示。

表 1-3　本书所涉及的大数据软件

大数据技术	大数据软件
数据采集	Flume、Kafka、Sqoop
数据存储与管理	HDFS、HBase、Redis、MongoDB
数据处理与分析	MapReduce、Spark、Hive
数据可视化	Easel. ly、D3、Tableau、魔镜、ECharts

针对表 1-3 中的每一款大数据软件，本书都会给出详细的安装和使用方法介绍，并讲解如何开展基础编程实践。