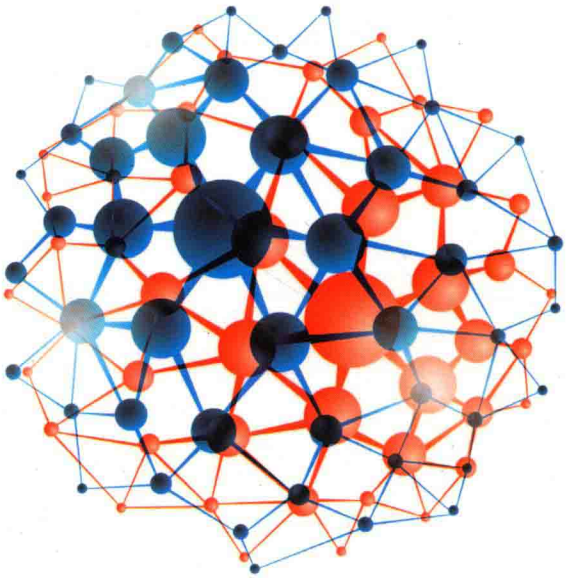
教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目
数据科学与大数据技术专业系列规划教材

华为信息与网络
技术学院指定教材

Spark
编程基础

林子雨 赖永炫 陶继平 ◉ 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
深入浅出，有效**降低**Spark技术学习**门槛**
资源全面，构建**全方位一站式**在线服务体系

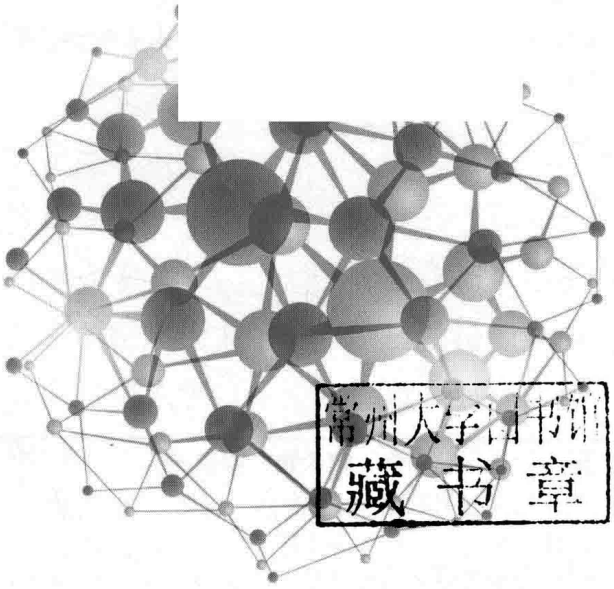
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

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

Spark
编程基础

林子雨 赖永炫 陶继平 ◉ 编著



人民邮电出版社
北京

图书在版编目（CIP）数据

Spark编程基础 / 林子雨，赖永炫，陶继平编著. -- 北京 : 人民邮电出版社，2018.7
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-47598-5

Ⅰ. ①S… Ⅱ. ①林… ②赖… ③陶… Ⅲ. ①数据处理软件－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第322133号

内 容 提 要

本书以 Scala 作为开发 Spark 应用程序的编程语言，系统介绍了 Spark 编程的基础知识。全书共 7 章，内容包括大数据技术概述、Spark 的设计与运行原理、Spark 环境搭建和使用方法、RDD 编程、Spark SQL、Spark Streaming 和 Spark MLlib。

本书每章都安排了入门级的编程实践操作，以便使读者能更好地学习和更牢固地掌握 Spark 编程方法。本书配套官网免费提供了全套的在线教学资源，包括讲义 PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

本书可以作为高等院校计算机、软件工程、数据科学与大数据技术等专业的进阶级大数据课程教材，用于指导 Spark 编程实践，也可供相关技术人员参考。

◆ 编　　著　林子雨　赖永炫　陶继平
责任编辑　邹文波
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：12.75　　2018 年 7 月第 1 版
字数：325 千字　　2018 年 7 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
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

大数据时代的来临，给各行各业带来了深刻的变革。大数据像能源、原材料一样，已经成为提升国家和企业竞争力的关键要素，被称为“未来的新石油”。正如电力技术的应用引发了生产模式的变革一样，基于互联网技术而发展起来的大数据技术的应用，将会为人们的生产和生活带来颠覆性的影响。

目前，大数据技术正处于快速发展之中，不断有新的技术涌现，Hadoop 和 Spark 等技术成为其中的佼佼者。在 Spark 流行之前，Hadoop 俨然已成为大数据技术的事实标准，在企业中得到了广泛的应用，但其本身还存在诸多缺陷，最主要的是 MapReduce 计算模型延迟过高，无法胜任实时、快速计算的需求，因而只适用于离线批处理的应用场景。Spark 在设计上充分吸收借鉴了 MapReduce 的精髓并加以改进，同时，采用了先进的 DAG 执行引擎，以支持循环数据流与内存计算，因此，在性能上比 MapReduce 有了大幅度的提升，从而迅速获得了学术界和业界的广泛关注。作为大数据计算平台的后起之秀，Spark 在 2014 年打破了 Hadoop 保持的基准排序纪录，此后逐渐发展成为大数据领域最热门的大数据计算平台之一。

随着大数据在企业应用的不断深化，企业对大数据人才的需求日益增长。为了有效地满足不断增长的大数据人才需求，国内高校从 2016 年开始设立“数据科学与大数据技术专业”，着力培养数据科学与工程领域的复合型高技术人才。课程体系的建设和课程教材的创作，是高校大数据专业建设的核心环节。

厦门大学数据库实验室在大数据教学领域辛勤耕耘、开拓创新，成为国内高校大数据教学资源的有力贡献者。实验室在积极践行 O2O 大数据教学理念的同时，提出了“以平台化思维构建全国高校大数据课程公共服务体系”的全新服务理念，成为推进国内高校大数据教学不断向前发展的一支重要力量，在全国高校之中形成了广泛的影响。2015 年 7 月，实验室编写出版了国内高校第一本系统性介绍大数据知识的专业教材——《大数据技术原理与应用》，受到了广泛的好评，目前已经成为国内众多高校的入门级大数据课程的开课教材。同时，实验室建设了国内高校首个大数据课程公共服务平台（网址：http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/），为全国高校教师和学生提供大数据教学资源一站式“免费”在线服务，包括课程教材、讲义 PPT、课程习题、实验指南、学习指南、备课指南、授课视频和技术资料等，自 2013 年 5 月建设以来，定位明确，进展顺利，目前平台每年访问量超过 100 万次，成为全国高校大数据教学的知名品牌。

《大数据技术原理与应用》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。高校在开设入门级课程以后，可以根据自己的实际情况，开设进阶级的大数据课程，继续深化对大数据技术的学习，而 Spark 是目前比较理想的大数据进阶课程学习内容。因此，厦门大学数据库实验室组织具有丰富经验的一线大数据教师精心编写了本教材。

为了确保教材质量，在出版纸质图书之前，实验室已经于 2016 年 10 月通过实验室官网免费共享了简化版的 Spark 在线教程和相关教学资源，同时，该在线教程也已经用于厦门大学计算机科学系研究生的大数据课程教学，并成为全国高校大数据课程教师培训交流班的授课内容。实验室根据读者对在线 Spark 教程的大量反馈意见以及在教学实践中发现的问题，对 Spark 在线教程进行了多次修正和完善，这些前期准备工作，都为纸质图书的编著出版打下了坚实的基础。

本书共 7 章，详细介绍了 Spark 的环境搭建和基础编程方法。第 1 章介绍大数据关键技术，帮助读者对大数据技术形成总体性认识以及了解 Spark 在其中所扮演的角色；第 2 章介绍 Spark 的设计与运行原理；第 3 章介绍 Spark 的环境搭建和使用方法，为开展 Spark 编程实践铺平道路；第 4 章介绍 RDD 编程，包括 RDD 的创建、操作 API、持久化、分区以及键值对 RDD 等，这章知识是开展 Spark 高级编程的基础；第 5 章介绍 Spark 中用于结构化数据处理的组件 Spark SQL，包括 DataFrame 数据模型、创建方法和常用操作等；第 6 章介绍 Spark Streaming，这是一种构建在 Spark 上的流计算框架，可以满足对流式数据进行实时计算的需求；第 7 章介绍 Spark 的机器学习库 MLlib，包括 MLlib 的基本原理、算法、模型选择和超参数调整方法等。

本书面向高校计算机、软件工程、数据科学与大数据技术等专业的学生，可以作为专业必修课或选修课教材。本书由林子雨、赖永炫和陶继平执笔，其中，林子雨负责全书规划、统稿、校对和在线资源创作，并撰写第 1、2、4、5、6 章的内容，赖永炫负责撰写第 7 章的内容，

陶继平负责撰写第3章的内容。在撰写过程中，厦门大学计算机科学系硕士研究生阮榕城、薛倩、魏亮、曾冠华、程璐、林哲等做了大量的辅助性工作，在此，向这些同学的辛勤工作表示衷心的感谢。同时，感谢夏小云老师在书稿校对过程中的辛勤付出。

本书配套的官方网站是 http://dblab.xmu.edu.cn/post/spark/，免费提供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，Spark 作为大数据进阶课程，在学习过程中会涉及大量相关的大数据基础知识以及各种大数据软件的安装和使用方法，因此，推荐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/），来获得必要的辅助学习内容。

本书在撰写过程中，参考了大量的网络资料和相关书籍，对 Spark 技术进行了系统梳理，有选择性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

厦门大学计算机科学系数据库实验室

2018 年 1 月

目 录 CONTENTS

第 1 章 大数据技术概述 1

1.1 大数据的概念与关键技术 2
1.1.1 大数据的概念 2
1.1.2 大数据关键技术 2
1.2 代表性大数据技术 4
1.2.1 Hadoop 4
1.2.2 Spark 8
1.2.3 Flink 10
1.2.4 Beam 11
1.3 编程语言的选择 12
1.4 在线资源 13
1.5 本章小结 14
1.6 习题 14
实验 1 Linux 系统的安装和常用命令 15
一、实验目的 15
二、实验平台 15
三、实验内容和要求 15
四、实验报告 16

第 2 章 Spark 的设计与运行原理 17

2.1 概述 18
2.2 Spark 生态系统 19
2.3 Spark 运行架构 20
2.3.1 基本概念 20
2.3.2 架构设计 21
2.3.3 Spark 运行基本流程 22
2.3.4 RDD 的设计与运行原理 23
2.4 Spark 的部署方式 32
2.5 本章小结 33
2.6 习题 34

第 3 章 Spark 环境搭建和使用方法 35

3.1 安装 Spark 36
3.1.1 基础环境 36
3.1.2 下载安装文件 36
3.1.3 配置相关文件 37
3.1.4 Spark 和 Hadoop 的交互 38
3.2 在 spark-shell 中运行代码 38
3.2.1 spark-shell 命令 39
3.2.2 启动 spark-shell 40
3.3 开发 Spark 独立应用程序 40
3.3.1 安装编译打包工具 41
3.3.2 编写 Spark 应用程序代码 42
3.3.3 编译打包 42
3.3.4 通过 spark-submit 运行程序 45
3.4 Spark 集群环境搭建 45
3.4.1 集群概况 46
3.4.2 搭建 Hadoop 集群 46
3.4.3 在集群中安装 Spark 47
3.4.4 配置环境变量 47
3.4.5 Spark 的配置 47
3.4.6 启动 Spark 集群 48
3.4.7 关闭 Spark 集群 48
3.5 在集群上运行 Spark 应用程序 49
3.5.1 启动 Spark 集群 49
3.5.2 采用独立集群管理器 49
3.5.3 采用 Hadoop YARN 管理器 50
3.6 本章小结 51
3.7 习题 52
实验 2 Spark 和 Hadoop 的安装 52
一、实验目的 52
二、实验平台 52

三、实验内容和要求……52
四、实验报告……53

第 4 章 RDD 编程……54

4.1 RDD 编程基础……55
4.1.1 RDD 创建……55
4.1.2 RDD 操作……56
4.1.3 持久化……62
4.1.4 分区……63
4.1.5 一个综合实例……67
4.2 键值对 RDD……69
4.2.1 键值对 RDD 的创建……69
4.2.2 常用的键值对转换操作……70
4.2.3 一个综合实例……74
4.3 数据读写……75
4.3.1 文件数据读写……76
4.3.2 读写 HBase 数据……78
4.4 综合实例……82
4.4.1 求 TOP 值……82
4.4.2 文件排序……84
4.4.3 二次排序……85
4.5 本章小结……87
实验 3 RDD 编程初级实践……87
一、实验目的……87
二、实验平台……87
三、实验内容和要求……87
四、实验报告……89

第 5 章 Spark SQL……90

5.1 Spark SQL 简介……91
5.1.1 从 Shark 说起……91
5.1.2 Spark SQL 架构……92
5.1.3 为什么推出 Spark SQL……93
5.2 DataFrame 概述……93
5.3 DataFrame 的创建……94
5.4 DataFrame 的保存……95
5.5 DataFrame 的常用操作……96
5.6 从 RDD 转换得到 DataFrame……97
5.6.1 利用反射机制推断 RDD 模式……98
5.6.2 使用编程方式定义 RDD 模式……99
5.7 使用 Spark SQL 读写数据库……101
5.7.1 通过 JDBC 连接数据库……101
5.7.2 连接 Hive 读写数据……103
5.8 本章小结……107
5.9 习题……107
实验 4 Spark SQL 编程初级实践……108
一、实验目的……108
二、实验平台……108
三、实验内容和要求……108
四、实验报告……109

第 6 章 Spark Streaming… 110

6.1 流计算概述……111
6.1.1 静态数据和流数据……111
6.1.2 批量计算和实时计算……112
6.1.3 流计算概念……112
6.1.4 流计算框架……113
6.1.5 流计算处理流程……114
6.2 Spark Streaming……115
6.2.1 Spark Streaming 设计……115
6.2.2 Spark Streaming 与 Storm 的对比……116
6.2.3 从“Hadoop+Storm”架构转向 Spark 架构……117
6.3 DStream 操作概述……118
6.3.1 Spark Streaming 工作机制……118
6.3.2 编写 Spark Streaming 程序的基本步骤……119
6.3.3 创建 StreamingContext 对象…119
6.4 基本输入源……120
6.4.1 文件流……120
6.4.2 套接字流……122
6.4.3 RDD 队列流……127
6.5 高级数据源……128
6.5.1 Kafka 简介……129
6.5.2 Kafka 准备工作……129

6.5.3 Spark 准备工作 ····· 130
6.5.4 编写 Spark Streaming 程序使用 Kafka 数据源 ····· 131
6.6 转换操作 ····· 135
6.6.1 DStream 无状态转换操作 ····· 135
6.6.2 DStream 有状态转换操作 ····· 136
6.7 输出操作 ····· 140
6.7.1 把 DStream 输出到文本文件中 ····· 140
6.7.2 把 DStream 写入到关系数据库中 ····· 141
6.8 本章小结 ····· 143
6.9 习题 ····· 143
实验 5 Spark Streaming 编程初级实践 ····· 144
一、实验目的 ····· 144
二、实验平台 ····· 144
三、实验内容和要求 ····· 144
四、实验报告 ····· 145
第 7 章 Spark Mllib ····· 146
7.1 基于大数据的机器学习 ····· 147
7.2 机器学习库 MLlib 概述 ····· 148
7.3 基本数据类型 ····· 149
7.3.1 本地向量 ····· 149
7.3.2 标注点 ····· 149
7.3.3 本地矩阵 ····· 150
7.4 机器学习流水线 ····· 151
7.4.1 流水线的概念 ····· 151
7.4.2 流水线工作过程 ····· 152
7.5 特征提取、转换和选择 ····· 153
7.5.1 特征提取 ····· 154
7.5.2 特征转换 ····· 156
7.5.3 特征选择 ····· 161
7.5.4 局部敏感哈希 ····· 162
7.6 分类算法 ····· 163
7.6.1 逻辑斯蒂回归分类器 ····· 163
7.6.2 决策树分类器 ····· 167
7.7 聚类算法 ····· 170
7.7.1 K-Means 聚类算法 ····· 171
7.7.2 GMM 聚类算法 ····· 173
7.8 协同过滤算法 ····· 175
7.8.1 推荐算法的原理 ····· 176
7.8.2 ALS 算法 ····· 176
7.9 模型选择和超参数调整 ····· 180
7.9.1 模型选择工具 ····· 180
7.9.2 用交叉验证选择模型 ····· 181
7.10 本章小结 ····· 183
7.11 习题 ····· 183
实验 6 Spark 机器学习库 MLlib 编程实践 ····· 184
一、实验目的 ····· 184
二、实验平台 ····· 184
三、实验内容和要求 ····· 184
四、实验报告 ····· 185
参考文献 ····· 186

01 第1章 大数据技术概述

大数据时代的来临，给各行各业带来了深刻的变革。大数据像能源、原材料一样，已经成为提升国家和企业竞争力的关键要素，被称为“未来的新石油”。正如电力技术的应用引发了生产模式的变革一样，基于互联网技术而发展起来的大数据应用，将会对人们的生产和生活产生颠覆性的影响。

本章首先介绍大数据的概念与关键技术，然后重点介绍有代表性的大数据技术，包括 Hadoop、Spark、Flink、Beam 等，最后探讨本教程编程语言的选择，并给出与本教材配套的相关在线资源。

1.1 大数据的概念与关键技术

随着大数据时代的到来，“大数据”已经成为互联网信息技术行业的流行词汇。本节介绍大数据的概念与关键技术。

1.1.1 大数据的概念

关于“什么是大数据”这个问题，学术界和业界比较认可关于大数据的“4V”说法。大数据的4个“V”，或者说是大数据的4个特点，包含4个层面：数据量大（Volume）、数据类型繁多（Variety）、处理速度快（Velocity）和价值密度低（Value）。

（1）数据量大。根据著名咨询机构IDC（Internet Data Center）做出的估测，人类社会产生的数据一直都在以每年50%的速度增长，这被称为“大数据摩尔定律”。这意味着，人类在最近两年产生的数据量相当于之前产生的全部数据量之和。预计到2020年，全球将总共拥有35ZB的数据量，数据量将增长到2010年数据的近30倍。

（2）数据类型繁多。大数据的数据类型丰富，包括结构化数据和非结构化数据，其中，前者占10%左右，主要是指存储在关系数据库中的数据，后者占90%左右，种类繁多，主要包括邮件、音频、视频、微信、微博、位置信息、链接信息、手机呼叫信息、网络日志等。

（3）处理速度快。大数据时代的很多应用，都需要基于快速生成的数据给出实时分析结果，用于指导生产和生活实践，因此，数据处理和分析的速度通常要达到秒级响应，这一点和传统的数据挖掘技术有着本质的不同，后者通常不要求给出实时分析结果。

（4）价值密度低。大数据价值密度却远远低于传统关系数据库中已经有的那些数据，在大数据时代，很多有价值的信息都是分散在海量数据中的。

1.1.2 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来看，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。

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

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
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，也是我们日常数据分析工作中非常常见的一类数据处理需求。比如，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。代表性的批处理框架包括 MapReduce、Spark 等。

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。比如，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。代表性的流计算框架包括 Twitter Storm、Yahoo! S4 等。Twitter Storm 是一个免费、开源的分布式实时计算系统，Storm 对于实时计算的意义类似于 Hadoop 对于批处理的意义，Storm 可以简单、高效、可靠地处理流数据，并支持多种编程语言。Storm 框架可以方便地与数据库系统进行整合，从而开发出强大的实时计算系统。Storm 可用于许多领域中，如实时分析、在线机器学习、持续计算、远程 RPC、数据提取加载转换等。由于 Storm 具有可扩展、高容错性、能可靠地处理消息等特点，目前已经被广泛应用于流计算应用中。

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。谷歌公司的 Pregel 是一种基于 BSP 模型实现的图计算框架。为了解决大型图的分布式计算问题，Pregel 搭建了一套可扩展的、有容错机制的平台，该平台提供了一套非常灵活的 API，可以描述各种各样的图计算。Pregel 作为分布式图计算的计算框架，主要用于图遍历、最短路径、PageRank 计算等。

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。典型的查询分析计算产品包括 Dremel、Hive、Cassandra、Impala 等。其中，Dremel 是一种可扩展的、交互式的实时查询系统，用于只读嵌套数据的分析。通过结合多级树状执行过程和列式数据结构，它能做到几秒内完成对万亿张表的聚合查询。系统可以扩展到成千上万的 CPU 上，满足谷歌上万用户操作 PB 级的数据，并且可以在 2～3 秒内完成 PB 级别数据的查询。Hive 是一个构建于 Hadoop 顶层的数据仓库工具，允许用户输入 SQL 语句进行查询。Hive 在某种程度上可以看作是用户编程接口，其本身并不