

教育部-邦飞产学合作协同育人项目

大数据与人工智能丛书

Hadoop+Spark 大数据技术

（微课版）

刘彬斌 主编

李永富 李柏章 周磊 编著

清华大学出版社
北 京

内 容 简 介

本书从初学者角度出发,通过丰富的实例,详细介绍了大数据开发环境和基本知识点的应用。全书内容包括:大数据系统基础篇、Hadoop 技术篇、Spark 技术篇和项目实战篇。大数据系统基础篇讲解 Linux 的安装、Linux 的使用和在 Linux 系统上安装并使用 MySQL;Hadoop 技术篇讲解 Hadoop 集群的搭建、Hadoop 两大核心的原理与使用、Hadoop 生态圈的工具原理与使用(Hive、HBase、Sqoop、Flume 等);Spark 技术篇讲解 Spark 集群的搭建、Scala 语言、RDD、Spark SQL、Spark streaming 和机器学习;项目实战篇将真实的“电力能源大数据分析”项目作为实战解读,帮助初学者快速入门。

本书所有知识点都结合具体实例和程序讲解,便于读者理解和掌握。本书适合作为高等院校计算机应用、大数据技术及相关专业的教材;也适合作为大数据开发入门者的自学用书,可快速提高开发技能。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

Hadoop+Spark 大数据技术:微课版/刘彬斌主编. —北京:清华大学出版社,2018

(大数据与人工智能丛书)

ISBN 978-7-302-51427-5

I. ①H… II. ①刘… III. ①数据处理软件—高等学校—教材 IV. ①TP274

中国版本图书馆 CIP 数据核字(2018)第 242160 号

责任编辑:付弘宇 薛 阳

封面设计:刘 键

责任校对:焦丽丽

责任印制:

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 刷 者:

装 订 者:

经 销:全国新华书店

开 本:185mm×260mm 印 张:22.5 字 数:545 千字

版 次:2018 年 11 月第 1 版 印 次:2018 年 11 月第 1 次印刷

印 数:

定 价: 元

产品编号:079663-01

前言

随着信息技术的不断发展，以及物联网、社交网络、移动终端等新兴技术与服务的不断涌现和广泛应用，数据种类日益增多，数据的规模急剧增大，大数据时代已悄然来临。由于大数据对政府决策、商业规划和危险预防等方面所起的重大作用，大数据逐渐成为一种重要的国家战略性资源，受到政府、能源及信息领域的普遍关注。大数据的多样性（Variety）、规模性（Volume）和高速性（Velocity）等特点，使得传统的数据存储、管理、分析技术已经无法满足大数据的处理要求。

时至今日，无论你是来自互联网、通信行业，还是来自金融业、服务业或零售业，相信你都不会对大数据感到陌生。调查显示，32.5%的公司正在搭建大数据平台，29.5%的公司已经在生产环境实践大数据技术，并有成功的用例/产品；24.5%的公司已经做了足够的了解，开发准备就绪；基本不了解的只占调查对象的13.5%。TalkingData 公司 CEO 助理、腾云大学执行校长杨慧在接受记者采访时表示，目前国内市场的 IT 人才缺口已经高达几十万，到 2025 年，这一数字还会增加至 200 万，“尤其是大数据技术方面的人才”。在智联、58 同城等大型招聘网站最新发布的招聘职位中，大数据相关岗位占比已经超过 50%，薪酬同比软件工程师高 10% 以上。由此可见，国内对大数据人才的需求缺口巨大。

1. 高校大数据人才培养的背景

（1）高校教育中，大数据人才培养存在起步晚、规模化不足的问题，而且高校学生从大学到研究生需要相当长的一段时间。本书从实用的角度出发，为高校快速培养大数据人才提供可行性。

（2）如前文所述，大数据人才紧缺的现象在全球越来越突出。在此背景下，本书旨在弥补高校大数据教材的不足，以模拟真实生产环境为教学目标，为企业培养“到岗就能用”的大数据实用型人才。

（3）经济社会的高速发展，对 IT 产业（尤其是软件产业）提出了更高的要求，对大数据开发人才从数量和质量方面提出了更高的要求。

（4）教育技术的进步和移动互联网时代的到来，打破了高校进行知识传播的技术壁垒。大量的资本和风险投资涌进 IT 培训产业。达内、东软、传智播客等实体 IT 培训机构，开课吧、慕课网、极客网等在线 IT 培养机构纷纷引入先进的教学理念、强大的技术支持，再加上商业化运作，对高校 IT 人才培养带来巨大的挑战和竞争压力。

（5）教学环境的变化：教室、实验室硬件配置齐全，实现了高速稳定的互联网接入。笔记本电脑和手机等互联网接入设备日渐普及，这些都为先进教学理念和教学模式（如微课）的实施提供了硬件和软件上的准备。

（6）教育参与者：教师正在树立“教育就是服务”的教育观念，正在贯彻工程教育的教育理念，从注重“教师教什么”转移到“学生学到了什么”。学生作为“数字原住民”，对新鲜事物、新技术、新教学方式（人性化学习、泛在学习等）有着天然的渴望。

综上所述, IT 产业、软件技术、软件人才培养竞争、教学理念、教学模式、教学环境、教学对象等因素的发展变化倒逼着高校进行教学改革, 教师必须围绕以上因素进行教学创新, 传统教材形式的革新也势在必行。

2. 本书内容

全书内容分为大数据系统基础、Hadoop 技术、Spark 技术和项目实战 4 部分。其中, Linux 是学习大数据技术的基础, 先从 Linux 入手, 打下坚实的基础, 之后才能更好地学习 Hadoop 和 Spark。4 部分内容分别介绍如下。

大数据系统基础篇通过大数据概述、Linux 系统安装、Linux 系统基础命令、Shell 编程和 MySQL 数据操作, 为以后编程奠定坚实的基础。

Hadoop 技术篇以 Hadoop 生态圈为中心, 详细介绍 Hadoop 高可用集群搭建、HDFS 技术、MapReduce 技术、Hive 技术, 为读者学习大数据开发技术提供便利, 并以实用的方式简单介绍 HBase、Sqoop、Flume 工具的使用, 使读者在精通一门技术的前提下, 能扩展了解相关知识, 真正成为一专多能的专业型人才。

Spark 技术篇从 Spark 概述、Scala 语言、环境搭建、RDD 核心技术、Spark SQL 和机器学习等多方面讲解 Spark 大数据的开发, 从基础的 Scala 语言开始学习, 并以 Hadoop 环境为基础搭建 Spark 大数据集群, 从最基础、最常用、最容易理解的思路出发, 帮助读者逐步掌握 Spark 大数据技术。

项目实战篇从真实项目“电力能源大数据分析”中抽取一部分业务作为实战解读, 通过简洁的流程讲解, 使读者了解大数据项目开发的整个过程。

3. 本书特色

本书不是对相关原理进行纯理论的阐述, 而是提供了丰富的上机实践操作和范例程序, 极大地降低了读者学习大数据技术的门槛。对于需要直接上机实践的读者而言, 本书更像是一本大数据学习的实践上机手册。书中首先展示了如何在单台 Windows 系统上通过 VirtualBox 虚拟机安装多台 Linux 虚拟机, 而后建立 Hadoop 集群, 再建立 Spark 开发环境。搭建这个上机实践的平台并不限制于单台实体计算机, 主要是考虑个人读者上机实践的实际条件和环境。对于有条件的公司和学校, 参照这个搭建过程, 同样可以将实践平台搭建在多台实体计算机上。

搭建好大数据上机实践的软硬件环境之后, 就可以在各个章节的学习中结合本书提供的范例程序逐一设置、修改、调试和运行, 从中体会大数据实践应用的真谛——对大数据进行高效的“加工”, 萃取大数据中蕴含的“智能和知识”, 实现数据的“增值”, 并最终将其应用于实际工作或者商业项目中。

4. 本书的使用

(1) 第 1 篇讲解 Linux 系统和 Linux 系统上的软件应用。本篇是学习大数据技术的第一步, 就如同你要学习 Java 开发, 必须先学会操作 Windows 系统一样。

(2) 第 2 篇讲解 Hadoop 大数据技术。Hadoop 大数据集群要求在 CentOS 6.9 版本的系统上搭建, JDK 版本为 JDK 1.8, Hadoop 版本为 Hadoop 2.6.5, Zookeeper 版本为 Zookeeper 3.4.10。

(3) 第 3 篇讲解在 Hadoop 大数据技术的基础上搭建 Spark 环境, 所以读者在学习本篇内容之前, 需要熟悉第 2 篇中的 Hadoop 大数据集群搭建的内容。

(4) 第 4 篇讲解电力大数据项目, 是基础 HDFS 的离线分析项目, 读者需要掌握 Java 知识、Hadoop 技术和 Web 前端知识。

5. 作者与致谢

本书由刘彬斌主编。参与本书的编写、资料整理、书稿校对、课件制作等工作的还有李永富、李柏章、周磊、汪磊等。另外，感谢清华大学出版社相关编辑专业和严谨的工作，为本书的顺利出版提供了宝贵的意见，并付出辛勤的劳动。

编 者

2018 年 3 月

目 录

第 1 篇 大数据系统基础

| | |
|---------------------|----|
| 第 1 章 大数据概述 | 3 |
| 1.1 数据的产生与发展 | 3 |
| 1.2 大数据的基础知识 | 5 |
| 1.3 大数据架构 | 5 |
| 第 2 章 系统的安装与使用 | 7 |
| 2.1 系统安装 | 7 |
| 2.1.1 安装 CentOS 6.x | 7 |
| 2.1.2 安装步骤 | 7 |
| 2.2 基本命令 | 18 |
| 2.2.1 cd 命令 | 18 |
| 2.2.2 打包和解压指令 | 19 |
| 2.2.3 其他常用命令 | 21 |
| 2.3 权限与目录 | 26 |
| 2.3.1 权限 | 26 |
| 2.3.2 目录 | 27 |
| 2.4 文件操作 | 28 |
| 2.4.1 文件与目录管理 | 28 |
| 2.4.2 用户和用户组管理 | 39 |
| 2.5 习题与思考 | 46 |
| 第 3 章 任务命令 | 47 |
| 3.1 脚本配置 | 47 |
| 3.1.1 Shell 脚本 | 47 |
| 3.1.2 Shell 变量 | 47 |
| 3.1.3 Shell 传递参数 | 48 |
| 3.1.4 Shell 数组 | 50 |
| 3.1.5 Shell 运算符 | 51 |
| 3.1.6 Shell echo 命令 | 55 |

| | | |
|--------------|-----------------|-----------|
| 3.1.7 | Shell printf 命令 | 57 |
| 3.1.8 | Shell test 命令 | 58 |
| 3.1.9 | Shell 流程控制 | 60 |
| 3.2 | 网络配置 | 67 |
| 3.3 | 习题与思考 | 70 |
| 第 4 章 | 数据库操作 | 71 |
| 4.1 | 数据库简介 | 71 |
| 4.1.1 | MySQL 数据库简介 | 71 |
| 4.1.2 | 安装 MySQL | 72 |
| 4.2 | 数据库基本操作 | 72 |
| 4.2.1 | MySQL 的 DDL 操作 | 72 |
| 4.2.2 | MySQL 的 DML 操作 | 80 |
| 4.3 | 数据库用户操作 | 83 |
| 4.3.1 | 创建用户 | 83 |
| 4.3.2 | 给用户授权 | 83 |
| 4.3.3 | 撤销授权 | 84 |
| 4.3.4 | 查看用户权限 | 85 |
| 4.3.5 | 删除用户 | 85 |
| 4.3.6 | 修改用户密码 | 86 |
| 4.4 | 数据库查询操作 | 86 |
| 4.5 | 习题与思考 | 90 |

第 2 篇 Hadoop 技术

| | | |
|--------------|-----------------------|------------|
| 第 5 章 | Hadoop 开发环境 | 95 |
| 5.1 | Hadoop 生态圈工具 | 95 |
| 5.2 | 环境搭建 | 97 |
| 5.2.1 | 步骤 1——虚拟机安装 | 97 |
| 5.2.2 | 步骤 2——安装 JDK 和 Hadoop | 97 |
| 5.2.3 | 步骤 3——复制虚拟机 | 113 |
| 5.2.4 | 步骤 4——设置免密 | 117 |
| 5.2.5 | 步骤 5——安装 Zookeeper | 119 |
| 5.2.6 | 步骤 6——启动 Hadoop 集群 | 122 |
| 5.2.7 | 正常启动顺序 | 125 |
| 5.3 | 常见问题汇总 | 127 |
| 5.4 | 习题与思考 | 128 |
| 第 6 章 | HDFS 技术 | 129 |
| 6.1 | HDFS 架构 | 129 |

| | | |
|--------------|----------------------------|------------|
| 6.2 | HDFS 命令 | 130 |
| 6.2.1 | version 命令 | 131 |
| 6.2.2 | dfsadmin 命令 | 131 |
| 6.2.3 | jar 命令 | 132 |
| 6.2.4 | fs 命令 | 132 |
| 6.3 | API 的使用 | 140 |
| 6.4 | 习题与思考 | 142 |
| 第 7 章 | MapReduce 技术 | 143 |
| 7.1 | MapReduce 工作原理 | 143 |
| 7.1.1 | MapReduce 作业运行流程 | 143 |
| 7.1.2 | 早期 MapReduce 架构存在的问题 | 144 |
| 7.2 | YARN 运行概述 | 144 |
| 7.2.1 | YARN 模块介绍 | 144 |
| 7.2.2 | YARN 工作流程 | 145 |
| 7.3 | MapReduce 编程模型 | 146 |
| 7.4 | MapReduce 数据流 | 148 |
| 7.4.1 | 输入文件 | 150 |
| 7.4.2 | 输入格式 | 150 |
| 7.4.3 | 数据片段 | 151 |
| 7.4.4 | 记录读取器 | 151 |
| 7.4.5 | Mapper | 151 |
| 7.4.6 | Shuffle | 152 |
| 7.4.7 | 排序 | 153 |
| 7.4.8 | 归约 | 153 |
| 7.4.9 | 输出格式 | 153 |
| 7.5 | MapReduce API 编程 | 154 |
| 7.5.1 | 词频统计 | 154 |
| 7.5.2 | 指定字段 | 156 |
| 7.5.3 | 求平均数 | 158 |
| 7.5.4 | 关联 | 160 |
| 7.6 | 习题与思考 | 163 |
| 第 8 章 | Hive 数据仓库 | 165 |
| 8.1 | Hive 模型 | 165 |
| 8.1.1 | Hive 架构与基本组成 | 165 |
| 8.1.2 | Hive 的数据模型 | 166 |
| 8.2 | Hive 的安装 | 167 |
| 8.2.1 | Hive 的基本安装 | 167 |

| | | |
|--------|--------------------------|-----|
| 8.2.2 | MySQL 的安装 | 168 |
| 8.2.3 | Hive 配置 | 169 |
| 8.3 | HQL 详解 | 170 |
| 8.3.1 | Hive 数据管理方式 | 170 |
| 8.3.2 | HQL 操作 | 174 |
| 8.4 | 习题与思考 | 182 |
| 第 9 章 | HBase 分布式数据库 | 183 |
| 9.1 | HBase 工作原理 | 183 |
| 9.1.1 | HBase 表结构 | 183 |
| 9.1.2 | 体系结构 | 184 |
| 9.1.3 | 物理模型 | 186 |
| 9.1.4 | HBase 读写流程 | 187 |
| 9.2 | HBase 完全分布式 | 189 |
| 9.2.1 | 安装前的准备 | 189 |
| 9.2.2 | 配置文件 | 189 |
| 9.2.3 | 集群启动 | 191 |
| 9.3 | HBase Shell | 192 |
| 9.3.1 | DDL 操作 | 192 |
| 9.3.2 | DML 操作 | 194 |
| 9.4 | 习题与思考 | 197 |
| 第 10 章 | Sqoop 工具 | 198 |
| 10.1 | Sqoop 安装 | 199 |
| 10.2 | Sqoop 的使用 | 200 |
| 10.2.1 | MySQL 的导入导出 | 200 |
| 10.2.2 | Oracle 的导入导出 | 201 |
| 10.3 | 习题与思考 | 202 |
| 第 11 章 | Flume 日志收集 | 203 |
| 11.1 | 体系架构 | 204 |
| 11.1.1 | Flume 内部结构 | 204 |
| 11.1.2 | Flume 事件 | 204 |
| 11.2 | Flume 的特点 | 205 |
| 11.3 | Flume 集群搭建 | 205 |
| 11.4 | Flume 实例 | 207 |
| 11.4.1 | 实例 1: 实时测试客户端传输的数据 | 207 |
| 11.4.2 | 实例 2: 监控本地文件夹并写入到 HDFS 中 | 208 |
| 11.5 | 习题与思考 | 210 |

第3篇 Spark 技术

| | |
|------------------------------|-----|
| 第 12 章 Spark 概述 | 213 |
| 12.1 Spark 框架原理 | 213 |
| 12.2 Spark 大数据处理 | 214 |
| 12.3 RDD 数据集 | 215 |
| 12.4 Spark 子系统 | 215 |
| 第 13 章 Scala 语言 | 216 |
| 13.1 Scala 语法基础 | 216 |
| 13.1.1 变量、常量与赋值 | 216 |
| 13.1.2 运算符与表达式 | 217 |
| 13.1.3 条件分支控制 | 217 |
| 13.1.4 循环流程控制 | 218 |
| 13.1.5 Scala 数据类型 | 218 |
| 13.2 Scala 运算与函数 | 219 |
| 13.3 Scala 闭包 | 220 |
| 13.4 Scala 数组与字符串 | 220 |
| 13.4.1 Scala 数组 | 220 |
| 13.4.2 Scala 字符串 | 221 |
| 13.5 Scala 迭代器 | 221 |
| 13.6 Scala 类和对象 | 222 |
| 13.7 习题与思考 | 223 |
| 第 14 章 Spark 高可用环境 | 224 |
| 14.1 环境搭建 | 224 |
| 14.1.1 准备工作 | 224 |
| 14.1.2 下载并安装 Spark | 224 |
| 14.2 常见问题汇总 | 226 |
| 第 15 章 RDD 技术 | 228 |
| 15.1 RDD 的实现 | 228 |
| 15.1.1 数据源 | 228 |
| 15.1.2 调度器 | 228 |
| 15.2 RDD 编程接口 | 229 |
| 15.3 RDD 操作 | 229 |
| 15.3.1 Spark 基于命令行的操作 | 229 |
| 15.3.2 Spark 基于应用作业的操作 | 231 |

| | | |
|--------|--------------------------|-----|
| 15.3.3 | Spark 操作的基础命令与开发工具介绍 | 231 |
| 15.3.4 | Spark 基于 YARN 的调度模式 | 231 |
| 15.3.5 | Spark 基于 Scala 语言的本地应用开发 | 234 |
| 15.3.6 | Spark 基于 Scala 语言的集群应用开发 | 235 |
| 15.3.7 | Spark 基于 Java 语言的应用开发 | 236 |
| 15.3.8 | Spark 基于 Java 语言的本地应用开发 | 237 |
| 15.3.9 | Spark 基于 Java 语言的集群应用开发 | 238 |
| 15.4 | 习题与思考 | 241 |

第 16 章 Spark SQL 242

| | | |
|---------|---------------------------------|-----|
| 16.1 | Spark SQL 架构原理 | 242 |
| 16.1.1 | Hive 的两种功能 | 242 |
| 16.1.2 | Spark SQL 的重要功能 | 242 |
| 16.1.3 | Spark SQL 的 DataFrame 特征 | 243 |
| 16.2 | Spark SQL 操作 Hive | 243 |
| 16.2.1 | 添加配置文件, 便于 Spark SQL 访问 Hive 仓库 | 243 |
| 16.2.2 | 安装 JDBC 驱动 | 243 |
| 16.2.3 | 启动 MySQL 服务及其 Hive 的元数据服务 | 243 |
| 16.2.4 | 启动 HDFS 集群和 Spark 集群 | 244 |
| 16.2.5 | 启动 Spark-Shell 并测试 | 244 |
| 16.3 | Spark SQL 操作 HDFS | 244 |
| 16.3.1 | 操作代码 | 244 |
| 16.3.2 | 工程文件 | 246 |
| 16.3.3 | 创建测试数据 | 246 |
| 16.3.4 | 运行 Job 并提交到集群 | 247 |
| 16.3.5 | 查看运行结果 | 247 |
| 16.4 | Spark SQL 操作关系数据库 | 248 |
| 16.4.1 | 添加访问 MySQL 的驱动包 | 248 |
| 16.4.2 | 添加必要的开发环境 | 248 |
| 16.4.3 | 使用 Spark SQL 操作关系数据库 | 248 |
| 16.4.4 | 初始化 MySQL 数据库服务 | 250 |
| 16.4.5 | 准备 Spark SQL 源数据 | 251 |
| 16.4.6 | 运行 Spark 代码 | 252 |
| 16.4.7 | 创建 dist 文件夹 | 252 |
| 16.4.8 | 安装数据库驱动 | 252 |
| 16.4.9 | 基于集群操作 | 253 |
| 16.4.10 | 打包工程代码到 dist 目录下 | 256 |
| 16.4.11 | 启动集群并提交 Job 应用 | 256 |
| 16.4.12 | 检查关系数据库中是否已有数据 | 258 |

| | | |
|---------------|--------------------------------|------------|
| 16.5 | 习题与思考 | 258 |
| 第 17 章 | Spark Streaming | 260 |
| 17.1 | 架构与原理 | 260 |
| 17.1.1 | Spark Streaming 中的离散流特征 | 260 |
| 17.1.2 | Spark Streaming 的应用场景 | 260 |
| 17.2 | KafKa 中间件 | 261 |
| 17.2.1 | KafKa 的特点 | 261 |
| 17.2.2 | ZeroCopy 技术 | 261 |
| 17.2.3 | KafKa 的通信原理 | 261 |
| 17.2.4 | KafKa 的内部存储结构 | 262 |
| 17.2.5 | KafKa 的下载 | 262 |
| 17.2.6 | KafKa 集群搭建 | 262 |
| 17.2.7 | 启动并使用 KafKa 集群 | 263 |
| 17.2.8 | 停止 KafKa 集群 | 264 |
| 17.2.9 | KafKa 集成 Flume | 264 |
| 17.3 | Socket 事件流操作 | 265 |
| 17.3.1 | netcat 网络 Socket 控制台工具 | 265 |
| 17.3.2 | 基于本地的 Spark Streaming 流式数据分析示例 | 266 |
| 17.3.3 | 基于集群的 Spark Streaming 流式数据分析示例 | 269 |
| 17.3.4 | 基于集群模式下的集群文件 I/O 流分析示例 | 272 |
| 17.4 | KafKa 事件流操作 | 275 |
| 17.4.1 | 基于 Receiver 模式的 KafKa 集成 | 275 |
| 17.4.2 | 基于 Direct 模式的 KafKa 集成 | 278 |
| 17.5 | I/O 文件事件流操作 | 280 |
| 17.5.1 | 基于路径扫描的 Spark Streaming | 281 |
| 17.5.2 | 打包至工程的 dist 目录 | 284 |
| 17.5.3 | 启动集群 | 284 |
| 第 18 章 | Spark 机器学习 | 289 |
| 18.1 | 机器学习原理 | 289 |
| 18.1.1 | 机器学习的概念 | 289 |
| 18.1.2 | 机器学习的分类 | 289 |
| 18.1.3 | Spark 机器学习的版本演变 | 290 |
| 18.1.4 | DataFrame 数据结构 | 290 |
| 18.1.5 | DataSet 数据结构 | 290 |
| 18.1.6 | 执行引擎的性能与效率 | 290 |
| 18.1.7 | Spark 2.x 的新特性 | 290 |
| 18.2 | 线性回归 | 291 |

| | | |
|--------|----------------------|-----|
| 18.2.1 | 线性回归分析过程 | 291 |
| 18.2.2 | 矩阵分析过程 | 291 |
| 18.2.3 | 基于本地模式的线性回归分析 | 291 |
| 18.2.4 | 基于集群模式的线性回归分析 | 294 |
| 18.3 | 聚类分析 | 300 |
| 18.3.1 | K-Means 聚类算法原理 | 300 |
| 18.3.2 | 聚类分析过程 | 300 |
| 18.3.3 | 基于本地模式的聚类算法分析 | 301 |
| 18.3.4 | 基于集群模式的聚类算法分析 | 305 |
| 18.4 | 协同过滤 | 312 |
| 18.4.1 | 个性化推荐算法 | 312 |
| 18.4.2 | 相关性推荐算法 | 312 |
| 18.4.3 | 基于本地的协同过滤算法分析 | 312 |
| 18.4.4 | 基于集群的协同过滤算法分析 | 317 |

第4篇 项目实战篇

| | | |
|--------|--------------------|-----|
| 第19章 | 基于电力能源的大数据实战 | 325 |
| 19.1 | 需求分析 | 325 |
| 19.2 | 项目设计 | 325 |
| 19.2.1 | 数据采集 | 325 |
| 19.2.2 | 数据处理 | 326 |
| 19.2.3 | 数据呈现 | 326 |
| 19.3 | 数据收集与处理 | 329 |
| 19.3.1 | 数据收集 | 329 |
| 19.3.2 | 数据处理 | 329 |
| 19.4 | 大数据呈现 | 341 |
| 19.4.1 | 数据传输 | 341 |
| 19.4.2 | 数据呈现 | 342 |
| 19.5 | 项目总结 | 343 |