过程完整：从基础配置过程一直到真正的商业项目开发
实例丰富：90个实例，1 个完整项目
原理清晰：重点概念、操作、思路都有图示，避免抽象讲解
代码详尽：所有实例都有详细的代码，所有代码都有详尽的解读
辐射面广：讲解了Spark与周边框架的交互

Broadview
www.broadview.com.cn

Spark 大数据分析

源码解析与实例详解

刘景泽 编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Spark 大数据分析

源码解析与实例详解

刘景泽 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书基于 Spark 2.3.x、Spark 2.4.x 系列版本，采用“理论+实践”的形式编写。全书共有 90 个实例，1 个完整项目。

第 1 篇“准备”，包括认识大数据和 Spark、安装与配置 Spark 集群、第 1 个 Spark 程序；第 2 篇“入门”，包括读写分布式数据、处理分布式数据；第 3 篇“进阶”，包括 RDD 的高级操作、用 SQL 语法分析结构化数据、实时处理流式数据；第 4 篇“高阶”，包括实时处理流式数据、Spark 的相关优化；第 5 篇“商业项目实战”，用 Spark 的各种组件实现一个学生学情分析商业项目。

本书结构清晰、实例丰富、通俗易懂、实用性强，特别适合 Spark 的初学者和进阶读者作为自学用书。另外，本书也适合社会培训学校作为培训教材，还适合大中专院校的相关专业作为教学参考书。

图书在版编目（CIP）数据

Spark 大数据分析：源码解析与实例详解/刘景泽编著. —北京：电子工业出版社，2019.7

ISBN 978-7-121-37051-9

Ⅰ. ①S… Ⅱ. ①刘… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 140115 号

责任编辑：吴宏伟
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：27.25　　字数：654 千字
版　　次：2019 年 7 月第 1 版
印　　次：2019 年 7 月第 1 次印刷
定　　价：89.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

前言

Spark 是一个较早实现完整生态圈的框架。它与 Hadoop 生态圈紧密结合，并能够独立运行。其高度完善的 API 及容错机制，能大大降低数据丢失或错乱的风险，从而让开发者将更多的精力放在数据分析的具体逻辑上。

本书特色

1. 利用实例贯穿理论知识，落地性强

本书提供了大量实例，让读者通过实例来透彻地理解理论知识。在实际开发过程中，可以将本书作为工具书反复查阅。

2. 丰富详细的图示说明

重点概念、重点操作（算子）、重点思路均配有详细的图示，避免抽象讲解。

3. 详尽的代码解读

杜绝只言片语的代码解读，详细解读所有重点代码和相关代码，使读者能够理解代码的含义。

4. 包含 Structured Streaming 详解

本书包含 Spark 2.x 版本之后新增的热门组件 Structured Streaming，并结合图示、实例详细讲解其功能。

5. 包含 Spark 商业实战项目

从 0 到 1 实现 Spark 项目开发，项目涵盖 Spark Core、Spark SQL、Spark Streaming、Structured Streaming、Redis、Kafka 等组件及框架。

6. 免费提供实例的源文件和素材

在网上免费提供实例的源文件和素材，方便读者直观、高效地学习本书内容。

7. 采用短段、短句，易于理解

本书采用短段、短句来讲解，让读者可以更流畅地阅读。

8. 图书配套讨论群

加入本书配套的 QQ 群，可与其他开发者共同探讨问题，共享资源。

9. 辐射面广

本书操作过程涉及与周边框架的交互，包含 HBase、Kafka、MySQL、Redis、HDFS、YARN、Hive，便于读者快速了解与周边框架的协作方式。

读者对象

- 大数据爱好者；
- 大数据分析工程师；
- 大数据挖掘工程师；
- 后台开发工程师；
- 大中专院校相关专业的学生。

致谢

诚挚感谢张皓先生为本书提供封面设计思路。

特别感谢家人、朋友的支持。

虽然我已对书中内容都进行了认真核实，并多次进行文字校对，但因作者水平所限，书中疏漏和错误在所难免，敬请广大读者批评与指正。

联系作者，可加入本书讨论 QQ 群 626608434，也可发 E-mail 到 616616769@qq.com。

联系编辑，请发 E-mail 到 wuhongwei@phei.com.cn。

刘景泽

2019 年 5 月

配套资源说明

配置资源

1. 书中实例的源文件

提供书中所有实例的源文件，如图 1 所示。

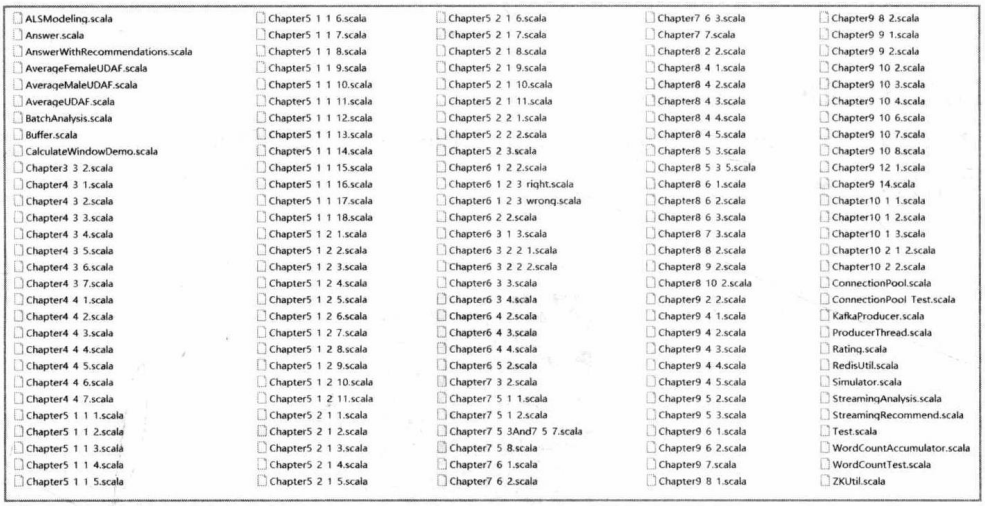

图 1　本书实例源文件

这些实例源文件被封装在与书中章节对应的 Maven 工程中，如图 2 所示。直接导入 Maven 工程即可方便地管理代码。

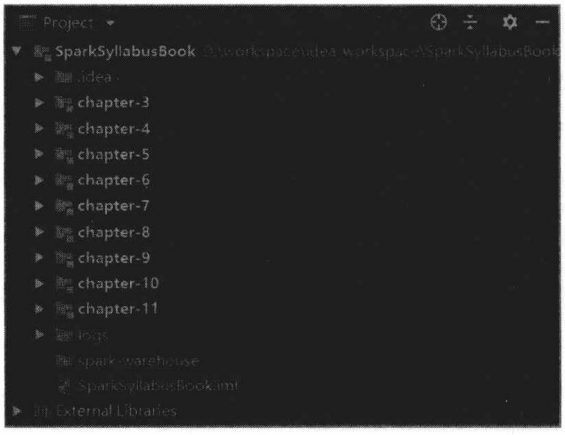

图 2　本书实例对应的 Maven 工程

2. 书中使用的素材

提供实例中所使用的全部素材，如图 3 所示。请读者在动手操作实例前先下载并参考这些素材。

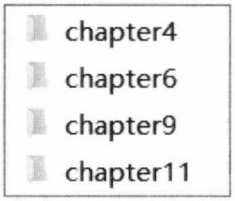

图 3　素材与样本

可通过以下方式获取源文件和素材。

- 本书配套代码的链接（Github）：https://github.com/Zjinji/SparkSyllabusBook
- 作者博客：http://jinjicode.com
- 微信公众号：

讨论群

如果在学习中遇到困难，可以加入本书配套的 QQ 群（626608434），使用手机 QQ 扫下方二维码即可加入。

目　录

第 1 篇　准备

第 2 篇 入门

第 3 篇 进阶

第 4 篇 高阶

第 5 篇 商业项目实战

第 1 篇　准备

本篇首先介绍大数据的基础知识，以及 Spark 框架的作用，这些内容能帮助读者宏观了解即将面临的挑战；然后，由浅入深地讲解如何部署 Spark 集群、如何运行 Spark 应用程序来检验部署效果；并详细讲解如何编写代码、调试应用程序、正式提交应用程序。

- 第 1 章　认识大数据和 Spark
- 第 2 章　安装与配置 Spark 集群
- 第 3 章　第 1 个 Spark 程序

第 1 章

认识大数据和Spark

本章非常轻松，不涉及太多专业知识。在简单介绍大数据的来源之后，将给读者普及一下众多流式处理框架之间的区别，以及在真正开始学习之前需要做哪些准备。现在，我们开始吧。

1.1 大数据的介绍

2008 年《自然》杂志专刊中提出一个概念，原文为："Researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching." 这句话表明，在 2008 年人们就需要开始调整技术体系和实践方式以适应越来越多的新增数据。从此正式提出"大数据"（BigData）概念。

在数据技术出现之前，人类活动产生的数据大多是基于传统的关系型数据库技术来存储的，并且可以基于关系型数据库技术实现数据检索、分析和挖掘。但随着社会的进步，人类需求日益增多并更加细化，传统的关系型数据库很难处理这类新出现的非结构化数据（例如：每一行数据的列数不相等，或者说并不是每一行数据的每一列都有值）。

2008 年前后，Google 先后公开发表了两篇论文影响世界——***Google File System***（被译为《谷歌文件系统》，简称 GFS）与 ***Google MapReduce***（被译为《谷歌分布式运算》，简称 GMR）。其中，GFS 负责做分布式存储，将大文件拆分为多个小文件分散存储到多台机器上；GMR 负责做分布式运算，因为只把数据保存下来是不够的，还需要对数据进行分析，从而得出有价值的结论，并以此结论服务社会生产。

2013 年被称为大数据元年，大数据技术开始辐射到商业的各个角落，例如：游戏、医疗、深度学习等，而且大数据应用需求在之后持续上升。越来越多的研发人员加入大数据行列，涌现出上百种框架，可谓百花齐放。