

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



电子工业出版社
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



电子工业出版社

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。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

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 zltts@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。

☐ ALSModeling.scala	☐ Chapter5 1 1 6.scala	☐ Chapter5 2 1 6.scala	☐ Chapter7 6 3.scala	☐ Chapter9 8 2.scala
☐ Answer.scala	☐ Chapter5 1 1 7.scala	☐ Chapter5 2 1 7.scala	☐ Chapter7 7.scala	☐ Chapter9 9 1.scala
☐ AnswerWithRecommendations.scala	☐ Chapter5 1 1 8.scala	☐ Chapter5 2 1 8.scala	☐ Chapter8 2 2.scala	☐ Chapter9 9 2.scala
☐ AverageFemaleUDAF.scala	☐ Chapter5 1 1 9.scala	☐ Chapter5 2 1 9.scala	☐ Chapter8 4 1.scala	☐ Chapter9 10 2.scala
☐ AverageMaleUDAF.scala	☐ Chapter5 1 1 10.scala	☐ Chapter5 2 1 10.scala	☐ Chapter8 4 2.scala	☐ Chapter9 10 3.scala
☐ AverageUDAF.scala	☐ Chapter5 1 1 11.scala	☐ Chapter5 2 1 11.scala	☐ Chapter8 4 3.scala	☐ Chapter9 10 4.scala
☐ BatchAnalysis.scala	☐ Chapter5 1 1 12.scala	☐ Chapter5 2 2 1.scala	☐ Chapter8 4 4.scala	☐ Chapter9 10 6.scala
☐ Buffer.scala	☐ Chapter5 1 1 13.scala	☐ Chapter5 2 2 2.scala	☐ Chapter8 4 5.scala	☐ Chapter9 10 7.scala
☐ CalculateWindowDemo.scala	☐ Chapter5 1 1 14.scala	☐ Chapter5 2 3.scala	☐ Chapter8 5 3.scala	☐ Chapter9 10 8.scala
☐ Chapter3 2.scala	☐ Chapter5 1 1 15.scala	☐ Chapter6 1 2 2.scala	☐ Chapter8 5 3 3.scala	☐ Chapter9 12 1.scala
☐ Chapter4 3 1.scala	☐ Chapter5 1 1 16.scala	☐ Chapter6 1 2 3 right.scala	☐ Chapter8 6 1.scala	☐ Chapter9 14.scala
☐ Chapter4 3 2.scala	☐ Chapter5 1 1 17.scala	☐ Chapter6 1 2 3 wrong.scala	☐ Chapter8 6 2.scala	☐ Chapter10 1 1.scala
☐ Chapter4 3 3.scala	☐ Chapter5 1 1 18.scala	☐ Chapter6 2 2.scala	☐ Chapter8 6 3.scala	☐ Chapter10 1 2.scala
☐ Chapter4 3 4.scala	☐ Chapter5 1 2 1.scala	☐ Chapter6 3 1 3.scala	☐ Chapter8 7 3.scala	☐ Chapter10 1 3.scala
☐ Chapter4 3 5.scala	☐ Chapter5 1 2 2.scala	☐ Chapter6 3 2 2 1.scala	☐ Chapter8 8 2.scala	☐ Chapter10 2 1 2.scala
☐ Chapter4 3 6.scala	☐ Chapter5 1 2 3.scala	☐ Chapter6 3 2 2 2.scala	☐ Chapter8 9 2.scala	☐ Chapter10 2 2.scala
☐ Chapter4 3 7.scala	☐ Chapter5 1 2 4.scala	☐ Chapter6 3 3.scala	☐ Chapter8 10 2.scala	☐ ConnectionPool.scala
☐ Chapter4 4 1.scala	☐ Chapter5 1 2 5.scala	☐ Chapter6 3 4.scala	☐ Chapter8 2 2.scala	☐ ConnectionPool Test.scala
☐ Chapter4 4 2.scala	☐ Chapter5 1 2 6.scala	☐ Chapter6 4 2.scala	☐ Chapter9 4 1.scala	☐ KafkaProducer.scala
☐ Chapter4 4 3.scala	☐ Chapter5 1 2 7.scala	☐ Chapter6 4 3.scala	☐ Chapter9 4 2.scala	☐ ProducerThread.scala
☐ Chapter4 4 4.scala	☐ Chapter5 1 2 8.scala	☐ Chapter6 4 4.scala	☐ Chapter9 4 3.scala	☐ Rating.scala
☐ Chapter4 4 5.scala	☐ Chapter5 1 2 9.scala	☐ Chapter6 5 2.scala	☐ Chapter9 4 4.scala	☐ RedisUtil.scala
☐ Chapter4 4 6.scala	☐ Chapter5 1 2 10.scala	☐ Chapter7 3 2.scala	☐ Chapter9 4 5.scala	☐ Simulator.scala
☐ Chapter4 4 7.scala	☐ Chapter5 1 2 11.scala	☐ Chapter7 5 1 1.scala	☐ Chapter9 5 2.scala	☐ StreamingAnalysis.scala
☐ Chapter5 1 1 1.scala	☐ Chapter5 2 1 1.scala	☐ Chapter7 5 1 2.scala	☐ Chapter9 5 3.scala	☐ StreamingRecommend.scala
☐ Chapter5 1 1 2.scala	☐ Chapter5 2 1 2.scala	☐ Chapter7 5 3And7 5 7.scala	☐ Chapter9 6 1.scala	☐ Test.scala
☐ Chapter5 1 1 3.scala	☐ Chapter5 2 1 3.scala	☐ Chapter7 5 8.scala	☐ Chapter9 6 2.scala	☐ WordCountAccumulator.scala
☐ Chapter5 1 1 4.scala	☐ Chapter5 2 1 4.scala	☐ Chapter7 6 1.scala	☐ Chapter9 7.scala	☐ WordCountTest.scala
☐ Chapter5 1 1 5.scala	☐ Chapter5 2 1 5.scala	☐ Chapter7 6 2.scala	☐ Chapter9 8 1.scala	☐ ZKUtil.scala

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。

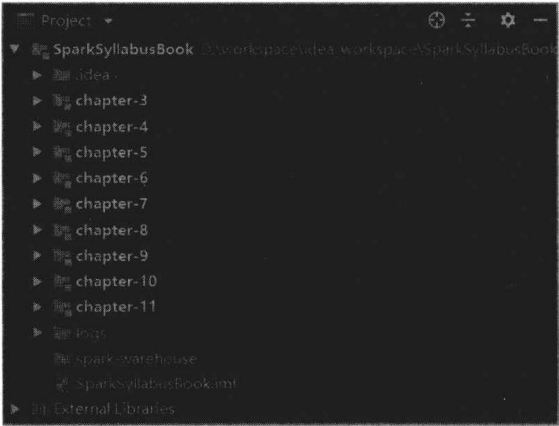


图 2 本书实例对应的 Maven 工程

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。

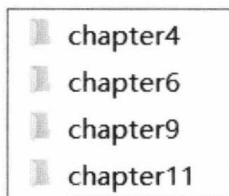


图 3 素材与样本

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 扫描下方二维码即可加入。



目 录

第 1 篇 准备

第 1 章 认识大数据和 Spark	2
1.1 大数据的介绍	2
1.2 Apache Spark 能做什么	3
1.3 其他分布式数据处理框架	4
1.4 如何使用本书	4
1.4.1 需要提前具备的基础	4
1.4.2 准备相关开发环境	4
1.4.3 如何学习本书	5
第 2 章 安装与配置 Spark 集群	6
2.1 下载 Spark 安装包	6
2.2 检查与准备集群环境	7
2.3 了解目前集群中已经部署的框架服务	11
2.4 部署 Spark 集群	12
2.4.1 实例 1: 基于 Standalone 模式部署 Spark 集群	12
2.4.2 实例 2: 部署 Spark 的历史服务——Spark History Server	16
2.4.3 实例 3: 基于 Standalone 模式部署高可用的 Master 服务	18
2.4.4 实例 4: 基于 YARN 模式部署 Spark 集群	20
2.4.5 Standalone 模式与 YARN 模式的特点	22
2.5 本章小结	23
第 3 章 第 1 个 Spark 程序	24
3.1 运行第 1 个 Spark 程序	24
3.1.1 实例 5: 基于 Standalone 模式运行第 1 个 Spark 程序	24
3.1.2 实例 6: 基于 YARN 模式运行第 1 个 Spark 程序	27

3.1.3	提交 Spark 程序时的参数规范	30
3.2	使用 spark-shell 编写并运行 WordCount 程序	30
3.2.1	实例 7: 启动 spark-shell	31
3.2.2	实例 8: 在 spark-shell 中编写 WordCount 程序	32
3.3	使用 IDEA 编写并运行 WordCount 程序	34
3.3.1	实例 9: 准备开发环境, 并构建代码工程	34
3.3.2	实例 10: 使用 IDEA 编写 WordCount 程序	41
3.3.3	实例 11: 在 IDEA 中本地运行 WordCount 程序	44
3.3.4	实例 12: 在 IDEA 中远程运行 WordCount 程序	46
3.3.5	实例 13: 打包程序并提交至集群运行	48
3.4	本章小结	49

第 2 篇 入门

第 4 章	读写分布式数据——基于 Spark Core	52
4.1	RDD 的诞生	52
4.2	进一步理解 RDD	53
4.2.1	数据存储	53
4.2.2	数据分析	55
4.2.3	程序调度	56
4.3	读取数据并生成 RDD	57
4.3.1	实例 14: 读取普通文本数据	58
4.3.2	实例 15: 读取 JSON 格式的数据	59
4.3.3	实例 16: 读取 CSV、TSV 格式的数据	61
4.3.4	实例 17: 读取 SequenceFile 格式的数据	62
4.3.5	实例 18: 读取 Object 格式的数据	64
4.3.6	实例 19: 读取 HDFS 中的数据——显式调用 Hadoop API	66
4.3.7	实例 20: 读取 MySQL 数据库中的数据	68
4.4	保存 RDD 中的数据到外部存储系统	70
4.4.1	实例 21: 保存成普通文本文件	70
4.4.2	实例 22: 保存成 JSON 文件	71
4.4.3	实例 23: 保存成 CSV、TSV 文件	73
4.4.4	实例 24: 保存成 SequenceFile 文件	74

4.4.5	实例 25: 保存成 Object 文件	75
4.4.6	实例 26: 保存成 HDFS 文件——显式调用 Hadoop API 的方式	76
4.4.7	实例 27: 写入 MySQL 数据库	78
4.5	本章小结	80
第 5 章	处理分布式数据——基于 Spark Core	81
5.1	RDD 的转换 (transformations) 操作——转换数据形态	81
5.1.1	实例 28: 基础转换操作	81
5.1.2	实例 29: 键值对转换操作	103
5.2	RDD 的行动 (actions) 操作——触发执行任务计划	115
5.2.1	实例 30: 基础行动操作	116
5.2.2	实例 31: 键值对行动操作	125
5.2.3	实例 32: 数值行动操作	127
5.3	本章小结	128

第 3 篇 进阶

第 6 章	RDD 的高级操作	130
6.1	缓存 RDD	130
6.1.1	缓存 RDD 的基础知识	130
6.1.2	实例 33: 缓存与释放 RDD	133
6.2	RDD 的检查点 (Checkpoint) 机制	139
6.2.1	了解 Checkpoint 机制	139
6.2.2	实例 34: 使用 Checkpoint 机制	141
6.2.3	Checkpoint 机制的工作流程	144
6.3	RDD 的依赖关系	145
6.3.1	窄依赖 (narrow dependencies)	145
6.3.2	宽依赖 (wide/shuffle dependencies)	148
6.3.3	实例 35: 让子 RDD 混合依赖多个父 RDD	151
6.3.4	实例 36: 词频统计——总结运算过程中涉及的概念	153
6.4	累加器 (Accumulator)	155
6.4.1	认识累加器	155
6.4.2	实例 37: 使用系统累加器 1——长整数、双精度浮点数累加器	156

6.4.3	实例 38: 使用系统累加器 2——集合累加器	159
6.4.4	实例 39: 自定义累加器	160
6.5	广播 (Broadcast) ——将数据块缓存到所有节点	164
6.5.1	认识广播	165
6.5.2	实例 40: 使用广播补全数据	165
6.6	本章小结	168
第 7 章	用 SQL 语法分析结构化数据——基于 Spark SQL	169
7.1	为什么会产生 Spark SQL	169
7.2	认识 DataFrame 与 Dataset 数据类型	170
7.2.1	认识 DataFrame	170
7.2.2	认识 Dataset	171
7.3	实例 41: 通过 Dataset、DataFrame 分析用户数据	172
7.3.1	用 spark-shell 编写程序	172
7.3.2	用 IDEA 编写程序	175
7.4	不同 Spark 版本的操作差异	177
7.4.1	认识 SQLContext 与 HiveContext	178
7.4.2	认识 SparkSession	178
7.5	DataFrame、Dataset 的基本操作	179
7.5.1	DSL 与 SQL 的语法风格	179
7.5.2	使用临时视图的注意事项	181
7.5.3	实例 42: 读取 JSON、CSV 格式的数据	183
7.5.4	实例 43: 读取 Parquet 格式的数据	185
7.5.5	实例 44: 读取代码中动态生成的数据	185
7.5.6	实例 45: 读取关系型数据库中的数据	188
7.5.7	实例 46: 输出 Dataset、DataFrame 中的数据	189
7.5.8	实例 47: RDD、DataFrame、Dataset 之间的相互转换	192
7.6	用户自定义函数	195
7.6.1	实例 48: 实现“一进一出”的 UDF	195
7.6.2	实例 49: 实现“多进一出”的 UDAF	198
7.6.3	实例 50: 实现“一进多出”的 UDTF	208
7.7	集成 Spark SQL 与 Hive	211
7.7.1	已经部署 Hive 框架	211

7.7.2	尚未部署 Hive 框架	215
7.8	本章小结	215
第 8 章	实时处理流式数据——基于 Spark Streaming	216
8.1	为什么会产生 Spark Streaming	216
8.2	第 1 个 Spark Streaming 程序	216
8.2.1	实例 51: 用 spark-shell 编写程序	216
8.2.2	实例 52: 用 IDEA 编写程序	221
8.3	什么是 DStream	222
8.3.1	认识 DStream	222
8.3.2	认识 DStreamGraph	223
8.4	读取数据到 DStream 中	227
8.4.1	实例 53: 读取 HDFS 文件夹中的数据	227
8.4.2	实例 54: 读取 RDD 组成的数据队列	229
8.4.3	实例 55: 实时读取 Flume 中的数据	230
8.4.4	实例 56: 用高阶 API 实时读取 Kafka 中的数据	235
8.4.5	实例 57: 用低阶 API 实时读取 Kafka 中的数据	242
8.5	Spark Streaming 中的几个时间概念	251
8.5.1	批处理间隔	251
8.5.2	窗口时间宽度与滑动时间宽度	252
8.5.3	实例 58: 使用窗口操作, 每两秒统计 10 秒内的平均温度	254
8.6	DStream 的操作总结	259
8.6.1	DStream 的操作说明	259
8.6.2	实例 59: 直接面向 DStream 中的 RDD 进行数据分析	261
8.6.3	实例 60: 将 DStream 中的数据实时输出至外部存储系统	263
8.6.4	实例 61: 对 Dstream 进行 join 操作	267
8.7	DStream 中的转换分类	269
8.7.1	无状态转换	269
8.7.2	有状态转换	270
8.7.3	实例 62: 用有状态转换做全局词频统计	270
8.8	在 Spark Streaming 中的缓存与 Checkpoint	272
8.8.1	认识 Spark Streaming 中的 Checkpoint	273
8.8.2	实例 63: 使用 Spark Streaming 中的 Checkpoint	273

8.9	Spark Streaming 中的累加器与广播变量	276
8.9.1	认识累加器与广播变量	276
8.9.2	实例 64: 自定义累加器, 并结合无状态转换, 实现实时的全局词频统计	276
8.10	关闭 Spark Streaming 程序	280
8.10.1	关闭程序的方案	281
8.10.2	合理关闭一个运行中的 Spark Streaming 程序	281
8.11	本章小结	284

第 4 篇 高阶

第 9 章	实时处理流式数据——基于 Structured Streaming	286
9.1	为什么会产生 Structured Streaming	286
9.2	第 1 个 Structured Streaming 程序	287
9.2.1	实例 65: 用 spark-shell 编写程序	287
9.2.2	实例 66: 用 IDEA 编写程序	289
9.3	Structured Streaming 的编程模型	291
9.4	输入数据——生成 Streaming Dataset、Streaming DataFrame	292
9.4.1	实例 67: 根据文件生成工作流	292
9.4.2	实例 68: 根据文件、文件夹生成自动分区的工作流	295
9.4.3	实例 69: 根据 Kafka 以 Streaming 模式生成工作流	297
9.4.4	实例 70: 以 Kafka 为数据源, 通过 Batch 模式生成工作流	300
9.4.5	实例 71: 根据指定速率生成工作流	304
9.5	基于事件时间的窗口操作	305
9.5.1	事件时间窗口的工作方式	305
9.5.2	实例 72: 事件时间窗口的生成规则	307
9.5.3	实例 73: 基于事件时间窗口实现词频统计	311
9.6	基于 Watermark 处理延迟数据	314
9.6.1	Watermark 的作用	314
9.6.2	实例 74: 基于 Update 模式实现词频统计, 并结合 Watermark 处理延迟数据	314
9.6.3	实例 75: 基于 Append 模式实现词频统计, 并结合 Watermark 处理延迟数据	320
9.6.4	Watermark 的底层工作原理	322

9.6.5	总结: Watermark 机制与输出模式	329
9.7	实例 76: 在处理流式数据时去除重复数据	330
9.8	Structured Streaming 中的 join 操作	332
9.8.1	实例 77: 在 Stream-Static 模式下的 inner join 操作	333
9.8.2	实例 78: 在 Stream-Stream 模式下的 inner join 操作	335
9.8.3	总结: 已经支持的 join 操作	340
9.9	在 Structured Streaming 中实现数据分组, 并手动维护分组状态	341
9.9.1	实例 79: 通过 mapGroupsWithState 实现数据分组, 并手动维护分组状态 ..	341
9.9.2	实例 80: 通过 flatMapGroupsWithState 实现数据分组, 并手动维护 分组状态	347
9.9.3	总结: 手动维护状态与 Watermark 的使用技巧	352
9.10	输出分析结果	353
9.10.1	输出模式 (Output Mode) 的使用场景	353
9.10.2	实例 81: 基于 File Sink 输出数据	354
9.10.3	实例 82: 基于 Kafka Sink, 以 Streaming 方式输出数据	356
9.10.4	实例 83: 基于 Kafka Sink, 以 Batch 方式输出数据	358
9.10.5	实例 84: 基于 Console Sink 输出数据	360
9.10.6	实例 85: 基于 Memory Sink 输出数据	360
9.10.7	实例 86: 基于 Foreach Sink 输出数据	362
9.10.8	实例 87: 基于 ForeachBatch Sink 输出数据	367
9.10.9	总结: 不同 Sink 所适用的输出模式	369
9.11	Trigger 触发器的分类	370
9.12	管理与监控工作流	370
9.12.1	管理工作流	370
9.12.2	监控工作流	372
9.13	Structured Streaming 中的 Checkpoint 机制	372
9.14	连续处理模式——Continuous Processing	373
9.15	本章小结	374
第 10 章	Spark 的相关优化	375
10.1	优化 Spark 程序	375
10.1.1	实例 88: 尽可能减少或避免出现 Shuffle 过程	375
10.1.2	实例 89: 使用 Kryo 作为序列化方案	377

10.1.3	尽可能批量操作数据	381
10.1.4	合理设置分区数	381
10.1.5	合理设置批处理间隔	381
10.2	优化数据	382
10.2.1	关于数据倾斜	382
10.2.2	实例 90: 使用自定义 Partitioner 缓解数据倾斜	383
10.2.3	关于数据补全	387
10.3	调优资源	388
10.4	本章小结	390

第 5 篇 商业项目实战

第 11 章	实战：学生学习情况分析系统	392
11.1	项目概述	392
11.1.1	业务背景	392
11.1.2	划分业务模块	392
11.2	开发环境说明	393
11.3	项目实施	394
11.3.1	构建工程	394
11.3.2	模拟数据	395
11.3.3	实时发送数据到 Kafka	399
11.3.4	实时分析平台答题数据	402
11.3.5	构建推荐模型	405
11.3.6	实时推荐题目	411
11.3.7	离线分析学习情况	415
11.4	本章小结	422

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。