



大数据创新人才  
培养系列

名校名师打造大数据领域精品力作

# Spark

## 编程基础

Python 版 | 第2版

林子雨◎主编 郑海山 赖永炫◎副主编

深入浅出，零基础入门

注重实战，贴近实际，详细介绍 Spark 编程基础知识

附  
微课视频

SPARK PROGRAMMING  
WITH PYTHON  
(2ND)



中国工信出版集团



人民邮电出版社  
POSTS & TELECOM PRESS

## 第1章 大数据技术概述 ..... 1

- 1.1 大数据概念与关键技术 ..... 1
  - 1.1.1 大数据概念 ..... 1
  - 1.1.2 大数据关键技术 ..... 2
- 1.2 代表性大数据技术 ..... 3
  - 1.2.1 Hadoop ..... 3
  - 1.2.2 Spark ..... 7
  - 1.2.3 Flink ..... 9
  - 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

## 第2章 Spark 的设计与运行原理 ..... 15

- 2.1 概述 ..... 15
- 2.2 Spark 生态系统 ..... 16
- 2.3 Spark 运行架构 ..... 18
  - 2.3.1 基本概念 ..... 18
  - 2.3.2 架构设计方法 ..... 18
  - 2.3.3 Spark 运行的基本流程 ..... 19
  - 2.3.4 RDD 的设计与运行原理 ..... 20
- 2.4 Spark 部署方式 ..... 29
- 2.5 本章小结 ..... 30
- 2.6 习题 ..... 30

## 第3章 大数据实验环境搭建 ..... 31

- 3.1 Linux 操作系统的安装 ..... 31
  - 3.1.1 下载安装文件 ..... 31
  - 3.1.2 Linux 操作系统的安装方式 ..... 32
  - 3.1.3 虚拟机和 Linux 操作系统的安装 ..... 33

## 3.2 Hadoop 的安装 ..... 39

- 3.2.1 Hadoop 简介 ..... 39
- 3.2.2 安装 Hadoop 前的准备工作 ..... 40
- 3.2.3 Hadoop 的 3 种安装模式 ..... 42
- 3.2.4 下载 Hadoop 安装文件 ..... 42
- 3.2.5 单机模式配置 ..... 43
- 3.2.6 伪分布式模式配置 ..... 43
- 3.2.7 分布式模式配置 ..... 47

## 3.3 MySQL 的安装 ..... 57

- 3.3.1 执行安装命令 ..... 57
- 3.3.2 启动 MySQL 服务 ..... 58
- 3.3.3 进入 MySQL Shell 界面 ..... 58
- 3.3.4 解决 MySQL 出现的中文乱码问题 ..... 58

## 3.4 Kafka 的安装 ..... 60

- 3.4.1 Kafka 简介 ..... 60
- 3.4.2 Kafka 的安装和使用 ..... 60

## 3.5 Anaconda 的安装和使用方法 ..... 61

## 3.6 本章小结 ..... 63

## 实验1 Linux、Hadoop 和 MySQL 的安装与使用 ..... 64

## 第4章 Spark 环境搭建和使用方法 ..... 66

## 4.1 安装 Spark (Local 模式) ..... 66

- 4.1.1 基础环境 ..... 66
- 4.1.2 下载安装文件 ..... 67
- 4.1.3 配置相关文件 ..... 67
- 4.1.4 验证 Spark 是否安装成功 ..... 68

## 4.2 在 PySpark 中运行代码 ..... 68

- 4.2.1 pyspark 命令 ..... 68
- 4.2.2 启动 PySpark ..... 69

## 4.3 使用 spark-submit 命令提交运行程序 ..... 70



4.4 Spark 集群环境搭建 (Standalone 模式) .....	70	5.1.4 分区 .....	105
4.4.1 集群概况 .....	71	5.1.5 综合实例 .....	109
4.4.2 搭建 Hadoop 集群 .....	71	5.2 键值对 RDD .....	110
4.4.3 安装 Anaconda3 .....	71	5.2.1 键值对 RDD 的创建 .....	111
4.4.4 在集群中安装 Spark .....	72	5.2.2 常用的键值对转换操作 .....	111
4.4.5 配置环境变量 .....	72	5.2.3 综合实例 .....	116
4.4.6 Spark 的配置 .....	72	5.3 数据读写 .....	117
4.4.7 启动 Spark 集群 .....	73	5.3.1 本地文件系统的数据读写 .....	117
4.4.8 关闭 Spark 集群 .....	74	5.3.2 分布式文件系统 HDFS 的数据读写 .....	118
4.5 在集群上运行 Spark 应用程序 .....	75	5.3.3 读写 MySQL 数据库 .....	119
4.5.1 启动 Spark 集群 .....	75	5.4 综合实例 .....	120
4.5.2 提交运行程序 .....	75	5.4.1 求 TOP 值 .....	120
4.6 Spark on YARN 模式 .....	76	5.4.2 文件排序 .....	124
4.6.1 概述 .....	76	5.4.3 二次排序 .....	126
4.6.2 Spark on YARN 模式的部署 .....	77	5.5 本章小结 .....	129
4.6.3 采用 YARN 模式运行 PySpark .....	77	5.6 习题 .....	129
4.6.4 通过 spark-submit 命令提交程序到 YARN 集群 .....	78	实验 3 RDD 编程初级实践 .....	130
4.6.5 Spark on YARN 的两种部署模式 .....	78	<b>第 6 章 Spark SQL .....</b>	<b>133</b>
4.7 安装 PySpark 类库 .....	79	6.1 Spark SQL 简介 .....	133
4.7.1 类库与框架的区别 .....	79	6.1.1 从 Shark 说起 .....	133
4.7.2 PySpark 类库的安装 .....	80	6.1.2 Spark SQL 架构 .....	135
4.8 开发 Spark 独立应用程序 .....	80	6.1.3 为什么推出 Spark SQL .....	135
4.8.1 编写程序 .....	80	6.1.4 Spark SQL 的特点 .....	136
4.8.2 通过 spark-submit 运行程序 .....	81	6.1.5 Spark SQL 简单编程实例 .....	136
4.9 PyCharm 的安装和使用 .....	81	6.2 结构化数据 DataFrame .....	137
4.9.1 安装 PyCharm .....	81	6.2.1 DataFrame 概述 .....	137
4.9.2 使用 PyCharm 开发 Spark 程序 .....	87	6.2.2 DataFrame 的优点 .....	138
4.10 本章小结 .....	89	6.3 DataFrame 的创建和保存 .....	139
4.11 习题 .....	89	6.3.1 Parquet .....	139
实验 2 Spark 的安装和使用 .....	89	6.3.2 JSON .....	139
<b>第 5 章 RDD 编程 .....</b>	<b>91</b>	6.3.3 CSV .....	140
5.1 RDD 编程基础 .....	91	6.3.4 文本文件 .....	141
5.1.1 RDD 创建 .....	91	6.3.5 序列集合 .....	141
5.1.2 RDD 操作 .....	93	6.4 DataFrame 的基本操作 .....	142
5.1.3 持久化 .....	104	6.4.1 DSL 语法风格 .....	142
		6.4.2 SQL 语法风格 .....	146
		6.5 从 RDD 转换得到 DataFrame .....	148
		6.5.1 利用反射机制推断 RDD 模式 .....	148

6.5.2 使用编程方式定义 RDD 模式	149
6.6 使用 Spark SQL 读写数据库	150
6.6.1 准备工作	150
6.6.2 读取 MySQL 数据库中的数据	151
6.6.3 向 MySQL 数据库写入数据	152
6.7 PySpark 和 pandas 的整合	153
6.7.1 PySpark 和 pandas 进行整合的可行性	153
6.7.2 pandas 数据结构	154
6.7.3 实例 1: 两种 DataFrame 之间的相互转换	155
6.7.4 实例 2: 使用自定义聚合函数	156
6.8 综合实例	157
6.9 本章小结	159
6.10 习题	159
实验 4 Spark SQL 编程初级实践	160

## 第 7 章 Spark Streaming ..... 162

7.1 流计算概述	162
7.1.1 静态数据和流数据	162
7.1.2 批量计算和实时计算	163
7.1.3 什么是流计算	164
7.1.4 流计算框架	164
7.1.5 流计算处理流程	165
7.2 Spark Streaming 概述	166
7.2.1 Spark Streaming 设计	167
7.2.2 Spark Streaming 与 Storm 的对比	168
7.2.3 从“Hadoop+Storm”架构转向 Spark 架构	168
7.3 DStream 操作概述	169
7.3.1 Spark Streaming 工作机制	169
7.3.2 编写 Spark Streaming 程序的基本步骤	170
7.3.3 创建 StreamingContext 对象	170
7.4 基本输入源	170
7.4.1 文件流	170
7.4.2 套接字流	172
7.4.3 RDD 队列流	176

7.5 转换操作	177
7.5.1 DStream 无状态转换操作	177
7.5.2 DStream 有状态转换操作	177
7.6 输出操作	182
7.6.1 把 DStream 输出到文本文件中	182
7.6.2 把 DStream 写入关系数据库中	183
7.7 本章小结	184
7.8 习题	185
实验 5 Spark Streaming 编程初级实践	185

## 第 8 章 Structured Streaming ..... 187

8.1 概述	187
8.1.1 基本概念	188
8.1.2 两种处理模型	189
8.1.3 Structured Streaming 和 Spark SQL、Spark Streaming 的关系	190
8.2 编写 Structured Streaming 程序的基本步骤	190
8.2.1 实现步骤	190
8.2.2 测试运行	192
8.3 输入源	194
8.3.1 File 源	194
8.3.2 Kafka 源	198
8.3.3 Socket 源	200
8.3.4 Rate 源	201
8.4 输出操作	202
8.4.1 启动流计算	203
8.4.2 输出模式	203
8.4.3 输出接收器	204
8.5 容错处理	205
8.5.1 从检查点恢复故障	206
8.5.2 故障恢复中的限制	206
8.6 迟到数据处理	206
8.6.1 事件时间	207
8.6.2 迟到数据	207

8.6.3 水印 .....	208
8.6.4 多水印规则 .....	209
8.6.5 处理迟到数据的实例 .....	210
8.7 查询的管理和监控 .....	213
8.7.1 管理和监控的方法 .....	213
8.7.2 一个监控的实例 .....	213
8.8 本章小结 .....	215
8.9 习题 .....	216
实验6 Structured Streaming 编程 实践 .....	216

## 第9章 Spark MLlib ..... 218

9.1 基于大数据的机器学习 .....	218
9.2 机器学习库 MLlib 概述 .....	219
9.3 基本的数据类型 .....	220
9.3.1 本地向量 .....	220
9.3.2 标注点 .....	221
9.3.3 本地矩阵 .....	222
9.3.4 数据源 .....	223
9.4 基本的统计分析工具 .....	224
9.4.1 相关性 .....	224
9.4.2 假设检验 .....	226
9.4.3 汇总统计 .....	227
9.5 机器学习流水线 .....	228
9.5.1 流水线的概念 .....	228
9.5.2 流水线的工作过程 .....	229

9.6 特征提取、特征转换、特征选择及 局部敏感散列 .....	230
9.6.1 特征提取 .....	231
9.6.2 特征转换 .....	234
9.6.3 特征选择 .....	239
9.6.4 局部敏感散列 .....	241
9.7 分类算法 .....	241
9.7.1 逻辑斯蒂回归分类算法 .....	242
9.7.2 决策树分类算法 .....	246
9.8 聚类算法 .....	250
9.8.1 K-Means 聚类算法 .....	250
9.8.2 GMM 聚类算法 .....	253
9.9 频繁模式挖掘算法 .....	256
9.9.1 FP-Growth 算法 .....	256
9.9.2 PrefixSpan 算法 .....	259
9.10 协同过滤算法 .....	261
9.10.1 协同过滤算法的原理 .....	262
9.10.2 ALS 算法 .....	262
9.11 模型选择 .....	266
9.11.1 模型选择工具 .....	266
9.11.2 用交叉验证选择模型 .....	267
9.12 本章小结 .....	269
9.13 习题 .....	270
实验7 Spark MLlib 编程初级实践 .....	270

## 参考文献 ..... 272