

目 录

第 1 章 Spark 概述 (1)

1.1 Spark 的背景 (1)

1.1.1 Spark 发展史 (1)

1.1.2 Spark 的特点 (2)

1.2 Spark 生态系统 (3)

1.2.1 Spark Core (3)

1.2.2 Spark SQL (4)

1.2.3 Spark Streaming (4)

1.2.4 GraphX (5)

1.2.5 MLBase/MLlib (5)

1.2.6 SparkR (5)

1.3 Spark 运行架构 (6)

1.3.1 相关术语 (6)

1.3.2 Spark 架构 (7)

1.3.3 执行步骤 (8)

1.3.4 Spark 运行模式 (10)

1.4 WordCount 示例 (13)

1.4.1 三种编程语言的示例程序 (13)

1.4.2 Scala 版本 WordCount 运行分析 (16)

1.4.3 WordCount 中的类调用关系 (18)

1.5 本章小结 (19)

思考与习题 (19)

第 2 章 搭建 Spark 开发环境 (20)

2.1 Spark 开发环境所需软件 (20)

2.2 安装 Spark (21)

2.2.1 spark-shell 下的实例 (25)

2.2.2 SparkWEB 的使用 (26)

2.3 IDEA (28)

2.3.1 安装 IDEA (28)

2.3.2 IDEA 的实例 (Scala) (32)

2.3.3	IDEA 打包运行	(37)
2.4	Eclipse	(40)
2.4.1	安装 Eclipse	(40)
2.4.2	Eclipse 的实例 (Scala)	(41)
2.5	本章小结	(46)
	思考与习题	(47)
第 3 章	Scala 语言基础	(48)
3.1	Scala 简介	(48)
3.1.1	Scala 特点	(48)
3.1.2	Scala 运行方式	(48)
3.2	变量与类型	(50)
3.2.1	变量的定义与使用	(50)
3.2.2	基本数据类型和操作	(56)
3.2.3	Range 操作	(61)
3.3	程序控制结构	(62)
3.3.1	if 条件表达式	(62)
3.3.2	循环表达式	(66)
3.3.3	匹配表达式	(70)
3.4	集合	(73)
3.4.1	数组	(73)
3.4.2	列表	(78)
3.4.3	集	(81)
3.4.4	映射	(85)
3.4.5	Option	(90)
3.4.6	迭代器与元组	(92)
3.5	函数式编程	(95)
3.5.1	函数	(95)
3.5.2	占位符语法	(97)
3.5.3	递归函数	(99)
3.5.4	嵌套函数	(101)
3.5.5	高阶函数	(102)
3.5.6	高阶函数的使用	(104)

3.6 本章小结	(108)
思考与习题	(108)
第4章 Scala 面向对象编程	(110)
4.1 类与对象	(110)
4.1.1 定义类	(110)
4.1.2 创建对象	(111)
4.1.3 类成员的访问	(112)
4.1.4 构造函数	(113)
4.1.5 常见对象类型	(116)
4.1.6 抽象类与匿名类	(118)
4.2 继承与多态	(120)
4.2.1 类的继承	(121)
4.2.2 构造函数执行顺序	(124)
4.2.3 方法重写	(125)
4.2.4 多态	(127)
4.3 特质 (trait)	(128)
4.3.1 特质的使用	(129)
4.3.2 特质与类	(132)
4.3.3 多重继承	(135)
4.4 导入和包	(137)
4.4.1 包	(137)
4.4.2 import 高级特性	(138)
4.5 本章小结	(141)
思考与习题	(141)
第5章 RDD 编程	(143)
5.1 RDD 基础	(143)
5.1.1 RDD 的基本特征	(143)
5.1.2 依赖关系	(144)
5.2 创建 RDD	(148)
5.2.1 从已有集合创建 RDD	(148)
5.2.2 从外部存储创建 RDD	(149)
5.3 RDD 操作	(150)

5.3.1 Transformation 操作	(151)
5.3.2 Action 操作	(159)
5.3.3 不同类型 RDD 之间的转换	(166)
5.4 数据的读取与保存	(168)
5.5 RDD 缓存与容错机制	(170)
5.5.1 RDD 的缓存机制 (持久化)	(170)
5.5.2 RDD 检查点容错机制	(173)
5.6 综合实例	(174)
5.7 本章小结	(179)
思考与习题	(180)
第 6 章 Spark SQL	(181)
6.1 Spark SQL 概述	(181)
6.1.1 Spark SQL 架构	(181)
6.1.2 程序主入口 SparkSession	(182)
6.1.3 DataFrame 与 RDD	(184)
6.2 创建 DataFrame	(185)
6.2.1 从外部数据源创建 DataFrame	(185)
6.2.2 RDD 转换为 DataFrame	(199)
6.3 DataFrame 操作	(203)
6.3.1 Transformation 操作	(204)
6.3.2 Action 操作	(216)
6.3.3 保存操作	(219)
6.4 Spark SQL 实例	(220)
6.5 本章小结	(226)
思考与习题	(226)
第 7 章 Spark Streaming	(228)
7.1 Spark Streaming 工作机制	(228)
7.1.1 Spark Streaming 工作流程	(228)
7.1.2 Spark Streaming 处理机制	(229)
7.2 DStream 输入源	(230)
7.2.1 基础输入源	(230)
7.2.2 高级输入源	(232)

7.3 DStream 转换操作	(233)
7.3.1 无状态转换操作	(233)
7.3.2 有状态转换操作	(234)
7.4 DStream 输出操作	(245)
7.5 Spark Streaming 处理流式数据	(246)
7.5.1 文件流	(246)
7.5.2 RDD 队列流	(248)
7.5.3 套接字流	(250)
7.5.4 Kafka 消息队列流	(251)
7.6 Spark Streaming 性能调优	(258)
7.6.1 减少批处理时间	(258)
7.6.2 设置适合的批次大小	(259)
7.6.3 优化内存使用	(259)
7.7 本章小结	(260)
思考与习题	(260)
第8章 Spark GraphX	(261)
8.1 GraphX 简介	(261)
8.2 GraphX 图存储	(262)
8.2.1 GraphX 的 RDD	(262)
8.2.2 GraphX 图分割	(264)
8.3 GraphX 图操作	(265)
8.3.1 构建图操作	(266)
8.3.2 基本属性操作	(268)
8.3.3 连接操作	(270)
8.3.4 转换操作	(271)
8.3.5 结构操作	(273)
8.3.6 聚合操作	(274)
8.3.7 缓存操作	(275)
8.3.8 Pregel API	(276)
8.4 内置的图算法	(279)
8.4.1 PageRank	(279)
8.4.2 计算三角形数	(282)

8.4.3	计算连通分量	(284)
8.4.4	标签传播算法	(285)
8.4.5	SVD+ +	(286)
8.5	GraphX 实现经典图算法	(288)
8.5.1	Dijkstra 算法	(288)
8.5.2	TSP 问题	(291)
8.5.3	最小生成树问题	(292)
8.6	GraphX 实例分析	(294)
8.6.1	寻找“最有影响力”论文	(294)
8.6.2	寻找社交媒体中的“影响力用户”	(296)
8.7	本章小结	(298)
	思考与习题	(299)
第 9 章	Spark 机器学习原理	(300)
9.1	Spark 机器学习简介	(300)
9.2	ML Pipeline	(301)
9.2.1	Pipeline 概念	(301)
9.2.2	Pipeline 工作过程	(302)
9.2.3	Pipeline 实例	(303)
9.3	Spark 机器学习数据准备	(310)
9.3.1	特征提取	(310)
9.3.2	特征转换	(314)
9.3.3	特征选择	(319)
9.4	算法调优	(326)
9.4.1	模型选择	(326)
9.4.2	交叉验证	(326)
9.4.3	TrainValidationSplit	(329)
9.5	本章小结	(331)
	思考与习题	(331)
第 10 章	Spark 机器学习模型	(332)
10.1	spark. ml 分类模型	(332)
10.1.1	spark. ml 分类模型简介	(332)
10.1.2	朴素贝叶斯分类器	(333)

10.1.3	朴素贝叶斯分类器程序示例	(335)
10.2	回归模型	(337)
10.2.1	spark. ml 回归模型简介	(338)
10.2.2	线性回归	(338)
10.2.3	线性回归程序示例	(341)
10.3	决策树	(343)
10.3.1	spark. ml 决策树模型简介	(343)
10.3.2	决策树分类	(345)
10.3.3	决策树分类程序示例	(347)
10.3.4	决策树回归	(350)
10.3.5	决策树回归程序示例	(354)
10.4	聚类模型	(357)
10.4.1	spark. ml 聚类模型简介	(358)
10.4.2	K-means 算法示例	(360)
10.4.3	K-means 程序示例	(362)
10.5	频繁模式挖掘	(363)
10.5.1	FP-Growth	(364)
10.5.2	FP-Growth 算法示例	(365)
10.5.3	FP-Growth 程序示例	(367)
10.6	本章小结	(369)
	思考与习题	(369)
	参考文献	(371)