

Spark大数据分析技术与实战



[Spark大数据分析技术与实战 下载链接1](#)

著者:董轶群

出版者:电子工业出版社

出版时间:2017-7

装帧:平装

isbn:9787121319037

Spark作为下一代大数据处理引擎，经过短短几年的飞跃式发展，正在以燎原之势席卷业界，现已成为大数据产业中的一股中坚力量。

《Spark大数据分析技术与实战》着重讲解了Spark内核、Spark GraphX、Spark SQL、Spark Streaming和Spark MLlib的核心概念与理论框架，并提供了相应的示例与解析。

《Spark大数据分析技术与实战》共分为8章，其中前4章介绍Spark内核，主要包括Spark简介、集群部署、工作原理、核心概念与操作等；后4章分别介绍Spark内核的核心组件，每章系统地介绍Spark的一个组件，并附以相应的案例分析。

《Spark大数据分析技术与实战》适合作为高等院校计算机相关专业的研究生学习参考资料，也适合大数据技术初学者阅读，还适合所有愿意对大数据技术有所了解并想要将大数据技术应用于本职工作的读者阅读。

作者介绍：

目录:	第1章 Spark导论	1
1.1	Spark的发展	2
1.2	什么是Spark	3
1.3	Spark主要特征	3
1.3.1	快速	3
1.3.2	简洁易用	5
1.3.3	通用	6
1.3.4	多种运行模式	8
第2章	Spark集群部署	9
2.1	运行环境说明	9
2.1.1	软硬件环境	9
2.1.2	集群网络环境	10
2.2	安装VMware Workstation	11
2.3	安装CentOS 6	16
2.4	安装Hadoop	21
2.4.1	克隆并启动虚拟机	21
2.4.2	网络基本配置	24
2.4.3	安装JDK	27
2.4.4	免密钥登录配置	28
2.4.5	Hadoop配置	29
2.4.6	配置从节点	33
2.4.7	配置系统文件	33
2.4.8	启动Hadoop集群	33
2.5	安装Scala	35
2.6	安装Spark	36
2.6.1	下载并解压Spark安装包	36
2.6.2	配置Spark-env.sh	37
2.6.3	配置Spark-defaults.conf	37
2.6.4	配置Slaves	38
2.6.5	配置环境变量	38
2.6.6	发送至Slave1、Slave2	39
2.7	启动Spark	39
第3章	RDD编程	42
3.1	RDD定义	42
3.2	RDD的特性	43
3.2.1	分区	43
3.2.2	依赖	44
3.2.3	计算	45

3.2.4 分区函数	45
3.2.5 优先位置	46
3.3 创建操作	46
3.3.1 基于集合的创建操作	47
3.3.2 基于外部存储的创建操作	47
3.4 常见执行操作	49
3.5 常见转换操作	49
3.5.1 一元转换操作	50
3.5.2 二元转换操作	53
3.6 持久化操作	56
3.7 存储操作	58
第4章 Spark调度管理与应用程序开发	59
4.1 Spark调度管理基本概念	59
4.2 作业调度流程	60
4.2.1 作业的生成与提交	61
4.2.2 阶段的划分	62
4.2.3 调度阶段的提交	62
4.2.4 任务的提交与执行	62
4.3 基于IntelliJ IDEA构建Spark应用程序	64
4.3.1 安装IntelliJ IDEA	64
4.3.2 创建Spark应用程序	70
4.3.3 集群模式运行Spark应用程序	81
第5章 GraphX	87
5.1 GraphX概述	87
5.2 GraphX基本原理	89
5.2.1 图计算模型处理流程	89
5.2.2 GraphX定义	90
5.2.3 GraphX的特点	90
5.2.4 GraphX设计与实现	91
5.3.1 弹性分布式属性图	91
5.3.2 图的数据模型	92
5.3.3 图的存储模型	94
5.3.4 GraphX模型框架	97
5.4 GraphX操作	97
5.4.1 创建图	97
5.4.2 基本属性操作	100
5.4.3 结构操作	102
5.4.4 转换操作	103
5.4.5 连接操作	105
5.4.6 聚合操作	106
5.5 GraphX案例解析	107
5.5.1 PageRank算法与案例解析	107
5.5.2 Triangle Count算法与案例解析	110
第6章 Spark SQL	113
6.1 Spark SQL概述	113
6.2 Spark SQL逻辑架构	116
6.2.1 SQL执行流程	116
6.2.2 Catalyst	117
6.3 Spark SQL CLI	117
6.3.1 硬软件环境	117
6.3.2 集群环境	118
6.3.3 结合Hive	118
6.3.4 启动Hive	118
6.4 DataFrame编程模型	119

6.4.1 DataFrame简介	119
6.4.2 创建DataFrames	120
6.4.3 保存DataFrames	126
6.5 DataFrame常见操作	127
6.5.1 数据展示	127
6.5.2 常用列操作	128
6.5.3 过滤	131
6.5.4 排序	132
6.5.5 其他常见操作	134
6.6 基于Hive的学生信息管理系统的SQL查询案例与解析	137
6.6.1 Spark SQL整合Hive	137
6.6.2 构建数据仓库	138
6.6.3 加载数据	141
6.6.4 查询数据	142
第7章 Spark Streaming	146
7.1 Spark Streaming概述	146
7.2 Spark Streaming基础概念	147
7.2.1 批处理时间间隔	147
7.2.2 窗口时间间隔	148
7.2.3 滑动时间间隔	148
7.3 DStream基本概念	149
7.4 DStream的基本操作	150
7.4.1 无状态转换操作	150
7.4.2 有状态转换操作	152
7.4.3 输出操作	153
7.4.4 持久化操作	154
7.5 数据源	154
7.5.1 基础数据源	154
7.5.2 高级数据源	155
7.6 Spark Streaming编程模式与案例分析	156
7.6.1 Spark Streaming编程模式	156
7.6.2 文本文件数据处理案例（一）	157
7.6.3 文本文件数据处理案例（二）	160
7.6.4 网络数据处理案例（一）	164
7.6.5 网络数据处理案例（二）	171
7.6.6 stateful应用案例	175
7.6.7 window应用案例	180
7.7 性能考量	185
7.7.1 运行时间优化	185
7.7.2 内存使用与垃圾回收	186
第8章 Spark MLlib	187
8.1 Spark MLlib概述	187
8.1.1 机器学习介绍	187
8.1.2 Spark MLlib简介	189
8.2 MLlib向量与矩阵	190
8.2.1 MLlib向量	190
8.2.2 MLlib矩阵	192
8.3 Spark MLlib分类算法	196
8.3.1 贝叶斯分类算法	197
8.3.2 支持向量机算法	201
8.3.3 决策树算法	204
8.4 MLlib线性回归算法	208
8.5 MLlib聚类算法	212
8.6 MLlib协同过滤	215

· · · · · (收起)

[Spark大数据分析技术与实战 下载链接1](#)

标签

大数据

算法

Spark

评论

干货不多，多数在堆砌函数式编程的语法。这种书随便看看就好了，还不如网上的教程靠谱。

[Spark大数据分析技术与实战 下载链接1](#)

书评

[Spark大数据分析技术与实战 下载链接1](#)