

Spark快速大数据分析



[Spark快速大数据分析_下载链接1](#)

著者:[美] Holden Karau

出版者:人民邮电出版社

出版时间:2015-10

装帧:

isbn:9787115403094

作者介绍:

Holden

Karau是Databricks的软件开发工程师，活跃于开源社区。她还著有《Spark快速数据处理》。

Andy Konwinski是Databricks联合创始人，Apache Spark项目技术专家，还是Apache Mesos项目的联合发起人。

Patrick Wendell是Databricks联合创始人，也是Apache Spark项目技术专家。他还负责维护Spark核心引擎的几个子系统。

Matei Zaharia是Databricks的CTO，同时也是Apache Spark项目发起人以及Apache基金会副主席。

目录: 目录

推荐序 xi

译者序 xiv

序 xvi

前言 xvii

第1章 Spark数据分析导论 1

1.1 Spark是什么 1

1.2 一个大一统的软件栈 2

1.2.1 Spark Core 2

1.2.2 Spark SQL 3

1.2.3 Spark Streaming 3

1.2.4 MLlib 3

1.2.5 GraphX 3

1.2.6 集群管理器 4

1.3 Spark的用户和用途 4

1.3.1 数据科学任务 4

1.3.2 数据处理应用 5

1.4 Spark简史 5

1.5 Spark的版本和发布 6

1.6 Spark的存储层次 6

第2章 Spark下载与入门 7

2.1 下载Spark 7

2.2 Spark中Python和Scala的shell 9

2.3 Spark 核心概念简介 12

2.4 独立应用 14

2.4.1 初始化SparkContext 15

2.4.2 构建独立应用 16

2.5 总结 19

第3章 RDD编程 21

3.1 RDD基础 21

3.2 创建RDD 23

3.3 RDD操作 24

3.3.1 转化操作 24

3.3.2 行动操作 26

3.3.3 惰性求值 27

3.4 向Spark传递函数 27

3.4.1 Python 27

3.4.2 Scala 28

- 3.4.3 Java 29
- 3.5 常见的转化操作和行动操作 30
 - 3.5.1 基本RDD 30
 - 3.5.2 在不同RDD类型间转换 37
- 3.6 持久化(缓存) 39
- 3.7 总结 40
- 第4章 键值对操作 41
 - 4.1 动机 41
 - 4.2 创建Pair RDD 42
 - 4.3 Pair RDD的转化操作 42
 - 4.3.1 聚合操作 45
 - 4.3.2 数据分组 49
 - 4.3.3 连接 50
 - 4.3.4 数据排序 51
 - 4.4 Pair RDD的行动操作 52
 - 4.5 数据分区 (进阶) 52
 - 4.5.1 获取RDD的分区方式 55
 - 4.5.2 从分区中获益的操作 56
 - 4.5.3 影响分区方式的操作 57
 - 4.5.4 示例: PageRank 57
 - 4.5.5 自定义分区方式 59
 - 4.6 总结 61
- 第5章 数据读取与保存 63
 - 5.1 动机 63
 - 5.2 文件格式 64
 - 5.2.1 文本文件 64
 - 5.2.2 JSON 66
 - 5.2.3 逗号分隔值与制表符分隔值 68
 - 5.2.4 SequenceFile 71
 - 5.2.5 对象文件 73
 - 5.2.6 Hadoop输入输出格式 73
 - 5.2.7 文件压缩 77
 - 5.3 文件系统 78
 - 5.3.1 本地/“常规”文件系统 78
 - 5.3.2 Amazon S3 78
 - 5.3.3 HDFS 79
 - 5.4 Spark SQL中的结构化数据 79
 - 5.4.1 Apache Hive 80
 - 5.4.2 JSON 80
 - 5.5 数据库 81
 - 5.5.1 Java数据库连接 81
 - 5.5.2 Cassandra 82
 - 5.5.3 HBase 84
 - 5.5.4 Elasticsearch 85
 - 5.6 总结 86
- 第6章 Spark编程进阶 87
 - 6.1 简介 87
 - 6.2 累加器 88
 - 6.2.1 累加器与容错性 90
 - 6.2.2 自定义累加器 91
 - 6.3 广播变量 91
 - 6.4 基于分区进行操作 94
 - 6.5 与外部程序间的管道 96
 - 6.6 数值RDD 的操作 99

6.7 总结	100
第7章 在集群上运行Spark	101
7.1 简介	101
7.2 Spark运行时架构	101
7.2.1 驱动器节点	102
7.2.2 执行器节点	103
7.2.3 集群管理器	103
7.2.4 启动一个程序	104
7.2.5 小结	104
7.3 使用spark-submit 部署应用	105
7.4 打包代码与依赖	107
7.4.1 使用Maven构建的用Java编写的Spark应用	108
7.4.2 使用sbt构建的用Scala编写的Spark应用	109
7.4.3 依赖冲突	111
7.5 Spark应用内与应用间调度	111
7.6 集群管理器	112
7.6.1 独立集群管理器	112
7.6.2 Hadoop YARN	115
7.6.3 Apache Mesos	116
7.6.4 Amazon EC2	117
7.7 选择合适的集群管理器	120
7.8 总结	121
第8章 Spark调优与调试	123
8.1 使用SparkConf配置Spark	123
8.2 Spark执行的组成部分：作业、任务和步骤	127
8.3 查找信息	131
8.3.1 Spark网页用户界面	131
8.3.2 驱动器进程和执行器进程的日志	134
8.4 关键性能考量	135
8.4.1 并行度	135
8.4.2 序列化格式	136
8.4.3 内存管理	137
8.4.4 硬件供给	138
8.5 总结	139
第9章 Spark SQL	141
9.1 连接Spark SQL	142
9.2 在应用中使用Spark SQL	144
9.2.1 初始化Spark SQL	144
9.2.2 基本查询示例	145
9.2.3 SchemaRDD	146
9.2.4 缓存	148
9.3 读取和存储数据	149
9.3.1 Apache Hive	149
9.3.2 Parquet	150
9.3.3 JSON	150
9.3.4 基于RDD	152
9.4 JDBC/ODBC服务器	153
9.4.1 使用Beeline	155
9.4.2 长生命周期的表与查询	156
9.5 用户自定义函数	156
9.5.1 Spark SQL UDF	156
9.5.2 Hive UDF	157
9.6 Spark SQL性能	158
9.7 总结	159

第10章 Spark Streaming	161
10.1 一个简单的例子	162
10.2 架构与抽象	164
10.3 转化操作	167
10.3.1 无状态转化操作	167
10.3.2 有状态转化操作	169
10.4 输出操作	173
10.5 输入源	175
10.5.1 核心数据源	175
10.5.2 附加数据源	176
10.5.3 多数据源与集群规模	179
10.6 24/7不间断运行	180
10.6.1 检查点机制	180
10.6.2 驱动器程序容错	181
10.6.3 工作节点容错	182
10.6.4 接收器容错	182
10.6.5 处理保证	183
10.7 Streaming用户界面	183
10.8 性能考量	184
10.8.1 批次和窗口大小	184
10.8.2 并行度	184
10.8.3 垃圾回收和内存使用	185
10.9 总结	185
第11章 基于MLlib的机器学习	187
11.1 概述	187
11.2 系统要求	188
11.3 机器学习基础	189
11.4 数据类型	192
11.5 算法	194
11.5.1 特征提取	194
11.5.2 统计	196
11.5.3 分类与回归	197
11.5.4 聚类	202
11.5.5 协同过滤与推荐	203
11.5.6 降维	204
11.5.7 模型评估	206
11.6 一些提示与性能考量	206
11.6.1 准备特征	206
11.6.2 配置算法	207
11.6.3 缓存RDD以重复使用	207
11.6.4 识别稀疏程度	207
11.6.5 并行度	207
11.7 流水线API	208
11.8 总结	209
作者简介	210
封面介绍	210
.	(收起)

[Spark快速大数据分析_下载链接1](#)

标签

大数据

spark

数据分析

Spark

计算机

bigdata

数据平台

技术

评论

入门书籍。很薄也很简洁。优点是把spark各个方面都介绍到了，缺点就是太简洁了，都没有很详细的分析个案例。

利用两个周末读完了，写得全面且易懂，spark入门的权威首选，第四作者是spark项目的发起人。spark是大数据分析的称手兵器，比hadoop mapreduce不知道高到哪去了，我准备跟它谈笑风生~

太好了，一本深入浅出，内容实用的中文Spark入门书籍。Learning Spark中译本

书是好书，就是版本有点儿旧，DataFrame之类的东西就跟进的不那么及时了。读完这

本书继续阅读官方文档，组合起来学习比较好。

大而全，就是代码有些旧跑不起来，偶调通在2.4.4下后重新整理的代码如下：
<https://github.com/greatabel/DataAnalysis/tree/master/i00Learning%20Spark>

入入门是很好的

简介，只能了解一部分：不讲分词，就不能说自己玩大数据？哈哈。spark是用来替代mapreduce的。

除了官方文档, 这是最好的入门教程

入门的书籍，感觉没有用一种语言围绕一个例子来讲解是一大遗憾，自己越往后看一些概念和方法理解起来越困难。但是总的来说，自己对spark的理解又深入了些。

不错的参考书

3.5 星，讲得比较浅显可以用来入门。看这书印象最深的就是函数式思想贯穿了 RDD 的设计与使用。scala 的表达力真得强，很多例子作者同时给出 scala java 两种语言写就的例程，对比强烈，once you go scala, you'll never go java.

Spark 扫盲书，让我对几个基本概念理解了

内容不错，但不是非常适合入门。相对于市面上其它书内容也比较新（大部分内容是 1.1 和 1.2 版本，但也有提到 1.4 里面的改变，现在最新版本是 1.5.2）。如果有些Scala/Java 和大数据基础理解起来不难，关于书中有些部分可以选读，比如部署方式有多种，你用

单机就看单机，用 Yarn 就看 Yarn 就好了，语言也是，用 Scala 也就不用关心 Java 和 python 的实现。由于我不做机器学习，所以最后一章就很粗略的翻了下了。

Spark介绍的很全面，文字通俗易懂，配套代码同时提供python, scala, java 3种语言，方便各类读者学习。看完对Spark有一定了解了。

入门不错，简单扼要

很清楚，入门不错

作为入门书不错，深度和广度的平衡性把握得比较好，对spark不熟悉的人可以通过这本书掌握spark的全貌。只是中译本的spark版本较老（v1.2），很多最新的进展没有包含在这本书里，有点遗憾。

入门好书，后面几章太浅了看看就行。

入门spark必备，市面上能看的入门书太少了。除了缺dataframe相关的，只能看官方文档了。

使用思路可以借鉴，接口版本老了一点

[Spark快速大数据分析_下载链接1](#)

书评

花了一天看完这本书，感觉这本书适合入门级人看，内容比较基础，没有阅读难度。给

