

《Spark大数据处理：技术、应用与性》

图书基本信息

书名：《Spark大数据处理：技术、应用与性能优化》

13位ISBN编号：9787111483863

出版时间：2014-11

作者：高彦杰

页数：268

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《Spark大数据处理：技术、应用与性》

内容概要

《Spark大数据处理：技术、应用与性能优化》根据最新技术版本，系统、全面、详细讲解Spark的各项功能使用、原理机制、技术细节、应用方法、性能优化，以及BDAS生态系统的相关技术。

作为一个基于内存计算的大数据并行计算框架，Spark不仅很好地解决了数据的实时处理问题，而且保证了高容错性和高可伸缩性。具体来讲，它有如下优势：

打造全栈多计算范式的高效数据流水线

轻量级快速处理

易于使用，支持多语言

与HDFS等存储层兼容

社区活跃度高

.....

Spark已经在全球范围内广泛使用，无论是Intel、Yahoo!、Twitter、阿里巴巴、百度、腾讯等国际互联网巨头，还是一些尚处于成长期的小公司，都在使用Spark。本书作者结合自己在微软和IBM实践Spark的经历和经验，编写了这本书。站着初学者的角度，不仅系统、全面地讲解了Spark的各项功能及其使用方法，而且较深入地探讨了Spark的工作机制、运行原理以及BDAS生态系统中的其他技术，同时还有一些可供操作的案例，能让没有经验的读者迅速掌握Spark。更为重要的是，本书还对Spark的性能优化进行了探讨。

《Spark大数据处理：技术、应用与性》

作者简介

高彦杰 毕业于中国人民大学，就职于IBM，精通Hadoop相关技术，较早接触并使用Spark，对Spark应用开发、Spark系统的运维和测试比较熟悉，深度阅读了Spark的源代码，了解Spark的运行机制，擅长Spark的查询优化。

书籍目录

前 言

第1章 Spark简介 1

- 1.1 Spark是什么 1
- 1.2 Spark生态系统BDAS 4
- 1.3 Spark架构 6
- 1.4 Spark分布式架构与单机多核架构的异同 9
- 1.5 Spark的企业级应用 10
 - 1.5.1 Spark在Amazon中的应用 11
 - 1.5.2 Spark在Yahoo!的应用 15
 - 1.5.3 Spark在西班牙电信的应用 17
 - 1.5.4 Spark在淘宝的应用 18
- 1.6 本章小结 20

第2章 Spark集群的安装与部署 21

- 2.1 Spark的安装与部署 21
 - 2.1.1 在Linux集群上安装与配置Spark 21
 - 2.1.2 在Windows上安装与配置Spark 30
- 2.2 Spark集群初试 33
- 2.3 本章小结 35

第3章 Spark计算模型 36

- 3.1 Spark程序模型 36
- 3.2 弹性分布式数据集 37
 - 3.2.1 RDD简介 38
 - 3.2.2 RDD与分布式共享内存的异同 38
 - 3.2.3 Spark的数据存储 39
- 3.3 Spark算子分类及功能 41
 - 3.3.1 Value型Transformation算子 42
 - 3.3.2 Key-Value型Transformation算子 49
 - 3.3.3 Actions算子 53
- 3.4 本章小结 59

第4章 Spark工作机制详解 60

- 4.1 Spark应用执行机制 60
 - 4.1.1 Spark执行机制总览 60
 - 4.1.2 Spark应用的概念 62
 - 4.1.3 应用提交与执行方式 63
- 4.2 Spark调度与任务分配模块 65
 - 4.2.1 Spark应用程序之间的调度 66
 - 4.2.2 Spark应用程序内Job的调度 67
 - 4.2.3 Stage和TaskSetManager调度方式 72
 - 4.2.4 Task调度 74
- 4.3 Spark I/O机制 77
 - 4.3.1 序列化 77
 - 4.3.2 压缩 78
 - 4.3.3 Spark块管理 80
- 4.4 Spark通信模块 93
 - 4.4.1 通信框架AKKA 94
 - 4.4.2 Client、Master和Worker间的通信 95

- 4.5 容错机制 104
 - 4.5.1 Lineage机制 104
 - 4.5.2 Checkpoint机制 108
- 4.6 Shuffle机制 110
- 4.7 本章小结 119
- 第5章 Spark开发环境配置及流程 120
 - 5.1 Spark应用开发环境配置 120
 - 5.1.1 使用Intellij开发Spark程序 120
 - 5.1.2 使用Eclipse开发Spark程序 125
 - 5.1.3 使用SBT构建Spark程序 129
 - 5.1.4 使用Spark Shell开发运行Spark程序 130
 - 5.2 远程调试Spark程序 130
 - 5.3 Spark编译 132
 - 5.4 配置Spark源码阅读环境 135
 - 5.5 本章小结 135
- 第6章 Spark编程实战 136
 - 6.1 WordCount 136
 - 6.2 Top K 138
 - 6.3 中位数 140
 - 6.4 倒排索引 141
 - 6.5 CountOnce 143
 - 6.6 倾斜连接 144
 - 6.7 股票趋势预测 146
 - 6.8 本章小结 153
- 第7章 Benchmark使用详解 154
 - 7.1 Benchmark简介 154
 - 7.1.1 Intel Hibench与Berkeley BigDataBench 155
 - 7.1.2 Hadoop GridMix 157
 - 7.1.3 Bigbench、BigDataBenchmark与TPC-DS 158
 - 7.1.4 其他Benchmark 161
 - 7.2 Benchmark的组成 162
 - 7.2.1 数据集 162
 - 7.2.2 工作负载 163
 - 7.2.3 度量指标 167
 - 7.3 Benchmark的使用 168
 - 7.3.1 使用Hibench 168
 - 7.3.2 使用TPC-DS 170
 - 7.3.3 使用BigDataBench 172
 - 7.4 本章小结 176
- 第8章 BDAS简介 177
 - 8.1 SQL on Spark 177
 - 8.1.1 使用Spark SQL的原因 178
 - 8.1.2 Spark SQL架构分析 179
 - 8.1.3 Shark简介 182
 - 8.1.4 Hive on Spark 184
 - 8.1.5 未来展望 185
 - 8.2 Spark Streaming 185
 - 8.2.1 Spark Streaming简介 186
 - 8.2.2 Spark Streaming架构 188

- 8.2.3 Spark Streaming原理剖析 189
- 8.2.4 Spark Streaming调优 198
- 8.2.5 Spark Streaming 实例 198
- 8.3 GraphX 205
 - 8.3.1 GraphX简介 205
 - 8.3.2 GraphX的使用 206
 - 8.3.3 GraphX架构 209
 - 8.3.4 运行实例 211
- 8.4 MLlib 215
 - 8.4.1 MLlib简介 217
 - 8.4.2 MLlib的数据存储 219
 - 8.4.3 数据转换为向量（向量空间模型VSM） 222
 - 8.4.4 MLlib中的聚类和分类 223
 - 8.4.5 算法应用实例 228
 - 8.4.6 利用MLlib进行电影推荐 230
- 8.5 本章小结 237
- 第9章 Spark性能调优 238
 - 9.1 配置参数 238
 - 9.2 调优技巧 239
 - 9.2.1 调度与分区优化 240
 - 9.2.2 内存存储优化 243
 - 9.2.3 网络传输优化 249
 - 9.2.4 序列化与压缩 251
 - 9.2.5 其他优化方法 253
 - 9.3 本章小结 255

精彩短评

- 1、 计算机科学
- 2、 2016 NO.7 还是有不少收获的，要是用 Java 代码就好了
- 3、 章节的安排，先难后易，抛出一大堆概念和原理。。有点谭浩强，太不注重实践了
- 4、 朋友的书，支持一下~
- 5、 真的是很差,配图和文字都有诸多错误.
- 6、 毕业纪念书籍
- 7、 分布式系统
- 8、 错误有点多 居然没有找到网上勘误
- 9、 第三，四两章不错，包含rdd操作还有spark的工作机制。
- 10、 初学者看一看还是蛮有帮助，不过还是看官方手册来的靠谱些。
- 11、 可以入门，但是编辑的细节错误不少
- 12、 这本书里边有比较多的细节错误或者描述不清。但是总体还行，读过之后可以对spark的某些方面的实现和优化，有一定的理解。
- 13、 比王家林什么的好太多了。
- 14、 书的内容还不错，就是内容顺序不太好，总感觉读起来不顺畅，有些章节顺序对调下，可能会比较好
- 15、 适合初学者学习Spark，讲的都比较浅显。
- 16、 很全面，也很技术。
- 17、 内容很少，且大多泛泛而谈，很多东西都是网上能够找到的东西，不推荐
- 18、 理论还要联系实践呀
- 19、 浏览了第4章 ...
- 20、 内容还不错，整体上把Spark的运行逻辑和注意点都理清了，但不少地方有些粗糙。
- 21、 了解spark系统原理的较好的入门书
- 22、 作为新手入门来说还不错，内容挺浅的。

《Spark大数据处理：技术、应用与性》

精彩书评

1、 有误，比如join，spark中的join是inner join，书中对着源码讲成了outer join..跑题，花了大量篇幅在FIFO，HASHMap的原理上...以上不过很多地方还是很详细的，而且通俗易懂

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:www.tushu000.com