

《Apache Spark源码剖析》

图书基本信息

书名：《Apache Spark源码剖析》

13位ISBN编号：9787121254204

出版时间：2015-3

作者：许鹏

页数：304

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

《Apache Spark源码剖析》

内容概要

《Apache Spark源码剖析》以Spark 1.02版本源码为切入点，着力于探寻Spark所要解决的主要问题及其解决办法，通过一系列精心设计的小实验来分析每一步背后的处理逻辑。

《Apache Spark源码剖析》第3~5章详细介绍了Spark Core中作业的提交与执行，对容错处理也进行了详细分析，有助读者深刻把握Spark实现机理。第6~9章对Spark Lib库进行了初步的探索。在对源码有了一定的分析之后，读者可尽快掌握Spark技术。

《Apache Spark源码剖析》对于Spark应用开发人员及Spark集群管理人员都有极好的学习价值；对于那些想从源码学习而又不知如何入手的读者，也不失为一种借鉴。

《Apache Spark源码剖析》

作者简介

许鹏长期致力于电信领域和互联网的软件开发，在数据处理方面积累了大量经验，对系统的可扩展性、可靠性方面进行过深入学习和研究。因此，累积了大量的源码阅读和分析的技巧与方法。目前在杭州同盾科技担任大数据平台架构师一职。对于Linux内核，作者也曾进行过深入的分析。

书籍目录

第一部分Spark概述1

第1章初识Spark 3

1.1 大数据和Spark 3

1.1.1 大数据的由来4

1.1.2 大数据的分析4

1.1.3 Hadoop 5

1.1.4 Spark简介6

1.2 与Spark的第一次亲密接触7

1.2.1 环境准备7

1.2.2 下载安装Spark 8

1.2.3 Spark下的WordCount 8

第二部分Spark核心概念13

第2章Spark整体框架 15

2.1 编程模型15

2.1.1 RDD 17

2.1.2 Operation 17

2.2 运行框架18

2.2.1 作业提交18

2.2.2 集群的节点构成18

2.2.3 容错处理19

2.2.4 为什么是Scala 19

2.3 源码阅读环境准备19

2.3.1 源码下载及编译19

2.3.2 源码目录结构21

2.3.3 源码阅读工具21

2.3.4 本章小结22

第3章SparkContext初始化 23

3.1 spark-shell 23

3.2 SparkContext的初始化综述27

3.3 Spark Repl综述30

3.3.1 Scala Repl执行过程31

3.3.2 Spark Repl 32

第4章Spark作业提交 33

4.1 作业提交33

4.2 作业执行38

4.2.1 依赖性分析及Stage划分39

4.2.2 Actor Model和Akka 46

4.2.3 任务的创建和分发47

4.2.4 任务执行53

4.2.5 Checkpoint和Cache 62

4.2.6 WebUI和Metrics 62

4.3 存储机制71

4.3.1 Shuffle结果的写入和读取71

4.3.2 Memory Store 80

4.3.3 存储子模块启动过程分析81

4.3.4 数据写入过程分析82

4.3.5 数据读取过程分析84

- 4.3.6 TachyonStore 88
- 第5章部署方式分析 91
- 5.1 部署模型91
- 5.2 单机模式local 92
- 5.3 伪集群部署local-cluster 93
- 5.4 原生集群Standalone Cluster 95
- 5.4.1 启动Master 96
- 5.4.2 启动Worker 97
- 5.4.3 运行spark-shell 102
- 5.4.4 容错性分析106
- 5.5 Spark On YARN 112
- 5.5.1 YARN的编程模型112
- 5.5.2 YARN中的作业提交112
- 5.5.3 Spark On YARN实现详解113
- 5.5.4 SparkPi on YARN 122
- 第三部分Spark Lib 129
- 第6章Spark Streaming 131
- 6.1 Spark Streaming整体架构131
- 6.1.1 DStream 132
- 6.1.2 编程接口133
- 6.1.3 Streaming WordCount 134
- 6.2 Spark Streaming执行过程135
- 6.2.1 StreamingContext初始化过程136
- 6.2.2 数据接收141
- 6.2.3 数据处理146
- 6.2.4 BlockRDD 155
- 6.3 窗口操作158
- 6.4 容错性分析159
- 6.5 Spark Streaming vs. Storm 165
- 6.5.1 Storm简介165
- 6.5.2 Storm和Spark Streaming对比168
- 6.6 应用举例168
- 6.6.1 搭建Kafka Cluster 168
- 6.6.2 KafkaWordCount 169
- 第7章SQL 173
- 7.1 SQL语句的通用执行过程分析175
- 7.2 SQL On Spark的实现分析178
- 7.2.1 SqlParser 178
- 7.2.2 Analyzer 184
- 7.2.3 Optimizer 191
- 7.2.4 SparkPlan 192
- 7.3 Parquet 文件和JSON数据集196
- 7.4 Hive简介197
- 7.4.1 Hive 架构197
- 7.4.2 HiveQL On MapReduce执行过程分析199
- 7.5 HiveQL On Sparki详解200
- 7.5.1 Hive On Spark环境搭建206
- 7.5.2 编译支持Hadoop 2.x的Spark 211
- 7.5.3 运行Hive On Spark测试用例213

第8章GraphX	215
8.1 GraphX简介	215
8.1.1 主要特点	216
8.1.2 版本演化	216
8.1.3 应用场景	217
8.2 分布式图计算处理技术介绍	218
8.2.1 属性图	218
8.2.2 图数据的存储与分割	219
8.3 Pregel计算模型	220
8.3.1 BSP	220
8.3.2 像顶点一样思考	220
8.4 GraphX图计算框架实现分析	223
8.4.1 基本概念	223
8.4.2 图的加载与构建	226
8.4.3 图数据存储与分割	227
8.4.4 操作接口	228
8.4.5 Pregel在GraphX中的源码实现	230
8.5 PageRank	235
8.5.1 什么是PageRank	235
8.5.2 PageRank核心思想	235
第9章MLLib	239
9.1 线性回归	239
9.1.1 数据和估计	240
9.1.2 线性回归参数求解方法	240
9.1.3 正则化	245
9.2 线性回归的代码实现	246
9.2.1 简单示例	246
9.2.2 入口函数train	247
9.2.3 最优化算法optimizer	249
9.2.4 权重更新update	256
9.2.5 结果预测predict	257
9.3 分类算法	257
9.3.1 逻辑回归	258
9.3.2 支持向量机	260
9.4 拟牛顿法	261
9.4.1 数学原理	261
9.4.2 代码实现	265
9.5 MLLib与其他应用模块间的整合	268
第四部分附录	271
附录A Spark源码调试	273
附录B 源码阅读技巧	283

《Apache Spark源码剖析》

精彩短评

- 1、代码贴太多，没看代码的我很快就过了一遍
- 2、有所收获，读源码的方式值得学习。篇幅比较短，贴了比较多的代码，有些地方解释的不够。domain相关的几个lib的细节等到做过一些实际应用后再对照或许会有更好的理解
- 3、贴代码太多，图例和讲解太少，但好在基本讲清楚，后面对Streaming、SQL、GraphX、Mllib基本讲解也不错。（2015.6.2jd）
- 4、我错了我再也不买国人写的书了
- 5、有几个问题：一，图太少，流程图类图这些几乎没有，描述很无力；二，贴代码不写明在哪个文件内，很难结合书和代码一起看；三，版本略久，和最新版代码差别较大
- 6、我的意见是很一般般！讲基本原理的部分没有讲太清楚，含含糊糊的，讲底层的时候全是大段的代码，让人云里雾里。
- 7、支持的是1.0版本，有点老，配合官方文档一起食用
- 8、有所收获，但是这本书的讲解确实太粗略了。我觉得如果三四百页都用来讲解spark-core差不多。简单地列出代码加上些许讲解，这能称为源码剖析？

《Apache Spark源码剖析》

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