

Spark最佳实践



[Spark最佳实践_下载链接1](#)

著者:陈 欢

出版者:人民邮电出版社

出版时间:2016-5

装帧:平装

isbn:9787115422286

本书是Spark实战指南，全书共分8章。前4章介绍Spark的部署、工作机制和内核，后4章分别通过实战项目介绍Spark SQL、Spark Streaming、Spark GraphX和Spark MLlib功能模块。此外，本书详细介绍了常见的实战问题，比如大数据环境下的配置设置、程序调优等。本书附带的一键安装脚本，更能为初学者提供很大帮助。

作者介绍:

陈欢

腾讯资深程序员，15年编码经验，曾任职网络安全、互联网金融等部门，亲手从零建设了财付通业务的Spark集群，并使之同时支持SQL、实时计算、机器学习等多种数据计算场景。他目前就职于腾讯社交与效果广告部，从事大数据分析工作。

林世飞

腾讯资深研究员，2005年加入腾讯，先后在无线产品、安全中心、搜索平台、开放平台、社交与效果广告部等部门从事开发和团队管理工作。他对网络安全、搜索引擎、数据挖掘、机器学习有一定了解，热衷知识传播和分享，曾获腾讯学院2009年年度优秀讲师。目前，他就职于社交与效果广告部，负责广告系统相关的研发工作。

目录: 第1章 Spark与大数据 1

1.1 大数据的发展及现状 1

1.1.1 大数据时代所面临的问题 1

1.1.2 谷歌的大数据解决方案 2

1.1.3 Hadoop生态系统 3

1.2 Spark应时而生 4

1.2.1 Spark的起源 4

1.2.2 Spark的特点 5

1.2.3 Spark的未来发展 6

第2章 Spark基础 8

2.1 Spark本地单机模式体验 8

2.1.1 安装虚拟机 8

2.1.2 安装JDK 19

2.1.3 下载Spark预编译包 21

2.1.4 本地体验Spark 22

2.2 高可用Spark分布式集群部署 25

2.2.1 集群总览 26

2.2.2 集群机器的型号选择 28

2.2.3 初始化集群机器环境 29

2.2.4 部署ZooKeeper集群 33

2.2.5 编译Spark 35

2.2.6 部署Spark Standalone集群 37

2.2.7 高可用Hadoop集群 40

2.2.8 让Spark运行在YARN上 40

2.2.9 一键部署高可用Hadoop + Spark集群 42

2.3 Spark编程指南 43

2.3.1 交互式编程 43

2.3.2 RDD创建 44

2.3.3 RDD操作 47

2.3.4 使用其他语言开发Spark程序 54

2.4 打包和提交 54

2.4.1 编译、链接、打包 54

2.4.2 提交 56

第3章 Spark工作机制 58

3.1 调度管理 58

3.1.1 集群概述及名词解释 58

3.1.2 Spark程序之间的调度	60
3.1.3 Spark程序内部的调度	63
3.2 内存管理	65
3.2.1 RDD持久化	65
3.2.2 共享变量	66
3.3 容错机制	67
3.3.1 容错体系概述	67
3.3.2 Master节点失效	68
3.3.3 Slave节点失效	69
3.4 监控管理	69
3.4.1 Web界面	69
3.4.2 REST API	72
3.4.3 Metrics指标体系	73
3.4.4 其他监控工具	73
3.5 Spark程序配置管理	73
3.5.1 Spark程序配置加载过程	74
3.5.2 环境变量配置	74
3.5.3 Spark属性项配置	74
3.5.4 查看当前的配置	76
3.5.5 配置Spark日志	76
第4章 Spark内核讲解	77
4.1 Spark核心数据结构RDD	77
4.1.1 RDD的定义	78
4.1.2 RDD的Transformation	80
4.1.3 RDD的Action	82
4.1.4 Shuffle	83
4.2 SparkContext	84
4.2.1 SparkConf配置	84
4.2.2 初始化过程	85
4.2.3 其他功能接口	87
4.3 DAG调度	87
4.3.1 DAGScheduler	87
4.3.2 TaskScheduler	90
第5章 Spark SQL与数据仓库	92
5.1 Spark SQL基础	93
5.1.1 分布式SQL引擎	93
5.1.2 支持的SQL语法	97
5.1.3 支持的数据类型	98
5.1.4 DataFrame	99
5.1.5 DataFrame数据源	103
5.1.6 性能调优	104
5.2 Spark SQL原理和运行机制	104
5.2.1 Spark SQL整体架构	105
5.2.2 Catalyst执行优化器	105
5.3 应用场景：基于淘宝数据建立电商数据仓库	110
5.3.1 电商数据仓库场景	111
5.3.2 数据准备和表设计	111
5.3.3 用Spark SQL来完成日常运营数据分析	115
5.3.4 Spark SQL在大规模数据下的性能表现	120
第6章 Spark流式计算	122
6.1 Spark Streaming基础知识	123
6.1.1 入门简单示例	123
6.1.2 基本概念	124
6.1.3 高级操作	129

6.2 深入理解Spark Streaming	132
6.2.1 DStream的两类操作	132
6.2.2 容错处理	134
6.2.3 性能调优	136
6.2.4 与Storm的对比	137
6.3 应用场景：一个类似百度统计的流式实时系统	139
6.3.1 Web log实时统计场景	139
6.3.2 日志实时采集	140
6.3.3 流式分析系统实现	140
第7章 Spark图计算	149
7.1 什么是图计算	149
7.1.1 图的基本概念	149
7.1.2 图计算的应用	150
7.2 Spark GraphX简介	151
7.2.1 GraphX实现	151
7.2.2 GraphX常用API介绍	152
7.3 应用场景：基于新浪微博数据的社交网络分析	153
7.3.1 社交网络分析的主要应用	153
7.3.2 社区发现算法简介	154
7.3.3 用GraphX实现Louvain算法	156
7.3.4 小试牛刀：谁是你的闺蜜	162
7.3.5 真实的场景：新浪微博关系分析	164
第8章 Spark MLlib	169
8.1 机器学习简介	169
8.1.1 什么是机器学习	169
8.1.2 机器学习示例	171
8.1.3 机器学习的基本方法	172
8.1.4 机器学习的常见技巧	173
8.1.5 机器学习参考资料	174
8.2 MLlib库简介	174
8.2.1 基础数据类型	174
8.2.2 主要的库	175
8.2.3 附带的示例程序	176
8.3 应用场景：搜索广告点击率预估系统	178
8.3.1 应用场景	178
8.3.2 逻辑回归	179
8.3.3 学习算法	181
8.3.4 模型评估	184
8.3.5 数据准备	186
8.3.6 模型训练	187
8.3.7 模型调优	195
附录 Scala语言参考	197
• • • • •	(收起)

[Spark最佳实践 下载链接1](#)

标签

spark

大数据

Spark

计算机

技术

没有含量

中国

~大数据

评论

这书写的我只能给三分了。看到连城推荐鹅厂的实践就下单，但是看到集群搭建standalone和yarn模式混为一潭顿时觉得索然无味，坚持读完觉得写的还算是全面，水平入门看看还算有零星收获，书名是入门必备可给四分，称为最佳实践的话只能三分了……

浪费钱。

刚看了前两章，还不错，有自己不知道或是遗漏的知识点 很扣细节啊。 === 多实践

书名最佳实践虽然起得比较虚，但还是有一定内容的，应用场景更丰富些会更好点。

出版社赠书，拖到今天随手翻完，感觉像是作者自己的学习笔记，该深不深该浅不浅。

也许这就是最佳实践？

最佳实践真谈不上，spark 一览还行

同样建议准备看这本的同学，还不如好好看看《Spark快速大数据分析》

很简单的，方便小白快速入门的书吧。作为编程类的书来说，200多页，就知道内容是什么样的了。

说是入门吧，却没把任何一点讲明白，贪大图全；说是高阶吧，却都只涉及皮毛，蜻蜓点水。国人写书真的是乏善可陈。如果没看过《Spark快速大数据分析》，就去看看，而不要看这本；如果已经看过了，就不要看这本。还有封底那些不负责任的推荐者，真不知道他们是不是看了书的内容。

快速的看完了，觉得spark实例部分还可以。

这本书写得并不好。。。怎么说呢？我不知道是作者本身不擅长写这种代码讲解的书，还是作者故意模糊了很多需要详细讲解的细节，整本书不通透，肥肠不适合初学者学习。大家看书都是从不懂到懂，为了学习，但是这本书很奇怪，一讲概念就罗里吧嗦；（1-4章难道不是应该一章就完吗？？？）一讲代码就几句话略过，甚至很多都只给结果，不给过程？？excuse me？谁想看你的结果？

[Spark最佳实践 下载链接1](#)

书评

[Spark最佳实践_下载链接1](#)