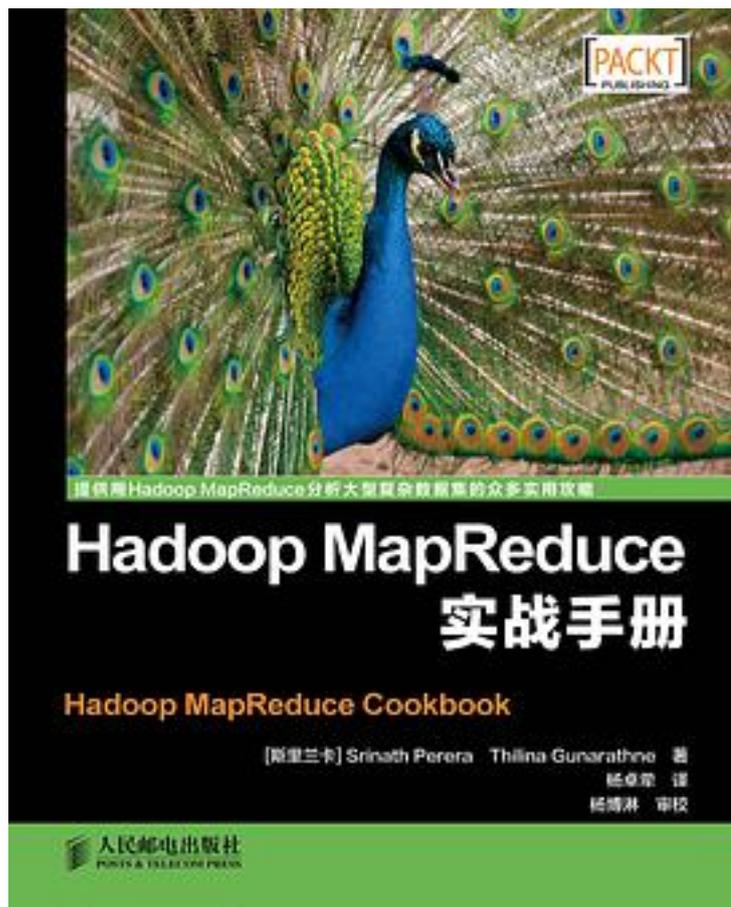


Hadoop MapReduce实战手册



[Hadoop MapReduce实战手册_下载链接1](#)

著者:[斯里兰卡] 萨那斯·佩雷拉 (Srinath Perera)

出版者:人民邮电出版社

出版时间:2015-3

装帧:

isbn:9787115384379

这是一本学习Hadoop MapReduce的一站式指南，完整介绍了Hadoop生态体系，包括Hadoop平台安装、部署、运维等，Hadoop生态系统成员Hive、Pig、HBase、Mahout等。最重要的是，书中包含丰富的示例和多样的实际应用场景，以一种简单而直接的方式呈现了90个实战攻

略，并给出一步步的指导。本书从获取Hadoop并在集群中运行讲起，依次介绍了高级HDFS，高级Hadoop MapReduce管理，开发复杂的Hadoop MapReduce应用程序，Hadoop的生态系统，统计分析，搜索与索引，聚类、推荐和寻找关联，海量文本数据处理，云部署等内容。

作者介绍:

作者介绍

Srinath

Perera是WSO2公司的高级软件架构师，与CTO一同全观整个WSO2平台架构。同时，他也是斯里兰卡软件基金会的一位研究科学家，并作为访问学者在莫勒图沃大学计算机科学与工程系授课。他是Apache Axis2开源软件项目的联合创始人，他自2002年以来一直参与Apache Web Service项目，并且是Apache软件基金会和Apache Web服务项目PMC的成员。Srinath也是Apache Axis、Axis2和Geronimo开源项目的committer。

他在美国印第安纳大学伯明顿分校获得博士和硕士学位，在斯里兰卡莫勒图沃大学获得了计算科学与工程学士学位。

Srinath已经撰写了许多技术文章和同行评审的研究文章，可以从他的个人网站找到更多细节。他还经常在技术会议上做演讲。

他长期研究大规模分布式系统。他的日常工作与大数据技术（如Hadoop和Cassandra）结合很紧密。他还在莫勒图沃大学研究生班教授并行计算，主要是基于Hadoop。

Thilina Gunarathne是印第安纳大学信息与计算学院博士。他在使用Apache Hadoop以及大规模数据密集型计算技术方面有着丰富的经验。他目前的主要工作是致力于研发在云环境执行可扩展的、高效的大规模数据密集型计算的技术。

Thilina发表了很多论文，并且同行评审了很多分布式计算和并行计算领域的研究论文，包括一些在云环境扩展MapReduce模型进行有效的数据挖掘和数据分析的论文。Thilina经常在学术界和工业界会议上发表演讲。

Thilina自2005年以来，在Apache软件基金会下贡献了若干个开源项目，并成为committer和PMC成员。在开始研究生学习之前，Thilina在WSO2公司担任高级软件工程师，专注于开源中间件开发。Thilina 2006年在斯里兰卡莫勒图沃大学获得计算机科学与工程学士学位，2009年在美国印第安纳大学伯明顿分校获得计算机科学硕士学位，2013年获得分布式和并行计算领域博士学位。

译者介绍

杨卓萃

阿里巴巴集团数据平台事业部资深研发工程师。2011年起，在阿里巴巴从事Hadoop五年，集团SQL on Hadoop负责人，Hadoop/Yarn/Hive contributor，开源软件爱好者。

目录: 第1章 搭建Hadoop并在集群中运行 1

1.1 简介 1

1.2 在你的机器上安装Hadoop 2

1.3 写WordCountMapReduce示例程序，打包并使用独立的Hadoop运行它 3

- 1.4 给WordCount MapReduce程序增加combiner步骤 8
- 1.5 安装HDFS 9
- 1.6 使用HDFS监控UI 14
- 1.7 HDFS的基本命令行文件操作 15
- 1.8 在分布式集群环境中设置Hadoop 17
- 1.9 在分布式集群环境中运行WordCount程序 22
- 1.10 使用MapReduce监控UI 24
- 第2章 HDFS进阶 26
 - 2.1 简介 26
 - 2.2 HDFS基准测试 27
 - 2.3 添加一个新的DataNode 28
 - 2.4 DataNode下架 30
 - 2.5 使用多个磁盘/卷以及限制HDFS的磁盘使用情况 32
 - 2.6 设置HDFS块大小 33
 - 2.7 设置文件冗余因子 34
 - 2.8 使用HDFS的Java API 35
 - 2.9 使用HDFS的C API (libhdfs) 40
 - 2.10 挂载HDFS (Fuse-DFS) 45
 - 2.11 在HDFS中合并文件 48
- 第3章 高级Hadoop MapReduce运维 49
 - 3.1 简介 49
 - 3.2 调优集群部署的Hadoop配置 49
 - 3.3 运行基准测试来验证Hadoop的安装 52
 - 3.4 复用Java虚拟机以提高性能 54
 - 3.5 容错和推测执行 54
 - 3.6 调试脚本——分析任务失败 55
 - 3.7 设置失败百分比以及跳过不良记录 59
 - 3.8 共享用户的Hadoop集群——使用公平调度器和其他调度器 61
 - 3.9 Hadoop的安全性——整合使用Kerberos 62
 - 3.10 使用Hadoop的工具接口 69
- 第4章 开发复杂的Hadoop MapReduce应用程序 72
 - 4.1 简介 72
 - 4.2 选择合适的Hadoop数据类型 73
 - 4.3 实现自定义的Hadoop Writable数据类型 75
 - 4.4 实现自定义Hadoop key类型 79
 - 4.5 从mapper中输出不同值类型的数据 83
 - 4.6 为输入数据格式选择合适的Hadoop InputFormat 87
 - 4.7 添加新的输入数据格式的支持——实现自定义的InputFormat 90
 - 4.8 格式化MapReduce计算的结果——使用Hadoop的OutputFormat 94
 - 4.9 Hadoop的中间 (map到reduce) 数据分区 96
 - 4.10 将共享资源传播和分发到MapReduce作业的任务中——Hadoop DistributedCache 98
 - 4.11 在Hadoop上使用传统应用程序——Hadoop Streaming 103
 - 4.12 添加MapReduce作业之间的依赖关系 106
 - 4.13 用于报告自定义指标的Hadoop计数器 108
- 第5章 Hadoop生态系统 110
 - 5.1 简介 110
 - 5.2 安装HBase 111
 - 5.3 使用Java客户端API随机存取数据 114
 - 5.4 基于HBase (表输入/输出) 运行MapReduce作业 116
 - 5.5 安装Pig 120
 - 5.6 运行第一条Pig命令 121
 - 5.7 使用Pig执行集合操作 (join, union) 与排序 123
 - 5.8 安装Hive 125

- 5.9 使用Hive运行SQL风格的查询 127
- 5.10 使用Hive执行join 129
- 5.11 安装Mahout 132
- 5.12 使用Mahout运行K-means 133
- 5.13 可视化K-means结果 136
- 第6章 分析 138
 - 6.1 简介 138
 - 6.2 使用MapReduce的简单分析 139
 - 6.3 使用MapReduce执行Group-By 143
 - 6.4 使用MapReduce计算频率分布和排序 146
 - 6.5 使用GNU Plot绘制Hadoop计算结果 148
 - 6.6 使用MapReduce计算直方图 151
 - 6.7 使用MapReduce计算散点图 154
 - 6.8 用Hadoop解析复杂的数据集 158
 - 6.9 使用MapReduce连接两个数据集 164
- 第7章 搜索和索引 170
 - 7.1 简介 170
 - 7.2 使用Hadoop MapReduce生成倒排索引 170
 - 7.3 使用Apache Nutch构建域内网络爬虫 175
 - 7.4 使用Apache Solr索引和搜索网络文档 180
 - 7.5 配置Apache HBase作为Apache Nutch的后端数据存储 182
 - 7.6 在Hadoop集群上部署Apache HBase 185
 - 7.7 使用Hadoop/HBase集群构建Apache Nutch全网爬虫服务 188
 - 7.8 用于索引和搜索的ElasticSearch 191
 - 7.9 生成抓取网页的内链图 193
- 第8章 聚类、推荐和关系发现 197
 - 8.1 简介 197
 - 8.2 基于内容的推荐 198
 - 8.3 层次聚类 204
 - 8.4 对亚马逊销售数据集进行聚类操作 208
 - 8.5 基于协同过滤的推荐 212
 - 8.6 使用朴素贝叶斯分类器的分类 216
 - 8.7 使用Adwords平衡算法给广告分配关键字 222
- 第9章 海量文本数据处理 231
 - 9.1 简介 231
 - 9.2 使用Hadoop Streaming和Python预处理数据（抽取、清洗和格式转换） 231
 - 9.3 使用Hadoop Streaming进行数据去重 235
 - 9.4 使用importtsv和批量加载工具把大型数据集加载到Apache HBase数据存储中 237
 - 9.5 创建用于文本数据的TF向量和TF-IDF向量 242
 - 9.6 聚类文本数据 246
 - 9.7 使用隐含狄利克雷分布（LDA）发现主题 249
 - 9.8 使用Mahout的朴素贝叶斯分类器分类文件 252
- 第10章 云端部署——在云上使用Hadoop 255
 - 10.1 简介 255
 - 10.2 使用亚马逊弹性MapReduce运行Hadoop MapReduce计算 256
 - 10.3 使用亚马逊EC2竞价实例来执行EMR作业流以节约开支 259
 - 10.4 使用EMR执行Pig脚本 261
 - 10.5 使用EMR执行Hive脚本 263
 - 10.6 使用命令行界面创建亚马逊EMR作业流 267
 - 10.7 使用EMR在亚马逊EC2云上部署Apache HBase集群 270
 - 10.8 使用EMR引导操作来配置亚马逊EMR作业的虚拟机 275
 - 10.9 使用Apache Whirr在云环境中部署Apache Hadoop集群 277
 - 10.10 使用Apache Whirr在云环境中部署Apache HBase集群 281

• • • • • [\(收起\)](#)

[Hadoop MapReduce实战手册_下载链接1](#)

标签

Hadoop

MapReduce

编程

程序设计

学习

千万别买这本书!!!

hadoop

Programming

评论

挺实在的，后面比较精彩。

自己当然要推荐一下自己编译的，打一个广告。

一本云计算领域值得推荐的好书，其理论联系实际，包括丰富案例。拟在本科生的云计算课中尝试使用。

这本书里的示例代码必须要翻墙，否则完全没办法下载，可没有示例代码你根本学不了
!!!!

[Hadoop MapReduce实战手册_下载链接1](#)

书评

书上的代码bin/hadoopjar hadoop-cookbook-chapter8.jar
chapter8.MostFrequentUserFinder/data/input1 /data/output1 但作者想表达的意思是
bin/hadoop(这里多出一个空格)jar hadoop-cookbook-chapter8.jar
chapter8.MostFrequentUserFinder(这里多出一个空格)/data/input1 /...

[Hadoop MapReduce实战手册_下载链接1](#)