《Hadoop实战（第2版）》

图书基本信息

书名：《Hadoop实战（第2版）》

13位ISBN编号：9787111395836

10位ISBN编号：7111395832

出版时间：2012-11

出版社：机械工业出版社华章公司

作者：陆嘉恒

页数：498

《Hadoop实战（第2版）》

前言

为什么写这本书计算技术已经改变了我们的工作、学习和生活。分布式的云计算技术是当下IT领域最热门的话题之一，它通过整合资源，为降低成本和能源消耗提供了一种简化、集中的计算平台。这种低成本、高扩展、高性能的特点促使其迅速发展，遍地开发，悄然改变着整个行业的面貌。社会各界对云计算的广泛研究和应用无疑证明了这一点：在学术界，政府和很多高校十分重视对云计算技术的研究和投入；在产业界，各大IT公司也在研究和开发相关的云计算产品上投入了大量的资源。这些研究和应用推动与云计算相关的新兴技术和产品不断涌现，传统的信息服务产品向云计算模式转型。Hadoop作为Apache基金会的开源项目，是云计算研究和应用最具代表性的产品。Hadoop分布式框架为开发者提供了一个分布式系统的基础架构，用户可以在不了解分布式系统底层细节的情况下开发分布式的应用，充分利用由Hadoop统一起来的集群存储资源、网络资源和计算资源，实现基于海量数据的高速运算和存储。在编写本书第一版时，鉴于Hadoop技术本身和应用环境较为复杂，入门和实践难度较大，而关于Hadoop的参考资料又非常少，笔者根据自己的实际研究和使用经历，理论与实践并重，从基础出发，为读者全面呈现了Hadoop的相关知识，旨在为Hadoop学习者提供一本工具书。但是时至今日，Hadoop的版本已从本书第一版介绍的0.20升级至正式版1.0，读者的需求也从入门发展到更加深入地了解Hadoop的实现细节，了解Hadoop的更新和发展的趋势，了解Hadoop在企业中的应用。虽然本书第一版受到广大Hadoop学习者的欢迎，但是为了保持对最新版Hadoop的支持，进一步满足读者的需求，继续推动Hadoop技术在国内的普及和发展，笔者不惜时间和精力，搜集资料，亲自实践，编写了本书第二版。第2版与第1版的区别基于Hadoop 1.0版本和相关项目的最新版，本书在第1版的基础上进行了更新和调整：每章都增加了新内容（如第1章增加了与Hadoop安全相关的知识，第2增加了在Max OS X系统上安装Hadoop的介绍，第9章增加了WebHDFS等）；部分章节深入剖析了Hadoop源码；增加了对Hadoop接口及实践方面的介绍（附录C和附录D）；增加了对下一代MapReduce的介绍（第8章）；将企业应用介绍移到本书最后并更新了内容（第19章）；增加了对Hadoop安装和代码执行的集中介绍（附录B）。本书面向的读者在编写本书时，笔者力图使不同背景、职业和层次的读者都能从这本书中获益。如果你是专业技术人员，本书将带领你深入云计算的世界，全面掌握Hadoop及其相关技术细节，帮助你使用Hadoop技术解决当前面临的问题。如果你是系统架构人员，本书将成为你搭建Hadoop集群、管理集群，并迅速定位和解决问题的工具书。如果你是高等院校计算机及相关专业的学生，本书将为你在课堂之外了解最新的IT技术打开了一扇窗户，帮助你拓宽视野，完善知识结构，为迎接未来的挑战做好知识储备。在学习本书之前，大家应该具有如下的基础：要有一定的分布式系统的基础知识，对文件系统的基本操作有一定的了解。要有一定的Linux操作系统的基础知识。有较好的编程基础和阅读代码的能力，尤其是要能够熟练使用Java语言。对数据库、数据仓库、系统监控，以及网络爬虫等知识最好也能有一些了解。如何阅读本书从整体内容上讲，本书包括19章和4个附录。前10章、第18章、第19章和4个附录主要介绍了Hadoop背景知识、Hadoop集群安装和代码执行、MapReduce机制及编程知识、HDFS实现细节及管理知识、Hadoop应用。第11章至第17章结合最新版本详细介绍了与Hadoop相关的其他项目，分别为Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa，以备读者扩展知识面之用。在阅读本书时，笔者建议大家先系统地学习Hadoop部分的理论知识（第1章、第3章、第6章至第10章），这样可对Hadoop的核心内容和实现机制有一个很好的理解。在此基础上，读者可进一步学习Hadoop部分的实践知识（第2章、第4章、第5章、第18章、第19章和4个附录），尝试搭建自己的Hadoop集群，编写并运行自己的MapReduce代码。对于本书中关于Hadoop相关项目的介绍，大家可以有选择地学习。在内容的编排上，各章的知识点是相对独立的，是并行的关系，因此大家可以有选择地进行学习。当然，如果时间允许，还是建议大家系统地学习全书的内容，这样能够对Hadoop系统的机制有一个完整而系统的理解，为今后深入地研究和实践Hadoop及云计算技术打下坚实的基础。另外，笔者希望大家在学习本书时能一边阅读，一边根据书中的指导动手实践，亲自实践本书中所给出的编程范例。例如，先搭建一个自己的云平台，如果条件受限，可以选择伪分布的方式。致谢在本书的编写过程中，很多Hadoop方面的实践者和研究者做了大量的工作，他们是冯博亮、程明、徐文韬、张林林、朱俊良、许翔、陈东伟、谭果、林春彬等，在此表示感谢。陆嘉恒2012年6月于北京

《Hadoop实战（第2版）》

内容概要

本书能满足读者全面学习最新的Hadoop技术及其相关技术（Hive、HBase等）的需求，是一本系统且极具实践指导意义的Hadoop工具书和参考书。第1版上市后广受好评，被誉为学习Hadoop技术的经典著作之一。与第1版相比，第2版技术更新颖，所有技术都针对最新版进行了更新；内容更全面，几乎每一个章节都增加了新内容，而且增加了新的章节；实战性更强，案例更丰富；细节更完美，对第1版中存在的缺陷和不足进行了修正。
本书内容全面，对Hadoop整个技术体系进行了全面的讲解，不仅包括HDFS、MapReduce、YARN等核心内容，而且还包括Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等与Hadoop技术相关的重要内容。实战性强，不仅为各个知识点精心设计了大量经典的小案例，而且还包括Yahoo!等多个大公司的企业级案例，可操作系极强。
全书一共19章：第1~2章首先对Hadoop进行了全方位的宏观介绍，然后介绍了Hadoop在三大主流操作系统平台上的安装与配置方法；第3~6章分别详细讲解了MapReduce计算模型、MapReduce的工作机制、MapReduce应用的开发方法，以及多个精巧的MapReduce应用案例；第7章全面讲解了Hadoop的I/O操作；第8章对YARN进行了介绍；第9章对HDFS进行了详细讲解和分析；第10章细致地讲解了Hadoop的管理；第11~17章对Hadoop大生态系统中的Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等技术进行了详细的讲解；第18章讲解了Hadoop的各种常用插件，以及Hadoop插件的开发方法；第19章分析了Hadoop在Yahoo!、eBay、百度、Facebook等企业中的应用案例。

《Hadoop实战（第2版）》

作者简介

陆嘉恒，资深数据库专家和云计算技术专家，对Hadoop及其相关技术有非常深入的研究，主持了多个分布式云计算项目的研究与实施，积累了丰富的实践经验。获得新加坡国立大学博士学位，美国加利福尼亚大学尔湾分校(University of California, Irvine) 博士后，现为中国人民大学教授，博士生导师。此外，他对数据挖掘和Web信息搜索等技术也有深刻的认识。

《Hadoop实战（第2版）》

书籍目录

目录
前言
第1章　Hadoop简介/1
1.1　什么是Hadoop/2
1.1.1　Hadoop概述/2
1.1.2　Hadoop的历史/2
1.1.3　Hadoop的功能与作用/2
1.1.4　Hadoop的优势/3
1.1.5　Hadoop应用现状和发展趋势/3
1.2　Hadoop项目及其结构/3
1.3　Hadoop体系结构/6
1.4　Hadoop与分布式开发/7
1.5　Hadoop计算模型—MapReduce/10
1.6　Hadoop数据管理/10
1.6.1　HDFS的数据管理/10
1.6.2　HBase的数据管理/12
1.6.3　Hive的数据管理/13
1.7　Hadoop集群安全策略/15
1.8　本章小结/17
第2章　Hadoop的安装与配置/19
2.1　在Linux上安装与配置Hadoop/20
2.1.1　安装JDK 1.6/20
2.1.2　配置SSH免密码登录/21
2.1.3　安装并运行Hadoop/22
2.2　在Mac OSX上安装与配置Hadoop/24
2.2.1　安装Homebrew/24
2.2.2　使用Homebrew安装Hadoop/25
2.2.3　配置SSH和使用Hadoop/25
2.3　在Windows上安装与配置Hadoop/25
2.3.1　安装JDK 1.6或更高版本/25
2.3.2　安装Cygwin/25
2.3.3　配置环境变量/26
2.3.4　安装sshd服务/26
2.3.5　启动sshd服务/26
2.3.6　配置SSH免密码登录/26
2.3.7　安装并运行Hadoop/26
2.4　安装和配置Hadoop集群/27
2.4.1　网络拓扑/27
2.4.2　定义集群拓扑/27
2.4.3　建立和安装Cluster /28
2.5　日志分析及几个小技巧/34
2.6　本章小结/35
第3章　MapReduce计算模型/36
3.1　为什么要用MapReduce/37
3.2　MapReduce计算模型/38
3.2.1　MapReduce Job/38
3.2.2　Hadoop中的Hello World程序/38

3.2.3　MapReduce的数据流和控制流/46
3.3　MapReduce任务的优化/47
3.4　Hadoop流/49
3.4.1　Hadoop流的工作原理/50
3.4.2　Hadoop流的命令/51
3.4.3　两个例子/52
3.5　Hadoop Pipes/54
3.6　本章小结/56
第4章　开发MapReduce应用程序/57
4.1　系统参数的配置/58
4.2　配置开发环境/60
4.3　编写MapReduce程序/60
4.3.1　Map处理/60
4.3.2　Reduce处理/61
4.4　本地测试/62
4.5　运行MapReduce程序/62
4.5.1　打包/64
4.5.2　在本地模式下运行/64
4.5.3　在集群上运行/64
4.6　网络用户界面/65
4.6.1　JobTracker页面/65
4.6.2　工作页面/65
4.6.3　返回结果/66
4.6.4　任务页面/67
4.6.5　任务细节页面/67
4.7　性能调优/68
4.7.1　输入采用大文件/68
4.7.2　压缩文件/68
4.7.3　过滤数据/69
4.7.4　修改作业属性/71
4.8　MapReduce工作流/72
4.8.1　复杂的Map和Reduce函数/72
4.8.2　MapReduce Job中全局共享数据/74
4.8.3　链接MapReduce Job/75
4.9　本章小结/77
第5章　MapReduce应用案例/79
5.1　单词计数/80
5.1.1　实例描述/80
5.1.2　设计思路/80
5.1.3　程序代码/81
5.1.4　代码解读/82
5.1.5　程序执行/83
5.1.6　代码结果/83
5.1.7　代码数据流/84
5.2　数据去重/85
5.2.1　实例描述/85
5.2.2　设计思路/86
5.2.3　程序代码/86
5.3　排序/87

5.3.1　实例描述/87
5.3.2　设计思路/88
5.3.3　程序代码/89
5.4　单表关联/91
5.4.1　实例描述/91
5.4.2　设计思路/92
5.4.3　程序代码/92
5.5　多表关联/95
5.5.1　实例描述/95
5.5.2　设计思路/96
5.5.3　程序代码/96
5.6　本章小结/98
第6章　MapReduce工作机制/99
6.1　MapReduce作业的执行流程/100
6.1.1　MapReduce任务执行总流程/100
6.1.2　提交作业/101
6.1.3　初始化作业/103
6.1.4　分配任务/104
6.1.5　执行任务/106
6.1.6　更新任务执行进度和状态/107
6.1.7　完成作业/108
6.2　错误处理机制 /108
6.2.1　硬件故障/109
6.2.2　任务失败/109
6.3　作业调度机制/110
6.4　Shuffle和排序/111
6.4.1　Map端/111
6.4.2　Reduce端/113
6.4.3　shuffle过程的优化/114
6.5　任务执行/114
6.5.1　推测式执行/114
6.5.2　任务JVM重用/115
6.5.3　跳过坏记录/115
6.5.4　任务执行环境/116
6.6　本章小结/117
第7章　Hadoop I/O操作/118
7.1　I/O操作中的数据检查/119
7.2　数据的压缩 /126
7.2.1　Hadoop对压缩工具的选择/126
7.2.2　压缩分割和输入分割/127
7.2.3　在MapReduce程序中使用压缩/127
7.3　数据的I/O中序列化操作/128
7.3.1　Writable类/128
7.3.2　实现自己的Hadoop数据类型/137
7.4　针对Mapreduce的文件类/139
7.4.1　SequenceFile类/139
7.4.2　MapFile类/144
7.4.3　ArrayFile、SetFile和BloomMapFile/146
7.5　本章小结/148

第8章　下一代MapReduce：YARN/149
8.1　MapReduce V2设计需求/150
8.2　MapReduce V2主要思想和架构/151
8.3　MapReduce V2设计细节/153
8.4　MapReduce V2优势/156
8.5　本章小结/156
第9章　HDFS详解/157
9.1　Hadoop的文件系统/158
9.2　HDFS简介/160
9.3　HDFS体系结构/161
9.3.1　HDFS的相关概念/161
9.3.2　HDFS的体系结构/162
9.4　HDFS的基本操作/164
9.4.1　HDFS的命令行操作/164
9.4.2　HDFS的Web界面/165
9.5　HDFS常用Java API详解/166
9.5.1　使用Hadoop URL读取数据/166
9.5.2　使用FileSystem API读取数据/167
9.5.3　创建目录/169
9.5.4　写数据/169
9.5.5　删除数据/171
9.5.6　文件系统查询/171
9.6　HDFS中的读写数据流/175
9.6.1　文件的读取/175
9.6.2　文件的写入/176
9.6.3　一致性模型/178
9.7　HDFS命令详解/179
9.7.1　通过distcp进行并行复制/179
9.7.2　HDFS的平衡/180
9.7.3　使用Hadoop归档文件/180
9.7.4　其他命令/183
9.8　WebHDFS/186
9.8.1　WebHDFS的配置/186
9.8.2　WebHDFS命令/186
9.9　本章小结/190
第10章　Hadoop的管理/191
10.1　HDFS文件结构/192
10.2　Hadoop的状态监视和管理工具/196
10.2.1　审计日志/196
10.2.2　监控日志/196
10.2.3　Metrics/197
10.2.4　Java管理扩展 /199
10.2.5　Ganglia/200
10.2.6　Hadoop管理命令/202
10.3　Hadoop集群的维护/206
10.3.1　安全模式/206
10.3.2　Hadoop的备份/207
10.3.3　Hadoop的节点管理/208
10.3.4　系统升级/210

10.4　本章小结/212
第11章　Hive详解/213
11.1　Hive简介/214
11.1.1　Hive的数据存储/214
11.1.2　Hive的元数据存储/216
11.2　Hive的基本操作/216
11.2.1　在集群上安装Hive/216
11.2.2　配置MySQL存储Hive元数据/218
11.2.3　配置Hive/220
11.3　Hive QL详解/221
11.3.1　数据定义（DDL）操作/221
11.3.2　数据操作（DML）/231
11.3.3　SQL操作/233
11.3.4　Hive QL使用实例/235
11.4　Hive网络（Web UI）接口/237
11.4.1　Hive网络接口配置/237
11.4.2　Hive网络接口操作实例/238
11.5　Hive的JDBC接口//241
11.5.1　Eclipse环境配置/241
11.5.2　程序实例/241
11.6　Hive的优化/244
11.7　本章小结/246
第12章　HBase详解/247
12.1　HBase简介/248
12.2　HBase的基本操作/249
12.2.1　HBase的安装/249
12.2.2　运行HBase /253
12.2.3　HBase Shell/255
12.2.4　HBase配置/258
12.3　HBase体系结构/260
12.3.1　HRegion/260
12.3.2　HRegion服务器/261
12.3.3　HBase Master服务器/262
12.3.4　ROOT表和META表/262
12.3.5　ZooKeeper/263
12.4　HBase数据模型/263
12.4.1　数据模型/263
12.4.2　概念视图/264
12.4.3　物理视图/264
12.5　HBase与RDBMS/265
12.6　HBase与HDFS/266
12.7　HBase客户端/266
12.8　Java API /267
12.9　HBase编程 /273
12.9.1　使用Eclipse开发HBase应用程序/273
12.9.2　HBase编程/275
12.9.3　HBase与MapReduce/278
12.10　模式设计/280
12.10.1　模式设计应遵循的原则/280

12.10.2　学生表/281
12.10.3　事件表/282
12.11　本章小结/283
第13章　Mahout详解/284
13.1　Mahout简介/285
13.2　Mahout的安装和配置/285
13.3　Mahout API简介/288
13.4　Mahout中的频繁模式挖掘/290
13.4.1　什么是频繁模式挖掘/290
13.4.2　Mahout中的频繁模式挖掘/290
13.5　Mahout中的聚类和分类/292
13.5.1　什么是聚类和分类/292
13.5.2　Mahout中的数据表示/293
13.5.3　将文本转化成向量/294
13.5.4　Mahout中的聚类、分类算法/295
13.5.5　算法应用实例/299
13.6　Mahout应用：建立一个推荐引擎/304
13.6.1　推荐引擎简介/304
13.6.2　使用Taste构建一个简单的推荐引擎/305
13.6.3　简单分布式系统下基于产品的推荐系统简介/307
13.7　本章小结/309
第14章　Pig详解/310
14.1　Pig简介/311
14.2　Pig的安装和配置 /311
14.2.1　Pig的安装条件/311
14.2.2　Pig的下载、安装和配置/312
14.2.3　Pig运行模式/313
14.3　Pig Latin语言/315
14.3.1　Pig Latin语言简介/315
14.3.2　Pig Latin的使用/316
14.3.3　Pig Latin的数据类型/318
14.3.4　Pig Latin关键字/319
14.4　用户定义函数 /323
14.4.1　编写用户定义函数/324
14.4.2　使用用户定义函数/325
14.5　Zebra简介 /326
14.5.1　Zebra的安装/326
14.5.2　Zebra的使用简介/327
14.6　Pig实例 /328
14.6.1　Local模式/328
14.6.2　MapReduce模式/330
14.7　Pig进阶/331
14.7.1　数据实例/331
14.7.2　Pig数据分析/332
14.8　本章小结/336
第15章　ZooKeeper详解/337
15.1　ZooKeeper简介/338
15.1.1　ZooKeeper的设计目标/338
15.1.2　数据模型和层次命名空间/339

15.1.3　ZooKeeper中的节点和临时节点/339
15.1.4　ZooKeeper的应用/340
15.2　ZooKeeper的安装和配置/340
15.2.1　安装ZooKeeper /340
15.2.2　配置ZooKeeper/346
15.2.3　运行ZooKeeper/348
15.3　ZooKeeper的简单操作/350
15.3.1　使用ZooKeeper命令的简单操作步骤/350
15.3.2　ZooKeeper API的简单使用/352
15.4　ZooKeeper的特性/355
15.4.1　ZooKeeper的数据模型/355
15.4.2　ZooKeeper会话及状态/356
15.4.3　ZooKeeper watches/357
15.4.4　ZooKeeper ACL/358
15.4.5　ZooKeeper的一致性保证/359
15.5　使用ZooKeeper进行Leader选举/359
15.6　ZooKeeper锁服务/360
15.6.1　ZooKeeper中的锁机制/360
15.6.2　ZooKeeper提供的一个写锁的实现/361
15.7　使用ZooKeeper创建应用程序 /363
15.7.1　使用Eclipse开发ZooKeeper应用程序/363
15.7.2　应用程序实例/365
15.8　BooKeeper/369
15.9　本章小结/371
第16章　Avro详解/372
16.1　Avro介绍/373
16.1.1　模式声明/374
16.1.2　数据序列化/378
16.1.3　数据排列顺序/380
16.1.4　对象容器文件 /381
16.1.5　协议声明/382
16.1.6　协议传输格式/383
16.1.7　模式解析/386
16.2　Avro的C/C++实现/387
16.3　Avro的Java实现/398
16.4　GenAvro（Avro IDL）语言/402
16.5　Avro SASL概述/406
16.6　本章小结/407
第17章　Chukwa详解/409
17.1　Chukwa简介/410
17.2　Chukwa架构/411
17.2.1　客户端及其数据模型/412
17.2.2　收集器/413
17.2.3　归档器和分离解析器/414
17.2.4　HICC/415
17.3　Chukwa的可靠性/415
17.4　Chukwa集群搭建/416
17.4.1　基本配置要求/416
17.4.2　Chukwa的安装/416

17.4.3　Chukwa的运行/419
17.5　Chukwa数据流的处理/424
17.6　Chukwa与其他监控系统比较/425
17.7　本章小结/426
本章参考资料/426
第18章　Hadoop的常用插件与开发/428
18.1　Hadoop Studio的介绍和使用/429
18.1.1　Hadoop Studio的介绍/429
18.1.2　Hadoop Studio的安装配置/430
18.1.3　Hadoop Studio的使用举例/430
18.2　Hadoop Eclipse的介绍和使用/436
18.2.1　Hadoop Eclipse的介绍/436
18.2.2　Hadoop Eclipse的安装配置/437
18.2.3　Hadoop Eclipse的使用举例/438
18.3　Hadoop Streaming的介绍和使用/440
18.3.1　Hadoop Streaming的介绍/440
18.3.2　Hadoop Streaming的使用举例/444
18.3.3　使用Hadoop Streaming常见的问题/446
18.4　Hadoop Libhdfs的介绍和使用/448
18.4.1　Hadoop Libhdfs的介绍/448
18.4.2　Hadoop Libhdfs的安装配置/448
18.4.3　Hadoop Libhdfs API简介/448
18.4.4　Hadoop Libhdfs的使用举例/449
18.5　本章小结/450
第19章　企业应用实例/452
19.1　Hadoop在Yahoo!的应用/453
19.2　Hadoop在eBay的应用/455
19.3　Hadoop在百度的应用/457
19.4　即刻搜索中的Hadoop/460
19.4.1　即刻搜索简介/460
19.4.2　即刻Hadoop应用架构/460
19.4.3　即刻Hadoop应用分析/463
19.5　Facebook中的Hadoop和HBase/463
19.5.1　Facebook中的任务特点/464
19.5.2　MySQL VS Hadoop+HBase/466
19.5.3　Hadoop和HBase的实现/467
19.6　本章小结/472
本章参考资料/472
附录A　云计算在线检测平台/474
附录B　Hadoop安装、运行与使用说明/484
附录C　使用DistributedCache的MapReduce程序/491
附录D　使用ChainMapper和ChainReducer的MapReduce程序/495

《Hadoop实战（第2版）》

章节摘录

第1章Hadoop简介本章内容什么是HadoopHadoop项目及其结构Hadoop体系结构Hadoop与分布式开发Hadoop计算模型—MapReduceHadoop数据管理Hadoop集群安全策略本章小结1.1 什么是Hadoop1.1.1 Hadoop概述Hadoop是Apache软件基金会旗下的一个开源分布式计算平台。以Hadoop分布式文件系统（Hadoop Distributed File System，HDFS）和MapReduce（Google MapReduce的开源实现）为核心的Hadoop为用户提供了系统底层细节透明的分布式基础架构。HDFS的高容错性、高伸缩性等优点允许用户将Hadoop部署在低廉的硬件上，形成分布式系统；MapReduce分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序。所以用户可以利用Hadoop轻松地组织计算机资源，从而搭建自己的分布式计算平台，并且可以充分利用集群的计算和存储能力，完成海量数据的处理。经过业界和学术界长达10年的锤炼，目前的Hadoop 1.0.1已经趋于完善，在实际的数据处理和分析任务中担当着不可替代的角色。

《Hadoop实战（第2版）》

媒体关注与评论

经过学术界和业界近10年的努力，Hadoop技术已经趋于完善而且应用广泛，几乎已经成为Big Data领域的事实标准。Hadoop技术本身比较复杂，而且还涉及Pig、ZooKeeper、Hive、HBase等一系列技术，学习门槛比较高，对于初学者和基础不太扎实的读者而言，有一本适合系统学习的Hadoop图书显得十分重要。本书即是专门为这两类读者量身定做的：第一，它的内容非常全面和前沿，不仅讲解了最新的Hadoop技术和第二代MapReduce，还讲解了涉及的所有周边技术，能满足系统学习的需求；第二，实战性非常强，不仅很多知识点配有精心设计的小案例，而且有完整的企业级案例，能满足操作实践的需求；第三，这一版在上一版的基础上根据最新的技术做了更新和补充，能满足读者学习最新技术的需求。本书第1版不仅取得了好的销量，而且广受好评，第2版在内容上有很大的提升，相信能让更多的读者从中受益。——EasyHadoop 国内专业的Hadooop社区，致力于让Hadoop大数据分析更简单

《Hadoop实战（第2版）》

编辑推荐

《Hadoop实战(第2版)》编辑推荐：第1版广受好评，第2版基于Hadoop及其相关技术最新版本撰写，从多角度做了全面的修订和补充。不仅详细讲解了新一代的Hadoop技术，而且全面介绍了Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等重要技术，是系统学习Hadoop技术的首选之作。

《Hadoop实战（第2版）》

精彩短评

1、大数据量应用实战，很值得看的一本书。
2、当当就是送货速度快！书还行，特别新
3、很全面的讲解，案头必备
4、作为cookbook已经足够了，离高手还有很远一段距离。
5、了解 Hadoop，HDFS，HBase，Hive基本原理和应用。
6、买来及看了一下这本书，内容比较全面，涵盖了hadoop下mapeduce和hdfs两个框架，既有原理的介绍，又有开发实践的介绍，还配套一个检测平台。整体来说比较值，能学到不少东西。
7、书很好，很基础，适合基础学者学习
8、这本书是进行Hadoop学习的不二之选，让我们可以从一个初学者逐步深入。
9、作为入门书很不错，即有高层的架构分析，也可以了解底层的代码实现，对一线开发很有帮助
10、新书还行，没有想象的那么深，得跟hadoop权威指南搭配起来看挺不错的，感觉书里面旧的api用法可以扔掉了，稳定版1.0.4已经发布后没必要讲解0.2版本的api了
11、不值这个价，讲得也太泛或太细，看不到核心
12、淘宝阅读下的。没注意，以为是之前另外那本。花了一个小时把优化和例子看完，最后的企业应用也翻阅了下。作为实战来说，内容深度还是欠缺。
13、Hadoop实战(第2版)很不错，全面，实战强。不过总的适合新手入门。
14、热门技术，给自己充电的，唉，还得看下JAVA充下电~~
15、Hadoop入门书，把写法、运行逻辑、管理、各种配套软件全覆盖，更深入的倒也没有了。2012年的书内容有些旧了。（2012.11.12京东预购）
16、一般，还是比较旧，没有主讲yarn。
17、当当网发货速度很快，这本书是老师推荐的，我当时看了看目录，果断买下
18、前半部分还不错，后面东西比较杂乱了。好久没读技术书籍了。
19、理论与实际结合，很不错的书，实战意味很强
20、还以为是Manning书的译本，其实是国内教授自己写的。确实有不一样的东西，但是真的质量很差。举个例子：单表join，一整个示例（二度关系）里边居然就没出现这个词。有我小时候还不懂英语，被国产技术书坑的感觉。
21、内容不错。书的后面十页压皱了，有点美中不足。
22、涉及的内容比较多，但都是点到为止，看了此书能多这个生态圈有一个基本认。
23、初学者角度来看还是很不错，很系统的。对大数据各个系统框架有了一个全面的认识
24、讲的面很广 但是都不细致 好像是在hbase 那部分知识讲的比较老了 不是很与时俱进 表示不太好
25、实用，印刷质量也很好。
26、还不错，内容还没看，回头再评价吧～～
27、非常好非常好的入门书
28、纸张质量不错，书的内容也不错。
29、能够指导进行相关开发。
30、还没看，质量还可以，看过了之后再来评论。
31、挺好的，物超所值，系统全面
32、畅销书全新升级、技术更新、内容更全、实战性更强、细节更完善！
33、别字很多，东拼西凑的。
34、内容比较的丰富，讲了很多的新技术，讲了hadoop及下面应用的技术很丰富，很有价值的
35、5折买的，正好用到
36、感觉很像hadoop权威指南, 不过毕竟是中国人写的, 所以读起来比翻译的权威指南更顺畅一些; 和权威指南各有优劣.
37、印刷还行，看看内容再说
38、商品质量不错，内容很充实
39、比较经典，适合入门学习
40、很适合初学者。

41、书的质量很好～值得读
42、本书相当实用
43、收到书了看到第二章，感觉还不错，入门级的。
44、正在看，大数据云时代必备啊
45、经过学术界和业界近10年的努力，Hadoop技术已经趋于完善而且应用广泛，几乎已经成为BigData领域的事实标准。Hadoop技术本身比较复杂，而且还涉及Pig、ZooKeeper、Hive、HBase等一系列技术，学习门槛比较高，对于初学者和基础不太扎实的读者而言，有一本适合系统学习的Hadoop图书显得十分重要。本书即是专门为这两类读者量身定做的：第一，它的内容非常全面和前沿，不仅讲解了最新的Hadoop技术和第二代MapReduce，还讲解了涉及的所有周边技术，能满足系统学习的需求；第二，实战性非常强，不仅很多知识点配有精心设计的小案例，而且有完整的企业级案例，能满足操作实践的需求；第三，这一版在上一版的基础上根据最新的技术做了更新和补充，能满足读者学习最新技术的需求。本书第1版不仅取得了好的销量，而且广受好评，第2版在内容上有很大的提升，相信能让更多的读者从中受益。
46、讲得通俗性易懂，很适合学习。
47、书很好，快递也很快！~
48、大数据处理是发展趋势，值得学习
49、不错 适合弄hadoop的人
50、很一般入门书，很多东西都没讲清楚，随便翻翻可以，根本不值这个价，快速浏览一遍就没有用了
51、Hadoop入门的一本书，还可以
52、体系讲解全面，不过对我这种只用写hive代码的屌丝来说有些不明觉厉
53、很好，性价比很高。实用性很强。
54、内容很好。好好学习
55、书不错，配合权威指南学习，作为学习hadoop的入门书籍。
56、知识细节描述较全面，实践性较强
57、初学还行吧，凑活看，有些地方确实不到位
58、经典教程之一。 里面的几个MAPREDUCE程序挺好。 同时比较注重实战。 还有对于第二代架构YARN的简介。 觉的学习HADOOP的必读教程之一
59、做过，易懂，慢慢读。
60、书送的挺快 质量也不错 内容 还在看，看目录 不错。国人 自己写的，先支持一下。
61、预售的书，但是很快就到货了。正在看，内容很经典。很难得机械工业的纸张这么好。
62、拿来救急！
63、用这本书基本入门没问题
64、推荐，在学习中
65、还不错，别人推荐的，当当便宜一些
66、实用性很强,可以跟着做实验
67、撒地方的萨芬撒的飞洒地方
68、书的质量不错，内容还没开始看
69、内容还可以，只是有一些过时了。不是太深入浅出型。
70、还不错 应该是正版
71、在研究云计算 非常好的一本书
72、从最基础的介绍到实践都有涉及，是一本不错的学习hadoop的书
73、匆匆扫了一遍，对hadoop以及hadoop配套的相关大数据分析、挖掘、处理的工具有了一个初步的了解。
74、是很新，质量不错，正在阅读中
75、很实用的书 简单易懂 深入浅出
76、太冗长了，而且太浅了，讲的完全是百度一下就能知道的，本来看书就是想看些系统性的原理性的东西，很失望
77、被同学极力推荐的，确实不错。

78、Hadoop实战（第2版） 实战技巧学习用书
79、入门书，比起hadoop权威指南内容要新一点，简要介绍了yarn和mahout等
80、很好，通过项目来引出知识点，就喜欢这样的书。好评
81、这本书讲的面比较广，虽然比不上那本权威指南，但也是不错的书。
82、还可以！新东西学习学习！
83、内容详尽 推荐
84、hadoop数据仓库
85、这本书不错，一天就到，物流给力！一个很严重的缺陷是：在运输过程中把我的书皮弄破了，不影响阅读就不换了~当当在包装书太简陋了，就一个塑料袋加塑料膜~
86、啥都简单介绍到了，命令也很详细，用做入门手册不错。这么快就出第二版似乎有点不厚道。
87、作为Hadoop入门的不错中文参考
88、国内Hadoop这个方向能看的书不多，这是一本好书！
89、非常不错一本书，例子给的很好
90、入门的好书
91、高房价蝴蝶结今年发动机的加发动机的国家的风景
92、入门书
93、开启分布式实战。
94、专业性很强，老公很喜欢
95、范围广而泛
96、朋友推荐的，很实用，还没有看，书的质量很好
97、还没细看，随便翻了翻，看起来不错
98、这本书第一版读过，内容通俗易懂，还配套检测系统。第二版内容更加丰富全面，深度上也有所提高。值得入手阅读
99、国人写的书，适合按步骤一步步地敲命令。只看了HDFS、HIVE、Zookeeper等工作中用到的部分
100、手边的参考书
101、希望我用心学习，从中学到东西

《Hadoop实战（第2版）》

精彩书评

1、本书虽然讲了很多Hadoop的框架，但是都讲得不够透彻，有的地方还有一些错误。在一些文件配置方面，作者给的建议是修改-default.xml,但是连源码中、配置文件的注释中都是不推荐修改-default.xml的文件，而是在另一个配置文件中添加该属性。在涉及一些原理方面也没有讲，这对于以后从事Hadoop行业的人来说，比较痛苦，因为出错了，你都不知道该改哪个地方。总体的感觉，就是这本书很臃肿，对于有Hadoop基础的人来说，我还是建议看看Hadoop权威指南，董西成的Hadoop技术内幕系列丛书以及官方文档。根据源码和书本来看。如果感觉还不行的话，请看外文书籍。
2、sdfgxgd楼去我lz我cry我了那是小JJ9429477up路我会怕lz婆婆你要求是YY来咯拿去心哦哦苏州哦TMD兔子XP马虎x5哦dry五orz呀啊
3、这本书是进行Hadoop学习的不二之选，让我们可以从一个初学者逐步深入。他也适合有一定基础的用户加深进步了解。随书附有的Map-Reduce在线测试平台，给了没有条件搭建一个分布式环境的用户运行代码的一个很好的平台。可见作者的用心之处。
4、这么书确实写得不怎么样，别看他那么厚，内容好像很丰富，但其实很多都是没必要的，罗罗嗦嗦一大堆，内容提炼提炼就那么点。看着看着就越来越像我的硕士论文那样，为了凑字得写很多，要四五十页，但如果发到期刊上，只需要两三页纸就能讲完。

《Hadoop实战（第2版）》

版权说明