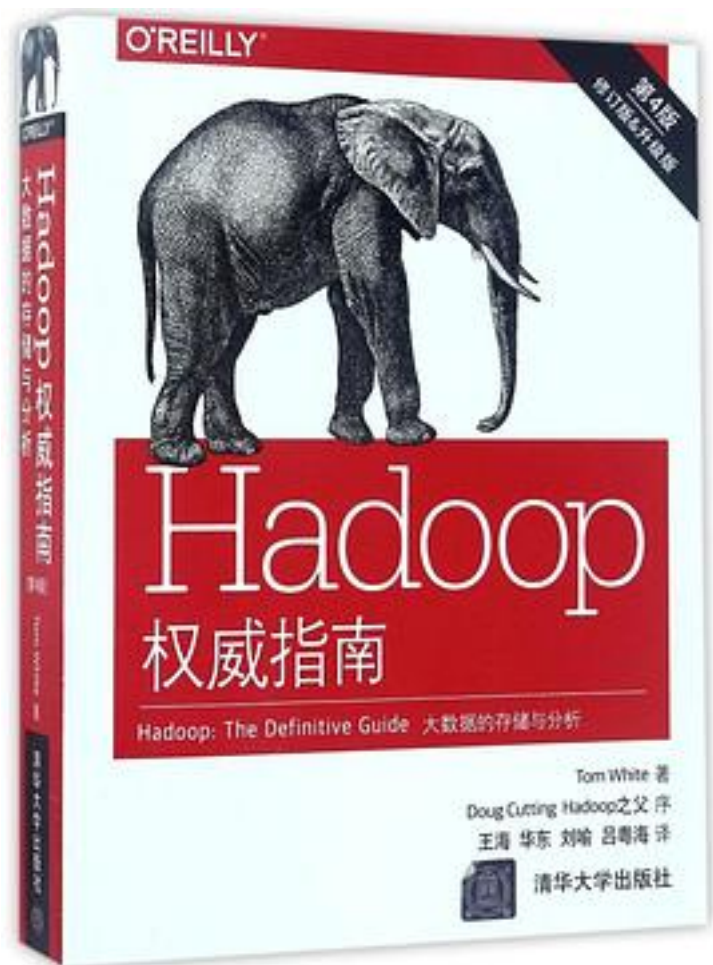


Hadoop权威指南:大数据的存储与分析(第4版)(修订版)(升级版)



[Hadoop权威指南:大数据的存储与分析\(第4版\)\(修订版\)\(升级版\) 下载链接1](#)

著者:Tom White

出版者:清华大学出版社

出版时间:2017-7

装帧:平装

isbn:9787302465133

本书结合理论和实践，由浅入深，全方位介绍了Hadoop

这一高性能的海量数据处理和分析平台。全书5部分24章，第Ⅰ部分介绍Hadoop基础知识，第Ⅱ部分介绍MapReduce,第Ⅲ部分介绍Hadoop的运维，第Ⅳ部分介绍Hadoop相关开源项目，第Ⅴ部分提供了三个案例，分别来自医疗卫生信息技术服务商塞纳(Cerner)、微软的人工智能项目ADAM(一种大规模分布式深度学习框架)和开源项目Cascading(一个新的针对MapReduce的数据处理API)。本书是一本专业、全面的Hadoop参考书和工具书，阐述了Hadoop生态圈的新发展和应用，程序员可以从中探索海量数据集的存储和分析，管理员可以从中了解Hadoop集群的安装和运维。

作者介绍:

Tom White是最杰出的Hadoop专家之一。自2007年2月以来，Tom White一直是Apache Hadoop的提交者(committer)，也是Apache软件基金会的成员。Tom是Cloudera的软件工程师，他是Cloudera的首批员工，对Apache和Cloudera做出了举足轻重的贡献。在此之前，他是一名独立的Hadoop顾问，帮助公司搭建、使用和扩展Hadoop。他是很多行业大会的专题演讲人，比如ApacheCon、OSCON和Strata。Tom在英国剑桥大学获得数学学士学位，在利兹大学获得科学哲学硕士学位。他目前与家人居住在威尔士。

译者简介

王海博士，解放军理工大学通信工程学院教授，博导，教研中心主任，长期从事无线自组网网络的设计与研发工作，主持国家自然科学基金、国家863计划课题等多项国家级课题，近5年获军队科技进步二等奖1项，三等奖6项，作为第1发明人申请国家发明专利十余项，发表学术论文50余篇。

华东博士，现任南京医科大学计算机教研室教师，一直致力于计算机辅助教学的相关技术研究，陆续开发了人体解剖学网络自主学习考试平台、诊断学自主学习平台和面向执业医师考试的预约化考试平台等系统，并在各个学科得到广泛的使用，获得全国高等学校计算机课件评比一等奖和三等奖各一项。主编、副主编教材两部，获发明专利一项、软件著作权多项。

刘喻博士，长期从事软件开发、软件测试和软件工程化管理工作，目前任教于清华大学软件所。

吕粤海，长期从事军事通信网络技术与软件开发工作，先后通过华为光网络高级工程师认证、思科网络工程师认证。

目录: 第Ⅰ部分 Hadoop基础知识

第1章 初识Hadoop 3

1.1 数据！数据！ 3

1.2 数据的存储与分析 5

1.3 查询所有数据 6

1.4 不仅仅是批处理 7

1.5 相较于其他系统的优势 8

1.5.1 关系型数据库管理系统 8

1.5.2 网格计算 10

1.5.3 志愿计算 11

1.6 Apache Hadoop发展简史 12

1.7 本书包含的内容 16

第2章 关于MapReduce	19
2.1 气象数据集	19
2.2 使用Unix工具来分析数据	21
2.3 使用Hadoop来分析数据	22
2.3.1 map和reduce	23
2.3.2 Java MapReduce	24
2.4 横向扩展	31
2.4.1 数据流	31
2.4.2 biner函数	35
2.4.3 运行分布式的MapReduce作业	37
2.5 Hadoop Streaming	37
2.5.1 Ruby版本	38
2.5.2 Python版本	40
第3章 Hadoop分布式文件系统	42
3.1 HDFS的设计	42
3.2 HDFS的概念	44
3.2.1 数据块	44
3.2.2 namenode和datanode	45
3.2.3 块缓存	46
3.2.4 联邦HDFS	47
3.2.5 HDFS的高可用性	47
3.3 命令行接口	50
3.4 Hadoop文件系统	52
3.5 Java接口	56
3.5.1 从Hadoop URL读取数据	56
3.5.2 通过FileSystem API读取数据	58
3.5.3 写入数据	61
3.5.4 目录	63
3.5.5 查询文件系统	63
3.5.6 删除数据	68
3.6 数据流	68
3.6.1 剖析文件读取	68
3.6.2 剖析文件写入	71
3.6.3 一致模型	74
3.7 通过distcp并行复制	76
第4章 关于YARN	78
4.1 剖析YARN应用运行机制	79
4.1.1 资源请求	80
4.1.2 应用生命期	81
4.1.3 构建YARN应用	81
4.2 YARN与MapReduce 1相比	82
4.3 YARN中的调度	85
4.3.1 调度选项	85
4.3.2 容量调度器配置	87
4.3.3 公平调度器配置	89
4.3.5 延迟调度	93
4.3.5 主导资源公平性	94
4.4 延伸阅读	95
第5章 Hadoop的I/O操作	96
5.1 数据完整性	96
5.1.1 HDFS的数据完整性	97
5.1.2 LocalFileSystem	98
5.1.3 ChecksumFileSystem	98
5.2 压缩	99

- 5.2.1 codec 100
- 5.2.2 压缩和输入分片 105
- 5.2.3 在MapReduce中使用压缩 106
- 5.3 序列化 109
 - 5.3.1 Writable接口 110
 - 5.3.2 Writable类 112
 - 5.3.3 实现定制的Writable集合 121
 - 5.3.4 序列化框架 125
- 5.4 基于文件的数据结构 127
 - 5.4.1 关于SequenceFile 127
 - 5.4.2 关于MapFile 135
 - 5.4.3 其他文件格式和面向列的格式 136
- 第 II 部分 关于MapReduce
- 第6章 MapReduce应用开发 141
 - 6.1 用于配置的API 142
 - 6.1.1 资源合并 143
 - 6.1.2 变量扩展 144
 - 6.2 配置开发环境 144
 - 6.2.1 管理配置 146
 - 6.2.2 辅助类GenericOptionsParser, Tool和ToolRunner 149
 - 6.3 用MRUnit来写单元测试 152
 - 6.3.1 关于Mapper 152
 - 6.3.2 关于Reducer 156
 - 6.4 本地运行测试数据 156
 - 6.4.1 在本地作业运行器上运行作业 156
 - 6.4.2 测试驱动程序 158
 - 6.5 在集群上运行 160
 - 6.5.1 打包作业 160
 - 6.5.2 启动作业 162
 - 6.5.3 MapReduce的Web界面 165
 - 6.5.4 获取结果 167
 - 6.5.5 作业调试 168
 - 6.5.6 Hadoop日志 171
 - 6.5.7 远程调试 173
 - 6.6 作业调优 174
 - 6.7 MapReduce的工作流 176
 - 6.7.1 将问题分解成MapReduce作业 177
 - 6.7.2 关于JobControl 178
 - 6.7.3 关于Apache Oozie 179
- 第7章 MapReduce的工作机制 184
 - 7.1 剖析MapReduce作业运行机制 184
 - 7.1.1 作业的提交 185
 - 7.1.2 作业的初始化 186
 - 7.1.3 任务的分配 187
 - 7.1.4 任务的执行 188
 - 7.1.5 进度和状态的更新 189
 - 7.1.6 作业的完成 191
 - 7.2 失败 191
 - 7.2.1 任务运行失败 191
 - 7.2.2 application master运行失败 193
 - 7.2.3 节点管理器运行失败 193
 - 7.2.4 资源管理器运行失败 194
 - 7.3 shuffle和排序 195
 - 7.3.1 map端 195

- 7.3.2 reduce端 197
- 7.3.3 配置调优 199
- 7.4 任务的执行 201
 - 7.4.1 任务执行环境 201
 - 7.4.2 推测执行 202
 - 7.4.3 关于OutputCommitters 204
- 第8章 MapReduce的类型与格式 207
 - 8.1 MapReduce的类型 207
 - 8.1.1 默认的MapReduce作业 212
 - 8.1.2 默认的Streaming作业 216
 - 8.2 输入格式 218
 - 8.2.1 输入分片与记录 218
 - 8.2.2 文本输入 229
 - 8.2.3 二进制输入 233
 - 8.2.4 多个输入 234
 - 8.2.5 数据库输入（和输出） 235
 - 8.3 输出格式 236
 - 8.3.1 文本输出 236
 - 8.3.2 二进制输出 237
 - 8.3.3 多个输出 237
 - 8.3.4 延迟输出 242
 - 8.3.5 数据库输出 242
- 第9章 MapReduce的特性 243
 - 9.1 计数器 243
 - 9.1.1 内置计数器 243
 - 9.1.2 用户定义的Java计数器 248
 - 9.1.3 用户定义的Streaming计数器 251
 - 9.2 排序 252
 - 9.2.1 准备 252
 - 9.2.2 部分排序 253
 - 9.2.3 全排序 255
 - 9.2.4 辅助排序 259
 - 9.3 连接 264
 - 9.3.1 map端连接 266
 - 9.3.2 reduce端连接 266
 - 9.4 边数据分布 270
 - 9.4.1 利用JobConf来配置作业 270
 - 9.4.2 分布式缓存 270
 - 9.5 MapReduce库类 276
- 第III部分 Hadoop的操作
- 第10章 构建Hadoop集群 279
 - 10.1 集群规范 280
 - 10.1.1 集群规模 281
 - 10.1.2 网络拓扑 282
 - 10.2 集群的构建和安装 284
 - 10.2.1 安装Java 284
 - 10.2.2 创建Unix 用户账号 284
 - 10.2.3 安装Hadoop 284
 - 10.2.4 SSH配置 285
 - 10.2.5 配置Hadoop 286
 - 10.2.6 格式化HDFS 文件系统 286
 - 10.2.7 启动和停止守护进程 286
 - 10.2.8 创建用户目录 288
 - 10.3 Hadoop配置 288

- 10.3.1 配置管理 289
- 10.3.2 环境设置 290
- 10.3.3 Hadoop守护进程的关键属性 293
- 10.3.4 Hadoop守护进程的地址和端口 300
- 10.3.5 Hadoop的其他属性 303
- 10.4 安全性 305
 - 10.4.1 Kerberos和Hadoop 306
 - 10.4.2 委托令牌 308
 - 10.4.3 其他安全性改进 309
- 10.5 利用基准评测程序测试Hadoop集群 311
 - 10.5.1 Hadoop基准评测程序 311
 - 10.5.2 用户作业 313
- 第11章 管理Hadoop 314
 - 11.1 HDFS 314
 - 11.1.1 永久性数据结构 314
 - 11.1.2 安全模式 320
 - 11.1.3 日志审计 322
 - 11.1.4 工具 322
 - 11.2 监控 327
 - 11.2.1 日志 327
 - 11.2.2 度量和JMX (Java管理扩展) 328
 - 11.3 维护 329
 - 11.3.1 日常管理过程 329
 - 11.3.2 委任和解除节点 331
 - 11.3.3 升级 334
- 第IV部分 Hadoop相关开源项目
- 第12章 关于Avro 341
 - 12.1 Avro数据类型和模式 342
 - 12.2 内存中的序列化和反序列化特定API 347
 - 12.3 Avro数据文件 349
 - 12.4 互操作性 351
 - 12.4.1 Python API 351
 - 12.4.2 Avro工具集 352
 - 12.5 模式解析 352
 - 12.6 排列顺序 354
 - 12.7 关于Avro MapReduce 356
 - 12.8 使用Avro MapReduce进行排序 359
 - 12.9 其他语言的Avro 362
- 第13章 关于Parquet 363
 - 13.1 数据模型 364
 - 13.2 Parquet文件格式 367
 - 13.3 Parquet的配置 368
 - 13.4 Parquet文件的读/写 369
 - 13.4.1 Avro、Protocol Buffers和Thrift 371
 - 13.4.2 投影模式和读取模式 373
 - 13.5 Parquet MapReduce 374
- 第14章 关于Flume 377
 - 14.1 安装Flume 378
 - 14.2 示例 378
 - 14.3 事务和可靠性 380
 - 14.4 HDFS Sink 382
 - 14.5 扇出 385
 - 14.5.1 交付保证 386
 - 14.5.2 复制和复用选择器 387

- 14.6 通过代理层分发 387
- 14.7 Sink组 391
- 14.8 Flume与应用程序的集成 395
- 14.9 组件编目 395
- 14.10 延伸阅读 397
- 第15章 关于Sqoop 398
 - 15.1 获取Sqoop 398
 - 15.2 Sqoop连接器 400
 - 15.3 一个导入的例子 401
 - 15.4 生成代码 404
 - 15.5 深入了解数据库导入 405
 - 15.5.1 导入控制 407
 - 15.5.2 导入和一致性 408
 - 15.5.3 增量导入 408
 - 15.5.4 直接模式导入 408
 - 15.6 使用导入的数据 409
 - 15.7 导入大对象 412
 - 15.8 执行导出 414
 - 15.9 深入了解导出功能 416
 - 15.9.1 导出与事务 417
 - 15.9.2 导出和SequenceFile 418
 - 15.10 延伸阅读 419
- 第16章 关于Pig 420
 - 16.1 安装与运行Pig 421
 - 16.1.1 执行类型 422
 - 16.1.2 运行Pig程序 423
 - 16.1.3 Grunt 424
 - 16.1.4 Pig Latin编辑器 424
 - 16.2 示例 425
 - 16.3 与数据库进行比较 428
 - 16.4 PigLatin 429
 - 16.4.1 结构 430
 - 16.4.2 语句 431
 - 16.4.3 表达式 436
 - 16.4.4 类型 437
 - 16.4.5 模式 438
 - 16.4.6 函数 443
 - 16.4.7 宏 445
 - 16.5 用户自定义函数 446
 - 16.5.1 过滤UDF 447
 - 16.5.2 计算UDF 450
 - 16.5.3 加载UDF 452
 - 16.6 数据处理操作 455
 - 16.6.1 数据的加载和存储 455
 - 16.6.2 数据的过滤 455
 - 16.6.3 数据的分组与连接 458
 - 16.6.4 数据的排序 463
 - 16.6.5 数据的组合和切分 465
 - 16.7 Pig实战 465
 - 16.7.1 并行处理 465
 - 16.7.2 匿名关系 466
 - 16.7.3 参数代换 467
 - 16.8 延伸阅读 468
- 第17章 关于Hive 469

- 17.1 安装Hive 470
 - Hive的shell环境 471
- 17.2 示例 472
- 17.3 运行Hive 473
 - 17.3.1 配置Hive 473
 - 17.3.2 Hive服务 476
 - 17.3.3 Metastore 478
- 17.4 Hive与传统数据库相比 480
 - 17.4.1 读时模式vs.写时模式 480
 - 17.4.2 更新、事务和索引 481
 - 17.4.3 其他SQL—on—Hadoop技术 482
- 17.5 HiveQL 483
 - 17.5.1 数据类型 484
 - 17.5.2 操作与函数 487
- 17.6 表 488
 - 17.6.1 托管表和外部表 488
 - 17.6.2 分区和桶 490
 - 17.6.3 存储格式 494
 - 17.6.4 导入数据 498
 - 17.6.5 表的修改 500
 - 17.6.6 表的丢弃 501
- 17.7 查询数据 501
 - 17.7.1 排序和聚集 501
 - 17.7.2 MapReduce脚本 502
 - 17.7.3 连接 503
 - 17.7.4 子查询 506
 - 17.7.5 视图 507
- 17.8 用户定义函数 508
 - 17.8.1 写UDF 510
 - 17.8.2 写UDAF 512
- 17.9 延伸阅读 516
- 第18章 关于Crunch 517
 - 18.1 示例 518
 - 18.2 Crunch核心API 521
 - 18.2.1 基本操作 522
 - 18.2.2 类型 527
 - 18.2.3 源和目标 530
 - 18.2.4 函数 532
 - 18.2.5 物化 535
 - 18.3 管线执行 537
 - 18.3.1 运行管线 538
 - 18.3.2 停止管线 539
 - 18.3.3 查看Crunch计划 540
 - 18.3.4 迭代算法 543
 - 18.3.5 给管线设置检查点 544
 - 18.4 Crunch库 545
 - 18.5 延伸阅读 547
- 第19章 关于Spark 548
 - 19.1 安装Spark 549
 - 19.2 示例 549
 - 19.2.1 Spark应用、作业、阶段和任务 551
 - 19.2.2 Scala独立应用 552
 - 19.2.3 Java示例 553
 - 19.2.4 Python示例 554

- 19.3 弹性分布式数据集 555
 - 19.3.1 创建 555
 - 19.3.2 转换和动作 557
 - 19.3.3 持久化 561
 - 19.3.4 序列化 563
- 19.4 共享变量 564
 - 19.4.1 广播变量 564
 - 19.4.2 累加器 565
- 19.5 剖析Spark作业运行机制 565
 - 19.5.1 作业提交 566
 - 19.5.2 DAG的构建 566
 - 19.5.3 任务调度 569
 - 19.5.4 任务执行 570
- 19.6 执行器和集群管理器 570
- 19.7 延伸阅读 574
- 第20章 关于HBase 575
 - 20.1 HBase基础 575
 - 20.2 概念 576
 - 20.2.1 数据模型的“旋风之旅” 576
 - 20.2.2 实现 578
 - 20.3 安装 581
 - 20.4 客户端 584
 - 20.4.1 Java 584
 - 20.4.2 MapReduce 588
 - 20.4.3 REST和Thrift 589
 - 20.5 创建在线查询应用 589
 - 20.5.1 模式设计 590
 - 20.5.2 加载数据 591
 - 20.5.3 在线查询 595
 - 20.6 HBase和RDBMS的比较 598
 - 20.6.1 成功的服务 599
 - 20.6.2 HBase 600
 - 20.7 Praxis 601
 - 20.7.1 HDFS 601
 - 20.7.2 用户界面 602
 - 20.7.3 度量 602
 - 20.7.4 计数器 602
 - 20.8 延伸阅读 602
- 第21章 关于ZooKeeper 604
 - 21.1 安装和运行ZooKeeper 605
 - 21.2 示例 607
 - 21.2.1 ZooKeeper中的组成员关系 608
 - 21.2.2 创建组 608
 - 21.2.3 加入组 611
 - 21.2.4 列出组成员 612
 - 21.2.5 删除组 614
 - 21.3 ZooKeeper服务 615
 - 21.3.1 数据模型 615
 - 21.3.2 操作 618
 - 21.3.3 实现 622
 - 21.3.4 一致性 624
 - 21.3.5 会话 626
 - 21.3.6 状态 628
 - 21.4 使用ZooKeeper来构建应用 629

21.4.1 配置服务 629
21.4.2 可复原的ZooKeeper应用 633
21.4.3 锁服务 637
21.4.4 更多分布式数据结构和协议 639
21.5 生产环境中的ZooKeeper 640
21.5.1 可恢复性和性能 641
21.5.2 配置 642
21.6 延伸阅读 643
第V部分 案例学习
第22章 医疗公司塞纳（Cerner）的可聚合数据 647
22.1 从多CPU到语义集成 647
22.2 进入Apache Crunch 648
22.3 建立全貌 649
22.4 集成健康医疗数据 651
22.5 框架之上的可组合性 654
22.6 下一步 655
第23章 生物数据科学：用软件拯救生命 657
23.1 DNA的结构 659
23.2 遗传密码：将DNA字符转译为蛋白质 660
23.3 将DNA想象成源代码 661
23.4 人类基因组计划和参考基因组 663
23.5 DNA测序和比对 664
23.6 ADAM，一个可扩展的基因组分析平台 666
23.7 使用Avro接口描述语言进行自然语言编程 666
23.8 使用Parquet进行面向列的存取 668
23.9 一个简单例子：用Spark和ADAM做k—mer计数 669
23.10 从个性化广告到个性化医疗 672
23.11 联系我们 673
第24章 开源项目Cascading 674
24.1 字段、元组和管道 675
24.2 操作 678
24.3 Taps，Schemes和Flows 680
24.4 Cascading实践应用 681
24.5 灵活性 684
24.6 ShareThis中的Hadoop和Cascading 685
24.7 总结 689
附录A 安装Apache Hadoop 691
附录B 关于CDH 697
附录C 准备NCDC气象数据 699
附录D 新版和旧版JavaMapReduce API 702
• • • • • ([收起](#))

[Hadoop权威指南:大数据的存储与分析\(第4版\)\(修订版\)\(升级版\) 下载链接1](#)

标签

大数据

Hadoop

计算机

hadoop

数据分析

编程

Java

数据存储

评论

这本书最大的问题就是想做大做全，但是碍于篇幅限制，什么都讲不清楚。hive、spark、HBASE、Pig、map reduce、Arvo、zookeeper.....统统都说，但也都点到为止。不仅晦涩难懂，内容也不深，看了等于白看系列。

web程序可以看做大数据分析的前台程序或者是数据采集和展示的程序。而hadoop可以看做一个框架或者数据库引擎。

概念清晰，内容全面

结构跟第一版没太大区别，有些内容稍微有点过时，例如HBase的API。

翻的什么玩儿？？

全而较浅。

系统的介绍了hadoop原理及其重要的组建，很经典的书籍。但是这些东西只有自己上手用之后，才能理解的更深。目前看的不是太懂

除了实战练习和第四部分只看了hive之外，其他看了写的不错

因架构设计选型的需要，快速撸了一遍，建立一个概念概要大图。

日常随使用用的话，没必要读这本，如果想深入了解Hadoop生态，以及做数据开发、架构选型，还是要仔细阅读一番的

列了一遍 api
菜名（这些容易变化的点）…明明就是按照文档过一遍就能获取的知识。内容也不够深度…如果作为提纲挈领…结构也显得混乱…翻译也有点问题

算是大数据入门指南，看完能理解啥叫MR

非常适合用于新手入门hadoop技术栈

出了pig和crunch，全部阅读完毕

书不错 翻译太烂 比较全面 深度一般 推荐给新手

选择性的读了感兴趣的章节，满足了自己的需求，是入门数据 介绍比较全面的书籍。

真有那么好? 反正也要处理掉了

适合入门，帮助了解Hadoop的基础原理及其生态，但全书最喜欢的一句话却与主题无关：我们这一代最杰出的脑袋都在拼命思考如何吸引人们点击更多的广告，这种感觉糟透了一Jeff Hammerbacher

为了准备面试囫圇吞枣读了遍。比较理论，适合当字典。

只读了第一部分，第三部分，第四部分的部分章节（第12章：关于Avro,第13章：关于Parquet,第14章：关于Flume,第15章：关于Sqoop,第17章：关于Hive,第19章：关于Spark)

[Hadoop权威指南:大数据的存储与分析\(第4版\)\(修订版\)\(升级版\) 下载链接1](#)

书评

买了第一版，时间太紧，没来得及看，后来出了个号称修订升级的第二版，毫不犹豫又买了，后来听说第二版比第一版翻译得好，心中窃喜，再后来看了第二版，我震惊了，我TM就是一傻子，放着好好的英文版不看，赶什么时髦买中文版呢。在这个神奇的国度，牛奶里放的是三聚氰胺，火腿...

其实也不算全部读完了，读它主要是为了技术选型，考虑升级持久层架构、提高系统可扩展性，仔细研读了前几章，对Hadoop、MapReduce、HDFS的模型、机制、使用场景有了一定了解。后面几章及其生态圈内的其他项目抱着了解的心态简单浏览了一下。整体感觉还行，至少从我看过的章节来...

中文版412页：

所以理论上，任何东西都可以表示成二进制形式，然后转化成为长整型的字符串或直接对数据结构进行序列化，来作为键值。原文460页： ..., so theoretically anything can serve as row key, from strings to binary representations of long or even serialized ...

-- china-pub 赠书活动 -- <http://www.douban.com/group/topic/20965935/>

一直比较忙，整本书还没读完，只是粗略翻了个大概，其中有两三章细读了一遍。先做个大体评价吧，有时间全部细读后再评论。

从书的内容上来讲，大致上与网上该书的内容介绍一致。简单点概括：这本书对...

参加豆瓣China-pub抽奖，比较幸运的得到这本Hadoop权威指南中文第二版，拿来与第一版相比，发现新加入了Hive和Sqoop章节，译文质量也提高了不少，并且保留了英文索引。

这本书对Hadoop的介绍还算全面，有实践冲动的朋友基本可以拿着书、配合Google百度马上实现梦想。个人感觉 “...

看了几章中文版的，各种错误，太低级，实在是看不下去了。建议还是看原版吧。译者们的脸皮可真厚，英文译不明白也就罢了，中文都组织的不通顺，好意思吗！！什么叫“但是，.....，但是”啊，“但是体”啊。

很好的Hadoop教程，比Apache和Yahoo

!网页版guide详细很多，很多想不明白的Hadoop实现细节都可以在这本书里找到。

详见：<http://www.cnblogs.com/aprilrain/archive/2013/03/07/2947664.html>

很多地方翻译的不行，需要对照英文看才能明白。。。不过对于快速学习，仍然是不错的选择。建议译者看看每部分内容的重要性，不重要的瞎翻翻就算了，重要的部分还是好好花点功夫，不要本末倒置了。比如第三章的数据流部分，这么经典的地方居然被翻译烂的一塌糊涂。不知道译者会...

书中没有透露太多实现架构方面的细节，更多的是从使用者的角度上介绍了Hadoop的各种知识，包括MapReduce, HDFS, Hive, Pig, HBase, ZooKeeper。几乎涉及了Hadoop的所有关于使用方面的知识，包括安装和使用。你甚至可以直接在自己的电脑上装上一个Hadoop，对着书中的例子实际演...

是我遇到过的翻译最烂的一本书，在译者的“妙语连珠”里折腾了半个钟头就再也没兴趣了。略举几例如下： P.6 任然 -> 仍然 P.21 输入键（为什么不像后面那样有个“的”？），输入的值，输出的键…… P. 27 “计数器” (Counter)，译文附原文；"Context Object"(上下文对象)，原...

专门登录来评论的，翻译也太烂了吧，真的真的建议强烈英语阅读能力好的人去读原版书，不要花冤枉钱在这上面，除了文字错误外，里边的图居然也有错，就比如260页的图最后两个年份应该是1901结果这里竟然是1900，我是真滴服了，一本神书被翻译成这样，作者得气死。zsbd zsbd zsbd...

你的履历添了一笔<hadoop权威指南>译者,但是你不配 这是我见过的最不用心的翻译,字里行间行文不通顺,请别勉强自己,map reduce shuffle机制都没翻译的好 虽然原作者写作功底也实在是一般 第 1 2 5 6 7 这几章 翻译的实在是太烂了 请不要呐Google翻译糊弄人阿 误人子弟 ...

首先，翻译太差，很多句子就是瞎翻，根本不通顺，很多时候你要停下来断句，慢慢去理解。然后，这本书是很多人去翻译的，很多人连代码都不懂，曾经一段代码看到我蒙圈，去看了一下源代码，好家伙，四行有五个错误。另外，从代码瞎缩进也可以看出这是群没写过代码的人翻的，而且...

Cobub Razor APP数据统计分析工具官网上有篇文章是讲Hadoop
Yarn调度器的选择和使用的，我觉得写的挺好的，推荐<http://www.cobub.com/the-selection-and-use-of-hadoop-yarn-scheduler/>

[Hadoop权威指南:大数据的存储与分析\(第4版\)\(修订版\)\(升级版\) 下载链接1](#)