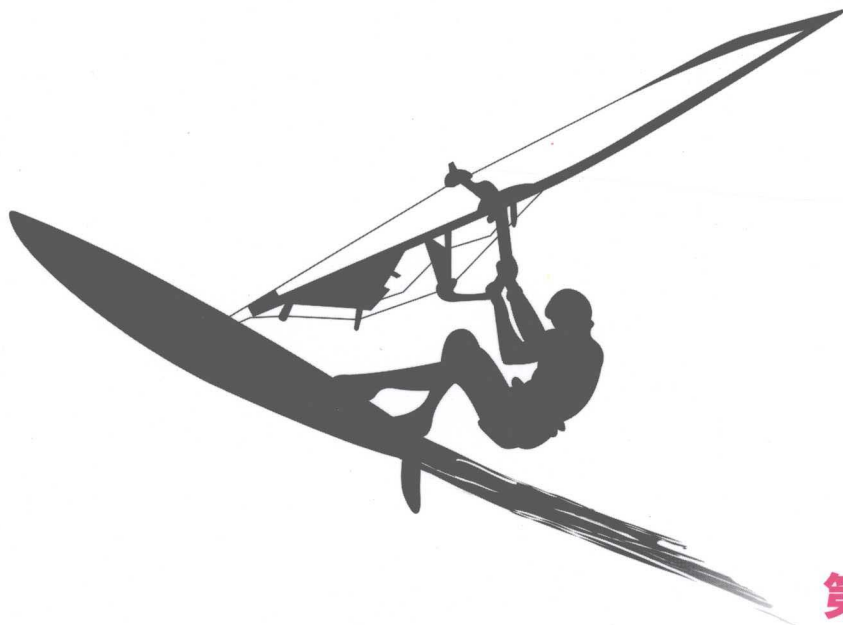


第1版广受好评，第2版基于Hadoop及其相关技术最新版本撰写，从多角度做了全面的修订和补充

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



第2版

陆嘉恒 著

Hadoop in Action Second Edition

Hadoop实战



机械工业出版社
China Machine Press



第2版

Hadoop in Action, Second Edition

Hadoop实战

陆嘉恒 著



机械工业出版社
China Machine Press

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。

封底无防伪标均为盗版

版权所有, 侵权必究

本书法律顾问 北京市展达律师事务所

图书在版编目 (CIP) 数据

Hadoop 实战 / 陆嘉恒著. —2 版. —北京: 机械工业出版社, 2012.11

ISBN 978-7-111-39583-6

I. H… II. 陆… III. 数据处理—应用软件 IV. TP274

中国版本图书馆 CIP 数据核字 (2012) 第 208678 号

机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码 100037)

责任编辑: 孙海亮

北京市荣盛彩色印刷有限公司印刷

2012 年 11 月第 2 版第 1 次印刷

186mm×240mm·32.25 印张

标准书号: ISBN 978-7-111-39583-6

定价: 79.00 元

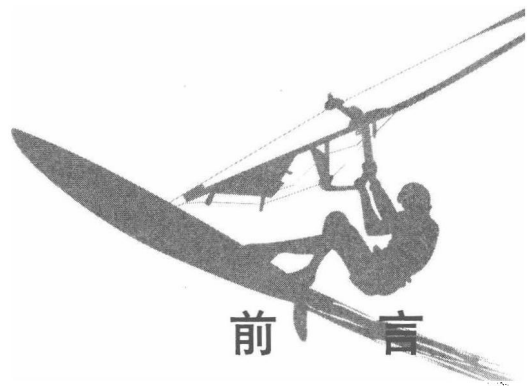
凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

客服热线: (010) 88378991; 88361066

购书热线: (010) 68326294; 88379649; 68995259

投稿热线: (010) 88379604

读者信箱: hzjsj@hzbook.com



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。

在线资源及勘误

在本书的附录中，提供了一个基于 Hadoop 的云计算在线测试平台（<http://cloud-computing.ruc.edu.cn>），大家可以先注册一个免费账户，然后即可体验 Hadoop 平台，通过该平台大家可在线编写 MapReduce 应用并进行自动验证。如果大家希望获得该平台的验证码，或者希望获得完全编程测试和理论测试的权限，请发邮件到 jiahenglu@gmail.com。读者也可访问 Hadoop 的官方网站（hadoop.apache.org）阅读官方介绍文档，下载学习示例代码。

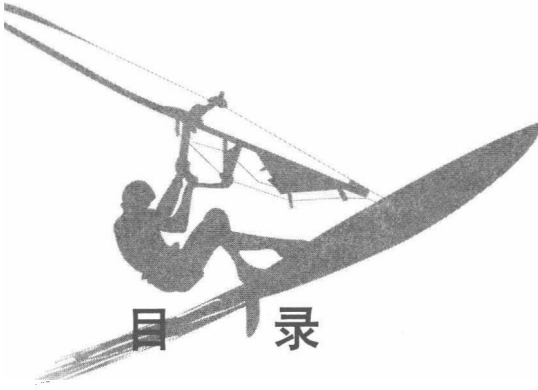
在本书的撰写和相关技术的研究中，尽管笔者投入了大量的精力、付出了艰辛的努力，但是受知识水平所限，书中存在不足和疏漏之处在所难免，恳请大家批评指正。如果有任何问题和建议，可发送电子邮件至 jiahenglu@gmail.com 或 jiahenglu@ruc.edu.cn。

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



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

VIII

- 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