

# 《Hadoop应用开发实战详解-深入云肌

## 图书基本信息

书名：《Hadoop应用开发实战详解-深入云计算》

13位ISBN编号：9787113161934

10位ISBN编号：7113161936

出版时间：2013-6

出版社：万川梅 中国铁道 (2013-06出版)

作者：万川梅

页数：397

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《Hadoop应用开发实战详解-深入云肌

## 内容概要

本书由浅入深，全面、系统地介绍了Hadoop这一高性能处理海量数据集的理想工具。本书的内容主要包括hdfs、mapreduce、hive、hbase、mahout、pig、zookeeper、avro、chukwa等与hadoop相关的子项目，各个知识点都精心设计了大量经典的小案例，实战型强，可操作性强。

## 书籍目录

第1篇 Hadoop技术篇第1章 初识Hadoop 1.1 Hadoop简介 1.1.1 Hadoop的起源 1.1.2 什么是Hadoop 1.1.3 Hadoop的核心技术是Google核心技术的开源实现 1.1.4 Hadoop的功能与优点 1.1.5 Hadoop的应用现状和发展趋势 1.2 Hadoop的体系结构 1.2.1 HDFS的体系结构 1.2.2 MapReduce的体系结构 1.3 Hadoop与分布式开发 1.4 Hadoop的数据管理 1.4.1 HDFS的数据管理 1.4.2 HBase的数据管理 1.4.3 Hive的数据管理 1.5 思考与总结 第2章 Hadoop的安装和配置 2.1 在Windows下安装与配置Hadoop 2.1.1 JDK的安装 2.1.2 Cygwin的安装 2.1.3 Hadoop的安装 2.2 在Linux下安装与配置Hadoop 2.2.1 Ubuntu的安装 2.2.2 JDK的安装 2.2.3 Hadoop的安装 2.3 Hadoop的执行实例 2.3.1 运行Hadoop 2.3.2 运行wordcount.java程序 2.4 Hadoop Eclipse简介和使用 2.4.1 Eclipse插件介绍 2.4.2 Eclipse插件开发配置 2.4.3 在Eclipse下运行WordCount程序 2.5 Hadoop的集群和优化 2.5.1 Hadoop的性能优化 2.5.2 Hadoop配置机架感知信息 2.6 思考与总结 第3章 HDFS海量存储 3.1 开源的GFS——HDFS 3.1.1 HDFS简介 3.1.2 HDFS的体系结构 3.1.3 HDFS的保障可靠性措施 3.2 HDFS的常用操作 3.2.1 HDFS下的文件操作 3.2.2 管理与更新 3.2.3 HDFS API详解 3.2.4 HDFS的读/写数据流 3.3 用HDFS存储海量的视频数据 3.3.1 场景分析 3.3.2 设计实现 3.4 思考与总结 第4章 初识MapReduce 4.1 MapReduce简介 4.1.1 MapReduce要解决什么问题 4.1.2 MapReduce的理论基础 4.1.3 MapReduce的编程模式 4.2 MapReduce的集群行为 4.3 Map/Reduce框架 4.4 样例分析：单词计数 4.4.1 WordCount实例的运行过程 4.4.2 WordCount的源码分析和程序处理过程 4.4.3 MapReduce常用类及其接口 4.5 实例：倒排索引 4.5.1 倒排索引的分析和设计 4.5.2 倒排索引完整源码 4.5.3 运行代码结果 4.6 MapReduce 在日志分析中数据去重案例 4.6.1 什么是数据去重 4.6.2 设计思路 4.6.3 程序代码 4.6.4 代码运行结果 4.7 数据排序实例 4.7.1 实例描述 4.7.2 设计思路 4.7.3 程序代码 4.8 思考与总结 第5章 分布式开源数据库HBase 5.1 HBase简介 5.1.1 HBase逻辑视图 5.1.2 HBase物理存储 5.1.3 子表Region服务器 5.1.4 Hmaster主服务器 5.1.5 元数据表 5.2 HBase的安装配置 5.2.1 HBase单机模式 5.2.2 HBase伪分布模式 5.2.3 HBase完全分布模式 5.3 学生成绩表实例 5.3.1 Shell的基本操作 5.3.2 代码实现 5.3.3 关于中文的处理 5.3.4 常用HBase的Shell操作 5.4 思考与总结 第6章 MapReduce进阶 6.1 API的配置 6.1.1 一个简单的配置文件 6.1.2 合并多个源文件 6.1.3 可变的扩展 6.2 配置开发环境 6.2.1 配置文件设置 6.2.2 设置用户标识 6.3 复合键值对的使用 6.3.1 小的键值对如何合并成大的键值对 6.3.2 巧用复合键让系统完成排序 6.4 用户定制数据类型 6.4.1 内置数据类型 6.4.2 用户自定义数据类型 6.5 用户定制输入/输出格式 6.5.1 内置数据的输入格式 6.5.2 用户定制数据输入格式与RecordReader 6.5.3 Hadoop内置的数据输出格式 6.5.4 Hadoop内置的数据输出格式与RecordWriter 6.6 用户定制Partitioner和Combiner 6.7 组合式的MapReduce作业 6.7.1 MapReduce作业运行机制 6.7.2 组合式MapReduce计算作业 6.8 DataJoin链接多数据源 6.9 思考与总结 第7章 Hive数据仓库 7.1 Hive简介 7.2 Hive安装与配置 7.3 Hive的服务 7.3.1 Hive shell 7.3.2 JDBC/ODBC 7.3.3 Thrift服务 7.3.4 Web接口 7.3.5 元数据服务 7.4 HiveQL查询语言 7.5 Hive实例 7.5.1 UDF 编程实例 7.5.2 UDAF 编程实例 7.5.3 Hive的日志数据统计实战 7.6 思考与总结 第8章 Pig开发应用 8.1 Pig简介 8.2 Pig的安装与配置 8.3 Pig的使用 8.3.1 Pig的MapReduce模式 8.3.2 Pig的运行方式 8.4 通过Grunt学习Pig Latin 8.4.1 Pig的数据模型 8.4.2 运算符 8.4.3 常用操作 8.4.4 各种SQL在Pig中的实现 8.4.5 Pig Latin实现 8.5 Pig使用的案例 8.6 思考与总结 第9章 Chukwa数据收集系统 9.1 Chukwa简介 9.1.1 Chukwa是什么 9.1.2 Chukwa主要解决什么问题 9.2 Chukwa的安装配置 9.2.1 Chukwa的安装 9.2.2 Chukwa的配置 9.2.3 Chukwa的启动 9.3 Chukwa的基本命令 9.3.1 Chukwa端的命令 9.3.2 Agent 端的命令 9.4 Chukwa在数据收集处理方面的运用 9.4.1 数据生成 9.4.2 数据收集 9.4.3 数据处理 9.4.4 数据析取 9.4.5 数据稀释 9.4.6 数据显示 9.5 思考与总结 第10章 ZooKeeper开发应用 10.1 ZooKeeper简介 10.1.1 ZooKeeper的设计目标 10.1.2 ZooKeeper主要解决什么问题 10.1.3 ZooKeeper的基本概念和工作原理 10.2 ZooKeeper的安装配置 10.2.1 单机模式 10.2.2 启动并测试ZooKeeper 10.2.3 集群模式 10.3 ZooKeeper提供的接口 10.4 ZooKeeper事件 10.5 ZooKeeper实例 10.5.1 实例1：一个简单的应用——分布式互斥锁 10.5.2 实例2：进程调度系统 10.6 思考与总结 第2篇 Hadoop管理和容错篇第11章 Hadoop管理 11.1 Hadoop权限管理 11.2 HDFS文件系统管理 11.3 Hadoop维护与管理 11.4 Hadoop常见问题及解决办法 11.5 思考与总结 第12章 Hadoop容错 12.1 Hadoop的可靠性 12.1.1 HDFS中的NameNode单点失效解决方案 12.1.2 HDFS数据块副本机制 12.1.3 HDFS心跳机制 12.1.4 HDFS负载均衡 12.1.5 MapReduce容错 12.2 Hadoop的SecondaryNameNode机制 12.2.1 磁盘镜像与日志文件 12.2.2 SecondaryNameNode更新镜像的流程

12.3 Avatar机制 12.3.1 Avatar机制简介 12.3.2 Avatars部署实战 12.4 Hadoop\_HBase容错 12.5 思考与总结 第3篇 Hadoop实战篇第13章 综合实战1：Hadoop中的数据库访问 13.1 DBInputFormat类访问数据库 13.1.1 在DBInputFormat类中包含的内置类 13.1.2 使用DBInputFormat读取数据库表中的记录 13.1.3 使用示例 13.2 使用DBOutputFormat向数据库中写记录 13.3 思考与总结 第14章 综合实战2：一个简单的分布式的Grep 14.1 分析与设计 14.2 实现代码 14.3 运行程序 14.4 思考与总结 第15章 综合实战3：打造一个搜索引擎 15.1 搜索引擎工作原理 15.2 网页搜集与信息提取 15.2.1 设计的主要思想 15.2.2 系统设计目标 15.3 网页信息的提取与存储 15.4 MapReduce的预处理 15.4.1 第一步：源数据过滤 15.4.2 第二步：生成倒排文件 15.4.3 第三步：建立二级索引 15.5 建立Web信息查询服务 15.6 思考与总结 第16章 综合实战4：移动通信信令监测与查询 16.1 分析与设计 16.1.1 CDR数据文件的检测与索引创建任务调度 16.1.2 从HDFS读取数据并创建索引 16.1.3 查询CDR信息 16.2 代码实现 16.2.1 CDR文件检测和索引创建任务程序 16.2.2 读取CDR数据和索引创建处理 16.2.3 CDR查询 16.3 思考与总结 附录A Hadoop命令大全 附录B HDFS命令大全

# 《Hadoop应用开发实战详解-深入云肌

## 编辑推荐

《深入云计算(Hadoop应用开发实战详解)》编著者万川梅、谢正兰。 重点推荐！深入云计算系列丛书：《深入云计算：Hadoop应用开发实战详解》《深入云计算：MongoDB管理与开发实战详解》《深入云计算：Hadoop源代码分析》 精准的内容梯度安排 遵循读者学习习惯和Hadoop技术应用实践，合理安排图书内容；精挑细选经典实例，嵌入完善的代码注释。 精炼的实用经验阐述 作者多年开发经验融入其中，读者在全面掌握Hadoop编程和开发技术的同时更能获得快速分析和解决实际问题的能力。

## 精彩短评

1、很多命令行没有空格，还出现linux中下载.exe文件这种错误.....

## 精彩书评

1、1.本书读者对象中提到的“编者也会定期在www.rzchina.net进行答疑解惑”,在百度中搜索“hadoop site:rzchina.net 搜不到任何内容”(2014年7月16日15:24分搜索结果)2.第10行,“在一个910个节点的上,Hadoop仅仅用了209秒(不到3分钟)就完成了1TB数据排序……”第一个状语从句成分残缺。3.第4页,第6行,“(前提是安装了cygwincygwin)”,正确的应该是cygwin仅仅开头四页就出现两处错误,让人觉得这不是一本严谨的书。抢占市场固然是优势,但仓促到连校对一下的时间都没有吗?太不负责任了。4.第7页图1-2 Hadoop的运用地域分布图,第9,第10页Hadoop发展趋势调研均没说明数据来源,也没说明调查年限,单纯地摆数据。5.hadoop版本用的是hadoop 1.0.1这是2013年出的书,我本来期待hadoop的版本会比前几年的书更新一点的。6.36页倒数第二行。“解压安装包adoop-0.21.0.tar.gz”虽然安装包名可以随便改的,感觉这里是印漏了h的可能性大些。事不过三,这本书看不完了。

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:[www.tushu000.com](http://www.tushu000.com)