《Hadoop应用开发技术详解》

图书基本信息

书名：《Hadoop应用开发技术详解》

13位ISBN编号：9787111452447

10位ISBN编号：7111452445

出版时间：2014-1-1

出版社：机械工业出版社

作者：刘刚

页数：408

《Hadoop应用开发技术详解》

内容概要

《大数据技术丛书：Hadoop应用开发技术详解》共12章。第1～2章详细地介绍了Hadoop的生态系统、关键技术以及安装和配置；第3章是MapReduce的使用入门，让读者了解整个开发过程；第4～5章详细讲解了分布式文件系统HDFS和Hadoop的文件I/O；第6章分析了MapReduce的工作原理；第7章讲解了如何利用Eclipse来编译Hadoop的源代码，以及如何对Hadoop应用进行测试和调试；第8～9章细致地讲解了MapReduce的开发方法和高级应用；第10～12章系统地讲解了Hive、HBase和Mahout。

《Hadoop应用开发技术详解》

书籍目录

前言
第1章 Hadoop概述
1.1 Hadoop起源
1.1.1 Google与Hadoop模块
1.1.2 为什么会有Hadoop
1.1.3 Hadoop版本介绍
1.2 Hadoop生态系统
1.3 Hadoop常用项目介绍
1.4 Hadoop在国内的应用
1.5 本章小结
第2章 Hadoop安装
2.1 Hadoop环境安装配置
2.1.1 安装VMware
2.1.2 安装Ubuntu
2.1.3 安装VMwareTools
2.1.4 安装JDK
2.2 Hadoop安装模式
2.2.1 单机安装
2.2.2 伪分布式安装
2.2.3 分布式安装
2.3 如何使用Hadoop
2.3.1 Hadoop的启动与停止
2.3.2 Hadoop配置文件
2.4 本章小结
第3章 MapReduce快速入门
3.1 WordCount实例准备开发环境
3.1.1 使用Eclipse创建一个Java工程
3.1.2 导入Hadoop的JAR文件
3.2 MapReduce代码的实现
3.2.1 编写WordMapper类
3.2.2 编写WordReducer类
3.2.3 编写WordMain驱动类
3.3 打包、部署和运行
3.3.1 打包成JAR文件
3.3.2 部署和运行
3.3.3 测试结果
3.4 本章小结
第4章 Hadoop分布式文件系统详解
4.1 认识HDFS
4.1.1 HDFS的特点
4.1.2 Hadoop文件系统的接口
4.1.3 HDFS的Web服务
4.2 HDFS架构
4.2.1 机架
4.2.2 数据块
4.2.3 元数据节点
4.2.4 数据节点

4.2.5 辅助元数据节点
4.2.6 名字空间
4.2.7 数据复制
4.2.8 块备份原理
4.2.9 机架感知
4.3 Hadoop的RPC机制
4.3.1 RPC的实现流程
4.3.2 RPC的实体模型
4.3.3 文件的读取
4.3.4 文件的写入
4.3.5 文件的一致模型
4.4 HDFS的HA机制
4.4.1 HA集群
4.4.2 HA架构
4.4.3 为什么会有HA机制
4.5 HDFS的Federation机制
4.5.1 单个NameNode的HDFS架构的局限性
4.5.2 为什么引入Federation机制
4.5.3 Federation架构
4.5.4 多个名字空间的管理问题
4.6 Hadoop文件系统的访问
4.6.1 安全模式
4.6.2 HDFS的Shell访问
4.6.3 HDFS处理文件的命令
4.7 JavaAPI接口
4.7.1 HadoopURL读取数据
4.7.2 FileSystem类
4.7.3 FileStatus类
4.7.4 FSDataInputStream类
4.7.5 FSDataOutputStream类
4.7.6 列出HDFS下所有的文件
4.7.7 文件的匹配
4.7.8 PathFilter对象
4.8 维护HDFS
4.8.1 追加数据
4.8.2 并行复制
4.8.3 升级与回滚
4.8.4 添加节点
4.8.5 删除节点
4.9 HDFS权限管理
4.9.1 用户身份
4.9.2 权限管理的原理
4.9.3 设置权限的Shell命令
4.9.4 超级用户
4.9.5 HDFS权限配置参数
4.10 本章小结
第5章 Hadoop文件I/O详解
5.1 Hadoop文件的数据结构
5.1.1 SequenceFile存储

5.1.2 MapFile存储
5.1.3 SequenceFile转换为MapFile
5.2 HDFS数据完整性
5.2.1 校验和
5.2.2 数据块检测程序
5.3 文件序列化
5.3.1 进程间通信对序列化的要求
5.3.2 Hadoop文件的序列化
5.3.3 Writable接口
5.3.4 WritableComparable接口
5.3.5 自定义Writable接口
5.3.6 序列化框架
5.3.7 数据序列化系统Avro
5.4 Hadoop的Writable类型
5.4.1 Writable类的层次结构
5.4.2 Text类型
5.4.3 NullWritable类型
5.4.4 ObjectWritable类型
5.4.5 GenericWritable类型
5.5 文件压缩
5.5.1 Hadoop支持的压缩格式
5.5.2 Hadoop中的编码器和解码器
5.5.3 本地库
5.5.4 可分割压缩LZO
5.5.5 压缩文件性能比较
5.5.6 Snappy压缩
5.5.7 gzip、LZO和Snappy比较
5.6 本章小结
第6章 MapReduce工作原理
6.1 MapReduce的函数式编程概念
6.1.1 列表处理
6.1.2 Mapping数据列表
6.1.3 Reducing数据列表
6.1.4 Mapper和Reducer如何工作
6.1.5 应用实例：词频统计
6.2 MapReduce框架结构
6.2.1 MapReduce模型
6.2.2 MapReduce框架组成
6.3 MapReduce运行原理
6.3.1 作业的提交
6.3.2 作业初始化
6.3.3 任务的分配
6.3.4 任务的执行
6.3.5 进度和状态的更新
6.3.6 MapReduce的进度组成
6.3.7 任务完成
6.4 MapReduce容错
6.4.1 任务失败
6.4.2 TaskTracker失败

6.4.3 JobTracker失败
6.4.4 子任务失败
6.4.5 任务失败反复次数的处理方法
6.5 Shuffle阶段和Sort阶段
6.5.1 Map端的Shuffle
6.5.2 Reduce端的Shuffle
6.5.3 Shuffle过程参数调优
6.6 任务的执行
6.6.1 推测执行
6.6.2 任务JVM重用
6.6.3 跳过坏的记录
6.6.4 任务执行的环境
6.7 作业调度器
6.7.1 先进先出调度器
6.7.2 容量调度器
6.7.3 公平调度器
6.8 自定义Hadoop调度器
6.8.1 Hadoop调度器框架
6.8.2 编写Hadoop调度器
6.9 YARN介绍
6.9.1 异步编程模型
6.9.2 YARN支持的计算框架
6.9.3 YARN架构
6.9.4 YARN工作流程
6.10 本章小结
第7章 Eclipse插件的应用
7.1 编译Hadoop源码
7.1.1 下载Hadoop源码
7.1.2 准备编译环境
7.1.3 编译common组件
7.2 Eclipse安装MapReduce插件
7.2.1 查找MapReduce插件
7.2.2 新建一个Hadooplocation
7.2.3 Hadoop插件操作HDFS
7.2.4 运行MapReduce的驱动类
7.3 MapReduce的Debug调试
7.3.1 进入Debug运行模式
7.3.2 Debug调试具体操作
7.4 单元测试框架MRUnit
7.4.1 认识MRUnit框架
7.4.2 准备测试案例
7.4.3 Mapper单元测试
7.4.4 Reducer单元测试
7.4.5 MapReduce单元测试
7.5 本章小结
第8章 MapReduce编程开发
8.1 WordCount案例分析
8.1.1 MapReduce工作流程
8.1.2 WordCount的Map过程

8.1.3 WordCount的Reduce过程
8.1.4 每个过程产生的结果
8.1.5 Mapper抽象类
8.1.6 Reducer抽象类
8.1.7 MapReduce驱动
8.1.8 MapReduce最小驱动
8.2 输入格式
8.2.1 InputFormat接口
8.2.2 InputSplit类
8.2.3 RecordReader类
8.2.4 应用实例：随机生成100个小数并求最大值
8.3 输出格式
8.3.1 OutputFormat接口
8.3.2 RecordWriter类
8.3.3 应用实例：把首字母相同的单词放到一个文件里
8.4 压缩格式
8.4.1 如何在MapReduce中使用压缩
8.4.2 Map作业输出结果的压缩
8.5 MapReduce优化
8.5.1 Combiner类
8.5.2 Partitioner类
8.5.3 分布式缓存
8.6 辅助类
8.6.1 读取Hadoop配置文件
8.6.2 设置Hadoop的配置文件属性
8.6.3 GenericOptionsParser选项
8.7 Streaming接口
8.7.1 Streaming工作原理
8.7.2 Streaming编程接口参数
8.7.3 作业配置属性
8.7.4 应用实例：抓取网页的标题
8.8 本章小结
第9章 MapReduce高级应用
9.1 计数器
9.1.1 默认计数器
9.1.2 自定义计数器
9.1.3 获取计数器
9.2 MapReduce二次排序
9.2.1 二次排序原理
9.2.2 二次排序的算法流程
9.2.3 代码实现
9.3 MapReduce中的Join算法
9.3.1 Reduce端Join
9.3.2 Map端Join
9.3.3 半连接SemiJoin
9.4 MapReduce从MySQL读写数据
9.4.1 读数据
9.4.2 写数据
9.5 Hadoop系统调优

9.5.1 小文件优化
9.5.2 Map和Reduce个数设置
9.6 本章小结
第10章 数据仓库工具Hive
10.1 认识Hive
10.1.1 Hive工作原理
10.1.2 Hive数据类型
10.1.3 Hive的特点
10.1.4 Hive下载与安装
10.2 Hive架构
10.2.1 Hive用户接口
10.2.2 Hive元数据库
10.2.3 Hive的数据存储
10.2.4 Hive解释器
10.3 Hive文件格式
10.3.1 TextFile格式
10.3.2 SequenceFile格式
10.3.3 RCFile文件格式
10.3.4 自定义文件格式
10.4 Hive操作
10.4.1 表操作
10.4.2 视图操作
10.4.3 索引操作
10.4.4 分区操作
10.4.5 桶操作
10.5 Hive复合类型
10.5.1 Struct类型
10.5.2 Array类型
10.5.3 Map类型
10.6 Hive的JOIN详解
10.6.1 JOIN操作语法
10.6.2 JOIN原理
10.6.3 外部JOIN
10.6.4 Map端JOIN
10.6.5 JOIN中处理NULL值的语义区别
10.7 Hive优化策略
10.7.1 列裁剪
10.7.2 MapJoin操作
10.7.3 GroupBy操作
10.7.4 合并小文件
10.8 Hive内置操作符与函数
10.8.1 字符串函数
10.8.2 集合统计函数
10.8.3 复合类型操作
10.9 Hive用户自定义函数接口
10.9.1 用户自定义函数UDF
10.9.2 用户自定义聚合函数UDAF
10.10 Hive的权限控制
10.10.1 角色的创建和删除

10.10.2 角色的授权和撤销
10.10.3 超级管理员权限
10.11 应用实例：使用JDBC开发Hive程序
10.11.1 准备测试数据
10.11.2 代码实现
10.12 本章小结
第11章 开源数据库HBase
11.1 认识HBase
11.1.1 HBase的特点
11.1.2 HBase访问接口
11.1.3 HBase存储结构
11.1.4 HBase存储格式
11.2 HBase设计
11.2.1 逻辑视图
11.2.2 框架结构及流程
11.2.3 Table和Region的关系
11.2.4 -ROOT-表和.META.表
11.3 关键算法和流程
11.3.1 Region定位
11.3.2 读写过程
11.3.3 Region分配
11.3.4 RegionServer上线和下线
11.3.5 Master上线和下线
11.4 HBase安装
11.4.1 HBase单机安装
11.4.2 HBase分布式安装
11.5 HBase的Shell操作
11.5.1 一般操作
11.5.2 DDL操作
11.5.3 DML操作
11.5.4 HBaseShell脚本
11.6 HBase客户端
11.6.1 JavaAPI交互
11.6.2 MapReduce操作HBase
11.6.3 向HBase中写入数据
11.6.4 读取HBase中的数据
11.6.5 Avro、REST和Thrift接口
11.7 本章小结
第12章 Mahout算法
12.1 Mahout的使用
12.1.1 安装Mahout
12.1.2 运行一个Mahout案例
12.2 Mahout数据表示
12.2.1 偏好Perference类
12.2.2 数据模型DataModel类
12.2.3 Mahout链接MySQL数据库
12.3 认识Taste框架
12.4 Mahout推荐器
12.4.1 基于用户的推荐器

12.4.2 基于项目的推荐器
12.4.3 SlopeOne推荐策略
12.5 推荐系统
12.5.1 个性化推荐
12.5.2 商品推荐系统案例
12.6 本章小结
附录A Hive内置操作符与函数
附录B HBase默认配置解释[1]
附录C Hadoop三个配置文件的参数含义说明

《Hadoop应用开发技术详解》

精彩短评

1、使用版本太低，但对Hadoop常用组件做了从使用到原理的较为全面的介绍。
2、写这本书的人还在智客传播把好想，上有视频的。 两个结合起来更好了。 但是视频不好找啊
3、字数不够，log来凑。。。copy框架图，根本讲不明白。。看不下去的烂，学校竟然拿它做课本！！！(≧０≦)
4、看了一半了，相比晦涩难懂的翻译，这确实算是国产书里面适合入门的hadoop书
5、没有源码坚决不买！

《Hadoop应用开发技术详解》

精彩书评

1、作者很不负责任，我在心里已经骂了你N次了,看一会就想骂，再继续看还想骂。为什么呢？我已经忍着读到第8章了，前面的很多问题不说了，我想问问第200页的public FindMaxValue InputSplit() 的方法，请问有这个FindMaxValue 类型吗？竟然连返回类型也没有，我真靠！写书的时候你能不能测试一下你的代码啊，这样误导多少人啊，我真怀疑平时你怎么写代码的啊。好吧，写错就写错了吧，那我看随书的源代码总行吧，我去华章官网下载，好！我了去，华章书网http://www.hzbook.com 竟然找不到源代码，哥，我彻底死心了，我看不下去了，太痛苦了。前面的内容很多都是东拼西凑的，而且读起来很痛苦，根本就不通顺，这又不是翻译过来的，自己写的怎么就不通顺呢？希望想买此书的读者，先掂量掂量再说吧 ....
2、先说优点：把Hadoop的各个部分都说了一遍，什么HDFS，IO，Map-Reduce等等，而且也有涉及原理的部分。然后。。。。我看这本书的时候，说了无数遍的X了狗了-，-感觉章节安排不合理，本书从刚开始装完Hadoop，写了个WordCount，就开始讲HDFS，IO，对初学者来说，根本不知道讲的啥。更不说章节里面的错误、语句不通顺了。虽然书的出版年比较新（冲着这个买的），但是这本书讲的还是老版的东西，对2.x只是有些许介绍。

《Hadoop应用开发技术详解》

版权说明