第一图书网, tushu007.com

<<Hadoop实战>>

图书基本信息

书名：<<Hadoop实战>>

13位ISBN编号：9787115264480

10位ISBN编号：7115264481

出版时间：2011-10

出版时间：人民邮电

作者：Chuck Lam

译者：韩冀中

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：http://www.tushu007.com

第一图书网, tushu007.com

<<Hadoop实战>>

内容概要

作为云计算所青睐的分布式架构，Hadoop 是一个用Java
语言实现的软件框架，在由大量计算机组成的集群中运行海量数据的分布式计算，是谷歌实现云计算的重要基石。
本书分为3
个部分，深入浅出地介绍了Hadoop 框架、编写和运行Hadoop 数据处理程序所需的实践技能及Hadoop之外更大的生态系统。

本书适合需要处理大量离线数据的云计算程序员、架构师和项目经理阅读参考。

第一图书网, tushu007.com

作者简介

作者：(美国)Chuck Lam 译者：韩冀中Chuck Lam 目前建立了一家名为Rollcall的移动社交网络公司，让活跃的个体用户拥有了一个社交助理。
他曾任RockYou的高级技术组长，开发了社交应用程序和数据处理基础架构，能够支撑上亿的用户。
在斯坦福大学攻读博士学位的时候，Chuck就对大数据产生了兴趣。
他的论文“ Computational Data Acquisition ”吸纳了开源软件和网络游戏等领域的思想，首创了可用于机器学习的数据采集方法。
韩冀中博士，中国科学院计算技术研究所副研究员，研究生导师，长期从事并行分布式计算领域的科研工作，国内早期的Hadoop使用者之一，有着丰富的相关应用开发经验。

第一图书网, tushu007.com

<<Hadoop实战>>

书籍目录

第一部分 Hadoop——一种分布式编程框架
第1章 Hadoop简介
1.1 为什么写《Hadoop 实战》
1.2 什么是Hadoop
1.3 了解分布式系统和Hadoop
1.4 比较SQL 数据库和Hadoop
1.5 理解MapReduce
1.5.1 动手扩展一个简单程序
1.5.2 相同程序在MapReduce中的扩展
1.6 用Hadoop统计单词——运行第一个程序
1.7 Hadoop历史
1.8 小结
1.9 资源
第2章 初识Hadoop
2.1 Hadoop 的构造模块
2.1.1 NameNode
2.1.2 DataNode
2.1.3 Secondary NameNode
2.1.4 JobTracker
2.1.5 TaskTracker
2.2 为Hadoop 集群安装SSH
2.2.1 定义一个公共账号
2.2.2 验证SSH安装
2.2.3 生成SSH密钥对
2.2.4 将公钥分布并登录验证
2.3 运行Hadoop
2.3.1 本地（单机）模式
2.3.2 伪分布模式
2.3.3 全分布模式
2.4 基于Web 的集群用户界面
2.5 小结
第3章 Hadoop组件
3.1 HDFS 文件操作
3.1.1 基本文件命令
3.1.2 编程读写HDFS
3.2 剖析MapReduce 程序
3.2.1 Hadoop数据类型
3.2.2 Mapper
3.2.3 Reducer
3.2.4 Partitioner：重定向Mapper输出
3.2.5 Combiner：本地reduce
3.2.6 预定义mapper和Reducer类的单词计数
3.3 读和写
3.3.1 InputFormat
3.3.2 OutputFormat

第一图书网, tushu007.com

3.4 小结
第二部分 实战
第4章 编写MapReduce基础程序
4.1 获得专利数据集
4.1.1 专利引用数据
4.1.2 专利描述数据
4.2 构建MapReduce 程序的基础模板
4.3 计数
4.4 适应Hadoop API 的改变
4.5 Hadoop 的Streaming
4.5.1 通过Unix命令使用Streaming
4.5.2 通过脚本使用Streaming
4.5.3 用Streaming处理键/值对
4.5.4 通过Aggregate包使用Streaming
4.6 使用combiner 提升性能
4.7 温故知新
4.8 小结
4.9 更多资源
第5章 高阶MapReduce
5.1 链接MapReduce 作业
5.1.1 顺序链接MapReduce作业
5.1.2 具有复杂依赖的MapReduce链接
5.1.3 预处理和后处理阶段的链接
5.2 联结不同来源的数据
5.2.1 Reduce侧的联结
5.2.2 基于DistributedCache的复制联结
5.2.3 半联结：map侧过滤后在reduce侧联结
5.3 创建一个Bloom filter
5.3.1 Bloom filter做了什么
5.3.2 实现一个Bloom filter
5.3.3 Hadoop 0.20 以上版本的Bloom filter
5.4 温故知新
5.5 小结
5.6 更多资源
第6章 编程实践
6.1 开发MapReduce 程序
6.1.1 本地模式
6.1.2 伪分布模式
6.2 生产集群上的监视和调试
6.2.1 计数器
6.2.2 跳过坏记录
6.2.3 用IsolationRunner重新运行出错的任务
6.3 性能调优
6.3.1 通过combiner来减少网络流量
6.3.2 减少输入数据量
6.3.3 使用压缩
6.3.4 重用JVM

第一图书网, tushu007.com

6.3.5 根据猜测执行来运行
6.3.6 代码重构与算法重写
6.4 小结
第7章 细则手册
7.1 向任务传递作业定制的参数
7.2 探查任务特定信息
7.3 划分为多个输出文件
7.4 以数据库作为输入输出
7.5 保持输出的顺序
7.6 小结
第8章 管理Hadoop
8.1 为实际应用设置特定参数值
8.2 系统体检
8.3 权限设置
8.4 配额管理
8.5 启用回收站
8.6 删减DataNode
8.7 增加DataNode
8.8 管理NameNode 和SNN
8.9 恢复失效的NameNode
8.10 感知网络布局和机架的设计
8.11 多用户作业的调度
8.11.1 多个JobTracker
8.11.2 公平调度器
8.12 小结
第三部分 Hadoop也疯狂
第9章 在云上运行Hadoop
9.1 Amazon Web Services 简介
9.2 安装AWS
9.2.1 获得AWS身份认证凭据
9.2.2 获得命令行工具
9.2.3 准备SSH密钥对
9.3 在EC2 上安装Hadoop
9.3.1 配置安全参数
9.3.2 配置集群类型
9.4 在EC2 上运行MapReduce 程序
9.4.1 将代码转移到Hadoop集群上
9.4.2 访问Hadoop集群上的数据
9.5 清空和关闭EC2 实例
9.6 Amazon Elastic MapReduce 和其他AWS 服务
9.6.1 Amazon Elastic MapReduce
9.6.2 AWS导入/导出
9.7 小结
第10章 用Pig编程
10.1 像Pig 一样思考
10.1.1 数据流语言
10.1.2 数据类型

第一图书网, tushu007.com

10.1.3 用户定义函数
10.2 安装Pig
10.3 运行Pig
10.4 通过Grunt 学习Pig Latin
10.5 谈谈Pig Latin
10.5.1 数据类型和schema
10.5.2 表达式和函数
10.5.3 关系型运算符
10.5.4 执行优化
10.6 用户定义函数
10.6.1 使用UDF
10.6.2 编写UDF
10.7 脚本
10.7.1 注释
10.7.2 参数替换
10.7.3 多查询执行
10.8 Pig 实战——计算相似专利的例子
10.9 小结
第11章 Hive及Hadoop群
11.1 Hive
11.1.1 安装与配置Hive
11.1.2 查询的示例
11.1.3 深入HiveQL
11.1.4 Hive小结
11.2 其他Hadoop 相关的部分
11.2.1 HBase
11.2.2 ZooKeeper
11.2.3 Cascading
11.2.4 Cloudera
11.2.5 Katta
11.2.6 CloudBase
11.2.7 Aster Data和Greenplum
11.2.8 Hama和Mahout
11.2.9 search-hadoop.com
11.3 小结
第12章 案例研究
12.1 转换《纽约时报》1100 万个库存图片文档
12.2 挖掘中国移动的数据
12.3 在StumbleUpon 推荐最佳网站
12.3.1 分布式StumbleUpon 的开端
12.3.2 HBase 和StumbleUpon
12.3.3 StumbleUpon 上的更多Hadoop 应用
12.4 搭建面向企业查询的分析系统——IBM的ES2 项目
12.4.1 ES2 系统结构
12.4.2 ES2 爬虫
12.4.3 ES2 分析
12.4.4 小结

第一图书网, tushu007.com

12.4.5 参考文献
附录A HDFS文件命令

第一图书网, tushu007.com

章节摘录

版权页：插图：通常情况下，扩展数据库涉及增加读操作的从节点以及系统的缓存。
只有当你的应用程序读多写少时，增加读操作的从节点才有作用。
如果你的数据集更改并不频繁，缓存才有作用。
即便如此，这些系统结构的特征也总是会在应用层增加巨大的复杂性。
HBase驻留在集群上任何一个机器的每个区域上（每个都是区域服务器）。
写操作涉及托管该区域的区域服务器，而HBase的区域服务器（默认情况下）写入3个HDFS数据节点。
基于一个大表和一个同样大的集群，写操作被分散到很多不同的机器上，从根本上避免了主／从数据存储所具有的单机写瓶颈问题。
这个特征可以帮助你使用传统关系数据库管理系统成本的很小一部分来获得扩展。
随着大型硬件系统与其所提供的实际性能相比越来越昂贵，这是一个影响相当深远而重要的能力。
对于在StumbleUpon的大型工作负载，单从字面上就可能节省数百万美元。
还有一些问题在单机系统中根本无法得到解决！
对于高度动态的数据集，我们经常读取刚刚写入的内容，这样系统中的缓存，如memcached，可能无法提供很多帮助。
HBase在写缓冲区中保存最近写入的数据。
读取的数据直接来自内存。
此操作可以完全避免使用缓存层。
高度动态的数据集的一个例子是事件计数器。
这是一个困难的问题，因为大多数高速解决方案往往是只有利用内存才能满足性能（比如memcached），但又无法满足持久性。
考虑HBase及其incrementColunnValue（）调用。
通过在磁盘中记录日志并缓冲到写缓冲区，读取可以直接来自写缓冲区，达到高性能和高持久性。
StumbleUpon利用HBase的能力来对网站的每个事件进行统计——单击、点击率、广告送达等。
此外，HBase为典型的分区方案提供了绝佳的选择。
大多数传统的分区方法需要对键空间的先验假设。
当散列函数分布不均匀时，或键的分布违背了分区的假设时，就会对性能造成严重的影响。

第一图书网, tushu007.com

媒体关注与评论

“本书是初学者的指路明灯，是高级用户的洞察力之源。
”　　——Philipp K Janert,Principal Value公司“为你全面阐释Hadoop的内容、成因和运行机理。
”　　——Paul Stusiak，Falcon技术公司“将Hadoop阐释清楚的最佳图书l”　　——Rick Wagner，Acxiom公司“全面覆盖Hadoop f他书无而本书有。
”　　——John S Griffin, Overstock.com“本书是对Hadoop和MapReduce的极佳介绍。
，”　　——Kenneth DeLong，BabyCenter公司

第一图书网, tushu007.com

<<Hadoop实战>>

编辑推荐

《Hadoop实战》纵情享受海量数据之美、揭开云计算的神秘面纱、深入分析，追本溯源。
ApacheHadoop是一个NoSQL应用程序框架，在分布式集群中运行，它适合于处理大数据集。
如果需要从数据中分析信息，那么Hadoop是你的最佳选择。
《Hadoop实战》是一本深受读者好评的专著，旨在教会你如何以MapReduce方式编写程序，其中包含MapReduce编程中的最佳实践及设计模式。
书中内容由浅入深，以几个简单的例子开始，继而转向Hadoop在较为复杂的数据分析中的应用.，此外，还介绍了StreamingAPI及Pig和Hive等工具。

第一图书网, tushu007.com

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:http://www.tushu007.com