<<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

<<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。

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。

<<Hadoop实战>>

书籍目录

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。

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

<<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。

版权说明