<<Hadoop权威指南(中文版)>>

图书基本信息

书名：<<Hadoop权威指南(中文版)>>

13位ISBN编号：9787302224242

10位ISBN编号：7302224242

出版时间：2010.5

出版时间：清华大学出版社

作者：(美) Tom White

页数：504

字数：769000

译者：周傲英,曾大聃

前言

马丁·加德纳（数学家和科学作家），曾经在一次采访中说道：“没有微积分，我的生命就失去了意义。
这是我成功的秘诀。
我花了如此长的时间了解我在写什么，所以我知道如何写作才能让大多数读者明白我的意思。”
在许多方面，这就是我对Hadoop的感觉。
它的内部工作机制是复杂的、相互依赖的，因为它运行在分布式系统的理论、实用技术和技术常识这些复杂的基础之上。
对于门外汉来说，Hadoop就像是异形一样难以理解。

但事实上并不是这样的。
剥离其核心，Hadoop提供给组件分布式系统的工具——如数据存储、数据分析和协调——是十分简单的。
如果有一个共同的主题，那么它将与提高抽象水平相关的——为程序员创建用于处理这些事情的基础架构，这些程序员中，或者正好有大量数据需要存储，或者有大量数据需要分析，或者有大量机器需要协调，或者没有时间、技能或兴趣成为分布式系统专家。

借由这样一个简单的、普遍适用的功能组合，在开始使用这个理当被广泛普及的Hadoop的时候，我的想法逐渐清晰起来。
然而，在当时（2006年初），设置、配置和编写程序来使用Hadoop称得上是一门艺术。
幸运的是，此后有了明显的进步，因为有更多的文件，更多的例子，一旦有疑问，还有那么多邮件地址可以发过去帮助你解惑。
但对大多数新手来说，最大的障碍是理解这项技术能做什么，它的长处何在，如何使用它。
这就是我写这本书的原因。

<<Hadoop权威指南(中文版)>>

内容概要

本书从Hadoop的缘起开始，由浅入深，结合理论和实践，全方位地介绍Hadoop这一高性能处理海量数据集的理想工具。
全书共14章，3个附录，涉及的主题包括：Haddoop简介；MapReduce简介；Hadoop分布式文件系统；Hadoop的I／O、MapReduce应用程序开发；MapReduce的工作机制；MapReduce的类型和格式；MapReduce的特性；如何安装Hadoop集群，如何管理Hadoop；Pig简介；Hbase简介；ZooKeeper简介，最后还提供了丰富的案例分析。
本书是Hadoop权威参考，程序员可从中探索如何分析海量数据集，管理员可以从中了解如何安装与运行Hadoop集群。

作者简介

怀特，2007年2月以来，一直担任Apache Hadoop项目负责人。
他是Apache软件基金会的成员之一，同时也是Cloudera的一名工程师。
Tome为IBM的developerWorks撰写过大量文章，并经常在很多行业大会上举行Hadoop主题演讲。
Loudera Cloudera为Hadoop提供商业支持并志愿贡献社区，不收取任何费用。
不管是打算在云中运行Hadoop，还是在自己的服务器上运行Hadoop Cloudera都能使其轻松实现。

<<Hadoop权威指南(中文版)>>

书籍目录

第1章 初识Hadoop第2章 MapReduce简介第3章 Hadoop分布式文件系统第4章 Hadoop的I/O第5章 MapReduce应用开发第6章 MapReduce的工作原理第7章 MapReduce的类型与格式第8章 MapReduce特性第9章 Hadoop集群的安装第10章 Hadoop的管理第11章 Pig简介第12章 Hbase简介第13章 ZooKeeper简介第14章 案例研究附录A Apache Hadoop的安装附录B Cloudera的Hadoop分发包附录C 预备NCDC气象资料

章节摘录

HDFS建立在这样一个思想上：一次写入、多次读取模式是最高效的。
一个数据集通常由数据源生成或复制，接着在此基础上进行各种各样的分析。
每个分析至少都会涉及数据集中的大部分数据（甚至全部），因此读取整个数据集的时间比读取第一条记录的延迟更为重要。
商甬硬件Hadoop不需要运行在昂贵并且高可靠性的硬件上。
它被设计运行在商用硬件（在各种零售店都能买到的普通硬件）的集群上，因此至少对于大的集群来说，节点故障的几率还是较高的。
HDFS在面对这种故障时，被设计为能够继续运行而让用户察觉不到明显的中断。
同时，那些并不适合HDFS的应用也是值得研究的。
在目前，HDFS还不太适用于某些领域，不过日后可能会有所改进。
低延迟数据访问需要低延迟访问数据在毫秒范周内的应用并不适HDFS。
HDFS是为达到高数据吞吐量而优化的，这有可能会以延迟为代价。
目前，对于低延迟访问，HBase（参见第12章）是更好的选择。
大量的小文件名称节A（namenode）存储着文件系统的元数据，因此文件数量的限制也由名称节点的内存量决定。
根据经验，每个文件，索引目录以及块占大约150个字节。
因此，举例来说，如果有一百万个文件，每个文件占一个块，就至少需要300MB的内存。
虽然存储上百万的文件是可行的，十亿或更多的文件就超出目前硬件的能力了。
多用声写入，任意修改文馋HDFS中的文件只有一个写入者，而且写操作总是在文件的末尾。
它不支持多个写入者，或是在文件的任意位置修改。
……

媒体关注与评论

“恭喜您有此良机向大师学习Hadoop，在享用技术本身的同时，您还能领略到大师的睿智及其令人如沐春风的写作风格。
”　　——Hadoop 创始人 Doug Cutting

<<Hadoop权威指南(中文版)>>

版权说明