<<Hadoop权威指南>>

图书基本信息

书名：<<Hadoop权威指南>>

13位ISBN编号：9787564126766

10位ISBN编号：7564126760

出版时间：2011-5

出版时间：东南大学出版社

作者：Tom White

页数：600

前言

据2011年4月圣地亚哥大学公布的报告，2008年全球两千七百万台服务器处理的数据量已达9.57ZB。
如何有效管理和高效处理这些海量数据已成为当前亟待解决的问题。
另外，三大类海量数据——商业数据、科学数据、网页数据——的异构性(结构化数据、半结构化数据以及非结构化数据)又进一步加剧了海量数据处理的难度。
2011年2月出版的《科学》杂志刊登专题“ Special Online Collection: Dealing with Data ”，围绕着目前各类数据量的激增展开讨论，认为海量数据的收集、维护和使用已成为科学研究的主要工作。
对许多学科而言，海量数据处理意味着更严峻的挑战，然而更好地管理和处理这些数据也将会获得意想不到的收获。
关系型数据库系统的研究在数据管理方面积累较多经验。
20世纪70年代，关系模型的提出以及IBM System R 和伯克利Ingres的成功开发，证明了关系型数据库系统处理商业数据的优越性。
20世纪80年代，由此模型派生出的IBM DB2，Sybase SQL Server、Oracle Database等以联机事务处理(OLTP)为主的数据库系统的蓬勃发展，使数据库系统得以充分的商业化。
20世纪90年代，W. H. Inmon提出的整合历史数据，通过在线分析(OLAP)和数据挖掘等方法实现商业规划、决策支持等商业智能服务的数据仓库系统，为数据库系统的应用翻开了崭新的篇章。
然而，面对当下的海量数据，这一近40年历史、一体适用(one size fits all)的数据库系统架构显得老态龙钟，力不从心，逐渐无法应对当前的需求。
自从2003年以来，谷歌陆续发布GFS和MapReduce等高可扩展、高性能的分布式海量数据处理框架，并证明了该框架在处理海量网页数据时的优越性。
该框架实现了更高应用层次的抽象，使用户无需关注复杂的内部工作机制，无需具备丰富的分布式系统知识及开发经验，即可实现大规模分布式系统的部署与海量数据的并行处理。
Apache Hadoop开源项目克隆了这一框架，推出了Hadoop系统。
该系统已受到学术界和工业界的广泛认可和采纳，并孵化出众多子项目(如Pig，Zookeeper和Hive等)，日益形成一个易部署、易开发、功能齐全、性能优良的系统。
华东师范大学海量计算研究所从2006年开始从事海量数据方面的研究，且在集群(288核，40TB存储)上部署了Hadoop系统，并成功完成多项研究。
多年来从事海量数据学术研究和项目实施的相关经历，使得我们对Hadoop系统及其开发有了较深入的理解和认识，并在Hadoop部署、调优和优化等方面积累了丰富的经验。
2010年，Hadoop项目负责人Tom White的《Hadoop权威指南》出版第2版。
这本书内容组织得很好，思路清晰，紧密结合了实际问题。

<<Hadoop权威指南>>

内容概要

揭示了Apache
Hadoop如何为你释放数据的力量。
这本内容全面的书籍展示了如何使用Hadoop架构搭建和维护可靠、可伸缩的分布式系统。
Hadoop架构是MapReduce算法的一种开源应用，是Google开创其帝国的重要基石。
程序员可从中探索如何分析海量数据集，管理员可以了解如何建立与运行Hadoop集群。

《Hadoop权威指南(影印版第2版修订版)》涵盖了Hadoop最近的更新，包括诸如Hive、Sqoop和Avro之类的新特性。
它也提供了案例学习来展示Hadoop如何解决特殊问题。
期待尽情享受你的数据？
这就是你要的书。
本身由Tom
White著。

<<Hadoop权威指南>>

作者简介

Tom White从2007年起就是Apache Hadoop的理事。
他是Apache软件基金会的成员和Cloudera的工程师。
Tom为oreilly．com，java．net~llBM的developerWorks撰文，并为业内会议演讲。

<<Hadoop权威指南>>

书籍目录

Foreword
Preface
1. Meet Hadoop
Data!
Data Storage and Analysis
Comparison with Other Systems
RDBMS
Grid Computing
Volunteer Computing
A Brief History of Hadoop
Apache Hadoop and the Hadoop Ecosystem
2. MapReduce
A Weather Dataset
Data Format
Analyzing the Data with Unix Tools
Analyzing the Data with Hadoop
Map and Reduce
Java MapReduce
Scaling Out
Data Flow
Combiner Functions
Running a Distributed MapReduce Job
Hadoop Streaming
Ruby
Python
Hadoop Pipes
Compiling and Running
3. The Hadoop Distributed Filesystem
The Design of HDFS
HDFS Concepts
Blocks
Namenodes and Datanodes
The Command-Line Interface
Basic Filesystem Operations
Hadoop Filesystems
Interfaces
The Java Interface
Reading Data from a Hadoop URL
Reading Data Using the FileSystem API
Writing Data
Directories
Querying the Filesystem
Deleting Data
Data Flow.
Anatomy of a File Read

Anatomy of a File Write
Coherency Model
Parallel Copying with distcp
Keeping an HDFS Cluster Balanced
Hadoop Archives
Using Hadoop Archives
Limitations
4. Hadoop I/0
Data Integrity
Data Integrity in HDFS
LocalFileSystem
ChecksumFileSystem
Compression
Codecs
Compression and Input Splits
Using Compression in MapReduce
Serialization
The Writable Interface
Writable Classes
Implementing a Custom Writable
Serialization Frameworks
Avro
File-Based Data Structures
SequenceFile
……

章节摘录

版权页：插图：Hadoop起源于Nutch项目。
我们曾尝试构建一个开源的Web搜索引擎，但是始终无法有效地将计算任务分配到多台(也就寥寥几台)计算机上。
直到谷歌公司发表GFS和MapReduce的相关论文，我们的思路才清晰起来。
他们设计的系统已可精准地解决我们在Nutch项目中面临的困境。
因此，我们(两个半天工作制的人)也尝试重建这些系统，将其作为Nutch的一部分。
我们成功地在20多台机器上运行了Nutch。
但是我们很快就意识到，只有在几千台机器上运行Nutch才能够应付Web的超大规模，但这个工作量远远不是两个半天工作制的开发人员能搞定的。
几乎就在那个时候，雅虎公司也对这项技术产生了浓厚的兴趣，并迅速组建了一支开发团队。
我有幸成为其中一员。
我们剥离了Nutch的分布式计算模块，将其称为Hadoop。
在雅虎的帮助下，Hadoop很快就能够真正处理Web数据了。
从2006年起，Tom White就对Hadoop贡献良多。
我很早以前通过他的一篇非常优秀的有关Nutch的论文认识了他，在这篇论文中，他以一种优美的笔调清晰地阐述了深刻的想法。
很快，我发现他开发的软件也同样优美且易于理解。
Tom从一开始就乐于站在用户和项目的角度来考虑问题。
与其他开源程序开发人员不同，Tom不会刻意调整系统使其更加符合他个人的需要，而是尽可能地让所有用户用起来都很方便。
Tom最初专注于如何让Hadoop在亚马逊的EC2和S3服务上运行良好。
之后，他转而解决更为广泛的难题，包括如何提高MapReduce API，如增加强网站，如何设计对象序列化框架，等等。
在所有工作中，Tom都非常精准地阐明了想法。
在很短的时间里，Tom进入了Hadoop委员会，并在不久之后成为Hadoop项目管理委员会的一员。
现在，Tom是一个受人尊敬的Hadoop开发社区的高级成员。
尽管他是这个项目多个技术领域的专家，但他的专长是使Hadoop易于理解和使用。
因此，当我得知Tom有意写一本关于Hadoop的书时，我非常高兴。
是的，谁能够比他更胜任呢？
现在，你们有机会向这位大师学习Hadoop——不单单是技术，也包括一些常识和通俗的笔调。

媒体关注与评论

“有了这本权威指南，读者有机会通过大师的手笔来学习Hadoop——在掌握技术的同时，领略作者的睿智和清晰的文风。
”　　——Hadoop创始人 Doug Cutting于Cloudera

<<Hadoop权威指南>>

编辑推荐

版权说明