

## <<Hadoop权威指南>>

### 图书基本信息

书名：<<Hadoop权威指南>>

13位ISBN编号：9787302257585

10位ISBN编号：7302257582

出版时间：2011-6

出版时间：清华大学出版社

作者：Tom White

页数：600

译者：周敏奇,王晓玲,金澈清,钱卫宁,周傲英

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>

## &lt;&lt;Hadoop权威指南&gt;&gt;

## 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

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16章，3个附录，涉及的主题包括：Hadoop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；如何构建Hadoop集群，如何管理Hadoop；Pig简介；Hbase简介；Hive简介；ZooKeeper简介；开源工具Sqoop，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。

## <<Hadoop权威指南>>

### 作者简介

作者：(美国)怀特 (Tom White) 译者：周敏奇 钱卫宁 金澈清 王晓玲 怀特(Tom White)，从2007年以来，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。

Tom为oreully网、java.net和IBM的developerWorks写过大量文章，并经常在很多行业大会上发表演讲。

## <<Hadoop权威指南>>

### 书籍目录

#### 第1章 初识Hadoop

数据！

数据！

数据存储与分析

与其他系统相比

关系型数据库管理系统

网格计算

志愿计算

1.3.4 Hadoop 发展简史

Apache Hadoop和Hadoop生态圈

#### 第2章 关于MapReduce

一个气象数据集

数据的格式

使用Unix工具进行数据分析

使用Hadoop分析数据

map阶段和reduce阶段

横向扩展

合并函数

运行一个分布式的MapReduce作业

Hadoop的Streaming

Ruby版本

Python版本

Hadoop Pipes

编译运行

#### 第3章 Hadoop分布式文件系统

HDFS的设计

HDFS的概念

数据块

namenode和datanode

命令行接口

基本文件系统操作

Hadoop文件系统

接口

Java接口

从Hadoop URL中读取数据

通过FileSystem API读取数据

写入数据

目录

查询文件系统

删除数据

数据流

文件读取剖析

文件写入剖析

一致模型

## <<Hadoop权威指南>>

- 通过 distcp并行拷贝
- 保持 HDFS 集群的均衡
- Hadoop的归档文件
- 使用Hadoop归档文件
- 不足

### 第4章 Hadoop I/O

- 数据完整性
- HDFS的数据完整性
- LocalFileSystem
- ChecksumFileSystem
- 压缩
- codec
- 压缩和输入切分
- 在MapReduce中使用压缩
- 序列化
- Writable接口
- Writable类
- 实现定制的Writable类型
- 序列化框架
- Avro
- 依据文件的数据结构
- 写入SequenceFile
- MapFile

### 第5章 MapReduce应用开发

- 配置API
- 合并多个源文件
- 可变的扩展
- 配置开发环境
- 配置管理
- 辅助类GenericOptionsParser, Tool和ToolRunner
- 编写单元测试
- mapper
- reducer
- 本地运行测试数据
- 在本地作业运行器上运行作业
- 测试驱动程序
- 在集群上运行
- 打包
- 启动作业
- MapReduce的Web界面
- 获取结果
- 作业调试
- 使用远程调试器
- 作业调优
- 分析任务
- MapReduce的工作流
- 将问题分解成MapReduce作业

## <<Hadoop权威指南>>

运行独立的作业

### 第6章 MapReduce的工作机制

剖析MapReduce作业运行机制

作业的提交

作业的初始化

任务的分配

任务的执行

进度和状态的更新

作业的完成

失败

任务失败

tasktracker失败

jobtracker失败

作业的调度

Fair Scheduler

Capacity Scheduler

shuffle和排序

map端

reduce端

配置的调优

任务的执行

推测式执行

重用JVM

跳过坏记录

任务执行环境

### 第7章 MapReduce的类型与格式

MapReduce的类型

默认的MapReduce作业

输入格式

输入分片与记录

文本输入

二进制输入

多种输入

数据库输入(和输出)

输出格式

文本输出

二进制输出

多个输出

延迟输出

数据库输出

### 第8章 MapReduce的特性

计数器

内置计数器

用户定义的Java计数器

用户定义的Streaming计数器

排序

准备

## <<Hadoop权威指南>>

部分排序

总排序

二次排序

联接

map端联接

reduce端联接

边数据分布

利用JobConf来配置作业

分布式缓存

MapReduce库类

### 第9章 构建Hadoop集群

集群规范

网络拓扑

集群的构建和安装

安装Java

创建Hadoop用户

安装Hadoop

测试安装

SSH配置

Hadoop配置

配置管理

环境设置

Hadoop守护进程的关键属性

Hadoop守护进程的地址和端口

Hadoop的其他属性

创建用户帐号

安全性

Kerberos和Hadoop

委托令牌

其他安全性改进

利用基准测试程序测试Hadoop集群

Hadoop基准测试程序

用户的作业

云上的Hadoop

Amazon EC2上的Hadoop

### 第10章 管理Hadoop

HDFS

永久性数据结构

安全模式

日志审计

工具

监控

日志

度量

Java管理扩展(JMX)

维护

日常管理过程



## <<Hadoop权威指南>>

委任节点和解除节点

升级

### 第11章 Pig简介

安装与运行Pig

执行类型

运行Pig程序

Grunt

Pig Latin编辑器

示例

生成示例

与数据库比较

PigLatin

结构

语句

表达式

1.4.4 类型

模式

函数

用户自定义函数

过滤UDF

计算UDF

加载UDF

数据处理操作

加载和存储数据

过滤数据

分组与连接数据

对数据进行排序

组合和分割数据

Pig实战

并行处理

参数代换

### 第12章 Hive

1.1 安装Hive

1.1.1 Hive外壳环境

1.2 示例

1.3 运行Hive

1.3.1 配置Hive

1.3.2 Hive服务

1.3.3 Metastore

1.4 和传统数据库进行比较

1.4.1 读时模式(Schema on Read)vs.写时模式(Schema on

Write)

1.4.2 更新、事务和索引

1.5 HiveQL

1.5.1 数据类型

1.5.2 操作和函数

1.6 表

## <<Hadoop权威指南>>

1.6.1 托管表(Managed Tables)和外部表(External Tables)

1.6.2 分区(Partitions)和桶(Buckets)

1.6.3 存储格式

1.6.4 导入数据

1.6.5 表的修改

1.6.6 表的丢弃

1.7 查询数据

1.7.1 排序(Sorting)和聚集(Aggregating)

1.7.2 MapReduce脚本

1.7.3 连接

1.7.4 子查询

1.7.5 视图(view)

1.8 用户定义函数(User-Defined Functions)

1.8.1 编写UDF

1.8.2 编写UDAF

### 第13章 HBase

2.1 HBasics

2.1.1 背景

2.2 概念

2.2.1 数据模型的“旋风之旅”

2.2.2 实现

2.3 安装

2.3.1 测试驱动

2.4 客户机

2.4.1 Java

2.4.2 Avro, REST, 以及Thrift

2.5 示例

2.5.1 模式

2.5.2 加载数据

2.5.3 Web查询

2.6 HBase和RDBMS的比较

2.6.1 成功的服务

2.6.2 HBase

2.6.3 实例：HBase在Streamy.com的使用

2.7 Praxis

2.7.1 版本

2.7.2 HDFS

2.7.3 用户接口(UI)

2.7.4 度量(metrics)

2.7.5 模式设计

2.7.6 计数器

2.7.7 批量加载(bulkloading)

### 第14章 ZooKeeper

安装和运行ZooKeeper

示例

ZooKeeper中的组成员关系

创建组

## <<Hadoop权威指南>>

加入组

列出组成员

ZooKeeper服务

数据模型

操作

实现

一致性

会话

状态

使用ZooKeeper来构建应用

配置服务

具有可恢复性的ZooKeeper应用

锁服务

生产环境中的ZooKeeper

可恢复性和性能

配置

### 第15章 开源工具Sqoop

获取Sqoop

一个导入的例子

生成代码

其他序列化系统

深入了解数据库导入

导入控制

导入和一致性

直接模式导入

使用导入的数据

导入的数据与Hive

导入大对象

执行导出

深入了解导出

导出与事务

导出和SequenceFile

### 第16章 实例分析

Hadoop 在Last.fm的应用

Last.fm：社会音乐史上的革命

Hadoop a Last.fm

用Hadoop产生图表

Track Statistics程序

总结

Hadoop和Hive在Facebook的应用

概要介绍

Hadoop a Facebook

假想的使用情况案例

Hive

问题与未来工作计划

Nutch 搜索引擎

背景介绍

## <<Hadoop权威指南>>

数据结构

Nutch系统利用Hadoop进行数据处理的精选实例

总结

Rackspace的日志处理

简史

选择Hadoop

收集和存储

日志的MapReduce模型

关于Cascading

字段、元组和管道

操作

Tap类, Scheme对象和Flow对象

Cascading实战

灵活性

Hadoop和Cascading在ShareThis的应用

总结

在Apache Hadoop上的TB字节数量级排序

使用Pig和Wukong来探索10亿数量级边的网络图

测量社区

每个人都在和我说话: Twitter回复关系图

degree(度)

对称链接

社区提取

附录A 安装Apache Hadoop

先决条件

安装

配置

本机模式

伪分布模式

全分布模式

附录B Cloudera 's Distribution for Hadoop

附录C 准备NCDC天气数据

## &lt;&lt;Hadoop权威指南&gt;&gt;

## 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。

## <<Hadoop权威指南>>

### 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

## <<Hadoop权威指南>>

### 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>