HZ BOOKS
华章科技

内容全面，涵盖Hadoop技术本身以及与其相关的Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等所有子项目

实战性强，为各个知识点精心设计了大量经典的案例，易于理解，可操作性强

實戰

陆嘉恒 著

Hadoop in Action

Hadoop实战

机械工业出版社
China Machine Press

實戰

Hadoop in Action

Hadoop实战

陆嘉恒 著

机械工业出版社
China Machine Press

本书是一本系统且极具实践指导意义的Hadoop工具书和参考书。内容全面，对Hadoop整个技术体系进行了全面的讲解，不仅包括HDFS和MapReduce这两大核心内容，而且还包括Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等与Hadoop相关的子项目的内容。实战性强，为各个知识点精心设计了大量经典的小案例，易于理解，可操作性强。

全书一共18章：第1章全面介绍了Hadoop的概念、优势、项目结构、体系结构，以及它与分布式计算的关系；第2章详细讲解了Hadoop集群的安装和配置，以及常用的日志分析技巧；第3章分析了Hadoop在Yahoo！、eBay、Facebook和百度的应用案例，以及Hadoop平台上海量数据的排序；第4～7章深入地讲解了MapReduce计算模型、MapReduce应用的开发方法、MapReduce的工作机制，同时还列出了多个MapReduce的应用案例，涉及单词计数、数据去重、排序、单表关联和多表关联等内容；第8～11章全面地阐述了Hadoop的I/O操作、HDFS的原理与基本操作，以及Hadoop的各种管理操作，如集群的维护等；第12～17章详细而系统地讲解了Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等所有与Hadoop相关的子项目的原理及使用，以及这些子项目与Hadoop的整合使用；第18章以实例的方式讲解了常用Hadoop插件的使用和Hadoop插件的开发。

本书既适合没有Hadoop基础的初学者系统地学习，又适合有一定Hadoop基础但是缺乏实践经验的读者实践和参考。

图书在版编目（CIP）数据

Hadoop实战 / 陆嘉恒著. —北京：机械工业出版社，2011.9
（云计算技术系列丛书）

ISBN 978-7-111-35944-9

I. H… II. 陆… III. 数据处理－应用软件－网络编程 IV. TP274

中国版本图书馆CIP数据核字（2011）第192844号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）
责任编辑：杨绣国　　陈佳媛
北京京师印务有限公司印刷
2011年10月第1版第1次印刷
186mm×240mm • 28.75印张
标准书号：ISBN 978-7-111-35944-9
定价：69.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991；88361066
购书热线：（010）68326294；88379649；68995259
投稿热线：（010）88379604
读者信箱：hzjsj@hzbook.com

目　录

第1章 Hadoop简介

本章内容

- ❑ 什么是Hadoop
- ❑ Hadoop项目及其结构
- ❑ Hadoop的体系结构
- ❑ Hadoop与分布式开发
- ❑ Hadoop计算模型——MapReduce
- ❑ Hadoop的数据管理
- ❑ 小结

1.1 什么是 Hadoop

1.1.1 Hadoop 概述

Hadoop 是 Apache 软件基金会旗下的一个开源分布式计算平台。以 Hadoop 分布式文件系统（HDFS，Hadoop Distributed Filesystem）和 MapReduce（Google MapReduce 的开源实现）为核心的 Hadoop 为用户提供了系统底层细节透明的分布式基础架构。HDFS 的高容错性、高伸缩性等优点允许用户将 Hadoop 部署在低廉的硬件上，形成分布式系统；MapReduce 分布式编程模型允许用户在不了解分布式系统底层细节的情况下开发并行应用程序。所以用户可以利用 Hadoop 轻松地组织计算机资源，从而搭建自己的分布式计算平台，并且可以充分利用集群的计算和存储能力，完成海量数据的处理。

1.1.2 Hadoop 的历史

Hadoop 的源头是 Apache Nutch，该项目开始于 2002 年，是 Apache Lucene 的子项目之一。2004 年，Google 在“操作系统设计与实现”（OSDI，Operating System Design and Implementation）会议上公开发表了题为“MapReduce：Simplified Data Processing on Large Clusters”（MapReduce：简化大规模集群上的数据处理）的论文，之后受到启发的 Doug Cutting 等人开始尝试实现 MapReduce 计算框架，并将它与 NDFS（Nutch Distributed File System）结合，以支持 Nutch 引擎的主要算法。由于 NDFS 和 MapReduce 在 Nutch 引擎中有着良好的应用，所以它们于 2006 年 2 月被分离出来，成为了一套完整而独立的软件，起名为 Hadoop。到了 2008 年年初，Hadoop 已成为 Apache 的顶级项目，它被包括 Yahoo! 在内的很多互联网公司所采用。现在，Hadoop 已经发展成为包含 HDFS、MapReduce、Pig、ZooKeeper 等子项目的集合，用于分布式计算。

1.1.3 Hadoop 的功能与作用

我们为什么需要 Hadoop 呢？众所周知，现代社会的信息量增长速度极快，这些信息里又积累着大量的数据，其中包括个人数据和工业数据。预计到 2020 年，每年产生的数字信息将会有超过 1/3 的内容驻留在云平台中或借助云平台处理。我们需要对这些数据进行分析和处理，以获取更多有价值的信息。那么我们如何高效地存储和管理这些数据，如何分析这些数据呢？这时可以选用 Hadoop 系统，它在处理这类问题时，采用了分布式存储方式，提高了读写速度，并扩大了存储容量。采用 MapReduce 来整合分布式文件系统上的数据，可保证分析和处理数据的高效。与此同时，Hadoop 还采用存储冗余数据的方式保证了数据的安全性。

Hadoop 中 HDFS 的高容错特性，以及它是基于 Java 语言开发的，这使得 Hadoop 可以部署在低廉的计算机集群中，同时不限于某个操作系统。Hadoop 中 HDFS 的数据管理能力，MapReduce 处理任务时的高效率，以及它的开源特性，使其在同类的分布式系统中大放异

彩，并在众多行业和科研领域中被广泛采用。

1.1.4　Hadoop 的优势

Hadoop 是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发和运行处理海量数据的应用程序。它主要有以下几个优点：

- 高可靠性。Hadoop 按位存储和处理数据的能力值得人们信赖。
- 高扩展性。Hadoop 是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。
- 高效性。Hadoop 能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此其处理速度非常快。
- 高容错性。Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

1.1.5　Hadoop 的应用现状和发展趋势

由于 Hadoop 优势突出，基于 Hadoop 的应用已经遍地开花，尤其是在互联网领域。Yahoo! 通过集群运行 Hadoop，以支持广告系统和 Web 搜索的研究；Facebook 借助集群运行 Hadoop，以支持其数据分析和机器学习；百度则使用 Hadoop 进行搜索日志的分析和网页数据的挖掘工作；淘宝的 Hadoop 系统用于存储并处理电子商务交易的相关数据；中国移动研究院基于 Hadoop 的“大云”（BigCloud）系统用于对数据进行分析和并对外提供服务。

2008 年 2 月，Hadoop 最大贡献者的 Yahoo！构建了当时规模最大的 Hadoop 应用，它们在 2000 个节点上面执行了超过 1 万个 Hadoop 虚拟机器来处理超过 5PB 的网页内容，分析大约 1 兆个网络连接之间的网页索引资料。这些网页索引资料压缩后超过 300TB。Yahoo！正是基于这些为用户提供了高质量的搜索服务。

Hadoop 目前已经取得了非常突出的成绩。随着互联网的发展，新的业务模式还将不断涌现，Hadoop 的应用也会从互联网领域向电信、电子商务、银行、生物制药等领域拓展。相信在未来，Hadoop 将会在更多的领域中扮演幕后英雄，为我们提供更加快捷优质的服务。

1.2　Hadoop 项目及其结构

现在 Hadoop 已经发展成为包含多个子项目的集合。虽然其核心内容是 MapReduce 和 Hadoop 分布式文件系统（HDFS），但 Hadoop 下的 Common、Avro、Chukwa、Hive、HBase 等子项目也是不可或缺的。它们提供了互补性服务或在核心层上提供了更高层的服务。图 1-1 展现了 Hadoop 的项目结构图。

下面将对 Hadoop 的各个子项目进行更详细的介绍。