深入云计算

Hadoop 源代码分析

修订版

全面升级 重装上市！

◎ 张鑫 著

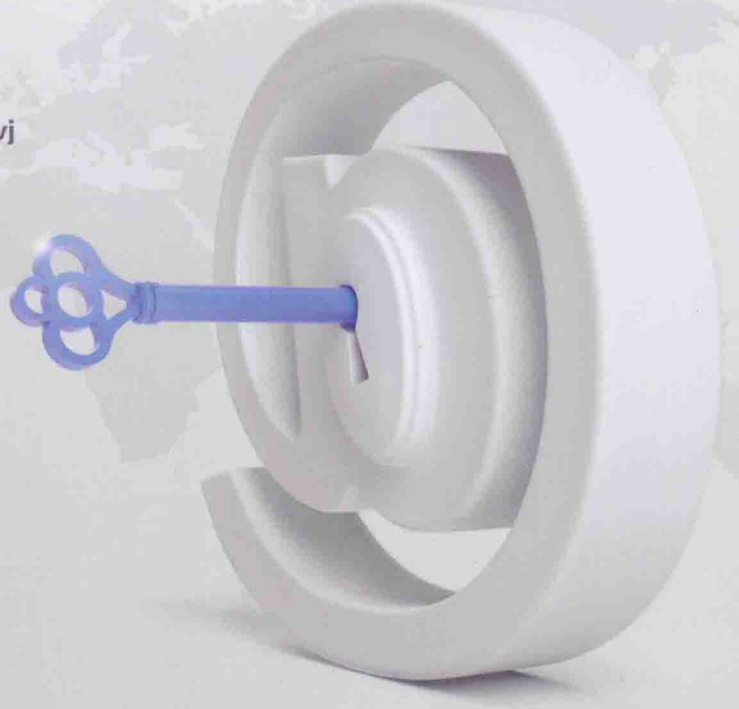
书中源代码下载地址：
http://www.tdpress.com/zyzx/tsscflwj

经典 Hadoop 源码分析图书

笔者悉心整理多年 Hadoop 架构分析与源码学习笔记，凝练丰富的实践开发经验，将辛勤劳动成果倾囊相送，以飨读者。

实践开发经验质的提升

深入解析 Hadoop 各部分的源代码，带领读者透彻理解 Hadoop 的结构和工作机理，深入了解Hadoop的底层工作机制。

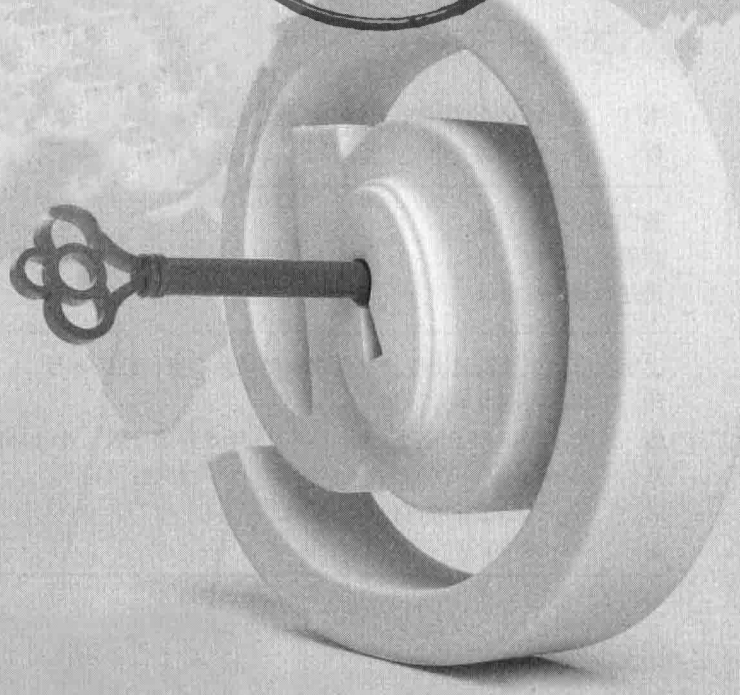


中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

深入云计算

Hadoop 源代码分析（修订版）

◎ 张鑫 著



中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书是一本全面细致介绍和分析 Hadoop 源码和内部工作机理的技术书籍，通过对 Hadoop 内部源码详细透彻的解析，使读者能够快速高效地掌握 Hadoop 的内部工作机制，了解 Hadoop 内部源码架构，对 Hadoop 有更加深刻的认识。

本书主要对 Hadoop 最核心的部分：HDFS 和 MapReduce 进行源码解析和说明。适合所有想全面学习 Hadoop 开发技术的人员阅读，也适用于使用 Hadoop 进行开发的工程技术人员，还可作为想深入了解 Hadoop 运行机制、源代码的开发人员的参考书籍。

图书在版编目（CIP）数据

Hadoop 源代码分析 / 张鑫著. — 2 版（修订本）
. — 北京 : 中国铁道出版社, 2014.8
（深入云计算）
ISBN 978-7-113-18624-1

Ⅰ. ①H… Ⅱ. ①张… Ⅲ. ①数据处理软件 Ⅳ.
①TP274

中国版本图书馆 CIP 数据核字(2014)第 115309 号

书　　名：深入云计算：Hadoop 源代码分析（修订版）
作　　者：张　鑫　著

责任编辑：荆　波　　**读者服务热线：010-63560056**
封面设计：付　巍　　**编辑助理：刘建玮**
责任印制：赵星辰

出版发行：中国铁道出版社（北京市西城区右安门西街 8 号　　邮政编码：100054）
印　　刷：三河市华业印装厂
版　　次：2013 年 6 月第 1 版　　2014 年 8 月第 2 版　　2014 年 8 月第 2 次印刷
开　　本：787mm×1 092mm　1/16　　印张：41　　字数：687 千
书　　号：ISBN 978-7-113-18624-1
定　　价：89.00 元

前　言　Foreword

在如今的数字时代，全球以电子方式存储的数据总量是巨大的。据资料显示，2006 年的数据总量为 0.18 ZB，在 2011 年数据总量则达到 1.8 ZB。而且数据的增长速度还在不断加快，人们已经面临着需要高效快速处理大量数据的问题。

想象一下，如果全球的网页都存储在你的计算机中，你要搜索其中一些网页出来，命令一旦发布，那么你的计算机要多久才能完成搜索任务？可是 Google 只需要不到一秒钟就能完成，这是因为 Google 把这些网页同时分布在不同的计算机上，每台计算机只搜索自己的这部分，一个命令，同时有多台计算机在执行，但是给我们的感觉是在使用一台计算机而已。Hadoop 就是模仿 Google 核心技术而衍生的分布式计算机系统框架。

Hadoop 是由 Apache Software Foundation 公司于 2005 年秋天作为 Lucene 的子项目 Nutch 的一部分正式引入的。它受到最先由 Google Lab 开发的 Map/Reduce 计算模型和 Google File System（GFS）分布式文件系统的启发。2006 年 3 月份，Map/Reduce 和 Nutch Distributed File System（NDFS）分别被纳入称为 Hadoop 的项目中。

Hadoop 是一个能够对大量数据进行分布式处理的软件框架。Hadoop 因具有高可靠性、高扩展性、高效性和高容错性等特性而深受广大用户的欢迎，并且迅速在大数据处理领域占领了一席之地。

编者从事 Hadoop 架构研究很久，对 Hadoop 有深入的理解。本书通过深入细致地解析 Hadoop 各个部分的源代码，带领读者快速高效地将 Hadoop 的结构和工作机理理解透彻，使读者在读完此书后能对 Hadoop 的底层工作机制有深入了解，能对开发 Hadoop 应用程序有质的提高。本书很大程度也是编者在学习 Hadoop 源码时候的笔记，这里将这些成果分享给大家。

本书特色

- **内容全面、系统、深入**

  本书全面介绍了 Hadoop 各个组件的源代码，包括 HDFS、MapReduce 以及 RPC 等部分。此外，每节都配有详细的 UML 模型图和流程图，以便读者对所讲内容有更加清晰的认识。

- **讲解细致，适合各个层次的读者阅读**

  本书从不同的层次结构依次讲解 Hadoop 的源代码，并深入到其结构内部，详细讲解每个部分。内容梯度从易到难，讲解由浅入深，循序渐进，适合各个层次的读者阅读。

- **全面而详细解析 Hadoop 源码**

  本书将 Hadoop 关键核心的代码一一展现给大家，做到浅显易懂，使读者能快速高效地对 Hadoop 有一个深入的理解。

- **提供技术支持，答疑解惑**

  读者阅读本书时若有任何疑问可到网站 http://www.rzchina.net 中的 Hadoop 相关论坛提问，以获得帮助。笔者会及时解答读者的各种问题。

  书中的程序源代码读者可到网站 http://www.tdpress.com/zyzx/tsscflwj 中下载。

本书内容及体系结构

第 1 篇　Hadoop 概述与安装（第 1 章）

本篇主要对 Hadoop 的组成部分进行概述，简要介绍了 HDFS 和 MapReduce 的架构以及 Hadoop 源代码的组织形式。此外，详细介绍了 Hadoop 集群的搭建步骤以及注意事项。通过本篇的学习，可以让读者对 Hadoop 有一个整体上的认识，明确 Hadoop 的地位和作用，并初步搭建起 Hadoop 集群环境。

第 2 篇　HDFS 分布式文件系统及模型（第 2～6 章）

本篇主要是对 Hadoop 的 HDFS 及 IO 进行介绍，详细讲解 HDFS 结构和分布式文件系统、Hadoop IO 相关的各种数据类型，对 HDFS 的重要组件 NameNode 和 DataNode 的源码进行了细致的介绍。通过本篇的学习，读者能很好地掌握 Hadoop HDFS 相关的源码知识和对 Hadoop 数据类型有更深层次的了解，为后续 Hadoop 的开发学习打下坚实的基础。

第 3 篇　MapReduce 计算框架及 RPC 通信模型（第 7～14 章）

本篇主要是对 Hadoop 的核心 MapReduce 框架和 RPC 协议进行介绍，详细讲解：MapReduce 的输入输出，Hadoop 中的 Context 和 ID，MapReduce 的计算模型，JobClient、JobTracker、TaskTracker 的执行过程分析和 Hadoop 的作业调度器，Hadoop RPC 协议，并对 Client 客户端和 Server 服务端进行了详细的分析。通过本篇的学习，读者能够对 Hadoop 有更加深入的了解，理解 Hadoop 的工作机制及各项工作细节，从而全面细致地对 Hadoop 进行认识。

本书读者对象

- 想全面学习 Hadoop 开发技术的人员；
- Hadoop 专业开发人员；
- 利用 Hadoop 做云计算平台的工程技术人员；
- Hadoop 云计算开发爱好者；
- 本科研究生等学习 Hadoop 的学生。

编　者

2014 年 3 月

目 录 Contents

第 1 篇 Hadoop 概述与安装

第 1 章 Hadoop 的简介和安装

第 2 篇 HDFS 分布式文件系统及 IO 模型

第 2 章 HDFS 架构和分布式文件系统

第 10 章 JobClient 的执行过程分析

第 11 章 JobTracker 的执行过程分析

第 12 章 Hadoop 的作业调度器

第 13 章 TaskTracker 的执行过程

第 14 章 Hadoop 的 RPC 协议

第 1 章　Hadoop 的简介和安装

想象一下，如果全球的网页都存储在你的计算机上，你要搜索其中一些网页出来，命令一旦发出，那么你的计算机要多久才能完成搜索任务？可是 Google 只需要不到一秒钟就能完成，这是因为 Google 把这些网页同时分布在不同的计算机上，每台计算机只搜索自己这部分，一个命令，同时有 n 多台计算机在执行，但是给我们的感觉是在用一台电脑而已。Hadoop 就是模仿 Google 核心技术而成的分布式计算机系统框架。

1.1　Hadoop 的简介

在如今正处在海量信息的时代，人们上传视频、拍摄照片、在网站上留言、浏览广告等都使信息容量在飞速增加，而机器在运行时产生的信息量更是巨大。海量的信息使得我们不能像以前一样处理数据了，如何从 TB、PB 级的数据集里快速挖掘出有用的数据，如何在信息的海洋里快速找到我们需要的信息，这就是 Hadoop 所要做的了。

Hadoop 是一个开源的分布式框架，是 Apache 下的一个开源项目。Hadoop 运行可以在成千上万个普通机器的节点组成的集群上，通过分布式的计算模型和存储模型来处理大数据集。Hadoop 具有高容错性、工作在普通的机器节点上扩展性强等众多的优点，是企业选择处理大数据集工具的不二“人”选。

Hadoop 主要包括如下组成部分：

- Hadoop Common：一些支持 Hadoop 其他子项目的通用工具集。
- HDFS：Hadoop 的一个高容错性的分布式文件系统，用于存储数据。
- MapReduce：Hadoop 的一个处理大数据集的分布式计算框架。

1.1.1　分布式文件系统 HDFS

Hadoop 的分布式文件系统 HDFS（Hadoop Distributed File System）是 Hadoop 主要的存储系统（Hadoop

还有一些其他的文件系统，后续会介绍），其对用户来看和其他的文件系统没有什么大的区别：可以创建文件、删除文件、移动文件、重命名文件，但其还是有很多的不同之处。

一个 HDFS 集群主要由 Namenode 和 Datanode 组成。其中 Namenode 只有一个，主要用于管理存储数据的元数据，而 Datanode 可以有多个，主要用于直接存储数据。

HDFS 是被设计为运行在商业计算机上的分布式文件系统，其和现有的许多分布式的文件系统有相似的地方，但是也有着重大的区别：HDFS 是一个高容错的系统，其可以部署在一般的普通商业机器上。HDFS 主要是面向于大数据集的，如 PB、TB 级的数据，其在面对数据可能损害出错时，不是采用使用更好的机器来防止数据出问题，而是提供了一种机制，使得在普通机器节点上的数据损坏出错后也能很好的处理。换句话说，HDFS 是面向一种数据高错率的一种解决方案。

HDFS 的文件模型是一次写入多次读取，一个 HDFS 上的文件一旦被创建，写入了内容，并且关闭了，那么以后则不需要对它再进行更改（但支持对文件进行追加）。这种做法大大提高 HDFS 处理文件的吞吐量，并且使 HDFS 简化。

HDFS 的核心：Namenode 和 Datanodes。HDFS 是使用的 master/slave 的设计架构，即管理者和服从者的架构。每个 HDFS 集群都是由一个 Namenode 和至少一个 Datanode 组成，Namenode 用来管理文件系统的命名空间，负责将文件的注册信息分发给客户端使其能顺利地读取和存储文件，任何对文件的改变也都会被 Namenode 所记录。Datanode 负责直接存储数据，听从 Namenode 的调度安排。

在实际工作中，Namenode 将需要存储的大数据集切分成很多小块，将这些小块的数据根据一定的规则分发到不同的 Namenode 上，并将分块信息保存（成为元数据）以供后续读取文件时将这些小的文件块组织成一个完整的数据文件。在 Datanode 存储小的数据块时，其不只存储一份，会默认根据一定的规则在机器中的不同 Datanode 中存储三份（可由用户指定多少份）数据块，以防止数据发生损坏而造成不可挽回的损失，这也是其具有高容错性的原因。

图 1.1 为 HDFS 的架构图。

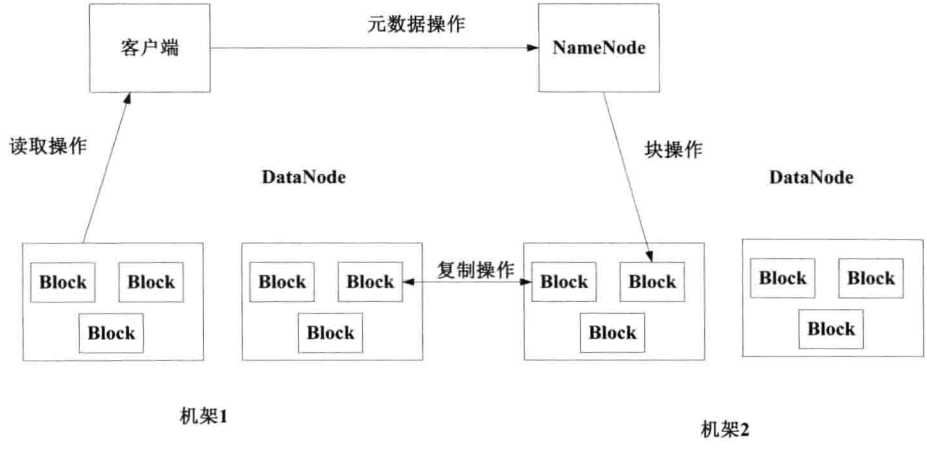


图 1.1　HDFS 的架构图

1.1.2　并行计算模型 MapReduce

Map/Reduce 是 Hadoop 的一个能轻松处理大数据集（TB、PB 数量级）的软件框架，能够轻松运行在成千上万个普通机器的节点上，并且有着很高的容错性。

Map/Reduce 分为 Map 任务和 Reduce 任务：一个 Map/Reduce 任务会首先将输入的数据集切分成独立的小块，并被 Map 任务（map tasks）平行地在 Datanode 上进行处理，实现分布式处理；之后 Map/Reduce 框架会将 Map 任务处理的结果汇总到 Reduce 任务（Reduce tasks）中做进一步汇总处理，最后将最终结果输出。

Hadoop 在处理大数据集的计算时是实现这样一个理念：将计算放在数据节点上，而不是将数据放在计算节点上。所以 Map/Reduce 框架和 HDFS 是运行在相同集群节点上，即计算节点和存储节点是同一个节点。这样就解决了在 Hadoop 进行处理数据时数据的移动所带来的各种问题，使在处理数据时有更高的效率。

Map/Reduce 是由一个 master（即 JobTracke）和一个 slaver（即 TaskTracke）组成。Master 会响应客户端分配的任务，并将任务按照一定的规则分配给 slaver，并且监视 slaver 的执行，如果发生了错误会重新选择新的 slaver 去执行。而 slaver 只负责执行 master 分配来的任务，并在执行完后输入给 Reduce 进行汇总。

Map/Reduce 框架在执行的时候是通过键值对的形式进行操作的。数据集以 Key、Value 的形式输入 Map 任务，Map 任务处理完成之后以 Key、Value 的形式输出给 Redcuce，Reduce 处理完成后将结果输出。其中要注意，Map 的输出是直接给 Reduce 的，所以 Map 输出的 Key、Value 值的类型必须与 Reduce 输入的 Key、Value 值的类型相对应。

1.2　Hadoop 的安装

我们打算在 Windows 平台上首先使用 Vmware Workstation 虚拟机来安装 Ubuntu 系统，然后在 Ubuntu 系统之上搭建 Hadoop 的运行环境。详细的安装配置步骤如下。

1.2.1　虚拟机以及 Ubuntu 的安装

如果我们想在 Windows 系统下进行 Hadoop 的开发和学习的话，可以安装 VMware 虚拟机，并在虚拟机中安装 Ubuntu 操作系统。

（1）Vmware Workstation 的安装

首先进行 Vmware Workstation 虚拟机软件的安装，本书选择的是 8.0.2.591240 版本。

（2）Ubuntu 的安装

本节采用的 Ubuntu 版本是 10.10。在前面安装好的 VM8 上安装 Ubuntu，如下，在虚拟机主界面上选择新建虚拟机选项，其主界面如图 1.2 所示：

图 1.2　新建虚拟机

之后进入虚拟机安装向导，其主界面如图 1.3 所示：

我们选择默认的标准配置类型，因为 VM 可以自动根据我们主机的硬件属性进行相应的配置处理，然后单击“下一步”按钮，进入到图 1.4 所示的主界面：

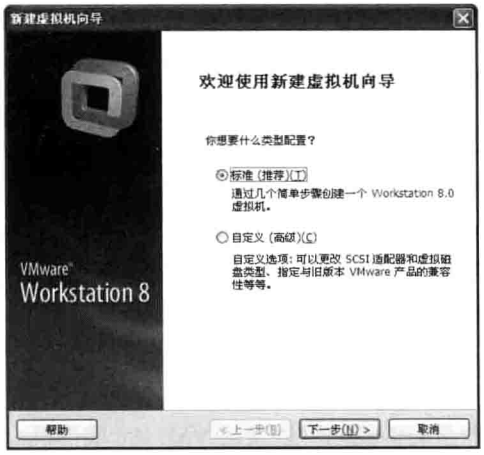


图 1.3　虚拟机安装向导

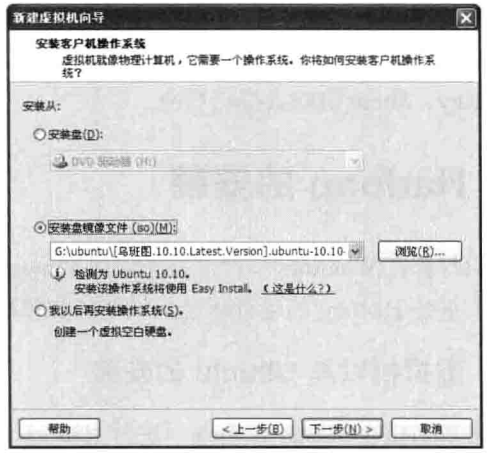


图 1.4　选择镜像文件位置

我们需要在虚拟机上安装 Ubuntu 操作系统，选择 Ubuntu 软件镜像文件，然后单击“下一步”按钮，进入到如图 1.5 所示界面：填写 Ubuntu 的系统名称、用户名和密码。然后单击“下一步”按钮，进入到如图 1.6 所示界面：填写虚拟机的名称，并选择 Ubuntu 系统安装的位置。然后单击“下一步”按钮，进入到如图 1.7 所示界面：指定磁盘容量大小，保存默认的选项即可。然后单击“下一步”按钮，进入到如图 1.8 所示界面：单击“完成”按钮，至此在 VM8 上安装 Ubuntu 操作系统成功完成。