深入云计算

Hadoop 应用开发实战详解

修订版

全面升级 重装上市！

◎ 万川梅　谢正兰　编著

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

精准的内容梯度安排

遵循读者学习习惯和Hadoop技术应用实践，合理安排图书内容；精挑细选经典实例，嵌入完善的代码注释。

精炼的实用经验阐述

作者多年开发经验融入其中，读者在全面掌握Hadoop编程和开发技术的同时更能获得快速分析和解决实际问题的能力。

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

深入云计算

Hadoop 应用开发实战详解（修订版）

◎ 万川梅　谢正兰　编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书由浅入深，全面、系统地介绍了 Hadoop 这一高性能处理大量数据集的理想工具。本书内容主要包括 HDFS、MapReduce、Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa 等与 Hadoop 相关的子项目，各个知识点都配有精心设计的大量经典的小案例，实战性和可操作性强。

本书旨在帮助云计算初学者迅速掌握 Hadoop 系统，提升读者在云计算实践中的应用和开发能力。同时本书极强的系统性和大量翔实的案例对于有一定基础的中高级用户有非常好的参考价值。

图书在版编目（CIP）数据

Hadoop 应用开发实战详解 / 万川梅，谢正兰编著
.—2 版（修订本）. — 北京 : 中国铁道出版社，
2014.8
（深入云计算）
ISBN 978-7-113-18625-8

Ⅰ. ①H… Ⅱ. ①万… ②谢… Ⅲ. ①数据处理软件
Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2014）第 115307 号

书　　名：深入云计算：Hadoop 应用开发实战详解（修订版）
作　　者：万川梅　谢正兰　编著

责任编辑：荆　波　　**读者服务热线：**010-63560056
责任印制：赵星辰　　**特邀编辑：**赵树刚
封面设计：多宝格 • 付　巍

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：三河市宏盛印务有限公司
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：**25.75　　**字数：**603 千
书　　号：ISBN 978-7-113-18625-8
定　　价：59.80 元

前　言　Foreword

为什么需要云计算？可以通过一个简单的案例说明：一台计算机处理一批数据需要 30 小时，比如处理地震预测、天气预报的数据，这样的计算速度实在是太慢了。提升单台计算机的速度是过去的办法，但 CPU 的速度不可能再大幅度提升了。人们一直希望通过增加计算机数量并行运算提升运算和数据处理速度，例如，希望通过同时在 300 台计算机上处理数据，让处理这批数据的速度变成 10 分钟，并且看起来是在一台计算机上处理的。当然，这是一种理想状态。实际上，人们已经开始设计这样的分布式系统，通过把众多的计算机通过集群方式并行同时运行，以此来提高处理的速度。这就是云计算的源起。

在开源云计算系统中，Hadoop 稳居第一。事实上，Hadoop 非常受欢迎，全球已经安装了数以万计的 Hadoop 系统。在诸多的云计算技术中，Hadoop 具有无与伦比的优势，越来越多的公司和组织选择使用 Hadoop 开源项目作为其解决方案。

Hadoop 是 Apache 基金会的开源项目，为开发者提供了一个分布式系统的基础架构，用户可以在不了解分布式系统的底层细节的情况下开发分布式的应用，充分利用集群的强大功能，实现高速运算和存储。Hadoop 项目中包括一个分布式的文件系统 HDFS，一个分布式的并行编程框架 MapReduce，以及包括 Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa 在内的诸多子项目。

本书的特点

1. 结构合理，内容系统全面；叙述翔实，例程丰富

在内容的安排上，根据读者的学习习惯和内容的梯度合理安排，更加适合读者学习。同时，本书有详细的例程，每个例子都经过精挑细选，有很强的针对性。书中的程序都有完整的代码（读者可在 http://tdpress.com/zyzx/tsscflwj 上的下载专区中下载使用），而且代码非常简洁和高效，便于读者学习和调试，读者也可以直接使用这些代码来解决自己的问题。

2. 基础知识和实践并重

本书不仅注重基础知识，而且非常注重实践，让读者快速上手，迅速掌握 Hadoop 知识。

3. 结合实际，实战项目贯穿其中

本书写作时特意给出了大量的实战项目，这些项目的灵活使用，将会让读者事半功倍。同时，为了便于读者高效、直观地学习本书内容，对每章的内容的学习都特意编写了思考与总结。

本书内容体系

全书总共 16 章，分为 3 篇，各篇对应的章节和具体内容如下。

第 1 篇　Hadoop 技术篇，包括 1~10 章，主要介绍了初识 Hadoop、Hadoop 的安装和配置、HDFS 海量存储、初识 MapReduce、分布式开源数据库 HBase、MapReduce 进阶、Hive 数据仓库、Pig 开发应用、Chukwa 数据收集系统、ZooKeeper 开发应用。

第 2 篇　Hadoop 管理和容错篇，包括 11~12 章，主要介绍了 Hadoop 管理和 Hadoop 容错。

第 3 篇　Hadoop 实战篇，包括 13~16 章，主要介绍了综合实战 1：Hadoop 中的数据访问；综合实战 2：一个简单的分布式的 Grep；综合实战 3：打造一个搜索引擎；综合实战 4：移动通信信令监测与查询。

本书读者对象

- Hadoop 初学者。
- 想全面、系统地学习 Hadoop 的人员。
- Hadoop 技术爱好者。
- 利用 Hadoop 进行编程和开发的技术人员。
- 大中专院校的学生和老师。
- 相关培训机构的学员。

由于编者水平有限，加之时间较紧，书中难免存在疏漏和错误之处，敬请读者批评指正，编者也会定期在 www.rzchina.net 进行答疑解惑。

编　者

2014 年 3 月

读者意见反馈表

亲爱的读者：

感谢您对中国铁道出版社的支持，您的建议是我们不断改进工作的信息来源，您的需求是我们不断开拓创新的基础。为了更好地服务读者，出版更多的精品图书，希望您能在百忙之中抽出时间填写这份意见反馈表发给我们。随书纸制表格请在填好后剪下寄到：北京市西城区右安门西街8号中国铁道出版社综合编辑部 荆波 收（邮编：100054）。或者采用传真（010-63549458）方式发送。此外，读者也可以直接通过电子邮件把意见反馈给我们，E-mail地址是：jb@163.jb18803242@yahoo.com.cn。我们将选出意见中肯的热心读者，赠送本社的其他图书作为奖励。同时，我们将充分考虑您的意见和建议，并尽可能地给您满意的答复。谢谢！

所购书名：________________

个人资料：

姓名：__________ 性别：__________ 年龄：__________ 文化程度：__________

职业：__________ 电话：__________ E-mail：__________

通信地址：________________________ 邮编：__________

您是如何得知本书的：

□书店宣传 □网络宣传 □展会促销 □出版社图书目录 □老师指定 □杂志、报纸等的介绍 □别人推荐

□其他（请指明）__________

您从何处得到本书的：

□书店 □邮购 □商场、超市等卖场 □图书销售的网站 □培训学校 □其他

影响您购买本书的因素（可多选）：

□内容实用 □价格合理 □装帧设计精美 □带多媒体教学光盘 □优惠促销 □书评广告 □出版社知名度

□作者名气 □工作、生活和学习的需要 □其他

您对本书封面设计的满意程度：

□很满意 □比较满意 □一般 □不满意 □改进建议

您对本书的总体满意程度：

从文字的角度 □很满意 □比较满意 □一般 □不满意

从技术的角度 □很满意 □比较满意 □一般 □不满意

您希望书中图的比例是多少：

□少量的图片辅以大量的文字 □图文比例相当 □大量的图片辅以少量的文字

您希望本书的定价是多少：

本书最令您满意的是：

1.

2.

您在使用本书时遇到哪些困难：

1.

2.

您希望本书在哪些方面进行改进：

1.

2.

您需要购买哪些方面的图书？对我社现有图书有什么好的建议？

您更喜欢阅读哪些类型和层次的计算机书籍（可多选）？

□入门类 □精通类 □综合类 □问答类 □图解类 □查询手册类 □实例教程类

您在学习计算机的过程中有什么困难？

您的其他要求：

目 录 Contents

第 1 篇 Hadoop 技术篇

第 1 章 初识 Hadoop

第 2 章 Hadoop 的安装和配置

第 6 章 MapReduce 进阶

第 9 章 Chukwa 数据收集系统

第 10 章 ZooKeeper 开发应用

第 2 篇　Hadoop 管理和容错篇

第 11 章　Hadoop 管理

第 12 章　Hadoop 容错

第 3 篇 Hadoop 实战篇

第 13 章 综合实战 1：Hadoop 中的数据库访问

第 14 章 综合实战 2：一个简单的分布式的 Grep

第 15 章 综合实战 3：打造一个搜索引擎

第 16 章　综合实战 4：移动通信信令监测与查询

第 1 篇　Hadoop 技术篇

Hadoop 是 Apache 开源组织的一个分布式计算框架，可以在大量低成本硬件设备组成的集群上运行应用程序，为应用程序提供了一组稳定可靠的接口，旨在构建一个具有高可靠性和良好扩展性的分布式系统。Hadoop 技术篇主要包括 Hadoop 的安装和配置、HDFS 海量存储、初识 MapReduce、分布式开源数据库 Hbase、MapReduce 应用开发、分布式开源数据库 Hbase、Hive 数据仓库、Pig 开发应用、Chukwa 数据收集系统等内容。

第 1 章 初识 Hadoop

云计算是 IT 界的第三次浪潮。在这次浪潮中，各大厂商面临着极大的挑战——他们需要从 TB 乃至 PB 级数据中挖掘出有用的数据，并对这些海量的数据进行更快捷、高效率的处理。于是 IT 厂商推出了自己的云计算平台。Google 的 MapReduce、GFS、BigTable 成为互联网的领头羊，然而它的技术是保密的，Google 公司并没有开源 MapReduce 的实现细节。Amazon 的 AWS、微软的 Azure 和 IBM 的蓝云等也是云计算的典型代表，但它们都是商业性平台，对想要继续研究和发展云计算技术的人员或科研团体来说，无法获得更多的了解。

Hadoop 作为 Apache 基金会资助的开源项目，模仿 Google 的核心技术，是一个分布式系统的基础架构，由 Doug Cutting 带领的团队进行开发。它的出现给研究者带来了希望，是最典型和最常见的云计算平台。未来，Hadoop 将作为一个幕后英雄，应用于越来越多的行业。

1.1 Hadoop 简介

Hadoop 起源于从 2002 年开始的 Apache Nutch，它是 Apache Lucene 的子项目之一。直到 2006 年，Hadoop 才逐渐成为一套完整而独立的软件，并被正式命名，最大支持者是 Yahoo!。2008 年初，Hadoop 开始应用到 Yahoo!以外的很多互联网公司。Hadoop 并不是一个缩写，而是一个虚构的名字，该项目的创建者 Doug Cutting 这样解释 Hadoop 的得名："这个名字是我的孩子给一个棕黄色的大象填充玩具命名的。我的命名标准就是简短，容易发音和拼写，没有太多的意义，并且不会被用于别处。小孩子是这方面的高手。"

1.1.1 Hadoop 的起源

MapReduce 编程思想是由 Google 工程师 Jeffrey Dean 于 2004 年提出来的，与此同时，Google 也发表了 GFS、BigTable 等底层系统以应用 MapReduce 模型。2007 年，Google 公司发布了 *Google's MapReduce*