HZ BOOKS
华章科技

第1版广受好评，第2版基于Hadoop及其相关技术最新版本撰写，从多角度做了全面的修订和补充

不仅详细讲解了新一代的Hadoop技术，而且全面介绍了Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等重要技术，是系统学习Hadoop技术的首选之作

實戰

第2版

陆嘉恒 著

Hadoop in Action Second Edition

Hadoop实战

机械工业出版社
China Machine Press

實戰

第2版

Hadoop in Action Second Edition

Hadoop实战

陆嘉恒 著

机械工业出版社
China Machine Press

本书能满足读者全面学习最新的Hadoop技术及其相关技术（Hive、HBase等）的需求，是一本系统且极具实践指导意义的Hadoop工具书和参考书。第1版上市后广受好评，被誉为学习Hadoop技术的经典著作之一。与第1版相比，第2版技术更新颖，所有技术都针对最新版进行了更新；内容更全面，几乎每一个章节都增加了新内容，而且增加了新的章节；实战性更强，案例更丰富；细节更完美，对第1版中存在的缺陷和不足进行了修正。

本书内容全面，对Hadoop整个技术体系进行了全面的讲解，不仅包括HDFS、MapReduce、YARN等核心内容，而且还包括Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等与Hadoop技术相关的重要内容。实战性强，不仅为各个知识点精心设计了大量经典的小案例，而且还包括Yahoo!等多个大公司的企业级案例，可操作系极强。

全书一共19章：第1~2章首先对Hadoop进行了全方位的宏观介绍，然后介绍了Hadoop在三大主流操作系统平台上的安装与配置方法；第3~6章分别详细讲解了MapReduce计算模型、MapReduce的工作机制、MapReduce应用的开发方法，以及多个精巧的MapReduce应用案例；第7章全面讲解了Hadoop的I/O操作；第8章对YARN进行了介绍；第9章对HDFS进行了详细讲解和分析；第10章细致地讲解了Hadoop的管理；第11~17章对Hadoop大生态系统中的Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa等技术进行了详细的讲解；第18章讲解了Hadoop的各种常用插件，以及Hadoop插件的开发方法；第19章分析了Hadoop在Yahoo!、eBay、百度、Facebook等企业中的应用案例。

图书在版编目（CIP）数据

Hadoop实战 / 陆嘉恒著. —2版. —北京：机械工业出版社，2012.11

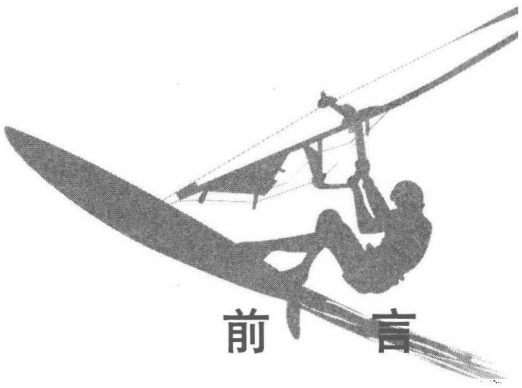
ISBN 978-7-111-39583-6

I. H…　II. 陆…　III. 数据处理－应用软件　IV. TP274

中国版本图书馆CIP数据核字（2012）第208678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：孙海亮
北京市荣盛彩色印刷有限公司印刷
2012年11月第2版第1次印刷
186mm×240mm • 32.25印张
标准书号：ISBN 978-7-111-39583-6
定价：79.00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



为什么写这本书

计算技术已经改变了我们的工作、学习和生活。分布式的云计算技术是当下IT领域最热门的话题之一，它通过整合资源，为降低成本和能源消耗提供了一种简化、集中的计算平台。这种低成本、高扩展、高性能的特点促使其迅速发展，遍地开发，悄然改变着整个行业的面貌。社会各界对云计算的广泛研究和应用无疑证明了这一点：在学术界，政府和很多高校十分重视对云计算技术的研究和投入；在产业界，各大IT公司也在研究和开发相关的云计算产品上投入了大量的资源。这些研究和应用推动与云计算相关的新兴技术和产品不断涌现，传统的信息服务产品向云计算模式转型。

Hadoop作为Apache基金会的开源项目，是云计算研究和应用最具代表性的产品。Hadoop分布式框架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底层细节的情况下开发分布式的应用，充分利用由Hadoop统一起来的集群存储资源、网络资源和计算资源，实现基于海量数据的高速运算和存储。

在编写本书第一版时，鉴于Hadoop技术本身和应用环境较为复杂，入门和实践难度较大，而关于Hadoop的参考资料又非常少，笔者根据自己的实际研究和使用经历，理论与实践并重，从基础出发，为读者全面呈现了Hadoop的相关知识，旨在为Hadoop学习者提供一本工具书。但是时至今日，Hadoop的版本已从本书第一版介绍的0.20升级至正式版1.0，读者的需求也从入门发展到更加深入地了解Hadoop的实现细节，了解Hadoop的更新和发展的趋势，了解Hadoop在企业中的应用。虽然本书第一版受到广大Hadoop学习者的欢迎，

但是为了保持对最新版 Hadoop 的支持，进一步满足读者的需求，继续推动 Hadoop 技术在国内的普及和发展，笔者不惜时间和精力，搜集资料，亲自实践，编写了本书第二版。

第 2 版与第 1 版的区别

基于 Hadoop 1.0 版本和相关项目的最新版，本书在第 1 版的基础上进行了更新和调整：

- 每章都增加了新内容（如第 1 章增加了与 Hadoop 安全相关的知识，第 2 增加了在 Max OS X 系统上安装 Hadoop 的介绍，第 9 章增加了 WebHDFS 等）；
- 部分章节深入剖析了 Hadoop 源码；
- 增加了对 Hadoop 接口及实践方面的介绍（附录 C 和附录 D）；
- 增加了对下一代 MapReduce 的介绍（第 8 章）；
- 将企业应用介绍移到本书最后并更新了内容（第 19 章）；
- 增加了对 Hadoop 安装和代码执行的集中介绍（附录 B）。

本书面向的读者

在编写本书时，笔者力图使不同背景、职业和层次的读者都能从这本书中获益。

如果你是专业技术人员，本书将带领你深入云计算的世界，全面掌握 Hadoop 及其相关技术细节，帮助你使用 Hadoop 技术解决当前面临的问题。

如果你是系统架构人员，本书将成为你搭建 Hadoop 集群、管理集群，并迅速定位和解决问题的工具书。

如果你是高等院校计算机及相关专业的学生，本书将为你在课堂之外了解最新的 IT 技术打开了一扇窗户，帮助你拓宽视野，完善知识结构，为迎接未来的挑战做好知识储备。

在学习本书之前，大家应该具有如下的基础：

- 要有一定的分布式系统的基础知识，对文件系统的基本操作有一定的了解。
- 要有一定的 Linux 操作系统的基础知识。
- 有较好的编程基础和阅读代码的能力，尤其是要能够熟练使用 Java 语言。
- 对数据库、数据仓库、系统监控，以及网络爬虫等知识最好也能有一些了解。

如何阅读本书

从整体内容上讲，本书包括 19 章和 4 个附录。前 10 章、第 18 章、第 19 章和 4 个附录主要介绍了 Hadoop 背景知识、Hadoop 集群安装和代码执行、MapReduce 机制及编程知识、HDFS 实现细节及管理知识、Hadoop 应用。第 11 章至第 17 章结合最新版本详细介绍了与 Hadoop 相关的其他项目，分别为 Hive、HBase、Mahout、Pig、ZooKeeper、Avro、Chukwa，以备读者扩展知识面之用。

在阅读本书时，笔者建议大家先系统地学习Hadoop部分的理论知识（第1章、第3章、第6章至第10章），这样可对Hadoop的核心内容和实现机制有一个很好的理解。在此基础上，读者可进一步学习Hadoop部分的实践知识（第2章、第4章、第5章、第18章、第19章和4个附录），尝试搭建自己的Hadoop集群，编写并运行自己的MapReduce代码。对于本书中关于Hadoop相关项目的介绍，大家可以有选择地学习。在内容的编排上，各章的知识点是相对独立的，是并行的关系，因此大家可以有选择地进行学习。当然，如果时间允许，还是建议大家系统地学习全书的内容，这样能够对Hadoop系统的机制有一个完整而系统的理解，为今后深入地研究和实践Hadoop及云计算技术打下坚实的基础。

另外，笔者希望大家在学习本书时能一边阅读，一边根据书中的指导动手实践，亲自实践本书中所给出的编程范例。例如，先搭建一个自己的云平台，如果条件受限，可以选择伪分布的方式。

在线资源及勘误

在本书的附录中，提供了一个基于Hadoop的云计算在线测试平台（http://cloud-computing.ruc.edu.cn），大家可以先注册一个免费账户，然后即可体验Hadoop平台，通过该平台大家可在线编写MapReduce应用并进行自动验证。如果大家希望获得该平台的验证码，或者希望获得完全编程测试和理论测试的权限，请发邮件到jiahenglu@gmail.com。读者也可访问Hadoop的官方网站（hadoop.apache.org）阅读官方介绍文档，下载学习示例代码。

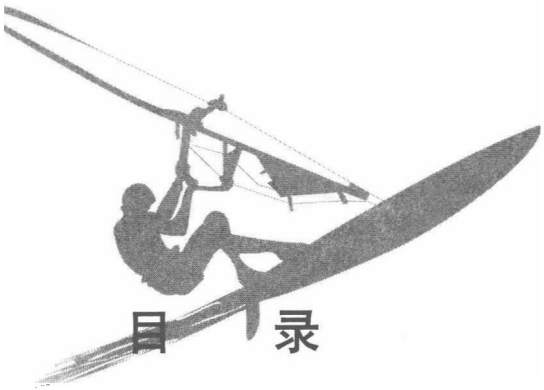
在本书的撰写和相关技术的研究中，尽管笔者投入了大量的精力、付出了艰辛的努力，但是受知识水平所限，书中存在不足和疏漏之处在所难免，恳请大家批评指正。如果有任何问题和建议，可发送电子邮件至jiahenglu@gmail.com或jiahenglu@ruc.edu.cn。

致谢

在本书的编写过程中，很多Hadoop方面的实践者和研究者做了大量的工作，他们是冯博亮、程明、徐文韬、张林林、朱俊良、许翔、陈东伟、谭果、林春彬等，在此表示感谢。

陆嘉恒

2012年6月于北京



目 录