Hadoop

海量数据处理

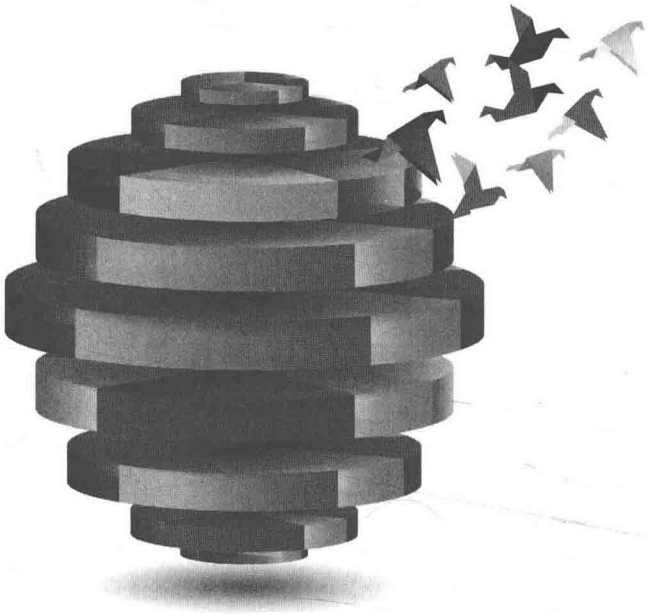
技术详解与项目实战

第2版

范东来／著

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS



Hadoop
海量数据处理
技术详解与项目实战
第2版

范东来／著

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

Hadoop海量数据处理 ：技术详解与项目实战 / 范东来著. -- 2版. -- 北京 ：人民邮电出版社，2016.8
ISBN 978-7-115-42746-5

Ⅰ. ①H… Ⅱ. ①范… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第151385号

内 容 提 要

本书介绍了 Hadoop 技术的相关知识，并将理论知识与实际项目相结合。全书共分为三个部分：基础篇、应用篇和总结篇。基础篇详细介绍了 Hadoop、YARN、MapReduce、HDFS、Hive、Sqoop 和 HBase，并深入探讨了 Hadoop 的运维和调优；应用篇则包含了一个具有代表性的完整的基于 Hadoop 的商业智能系统的设计和实现；结束篇对全书进行总结，并对技术发展做了展望。

本书结构针对学习曲线进行了优化，由浅至深，从理论基础到项目实战，适合 Hadoop 的初学者阅读，也适合作为高等院校相关课程的教学参考书。

◆ 著　　范东来
责任编辑　杨海玲
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：23.25
字数：524 千字　　2016 年 8 月第 2 版
印数：3 801 – 6 800　　2016 年 8 月北京第 1 次印刷

定价：59.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

第 2 版序

作为作者的老师，我很欣喜地看到作者的成长，正如这本书一样。

本书第 1 版问世后，得到了读者认可，市场反响也不错，而且其台湾繁体字版也于 2015 年问世。时间过得真快，我依然记得作者抱着第 1 版初稿到我办公室的那个下午，我们聊了很久。就一两年的时间，大数据技术已发生了巨大的变化，本书第 2 版的出版也就成了顺理成章的事。第 2 版根据最新技术做了全面修订，并新增了 YARN 和 HBase 的章节，更加全面和实用。

大数据本质上是一种思想，代表了数据的深度和广度。读者当然可以从本书学到 Hadoop 相关的技术，但这并不是最重要的，最重要的是能够用大数据的思想来思考和解决问题。

大数据天然地能够和几乎任意行业相结合，如金融、医疗、电商，但是这些都需要行业应用来支撑，需要各位读者来积极投身其中。中国的互联网土壤非常肥沃，这给了大数据非常好的发展基础。目前国内的大数据活跃程度丝毫不亚于国外，在大数据方面，中国很有机会弯道超车，成为世界一流。这是大数据最好的时代，充满了机遇与挑战。

但是，“路漫漫其修远兮”，希望这本书能够帮到更多的人。

北京软件行业协会执行会长
北京航空航天大学软件学院教授
2016 年 6 月，北京

第1版序

这是一本大数据工程师和 Hadoop 工程师的必备书。

近年来，由于移动互联网的高速发展和智能移动设备的普及，数据累积的速率已超过以往任何时候，这个世界已经进入了大数据时代。如何高效地存储、处理这些海量、多种类、高速流动的数据已成为亟待解决的问题。

Hadoop 最早来源于全球云计算技术的领导者谷歌在 2003 年至 2006 年间发表的三篇论文。得益于学术界和工业界的大力支持，Hadoop 目前已成为最为成熟的大数据处理技术。Hadoop 利用了“分而治之”的朴素思想为大数据处理提供了一整套新的解决方案，如分布式文件系统 HDFS、分布式计算框架 MapReduce、NoSQL 数据库 HBase、数据仓库工具 Hive 等。Hadoop 打破了传统数据处理技术的瓶颈，如样本容量、样本种类，让大数据真正成为了生产力。Hadoop 目前已广泛应用于各行各业，行业巨头也纷纷推出自己的基于 Hadoop 的解决方案。今天，Hadoop 已经在电信业、能源业等有了一定的用户基础，传统数据分析架构也逐渐在向 Hadoop 进行过渡。

大数据和大数据处理技术在相互促进，大数据刺激了大数据处理技术的发展，而大数据处理技术又加速了大数据应用落地。大数据催生了一批新的产业，并产生了对 Hadoop 工程师的庞大迫切需求，而目前有关 Hadoop 的书籍和在线材料仍然太少，这更进一步加大了人才缺口。

本书章节安排合理，结构清晰，内容由浅入深，循序渐进。作者是我的学生，作为一个奋战在大数据第一线的工程师，经验非常丰富，能够更加理解并贴近开发者和读者的需求。全书涵盖了 HDFS、MapReduce、Hive、Sqoop 等内容，尤其宝贵的是包含了大量动手实例和一个完备的 Hadoop 项目实例。我相信本书对于希望学习 Hadoop 的读者来说，是一个不错的选择。

北京软件行业协会执行会长
北京航空航天大学软件学院教授、院长
2014 年 12 月，北京

前言

为什么要写这本书

2013 年被称为“大数据元年”，标志着世界正式进入了大数据时代，而就在这一年，我加入了清华大学苏州汽车研究院大数据处理中心，从事 Hadoop 的开发、运维和数据挖掘等方面的工作。从出现之日起，Hadoop 就深刻地改变了人们处理数据的方式。作为一款开源软件，Hadoop 能让所有人享受到大数据红利，让所有人在大数据时代站在了同一起跑线上。Hadoop 很好地诠释了什么是“大道至简，衍化至繁”，Hadoop 来源于非常朴素的思想，但是却衍生出大量的组件，让初学者难以上手。

我在学习和工作的过程中，走过很多弯路也做过很多无用功，尽管这是学习新技术的必由之路，但却浪费了大量的时间。我将自己学习和工作的心得记录下来，为了帮助更多像我当年一样的 Hadoop 学习者，我决定写一本书，一本自己开始 Hadoop 职业生涯的时候也想读到的书。

本书特点有哪些

本书结构针对学习曲线进行了优化，本书由浅至深，从理论基础到项目实战。

本书最大的特点是面向实践。基础篇介绍了 Hadoop 及相关组件，包含了大量动手实例，而应用篇则包含了一个具有代表性的基于 Hadoop 的项目完整实例，该实例脱胎于生产环境的真实项目，在通过基础篇的学习后，读者将在应用篇得到巩固和升华，并对 Hadoop 有一个更加清晰和完整的认识，这也符合实践出真知的规律。

本书介绍了 Hadoop 主要组件，如 HDFS、YARN、MapReduce、Hive、Sqoop、HBase 等，还介绍了 Hadoop 生产环境下的调优和运维、机器学习算法等高级主题。

读者对象是哪些

全书内容由浅入深，既适合初学者入门，也适合有一定基础的技术人员进一步提高技术水平，本书特别适合循序渐进地学习。本书的读者对象包括：

- 准备学习 Hadoop 的开发人员；
- 准备学习 Hadoop 的数据分析师；
- 希望将 Hadoop 运用到实际项目中的开发人员和管理人员；

- 计算机相关专业的高年级本科生和研究生；
- 具有一定的 Hadoop 使用经验，并想进一步提高的使用者。

为什么要写第 2 版

本书第 1 版于 2014 年下半年完稿，次年年初出版，当时 Hadoop 正值急速发展的时期，很多组件都没有实现自己的最终形态。到目前为止，Hadoop 离我完成本书第 1 版之时已有较大变化并已基本稳定。

很多读者在阅读过程中给我来信，说书中使用的 Hadoop 版本较老（CDH3），而且没有 YARN 和 HBase 的内容。收到读者的反馈，加之 HBase 1.0 版已经发布，CDH5 已经基本普及，我就开始本书第 2 版的写作工作，于是呈现在读者面前的就是全新的《Hadoop 海量数据处理》。第 2 版主要做了以下修改和内容增补：

- 全书所有内容根据最新技术进行了修改，并修订了全书；
- 新增了 HDFS 新特性；
- 新增了关于 YARN 的一章；
- 新增了关于 HBase 的一章；
- 新增了 HBase 调优内容；
- 在项目实战中新增了有关 HBase 的内容；
- 根据最新趋势重写了技术展望的内容。

如何阅读本书

本书在章节的安排上旨在引导读者以最快的速度上手 Hadoop，而省去其他不必要的学习过程。如果你是一个有经验的 Hadoop 工程师或者是项目经理，也可以直接进入应用篇，关注项目的设计和实现；如果不是，还是建议你循序渐进地阅读本书方能获得最好的学习效果。

本书一共分为基础篇、应用篇和结束篇 3 个部分，一共 18 章。

基础篇从第 1 章至第 9 章，其中第 1 章为绪论，第 2 章为环境准备，第 3 章至第 8 章主要介绍 HDFS、YARN、MapReduce、Hive、Sqoop、HBase 的原理和使用，在此之上，第 9 章介绍 Hadoop 的性能调优和运维。读者将从基础篇获得 Hadoop 工程师的理论基础。

应用篇从第 9 章至 19 章，主要内容为一个基于 Hadoop 的在线图书销售商业智能系统的设计和实现，包含了系统需求说明、总体设计和完整的实现。应用篇会运用基础篇的知识，巩固并升华基础篇的学习效果。此外，应用篇的项目架构可以进行一些改动并推而广之，有一定的参考价值。读者将从应用篇获得 Hadoop 工程师的项目经验。

结束篇为第 20 章，将对全书进行总结，并对技术发展做了展望。

勘误和支持

写书就像是跳水，高高跳起跃入水中，但在浮出水面之前，运动员却无法知道评委的给分，而我期待读者的评价。由于作者水平有限，编写时间仓促，书中难免会出现一些错误，恳请读者批评指正。读者可以将对本书的反馈和疑问发到 ddna_1022@163.com，我将尽力为读者提供满意的回复。

致谢

感谢电子科技大学的赵勇教授、北京航空航天大学的孙伟教授和邵兵副教授，从您们身上我学到了严谨的学术精神和做人的道理。

感谢清华大学苏州汽车研究院大数据处理中心的林辉主任，您的锐意进取精神一直深留我心。

感谢周俊琨、肖宇、赵虎、李为、黄普、朱游强、熊荣、江彦平，没有你们的帮助和努力，本书不可能完成。

感谢我的父母和外婆这些年来在生活上对我无微不至的关怀和无时无刻的支持，你们辛苦了；感谢吴静宜和她的家人对我的支持；感谢范若云哥哥，是你改变了我。

感谢人民邮电出版社的杨海玲编辑在本书出版过程中给予我的指导和一如既往的信任，感谢庞燕博士为审阅本书第 1 版付出的辛勤劳动。

感谢所有在我求学路上帮助过我的人。

范东来

2016 年 6 月于成都

目 录

基础篇：Hadoop 基础

应用篇：商业智能系统项目实战

结束篇：总结和展望

基础篇：Hadoop 基础

本书的第一部分相当于工具的使用手册，将会介绍 Hadoop 的核心组件：HDFS、YARN、MapReduce、Hive、Sqoop 和 HBase，并在此基础上，进一步学习 Hadoop 性能调优和运维。通过这部分的学习，读者将获得 Hadoop 工程师的理论基础。

第 1 章

绪论

这是最好的时代，这是最坏的时代；这是智慧的时代，这是愚蠢的时代；这是信仰的时期，这是怀疑的时期；这是光明的季节，这是黑暗的季节；这是希望之春，这是失望之冬……

——狄更斯《双城记》

本章作为绪论，目的是在学习 Hadoop 之前，让读者理清相关概念以及这些概念之间的联系。

1.1 Hadoop 和云计算

Hadoop 从问世之日起，就和云计算有着千丝万缕的联系。本节将在介绍 Hadoop 的同时，介绍 Hadoop 和云计算之间的关系，为后面的学习打下基础。

1.1.1 Hadoop 的电梯演讲

如果你是一名创业者或者是一名项目经理，那么最好准备一份“电梯演讲”。所谓电梯演讲，是对自己产品的简单介绍，通常都是 1～2 分钟（电梯从 1 层～30 层的时间），以便如果你恰巧和投资人挤上同一部电梯的时候，能够说服他投资你的项目或者产品。

在做 Hadoop 的电梯演讲之前，先来恶补一下 Hadoop 的有关知识。来看看 Hadoop 的发布者 Apache 软件基金会（ASF）对 Hadoop 的定义：Hadoop 软件库是一个框架，允许在集群中使用简单的编程模型对大规模数据集进行分布式计算。它被设计为可以从单一服务器扩展到数以千计的本地计算和存储的节点，并且 Hadoop 会在应用层面监测和处理错误，而不依靠硬件的高可用性，所以 Hadoop 能够在一个每个节点都有可能出错的集群之上提供一个高可用服务。

从上面的定义可以看出 Hadoop 的如下几个特点。