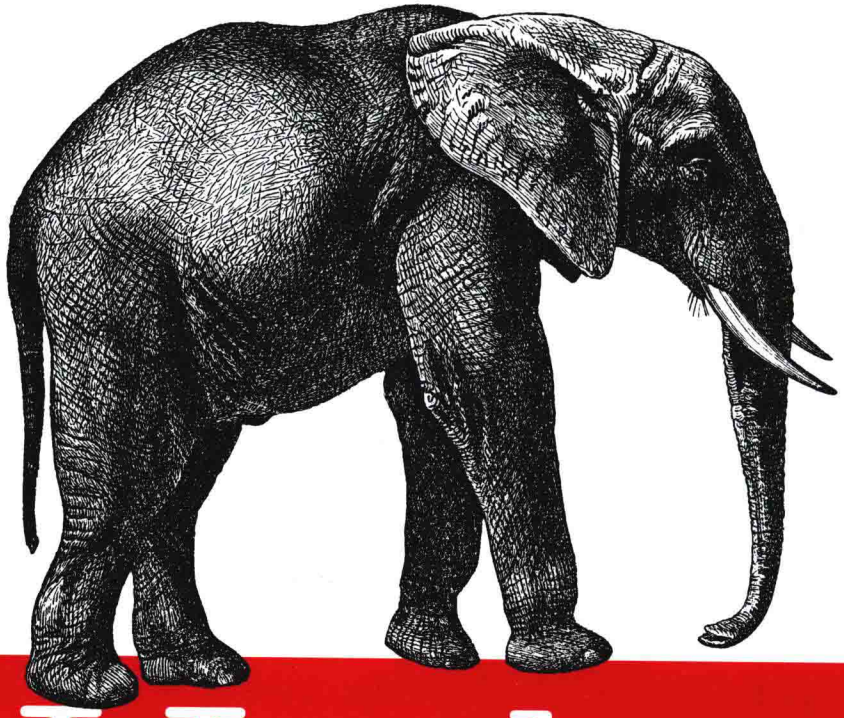
O'REILLY®

第4版
修订版&升级版



Hadoop 权威指南

Hadoop: The Definitive Guide 大数据的存储与分析

Tom White 著
Doug Cutting Hadoop之父 序
王海 华东 刘喻 吕粤海 译

清华大学出版社

Hadoop 权威指南

大数据的存储与分析(第 4 版)

Hadoop: The Definitive Guide Storage and Analysis at Internet Scale

Tom White 著

王 海 华 东 刘 喻 吕粤海 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc. 授权清华大学出版社出版

清华大学出版社

北 京

内 容 简 介

本书结合理论和实践，由浅入深，全方位介绍了 Hadoop 这一高性能的海量数据处理和分析平台。全书 5 部分 24 章，第 I 部分介绍 Hadoop 基础知识，主题涉及 Hadoop、MapReduce、Hadoop 分布式文件系统、YARN、Hadoop 的 I/O 操作。第 II 部分介绍 MapReduce,主题包括 MapReduce 应用开发；MapReduce 的工作机制、MapReduce 的类型与格式、MapReduce 的特性。第III部分介绍 Hadoop 的运维，主题涉及构建 Hadoop 集群、管理 Hadoop。第IV部分介绍 Hadoop 相关开源项目，主题涉及 Avro、Parquet、Flume、Sqoop、Pig、Hive、Crunch、Spark、HBase、ZooKeeper。第 V 部分提供了三个案例，分别来自医疗卫生信息技术服务商塞纳(Cerner)、微软的人工智能项目 ADAM(一种大规模分布式深度学习框架)和开源项目 Cascading(一个新的针对 MapReduce 的数据处理 API)。

本书一本权威、全面的 Hadoop 参考与工具书，阐述了 Hadoop 生态圈的最新发展和应用，程序员可以从中探索海量数据集的存储和分析，管理员可以从中了解 Hadoop 集群的安装和运维。

Copyright © 2016 Tom White. All rights reserved.
Authorized Simplified Chinese translation edition, by O'Reilly Media, Inc., is published by Tsinghua University Press, 2017. Authorized translation of the original English edition, 2012 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.
本书之英文原版由 O'Reilly Media, Inc. 于 2016 年出版。
本书中文简体版由 O'Reilly Media,Inc.授权清华大学出版社 2017 年出版。此翻译版的出版和销售得到出版权和销售权的所有者——O'Reilly Media,Inc.的许可。

版权所有，未经书面许可，本书的任何部分和全部不得以任何形式复制。

北京市版权局著作权合同登记号 图字：**01-2015-2862**
本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：**010-62782989 13701121933**

图书在版编目(CIP)数据

Hadoop 权威指南/(美)汤姆·怀特(Tom White)著；王海，华东，刘喻，吕粤海译. —4 版. —北京：清华大学出版社，2017

书名原文：Hadoop: The Definitive Guide

ISBN 978-7-302-46513-3

Ⅰ. ①H… Ⅱ. ①汤… ②王… ③华… ④刘… ⑤吕… Ⅲ. ①数据处理软件－指南 Ⅳ. ①TP274-62

中国版本图书馆 CIP 数据核字(2017)第 025689 号

责任编辑：文开琪
封面设计：Karen Montgomery 张 健
责任校对：周剑云
责任印制：刘海龙

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社总机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市铭诚印务有限公司
经 销：全国新华书店
开 本：178mm×233mm **印 张**：46 **插 页**：1 **字 数**：594 千字
版 次：2017 年 7 月第 4 版 **印 次**：2017 年 7 月第1 次印刷
定 价：148.00 元

产品编号：063443-01

O'Reilly Media, Inc.介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站(GNN)；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，还是在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——*Wired*

“O'Reilly 凭借一系列(真希望当初我也想到了)非凡想法建立了数百万美元的业务。”

——*Business* 2.0

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是一位特立独行的人，他不光放眼于最长远、最广阔的视野并且切实地按照尤吉·贝拉的建议去做了：‘如果你在路上遇到岔路口，就选择走小路。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

推荐序一

Doug Cutting@加州院内小屋

Hadoop 起源于 Nutch 项目。我们几个人有一段时间一直在尝试构建一个开源的 Web 搜索引擎，但始终无法有效地将计算任务分配到多台计算机上，即使就只是屈指可数的几台。直到谷歌发表 GFS 和 MapReduce 的相关论文之后，我们的思路才清晰起来。他们设计的系统已经可以精准地解决我们在 Nutch 项目中面临的困境。于是，我们(两个半天工作制的人)开始尝试重建这些系统，并将其作为 Nutch 的一部分。

我们终于让 Nutch 可以在 20 台机器上平稳运行，但很快又意识一点：要想应对大规模的 Web 数据计算，还必须得让 Nutch 能在几千台机器上运行，不过这个工作远远不是两个半天工作制的开发人员能够搞定的。

差不多就在那个时候，雅虎也对这项技术产生了浓厚的兴趣并迅速组建了一个开发团队。我有幸成为其中一员。我们剥离出 Nutch 的分布式计算模块，将其称为“Hadoop”。在雅虎的帮助下，Hadoop 很快就能够真正处理海量的 Web 数据了。

从 2006 年起，Tom White 就开始为 Hadoop 做贡献。很早以前，我便通过他的一篇非常优秀的 Nutch 论文认识了他。在这篇论文中，他以一种优美的文风清晰地阐述复杂的思路。很快，我还得知他开发的软件一如他的文笔，优美易懂。

从一开始，Tom 对 Hadoop 所做的贡献就体现出他对用户和项目的关注。与大多数开源贡献者不同，Tom 并没有兴致勃勃地调整系统使其更符合自己个人的需要，而是尽可能地使其方便所有人使用。

最开始，Tom 专攻如何使 Hadoop 在亚马逊的 EC2 和 S3 服务上高效运行。随后，他转向解决更广泛的各种各样的难题，包括如何改进 MapReduce

试读结束：需要全本请在线购买：www.ertongbook.com

API，如何增强网站特色，如何精心构思对象序列化框架，如此等等，不一而举。在所有这些工作中，Tom 都非常清晰、准确地阐明了自己的想法。在很短的时间里，Tom 就赢得大家的认可，拥有 Hadoop 提交者(committer)的权限并很快顺理成章地成为 Hadoop 项目管理委员会的成员。

现在的 Tom，是 Hadoop 开发社区中受人尊敬的资深成员。他精通 Hadoop 项目的若干个技术领域，但他更擅长于 Hadoop 的普及，使其更容易理解和使用。

基于我对 Tom 的这些了解，所以当我得知 Tom 打算写一本 Hadoop 的书之时，别提有多高兴了。是的，谁比他更有资格呢？！现在，你们有机会向这位年青的大师学习 Hadoop，不单单是技术，还有一些必知必会的常识，以及他化繁为简、通俗易懂的写作风格。

推荐序二

周立柱@清华园

在这本《Hadoop 权威指南(第 4 版)》即将出版之际，我十分高兴地再次向广大读者推荐这本书，并期待着它成为我国从事大数据系统研究与开发的科研人员、工程师的一本有价值的参考书。

迄今为止，Hadoop 的发展已经经历了两代，分别为 Hadoop 1.0 和 Hadoop 2.0。与《Hadoop 权威指南(第 3 版)》相比，第 4 版在重点介绍 Hadoop 2.0 的基础上，新增了对当前热门的 Hadoop 技术(如 YARN、Parquet、Flume、Crunch 和 Spark)的专门讲解，有助于 Hadoop 开发者更好地理解相关技术的背景、原理及使用。此外，第 4 版还引入了 Hadoop 在医疗健康领域和分子生物学领域的最新应用成果，并为此新增了相关的实例学习，这对广大 Hadoop 用户而言，具有更好的实践指导意义。

今天，Hadoop 开源项目已经成为研究大数据、开发大数据应用的重要平台，在我国已经形成一个庞大的 Hadoop 用户社群，他们对学习、掌握和提高 Hadoop 提出了很高的需求，《Hadoop 权威指南》系列版本的推出恰好可以满足这样的需要。该书从第 1 版发行以来，历次再版后的畅销也证明了它的用途和价值。

原著的内容组织得当，思路清晰，从原著第 4 版的大幅更新可以看出作者 Tom White 认真、严谨的态度以及对技术的尊重。几位译者在本书翻译过程中，也力求做到清晰、准确和忠实于原著，并为此付出了宝贵的时间和艰辛的劳动。

译者序

自 2006 年面世以来，Hadoop 技术发展迅猛，其技术生态圈也日益壮大，从最初只有 HDFS 和 MapReduce 两个组件，发展到当前的六十多个组件，覆盖了从数据存储、执行引擎到数据访问框架等各个层面。Hadoop 的本地化计算理念、弹性的多层级架构、高效的分布式计算框架，在提供了前所未有的计算能力的同时，也大大降低了计算成本，使其在大规模数据处理分析上的表现远远超过其他产品，不但被广泛应用于各行各业的数据分析和处理，更已成为各大企业数据平台的首选。

伴随着 Hadoop 的发展壮大，本书作者 Tom White 的《Hadoop 权威指南》从 2009 年初版发布至今，也在不断地修订、更新和进一步完善，目前已经推出了第 4 版。该书被业内誉为 Hadoop 圣经，见证了 Hadoop 发展过程中的每一次飞跃。该书的每一新版本都会将 Hadoop 的最新技术和最新应用实践分享给广大读者。

第 4 版完全围绕 Hadoop 2.0 展开描述和讨论。在这一版中，读者可以学习 MapReduce、HDFS 和 YARN 等基础组件的概念和使用步骤；学会设置和维护一个在 YARN 上运行 HDFS 和 MapReduce 的 Hadoop 集群；学习 Avro(用于数据序列化)和 Parquet 两种数据格式(用于嵌套数据)；学习使用数据注入工具，如 Flume(用于序列化数据)和 Sqoop(用于块数据传输)；深入了解高层数据处理工具如 Pig、Hive、Crunch 和 Spark 与 Hadoop 的配合使用；学习 HBase 分布式数据库和 ZooKeeper 分布式配置服务。相应的，作者专门分章介绍 YARN 及几个 Hadoop 相关联的项目如 Parquet、Flume、Crunch 和 Spark。通过第 4 版，读者可以学习到 Hadoop 的最新变化，以及 Hadoop 在健康医疗系统和基因数据处理中的新应用实践。本书非常适用于从事任意规模数据集分析的编程者，对于设置、运行并维护 Hadoop 集群的管理者来说也相当有指导意义。

作为经典书籍，这本书内容丰富、结构清晰，理论与实际结合紧密。2015年，Tom· White 根据 Hadoop 的新版本推出第 4 版，2016 年，在 Hadoop 诞生十年之际，我们非常有幸能够承担本书的翻译工作。希望新版本能够给读者带来更高的技术含量，更好的阅读感受。

全书包含 24 章和 4 个附录。翻译和审校工作由王海教授组织完成。参加翻译工作的有刘喻(第 1 章到第 5 章)、华东(第 6 章到第 11 章)、吕粤海(第 12 章到第 14 章)、张娟(第 15 章到第 17 章)、米志超(第 18 章到第 20 章)、牛大伟(第 21 章到第 23 章)、董超(第 24 章及 4 个附录)。以下老师和同学也对本书的翻译和审校做出了贡献：于卫波、郭晓、李艾静、陈强、刘庆和徐晶。

为确保读者阅读的连贯性及与《Hadoop 权威指南》(第 3 版)的一致性，本书对于术语的翻译处理遵从实践做法；对于新出现的术语，从实践人员的实际使用情况考虑，采取翻译或保留英文名称两种处理方式，希望可以增加读者对新术语的理解。由于译者水平有限，译文中的不当之处也在所难免，真诚地希望广大读者批评与指正。

前言

数学科普作家马丁·加德纳(Martin Gardner)曾经在一次采访中谈到：

> “在我的世界里，只有微积分。这是我的专栏取得成功的奥秘。我花了很多时间才明白如何以大多数读者都能明白的方式将自己所知道的东西娓娓道来。”[①]

这也是我对 Hadoop 的诸多感受。它的内部工作机制非常复杂，是一个集分布式系统理论、实际工程和常识于一体的系统。而且，对门外汉而言，Hadoop 更像是“天外来客”。

但 Hadoop 其实并没有那么让人费解，抽丝剥茧，我们来看看它的“庐山真面目”。Hadoop 提供的用于处理大数据的工具都非常简单。如果说这些工具有一个共同的主题，那就是它们更抽象，为(有大量数据需要存储和分析却没有足够的时间、技能或者不想成为分布式系统专家的)程序员提供一套组件，使其能够利用 Hadoop 来构建一个处理数据的基础平台。

这样一个简单、通用的特性集，促使我在开始使用 Hadoop 时便明显感觉到 Hadoop 真的值得推广。但最开始的时候(2006 年初)，安装、配置和 Hadoop 应用编程是一门高深的艺术。之后，情况确实有所改善：文档增多了；示例增多了；碰到问题时，可以向大量活跃的邮件列表发邮件求助。对新手而言，最大的障碍是理解 Hadoop 有哪些能耐，它擅长什么，它如何使用。这些问题使我萌发了写作本书的动机。

Apache Hadoop 社区的发展来之不易。从本书的第 1 版发行以来，Hadoop 项目如雨后春笋般发展兴旺。“大数据”已成为大家耳熟能详的名词术

① 摘自“The science of fun”，网址为 *http://bit.ly/science_of_fun*。此文 2008 年 5 月 31 日发表于《卫报》。

语。[2]当前，软件在可用性、性能、可靠性、可扩展性和可管理性方面都实现了巨大的飞跃。在 Hadoop 平台上搭建和运行的应用增长迅猛。事实上，对任何一个人来说，跟踪这些发展动向都很困难。但为了让更多的人采用 Hadoop，我认为我们要让 Hadoop 更好用。这需要创建更多新的工具，集成更多的系统，创建新的增强型 API。我希望自己能够参与，同时也希望本书能够鼓励并吸引其他人也参与 Hadoop 项目。

说明

在文中讨论特定的 Java 类时，我常常会忽略包的名称以免啰嗦杂乱。如果想知道一个类在哪个包内，可以查阅 Hadoop 或相关项目的 Java API 文档(Apache Hadoop 主页 *http://hadoop.apache.org* 上有链接可以访问)。如果使用 IDE 编程，其自动补全机制(也称“自动完成机制”)能够帮助你找到你需要的东西。

与此类似，尽管偏离传统的编码规范，但如果要导入同一个包的多个类，程序可以使用星号通配符来节省空间(例如 `import org.apache.hadoop.io.*`)。

本书中的示例代码可以从本书网站下载，网址为 *http://www.hadoopbook.com/*。可以根据网页上的指示获取本书示例所用的数据集以及运行本书示例的详细说明、更新链接、额外的资源与我的博客。

第 4 版新增内容

第 4 版的主题是 Hadoop 2。Hadoop 2 系列发行版本是当前应用最活跃的系列，且包含 Hadoop 的最稳定的版本。

第 4 版新增的章节包括 YARN(第 4 章)、Parquet(第 13 章)、Flume(第 14 章)、Crunch(第 18 章)和 Spark(第 19 章)。此外，为了帮助读者更方便地阅读本书，第 1 章新增了一节“本书包含的内容”(参见 1.7 节)。

② 术语“大数据”在 2013 年被收入《牛津英语辞典》(Oxford English Dictionary)，网址为 *http://bit.ly/6_13_oed_update*。

试读结束：需要全本请在线购买：www.ertongbook.com

第 4 版包括两个新的实例学习(第 22 章和第 23 章)：一个是关于 Hadoop 如何应用于医疗健康系统，另一个是关于将 Hadoop 技术如何应用于基因数据处理。旧版本中的实例学习可以在线查到，网址为 *http:/bit.ly/hadoop_tdg_prev*。

为了和 Hadoop 最新发行版本及其相关项目同步，第 4 版对原有章节进行了修订、更新和优化。

第 3 版新增内容

第 3 版概述 Apache Hadoop 1.x(以前的 0.20)系列发行版本，以及新近的 0.22 和 2.x(以前的 0.23)系列。除了少部分(文中有说明)例外，本书包含的所有范例都在这些版本上运行过。

第 3 版的大部分范例代码都使用了新的 MapReduce API。因为旧的 API 仍然应用很广，所以文中在讨论新的 API 时我们还会继续讨论它，使用旧 API 的对应范例代码可以到本书的配套网站下载。

Hadoop 2.0 最主要的变化是新增的 MapReduce 运行时 MapReduce 2，它建立在一个新的分布式资源管理系统之上，该系统称为 YARN。针对建立在 YARN 之上的 MapReduce，第 3 版增加了相关的介绍，包括它的工作机制(第 7 章)及如何运行(第 10 章)。

第 3 版还增加了更多对 MapReduce 的介绍，包括丰富的开发实践，比如用 Maven 打包 MapReduce 作业，设置用户的 Java 类路径，用 MRUnit 写测试等(这些内容都请参见第 6 章)。第 3 版还深入介绍了一些特性，如输出 committer 和分布式缓存(第 9 章)，任务内存监控(第 10 章)。第 3 版还新增了两小节内容，一节是关于如何写 MapReduce 作业来处理 Avro 数据(参见第 12 章)，另一节是关于如何在 Oozie 中运行一个简单的 MapReduce 工作流(参见第 6 章)。

关于 HDFS 的章节(第 3 章)，新增了对高可用性、联邦 HDFS、新的 WebHDFS 和 HttpFS 文件系统的介绍。

对 Pig，Hive，Sqoop 和 ZooKeeper 的相关介绍，第 3 版全部进行了相应的扩展，广泛介绍其最新发行版本中的新特性和变化。

此外，第 3 版还对第 2 版进行了彻底的更新、修订和优化。

第 2 版新增内容

《Hadoop 权威指南》(第 2 版)新增两章内容，分别介绍 Sqoop 和 Hive(第 15 章和第 17 章)，新增一个小节专门介绍 Avro(参见第 12 章)，补充了关于 Hadoop 新增安全特性的介绍(参见第 10 章)以及一个介绍如何使用 Hadoop 来分析海量网络图的新实例分析。

第 2 版继续介绍 Apache Hadoop 0.20 系列发行版本，因为当时最新、最稳定的发行版本。书中有时会提到一些最新发行版本中的一些新特性，但在首次介绍这些特性时，有说明具体的 Hadoop 版本号。

本书采用的约定

本书采用以下排版约定。

斜体

用于表明新的术语、URL、电子邮件地址、文件名和文件扩展名。

等宽字体 Consolas

用于程序清单，在正文段落中出现的程序元素(如变量或函数名)、数据库、数据类型、环境变量、语句和关键字也采用这样的字体。

等宽字体 Consolas+加粗

用于显示命令或应该由用户键入的其他文本。

等宽字体 Consolas+斜体

表明这里的文本需要替换为用户提供的值或其他由上下文确定的值。

这个图标表示通用的说明。

这个图标表示重要的指示或建议。

这个图标表示警告或需要注意的问题。

示例代码的使用

本书的补充材料（代码、示例及练习等）可以从本书网站(*http://www.hadoopbook.com*)或 GitHub(*https://github.com/tomwhite/hadoop-book/*)下载。

本书的目的是帮助读者完成工作。通常情况下，可以在你的程序或文档中使用本书中给出的代码。不必联系我们获得代码使用授权，除非你需要使用大量的代码。例如，在写程序的时候引用几段代码不需要向我们申请许可。但以光盘方式销售或重新发行 O'Reilly 书中的示例的确需要获得许可。引用本书或引用本书中的示例代码来回答问题也不需要申请许可。但是，如果要将本书中的大量范例代码加入你的产品文档，则需要申请许可。

我们欣赏你在引用时注明出处，但不强求。引用通常包括书名、作者、出版社和 ISBN，如“*Hadoop*: *The Definitive Guide*, Fourth Edition, by Tom White(O'Reilly).Copyright © 2015 Tom White，978-1-491-90163-2”。

如果觉得使用示例代码的情况不属于前面列出的合理使用或许可范围，请通过电子邮件联系我们，邮箱地址为 *permissions@oreilly.com*。

Safari Books Online

Safari Books Online(*www.safaribooksonline.com*)是一个按需定制的数字图书馆，以图书和视频的形式提供全球技术领域和经管领域内知名作者的专业作品。

专业技术人员、软件开发人员、网页设计人员、商务人员和创意专家将 Safari Books Online 用作自己开展研究、解决问题、学习和完成资格认证培训的重要来源。

Safari Books Online 为企业、政府部门、教育机构和个人提供广泛、灵活的计划和定价。

在这里，成员们通过一个可以全文检索的数据库中就能够访问数千种图书、培训视频和正式出版之前的书稿，这些内容提供商有 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 及其他上百家出版社。欢迎访问 Safari Books Online，了解更多详情。

联系我们

对于本书，如果有任何意见或疑问，请通过以下地址联系出版商：

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室(100035)
奥莱利技术咨询(北京)有限公司

本书也有相关的网页，我们在上面列出了勘误表、范例以及其他一些信息。网址如下：

http://bit.ly/hadoop_tdg_4e(英文版)
http://www.oreilly.com.cn/book.php?bn=978-7-302-46513-3(中文版)

对本书做出评论或者询问技术问题，请发送 E-mail 至以下邮箱：

bookquestions@oreilly.com

如果希望获得关于本书、会议、资源中心和 O'Reilly 的更多信息，请访问以下网址：

http://www.oreilly.com
http://www.oreilly.com.cn

致谢

在本书写作期间，我仰赖于许多人的帮助，直接的或间接的。感谢 Hadoop 社区，我从中学到很多，这样的学习仍将继续。

特别感谢 Michael Stack 和 Jonathan Gray，HBase 这一章的内容就是他们写的。我还要感谢 Adrian Woodhead，Marc de Palol，Joydeep Sen Sarma，Ashish Thusoo，Andrzej Białecki，Stu Hood，Chris K. Wensel 和 Owen O'Malley，他们提供了学习实例。

感谢为草稿提出有用建议和改进建议的评审人：Raghu Angadi，Matt Biddulph，Christophe Bisciglia，Ryan Cox，Devaraj Das，Alex Dorman，Chris Douglas，Alan Gates，Lars George，Patrick Hunt，Aaron Kimball，Peter Krey，Hairong Kuang，Simon Maxen，Olga Natkovich，Benjamin Reed，Konstantin Shvachko，Allen Wittenauer，Matei Zaharia 和 Philip Zeyliger。Ajay Anand 组织本书的评审并使其顺利完成。Philip（“flip”）Komer 帮助我获得了 NCDC 气温数据，使本书示例很有特色。特别感谢 Owen O'Malley 和 Arun C. Murthy，他们为我清楚解释了 MapReduce 中 shuffle 的复杂过程。当然，如果有任何错误，得归咎于我。

对于第 2 版，我特别感谢 Jeff Bean，Doug Cutting，Glynn Durham，Alan Gates，Jeff Hammerbacher，Alex Kozlov，Ken Krugler，Jimmy Lin，Todd Lipcon，Sarah Sproehnle，Vinithra Varadharajan 和 Ian Wrigley，感谢他们仔细审阅本书，并提出宝贵的建议，同时也感谢对本书第 1 版提出勘误建议的读者。我也想感谢 Aaron Kimball 对 Sqoop 所做的贡献和 Philip（“flip”）Kromer 对图处理实例分析所做的贡献。

对于第 3 版，我想感谢 Alejandro Abdelnur，Eva Andreasson，Eli Collins，Doug Cutting，Patrick Hunt，Aaron Kimball，Aaron T. Myers，Brock Noland，Arvind Prabhakar，Ahmed Radwan 和 Tom Wheeler，感谢他们的反馈意见和建议。Rob Weltman 友善地对整本书提出了非常详细的反馈意见，这些意见和建议使得本书终稿的质量得以更上一层楼。此外，我还要向提交第 2 版勘误的所有读者表达最真挚的谢意。

对于第 4 版，我想感谢 Jodok Batlogg，Meghan Blanchette，Ryan Blue，Jarek Jarcec Cecho，Jules Damji，Dennis Dawson，Matthew Gast，Karthik Kambatla，Julien Le Dem，Brock Noland，Sandy Ryza，Akshai Sarma，Ben Spivey，Michael Stack，Kate Ting，Josh Walter，Josh Wills 和 Adrian Woodhead，感谢他们所有人非常宝贵的审阅反馈。Ryan Brush，Micah Whitacre 和 Matt Massie kindly 为第 4 版友情提供新的实例学习。再次感谢提交勘误的所有读者。

特别感谢 Doug Cutting 对我的鼓励、支持、友谊以及他为本书所写的序。

我还要感谢在本书写作期间以对话和邮件方式进行交流的其他人。

在本书第 1 版写到一半的时候，我加入了 Cloudera，我想感谢我的同事，他们为我提供了大量的帮助和支持，使我有充足的时间好好写书，并能及时交稿。

非常感谢我的编辑 Mike Loukides、Meghan Blanchette 及其 O'Reilly Media 的同事，他们在本书的准备阶段为我提供了很多帮助。Mike 和 Meghan 一直为我答疑解惑、审读我的初稿并帮助我如期完稿。

最后，写作是一项艰巨的任务，如果没有家人一如既往地支持，我是不可能完成这本的。我的妻子 Eliane，她不仅操持着整个家庭，还协助我，参与本书的审稿、编辑和跟进案例学习。还有我的女儿 Emilia 和 Lottie，她们一直都非常理解并支持我的工作，我期待有更多时间好好陪陪她们。

试读结束：需要全本请在线购买：www.ertongbook.com