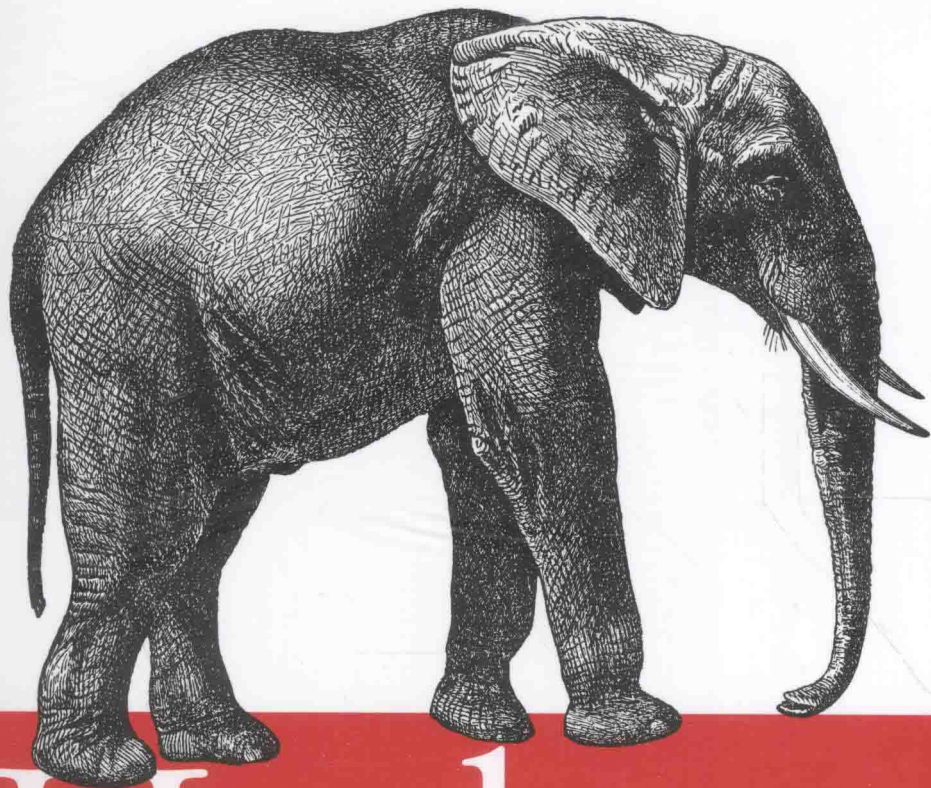


Hadoop: The Definitive Guide

谁说大象不能跳舞?!

——挑战互联网规模的数据存储与分析!

第3版
修订版



Hadoop

权威指南

Tom White 著

Hadoop之父 Doug Cutting 序

华东师范大学数据科学与工程学院 译

O'REILLY®



清华大学出版社

Hadoop 权威指南

(第 3 版)

(美)Tom White 著

华东师范大学数据科学与工程学院 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

本书从 Hadoop 的缘起开始,由浅入深,结合理论和实践,全方位地介绍 Hadoop 这一高性能处理海量数据集的理想工具。全书共 16 章,3 个附录,涉及的主题包括: Hadoop; MapReduce; Hadoop 分布式文件系统; Hadoop 的 I/O、MapReduce 应用程序开发; MapReduce 的工作机制; MapReduce 的类型和格式; MapReduce 的特性; 如何构建 Hadoop 集群, 如何管理 Hadoop; Pig; HBase; Hive; ZooKeeper; 开源工具 Sqoop, 最后还提供了丰富的案例分析。

本书是 Hadoop 权威参考,程序员可从中探索如何分析海量数据集,管理员可以从中了解如何安装与运行 Hadoop 集群。

Copyright © 2012 Tom White. All rights reserved.
Authorized Simplified Chinese translation edition, by O'Reilly Media, Inc., is published by Tsinghua University Press, 2014. 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 2012 年出版。

本书中文简体版由 O'Reilly Media, Inc. 授权清华大学出版社 2014 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: 01-2012-8670

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(第 3 版)/(美)怀特(White, T.)著; 华东师范大学数据科学与工程学院译.
—北京: 清华大学出版社, 2015

书名原文: Hadoop: The Definitive Guide, 3E

ISBN 978-7-302-37085-7

I. ①H… II. ①怀… ②华… 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(2014)第 145980 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 刷 者: 清华大学印刷厂

装 订 者: 三河市新茂装订有限公司

经 销: 全国新华书店

开 本: 178mm×233mm 印 张: 44.5 插 页: 1 字 数: 792 千字

版 次: 2010 年 5 月第 1 版 2015 年 1 月第 3 版 印 次: 2015 年 1 月第 1 次印刷

印 数: 1~3500

定 价: 99.00 元

Doug Cutting推荐序

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)开始着手尝试重建这些系统，将其作为Nutch的一部分。

我们终于让Nutch在20台机器上得以平稳运行，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开发人员能够搞定的。

几乎就在那个时候，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MapReduceAPI，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，不一而足。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，还有他那化繁为简、通俗易懂的写作风格。

Doug Cutting
加州，院内小屋

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 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路(岔路)。’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

在这本《Hadoop 权威指南(第 3 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。

一本书的价值最重要的是取决于它的用途。今天，开源的 Hadoop 已经成为研究大数据十分重要的平台，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(第 3 版)》恰好满足了这样的需要，其用途和价值不言而喻。这一点也可以从下面的事实中得到佐证：从 2011 年年底出版至今，本书的第 2 版已经实现了 12 次印刷，拥有近 3.5 万名读者并连续两年位列专业图书畅销榜榜首。

一本书的价值还来自于它的内容。原著是用英文写作的，它的内容组织得当，思路清晰，紧密结合实际。但是要把它翻译成中文介绍给中国的读者，并非易事。它不单单要求译者能够熟练地掌握英文，还要求他们对书中的技术性内容有深入、准确的了解和掌握。从这两点来审视，本书的译者团队是胜任有余的。作为大学老师，他们在大数据领域从事一线的教学、科研和指导研究生的工作。从 2006 年开始，他们就在计算机集群上部署了 Hadoop 并成功完成了很多项研究。在这几年的工作过程中，他们对 Hadoop 及其应用开发有着深入的理解和认识，这是本书翻译质量得以保证的重要前提。

经过再一次升级、修订和更新，值此《Hadoop 权威指南(第 3 版)》出版之际，我衷心地希望这本书继续为广大读者带来更高的学习价值，更友好的阅读体验。

周立柱@清华园

2014 年秋

译者序

互联网改变了一切，更改变了信息技术的发展模式。简单看来，互联网企业发展的基本路线图是这样的：通过免费向人们提供搜索、娱乐、社交等服务，吸引用户上线；产生流量之后，再通过“流量变现”来获取收益。这也就是说，通过构筑虚拟世界，通过“制造需求”，从而产生效益。这一事实也说明了为什么人们会用“眼球经济”来形容互联网经济。互联网提供的各类服务是直接面向广大网民，也就是面向消费者的(Consumer-oriented)。这和面向企业/机构的(Enterprise-oriented)传统信息技术和系统形成了鲜明的对比。互联网所需要的技术和系统不仅必须具备良好的扩展性和弹性，还必须费用低、可用性高。具体到数据管理，以关系数据库为代表的传统数据库管理系统(DBMS)是无法达到这个要求的，以 Hadoop 为代表的数据处理技术和系统因应这个阶段的数据管理需求而生。2003 年、2004 年和 2006 年，谷歌先后发表 GFS、MapReduce 和 BigTable 三篇相关论文，此后以 Hadoop 及其相关系统为代表的开源数据管理平台如雨后春笋般涌现，迅速形成互联网数据处理的技术生态圈，其典型特征就是通过搭建无共享的 MPP 集群获得分布并行的数据处理能力，进而拥有高可用性和高性能的优势。

在 2008 年，学术界曾经爆发过 MapReduce 技术是否是一个倒退的争论。一方面，以 Michale Stonebraker 和 David Dewitt 为主导的数据库大师指出，MapReduce 数据处理模型存在有一些不足和局限，例如，缺乏对过去四十多年以来数据库系统领域所取得成就的重视(包括数据建模、数据库模式与应用分离、高级查询语言等)，而是采用蛮力进行数据扫描，并不重视索引在数据处理方面的作用。而且，MapReduce 所采用的并行处理模型在早期的并行数据库 Gamma 系统中业已采用，并得到了深入的研究，所以也算不上什么重大的创新。另一方面，以谷歌云计算专家 Jeff Dean 和 Sanjay

Ghemawat 为首的技术新秀(2013 年 ACM Infosys 基金会计算机科学奖获得者)则认为, MapReduce 最大的作用是它提供了一种便捷的大数据处理方式,大幅简化了企事业单位安装、部署和配置大数据处理系统的工作,大幅降低了大数据处理成本。这场争论最后以并行数据库系统和 MapReduce 技术各自有其适合的应用场合而结束。

以 MapReduce 技术为主要技术蓝本的 Hadoop 系统在我国拥有者庞大的用户群,形成了一个活跃的开源系统生态圈。互联网企业(例如百度、阿里和腾讯等)、金融行业(例如工商银行和建设银行等)、IT 厂商(例如华为和中兴等)以及不计其数的中小型企业,都是 Hadoop 系统的拥趸。从每年召开的大数据技术大会上可以看出,我国拥有着庞大的 Hadoop 系统开发群体,他们现在已经构建了 Hadoop 之上大量的应用软件,每天处理着类型各异的大规模数据。因此, Hadoop 及其周边技术和系统在我国仍然有广泛的影响和应用市场。

从 2006 年开始,我们团队就在拥有 288 核和 40 TB 存储的计算机集群上部署了 Hadoop 系统,成功完成了多项研究。通过多年的学术研究和项目实施,我们对 Hadoop 系统及其上的应用开发有了较深入的理解和认识,积累了系统部署、调优、优化等方面的诸多经验。在本书作者 Tom White 于 2010 年出版《Hadoop 权威指南》(第 2 版)之后,我们受邀进行翻译工作。这本书的内容组织得很好,思路也很清晰,紧密结合实际问题。第 2 版的翻译是成功的,深受广大读者喜爱和欢迎。在 2013 年, Tom White 根据 Hadoop 的新版本推出第 3 版之后,我们出于对这本书的喜爱,再一次着手修订、翻译和完善。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 16 章和 3 个附录。翻译和审校工作由周傲英教授组织完成。参加翻译工作的有:周敏奇,第 1 章到第 4 章;王晓玲,第 5 章到第 7 章;金澈清,第 8 章到第 10 章及附录 A、B、C;钱卫宁,第 11 章到第 13 章;宫学庆,第 14 章到第 15 章;张蓉,第 16 章。

由于本书涉及面广,许多术语目前尚无固定译法。有时,为一个术语选择一个简洁、达意的译法,译者团队虽然经过反复推敲和讨论,但仍然可能出现词不达意的情况。对于有些确信不会对读者造成困扰的专有名称(如 datanode, comparator 等),由于在实践过程中都习惯于沿用这些外来语,所以我们在本书的表述中也保留了部分名称,希望不会对您的工作和学习造

成困扰。由于译者水平有限，译文中的不当之处也在所难免，真诚希望广大同行和读者不吝赐教。如果能将您的意见和建议发往{mqzhou, wqian, cqjin, xlwang, xqgong, rzhang}@sei.ecnu.edu.cn，我们将不胜感激。

华东师范大学数据科学与工程研究院

2014年11月

前言

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

“在我的世界里，只有微积分。这是我的专栏^①取得成功奥秘。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。”^②

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，我们来看它的“庐山真面目”。它提供的用于构建分布式系统的每个工具(用于数据存储、数据分析和协调处理)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，对偶尔有大量数据需要存储的程序员、有大量数据需要分析的程序员、有大量计算机需要管理的程序员同时却没有足够的时间、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

① 1957 年，加德纳在《科学美国人》杂志开设了“数学游戏”专栏，当然也介绍最新的数学研究成果。这个专栏深受读者欢迎，一直延续到了 1981 年，它成就了加德纳在趣味数学这个领域的大师地位

② 作者是《卫报》记者兼科学专栏作家 Alex Bellos。此文发表于 2008 年 5 月 31 日，网址为 <http://www.theguardian.com/science/2008/may/31/maths.science>。

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。在过去的三年多时间里，Hadoop 项目开花结果并孵化出大约半打子项目。到目前，它在性能、可靠性、可扩展性和可管理性方面都实现了巨大的飞跃。但是，为了让更多人采用 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、改进的 API。我希望我自己能够参与，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，要想查阅相关子项目的 Hadoop Java API 文档，可以访问 Apache Hadoop 主页(<http://hadoop.apache.org>)。如果使用 IDE 编程，可以充分利用其自动补全机制(也称“自动完成机制”)。

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。

本书特色内容

本书是这样组织的。第 1 章强调为什么需要 Hadoop，然后概述项目发展历史。第 2 章简要介绍 MapReduce。第 3 章深入剖析 Hadoop 文件系统，特别是 HDFS。第 4 章包含 Hadoop 的基本 I/O 操作：数据完整性、压缩、序列化及基于文件的数据结构。

接下来的第 5 章~第 8 章深入剖析 MapReduce。第 5 章全景呈现了 MapReduce 应用开发所涉及的具体步骤。第 6 章从用户的角度来看如何在 Hadoop 中实现 MapReduce。第 7 章主要包含 MapReduce 编程模型和 MapReduce 可以使用的各种数据格式。第 8 章是 MapReduce 高级主题，包括排序和数据连接。

第 9 章和第 10 章主要面向 Hadoop 管理员，主要描述如何在 Hadoop 集群上设置和维护运行 HDFS 和 MapReduce。

第 11 章~第 15 章专门介绍在 Hadoop 上构建的特定项目或相关内容。第 11 章和第 12 章描述的是 Pig 和 Hive，这两个分析平台构建在 HDFS 和 MapReduce 之上，而第 13 章、第 14 章和第 15 章分别介绍 HBase、ZooKeeper 和 Sqoop。

最后，第 16 章收集了 Apache Hadoop 社区成员提供的一系列实例。

第 2 版新增内容

《Hadoop 权威指南》(第 2 版)新增两章内容(第 12 章和第 15 章)，分别介绍 Hive 和 Sqoop。第 4 章新增一个小节专门介绍 Avro，第 9 章概述 Hadoop 新增的安全特性，第 16 章新增一个新的实例分析，介绍如何使用 Hadoop 来分析海量网络图。

第 2 版继续介绍 Apache Hadoop 0.20 系列发行版本，因为这是本书写作期间最新、最稳定的发行版本。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，我会说明具体的 Hadoop 版本号。

第 3 版新增内容

第 3 版概述了 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 1 章的 1.6 节将进一步描述每一个发行版本的特性。

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，第 6 章介绍它的工作机制，第 9 章介绍如何运行它。

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 5 章)。第 3 版还深入介绍了一些特性，第 8 章介绍输出 committer 和分布式缓存，第 9 章描述任务内存监控。我们还在第 4 章新增了如何写 MapReduce 作业来处理 Avro 数据的相关介绍，在第 5 章介绍如何在 Oozie 中运行一个简单的 MapReduce 工作流。

第 3 章的主题是 HDFS，现在新增了对高可用性、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，在第 3 版中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 版进行了彻底的纠错、修订和优化。

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。



这个图标表示重要的指示、建议或通用的说明。



这个图标表示警告或需要注意。

示例代码的使用

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。

我们欣赏引用时注明出处的做法，但不强求。引用通常包括书名、作者、出版社和 ISBN，例如“Hadoop: The Definitive Guide, Third Edition, by Tom White. Copyright 2013 Tom White, 978-1-449-31152-0”。

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 为组织机构、政府部门和个人提供广泛、灵活的产品组合和定价方式。在这里，订阅者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 及其他几十家出版社。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://www.oreilly.com/catalog/9781449311520>(英文版)

<http://www.oreilly.com.cn/book.php?bn=978-7-302-37085-7>(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 Biał ecki, Stu Hood, Chris K. Wensel 和 Owen O'Malley，他们为第 16 章提供了实例学习。

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”) Komer 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, 我还要向所有提交第 2 版勘误的读者表达最真挚的谢意。

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言。

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, 使我有充足的时间写书, 并能及时交稿。

非常感谢我的编辑 Mike Loukides 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 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, 如果没有家庭一如既往地支持, 我是不可能完成的。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, 一直都非常善解人意, 期待我有更多时间好好陪陪她们。