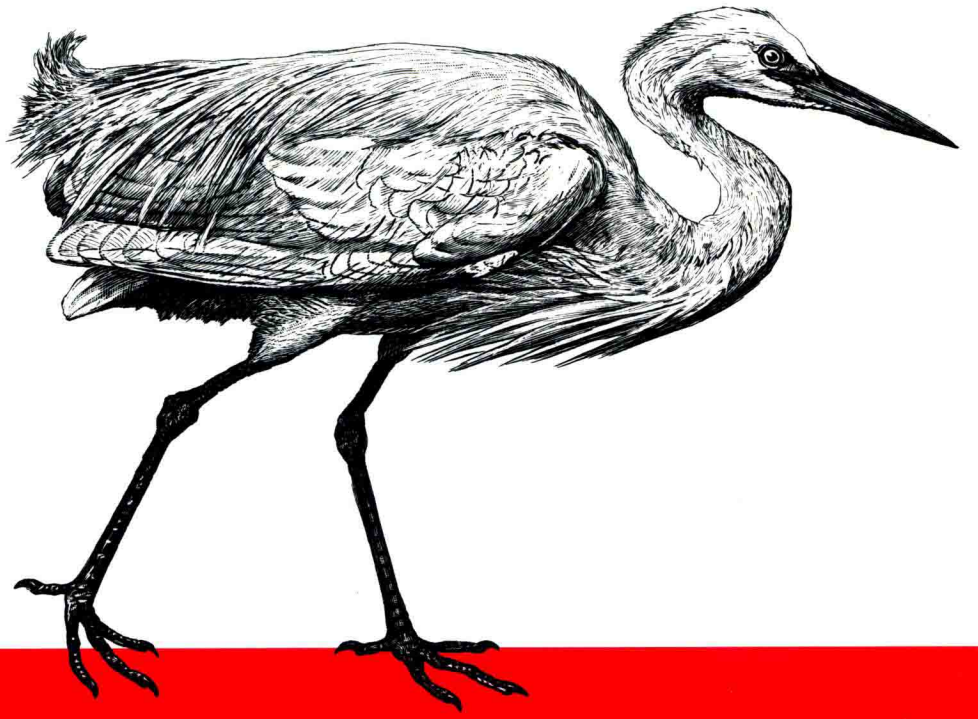
O'REILLY

TURING 图灵程序设计丛书



Hadoop数据分析

Data Analytics with Hadoop

针对数据分析介绍分布式计算涉及的大量概念、工具和技术，纵览Hadoop生态系统

[美] Benjamin Bengfort Jenny Kim 著

王纯超 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Hadoop数据分析

Data Analytics with Hadoop

[美] Benjamin Bengfort Jenny Kim 著
王纯超 译

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®
O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

Hadoop数据分析 / （美）本杰明·班福特（Benjamin Bengfort），（美）珍妮·基姆（Jenny Kim）著；王纯超译. -- 北京：人民邮电出版社，2018.4
（图灵程序设计丛书）
ISBN 978-7-115-47964-8

Ⅰ. ①H… Ⅱ. ①本… ②珍… ③王… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第036682号

内 容 提 要

通过提供分布式数据存储和并行计算框架，Hadoop已经从一个集群计算的抽象演化成了一个大数据的操作系统。本书旨在通过以可读且直观的方式提供集群计算和分析的概览，为数据科学家深入了解特定主题领域铺平道路，从数据科学家的视角介绍Hadoop集群计算和分析。本书分为两大部分，第一部分从非常高的层次介绍分布式计算，讨论如何在集群上运行计算；第二部分则重点关注数据科学家应该了解的工具和技术，意在为各种分析和大规模数据管理提供动力。

本书适合数据科学领域的从业人员，以及对数据分析感兴趣的研究人员。

◆ 著　[美] Benjamin Bengfort　Jenny Kim
译　王纯超
责任编辑　朱　巍
执行编辑　夏静文
责任印制　周昇亮

◆ 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
三河市潮河印业有限公司印刷

◆ 开本：800×1000　1/16
印张：14.25
字数：337千字　2018年4月第1版
印数：1-3 500册　2018年4月河北第1次印刷
著作权合同登记号　图字：01-2017-6475号

定价：69.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

站在巨人的肩上
Standing on Shoulders of Giants

iTuring.cn

版权声明

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。无论是通过图书出版、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*

前言

大数据已经成为一个流行词。人们用它来描述数据驱动型应用程序中的那些令人兴奋的新工具和新技术。这些应用程序正为我们带来崭新的计算方式。令统计学家懊恼的是，这一词语似乎被随意使用，其范围甚至包括在大型数据集上使用众所周知的统计技术进行预测。虽然大数据已经成为流行词，但事实上，现代分布式计算技术能分析的数据集远比过去那些“典型”方式能应对的数据集大得多，结果也更令人震撼。

然而，单纯的分布式计算并不等于数据科学。互联网带来了快速增长的数据集，这些数据集又能驱动预测模型（“更多的数据优于更好的算法”[1]），数据产品也因此成为了一种新型的经济范式。为大型跨域异构数据集建模所取得的巨大成功（例如 Nate Silver 通过大数据技术像使用魔法一般预测了 2008 年的美国大选结果），使很多人认识到了数据科学的价值，也为这个领域吸引了大量从业者。

通过提供分布式数据存储和并行计算框架，Hadoop 已经从集群计算的抽象演变成了大数据操作系统。Spark 正是基于这一理念构建的，它使数据科学家能更轻松地使用集群计算。然而，不了解分布式计算的数据科学家和分析人员可能会觉得这些工具是面向**程序员**的，而不是面向**分析人员**的。这是因为，我们需要从根本上转变管理数据和计算数据的思维方式，这样才能从串行模式转换到**并行**模式。

本书旨在通过**可读**且直观的方式介绍集群计算和分析，帮助数据科学家完成这一思维转换。我们将针对数据分析介绍分布式计算涉及的大量概念、工具和技术，为深入了解特定领域铺平道路。

本书目标

本书不会详细讲解 Hadoop（推荐 Tom White 的《Hadoop 权威指南》），也不是 Spark 入门资料（推荐 Holden Karau 等人所著的《Spark 快速大数据分析》[2]），当然更不是为了教你如

注 1：Anand Rajaraman, “More data usually beats better algorithms” (http://anand.typepad.com/datawocky/2008/03/more-data-usual.html), Datawocky, March 24, 2008.

注 2：本书已由人民邮电出版社出版，http://www.ituring.com.cn/book/1558。——编者注

何进行分布式计算。本书将纵览 Hadoop 生态系统和分布式计算，旨在武装数据科学家、统计学家、程序员和对 Hadoop 感兴趣（但是对 Hadoop 的了解十分有限）的人。希望本书能成为你深入 Hadoop 世界的向导，助你找到最感兴趣的工具和技术。这可能是 Spark、Hive、机器学习、ETL（抽取、转换和加载）操作、关系数据库或者众多与集群计算相关的主题之一。

目标读者

人们经常把数据科学与大数据混为一谈。虽然为了达到良好的泛化效果，许多机器学习模型确实需要大型数据集，但即使是小型数据集也能支持模式识别。因此，数据科学软件类的图书大多关注易于在一台机器（尤其是内存容量多达几吉字节的机器）上分析的数据集。尽管大数据和数据科学非常适合协同工作，但是直到今天，与计算相关的图书还是将它们分开讨论。

本书以数据科学家为目标读者，旨在弥补这一隔阂。它将以数据科学的视角介绍 Hadoop 集群计算和分析。本书的关注点不是部署、运维或软件开发，而是常用分析、数据仓储技术和高阶数据流。

那么，什么人算是数据科学家呢？本书所说的数据科学家是指具有高超统计技能的软件开发人员，或者具有强大软件开发能力的统计学家。通常情况下，数据团队由三类数据科学家组成，分别是数据工程师、数据分析师和领域专家。

数据工程师指能构建或者使用高级计算系统的程序员或者计算机科学家。他们通常使用 Python、Java 或者 Scala 编程，熟悉 Linux、服务器、网络、数据库和应用程序部署。如果你是数据工程师，本书假设你能适应多进程编程、数据整理和数值计算。希望你在阅读本书后，能更了解如何在集群上部署应用程序，学会如何处理比单机在足够时间内能处理的数据集还要大得多的数据集。

数据分析师主要关注统计建模和探索性数据分析。在日常工作中，他们通常使用 R、Python 或者 Julia，熟悉数据挖掘和机器学习技术，比如回归、聚类和分类问题。数据分析师很可能通过采样处理过更大的数据集。我们将在本书中展示数据统计技术，处理比以往获取的数据量大得多的数据，从而构建预测能力既有广度又有深度的模型。

领域专家是团队里富有影响力、面向业务的成员。他们深入了解数据类型和所碰到的问题，理解数据带来的特定挑战，并寻求通过更好的方式利用数据应对新挑战。希望本书能够为他们提供一些业务决策思路，让当前的数据流更加灵活，并帮助他们理解怎样使用通用的计算框架来应对特定的领域挑战。

阅读方式

至今为止，Hadoop 已经有十多年的历史了，就技术而言，这已经是很长一段时间了。然而，摩尔定律仍然没有慢下来。10 年前，在数据中心使用廉价的机器集群远比为超级计算机编程简单。但现在，同样的廉价服务器要比以前强大约 32 倍，内存计算的开销也降低了很多。Hadoop 成为了大数据的操作系统，支持图形处理、类 SQL 查询和流处理等多种

计算框架。但这也给想要学习 Hadoop 的人带来了巨大的挑战——该从何学起？

本书篇幅简短的原因只有一个——想要尽可能简洁地覆盖多个方面。我们希望你通过两种方式阅读本书：一是快速通读全书，对 Hadoop 和分布式数据分析有大致了解；二是选择感兴趣的章节深入学习。本书以**易懂**为目的。我们通过简单的代码示例进行讲解，不一定需要你亲自实现和运行代码。本书是 Hadoop 和 Spark 领域的指导手册，对分析人员尤其如此。

内容概述

本书旨在带领你了解 Hadoop 生态系统，书中内容分为两部分：第一部分（第 1 章至第 5 章）宏观地介绍分布式计算，讨论**如何**在集群上运行计算；第二部分（第 6 章至第 10 章）侧重于介绍数据科学家应该具体了解的工具和技术，意在为各种分析和大规模数据管理提供**动力**。（第 5 章将从对分布式计算的讨论过渡到更加具体的工具和大数据科学流水线的实现。）每章的内容概述如下。

第 1 章 数据产品时代

介绍大数据和数据科学的结晶——数据产品，讨论创建数据产品背后的流程，说明数据分析的串行模型如何与分布式计算相契合。

第 2 章 大数据操作系统

概述 Hadoop 背后的核心概念，讲解为何集群计算既有益又复杂；主要着眼于 YARN 和 HDFS，详细讨论 Hadoop 体系架构，讲解与分布式存储系统的交互，为分析大型数据集作准备。

第 3 章 Python 框架和 Hadoop Streaming

介绍分布式计算的基本编程抽象 MapReduce。然而，MapReduce 的 API 是用 Java 编写的，这不是一种在数据科学家间流行的编程语言。因此，这一章专注于介绍如何通过 Hadoop Streaming 使用 Python 编写 MapReduce 作业。

第 4 章 Spark 内存计算

虽然理解 MapReduce 对理解分布式计算和编写高性能的批处理作业（如 ETL）十分重要，但是 Hadoop 集群上的日常交互和分析却通常都是使用 Spark 完成的。这一章将介绍 Spark，以及如何使用 Python 编写 Spark 应用程序，并通过 PySpark 以交互方式在 YARN 上运行，或者在集群模式下运行。

第 5 章 分布式分析和模式

通过展示设计模式和并行分析算法，从实践的角度研究怎样编写分布式数据分析作业。开始阅读这一章之前，你应该已经了解编写 Spark 和 MapReduce 作业的原理。读完这一章，你应该能轻松实现它们。

第 6 章 数据挖掘和数据仓储

介绍分布式环境下的数据管理、数据挖掘和数据仓储，特别是与传统数据库系统密切相关的方面。这一章重点介绍 Hive 和 HBase，它们分别是 Hadoop 最流行的基于 SQL 的查询引擎和 NoSQL 数据库。数据整理是数据科学流水线的第二步，但是数据需要被采集到某处。这一章还将探索怎样管理大型数据集。

第 7 章 数据采集

考虑到数据的容量和速度，如何将数据导入分布式系统并用于计算可能才是最大的挑战之一。这一章将研究从关系数据库获取数据的批量加载工具 Sqoop 以及更灵活的 Apache Flume，后者用于获取日志和来自网络的其他非结构化数据。

第 8 章 使用高级 API 进行分析

研究用于编写复杂 Hadoop 和 Spark 应用程序的高阶工具，尤其是 Apache Pig 和 Spark 的 DataFrame API。第一部分将讨论 MapReduce 和 Spark 分布式作业的实现过程，以及怎样从数据流的角度看待算法和数据流水线。Pig 让你无须使用 MapReduce 实现底层细节，从而能更轻松地描述数据流。Spark 提供了多个集成模块，能无缝结合过程式处理与关系查询，为强大的分析定制打开了大门。

第 9 章 机器学习

大数据的多数益处都是在机器学习中得以实现的——更加广泛的特征和输入空间让模式识别技术更加有效和个性化。这一章将介绍分类、聚类和协同过滤，但并不会详细讨论建模，而是使用 Spark 的 MLlib 让你上手可扩展机器学习技术。

第 10 章 总结：分布式数据科学实战

完整呈现分布式数据科学，把前面章节中单独讨论的工具与技术结合起来。数据科学不是单一的活动，而是一个生命周期，涉及数据的采集、整理、建模、计算和操作化。这一章将从整体上讨论分布式数据科学的架构和工作流。

附录 A 创建 Hadoop 伪分布式开发环境

附录 A 将指导你在本机上搭建一个开发环境，从而编写分布式作业。如果你没有集群可用，附录 A 是运行本书示例至关重要的准备工作。

附录 B 安装 Hadoop 生态系统产品

附录 B 是附录 A 的延伸，将提供本书讨论的众多生态系统工具和产品的安装指导。尽管附录 A 提供了安装服务的常用方法，但附录 B 专门为安装服务（用来运行书中示例，你在阅读的过程中会遇到它们）的过程中会遇到的问题提供指导。

你看，这么薄的一本书却涵盖了这么多主题。希望以上这些内容足以吸引你继续阅读下去。

编程和示例代码

随着 Hadoop 的分布式计算变得更加成熟和集成化，并行计算正在向更丰富的分析体验转变。例如，大数据生态系统的最新成员 Spark 提供了 4 种语言的编程 API，更方便那些习惯于使用数据框、交互式 notebook 和解释型语言的数据科学家使用。Hive 和 SparkSQL 以 SQL 语法形式提供了另外一种为人们所熟知的领域专用语言（domain-specific language，DSL），专门针对分布式集群上的数据查询。

因为本书的目标读者是数据科学家，所以我们大多选择使用 Python 来实现示例。Python 是通用的编程语言，拥有丰富的分析包（例如 Pandas 和 Scikit-Learn），在数据科学领域占有一席之地。不幸的是，Hadoop 的主要 API 通常都是以 Java 编写的，那些 Python 示例让我

们大费周章，但是大多数时候，我们会用更实际的方式来阐明思想。因此，本书中的代码要么是使用 Python 和 Hadoop Streaming 的 MapReduce，要么是使用 PySpark API 的 Spark 代码，或者是讨论 Hive、Spark SQL 时的 SQL 代码。希望这能让更多读者感觉简明易懂。

GitHub仓库

你可以在我们的 GitHub 仓库（https://github.com/bbengfort/hadoop-fundamentals）找到本书完整且可执行的示例代码。这个仓库也包含了我们的 Hadoop 视频教程“Hadoop Fundamentals for Data Scientists”[3] 的代码。

为了在纸质版中呈现代码并更清楚地解释过程，我们走了捷径，省略了代码的细节。例如，通常都省略了 `import` 语句——这意味着简单的复制粘贴无法奏效。然而，你可以使用仓库中的例子，它们是完整且可执行的代码，并附有相应的注释。

但要注意，仓库是持续更新的，可以查阅 README 文件以了解更新情况。你当然可以 fork 仓库，更改代码以在你自己的环境中运行——我们强烈推荐你这样做！

执行分布式作业

Hadoop 开发人员通常在“伪分布式模式”下使用“单节点集群”进行开发任务。该集群通常是一个运行着虚拟服务器环境的虚拟机，环境中运行着多个 Hadoop 守护进程。你可以在主开发工具里使用 SSH 访问该虚拟机，就像访问 Hadoop 集群一样。为了创建虚拟环境，你需要某种虚拟化软件，例如 VirtualBox（https://www.virtualbox.org）、VMWare（http://www.vmware.com/products/desktop-virtualization）或者 Parallels（http://www.parallels.com）。

附录 A 讨论怎样设置以伪分布式模式运行 Hadoop、Hive 和 Spark 的 Ubuntu x64 虚拟机。你也可以使用一些 Hadoop 发行版（例如 Cloudera 和 Hortonworks）提供的预先配置好的虚拟环境。如果你有想用的虚拟机环境，那么我们建议你下载它。如果你想了解更多的 Hadoop 操作，就请自己配置吧！

还有一点，因为 Hadoop 集群是在开源软件上运行的，所以你需要了解 Linux 和命令行。本书讨论的虚拟机通常都是通过命令行访问的，书中的许多例子都描述了通过命令行与 Hadoop、Spark、Hive 和其他工具交互的过程。命令行是分析人员不愿使用这些工具的一个主要原因。然而，学习命令行对你大有帮助，它也并不可怕。我们建议你学习一下！

使用示例代码

本书是要帮你完成工作的。一般来说，如果本书提供了示例代码，你可以把它用在你的程序或文档中。除非你使用了很大一部分代码，否则无须联系我们获得许可。比如，用本书的几个代码片段写一个程序就无须获得许可，销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用本书中的示例代码回答问题无须获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

注 3：http://shop.oreilly.com/product/0636920035183.do.

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和ISBN，比如“*Data Analytics with Hadoop* by Benjamin Bengfort and Jenny Kim (O'Reilly). Copyright 2016 Benjamin Bengfort and Jenny Kim, 978-1491-91370-3”。

如果你觉得自己对示例代码的用法超出了上述许可的范围，欢迎你通过permissions@oreilly.com与我们联系。

反馈及作者联系方式

关于本书的评论和技术性问题，请发送电子邮件至bookquestions@oreilly.com。

工具和技术的变化非常快，在大数据领域尤其如此。不幸的是，很难保证一本书（特别是纸质版）时刻跟上潮流。我们希望本书能够在未来继续为你服务，但如果你发现有变更让书中的示例无法运行或导致代码问题，请联系我们。

如果有关于代码或示例的问题，请在GitHub上（https://github.com/bbengfort/hadoop-fundamentals/issues/）提交问题，这是与我们联系的最佳方式。你也可以发送电子邮件到hadoopfundamentals@gmail.com，我们会尽快回复。非常感谢你提供积极且具有建设性的反馈！

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）是应运而生的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。技术专家、软件开发人员、Web设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将Safari Books Online视作获取资料的首选渠道。

对于组织团体、政府机构和个人，Safari Books Online提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问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以及其他几百家出版社的上千种图书、培训视频和正式出版之前的书稿。要了解Safari Books Online的更多信息，我们网上见。

联系我们

请把对本书的评价和问题发给出版社。

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

O'Reilly 的每一本书都有专属网页，你可以在那儿找到本书的相关信息，包括勘误表、示例代码以及其他信息。[4] 本书的网页是：

http://shop.oreilly.com/product/0636920035275.do

对于本书的评论和技术性问题，请发送电子邮件到：

bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：

http://www.oreilly.com

我们在 Facebook 的地址如下：

http://facebook.com/oreilly

请关注我们的 Twitter 动态：

http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：

http://www.youtube.com/oreillymedia

致谢

我们要感谢本书的审校者，你们在漫长的写作过程中持续提出颇具建设性的反馈和批评。感谢 Marck Vaisman，你从向数据科学家讲授 Hadoop 的角度阅读本书；特别感谢 Konstantinos Xirogiannopoulos 在忙碌的研究之余，志愿向我们提供清晰、有益且积极的评论，我们很高兴收到这样的评论。

我们还要感谢 O'Reilly 每位富有耐心、坚持不懈的编辑。在本书写作之初，我们走了一些弯路，是 Meghan Blanchette 带领我们一路走了过来，她一直支持着我们。很遗憾，在本书还未截稿时，她已离开 O'Reilly 去追求更好的事业。当 Nicole Tache 接替她并成功将我们带回正轨时，我们特别高兴。Nicole 引导我们完成写作，没有她，就没有本书。她有一项特殊的本领，总能在关键时刻发来让人看了就舒心的邮件，让工作如期完成。与 O'Reilly 每个人的合作都很愉快，特别是 Marie Beaugureau、Amy Jollymore、Ben Lorica 和 Mike Loukides，感谢你们给予的建议和鼓励。

在华盛顿，District Data Labs 的工作人员给予了我们巨大的支持。我们要特别提到 Tony Ojeda、Rebecca Bilbro、Allen Leis 和 Selma Gomez Orr，你们以各种方式支持本书，包括第一个购买早期版本、提供反馈、审查代码并询问完成时间，这都鼓励我们专心写作！

如果没有 Hadoop 社区中各位了不起的贡献，本书就不可能诞生，Jenny 更是有幸每天在 Cloudera 与这些志同道合者一起工作。特别感谢 Hue 团队，他们为提供最好的 Hadoop 用

注 4：也可以通过图灵社区获取相关信息：http://www.ituring.com.cn/book/1944。——编者注

户体验所做出的贡献和表现出的热情既超乎寻常又鼓舞人心。

感谢我们的家人，特别是 Benjamin 的父母 Randy Bengfort 和 Lily Bengfort，以及 Jenny 的父母 Wung Kim 和 Namoak Kim，感谢你们一直以来的鼓励、爱与支持。我们的父母向我们灌输了相互学习和探索的热情，因此我们钻研了许多方面。他们使我们拥有坚韧的精神和恒心，总能让我们找到实现目标的方法。

最后，感谢我们的伴侣 Patrick 和 Jacquelyn，感谢你们一直支持我们。我们中的谁可能说过："再写一本书，我的婚姻就要结束了。"诚然，在写作的最后阶段，他们都不愿意听到我们仍然在努力。但如果没有他们，我们的书就无法写成（"婚姻不在，书也就不在了"）。在我们通过视频电话商定细节和重写时，Patrick 和 Jacquelyn 总是和颜悦色。他们甚至阅读了部分内容，提供建议，在各方面都提供了帮助。在这之前，我们都没有写过书，也不知道会面临什么问题。现在我们知道了，也很高兴他们一直在身边支持我们。

电子书

扫描如下二维码，即可购买本书电子版。

目录

第二部分　大数据科学的工作流和工具