O'REILLY®

用Hadoop创建数据分析应用

敏捷数据科学

AGILE DATA SCIENCE

[美] Russell Jurney 著
冯文中 朱洪波 译

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

O'REILLY®

敏捷数据科学

AGILE DATA SCIENCE

[美] Russell Jurney 著
冯文中 朱洪波 译

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书面向大数据挖掘，以敏捷视角呈现高效构建数据模型的全程实践和思路。在一组以一个真实电子邮箱数据挖掘为例的数据 - 价值金字塔进阶模式中，你将学到：一整套实用工具及其方法论，可快速实现在 Hadoop 上构建数据分析应用；用 Python、Apache Pig 及 D3.js 等轻量级工具创建用于探索数据的敏捷环境；一种可根据数据中信息快速切换，进行不同类型数据分析的迭代式开发方法。

本书适合所有与数据工作相关的从业者，同时也适合有志成为数据科学工作者的广大读者作为入门读物。

版权贸易合同登记号 图字：01-2013-9391

图书在版编目（CIP）数据

敏捷数据科学：用 Hadoop 创建数据分析应用 /（美）朱尔尼（Jurney,R.）著；冯文中，朱洪波译 . —北京：电子工业出版社，2014.7
书名原文：Agile data science
ISBN 978-7-121-23619-8
Ⅰ . ①敏… Ⅱ . ①朱… ②冯… ③朱… Ⅲ . ①数据采集 Ⅳ . ① TP274
中国版本图书馆 CIP 数据核字 (2014) 第 135213 号

责任编辑：张春雨
印　　刷：北京丰源印刷厂
装　　订：三河市鹏成印业有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：787×980　1/16　印张：11.5　字数：240千字
版　　次：2014年7月第1版
印　　次：2014年7月第1次印刷
定　　价：49.00元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线：（010）88258888。

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、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

译者序

大数据时代已经到来。或者至少可以在概念层面上说，它已经到来。各行各业，都开始利用数据驱动决策。然而读者朋友是否想过，数据的核心价值来自哪里，又将我们引向何方？

我认为，核心价值是信息量。数据中蕴含的信息量，赋予我们进一步洞察这个复杂世界的“可能性”。数据是现实世界的快照——通过分析数据，我们可以对世界有更深入、更准确的理解。当我们能越快、越好地理解这个世界，就越有可能在行动中占得先机。

为了更好地开发数据带来的价值，我们开始构建数据分析应用程序，通过搜集、清洗、聚合、存储、分析、学习数据，挖掘出隐藏在数据内部的事物之间的关联。然而，构建数据分析应用程序是一项艰苦的工程——构建过程中，用户需求在变化，系统负载在增大，数据质量经常难以保证，改动不断，困难重重。究竟应该如何用有限的资源来构建出一条真正带来价值的数据流水线呢？

现在你拿在手里的，就是本书作者给出的答案——一份帮助读者高效构建数据分析产品，以更快更好地洞察这个复杂世界的实践指南。

作者结合自身的数据产品构建经验和大量例子，介绍如何用现代化的工具和平台，如Pig、MongoDB、Python、Elastic Search、D3.js、AWS等等，构建一个完整、可扩展的数据分析应用。读过本书，读者朋友就大概可以知道在构建数据分析应用时可能遇到的问题，解决这些问题时可以有哪些选择，设计中有哪些陷阱与反模式，以及如何利用开源项目组合出简洁优雅的解决方案。

这是一本难能可贵的覆盖构建数据分析产品方方面面的实战佳作，因此：

- 架构师，可以参考书中介绍的技术，改进系统设计和新增特性。

- 数据科学家，可以掌握更多的数据操作工具来处理和展示数据。
- 项目经理，可以认识到该如何构建团队、分解任务，如何使项目开发流程变得更加敏捷，如何向客户和团队设立合理的预期。
- 有志于向全栈方向发展的工程师或者学生，也能够进一步开阔眼界，了解自己熟悉的领域之外的生态系统。

2013 年秋天我与本书不期而遇，刚好对书中内容有所涉猎，一时技痒，向张春雨老师争取到与阿里朱洪波老师一起翻译的机会。虽然我和两位老师素未谋面，但依然在合作中获益良多。翻译过程中晓风老师悉心指导，在初稿完成后又字斟句酌地对译文进行了细致的审校和润色，使译文更为流畅自然。张春雨老师也在其间不断给予我们专业的反馈和帮助，让本书得以早日面世。在此，一并致以衷心的感谢。

翻译过程中，不由想起一个程序员自嘲的段子——“我不生产代码，我只是 Github 的搬运工”。大致意思是说，现今的程序员仅仅利用 Github 上的开源代码，就可以完成很多任务。从正面的角度来看，找到正确的方法和合适的工具，将生产效率提升十倍甚至百倍，不正是 IT 的价值所在？读罢本书，这种感受更深。

稍有疑问的是，这本书并未介绍单元测试，似乎与我们通常理解的敏捷略有不同。此外，由于译者的阅历和水平所限，本书的翻译难免存在疏漏和错误，还请读者朋友不吝指正。

最后，感谢你对本书的兴趣，相信你一定会有所收获。

文中

2014 年 6 月 于北京

前言

我写这本书是为了帮助大家避免重复我犯过的错误，进而防止失败项目的产生。在这本书里面，描述、反映了我在两个不同的 Hadoop 服务上构建数据分析应用的经验。

《敏捷数据科学》这本书有三个目标：提供一个用 Hadoop 构建数据分析应用的操作指南；帮助团队在大数据项目中以敏捷的形式进行更好的协作；提出一个进行敏捷式大数据分析的先进结构。

读者是谁

《敏捷数据科学》是一门帮助大数据的入门者以及萌芽中的数据科学家，成为数据科学与数据分析团队中更有生产力的成员的课程。它的目标是帮助工程师，分析师和数据科学家以敏捷的形式在 Hadoop 上处理大数据。它介绍了一种非常适合大数据的敏捷开发方式。

这本书同样针对需要处理数据并开发软件的程序员。设计师和项目经理可能会特别喜欢本书第一、二、五章，这些章节主要介绍一些敏捷的流程，而没有把关注点放在具体的代码上。

本书假设读者在一个 *nix 环境中工作，对于 Windows 用户，我们没有提供相应的例子，但他们可以使用 Cygwin。一个由用户贡献而且包含所有前置依赖的 Linux 镜像，可以从这（*https://github.com/charlesflynn/agiledata*）获取。读者可以通过这个工具在 Virtual Box 里面快速启动一台 Linux 机器。

本书如何组织

这本书包含两部分。第一部分介绍数据以及将在本书第二部分用到的工具集。我特意将第一部分写得比较简短，只用了较少的篇幅来介绍这些工具。假如感觉第一部分内容太过简略，也不要担心，本书第二部分将会深入地探索这些工具的用法。下面的章节组成了本书的第一部分：

第一章 理论

介绍敏捷大数据的工作方法。

第二章 数据

介绍本书中将要使用的数据，以及简单的预测方法。

第三章 敏捷开发工具

介绍工具集，并帮助读者将它们在机器上安装好。

第四章 在云端

带领读者将第三章中介绍的工具集扩展到云端，以支持 PB 级的数据规模。

本书第二部分是一个利用敏捷大数据的方式来构建数据分析应用的教程。这是一个笔记本形式的指南。在数据－价值金字塔的每一次上升都遵循着敏捷的原则。我会阐述如何在小的敏捷开发周期里面逐步创造价值。第二部分包含如下的章节：

第五章 收集和展示数据

帮助读者下载电子邮件收件箱数据并将邮件连接到一个 web 程序上。

第六章 使用图表可视化数据

让读者逐步在 web 程序中创建简单的图表来操纵数据。

第七章 利用报表探索数据

展示如何从数据里面提取实体，并将它们连接在一起，创建可交互的数据报告。

第八章 预测

帮助读者利用之前的成果预测邮件收到回复的概率。

第九章 驱动行动

介绍如何将已有的预测功能扩展成一个完整的实时分类器，来帮助用户写出会被回复的邮件。

本书所使用的约定

以下是本书所使用的排版约定：

斜体（*Italic*）

表示新的条目、网址、电子邮件地址、文件名和文件扩展名。

等宽字体（Constant width）

在程序代码中使用，同时也会出现在段落内的那些引用程序元素如函数名、数据库、数据类型、环境变量、声明和关键字等，还会在API、组件及模块名里遇到。

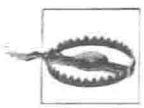
等宽粗字体（**Constant with bold**）

表示命令或由用户输入的文本。

等宽斜体字（*Constant with italic*）

表示应该由用户提供的值来代替上下文决定的值的那些文本。

这个图标标志是提示、建议或一般说明。



这个图标表示警告或告示。

使用代码实例

补充材料（代码示例，练习等）请登录 *https://github.com/rjurney/Agile_Date_Code* 自行下载。

本书就是要帮读者完成工作的。通常，如果本书包含了代码示例，你可以在你的程序和文档中使用本书中的代码。除非你复制了大段的代码，否则你无须联系我们来取得许可。举个例子，在编写程序时使用了本书中的数块代码是不需要经过许可的。出售或分发来自O'Reilly图书的示例CD-ROM是必须经过许可的。引用本书及本书中的示例代码来回答问题是不需要经过许可的。将大量的示例代码整合到你的产品文档中必须经过许可。

我们很感谢但不要求注明出处。出处的格式一般包括标题、作者、出版社和ISBN，例如“由Russell Jurney写的Agile Data Science (O'Reilly). Copyright 2014 Data Syndrome LLC, 978-1-449-32626-5。”

如果你觉得没有在正常范围内使用代码例子，并且不知是否在上面所说的许可范围内，请随时联系我们：

permissions@oreilly.com

Safari® Books Online

Safari Books Online（*www.safaribooksonline.com*）是一个发布来自全球技术和商业领域的顶尖作者写的书和视频等优质内容的按需数字化图书馆。

技术专业人士、软件开发者、网站设计师及商业和创意专业都用 Safari Books Online 作为他们的主要研究、解决问题、学习和认证培训资源。

Safari Books Online 提供了一系列产品及针对组织、政府和个人不同的定价方案，订阅者可以访问到成千上万的图书、培训视频及出版前的手稿，这些内容都可以从出版社，如 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 等数据库中搜索而得到，想了解更多关于 Safari Books Online 的信息，请在线访问我们。

联系我们

对于本书的评论或问题请联系出版商：

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

我们为本书制作了一个 Web 页面，页面中包含了简介、样章，以及其他信息。可以从这里访问这个页面：

http://oreil.ly/agile-data-science

如果要留言或者提交关于本书的技术问题的反馈，请发邮件至：

bookquestions@oreilly.com

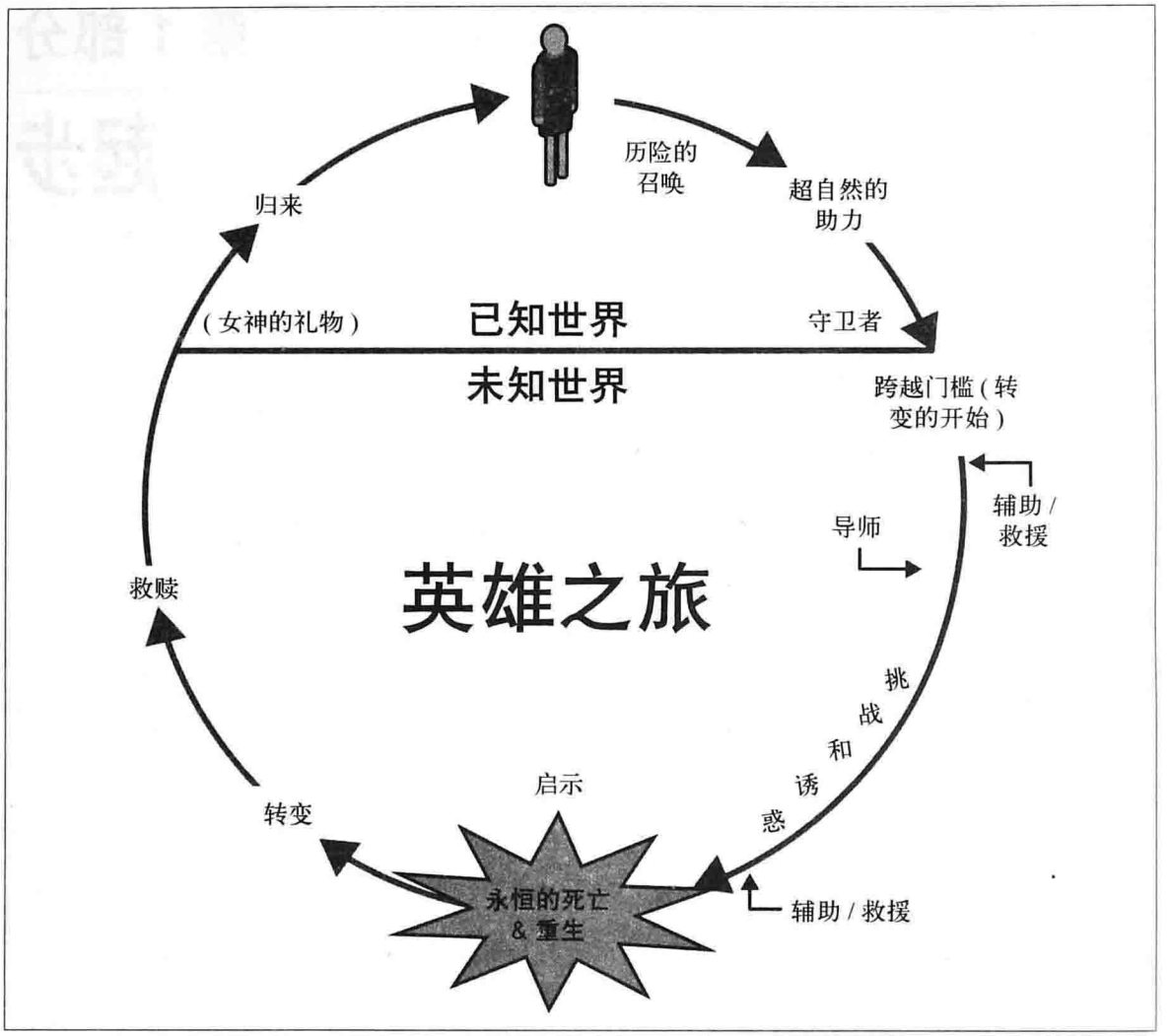
本书的更多信息、资源、参考文献和新闻，请登录出版社官网：http://www.oreilly.com。

Facebook: http://facebook.com/oreilly
Twitter: http://twitter.com/oreillymedia
YouTube: http://www.youtube.com/oreillymedia

目录

第 1 部分

起步



图I-1．英雄之旅，[译注1]来源：维基百科

译注 1：美国神话学家 Joseph Campbell 在其著作 *The Hero With a Thousand Faces* 中对各种神话故事剧情加以总结，提出“英雄之旅（The Henro's Journey）”剧情发展理论。作者引用“英雄之旅”隐喻本书的敏捷数据科学开发流程。更多内容请参考：*http://en. wikipedix.org/wiki/Monomyth*