O'REILLY®

Broadview®
www.b

Spark 全栈数据分析

Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark

[美] Russell Jurney 著
王道远 译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

O'REILLY®

Spark 全栈数据分析

Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark

[美] Russell Jurney 著

王道远 译

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

内 容 简 介

本书介绍了作者提出的基于 Spark 的敏捷数据科学方法论，结合作者在行业中多年的实际工作经验，为数据科学团队提供了一套以类似敏捷开发的方法开展数据科学研究的实践方法。书中展示了工业界一些常见工具的使用，包括从前端显示到后端处理的各个环节，手把手地帮助数据科学家快速将理论转化为真正面向用户的应用程序，从而让读者在利用数据创造真正价值的同时，也能不断完善自己的研究。

本书适合初学者阅读，数据科学家、工程师、分析师都能在本书中有所收获。

版权贸易合同登记号 图字：01-2017-5709

图书在版编目（CIP）数据

Spark 全栈数据分析 /（美）罗素 • 朱尼（Russell Jurney）著；王道远译 .
—北京：电子工业出版社，2018.11
书名原文：Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark
ISBN 978-7-121-35166-2

Ⅰ. ① S… Ⅱ. ①罗… ②王… Ⅲ. ①数据处理软件 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字（2018）第 227785 号

策划编辑：刘恩惠
责任编辑：牛 勇 特约编辑：顾慧芳
封面设计：Karen Montgomery 张 健
印 刷：三河市君旺印务有限公司
装 订：三河市君旺印务有限公司
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：21.5 字数：413千字
版 次：2018 年 11 月第 1 版
印 次：2018 年 11 月第 1 次印刷
定 价：99.00 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）88254888，88258888。

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。

本书咨询联系方式：010-51260888-819 fag@phei.com.cn。

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

这几年，大数据、人工智能都是计算机学科中非常热门的话题，数据科学也越来越受到各公司的关注。我接触到的很多不同行业的公司都对大数据与人工智能的应用跃跃欲试，甚至部分公司早已尝到其中的甜头。不过还有很多公司并没有真正用上它们那些本应蕴含着无穷价值的数据，也有很多公司已经学会从数据中挖掘信息，但得到的信息无法及时转化为真正的价值。毕竟大数据还是比较新的技术，大多数公司还在探索中前进。很多公司早就拥有了自己的应用开发团队，雇佣一些数据科学的专家也并不难，难的是如何让开发工程师和数据科学家理解对方的工作，把他们整合到一个团队中，从而真正创造出价值。

本书作者对数据有天生的热情，且通过在各种行业的工作经历形成了对不同角色的理解，更拥有多年的实际数据分析应用开发经验。在他的职业生涯中，也曾遇到过很多公司在尝试运用数据分析时会遇到的各种问题。如今，他在咨询公司工作，专门帮助各种公司进行大数据方面的数据分析。本书正是作者多年经验的总结与升华，涵盖了从团队建设、工作制度到工具选择、任务划分与执行的方方面面。本书还以一个完整的项目为例，贯穿全书，展示了敏捷数据科学的整个流程，这也是我最喜欢本书的地方。从具体案例出发，让有需求的读者能够更快地依葫芦画瓢，也让初学者能够从做中学，让读者能根据自己的感受，更好地领会作者提出的“敏捷数据科学”的精髓。

Spark 是当前大数据领域最为主流的项目，有着远超 Hadoop MapReduce 的性能，可以说是大数据领域的事实标准。能有今天的地位，Spark 的易用性功不可没。正因此，作者在本书中选择了 Spark 作为大数据处理框架。易于上手的 Spark 确实是敏捷项目的不二选择。不过，Spark 虽然兼具易用与高性能的特点，但不代表 Spark 的性能不会在实际应用中出现问题。事实上，随着业务日趋复杂，Spark 应用也会遇到各种不能通过扩展集群规模来解

决的问题，并不是用上 Spark 就代表算法能够适用于海量数据的场景了，这也是本书缺失的部分。不过不用过于担心，高性能的算法与集群架构也都是慢慢演进出来的，不妨让我们在下一个敏捷冲刺中不断完善我们的应用吧！

翻译本书的时候我还在英特尔工作，英特尔是我工作的第一家公司，在这里我有幸从 2012 年起接触大数据，并从 2014 年起就接触到 Spark。虽然我已经离开，但我会始终感激和怀念在英特尔成长的时光。本书的翻译主要在周末和假期完成，感谢家人和朋友们对我的关心和理解。感谢博文视点的刘恩惠老师和顾慧芳老师在本书审校工作中的辛勤付出，感谢张玲老师引荐我翻译本书，也感谢刘恩惠老师和张玲老师在我拖稿时对我的宽容和鼓励。

这是一个发展迅猛的领域，本书出版时，书中的许多工具（比如 Spark）可能已经又有了很多更新；本书所涉及的远远不止大数据和数据分析，还包括前端开发、团队管理等内容。由于我水平有限，难免有纰漏之处，希望读者能不吝指正，有疑惑之处，不妨也与我探讨。我的邮箱是 me@daoyuan.wang。

王道远

2018 年夏

前言

写作本书第 1 版的那段日子里，我刚好因为一次车祸而残疾，每天忍受疼痛折磨，双手也有些不听使唤。当时，一个叫作“职业浏览器”的项目的失败经历正困扰着我，为了从阴影中走出来，我用 iPad 在床上和沙发上写完了本书，尽管那时我的手都没办法切菜了。我在那个项目发布前几周受了伤，还想着坚持把项目做上线，日夜奋战，非常痛苦。在做项目的过程中，我们犯了许多低级错误，让我一直垂头丧气。最终产品糟透了。项目失败的挫折感不时让我难受，而我背部的慢性疼痛更是很少放过我。我的心脏也出了一些问题，心率下降了三分之一，记忆力也出现了衰退。我仿佛进入了一个幽暗的空间，难以找到出路。我要恢复起来，与失败抗争。说来有些奇怪，为了让自己恢复，我写了第 1 版书。我要把我能给团队同事的指导写下来，确保下一个项目成功。我想让自己摆脱这段经历。更重要的是，我想通过帮助别人，让我的人生重新获得意义，不让自己被残疾击垮。这样一件为大众服务以确保其他人不会重复我的错误的好事，我认为是值得去做的。那个失败项目暴露出了一个比我自身的处境更严重的问题，那就是大多数研究都停留在纸面上，从未让能够获益的人实际使用到。这本书就是一剂良方，是应用性研究的方法论，让研究成果能以产品的形式真正面世。

虽然听起来有些戏剧性，但我还是想在介绍第 2 版之前提一提写第 1 版时的个人情况。尽管那一版书对我来说有特殊的意义，但对于数据科学这个欣欣向荣的领域而言只做出了很小的贡献。但是我为它而自豪。我在那本书中获得了救赎，它让我重新找回了感觉，让我及时从病痛中恢复，让我摆脱失败的痛苦而获得了成就的喜悦，这就是第 1 版的情况。

在第 2 版中，我希望能做到更多。简单地说，我希望能引导初出茅庐的数据科学家，让其快速成长为数据分析应用开发者。我把自己在三个 Hadoop 团队与一个 Spark 团队中获得的构建分析应用的经验进行了总结和提炼。这次改版中，编程语言使用的是数据科学的通

用语言 Python，而选择的大数据平台是 Spark。希望本书能成为读者的必备指南，让读者快速学会如何构建足以应对各种数据规模的分析应用。

Spark 取代 Hadoop/MapReduce 成为了处理大规模数据的主流方式，因此我们在这一版中使用 Spark 来讲解。不仅如此，根据我们团队在工作中对敏捷数据科学的进一步理解，本书对敏捷数据科学方法论的理论和发展也做了进一步完善。希望第 1 版的读者还可以从第 2 版中获得提高，也希望比起相对更适合 Hadoop 用户阅读的第 1 版，这一版能更好地服务于 Spark 用户。

*敏捷数据科学*有两大目标：一是为了使用 Python 和 Spark 搭建出任意规模的数据分析应用，二是帮助产品团队学会使用敏捷的方式协作开发分析应用来保障工作成效。

敏捷数据科学的邮件列表

你可以在邮件列表（*agile-data-science@googlegroups.com*）或网页（*https://groups.google.com/d/forum/agile-data-science*）中学到最新的敏捷数据科学知识。

我为本书维护了一个网页（*http://datasyndrome.com/book*），里面有最新的更新，以及为读者准备的相关资料。

产品分析咨询公司 Data Syndrome

我创办了一家叫作 Data Syndrome 的咨询机构（见图 P-1）来推广本书中的方法论和技术栈。如果你要在你的公司里实践敏捷数据科学并且需要这方面的帮助，或者是需要构建数据产品方面的帮助，又或者需要“大数据”方面的培训，你可以通过我的邮箱（*rjurney@datasyndrome.com*）或网站（*http://llc.datasyndrome.com/*）来联系我。

Data Syndrome 提供视频课程《使用 Kafka、PySpark、Spark MLlib 和 Spark Streaming 进行实时预测分析》（*Realtime Predictive Analytics with Kafka,PySpark, Spark MLlib and Spark Streaming. http://datasyndrome.com/video*），使用了第 7 章和第 8 章的材料，教观看者如何用 Kafka、Spark Streaming 及网络应用的前端页面构建出整套的实时预测系统（见图 P-2）。如果想进一步了解，请访问 *http://datasyndrome.com/video* 或联系 *rjurney@datasyndrome.com*。

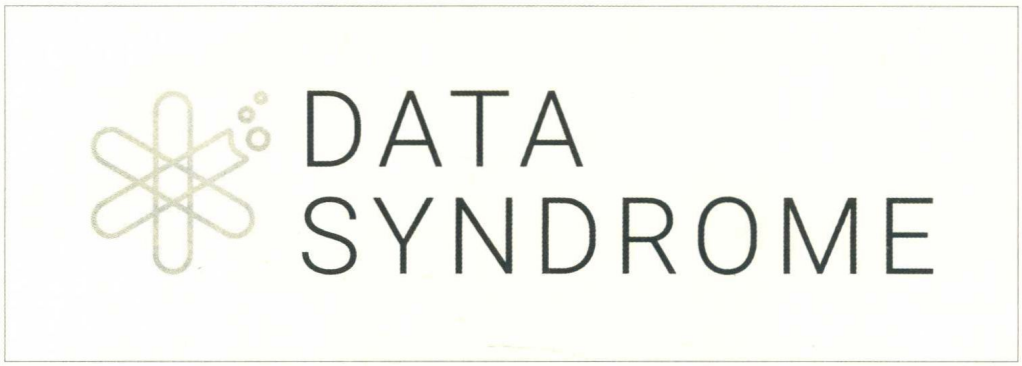

图 P-1　Data Syndrome

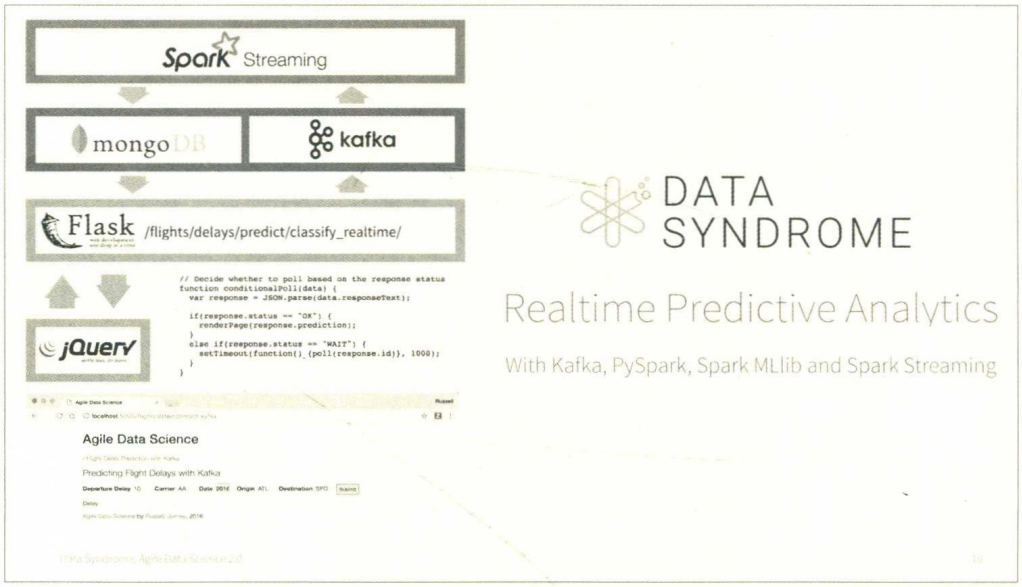

图 P-2　实时预测分析视频课程

在线培训

Data Syndrome 正在研制针对数据科学团队和数据工程团队的全套在线大数据培训课程。目前提供的课程可以根据需求进行自定义，包括以下几个主题。

敏捷数据科学

持续三天的课程，涵盖了全栈分析应用的构建。在内容上与本书相近，可以将数据科学家训练为全栈应用开发者。

实时预测分析

一天即可完成，时长总共 6 小时。包括如何使用 Kafka 和 Spark Streaming 及网络应用前端页面构建整套的实时预测系统。

PySpark 介绍

一天即可完成，时长 3 小时。向参与者介绍如何使用 Spark 的 Python 接口进行基本的数据处理。最终教会参与者如何使用 Spark MLlib 构建一个分类器模型来预测航班延误。

详情请访问 *http://datasyndrome.com/training* 或联系 *rjurney@datasyndrome.com*。

本书目标读者

本书的目的是帮助初学者和初出茅庐的数据科学家成长为数据科学与数据分析团队的主力成员。本书想要帮助工程师、分析师、数据科学家以敏捷的方式来使用 Hadoop 在大数据上进行工作。本书介绍的敏捷方法论很适合大数据领域。

本书是为需要开发软件来分析数据的程序员而写的。设计师和产品经理可能更适合第 1 章、第 2 章和第 5 章，这些章节主要作为敏捷过程的导论，没有专注于编码运行。

本书假设你在类 UNIX 环境中工作，没有为 Windows 用户提供示例，不过 Windows 用户可以使用 Cygwin 尝试。

本书主要结构

本书分为两个部分。第 I 部分介绍的是我们在第 II 部分中需要用到的数据集和工具集。第 I 部分故意写得简明扼要，只是为了尽可能快地介绍这些工具。第 II 部分会更深入地探讨这些工具的使用，所以如果在读第 I 部分时感觉有些不知所措也不用担心。第 I 部分的章节如下。

第 1 章 理论

介绍敏捷数据科学的方法论。

第 2 章 敏捷工具

介绍要用的工具集，并且讲解工具如何上手与安装。

第 3 章 数据

描述本书中使用的数据集。

第Ⅱ部分是我们使用敏捷数据科学来构建一个分析应用的教程。这是一份笔记本式的分析应用构建指南。我们逐层攀登数据价值金字塔，始终应用敏捷的原则。这一部分会展示在敏捷迭代进程中一步一步发掘数据价值的方法。第Ⅱ部分由以下所列章节组成。

第 4 章 记录收集与展示

帮你下载航班数据，并且通过网络应用展示航班记录。

第 5 章 使用图表进行数据可视化

一步步引导你如何在网络应用中加入一些简单的图表来展示数据。

第 6 章 通过报表探索数据

教你如何从数据中提取出实体关系，将其参数化并相互关联以创建交互式的报表。

第 7 章 进行预测

在先前所做的基础上对某一航班准点与否进行预测。

第 8 章 部署预测系统

展示如何部署预测系统来确保真正发挥作用。

第 9 章 改进预测结果

不断迭代提高我们的准点航班预测应用的表现。

附录 A 安装手册

展示如何安装所需工具。

本书样式约定

本书使用以下所列样式约定。

斜体（*Italic*）

表示新术语、URL、电子邮箱地址、文件名、文件扩展。

等宽字体（`Constant width`）

用于程序示例，还有在文字中引用的程序中的内容，比如变量名或函数名、数据库、数据类型、环境变量、语句，还有关键字。

加粗等宽字体（**`Constant width bold`**）

表示需要用户按字面输入的命令或其他文本。

等宽斜体（*`Constant width italic`*）

表示需要以用户提供的值或上下文决定的值替换的文本。

本图标表示一个提示、建议或注释。

本图标表示一个警告或提醒。

代码示例的使用

补充材料(代码示例、练习等)可以在 *https://github.com/rjurney/Agile_Data_Code_2* 中下载到。

本书是要帮你解决问题的。总的来说，你可以在你的程序或文档中直接使用本书所提供的示例代码。除非要大部分照搬代码，否则你不需要联系我们获取许可。举例来说，写一个程序时从本书中抄几段代码片段不需要许可。出售或分发 O'Reilly 书籍代码示例的光碟也不需要许可。在回答问题时引用本书并且摘抄示例代码不需要许可。但是如果要在你的产品文档中大量使用示例代码，你就需要申请许可。

我们鼓励但是不要求标明出处。通常情况下，引用要包含题目、作者、出版商、ISBN 码等信息。你可以这样标明对本书的引用："*Agile Data Science 2.0* by Russell Jurney (O'Reilly). Copyright 2017 Data Syndrome LLC, 978-1-491-96011-0."

如果你觉得你对示例代码的使用超出了正常的范围或上面列出的许可范围，请通过 *permissions@oreilly.com* 联系我们。

O'Reilly Safari

Safari (前身为 Safari Books Online) 是一个会员制的培训与参考平台，为企业、政府、教育机构和个人提供服务。

会员可以访问数以千计的书籍、培训视频、学习路径、交互式指南、编排的播放列表等来自超过 250 家出版商的资料。这些出版商包括 O'Reilly Media、Harvard Business Review、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Adobe、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 及其他一些出版商。

详情请访问 *http://oreilly.com/safari*。

如何联系我们

关于本书的意见和建议，可以通过以下方式联系出版商：

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

对于本书的评论或技术性问题，你可以发邮件到 *bookquestions@oreilly.com*。

关于我们的书籍、课程、会议、新闻的更多信息，请访问我们的网站 *http://www.oreilly.com*。

在 Facebook 上找到我们：*http://facebook.com/oreilly*。

在 Twitter 上关注我们：*http://twitter.com/oreillymedia*。

在 YouTube 上观看我们：*http://www.youtube.com/oreillymedia*。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **下载资源**：本书如提供示例代码及资源文件，均可在下载资源处下载。
- **提交勘误**：您对书中内容的修改意见可在提交勘误处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方读者评论处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/35166

目录

第 II 部分　攀登金字塔