O'REILLY®

TURING 图灵程序设计丛书

Spark

高级数据分析

Advanced Analytics with Spark

[美] Sandy Ryza [美] Uri Laserson
[英] Sean Owen [美] Josh Wills 著
龚少成 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

Spark高级数据分析

Advanced Analytics with Spark

[美] Sandy Ryza [美] Uri Laserson [英] Sean Owen [美] Josh Wills 著
龚少成 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

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

Spark高级数据分析 /（美）里扎（Ryza, S.）等著；
龚少成译. -- 北京：人民邮电出版社，2015.11
（图灵程序设计丛书）
ISBN 978-7-115-40474-9

Ⅰ. ①S… Ⅱ. ①里… ②龚… Ⅲ. ①数据处理软件
Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2015)第225780号

内 容 提 要

本书是使用 Spark 进行大规模数据分析的实战宝典，由著名大数据公司 Cloudera 的数据科学家撰写。四位作者首先结合数据科学和大数据分析的广阔背景讲解了 Spark，然后介绍了用 Spark 和 Scala 进行数据处理的基础知识，接着讨论了如何将 Spark 用于机器学习，同时介绍了常见应用中几个最常用的算法。此外还收集了一些更加新颖的应用，比如通过文本隐含语义关系来查询 Wikipedia 或分析基因数据。

本书适合从事大数据分析的各类专业人员阅读。

◆ 著　[美] Sandy Ryza [美] Uri Laserson
[英] Sean Owen [美] Josh Wills
译　龚少成
责任编辑　岳新欣
执行编辑　李松峰
责任印制　杨林杰

◆ 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
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：15.25
字数：360千字　2015年11月第1版
印数：1－4 000册　2015年11月北京第1次印刷
著作权合同登记号　图字：01-2015-3956号

定价：59.00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：京崇工商广字第 0021 号

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*

推荐序

数据的爆炸式增长和隐藏在这些数据背后的商业价值催生了一代又一代的大数据处理技术。十年前 Hadoop 横空出世，Cloudera 首席架构师 Doug Cutting 先生将 Google 的 MapReduce 思想用开源的方式实现出来，由此拉开了基于 MapReduce 的大数据处理框架在企业中应用的序幕。最近几年，Hadoop 生态系统又发展出以 Spark 为代表的新计算框架。相比 MapReduce，Spark 速度快，开发简单，并且能同时兼顾批处理和实时数据分析。Spark 起源于加州大学伯克利分校的 AMPLab，Cloudera 公司作为大数据市场上的翘楚很早就开始将 Spark 推广到广大企业级客户并积累了大量的经验。Advanced Analysis with Spark 一书正是这些经验的结晶。另一方面，企业级用户在引入 Spark 技术时碰到的最大难题之一就是能够灵活应用 Spark 技术的人才匮乏。听闻 Cloudera 中国公司的龚少成在与图灵公司一起为 Advanced Analysis with Spark 一书的中文版在日夜奋战，我便欣然作序，也算是为国内企业更好地应用 Spark 技术尽自己的一份力量！

本书开篇介绍了 Spark 的基础知识，然后详细介绍了如何将 Spark 应用到各个行业。与许多书籍只着重描述最终方案不同，本书作者在介绍案例时把解决问题的整个过程也展现了出来。在介绍一个主题时，并不是一开始就给出最终方案，而是先给出一个最初并不完善的方案，然后指出方案的不足，引导读者思考并逐步改进，最终得出一个相对完善的方案。这体现了工程问题的解决思路，也体现了大数据分析是一个迭代的过程，这样的论述方式更能激发读者的思考，这一点实在难能可贵。

本书英文版自出版以来在亚马逊网站大数据分析类书籍中一直名列前茅，而且获得的多为五星级评价，可见国外读者对该书的喜爱。本书中文版译者龚少成技术扎实，在英特尔和 Cloudera 工作期间带领团队成功实施过许多大数据平台项目，而且其英语功底也相当扎实，此外我偶然得知他还是国内少数通过高级口译考试的专业人才。所以本书的中文版交给龚少成翻译实在是件让人欣慰的事情。本书中文版初稿也证实了我的判断，不仅保持了英文版的风格，而且语言也十分流畅。如果你了解 Scala 语言，还有一些统计学和机器学习基础，那么本书是你学习 Spark 时必备的书籍之一！

——苗凯翔，Cloudera 公司副总裁

译者序

大数据是这几年科技和应用领域炙手可热的话题，而 Spark 又是大数据领域里最活跃的技术。对 Spark 这个技术，国内研究比较多的是原理和源代码，而许多客户抱怨 Spark 应用落地难。造成这一现象的一个主要原因是 Spark 技术比较新，许多应用还处在探索阶段。Cloudera 公司作为全球大数据领域的领头羊，在给全球客户提供最高质量大数据平台的同时，也积累了许多 Spark 应用方面的宝贵经验。本书四位作者均为 Cloudera 公司的数据科学家，也长期为客户提供专业的数据分析服务。可以说，本书的出版将为 Spark 数据分析项目的落地起到巨大的推动作用。

同时我也注意到，国内 Spark 数据分析方面的书籍少，而且许多书籍都停留在源代码研究的层面上。当然，这些书中也不乏非常优秀的作品，但我认为 Spark 真正的力量在于其开发的大数据应用。所以早在本书还处于初期编写过程中时，我就自告奋勇和作者联系中文版事宜，希望以此为中国的大数据分析事业略尽绵力。

本书在翻译过程中得到了许多人的帮助。首先要感谢我在 Cloudera 公司的同事，也就是本书的四位作者。在本书的翻译过程中，由于不同语言的习惯问题，四位作者 Sandy Ryza、Uri Laserson、Sean Owen 和 Josh Wills 花了许多时间和我交流。本人之所以有幸负责本书的中文版翻译，也是承蒙 Sean Owen 的引荐。感谢 Cloudera 公司全球副总裁凌琦先生和苗凯翔博士，没有两位领导的努力，Cloudera 中国区团队不可能如此迅速组建并形成如此强大的战斗力，我也无法参与到轰轰烈烈的大数据事业中。感谢我的同事田占凤博士和陈建忠的鼓励，中文版的翻译工作才得以开始。英特尔亚太研发公司工程师邱鑫对本书初稿的修改贡献了许多宝贵建议。同时本书在翻译过程中还得到了 Cloudera 公司中国区同事刘贺峰、糜君、陈飚、陈新江、李大超和张莉苹的鼎力帮助。感谢图灵公司的李松峰编辑和岳新欣编辑在翻译过程中的指导和仔细审阅。由于本书的翻译都是在周末完成的，所以要特别感谢我的妻子周幼琼在每个周末对我的照顾。

由于本人的水平有限，同时本书涉及许多课题，所以现有译文中难免存在纰漏之处。希望读者能够不吝赐教，发现问题时麻烦和我联系。邮件请发送至 gongshaocheng@gmail.com。

龚少成

2015 年 7 月于上海

序

自从在加州大学伯克利分校创立 Spark 项目起，我就时常心潮澎湃。不仅因为 Spark 可以帮助人们快速构建并行系统，更因为 Spark 帮助了越来越多的人使用大规模计算。因此看到这本介绍 Spark 高级分析的书，我非常欣慰！该书由数据科学领域四位专家 Sandy、Uri、Sean 和 Josh 携手打造。四位作者研习 Spark 已久，他们在本书中跟读者分享了关于 Spark 的大量精彩内容，同时本书的案例部分同样出众！

对于这本书，我最钟爱的是它强调案例，而且这些案例都源于现实数据和实际应用。找到一个像样的、能在笔记本电脑上运行的大数据案例已经很难，更遑论十个了。但本书作者做到了！作者为大家准备好了一切，只等你在 Spark 中运行它们。更难能可贵的是，作者不仅讨论了核心算法，更倾心于数据准备和模型调优，没有这些工作，实际项目中就无法得到好的结果。认真研读此书，你应该可以吸收这些案例中的概念并直接将其运用在自己的项目中！

大数据处理无疑是当今计算领域最激动人心的方向之一，发展非常迅猛，新思想层出不穷。愿本书能帮助你在这个崭新的领域中扬帆启航！

Matei Zaharia

Databricks 公司 CTO 兼 Apache Spark 项目副总裁

前言

作者：Sandy Ryza

我不想我的人生有很多遗憾。2011 年的某个慵懒的时刻，我在正绞尽脑汁地想如何把高难度的离散优化问题最优地分配给计算机集群处理，真是很难想到有什么好方法。我的导师跟我讲，他听说有个叫 Spark 的新技术，可我基本上没当回事。Spark 的想法太好了，让人觉得有点儿不靠谱。就这样，我很快又回去接着写 MapReduce 的本科毕业论文了。时光荏苒，Spark 和我都渐渐成熟，但我们之间有一个已然成为冉冉之星，这让人不禁感叹“Spark”（星星之火）这个双关语是多么贴切。两年后，Spark 的价值举世皆知！

Spark 的前辈有很多，从 MPI 到 MapReduce。利用这些计算框架，我们写的程序可以充分利用大量资源，但不需要关心分布式系统的实现细节。数据处理的需求促进了这些技术框架的发展。同样，大数据领域也和这些框架关系密切，这些框架界定了大数据的范围。Spark 有望更进一步，让写分布式程序就像写普通程序一样。

Spark 能大大提升 ETL 流水作业的性能，并把 MapReduce 程序员从每天问天天不灵、问地地不应的绝望痛苦中解救出来。对我而言，Spark 的激动人心之处在于，它真正打开了复杂数据分析的大门。Spark 带来了支持迭代式计算和交互式探索的模式。利用这一开源计算框架，数据科学家终于可以在大数据集上高效地工作了。

我认为数据科学教学最有效的方法是利用实例。为此，我和同事一起编撰了这本关于实际应用的书，希望它能涵盖大规模数据分析中最常用的算法、数据集和设计模式。阅读本书时不必一页一页地看，可以根据工作需要或按兴趣直接翻到相关章节。

本书内容

第 1 章结合数据科学和大数据分析的广阔背景来讨论 Spark。随后各章在介绍 Spark 数据

分析时都自成一体。第 2 章通过数据清洗这一使用场景来介绍用 Spark 和 Scala 进行数据处理的基础知识。接下来几章深入讨论如何将 Spark 用于机器学习，介绍了常见应用中几个最常用的算法。其余几章则收集了一些更新颖的应用，比如通过文本隐含语义关系来查询 Wikipedia 或分析基因数据。

使用代码示例

补充材料（代码示例、练习、勘误表等）可以从 https://github.com/sryza/aas 下载。

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，否则无需联系我们获得许可。比如，用本书的几个代码片段写一个程序就无需获得许可，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需获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

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："*Advanced Analytics with Spark* by Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills (O'Reilly). Copyright 2015 Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills, 978-1-491-91276-8."

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。

Safari® Books Online

Safari Books Online（http://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 以及其他几十家出版社（https://www.safaribooksonline.com/our-library/）

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。本书的网站地址是：

http://shop.oreilly.com/product/0636920035091.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

如果没有 Apache Spark 和 MLlib，就没有本书。所以我们应该感谢开发了 Spark 和 MLlib 并将其开源的团体，也要感谢那些添砖加瓦的数以百计的代码贡献者。

我们还要感谢本书的每一位审阅者，感谢他们花费了大量的时间来审阅本书的内容，感谢他们的专业视角，他们是 Michael Bernico、Ian Buss、Jeremy Freeman、Chris Fregly、Debashish Ghosh、Juliet Hougland、Jonathan Keebler、Frank Nothaft、Nick Pentreath、Kostas Sakellis、Marcelo Vanzin 和另一位 Juliet Hougland。谢谢你们所有人！我们欠你们一个大人情！你们的努力大大改进了本书的结构和质量。

我（Sandy）还要感谢 Jordan Pinkus 和 Richard Wang，你们帮助我完成了风险分析章节的原理部分。

感谢 Marie Beaugureau 和 O'Reilly 出版社在本书出版和发行过程中贡献的宝贵经验和大力支持！

目录

推荐序 ····· ix

译者序 ····· xi

序 ····· xiii

前言 ····· xv

第 1 章　大数据分析 ····· 1

1.1　数据科学面临的挑战 ····· 2
1.2　认识 Apache Spark ····· 4
1.3　关于本书 ····· 5

第 2 章　用 Scala 和 Spark 进行数据分析 ····· 7

2.1　数据科学家的 Scala ····· 8
2.2　Spark 编程模型 ····· 9
2.3　记录关联问题 ····· 9
2.4　小试牛刀：Spark shell 和 SparkContext ····· 10
2.5　把数据从集群上获取到客户端 ····· 15
2.6　把代码从客户端发送到集群 ····· 18
2.7　用元组和 case class 对数据进行结构化 ····· 19
2.8　聚合 ····· 23
2.9　创建直方图 ····· 24
2.10　连续变量的概要统计 ····· 25
2.11　为计算概要信息创建可重用的代码 ····· 26
2.12　变量的选择和评分简介 ····· 30
2.13　小结 ····· 31

第 3 章　音乐推荐和 Audioscrobbler 数据集……33
3.1　数据集……34
3.2　交替最小二乘推荐算法……35
3.3　准备数据……37
3.4　构建第一个模型……39
3.5　逐个检查推荐结果……42
3.6　评价推荐质量……43
3.7　计算 AUC……44
3.8　选择超参数……46
3.9　产生推荐……48
3.10　小结……49
第 4 章　用决策树算法预测森林植被……51
4.1　回归简介……52
4.2　向量和特征……52
4.3　样本训练……53
4.4　决策树和决策森林……54
4.5　Covtype 数据集……56
4.6　准备数据……57
4.7　第一棵决策树……58
4.8　决策树的超参数……62
4.9　决策树调优……63
4.10　重谈类别型特征……65
4.11　随机决策森林……67
4.12　进行预测……69
4.13　小结……69
第 5 章　基于 *K* 均值聚类的网络流量异常检测……71
5.1　异常检测……72
5.2　*K* 均值聚类……72
5.3　网络入侵……73
5.4　KDD Cup 1999 数据集……73
5.5　初步尝试聚类……74
5.6　*K* 的选择……76
5.7　基于 R 的可视化……79
5.8　特征的规范化……81
5.9　类别型变量……83
5.10　利用标号的熵信息……84
5.11　聚类实战……85
5.12　小结……86

第 6 章　基于潜在语义分析算法分析维基百科 ····· 89
6.1　词项 – 文档矩阵 ····· 90
6.2　获取数据 ····· 91
6.3　分析和准备数据 ····· 92
6.4　词形归并 ····· 93
6.5　计算 TF-IDF ····· 94
6.6　奇异值分解 ····· 97
6.7　找出重要的概念 ····· 98
6.8　基于低维近似的查询和评分 ····· 101
6.9　词项 – 词项相关度 ····· 102
6.10　文档 – 文档相关度 ····· 103
6.11　词项 – 文档相关度 ····· 105
6.12　多词项查询 ····· 106
6.13　小结 ····· 107
第 7 章　用 GraphX 分析伴生网络 ····· 109
7.1　对 MEDLINE 文献引用索引的网络分析 ····· 110
7.2　获取数据 ····· 111
7.3　用 Scala XML 工具解析 XML 文档 ····· 113
7.4　分析 MeSH 主要主题及其伴生关系 ····· 114
7.5　用 GraphX 来建立一个伴生网络 ····· 116
7.6　理解网络结构 ····· 119
7.6.1　连通组件 ····· 119
7.6.2　度的分布 ····· 122
7.7　过滤噪声边 ····· 124
7.7.1　处理 EdgeTriplet ····· 125
7.7.2　分析去掉噪声边的子图 ····· 126
7.8　小世界网络 ····· 127
7.8.1　系和聚类系数 ····· 128
7.8.2　用 Pregel 计算平均路径长度 ····· 129
7.9　小结 ····· 133
第 8 章　纽约出租车轨迹的空间和时间数据分析 ····· 135
8.1　数据的获取 ····· 136
8.2　基于 Spark 的时间和空间数据分析 ····· 136
8.3　基于 JodaTime 和 NScalaTime 的时间数据处理 ····· 137
8.4　基于 Esri Geometry API 和 Spray 的地理空间数据处理 ····· 138
8.4.1　认识 Esri Geometry API ····· 139
8.4.2　GeoJSON 简介 ····· 140
8.5　纽约市出租车客运数据的预处理 ····· 142

8.5.1 大规模数据中的非法记录处理 ····· 143
8.5.2 地理空间分析 ····· 147
8.6 基于 Spark 的会话分析 ····· 149
8.7 小结 ····· 153
第 9 章 基于蒙特卡罗模拟的金融风险评估 ····· 155
9.1 术语 ····· 156
9.2 VaR 计算方法 ····· 157
9.2.1 方差 – 协方差法 ····· 157
9.2.2 历史模拟法 ····· 157
9.2.3 蒙特卡罗模拟法 ····· 157
9.3 我们的模型 ····· 158
9.4 获取数据 ····· 158
9.5 数据预处理 ····· 159
9.6 确定市场因素的权重 ····· 162
9.7 采样 ····· 164
9.8 运行试验 ····· 167
9.9 回报分布的可视化 ····· 170
9.10 结果的评估 ····· 171
9.11 小结 ····· 173
第 10 章 基因数据分析和 BDG 项目 ····· 175
10.1 分离存储与模型 ····· 176
10.2 用 ADAM CLI 导入基因学数据 ····· 178
10.3 从 ENCODE 数据预测转录因子结合位点 ····· 185
10.4 查询 1000 Genomes 项目中的基因型 ····· 191
10.5 小结 ····· 193
第 11 章 基于 PySpark 和 Thunder 的神经图像数据分析 ····· 195
11.1 PySpark 简介 ····· 196
11.2 Thunder 工具包概况和安装 ····· 199
11.3 用 Thunder 加载数据 ····· 200
11.4 用 Thunder 对神经元进行分类 ····· 207
11.5 小结 ····· 211
附录 A Spark 进阶 ····· 213
附录 B 即将发布的 MLlib Pipelines API ····· 221
作者介绍 ····· 226
封面介绍 ····· 226