HZ BOOKS
华章教育
Pearson

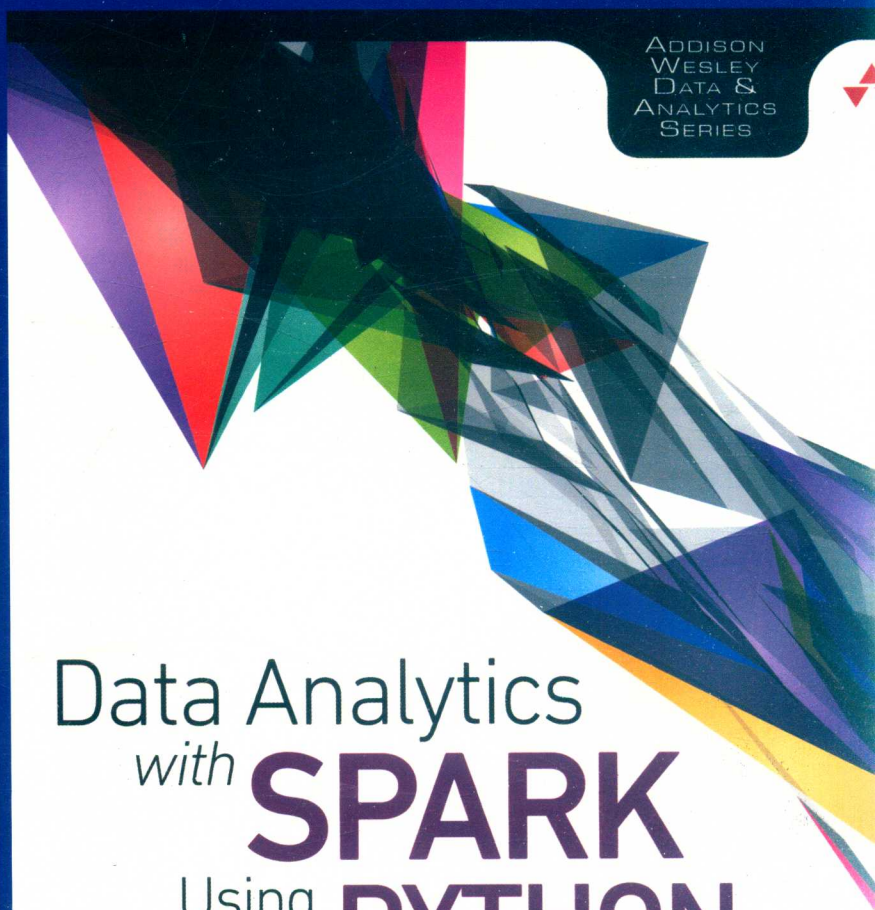
经 典 原 版 书 库

Spark数据分析

基于Python语言

[澳] 杰夫瑞·艾文（Jeffrey Aven） 著

（英文版）



Data Analytics with SPARK Using PYTHON

JEFFREY AVEN

机械工业出版社
China Machine Press

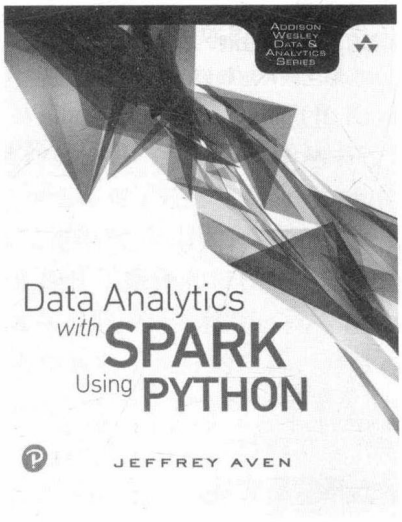
经 库

Spark数据分析

基于Python语言

（英文版）

Data Analytics with Spark Using Python



[澳] 杰夫瑞·艾文（Jeffrey Aven） 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Spark 数据分析：基于 Python 语言（英文版）/（澳）杰夫瑞·艾文（Jeffrey Aven）著．—北京：机械工业出版社，2019.4
（经典原版书库）
书名原文：Data Analytics with Spark Using Python

ISBN 978-7-111-62003-7

Ⅰ. S… Ⅱ. 杰… Ⅲ. 数据处理软件－英文 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2019）第 031718 号

本书版权登记号：图字 01-2018-7390

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：张梦玲　　责任校对：李秋荣
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2019 年 3 月第 1 版第 1 次印刷
开　　本：186mm×240mm　1/16　　印　　张：18.25
书　　号：ISBN 978-7-111-62003-7　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

出版者的话

文艺复兴以来，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的优势，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson、McGraw-Hill、Elsevier、MIT、John Wiley & Sons、Cengage 等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出 Andrew S. Tanenbaum、Bjarne Stroustrup、Brian W. Kernighan、Dennis Ritchie、Jim Gray、Afred V. Aho、John E. Hopcroft、Jeffrey D. Ullman、Abraham Silberschatz、William Stallings、Donald E. Knuth、John L. Hennessy、Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力相助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专门为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近 500 个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www.hzbook.com

电子邮件：hzjsj@hzbook.com

联系电话：（010）88379604

联系地址：北京市西城区百万庄南街 1 号

邮政编码：100037

华章教育

华章科技图书出版中心

前　言

Spark 在这场由大数据与开源软件掀起的颠覆性革命中处于核心位置。不论是尝试 Spark 的意向还是实际用例的数量都在以几何级数增长，而且毫无衰退的迹象。本书将手把手引导你在大数据分析领域中收获事业上的成功。

本书重点

本书重点关注 Spark 项目的基本知识，从 Spark 核心开始，然后拓展到各种 Spark 扩展、Spark 相关项目及子项目，以及 Spark 所处的丰富的生态系统里各种别的开源技术，比如 Hadoop、Kafka、Cassandra 等。

尽管对于本书所介绍的 Spark 基本概念（包括运行环境、集群架构、应用架构等）的理解是与编程语言无关且透明的，本书中大多数示例程序和练习是用 Python 实现的。Spark 的 Python API（PySpark）为数据分析师、数据工程师、数据科学家等提供了易用的编程环境，让开发者能在获得 Python 语言的灵活性和可扩展性的同时，获得 Spark 的分布式处理能力和伸缩性。

本书所涉及的范围非常广泛，涵盖了从基本的 Spark 核心编程到 Spark SQL、Spark Streaming、机器学习等方方面面的内容。本书对于每个主题都给出了良好的介绍和概览，足以让你以 Spark 项目为基础构建出针对任何特定领域或学科的平台。

目标读者

本书是为有志进入大数据领域或是已经入门想要进一步巩固大数据领域知识的数据分析师和工程师而写的。当前市场对于具备大数据技能、懂得大数据领域优秀处理框架 Spark 的工程师的需求特别大。本书的目标是针对这一不断增长的雇佣市场需求培训读者，使得读者获得雇主亟需的技能。

对于阅读本书来说，有 Python 使用经验是有帮助的，没有的话也没关系，毕竟 Python 对于任何有编程经验的人来说都非常直观易懂。读者最好对数据分析和数据处理有一定了解。这本书尤其适合有兴趣进入大数据领域的数据仓库技术人员阅读。

如何使用本书

本书分为两大部分共 8 章。第一部分“Spark 基础”包括 4 章，会让读者深刻理解 Spark 是什么，如何部署 Spark，如何使用 Spark 进行基本的数据处理操作。

第 1 章提供了大数据生态圈的概览，包括 Spark 项目的起源和演进过程。本章讨论了 Spark 项目的关键属性，包括 Spark 是什么，用起来如何，以及 Spark 与 Hadoop 项目之间的关系。

第 2 章展示了如何部署一个 Spark 集群，包括 Spark 集群的各种部署模式，以及调用

Spark 的各种方法。

第 3 章讨论 Spark 集群和应用是如何运作的，让读者深刻理解 Spark 是如何工作的。

第 4 章关注使用弹性分布式数据集（RDD）进行 Spark 初级编程的基础。

第二部分“基础拓展”包括后四章的内容，扩展到 Spark 的核心模块以外，包括 SQL 和 NoSQL 系统、流处理应用、数据科学与机器学习中 Spark 的使用：

第 5 章讲解了用来扩展、加速和优化常规 Spark 例程的高级元件，包括各种共享变量和 RDD 存储，以及分区的概念及其实现。

第 6 章讨论 Spark 与广袤的 SQL 土壤的整合，还有 Spark 与非关系型存储的整合。

第 7 章介绍了 Spark 的 Streaming 子项目，以及 Streaming 中最基本的 DStream 对象。本章还涵盖了 Spark 对于 Apache Kafka 这样的常用消息系统的使用。

第 8 章介绍了使用 R 语言使用 Spark 建立预测模型，以及 Spark 中用来实现机器学习的子项目 MLlib。

本书代码

本书中各个练习的示例数据和源代码可以从 http://sparkusingpython.com 下载。你也可以从 https://github.com/sparktraining/spark_using_python 查看或者下载本书的 GitHub 代码仓库。

引　言

Spark 是用于大数据的一流的数据处理平台和编程接口，与大数据技术浪潮密不可分。在本书撰写时，Spark 是 Apache 软件基金会（ASF）框架下最活跃的开源项目之一，也是有史以来最活跃的开源大数据项目之一。

数据分析、数据处理以及数据科学社区都对 Spark 有着浓厚兴趣，因而理解 Spark 是什么，Spark 是用来干什么的，Spark 的优势在哪里，以及如何利用 Spark 进行大数据分析都是很重要的。这本书涵盖了上述内容的方方面面。

与其他专门针对 Spark 的出版物不同，后者几乎仅使用 Scala API，而本书专注于 Spark 的 Python API，也就是 PySpark。选择 Python 作为本书基础语言是因为 Python 是一门直观易懂的解释型语言，广为人知而且对于新手也极易上手。更何况 Python 对于数据科学家而言是一门非常常用的编程语言，而数据科学家在 Spark 社区也是相当大的一个群体。

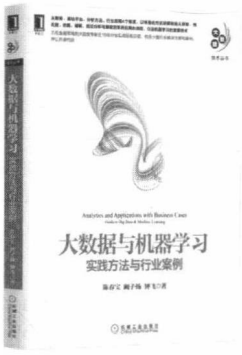
本书会从零开始讲大数据和 Spark，因此无论你是从未接触 Spark 和 Hadoop，还是已经有所接触，正在寻求全面了解 Spark 的运作方式和如何充分利用 Spark 丰富的功能，这本书都是合适的。

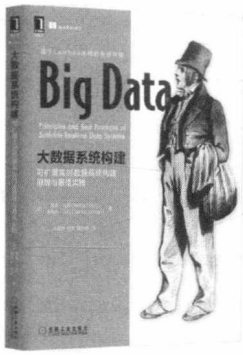
在阅读本书的过程中，你会了解一些相近的或者相互协作的平台、项目以及技术，比如 Hadoop、HBase、Kafka 等，并且看到它们如何与 Spark 交互与集成。

过去的几年中，我的工作都集中在这个领域，包括针对大数据分析进行授课和向客户提供咨询服务。我经历了 Spark 和大数据乃至整个开源运动的出现和成熟，并且参与到了开源软件融入企业使用的进程中。我尽量把个人的学习历程总结到了本书中。

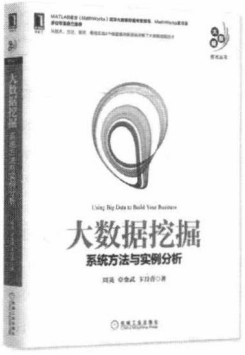
希望这本书能助你开启大数据和 Spark 从业人员的旅程。

推荐阅读

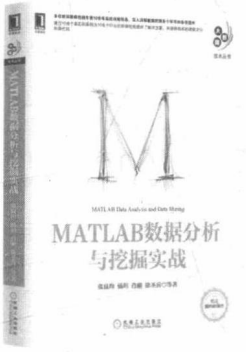

大数据与机器学习
实践方法与行业案例


Big Data
大数据系统构建


BIGDATA
企业大数据系统构建实战
技术、架构、实施与应用


大数据挖掘
系统方法与实例分析


数据挖掘核心技术揭秘


MATLAB数据分析与挖掘实战


Hadoop大数据分析与挖掘实战


Python数据分析与挖掘实战


数据实践之美
31位大数据专家的方法、技术与思想

目　录

第一部分　Spark 基础

第二部分 基础拓展

Contents

II: Beyond the Basics

Spark Foundations