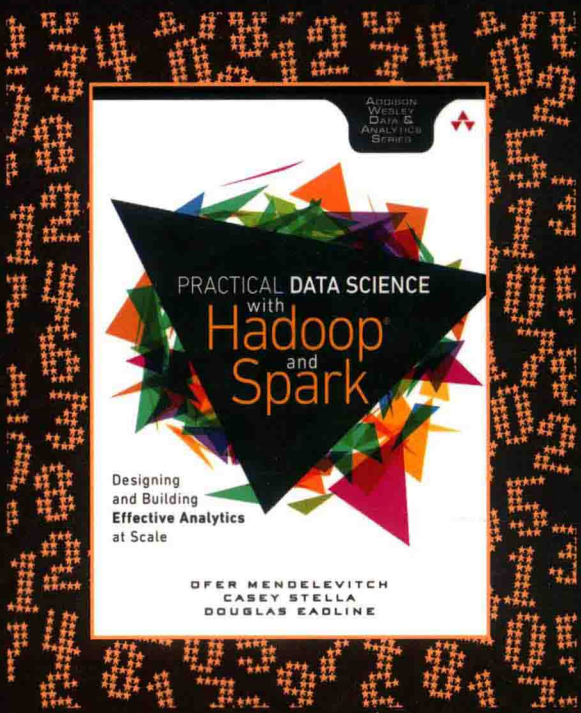
HZ Books 华章IT
Pearson
数据科学与工程技术丛书

数据科学与大数据技术导论

奥弗·曼德勒维奇（Ofer Mendelevitch）
[美] 凯西·斯特拉（Casey Stella） 著
道格拉斯·伊德理恩（Douglas Eadline）
唐金川 译



PRACTICAL DATA SCIENCE WITH HADOOP AND SPARK

DESIGNING AND BUILDING EFFECTIVE ANALYTICS AT SCALE

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PRACTICAL DATA SCIENCE WITH HADOOP AND SPARK

DESIGNING AND BUILDING EFFECTIVE ANALYTICS AT SCALE

数据科学与大数据技术导论

[美] 奥弗·曼德勒维奇（Ofer Mendelevitch）
凯西·斯特拉（Casey Stella） 著
道格拉斯·伊德理恩（Douglas Eadline）

唐金川 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据科学与大数据技术导论 /（美）奥弗·曼德勒维奇（Ofer Mendelevitch）等著；唐金川译．
—北京：机械工业出版社，2018.6
（数据科学与工程技术丛书）
书名原文：Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale

ISBN 978-7-111-60034-3

I. 数… II. ① 奥… ② 唐… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 110207 号

本书版权登记号：图字 01-2017-0908

本书由 3 位资深数据科学家合作撰写，非常适合用来入门数据科学。全书共分三部分，12 章。第一部分（第 1～3 章）概述了数据科学及其历史演变，Hadoop 及其演进史，以及 Hadoop 生态系统中的各种工具；第二部分（第 4～6 章）讨论了将数据集从外部源导入 Hadoop 的各种工具和技术，使用 Hadoop 进行数据再加工，以及大数据的可视化；第三部分（第 7～12 章）介绍了对机器学习的高层次理解，预测建模的基本算法和各种 Hadoop 工具，各种聚类分析，异常检测的各种方法和算法，将数据科学应用于自然语言处理，以及 Hadoop 环境下数据科学的未来。

本书可作为高等院校数据科学专业相关课程的参考教材，也可供数据科学家、数据工程师、开发人员和项目利益相关者参考使用。

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：郎亚妹　　责任校对：殷　虹
印　　刷：中国电影出版社印刷厂　　版　　次：2018 年 6 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：12
书　　号：ISBN 978-7-111-60034-3　　定　　价：59.00 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

译 者 序

互联网，特别是移动互联网的发展，催生了海量的数据（如用户的行为数据、博文、照片、视频等），“大数据”概念应运而生。早些年提出的“数据挖掘”“机器学习”以及如今火热的“人工智能”，都致力于让这些“大数据”发挥越来越大的价值。

让大数据发挥巨大潜力的职位，国外更多叫作“数据科学家”，而国内则更多地细分为数据工程师、数据挖掘工程师、机器学习工程师，抑或泛称算法工程师。虽然各个公司技术栈不尽相同，但 Hadoop 与 Spark 的使用颇为广泛。

本书囊括的内容为 Hadoop 及 Spark 应用方面的从业者提供了比较全面的入门指南。全书分为三部分。

第一部分：数据科学概述及实例介绍，Hadoop 生态环境及相关工具介绍。

第二部分：数据的获取、存储、再加工、探索和可视化。

第三部分：应用数据，内容包括机器学习、预测模型、聚类、异常检测和 NLP。

本书所涵盖的内容有助于读者具备数据科学家的能力。在阅读过程中，如果对某些部分或章节已经了然于胸，则可跳过进而阅读后续内容。本书的不同章节，也可作为专项实践能力锻炼时的参考资料。

此书付梓之际，非常感谢吴怡、关敏两位编辑的指导和督促，也衷心感激爱妻李珂欣在我翻译期间给予我鼓励、体谅和帮助。

书中不少英文术语，国内业界人士也惯用英文，而对应的中文翻译则未形成统一的规范。例如“true negative”一词，“真负”和“真阴”的译法都有。原书作者英文表达行云流水，措辞变换也颇为丰富，本人英文才疏，翻译过程中未必能尽达作者之意。凡此种种都增添了翻译的难度。此书又是本人第一本译作，虽经反复校对，但也不免有疏漏、错误之处。在此，热切欢迎广大读者不吝指正。

唐金川

于 2018 年清明前夜

序

过去 5 年来，Hadoop 和数据科学分别受到追捧。然而，很少有出版物试图将两者结合在一起，即在 Hadoop 环境下讲授数据科学。对于既想入门数据科学又想用 Hadoop 及相关工具解决大规模数据问题的从业者来说，本书将是一个很好的资源。

数据科学涉及的主题包括数据摄取、数据再加工（data munging，通常包含数据清洗和整合）、特征提取、机器学习、预测建模、异常检测和自然语言处理。Hadoop、Spark 以及 Hadoop 生态系统的其他模块为前面这些主题提供了良好的实现用例。它们都是值得选择的平台。数据科学覆盖范围广泛，为此，本书提供具体示例，以帮助工程师解决实际工作中的问题。对于已经熟悉数据科学的读者而言，如果希望掌握超大数据集和 Hadoop 的相关技能，本书也是一块很好的敲门砖。

本书侧重于具体的例子，并通过不同方式来提供对业务价值的洞察。第 5 章提供了特别实用的实例：使用 Hadoop 准备大型数据集，用于常见机器学习和数据科学任务。第 10 章是关于异常检测的，对于重要的大型数据集的监控和报警特别有用。第 11 章是关于自然语言处理的，想研究聊天机器人的读者会比较感兴趣。

Ofer Mendelevitch 是 Lendup 公司的数据科学副总裁，他之前是 Hortonworks 的数据科学总监。在数据科学和 Hadoop 结合的本书中，还有其他几位重要作者。与 Ofer 一起参与本书写作的还有其前同事、Hortonworks 的首席数据科学家 Casey Stella。在这些数据科学和 Hadoop 专家中还有 Douglas Eadline，他也是 Addison-Wesley 的数据和分析系列图书《Hadoop Fundamentals Live Lessons》《Apache Hadoop 2 Quick-Start Guide》和《Apache Hadoop YARN》的贡献者。总的来说，这个作者团队有超过十年的 Hadoop 经验。能有如此丰富的数据科学和 Hadoop 经验的人屈指可数。

本书能加入数据和分析系列图书中令人欣喜。在产品系统中针对大规模数据集创建数据科学解决方案是一种必备技能。本书将助你在部署和执行大规模数据科学解决方案时游刃有余。

Paul Dix
图书系列编辑

前　言

数据科学和机器学习作为许多创新技术和产品的核心，预计在可预见的未来将继续颠覆全球许多行业和商业模式。早几年，这些创新大多受限于数据的可用性。

随着 Apache Hadoop 的引入，所有这一切都发生了变化。Hadoop 提供了一个平台，可以廉价且大规模地存储、管理和处理大型数据集，从而使大数据集的数据科学分析变得实际可行。在这个大规模数据深层分析的新世界，数据科学是核心竞争力，它使公司或组织得以超越传统的商业模式，并在竞争和创新方面保持优势。在 Hortonworks 工作期间，我们有机会看到各种公司和组织如何利用这些新的机会，帮助它们使用 Hadoop 和 Spark 进行规模化数据科学实现。在本书中，我们想分享一些这样的经验。

另外值得强调的是，Apache Hadoop 已经从早期的初始形态演变成整体强大的 MapReduce 引擎（Hadoop 版本 1），再到目前可运行在 YARN 上的多功能数据分析平台（Hadoop 版本 2）。目前 Hadoop 不仅支持 MapReduce，还支持 Tez 和 Spark 作为处理引擎。当前版本的 Hadoop 为许多数据科学应用程序提供了一个强大而高效的平台，并为以前不可想象的新业务开辟了大有可为的新天地。

本书重点

本书着重于在 Hadoop 和 Spark 环境中数据科学的实际应用。由于数据科学的范围非常广泛，而且其中的每一个主题都是深入且复杂的，所以全面阐述数据科学极其困难。为此，我们尝试在每个用例中覆盖理论并在实际实现时辅以样例，以期在理论和实践之间达到平衡。

本书的目的不是深入了解每个机器学习或统计学方法的诸多数学细节，而是提供重要概念的高级描述以及在业务问题背景下践行的指导原则。我们提供了一些参考文献，这些参考文献对书中技术的数学细节进行了更深入的介绍，附录 C 中还提供了相关资源列表。

在学习 Hadoop 时，访问 Hadoop 集群环境可能会成为一个问题。找到一种有效的方式来“把玩”Hadoop 和 Spark 对有些人来说可能是一个挑战。如果要搭建最基础的环境，建议使用 Hortonworks 虚拟机上的沙箱（sandbox），以便轻松开始使用 Hadoop。沙箱是在虚拟机内部可运行的完整的单节点 Hadoop。虚拟机可以在 Windows、Mac OS 和 Linux 下运行。有关如何下载和安装沙箱的更多信息，请参阅 http://hortonworks.com/products/sandbox。有

关 Hadoop 的进一步帮助信息，建议阅读《 Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computation in the Apache Hadoop 2 Ecosystem》一书并查看相关视频，在附录 C 中也可以找到这些信息。

谁应该读这本书

本书面向那些有兴趣了解数据科学且有意涉猎大规模数据集下的应用的读者。如果读者想要更多地了解如何实现各种用例，找到最适合的工具和常见架构，本书也提供了强大的技术基础。本书还提供了一个业务驱动的观点，即何时何地在大型数据集上应用数据科学更有利，这可以帮助利益相关者了解自己的公司能产生什么样的价值，以及在何处投资资源来进行大规模机器学习。

本书需要读者有一定的经验。对于不熟悉数据科学的人来说，需要一些基本知识以了解不同的方法，包括统计概念（如均值和标准差），也需要一些编程背景（主要是 Python，一点点 Java 或 Scala）以理解书中的例子。

对于有数据科学背景的人员，可能会碰到一些如熟悉众多 Apache 项目的实际问题，但是大体上应该对书中的内容游刃有余。此外，所有示例都是基于文本的，并且需要熟悉 Linux 命令行。需要特别注意的是，我们没有使用（或测试）Windows 环境的示例。但是，没有理由假定它们不会在其他环境中正常运行（Hortonworks 支持 Windows）。

在具体的 Hadoop 环境方面，所有示例和代码都是在 Hortonworks HDP Linux Hadoop 版本（笔记本电脑或集群都适用）下运行的。开发环境在发布版本（Cloudera、MapR、Apache Source）或操作系统（Windows）上可能有所不同。但是，所有这些工具在两种环境中都可使用。

如何使用本书

本书有几种不同类型的读者：

- 数据科学家
- 开发人员 / 数据工程师
- 商业利益相关者

虽然这些想参与 Hadoop 分析的读者具有不同背景，但他们的目标肯定是相同的：使用 Hadoop 和 Spark 处理大规模的数据分析。为此，我们设计了后续章节，以满足所有读者的需求。因此，对于在某领域具有良好实践经验的读者，可以选择跳过相应的章节。最后，我们也希望新手读者将本书作为理解规模化的数据科学的第一步。我们相信，即使你看得一头雾水，书中的例子也是有价值的。可以参考后面的背景材料来加深理解。

第一部分包括前 3 章。

第 1 章概述了数据科学及其历史演变，阐述了常见的数据科学家成长之路。对于那些不

熟悉数据科学的人，该章将帮助你了解为什么数据科学会发展成为一个强大的学科，并深入探讨数据科学家是如何设计和优化项目的。该章还会讨论是什么造就了数据科学家，以及如何规划这个方向的职业发展。

第 2 章概述了业务用例如何受现代数据流量、多样性和速度的影响，并涵盖了一些现实的数据科学用例，以帮助读者了解其在各个行业和各种应用中的优势。

第 3 章快速概述了 Hadoop 及其演变历史，以及 Hadoop 生态系统中的各种工具等。对于第一次使用 Hadoop 的用户，该章可能有点难以理解。该章引入了许多新概念，包括 Hadoop 文件系统（HDFS）、MapReduce、Hadoop 资源管理器（YARN）和 Spark。虽然 Hadoop 生态系统的子项目（有些子项目名称比较奇怪）的数量看起来令人生畏，但并不是每个项目都在同一时间使用，而后续章节中的应用通常仅仅集中在其中一小部分工具上。

第二部分包括接下来的 3 章。

第 4 章重点介绍数据摄取，讨论将数据集从外部源导入 Hadoop 的各种工具和技术，这对后续章节很有用。我们从描述 Hadoop 数据湖（data lake）概念开始，介绍了 Hadoop 平台可以使用的各种数据。数据摄取主要使用两个更受欢迎的 Hadoop 工具：Hive 和 Spark。该章重点介绍代码和实操解决方案，如果你是 Hadoop 的新手，可以参考附录 B，以便快速了解 HDFS 文件系统。

第 5 章重点介绍如何使用 Hadoop 进行数据再加工：如何识别和处理数据质量问题，如何预处理数据并进行建模准备。该章将介绍数据的完整性、有效性、一致性、及时性和准确性的概念，接着提供实际数据集的特征生成示例。该章对所有类型的后续分析都是有用的，与第 4 章一样，该章是后续章节中提到的许多技术的铺垫。

数据再加工过程中的一个重要工具就是可视化。第 6 章讨论了使用大数据进行可视化的意义。该章作为背景有助于加强对数据可视化背后一些基本概念的理解。该章中提供的图表是使用 R 生成的。所有图表的源代码都可用，因此读者可以使用自己的数据来尝试生成这些图表。

第三部分包括后 6 章。

第 7 章概述了机器学习，涵盖了机器学习的主要任务（如分类和回归、聚类和异常检测）。对于每个任务类型，我们会探究问题实质并找出解决问题的主要方法。

第 8 章讨论了预测建模的基本算法和各种 Hadoop 工具。该章包括使用 Hive 和 Spark 构建 Twitter 文本情感分析预测模型的端到端示例。

第 9 章深入讲解聚类分析，这也是数据科学中非常普遍的技术。该章介绍了各种聚类技术和相似度计算技术，这些功能都是聚类的核心。随后，该章展示了使用 Hadoop 和 Spark 在大型文档语料库上使用主题模型建模的实例。

第 10 章讨论异常检测，描述了各种方法和算法，以及如何对各种数据集执行大规模异常检测。然后展示了如何使用 Spark 为 KDD99 数据集构建异常检测系统。

第 11 章介绍了使用一套通常称为自然语言处理（NLP）的技术将数据科学应用于人类语言的特定领域。该章谈及 NLP 的各种方法、在各种 NLP 任务中有效的开源工具，以及如

何使用 Hadoop、Pig 和 Spark 将 NLP 应用于大规模语料库。该章用一个端到端的例子展示了在 Spark 中使用 NLP 进行情感分析的高级方法。

第 12 章讨论了 Hadoop 环境下数据科学的未来，涵盖了高级数据发现技术和深入学习。

可参阅附录 A，以查看本书相关网页（网页提供了问题和答案论坛）和代码库。如前所述，附录 B 提供了新用户快速入门 HDFS 的基本方法，附录 C 提供了深入学习 Hadoop、Spark、HDFS、机器学习等许多主题的参考文献。

致　谢

本书中的一些图表和例子来源于以下网站：雅虎（yahoo.com）、Apache 软件基金会（http://www.apache.org）和 Hortonworks（http://hortonworks.com）。任何复制内容都已经过作者的许可或者根据公开分享许可协议可用。

许多人在幕后工作才使这本书得以出版。感谢花时间仔细阅读初稿的审稿人：Fabricio Cannini、Brian D. Davison、Mark Fenner、Sylvain Jaume、Joshua Mora、Wendell Smith 和 John Wilson。

Ofer Mendelevitch

我想感谢 Jeff Needham 和 Ron Lee，是他们鼓励我开始写这本书的。Hortonworks 公司的许多人给了很多建设性的反馈和建议：John Wilson 提供了很有建设性的反馈和行业视角，Debra Williams Cauley 提供了愿景和支持。最后必须要说的是，我美丽的妻子 Noa 一路鼓舞和支持我，我的儿子 Daniel 和 Jordan 也让我觉得辛苦是值得的。如果没有他们给予的爱和鼓舞，此书将难以付梓。

Casey Stella

我要感谢我颇具耐心的爱妻 Leah，以及孩子 William 和 Sylvia，没有他们我不会有时间投入到这样一个耗时也如此有益的事情上来。我要感谢我的母亲和祖母，是她们的谆谆教诲使我至今拥有钟爱学习的良好品质。我还要感谢路易斯安那州的纳税人给我提供机会以接受大学教育，并能接触到图书馆、公共广播和电视资源。没有这些，我就没有如今的能力、学识和勇气。最后，我还要感谢 Addison-Wesley 的 Debra Williams Cauley，他在整个过程中偏好使用胡萝卜而不是大棒。

Douglas Eadline

感谢 Addison-Wesley 的 Debra Williams Cauley，感谢他的辛勤努力，他在 GCT 牡蛎酒吧的办公室使本书写作过程特别放松。感谢我的后勤团 Emily、Carla 和 Taylor，这是另外一本你们一无所知的书。当然，我不能忘记我的办公室伙伴，Marlee 和另外两个男孩。最后，感谢我的贤妻 Maddy 持续不断的支持。

关于作者

Ofer Mendelevitch 是 Lendup 公司的数据科学副总裁，领导 Lendup 的机器学习和高级分析小组。在加入 Lendup 之前，Ofer 是 Hortonworks 的数据科学总监，负责帮助 Hortonwork 的客户使用 Hadoop 和 Spark 将数据科学应用于医疗保健、金融、零售和其他行业。在 Hortonworks 之前，Ofer 曾先后是 XSeed Capital 的驻场企业家、Nor1 的工程副总裁、雅虎的工程总监。

Casey Stella 是 Hortonworks 的首席数据科学家。Hortonworks 提供了一个开源的 Hadoop 版本。Casey 的主要职责是领导开源的 Apache Metron 网络安全项目的分析和数据科学团队。在 Hortonworks 之前，Casey 是 Explorys 公司的架构师，该公司是克利夫兰诊所的一家医疗信息创业公司。更早时，Casey 曾是 Oracle 的开发人员、ION 地球物理研究所的地球物理学专家，并在德州农工大学获得数学学士学位。

Douglas Eadline 博士最初是一名分析化学家，并对计算机方法感兴趣。Douglas 从第一个 Beowulf 的入门文档开始，撰写了数百篇文章、白皮书和教学文件，涵盖了高性能计算（HPC）和 Hadoop 计算的各个方面。在 2005 年创立并编辑流行的 ClusterMonkey.net 网站之前，他曾担任《 ClusterWorld Magazine 》的主编，并且是《 Linux Magazine 》高性能计算的资深编辑。他在高性能计算和 Apache Hadoop 的许多方面具有实践经验，包括硬件和软件设计、基准测试、存储、GPU、云计算和并行计算。目前，他是高性能计算和分析行业的作家兼顾问，也是 Limulus Personal Cluster 项目的负责人（http://limulus.basement-supercomputing.com）。他是 Pearson 出版的《 Hadoop Fundamentals LiveLessons 》和《 Apache Hadoop YARN Fundamentals LiveLessons 》视频的作者，Addison-Wesley 出版的《 Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2》的联合作者，Addison-Wesley 出版的《 Hadoop 2 Quick Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem 》和《 High Performance Computing for Dummies 》的作者。

目　录

第一部分

Hadoop 中的数据科学概览

第 1 章

数据科学概述

我一直认为，未来十年最性感的职业将是统计学家；此非戏言。

——Hal Varian，谷歌首席经济学家

本章将介绍：

- 数据科学定义及其演进历史
- 数据科学家的成长之路
- 数据科学团队的组建
- 数据科学项目的生命周期
- 数据科学项目的管理

近来，数据科学几乎已成为所有数据驱动公司的常见话题。乘着“大数据”的东风，“数据科学”的热度也以令人难以置信的速度飙升。

那么数据科学究竟是什么？为什么它突然变得如此重要呢？

在本章中，我们会从从业者的角度来介绍数据科学，解释相关术语，并阐述数据科学家在大数据时代所扮演的角色。

1.1 数据科学究竟是什么

如果在谷歌或微软必应上搜索“数据科学”一词，那么会看到众说纷纭的定义或解释。大家对这个词的定义似乎并未达成明确的共识，而对于该词何时诞生就更难达成一致了。

本节我们不会重述这些定义，也不会尝试选择我们认为最正确或最准确的定义。相反，我们会从从业者的角度，给出我们自己的定义：

数据科学是通过科学的方法探索数据，以发现有价值的洞察，并在业务环境中运用这些有价值的洞察来构建软件系统。

这个定义强调了两个关键方面。