CMP BOOKS
机工IT

《计算机科学先进技术译丛》

MANNING

Spark 实战

Spark IN ACTION

[美] 彼得·泽斯维奇（Petar Zečević）
[美] 马可·波纳奇（Marko Bonaći）| 著
郑美珠 田华 王佐兵 | 译

在线提供源代码

机械工业出版社
CHINA MACHINE PRESS

计算机科学先进技术译丛

Spark 实战

[美] 彼得·泽斯维奇（Petar Zečević）
[美] 马可·波纳奇（Marko Bonaći）
著

郑美珠　田　华　王佐兵　译

机械工业出版社

本书介绍了 Spark 应用程序及更高级应用的工作流程，主要从使用角度进行了描述，每个具体内容都有对应的代码。本书涵盖了 Apache Spark 和它丰富的 API，构成 Spark 的组件（包括 Spark SQL、Spark Streaming、Spark MLlib 和 Spark GraphX），在 Spark standalone、Hadoop YARN 以及 Mesos clusters 上运行 Spark 应用程序的部署和安装。通过对应的实例全面、详细地介绍了整个 Spark 实战开发的流程。最后，还介绍了 Spark 的高级应用，包括 Spark 流应用程序及可扩展和快速的机器学习框架 H2O。

本书可以作为高等院校计算机、软件工程、数据科学与大数据技术等专业的大数据课程材料，可用于指导 Spark 编程实践，也可供相关技术人员参考使用。

北京市版权局著作权合同登记 图字：01-2017-0154 号

图书在版编目(CIP)数据

Spark 实战/(美)彼得·泽斯维奇,(美)马可·波纳奇著;郑美珠,田华,王佐兵译. —北京:机械工业出版社,2019.2
(计算机科学先进技术译丛)
书名原文：Spark in Action
ISBN 978-7-111-61748-8

Ⅰ. ①S… Ⅱ. ①彼… ②马… ③郑… ④田… ⑤王… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 144373 号

机械工业出版社(北京市百万庄大街 22 号 邮政编码 100037)
策划编辑：李培培 责任编辑：李培培
责任校对：张艳霞 责任印制：郜 敏

北京圣夫亚美印刷有限公司印刷

2019 年 8 月第 1 版·第 1 次印刷
184 mm×260 mm·24.5 印张·605 千字
0001-3000 册
标准书号：ISBN 978-7-111-61748-8
定价：99.00 元

电话服务
客服电话：010-88361066
010-88379833
010-68326294
封底无防伪标均为盗版

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
金 书 网：www.golden-book.com
机工教育服务网：www.cmpedu.com

译者序

生活离不开水，同样离不开数据，我们被数据包围，在数据中生活。当数据越来越多时，就成了大数据。

想要理解大数据，就需要理解大数据相关的查询、处理、机器学习、图计算和统计分析等，Spark 作为新一代轻量级大数据快速处理平台，集成了大数据相关的各种能力，是理解大数据的首选。

Spark 作为 Apache 顶级的开源项目，是一个快速、通用的大规模数据处理引擎，和 Hadoop 的 MapReduce 计算框架类似。但是相对于 MapReduce，Spark 凭借其可伸缩、基于内存计算等特点，以及可以直接读/写 Hadoop 上任何格式数据的优势，在进行批处理时更加高效，并有更低的延迟。相对于“one stack to rule them all”的目标，Spark 实际上已经成为轻量级大数据快速处理的统一平台，各种不同的应用，如实时流处理、机器学习、交互式查询等，都可以通过 Spark 建立在不同的存储和运行系统上。

本书全面、详细地介绍了 Spark 的相关知识及相关应用。它将 Spark 的基础知识、基础应用及更高级应用娓娓道来，给学习者指明了道路。

本书由郑美珠、田华、王佐兵共同翻译。由于译者水平有限，翻译中不当之处在所难免，请读者和同行不吝指正。

感谢丈夫张蕾和儿子祎祎，虽然在翻译本书的过程中减少了陪伴你们的时间，但你们依然理解和支持我的工作。感谢父母在生活上的帮助，使我可以全身心地投入到本书的翻译工作中。

郑美珠
烟台南山学院

致谢

我们的技术校对 Michiel Trimpe 提出了无数宝贵的建议。同时感谢 Robert Ormandi 审查第 7 章，还要感谢 Spark in Action 的审稿人，包括 Andy Kirsch、Davide Fiorentino、lo Regio、Dimitris Kouzis-Loukas、Gaurav Bhardwaj、Ian Stirk、Jason Kolter、Jeremy Gailor、John Guthrie、Jonathan Miller、Jonathan Sharley、Junilu Lacar、Mukesh Kumar、Peter J. KreyJr. 、Pranay Srivastava、Robert Ormandi、Rodrigo Abreu、Shobha Iyer、Sumit Pal。

我们要感谢 Manning 出版社的工作人员使本书能够出版：他们是出版商 Marjan Bace、Manning 审稿人和编辑团队，特别是 Marina Michaels 指导我们如何编写高质量的书。我们还要感谢生产团队在完成项目的过程中所做的工作。

Petar Zečević

我要感谢我的妻子在我工作时给予的支持和耐心。我要感谢我的父母用爱抚养我，尽可能给我最好的学习环境。最后，我要感谢我的公司 SV 集团为我写本书提供所需的资源和时间。

Marko Bonači

我要感谢我的合作者 Petar。没有他的坚持，这本书不会写出来。

前言

回顾过去一年半，笔者不禁想到：笔者在这个地球上是如何生存的，这是笔者生命中最繁忙的18个月！自从Manning出版社让笔者和Marko写一本关于Spark的书，笔者花了大部分空闲时间在Apache Spark上。笔者在这段时间过得很充实，学到了很多，并且觉得这是值得的。

如今，Spark是一个非常热门的话题，它于2009年由Matei Zaharia在加利福尼亚州的伯克利提出（最初是试图证明Mesos执行平台的可行性），在2010年开源。2013年Spark被捐赠给了Apache软件基金会，从那以后它以闪电般的速度发展。2015年，Spark是最活跃的Apache项目之一，有超过1000个贡献者（投稿人、捐助人）。今天，Spark是所有主要Hadoop发行版的一部分，并被许多组织使用，广泛应用于各种或大或小的程序中。

写一本关于Spark的书的挑战在于它发展很快。自从笔者开始写Spark in Action，笔者看到了6个版本的Spark，有许多新的、重要的功能需要覆盖。第一个主要版本（2.0版本）是在笔者完成了大部分书的写作后推出的，笔者不得不延迟出版计划以涵盖它附带的新功能。

写Spark的另一个挑战是主题的广度：Spark更多的是一个平台，而不是一个框架。用户可以使用它来编写各种应用程序（用4种语言），包括批处理作业、实时处理系统和Web应用程序执行Spark作业、用SQL处理结构化数据和使用传统编程技术处理非结构化数据、各种机器学习和数据修改任务、与分布式文件系统交互、各种关系和无SQL数据库、实时系统等。安装、配置和运行Spark，这些运行时的工作也同样重要。

笔者详细地介绍了Spark中的重要内容并且使本书成为使用Spark的指南，希望用户能够喜欢本书。

关于本书

Apache Spark 是一种通用的数据处理框架，这意味着用户可以在各种计算任务中使用它，任何关于 Apache Spark 的书都需要涵盖很多不同的主题。笔者试图描述使用 Spark 的各个方面：从配置运行时选项、运行独立和交互作业，到编写批处理、流式处理或机器学习应用程序。本书中的示例和示例数据集可以在个人计算机上运行，它们很容易理解，并且很好地说明了 Spark 的相关概念。

笔者希望用户能够找到本书和示例，以便了解如何使用和运行 Spark，并且它将帮助用户编写有应用前景的、可付诸生产的 Spark 应用程序。

谁应该读这本书

虽然本书包含了许多适合商业用户和管理者的资料，但它主要面向开发人员，或者更确切地说，面向的是能够理解和执行代码的人。Spark API 可以用 4 种语言：Scala、Java、Python、R。本书中主要的例子是用 Scala 编写的（Java 和 Python 的版本可以在本书的网站 www. manning. com/books/spark-in-action，以及在线 GitHub 存储库 https://github. com/spark-in-action/first-edition 获得）。本书对 Scala 的具体细节进行了解释，所以用户在读本书之前可以没有任何 Scala 的知识。但是如果掌握 Java 或 Scala 的技术，那么用户会更容易理解本书。第 2 章会详细介绍 Spark 的基础知识。

Spark 可以与许多系统交互，其中有一些会在本书中介绍。为了充分理解内容，以下主题的知识是首选的（但不是必需的）：

- SQL 和 JDBC（第 5 章）
- Amazon EC2（第 11 章）
- Hadoop（HDFS 和 YARN 第 5 章和第 12 章）
- 线性代数的基础知识和理解数学公式的能力（第 7 章和第 8 章）
- Kafka（第 6 章）
- Mesos（第 12 章）

本书准备了一个虚拟机，可以让用户轻松运行本书中的示例。要使用该虚拟机，计算机应满足第 1 章中列出的软件和硬件要求。

本书内容安排

本书共有 14 章，分为 4 个部分。

第 1 部分介绍了 Apache Spark 和它丰富的 API。理解这些信息对于编写高质量的 Spark 程序非常重要，也是本书其余部分的基础。

第 1 章大致描述了 Spark 的主要特点，并与 Hadoop's MapReduce 和 Hadoop 的生态系统的其他工具进行对比，还包括对 spark-in-action 虚拟机的描述，用户可以使用它来运行书中的示例。

第 2 章进一步探讨虚拟机，介绍如何使用 Spark 的命令行界面（spark-shell），并用几个例子来解释弹性分布式数据集（RDD），即 Spark 中的中心抽象。

第 3 章介绍了如何将 Eclipse 设置为编写独立的 Spark 应用程序。用户将按照书中内容编写一个用于分析 GitHub 日志的应用程序，并通过将其提交到 Spark 集群来执行该应用程序。

第 4 章更详细地探讨了 Spark 核心 API，展示了如何使用键值对，并解释了 Spark 中数据分区和混排的工作原理，介绍了如何分组、排序和连接数据，以及如何使用累加器和广播变量。

第 2 部分介绍了构成 Spark 的其他组件，包括 Spark SQL、Spark Streaming、Spark MLlib 和 Spark GraphX。

第 5 章介绍了 Spark SQL，详细介绍了如何创建和使用 DataFrame、如何使用 SQL 查询 DataFrame 数据，以及如何将数据加载到外部数据源并从中保存。还介绍了优化 Spark 的 SQL Catalyst 优化引擎和 Tungsten 项目引入的性能改进。

第 6 章介绍了 Spark Streaming，它是 Spark 家族中最受欢迎的成员之一。本章介绍了会在流应用程序运行时定期生成 RDD 的离散流、如何随时间保存计算状态和如何使用窗口操作，还介绍了连接 Kafka 的方法以及如何从流媒体作业中获得良好的性能，并且还介绍了结构化流媒体，它是 Spark 2.0 中的一个新概念。

第 7 章和第 8 章是关于机器学习的介绍，特别是关于 Spark MLlib 和 Spark ML Spark API 部分的内容。机器学习包括线性回归、逻辑回归、决策树、随机森林和 k 均值聚类。在此过程中，用户会使用正则化以及训练和评估机器学习模型来扩展和规范化功能。这两章还将解释 Spark ML 带来的 API 标准化。

第 9 章探讨了如何使用 Spark 的 GraphX API 构建图形，用户将使用图形算法转换和连接图形，并使用 GraphX API 实现 A* 搜索算法。

使用 Spark 不仅仅是编写和运行 Spark 应用程序，也是配置 Spark 集群和系统资源以供应用程序高效使用。第 3 部分解释了在 Spark standalone、Hadoop YARN 和 Mesos clusters 上运行 Spark 应用程序的必要概念和配置选项。

第 10 章探讨了 Spark 运行时组件、Spark 集群类型、作业和资源调度、配置 Spark 和 Spark Web UI。这些是 Spark 可以运行的所有集群管理器共同的概念：Spark standalone 集群、YARN 和 Mesos。

第 11 章介绍了 Spark standalone 集群，包括如何启动它并在其上运行应用程序，以及如何使用其 Web UI。还讨论了 Spark History 服务器，它保存有关以前运行的作业的详细信息。

最后，介绍了如何在 Amazon EC2 上使用 Spark 的脚本启动 Spark standalone 集群。

第 12 章详细介绍了如何设置、配置和使用 YARN 和 Mesos 集群来运行 Spark 应用程序。

第 4 部分介绍了使用 Spark 的高级应用。

第 13 章将所有内容汇总在一起，并探讨了一个 Spark 流应用程序，用于分析日志文件，并在实时仪表板上显示结果。本章中实现的应用程序可作为用户在未来编写应用程序的基础。

第 14 章介绍了 H2O，这是一个可扩展的快速机器学习框架，它实现了许多机器学习算法，最著名的是深度学习，而这正是 Spark 缺乏的。Sparkling Water 将 H2O 和 Spark 相结合，用户可以启动和使用 Spark 的 H2O 集群。通过 Sparkling Water，用户可以使用 Spark 的 Core、SQL、Streaming 和 GraphX 组件来获取、准备和分析数据，并将其传输到 H2O，以用于 H2O 的深度学习算法。然后用户可以将结果传回 Spark，并在后续计算中使用它们。

附录 A 给出了安装 Spark 的说明。附录 B 提供了一个简短的 MapReduce 视图。附录 C 是关于线性代数的简短引用。

关于代码

书中的所有源代码以单间隔字体呈现，这样可以将其与其他内容区别开来。在许多列表中，代码被注释以指出关键概念，并且有时在文本中编号项目符号以提供有关代码的其他信息。

Scala、Java 和 Python 语言的源代码以及示例中使用的数据文件可以从发布商的网站 www. manning. com/books/spark-in-action 下载，也可以从在线存储库 https://github. com/spark-in-action/first-edition 获取。这些示例是为 Spark 2. 0 编写和测试的。

作者在线

购买 Spark in Action 可以免费访问 Manning 出版社运行的私人网络论坛，在那里读者可以对这本书发表评论、提出技术问题，并接受主要作者和其他用户的帮助。

如果要访问论坛并订阅该论坛，则可以登录 www. manning. com/books/spark-in-action。此页面提供了如何在注册后访问论坛、提供什么样的帮助以及论坛上的行为规则等信息。

Manning 对读者的承诺是提供一个场所，在这里个体读者之间以及读者和作者之间可以进行有意义的对话。这不是对作者的任何具体参与的承诺，作者对在线论坛的贡献仍然是自愿的（并且是无偿的）。我们建议读者尝试向作者提出一些感兴趣的、具有挑战性的问题。只要本书出版，论坛和之前讨论的资料就可以从出版商的网站上获取。

关于作者

Petar Zečević 在软件行业工作超过了 15 年。他一开始是 Java 开发人员，后来作为全职开发人员、顾问、分析师和团队领导参加过很多项目。他目前担任 SV 集团的 CTO 角色。SV 集团是一家克罗地亚软件公司，为克罗地亚大型银行、政府机构和私人公司工作。Petar 每月都会组织 Apache Spark Zagreb 聚会，定期在会议上发言，他身后有几个 Apache Spark 项目。

Marko Bonači 已经从事 Java 工作 13 年。他作为 Spark 开发人员和顾问为 Sematext 工作。在此之前，他是 SV 集团 IBM 企业内容管理团队的团队领导。

关于封面

《Spark 实战》封面插图标题为“Hollandais”（荷兰人）。插图取自各种各样国家服装服饰的收藏者 Jacques Grasset de Saint-Sauveur（1757-1810）于 1797 年在法国出版的 Costumes de Différents Pays。其中的每幅图都经过精心绘制并手工着色。

Grasset de Saint-Sauveur 丰富的收藏生动地展示了 200 年前不同城市和地区的文化差异。由于相互隔离，人们说着不同的方言和语言。无论是在城市的街道、小城镇或乡村，都可以很容易地通过他们的穿着分辨出他们在哪里生活以及他们的生活习惯。

服饰密码从那时起已经改变，那个时候的人们根据区域和阶级的不同拥有的服饰特色现在已经逐渐消失。现在人们已经很难通过服饰区分不同大洲的居民，更不用说不同的城镇或地区了。也许人们已经将文化多样性换成了一种更加多样化的个人生活——当然是为了更加多样化和快节奏的科技生活。

当计算机图书多到无法区别时，本书采用 Grasset de Saint-Sauveur 两世纪前区域生活的多样性图片作为图书封面的方式，庆祝 Manning 计算机图书的创造性和主动性。

目录

第 2 部分 认识 Spark 家族

第3部分 Spark ops