HZ BOOKS 华章IT

[PACKT] PUBLISHING

PySpark的实用参考指南，深度挖掘Python+Spark的强大功能

系统讲解如何在Spark 2.0上高效运用Python来处理数据并建立机器学习模型，帮助读者开发并部署高效可拓展的实时Spark解决方案

大数据 技术丛书

Learning PySpark

PySpark实战指南

利用Python和Spark构建数据密集型应用并规模化部署

[美] 托马兹·卓巴斯 (Tomasz Drabas)
丹尼·李 (Denny Lee) 著

栾云杰 陈瑶 刘旭斌 译

机械工业出版社
China Machine Press

大数据技术丛书

Learning PySpark

PySpark实战指南

利用Python和Spark构建数据密集型应用并规模化部署

[美] 托马兹·卓巴斯（Tomasz Drabas）
丹尼·李（Denny Lee） 著

栾云杰 陈瑶 刘旭斌 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

PySpark 实战指南：利用 Python 和 Spark 构建数据密集型应用并规模化部署 /（美）托马兹·卓巴斯（Tomasz Drabas），（美）丹尼·李（Denny Lee）著；栾云杰，陈瑶，刘旭斌译 . —北京：机械工业出版社，2017.9
（大数据技术丛书）
书名原文：Learning PySpark

ISBN 978-7-111-58238-0

I.P… II.①托… ②丹… ③栾… ④陈… ⑤刘… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 251424 号

本书版权登记号：图字 01-2017-3409

PySpark 实战指南

利用 Python 和 Spark 构建数据密集型应用并规模化部署

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：张志铭　　责任校对：李秋荣
印　刷：北京文昌阁彩色印刷有限责任公司　　版　次：2017 年 11 月第 1 版第 1 次印刷
开　本：186mm×240mm 1/16　　印　张：12.5
书　号：ISBN 978-7-111-58238-0　　定　价：49.00 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）88379426 88361066　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

The Translator's Words 译 者 序

为什么要翻译这本书

年初的时候我们从机械工业出版社华章公司那里知道有一本关于大数据的书正在征集翻译，在看过英文版并翻译了样章后，我们几位志同道合的软件工程师一块儿接受了《Learning PySpark》的翻译工作。我们都非常兴奋，因为作为软件工程师，能有机会把当前最热最新的技术介绍给大家是何其荣幸。

Python 是数据分析最常用的语言之一，而 Apache Spark 是一个开源的强大的分布式查询和处理引擎。本书用详尽的例子介绍了如何使用 Python 来调用 Spark 的新特性，如何处理结构化和非结构化的数据，如何使用 PySpark 中一些基本的可用数据类型，如何生成机器学习模型、操作图像、阅读串流数据以及在云上部署你的模型。

数据是每个人身边都存在的，理解学习比较容易，但是数据量足够大才是一个相对准确的学习平台。在实践中，如何确定训练集合、如何将脏数据处理为清洁数据、如何填充数据等等，需要结合本书的知识理论，清楚了解待处理的大数据特性。每一种数据的特征或特性都不一致，所以前期的准备和调研必不可少。本书不仅仅是一本工具书，也是一本能深入浅出、结合简单实例来介绍 PySpark 语言的书。不管使用什么语言和工具，万变不离其宗。希望阅读此书的人，除了看懂示例，还能够结合实际经验来推敲，这样就能明白作者举这些例子的良苦用心。

希望大家喜欢这本书，因为译者的水平有限，翻译中的错漏缺点在所难免，希望读者批评指正。

读者对象

本书适合以下几类读者阅读：

- 对大数据的前沿技术非常感兴趣的人。
- 有志于成为一名数据科学家的从业人员。
- 有一定算法和编程基础的技术爱好者。

译者分工

本书由来自 IBM 中国开发中心的软件工程师联合翻译完成。其中：

- 栾云杰（目前就职于 IBM 中国开发中心）翻译了第 5 章、第 6 章。
- 陈瑶（原 IBM 工程师，现就职于某大数据公司）翻译了第 2 章、第 3 章、第 4 章、第 11 章。
- 刘旭斌（目前就职于 IBM 中国开发中心）翻译了第 7 章、第 8 章、第 9 章。

另外，第 1 章由栾云杰、陈瑶、刘旭斌三人共同翻译，第 10 章由栾云杰、陈瑶两人共同翻译。

致谢

感谢华章公司引进了该书的中译本版权，这是该中译本得以面市的最核心要素。

感谢华章公司的和静老师给予我们的支持和信任。因为这份信任，我们才有机会来翻译这本关于大数据和 Apache Spark 的书籍。

感谢本次翻译组的小伙伴们。翻译本书的过程，是一种学习与思考的结合，也是和伙伴合作与交流的经历。非常庆幸遇到了睿智又勤奋的伙伴，即使在繁忙的工作和节奏极快的生活中，也努力完成了翻译和审阅计划。

另外，也要感谢我们的家人对我们的支持，正是有他们的支持和鼓励，我们才能坚持下来。

Foreword 序

感谢你选择这本书开始 PySpark 之旅，希望你像我一样兴奋。当 Denny Lee 第一次告诉我这本新书的时候，我非常高兴。Apache Spark 既支持 Java、Scala、JVM 世界，又支持 Python（以及近来的 R）世界，这是它成为一个如此非凡的平台最为重要的原因。以前很多书籍都集中于核心语言，或者主要关注在 JVM 语言上，所以很高兴看到由如此有经验的 Spark 教育工作者来专门为 PySpark 出书，使 PySpark 有机会绽放光芒。PySpark 通过支持这两个不同的世界，使我们能够成为更高效的数据科学家和数据工程师，同时得以借鉴彼此社区的那些绝佳想法。

很荣幸有机会浏览这本书的早期版本，这使我对该项目的兴趣更为浓厚。我曾有幸参加过一些类似的会议和聚会，看着作者将 Spark 世界的新概念介绍给不同的观众（从新人到经验丰富的老手），并且他们提取自身的经验写出这本书，他们真是太棒了。从阐述知识到各个主题的覆盖，无一不体现了作者们的丰富经验。除了简单介绍 PySpark 之外，他们还花时间从社区中找来了日渐重要的包，如 GraphFrames 和 TensorFrames。

在决定使用哪些工具时，我觉得社区是经常被忽视的一部分，Python 拥有一个很棒的社区，我期待着你加入 Python Spark 社区。所以，来享受你的冒险之旅吧；我知道你会和 Denny Lee 以及 Tomek Drabas 有很好的联系。 我真的相信，通过拥有多样化的 Spark 用户社区，我们将能够创造出对每个人都有用的更好的工具，所以我希望能够在某个会议、聚会或邮件列表中见到你！

Holden Karau

前　言 Preface

据估计，2013 年全世界产生了大约 4.4ZB（词头 Z 代表 10^{21}）信息量的数据！而到 2020 年，预计人类将会产生 10 倍于 2013 年的数据量。随着字面上的数字越来越大，加上人们需求的日益增长，为了使这些数据更有意义，2004 年来自 Google 的 Jeffrey Dean 和 Sanjay Ghemawat 发表了一篇开创性的论文《MapReduce：Simplified Data Processing on Large Clusters》。至此，利用这一概念的技术开始快速增多，Apache Hadoop 也开始迅速变得流行起来，最终创建了一个 Hadoop 的生态系统，包括抽象层的 Pig、Hive 和 Mahout，都是利用了 map 和 reduce 的简单概念。

然而，即使拥有每天都分析过滤海量数据的能力，MapReduce 始终是一个限制相当严格的编程框架。此外，大多数的任务还要求读取、写入磁盘。认识到这些缺点，2009 年 Matei Zaharia 将 Spark 作为他博士课题的一部分开始研究。Spark 在 2012 年首次发布。虽然 Spark 是基于和 MapReduce 相同的概念，但其先进的数据处理方法和任务组织方式使得它比 Hadhoop 要快 100 倍（对于内存计算）。

在这本书中，我们将指导你使用 Python 了解 Apache Spark 的最新性能，包括如何看懂结构化和非结构化的数据，如何使用 PySpark 中一些基本的可用数据类型，生成机器学习模型，图像操作，阅读串流数据，在云上部署模型。每一章力求解决不同的问题，并且我们也希望看完这本书之后，你可以掌握足够的知识来解决其他我们还没来得及在书中讲解的问题。

本书的主要内容

第 1 章通过技术和作业的组织等概念提供了对 Spark 的介绍。

第 2 章介绍了 RDD、基本原理、PySpark 中可用的非模式化数据结构。

第 3 章详细介绍了 DataFrame 数据结构，它可以弥合 Scala 和 Python 之间在效率方面的差距。

第 4 章引导读者了解 Spark 环境中的数据清理和转换的过程。

第 5 章介绍了适用于 RDD 的机器学习库，并回顾了最有用的机器学习模型。

第 6 章涵盖了当前主流的机器学习库，并且提供了目前可用的所有模型的概述。

第 7 章引导你了解能轻松利用图解决问题的新结构。

第 8 章介绍了 Spark 和张量流（TensorFlow）领域中深度学习（Deep Learning）的连接桥梁。

第 9 章描述 Blaze 是如何跟 Spark 搭配使用以更容易地对多源数据进行抽象化的。

第 10 章介绍了 PySpark 中可用的流工具。

第 11 章一步步地指导你运用命令行界面完成代码模块化并提交到 Spark 执行。

其他一些详细信息，我们提供了以下额外的章节：

安装 Spark：https://www.packtpub.com/sites/default/files/downloads/InstallingSpark.pdf。

免费提供 Spark Cloud：https://www.packtpub.com/sites/default/files/downloads/FreeSparkCloudOffering.pdf。

本书需要的软 / 硬件支持

阅读本书，需要准备一台个人电脑（Windows、Mac 或者 Linux 任一系统都行）。运行 Apache Spark，需要 Java 7+ 并且安装配置 Python 2.6+ 版本或者 3.4+ 版本的环境；本书中使用的是 Anaconda Python3.5 版本，可以在 https://www.continuum.io/downloads 下载。

本书中我们随意使用了 Anaconda 的预装版 Python 模块。GraphFrames 和 TensorFrames 也可以在启动 Spark 实例时动态加载：载入时你的电脑需要联网。如果有的模块尚未安装到你的电脑里，也没有关系，我们会指导你完成安装过程。

本书的读者对象

想要学习大数据领域发展最迅速的技术即 Apache Spark 的每一个人，都可以阅读此书。我们甚至希望还有来自于数据科学领域更高级的从业人员，能够找到一些令人耳目一新的例子以及更有趣的主题。

本书约定

警告或重要的笔记

提示和技巧

下载代码示例

你可以从 http://www.packtpub.com 下载代码文件。你也可以访问华章图书官网：http://www.hzbook.com, 通过注册并登录个人账号，下载本书的源代码。

下载本书彩图

我们还提供了一个 PDF 文件，其中包含本书中使用的截图和彩图，可以帮助读者更好地了解输出的变化。您可以从此下载文件 https://www.packtpub.com/sites/default/files/downloads/LearningPySpark_ColorImages.pdf。

About the Authors 关于作者

托马兹·卓巴斯（Tomasz Drabas）工作于微软，是一名数据科学家，现居住在西雅图地区。他拥有超过 13 年的数据分析和数据科学经验：在欧洲、澳大利亚和北美洲三大洲期间，工作领域遍及先进技术、航空、电信、金融和咨询。在澳大利亚期间，托马兹一直致力于运筹学博士学位，重点是航空业中的选择建模和收入管理应用。

在微软，托马兹每天都和大数据打交道，解决机器学习问题，如异常检测、流失预测和使用 Spark 的模式识别。

托马兹还撰写了《 Practical Data Analysis Cookbook 》，该书由 Packt Publishing 于 2016 年出版。

我要感谢我的家人 Rachel、Skye 和 Albert，你们是我生命中的挚爱，我很珍惜与你们度过的每一天！谢谢你们永远站在我身边，鼓励我一步步接近我的职业目标。另外，感谢所有的亲人们。

多年来，还有很多人影响了我，我得再写一本书来感谢他们。你们知道，我从心底谢谢你们！

不过，如果不是 Czesia Wieruszewska，我不会得到博士学位。还有 Krzys Krzysztoszek，你一直相信我！谢谢！

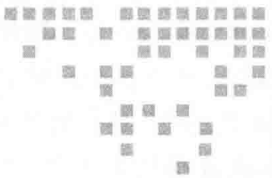
丹尼·李（Denny Lee）是微软 Azure DocumentDB 团队的首席项目经理，该团队致力于为微软发展高效性、重量级的托管文档存储服务。他是一名喜欢实践的分布式系统和数据科学工程师，拥有超过 18 年的互联网级别基础架构、数据平台和预测分析系统的开发经验，这些经验可用于内部部署和云环境。

他在组建新团队以及促进转型、改革方面拥有丰富的经验。在加入 Azure DocumentDB

团队之前，丹尼曾担任 Databricks 的技术传播专员，他从 Apache Spark 0.5 时就一直在使用 Apache Spark。他还是 Concur 数据科学工程的高级总监，曾就职于构建了微软 Windows 和 Azure 服务（目前称为 HDInsight）的 Hadoop 的孵化团队。丹尼还拥有俄勒冈州健康和科学大学的生物医学信息硕士学位，并在过去 15 年中为企业医疗保健客户构建和实施了强大的数据解决方案。

我要感谢我的好妻子 Hua-Ping，还有我出色的女儿 Isabella 和 Samantha。是你们让我保持清醒，帮我实现了梦寐以求的愿望！

Contents 目　录



第 1 章 Chapter 1

了解 Spark

Apache Spark 是一个强大的开源处理引擎，最初由 Matei Zaharia 开发，是他在加州大学伯克利分校的博士论文的一部分。Spark 的第一个版本于 2012 年发布。从那时起，Zaharia 于 2013 年与人合作创立了 Databricks 并担任 CTO，他还在斯坦福大学担任教授职务（之前在麻省理工学院）。同时，Spark 代码库被捐赠给 Apache Software Foundation，并已成为其旗舰项目。

Apache Spark 是快速、易于使用的框架，允许你解决各种复杂的数据问题，无论是半结构化、结构化、流式，或机器学习、数据科学。它也已经成为大数据方面最大的开源社区之一，拥有来自 250 多个组织的超过 1000 个贡献者，以及遍布全球 570 多个地方的超过 30 万个 Spark Meetup 社区成员。

在本章中，我们将提供一个了解 Apache Spark 的切入点。我们将解释 Spark Job 和 API 背后的概念，介绍 Spark 2.0 架构，并探索 Spark 2.0 的特性。

1.1 什么是 Apache Spark

Apache Spark 是一个开源的、强大的分布式查询和处理引擎。它提供 MapReduce 的灵活性和可扩展性，但速度明显更高：当数据存储在内存中时，它比 Apache Hadoop 快 100 倍，访问磁盘时高达 10 倍。

Apache Spark 允许用户读取、转换、聚合数据，还可以轻松地训练和部署复杂的统计模型。Java、Scala、Python、R 和 SQL 都可以访问 Spark API。Apache Spark 可用于构建应用程序，或将其打包成为要部署在集群上的库，或通过笔记本（notebook）（例如 Jupyter、Spark-Notebook、Databricks notebooks 和 Apache Zeppelin）交互式执行快速的分析。