O'REILLY®

Learning Spark

学习Spark（影印版）

Holden Karau, Andy Konwinski,
Patrick Wendell, Matei Zaharia 著

東南大學出版社

学习Spark（影印版）

Learning Spark

Holden Karau, Andy Konwinski,
Patrick Wendell, Matei Zaharia 著

常州大学山书馆
藏 书 章

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

O'Reilly Media, Inc.授权东南大学出版社出版

南京 东南大学出版社

图书在版编目(CIP)数据

学习 Spark:英文/(美)卡劳(Karau, H.)等著. —影印本. —南京:东南大学出版社,2015.8

书名原文:Learning Spark

ISBN 978-7-5641-5921-4

Ⅰ. ①学… Ⅱ. ①卡… Ⅲ. ①数据处理软件-英文 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 165450 号

江苏省版权局著作权合同登记
图字:10-2014-157 号

© 2015 by O'Reilly Media, Inc.

Reprint of the English Edition, jointly published by O'Reilly Media, Inc. and Southeast University Press, 2015. Authorized reprint of the original English edition, 2014 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc.出版 2015。

英文影印版由东南大学出版社出版 2015。此影印版的出版和销售得到出版权和销售权的所有者—— O'Reilly Media, Inc.的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

学习 Spark(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼 2 号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787 毫米×980 毫米　16 开本
印　　张:17.25
字　　数:338 千字
版　　次:2015 年 8 月第 1 版
印　　次:2015 年 8 月第 1 次印刷
书　　号:ISBN 978-7-5641-5921-4
定　　价:54.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Foreword

In a very short time, Apache Spark has emerged as the next generation big data processing engine, and is being applied throughout the industry faster than ever. Spark improves over Hadoop MapReduce, which helped ignite the big data revolution, in several key dimensions: it is much faster, much easier to use due to its rich APIs, and it goes far beyond batch applications to support a variety of workloads, including interactive queries, streaming, machine learning, and graph processing.

I have been privileged to be closely involved with the development of Spark all the way from the drawing board to what has become the most active big data open source project today, and one of the most active Apache projects! As such, I'm particularly delighted to see Matei Zaharia, the creator of Spark, teaming up with other longtime Spark developers Patrick Wendell, Andy Konwinski, and Holden Karau to write this book.

With Spark's rapid rise in popularity, a major concern has been lack of good reference material. This book goes a long way to address this concern, with 11 chapters and dozens of detailed examples designed for data scientists, students, and developers looking to learn Spark. It is written to be approachable by readers with no background in big data, making it a great place to start learning about the field in general. I hope that many years from now, you and other readers will fondly remember this as *the* book that introduced you to this exciting new field.

—Ion Stoica, CEO of Databricks and Co-director, AMPlab, UC Berkeley

Preface

As parallel data analysis has grown common, practitioners in many fields have sought easier tools for this task. Apache Spark has quickly emerged as one of the most popular, extending and generalizing MapReduce. Spark offers three main benefits. First, it is easy to use—you can develop applications on your laptop, using a high-level API that lets you focus on the content of your computation. Second, Spark is fast, enabling interactive use and complex algorithms. And third, Spark is a *general* engine, letting you combine multiple types of computations (e.g., SQL queries, text processing, and machine learning) that might previously have required different engines. These features make Spark an excellent starting point to learn about Big Data in general.

This introductory book is meant to get you up and running with Spark quickly. You'll learn how to download and run Spark on your laptop and use it interactively to learn the API. Once there, we'll cover the details of available operations and distributed execution. Finally, you'll get a tour of the higher-level libraries built into Spark, including libraries for machine learning, stream processing, and SQL. We hope that this book gives you the tools to quickly tackle data analysis problems, whether you do so on one machine or hundreds.

Audience

This book targets data scientists and engineers. We chose these two groups because they have the most to gain from using Spark to expand the scope of problems they can solve. Spark's rich collection of data-focused libraries (like MLlib) makes it easy for data scientists to go beyond problems that fit on a single machine while using their statistical background. Engineers, meanwhile, will learn how to write general-purpose distributed programs in Spark and operate production applications. Engineers and data scientists will both learn different details from this book, but will both be able to apply Spark to solve large distributed problems in their respective fields.

试读结束，需要全本PDF请购买 www.ertongbook.com

Data scientists focus on answering questions or building models from data. They often have a statistical or math background and some familiarity with tools like Python, R, and SQL. We have made sure to include Python and, where relevant, SQL examples for all our material, as well as an overview of the machine learning and library in Spark. If you are a data scientist, we hope that after reading this book you will be able to use the same mathematical approaches to solve problems, except much faster and on a much larger scale.

The second group this book targets is software engineers who have some experience with Java, Python, or another programming language. If you are an engineer, we hope that this book will show you how to set up a Spark cluster, use the Spark shell, and write Spark applications to solve parallel processing problems. If you are familiar with Hadoop, you have a bit of a head start on figuring out how to interact with HDFS and how to manage a cluster, but either way, we will cover basic distributed execution concepts.

Regardless of whether you are a data scientist or engineer, to get the most out of this book you should have some familiarity with one of Python, Java, Scala, or a similar language. We assume that you already have a storage solution for your data and we cover how to load and save data from many common ones, but not how to set them up. If you don't have experience with one of those languages, don't worry: there are excellent resources available to learn these. We call out some of the books available in "Supporting Books" on page xii.

How This Book Is Organized

The chapters of this book are laid out in such a way that you should be able to go through the material front to back. At the start of each chapter, we will mention which sections we think are most relevant to data scientists and which sections we think are most relevant for engineers. That said, we hope that all the material is accessible to readers of either background.

The first two chapters will get you started with getting a basic Spark installation on your laptop and give you an idea of what you can accomplish with Spark. Once we've got the motivation and setup out of the way, we will dive into the Spark shell, a very useful tool for development and prototyping. Subsequent chapters then cover the Spark programming interface in detail, how applications execute on a cluster, and higher-level libraries available on Spark (such as Spark SQL and MLlib).

Supporting Books

If you are a data scientist and don't have much experience with Python, the books *Learning Python* and *Head First Python* (both O'Reilly) are excellent introductions. If you have some Python experience and want more, *Dive into Python* (*http://*

www.diveintopython.net/) (Apress) is a great book to help you get a deeper understanding of Python.

If you are an engineer and after reading this book you would like to expand your data analysis skills, *Machine Learning for Hackers* and *Doing Data Science* are excellent books (both O'Reilly).

This book is intended to be accessible to beginners. We do intend to release a deep-dive follow-up for those looking to gain a more thorough understanding of Spark's internals.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

`Constant width bold`
: Shows commands or other text that should be typed literally by the user.

`Constant width italic`
: Shows text that should be replaced with user-supplied values or by values determined by context.

This element signifies a tip or suggestion.

This element indicates a warning or caution.

Code Examples

All of the code examples found in this book are on GitHub. You can examine them and check them out from *https://github.com/databricks/learning-spark*. Code examples are provided in Java, Scala, and Python.

> Our Java examples are written to work with Java version 6 and higher. Java 8 introduces a new syntax called *lambdas* that makes writing inline functions much easier, which can simplify Spark code. We have chosen not to take advantage of this syntax in most of our examples, as most organizations are not yet using Java 8. If you would like to try Java 8 syntax, you can see the Databricks blog post on this topic (*http://bit.ly/1ywZBs4*). Some of the examples will also be ported to Java 8 and posted to the book's GitHub site.

This book is here to help you get your job done. In general, if example code is offered with this book, you may use it in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Learning Spark* by Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia (O'Reilly). Copyright 2015 Databricks, 978-1-449-35862-4."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com*.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like 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, and hundreds more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://bit.ly/learning-spark*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

The authors would like to thank the reviewers who offered feedback on this book: Joseph Bradley, Dave Bridgeland, Chaz Chandler, Mick Davies, Sam DeHority, Vida Ha, Andrew Gal, Michael Gregson, Jan Joeppen, Stephan Jou, Jeff Martinez, Josh Mahonin, Andrew Or, Mike Patterson, Josh Rosen, Bruce Szalwinski, Xiangrui Meng, and Reza Zadeh.

The authors would like to extend a special thanks to David Andrzejewski, David Buttler, Juliet Hougland, Marek Kolodziej, Taka Shinagawa, Deborah Siegel, Dr. Normen Müller, Ali Ghodsi, and Sameer Farooqui. They provided detailed feedback on the majority of the chapters and helped point out many significant improvements.

We would also like to thank the subject matter experts who took time to edit and write parts of their own chapters. Tathagata Das worked with us on a very tight schedule to finish Chapter 10. Tathagata went above and beyond with clarifying examples, answering many questions, and improving the flow of the text in addition to his technical contributions. Michael Armbrust helped us check the Spark SQL chapter for correctness. Joseph Bradley provided the introductory example for MLlib in Chapter 11. Reza Zadeh provided text and code examples for dimensionality reduction. Xiangrui Meng, Joseph Bradley, and Reza Zadeh also provided editing and technical feedback for the MLlib chapter.

Table of Contents

试读结束，需要全本PDF请购买 www.ertongbook.com

CHAPTER 1

Introduction to Data Analysis with Spark

This chapter provides a high-level overview of what Apache Spark is. If you are already familiar with Apache Spark and its components, feel free to jump ahead to Chapter 2.

What Is Apache Spark?

Apache Spark is a cluster computing platform designed to be *fast* and *general-purpose*.

On the speed side, Spark extends the popular MapReduce model to efficiently support more types of computations, including interactive queries and stream processing. Speed is important in processing large datasets, as it means the difference between exploring data interactively and waiting minutes or hours. One of the main features Spark offers for speed is the ability to run computations in memory, but the system is also more efficient than MapReduce for complex applications running on disk.

On the generality side, Spark is designed to cover a wide range of workloads that previously required separate distributed systems, including batch applications, iterative algorithms, interactive queries, and streaming. By supporting these workloads in the same engine, Spark makes it easy and inexpensive to *combine* different processing types, which is often necessary in production data analysis pipelines. In addition, it reduces the management burden of maintaining separate tools.

Spark is designed to be highly accessible, offering simple APIs in Python, Java, Scala, and SQL, and rich built-in libraries. It also integrates closely with other Big Data tools. In particular, Spark can run in Hadoop clusters and access any Hadoop data source, including Cassandra.

试读结束，需要全本PDF请购买 www.ertongbook.com