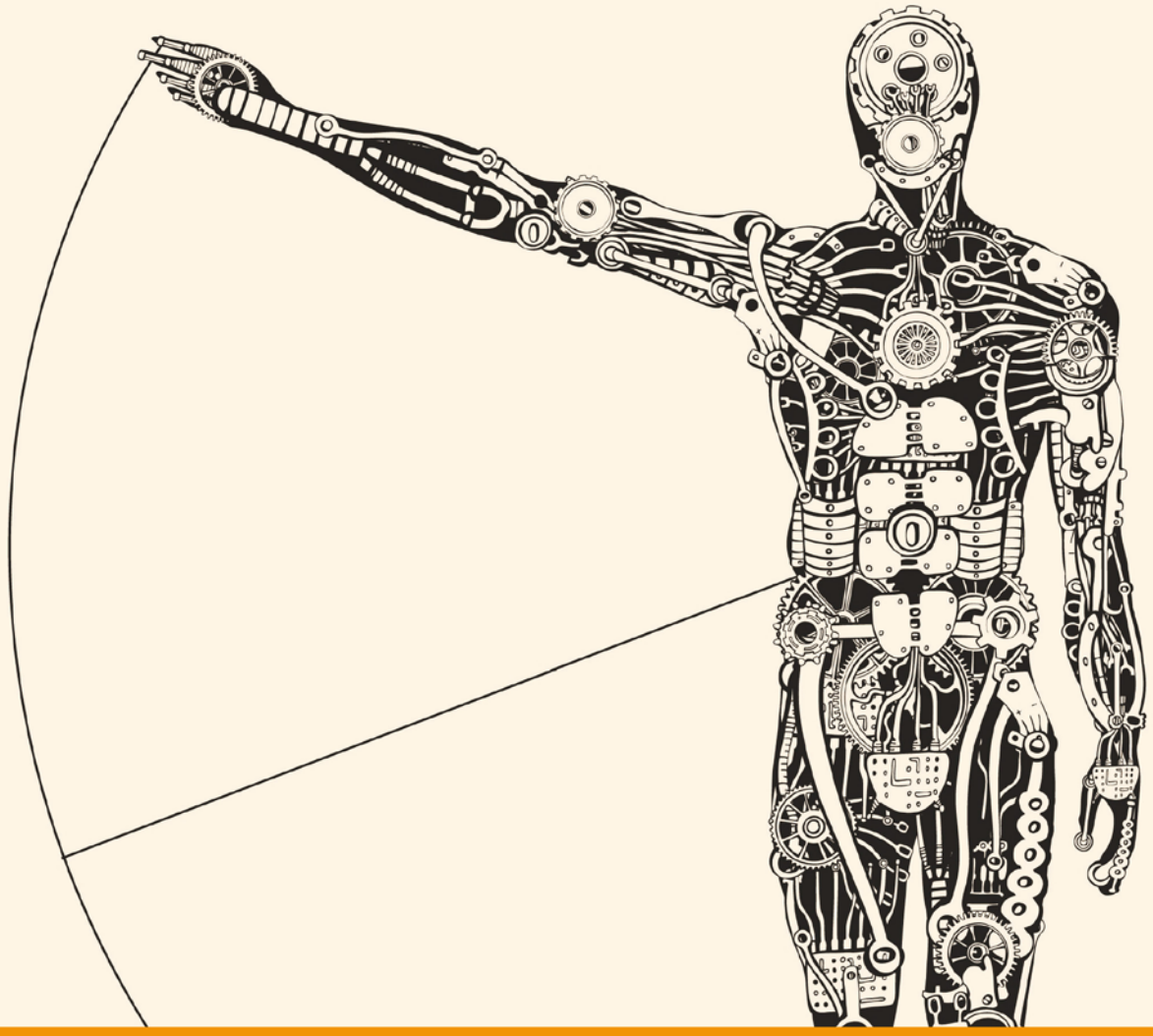

Community Experience Distilled

Fast Data Processing with Spark

Second Edition

Perform real-time analytics using Spark in a fast, distributed, and scalable way

Krishna Sankar
Holden Karau

[PACKT] PUBLISHING open source* community experience distilled

Fast Data Processing with Spark

Second Edition

Perform real-time analytics using Spark in a fast, distributed, and scalable way

Krishna Sankar

Holden Karau

[PACKT] PUBLISHING open source community experience distilled

BIRMINGHAM - MUMBAI

Fast Data Processing with Spark
Second Edition

First published: October 2013

Second edition: March 2015

Production reference: 1250315

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78439-257-4

www.packtpub.com

Credits

Authors

Krishna Sankar

Holden Karau

Reviewers

Robin East

Toni Verbeiren

Lijie Xu

Commissioning Editor

Akram Hussain

Acquisition Editors

Shaon Basu

Kunal Parikh

Content Development Editor

Arvind Koul

Technical Editors

Madhunikita Sunil Chindarkar

Taabish Khan

Copy Editor

Hiral Bhat

Project Coordinator

Neha Bhatnagar

Proofreaders

Maria Gould

Ameesha Green

Joanna McMahon

Indexer

Tejal Soni

Production Coordinator

Nilesh R. Mohite

Cover Work

Nilesh R. Mohite

About the Authors

Krishna Sankar is a chief data scientist at http://www.blackarrow.tv/, where he focuses on optimizing user experiences via inference, intelligence, and interfaces. His earlier roles include principal architect, data scientist at Tata America Intl, director of a data science and bioinformatics start-up, and a distinguished engineer at Cisco. He has spoken at various conferences, such as Strata-Sparkcamp, OSCON, Pycon, and Pydata about predicting NFL (http://goo.gl/movfds), Spark (http://goo.gl/E4kqMD), data science (http://goo.gl/9pyJMH), machine learning (http://goo.gl/SXF53n), and social media analysis (http://goo.gl/D9YpVQ). He was a guest lecturer at Naval Postgraduate School, Monterey. His blogs can be found at https://doubleclix.wordpress.com/. His other passion is Lego Robotics. You can find him at the St. Louis FLL World Competition as the robots design judge.

The credit goes to my coauthor, Holden Karau, the reviewers, and the editors at Packt Publishing. Holden wrote the first edition, and I hope I was able to contribute to the same depth. I am deeply thankful to the reviewers Lijie, Robin, and Toni. They spent time diligently reviewing the material and code. They have added lots of insightful tips to the text, which I have gratefully included. In addition, their sharp eyes caught tons of errors in the code and text. Thanks to Arvind Koul, who has been the chief force behind the book. A great editor is absolutely essential for the completion of a book, and I was lucky to have Arvind. I also want to thank the editors at Packt Publishing: Anila, Madhunikita, Milton, Neha, and Shaon, with whom I had the fortune to work with at various stages. The guidance and wisdom from Joe Matarese, my boss at http://www.blackarrow.tv/, and from Paco Nathan at Databricks are invaluable. My spouse, Usha and son Kaushik, were always with me, cheering me on for any endeavor that I embark upon—mostly successful, like this book, and occasionally foolhardy efforts! I dedicate this book to my mom, who unfortunately passed away last month; she was always proud to see her eldest son as an author.

Holden Karau is a software development engineer and is active in the open source sphere. She has worked on a variety of search, classification, and distributed systems problems at Databricks, Google, Foursquare, and Amazon. She graduated from the University of Waterloo with a bachelor's of mathematics degree in computer science. Other than software, she enjoys playing with fire and hula hoops, and welding.

About the Reviewers

Robin East has served a wide range of roles covering operations research, finance, IT system development, and data science. In the 1980s, he was developing credit scoring models using data science and big data before anyone (including himself) had even heard of those terms! In the last 15 years, he has worked with numerous large organizations, implementing enterprise content search applications, content intelligence systems, and big data processing systems. He has created numerous solutions, ranging from swaps and derivatives in the banking sector to fashion analytics in the retail sector.

Robin became interested in Apache Spark after realizing the limitations of the traditional MapReduce model with respect to running iterative machine learning models. His focus is now on trying to further extend the Spark machine learning libraries, and also on teaching how Spark can be used in data science and data analytics through his blog, Machine Learning at Speed (`http://mlspeed.wordpress.com`).

Before NoSQL databases became the rage, he was an expert on tuning Oracle databases and extracting maximum performance from EMC Documentum systems. This work took him to clients around the world and led him to create the open source profiling tool called DFCprof that is used by hundreds of EMC users to track down performance problems. For many years, he maintained the popular Documentum internals and tuning blog, Inside Documentum (`http://robineast.wordpress.com`), and contributed hundreds of posts to EMC support forums. These community efforts bore fruit in the form of the award of EMC MVP and acceptance into the EMC Elect program.

Toni Verbeiren graduated as a PhD in theoretical physics in 2003. He used to work on models of artificial neural networks, entailing mathematics, statistics, simulations, (lots of) data, and numerical computations. Since then, he has been active in the industry in diverse domains and roles: infrastructure management and deployment, service management, IT management, ICT/business alignment, and enterprise architecture. Around 2010, Toni started picking up his earlier passion, which was then named data science. The combination of data and common sense can be a very powerful basis to make decisions and analyze risk.

Toni is active as an owner and consultant at Data Intuitive (`http://www.data-intuitive.com/`) in everything related to big data science and its applications to decision and risk management. He is currently involved in Exascience Life Lab (`http://www.exascience.com/`) and the Visual Data Analysis Lab (`http://vda-lab.be/`), which is concerned with scaling up visual analysis of biological and chemical data.

I'd like to thank various employers, clients, and colleagues for the insight and wisdom they shared with me. I'm grateful to the Belgian and Flemish governments (FWO, IWT) for financial support of the aforementioned academic projects.

Lijie Xu is a PhD student at the Institute of Software, Chinese Academy of Sciences. His research interests focus on distributed systems and large-scale data analysis. He has both academic and industrial experience in Microsoft Research Asia, Alibaba Taobao, and Tencent. As an open source software enthusiast, he has contributed to Apache Spark and written a popular technical report, named *Spark Internals*, in Chinese at `https://github.com/JerryLead/SparkInternals/tree/master/markdown`.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.PacktPub.com`.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

Apache Spark has captured the imagination of the analytics and big data developers, and rightfully so. In a nutshell, Spark enables distributed computing on a large scale in the lab or in production. Till now, the pipeline collect-store-transform was distinct from the Data Science pipeline reason-model, which was again distinct from the deployment of the analytics and machine learning models. Now, with Spark and technologies, such as Kafka, we can seamlessly span the data management and data science pipelines. We can build data science models on larger datasets, requiring not just sample data. However, whatever models we build can be deployed into production (with added work from engineering on the "ilities", of course). It is our hope that this book would enable an engineer to get familiar with the fundamentals of the Spark platform as well as provide hands-on experience on some of the advanced capabilities.

What this book covers

Chapter 1, *Installing Spark and Setting up your Cluster*, discusses some common methods for setting up Spark.

Chapter 2, *Using the Spark Shell*, introduces the command line for Spark. The Shell is good for trying out quick program snippets or just figuring out the syntax of a call interactively.

Chapter 3, *Building and Running a Spark Application*, covers Maven and sbt for compiling Spark applications.

Chapter 4, *Creating a SparkContext*, describes the programming aspects of the connection to a Spark server, for example, the SparkContext.

Chapter 5, *Loading and Saving Data in Spark*, deals with how we can get data in and out of a Spark environment.

Chapter 6, *Manipulating your RDD*, describes how to program the Resilient Distributed Datasets, which is the fundamental data abstraction in Spark that makes all the magic possible.

Chapter 7, *Spark SQL*, deals with the SQL interface in Spark. Spark SQL probably is the most widely used feature.

Chapter 8, *Spark with Big Data*, describes the interfaces with Parquet and HBase.

Chapter 9, *Machine Learning Using Spark MLlib*, talks about regression, classification, clustering, and recommendation. This is probably the largest chapter in this book. If you are stranded on a remote island and could take only one chapter with you, this should be the one!

Chapter 10, *Testing*, talks about the importance of testing distributed applications.

Chapter 11, Tips and Tricks, distills some of the things we have seen. Our hope is that as you get more and more adept in Spark programming, you will add this to the list and send us your gems for us to include in the next version of this book!

What you need for this book

Like any development platform, learning to develop systems with Spark takes trial and error. Writing programs, encountering errors, agonizing over pesky bugs are all part of the process. We expect a basic level of programming skills—Python or Java—and experience in working with operating system commands. We have kept the examples simple and to the point. In terms of resources, we do not assume any esoteric equipment for running the examples and developing the code. A normal development machine is enough.

Who this book is for

Data scientists and data engineers would benefit more from this book. Folks who have an exposure to big data and analytics will recognize the patterns and the pragmas. Having said that, anyone who wants to understand distributed programming would benefit from working through the examples and reading the book.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "While the methods for loading an RDD are largely found in the `SparkContext` class, the methods for saving an RDD are defined on the RDD classes."

A block of code is set as follows:

```
//Next two lines only needed if you decide to use the assembly plugin
import AssemblyKeys._assemblySettings

scalaVersion := "2.10.4"

name := "groupbytest"

libraryDependencies ++= Seq(
   "org.spark-project" % "spark-core_2.10" % "1.1.0"
)
```

Any command-line input or output is written as follows:

```
scala> val inFile = sc.textFile("./spam.data")
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: " Select **Source Code** from option **2. Choose a package type** and either download directly or select a mirror."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at `http://www.packtpub.com` for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at `questions@packtpub.com`, and we will do our best to address the problem.