Community Experience Distilled

Fast Data Processing with Spark

High-speed distributed computing made easy with Spark

Holden Karau

[PACKT] open source* PUBLISHING
community experience distilled

Fast Data Processing with Spark

High-speed distributed computing made easy with Spark

Holden Karau

[PACKT] PUBLISHING
open source
community experience distilled

BIRMINGHAM - MUMBAI

Fast Data Processing with Spark

First published: October 2013

Production Reference: 1151013

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

ISBN 978-1-78216-706-8

www.packtpub.com

Cover Image by Suresh Mogre (suresh.mogre.99@gmail.com)

Credits

Author
Holden Karau

Reviewers
Wayne Allan
Andrea Mostosi
Reynold Xin

Acquisition Editor
Kunal Parikh

Commissioning Editor
Shaon Basu

Technical Editors
Krutika Parab
Nadeem N. Bagban

Project Coordinator
Amey Sawant

Copy Editors
Brandt D'Mello
Kirti Pai
Lavina Pereira
Tanvi Gaitonde
Dipti Kapadia

Proofreader
Jonathan Todd

Indexer
Rekha Nair

Production Coordinator
Manu Joseph

Cover Work
Manu Joseph

About the Author

Holden Karau is a transgendered software developer from Canada currently living in San Francisco. Holden graduated from the University of Waterloo in 2009 with a Bachelors of Mathematics in Computer Science. She currently works as a Software Development Engineer at Google. She has worked at Foursquare, where she was introduced to Scala. She worked on search and classification problems at Amazon. Open Source development has been a passion of Holden's from a very young age, and a number of her projects have been covered on Slashdot. Outside of programming, she enjoys playing with fire, welding, and dancing. You can learn more at her website (`http://www.holdenkarau.com`), blog (`http://blog.holdenkarau.com`), and github (`https://github.com/holdenk`).

I'd like to thank everyone who helped review early versions of this book, especially Syed Albiz, Marc Burns, Peter J. J. MacDonald, Norbert Hu, and Noah Fiedel.

About the Reviewers

Andrea Mostosi is a passionate software developer. He started software development in 2003 at high school with a single-node LAMP stack and grew with it by adding more languages, components, and nodes. He graduated in Milan and worked on several web-related projects. He is currently working with data, trying to discover information hidden behind huge datasets.

I would like to thank my girlfriend, Khadija, who lovingly supports me in everything I do, and the people I collaborated with—for fun or for work—for everything they taught me. I'd also like to thank Packt Publishing and its staff for the opportunity to contribute to this book.

Reynold Xin is an Apache Spark committer and the lead developer for Shark and GraphX, two computation frameworks built on top of Spark. He is also a co-founder of Databricks which works on transforming large-scale data analysis through the Apache Spark platform. Before Databricks, he was pursuing a PhD in the UC Berkeley AMPLab, the birthplace of Spark.

Aside from engineering open source projects, he frequently speaks at Big Data academic and industrial conferences on topics related to databases, distributed systems, and data analytics. He also taught Palestinian and Israeli high-school students Android programming in his spare time.

www.PacktPub.com

Support files, eBooks, discount offers and more

You might want to visit `www.PacktPub.com` for support files and downloads related to your book.

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.

`http://PacktLib.PacktPub.com`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across 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 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 nine entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface

As programmers, we are frequently asked to solve problems or use data that is too much for a single machine to practically handle. Many frameworks exist to make writing web applications easier, but few exist to make writing distributed programs easier. The Spark project, which this book covers, makes it easy for you to write distributed applications in the language of your choice: Scala, Java, or Python.

What this book covers

Chapter 1, *Installing Spark and Setting Up Your Cluster*, covers how to install Spark on a variety of machines and set up a cluster—ranging from a local single-node deployment suitable for development work to a large cluster administered by a Chef to an EC2 cluster.

Chapter 2, *Using the Spark Shell*, gets you started running your first Spark jobs in an interactive mode. Spark shell is a useful debugging and rapid development tool and is especially handy when you are just getting started with Spark.

Chapter 3, *Building and Running a Spark Application*, covers how to build standalone jobs suitable for production use on a Spark cluster. While the Spark shell is a great tool for rapid prototyping, building standalone jobs is the way you will likely find most of your interaction with Spark to be.

Chapter 4, *Creating a SparkContext*, covers how to create a connection a Spark cluster. SparkContext is the entry point into the Spark cluster for your program.

Chapter 5, *Loading and Saving Your Data*, covers how to create and save RDDs (Resilient Distributed Datasets). Spark supports loading RDDs from any Hadoop data source.