Community Experience Distilled

Apache Spark Graph Processing

Build, process, and analyze large-scale graphs with Spark

Foreword by Denny Lee, Technology Evangelist, Databricks
Advisor, WearHacks

Rindra Ramamonjison

[PACKT] PUBLISHING open source* community experience distilled

Apache Spark Graph Processing

Build, process, and analyze large-scale graphs with Spark

Rindra Ramamonjison

[PACKT] open source
community experience distilled
PUBLISHING
BIRMINGHAM - MUMBAI

Apache Spark Graph Processing

First published: September 2015

Production reference: 1040915

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

ISBN 978-1-78439-180-5

www.packtpub.com

Credits

Author
Rindra Ramamonjison

Reviewer
Thomas W. Dinsmore
Ryan Mccune
Francoise Provencher

Commissioning Editor
Amit Ghodke

Acquisition Editor
Larissa Pinto

Content Development Editor
Dharmesh Parmar

Technical Editor
Prajakta Mhatre

Copy Editor
Yesha Gangani

Project Coordinator
Nikhil Nair

Proofreader
Safis Editing

Indexer
Tejal Soni

Production Coordinator
Aparna Bhagat

Cover Work
Aparna Bhagat

Foreword

Apache Spark is one of the most compelling technologies in the big data space and for good reason. It allows data scientists and data engineers alike to work in their language of choice (Java, Scala, Python, SQL, and R as of this writing) to make sense of their data. As ReynoldXin noted, Apache Spark is the Swiss Army Knife of big data analytics tools. It allows you to use one tool to do many things from real-time streaming to advanced analytics. And in no small part, the versatility and power of GraphX has helped Spark propel forward.

Apache Spark Graph Processing follows Rindra's journey into solving complex analytics problems. As a PhD graduate in electrical engineering from the University of British Columbia, he focused on applying learning and optimization algorithms to achieve energy-efficient wireless networks. As he dove further into these problems, he realized the ease of which he could solve graph-processing problems by using Apache Spark GraphX. With a tutorial style and hands-on projects with interesting datasets, this book is a reflection of his path from getting started with Apache Spark GraphX to iterative graph parallel processing to learning graph structures.

This book is a great jump-start into GraphX, a practical guide for large-scale graph processing, and a testament to the author's enthusiasm for the Spark community (and the community as a whole).

Denny Lee

Technology Evangelist, Databricks

Advisor, WearHacks

About the Author

Rindra Ramamonjison is a fourth year PhD student of electrical engineering at the University of British Columbia, Vancouver. He received his master's degree from Tokyo Institute of Technology. He has played various roles in many engineering companies, within telecom and finance industries. His primary research interests are machine learning, optimization, graph processing, and statistical signal processing. Rindra is also the co-organizer of the Vancouver Spark Meetup.

About the Reviewer

Thomas W. Dinsmore is a consultant and author with more than 30 years of service to enterprises around the world. He is an expert in business analytics, and has working experience with the leading analytic tools, languages, and databases. In his practice, Thomas helps organizations streamline analytics for improved performance and time to value.

Previously, Thomas served with The Boston Consulting Group, IBM, PriceWaterhouseCoopers and SAS, as well as several startups.

Thomas coauthored *Modern Analytics Methodologies and Advanced Analytics Methodologies*, published in 2014 by FT Press. He is currently under contract to publish a book on disruptive technologies in business analytics, scheduled for publication in Q2 2016.

I would like to thank the entire editorial and production team at Packt Publishing, who work tirelessly to bring quality books to the public.

---

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

This book is intended to present the GraphX library for Apache Spark and to teach the fundamental techniques and recipes to process graph data at scale. It is intended to be a self-study step-by-step guide for anyone new to Spark with an interest in or need for large-scale graph processing.

Distinctive features

The focus of this book is on large-scale graph processing with Apache Spark. The book teaches a variety of graph processing abstractions and algorithms and provides concise and sufficient information about them. You can confidently learn all of it and put it to use in different applications.

- **Step-by-step guide**: Each chapter teaches important techniques for every stage of the pipeline, from loading and transforming graph data to implementing graph-parallel operations and machine learning algorithms.
- **Hands-on approach**: We show how each technique works using the Scala REPL with simple examples and by building standalone Spark applications.
- **Detailed code**: All the Scala code in the book is available for download from the book webpage of Packt Publishing.
- **Real-world examples**: We apply these techniques on open datasets collected from a broad variety of applications ranging from social networks to food science and sports analytics.

What this book covers

This book consists of seven chapters. The first three chapters help you to get started quickly with Spark and GraphX. Then, the next two chapters teach the core techniques and abstractions to manipulate and aggregate graph data. Finally, the last two chapters of this book cover more advanced topics such as graph clustering, implementing graph-parallel iterative algorithms with Pregel, and learning methods from graph data.

Chapter 1, *Getting Started with Spark and GraphX*, begins with an introduction to the Spark system, its libraries, and the Scala Build Tool. It explains how to install and leverage Spark on the command line and in a standalone Scala program.

Chapter 2, *Building and Exploring Graphs*, presents the methods for building Spark graphs using illustrative network datasets.

Chapter 3, *Graph Analysis and Visualization*, walks you through the process of exploring, visualizing, and analyzing different network characteristics.

Chapter 4, *Transforming and Shaping Up Graphs to Your Needs*, teaches you how to transform raw datasets into a usable form that is appropriate for later analysis.

Chapter 5, *Creating Custom Graph Aggregation Operators*, teaches you how to create custom graph operations that are tailored to your specific needs with efficiency in mind, using the powerful message-passing aggregation operator in Spark.

Chapter 6, *Iterative Graph-Parallel Processing with Pregel*, explains the inner workings of the Pregel computational model and describes some use cases.

Chapter 7, *Learning Graph Structures*, introduces graph clustering, which is useful for detecting communities in graphs and applies it to a social music database.

What you need for this book

To learn effectively from this book, it is helpful to have a beginner-level programming experience with Scala. However, intermediate functional constructs or Scala-specific syntax are highlighted and explained as they appear in the book. Prior experience with Spark's core API or with the MapReduce framework is beneficial but not required.

It is also beneficial to follow along with the examples, using a Windows or Unix computer with a Java Development Kit environment. More details on the system requirements are described in the first chapter.