



C o m m u n i t y E x p e r i e n c e D i s t i l l e d

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] open source*
PUBLISHING community experience distilled

Apache Spark Graph Processing

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

Rindra Ramamonjison



BIRMINGHAM - MUMBAI

Apache Spark Graph Processing

Copyright © 2015 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

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

Project Coordinator

Nikhil Nair

Reviewer

Thomas W. Dinsmore

Ryan Mccune

Francoise Provencher

Proofreader

Safis Editing

Indexer

Tejal Soni

Commissioning Editor

Amit Ghodke

Production Coordinator

Aparna Bhagat

Acquisition Editor

Larissa Pinto

Cover Work

Aparna Bhagat

Content Development Editor

Dharmesh Parmar

Technical Editor

Prajakta Mhatre

Copy Editor

Yesha Gangani

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	v
Chapter 1: Getting Started with Spark and GraphX	1
Downloading and installing Spark 1.4.1	1
Experimenting with the Spark shell	3
Getting started with GraphX	5
Building a tiny social network	5
Loading the data	6
The property graph	6
Transforming RDDs to VertexRDD and EdgeRDD	7
Introducing graph operations	9
Building and submitting a standalone application	10
Writing and configuring a Spark program	10
Building the program with the Scala Build Tool	14
Deploying and running with spark-submit	15
Summary	16
Chapter 2: Building and Exploring Graphs	17
Network datasets	17
The communication network	18
Flavor networks	18
Social ego networks	19
Graph builders	19
The Graph factory method	19
edgeListFile	20
fromEdges	20
fromEdgeTuples	21
Building graphs	21
Building directed graphs	21
Building a bipartite graph	22
Building a weighted social ego network	26

Computing the degrees of the network nodes	30
In-degree and out-degree of the Enron email network	30
Degrees in the bipartite food network	31
Degree histogram of the social ego networks	32
Summary	33
Chapter 3: Graph Analysis and Visualization	35
Network datasets	36
The graph visualization	36
Installing the GraphStream and BreezeViz libraries	36
Visualizing the graph data	37
Plotting the degree distribution	41
The analysis of network connectedness	43
Finding the connected components	45
Counting triangles and computing clustering coefficients	46
The network centrality and PageRank	49
How PageRank works	49
Ranking web pages	50
Scala Build Tool revisited	51
Organizing build definitions	51
Managing library dependencies	52
A preview of the steps	53
Running tasks with SBT commands	58
Summary	58
Chapter 4: Transforming and Shaping Up Graphs to Your Needs	59
Transforming the vertex and edge attributes	59
mapVertices	60
mapEdges	61
mapTriplets	61
Modifying graph structures	61
The reverse operator	62
The subgraph operator	62
The mask operator	63
The groupEdges operator	63
Joining graph datasets	64
joinVertices	64
outerJoinVertices	64
Example – Hollywood movie graph	65

Data operations on VertexRDD and EdgeRDD	69
Mapping VertexRDD and EdgeRDD	69
Filtering VertexRDDs	70
Joining VertexRDDs	71
Joining EdgeRDDs	72
Reversing edge directions	72
Collecting neighboring information	74
Example – from food network to flavor pairing	74
Summary	78
Chapter 5: Creating Custom Graph Aggregation Operators	79
NCAA College Basketball datasets	79
The aggregateMessages operator	83
EdgeContext	83
Abstracting out the aggregation	85
Keeping things DRY	86
Coach wants more numbers	88
Calculating average points per game	90
Defense stats – D matters as in direction	91
Joining average stats into a graph	92
Performance optimization	95
The MapReduceTriplets operator	98
Summary	98
Chapter 6: Iterative Graph-Parallel Processing with Pregel	99
The Pregel computational model	99
Example – iterating towards the social equality	100
The Pregel API in GraphX	103
Community detection through label propagation	104
The Pregel implementation of PageRank	105
Summary	106
Chapter 7: Learning Graph Structures	107
Community clustering in graphs	107
Spectral clustering	108
Power iteration clustering	108
Applications – music fan community detection	110
Step 1 – load the data into a Spark graph property	111
Step 2 – extract the features of nodes	112
Step 3 – define a similarity measure between two nodes	114

Table of Contents

Step 4 – create an affinity matrix	114
Step 5 – run k-means clustering on the affinity matrix	116
Exercise – collaborative clustering through playlists	120
Summary	120
Appendix: References	121
Chapter 2, Building and Exploring Graphs	121
Chapter 3, Graph Analysis and Visualization	122
Chapter 7, Learning Graph Structures	122
Index	123

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.