PySpark Recipes

A Problem-Solution Approach with PySpark2

Raju Kumar Mishra

Apress®

PySpark Recipes

Raju Kumar Mishra
Bangalore, Karnataka, India

ISBN-13 (pbk): 978-1-4842-3140-1 ISBN-13 (electronic): 978-1-4842-3141-8
https://doi.org/10.1007/978-1-4842-3141-8

Library of Congress Control Number: 2017962438

Cover image by Freepik (www.freepik.com)

Managing Director: Welmoed Spahr
Editorial Director: Todd Green
Acquisitions Editor: Celestin Suresh John
Development Editor: Laura Berendson
Technical Reviewer: Sundar Rajan
Coordinating Editor: Sanchita Mandal
Copy Editor: Sharon Wilkey
Compositor: SPi Global
Indexer: SPi Global
Artist: SPi Global

Distributed to the book trade worldwide by Springer Science + Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC, and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com. For more detailed information, please visit www.apress.com/source-code.

Printed on acid-free paper

To the Almighty, who guides me in every aspect of my life.
And to my mother, Smt. Savitri Mishra, and
my lovely wife, Smt. Smita Rani Pathak.

Contents

About the Author xvii

About the Technical Reviewer xix

Acknowledgments xxi

Introduction xxiii

Chapter 1: The Era of Big Data, Hadoop, and Other Big Data Processing Frameworks 1

Big Data 2

Volume 2

Velocity 3

Variety 3

Veracity 3

Hadoop 3

HDFS 4

MapReduce 5

Apache Hive 6

Apache Pig 7

Apache Kafka 8

Producer 8

Broker 8

Consumer 8

Apache Spark 9
Cluster Managers 10
Standalone Cluster Manager 11
Apache Mesos Cluster Manager 11
YARN Cluster Manager 11
PostgreSQL 12
HBase 12
Chapter 2: Installation 15
Recipe 2-1. Install Hadoop on a Single Machine 16
Problem 16
Solution 16
How It Works 16
Recipe 2-2. Install Spark on a Single Machine 23
Problem 23
Solution 23
How It Works 23
Recipe 2-3. Use the PySpark Shell 25
Problem 25
Solution 25
How It Works 25
Recipe 2-4. Install Hive on a Single Machine 27
Problem 27
Solution 27
How It Works 27
Recipe 2-5. Install PostgreSQL 30
Problem 30
Solution 30
How It Works 30

Recipe 2-6. Configure the Hive Metastore on PostgreSQL 31
Problem 31
Solution 31
How It Works 32
Recipe 2-7. Connect PySpark to Hive 37
Problem 37
Solution 37
How It Works 37
Recipe 2-8. Install Apache Mesos 38
Problem 38
Solution 38
How It Works 38
Recipe 2-9. Install HBase 42
Problem 42
Solution 42
How It Works 42
Chapter 3: Introduction to Python and NumPy 45
Recipe 3-1. Create Data and Verify the Data Type 46
Problem 46
Solution 46
How It Works 46
Recipe 3-2. Create and Index a Python String 48
Problem 48
Solution 48
How It Works 49
Recipe 3-3. Typecast from One Data Type to Another 51
Problem 51
Solution 51
How It Works 51

Recipe 3-4. Work with a Python List 54
Problem 54
Solution 54
How It Works 54
Recipe 3-5. Work with a Python Tuple 58
Problem 58
Solution 58
How It Works 58
Recipe 3-6. Work with a Python Set 60
Problem 60
Solution 60
How It Works 60
Recipe 3-7. Work with a Python Dictionary 62
Problem 62
Solution 62
How It Works 63
Recipe 3-8. Work with Define and Call Functions 64
Problem 64
Solution 64
How It Works 65
Recipe 3-9. Work with Create and Call Lambda Functions 66
Problem 66
Solution 66
How It Works 66
Recipe 3-10. Work with Python Conditionals 67
Problem 67
Solution 67
How It Works 67

Recipe 3-11. Work with Python "for" and "while" Loops 68
Problem 68
Solution 68
How It Works 69
Recipe 3-12. Work with NumPy 70
Problem 70
Solution 70
How It Works 71
Recipe 3-13. Integrate IPython and IPython Notebook with PySpark 78
Problem 78
Solution 79
How It Works 79
Chapter 4: Spark Architecture and the Resilient Distributed Dataset 85
Recipe 4-1. Create an RDD 89
Problem 89
Solution 89
How It Works 89
Recipe 4-2. Convert Temperature Data 91
Problem 91
Solution 91
How It Works 92
Recipe 4-3. Perform Basic Data Manipulation 94
Problem 94
Solution 94
How It Works 95

Recipe 4-4. Run Set Operations 99
Problem 99
Solution 99
How It Works 100
Recipe 4-5. Calculate Summary Statistics 103
Problem 103
Solution 103
How It Works 104
Recipe 4-6. Start PySpark Shell on Standalone Cluster Manager 109
Problem 109
Solution 109
How It Works 109
Recipe 4-7. Start PySpark Shell on Mesos 113
Problem 113
Solution 113
How It Works 113
Chapter 5: The Power of Pairs: Paired RDDs 115
Recipe 5-1. Create a Paired RDD 115
Problem 115
Solution 115
How It Works 116
Recipe 5-2. Aggregate data 119
Problem 119
Solution 119
How It Works 120
Recipe 5-3. Join Data 126
Problem 126
Solution 127
How It Works 128

Recipe 5-4. Calculate Page Rank 132
Problem 132
Solution 132
How It Works 133

Chapter 6: I/O in PySpark 137

Recipe 6-1. Read a Simple Text File 137
Problem 137
Solution 138
How It Works 138

Recipe 6-2. Write an RDD to a Simple Text File 141
Problem 141
Solution 141
How It Works 142

Recipe 6-3. Read a Directory 143
Problem 143
Solution 143
How It Works 144

Recipe 6-4. Read Data from HDFS 145
Problem 145
Solution 145
How It Works 145

Recipe 6-5. Save RDD Data to HDFS 146
Problem 146
Solution 146
How It Works 146

Recipe 6-6. Read Data from a Sequential File 147
Problem 147
Solution 147
How It Works 148
Recipe 6-7. Write Data to a Sequential File 148
Problem 148
Solution 148
How It Works 149
Recipe 6-8. Read a CSV File 150
Problem 150
Solution 150
How It Works 151
Recipe 6-9. Write an RDD to a CSV File 152
Problem 152
Solution 152
How It Works 152
Recipe 6-10. Read a JSON File 154
Problem 154
Solution 154
How It Works 155
Recipe 6-11. Write an RDD to a JSON File 156
Problem 156
Solution 156
How It Works 157
Recipe 6-12. Read Table Data from HBase by Using PySpark 159
Problem 159
Solution 159
How It Works 160

Chapter 7: Optimizing PySpark and PySpark Streaming..... 163

Recipe 7-1. Optimize the Page-Rank Algorithm by Using PySpark Code 164

Problem 164

Solution..... 164

How It Works..... 164

Recipe 7-2. Implement the k-Nearest Neighbors Algorithm by Using PySpark 166

Problem 166

Solution..... 166

How It Works..... 171

Recipe 7-3. Read Streaming Data from the Console Using PySpark Streaming..... 174

Problem 174

Solution..... 174

How It Works..... 175

Recipe 7-4. Integrate PySpark Streaming with Apache Kafka, and Read and Analyze the Data..... 178

Problem 178

Solution..... 178

How It Works..... 179

Recipe 7-5. Execute a PySpark Script in Local Mode 182

Problem 182

Solution..... 182

How It Works..... 183

Recipe 7-6. Execute a PySpark Script Using Standalone Cluster Manager and Mesos Cluster Manager 184

Problem 184

Solution..... 184

How It Works..... 185

Chapter 8: PySparkSQL 187

Recipe 8-1. Create a DataFrame 188

Problem 188

Solution 188

How It Works 188

Recipe 8-2. Perform Exploratory Data Analysis on a DataFrame 195

Problem 195

Solution 195

How It Works 195

Recipe 8-3. Perform Aggregation Operations on a DataFrame 200

Problem 200

Solution 200

How It Works 201

Recipe 8-4. Execute SQL and HiveQL Queries on a DataFrame 207

Problem 207

Solution 207

How It Works 207

Recipe 8-5. Perform Data Joining on DataFrames 210

Problem 210

Solution 210

How It Works 210

Recipe 8-6. Perform Breadth-First Search Using GraphFrames 220

Problem 220

Solution 221

How It Works 222

Recipe 8-7. Calculate Page Rank Using GraphFrames 226
Problem 226
Solution 226
How It Works 226
Recipe 8-8. Read Data from Apache Hive 230
Problem 230
Solution 230
How It Works 232
Chapter 9: PySpark MLlib and Linear Regression 235
Recipe 9-1. Create a Dense Vector 236
Problem 236
Solution 236
How It Works 236
Recipe 9-2. Create a Sparse Vector 237
Problem 237
Solution 237
How It Works 238
Recipe 9-3. Create Local Matrices 239
Problem 239
Solution 239
How It Works 239
Recipe 9-4. Create a Row Matrix 241
Problem 241
Solution 241
How It Works 241
Recipe 9-5. Create a Labeled Point 242
Problem 242
Solution 242
How It Works 242

Recipe 9-6. Apply Linear Regression 243
Problem 243
Solution 243
How It Works 244
Recipe 9-7. Apply Ridge Regression 251
Problem 251
Solution 251
How It Works 252
Recipe 9-8. Apply Lasso Regression 257
Problem 257
Solution 257
How It Works 258

Index 261

About the Author

Raju Kumar Mishra has a strong interest in data science and systems that have the capability of handling large amounts of data and operating complex mathematical models through computational programming. He was inspired to pursue a Master of Technology degree in computational sciences from the Indian Institute of Science in Bangalore, India. Raju primarily works in the areas of data science and its various applications. Working as a corporate trainer, he has developed unique insights that help him in teaching and explaining complex ideas with ease. Raju is also a data science consultant who solves complex industrial problems. He works on programming tools such as R, Python, scikit-learn, Statsmodels, Hadoop, Hive, Pig, Spark, and many others.

About the Technical Reviewer

Sundar Rajan Raman is an artificial intelligence practitioner currently working for Bank of America. He holds a Bachelor of Technology degree from the National Institute of Technology in India. Being a seasoned Java and J2EE programmer, he has worked at companies such as AT&T, Singtel, and Deutsche Bank. He is a messaging platform specialist with vast experience on SonicMQ, WebSphere MQ, and TIBCO software, with respective certifications. His current focus is on artificial intelligence, including machine learning and neural networks. More information is available at https://in.linkedin.com/pub/sundar-rajan-raman/7/905/488.

I would like to thank my wife, Hema, and my daughter, Shriya, for their patience during the review process.

Acknowledgments

My heartiest thanks to the Almighty. I also would like to thank my mother, Smt. Savitri Mishra; my sisters, Mitan and Priya; my cousins, Suchitra and Chandni; and my maternal uncle, Shyam Bihari Pandey; for their support and encouragement. I am very grateful to my sweet and beautiful wife, Smt. Smita Rani Pathak, for her continuous encouragement and love while I was writing this book. I thank my brother-in-law, Mr. Prafull Chandra Pandey, for his encouragement to write this book. I am very thankful to my sisters-in-law, Rinky, Reena, Kshama, Charu, Dhriti, Kriti, and Jyoti for their encouragement as well.
I am grateful to Anurag Pal Sehgal, Saurabh Gupta, Devendra Mani Tripathi, and all my friends. Last but not least, thanks to Coordinating Editor Sanchita Mandal, Acquisitions Editor Celestin Suresh John, and Development Editor Laura Berendson at Apress; without them, this book would not have been possible.

Introduction

This book will take you on an interesting journey to learn about PySpark and big data through a problem-solution approach. Every problem is followed by a detailed, step-by-step answer, which will improve your thought process for solving big data problems with PySpark. This book is divided into nine chapters. Here's a brief description of each chapter:

Chapter 1, "The Era of Big Data, Hadoop, and Other Big Data Processing Frameworks," covers many big data processing tools such as Apache Hadoop, Apache Pig, Apache Hive, and Apache Spark. The shortcomings of Hadoop and the evolution of Spark are discussed. Apache Kafka is explained as a publish-subscribe system. This chapter also sheds light on HBase, a NoSQL database.

Chapter 2, "Installation," will take you to the real battleground. You'll learn how to install many big data processing tools such as Hadoop, Hive, Spark, Apache Mesos, and Apache HBase.

Chapter 3, "Introduction to Python and NumPy," is for newcomers to Python. You will learn about the basics of Python and NumPy by following a problem-solution approach. Problems in this chapter are data-science oriented.

Chapter 4, "Spark Architecture and the Resilient Distributed Dataset," explains the architecture of Spark and introduces resilient distributed datasets. You'll learn about creating RDDs and using data-analysis algorithms for data aggregation, data filtering, and set operations on RDDs.

Chapter 5, "The Power of Pairs: Paired RDD," shows how to create paired RDDs and how to perform data aggregation, data joining, and other algorithms on these paired RDDs.

Chapter 6, "I/O in PySpark," will teach you how to read data from various types of files and save the result as an RDD.

Chapter 7, "Optimizing PySpark and PySpark Streaming," is one of the most important chapters. You will start by optimizing a page-rank algorithm. Then you'll implement a *k*-nearest neighbors algorithm and optimize it by using broadcast variables provided by the PySpark framework. Learning PySpark Streaming will finally lead us into integrating Apache Kafka with the PySpark Streaming framework.

Chapter 8, "PySparkSQL," is paradise for readers who use SQL. But newcomers will also learn PySparkSQL in order to write SQL-like queries on DataFrames by using a problem-solution approach. Apart from DataFrames, we will also implement the graph algorithms breadth-first search and page rank by using the GraphFrames library.

Chapter 9, "PySpark MLlib and Linear Regression," describes PySpark's machine-learning library, MLlib. You will see many recipes on various data structures provided by PySpark MLlib. You'll also implement linear regression. Recipes on lasso and ridge regression are included in the chapter.