Hadoop MapReduce v2 参考手册 第2版（影印版）

adoop MapReduce v2 Cookbook
econd Edition

Thilina Gunarathne 著

[PACKT] PUBLISHING

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Hadoop MapReduce v2 参考手册

第 2 版

(影印版)

Thilina Gunarathne 著

南京　东南大学出版社

图书在版编目(CIP)数据

Hadoop MapReduce v2 参考手册:第 2 版:英文/(美)冈纳拉森(Gunarathne,T.)著.—影印本. --南京:东南大学出版社,2016.1

书名原文:Hapdoop MapReduce v2 Cookbook, Second Edition

ISBN 978-7-5641-6089-0

Ⅰ.①H… Ⅱ.①冈… Ⅲ.①数据处理软件—手册—英文 ②软件工具—程序设计—手册—英文 Ⅳ.①TP274-62 ②TP311.56-62

中国版本图书馆 CIP 数据核字(2015)第 256601 号

Hadoop MapReduce v2 参考手册 第 2 版(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼 2 号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787 毫米×980 毫米　　16 开本
印　　张:20
字　　数:392 千字
版　　次:2016 年 1 月第 1 版
印　　次:2016 年 1 月第 1 次印刷
书　　号:ISBN 978-7-5641-6089-0
定　　价:64.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Credits

Authors
Thilina Gunarathne
Srinath Perera

Reviewers
Skanda Bhargav
Randal Scott King
Dmitry Spikhalskiy
Jeroen van Wilgenburg
Shinichi Yamashita

Commissioning Editor
Edward Gordon

Acquisition Editors
Joanne Fitzpatrick

Content Development Editor
Shweta Pant

Technical Editors
Indrajit A. Das
Pankaj Kadam

Copy Editors
Puja Lalwani
Alfida Paiva
Laxmi Subramanian

Project Coordinator
Shipra Chawhan

Proofreaders
Bridget Braund
Maria Gould
Paul Hindle
Bernadette Watkins

Indexer
Priya Sane

Production Coordinator
Nitesh Thakur

Cover Work
Nitesh Thakur

About the Author

Thilina Gunarathne is a senior data scientist at KPMG LLP. He led the Hadoop-related efforts at Link Analytics before its acquisition by KPMG LLP. He has extensive experience in using Apache Hadoop and its related technologies for large-scale data-intensive computations. He coauthored the first edition of this book, *Hadoop MapReduce Cookbook*, with Dr. Srinath Perera.

Thilina has contributed to several open source projects at Apache Software Foundation as a member, committer, and a PMC member. He has also published many peer-reviewed research articles on how to extend the MapReduce model to perform efficient data mining and data analytics computations in the cloud. Thilina received his PhD and MSc degrees in computer science from Indiana University, Bloomington, USA, and received his bachelor of science degree in computer science and engineering from University of Moratuwa, Sri Lanka.

Acknowledgments

I would like to thank my wife, Bimalee, my son, Kaveen, and my daughter, Yasali, for putting up with me for all the missing family time and for providing me with love and encouragement throughout the writing period. I would also like to thank my parents and siblings. Without their love, guidance, and encouragement, I would not be where I am today.

I really appreciate the contributions from my coauthor, Dr. Srinath Perera, for the first edition of this book. Many of his contributions from the first edition of this book have been adapted to the current book even though he wasn't able to coauthor this book due to his work and family commitments.

I would like to thank the Hadoop, HBase, Mahout, Pig, Hive, Sqoop, Nutch, and Lucene communities for developing great open source products. Thanks to Apache Software Foundation for fostering vibrant open source communities.

Big thanks to the editorial staff at Packt for providing me with the opportunity to write this book and feedback and guidance throughout the process. Thanks to the reviewers of this book for the many useful suggestions and corrections.

I would like to express my deepest gratitude to all the mentors I have had over the years, including Prof. Geoffrey Fox, Dr. Chris Groer, Dr. Sanjiva Weerawarana, Prof. Dennis Gannon, Prof. Judy Qiu, Prof. Beth Plale, and all my professors at Indiana University and University of Moratuwa for all the knowledge and guidance they gave me. Thanks to all my past and present colleagues for the many insightful discussions we've had and the knowledge they shared with me.

About the Author

Srinath Perera (coauthor of the first edition of this book) is a senior software architect at WSO2 Inc., where he overlooks the overall WSO2 platform architecture with the CTO. He also serves as a research scientist at Lanka Software Foundation and teaches as a member of the visiting faculty at Department of Computer Science and Engineering, University of Moratuwa. He is a cofounder of Apache Axis2 open source project, and he has been involved with the Apache Web Service project since 2002 and is a member of Apache Software foundation and Apache Web Service project PMC. Srinath is also a committer of Apache open source projects Axis, Axis2, and Geronimo.

Srinath received his PhD and MSc in computer science from Indiana University, Bloomington, USA, and his bachelor of science in computer science and engineering from University of Moratuwa, Sri Lanka.

Srinath has authored many technical and peer-reviewed research articles; more details can be found on his website. He is also a frequent speaker at technical venues.

Srinath has worked with large-scale distributed systems for a long time. He closely works with big data technologies such as Hadoop and Cassandra daily. He also teaches a parallel programming graduate class at University of Moratuwa, which is primarily based on Hadoop.

I would like to thank my wife, Miyuru, and my parents, whose never-ending support keeps me going. I would also like to thank Sanjiva from WSO2 who encouraged us to make our mark even though project such as these are not in the job description. Finally, I would like to thank my colleagues at WSO2 for ideas and companionship that have shaped the book in many ways.

About the Reviewers

Skanda Bhargav is an engineering graduate from Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India. He did his majors in computer science engineering. He is currently employed with Happiest Minds Technologies, an MNC based out of Bangalore. He is a Cloudera-certified developer in Apache Hadoop. His interests are big data and Hadoop.

He has been a reviewer for the following books and a video, all by Packt Publishing:

- *Instant MapReduce Patterns – Hadoop Essentials How-to*
- *Hadoop Cluster Deployment*
- *Building Hadoop Clusters [Video]*
- *Cloudera Administration Handbook*

I would like to thank my family for their immense support and faith in me throughout my learning stage. My friends have brought the confidence in me to a level that makes me bring out the best in myself. I am happy that God has blessed me with such wonderful people, without whom I wouldn't have tasted the success that I've achieved today.

Randal Scott King is a global consultant who specializes in big data and network architecture. His 15 years of experience in IT consulting has resulted in a client list that looks like a "Who's Who" of the Fortune 500. His recent projects include a complete network redesign for an aircraft manufacturer and an in-store video analytics pilot for a major home improvement retailer.

He lives with his children outside Atlanta, GA. You can visit his blog at `www.randalscottking.com`.

Dmitry Spikhalskiy currently holds the position of software engineer in a Russian social network service, Odnoklassniki, and is working on a search engine, video recommendation system, and movie content analysis.

Previously, Dmitry took part in developing Mind Labs' platform, infrastructure, and benchmarks for a high-load video conference and streaming service, which got "The biggest online-training in the world" Guinness world record with more than 12,000 participants. As a technical lead and architect, he launched a mobile social banking start-up called Instabank. He has also reviewed *Learning Google Guice* and *PostgreSQL 9 Admin Cookbook*, both by Packt Publishing.

Dmitry graduated from Moscow State University with an MSc degree in computer science, where he first got interested in parallel data processing, high-load systems, and databases.

Jeroen van Wilgenburg is a software craftsman at JPoint (`http://www.jpoint.nl`), a software agency based in the center of the Netherlands. Their main focus is on developing high-quality Java and Scala software with open source frameworks.

Currently, Jeroen is developing several big data applications with Hadoop, MapReduce, Storm, Spark, Kafka, MongoDB, and Elasticsearch.

Jeroen is a car enthusiast and likes to be outdoors, usually training for a triathlon. In his spare time, Jeroen writes about his work experience at `http://vanwilgenburg.wordpress.com`.

Shinichi Yamashita is a solutions architect at System Platform Sector in NTT DATA Corporation, Japan. He has more than 9 years of experience in software and middleware engineering (Apache, Tomcat, PostgreSQL, Hadoop Ecosystem, and Spark). Shinichi has written a few books on Hadoop in Japanese.

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

We are currently facing an avalanche of data, and this data contains many insights that hold the keys to success or failure in the data-driven world. Next generation Hadoop (v2) offers a cutting-edge platform to store and analyze these massive data sets and improve upon the widely used and highly successful Hadoop MapReduce v1. The recipes that will help you analyze large and complex datasets with next generation Hadoop MapReduce will provide you with the skills and knowledge needed to process large and complex datasets using the next generation Hadoop ecosystem.

This book presents many exciting topics such as MapReduce patterns using Hadoop to solve analytics, classifications, and data indexing and searching. You will also be introduced to several Hadoop ecosystem components including Hive, Pig, HBase, Mahout, Nutch, and Sqoop.

This book introduces you to simple examples and then dives deep to solve in-depth big data use cases. This book presents more than 90 ready-to-use Hadoop MapReduce recipes in a simple and straightforward manner, with step-by-step instructions and real-world examples.

What this book covers

Chapter 1, *Getting Started with Hadoop v2*, introduces Hadoop MapReduce, YARN, and HDFS, and walks through the installation of Hadoop v2.

Chapter 2, *Cloud Deployments – Using Hadoop Yarn on Cloud Environments*, explains how to use Amazon Elastic MapReduce (EMR) and Apache Whirr to deploy and execute Hadoop MapReduce, Pig, Hive, and HBase computations on cloud infrastructures.

Chapter 3, *Hadoop Essentials – Configurations, Unit Tests, and Other APIs*, introduces basic Hadoop YARN and HDFS configurations, HDFS Java API, and unit testing methods for MapReduce applications.

Chapter 4, *Developing Complex Hadoop MapReduce Applications*, introduces you to several advanced Hadoop MapReduce features that will help you develop highly customized and efficient MapReduce applications.

Chapter 5, Analytics, explains how to perform basic data analytic operations using Hadoop MapReduce.

Chapter 6, Hadoop Ecosystem – Apache Hive, introduces Apache Hive, which provides data warehouse capabilities on top of Hadoop, using a SQL-like query language.

Chapter 7, Hadoop Ecosystem II – Pig, HBase, Mahout, and Sqoop, introduces the Apache Pig data flow style data-processing language, Apache HBase NoSQL data storage, Apache Mahout machine learning and data-mining toolkit, and Apache Sqoop bulk data transfer utility to transfer data between Hadoop and the relational databases.

Chapter 8, Searching and Indexing, introduces several tools and techniques that you can use with Apache Hadoop to perform large-scale searching and indexing.

Chapter 9, Classifications, Recommendations, and Finding Relationships, explains how to implement complex algorithms such as classifications, recommendations, and finding relationships using Hadoop.

Chapter 10, Mass Text Data Processing, explains how to use Hadoop and Mahout to process large text datasets and how to perform data preprocessing and loading of operations using Hadoop.

What you need for this book

You need a moderate knowledge of Java and access to the Internet and a computer that runs a Linux operating system.

Who this book is for

If you are a big data enthusiast and wish to use Hadoop v2 to solve your problems, then this book is for you. This book is for Java programmers with little to moderate knowledge of Hadoop MapReduce. This is also a one-stop reference for developers and system admins who want to quickly get up to speed with using Hadoop v2. It would be helpful to have a basic knowledge of software development using Java and a basic working knowledge of Linux.

Conventions

In this book, you will find a number of styles of text 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: "The following are the descriptions of the properties we used in the `hadoop.properties` file."

A block of code is set as follows:

```
Path file = new Path("demo.txt");
FSDataOutputStream outStream = fs.create(file);
outStream.writeUTF("Welcome to HDFS Java API!!!");
outStream.close();
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
Job job = Job.getInstance(getConf(), "MLReceiveReplyProcessor");
job.setJarByClass(CountReceivedRepliesMapReduce.class);
job.setMapperClass(AMapper.class);
job.setReducerClass(AReducer.class);
job.setNumReduceTasks(numReduce);

job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
job.setInputFormatClass(MBoxFileInputFormat.class);
FileInputFormat.setInputPaths(job, new Path(inputPath));
FileOutputFormat.setOutputPath(job, new Path(outputPath));

int exitStatus = job.waitForCompletion(true) ? 0 : 1;
```

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

```
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET
/shuttle/countdown/countdown.html HTTP/1.0" 200 3985
```

New terms and **important words** are shown in bold. Words that you see on the screen, in menus or dialog boxes for example, appear in the text like this: "Select **Custom Action** in the **Add Bootstrap Actions** drop-down box. Click on **Configure and add**."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.