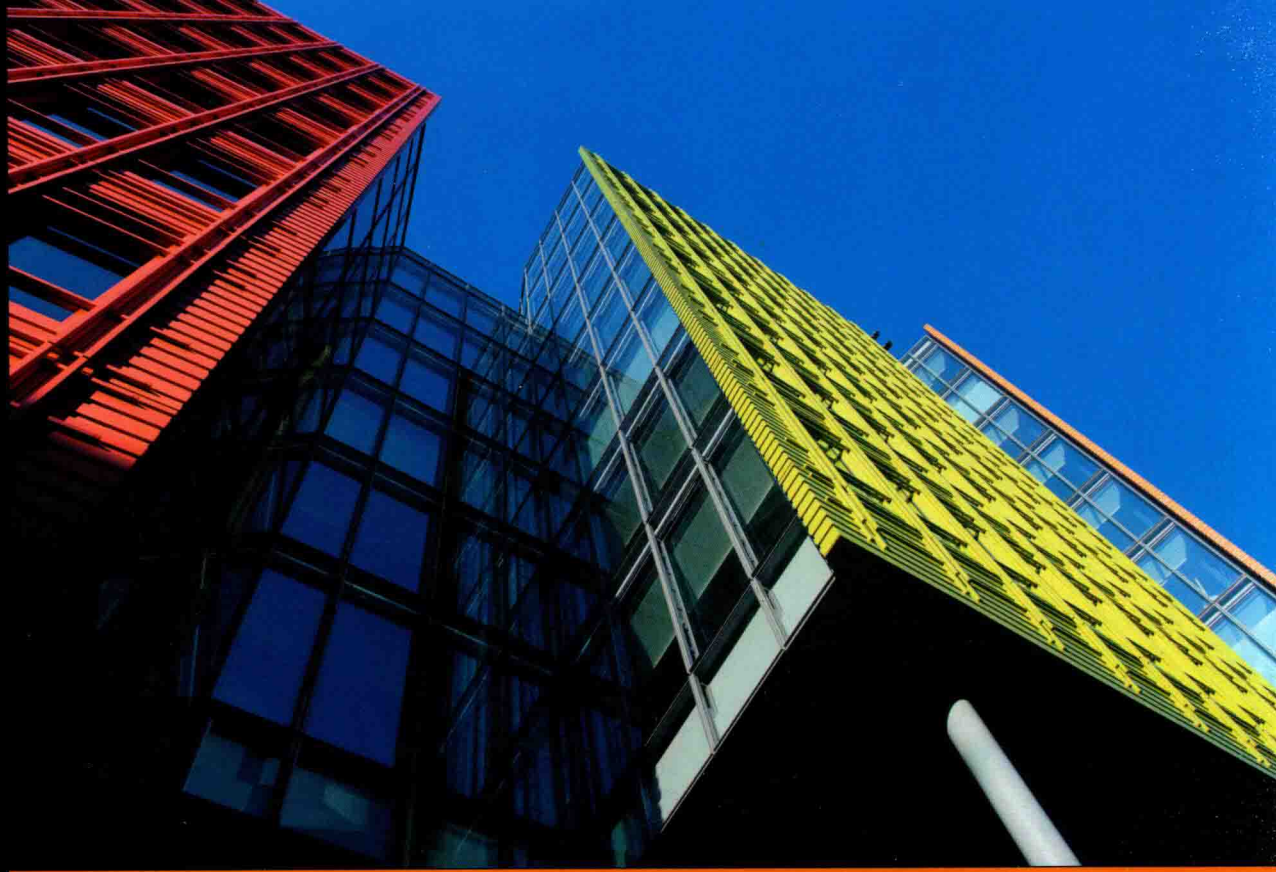

大数据分析：R语言实现

（影印版）

Big Data Analytics with R

Simon Walkowiak 著

[PACKT] PUBLISHING

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

大数据分析：R语言实现(影印版)

Big Data Analytics With R

Simon Walkowiak 著

南京　东南大学出版社

图书在版编目(CIP)数据

大数据分析:R 语言实现:英文/(英)西蒙·沃克威克(Simon Walkowiak)著. —影印本. —南京:东南大学出版社,2017.10

书名原文:Big Data Analytics With R

ISBN 978-7-5641-7361-6

Ⅰ.①大… Ⅱ.①西… Ⅲ.①程序语言—程序设计—英文 Ⅳ.①TP312

中国版本图书馆 CIP 数据核字(2017)第 192628 号

图字:10-2017-115 号

大数据分析:R 语言实现(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:31.5
字　　数:617 千字
版　　次:2017 年 10 月第 1 版
印　　次:2017 年 10 月第 1 次印刷
书　　号:ISBN 978-7-5641-7361-6
定　　价:94.00 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

Simon Walkowiak

Reviewer

Zacharias Voulgaris

Dipanjan Sarkar

Commissioning Editor

Akram Hussain

Acquisition Editor

Sonali Vernekar

Content Development Editor

Onkar Wani

Technical Editor

Sushant S Nadkar

Copy Editor

Safis Editing

Project Coordinator

Ulhas Kambali

Proofreader

Safis Editing

Indexer

Tejal Daruwale Soni

Graphics

Kirk D'Penha

Production Coordinator

Arvindkumar Gupta

About the Author

Simon Walkowiak is a cognitive neuroscientist and a managing director of Mind Project Ltd – a Big Data and Predictive Analytics consultancy based in London, United Kingdom. As a former data curator at the UK Data Service (UKDS, University of Essex) – European largest socio-economic data repository, Simon has an extensive experience in processing and managing large-scale datasets such as censuses, sensor and smart meter data, telecommunication data and well-known governmental and social surveys such as the British Social Attitudes survey, Labour Force surveys, Understanding Society, National Travel survey, and many other socio-economic datasets collected and deposited by Eurostat, World Bank, Office for National Statistics, Department of Transport, NatCen and International Energy Agency, to mention just a few. Simon has delivered numerous data science and R training courses at public institutions and international companies. He has also taught a course in *Big Data Methods in R* at major UK universities and at the prestigious Big Data and Analytics Summer School organized by the Institute of Analytics and Data Science (IADS).

Acknowledgement

The inspiration for writing this book came directly from the brilliant work and dedication of many R developers and users, whom I would like to thank first for creating a vibrant and highly-supportive community that nourishes the progress of publicly accessible data analytics and development of R language. However, this book would never be completed if I wasn't surrounded with love and unconditional support from my partner Ignacio, who always knew how to encourage and motivate me, particularly in moments of my weakness and when I lacked creativity.

I would also like to thank other members of my family, especially my father Peter, who despite not sharing my excitement of data science, always listens patiently to my stories about emerging Big Data technologies and their use cases.

Also, I dedicate this book to my friends and former colleagues from UK Data Service at the University of Essex, where I had an opportunity to work with amazing individuals and experience the best practices in robust data management and processing.

Finally, I highly appreciate the hard work, expertise and feedback offered by many people involved in the creation of this book at Packt Publishing – especially my content development editor Onkar Wani, publishers, and the reviewers, who kindly shared their knowledge with me in order to create a quality and well-received publication.

About the Reviewers

Dr. Zacharias Voulgaris was born in Athens, Greece. He studied Production Engineering and Management at the Technical University of Crete, shifted to Computer Science through a Masters in Information Systems & Technology (City University, London), and then to Data Science through a PhD on Machine Learning (University of London). He has worked at Georgia Tech as a Research Fellow, at an e-marketing startup in Cyprus as an SEO manager, and as a Data Scientist in both Elavon (GA) and G2 (WA). He also was a Program Manager at Microsoft, on a data analytics pipeline for Bing.

Zacharias has authored two books and several scientific articles on Machine Learning and as well as a couple of articles on AI topics. His first book, Data Scientist - The Definitive Guide to Becoming a Data Scientist (Technics Publications), has been translated into Korean and Chinese, while his latest one, Julia for Data Science (Technics Publications) is coming out this September. He has also reviewed a number of data science books (mainly on Python and R) and has a passion for new technologies, literature, and music.

I'd like to thank the people at Packt for inviting me to review this book and for promoting Data Science and particularly Julia through their books. Also, a big thanks to all the great authors out there who choose to publish their work through the lesser-known publishers, keeping the whole process of sharing knowledge a democratic endeavor.

Dipanjan Sarkar is a Data Scientist at Intel, the world's largest silicon company which is on a mission to make the world more connected and productive. He primarily works on analytics, business intelligence, application development and building large scale intelligent systems. He received his Master's degree in Information Technology from the International Institute of Information Technology, Bangalore. His area of specialization includes software engineering, data science, machine learning and text analytics.

Dipanjan's interests include learning about new technology, disruptive start-ups, data science and more recently deep learning. In his spare time he loves reading, writing, gaming and watching popular sitcoms. He has authored a book on Machine Learning titled *R Machine Learning by Example, Packt Publishing* and also acted as a technical reviewer for several books on Machine Learning and Data Science from Packt Publishing.

www.PacktPub.com

eBooks, discount offers, and more

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 `customercare@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

Table of Contents

Preface

We live in times of Internet of Things—a large, world-wide network of interconnected devices, sensors, applications, environments, and interfaces. They generate, exchange, and consume massive amounts of data on a daily basis, and the ability to harness these huge quantities of information can provide us with novel understanding of physical and social phenomena.

The recent rapid growth of various open source and proprietary big data technologies allows deep exploration of these vast amounts of data. However, many of them are limited in terms of their statistical and data analytics capabilities. Some others implement techniques and programming languages that many classically educated statisticians and data analysts are simply unfamiliar with and find them difficult to apply in real-world scenarios.

R programming language—an open source, free, extremely versatile statistical environment, has a potential to fill this gap by providing users with a large variety of highly optimized data processing methods, aggregations, statistical tests, and machine learning algorithms with a relatively user-friendly and easily customizable syntax.

This book challenges traditional preconceptions about R as a programming language that does not support big data processing and analytics. Throughout the chapters of this book, you will be exposed to a variety of core R functions and a large array of actively maintained third-party packages that enable R users to benefit from most recent cutting-edge big data technologies and frameworks, such as Hadoop, Spark, H2O, traditional SQL-based databases, such as SQLite, MariaDB, and PostgreSQL, and more flexible NoSQL databases, such as MongoDB or HBase, to mention just a few. By following the exercises and tutorials contained within this book, you will experience firsthand how all these tools can be integrated with R throughout all the stages of the Big Data Product Cycle, from data import and data management to advanced analytics and predictive modeling.

What this book covers

`Chapter 1`, *The Era of "Big Data"*, gently introduces the concept of Big Data, the growing landscape of large-scale analytics tools, and the origins of R programming language and the statistical environment.

`Chapter 2`, *Introduction to R Programming Language and Statistical Environment*, explains the most essential data management and processing functions available to R users. This chapter also guides you through various methods of Exploratory Data Analysis and hypothesis testing in R, for instance, correlations, tests of differences, ANOVAs, and Generalized Linear Models.

`Chapter 3`, *Unleashing the Power of R From Within*, explores possibilities of using R language for large-scale analytics and out-of-memory data on a single machine. It presents a number of third-party packages and core R methods to address traditional limitations of Big Data processing in R.

`Chapter 4`, *Hadoop and MapReduce Framework for R*, explains how to create a cloud-hosted virtual machine with Hadoop and to integrate its HDFS and MapReduce frameworks with R programming language. In the second part of the chapter, you will be able to carry out a large-scale analysis of electricity meter data on a multinode Hadoop cluster directly from the R console.

`Chapter 5`, *R with Relational Database Management Systems (RDBMSs)*, guides you through the process of setting up and deploying traditional SQL databases, for example, SQLite, PostgreSQL and MariaDB/MySQL, which can be easily integrated with their current R-based data analytics workflows. The chapter also provides detailed information on how to build and benefit from a highly scalable Amazon Relational Database Service instance and query its records directly from R.

`Chapter 6`, *R with Non-Relational (NoSQL) Databases*, builds on the skills acquired in the previous chapters and allows you to connect R with two popular nonrelational databases a.) a fast and user-friendly MongoDB installed on a Linux-run virtual machine, and b.) HBase database operated on a Hadoop cluster run as part of the Azure HDInsight service.

`Chapter 7`, *Faster than Hadoop: Spark with R*, presents a practical example and a detailed explanation of R integration with the Apache Spark framework for faster Big Data manipulation and analysis. Additionally, the chapter shows how to use Hive database as a data source for Spark on a multinode cluster with Hadoop and Spark installed.

`Chapter 8`, *Machine Learning Methods for Big Data in R*, takes you on a journey through the most cutting-edge predictive analytics available in R. Firstly, you will perform fast and highly optimized Generalized Linear Models using Spark MLlib library on a multinode Spark HDInsight cluster. In the second part of the chapter, you will implement Naïve Bayes and multilayered Neural Network algorithms using R's connectivity with H2O-an award-winning, open source, big data distributed machine learning platform.

`Chapter 9`, *The Future of R: Big, Fast and Smart Data*, wraps up the contents of the earlier chapters by discussing potential areas of development for R language and its opportunities in the landscape of emerging Big Data tools.

Online Chapter, Pushing R Further, available at `https://www.packtpub.com/sites/default/files/downloads/5396_6457OS_PushingRFurther.pdf`, enables you to configure and deploy their own scaled-up and Cloud-based virtual machine with fully operational R and RStudio Server installed and ready to use.

What you need for this book

All the code snippets presented in the book have been tested on a Mac OS X (Yosemite) running on a personal computer equipped with 2.3 GHz Intel Core i5 processor, 1 TB Solid State hard drive, and 16 GB of RAM. It is recommended that readers run the scripts on a Mac OS X or Windows machine with at least 4 GB of RAM. In order to benefit from the instructions presented throughout the book, it is advisable that readers install most recent R and RStudio on their machines as well as at least one of the popular web browsers: Mozilla Firefox, Chrome, Safari, or Internet Explorer.

Who this book is for

This book is intended for middle level data analysts, data engineers, statisticians, researchers, and data scientists, who consider and plan to integrate their current or future big data analytics workflows with R programming language.

It is also assumed that readers will have some previous experience in data analysis and the understanding of data management and algorithmic processing of large quantities of data. However, they may lack specific R skills related to particular open source big data tools.

Conventions

In this book, you will find a number of text styles 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 `-getmerge` option allows to merge all data files from a specified directory on HDFS."

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

```
$ sudo -u hdfs hadoop fs -ls /user
```

New terms and important words are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "Clicking the **Next** button moves you to the next screen."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files for this book from your account at `http://www.packtpub.com`. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.