R for Cloud Computing

A Ohri

R for Cloud Computing

An Approach for Data Scientists

Springer

A Ohri
Decisionstats.com
Delhi, India

ISBN 978-1-4939-1701-3 ISBN 978-1-4939-1702-0 (eBook)
DOI 10.1007/978-1-4939-1702-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014950411

© Springer Science+Business Media New York 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*Dedicated to my six-year-old son Kush.
All that is good in me comes from your love*

Preface

I started writing this book on October 10, 2012, and finished its first draft on July 30, 2014, with 54,000 words. I would like to thank my editors and team members at Springer for their patience.

Origin I started learning analytics in 2004, and it was just known as analytics back then. We had moved on from knowledge discovery in databases, and we had not invented the buzzword called data science or Big Data yet. Analytics had two parts, reporting (pulling data, aggregating it, slicing it for a custom view, presenting it in a spreadsheet) and modeling (predictive with regression models and forecasting models). Analytics demanded a secure data environment, and all my clients insisted on retaining data in their premises. Once in a while, I used PGP for encrypting data transfers, and sometimes we used Remote Desktop to connect to remote servers. While working with sensitive data, we used remote submit on remote data warehouses using the analytics stack by the primary analytics player, the SAS Institute.

In the mid-2000s, I came across “cloud computing” as a paradigm in renting hardware and computing via the Internet. A friend of mine, a respected MBA, once called cloud computing as a lot of servers sitting together, and it took me many years to comprehend that simplistic impression. I first came to be involved with the R language in 2007 as I created my own startup, Decisionstats.com, in business analytics writing and consulting. The R language seemed to have a steep learning curve, but it was truly of immense benefit to my fledgling consulting practice.

Once I trained my mind to deal with the vagaries of lists, arrays, and data frames of R, I started off with analyzing projects and doing research. Of course, like anyone who has created a startup, I had help, a lot of help, from some of the best brains on the planet. The developments in technology and in open source software also proved to be of much help. I thank the developers of the R project for this as well as the broader community. One you get familiar with the personalities and their mannerisms, it will prove to be one of the best communities in the technology world. R was associated with some drawbacks in the 2007–2008 era, mainly that it only used data in its memory RAM. This meant it was limited to around 2–3 GB of data, unless one used a database (*those days it meant RDBMS, but now we have*

many kinds). The second flaw in my opinion was the speed of processing, especially compared to mainstream analytics software used for business processes.

Around 2009 I first came across the word Big Data. It basically meant data that had bigger volume, faster velocities of refresh and creation, and variety in terms of formats. Volume, Velocity, and Variety are what separate the Big Data people from the rest even today.

Around September 2010, I created an easy-to-use tutorial on my Decision-stats.com blog on using R from Amazon's cloud. While the very first tutorial on the topic was created by Robert Grossman, I simplified it further with a few screenshots. This was primarily for my own use and remembering.

I was a business analytics user, and I sometimes got confused in the online documentation of R, and I found that many other people had exactly the same issue—there was no proper indexed R Online Doc Version 9 to help people like me. It is the huge and ongoing traffic to these tutorials that motivated me to write a book hoping to present a collection of DIY cloud and R for the average common tech geek.

Scope The purpose of the book is to introduce R and cloud computing to the professional practitioner and turn them into data scientists in the process. Chapter 1 gives an Introduction. Chapter 2 describes an approach for people to think like data scientists. Chapter 3 presents choices (some of them confusing) that confront a person navigating R on the cloud. Chapter 4 deals with setting up R on the cloud infrastructure and offers different perspectives and interfaces. Chapter 5 deals with working in R and is aimed at people new to R. Chapter 6 deals with R and Big Data and introduces the reader to the various paradigms in it. Chapter 7 moves beyond Chap. 4's infrastructure as a service and deals with how the R system can interface with Cloud Applications and Services. Chapter 7 is actually a use case chapter for using R on the cloud. Chapter 8 reviews ways to secure cloud as this is a constant insecurity in transitioning to the cloud. Chapter 9 deals with training literature to further help the reader.

Purpose The book has been written from a practical use case perspective. I find that information asymmetry and brand clutter have managed to confuse audiences about the true benefits and costs of a cloud-hosted open source analytics environment. The book is written for R because at the time of publication it remains the most widely used statistical language in the world with the biggest library of open source packages in statistics related to business analytics and the cheapest total cost of ownership including training, transition, and license costs. An earlier book on *R for Business Analytics* has also been written by me and is available at <http://www.springer.com/statistics/book/978-1-4614-4342-1>.

With over 7000 packages (a very dynamic number) and over 2 million users (a rapidly increasing number), R has a significant lead over other statistical software in terms of a broad library of packages for all analytical needs, and that lead continues to grow, thanks to a highly motivated team of volunteers and developers. R is not going anywhere but up, and for the analytics shop it pays to diversify a bit into the R space.

Plan I will continue to use screenshots as a tutorial device and easy-to-use methods (like graphical User interfaces) to help you analyze much more data at much lower cost. If you use R on the cloud, in an optimum manner, nothing will ever come close in terms of speed, reliability, security, as well as lowered costs (okay, maybe Python will come close—I grant you that, but Pythonistas are still catching up on the statistical libraries). I have tried to make it easy for readers to navigate both R and the cloud. In doing so, I have deliberately adopted a readable everyday conversational style. Every chapter has cited references as well as a few do-it-yourself tutorials.

Intended Audience This is a book for business analysts who are curious about using the cloud. It is also a book for people who use cloud computing and wonder what the buzz on R is all about. Some interviews of well-known practitioners have been included, and these will help decision makers at the CIO level to fine-tune their managerial perspectives on the choices and dilemmas associated with changing to cloud-hosted open source analytics.

Afterthoughts I have focussed on practical solutions. I will therefore proceed on the assumption that the user wants to perform analytics at the lowest cost and greatest accuracy, robustness, and ease possible. I would thus not suggest purely open source solutions when I feel the user is better off with existing software from vendors. I would also recommend in my blog writing and consulting some alternative languages including Python, Scala, Julia, and even SAS for specific use cases. This is because I believe no one software can be suitable for all the user needs, and each user may have their own need based on usage, context, constraints (time, money, training), and flexibility in transitioning to a new solution. I believe new innovation will constantly replace old, legacy solutions, and Andy Grove’s advice on “where only the paranoid survive” will be of use to people embarking on a lifelong journey of learning cutting-edge technology which business analytics commits them to.

The instructions and tutorials within this book have no warranty, and using them will be at your own risk. Yet, this is needed because cloud computing can have both existing as well as new security issues as well as constant changes in user interfaces, licensing conditions, regulatory restrictions including those based on user data geography, and price changes besides the usual tendency of unexpected technological changes. One reason I have taken longer to write this book compared to my earlier book is my desire to constantly eliminate what is not needed anymore in the current scenario for enterprises and students wishing to use R and the cloud. As a special note on the formatting of this manuscript, I mostly write on Google Docs, but here I am writing using the GUI Lyx for the typesetting software LaTeX, and I confess I am not very good at it. However, having written one earlier book on R, *R for Business Analytics* (Springer 2012); 1,900 blog posts; and 100+ paid articles (including 150+ interviews on technology), I hope this book is better formatted and readable than the last one. And yet one more thing—due to a huge number of PDF downloads and torrents of my previous book, I have deliberately made some (but not all) codes within screenshots. This effectively means the print

copy of the book would be much better formatted and more easily readable than the electronic version.

I do hope the book is read by both Chief Technology Officers keen to move to open source analytics based on the cloud and students wishing to enter a potentially lucrative career as data scientists. The R Web, the R Apache projects, and demonstration sites by UCLA, Dataspora, and Vanderbilt University are the well-known early implementations of cloud computing and R—as of today, this space is taken by Revolution Analytics (EC2 licensing of their product, RevoDeployR), Open CPU, and RStudio Servers including the Shiny Project.

R is well known for excellent graphics but is not so suitable for bigger data sets in its native easy-to-use open source version. Using cloud computing removes this hurdle very easily; you just increase the RAM on your instance. The enterprise CTO can thus reduce costs incredibly by shrinking software and hardware costs. The book reflects the current landscape of cloud computing providers, so much attention is devoted to Amazon, then to Google, then to Microsoft Azure, with a small mention of IBM and Oracle's efforts as well.

Delhi, India

A Ohri

Acknowledgments I am grateful to many people working in both the cloud and R communities for making this book possible. I would like to thank Anne Milley, JMP; Bob Muenchen, author of *R for SAS and SPSS Users*; Jerooen Ooms; Jan De Leeuw; Gregory Piatetsky Shapiro; Markus Schmidberger; Dr Ingo Mieressa; all readers of Decisionstats.com; Gergely Rapporter; all the R package creators; everyone interviewed by me in DecisionStats.com; and Vignesh Prajapathi and his Decisionstats.com intern Chandan. A writer is a helpless baby in terms of material and practical needs. I would also like to thank my bhais (Kala, Sonu, Micky), my friends (Namy, Tammy, Sam), and my loving family in Delhi, Mumbai, and Calgary.

Contents

1	Introduction to R for Cloud Computing	1
1.1	What Is R Really?	1
1.2	What Is Cloud Computing?	2
1.2.1	What Is SaaS, PaaS, and IaaS?	3
1.3	Why Should Cloud Users Learn More About R?	4
1.4	Big Data and R	5
1.5	Why Should R Users Learn More About the Cloud?	6
1.6	How Can You Use R on the Cloud Infrastructure?	6
1.7	How Long Does It Take to Learn and Apply R on a Cloud?	7
1.8	Commercial and Enterprise Versions of R	7
1.9	Other Versions of R	9
1.10	Cloud Specific Packages of R	10
1.11	Examples of R and Cloud Computing	10
1.12	How Acceptable Is R and the Cloud to Enterprises?	12
1.12.1	SAP Hana with R Interview 1 (7 June 2012)	12
1.12.2	SAP Hana with R Interview 2–14 June 2012	14
1.12.3	TIBCO TERR	14
1.13	What Is the Extent of Data that Should Be on Cloud Versus Local Machines?	16
1.14	Alternatives to R	17
1.15	Interview of Prof Jan De Leeuw, Founder of JSS	18
2	An Approach for Data Scientists	21
2.1	Where to Get Data?	22
2.2	Where to Process Data?	23
2.2.1	Cloud Processing	24
2.3	Where to Store Data?	24
2.3.1	Cloud Storage	25
2.3.2	Databases on the Cloud	26
2.4	Basic Statistics for Data Scientists	27
2.5	Basic Techniques for Data Scientists	28

2.6	How to Store Output?	28
2.7	How to Share Results and Promote Yourself?.....	29
2.8	How to Move or Query Data?.....	29
2.8.1	Data Transportation on the Cloud	29
2.9	How to Analyse Data?.....	30
2.9.1	Data Quality—Tidy Data	30
2.9.2	Data Quality—Treatment	30
2.9.3	Data Quality—Transformation	31
2.9.4	Split Apply Combine	31
2.10	How to Manipulate Data?	31
2.10.1	Data Visualization.....	32
2.11	Project Methodologies.....	32
2.12	Algorithms for Data Science	33
2.13	Interview of John Myles White, co-author of Machine Learning for Hackers	36
3	Navigating the Choices in R and Cloud Computing	37
3.1	Which Version of R to Use?.....	37
3.1.1	Renjin.....	37
3.1.2	pqR	37
3.2	Which Interface of R to Use	38
3.3	Using R from the Browser	40
3.3.1	Statace	40
3.3.2	R Fiddle	42
3.3.3	RShiny and RStudio	45
3.4	The Cloud Computing Services Landscape	48
3.4.1	Choosing Infrastructure Providers	49
3.4.1.1	Amazon Cloud	49
3.4.1.2	Other Components of Amazon Cloud	50
3.4.1.3	Google Cloud Services	51
3.4.1.4	Windows Azure	53
3.5	Interview Ian Fellows Deducer	54
3.6	Notable R Projects	57
3.6.1	Installr Package	57
3.6.2	Rserve	58
3.6.3	RApache	58
3.6.4	Rook	58
3.6.5	RJ and Rservi.....	58
3.6.6	R and Java	59
3.7	Creating Your Desktop on the Cloud	59

4	Setting Up R on the Cloud	61
4.1	Connecting to the Cloud Instance Easily	62
4.2	Setting Up R in Amazon's Cloud	64
4.2.1	Local Computer (Windows) Cloud (Linux)	64
4.3	Using Revolution R on the Amazon Cloud	75
4.4	Using the BioConductor Cloud AMI	76
4.5	R in Windows Azure Cloud	80
4.6	R in the Google Cloud	93
4.6.1	Google Big Query	105
4.6.2	Google Fusion Tables API	105
4.6.3	Google Cloud SQL	107
4.6.4	Google Prediction API	107
4.7	R in IBM SmartCloud	107
4.7.1	IBM SmartCloud Enterprise	107
4.7.2	IBM Softlayer	124
4.7.3	Using R with IBM Bluemix	125
5	Using R	127
5.1	A R Tutorial for Beginners	127
5.2	Doing Data Mining with R	131
5.3	Web Scraping Using R	142
5.4	Social Media and Web Analytics Using R	144
5.4.1	Facebook Network Analytics Using R	144
5.4.2	Twitter Analysis with R	149
5.4.3	Google Analytics API with R	150
5.4.4	R for Web Analytics	157
5.5	Doing Data Visualization Using R	157
5.5.1	Using Deducer for Faster and Better Data Visualization in R	158
5.5.2	R for Geographical Based Analysis	170
5.6	Creating a Basic Forecasting Model with R	174
5.7	Easy Updating R	176
5.8	Other R on the Web Projects	177
5.8.1	Concerto	177
5.8.2	Rapporter	179
5.8.3	R Service Bus	186
5.9	The Difference Between Open Source and Closed Source Enterprise Software	188

6	Using R with Data and Bigger Data	193
6.1	Big Data Interview	193
6.2	The Hadoop Paradigm	197
6.2.1	What Is Hadoop All About	197
6.2.2	The Ecosystem of Hadoop	197
6.2.3	Current Developments in Hadoop	200
6.3	Commercial Distributions of Hadoop	201
6.3.1	Hadoop on the Cloud as a Service	202
6.3.2	Learning Hadoop	202
6.4	Learning Hadoop	202
6.5	RHadoop and Other R Packages	203
6.6	Interview with Antonio Piccolboni, Consultant on RHadoop	204
6.7	Big Data R Packages	206
6.8	Databases and CAP Theorem	208
6.8.1	CAP Theorem Visually	209
6.9	ACID and BASE for Databases	209
6.10	Using R with MySQL	210
6.11	Using R with NoSQL	211
6.11.1	MongoDB	211
6.11.2	jsonlite (prettify)	213
6.11.3	CouchDB	214
6.11.4	MonetDB	214
6.12	Big Data Visualization	215
7	R with Cloud APIs	217
7.1	Everything Has an API	217
7.2	REST	218
7.3	rOpenSci	219
7.4	What Is Curl	221
7.5	Navigating OAuth2	221
7.6	Machine Learning as a Service	222
7.6.1	Google Prediction API and R	222
7.6.2	BigML with R	223
7.6.3	Azure Machine Learning with R	223
7.7	Examples of R with APIs	223
7.7.1	Quandl with R	223
7.7.2	Data Visualization APIs with R	224
7.7.2.1	Plotly	224
7.7.2.2	googleVis	227
7.7.2.3	tabplotd3	228
7.7.2.4	Yhat and R	229
7.7.3	OpenCPU and R	229
7.7.3.1	Interview Jeroen Ooms OpenCPU	230
7.8	RForcecom by Takekatsu Hiramura	232

8 Securing Your R cloud	237
8.1 Ensuring R Code Does not Contain Your Login Keys	237
8.2 Setting Up Access Control (User Management Rights)	238
8.2.1 Amazon User Management	238
8.3 Setting Up Security Control Groups for IP Address Level Access (Security Groups)	240
8.3.1 A Note on Passwords and Passphrases	241
8.3.2 A Note on Social Media's Impact on Cyber Security	242
8.4 Monitoring Usage for Improper Access	243
8.5 Basics of Encryption for Data Transfer (PGP- Public Key, Private Key)	243
8.5.1 Encryption Software	243
9 Training Literature for Cloud Computing and R	247
9.1 Challenges and Tips in Transitioning to R	247
9.2 Learning R for Free Online	252
9.3 Learn More Linux	253
9.4 Learn Git	254
9.5 Reference Cards for Data Scientists	254
9.6 Using Public Datasets from Amazon	255
9.7 Interview Vivian Zhang Co-founder SupStat	256
9.8 Interview EODA	258
9.9 Rpubs	260
Appendix	261
A.1 Creating a Graphical Cloud Desktop on a Linux OS	261
References	263

About the Author

A Ohri is the founder of the analytics startup Decisionstats.com. He has pursued graduate courses at the University of Tennessee, Knoxville, and completed his master's from the Indian Institute of Management, Lucknow. Ohri also has a mechanical engineering degree from the Delhi College of Engineering. He has interviewed more than 150 practitioners in analytics, including leading members from all the analytics software vendors. Ohri has written almost 2,000 articles on his blog, in addition to writing about APIs for influential websites like ProgrammableWeb. Ohri's current research interests include spreading open source analytics, analyzing social media manipulation with mechanism design, creating simpler interfaces to cloud computing, and investigating climate change manipulation and unorthodox cryptography including visual and quantum. He is currently advising multiple startups in analytics offshoring, analytics services, and analytics education as well as using social media to enhance the buzz for analytics products. Ohri works with R, SAS, Julia, and Python languages and finds beauty in all of them. He is the author of another book on R called *R for Business Analytics* (Springer 2013). *R for Business Analytics* is being translated in Chinese.