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 Miereswa; 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.

http://www.springer.com/978-1-4939-1701-3

R for Cloud Computing
An Approach for Data Scientists
Ohri, A.
2014, XVII, 267 p. 255 illus., 160 illus. in color.,
Softcover
ISBN: 978-1-4939-1701-3