

Nataraj Dasgupta

Practical Big Data Analytics

Hands-on techniques to implement enterprise analytics
and machine learning using Hadoop, Spark, NoSQL and R



Packt>

Practical Big Data Analytics

Big Data analytics relates to the strategies used by organizations to collect, organize, and analyze large amounts of data to uncover valuable business insights that cannot be analyzed through traditional systems. Crafting an enterprise-scale cost-efficient Big Data and machine learning solution to uncover insights and value from your organization's data is a challenge. Today, with hundreds of new Big Data systems, machine learning packages, and BI tools, selecting the right combination of technologies is an even greater challenge. This book will help you do that.

With the help of this guide, you will be able to bridge the gap between the theoretical world of technology and the practical reality of building corporate Big Data and data science platforms. You will get hands-on exposure to Hadoop and Spark, build machine learning dashboards using R and R Shiny, create web-based apps using NoSQL databases such as MongoDB, and even learn how to write R code for neural networks.

By the end of the book, you will have a very clear and concrete understanding of what Big Data analytics means, how it drives revenues for organizations, and how you can develop your own Big Data analytics solution using the different tools and methods articulated in this book.

Things you will learn:

- Get a 360-degree view of the world of Big Data, data science, and machine learning
- Go through a broad range of technical and business Big Data analytics topics that caters to the interests of technical experts as well as corporate IT executives
- Get hands-on experience with industry-standard Big Data and machine learning tools such as Hadoop, Spark, MongoDB, kdb+, and R
- Create production-grade machine learning BI dashboards using R and R Shiny with step-by-step instructions
- Learn how to combine open-source Big Data, machine learning, and BI tools to create low-cost business analytics applications
- Understand corporate strategies for successful Big Data and data science projects
- Go beyond general-purpose analytics to develop cutting-edge Big Data applications using emerging technologies

Practical Big Data Analytics

Nataraj Dasgupta



Practical Big Data Analytics

Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R

Nataraj Dasgupta

Packt >

BIRMINGHAM - MUMBAI

Practical Big Data Analytics

Copyright © 2018 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

Commissioning Editor: Veena Pagare
Acquisition Editor: Vinay Argekar
Content Development Editor: Tejas Limkar
Technical Editor: Dinesh Chaudhary
Copy Editor: Safis Editing
Project Coordinator: Manthan Patel
Proofreader: Safis Editing
Indexer: Pratik Shirodkar
Graphics: Tania Dutta
Production Coordinator: Aparna Bhagat

First published: January 2018

Production reference: 1120118

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham
B3 2PB, UK.

ISBN 978-1-78355-439-3

www.packtpub.com



mapt.io

Mapt is an online digital library that gives you full access to over 5,000 books and videos, as well as industry leading tools to help you plan your personal development and advance your career. For more information, please visit our website.

Why subscribe?

- Spend less time learning and more time coding with practical eBooks and Videos from over 4,000 industry professionals
- Improve your learning with Skill Plans built especially for you
- Get a free eBook or video every month
- Mapt is fully searchable
- Copy and paste, print, and bookmark content

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.

Contributors

About the author

Nataraj Dasgupta is the vice president of Advanced Analytics at RxDataScience Inc. Nataraj has been in the IT industry for more than 19 years and has worked in the technical and analytics divisions of Philip Morris, IBM, UBS Investment Bank and Purdue Pharma. He led the data science division at Purdue Pharma L.P. where he developed the company's award-winning big data and machine learning platform. Prior to Purdue, at UBS, he held the role of associate director working with high frequency and algorithmic trading technologies in the Foreign Exchange trading division of the bank.

I'd like to thank my wife, Suraiya, for her caring, support, and understanding as I worked during long weekends and evening hours and to my parents, in-laws, sister and grandmother for all the support, guidance, tutelage and encouragement over the years.

I'd also like to thank Packt, especially the editors, Tejas, Dinesh, Vinay, and the team whose persistence and attention to detail has been exemplary.

About the reviewer

Giancarlo Zaccone has more than 10 years experience in managing research projects both in scientific and industrial areas. He worked as a researcher at the C.N.R, the National Research Council, where he was involved in projects on parallel numerical computing and scientific visualization.

He is a senior software engineer at a consulting company, developing and testing software systems for space and defense applications.

He holds a master's degree in physics from the Federico II of Naples and a second level postgraduate master course in scientific computing from La Sapienza of Rome.

Packt is searching for authors like you

If you're interested in becoming an author for Packt, please visit authors.packtpub.com and apply today. We have worked with thousands of developers and tech professionals, just like you, to help them share their insight with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Table of Contents

Preface	1
<hr/>	
Chapter 1: Too Big or Not Too Big	9
<hr/>	
What is big data?	10
A brief history of data	10
Dawn of the information age	10
Dr. Alan Turing and modern computing	11
The advent of the stored-program computer	11
From magnetic devices to SSDs	11
Why we are talking about big data now if data has always existed	12
Definition of big data	14
Building blocks of big data analytics	14
Types of Big Data	15
Structured	16
Unstructured	16
Semi-structured	17
Sources of big data	18
The 4Vs of big data	19
When do you know you have a big data problem and where do you start your search for the big data solution?	19
Summary	21
<hr/>	
Chapter 2: Big Data Mining for the Masses	23
<hr/>	
What is big data mining?	24
Big data mining in the enterprise	24
Building the case for a Big Data strategy	24
Implementation life cycle	26
Stakeholders of the solution	27
Implementing the solution	27
Technical elements of the big data platform	28
Selection of the hardware stack	29
Selection of the software stack	31
Summary	34
<hr/>	
Chapter 3: The Analytics Toolkit	35
<hr/>	
Components of the Analytics Toolkit	36
System recommendations	36

Installing on a laptop or workstation	37
Installing on the cloud	37
Installing Hadoop	37
Installing Oracle VirtualBox	38
Installing CDH in other environments	46
Installing Packt Data Science Box	47
Installing Spark	51
Installing R	51
Steps for downloading and installing Microsoft R Open	52
Installing RStudio	56
Installing Python	57
Summary	60
Chapter 4: Big Data With Hadoop	61
<hr/>	
The fundamentals of Hadoop	62
The fundamental premise of Hadoop	63
The core modules of Hadoop	64
Hadoop Distributed File System - HDFS	64
Data storage process in HDFS	67
Hadoop MapReduce	69
An intuitive introduction to MapReduce	70
A technical understanding of MapReduce	70
Block size and number of mappers and reducers	72
Hadoop YARN	73
Job scheduling in YARN	77
Other topics in Hadoop	77
Encryption	78
User authentication	78
Hadoop data storage formats	78
New features expected in Hadoop 3	80
The Hadoop ecosystem	80
Hands-on with CDH	82
WordCount using Hadoop MapReduce	82
Analyzing oil import prices with Hive	92
Joining tables in Hive	100
Summary	105
Chapter 5: Big Data Mining with NoSQL	107
<hr/>	
Why NoSQL?	108
The ACID, BASE, and CAP properties	108
ACID and SQL	108
The BASE property of NoSQL	109
The CAP theorem	110

The need for NoSQL technologies	111
Google Bigtable	112
Amazon Dynamo	112
NoSQL databases	113
In-memory databases	113
Columnar databases	116
Document-oriented databases	121
Key-value databases	125
Graph databases	128
Other NoSQL types and summary of other types of databases	130
Analyzing Nobel Laureates data with MongoDB	131
JSON format	131
Installing and using MongoDB	132
Tracking physician payments with real-world data	148
Installing kdb+, R, and RStudio	149
Installing kdb+	150
Installing R	155
Installing RStudio	155
The CMS Open Payments Portal	158
Downloading the CMS Open Payments data	159
Creating the Q application	166
Loading the data	166
The backend code	168
Creating the frontend web portal	170
R Shiny platform for developers	171
Putting it all together - The CMS Open Payments application	184
Applications	187
Summary	189
Chapter 6: Spark for Big Data Analytics	191
<hr/>	
The advent of Spark	192
Limitations of Hadoop	192
Overcoming the limitations of Hadoop	193
Theoretical concepts in Spark	194
Resilient distributed datasets	195
Directed acyclic graphs	195
SparkContext	195
Spark DataFrames	196
Actions and transformations	196
Spark deployment options	197
Spark APIs	197
Core components in Spark	198

Spark Core	198
Spark SQL	198
Spark Streaming	198
GraphX	199
MLlib	199
The architecture of Spark	200
Spark solutions	201
Spark practicals	201
Signing up for Databricks Community Edition	202
Spark exercise - hands-on with Spark (Databricks)	211
Summary	216
Chapter 7: An Introduction to Machine Learning Concepts	217
What is machine learning?	218
The evolution of machine learning	219
Factors that led to the success of machine learning	220
Machine learning, statistics, and AI	221
Categories of machine learning	223
Supervised and unsupervised machine learning	224
Supervised machine learning	224
Vehicle Mileage, Number Recognition and other examples	225
Unsupervised machine learning	227
Subdividing supervised machine learning	229
Common terminologies in machine learning	231
The core concepts in machine learning	233
Data management steps in machine learning	233
Pre-processing and feature selection techniques	233
Centering and scaling	234
The near-zero variance function	235
Removing correlated variables	236
Other common data transformations	238
Data sampling	238
Data imputation	242
The importance of variables	246
The train, test splits, and cross-validation concepts	249
Splitting the data into train and test sets	249
The cross-validation parameter	250
Creating the model	254
Leveraging multicore processing in the model	257
Summary	260
Chapter 8: Machine Learning Deep Dive	261
The bias, variance, and regularization properties	262

The gradient descent and VC Dimension theories	270
Popular machine learning algorithms	270
Regression models	271
Association rules	273
Confidence	274
Support	275
Lift	275
Decision trees	276
The Random forest extension	281
Boosting algorithms	283
Support vector machines	286
The K-Means machine learning technique	289
The neural networks related algorithms	292
Tutorial - associative rules mining with CMS data	296
Downloading the data	297
Writing the R code for Apriori	298
Shiny (R Code)	299
Using custom CSS and fonts for the application	303
Running the application	304
Summary	306
Chapter 9: Enterprise Data Science	307
<hr/>	
Enterprise data science overview	308
A roadmap to enterprise analytics success	313
Data science solutions in the enterprise	315
Enterprise data warehouse and data mining	316
Traditional data warehouse systems	316
Oracle Exadata, Exalytics, and TimesTen	316
HP Vertica	317
Teradata	317
IBM data warehouse systems (formerly Netezza appliances)	318
PostgreSQL	319
Greenplum	319
SAP Hana	320
Enterprise and open source NoSQL Databases	320
Kdb+	320
MongoDB	321
Cassandra	322
Neo4j	323
Cloud databases	323
Amazon Redshift, Redshift Spectrum, and Athena databases	323
Google BigQuery and other cloud services	325

Azure CosmosDB	326
GPU databases	327
Brytlyt	328
MapD	328
Other common databases	328
Enterprise data science – machine learning and AI	329
The R programming language	329
Python	330
OpenCV, Caffe, and others	331
Spark	331
Deep learning	332
H2O and Driverless AI	333
Datarobot	334
Command-line tools	334
Apache MADlib	334
Machine learning as a service	335
Enterprise infrastructure solutions	336
Cloud computing	337
Virtualization	337
Containers – Docker, Kubernetes, and Mesos	339
On-premises hardware	340
Enterprise Big Data	341
Tutorial – using RStudio in the cloud	342
Summary	367
Chapter 10: Closing Thoughts on Big Data	369
Corporate big data and data science strategy	370
Ethical considerations	373
Silicon Valley and data science	374
The human factor	375
Characteristics of successful projects	376
Summary	377
Appendix: External Data Science Resources	379
Big data resources	379
NoSQL products	380
Languages and tools	380
Creating dashboards	380
Notebooks	381
Visualization libraries	381

Courses on R	381
Courses on machine learning	382
Machine learning and deep learning links	382
Web-based machine learning services	383
Movies	383
Machine learning books from Packt	384
Books for leisure reading	384
Other Books You May Enjoy	385
Leave a review - let other readers know what you think	387
Index	389
