EMC²

Data Science and Big Data Analytics

Discovering, Analyzing, Visualizing and Presenting Data

EMC Education Services

Data Science & Big Data Analytics

Data Science & Big Data Analytics

Discovering, Analyzing, Visualizing and Presenting Data

EMC Education Services

WILEY

Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Published simultaneously in Canada

ISBN: 978-1-118-87613-8
ISBN: 978-1-118-87622-0 (ebk)
ISBN: 978-1-118-87605-3 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2014946681

Credits

Executive Editor
Carol Long

Project Editor
Kelly Talbot

Production Manager
Kathleen Wisor

Copy Editor
Karen Gill

Manager of Content Development and Assembly
Mary Beth Wakefield

Marketing Director
David Mayhew

Marketing Manager
Carrie Sherrill

Professional Technology and Strategy Director
Barry Pruett

Business Manager
Amy Knies

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Patrick Redmond

Proofreader
Nancy Carrasco

Indexer
Johnna VanHoose Dinse

Cover Designer
Mallesh Gurram

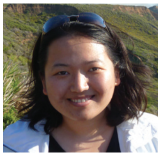
About the Key Contributors

David Dietrich heads the data science education team within EMC Education Services, where he leads the curriculum, strategy and course development related to Big Data Analytics and Data Science. He co-authored the first course in EMC's Data Science curriculum, two additional EMC courses focused on teaching leaders and executives about Big Data and data science, and is a contributing author and editor of this book. He has filed 14 patents in the areas of data science, data privacy, and cloud computing.

David has been an advisor to several universities looking to develop academic programs related to data analytics, and has been a frequent speaker at conferences and industry events. He also has been a a guest lecturer at universities in the Boston area. His work has been featured in major publications including Forbes, Harvard Business Review, and the 2014 Massachusetts Big Data Report, commissioned by Governor Deval Patrick.

Involved with analytics and technology for nearly 20 years, David has worked with many Fortune 500 companies over his career, holding multiple roles involving analytics, including managing analytics and operations teams, delivering analytic consulting engagements, managing a line of analytical software products for regulating the US banking industry, and developing Software-as-a-Service and BI-as-a-Service offerings. Additionally, David collaborated with the U.S. Federal Reserve in developing predictive models for monitoring mortgage portfolios.

Barry Heller is an advisory technical education consultant at EMC Education Services. Barry is a course developer and curriculum advisor in the emerging technology areas of Big Data and data science. Prior to his current role, Barry was a consultant research scientist leading numerous analytical initiatives within EMC's Total Customer Experience organization. Early in his EMC career, he managed the statistical engineering group as well as led the data warehousing efforts in an Enterprise Resource Planning (ERP) implementation. Prior to joining EMC, Barry held managerial and analytical roles in reliability engineering functions at medical diagnostic and technology companies. During his career, he has applied his quantitative skill set to a myriad of business applications in the Customer Service, Engineering, Manufacturing, Sales/Marketing, Finance, and Legal arenas. Underscoring the importance of strong executive stakeholder engagement, many of his successes have resulted from not only focusing on the technical details of an analysis, but on the decisions that will be resulting from the analysis. Barry earned a B.S. in Computational Mathematics from the Rochester Institute of Technology and an M.A. in Mathematics from the State University of New York (SUNY) New Paltz.



Beibei Yang is a Technical Education Consultant of EMC Education Services, responsible for developing several open courses at EMC related to Data Science and Big Data Analytics. Beibei has seven years of experience in the IT industry. Prior to EMC she worked as a software engineer, systems manager, and network manager for a Fortune 500 company where she introduced new technologies to improve efficiency and encourage collaboration. Beibei has published papers to prestigious conferences and has filed multiple patents. She received her Ph.D. in computer science from the University of Massachusetts Lowell. She has a passion toward natural language processing and data mining, especially using various tools and techniques to find hidden patterns and tell stories with data.

Data Science and Big Data Analytics is an exciting domain where the potential of digital information is maximized for making intelligent business decisions. We believe that this is an area that will attract a lot of talented students and professionals in the short, mid, and long term.

Acknowledgments

EMC Education Services embarked on learning this subject with the intent to develop an "open" curriculum and certification. It was a challenging journey at the time as not many understood what it would take to be a true data scientist. After initial research (and struggle), we were able to define what was needed and attract very talented professionals to work on the project. The course, "Data Science and Big Data Analytics," has become well accepted across academia and the industry.

Led by EMC Education Services, this book is the result of efforts and contributions from a number of key EMC organizations and supported by the office of the CTO, IT, Global Services, and Engineering. Many sincere thanks to many key contributors and subject matter experts **David Dietrich, Barry Heller,** and **Beibei Yang** for their work developing content and graphics for the chapters. A special thanks to subject matter experts **John Cardente** and **Ganesh Rajaratnam** for their active involvement reviewing multiple book chapters and providing valuable feedback throughout the project.

We are also grateful to the following experts from EMC and Pivotal for their support in reviewing and improving the content in this book:

- Aidan O'Brien
- Alexander Nunes
- Bryan Miletich
- Dan Baskette
- Daniel Mepham
- Dave Reiner
- Deborah Stokes
- Ellis Kriesberg
- Frank Coleman
- Hisham Arafat
- Ira Schild
- Jack Harwood
- Jim McGroddy
- Joe Kambourakis
- Joe Milardo
- John Sopka
- Kathryn Stiles
- Ken Taylor
- Lanette Wells
- Michael Hancock
- Michael Vander Donk
- Narayanan Krishnakumar
- Richard Moore
- Ron Glick
- Stephen Maloney
- Steve Todd

Jody Goncalves	Suresh Thankappan
Joe Dery	Tom McGowan

We also thank Ira Schild and Shane Goodrich for coordinating this project, Mallesh Gurram for the cover design, Chris Conroy and Rob Bradley for graphics, and the publisher, John Wiley and Sons, for timely support in bringing this book to the industry.

Nancy Gessler
Director, Education Services, EMC Corporation
Alok Shrivastava
Sr. Director, Education Services, EMC Corporation

Contents

Introduction . . . . . . *xvii*

Chapter 1 • Introduction to Big Data Analytics **1**

1.1 Big Data Overview 2

1.1.1 Data Structures *5*

1.1.2 Analyst Perspective on Data Repositories *9*

1.2 State of the Practice in Analytics 11

1.2.1 BI Versus Data Science *12*

1.2.2 Current Analytical Architecture *13*

1.2.3 Drivers of Big Data *15*

1.2.4 Emerging Big Data Ecosystem and a New Approach to Analytics *16*

1.3 Key Roles for the New Big Data Ecosystem 19

1.4 Examples of Big Data Analytics 22

Summary 23

Exercises 23

Bibliography 24

Chapter 2 • Data Analytics Lifecycle **25**

2.1 Data Analytics Lifecycle Overview 26

2.1.1 Key Roles for a Successful Analytics Project *26*

2.1.2 Background and Overview of Data Analytics Lifecycle *28*

2.2 Phase 1: Discovery 30

2.2.1 Learning the Business Domain *30*

2.2.2 Resources *31*

2.2.3 Framing the Problem *32*

2.2.4 Identifying Key Stakeholders *33*

2.2.5 Interviewing the Analytics Sponsor *33*

2.2.6 Developing Initial Hypotheses *35*

2.2.7 Identifying Potential Data Sources *35*

2.3 Phase 2: Data Preparation 36

2.3.1 Preparing the Analytic Sandbox *37*

2.3.2 Performing ETLT *38*

2.3.3 Learning About the Data *39*

2.3.4 Data Conditioning *40*

2.3.5 Survey and Visualize *41*

2.3.6 Common Tools for the Data Preparation Phase *42*

2.4 Phase 3: Model Planning 42

2.4.1 Data Exploration and Variable Selection *44*

2.4.2 Model Selection *45*

2.4.3 Common Tools for the Model Planning Phase *45*

2.5 Phase 4: Model Building . . . 46
2.5.1 Common Tools for the Model Building Phase . . . 48
2.6 Phase 5: Communicate Results . . . 49
2.7 Phase 6: Operationalize . . . 50
2.8 Case Study: Global Innovation Network and Analysis (GINA) . . . 53
2.8.1 Phase 1: Discovery . . . 54
2.8.2 Phase 2: Data Preparation . . . 55
2.8.3 Phase 3: Model Planning . . . 56
2.8.4 Phase 4: Model Building . . . 56
2.8.5 Phase 5: Communicate Results . . . 58
2.8.6 Phase 6: Operationalize . . . 59
Summary . . . 60
Exercises . . . 61
Bibliography . . . 61

Chapter 3 • Review of Basic Data Analytic Methods Using R . . . 63
3.1 Introduction to R . . . 64
3.1.1 R Graphical User Interfaces . . . 67
3.1.2 Data Import and Export . . . 69
3.1.3 Attribute and Data Types . . . 71
3.1.4 Descriptive Statistics . . . 79
3.2 Exploratory Data Analysis . . . 80
3.2.1 Visualization Before Analysis . . . 82
3.2.2 Dirty Data . . . 85
3.2.3 Visualizing a Single Variable . . . 88
3.2.4 Examining Multiple Variables . . . 91
3.2.5 Data Exploration Versus Presentation . . . 99
3.3 Statistical Methods for Evaluation . . . 101
3.3.1 Hypothesis Testing . . . 102
3.3.2 Difference of Means . . . 104
3.3.3 Wilcoxon Rank-Sum Test . . . 108
3.3.4 Type I and Type II Errors . . . 109
3.3.5 Power and Sample Size . . . 110
3.3.6 ANOVA . . . 110
Summary . . . 114
Exercises . . . 114
Bibliography . . . 115

Chapter 4 • Advanced Analytical Theory and Methods: Clustering . . . 117
4.1 Overview of Clustering . . . 118
4.2 K-means . . . 118
4.2.1 Use Cases . . . 119
4.2.2 Overview of the Method . . . 120
4.2.3 Determining the Number of Clusters . . . 123
4.2.4 Diagnostics . . . 128

4.2.5 Reasons to Choose and Cautions *130*
4.3 Additional Algorithms 134
Summary 135
Exercises 135
Bibliography 136

Chapter 5 • Advanced Analytical Theory and Methods: Association Rules 137
5.1 Overview 138
5.2 Apriori Algorithm 140
5.3 Evaluation of Candidate Rules 141
5.4 Applications of Association Rules 143
5.5 An Example: Transactions in a Grocery Store. 143
5.5.1 The Groceries Dataset *144*
5.5.2 Frequent Itemset Generation *146*
5.5.3 Rule Generation and Visualization *152*
5.6 Validation and Testing 157
5.7 Diagnostics. 158
Summary 158
Exercises 159
Bibliography 160

Chapter 6 • Advanced Analytical Theory and Methods: Regression 161
6.1 Linear Regression 162
6.1.1 Use Cases *162*
6.1.2 Model Description *163*
6.1.3 Diagnostics. *173*
6.2 Logistic Regression 178
6.2.1 Use Cases. *179*
6.2.2 Model Description *179*
6.2.3 Diagnostics *181*
6.3 Reasons to Choose and Cautions 188
6.4 Additional Regression Models 189
Summary 190
Exercises 190

Chapter 7 • Advanced Analytical Theory and Methods: Classification 191
7.1 Decision Trees 192
7.1.1 Overview of a Decision Tree. *193*
7.1.2 The General Algorithm *197*
7.1.3 Decision Tree Algorithms *203*
7.1.4 Evaluating a Decision Tree. *204*
7.1.5 Decision Trees in R *206*
7.2 Naïve Bayes 211
7.2.1 Bayes' Theorem *212*
7.2.2 Naïve Bayes Classifier *214*

7.2.3 Smoothing 217
7.2.4 Diagnostics 217
7.2.5 Naïve Bayes in R 218
7.3 Diagnostics of Classifiers 224
7.4 Additional Classification Methods 228
Summary 229
Exercises 230
Bibliography 231

Chapter 8 • Advanced Analytical Theory and Methods: Time Series Analysis 233
8.1 Overview of Time Series Analysis 234
8.1.1 Box-Jenkins Methodology 235
8.2 ARIMA Model 236
8.2.1 Autocorrelation Function (ACF) 236
8.2.2 Autoregressive Models 238
8.2.3 Moving Average Models 239
8.2.4 ARMA and ARIMA Models 241
8.2.5 Building and Evaluating an ARIMA Model 244
8.2.6 Reasons to Choose and Cautions 252
8.3 Additional Methods 253
Summary 254
Exercises 254

Chapter 9 • Advanced Analytical Theory and Methods: Text Analysis 255
9.1 Text Analysis Steps 257
9.2 A Text Analysis Example 259
9.3 Collecting Raw Text 260
9.4 Representing Text 264
9.5 Term Frequency—Inverse Document Frequency (TFIDF) 269
9.6 Categorizing Documents by Topics 274
9.7 Determining Sentiments 277
9.8 Gaining Insights 283
Summary 290
Exercises 290
Bibliography 291

Chapter 10 • Advanced Analytics—Technology and Tools: MapReduce and Hadoop 295
10.1 Analytics for Unstructured Data 296
10.1.1 Use Cases 296
10.1.2 MapReduce 298
10.1.3 Apache Hadoop 300
10.2 The Hadoop Ecosystem 306
10.2.1 Pig 306
10.2.2 Hive 308
10.2.3 HBase 311
10.2.4 Mahout 319

10.3 NoSQL 322
Summary 323
Exercises 324
Bibliography 324

Chapter 11 • Advanced Analytics—Technology and Tools: In-Database Analytics 327
11.1 SQL Essentials 328
11.1.1 Joins 330
11.1.2 Set Operations 332
11.1.3 Grouping Extensions 334
11.2 In-Database Text Analysis 338
11.3 Advanced SQL 343
11.3.1 Window Functions 343
11.3.2 User-Defined Functions and Aggregates 347
11.3.3 Ordered Aggregates 351
11.3.4 MADlib 352
Summary 356
Exercises 356
Bibliography 357

Chapter 12 • The Endgame, or Putting It All Together 359
12.1 Communicating and Operationalizing an Analytics Project 360
12.2 Creating the Final Deliverables 362
12.2.1 Developing Core Material for Multiple Audiences 364
12.2.2 Project Goals 365
12.2.3 Main Findings 367
12.2.4 Approach 369
12.2.5 Model Description 371
12.2.6 Key Points Supported with Data 372
12.2.7 Model Details 372
12.2.8 Recommendations 374
12.2.9 Additional Tips on Final Presentation 375
12.2.10 Providing Technical Specifications and Code 376
12.3 Data Visualization Basics 377
12.3.1 Key Points Supported with Data 378
12.3.2 Evolution of a Graph 380
12.3.3 Common Representation Methods 386
12.3.4 How to Clean Up a Graphic 387
12.3.5 Additional Considerations 392
Summary 393
Exercises 394
References and Further Reading 394
Bibliography 394
Index 397

Foreword

Technological advances and the associated changes in practical daily life have produced a rapidly expanding "parallel universe" of new content, new data, and new information sources all around us. Regardless of how one defines it, the phenomenon of Big Data is ever more present, ever more pervasive, and ever more important. There is enormous value potential in Big Data: innovative insights, improved understanding of problems, and countless opportunities to predict—and even to shape—the future. Data Science is the principal means to discover and tap that potential. Data Science provides ways to deal with and benefit from Big Data: to see patterns, to discover relationships, and to make sense of stunningly varied images and information.

Not everyone has studied statistical analysis at a deep level. People with advanced degrees in applied mathematics are not a commodity. Relatively few organizations have committed resources to large collections of data gathered primarily for the purpose of exploratory analysis. And yet, while applying the practices of Data Science to Big Data is a valuable differentiating strategy at present, it will be a standard core competency in the not so distant future.

How does an organization operationalize quickly to take advantage of this trend? We've created this book for that exact purpose.

EMC Education Services has been listening to the industry and organizations, observing the multi-faceted transformation of the technology landscape, and doing direct research in order to create curriculum and content to help individuals and organizations transform themselves. For the domain of Data Science and Big Data Analytics, our educational strategy balances three things: *people*—especially in the context of data science teams, *processes*—such as the analytic lifecycle approach presented in this book, and *tools and technologies*—in this case with the emphasis on proven analytic tools.

So let us help you capitalize on this new "parallel universe" that surrounds us. We invite you to learn about Data Science and Big Data Analytics through this book and hope it significantly accelerates your efforts in the transformational process.

Thomas P. Clancy
Vice President, Education Services, EMC Corporation
January 2015

Introduction

Big Data is creating significant new opportunities for organizations to derive new value and create competitive advantage from their most valuable asset: information. For businesses, Big Data helps drive efficiency, quality, and personalized products and services, producing improved levels of customer satisfaction and profit. For scientific efforts, Big Data analytics enable new avenues of investigation with potentially richer results and deeper insights than previously available. In many cases, Big Data analytics integrate structured and unstructured data with real-time feeds and queries, opening new paths to innovation and insight.

This book provides a practitioner's approach to some of the key techniques and tools used in Big Data analytics. Knowledge of these methods will help people become active contributors to Big Data analytics projects. The book's content is designed to assist multiple stakeholders: business and data analysts looking to add Big Data analytics skills to their portfolio; database professionals and managers of business intelligence, analytics, or Big Data groups looking to enrich their analytic skills; and college graduates investigating data science as a career field.

The content is structured in twelve chapters. The first chapter introduces the reader to the domain of Big Data, the drivers for advanced analytics, and the role of the data scientist. The second chapter presents an analytic project lifecycle designed for the particular characteristics and challenges of hypothesis-driven analysis with Big Data.

Chapter 3 examines fundamental statistical techniques in the context of the open source R analytic software environment. This chapter also highlights the importance of exploratory data analysis via visualizations and reviews the key notions of hypothesis development and testing.

Chapters 4 through 9 discuss a range of advanced analytical methods, including clustering, classification, regression analysis, time series and text analysis.

Chapters 10 and 11 focus on specific technologies and tools that support advanced analytics with Big Data. In particular, the MapReduce paradigm and its instantiation in the Hadoop ecosystem, as well as advanced topics in SQL and in-database text analytics form the focus of these chapters.

Chapter 12 provides guidance on operationalizing Big Data analytics projects. This chapter focuses on creating the final deliverables, converting an analytics project to an ongoing asset of an organization's operation, and creating clear, useful visual outputs based on the data.

EMC Academic Alliance

University and college faculties are invited to join the Academic Alliance program to access unique "open" curriculum-based education on the following topics:

- Data Science and Big Data Analytics
- Information Storage and Management
- Cloud Infrastructure and Services
- Backup Recovery Systems and Architecture

The program provides faculty with course resources to prepare students for opportunities that exist in today's evolving IT industry at no cost. For more information, visit `http://education.EMC.com/academicalliance`.

EMC Proven Professional Certification

EMC Proven Professional is a leading education and certification program in the IT industry, providing comprehensive coverage of information storage technologies, virtualization, cloud computing, data science/Big Data analytics, and more.

Being proven means investing in yourself and formally validating your expertise.

This book prepares you for Data Science Associate (EMCDSA) certification. Visit `http://education.EMC.com` for details.

1

Introduction to Big Data Analytics

Key Concepts

Big Data overview
State of the practice in analytics
Business Intelligence versus Data Science
Key roles for the new Big Data ecosystem
The Data Scientist
Examples of Big Data analytics

Much has been written about Big Data and the need for advanced analytics within industry, academia, and government. Availability of new data sources and the rise of more complex analytical opportunities have created a need to rethink existing data architectures to enable analytics that take advantage of Big Data. In addition, significant debate exists about what Big Data is and what kinds of skills are required to make best use of it. This chapter explains several key concepts to clarify what is meant by Big Data, why advanced analytics are needed, how Data Science differs from Business Intelligence (BI), and what new roles are needed for the new Big Data ecosystem.

1.1 Big Data Overview

Data is created constantly, and at an ever-increasing rate. Mobile phones, social media, imaging technologies to determine a medical diagnosis—all these and more create new data, and that must be stored somewhere for some purpose. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time. Merely keeping up with this huge influx of data is difficult, but substantially more challenging is analyzing vast amounts of it, especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information. These challenges of the data deluge present the opportunity to transform business, government, science, and everyday life.

Several industries have led the way in developing their ability to gather and exploit data:

- Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.
- Mobile phone companies analyze subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.
- For companies such as LinkedIn and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows.

Three attributes stand out as defining Big Data characteristics:

- **Huge volume of data:** Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
- **Complexity of data types and structures:** Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
- **Speed of new data creation and growth:** Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Although the volume of Big Data tends to attract the most attention, generally the variety and velocity of the data provide a more apt definition of Big Data. (Big Data is sometimes described as having 3 Vs: volume, variety, and velocity.) Due to its size or structure, Big Data cannot be efficiently analyzed using only traditional databases or methods. Big Data problems require new tools and technologies to store, manage, and realize the business benefit. These new tools and technologies enable creation, manipulation, and management of large datasets and the storage environments that house them. Another definition of Big Data comes from the McKinsey Global report from 2011:

> ***Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.***
>
> McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity [1]

McKinsey's definition of Big Data implies that organizations will need new data architectures and analytic sandboxes, new tools, new analytical methods, and an integration of multiple skills into the new role of the data scientist, which will be discussed in Section 1.3. Figure 1-1 highlights several sources of the Big Data deluge.

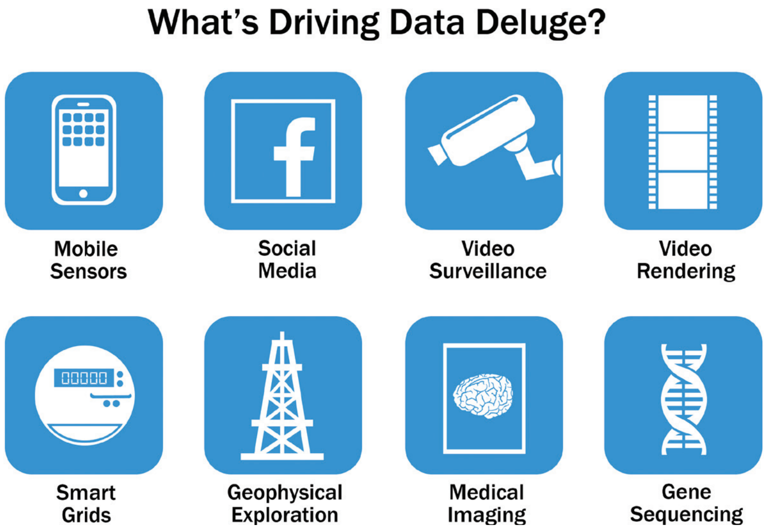


FIGURE 1-1 *What's driving the data deluge*

The rate of data creation is accelerating, driven by many of the items in Figure 1-1.

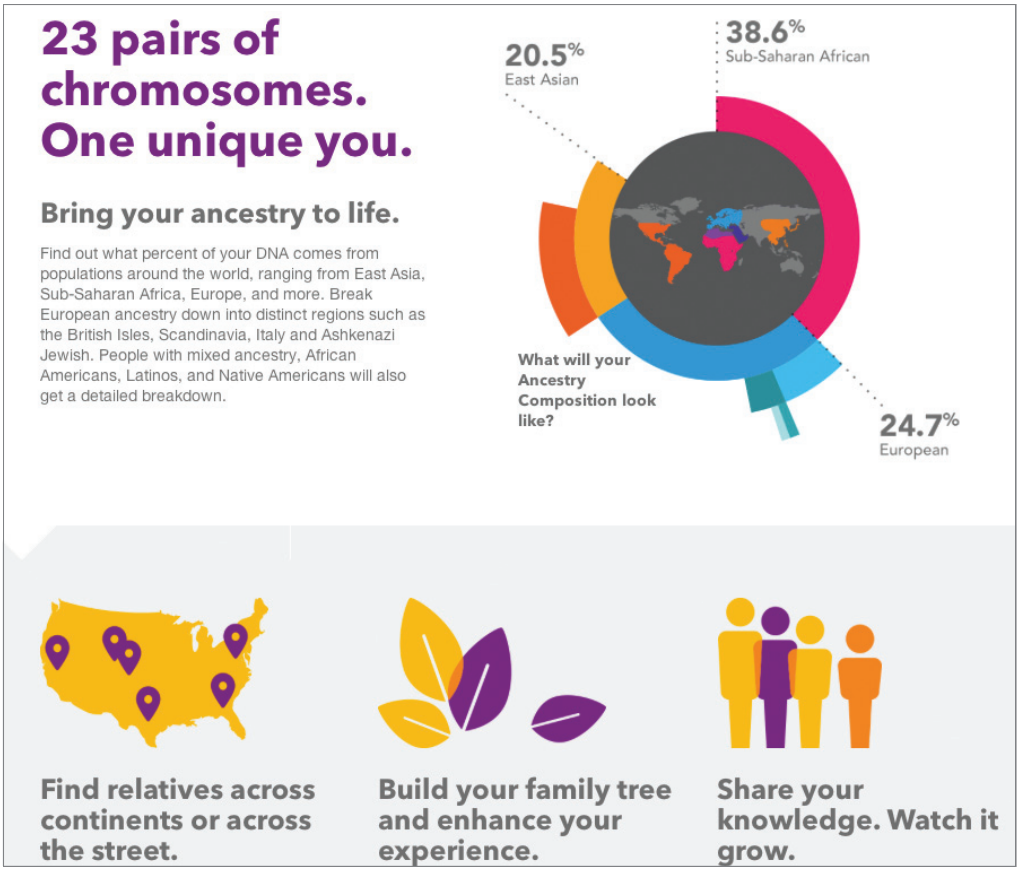
Social media and genetic sequencing are among the fastest-growing sources of Big Data and examples of untraditional sources of data being used for analysis.

For example, in 2012 Facebook users posted 700 status updates per second worldwide, which can be leveraged to deduce latent interests or political views of users and show relevant ads. For instance, an update in which a woman changes her relationship status from "single" to "engaged" would trigger ads on bridal dresses, wedding planning, or name-changing services.

Facebook can also construct social graphs to analyze which users are connected to each other as an interconnected network. In March 2013, Facebook released a new feature called "Graph Search," enabling users and developers to search social graphs for people with similar interests, hobbies, and shared locations.

Another example comes from genomics. Genetic sequencing and human genome mapping provide a detailed understanding of genetic makeup and lineage. The health care industry is looking toward these advances to help predict which illnesses a person is likely to get in his lifetime and take steps to avoid these maladies or reduce their impact through the use of personalized medicine and treatment. Such tests also highlight typical responses to different medications and pharmaceutical drugs, heightening risk awareness of specific drug treatments.

While data has grown, the cost to perform this work has fallen dramatically. The cost to sequence one human genome has fallen from $100 million in 2001 to $10,000 in 2011, and the cost continues to drop. Now, websites such as 23andme (Figure 1-2) offer genotyping for less than $100. Although genotyping analyzes only a fraction of a genome and does not provide as much granularity as genetic sequencing, it does point to the fact that data and complex analysis is becoming more prevalent and less expensive to deploy.



23 pairs of chromosomes. One unique you.

Bring your ancestry to life.

Find out what percent of your DNA comes from populations around the world, ranging from East Asia, Sub-Saharan Africa, Europe, and more. Break European ancestry down into distinct regions such as the British Isles, Scandinavia, Italy and Ashkenazi Jewish. People with mixed ancestry, African Americans, Latinos, and Native Americans will also get a detailed breakdown.

Find relatives across continents or across the street.

Build your family tree and enhance your experience.

Share your knowledge. Watch it grow.

FIGURE 1-2 *Examples of what can be learned through genotyping, from 23andme.com*

As illustrated by the examples of social media and genetic sequencing, individuals and organizations both derive benefits from analysis of ever-larger and more complex datasets that require increasingly powerful analytical capabilities.

1.1.1 Data Structures

Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings. Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze. [2] Distributed computing environments and massively parallel processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data.

With this in mind, this section takes a closer look at data structures.

Figure 1-3 shows four types of data structures, with 80–90% of future data growth coming from non-structured data types. [2] Though different, the four are commonly mixed. For example, a classic Relational Database Management System (RDBMS) may store call logs for a software support call center. The RDBMS may store characteristics of the support calls as typical structured data, with attributes such as time stamps, machine type, problem type, and operating system. In addition, the system will likely have unstructured, quasi- or semi-structured data, such as free-form call log information taken from an e-mail ticket of the problem, customer chat history, or transcript of a phone call describing the technical problem and the solution or audio file of the phone call conversation. Many insights could be extracted from the unstructured, quasi- or semi-structured data in the call center data.

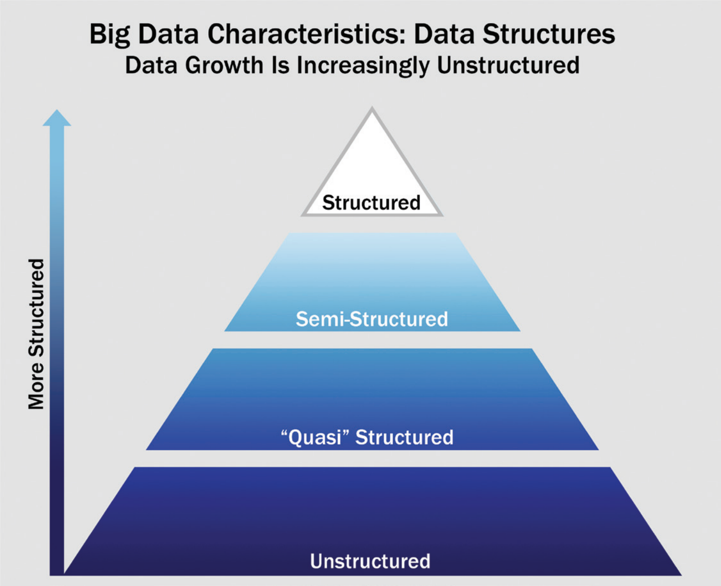
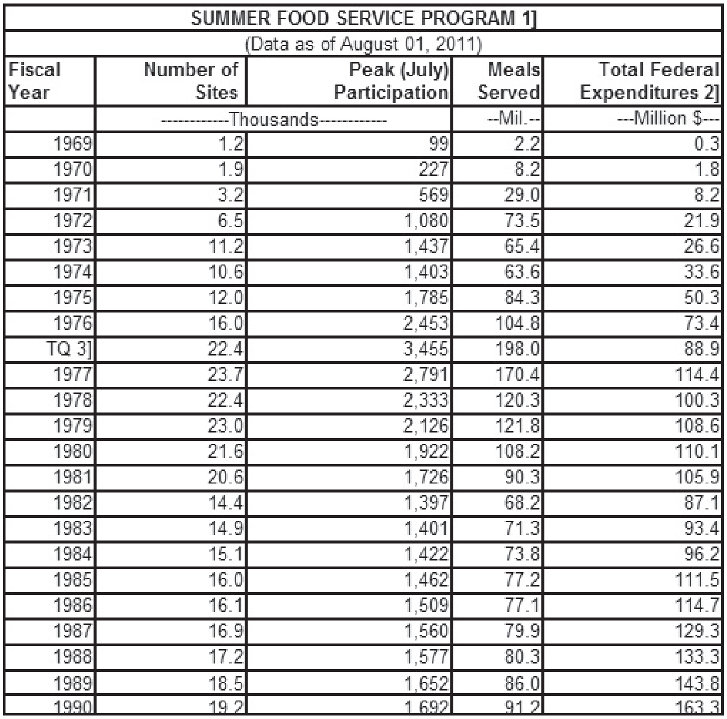

Big Data Characteristics: Data Structures
Data Growth Is Increasingly Unstructured
More Structured
Structured
Semi-Structured
"Quasi" Structured
Unstructured

Figure 1-3 *Big Data Growth is increasingly unstructured*

Although analyzing structured data tends to be the most familiar technique, a different technique is required to meet the challenges to analyze semi-structured data (shown as XML), quasi-structured (shown as a clickstream), and unstructured data.

Here are examples of how each of the four main types of data structures may look.

- **Structured data:** Data containing a defined data type, format, and structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets). See Figure 1-4.



SUMMER FOOD SERVICE PROGRAM 1]

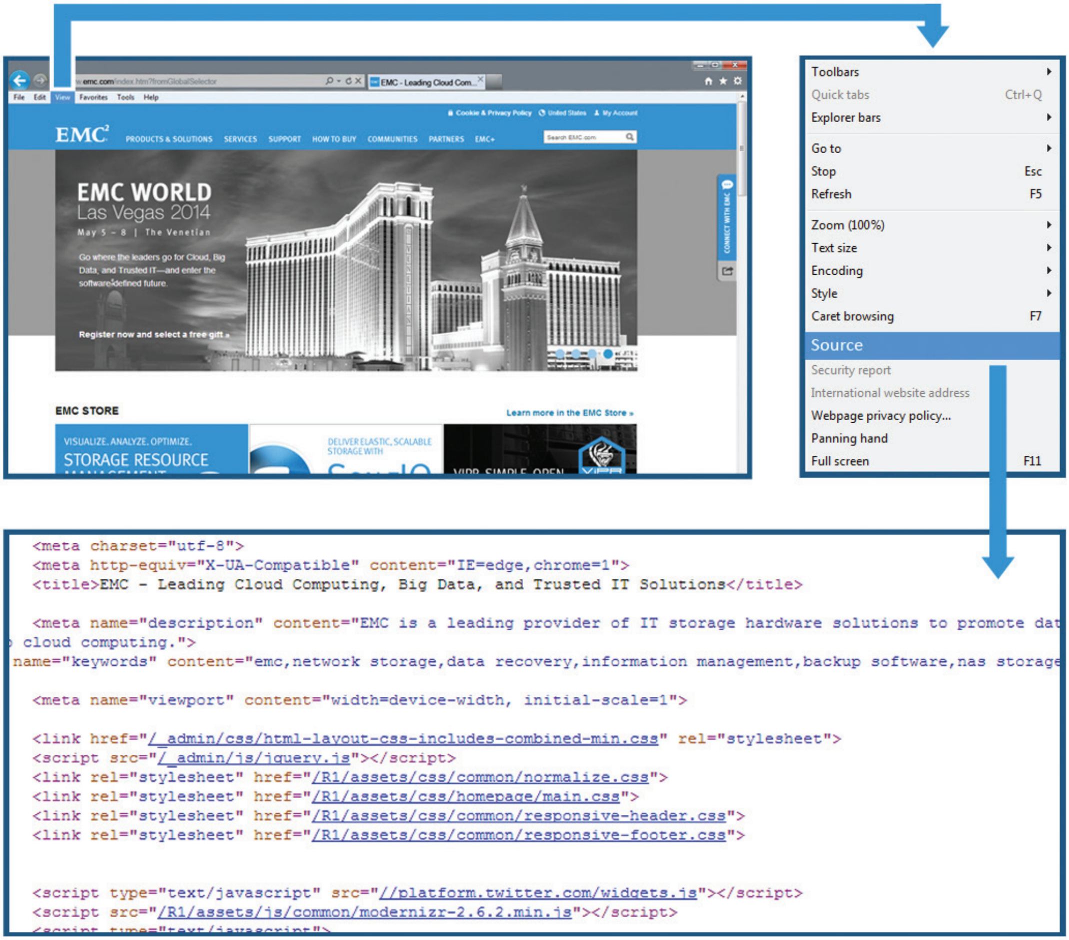
(Data as of August 01, 2011)

Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	------------Thousands------------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Figure 1-4 *Example of structured data*

- **Semi-structured data:** Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML schema). See Figure 1-5.
- **Quasi-structured data:** Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats). See Figure 1-6.
- **Unstructured data:** Data that has no inherent structure, which may include text documents, PDFs, images, and video. See Figure 1-7.

Quasi-structured data is a common phenomenon that bears closer scrutiny. Consider the following example. A user attends the EMC World conference and subsequently runs a Google search online to find information related to EMC and Data Science. This would produce a URL such as `https://www.google.com/#q=EMC+ data+science` and a list of results, such as in the first graphic of Figure 1-5.


emc.com/index.htm?fromGlobalSelector
EMC - Leading Cloud Com...
File Edit View Favorites Tools Help
EMC²
PRODUCTS & SOLUTIONS SERVICES SUPPORT HOW TO BUY COMMUNITIES PARTNERS EMC+
EMC WORLD
Las Vegas 2014
May 5 – 8 | The Venetian
Go where the leaders go for Cloud, Big Data, and Trusted IT—and enter the software-defined future.
Register now and select a free gift »
EMC STORE
Learn more in the EMC Store »
VISUALIZE. ANALYZE. OPTIMIZE.
STORAGE RESOURCE
Toolbars
Quick tabs Ctrl+Q
Explorer bars
Go to
Stop Esc
Refresh F5
Zoom (100%)
Text size
Encoding
Style
Caret browsing F7
Source
Security report
International website address
Webpage privacy policy...
Panning hand
Full screen F11

```
  <meta charset="utf-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
  <title>EMC - Leading Cloud Computing, Big Data, and Trusted IT Solutions</title>

  <meta name="description" content="EMC is a leading provider of IT storage hardware solutions to promote dat
 cloud computing.">
name="keywords" content="emc,network storage,data recovery,information management,backup software,nas storage

  <meta name="viewport" content="width=device-width, initial-scale=1">

  <link href="/_admin/css/html-layout-css-includes-combined-min.css" rel="stylesheet">
  <script src="/_admin/js/jquery.js"></script>
  <link rel="stylesheet" href="/R1/assets/css/common/normalize.css">
  <link rel="stylesheet" href="/R1/assets/css/homepage/main.css">
  <link rel="stylesheet" href="/R1/assets/css/common/responsive-header.css">
  <link rel="stylesheet" href="/R1/assets/css/common/responsive-footer.css">


  <script type="text/javascript" src="//platform.twitter.com/widgets.js"></script>
  <script src="/R1/assets/js/common/modernizr-2.6.2.min.js"></script>
  <script type="text/javascript">
```

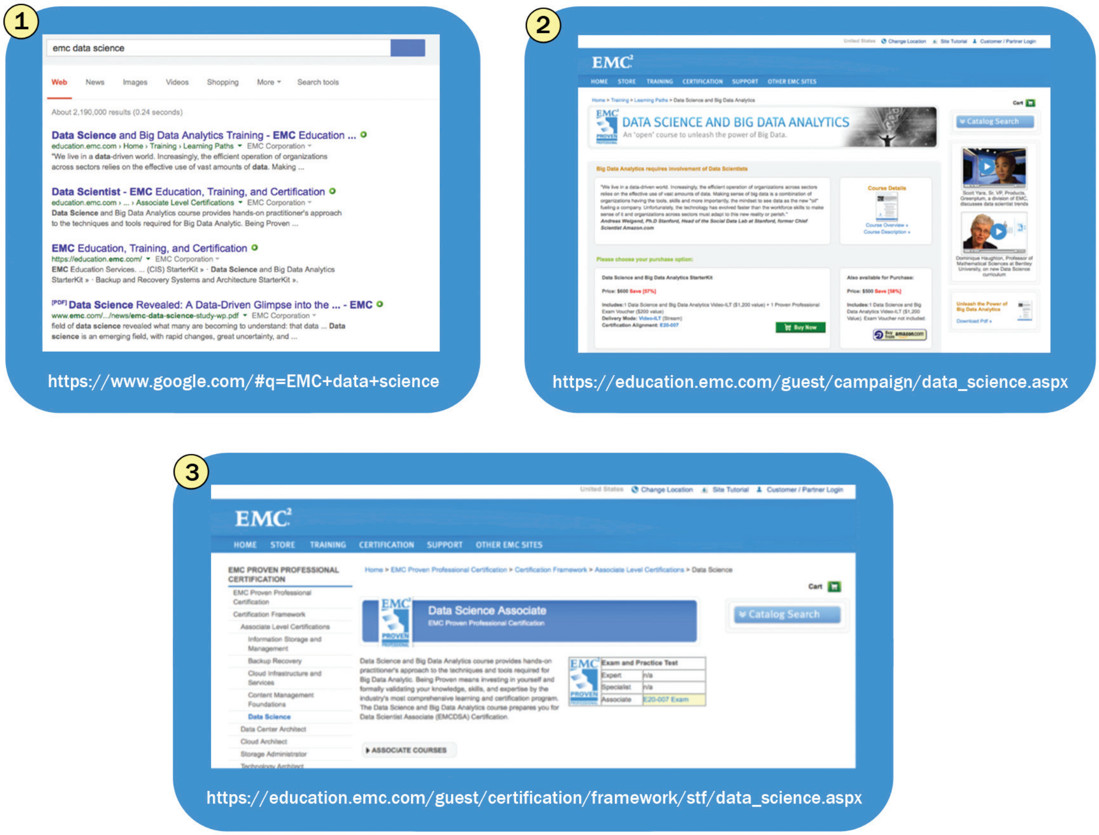
FIGURE 1-5 *Example of semi-structured data*

After doing this search, the user may choose the second link, to read more about the headline "Data Scientist—EMC Education, Training, and Certification." This brings the user to an `emc.com` site focused on this topic and a new URL, `https://education.emc.com/guest/campaign/data_science`

`.aspx`, that displays the page shown as (2) in Figure 1-6. Arriving at this site, the user may decide to click to learn more about the process of becoming certified in data science. The user chooses a link toward the top of the page on Certifications, bringing the user to a new URL: `https://education.emc.com/guest/certification/framework/stf/data_science.aspx`, which is (3) in Figure 1-6.

Visiting these three websites adds three URLs to the log files monitoring the user's computer or network use. These three URLs are:

```
https://www.google.com/#q=EMC+data+science
https://education.emc.com/guest/campaign/data_science.aspx
https://education.emc.com/guest/certification/framework/stf/data_
science.aspx
```



About 2,190,000 results (0.24 seconds)

Data Science and Big Data Analytics Training - **EMC** Education ...
education.emc.com › Home › Training › Learning Paths ▾ EMC Corporation ▾
"We live in a **data**-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of **data**. Making ...

Data Scientist - **EMC** Education, Training, and Certification
education.emc.com › ... › Associate Level Certifications ▾ EMC Corporation ▾
Data Science and Big Data Analytics course provides hands-on practitioner's approach to the techniques and tools required for Big Data Analytic. Being Proven ...

EMC Education, Training, and Certification
https://education.**emc**.com/ ▾ EMC Corporation ▾
EMC Education Services. ... (CIS) StarterKit » · **Data Science** and Big Data Analytics StarterKit » · Backup and Recovery Systems and Architecture StarterKit ».

[PDF] **Data Science** Revealed: A Data-Driven Glimpse into the ... - **EMC**
www.**emc**.com/.../news/**emc-data-science**-study-wp.pdf ▾ EMC Corporation ▾
field of **data science** revealed what many are becoming to understand: that data ... **Data science** is an emerging field, with rapid changes, great uncertainty, and ...

https://www.google.com/#q=EMC+data+science
2
United States
Change Location
Site Tutorial
Customer / Partner Login
EMC²
HOME
STORE
TRAINING
CERTIFICATION
SUPPORT
OTHER EMC SITES
Home > Training > Learning Paths > Data Science and Big Data Analytics
Cart
DATA SCIENCE AND BIG DATA ANALYTICS
An 'open' course to unleash the power of Big Data.
Catalog Search

Big Data Analytics requires involvement of Data Scientists

"We live in a data-driven world. Increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Making sense of big data is a combination of organizations having the tools, skills and more importantly, the mindset to see data as the new "oil" fueling a company. Unfortunately, the technology has evolved faster than the workforce skills to make sense of it and organizations across sectors must adapt to this new reality or perish."
Andreas Weigend, Ph.D Stanford, Head of the Social Data Lab at Stanford, former Chief Scientist Amazon.com

Course Details
Course Overview »
Course Description »

Scott Yara, Sr. VP, Products, Greenplum, a division of EMC, discusses data scientist trends

Dominique Haughton, Professor of Mathematical Sciences at Bentley University, on new Data Science curriculum

Please choose your purchase option:

Data Science and Big Data Analytics StarterKit

Price: $600 **Save [57%]**

Includes: 1 Data Science and Big Data Analytics Video-ILT ($1,200 value) + 1 Proven Professional Exam Voucher ($200 value)
Delivery Mode: Video-ILT (Stream)
Certification Alignment: E20-007

Buy Now

Also available for Purchase:

Price: $500 **Save [58%]**

Includes: 1 Data Science and Big Data Analytics Video-ILT ($1,200 Value). Exam Voucher not included.

Buy from amazon.com

Unleash the Power of Big Data Analytics
Download Pdf »

https://education.emc.com/guest/campaign/data_science.aspx
3
United States
Change Location
Site Tutorial
Customer / Partner Login
EMC²
HOME
STORE
TRAINING
CERTIFICATION
SUPPORT
OTHER EMC SITES
EMC PROVEN PROFESSIONAL CERTIFICATION
EMC Proven Professional Certification
Certification Framework
Associate Level Certifications
Information Storage and Management
Backup Recovery
Cloud Infrastructure and Services
Content Management Foundations
Data Science
Data Center Architect
Cloud Architect
Storage Administrator
Technology Architect
Home > EMC Proven Professional Certification > Certification Framework > Associate Level Certifications > Data Science
Cart
Data Science Associate
EMC Proven Professional Certification
Catalog Search

Data Science and Big Data Analytics course provides hands-on practitioner's approach to the techniques and tools required for Big Data Analytic. Being Proven means investing in yourself and formally validating your knowledge, skills, and expertise by the industry's most comprehensive learning and certification program. The Data Science and Big Data Analytics course prepares you for Data Scientist Associate (EMCDSA) Certification.

Exam and Practice Test	
Expert	n/a
Specialist	n/a
Associate	E20-007 Exam

ASSOCIATE COURSES

FIGURE 1-6 *Example of EMC Data Science search results*

FIGURE 1-7 *Example of unstructured data: video about Antarctica expedition [3]*

This set of three URLs reflects the websites and actions taken to find Data Science information related to EMC. Together, this comprises a *clickstream* that can be parsed and mined by data scientists to discover usage patterns and uncover relationships among clicks and areas of interest on a website or group of sites.

The four data types described in this chapter are sometimes generalized into two groups: structured and unstructured data. Big Data describes new kinds of data with which most organizations may not be used to working. With this in mind, the next section discusses common technology architectures from the standpoint of someone wanting to analyze Big Data.

1.1.2 Analyst Perspective on Data Repositories

The introduction of spreadsheets enabled business users to create simple logic on data structured in rows and columns and create their own analyses of business problems. Database administrator training is not required to create spreadsheets: They can be set up to do many things quickly and independently of information technology (IT) groups. Spreadsheets are easy to share, and end users have control over the logic involved. However, their proliferation can result in "many versions of the truth." In other words, it can be challenging to determine if a particular user has the most relevant version of a spreadsheet, with the most current data and logic in it. Moreover, if a laptop is lost or a file becomes corrupted, the data and logic within the spreadsheet could be lost. This is an ongoing challenge because spreadsheet programs such as Microsoft Excel still run on many computers worldwide. With the proliferation of data islands (or spreadmarts), the need to centralize the data is more pressing than ever.

As data needs grew, so did more scalable data warehousing solutions. These technologies enabled data to be managed centrally, providing benefits of security, failover, and a single repository where users could rely on getting an "official" source of data for financial reporting or other mission-critical tasks. This structure also enabled the creation of OLAP cubes and BI analytical tools, which provided quick access to a set of dimensions within an RDBMS. More advanced features enabled performance of in-depth analytical techniques such as regressions and neural networks. Enterprise Data Warehouses (EDWs) are critical for reporting and BI tasks and solve many of the problems that proliferating spreadsheets introduce, such as which of multiple versions of a spreadsheet is correct. EDWs—and a good BI strategy—provide direct data feeds from sources that are centrally managed, backed up, and secured.

Despite the benefits of EDWs and BI, these systems tend to restrict the flexibility needed to perform robust or exploratory data analysis. With the EDW model, data is managed and controlled by IT groups and database administrators (DBAs), and data analysts must depend on IT for access and changes to the data schemas. This imposes longer lead times for analysts to get data; most of the time is spent waiting for approvals rather than starting meaningful work. Additionally, many times the EDW rules restrict analysts from building datasets. Consequently, it is common for additional systems to emerge containing critical data for constructing analytic datasets, managed locally by power users. IT groups generally dislike existence of data sources outside of their control because, unlike an EDW, these datasets are not managed, secured, or backed up. From an analyst perspective, EDW and BI solve problems related to data accuracy and availability. However, EDW and BI introduce new problems related to flexibility and agility, which were less pronounced when dealing with spreadsheets.

A solution to this problem is the analytic sandbox, which attempts to resolve the conflict for analysts and data scientists with EDW and more formally managed corporate data. In this model, the IT group may still manage the analytic sandboxes, but they will be purposefully designed to enable robust analytics, while being centrally managed and secured. These sandboxes, often referred to as *workspaces,* are designed to enable teams to explore many datasets in a controlled fashion and are not typically used for enterprise-level financial reporting and sales dashboards.

Many times, analytic sandboxes enable high-performance computing using in-database processing—the analytics occur within the database itself. The idea is that performance of the analysis will be better if the analytics are run in the database itself, rather than bringing the data to an analytical tool that resides somewhere else. In-database analytics, discussed further in Chapter 11, "Advanced Analytics—Technology and Tools: In-Database Analytics," creates relationships to multiple data sources within an organization and saves time spent creating these data feeds on an individual basis. In-database processing for deep analytics enables faster turnaround time for developing and executing new analytic models, while reducing, though not eliminating, the cost associated with data stored in local, "shadow" file systems. In addition, rather than the typical structured data in the EDW, analytic sandboxes can house a greater variety of data, such as raw data, textual data, and other kinds of unstructured data, without interfering with critical production databases. Table 1-1 summarizes the characteristics of the data repositories mentioned in this section.

Table 1-1 *Types of Data Repositories, from an Analyst Perspective*

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.