

# Contents at a Glance

<b><i>Introduction</i></b> .....	<b>1</b>
<b><i>Part I: Getting Started with Big Data</i></b> .....	<b>7</b>
Chapter 1: Grasping the Fundamentals of Big Data.....	9
Chapter 2: Examining Big Data Types .....	25
Chapter 3: Old Meets New: Distributed Computing.....	37
<b><i>Part II: Technology Foundations for Big Data</i></b> .....	<b>45</b>
Chapter 4: Digging into Big Data Technology Components .....	47
Chapter 5: Virtualization and How It Supports Distributed Computing.....	61
Chapter 6: Examining the Cloud and Big Data .....	71
<b><i>Part III: Big Data Management</i></b> .....	<b>83</b>
Chapter 7: Operational Databases.....	85
Chapter 8: MapReduce Fundamentals .....	101
Chapter 9: Exploring the World of Hadoop .....	111
Chapter 10: The Hadoop Foundation and Ecosystem.....	121
Chapter 11: Appliances and Big Data Warehouses .....	129
<b><i>Part IV: Analytics and Big Data</i></b> .....	<b>139</b>
Chapter 12: Defining Big Data Analytics .....	141
Chapter 13: Understanding Text Analytics and Big Data.....	153
Chapter 14: Customized Approaches for Analysis of Big Data.....	167
<b><i>Part V: Big Data Implementation</i></b> .....	<b>179</b>
Chapter 15: Integrating Data Sources.....	181
Chapter 16: Dealing with Real-Time Data Streams and Complex Event Processing .....	193
Chapter 17: Operationalizing Big Data.....	201
Chapter 18: Applying Big Data within Your Organization .....	211
Chapter 19: Security and Governance for Big Data Environments .....	225

<b><i>Part VI: Big Data Solutions in the Real World</i></b> .....	<b>235</b>
Chapter 20: The Importance of Big Data to Business .....	237
Chapter 21: Analyzing Data in Motion: A Real-World View .....	245
Chapter 22: Improving Business Processes with Big Data Analytics: A Real-World View .....	255
<b><i>Part VII: The Part of Tens</i></b> .....	<b>263</b>
Chapter 23: Ten Big Data Best Practices .....	265
Chapter 24: Ten Great Big Data Resources .....	271
Chapter 25: Ten Big Data Do's and Don'ts.....	275
<b><i>Glossary</i></b> .....	<b>279</b>
<b><i>Index</i></b> .....	<b>295</b>

# Table of Contents

---

## ***Introduction*** ..... 1

About This Book .....	2
Foolish Assumptions .....	2
How This Book Is Organized .....	3
Part I: Getting Started with Big Data.....	3
Part II: Technology Foundations for Big Data .....	3
Part III: Big Data Management .....	3
Part IV: Analytics and Big Data .....	4
Part V: Big Data Implementation.....	4
Part VI: Big Data Solutions in the Real World.....	4
Part VII: The Part of Tens.....	4
Glossary .....	4
Icons Used in This Book .....	5
Where to Go from Here.....	5

## ***Part 1: Getting Started with Big Data***..... 7

### **Chapter 1: Grasping the Fundamentals of Big Data** ..... 9

The Evolution of Data Management .....	10
Understanding the Waves of Managing Data .....	11
Wave 1: Creating manageable data structures.....	11
Wave 2: Web and content management .....	13
Wave 3: Managing big data .....	14
Defining Big Data.....	15
Building a Successful Big Data Management Architecture .....	16
Beginning with capture, organize, integrate, analyze, and act .....	16
Setting the architectural foundation .....	17
Performance matters.....	20
Traditional and advanced analytics .....	22
The Big Data Journey .....	23

### **Chapter 2: Examining Big Data Types** ..... 25

Defining Structured Data .....	26
Exploring sources of big structured data .....	26
Understanding the role of relational databases in big data .....	27
Defining Unstructured Data.....	29
Exploring sources of unstructured data .....	29
Understanding the role of a CMS in big data management .....	31

Looking at Real-Time and Non-Real-Time Requirements .....	32
Putting Big Data Together .....	33
Managing different data types.....	33
Integrating data types into a big data environment .....	34
<b>Chapter 3: Old Meets New: Distributed Computing . . . . .</b>	<b>37</b>
A Brief History of Distributed Computing .....	37
Giving thanks to DARPA.....	38
The value of a consistent model .....	39
Understanding the Basics of Distributed Computing .....	40
Why we need distributed computing for big data.....	40
The changing economics of computing .....	40
The problem with latency.....	41
Demand meets solutions.....	41
Getting Performance Right .....	42
 <b>Part 11: Technology Foundations for Big Data .....</b>	 <b>45</b>
<b>Chapter 4: Digging into Big Data Technology Components . . . . .</b>	<b>47</b>
Exploring the Big Data Stack .....	48
Layer 0: Redundant Physical Infrastructure .....	49
Physical redundant networks .....	51
Managing hardware: Storage and servers .....	51
Infrastructure operations .....	51
Layer 1: Security Infrastructure.....	52
Interfaces and Feeds to and from Applications and the Internet.....	53
Layer 2: Operational Databases.....	54
Layer 3: Organizing Data Services and Tools.....	56
Layer 4: Analytical Data Warehouses .....	56
Big Data Analytics.....	58
Big Data Applications .....	58
 <b>Chapter 5: Virtualization and How It Supports Distributed Computing . . . . .</b>	 <b>61</b>
Understanding the Basics of Virtualization .....	61
The importance of virtualization to big data .....	63
Server virtualization .....	64
Application virtualization .....	65
Network virtualization.....	66
Processor and memory virtualization.....	66
Data and storage virtualization.....	67
Managing Virtualization with the Hypervisor .....	68
Abstraction and Virtualization .....	69
Implementing Virtualization to Work with Big Data .....	69

**Chapter 6: Examining the Cloud and Big Data . . . . . 71**

Defining the Cloud in the Context of Big Data .....	71
Understanding Cloud Deployment and Delivery Models .....	72
Cloud deployment models.....	73
Cloud delivery models .....	74
The Cloud as an Imperative for Big Data.....	75
Making Use of the Cloud for Big Data .....	77
Providers in the Big Data Cloud Market .....	78
Amazon's Public Elastic Compute Cloud.....	78
Google big data services .....	79
Microsoft Azure.....	80
OpenStack.....	80
Where to be careful when using cloud services .....	81

**Part III: Big Data Management ..... 83****Chapter 7: Operational Databases . . . . . 85**

RDBMSs Are Important in a Big Data Environment.....	87
PostgreSQL relational database.....	87
Nonrelational Databases.....	88
Key-Value Pair Databases .....	89
Riak key-value database.....	90
Document Databases .....	91
MongoDB.....	92
CouchDB .....	93
Columnar Databases .....	94
HBase columnar database .....	94
Graph Databases.....	95
Neo4J graph database .....	96
Spatial Databases.....	97
PostGIS/OpenGEO Suite.....	98
Polyglot Persistence.....	99

**Chapter 8: MapReduce Fundamentals . . . . . 101**

Tracing the Origins of MapReduce.....	101
Understanding the map Function.....	103
Adding the reduce Function.....	104
Putting map and reduce Together .....	105
Optimizing MapReduce Tasks .....	108
Hardware/network topology .....	108
Synchronization .....	108
File system .....	108

<b>Chapter 9: Exploring the World of Hadoop</b> .....	<b>111</b>
Explaining Hadoop .....	111
Understanding the Hadoop Distributed File System (HDFS) .....	112
NameNodes .....	113
Data nodes .....	114
Under the covers of HDFS .....	115
Hadoop MapReduce .....	116
Getting the data ready .....	117
Let the mapping begin .....	118
Reduce and combine .....	118
<b>Chapter 10: The Hadoop Foundation and Ecosystem</b> .....	<b>121</b>
Building a Big Data Foundation with the Hadoop Ecosystem .....	121
Managing Resources and Applications with Hadoop YARN .....	122
Storing Big Data with HBase .....	123
Mining Big Data with Hive .....	124
Interacting with the Hadoop Ecosystem .....	125
Pig and Pig Latin .....	125
Sqoop .....	126
Zookeeper .....	127
<b>Chapter 11: Appliances and Big Data Warehouses</b> .....	<b>129</b>
Integrating Big Data with the Traditional Data Warehouse .....	129
Optimizing the data warehouse .....	130
Differentiating big data structures from data warehouse data ...	130
Examining a hybrid process case study .....	131
Big Data Analysis and the Data Warehouse .....	133
The integration lynchpin .....	134
Rethinking extraction, transformation, and loading .....	134
Changing the Role of the Data Warehouse .....	135
Changing Deployment Models in the Big Data Era .....	136
The appliance model .....	136
The cloud model .....	137
Examining the Future of Data Warehouses .....	137
<b>Part IV: Analytics and Big Data</b> .....	<b>139</b>
<b>Chapter 12: Defining Big Data Analytics</b> .....	<b>141</b>
Using Big Data to Get Results .....	142
Basic analytics .....	142
Advanced analytics .....	143
Operationalized analytics .....	146
Monetizing analytics .....	146

Modifying Business Intelligence Products to Handle Big Data.....	147
Data.....	147
Analytical algorithms .....	148
Infrastructure support .....	148
Studying Big Data Analytics Examples.....	149
Orbitz.....	149
Nokia.....	150
NASA.....	150
Big Data Analytics Solutions .....	151

### **Chapter 13: Understanding Text Analytics and Big Data . . . . . 153**

Exploring Unstructured Data .....	154
Understanding Text Analytics .....	155
The difference between text analytics and search.....	156
Analysis and Extraction Techniques.....	157
Understanding the extracted information.....	159
Taxonomies .....	160
Putting Your Results Together with Structured Data.....	160
Putting Big Data to Use .....	161
Voice of the customer .....	161
Social media analytics.....	162
Text Analytics Tools for Big Data.....	164
Attensity.....	164
Clarabridge .....	165
IBM.....	165
OpenText .....	165
SAS .....	166

### **Chapter 14: Customized Approaches for Analysis of Big Data . . . . 167**

Building New Models and Approaches to Support Big Data.....	168
Characteristics of big data analysis .....	168
Understanding Different Approaches to Big Data Analysis .....	170
Custom applications for big data analysis .....	171
Semi-custom applications for big data analysis.....	173
Characteristics of a Big Data Analysis Framework .....	174
Big to Small: A Big Data Paradox .....	177

## ***Part V: Big Data Implementation..... 179***

### **Chapter 15: Integrating Data Sources. . . . . 181**

Identifying the Data You Need .....	181
Exploratory stage.....	182
Codifying stage.....	184
Integration and incorporation stage .....	184

Understanding the Fundamentals of Big Data Integration .....	186
Defining Traditional ETL.....	187
Data transformation .....	188
Understanding ELT — Extract, Load, and Transform.....	189
Prioritizing Big Data Quality.....	189
Using Hadoop as ETL.....	191
Best Practices for Data Integration in a Big Data World.....	191

## **Chapter 16: Dealing with Real-Time Data Streams and Complex Event Processing . . . . . 193**

Explaining Streaming Data and Complex Event Processing.....	194
Using Streaming Data.....	194
Data streaming .....	195
The need for metadata in streams.....	196
Using Complex Event Processing .....	198
Differentiating CEP from Streams .....	199
Understanding the Impact of Streaming Data and CEP on Business ....	200

## **Chapter 17: Operationalizing Big Data . . . . . 201**

Making Big Data a Part of Your Operational Process .....	201
Integrating big data.....	202
Incorporating big data into the diagnosis of diseases .....	203
Understanding Big Data Workflows .....	205
Workload in context to the business problem.....	206
Ensuring the Validity, Veracity, and Volatility of Big Data.....	207
Data validity.....	207
Data volatility .....	208

## **Chapter 18: Applying Big Data within Your Organization. . . . . 211**

Figuring the Economics of Big Data .....	212
Identification of data types and sources.....	212
Business process modifications or new process creation .....	215
The technology impact of big data workflows .....	215
Finding the talent to support big data projects .....	216
Calculating the return on investment (ROI) from big data investments.....	216
Enterprise Data Management and Big Data.....	217
Defining Enterprise Data Management.....	217
Creating a Big Data Implementation Road Map.....	218
Understanding business urgency .....	218
Projecting the right amount of capacity .....	219
Selecting the right software development methodology.....	219
Balancing budgets and skill sets .....	219
Determining your appetite for risk.....	220
Starting Your Big Data Road Map.....	220



**Chapter 19: Security and Governance for Big Data Environments . . . 225**

Security in Context with Big Data.....	225
Assessing the risk for the business .....	226
Risks lurking inside big data.....	226
Understanding Data Protection Options .....	227
The Data Governance Challenge .....	228
Auditing your big data process.....	230
Identifying the key stakeholders.....	231
Putting the Right Organizational Structure in Place .....	231
Preparing for stewardship and management of risk.....	232
Setting the right governance and quality policies.....	232
Developing a Well-Governed and Secure Big Data Environment .....	233

***Part VI: Big Data Solutions in the Real World..... 235*****Chapter 20: The Importance of Big Data to Business . . . . . 237**

Big Data as a Business Planning Tool .....	238
Stage 1: Planning with data.....	238
Stage 2: Doing the analysis .....	239
Stage 3: Checking the results.....	239
Stage 4: Acting on the plan .....	240
Adding New Dimensions to the Planning Cycle.....	240
Stage 5: Monitoring in real time .....	240
Stage 6: Adjusting the impact.....	241
Stage 7: Enabling experimentation .....	241
Keeping Data Analytics in Perspective .....	241
Getting Started with the Right Foundation .....	242
Getting your big data strategy started .....	242
Planning for Big Data.....	243
Transforming Business Processes with Big Data .....	244

**Chapter 21: Analyzing Data in Motion: A Real-World View. . . . . 245**

Understanding Companies' Needs for Data in Motion .....	246
The value of streaming data.....	247
Streaming Data with an Environmental Impact .....	247
Using sensors to provide real-time information about rivers and oceans .....	248
The benefits of real-time data .....	249
Streaming Data with a Public Policy Impact .....	249
Streaming Data in the Healthcare Industry .....	251
Capturing the data stream.....	251

Streaming Data in the Energy Industry .....	252
Using streaming data to increase energy efficiency .....	252
Using streaming data to advance the production of alternative sources of energy .....	252
Connecting Streaming Data to Historical and Other Real-Time Data Sources .....	253

## **Chapter 22: Improving Business Processes with Big Data Analytics: A Real-World View . . . . . 255**

Understanding Companies' Needs for Big Data Analytics .....	256
Improving the Customer Experience with Text Analytics .....	256
The business value to the big data analytics implementation ....	257
Using Big Data Analytics to Determine Next Best Action .....	257
Preventing Fraud with Big Data Analytics .....	260
The Business Benefit of Integrating New Sources of Data .....	262

## ***Part VII: The Part of Tens* .....** 263

### **Chapter 23: Ten Big Data Best Practices . . . . . 265**

Understand Your Goals .....	265
Establish a Road Map .....	266
Discover Your Data .....	266
Figure Out What Data You Don't Have .....	267
Understand the Technology Options .....	267
Plan for Security in Context with Big Data .....	268
Plan a Data Governance Strategy .....	268
Plan for Data Stewardship .....	268
Continually Test Your Assumptions .....	269
Study Best Practices and Leverage Patterns .....	269

### **Chapter 24: Ten Great Big Data Resources . . . . . 271**

Hurwitz & Associates .....	271
Standards Organizations .....	271
The Open Data Foundation .....	272
The Cloud Security Alliance .....	272
National Institute of Standards and Technology .....	272
Apache Software Foundation .....	273
OASIS .....	273
Vendor Sites .....	273
Online Collaborative Sites .....	274
Big Data Conferences .....	274

**Chapter 25: Ten Big Data Do's and Don'ts ..... 275**

Do Involve All Business Units in Your Big Data Strategy ..... 275  
Do Evaluate All Delivery Models for Big Data ..... 276  
Do Think about Your Traditional Data Sources as Part of  
Your Big Data Strategy ..... 276  
Do Plan for Consistent Metadata ..... 276  
Do Distribute Your Data ..... 277  
Don't Rely on a Single Approach to Big Data Analytics ..... 277  
Don't Go Big Before You Are Ready ..... 277  
Don't Overlook the Need to Integrate Data ..... 277  
Don't Forget to Manage Data Securely ..... 278  
Don't Overlook the Need to Manage the Performance of Your Data .... 278

***Glossary* ..... 279**

***Index* ..... 295**

