CONTENTS

Preface 1

The Fourth Industrial Revolution and Big Data 1
About the Book 2
Features of the Book 4
Acknowledgments 4
About the Author 5
Trademark Information 5

Chapter 1 Big Data and Analytics 7

Big Data Application: COVID-19 7
Introduction 7
Big Data: A Historical Perspective 8
Big Data Definition and Technologies 9
Big Data Analytics in Action 11
Implementing Big Data Projects 14
Summary and Takeaways 16
Review Questions 17
Practice Exercises 18

Chapter 2 Cloud Computing and Big Data 19

Big Data Application: Amazon Web Services 19
Introduction 19
What Is Cloud Computing? 19
Advantages of Using Cloud Computing for Big Data Analytics 21
Challenges of Cloud Computing for Big Data Analytics 22
Tutorial A: Setting Up a Hadoop Cluster on AWS 22
Summary and Takeaways 33
Review Questions 35
Practice Exercises 35

Chapter 3 Distributed File Systems 37

Big Data Application: Airbnb 37
Introduction 37
Why Distributed File Systems? 37
Complications of Traditional Distributed File Systems 38
Consistency and Availability 39
Hadoop and Its Distributed File System 41
HDFS Architecture 42
Write and Read Operations in HDFS 43
The MapReduce Layer 45
Summary and Takeaways 46
Review Questions 47
Practice Exercises 47

Chapter 4 Anatomy of MapReduce 49

Big Data Application: Social Media Influencers 49
Introduction 49
The MapReduce Concept 50

The Main Components of MapReduce 50
The Pseudocode of MapReduce 52
MapReduce Patterns and Examples 55
Variations of the WordCount Algorithm 58
The MapReduce in Java and Python 59
Tutorial A: Executing a MapReduce Job with Java 62
Tutorial B: Executing a MapReduce Job with Python in the Hadoop Cluster 64
Summary and Takeaways 67
Review Questions 68
Practice Exercises 68

Chapter 5 Apache Pig and Pig Latin 71

Big Data Application: LinkedIn 71
Introduction 71
Pig vs. MapReduce 72
Pig Components and the Execution Modes 72
Pig Latin Data Types 73
Pig Latin Operators 74
Tutorial A: Website Visitors 74
Tutorial B: Fast-Food Employees 78
Tutorial C: The WordCount Problem 82
Additional Considerations When Using Pig and Pig Latin 84
Summary and Takeaways 84
Review Questions 85
Practice Exercises 85

Chapter 6 Apache Hive and HiveQL 89

Big Data Application: Uber 89
Introduction 89
Comparing Hive to RDBMS and Pig 89
Hive Architecture and Components 90
Hive Data Models and Units 91
Hive Data Types 92
Tutorial A: Hive in Action 93
Tutorial B: Performing Data Analysis with Hive 98
Summary and Takeaways 101
Review Questions 102
Practice Exercises 102

Chapter 7 Moving Data with Sqoop 105

Big Data Application: Wells Fargo 105
Introduction 105
Overview of Apache Sqoop 106
Tutorial A: Moving Data Between MySQL and HDFS 106
Tutorial B: Importing Data from MySQL to Hive 111
Tutorial C: Moving Data Between a Local Database and HDFS 113
Summary and Takeaways 115
Review Questions 115
Practice Exercises 116

Chapter 8 NoSQL Databases 119

Big Data Application: The Weather Channel 119
Introduction 119

Unstructured Data and NoSQL Technology 120
Key-Value Databases 121
Document-Based Databases 122
Column-Based Databases 124
Graph-Based NoSQL Databases 126
Differences Between Relational and NoSQL Databases 129
Summary and Takeaways 130
Review Questions 132
Practice Exercises 132

Chapter 9 BigTable and HBase 135

Big Data Application: Monster.com 135
Introduction 135
The HBase Architecture 136
The BigTable Data Model 136
Tutorial A: Basic HBase Shell Commands 140
Tutorial B: Creating and Populating Tables with HBase 144
Tutorial C: Uploading and Downloading Data Between HDFS and HBase 147
Tutorial D: Data Manipulation in HBase 149
Summary and Takeaways 151
Review Questions 151
Practice Exercises 151

Chapter 10 Introduction to Spark 153

Big Data Application: CrowdStrike 153
Introduction 153
Essential Components of Apache Spark 154
Spark DataFrames 154
Tutorial A: Exploring DataFrames 155
Tutorial B: Analyzing Data with DataFrames 161
Tutorial C: Performing Joins with DataFrames 162
Spark Datasets 163
Summary and Takeaways 164
Review Questions 165
Practice Exercises 166

Chapter 11 Resilient Distributed Datasets 169

Big Data Application: Walmart 169
Introduction 169
Loading Data into RDDs and Saving RDDs to Files 170
RDD Action Operations 172
Tutorial A: Transforming Data with RDDs 173
Tutorial B: Creating a DataFrame from an RDD 176
Pair RDDs and MapReduce 178
Tutorial C: WordCount with Pair RDDs 179
Summary and Takeaways 181
Review Questions 182
Practice Exercises 183

Chapter 12 Applications with Spark 187

Big Data Application: Experian 187
Introduction 187
Spark Applications 187

Tutorial A: Creating and Running a Scala Application 188
Tutorial B: Running a Python Application 196
Summary and Takeaways 197
Review Questions 197
Practice Exercises 197

Chapter 13 Iterative Processing and Data Streaming with Spark 199

Big Data Application: Verizon Media Group 199
Introduction 199
Overview of Iterative Processing Techniques 200
Tutorial A: Running the PageRank Algorithm in the Interactive Scala Shell 201
Tutorial B: Running the PageRank Algorithm with a Spark Application 202
Spark Streaming Overview 205
Tutorial C: Running a Streaming Application with a Spark Application 205
Summary and Takeaways 207
Review Questions 208
Practice Exercises 208

Appendix A: Shell Commands in Linux and HDFS 211

Part 1: Basic Shell Commands 211
Part 2: HDFS Commands 216

Appendix B: Creating a Database Instance in AWS 221

Part1: Creating a Role with RDS Access Privileges 221
Part 2: Creating a MySQL Database Instance 225

Bibliography 229

Index 237