

Contents

Preface	vii
The Fourth Industrial Revolution and Big Data	vii
About the Book	viii
About the Second Edition of the Book	x
Features of the Book	xi
Acknowledgments	xii
About the Authors	xiii
Trademark Information	xiii
Chapter 1 Big Data and Analytics	1
Introduction	1
Big Data: A Historical Perspective	1
Big Data Definition and Technologies	3
Big Data Analytics in Action	6
Implementing Big Data Projects	8
Summary and Takeaways	10
Chapter 2 Cloud Computing and Big Data	13
Introduction	13
What Is Cloud Computing?	14
Advantages of Using Cloud Computing for Big Data Analytics	15
Challenges of Cloud Computing for Big Data Analytics	16
Tutorial: Setting Up a Hadoop Cluster on AWS	17
Summary and Takeaways	18
Chapter 3 Understanding and Working with Data	21
Introduction	21
What Is Data?	21
Data Sources	23
Data Formats	24
Cleaning and Preparing Data	28
Processing and Parsing Data	29
Summary and Takeaways	30
Chapter 4 Distributed File Systems	32
Introduction	32
Why Distributed File Systems?	33
Complications of Traditional Distributed File Systems	33
Hadoop and Its Distributed File System	36
The Hadoop System	37

Hadoop Architecture Components	38
Write and Read Operations in the HDFS	40
Summary and Takeaways	42
Chapter 5 Anatomy of MapReduce	45
Introduction	45
The MapReduce Concept	46
The MapReduce Layer	46
The Main Components of MapReduce	47
The Pseudocode of MapReduce	49
MapReduce Patterns and Examples	52
Variations of the WordCount Algorithm	53
MapReduce in Python	55
Tutorial: Executing a MapReduce Job with Python in the Hadoop Cluster	55
Summary and Takeaways	57
Chapter 6 Apache Pig and Pig Latin	61
Introduction	61
Pig vs. MapReduce	62
Pig Components and the Execution Modes	62
Pig Latin Data Types	63
Pig Latin Operators	64
Tutorial A: Website Visitors	65
Tutorial B: Fast-Food Employees	65
Tutorial C: The WordCount Problem	65
Additional Considerations When Using Pig and Pig Latin	65
Summary and Takeaways	65
Chapter 7 Apache Hive and HiveQL	70
Introduction	70
Comparing Hive to RDBMS and Pig	70
Hive Architecture and Components	71
Hive Data Models and Units	72
Hive Data Types	73
Tutorial A: Hive in Action	74
Tutorial B: Performing Data Analysis with Hive	74
Summary and Takeaways	74
Chapter 8 NoSQL Databases	78
Introduction	78
Unstructured Data and NoSQL Technology	79
Key-Value Databases	79
Document-Based Databases	81
Column-Based Databases	82
Graph-Based NoSQL Databases	84
Differences Between Relational and NoSQL Databases	88
Summary and Takeaways	89

Chapter 9 BigTable and HBase	93
Introduction	93
The HBase Architecture	94
The BigTable Data Model	95
Tutorial A: Basic HBase Shell Commands	99
Tutorial B: Creating and Populating Tables with HBase	99
Tutorial C: Uploading and Downloading Data Between the HDFS and HBase	99
Tutorial D: Data Manipulation in HBase	99
Summary and Takeaways	99
Chapter 10 Introduction to Spark	102
Introduction	102
Essential Components of Apache Spark	103
Tutorial A: Exploring DataFrames	104
Tutorial B: Analyzing Data with DataFrames	104
Tutorial C: Performing Joins with DataFrames	104
Spark DataFrames	104
Spark Datasets	104
Summary and Takeaways	106
Chapter 11 Resilient Distributed Datasets	110
Introduction	110
Loading Data into RDDs and Saving RDDs to Files	111
RDD Action Operations	113
Tutorial A: Transforming Data with RDDs	114
Tutorial B: Creating a DataFrame from an RDD	114
Pair RDDs and MapReduce	114
Tutorial C: WordCount with Pair RDDs	115
Summary and Takeaways	115
Chapter 12 Applications, Iterative Processing, and Data Streaming with Spark	120
Introduction	120
Spark Applications	121
Tutorial A: Running a Python Application	121
Overview of Iterative Processing Techniques	122
Tutorial B: Running the PageRank Algorithm with a Spark Application	123
Spark Streaming Overview	123
Tutorial C: Running a Streaming Application with a Spark Application	123
Summary and Takeaways	123
Bibliography	129
Index	137