## Contents

### Preface

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>xi</td>
</tr>
</tbody>
</table>

### Part I Understanding and Dealing with Big Data

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>1.1</td>
<td>What Is Big Data?</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Success (and Failure) of Big Data</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Big Data Evolution and Definition</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1</td>
<td>The Faces of Big Data</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>How to Deal with Big Data?</td>
<td>9</td>
</tr>
<tr>
<td>1.4.1</td>
<td>High Performance Computing versus Big Data Computing</td>
<td>11</td>
</tr>
<tr>
<td>1.4.2</td>
<td>Distributed Systems for Data-Intensive Jobs</td>
<td>15</td>
</tr>
<tr>
<td>1.5</td>
<td>Take-Home Message</td>
<td>16</td>
</tr>
<tr>
<td>1.6</td>
<td>To Learn More</td>
<td>17</td>
</tr>
<tr>
<td>1.7</td>
<td>Exercises</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>MapReduce</td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>What Is MapReduce?</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>Map and Reduce Functions in Python</td>
<td>20</td>
</tr>
<tr>
<td>2.3</td>
<td>“Hello World” in Big Data: Word Count</td>
<td>23</td>
</tr>
<tr>
<td>2.4</td>
<td>Working with Key-Value Pairs</td>
<td>27</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Map</td>
<td>28</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Shuffle</td>
<td>29</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Reduce</td>
<td>30</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Implementing Word Count with MapReduce</td>
<td>31</td>
</tr>
<tr>
<td>2.5</td>
<td>Combiners</td>
<td>33</td>
</tr>
<tr>
<td>2.6</td>
<td>Internal Working</td>
<td>34</td>
</tr>
<tr>
<td>2.7</td>
<td>Take-Home Message</td>
<td>35</td>
</tr>
<tr>
<td>2.8</td>
<td>To Learn More</td>
<td>36</td>
</tr>
<tr>
<td>2.9</td>
<td>Solutions to Challenges</td>
<td>36</td>
</tr>
<tr>
<td>2.10</td>
<td>Exercises</td>
<td>39</td>
</tr>
</tbody>
</table>
# Table of Contents

## Part II Big Data Frameworks

### 3 Hadoop

3.1 What Is Hadoop? 45
3.2 Word Count Using Hadoop MapReduce 46
3.3 Resource Negotiator: Internal Working
   - 3.3.1 Components 48
   - 3.3.2 Executing a MapReduce Process on YARN 51
3.4 Hadoop Distributed File System
   - 3.4.1 What is the HDFS? 55
   - 3.4.2 How HDFS Works 57
3.5 Limitations: Technologies and Frameworks beyond Hadoop 60
3.6 Take-Home Message 62
3.7 To Learn More 63
3.8 Solutions to Challenges 64
3.9 Exercises 66

### 4 Spark

4.1 What Is Spark? 69
   - 4.1.1 The Goal of Apache Spark 70
4.2 Basic Concepts 71
4.3 Resilient Distributed Datasets
   - 4.3.1 Creating RDDs 73
   - 4.3.2 Operations with RDDs 75
   - 4.3.3 Transformations 75
   - 4.3.4 Actions on RDDs 81
   - 4.3.5 Key–Value Transformations 83
   - 4.3.6 Key–Value Actions 87
   - 4.3.7 File I/O 88
   - 4.3.8 RDD Lineage 90
   - 4.3.9 Cache Your RDDs 91
4.4 Advanced Concepts
   - 4.4.1 Shared Variables 92
   - 4.4.2 Partitions 95
   - 4.4.3 Operations with Numeric RDDs 97
4.5 Internal Workings 98
   - 4.5.1 Anatomy of a Spark Application 100
4.6 Take-Home Message 101
4.7 To Learn More 102
4.8 Solutions to Challenges 102
4.9 Exercises 106

### 5 Spark SQL and DataFrames

5.1 What Is Spark SQL? 110
5.2 Basic Concepts: From RDDs to Structured Data 110
  5.2.1 DataFrame and Dataset APIs 110
  5.2.2 Optimizations: DataFrames and Datasets versus RDDs 111
  5.2.3 Spark Session: A New Entry Point to Spark 112
5.3 DataFrames 113
  5.3.1 Creating a DataFrame 115
  5.3.2 Operations with DataFrames 125
  5.3.3 Transformations 125
  5.3.4 Actions on DataFrames 152
  5.3.5 Caching DataFrames 155
5.4 Advanced Concepts 156
5.5 Take-Home Message 163
5.6 To Learn More 164
5.7 Solutions to Challenges 165
5.8 Exercises 171

Part III Machine Learning for Big Data 175

6 Machine Learning with Spark 177
  6.1 Machine Learning Basics and Key Terminology 177
    6.1.1 Supervised Learning: Classification and Regression 178
    6.1.2 Unsupervised Learning: Dimensionality Reduction and Clustering 179
    6.1.3 Machine Learning Life Cycle 179
    6.1.4 Data Preparation and Preprocessing 180
    6.1.5 Model Evaluation and Selection 182
    6.1.6 Machine Learning Pipelines 183
  6.2 Machine Learning with Big Data 184
  6.3 The Spark Machine Learning Library 185
    6.3.1 Basic Concepts: Transformers, Estimators, and Pipelines 187
    6.3.2 An Example with Linear Regression 188
    6.3.3 Hyperparameter Tuning 198
  6.4 Take-Home Message 201
  6.5 To Learn More 202
  6.6 Solutions to Challenges 203
  6.7 Exercises 210

7 Machine Learning for Big Data 212
  7.1 Designing Machine Learning Methods for Big Data 212
    7.1.1 Example with Decision Trees 213
    7.1.2 Definitions: Global versus Local 227
  7.2 Scalability Measures: Is My Algorithm Scalable? 229
    7.2.1 Speed-Up 229
    7.2.2 Size-Up 230
    7.2.3 Scale-Up 230
    7.2.4 A Test on Decision Trees 231
# Table of Contents

7.3 Take-Home Message 242  
7.4 To Learn More 244  
7.5 Solutions to Challenges 244  
7.6 Exercises 251  

8 Implementing Classical Methods: \(k\)-Means and Linear Regression 254  
8.1 Clustering: \(k\)-Means 254  
8.1.1 Non-distributed Implementation 257  
8.1.2 Distributed Implementation 260  
8.2 Linear Regression 265  
8.2.1 Closed-Form Solution 266  
8.2.2 Gradient Descent 279  
8.3 Take-Home Message 290  
8.4 To Learn More 292  
8.5 Solutions to Challenges 293  
8.6 Exercises 303  

9 Advanced Examples: Semi-supervised, Ensembles, Deep Learning Model Deployment 305  
9.1 Semi-supervised Classification: Self-Labeling Approaches 306  
9.1.1 Self-Training for Classification 306  
9.1.2 A Global Solution for Self-Training 308  
9.2 Ensemble Learning: Bagging and Boosting 320  
9.2.1 A Global Solution for Bagging 321  
9.2.2 A Global Solution for Boosting: AdaBoost 327  
9.3 Model Deployment: Case Study in Image Segmentation with Deep Learning 336  
9.3.1 Case Study on Remote Sensing: Building Semantic Segmentation 336  
9.3.2 Non-distributed Implementation 346  
9.3.3 Distributed Implementation 350  
9.4 Take-Home Message 354  
9.5 To Learn More 356  
9.6 Solutions to Challenges 357  
9.7 Exercises 367  

Bibliography 369  
Index 373