

---

# Table of Contents

|              |     |
|--------------|-----|
| Preface..... | vii |
|--------------|-----|

---

## Part I. Introduction to Distributed Computing

|  |           |
|--|-----------|
| <b>1. The Age of the Data Product.....</b>         | <b>3</b>  |
| What Is a Data Product?                            | 4         |
| Building Data Products at Scale with Hadoop        | 5         |
| Leveraging Large Datasets                          | 6         |
| Hadoop for Data Products                           | 7         |
| The Data Science Pipeline and the Hadoop Ecosystem | 8         |
| Big Data Workflows                                 | 10        |
| Conclusion   | 11        |
| <b>2. An Operating System for Big Data.....</b>    | <b>13</b> |
| Basic Concepts                                     | 14        |
| Hadoop Architecture                                | 15        |
| A Hadoop Cluster                                   | 17        |
| HDFS   | 20        |
| YARN   | 21        |
| Working with a Distributed File System             | 22        |
| Basic File System Operations                       | 23        |
| File Permissions in HDFS                           | 25        |
| Other HDFS Interfaces                              | 26        |
| Working with Distributed Computation               | 27        |
| MapReduce: A Functional Programming Model          | 28        |
| MapReduce: Implemented on a Cluster                | 30        |
| Beyond a Map and Reduce: Job Chaining              | 37        |

|  |           |
|--|-----------|
| Submitting a MapReduce Job to YARN                         | 38        |
| Conclusion   | 40        |
| <b>3. A Framework for Python and Hadoop Streaming.....</b> | <b>41</b> |
| Hadoop Streaming   | 42        |
| Computing on CSV Data with Streaming                       | 45        |
| Executing Streaming Jobs                                   | 50        |
| A Framework for MapReduce with Python                      | 52        |
| Counting Bigrams   | 55        |
| Other Frameworks   | 59        |
| Advanced MapReduce   | 60        |
| Combiners  | 60        |
| Partitioners   | 61        |
| Job Chaining   | 62        |
| Conclusion   | 65        |
| <b>4. In-Memory Computing with Spark.....</b>              | <b>67</b> |
| Spark Basics   | 68        |
| The Spark Stack  | 70        |
| Resilient Distributed Datasets                             | 72        |
| Programming with RDDs                                      | 73        |
| Interactive Spark Using PySpark                            | 77        |
| Writing Spark Applications                                 | 79        |
| Visualizing Airline Delays with Spark                      | 81        |
| Conclusion   | 87        |
| <b>5. Distributed Analysis and Patterns.....</b>           | <b>89</b> |
| Computing with Keys  | 91        |
| Compound Keys  | 92        |
| Keyspace Patterns  | 96        |
| Pairs versus Stripes                                       | 100       |
| Design Patterns  | 104       |
| Summarization  | 105       |
| Indexing   | 110       |
| Filtering  | 117       |
| Toward Last-Mile Analytics                                 | 123       |
| Fitting a Model  | 124       |
| Validating Models  | 125       |
| Conclusion   | 127       |

---

## Part II. Workflows and Tools for Big Data Science

|   |            |
|---|------------|
| <b>6. Data Mining and Warehousing.....</b>      | <b>131</b> |
| Structured Data Queries with Hive               | 132        |
| The Hive Command-Line Interface (CLI)           | 133        |
| Hive Query Language (HQL)                       | 134        |
| Data Analysis with Hive                         | 139        |
| HBase   | 144        |
| NoSQL and Column-Oriented Databases             | 145        |
| Real-Time Analytics with HBase                  | 148        |
| Conclusion                                      | 156        |
| <b>7. Data Ingestion.....</b>                   | <b>157</b> |
| Importing Relational Data with Sqoop            | 158        |
| Importing from MySQL to HDFS                    | 159        |
| Importing from MySQL to Hive                    | 161        |
| Importing from MySQL to HBase                   | 163        |
| Ingesting Streaming Data with Flume             | 165        |
| Flume Data Flows                                | 166        |
| Ingesting Product Impression Data with Flume    | 169        |
| Conclusion                                      | 173        |
| <b>8. Analytics with Higher-Level APIs.....</b> | <b>175</b> |
| Pig   | 175        |
| Pig Latin                                       | 177        |
| Data Types                                      | 181        |
| Relational Operators                            | 182        |
| User-Defined Functions                          | 182        |
| Wrapping Up                                     | 184        |
| Spark's Higher-Level APIs                       | 184        |
| Spark SQL                                       | 186        |
| DataFrames                                      | 189        |
| Conclusion                                      | 195        |
| <b>9. Machine Learning.....</b>                 | <b>197</b> |
| Scalable Machine Learning with Spark            | 197        |
| Collaborative Filtering                         | 199        |
| Classification                                  | 206        |
| Clustering                                      | 208        |
| Conclusion                                      | 212        |

|  |            |
|--|------------|
| <b>10. Summary: Doing Distributed Data Science. ....</b>                     | <b>213</b> |
| Data Product Lifecycle   | 214        |
| Data Lakes   | 216        |
| Data Ingestion   | 218        |
| Computational Data Stores  | 220        |
| Machine Learning Lifecycle   | 222        |
| Conclusion   | 224        |
| <b>A. Creating a Hadoop Pseudo-Distributed Development Environment. ....</b> | <b>227</b> |
| <b>B. Installing Hadoop Ecosystem Products. ....</b>                         | <b>237</b> |
| <b>Glossary. ....</b>  | <b>247</b> |
| <b>Index. ....</b>   | <b>263</b> |