

---

# Table of Contents

|                                |             |
|--------------------------------|-------------|
| <b>Foreword.....</b>           | <b>xi</b>   |
| <b>Preface.....</b>            | <b>xiii</b> |
| <b>1. Introduction.....</b>    | <b>1</b>    |
| Overview                       | 2           |
| Hadoop                         | 3           |
| Spark                          | 4           |
| R                              | 9           |
| sparklyr                       | 13          |
| Recap                          | 14          |
| <b>2. Getting Started.....</b> | <b>15</b>   |
| Overview                       | 15          |
| Prerequisites                  | 16          |
| Installing sparklyr            | 17          |
| Installing Spark               | 17          |
| Connecting                     | 18          |
| Using Spark                    | 19          |
| Web Interface                  | 20          |
| Analysis                       | 22          |
| Modeling                       | 23          |
| Data                           | 25          |
| Extensions                     | 25          |
| Distributed R                  | 26          |
| Streaming                      | 26          |
| Logs                           | 27          |
| Disconnecting                  | 28          |

|                               |           |
|-------------------------------|-----------|
| Using RStudio                 | 28        |
| Resources                     | 30        |
| Recap                         | 31        |
| <b>3. Analysis.....</b>       | <b>33</b> |
| Overview                      | 33        |
| Import                        | 36        |
| Wrangle                       | 37        |
| Built-in Functions            | 39        |
| Correlations                  | 40        |
| Visualize                     | 42        |
| Using ggplot2                 | 42        |
| Using dbplot                  | 44        |
| Model                         | 46        |
| Caching                       | 48        |
| Communicate                   | 49        |
| Recap                         | 52        |
| <b>4. Modeling.....</b>       | <b>53</b> |
| Overview                      | 54        |
| Exploratory Data Analysis     | 56        |
| Feature Engineering           | 63        |
| Supervised Learning           | 66        |
| Generalized Linear Regression | 70        |
| Other Models                  | 72        |
| Unsupervised Learning         | 72        |
| Data Preparation              | 73        |
| Topic Modeling                | 74        |
| Recap                         | 76        |
| <b>5. Pipelines.....</b>      | <b>77</b> |
| Overview                      | 77        |
| Creation                      | 79        |
| Use Cases                     | 80        |
| Hyperparameter Tuning         | 82        |
| Operating Modes               | 83        |
| Interoperability              | 84        |
| Deployment                    | 86        |
| Batch Scoring                 | 87        |
| Real-Time Scoring             | 88        |
| Recap                         | 90        |

|                            |            |
|----------------------------|------------|
| <b>6. Clusters.....</b>    | <b>93</b>  |
| Overview                   | 93         |
| On-Premises                | 95         |
| Managers                   | 96         |
| Distributions              | 101        |
| Cloud                      | 103        |
| Amazon                     | 104        |
| Databricks                 | 106        |
| Google                     | 107        |
| IBM                        | 108        |
| Microsoft                  | 109        |
| Qubole                     | 110        |
| Kubernetes                 | 111        |
| Tools                      | 112        |
| RStudio                    | 113        |
| Jupyter                    | 114        |
| Livy                       | 114        |
| Recap                      | 115        |
| <br>                       |            |
| <b>7. Connections.....</b> | <b>117</b> |
| Overview                   | 118        |
| Edge Nodes                 | 119        |
| Spark Home                 | 120        |
| Local                      | 120        |
| Standalone                 | 121        |
| YARN                       | 122        |
| YARN Client                | 122        |
| YARN Cluster               | 123        |
| Livy                       | 124        |
| Mesos                      | 126        |
| Kubernetes                 | 127        |
| Cloud                      | 128        |
| Batches                    | 128        |
| Tools                      | 129        |
| Multiple Connections       | 129        |
| Troubleshooting            | 130        |
| Logging                    | 130        |
| Spark Submit               | 130        |
| Windows                    | 132        |
| Recap                      | 132        |

|                       |            |
|-----------------------|------------|
| <b>8. Data.....</b>   | <b>135</b> |
| Overview              | 135        |
| Reading Data          | 137        |
| Paths                 | 137        |
| Schema                | 138        |
| Memory                | 140        |
| Columns               | 141        |
| Writing Data          | 141        |
| Copying Data          | 143        |
| File Formats          | 144        |
| CSV                   | 145        |
| JSON                  | 146        |
| Parquet               | 147        |
| Others                | 148        |
| File Systems          | 149        |
| Storage Systems       | 150        |
| Hive                  | 150        |
| Cassandra             | 151        |
| JDBC                  | 152        |
| Recap                 | 152        |
| <br>                  |            |
| <b>9. Tuning.....</b> | <b>155</b> |
| Overview              | 155        |
| Graph                 | 157        |
| Timeline              | 159        |
| Configuring           | 160        |
| Connect Settings      | 161        |
| Submit Settings       | 162        |
| Runtime Settings      | 163        |
| sparklyr Settings     | 164        |
| Partitioning          | 167        |
| Implicit Partitions   | 167        |
| Explicit Partitions   | 168        |
| Caching               | 169        |
| Checkpointing         | 170        |
| Memory                | 171        |
| Shuffling             | 172        |
| Serialization         | 172        |
| Configuration Files   | 172        |
| Recap                 | 173        |

|                               |            |
|-------------------------------|------------|
| <b>10. Extensions.....</b>    | <b>175</b> |
| Overview                      | 176        |
| H2O                           | 178        |
| Graphs                        | 182        |
| XGBoost                       | 187        |
| Deep Learning                 | 189        |
| Genomics                      | 192        |
| Spatial                       | 195        |
| Troubleshooting               | 197        |
| Recap                         | 197        |
| <b>11. Distributed R.....</b> | <b>199</b> |
| Overview                      | 200        |
| Use Cases                     | 202        |
| Custom Parsers                | 202        |
| Partitioned Modeling          | 204        |
| Grid Search                   | 205        |
| Web APIs                      | 206        |
| Simulations                   | 207        |
| Partitions                    | 209        |
| Grouping                      | 210        |
| Columns                       | 211        |
| Context                       | 212        |
| Functions                     | 213        |
| Packages                      | 214        |
| Cluster Requirements          | 215        |
| Installing R                  | 216        |
| Apache Arrow                  | 217        |
| Troubleshooting               | 219        |
| Worker Logs                   | 220        |
| Resolving Timeouts            | 221        |
| Inspecting Partitions         | 222        |
| Debugging Workers             | 223        |
| Recap                         | 223        |
| <b>12. Streaming.....</b>     | <b>225</b> |
| Overview                      | 225        |
| Transformations               | 228        |
| Analysis                      | 229        |
| Modeling                      | 230        |
| Pipelines                     | 231        |
| Distributed R                 | 232        |

|  |            |
|--|------------|
| Kafka  | 233        |
| Shiny  | 235        |
| Recap  | 237        |
| <b>13. Contributing</b> .....                | <b>239</b> |
| Overview                                     | 240        |
| The Spark API                                | 242        |
| Spark Extensions                             | 243        |
| Using Scala Code                             | 245        |
| Recap  | 247        |
| <b>A. Supplemental Code References</b> ..... | <b>249</b> |
| <b>Index</b> .....                           | <b>267</b> |