

# Table of Contents

<b>Foreword.....</b>	<b>ix</b>
<b>Preface.....</b>	<b>xi</b>
<b>1. Introduction to Data Analysis with Spark.....</b>	<b>1</b>
What Is Apache Spark?	1
A Unified Stack	2
Spark Core	3
Spark SQL	3
Spark Streaming	3
MLlib	4
GraphX	4
Cluster Managers	4
Who Uses Spark, and for What?	4
Data Science Tasks	5
Data Processing Applications	6
A Brief History of Spark	6
Spark Versions and Releases	7
Storage Layers for Spark	7
<b>2. Downloading Spark and Getting Started.....</b>	<b>9</b>
Downloading Spark	9
Introduction to Spark's Python and Scala Shells	11
Introduction to Core Spark Concepts	14
Standalone Applications	17
Initializing a SparkContext	17
Building Standalone Applications	18
Conclusion	21

<b>3. Programming with RDDs.....</b>	<b>23</b>
RDD Basics	23
Creating RDDs	25
RDD Operations	26
Transformations	27
Actions	28
Lazy Evaluation	29
Passing Functions to Spark	30
Python	30
Scala	31
Java	32
Common Transformations and Actions	34
Basic RDDs	34
Converting Between RDD Types	42
Persistence (Caching)	44
Conclusion	46
<b>4. Working with Key/Value Pairs.....</b>	<b>47</b>
Motivation	47
Creating Pair RDDs	48
Transformations on Pair RDDs	49
Aggregations	51
Grouping Data	57
Joins	58
Sorting Data	59
Actions Available on Pair RDDs	60
Data Partitioning (Advanced)	61
Determining an RDD's Partitioner	64
Operations That Benefit from Partitioning	65
Operations That Affect Partitioning	66
Example: PageRank	66
Custom Partitioners	68
Conclusion	70
<b>5. Loading and Saving Your Data.....</b>	<b>71</b>
Motivation	71
File Formats	72
Text Files	73
JSON	74
Comma-Separated Values and Tab-Separated Values	77
SequenceFiles	80
Object Files	83

Hadoop Input and Output Formats	83
File Compression	87
Filesystems	89
Local/“Regular” FS	89
Amazon S3	89
HDFS	90
Structured Data with Spark SQL	90
Apache Hive	91
JSON	92
Databases	92
Java Database Connectivity	93
Cassandra	94
HBase	96
Elasticsearch	96
Conclusion	98
<b>6. Advanced Spark Programming.....</b>	<b>99</b>
Introduction	99
Accumulators	100
Accumulators and Fault Tolerance	103
Custom Accumulators	103
Broadcast Variables	104
Optimizing Broadcasts	106
Working on a Per-Partition Basis	107
Piping to External Programs	109
Numeric RDD Operations	113
Conclusion	115
<b>7. Running on a Cluster.....</b>	<b>117</b>
Introduction	117
Spark Runtime Architecture	117
The Driver	118
Executors	119
Cluster Manager	119
Launching a Program	120
Summary	120
Deploying Applications with spark-submit	121
Packaging Your Code and Dependencies	123
A Java Spark Application Built with Maven	124
A Scala Spark Application Built with sbt	126
Dependency Conflicts	128
Scheduling Within and Between Spark Applications	128

Cluster Managers	129
Standalone Cluster Manager	129
Hadoop YARN	133
Apache Mesos	134
Amazon EC2	135
Which Cluster Manager to Use?	138
Conclusion	139
<b>8. Tuning and Debugging Spark.....</b>	<b>141</b>
Configuring Spark with SparkConf	141
Components of Execution: Jobs, Tasks, and Stages	145
Finding Information	150
Spark Web UI	150
Driver and Executor Logs	154
Key Performance Considerations	155
Level of Parallelism	155
Serialization Format	156
Memory Management	157
Hardware Provisioning	158
Conclusion	160
<b>9. Spark SQL.....</b>	<b>161</b>
Linking with Spark SQL	162
Using Spark SQL in Applications	164
Initializing Spark SQL	164
Basic Query Example	166
DataFrames	166
Caching	169
Loading and Saving Data	170
Apache Hive	171
Data Sources/Parquet	172
JSON	173
From RDDs	175
JDBC/ODBC Server	176
Working with Beeline	178
Long-Lived Tables and Queries	179
User-Defined Functions	179
Spark SQL UDFs	180
Hive UDFs	181
Spark SQL Performance	181
Performance Tuning Options	182
Conclusion	183

<b>10. Spark Streaming.....</b>	<b>185</b>
A Simple Example	186
Architecture and Abstraction	188
Transformations	191
Stateless Transformations	192
Stateful Transformations	194
Output Operations	199
Input Sources	201
Core Sources	201
Additional Sources	202
Multiple Sources and Cluster Sizing	207
24/7 Operation	208
Checkpointing	208
Driver Fault Tolerance	209
Worker Fault Tolerance	210
Receiver Fault Tolerance	210
Processing Guarantees	211
Streaming UI	212
Performance Considerations	212
Batch and Window Sizes	212
Level of Parallelism	213
Garbage Collection and Memory Usage	213
Conclusion	214
<b>11. Machine Learning with MLlib.....</b>	<b>215</b>
Overview	215
System Requirements	216
Machine Learning Basics	217
Example: Spam Classification	218
Data Types	220
Working with Vectors	221
Algorithms	223
Feature Extraction	223
Statistics	225
Classification and Regression	226
Clustering	232
Collaborative Filtering and Recommendation	233
Dimensionality Reduction	234
Model Evaluation	236
Tips and Performance Considerations	237
Preparing Features	237
Configuring Algorithms	237

Caching RDDs to Reuse	237
Recognizing Sparsity	238
Level of Parallelism	238
Pipeline API	238
Conclusion	240
<b>Index.....</b>	<b>241</b>