

Learning Spark

*Holden Karau, Andy Konwinski, Patrick Wendell, and
Matei Zaharia*

Beijing • Cambridge • Farnham • Ktiln • Sebastopol • Tokyo

O'REILLY

Table of Contents

Foreword	ix
Preface	xi
1. Introduction to Data Analysis with Spark	1
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
2. Downloading Spark and Getting Started	9
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

3. Programming with RDDs	23
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
4. Working with Key/Value Pairs	47
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	65
Example: PageRank	66
Custom Partitioners	68
Conclusion	70
5. Loading and Saving Your Data	71
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	84
File Compression	88
Filesystems	89
Local/"Regular" FS	89
Amazon S3	90
HDFS	90
Structured Data with Spark SQL	91
Apache Hive	91
JSON	92
Databases	93
Java Database Connectivity	93
Cassandra	94
HBase	96
Elasticsearch	97
Conclusion	98
6. Advanced Spark Programming	99
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
7. Running on a Cluster	117
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
8. Tuning and Debugging Spark	141
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
9. Spark SQL	161
Linking with Spark SQL	162
Using Spark SQL in Applications	164
Initializing Spark SQL	164
Basic Query Example	165
SchemaRDDs	166
Caching	169
Loading and Saving Data	170
Apache Hive	170
Parquet	171
JSON	172
From RDDs	174
JDBC/ODBC Server	175
Working with Beeline	177
Long-Lived Tables and Queries	178
User-Defined Functions	178
Spark SQL UDFs	178
Hive UDFs	179
Spark SQL Performance	180
Performance Tuning Options	180
Conclusion	182

10. Spark Streaming	183
A Simple Example	184
Architecture and Abstraction	186
Transformations	189
Stateless Transformations	190
Stateful Transformations	192
Output Operations	197
Input Sources	199
Core Sources	199
Additional Sources	200
Multiple Sources and Cluster Sizing	204
24/7 Operation	205
Checkpointing	205
Driver Fault Tolerance	206
Worker Fault Tolerance	207
Receiver Fault Tolerance	207
Processing Guarantees	208
Streaming UI	208
Performance Considerations	209
Batch and Window Sizes	209
Level of Parallelism	210
Garbage Collection and Memory Usage	210
Conclusion	211
11. Machine Learning with MLlib	213
Overview	213
System Requirements	214
Machine Learning Basics	215
Example: Spam Classification	216
Data Types	218
Working with Vectors	219
Algorithms	220
Feature Extraction	221
Statistics	223
Classification and Regression	224
Clustering	229
Collaborative Filtering and Recommendation	230
Dimensionality Reduction	232
Model Evaluation	234
Tips and Performance Considerations	234
Preparing Features	234
Configuring Algorithms	235

Caching RDDs to Reuse	235
Recognizing Sparsity	235
Level of Parallelism	236
Pipeline API	236
Conclusion	237
Index	239