

SECOND EDITION

Learning Spark

Lightning-Fast Data Analytics

*Jules S. Damji, Brooke Wenig,
Tathagata Das, and Denny Lee*

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Table of Contents

Foreword	xiii
Preface	xv
1. Introduction to Apache Spark: A Unified Analytics Engine	1
The Genesis of Spark	1
Big Data and Distributed Computing at Google	1
Hadoop at Yahoo!	2
Spark's Early Years at AMPLab	3
What Is Apache Spark?	4
Speed	4
Ease of Use	5
Modularity	5
Extensibility	5
Unified Analytics	6
Apache Spark Components as a Unified Stack	6
Apache Spark's Distributed Execution	10
The Developer's Experience	14
Who Uses Spark, and for What?	14
Community Adoption and Expansion	16
2. Downloading Apache Spark and Getting Started	19
Step 1: Downloading Apache Spark	19
Spark's Directories and Files	21
Step 2: Using the Scala or PySpark Shell	22
Using the Local Machine	23
Step 3: Understanding Spark Application Concepts	25
Spark Application and SparkSession	26

Spark Jobs	27
Spark Stages	28
Spark Tasks	28
Transformations, Actions, and Lazy Evaluation	28
Narrow and Wide Transformations	30
The Spark UI	31
Your First Standalone Application	34
Counting M&Ms for the Cookie Monster	35
Building Standalone Applications in Scala	40
Summary	42
3. Apache Spark's Structured APIs.....	43
Spark: What's Underneath an RDD?	43
Structuring Spark	44
Key Merits and Benefits	45
The DataFrame API	47
Spark's Basic Data Types	48
Spark's Structured and Complex Data Types	49
Schemas and Creating DataFrames	50
Columns and Expressions	54
Rows	57
Common DataFrame Operations	58
End-to-End DataFrame Example	68
The Dataset API	69
Typed Objects, Untyped Objects, and Generic Rows	69
Creating Datasets	71
Dataset Operations	72
End-to-End Dataset Example	74
DataFrames Versus Datasets	74
When to Use RDDs	75
Spark SQL and the Underlying Engine	76
The Catalyst Optimizer	77
Summary	82
4. Spark SQL and DataFrames: Introduction to Built-in Data Sources.....	83
Using Spark SQL in Spark Applications	84
Basic Query Examples	85
SQL Tables and Views	89
Managed Versus Unmanaged Tables	89
Creating SQL Databases and Tables	90
Creating Views	91
Viewing the Metadata	93

Caching SQL Tables	93
Reading Tables into DataFrames	93
Data Sources for DataFrames and SQL Tables	94
DataFrameReader	94
DataFrameWriter	96
Parquet	97
JSON	100
CSV	102
Avro	104
ORC	106
Images	108
Binary Files	110
Summary	111
5. Spark SQL and DataFrames: Interacting with External Data Sources.....	113
Spark SQL and Apache Hive	113
User-Defined Functions	114
Querying with the Spark SQL Shell, Beeline, and Tableau	119
Using the Spark SQL Shell	119
Working with Beeline	120
Working with Tableau	122
External Data Sources	129
JDBC and SQL Databases	129
PostgreSQL	132
MySQL	133
Azure Cosmos DB	134
MS SQL Server	136
Other External Sources	137
Higher-Order Functions in DataFrames and Spark SQL	138
Option 1: Explode and Collect	138
Option 2: User-Defined Function	138
Built-in Functions for Complex Data Types	139
Higher-Order Functions	141
Common DataFrames and Spark SQL Operations	144
Unions	147
Joins	148
Windowing	149
Modifications	151
Summary	155
6. Spark SQL and Datasets.....	157
Single API for Java and Scala	157

Scala Case Classes and JavaBeans for Datasets	158
Working with Datasets	160
Creating Sample Data	160
Transforming Sample Data	162
Memory Management for Datasets and DataFrames	167
Dataset Encoders	168
Spark's Internal Format Versus Java Object Format	168
Serialization and Deserialization (SerDe)	169
Costs of Using Datasets	170
Strategies to Mitigate Costs	170
Summary	172
7. Optimizing and Tuning Spark Applications.....	173
Optimizing and Tuning Spark for Efficiency	173
Viewing and Setting Apache Spark Configurations	173
Scaling Spark for Large Workloads	177
Caching and Persistence of Data	183
DataFrame.cache()	183
DataFrame.persist()	184
When to Cache and Persist	187
When Not to Cache and Persist	187
A Family of Spark Joins	187
Broadcast Hash Join	188
Shuffle Sort Merge Join	189
Inspecting the Spark UI	197
Journey Through the Spark UI Tabs	197
Summary	205
8. Structured Streaming.....	207
Evolution of the Apache Spark Stream Processing Engine	207
The Advent of Micro-Batch Stream Processing	208
Lessons Learned from Spark Streaming (DStreams)	209
The Philosophy of Structured Streaming	210
The Programming Model of Structured Streaming	211
The Fundamentals of a Structured Streaming Query	213
Five Steps to Define a Streaming Query	213
Under the Hood of an Active Streaming Query	219
Recovering from Failures with Exactly-Once Guarantees	221
Monitoring an Active Query	223
Streaming Data Sources and Sinks	226
Files	226
Apache Kafka	228

Custom Streaming Sources and Sinks	230
Data Transformations	234
Incremental Execution and Streaming State	234
Stateless Transformations	235
Stateful Transformations	235
Stateful Streaming Aggregations	238
Aggregations Not Based on Time	238
Aggregations with Event-Time Windows	239
Streaming Joins	246
Stream–Static Joins	246
Stream–Stream Joins	248
Arbitrary Stateful Computations	253
Modeling Arbitrary Stateful Operations with mapGroupsWithState()	254
Using Timeouts to Manage Inactive Groups	257
Generalization with flatMapGroupsWithState()	261
Performance Tuning	262
Summary	264
9. Building Reliable Data Lakes with Apache Spark	265
The Importance of an Optimal Storage Solution	265
Databases	266
A Brief Introduction to Databases	266
Reading from and Writing to Databases Using Apache Spark	267
Limitations of Databases	267
Data Lakes	268
A Brief Introduction to Data Lakes	268
Reading from and Writing to Data Lakes using Apache Spark	269
Limitations of Data Lakes	270
Lakehouses: The Next Step in the Evolution of Storage Solutions	271
Apache Hudi	272
Apache Iceberg	272
Delta Lake	273
Building Lakehouses with Apache Spark and Delta Lake	274
Configuring Apache Spark with Delta Lake	274
Loading Data into a Delta Lake Table	275
Loading Data Streams into a Delta Lake Table	277
Enforcing Schema on Write to Prevent Data Corruption	278
Evolving Schemas to Accommodate Changing Data	279
Transforming Existing Data	279
Auditing Data Changes with Operation History	282
Querying Previous Snapshots of a Table with Time Travel	283
Summary	284

10. Machine Learning with MLlib.....	285
What Is Machine Learning?	286
Supervised Learning	286
Unsupervised Learning	288
Why Spark for Machine Learning?	289
Designing Machine Learning Pipelines	289
Data Ingestion and Exploration	290
Creating Training and Test Data Sets	291
Preparing Features with Transformers	293
Understanding Linear Regression	294
Using Estimators to Build Models	295
Creating a Pipeline	296
Evaluating Models	302
Saving and Loading Models	306
Hyperparameter Tuning	307
Tree-Based Models	307
k-Fold Cross-Validation	316
Optimizing Pipelines	320
Summary	321
11. Managing, Deploying, and Scaling Machine Learning Pipelines with Apache Spark. .	323
Model Management	323
MLflow	324
Model Deployment Options with MLlib	330
Batch	332
Streaming	333
Model Export Patterns for Real-Time Inference	334
Leveraging Spark for Non-MLlib Models	336
Pandas UDFs	336
Spark for Distributed Hyperparameter Tuning	337
Summary	341
12. Epilogue: Apache Spark 3.0.....	343
Spark Core and Spark SQL	343
Dynamic Partition Pruning	343
Adaptive Query Execution	345
SQL Join Hints	348
Catalog Plugin API and DataSourceV2	349
Accelerator-Aware Scheduler	351
Structured Streaming	352
PySpark, Pandas UDFs, and Pandas Function APIs	354
Redesigned Pandas UDFs with Python Type Hints	354

Iterator Support in Pandas UDFs	355
New Pandas Function APIs	356
Changed Functionality	357
Languages Supported and Deprecated	357
Changes to the DataFrame and Dataset APIs	357
DataFrame and SQL Explain Commands	358
Summary	360
Index.....	361