Table of Contents

Preface. xxi

Part I. Gentle Overview of Big Data and Spark

1. What Is Apache Spark?. 3
Apache Spark's Philosophy 4
Context: The Big Data Problem 6
History of Spark 7
The Present and Future of Spark 8
Running Spark 9
Downloading Spark Locally 9
Launching Spark's Interactive Consoles 10
Running Spark in the Cloud 11
Data Used in This Book 11

2. A Gentle Introduction to Spark. 13
Spark's Basic Architecture 13
Spark Applications 14
Spark's Language APIs 15
Spark's APIs 16
Starting Spark 16
The SparkSession 17
DataFrames 17
Partitions 18
Transformations 19
Lazy Evaluation 20
Actions 20
Spark UI 21
An End-to-End Example 22

DataFrames and SQL 25
Conclusion 30

3. A Tour of Spark's Toolset. 31
Running Production Applications 32
Datasets: Type-Safe Structured APIs 33
Structured Streaming 34
Machine Learning and Advanced Analytics 38
Lower-Level APIs 43
SparkR 44
Spark's Ecosystem and Packages 45
Conclusion 45

Part II. Structured APIs—DataFrames, SQL, and Datasets

4. Structured API Overview. 49
DataFrames and Datasets 50
Schemas 50
Overview of Structured Spark Types 51
DataFrames Versus Datasets 51
Columns 52
Rows 52
Spark Types 52
Overview of Structured API Execution 56
Logical Planning 56
Physical Planning 57
Execution 58
Conclusion 58

5. Basic Structured Operations. 59
Schemas 60
Columns and Expressions 61
Columns 62
Expressions 63
Records and Rows 64
Creating Rows 65
DataFrame Transformations 66
Creating DataFrames 66
select and selectExpr 68
Converting to Spark Types (Literals) 71
Adding Columns 72

Renaming Columns 72
Reserved Characters and Keywords 73
Case Sensitivity 74
Removing Columns 74
Changing a Column's Type (cast) 74
Filtering Rows 74
Getting Unique Rows 76
Random Samples 76
Random Splits 77
Concatenating and Appending Rows (Union) 77
Sorting Rows 78
Limit 79
Repartition and Coalesce 80
Collecting Rows to the Driver 81
Conclusion 81

6. Working with Different Types of Data. 83
Where to Look for APIs 83
Converting to Spark Types 85
Working with Booleans 85
Working with Numbers 88
Working with Strings 92
Regular Expressions 93
Working with Dates and Timestamps 97
Working with Nulls in Data 102
Coalesce 102
ifnull, nullIf, nvl, and nvl2 102
drop 103
fill 103
replace 104
Ordering 104
Working with Complex Types 105
Structs 105
Arrays 105
split 106
Array Length 106
array_contains 107
explode 107
Maps 108
Working with JSON 109
User-Defined Functions 111
Conclusion 114

7. Aggregations.. 117
Aggregation Functions 119
count 119
countDistinct 120
approx_count_distinct 120
first and last 121
min and max 121
sum 121
sumDistinct 122
avg 122
Variance and Standard Deviation 123
skewness and kurtosis 124
Covariance and Correlation 124
Aggregating to Complex Types 125
Grouping 125
Grouping with Expressions 126
Grouping with Maps 127
Window Functions 127
Grouping Sets 130
Rollups 132
Cube 133
Grouping Metadata 134
Pivot 134
User-Defined Aggregation Functions 135
Conclusion 137

8. Joins.. 139
Join Expressions 139
Join Types 140
Inner Joins 141
Outer Joins 142
Left Outer Joins 143
Right Outer Joins 143
Left Semi Joins 144
Left Anti Joins 144
Natural Joins 145
Cross (Cartesian) Joins 145
Challenges When Using Joins 146
Joins on Complex Types 146
Handling Duplicate Column Names 147
How Spark Performs Joins 148
Communication Strategies 149

Conclusion 151

9. Data Sources. 153
The Structure of the Data Sources API 154
Read API Structure 154
Basics of Reading Data 154
Write API Structure 155
Basics of Writing Data 156
CSV Files 156
CSV Options 157
Reading CSV Files 158
Writing CSV Files 160
JSON Files 160
JSON Options 161
Reading JSON Files 162
Writing JSON Files 162
Parquet Files 163
Reading Parquet Files 163
Writing Parquet Files 164
ORC Files 164
Reading Orc Files 164
Writing Orc Files 165
SQL Databases 165
Reading from SQL Databases 167
Query Pushdown 169
Writing to SQL Databases 172
Text Files 173
Reading Text Files 173
Writing Text Files 173
Advanced I/O Concepts 174
Splittable File Types and Compression 174
Reading Data in Parallel 174
Writing Data in Parallel 174
Writing Complex Types 176
Managing File Size 176
Conclusion 177

10. Spark SQL. 179
What Is SQL? 179
Big Data and SQL: Apache Hive 180
Big Data and SQL: Spark SQL 180
Spark's Relationship to Hive 180

How to Run Spark SQL Queries 181
Spark SQL CLI 181
Spark's Programmatic SQL Interface 181
SparkSQL Thrift JDBC/ODBC Server 182
Catalog 183
Tables 184
Spark-Managed Tables 184
Creating Tables 184
Creating External Tables 186
Inserting into Tables 186
Describing Table Metadata 186
Refreshing Table Metadata 187
Dropping Tables 187
Caching Tables 188
Views 188
Creating Views 188
Dropping Views 189
Databases 189
Creating Databases 190
Setting the Database 190
Dropping Databases 190
Select Statements 190
case…when…then Statements 191
Advanced Topics 191
Complex Types 191
Functions 193
Subqueries 194
Miscellaneous Features 195
Configurations 195
Setting Configuration Values in SQL 196
Conclusion 196

11. Datasets. 197
When to Use Datasets 198
Creating Datasets 199
In Java: Encoders 199
In Scala: Case Classes 199
Actions 200
Transformations 200
Filtering 201
Mapping 202
Joins 202

Grouping and Aggregations 203
Conclusion 205

Part III. Low-Level APIs

12. Resilient Distributed Datasets (RDDs). 209
What Are the Low-Level APIs? 209
When to Use the Low-Level APIs? 209
How to Use the Low-Level APIs? 210
About RDDs 210
Types of RDDs 211
When to Use RDDs? 212
Datasets and RDDs of Case Classes 212
Creating RDDs 213
Interoperating Between DataFrames, Datasets, and RDDs 213
From a Local Collection 214
From Data Sources 214
Manipulating RDDs 215
Transformations 215
distinct 215
filter 215
map 216
sort 217
Random Splits 217
Actions 217
reduce 217
count 218
first 219
max and min 219
take 220
Saving Files 220
saveAsTextFile 220
SequenceFiles 221
Hadoop Files 221
Caching 221
Checkpointing 221
Pipe RDDs to System Commands 222
mapPartitions 222
foreachPartition 223
glom 224
Conclusion 224

13. **Advanced RDDs** **225**
Key-Value Basics (Key-Value RDDs) 226
keyBy 226
Mapping over Values 226
Extracting Keys and Values 227
lookup 227
sampleByKey 227
Aggregations 228
countByKey 229
Understanding Aggregation Implementations 229
Other Aggregation Methods 230
CoGroups 232
Joins 233
Inner Join 233
zips 233
Controlling Partitions 234
coalesce 234
repartition 234
repartitionAndSortWithinPartitions 235
Custom Partitioning 235
Custom Serialization 237
Conclusion 238

14. **Distributed Shared Variables** **239**
Broadcast Variables 239
Accumulators 241
Basic Example 242
Custom Accumulators 245
Conclusion 246

Part IV. Production Applications

15. **How Spark Runs on a Cluster** **249**
The Architecture of a Spark Application 249
Execution Modes 251
The Life Cycle of a Spark Application (Outside Spark) 252
Client Request 253
Launch 254
Execution 254
Completion 255
The Life Cycle of a Spark Application (Inside Spark) 255

The SparkSession 256
Logical Instructions 257
A Spark Job 258
Stages 259
Tasks 260
Execution Details 260
Pipelining 260
Shuffle Persistence 261
Conclusion 261

16. Developing Spark Applications. 263
Writing Spark Applications 263
A Simple Scala-Based App 263
Writing Python Applications 266
Writing Java Applications 267
Testing Spark Applications 268
Strategic Principles 268
Tactical Takeaways 269
Connecting to Unit Testing Frameworks 270
Connecting to Data Sources 270
The Development Process 271
Launching Applications 271
Application Launch Examples 273
Configuring Applications 274
The SparkConf 275
Application Properties 276
Runtime Properties 277
Execution Properties 277
Configuring Memory Management 278
Configuring Shuffle Behavior 278
Environmental Variables 278
Job Scheduling Within an Application 279
Conclusion 280

17. Deploying Spark. 281
Where to Deploy Your Cluster to Run Spark Applications 282
On-Premises Cluster Deployments 282
Spark in the Cloud 283
Cluster Managers 284
Standalone Mode 284
Spark on YARN 286
Configuring Spark on YARN Applications 287

Spark on Mesos 288
Secure Deployment Configurations 289
Cluster Networking Configurations 289
Application Scheduling 290
Miscellaneous Considerations 291
Conclusion 292

18. Monitoring and Debugging. 293
The Monitoring Landscape 293
What to Monitor 294
Driver and Executor Processes 295
Queries, Jobs, Stages, and Tasks 295
Spark Logs 295
The Spark UI 296
Spark REST API 303
Spark UI History Server 303
Debugging and Spark First Aid 304
Spark Jobs Not Starting 304
Errors Before Execution 305
Errors During Execution 305
Slow Tasks or Stragglers 306
Slow Aggregations 308
Slow Joins 309
Slow Reads and Writes 310
Driver OutOfMemoryError or Driver Unresponsive 310
Executor OutOfMemoryError or Executor Unresponsive 311
Unexpected Nulls in Results 312
No Space Left on Disk Errors 313
Serialization Errors 313
Conclusion 314

19. Performance Tuning. 315
Indirect Performance Enhancements 316
Design Choices 316
Object Serialization in RDDs 317
Cluster Configurations 318
Scheduling 318
Data at Rest 319
Shuffle Configurations 322
Memory Pressure and Garbage Collection 322
Direct Performance Enhancements 324
Parallelism 324

Improved Filtering 325
Repartitioning and Coalescing 325
User-Defined Functions (UDFs) 325
Temporary Data Storage (Caching) 326
Joins 328
Aggregations 329
Broadcast Variables 329
Conclusion 329

Part V. Streaming

20. Stream Processing Fundamentals. 333
What Is Stream Processing? 334
Stream Processing Use Cases 334
Advantages of Stream Processing 336
Challenges of Stream Processing 337
Stream Processing Design Points 338
Record-at-a-Time Versus Declarative APIs 338
Event Time Versus Processing Time 339
Continuous Versus Micro-Batch Execution 339
Spark's Streaming APIs 341
The DStream API 341
Structured Streaming 341
Conclusion 342

21. Structured Streaming Basics. 343
Structured Streaming Basics 343
Core Concepts 344
Transformations and Actions 345
Input Sources 345
Sinks 345
Output Modes 346
Triggers 346
Event-Time Processing 346
Structured Streaming in Action 347
Transformations on Streams 351
Selections and Filtering 351
Aggregations 352
Joins 353
Input and Output 353
Where Data Is Read and Written (Sources and Sinks) 354

Reading from the Kafka Source 355
Writing to the Kafka Sink 356
How Data Is Output (Output Modes) 360
When Data Is Output (Triggers) 361
Streaming Dataset API 362
Conclusion 363

22. Event-Time and Stateful Processing. 365
Event Time 365
Stateful Processing 367
Arbitrary Stateful Processing 367
Event-Time Basics 368
Windows on Event Time 369
Tumbling Windows 369
Handling Late Data with Watermarks 372
Dropping Duplicates in a Stream 375
Arbitrary Stateful Processing 376
Time-Outs 377
Output Modes 378
mapGroupsWithState 379
flatMapGroupsWithState 383
Conclusion 386

23. Structured Streaming in Production. 387
Fault Tolerance and Checkpointing 387
Updating Your Application 389
Updating Your Streaming Application Code 389
Updating Your Spark Version 389
Sizing and Rescaling Your Application 390
Metrics and Monitoring 390
Query Status 390
Recent Progress 391
Spark UI 393
Alerting 393
Advanced Monitoring with the Streaming Listener 393
Conclusion 395

Part VI. Advanced Analytics and Machine Learning

24. Advanced Analytics and Machine Learning Overview. 399
A Short Primer on Advanced Analytics 400

Supervised Learning 401
Recommendation 403
Unsupervised Learning 403
Graph Analytics 404
The Advanced Analytics Process 405
Spark's Advanced Analytics Toolkit 408
What Is MLlib? 408
High-Level MLlib Concepts 409
MLlib in Action 412
Feature Engineering with Transformers 413
Estimators 415
Pipelining Our Workflow 417
Training and Evaluation 419
Persisting and Applying Models 421
Deployment Patterns 422
Conclusion 423

25. Preprocessing and Feature Engineering. 425
Formatting Models According to Your Use Case 425
Transformers 427
Estimators for Preprocessing 428
Transformer Properties 429
High-Level Transformers 429
RFormula 430
SQL Transformers 431
VectorAssembler 432
Working with Continuous Features 433
Bucketing 433
Scaling and Normalization 435
StandardScaler 436
Working with Categorical Features 439
StringIndexer 439
Converting Indexed Values Back to Text 441
Indexing in Vectors 441
One-Hot Encoding 442
Text Data Transformers 443
Tokenizing Text 443
Removing Common Words 445
Creating Word Combinations 446
Converting Words into Numerical Representations 447
Word2Vec 450
Feature Manipulation 452

PCA 452
Interaction 453
Polynomial Expansion 453
Feature Selection 454
ChiSqSelector 454
Advanced Topics 455
Persisting Transformers 455
Writing a Custom Transformer 456
Conclusion 457

26. Classification. 459
Use Cases 459
Types of Classification 460
Binary Classification 460
Multiclass Classification 460
Multilabel Classification 460
Classification Models in MLlib 460
Model Scalability 461
Logistic Regression 462
Model Hyperparameters 462
Training Parameters 463
Prediction Parameters 463
Example 464
Model Summary 465
Decision Trees 465
Model Hyperparameters 466
Training Parameters 467
Prediction Parameters 467
Random Forest and Gradient-Boosted Trees 467
Model Hyperparameters 468
Training Parameters 469
Prediction Parameters 469
Naive Bayes 469
Model Hyperparameters 470
Training Parameters 470
Prediction Parameters 470
Evaluators for Classification and Automating Model Tuning 471
Detailed Evaluation Metrics 471
One-vs-Rest Classifier 472
Multilayer Perceptron 473
Conclusion 473

27. Regression. 475
Use Cases 475
Regression Models in MLlib 476
Model Scalability 476
Linear Regression 477
Model Hyperparameters 477
Training Parameters 477
Example 477
Training Summary 478
Generalized Linear Regression 478
Model Hyperparameters 479
Training Parameters 480
Prediction Parameters 480
Example 480
Training Summary 481
Decision Trees 481
Model Hyperparameters 481
Training Parameters 482
Example 482
Random Forests and Gradient-Boosted Trees 482
Model Hyperparameters 482
Training Parameters 482
Example 483
Advanced Methods 483
Survival Regression (Accelerated Failure Time) 483
Isotonic Regression 484
Evaluators and Automating Model Tuning 484
Metrics 485
Conclusion 486

28. Recommendation. 487
Use Cases 487
Collaborative Filtering with Alternating Least Squares 488
Model Hyperparameters 488
Training Parameters 489
Prediction Parameters 490
Example 490
Evaluators for Recommendation 492
Metrics 492
Regression Metrics 492
Ranking Metrics 493
Frequent Pattern Mining 494

Conclusion 495

29. **Unsupervised Learning. 497**
Use Cases 497
Model Scalability 498
k-means 499
Model Hyperparameters 499
Training Parameters 499
Example 500
k-means Metrics Summary 500
Bisecting k-means 501
Model Hyperparameters 501
Training Parameters 501
Example 501
Bisecting k-means Summary 502
Gaussian Mixture Models 502
Model Hyperparameters 503
Training Parameters 503
Example 503
Gaussian Mixture Model Summary 503
Latent Dirichlet Allocation 504
Model Hyperparameters 504
Training Parameters 505
Prediction Parameters 506
Example 506
Conclusion 507

30. **Graph Analytics. 509**
Building a Graph 511
Querying the Graph 513
Subgraphs 514
Motif Finding 514
Graph Algorithms 516
PageRank 516
In-Degree and Out-Degree Metrics 517
Breadth-First Search 519
Connected Components 520
Strongly Connected Components 522
Advanced Tasks 522
Conclusion 522

31. Deep Learning. 525
What Is Deep Learning? 525
Ways of Using Deep Learning in Spark 527
Deep Learning Libraries 528
MLlib Neural Network Support 528
TensorFrames 528
BigDL 529
TensorFlowOnSpark 529
DeepLearning4J 529
Deep Learning Pipelines 529
A Simple Example with Deep Learning Pipelines 530
Setup 531
Images and DataFrames 531
Transfer Learning 532
Applying Popular Models 533
Conclusion 534

Part VII. Ecosystem

32. Language Specifics: Python (PySpark) and R (SparkR and sparklyr). 539
PySpark 540
Fundamental PySpark Differences 540
Pandas Integration 540
R on Spark 541
SparkR 541
sparklyr 548
Conclusion 551

33. Ecosystem and Community. 553
Spark Packages 553
An Abridged List of Popular Packages 554
Using Spark Packages 554
External Packages 555
Community 556
Spark Summit 556
Local Meetups 556
Conclusion 557

Index. 559