
Table of Contents

Foreword.....	xiii
---------------	------

Preface.....	xv
--------------	----

Part I. Fundamentals

1. Introduction to Spark and PySpark.....	1
Why Spark for Data Analytics	2
The Spark Ecosystem	5
Spark Architecture	6
The Power of PySpark	12
PySpark Architecture	15
Spark Data Abstractions	17
RDD Examples	17
Spark RDD Operations	18
DataFrame Examples	21
Using the PySpark Shell	24
Launching the PySpark Shell	25
Creating an RDD from a Collection	26
Aggregating and Merging Values of Keys	26
Filtering an RDD's Elements	28
Grouping Similar Keys	28
Aggregating Values for Similar Keys	29
ETL Example with DataFrames	30
Extraction	31
Transformation	32
Loading	33

Summary	33
2. Transformations in Action.....	35
The DNA Base Count Example	36
The DNA Base Count Problem	38
FASTA Format	39
Sample Data	39
DNA Base Count Solution 1	40
Step 1: Create an RDD[String] from the Input	41
Step 2: Define a Mapper Function	42
Step 3: Find the Frequencies of DNA Letters	44
Pros and Cons of Solution 1	47
DNA Base Count Solution 2	47
Step 1: Create an RDD[String] from the Input	49
Step 2: Define a Mapper Function	49
Step 3: Find the Frequencies of DNA Letters	51
Pros and Cons of Solution 2	52
DNA Base Count Solution 3	52
The mapPartitions() Transformation	52
Step 1: Create an RDD[String] from the Input	60
Step 2: Define a Function to Handle a Partition	60
Step 3: Apply the Custom Function to Each Partition	62
Pros and Cons of Solution 3	64
Summary	64
3. Mapper Transformations.....	65
Data Abstractions and Mappers	65
What Are Transformations?	67
Lazy Transformations	72
The map() Transformation	73
DataFrame Mapper	78
The flatMap() Transformation	80
map() Versus flatMap()	85
Apply flatMap() to a DataFrame	86
The mapValues() Transformation	89
The flatMapValues() Transformation	90
The mapPartitions() Transformation	91
Handling Empty Partitions	95
Benefits and Drawbacks	98
DataFrames and mapPartitions() Transformation	99
Summary	102

4. Reductions in Spark.....	103
Creating Pair RDDs	104
Reduction Transformations	105
Spark's Reductions	108
Simple Warmup Example	110
Solving with reduceByKey()	111
Solving with groupByKey()	112
Solving with aggregateByKey()	112
Solving with combineByKey()	113
What Is a Monoid?	115
Monoid and Non-Monoid Examples	117
The Movie Problem	118
Input Dataset to Analyze	121
The aggregateByKey() Transformation	122
First Solution Using aggregateByKey()	124
Second Solution Using aggregateByKey()	127
Complete PySpark Solution Using groupByKey()	129
Complete PySpark Solution Using reduceByKey()	131
Complete PySpark Solution Using combineByKey()	134
The Shuffle Step in Reductions	137
Shuffle Step for groupByKey()	138
Shuffle Step for reduceByKey()	139
Summary	140

Part II. Working with Data

5. Partitioning Data.....	145
Introduction to Partitions	146
Partitions in Spark	146
Managing Partitions	150
Default Partitioning	151
Explicit Partitioning	152
Physical Partitioning for SQL Queries	153
Physical Partitioning of Data in Spark	156
Partition as Text Format	156
Partition as Parquet Format	157
How to Query Partitioned Data	158
Amazon Athena Example	158
Summary	160

6. Graph Algorithms.....	161
Introduction to Graphs	162
The GraphFrames API	164
How to Use GraphFrames	165
GraphFrames Functions and Attributes	168
GraphFrames Algorithms	169
Finding Triangles	169
Motif Finding	172
Real-World Applications	181
Gene Analysis	181
Social Recommendations	183
Facebook Circles	187
Connected Components	191
Analyzing Flight Data	193
Summary	202
7. Interacting with External Data Sources.....	203
Relational Databases	204
Reading from a Database	205
Writing a DataFrame to a Database	213
Reading Text Files	218
Reading and Writing CSV Files	220
Reading CSV Files	220
Writing CSV Files	224
Reading and Writing JSON Files	225
Reading JSON Files	226
Writing JSON Files	227
Reading from and Writing to Amazon S3	228
Reading from Amazon S3	229
Writing to Amazon S3	231
Reading and Writing Hadoop Files	232
Reading Hadoop Text Files	233
Writing Hadoop Text Files	236
Reading and Writing HDFS SequenceFiles	238
Reading and Writing Parquet Files	239
Writing Parquet Files	239
Reading Parquet Files	241
Reading and Writing Avro Files	242
Reading Avro Files	242
Writing Avro Files	242
Reading from and Writing to MS SQL Server	243
Writing to MS SQL Server	243

Reading from MS SQL Server	244
Reading Image Files	244
Creating a DataFrame from Images	244
Summary	246
8. Ranking Algorithms.....	247
Rank Product	248
Calculation of the Rank Product	249
Formalizing Rank Product	249
Rank Product Example	250
PySpark Solution	251
PageRank	257
PageRank's Iterative Computation	259
Custom PageRank in PySpark Using RDDs	261
Custom PageRank in PySpark Using an Adjacency Matrix	263
PageRank with GraphFrames	266
Summary	267

Part III. Data Design Patterns

9. Classic Data Design Patterns.....	271
Input-Map-Output	272
RDD Solution	273
DataFrame Solution	275
Flat Mapper functionality	277
Input-Filter-Output	278
RDD Solution	279
DataFrame Solution	280
DataFrame Filter	280
Input-Map-Reduce-Output	282
RDD Solution	282
DataFrame Solution	285
Input-Multiple-Maps-Reduce-Output	287
RDD Solution	288
DataFrame Solution	290
Input-Map-Combiner-Reduce-Output	291
Input-MapPartitions-Reduce-Output	294
Inverted Index	298
Problem Statement	298
Input	298
Output	299

PySpark Solution	299
Summary	302
10. Practical Data Design Patterns.....	303
In-Mapper Combining	304
Basic MapReduce Algorithm	305
In-Mapper Combining per Record	307
In-Mapper Combining per Partition	309
Top-10	312
Top-N Formalized	314
PySpark Solution	316
Finding the Bottom 10	318
MinMax	319
Solution 1: Classic MapReduce	319
Solution 2: Sorting	319
Solution 3: Spark's mapPartitions()	320
The Composite Pattern and Monoids	323
Monoids	324
Monoidal and Non-Monoidal Examples	328
Non-Monoid MapReduce Example	331
Monoid MapReduce Example	332
PySpark Implementation of Monoidal Mean	334
Functors and Monoids	336
Conclusion on Using Monoids	338
Binning	338
Sorting	342
Summary	342
11. Join Design Patterns.....	345
Introduction to the Join Operation	345
Join in MapReduce	348
Map Phase	348
Reducer Phase	349
Implementation in PySpark	350
Map-Side Join Using RDDs	351
Map-Side Join Using DataFrames	355
Step 1: Create Cache for Airports	357
Step 2: Create Cache for Airlines	357
Step 3: Create Facts Table	358
Step 4: Apply Map-Side Join	358
Efficient Joins Using Bloom Filters	359
Introduction to Bloom Filters	359

A Simple Bloom Filter Example	361
Bloom Filters in Python	362
Using Bloom Filters in PySpark	362
Summary	363
12. Feature Engineering in PySpark	365
Introduction to Feature Engineering	366
Adding New Features	368
Applying UDFs	369
Creating Pipelines	370
Binarizing Data	372
Imputation	373
Tokenization	375
Tokenizer	376
RegexTokenizer	376
Tokenization with a Pipeline	377
Standardization	377
Normalization	380
Scaling a Column Using a Pipeline	382
Using MinMaxScaler on Multiple Columns	383
Normalization Using Normalizer	384
String Indexing	385
Applying StringIndexer to a Single Column	385
Applying StringIndexer to Several Columns	386
Vector Assembly	386
Bucketing	387
Bucketizer	388
QuantileDiscretizer	389
Logarithm Transformation	390
One-Hot Encoding	391
TF-IDF	397
FeatureHasher	401
SQLTransformer	402
Summary	403
Index	405

