

Table of Contents

Preface.....	xi
1. Distributed Machine Learning Terminology and Concepts.....	1
The Stages of the Machine Learning Workflow.....	4
Tools and Technologies in the Machine Learning Pipeline.....	6
Distributed Computing Models.....	8
General-Purpose Models.....	8
Dedicated Distributed Computing Models.....	10
Introduction to Distributed Systems Architecture.....	11
Centralized Versus Decentralized Systems.....	11
Interaction Models.....	12
Communication in a Distributed Setting.....	13
Introduction to Ensemble Methods.....	14
High Versus Low Bias.....	15
Types of Ensemble Methods.....	15
Distributed Training Topologies.....	16
The Challenges of Distributed Machine Learning Systems.....	18
Performance.....	18
Resource Management.....	21
Fault Tolerance.....	22
Privacy.....	23
Portability.....	24
Setting Up Your Local Environment.....	24
Chapters 2–6 Tutorials Environment.....	25
Chapters 7–10 Tutorials Environment.....	27
Summary.....	28

2. Introduction to Spark and PySpark.....	29
Apache Spark Architecture	30
Intro to PySpark	32
Apache Spark Basics	33
Software Architecture	33
PySpark and Functional Programming	39
Executing PySpark Code	40
pandas DataFrames Versus Spark DataFrames	41
Scikit-Learn Versus MLlib	42
Summary	43
3. Managing the Machine Learning Experiment Lifecycle with MLflow.....	45
Machine Learning Lifecycle Management Requirements	46
What Is MLflow?	47
Software Components of the MLflow Platform	48
Users of the MLflow Platform	49
MLflow Components	50
MLflow Tracking	50
MLflow Projects	54
MLflow Models	54
MLflow Model Registry	55
Using MLflow at Scale	57
Summary	60
4. Data Ingestion, Preprocessing, and Descriptive Statistics.....	61
Data Ingestion with Spark	62
Working with Images	63
Working with Tabular Data	65
Preprocessing Data	66
Preprocessing Versus Processing	66
Why Preprocess the Data?	67
Data Structures	68
MLlib Data Types	68
Preprocessing with MLlib Transformers	70
Preprocessing Image Data	77
Save the Data and Avoid the Small Files Problem	80
Descriptive Statistics: Getting a Feel for the Data	81
Calculating Statistics	82
Descriptive Statistics with Spark Summarizer	83
Data Skewness	85
Correlation	86
Summary	90

5. Feature Engineering.....	91
Features and Their Impact on Models	93
MLlib Featurization Tools	96
Extractors	96
Selectors	97
Example: Word2Vec	98
The Image Featurization Process	99
Understanding Image Manipulation	101
Extracting Features with Spark APIs	103
The Text Featurization Process	109
Bag-of-Words	110
TF-IDF	110
N-Gram	111
Additional Techniques	112
Enriching the Dataset	112
Summary	113
6. Training Models with Spark MLlib.....	115
Algorithms	116
Supervised Machine Learning	117
Classification	117
Regression	122
Unsupervised Machine Learning	127
Frequent Pattern Mining	127
Clustering	127
Evaluating	131
Supervised Evaluators	131
Unsupervised Evaluators	134
Hyperparameters and Tuning Experiments	135
Building a Parameter Grid	135
Splitting the Data into Training and Test Sets	135
Cross-Validation: A Better Way to Test Your Models	137
Machine Learning Pipelines	138
Constructing a Pipeline	140
How Does Splitting Work with the Pipeline API?	141
Persistence	141
Summary	142
7. Bridging Spark and Deep Learning Frameworks.....	143
The Two Clusters Approach	147
Implementing a Dedicated Data Access Layer	148
Features of a DAL	148

Selecting a DAL	150
What Is Petastorm?	151
SparkDatasetConverter	152
Petastorm as a Parquet Store	156
Project Hydrogen	158
Barrier Execution Mode	158
Accelerator-Aware Scheduling	160
A Brief Introduction to the Horovod Estimator API	161
Summary	162
8. TensorFlow Distributed Machine Learning Approach.....	163
A Quick Overview of TensorFlow	164
What Is a Neural Network?	166
TensorFlow Cluster Process Roles and Responsibilities	168
Loading Parquet Data into a TensorFlow Dataset	169
An Inside Look at TensorFlow's Distributed Machine Learning Strategies	171
ParameterServerStrategy	173
CentralStorageStrategy: One Machine, Multiple Processors	175
MirroredStrategy: One Machine, Multiple Processors, Local Copy	175
MultiWorkerMirroredStrategy: Multiple Machines, Synchronous	177
TPUStrategy	181
What Things Change When You Switch Strategies?	181
Training APIs	182
Keras API	182
Custom Training Loop	186
Estimator API	188
Putting It All Together	189
Troubleshooting	191
Summary	192
9. PyTorch Distributed Machine Learning Approach.....	193
A Quick Overview of PyTorch Basics	194
Computation Graph	194
PyTorch Mechanics and Concepts	196
PyTorch Distributed Strategies for Training Models	200
Introduction to PyTorch's Distributed Approach	201
Distributed Data-Parallel Training	202
RPC-Based Distributed Training	203
Communication Topologies in PyTorch (c10d)	212
What Can We Do with PyTorch's Low-Level APIs?	220
Loading Data with PyTorch and Petastorm	221

Troubleshooting Guidance for Working with Petastorm and Distributed PyTorch	224
The Enigma of Mismatched Data Types	224
The Mystery of Straggling Workers	226
How Does PyTorch Differ from TensorFlow?	227
Summary	228
10. Deployment Patterns for Machine Learning Models.....	229
Deployment Patterns	230
Pattern 1: Batch Prediction	230
Pattern 2: Model-in-Service	231
Pattern 3: Model-as-a-Service	232
Determining Which Pattern to Use	233
Production Software Requirements	234
Monitoring Machine Learning Models in Production	238
Data Drift	239
Model Drift, Concept Drift	242
Distributional Domain Shift (the Long Tail)	242
What Metrics Should I Monitor in Production?	243
How Do I Measure Changes Using My Monitoring System?	244
What It Looks Like in Production	245
The Production Feedback Loop	246
Deploying with MLlib	247
Production Machine Learning Pipelines with Structured Streaming	248
Deploying with MLflow	249
Defining an MLflow Wrapper	250
Deploying the Model as a Microservice	253
Loading the Model as a Spark UDF	254
How to Develop Your System Iteratively	254
Summary	256
Index.....	257