

**Practical Machine Learning with Apache Spark**

**Course Number:** PYTH-250WA  
**Duration:** 3 days

**Overview**

This Practical Machine Learning with Apache Spark training course teaches attendees how to integrate Python's capabilities for scaling data with machine learning (ML) on the Apache Spark platform. In addition, attendees learn the terminology, concepts, and algorithms used in ML.

**Prerequisites**

All attendees must have basic knowledge of statistics and programming.

**Materials**

All Machine Learning training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Windows, Mac, or Linux with at least 8 GB RAM
  + Most class activities will create Spark code and visualizations in a browser-based notebook environment. The class also details how to export these notebooks and how to run code outside of this environment.
* A current version of Anaconda for Python 3.x
* Related lab files that Accelebrate will provide
* Internet access

**Objectives**

* Understand the elements of Functional Programming with Python
* Use the Spark Shell
* Use the spark-submit Tool
* Understand the DataFrame Object
* Transform data with PySpark
* Switch to PySpark Jupyter Notebooks
* Use matplotlib for data visualization
* Work with descriptive statistics and EDA
* Use PySpark for data repair and normalization
* Understand linear regression
* Work with logistic regression
* Perform classification with Naive Bayes
* Work with Random Forest Classification
* Support Vector Machine Classification
* Use kMeans Algorithm

**Outline**

* Introduction
* Defining Data Science
  + Data Science, Machine Learning, AI?
  + The Data-Related Roles
  + Data Science Ecosystem
  + Business Analytics vs. Data Science
  + Who is a Data Scientist?
  + The Break-Down of Data Science Project Activities
  + Data Scientists at Work
  + The Data Engineer Role
  + What is Data Wrangling (Munging)?
  + Examples of Data Science Projects
  + Data Science Gotchas
* Machine Learning Life-cycle Phases
  + Data Analytics Pipeline
  + Data Discovery Phase
  + Data Harvesting Phase
  + Data Priming Phase
  + Data Cleansing
  + Feature Engineering
  + Data Logistics and Data Governance
  + Exploratory Data Analysis
  + Model Planning Phase
  + Model Building Phase
  + Communicating the Results
  + Production Roll-out
* Quick Introduction to Python Programming
  + Module Overview
  + Some Basic Facts about Python
  + Dynamic Typing Examples
  + Code Blocks and Indentation
  + Importing Modules
  + Lists and Tuples
  + Dictionaries
  + List Comprehension
  + What is Functional Programming (FP)?
  + Terminology: Higher-Order Functions
  + A Short List of Languages that Support FP
  + Lambda
  + Common High-Order Functions in Python 3
* Introduction to Apache Spark
  + What is Apache Spark
  + Where to Get Spark?
  + The Spark Platform
  + Spark Logo
  + Common Spark Use Cases
  + Languages Supported by Spark
  + Running Spark on a Cluster
  + The Driver Process
  + Spark Applications
  + Spark Shell
  + The spark-submit Tool
  + The spark-submit Tool Configuration
  + The Executor and Worker Processes
  + The Spark Application Architecture
  + Interfaces with Data Storage Systems
  + Limitations of Hadoop's MapReduce
  + Spark vs. MapReduce
  + Spark as an Alternative to Apache Tez
  + The Resilient Distributed Dataset (RDD)
  + Datasets and DataFrames
  + Spark SQL
  + Spark Machine Learning Library
  + GraphX
* The Spark Shell
  + The Spark Shell
  + The Spark v.2 + Shells
  + The Spark Shell UI
  + Spark Shell Options
  + Getting Help
  + The Spark Context (sc) and Spark Session (spark)
  + The Shell Spark Context Object (sc)
  + The Shell Spark Session Object (spark)
  + Loading Files
  + Saving Files
* Quick Intro to Jupyter Notebooks
  + Python Dev Tools and REPLs
  + IPython
  + Jupyter
  + Jupyter Operation Modes
  + Basic Edit Mode Shortcuts
  + Basic Command Mode Shortcuts
* Data Visualization in Python using matplotlib
  + Data Visualization
  + What is matplotlib?
  + Getting Started with matplotlib
  + The matplotlib.pyplot.plot() Function
  + The matplotlib.pyplot.scatter() Function
  + Labels and Titles
  + Styles
  + The matplotlib.pyplot.bar() Function
  + The matplotlib.pyplot.hist () Function
  + The matplotlib.pyplot.pie () Function
  + The Figure Object
  + The matplotlib.pyplot.subplot() Function
  + Selecting a Grid Cell
  + Saving Figures to a File
* Data Science and ML Algorithms with PySpark
  + In-Class Discussion
  + Types of Machine Learning
  + Supervised vs Unsupervised Machine Learning
  + Supervised Machine Learning Algorithms
  + Classification (Supervised ML) Examples
  + Unsupervised Machine Learning Algorithms
  + Clustering (Unsupervised ML) Examples
  + Choosing the Right Algorithm
  + Terminology: Observations, Features, and Targets
  + Representing Observations
  + Terminology: Labels
  + Terminology: Continuous and Categorical Features
  + Continuous Features
  + Categorical Features
  + Common Distance Metrics
  + The Euclidean Distance
  + What is a Model
  + Model Evaluation
  + The Classification Error Rate
  + Data Split for Training and Test Data Sets
  + Data Splitting in PySpark
  + Hold-Out Data
  + Cross-Validation Technique
  + Spark ML Overview
  + DataFrame-based API is the Primary Spark ML API
  + Estimators, Models, and Predictors
  + Descriptive Statistics
  + Data Visualization and EDA
  + Correlations
  + Feature Engineering
  + Scaling of the Features
  + Feature Blending (Creating Synthetic Features)
  + The 'One-Hot' Encoding Scheme
  + Example of 'One-Hot' Encoding Scheme
  + Bias-Variance (Underfitting vs Overfitting) Trade-off
  + The Modeling Error Factors
  + One Way to Visualize Bias and Variance
  + Underfitting vs Overfitting Visualization
  + Balancing Off the Bias-Variance Ratio
  + Linear Model Regularization
  + ML Model Tuning Visually
  + Linear Model Regularization in Spark
  + Regularization, Take Two
  + Dimensionality Reduction
  + PCA and isomap
  + The Advantages of Dimensionality Reduction
  + Spark Dense and Sparse Vectors
  + Labeled Point
  + Python Example of Using the LabeledPoint Class
  + The LIBSVM format
  + LIBSVM in PySpark
  + Example of Reading a File In LIBSVM Format
  + Life-cycles of Machine Learning Development
  + Regression Analysis
  + Regression vs Correlation
  + Regression vs Classification
  + Simple Linear Regression Model
  + Linear Regression Illustration
  + Least-Squares Method (LSM)
  + Gradient Descent Optimization
  + Locally Weighted Linear Regression
  + Regression Models in Excel
  + Multiple Regression Analysis
  + Evaluating Regression Model Accuracy
  + The R2 Model Score
  + The MSE Model Score
  + Linear Logistic (Logit) Regression
  + Interpreting Logistic Regression Results
  + Hands-on Exercise
  + Naive Bayes Classifier (SL)
  + Naive Bayesian Probabilistic Model in a Nutshell
  + Bayes Formula
  + Classification of Documents with Naive Bayes
  + Decision Trees
  + Decision Tree Terminology
  + Properties of Decision Trees
  + Decision Tree Classification in the Context of Information Theory
  + The Simplified Decision Tree Algorithm
  + Using Decision Trees
  + Random Forests
  + Support Vector Machines (SVMs)
  + Unsupervised Learning Type: Clustering
  + k-Means Clustering (UL)
  + k-Means Clustering in a Nutshell
  + k-Means Characteristics
  + Global vs. Local Minimum Explained
  + Time-Series Analysis
  + Decomposing Time-Series
  + A Better Algorithm or More Data?
* Conclusion