Databricks - Apache Spark for Machine Learning and Data Science

This 3-day course is primarily for data scientists but is directly applicable to analysts, architects, software engineers, and technical managers interested in a thorough, hands-on overview of Apache Spark and its applications to Machine Learning.

The course covers the fundamentals of Apache Spark including Spark’s architecture and internals, the core APIs for using Spark, SQL and other high-level data access tools, Spark’s streaming capabilities and a heavy focus on Spark’s machine learning APIs. The class is a mixture of lecture and hands-on labs.

Each topic includes lecture content along with hands-on labs in the Databricks notebook environment. Students may keep the notebooks and continue to use them after the class ends; all examples are guaranteed to run in the environment the class was taught on (Azure Databricks or CE)

Skills Gained
- Understand when and where to use Spark
- Articulate the difference between an RDD, DataFrame, and Dataset
- Explain supervised vs unsupervised machine learning, and typical applications of both
- Build a Machine Learning Pipeline using a combination of Transformers and Estimators: Save/Restore Models and Apply models to streaming data
- Perform hyperparameter tuning with cross-validation
- Analyze Spark query performance using the Spark UI
- Train models with 3rd party libraries such as XGBoost
- Perform hyperparameter search in parallel using single node algorithms such as scikit-learn
- Gain familiarity with Decision Trees, Random Forests, Gradient Boosted Trees, Linear Regression, Collaborative Filtering, and K-Means
- Explain options for putting models into production

Who Can Benefit
Data scientists, analysts, architects, software engineers, and technical managers with experience in machine learning who want to adapt traditional machine learning tasks to run at scale using Apache Spark.

Prerequisites
- Some familiarity with Apache Spark is helpful but not required.
- Some familiarity with Machine Learning and Data Science concepts are highly recommended but not required.
- Basic programming experience in an object-oriented or functional language is required. The class can be taught concurrently in Python and Scala.
Course Details

Lab Requirements
- A computer or laptop
- Chrome or Firefox Web Browser Internet explorer and Safari are not supported
- Internet access with unfettered connections to the following domains:
  - 1. *.databricks.com - required
  - 2. *.slack.com - highly recommended
  - 3. spark.apache.org - required
  - 4. drive.google.com - helpful but not required

Topics
Spark Overview

In-depth discussion of Spark SQL and DataFrames, including:
- RDD vs DataFrame vs Dataset API
- Spark SQL
- Data Aggregation
- Column Operations
- The Functions API: date/time, string manipulation, aggregation
- Caching and caching storage levels
- Use of the Spark UI to analyze behavior and performance

Overview of Spark internals
- Cluster Architecture
- How Spark schedules and executes jobs and tasks
- Shuffling, shuffle files, and performance
- The Catalyst query optimizer

Spark Structured Streaming
- Sources and sinks
- Structured Streaming APIs
- Windowing & Aggregation
- Checkpointing & Watermarking
- Reliability and Fault Tolerance

In-depth overview of Spark’s MLlib Pipeline API for Machine Learning
- Build machine learning pipelines for both supervised and unsupervised learning
- Transformer/Estimator/Pipeline API
- Use transformers to perform pre-processing on a dataset prior to training
- Train analytical models with Spark ML’s DataFrame-based estimators including Decision Trees, Random Forests, Gradient
Boosted Trees, Linear Regression, K-Means, and Alternating Least Squares

- Tune hyperparameters via cross-validation and grid search
- Evaluate model performance

MLflow

- Track and benchmark model performance

3rd Party Library Integrations

- XGBoost
  - How to distribute single-node algorithms (like scikit-learn) with Spark
  - Spark-Sklearn: Perform scikit-learn hyperparameter search in parallel

Production Discussion