Apache Spark for Machine Learning and Data Science (DB 301)

COURSE OVERVIEW:

This 2.5-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 with the free Databricks Community Edition offering after the class ends; all examples are guaranteed to run in that environment.

WHO WILL BENEFIT FROM THIS COURSE?

This course is designed for 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 OBJECTIVES:

After completion of this course, students will be able to:

- Use the core Spark APIs to operate on data
- Articulate and implement typical use cases for Spark
- Build data pipelines and query large data sets using Spark SQL and DataFrames
- Analyze Spark jobs using the administration UIs inside Databricks
- Create Structured Streaming jobs
- Understand the basics of Spark's internals
- Work with relational data using the GraphFrames APIs
- Understand how a Machine Learning pipeline works
- Use various ML algorithms to perform clustering, regression and classification tasks.
- Train & export ML models
- How to train models with 3rd-party libraries like scikit-learn

- Create and transform DataFrames to query large datasets.
- Improve performance through judicious use of caching and applying best practices.
- Visualize how jobs are broken into stages and tasks and executed within Spark.
- Troubleshoot errors and program crashes using Spark UI, executor logs, driver stack traces, and local-mode runtimes.
- Find answers to common Spark and Databricks questions using the documentation and other resources.

COURSE OUTLINE:

Module 1: Spark Overview

Lecture

- Databricks Overview
- Spark Capabilities
- Spark Ecosystem
- Basic Spark Components

Hands-On

- Databricks Lab Environment
- Working with Notebooks
- Spark Clusters and Files

Module 2: Spark SQL and DataFrames

Lecture

- Use of Spark SQL
- Use of DataFrames / DataSets
- Reading & Writing Data
- DataFrame, DataSet and SQL APIs
- Catalyst Query Optimization
- Tungsten
- ETL

Hands-On

- Creating DataFrames
- Querying with DataFrames
- Querying with SQL
- ETL with DataFrames
- Caching
- Visualization

Module 3: Spark Internals

Lecture

- Jobs, Stages, and Tasks
- Partitions and Shuffling
- Job Performance

Hands-On

- Visualizing SQL Queries
- Observing Task Execution
- Understanding Performance
- Measuring Memory Use

Module 4: Machine Learning

Lecture

- Spark MLlib Pipeline API
- Built-in Featurizing and Algorithms
- Cross-Validation and Grid Search for Hyperparameter Tuning
- Evaluation Metrics
- Data Partitioning Strategies
- Spark integration with Scikit-learn

Hands-On

- NLP/Text Classification with Logistic Regression
- Decision Tree vs. Random Forest
- Data imputation with Alternating Least Squares
- Clustering with K-Means
- Neural Networks
- Spark-sklearn

Module 5: Structured Streaming

Lecture

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

Hands-On

- Reading from TCP
- Continuous Visualization

Module 6: Graph Processing with GraphFrames

Lecture

- Basic Graph Analysis
- GraphFrames API

Hands-On

- GraphFrames ETL
- Pagerank and Label Propagation with GraphFrames

SUNSET LEARNING INSTITUTE (SLI) DIFFERENTIATORS:

Sunset Learning Institute (SLI) has been an innovative leader in developing and delivering authorized technical training since 1996. Our goal is to help our customers optimize their cloud technology investments by providing convenient, high quality technical training that our customers can rely on. We empower students to master their desired technologies for their unique environments.

What sets SLI apart is not only our immense selection of trainings options, but our convenient and consistent delivery system. No matter how complex your environment is or where you are located, SLI is sure to have a training solution that you can count on!

Premiere World Class Instruction Team

- All SLI instructors have a four-year technical degree, instructor level certifications and field consulting work experience.
- Sunset Learning has won numerous Instructor Excellence and Instructor Quality Distinction awards since 2012

Enhanced Learning Experience

- The goal of our instructors during class is ensure students understand the material, guide them through our labs and encourage questions and interactive discussions.

Convenient and Reliable Training Experience

- You have the option to attend classes at any of our established training facilities or from the convenience of your home or office with the use of our HD-ILT network (High Definition Instructor Led Training)
- All Sunset Learning Institute classes are guaranteed to run – you can count on us to deliver the training you need when you need it!

Outstanding Customer Service

- Dedicated account manager to suggest the optimal learning path for you and your team
- Enthusiastic Student Services team available to answer any questions and ensure a quality training experience