Apache Spark for Machine Learning & Data Science

Overview:

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.

Learning Objectives:

After taking this class, 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

Topics

- ✓ Spark overviewIn-depth discussion of Spark SQL and DataFrames, including:

  - The DataFrames / Datasets API
  - Spark SQL
  - Data Aggregation
  - Column Operations
  - The Functions API: dat / time, string manipulation, aggregation
  - Caching and caching storage levels
  - Use of the Spark UI to analyze behaviour and performance

Apache Spark for Machine Learning & Data Science

✓ Overview of Spark internals

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

✓ An 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 Linear Regression, Logistic Regression, Decision Trees + Random Forests, Boosted Trees, K-Means, Alternating Least Squares, and Neural Nets
- Tunehyperparameters via cross-validation and grid search
- Evaluate model performance

✓ Spark-sklearn

- How to distribute single-node algorithms (like scikit-learn) with Spark
- Partitioning data concerns

✓ Spark Structured Streaming

- Sources and sinks
- Structured Streaming APIs
- Windowing & Aggregation
- Checkpointing & Watermaking
- Reliability andFault Tolerance

✓ Graph processing with GraphFrames

- Transforming DataFrames into a graph
- Perform graph analysis, including Label Propagation, PageRank, and ShortestPaths

Target Audience

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.

Lab Requirements

✓ A computer or laptop
✓ Chrome or Firefox web browser – Internet Explorer and Safari are not supported
✓ Internet access

Course Outline

Module: Spark Overview

Lecture:

✓ Overview of Databricks
✓ Spark Capabilities
✓ Spark Ecosystem
✓ Basic Spark Components

Hands-On:

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

Apache Spark for Machine Learning & Data Science

Module: 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: 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: 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: 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: Graph Processing with GraphFrames

Lecture:

✓ Basic Graph Analysis
✓ GraphFrames API

Hands-On:

✓ GraphFrames ETL
✓ Pagerank and Label Propagation with GraphFrames

Reading homework is required the first and second evenings of the course.

Contact: info@unicom.co.uk | + 44 (0) 1895 256 484 | www.unicom.co.uk

Contact Us

UNICOM

www.unicom.co.uk
@UNICOMSeminars
info@unicom.co.uk

+44 1895 256 484 (UK)
www.youtube.com/unicomseminars
www.linkedin.com/UNICOMSeminars