This 1-day course is for data engineers, analysts, architects, data scientist, software engineers, IT operations, and technical managers interested in a brief hands-on overview of Apache Spark.

The course provides an introduction to the Spark architecture, some of the core APIs for using Spark, SQL and other high-level data access tools, as well as Spark’s streaming capabilities and 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.

Skills Gained
After taking this class, students will be able to:

- Use a subset of the core Spark APIs to operate on data.
- Articulate and implement simple use cases for Spark
- Build data pipelines and query large data sets using Spark SQL and DataFrames
- Create Structured Streaming jobs
- Understand how a Machine Learning pipeline works
- Understand the basics of Spark’s internals

Who Can Benefit
Data engineers, analysts, architects, data scientist, software engineers, and technical managers who want a quick introduction into how to use Apache Spark to streamline their big data processing, build production Spark jobs, and understand and debug running Spark applications.

Prerequisites
Some familiarity with Apache Spark is helpful but not required. Knowledge of SQL is helpful. Basic programming experience in an object-oriented or functional language is highly recommended but not required. The class can be taught concurrently in Python and Scala.

Course Details

Topics
Spark Overview

Introduction to Spark SQL and DataFrames, including:

- Reading & Writing Data
- The DataFrames/Datasets API
- Spark SQL
- Caching and caching storage levels

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

Overview of Spark’s MLlib Pipeline API for Machine Learning

- Transformer/Estimator/Pipeline API
- Perform feature preprocessing
- Evaluate and apply ML models