About

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

The course covers the core APIs for using Spark, fundamental mechanisms and basic internals of the framework, SQL and other high-level data access tools, as well as Spark's streaming capabilities and machine learning APIs.

Each topic includes slide and lecture content along with hands-on use of Spark through an elegant web-based notebook environment. Inspired by tools like IPython/Jupyter, notebooks allow attendees to code jobs, data analysis queries, and visualizations using their own Spark cluster, accessed through a web browser. All class code is directly usable with pure open-source Spark or any commercial Spark distribution.

Objectives

After taking this class you will be able to:

• Describe Spark’s fundamental mechanics
• Use the core Spark APIs to operate on data
• Articulate and implement typical use cases for Spark
• Build data pipelines with SparkSQL and DataFrames
• Analyze Spark jobs using the UIs and logs
• Create Streaming and Machine Learning jobs

Modules

• Spark Overview
• RDD Fundamentals
• SparkSQL and DataFrames
• Spark Job Execution
• Cluster Architectures for Spark
• Intro to Spark Streaming
• Machine Learning Basics

Details

DURATION: 3 Days
HOURS: 9:00 a.m. – 5:00 p.m.

Target Audience

• Data Analysts
• Software Developers

Prerequisites

• Basic Python or Scala required
• SQL is helpful but not required

Lab Requirements

• Chrome Browser
  Internet Explorer and Safari are not supported
• Internet (Web Access)
## Course Syllabus

### MODULE 1
**Spark Overview**

**LECTURE**
- Spark Opportunity and Solution
- Capabilities and Ecosystem
- Spark Components vs Hadoop

**HANDS-ON**
- Databricks Lab Environment
- Working with Notebooks
- Spark Clusters and Files

### MODULE 2
**RDD Fundamentals**

**LECTURE**
- Purpose and Structure of RDDs
- Transformations, Actions, and DAG
- RDD programming API

**HANDS-ON**
- Creating RDDs from Data Files
- Reshaping Data to Add Structure
- Interactive Queries Using RDDs

### MODULE 3
**Spark SQL / Dataframes**

**LECTURE**
- Spark SQL and DataFrame Uses
- DataFrame / SQL APIs
- Catalyst Query Optimization

**HANDS-ON**
- Creating DataFrames
- Query with DataFrame API and SQL
- Caching and Re-using DataFrames
- Generating Graphics and Reports

### MODULE 4
**Spark Job Execution**

**LECTURE**
- Jobs, Stages, and Tasks
- Partitions and Shuffles
- Data Locality
- Job Performance

**HANDS-ON**
- Visualizing DAG Execution
- Observing Task Scheduling
- Understanding Performance
- Measuring Memory Usage

### MODULE 5
**Clustering Architecture**

**LECTURE**
- Cluster Managers for Spark: Spark Standalone, YARN, and Mesos
- Understanding Spark on YARN

**HANDS-ON**
- Tracking Jobs through the Cluster UI
- Understanding Deploy Modes
- Specifying Executors, Cores, Memory

### MODULE 6
**Spark Streaming**

**LECTURE**
- Streaming Sources and Tasks
- DStream APIs and Stateful Streams
- Reliability and Fault Recovery

**HANDS-ON**
- Creating DStreams from Sources
- Operating on DStream Data
- Viewing Streaming Jobs in the Web UI

### MODULE 7
**Machine Learning**

**LECTURE**
- Basic Principles of Machine Learning
- Spark ML API Patterns
- Built-in Featurizing and Algorithm APIs

**LECTURE**
- Featurizing and Learning with RDDs
- ML Using Pipelines and DataFrames