Spark Programming

Overview:

This 3-day hands-on workshop will introduce you to Apache Spark with coding exercises and lectures. Spark is a unified framework for big data analytics. Spark provides one integrated API for use by developers, data scientists, and analysts to perform diverse tasks that would have previously required separate processing engines such as batch analytics, stream processing and statistical modeling. Spark supports a wide range of popular languages including Python, R, Scala, SQL, and Java. Spark can read from diverse data sources and scale to thousands of of nodes.

In this class, you will learn how to build and manage Spark applications using Spark's core programming APIs and its standard Libraries.

Duration:

3 days.

Who is the course for:

Engineers, Data Scientists, and Analysts.

Prerequisites:

Students should arrive to class with:

- ✓ A basic understanding of software development
- ✓ Some experience coding in Python, Java, SQL, or Scala
- ✓ A laptop with a modern operating system (Windows, OS X, Linux), browser (Internet Explorer not supported), and Internet access

What you will learn?

After taking this class you will be able to:

- ✓ Build a data pipeline using Spark DataFrames and Spark SQL
- ✓ Understand Spark concepts, architecture, and applications
- ✓ Execute SQL queries on large scale data using Spark
- ✓ Explore and visualize your data by entering and running code in Notebooks
- ✓ Train, and use an ML model on real data with Spark's Machine Learning library MLlib
- ✓ Tune Spark job performance and troubleshoot errors using logs and administration UIs
- ✓ Find answers to common questions using Spark documentation and discussion forums
- ✓ Write and monitor a Spark Streaming job to analyze data with sub-second latency
- ✓ Understand common use-cases and business applications of Spark

Spark Programming

Course Outline

Day 1:

- ✓ History of Big Data & Apache Spark
- ✓ Databricks Overview
- ✓ Spark Capabilities and Ecosystem
- ✓ Basic Spark Components
- ✓ Spark SQL and DataFrame Uses
- ✓ DataFrame / SQL APIs
- ✓ Catalyst Query Optimization
- ✓ ETL

Day 2:

- ✓ Data Sources: reading from Parquet, S3, Cassandra, HDFS, and your local file system
- ✓ Memory & Persistence
- ✓ Jobs, Stages and Tasks
- ✓ Partitions and Shuffling
- ✓ Data Locality
- ✓ Spark's Architecture

Day 3:

- ✓ Structured streaming APIs
- ✓ Windowing
- ✓ Checkpointing and watermarking
- ✓ Streaming DataFrames
- ✓ Reliability and fault tolerance in Spark Streaming
- ✓ Spark MLlib Pipeline API
- ✓ Built-in featurizing and algorithms
- ✓ Basic graph analysis
- ✓ GraphFrames API
- ✓ GraphFrames motif finding
- ✓ Persisting graph data

Format

50% Lecture.

50% Labs.

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