Working with Apache Spark (for JPMC) | 2021

Course: TTDS6522-GKJ Working with Apache Spark (and Spark SQL) (for JPMC)
Duration: 2 days
Skill Level: Introductory
Targeted Audience: Typical attendees would include systems administrators, testers or technical data related roles who need to learn to use Spark for data analysis or processing data.
Hands-on Learning: This course combines engaging lecture, demos, group activities and discussions with machine-based student labs and exercises. Student machines are required.

Course Overview

Apache Spark is a powerful, open-source processing engine for data in the Hadoop cluster, optimized for speed, ease of use, and sophisticated analytics. The Spark framework supports streaming data processing and complex, iterative algorithms, enabling applications to run up to 100x faster than traditional Hadoop MapReduce programs. With Spark, you can write sophisticated parallel applications to execute faster decisions, better decisions, and real-time actions, applied to a wide variety of use cases, architectures, and industries.

This hands-on course is geared for technical business professional who wish to solve real-world data related problems using Apache Spark. This course explores using Apache Spark for common data related activities.

Audience: Who Should Attend?

This course is an **Introductory level and beyond** course. Typical attendees would include systems administrators, testers or technical data related roles who need to learn to use Spark for data analysis or processing data.

Attending students should have the following background:
- Introduction to Scala Programming (at least exposure to basic Scala syntax in support of Spark labs)
- Basic knowledge of Statistics and Probability
- Data Science background

Course Topics Covered

Spark Introduction
- Big Data, Hadoop, Spark
- Spark concepts and architecture
- Spark components overview
- Labs: Installing and running Spark

Spark and Hadoop
- Hadoop Primer: HDFS / YARN
- Hadoop + Spark architecture
- Running Spark on Hadoop YARN
- Processing HDFS files using Spark
- Spark & Hive

First Look at Spark
- Spark shell
- Spark web UIs
- Analyzing dataset – part 1
- Labs: Spark shell exploration

Spark Data Structures
- Partitions
- Distributed execution
- Operations: transformations and actions
- Labs: Unstructured data analytics using RDDs

Caching
- Caching overview
- Various caching mechanisms available in Spark
- In memory file systems
- Caching use cases and best practices
- Labs: Benchmark of caching performance

DataFrames / Datasets
- DataFrames Intro
- Loading structured data (json, CSV) using DataFrames
- Using schema
- Specifying schema for DataFrames
- Labs: DataFrames, Datasets, Schema

Spark SQL
- Spark SQL concepts and overview
- Defining tables and importing datasets
- Querying data using SQL
- Handling various storage formats: JSON / Parquet /

ORC
Labs: querying structured data using SQL; evaluating data formats