Introduction to Apache Spark and Spark SQL | JPMC 2022

* **Course: Introduction to Apache Spark and Spark SQL (TTSK7502-GKJ)**
* **Duration**: Four half day sessions; four hours each
* **Skill Level: This introductory-level course is for intermediate skilled attendees new to Apache Spark**.
* **Targeted Audience**: Typical attendees would include developers, analysts and business analysts who need to learn to use Spark for data analysis or processing data. Basic Python experience is recommended for the hands-on labs.
* **Hands-on Learning:** This course is *approximately* 50% hands-on lab to lecture ratio, combining 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. Thishands-on course is geared for technical business professional who wish to solve real-world data related problems using Apache Spark.

Learning Objectives

This course explores using Apache Spark for common data related activities. Throughout the course students will be led through a series of progressively advanced topics, where each topic consists of lecture, group discussion, comprehensive hands-on lab exercises, and lab review.

Throughout the course you’ll explore:

* Spark Essentials
* Spark Data structures
* Caching
* Dataframes / Datasets
* Spark SQL
* Spark and Hadoop
* Spark API

Audience: Who Should Attend?

This course is an **Introductory level and beyond** course. Typical attendees would include developers, analysts and business analysts who need to learn to use Spark for data analysis or processing data. Basic Python experience is recommended for the hands-on labs.

Attending students should have the following background:

* Introduction to Python Programming (at least exposure to basic Python syntax in support of Spark labs). Students without incoming Python experience can cut and paste labs or follow along, and plan to review practice further later after class.
* Basic knowledge of Statistics and Probability
* Basic Data Science background

Course Topics

NOTE: The topics, tools and skills in this course have been selected by JPMC management. Timing below is **approximate** and may require adjustment during live class delivery based on audience skill-level, needs and participation.

Session 1

Spark Introduction

* Big Data stacks, Hadoop, Spark
* Spark-3 new features
* Spark concepts and architecture
* Spark components overview
* Lab: Installing and running Spark

First Look at Spark

* Spark shell
* Spark web UIs
* Analyzing dataset – part 1
* Labs: Spark shell exploration

Session 2

Spark Data structures

* Partitions
* Distributed execution
* Operations : transformations and actions
* Labs : Unstructured data analytics

Session 3

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
* Specifying schema for Dataframes
* Labs : Dataframes, Datasets, Schema

Session 4

Spark SQL

* Spark SQL concepts and overview
* Defining tables and importing datasets
* Querying data using SQL
* Handling various storage formats : JSON / Parquet / ORC
* Adaptive Query Engine (AQE) (Spark 3 feature)
* Labs : querying structured data using SQL; evaluating data formats

Spark and Hadoop

* Hadoop Primer
* Hadoop + Spark architecture
* Running Spark on Hadoop
* Processing HDFS files using Spark
* Spark & Hive
* Labs: Spark and Hive

Bonus Content / Time Permitting

*These topics may or may not be presented during the live class depending on the pace of the course and attendee skill level and participation.*

Spark API

* Overview of Spark APIs in Python
* Life cycle of a Spark application
* Spark APIs
* Deploying Spark applications on YARN
* Labs : Developing and deploying a Spark application