Apache Spark & Scala

Duration: 6/7 Weekend

Prerequisites

- There are no pre-requisites for this course.
- Basic knowledge of Core Java and SQL is advantageous.

Course Content

1. **Introduction to Bigdata**
   - Introduction and relevance
   - Uses of Big Data analytics in various industries like Telecom, E- commerce, Finance and Insurance etc.
   - Problems with Traditional Large-Scale Systems
2. **Hadoop (Big Data) Ecosystem**
   - Motivation for Hadoop
   - Different types of projects by Apache
   - Role of projects in the Hadoop Ecosystem
   - Key technology foundations required for Big Data
   - Limitations and Solutions of existing Data Analytics Architecture
   - Comparison of traditional data management systems with Big Data management systems
   - Evaluate key framework requirements for Big Data analytics
   - Hadoop Ecosystem & Hadoop 2.x core components
   - Explain the relevance of real-time data
   - Explain how to use big and real-time data as a Business planning tool
3. **Hadoop Cluster Architecture – Configuration Files**
   - Hadoop Master-Slave Architecture
   - The Hadoop Distributed File System - data storage
   - Explain different types of cluster setups (Fully distributed/Pseudo etc.)
   - Hadoop Cluster set up - Installation
   - Hadoop 2.x Cluster Architecture
   - A Typical enterprise cluster – Hadoop Cluster Modes
4. **Data Analysis using HIVE**
   - Introduction to Hive - HiveHive Use Cases
   - Discuss the Hive data storage principle
   - Explain the File formats and Records formats supported by the Hive environment
   - Perform operations with data in Hive
   - Hive QL: Joining Tables, Dynamic Partitioning, Custom MapReduce Scripts
   - Hive Script, Hive UDF
5. **Scala Basics**
   - What is Scala?
   - Why Scala for Spark?
   - Scala in other Frameworks
   - Introduction to Scala REPL
   - Basic Scala Operations
   - Variable Types in Scala
   - Control Structures in Scala
   - Foreach loop, Functions and Procedures
   - Collections in Scala- Array
   - ArrayBuffer, Map, Tuples, Lists, and more
   - Scala Advance
6. **Advance Scala**
   - Functional Programming
   - Higher Order Functions
   - Anonymous Functions
   - Class in Scala
   - Getters and Setters
   - Custom Getters and Setters
   - Properties with only Getters
   - Auxiliary Constructor and Primary Constructor
   - Singletons
   - Extending a Class
   - Overriding Methods
   - Traits as Interfaces and Layered Traits
7. **Spark**
   - What is Apache Spark?
   - Using the Spark Shell
   - RDDs (Resilient Distributed Datasets)
   - Functional Programming in Spark
   - Working with RDDs in Spark
   - A Closer Look at RDDs
   - Key-Value Pair RDDs
   - MapReduce
   - Other Pair RDD Operations
8. **Apache Kafka**
   - Need for Kafka
   - What is Kafka?
   - Core Concepts of Kafka
   - Kafka Architecture

- Where is Kafka Used?
- Understanding the Components of Kafka Cluster
- Configuring Kafka Cluster
- Kafka Producer and Consumer Java API

9. Spark SQL

- Need for Spark SQL
- What is Spark SQL?
- Spark SQL Architecture
- SQL Context in Spark SQL
- User Defined Functions
- Data Frames & Datasets
- Interoperating with RDDs
- JSON and Parquet File Formats
- Loading Data through Different Sources

10. Spark Streaming

- What is Spark Streaming?
- Spark Streaming Features
- Spark Streaming Workflow
- How Uber Uses Streaming Data
- Streaming Context & DStreams
- Transformations on DStreams
- Using a Kafka Direct Data Source for spark streaming

11. Project

- Multiple Real World Use Case Scenarios