**Apache Spark Training**

Course Content Duration: 18 hours

Course Outline:

## Big Data Conceptuals

* What is Big Data?
* The need for Big Data
* Why Big Data now?
* Myths of Big Data
* Tabular representation of data unit measurement.
* Is one petabyte big data ? Types of Architectures in Big Data
* Lambda Architecture
* Kappa Architecture
* Zeta Architecture
* Seda Architecture
* NoSQL Store and high throughput messaging system Illustration about CAP theorem

Problems with large-scale systems

## HDFS

* Why HDFS ?
* HDFS Architecture
* Using HDFS and hdfs commands

## Spark (The Spark version covered is the latest version of Spark - 1.6) JDK 8 - Quick Introduction

* Functional Programming with Java
* Lambda expressions and Functional Interfaces in Java

## Scala - Introduction

* Objects and Classes
* val, var, functions, currying, implicits
* traits, actors and file manipulations

## Core Spark

* Introduction to Apache Spark
* What is Spark ? Explain about the modules in spark
* Spark-Shell - scala and python REPL
* Spark Internals - The Driver program, master, workers, executors and the tasks
* SparkSession- The Umbrella API for all context
* Running spark in a standalone mode
* Spark UI and monitoring a job
* Functional programming with Spark
* Map-reduce and Spark advantages over Map-reduce.

## RDD

* What is an RDD ?
* Laziness in RDD Evaluation
* Different ways of creating an RDD
* Types of RDD’s - PairRDD, DoubleRDD
* RDD Operations
* Partitions - The core of RDD
* textFiles, wholeFiles

## Running Spark on a Cluster

* Overview
* A Spark Standalone Cluster
* The Spark Standalone Web UI
* Installing and configuring a cluster

## Operations in Spark

* Spark Configuration and the Spark Context
* Configuring spark properties

RDD Operations - Transformation and Actions

* map, flatMap, repartition, coalesce, glom, reduce, cartesian, pipe, sample,
* distinct, mapPartitions, mapPartitionsWithIndex
* Map, filter, distinct, collect, take operations Joining two RDD’s

Storage levels supported in spark

Programming with a partition and use of custom partitioners Accumulators and Broadcast variables

Checkpointing an RDD Spark deployment plans Spark History Server

## Reading Data from External Sources

* JdbcRdd - Read data from mysql
* Connecting and reading data from mongodb

## Caching and Persistence

* RDD Lineage
* Caching Overview
* Distributed Persistence

## SparkSQL

* The DataFrame Abstraction
* Elucidate on SparkSQL
* Dataframe manipulation on top of json
* The temp table abstraction on top of DataFrame Schema
* SQL manipulation on top of parquest files
* Dataframes caching
* Connecting dataframes to relational database

## Spark Streaming

* Kafka and the need
* Basic read from a socket
* Spark Streaming from kafka
* Windowing operation in streaming
* Developing streaming applications
* Writing a custom receiver
* Spark Structured Streaming

## Spark Machine Learning

* Decision Trees
* Linear Regression
* Bayesian Classification

## Advanced Topics

* Spark SQL with Hive
* The new Dataset API
* Connecting Spark with HBASE
* Working with nested data
* Spark with Alluxio
* Custom Accumulators
* Writing custom RDD
* Writing custom partitioner
* Internals of persistence API. How spark manages persistence internally.
* (Drilling down the source code)
* Connecting spark with cassandra and ingesting data into cassandra

## Spark Performance Tuning

* Various strategies to adopt to performance tune your spark application.
* Introduction to various variables in Spark like shared variables.
* Broadcast variables and learning about accumulators.
* Common performance issues and troubleshooting the performance problems.
* Maven would be used as the build tool to download the dependencies. IntelliJ would be the IDE to develop the applications and examples.

## Project: A live project of how each of the API’s are used in the industry.

Use cases covered:

* A csv file format of three hundred columns will be used as a dataset.
* Consuming and operating two csv files (each of 3 MB) that are produced every second
* through spark streaming.
* Ten to fifteen transformation on a single job. Efficiently optimize and fine tune on all the
* transformations.
* Architectural sharing of data between spark jobs.

## Hands-on/Lecture Ratio:

The course is 60 % hands-on, 40 % discussion, with the longest discussion segments lasting 20 minutes.