Course Title: Big Data with Spark

Duration: 60 hours

Part – I: Big Data (Eco System) – 45 hours

Module – 1: Overview of Big Data and Analytics

- Big data Engineer and Data scientist
- Hadoop and its components
- NoSQL

Module – 2a: Setup

- Setting up a Single Node Hadoop Cluster (Local Machine)
- Setting up a Single Node Hadoop Cluster (Cloud - Optional)
- Setting up a Multi Node Hadoop Cluster (Local Machine and Cloud - Optional)

Module – 2b: HDFS

- HDFS daemons
- HDFS Read
- HDFS write
- Namenode & Datanode Architecture
- HDFS Federation

Module – 3: Map Reduce

- Map Reduce 1 Daemons
- Map Reduce 1 job work flow
- YARN
- Map Reduce on YARN job work flow
- Spark on YARN Overview
- Word count walkthrough

Module – 4: Pig & Hive

- What is Pig and Pig Latin?
- Pig Latin Constructs/Instructions
- Pig Latin Joins, Complex Types
- What is Hive? Hive Vs. Pig!
- Hive data types, Hive-QL, JOINS
- Partitioning, User Defined Functions
- Row Format and File format

Module – 5: HBase

- CAP theorem
- Log structured Merge Trees
- NoSQL Data model
- HBASE read and write
- HBase Storage

Module – 6: Introduction to Scala, Spark and additional tools

- Introduction to Scala
- Basic constructs required
- What is Spark and Why?
- Who wins – Hadoop Map-Reduce or Spark?
- Spark or Hadoop Map-Reduce – Which one to choose and when?
- Sqoop and Flume Concepts with hands on.

Module – 7: Hands-on Project

- Requirements
- Architecture
- Implementation
- Industry connect

Part – II: Scala and Spark – 15 hours (Optional i.e. Extension to Part I)

Module – 1: Scala

- The Basics
- Control structure and Functions
- Working with Arrays
- Maps and Tuples
- Classes
- Objects
- Packages and Imports
- Inheritance

Module – 2: Spark

- Introduction to Apache Spark
- Load & Inspect Data
- Build a Simple Spark Application
- Work with Pair RDD
- Work with DataFrames
- Monitor Apache Spark Applications
- Introduction to Apache Spark Data Pipelines
- Create an Apache Spark Streaming Application