**Apache SPARK**

**Duration : 30 Hours**

**Introduction to Big Data and Spark**

• Overview of BigData and Spark

• MapReduce limitations

• Spark History

• Spark Architecture

• Spark and Hadoop Advantages

• Benefits of Spark + Hadoop

• Introduction to Spark Eco-system

**Foundation to Spark**

• Spark Shell

• Basic operations on Shell

• Spark Context and Spark Properties

• Persistence in Spark

• HDFS data from Spark

**Working with Resilient Distributed DataSets (RDD)**

• Understanding RDD

• Loading data into RDD

• Scala RDD, Paired RDD, Double RDD & General RDD Functions

• Transformations, Actions and Shared Variables

• Spark Operations

* Spark Dataformats
* Data Formats & Management
* InputFormat and InputSplit
* JSON
* XML
* AVRO
* How to store many small files - SequenceFile?
* Parquet
* Comparing Compressions
* Understanding Row Oriented and Column Oriented Formats - RCFile?

**Spark Eco-system - Spark Streaming & Spark SQL**

• Introduction to Spark Streaming

• Introduction to Spark SQL

• Querying Files as Tables

• Text file Format

• JSON file Format

• Hive and Spark SQL Architecture

**Spark Best Practises**