**Apache SPARK & Kafka**

**Duration : 40 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 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**

**Kafka introduction**

• What is Apache Kafka

• Kafka Features and terminologies

• High level Kafka Architecture

• Real life Kafka Case Studies

**Kafka Architecture**

• Internals of architecture and core concepts

• Kafka components - Broker, Producer, Consumer, Topics, Partitions

• Different versions of Kafka

**Understanding Broker**

• Working of Broker

• Broker Deployment

• Multiple brokers on single machine

• Decommissioning Brokers

**Understanding Producer**

• Basics of producer

• Producer Architecture

• Producer partition- Custom, Round Robin, Field Based Partition

• Producer Java API

• Types of Producer - sync,async

• Different Producer Configurations

• Sync and async producer hands on

**Understanding Consumer**

• Basics of Consumer

• Consumer Queuing, Consumer Group

• Consumer Java API

• Producer and Consumer Hands On

**Mirroring Kafka**

• What is mirroring

• How mirroring works

• Mirror Maker and its role

• Kafka Mirroring Hands on

• Mirror Maker producer and consumer

**Topics and Partitions**

• Working with topics

• Using Partitions and distribution of partitions

**Replication in Kafka**

• High availability and reliability using Replication

• ISR - In Sync Replication

• Topic, Partition and Replication Hands on

**Kafka Best Practises**