**Apache SPARK**

**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

**Scala**

Introduction

Basics of Scala

data Types

Operators

The control instructions and loops

Object-Oriented Programming

Classes

Fields

Methods

Objects and instances

Inheritance, abstraction, encapsulation, polymorphism

Features

Functional Programming

A function declaration

function Arguments

closures

anonymous functions

recursion

**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 Streaming**

Introduction to Spark Streaming . Spark Structured Streaming

Windowing

Delta Lakes

Streaming Architectures, Lambda Architecture

Differentiating discretized and structured streaming

Hands-on

**Spark SQL**

• Introduction to Spark SQL

• Architecture and flow

• Querying Files as Tables

• Grouping,Joins,Aggregations

• Text file Format

• JSON file Format

• Hive and Spark SQL Architecture

Hands-on

**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

• Working of Broker

• Broker Deployment

• Multiple brokers on single machine

• Decommissioning Brokers

 • Basics of producer

• Producer Architecture

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

• Producer Java API

• Basics of Consumer

• Consumer Queuing, Consumer Group

• Working with topics

• Using Partitions and distribution of partitions

• High availability and reliability using Replication

• ISR - In Sync Replication

• Topic, Partition and Replication Hands on