**Apache SPARK & Kafka**

**Duration : 40 Hours**

 Scala : 2 day

Spark : 3 days

**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

What is Scala

The paradigms of object-oriented programming

Functional programming paradigms

Scala and Java

Language runtime environments

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 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**

**Optimizing and Tuning Spark for Efficiency**

Optimizing and Tuning Spark for Efficiency

Viewing and Setting Apache Spark Configurations

Scaling Spark for Large Workloads

Caching and Persistence of Data

DataFrame.cache()

DataFrame.persist()

When to Cache and Persist

When Not to Cache and Persist

A Family of Spark Joins

Broadcast Hash Join

Shuffle Sort Merge Join

Inspecting the Spark UI

Journey Through the Spark UI Tabs

Spark Tuning and Cluster Sizing

How to Adjust Spark Settings

How to Determine the Relevant Information About Your Cluster

Basic Spark Core Settings: How Many Resources to Allocate to the Spark Application?

Calculating Executor and Driver Memory Overhead

How Large to Make the Spark Driver

A Few Large Executors or Many Small Executors?

Allocating Cluster Resources and Dynamic Allocation

Dividing the Space Within One Executor

Number and Size of Partitions

Serialization Options

Debugging Techniques