Getting Started with Big Data

MODULE 1: BIG DATA INTRODUCTION

- Big Data Overview
- Five Vs of Big Data
- What is Big Data and Hadoop
- Introduction to Hadoop
- Components of Hadoop Ecosystem
- Big Data Analytics Introduction

MODULE 2: HDFS AND MAP REDUCE

- HDFS – Big Data Storage
- Distributed Processing with Map Reduce
- Mapping and reducing stages concepts
- Key Terms: Output Format, Partitioners, Combiners, Shuffle, and Sort
- Hands-on Map Reduce task

MODULE 3: PYSPARK FOUNDATION

- PySpark Introduction
- Spark Configuration
- Resilient distributed datasets (RDD)
- Working with RDDs in PySpark
- Aggregating Data with Pair RDDs

MODULE 4: SPARK SQL and HADOOP HIVE

- Introducing Spark SQL
- Spark SQL vs Hadoop Hive
- Working with Spark SQL Query Language

MODULE 5: MACHINE LEARNING WITH SPARK ML

- Introduction to MLlib Various ML algorithms supported by Mlib
- ML model with Spark ML.
- Linear regression
- logistic regression
- Random forest

MODULE 6: KAFKA and Spark

- Kafka architecture
- Kafka workflow
- Configuring Kafka cluster
- Operations