Big Data – Spark and Scala

Course Modules

Big Data

- Understanding Data & Hadoop: Basic Concepts
- What is BigData
- Characteristics of BigData
- Challenges with Traditional Systems
- Problems with BigData
- Handling BigData

HADOOP Core Concepts

- Problems with Existing Distributed Systems to deal Big Data
- Why Hadoop and An Overview and History of Hadoop
- Requirements of New Approach
- The Hadoop Project and Hadoop Components

Scala Basics

- Scala Installation
- Know the concepts of classes in scala
- Object orientation in scala
- Primitive Datatypes
- Scala simple build tool – SBT
- Functional programming in scala – Closures,Currying,Anonymous functions
- Exploring mutable and immutable variables
- Execution of Scala code through REPL or CLI
- Working on basic programming constructs
- Collections – array,set

Apache Spark

- Introduction to Apache Spark
- Hadoop vs Spark
- Why Spark
- Spark Vs Mapreduce
- Batch Vs. Real Time Big Data Analytics
- Spark Installation and Configuration
- Spark Execution Architecture
- Components of Spark – SQL,Streaming,Storm,GraphX
- Understanding Spark Context
- Resilient Distributed Data (RDD) – Partitions,Features ,Parallelism

Working with RDD's

- RDD operations – Transformations and Actions
- RDD - DeepDive,Persistance/Caching,Lineage
- Types of RDD -Pair RDD,chain RDD
- Spark API programming
- Executing spark program with SBT and spark-assembly
- Understanding spark-submit.
- Running spark program in local mode and in cluster

Spark SQL – structure data (Hive with spark sql) – batch processing

- Spark SQL overview
- Understanding Dataframes,Datasets.
- Dataframes Vs RDD's
- Processing data using Dataframes
- Hive Context
- Custom case classes
- Temp tables Vs Persistent tables
- Inferring Schema programmatically
- Querying files as tables – CSV,Text,JSON,Parquet
- Standard transformations in querying
- Analytics and Window functions in sql
- Working of Spark SQL in Native and Hive context

Spark Streaming – unstructured data , real time processing

- Features of Spark Streaming
- Understanding Dstreams
- Use case 1:- Streaming data from netcat server
- Use case 2:- Flume and spark streaming integration
- Use case 3:- Kafka and Spark streaming integration (kafka -messaging service)
- Sliding window operations
- Transformers and Estimators

Kafka

- What is Kafka
- Architecture
- How messaging works
- Kafka Cluster
- Kafka Partitioning
- Command Line Broadcasting
- Integrating kafka with flume for broadcasting

Assessments will happen for every week to analyse the understanding of the students on the ongoing topics.

Once all the workarounds and assessments are done, the students will be given five real time projects based on what they have learned in the course.