Apache Spark - 40Hrs

1.0Introduction to Big Data and Apache Spark

Topics - Introduction to big data, challenges with big data, Batch Vs. Real Time big data analytics, Batch Analytics - Hadoop Ecosystem Overview, Real-time Analytics Options, Streaming Data - Spark, In-memory data - Spark, What is Spark?, Spark Ecosystem, modes of Spark, Spark installation demo, overview of Spark on a cluster, Spark Standalone cluster, Spark Web UI.

2.0. Spark Common Operations

Topics - Invoking Spark Shell, creating the Spark Context, loading a file in Shell, performing basic Operations on files in Spark Shell, Overview of SBT, building a Spark project with SBT, running Spark project with SBT, local mode, Spark mode, caching overview, Distributed Persistence.

3.0. Playing with RDDs

Topics - RDDs, transformations in RDD, actions in RDD, loading data in RDD, saving data through RDD, Key-Value Pair RDD, MapReduce and Pair RDD Operations, Spark and Hadoop Integration-HDFS, Spark and Hadoop Integration-Yarn, Handling Sequence Files, Partitioner.

4.0. Spark Streaming

- Spark Streaming Architecture,

- First Spark Streaming Program,
- Transformations in Spark Streaming,
- Fault tolerance in Spark Streaming
- check pointing
- TCP Streams
- File Streams
- FLUME
- Kafka

6.0.Real Time ETL & Analytics With Spark

- ✓ **First Streaming Spark SQL Application**
- ✓ **Apache Spark SQL :**
    - Data Frame Creation
    - SQL Execution
    - Configuration
    - Processing The Text File
    - Processing The JSON Files
    - Processing The Paraquet files
    - Using SQL
    - User defined functions

- ✓ **Data Frames :**
    - Types
    - Query Transformation
    - Actions
    - RDD Operation
    - Persistence

7.0.SparkR

First SparR Application

Execution

Streaming SparkR

8.0 . Spark Hive

- Hive Context
- Local Hive Meta Store Server
- A Hive Based Metastore Server

9.0. Machine Learning at Scale

- **Introduction**
- **Machine Learning Applications**
  - **classification**
  - **Regression**
  - **Clustering**
  - **Anomaly Detection**
  - **Recommendation**
  - Dimensionality Reduction

- **Architecture**
- **Development Environment**
- **Classification with Naive Bayes**
- **Clustering**
  - K-Means
  - Streaming K-means
  - Gaussian Mixture
- **Artificial Neural Network(ANN)**

✓ **Feature Selection & Extraction Algorithm**

- Chi-Square Selection
- Principal Component Analysis (PCA)

✓ **Recommendation Algorithm :**

    - **Collaborative Filtering Algorithm**
    - Collaborative Filtering with Alternating Least Square (ALS)
- ✓ **Streaming MLib Application**

10.0.Apache Spark GraphX

- Inroduction GraphX
- GraphX Coding
- Enviroment
- Creating a graph
- Example1 - Counting
- Example 2 - Filtering
- Exampe 3 - PageRank
- Example 4 Traingle Counting
- Example 5 connected components

11.0.Apache Spark with H2O

- Installing H2O
- The Build Environment
- Architecture
- Sourcing the data
- The Data Quality
- Performance Tuning

12.Cluster Managers

Scala Programming : 16hrs

- What is Scala?
- Why Scala for Spark?
- Scala in other frameworks,
- introduction to Scala REPL,
- basic Scala operations,
- Variable Types in Scala,
- Control Structures in Scala,
- Foreach loop,
- Functions,
- Procedures,
- Class in Scala,
- Getters and Setters,
- Custom Getters and Setters,
- Properties with only Getters,
- Auxiliary Constructor,
- Primary Constructor,
- Singletons, Companion Objects,
- Extending a Class,
- Overriding Methods,
- Traits as Interfaces,
- Layered Traits,
- Functional Programming,
- Higher Order Functions,
- Anonymous Functions, and more.
- Curry Function

- **Collections** in Scala- Array, ArrayBuffer, Map, Tuples, Lists, and more.
- File Handling
- Exception in Scala
- Multithreading In Scala

APACHE KAFKA : 16-20hrs

Apache Kafka Training Topics	
Introduction of Kafka Basics and Messaging System	1. Messaging System 2. Distributed Messaging System 3. Point to Point Messaging system 4. Publisher and Subscriber Messaging System 5. Introduction Event Processing and CEP 6. Use of Kafka 7. What is Kafka 8. Kafka Architecture 9. Different components in the Kafka architecture. 10. Role of zookeeper, Kafka Broker, Kafka Cluster, Producers and Consumers.
Download, Installation and Configuration	1. Download Kafka, Zookeeper , install and configure
Core Internals of Apache Kafka	1. Topics 2. Partitions 3. Consumers 4. Producers 5. Working with Topic, producer and consumers 6. Analysing Commit Log
Kafka APIs and Usage	1. Core API's and its usage
Kafka Brokers	1. What is Broker 2. Running multiple brokers 3. Working with Producer, Consumer and Broker 4. Leader, Replica and ISR attributes
Kafka Producers	1. Kafka Producer 2. Working with Producer to connect with Kafka Cluster. 3. implement the Kafka producer using Java

	4. Different partitioning mechanism 5. Configuration of Topic
Kafka Consumer	1. Kafka Consumer 2. implement the Kafka Consumer 3. Offset Management in Kafka Consumer 4. Automatic and Manual Offset Management 5. Consumer Groups its usage and Advantages. 6. Consumer Group id and its benefits. 7. Implementation of consumer group 8. Resetting offset in consumer
Kafka Client - GUI Tool	1. Download, install and demo about the Kafka GUI client tool that will be used to connect and Manage the Kafka cluster.
Apache Kafka Security and Authentication	1. Kafka security implementation , 2. Enabling SSL in Kafka Broker, 3. Accessing SSL secured broker using Console Consumer/Producer 4. Configure SSL in Kafka Producer , Consumer 5. Deleting a topic
Kafka End to End Process Flow Kafka Spark Integration Kafka Storm Integration Kafka High Availability and Consistency	