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 <MLIB>

- **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