Bigdata - Hadoop 2.X

Duration: 40-50 Hours

Prerequisites

- There are no pre-requisites for this course.
- Basic knowledge of Core Java and SQL is advantageous.

Course Content

1. Core Java

- Overview of Java
- Classes and Objects
- Garbage Collection and Modifiers
- Inheritance, Aggregation, Polymorphism
- Command line argument
- Abstract class and Interfaces
- String Handling
- Exception Handling, Multithreading
- Serialization and Advanced Topics
- Collection Framework, GUI, JDBC

Bigdata – Hadoop 2.X

1. Introduction to Bigdata

- Introduction and relevance
- Uses of Big Data analytics in various industries like Telecom, E- commerce, Finance and Insurance etc.
- Problems with Traditional Large-Scale Systems

2. Hadoop (Big Data) Ecosystem

- Motivation for Hadoop
- Key technology required for Big Data
- Limitations and Solutions of existing Data Analytics Architecture
- Comparison of traditional data management systems with Big Data management systems
- Evaluate key framework requirements for Big Data analytics
- Explain the relevance of real-time data
- Introduction to **Apache Hadoop**

3. Building Blocks

- Quick tour of Java (As Hadoop is Written in Java , so it will help us to understand it better)
- Quick tour of Linux commands (Basic Commands to traverse the Linux OS)
- Quick Tour of RDBMS Concepts (to use HIVE and Impala)
- Quick hands on experience of SQL.
- Introduction to Cloudera VM and usage instructions

4. Hadoop Cluster Architecture – Configuration Files

- Hadoop Master-Slave Architecture
- The Hadoop Distributed File System - data storage
- Explain different types of cluster setups (Fully distributed/Pseudo etc.)
- Hadoop Cluster set up - Installation
- Hadoop 2.x Cluster Architecture
- A Typical enterprise cluster – Hadoop Cluster Modes

5. Hadoop Core Components – HDFS & Map Reduce (YARN)

6. HDFS Overview & Data storage in HDFS

- Get the data into Hadoop from local machine (Data Loading Techniques) - vice versa
- MapReduce Overview (Traditional way Vs. MapReduce way)
- Concept of Mapper & Reducer
- Understanding MapReduce program skeleton
- Develop MapReduce Program in JAVA
- Test and debug a MapReduce program in the design time
- How Partitioners and Reducers Work Together
- Writing Customer Partitioners Data Input and Output

7. **Data Integration Using Sqoop and Flume**
   - Integrating Hadoop into an existing Enterprise
   - Loading Data from an RDBMS into HDFS by Using Sqoop
   - Managing Real-Time Data Using Flume
   - Accessing HDFS from Legacy Systems with FuseDFS and HttpFS
8. **Data Analysis using HIVE**
   - Introduction to Hive
   - Discuss the Hive data storage principle
   - Explain the File formats and Records formats supported by the Hive environment
   - Perform operations with data in Hive
   - Hive QL: Joining Tables, Dynamic Partitioning, Custom MapReduce Scripts
   - Hive Script, Hive UDF
9. **Data Analysis Using Impala**
   - Introduction to Impala & Architecture
   - How Impala executes Queries and its importance
   - Hive vs. Impala
   - Extending Impala with User Defined functions
   - Improving Impala performance
10. **NoSQL Database – Hbase**
    - Introduction to NoSQL Databases and Hbase
    - HBase v/s RDBMS, HBase Components, HBase Architecture
    - HBase Cluster Deployment
11. **Other Apache Projects**
    - Introduction to Zookeeper - ZooKeeper Data Model, Zookeeper Service
    - Introduction to Spark
12. **Spark**
    - What is Apache Spark?
    - Using the Spark Shell
    - RDDs (Resilient Distributed Datasets)
    - Functional Programming in Spark
    - Working with RDDs in Spark
    - A Closer Look at RDDs
    - Key-Value Pair RDDs
    - MapReduce
    - Other Pair RDD Operations
    - Introduction to Spark Core
    - Introduction to Spark SQL
    - Introduction to Spark Storm

Final Project

- ❖ Real World Use Case Scenarios
- ❖ Understand the implementation of Hadoop in Real World and its benefits.
- ❖ Final project including integration various key components
- ❖ Follow-up session: Tips and tricks for projects, certification and interviews etc