EDUPRISTINE
Empowering Professionals

Big Data Hadoop

Duration 60 - 80 Hours

Topics covered under this program are:

1. Introduction to Big Data Hadoop
    - Overview of Big Data Terminologies and its role in analytics
    - Business Analytics Vs. Big Data Hadoop
    - Data Science Vs. Data Engineering
    - Big data challenges and solutions
2. Unix and Python
    - Case Study: Telecom Sector
        - Setting up, accessing and verifying Linux server access over SSH
        - Transferring files over FTP or SFTP
        - Creating directory structure and Setting up permissions
        - Understanding File name pattern and move using regular expressions
        - Changing file owners, permissions
    - Case Study: Generating data by Python
        - Preparing configuration file that can be changed to fit any req.
        - Developing python script to generate mock data

3. JAVA

Case: Design and Develop Phone Book in Java

- Implementing design to Java code by Eclipse
- Compiling and executing Java Program
- Inheritance, Method overlapping

4. Introduction to Hadoop Distributed File System (HDFS).

- Design of HDFS and its Concepts
- Command Line Interface
- Hadoop File Systems and Java Interface
- Data Flow (Anatomy of a File Read, Anatomy of a File Write, Coherency Model)
- Parallel Copying with DISTCP
- Hadoop Archives

5. Sqoop

Case Study: Develop automation utility to migrate huge RDBMS warehouse implemented in MySQL to Hadoop cluster

- Creating and loading data into RDBMS table to understand RDBMS setup
- Preparing data to experiment with Sqoop imports
- Importing using Sqoop Command in HDFS file system to understand simple imports
- Importing using Sqoop command in Hive table to import data into Hive partitioned table and perform ETL
- Exporting using Sqoop from Hive/ HDFS to RDBM to store the output of Hive ETL into RDBMS
- Wrapping Sqoop commands into Unix Shell Script To be able to build and use automated utility for day to day use

6. Map Reduce

Case Study: Processing 4G usage data of a Telecom Operator to find out potential customers for various promotional offers

- Cleaning data, ETL and Aggregation
- Exploring data set using known tools like Linux commands to understand the nature of data
- Setting up Eclipse project, maven dependencies to add required Map Reduce Libraries
- Coding, packaging and deploying project on Hadoop cluster to understand how to deploy/ run Map Reduce on Hadoop Cluster

7. HIVE

Case Study: Process a structured data set to find some insights

- Finding out per driver total miles and hours driven
- Creating Table, Loading Data, Selecting Query to load, query and cleaning of data
- Which driver has driven maximum & minimum miles
- Joining Tables, Saving Query results to table to explore and use right type of table type, partition schema, buckets
- Discussing optimum file format for hive table
- Using right file format, type of table, partition scheme to optimize query performance
- Using UDFs to reuse domain specific implementations

8. PIG

Case Study: Perform ETL processing on Data Set to find some insights

- To learn what is Pig, where we can use Pig, how Pig is tightly coupled with Map-Reduce
- Installing and Running Pig, Grunt, Pig's Data Model
- Pig Latin, Developing & Testing Pig Latin Scripts

- Writing Evaluation, Filter, Load & Store Functions.
- Aggregating and looping
- Dumping, storing, joining, sorting etc.

9. SPARK

Case Study: Build a model to predict production error

- Aggregating data based on Response Code to find out server' performance from logs
- Filtering, Joining and aggregating data to find top 20 Frequent Hosts that generates errors
- Spark Project: Build a model (using Python) to predict production error/ failure (huge servers - applications/ software) with good speed by using computation power efficiently while considering processor challenges

10. Oozie

Case Study: Setting up Data processing pipeline to work as per schedule in Hadoop Eco System comprising of multiple components like sqoop job, hive scripts, pig scripts, spark jobs etc.

- Setting up Oozie workflow to trigger a script, then Sqoop Job followed by Hive Job
- Executing workflow to run complete ETL pipeline

11. HBASE

Case Study: Find out top 10 customers by expenditure, top 10 most buying brands, and monthly sales from data stored in HBASE which is in Key value pair

- Introduction, Client API - Basics, Client API - Advanced Features
- Client API - Administrative Features, Available Client, Architecture
- Map/Reduce Integration, Advanced Usage, Advance Indexing

12. Zookeeper

- The Zookeeper Service (Data Modal, Operations, Implementation, Consistency, Sessions, States)
- Building Applications with Zookeeper (Zookeeper in Production)

13. SQOOP

- Database Imports, Working with Imported Data, Importing Large Objects
- Performing Exports, Exports - A Deeper Look