EDUPRISTINE
Empowering Professionals

BIG DATA & HADOOP PROGRAM

Hadoop is a software framework for storing and processing Big Data. It is an open-source tool build on java platform and focuses on improved performance in terms of data processing on clusters of commodity hardware.

Hadoop consists of multiple concepts and modules like HDFS, Map-Reduce, HBASE, PIG, HIVE, SQOOP and ZOOKEEPER to perform the easy and fast processing of huge data. Hadoop is conceptually different from Relational databases and can process the high volume, high velocity and high variety of data to generate value.

Career Opportunities

- Companies like Google, EMC Corporation, Yahoo, Apple, HortonWorks, Oracle, Amazon, Cloudera, IBM, Cisco, Microsoft and many more have opened their doors for Hadoop professionals.
- Various positions like product managers, Hadoop developers, software testers, database administrators, senior Hadoop developers and alike are open.
- Companies are searching for experienced candidates as well as freshers.

Who Can Join this Course?

- Software Engineers, who are into ETL/Programming and exploring for great job opportunities in Hadoop
- Managers, who are looking for the latest technologies to be implemented in their organization, to meet the current and upcoming challenges of data management
- Any Graduate/Post-Graduate, who is aspiring a great career towards the cutting edge technologies

Course Structure

Introduction to Big Data

- Overview of Big Data Technologies and its role in Analytics
- Big Data challenges & solutions
- Data Science vs Data Engineering
- Job Roles, Skills & Tools

Development Environment

- Setting up Development environment on User's laptop to be able to develop and execute programs
- Setting up Eclipse (Basics of Eclipse like, import, create project, add JARs) to understand basics of Eclipse for Map Reduce and Spark development
- Installing Maven & Gradle to understand building tools
- Installing Putty, FileZilla/WinSCP to get ready to access EduPristine's Big Data Cloud

Unix, Python & Java

- 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
- Reviewing mock file generator utility written in Shell Script, enhancing it to be more useful
- Understand provided data set
- clean data as per given process
- Developing python script which calculates loan eligibility process
- Identifying Classes and Methods for Phone Book
- Implementing design into Java Code using Eclipse
- Compiling and Executing Java Program
- Enhancing the code with each learnings, like Inheritance, Method overloading
- Further enhancing the code to initialize Phonebook from a Text File by using Java file reading

HDFS

- Understanding the problem statement and challenges persisting to such large data to perceive the need of Distributed File System
- Understanding HDFS architecture to solve problems
- Understanding configuration and creating directory structure to get a solution of the given problem statement
- setup appropriate permissions to secure data for appropriate users
- Setting up Java Development with HDFS libraries to use HDFS Java APIs
- Develop utility in Java which works like "ls" command but returns files older than given number of days

SQOOP

- 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

Map Reduce

- 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

HIVE

- 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

PIG

- Loading and exploring Movie - 100K data set to load data set, explore it and associate schema to it
- Using grunt, Loading data set, defining schema
- Finding Simple Statistic from given Data Set to clean up the data
- Filtering and modifying data schema
- Finding gender distribution in users
- Aggregating and looping
- Finding top 25 movies by rating, joining data sets and saving to HDFS to perform aggregation
- Dumping, Storing, joining, sorting
- Filtering function for complex condition to reuse domain specific functionalities & avoid rewriting code
- Using UDFs

SPARK

- Loading and performing pre-processing to convert unstructured data to some structured data format
- Cleaning data, filtering out bad records, converting data to more usable format
- 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

OOZIE

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

HBase

- Designing HBase Table Schema to model table structure, decide families in table as per data
- Deciding families in table as per data
- Bulk Loading & Programmatically Loading data using Java APIs to populate data into HBase table
- Querying and Showing data on UI to integrate HBase with UI/Reporting

R integration with Hadoop

- Real Time Analytics, Unstructured Data Ingestion

Mongo DB

- An open source database that uses a document-oriented data model

Resume Building and Cloudera Exam Guidance

- An open source database that uses a document-oriented data model

Live Projects

- ETL processing of retail logs
- Customer 360 degree
- Twitter Sentiment Analytics
- Developing a Chat-bot to offer an artificially intelligent customer help desk for an insurance company

Training Highlights

Classroom Training

15 days Classroom (75 hours) + 2 days Online Training (12 Hours) (Java, Unix & Python)

Live Project

4 live projects included with case studies and take away home assignment

Online Materials

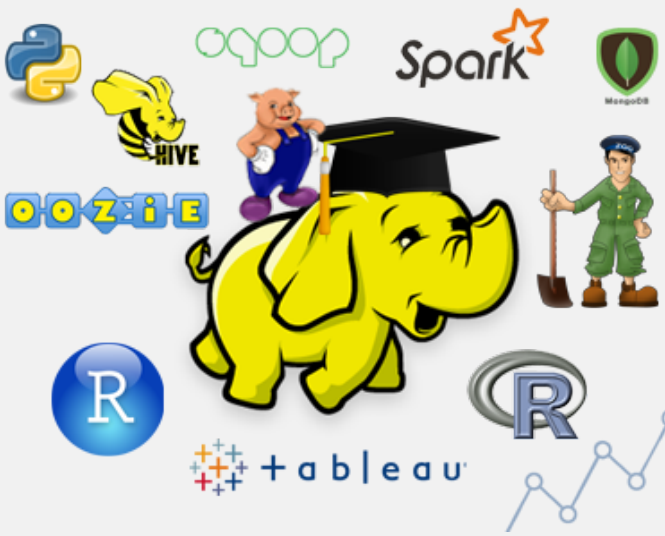
Topic Wise study material in the form of Presentation and Case Studies - PowerPoint Presentation covering all classes - Code files for each case study
- Recorded Videos of Live Instructor based Training
- Recorded Videos Covering all classes
- Quiz/Assignment with detailed answers and explanation
- Job Oriented Questions to prepare for Certification Exams
- Doubt solving forum to interact with faculty & fellow students

Complimentary Course

"Java Essentials for Hadoop", Python and UNIX session

24x7 Online Access

24x7 Online Access to Course Materials.



Course Fees: INR 39,500 /-

About EduPristine

EduPristine is one of the largest exam preparation and certifications providers for finance, accounting, analytics, marketing, health care like CFA® Program, FRM, Financial modeling, CPA, US CMA, ACCA, Business accounting and taxation, Business analytics, Big data and Hadoop, Data science, Digital Marketing, USMLE. Our online Learning Management System conducts online and classroom trainings for individuals and corporates.

Testimonials

> *The faculty of EduPristine has in depth knowledge and are able to deliver it very well. They are able to solve all the doubts of the students. The content for Hadoop is also good. I would rate the Hadoop course at EduPristine 9/10*
>
> **- Dipti Poojari**

> *The trend and concept of managing huge data and the course structure of EduPristine, First of all "Classroom training" gives amazing practical experience which everyone seek and the second best part is the course structure and faculty, It helps me in getting out of declining mainframe trend, Faculty is amazing and very knowledgeable. They teach with great interest. I would like to rate it as 8/10.*
>
> **- Hardik Joshi**

Contact Us:

TOLL FREE – 1800 200 5835

Bangalore | Chennai | Delhi | Hyderabad | Kolkata | Mumbai | Pune | Online