SEVENMENTOR TRAINING PVT.LTD

HadoopSyllabus

ABOUT BIG DATA & HADOOP PROGRAM

Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model.
It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures
are common occurrences and should be automatically handled by the framework.
The core part of Hadoop consist of a Storage part, known as Hadoop Distributed File system(HDFS),and a processing part which is a mapreduce programming model.

Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HBase, ZooKeeper, Oozie, Pig, and Sqoop.

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, testers, database administrators, senior hadoop developers and alike are open.
Companies are searching for experienced candidates as well as freshers.

Course Structure and Pre-Requisites

1.Software Developers/Programmers/Engineers,who are into Database/Programming and exploring for great job opportunities in Hadoop

2.Managers, who are looking for the latest technologies to be implemented in their organization, to meet the current and upcoming challenges of data management

3.Any Graduate/Post-Graduate, who is aspiring a great career towards the cutting edge technologies

Office 21-25/A, First Floor Shreenath Plaza,
Dnyaneshwar Paduka Chowk, Pune, Maharashtra 411005
Mob : +91-7798058777

SEVENMENTOR TRAINING PVT.LTD

Topics and Structure PRO package:

A)RDBMS Vs Hadoop

Comparison in between Mysql and Hadoop
Why Hadoop is better that Mysql??

B)Introduction to Java

Basics of Java required for Hadoop i.e.Core Java

C)Introduction to HDFS & Understanding cluster environment

- NameNode and DataNodes.
- HDFS has a master/slave architecture.
- An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients.

D)Understanding Map-Reduce Basics, Types & Formats

- MapReduce is mainly used for parallel processing of large sets of data stored in Hdfs.
- Initially, it is a hypothesis specially designed by Google to provide parallelism, data distribution and fault-tolerance

E)HIVE

- Hive is data warehousing software that addresses how data is structured and queried in distributed Hadoop clusters.
- Itís a system that gives users the tools to make powerful queries and get results often in real-time.
- Hive is popular development environment that is used to write queries for data in the Hadoop environment.
- Hive is a declarative language that is used to develop applications for the Hadoop environment, however it does not support real-time queries.

Hive has several components, including:

1. HCatalog
2. WebHCat
3. HiveQL

F)PIG

- Pig is a platform with a high-level query language built to handle large data sets.
- Pig is a procedural language for developing parallel processing applications for large data sets in the Hadoop environment.
- Pig is an alternative to Java programming for MapReduce, and automatically generates MapReduce functions.
- Pig is popular because it automates some of the complexity in MapReduce development.

G)SQOOP

- Think of Sqoop as a front-end loader for big data. Sqoop is a command-line interface that facilitates moving bulk data from Hadoop into relational databases and other structured data stores.
- Using Sqoop replaces the need to develop scripts to export and import data.
- One common use case is to move data from an enterprise data warehouse to a Hadoop cluster for ETL processing.
- Performing ETL on the commodity Hadoop cluster is resource efficient, while Sqoop provides a practical transfer method.

H)Oozie

- Oozie is the workflow scheduler that was developed as part of the Apache Hadoop project.
- It manages how workflows start and execute, and also controls the execution path.
- Oozie is a server-based Java web application that uses workflow definitions written in hPDL, which is an XML Process Definition Language similar to JBOSS JBPM jPDL.
- Oozie only supports specific workflow types, so other workload schedulers are commonly used instead of or in addition to Oozie in Hadoop environments.

I)HBASE

- HBase is a non-relational database management system that runs on top of HDFS. It is built to handle sparse data sets common to big data projects.
- It was designed to store structured data in tables that could have billions of rows and millions of columns.
- It has been deployed to power historical searches through large data sets, especially when the desired data is contained within a large amount of unimportant or irrelevant data (also known as sparse data sets).

J)Basics of Spark

- Apache Spark is a general compute engine that offers fast data analysis on a large scale.
- Spark is built on HDFS but bypasses MapReduce and instead uses its own data processing framework.

K)Basics of MongoDB

- Installation of mongodb in ubuntu

L)Product based Ecommerce application Project(Demo project)