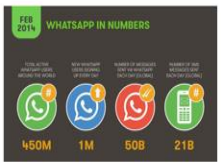
E2MATRIX

Training and Research Institute

www.e2matrix.com

HADOOP COURSE STRUCTURE

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is now an Apache Hadoop subproject.



Apache Hadoop is a framework that allows for distributed processing of large data sets across clusters of commodity computers using a simple programming model. It is used as a Processing Platform for Big Data processing by using the "Map Reduce" Processing Paradigm. The main purpose of this "Hands-on Training on

HADOOP" is to create awareness and enrich knowledge for research scholars, faculty and students in the area of Big Data using Hadoop.

1. **Introduction to Hadoop**
    a. What is Hadoop
    b. Current data challenges
    c. Big Data and its aspects
    d. Difference between Big Data and Traditional Framework
    e. Example: where to use Big Data
    f. Big Data Business Opportunities
2. **Hadoop and HDFS Architecture**
    a. Hadoop architecture
    b. Hadoop Core Components
    c. HDFS Architecture
    d. Modes of Installation
    e. Hadoop Daemons
    f. Rack Awareness
    g. Data Replication
3. **Overview of Linux and Hadoop Installation**
    a. Installation of VMware and Linux
    b. Basic commands of Linux
    c. Hadoop Installation and configuration
    d. Practical Examples
    e. Hadoop Admin commands
4. **Map Reduce implementation and Hadoop Ecosystem**
    a. Map Reduce Flowchart

b. Map Reduce Demons and Architecture
c. Understanding of sorting and shuffling phase
d. Map slot and Reducer slot in cluster
e. Writing and executing Map Reduce Program
f. Map Ecosystem Architecture

5. PIG Implementation

a. Introduction to Pig
b. Installation to Pig
c. Data Loading and Data Extraction in Pig
d. Data Transformation in Pig
e. Understanding of functions/methods in Pig
f. Practical exercise for working with Pig

6. HIVE Implementation

a. Introduction to Hive
b. Installation to Hive
c. Overview of Hive Query Language
d. DDL and DML manipulations in Hive
e. Partitions in Hive
f. Practical exercise for working with Hive