DATA ANALYTICS ASSIGNMENT

HADOOP

INTRODUCTION :

Hadoop is an open-source software framework designed for storing and processing vast amounts of data within a distributed computing environment. It is specifically built to handle big data and operates on the MapReduce programming model, enabling parallel processing of large datasets. The framework primarily utilizes Java programming, supplemented by native code written in C and shell scripts.

1.1 HISTORY OF HADOOP :

Hadoop was developed by the Apache Software Foundation, with Doug Cutting and Mike Cafarella as its co-founders. Doug Cutting named the project after his son’s toy elephant. The journey began in October 2003 with the release of the first paper on the Google File System. In January 2006, development on MapReduce started, integrated into Apache Nutch, involving around 6,000 lines of code for MapReduce and 5,000 lines for the Hadoop Distributed File System (HDFS). By April 2006, Hadoop version 0.1.0 was officially released.

1.2 VERSIONS OF HADOOP:

Important versions of Hadoop are :

1. Hadoop 3.3.6 (August 2024)
2. Hadoop 3.3.5 (April 2024)
3. Hadoop 3.3.0 (June 2021)
4. Hadoop 3.2.2 (November 2020)
5. Hadoop 2.10.1 (October 2020)
6. Hadoop 2.9.2 (April 2019)
7. Hadoop 2.8.5 (October 2018)
8. Hadoop 2.7.7 (June 2018)
9. Hadoop 1.2.2 (June 2014)

1.3 SYSTEM REQUIREMENTS :

1. Hardware Requirements:

Single-Node Setup (for Development/Testing):

Processor: Dual-core or higher (64-bit architecture recommended).

RAM: Minimum 4 GB (8 GB or more recommended for smoother operation).

Disk Space: At least 25 GB of free disk space (more space is required based on your dataset size).

Network: A reliable internet connection (for downloading dependencies and updates).

Multi-Node Cluster Setup (for Production):

Processor: Multi-core processors (Quad-core or higher, 64-bit architecture).

RAM: Minimum 8 GB per node (16 GB or more per node is recommended).

Disk Space: At least 100 GB per node (preferably on SSDs for better performance).

Network: Gigabit Ethernet or higher for inter-node communication.

Number of Nodes: Depends on the scale; typically, a minimum of 3 nodes is recommended for redundancy and fault tolerance.

2. Software Requirements:

Operating System:

Linux: Preferred for production environments (e.g., Ubuntu, CentOS).

Windows: Supported for development and testing, but not recommended for production.

macOS: Can be used for development and testing.

Java Development Kit (JDK):

Version: JDK 8 or later.

1.4 INSTALLATION STEPS OF HADOOP :

Step 1: Java installation

**Java Development Kit (JDK):**

* Hadoop requires Java to run. Download and install the latest JDK from the [Oracle website](https://www.oracle.com/java/technologies/javase-downloads.html).
* Set the JAVA\_HOME environment variable to the path where Java is installed.

Step 2: Extract Hadoop

Extract the downloaded Hadoop to a directory like C:\Hadoop.

Step 3: Set Environment Variables

Add Hadoop bin directory to PATH:

Add C:\Hadoop\bin to your PATH environment variable.

Set Hadoop environment variables:

Create a new system variable HADOOP\_HOME and set it to C:\Hadoop.

Add HADOOP\_HOME\bin to your PATH.

Step 4: Editing Hadoop files

Create a folder data in the Hadoop directory and two sub folders , namenode and datanode.

These folders are important because files on HDFS reside inside these datanode.

Step 5 : Editing configuration files

Configure core-site.xml

Configure hdfs-site.xml

Configure mapred-site.xml

Configure yarn-site.xml

Configure Hadoop-env.cmd

1.5 INSTALLATION SCREENSHOTS:











