DATA ANALYTICS ASSIGNMENT

HADOOP

INTRODUCTION :

Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets . Its framework is based on Java programming with some native code in C and shell scripts.

1.1 HISTORY OF HADOOP :

**Apache Software Foundation** is the developers of Hadoop, and it’s co-founders are **Doug Cutting** and **Mike Cafarella**. It’s co-founder Doug Cutting named it on his son’s toy elephant. In October 2003 the first paper release was Google File System. In January 2006, MapReduce development started on the Apache Nutch which consisted of around 6000 lines coding for it and around 5000 lines coding for HDFS. In April 2006 Hadoop 0.1.0 was 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 :

Hardware Requirements:

Processor:

Minimum: Multi-core processor (e.g., Intel Core i3 or equivalent)

Recommended: Multi-core processor with higher clock speed (e.g., Intel Xeon or equivalent)

Memory (RAM):

Minimum: 4 GB per node

Recommended: 8 GB or more per node

Storage:

Minimum: At least 500 GB of disk space per node

Recommended: 1 TB or more per node, preferably with SSDs for better performance

Network:

Minimum: 1 Gbps Ethernet

Recommended: 10 Gbps Ethernet or higher for large clusters

Software Requirements:

Operating System:

Supported: Linux (various distributions like CentOS, Ubuntu, Debian), Unix

Unsupported: Windows is not officially supported, but Hadoop can run on Windows with some additional configuration

Java:

Minimum: Java 8 (for most Hadoop versions)

Recommended: Java 11 or later, depending on the Hadoop version

Dependencies:

Python: Some Hadoop components may require Python (typically version 2.7 or 3.x)

SSH: SSH should be configured for passwordless login between nodes in the cluster.

1.4 INSTALLATION STEPS:

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:













