**Big Data Analytics**

**BAD601**

**Experiment 1:**

Install Hadoop and Implement the following file management tasks in Hadoop:

Adding files and directories

Retrieving files

Deleting files and directories.

Hint: A typical Hadoop workflow creates data files (such as log files) elsewhere and copies them into HDFS using one of the above command line utilities.

* **Install Hadoop**

**Download and Install Java**

Hadoop requires Java to run. Install Java if you haven’t already.

Check Java Installation

Open Command Prompt (cmd) and type:

java -version

If Java is not installed, download and install Java JDK (8 or 11) from:

Official Oracle Java Download

OpenJDK Download

**Set Java Environment Variables:**

Go to Control Panel > System > Advanced System Settings > Environment Variables.

Under System Variables, click New and set:

Variable name: JAVA\_HOME

Variable value: C:\Program Files\Java\jdk-<version> (replace <version> with your installed JDK version).

Edit the Path variable and add:

%JAVA\_HOME%\bin

**Download and Configure Hadoop**

Download Hadoop binary for Windows from:

Apache Hadoop Releases

Extract the zip file to C:\hadoop or any directory.

Set Hadoop Environment Variables:

Add a new System Variable:

Variable name: HADOOP\_HOME

Variable value: C:\hadoop

Edit the Path variable and add:

%HADOOP\_HOME%\bin

Configure Hadoop for Windows:

Inside the Hadoop folder (C:\hadoop\etc\hadoop), edit these configuration files:

Edit core-site.xml:

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

Edit hdfs-site.xml:

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:///C:/hadoop/data/namenode</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:///C:/hadoop/data/datanode</value>

</property>

</configuration>

Edit mapred-site.xml:

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

Edit yarn-site.xml:

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

</configuration>

Format the Namenode (Run in Command Prompt):

hdfs namenode -format

**Start Hadoop Services**

Open Command Prompt (cmd) as Administrator.

Navigate to Hadoop directory:

cd C:\hadoop\sbin

Start Hadoop services:

start-dfs.cmd

start-yarn.cmd

jps

Verify HDFS is running:

Open a browser and go to:

<http://localhost:9870>

<http://localhost:8088/>

http://localhost:9870/explorer.html

You should see the Hadoop Web UI.

**Task 1: Adding Files and Directories to HDFS**

**hdfs dfs -mkdir /mydata**

**hdfs dfs -put C:\Users\YourUser\Documents\sample.txt /mydata**

**Explanation:**

The -mkdir command creates a directory inside HDFS (/mydata).

The -put command uploads a local file (sample.txt) to HDFS (/mydata).

**Steps:**

You first create a directory in HDFS where your data will be stored.

Then, you upload the data (for example, log files or text files) into that directory.

**Task 2: Retrieving Files from HDFS**

**hdfs dfs -get /mydata/sample.txt C:\Users\YourUser\Downloads**

**Explanation:**

The -get command downloads a file from HDFS to your local directory.

Here, the file sample.txt from /mydata is downloaded to the local directory C:\Users\YourUser\Downloads.

**Steps:**

You can use this command to retrieve files after processing or analysis on Hadoop.

**Task 3: Deleting Files and Directories in HDFS**

**hdfs dfs -rm /mydata/sample.txt**

Command (for directories):

**hdfs dfs -rm -r /mydata**

**Explanation:**

The -rm command deletes a specific file (sample.txt) from HDFS.

The -rm -r command recursively deletes a directory (/mydata) and all of its contents.

**Steps:**

Use this to clean up data after processing or if the directory is no longer needed.