EXPERIMENT NO. 2

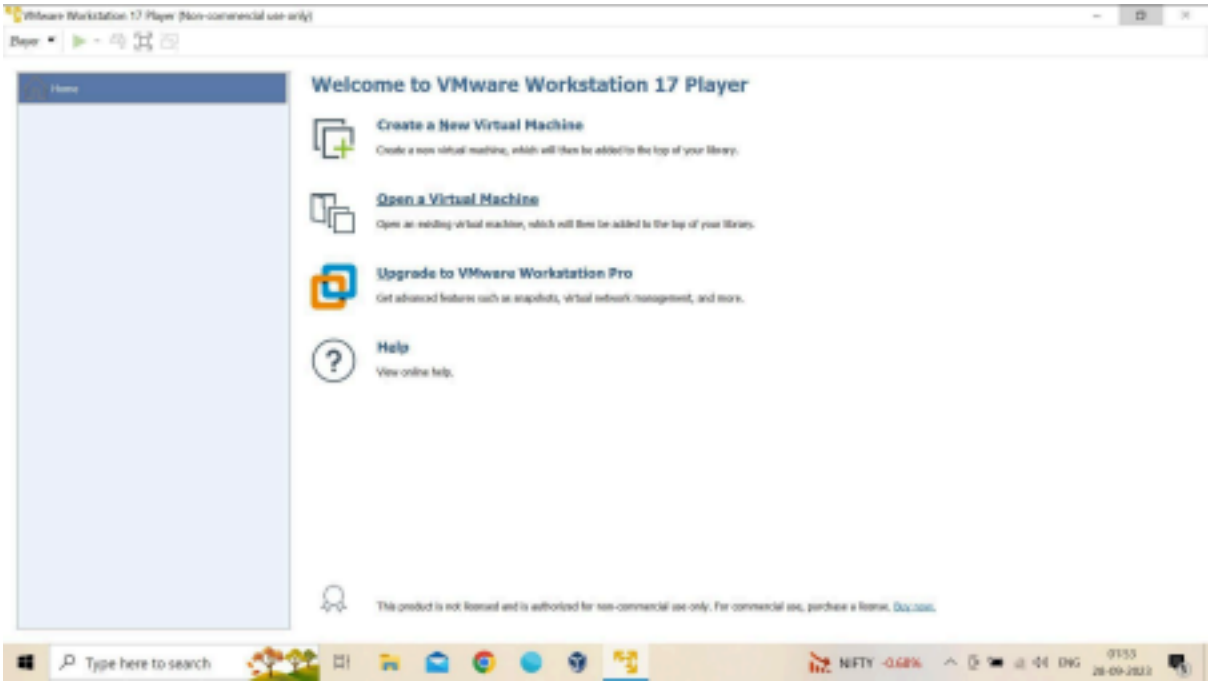
Aim: Installation of Hadoop on a single-node cluster

Post-lab questions: Explain the architecture of Hadoop in detail.

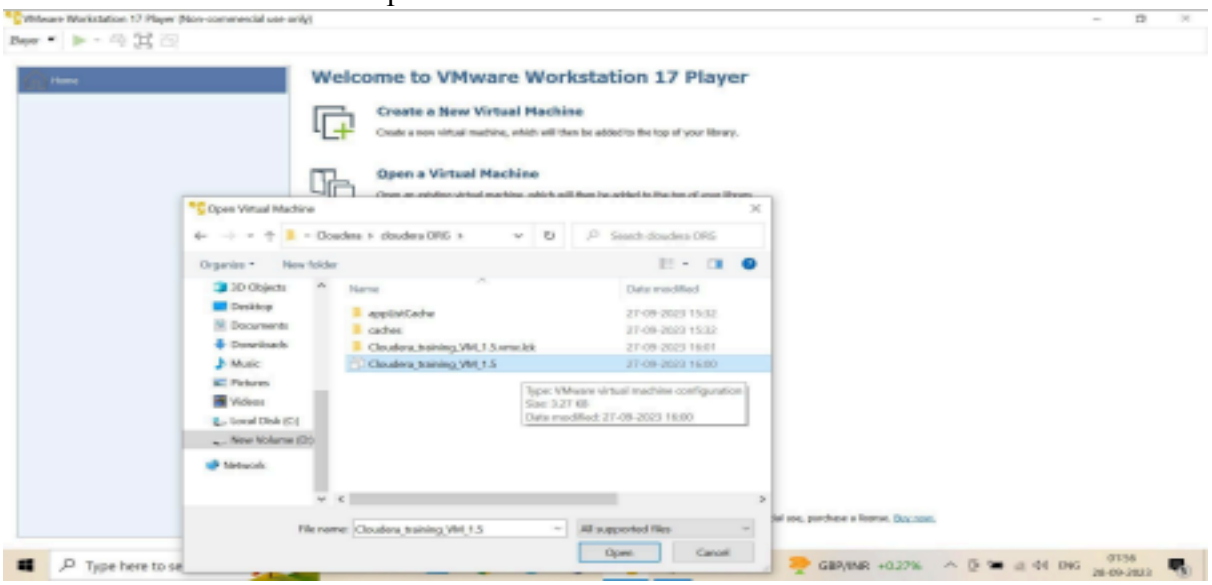
Two methods to install Hadoop on Windows OS.

Installation of Hadoop using Cloudera 1.5 System.

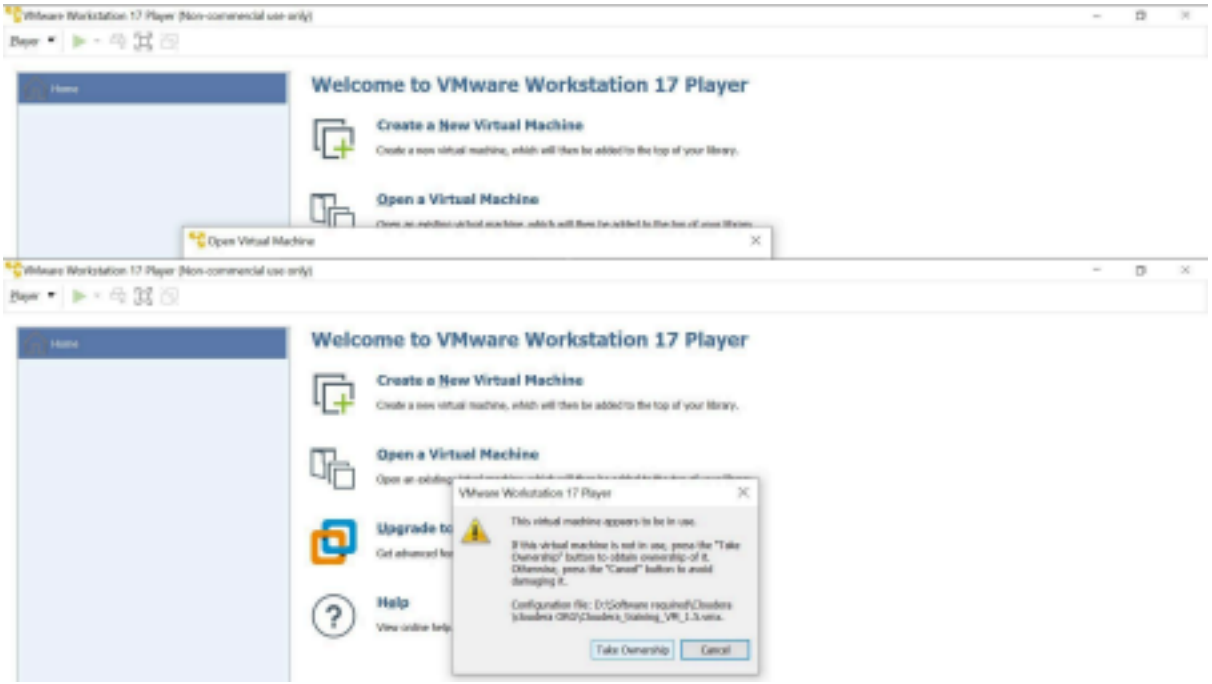
1. Click on Open a Virtual Machine.



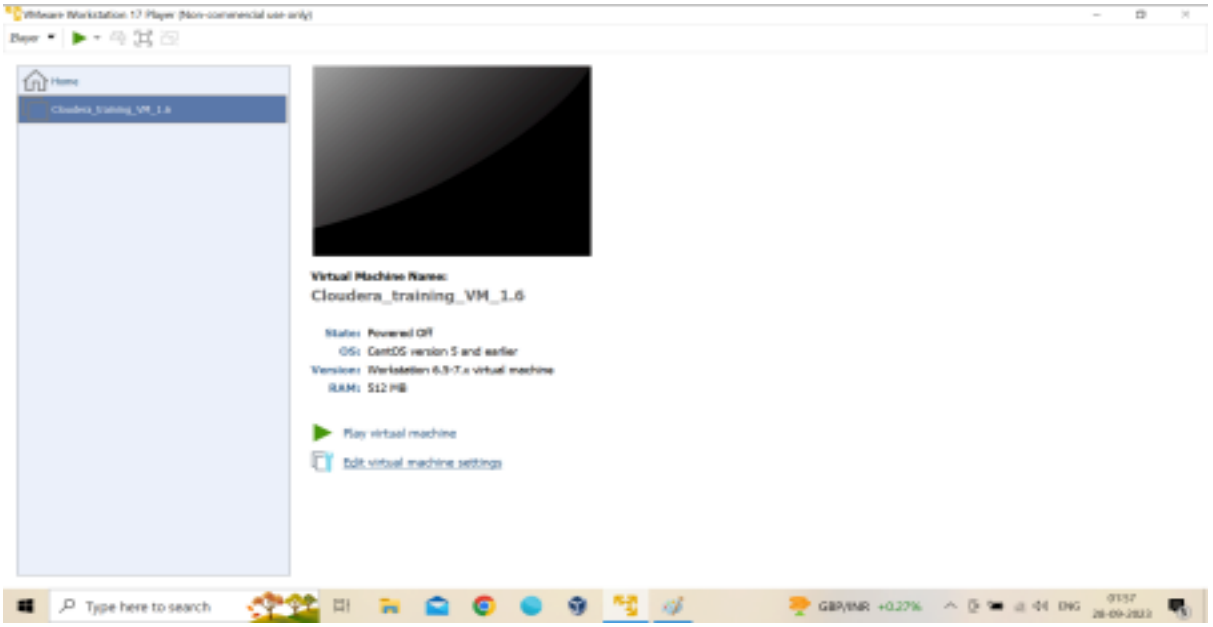
2. Select the Cloudera 1.5 setup file.



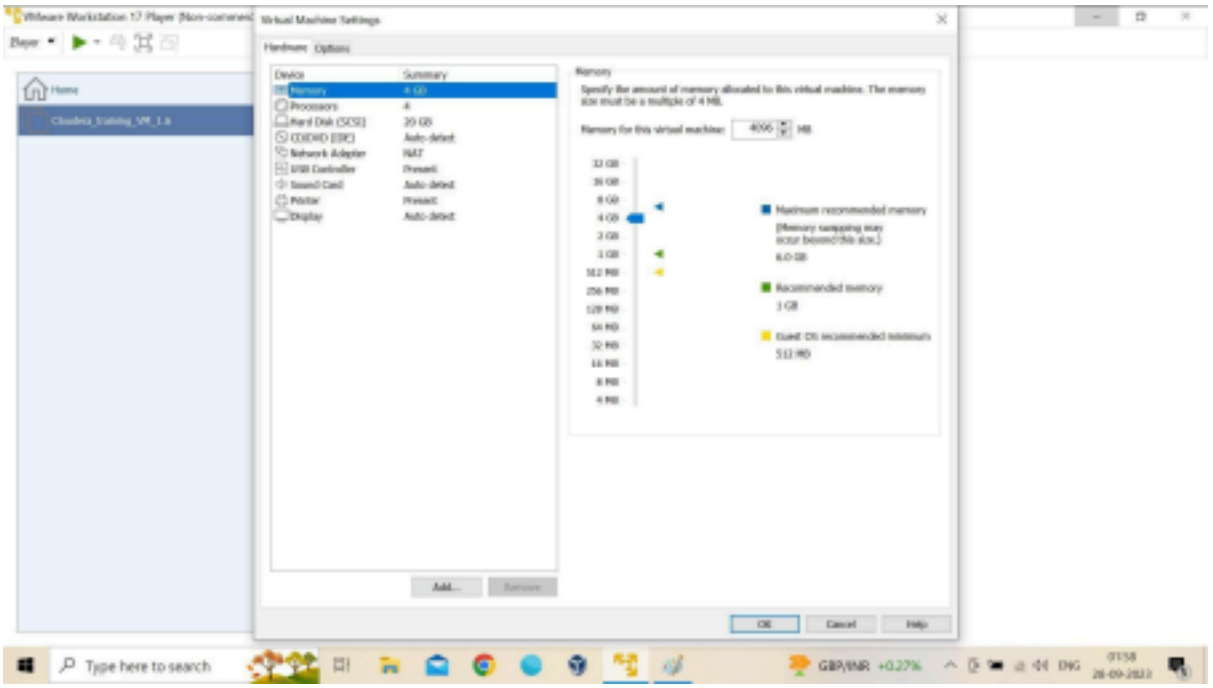
3.Once selection done next click on the Take Ownership button.



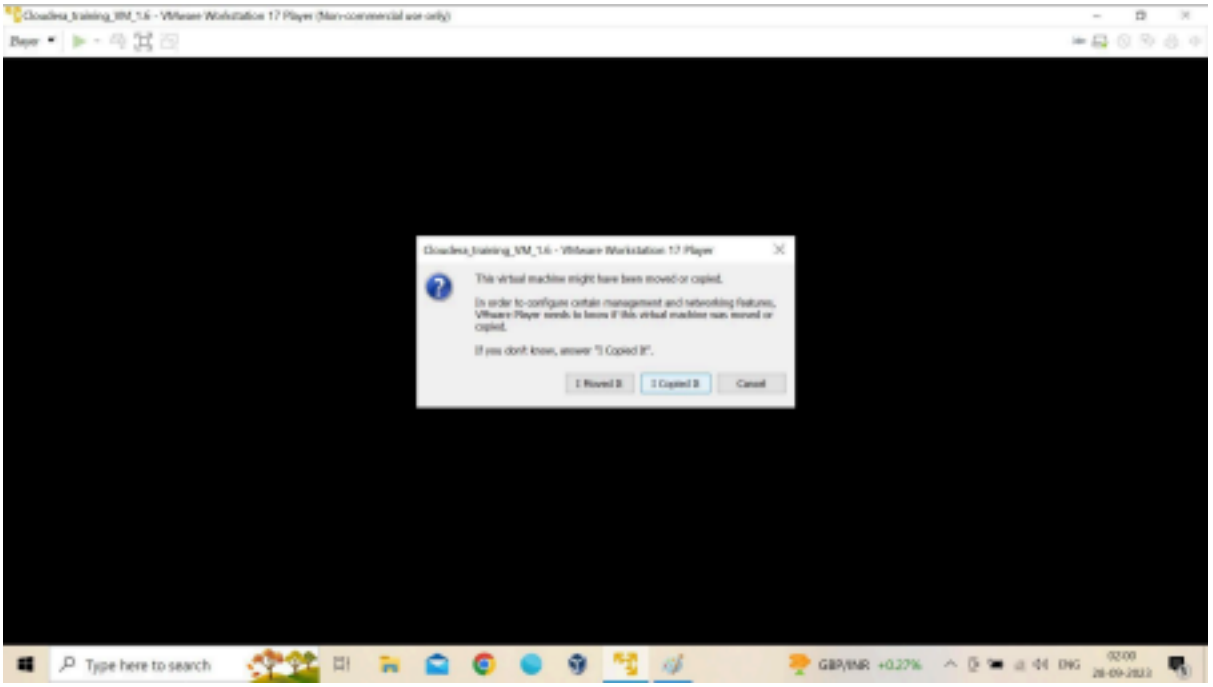
4.Click on the Edit Virtual Machine option.



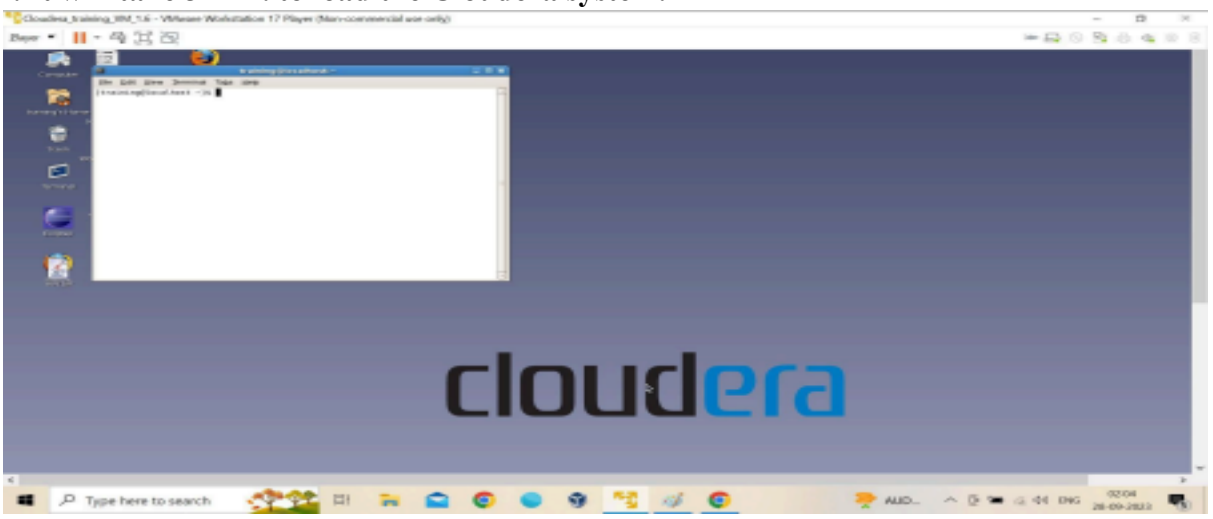
5.Edit Ram-> 4GB and Processor →2 core.. Click on ok button



6. Click on the I Copied It button.



7. It will take 3 min. to load the Cloudera system.



Setup Hadoop on Windows 10 machines

Required tools

1. Java JDK - used to run the Hadoop since it's built using Java
2. 7Zip or WinRAR - unzip Hadoop binary package; anything that unzips tar.gz 3. CMD or Powershell - used to test environment variables and run Hadoop

Step 1 - Download and extract Hadoop

1. Download Hadoop: Hadoop 3.2.4.

2. If there are permission errors, run your unzipping program as administrator and unzip again.

Step 2 - Install Java JDK

Java JDK is required to run Hadoop, so if you haven't installed it, install it.

1. download it from for example here (JDK 8u261).
2 .Run the installation file and the default installation directory will be C:\Program Files\Java\jdk1.8.0_202\bin

Step 3 - Configure environment variables

1. Open the Start Menu type in 'environment' and press enter.
2. A new window with System Properties should open up.
3. Click the Environment Variables button near the bottom right.

4.JAVA_HOME environment variable

1. From step 3, find the location of where you installed Java. In this example, the default directory is C:\Program Files\Java\jdk1.8.0_202
2. Create a new User variable with the variable name as JAVA_HOME and the value as C:\Program Files\Java\jdk1.8.0_202\bin

5.System Variable

1.Select Path →Edit→ add C:\Program Files\Java\jdk1.8.0_202

6.**After installation, open up CMD or Powershell and confirm Java is installed:** $ java -version

```
java version "1.8.0_261"
Java(TM) SE Runtime Environment (build 1.8.0_261-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.261-b12, mixed mode)
```

7.HADOOP_HOME environment variable

1. From step 1, copy the directory you extracted the Hadoop binaries to. In this example, the directory is C:\hadoop-3.2.1\bin
2. Create a new User variable with the variable name as HADOOP_HOME and the value as C:\hadoop-3.2.1\bin

8.System Variable

1.Select Path →Edit→ add C:\hadoop-3.2.1\bin

9.Hadoop environment

etc\hadoop\hadoop-env.cmd

Hadoop complains about the directory if the JAVA_HOME directory has spaces. In the default installation directory, Program Files has a problematic space. To fix this, open the %HADOOP_HOME%\etc\hadoop\hadoop-env.cmd and change the JAVA_HOME line to the following:

set JAVA_HOME=C:\PROGRA~1\Java\jdk1.8.0_261.

10.After setting those environment variables, you reopen CMD or Powershell and verify that the hadoop command is available:

$ hadoop -version

java version "1.8.0_261"
Java(TM) SE Runtime Environment (build 1.8.0_261-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.261-b12, mixed mode)

Step 4. Editing Hadoop files

Once we have configured the environment variables, the next step is to configure Hadoop.

1 Creating Folders

We need to create a folder data in the hadoop directory, and 2 sub folders namenode and datanode

- Once DATA folder is created, we need to create 2 new folders namely, **namenode and datanode** inside the data folder
- These folders are important because files on **HDFS reside inside the datanode.**

2.**Editing Configuration Files**

Now we need to edit the following config files in hadoop for configuring it :- (We can find these files in Hadoop -> etc -> hadoop)

* core-site.xml
* hdfs-site.xml
* mapred-site.xml
* yarn-site.xml

2.1 Editing core-site.xml

1. This file informs Hadoop daemon where NameNode runs in the cluster. 2. It contains the configuration settings for Hadoop Core such as I/O settings that are common to HDFS and MapReduce.
3. Location of namenode is specified by fs.defaultFS property
4. namenode running at 9000 ports on localhost.

```
<configuration>
<property>
 <name>fs.defaultFS</name>
 <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

2.2 Editing hdfs-site.xml (Note: Put address of Namenode and datanode)

1. It is one of the important configuration files which is required for runtime environment settings of a Hadoop.
2. It contains the configuration settings for NAMENODE, DATANODE, SECONDARY NODE.
3. Replication factor is specified by dfs.replication property;
4. it is a single node cluster hence we will set replication to 1.
5. *Also replace PATH~1 and PATH~2 with the path of the namenode and datanode folder that we created recently(step 4.1).*

```
<configuration>
<property>
 <name>dfs.replication</name>
 <value>1</value>
</property>
<property>
 <name>dfs.namenode.name.dir</name>
   <value>C:\hadoop-3.2.4\Data folder\namenode</value>
 <final>true</final>
</property>
<property>
 <name>dfs.datanode.data.dir</name>
   <value>C:\hadoop-3.2.4\Data folder\datanode</value>
 <final>true</final>
</property>
</configuration>
```

2.3 Editing mapred-site.xml

1. It contains the configuration settings for MapReduce .
2. In this file, we specify a framework name for MapReduce, by setting the MapReduce.framework.name.

```
<configuration>
<property>
   <name>mapreduce.framework.name</name>
 <value>yarn</value>
</property>
</configuration>
```

2.4 Editing yarn-site.xml

1. In order to specify an auxiliary service that needs to run with nodemanager“yarn.nodemanager.aux-services” property is used.
2. Here Shuffling is used as an auxiliary service. And in order to know the class that should be used for shuffling we user “yarn.nodemanager.aux-services.mapreduce.shuffle.class”


```
<configuration>
<property>
  <name>yarn.nodemanager.aux-services</name>
 <value>mapreduce_shuffle</value>
</property>
<property>

<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<!-- Site specific YARN configuration properties -->
</configuration>
```

Step 5. Replacing bin

Replace Bin folder
Delete old bin file and replace above given link bin file.

Note:- If you are using different version of Hadoop then please search for its respective bin folder and download it

Step 6.Testing Setup

6.1 Formatting Namenode
Before starting hadoop we need to format the namenode for this we need to start a NEW Command Prompt and run below command

C:\Users\admin> hadoop namenode -format

Note:- This command formats all the data in namenode. So, it's advisable to use only at the start and do not use it every time while starting a Hadoop cluster to avoid data loss.

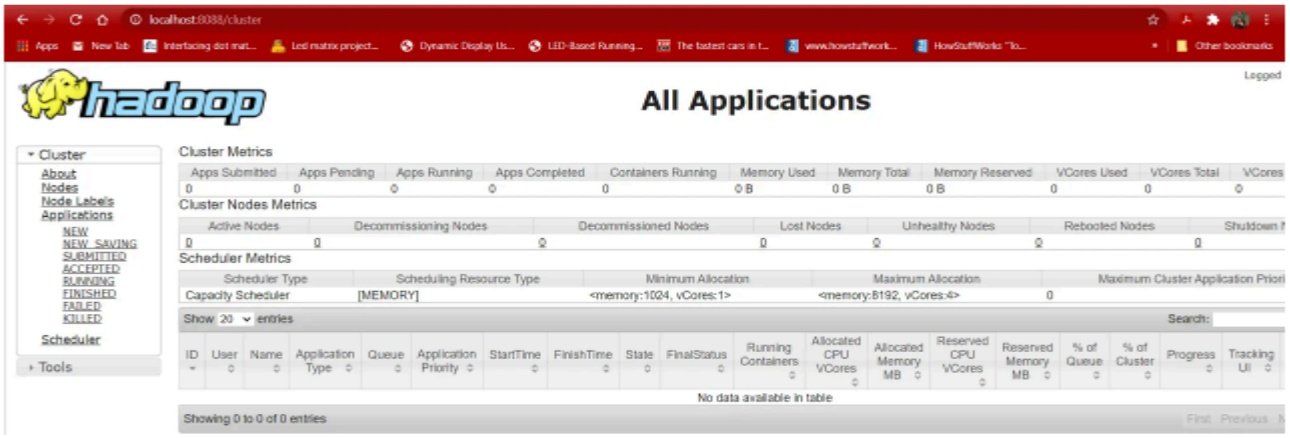
6.2 Launching Hadoop
Now we need to start a new Command Prompt remember to run it as administrator to avoid

permission issues and execute below commands

C:\>cd hadoop-3.2.4
C:\hadoop-3.2.4>cd sbin
C:\hadoop-3.2.4\sbin>*start-all.cmd*
4 Windows get open
or
C:\hadoop-3.2.4\sbin>start-dfs.cmd
2 Windows get open
C:\hadoop-3.2.4\sbin>start-yarn.cmd
2 windows get open

Step7.Enter in the URL

Localhost:8088/9870 where you can get the status of your dfs, name node and datanode.



For data node: portno:9870.
For resource manager: 8088

Link:https://www.youtube.com/watch?v=g7Qpnmi0Q-s&t=940s How to Install Hadoop on Windows 10 | Easy Steps to Install Hadoop | Hadoop Tutorial | Edureka

Connection refused

- Stop the service.
- Check properties in configuration files like **core-site.xml, hdfs-site.xml, hadoop-env.sh file**.
- Write your namenode IP in **/etc/hosts** file.
- Format namenode.
- Stop firewall.
- Start your service again.

Conclusion: hence we study how two install Hadoop on Window system