EX.NO: 1 **INSTALLATION OF HADOOP**

DATE:

AIM:

To Download and install Hadoop; Understand different Hadoop modes. Startup scripts, Configuration files.

THEORY:

Hadoop is a Java-based programming framework that supports the processing and storage of extremely large datasets on a cluster of inexpensive machines. It was the first major open-source project in the big data playing field and is sponsored by the Apache Software Foundation.

Hadoop-2.8.0 is comprised of four main layers:

- **Hadoop Common** is the collection of utilities and libraries that support other Hadoop modules.
- **HDFS**, which stands for Hadoop Distributed File System, is responsible for persisting data to disk.
- **YARN**, short for Yet Another Resource Negotiator, is the "operating system" for HDFS.
- **MapReduce** is the original processing model for Hadoop clusters. It distributes work within the cluster or map, then organizes and reduces the results from the nodes into a response to a query. Many other processing models are available for the 2.x version of Hadoop.

Hadoop clusters are relatively complex to set up, so the project includes a stand-alone mode which is suitable for learning about Hadoop, performing simple operations, and debugging.

PREPARE:

These softwares should be prepared to install Hadoop 2.8.0 on window 10 64bit

1. Download Hadoop 2.8.0
2. Java JDK 1.8.0.zip

PROCEDURE:

Procedure to Run Hadoop

1. Install Apache Hadoop 2.8.0 in Microsoft Windows OS

   If Apache Hadoop 2.8.0 is not already installed then follow the post Build, Install, Configure and Run Apache Hadoop 2.8.0 in Microsoft Windows OS.

2. Start HDFS (Namenode and Datanode) and YARN (Resource Manager and Node Manager)

Run following commands.

Command Prompt

C:\Users\> hdfs namenode –format

C:\hadoop\sbin>start-dfs

C:\hadoop\sbin>start-yarn

C:\hadoop\sbin>start-all.cmd

C:\hadoop\sbin>jps (used to check how many nodes are running in background of Hadoop)

Namenode, **Datanode**, **Resource Manager** and **Node Manager** will be started in few minutes and ready to execute Hadoop **MapReduce** job in the Single Node (pseudo-distributed mode) cluster.

PREREQUISITES:

Step1: Installing Java 8 version.

Openjdk version "1.8.0_91"

OpenJDK Runtime Environment (build 1.8.0_91-8u91-b14-3ubuntu1~16.04.1-b14)

OpenJDK 64-Bit Server VM (build 25.91-b14, mixed mode)

This output verifies that OpenJDK has been successfully installed.

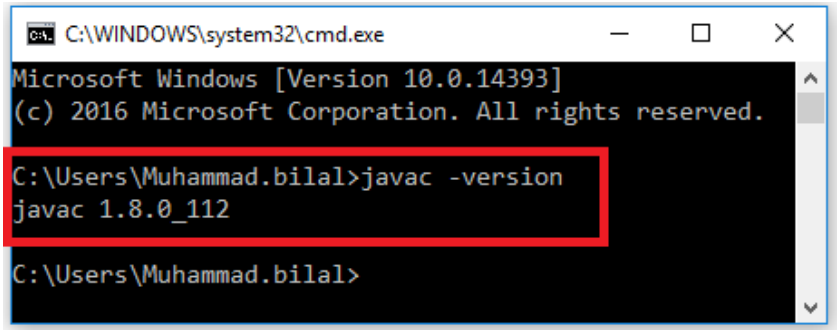
Note: To set the path for environment variables. i.e. JAVA_HOME

Step2: Installing Hadoop

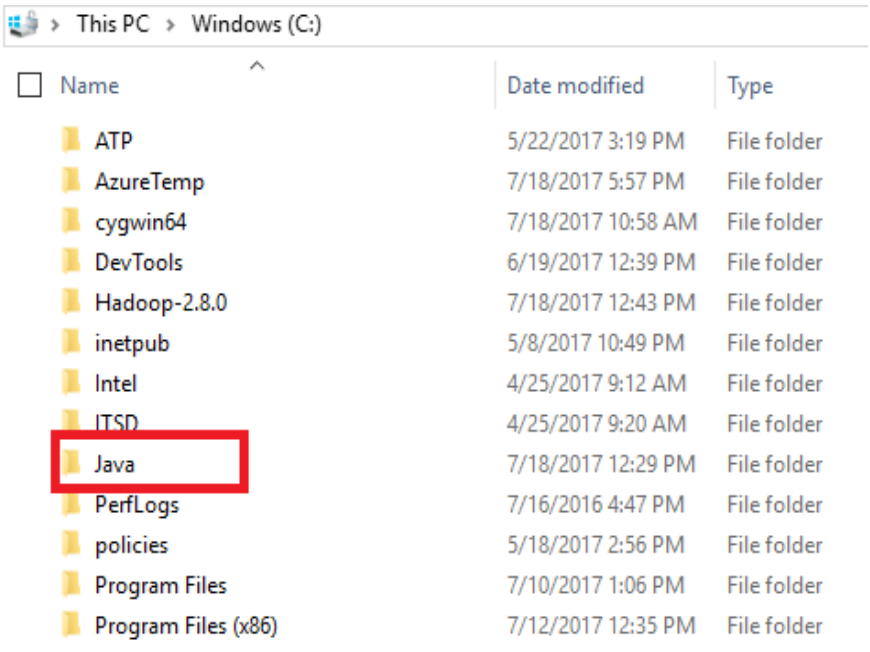
With Java in place, we'll visit the Apache Hadoop Releases page to find the mostrecent stable release. Follow the binary for the current release:

Set up

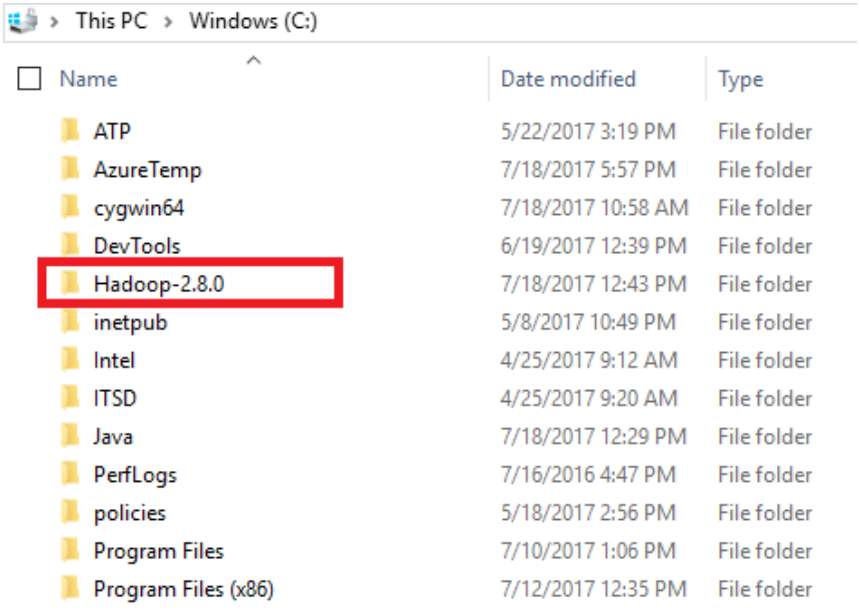
1. Check either Java 1.8.0 is already installed on your system or not, use **"Javac -version"** to check



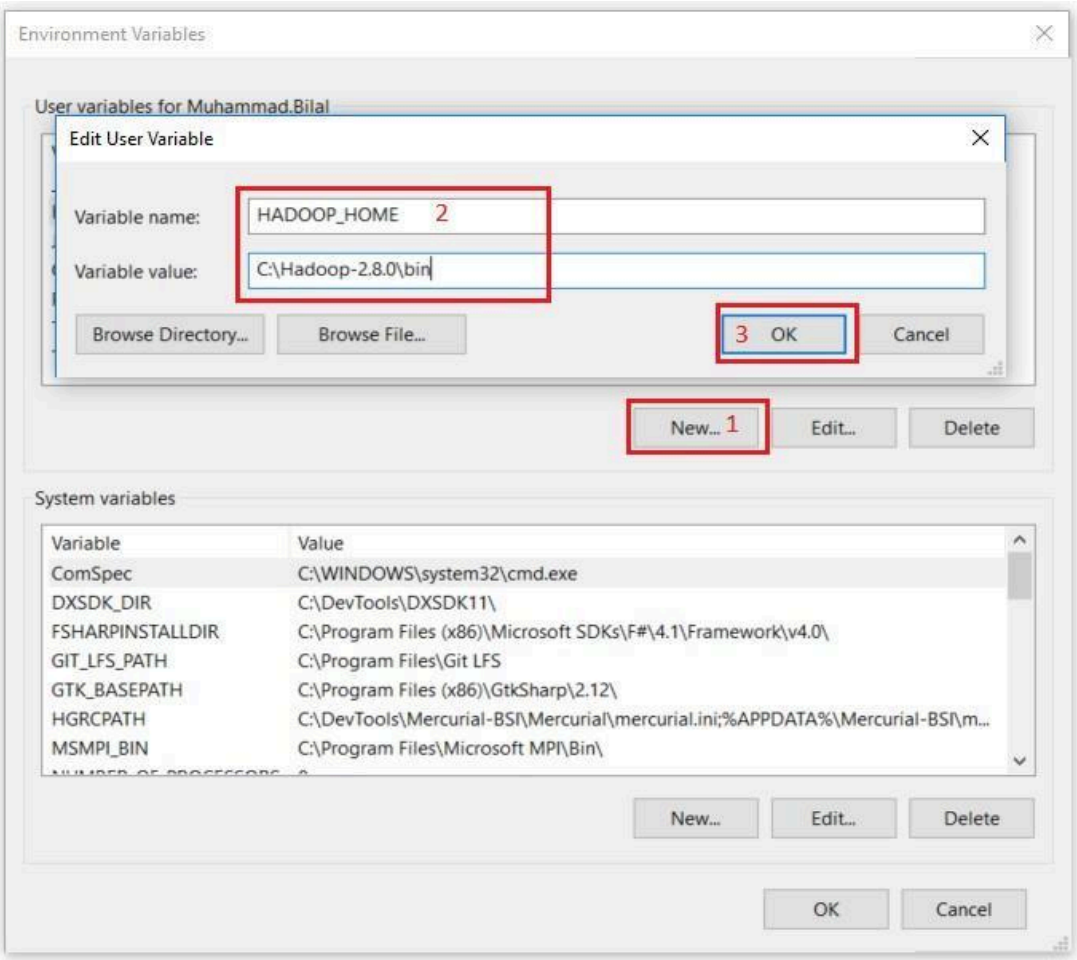
2. If Java is not installed on your system then first install java under **C:\JAVA**



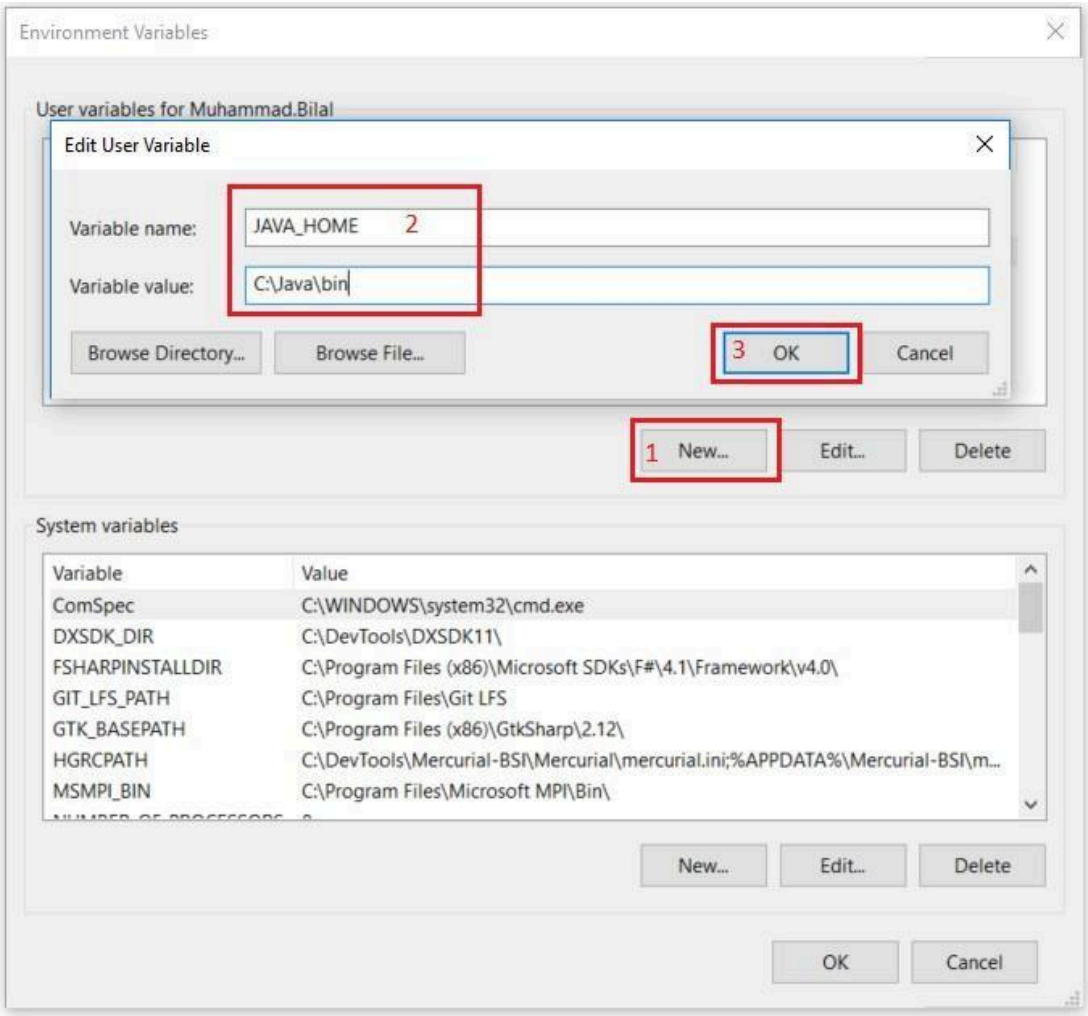
3. Extract file Hadoop 2.8.0.tar.gz or Hadoop-2.8.0.zip and place under **"C:\Hadoop-2.8.0"**



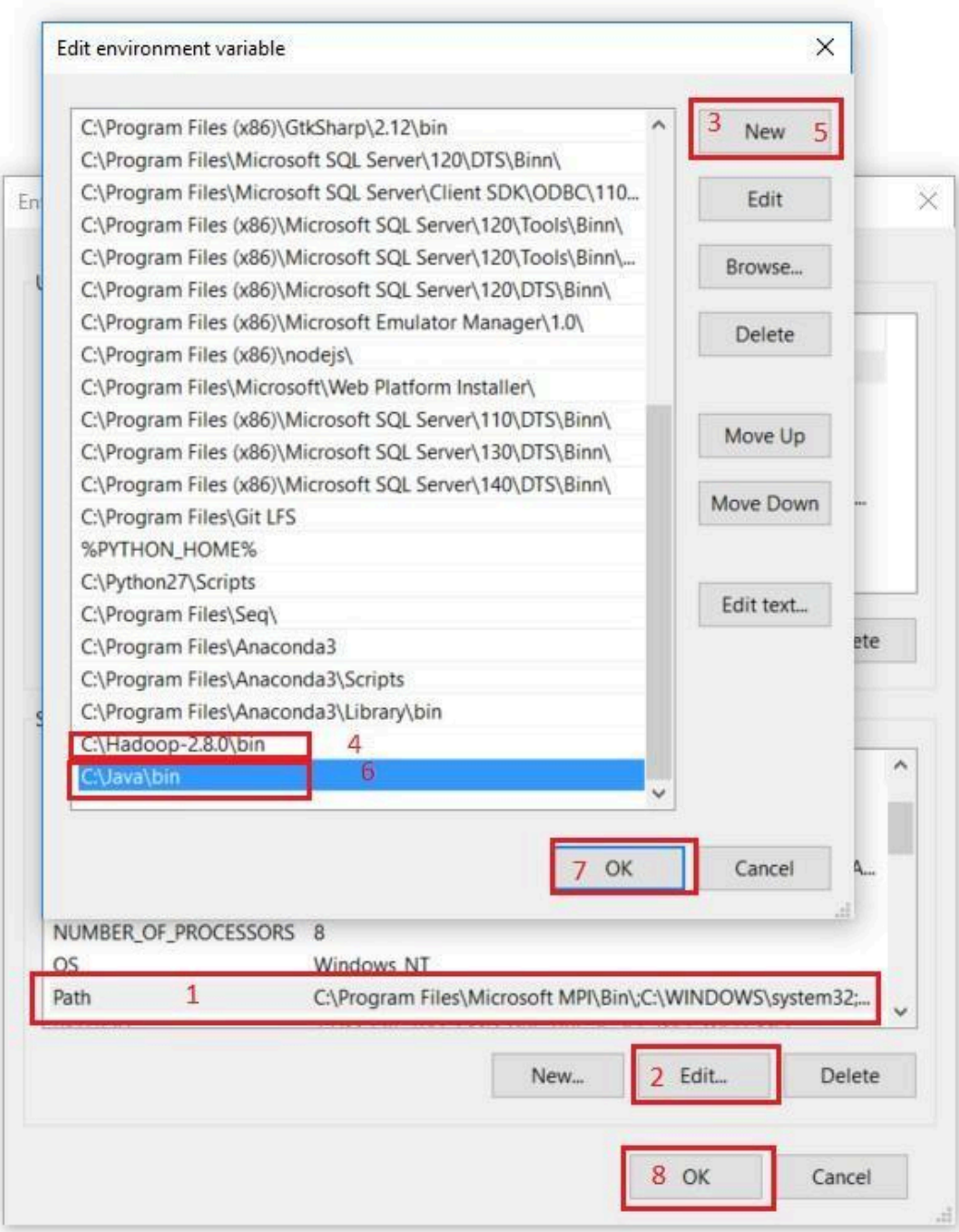
4. Set the path HADOOP_HOME Environment variable on windows 10(see Step 1,2,3 and 4 below)



5. Set the path JAVA_HOME Environment variable on windows 10(see Step 1,2,3 and 4 below)



6. Next, we set the Hadoop bin directory path and JAVA bin path



CONFIGURATION

1. Edit file **C:/Hadoop-2.8.0/etc/hadoop/core-site.xml**, paste below xml paragraph and save this file.

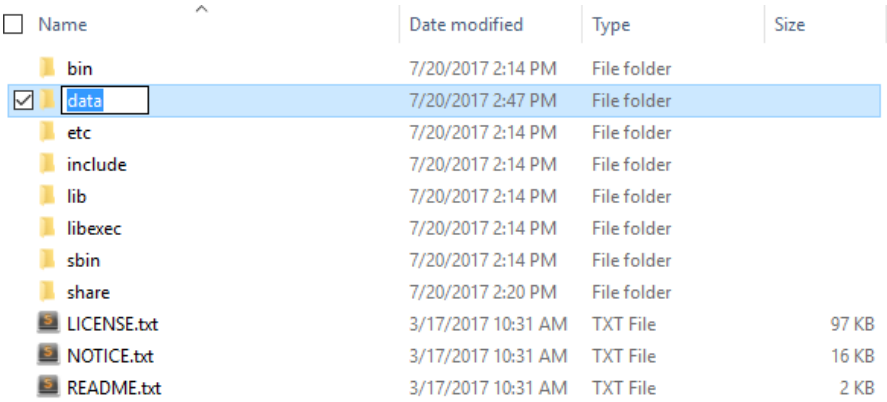
```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

2. Rename "mapred-site.xml.template" to "mapred-site.xml" and edit this file **C:/Hadoop-2.8.0/etc/hadoop/mapred-site.xml**, paste below xml paragraph and save this file.

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

3. Create folder **"data"** under **"C:\Hadoop-2.8.0"**

- Create folder **"datanode"** under **"C:\Hadoop-2.8.0\data"**
- Create folder **"namenode"** under **"C:\Hadoop-2.8.0\data"**



Name	Date modified	Type	Size
bin	7/20/2017 2:14 PM	File folder	
data	7/20/2017 2:47 PM	File folder	
etc	7/20/2017 2:14 PM	File folder	
include	7/20/2017 2:14 PM	File folder	
lib	7/20/2017 2:14 PM	File folder	
libexec	7/20/2017 2:14 PM	File folder	
sbin	7/20/2017 2:14 PM	File folder	
share	7/20/2017 2:20 PM	File folder	
LICENSE.txt	3/17/2017 10:31 AM	TXT File	97 KB
NOTICE.txt	3/17/2017 10:31 AM	TXT File	16 KB
README.txt	3/17/2017 10:31 AM	TXT File	2 KB

4. Edit file **C:\Hadoop-2.8.0/etc/hadoop/hdfs-site.xml**, paste below xml paragraph and save this file.

```
<configuration>
  <property>
    <name>dfs.replication</name>
```

```
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/hadoop-2.8.0/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/hadoop-2.8.0/data/datanode</value>
  </property>
</configuration>
```

5. Edit file **C:/Hadoop-2.8.0/etc/hadoop/yarn-site.xml**, paste below xml paragraph and save this file.

```
<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>
</configuration>
```

6. Edit file **C:/Hadoop-2.8.0/etc/hadoop/hadoop-env.cmd** by closing the command line **"JAVA_HOME=%JAVA_HOME%"** instead of set **"JAVA_HOME=C:\Java"** (On C:\java this is path to file jdk.18.0)

```
@rem The java implementation to use.  Required.
@rem set JAVA_HOME=%JAVA_HOME%
set JAVA_HOME=C:\java
```

Hadoop Configuration

1. Dowload file Hadoop Configuration.zip
2. Delete file bin on C:\Hadoop-2.8.0\bin, replaced by file bin on file just download (from Hadoop Configuration.zip).
3. Open cmd and typing command **"hdfs namenode –format"** . You will see

```
C:\WINDOWS\system32\cmd.exe
17/07/20 15:38:21 INFO util.GSet: VM type       = 64-bit
17/07/20 15:38:21 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
17/07/20 15:38:21 INFO util.GSet: capacity      = 2^18 = 262144 entries
17/07/20 15:38:21 INFO namenode.FSNamesystem: dfs.namenode.safemode.threshold-pct = 0.9990000128746033
17/07/20 15:38:21 INFO namenode.FSNamesystem: dfs.namenode.safemode.min.datanodes = 0
17/07/20 15:38:21 INFO namenode.FSNamesystem: dfs.namenode.safemode.extension     = 30000
17/07/20 15:38:21 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
17/07/20 15:38:21 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
17/07/20 15:38:21 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
17/07/20 15:38:21 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
17/07/20 15:38:21 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time
is 600000 millis
17/07/20 15:38:21 INFO util.GSet: Computing capacity for map NameNodeRetryCache
17/07/20 15:38:21 INFO util.GSet: VM type       = 64-bit
17/07/20 15:38:21 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
17/07/20 15:38:21 INFO util.GSet: capacity      = 2^15 = 32768 entries
17/07/20 15:38:26 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1080504939-192.168.68.1-1500547106632
17/07/20 15:38:26 INFO common.Storage: Storage directory C:\Hadoop-2.8.0\data\namenode has been successfully formatted.
17/07/20 15:38:26 INFO namenode.FSImageFormatProtobuf: Saving image file C:\Hadoop-2.8.0\data\namenode\current\fsimage.c
kpt_0000000000000000000 using no compression
17/07/20 15:38:26 INFO namenode.FSImageFormatProtobuf: Image file C:\Hadoop-2.8.0\data\namenode\current\fsimage.ckpt_000
0000000000000000 of size 330 bytes saved in 0 seconds.
17/07/20 15:38:26 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
17/07/20 15:38:26 INFO util.ExitUtil: Exiting with status 0
17/07/20 15:38:26 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at [redacted]
************************************************************/

C:\Users\Muhammad.bilal>
```

Testing

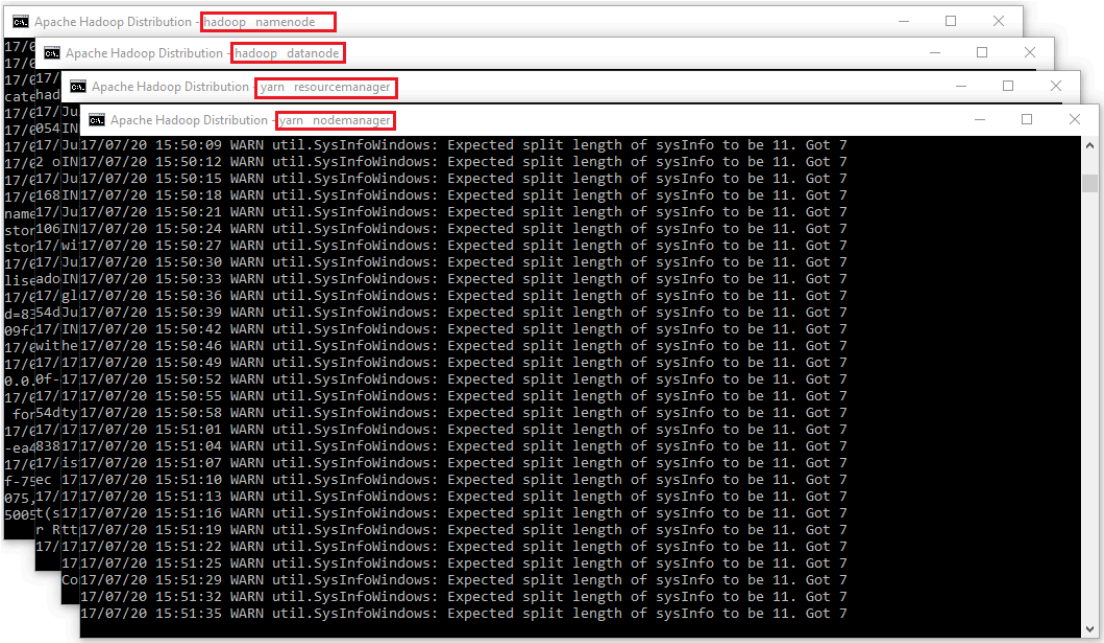
1. Open cmd and change directory to "C:\Hadoop-2.8.0\sbin" and type **"start-all.cmd"** to start Hadoop

```
Select C:\WINDOWS\system32\cmd.exe
C:\>cd Hadoop-2.8.0\sbin

C:\Hadoop-2.8.0\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Hadoop-2.8.0\sbin>
```

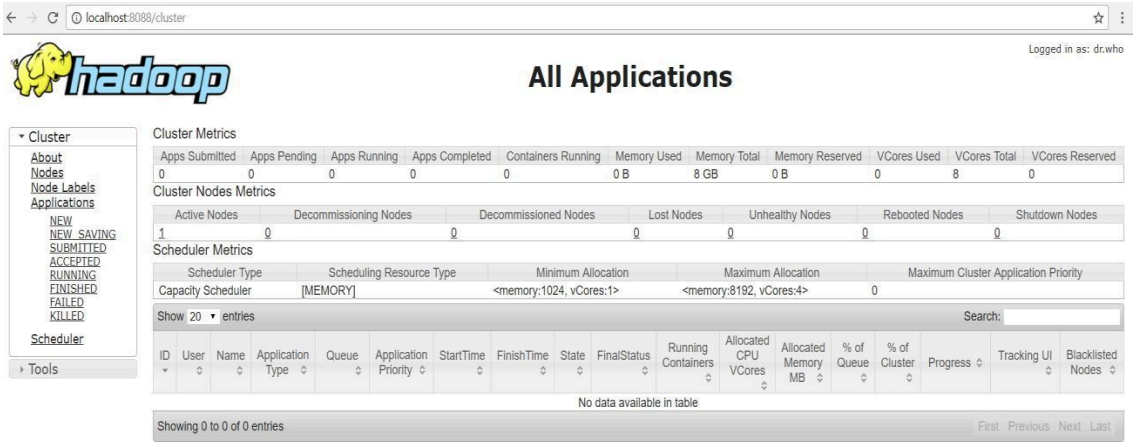
2. Make sure these apps are running
   - Hadoop Namenode
   - Hadoop datanode
   - YARN Resourc Manager
   - YARN Node Manager



3. OUTPUT:

Open:

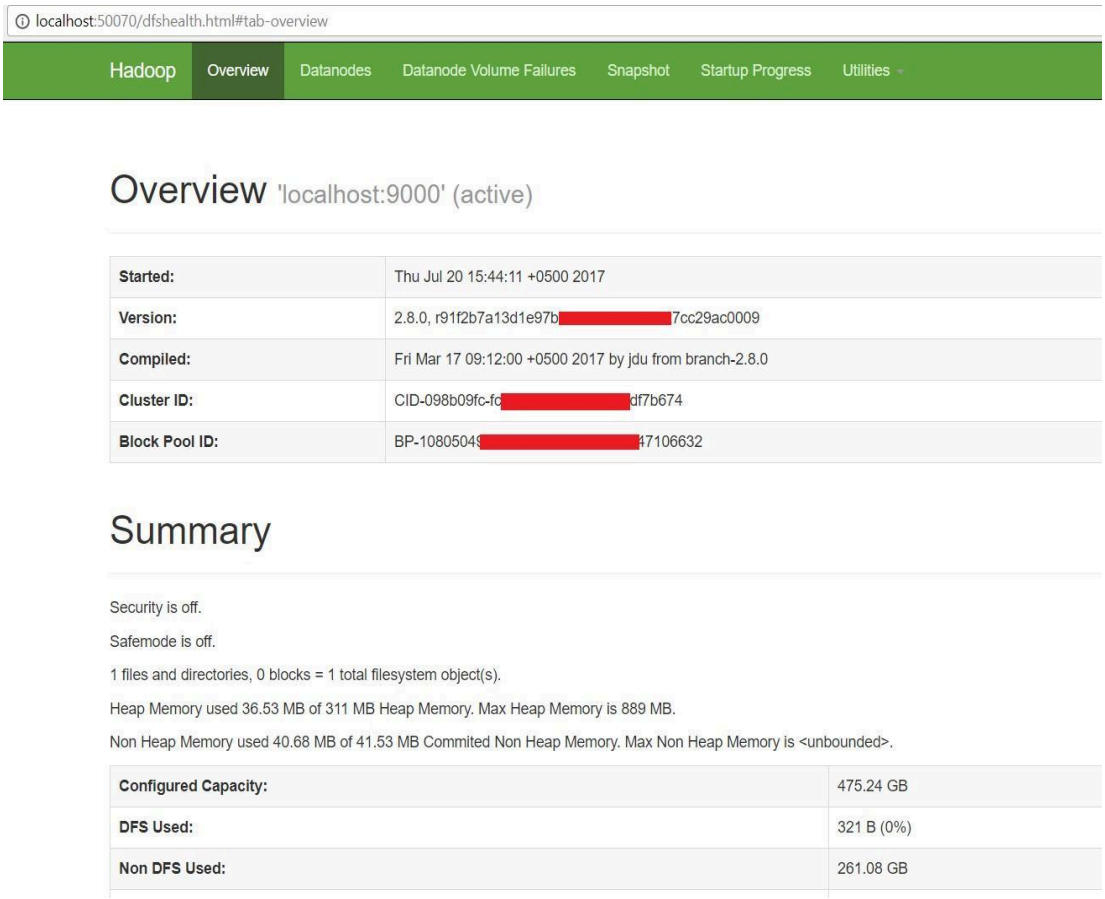
http://localhost:8088



4. **OUTPUT:**

Open:

http://localhost:50070



RESULT:

Thus, a procedure to installation of Hadoop cluster was successfully executed.