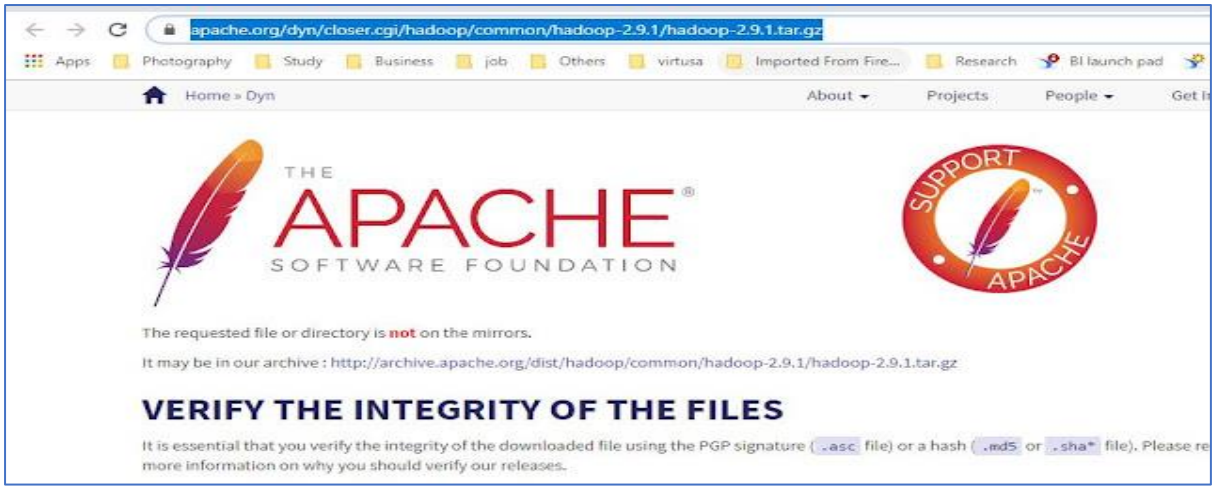
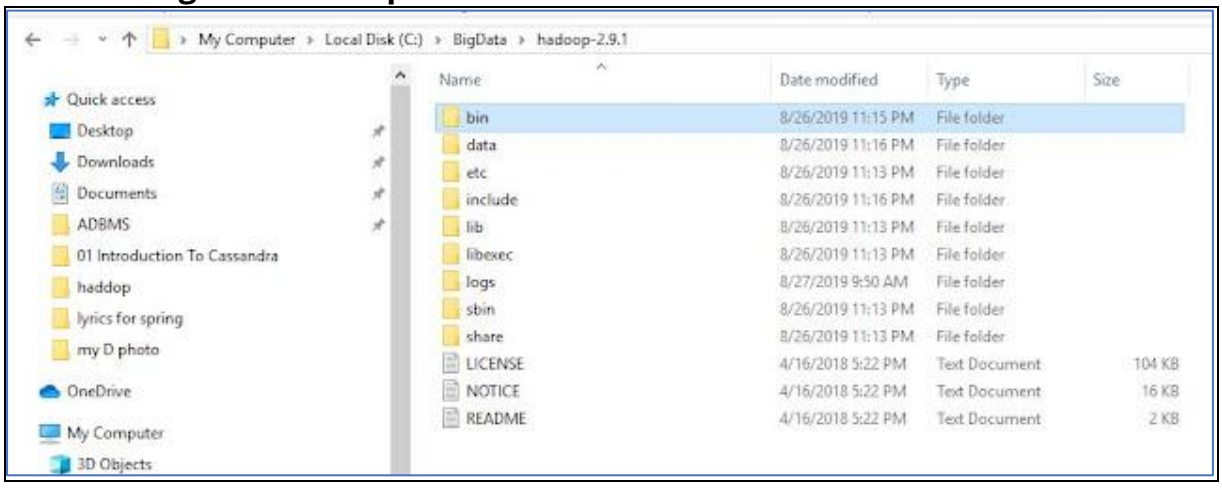
Install Hadoop 2.9.1 on Windows 10

Here, I am going to install Apache Hadoop 2.9.1 on Windows 10 platform.

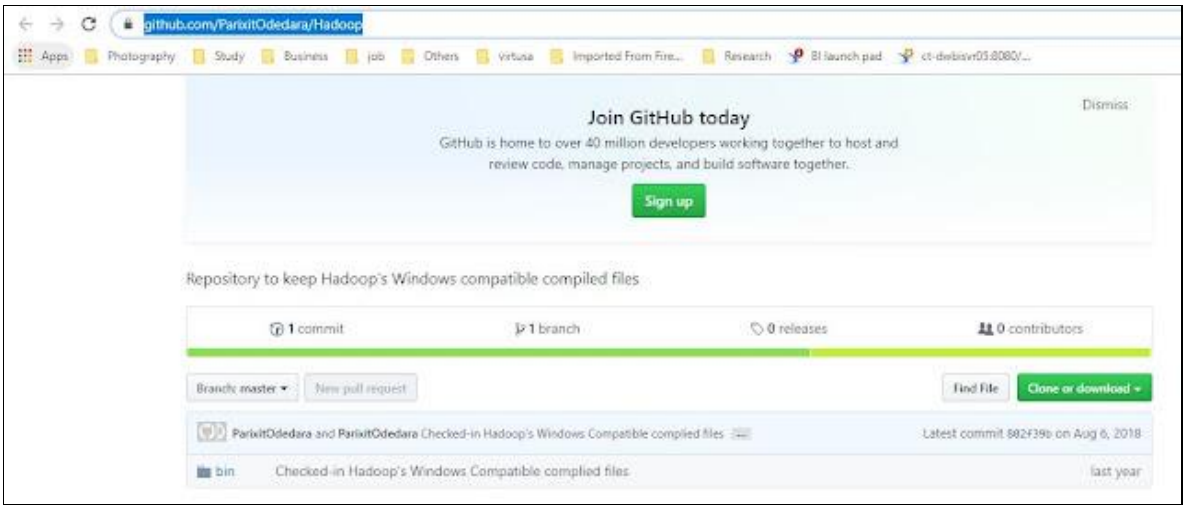
First download the **Hadoop 2.9.1** from the below link.
https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.9.1/hadoop-2.9.1.tar.gz



Create a folder path as below and copy the downloaded msi into this folder.
Path:- '**C:/BigData/hadoop-2.9.1**'

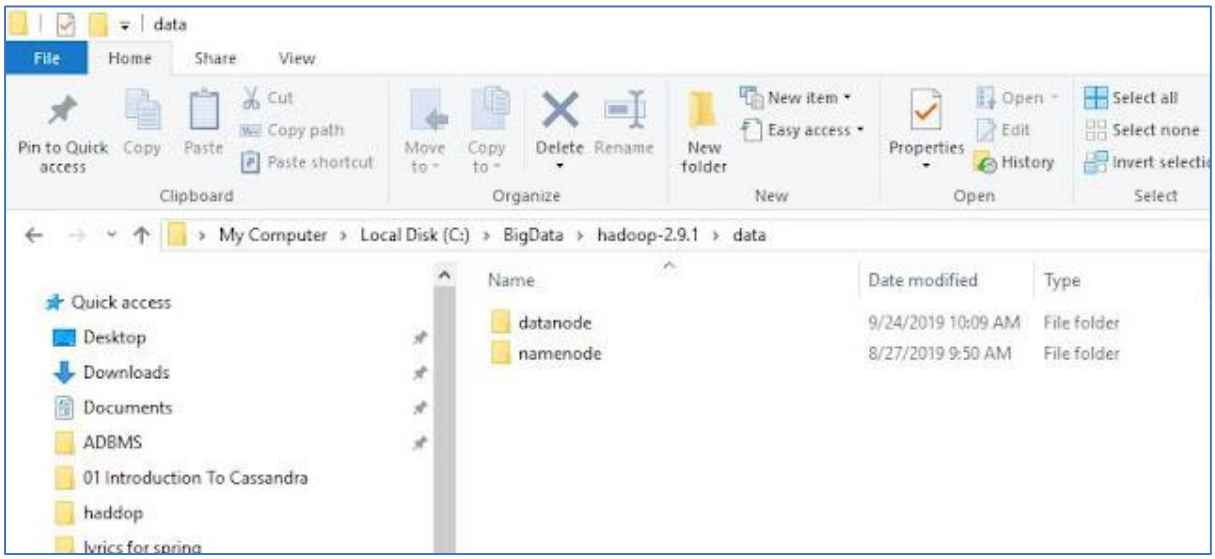


Then download the windows compatible binaries from the git hub repo.
Link:- https://github.com/ParixitOdedara/Hadoop



Extract the zip and copy all the files present under bin folder to C:\BigData\hadoop-2.9.1\bin. Replace the existing files as well.

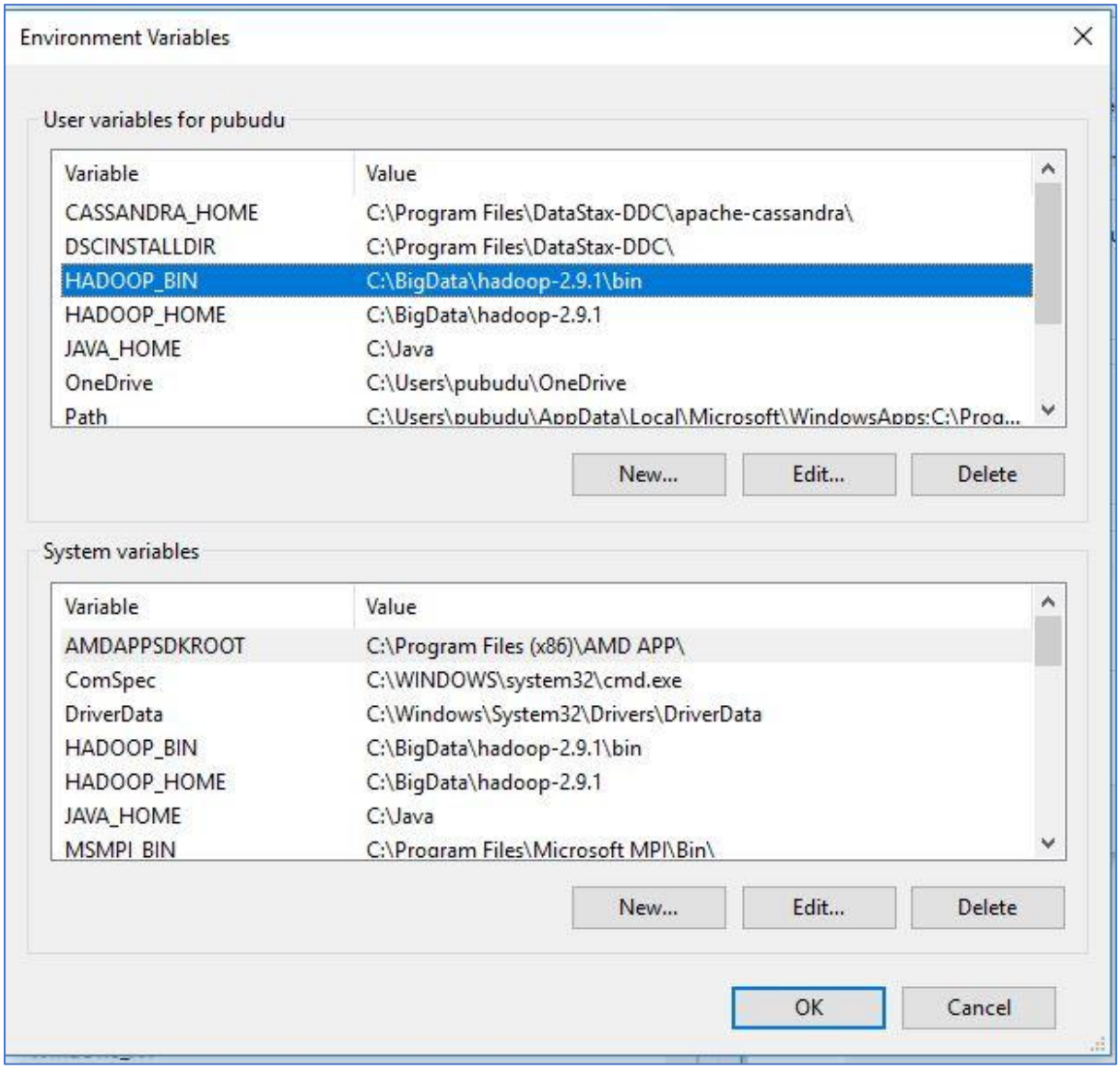
Go to **C:/BigData/ adoop-2.9.1** and create a folder '**data**'. Inside the 'data' folder create two folders '**datanode**' and '**namenode**'.



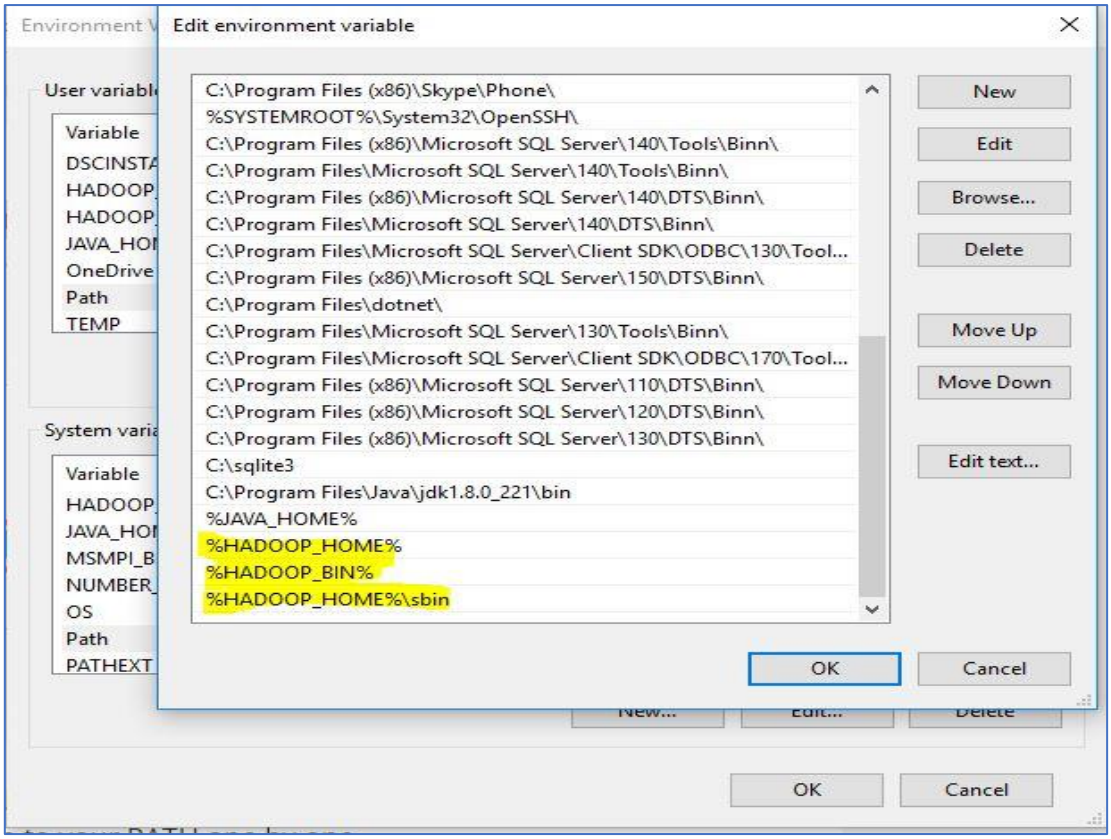
Then Set Hadoop Environment Variables

HADOOP_HOME="C:\BigData\hadoop-2.9.1"
HADOOP_BIN="C:\BigData\hadoop-2.9.1\bin"
JAVA_HOME=<JDK installation location>"

To set these variables, go to My Computer or This PC. Right click --> Properties --> Advanced System settings --> Environment variables. Click New to create a new environment variables.



Then edit PATH Environment Variable

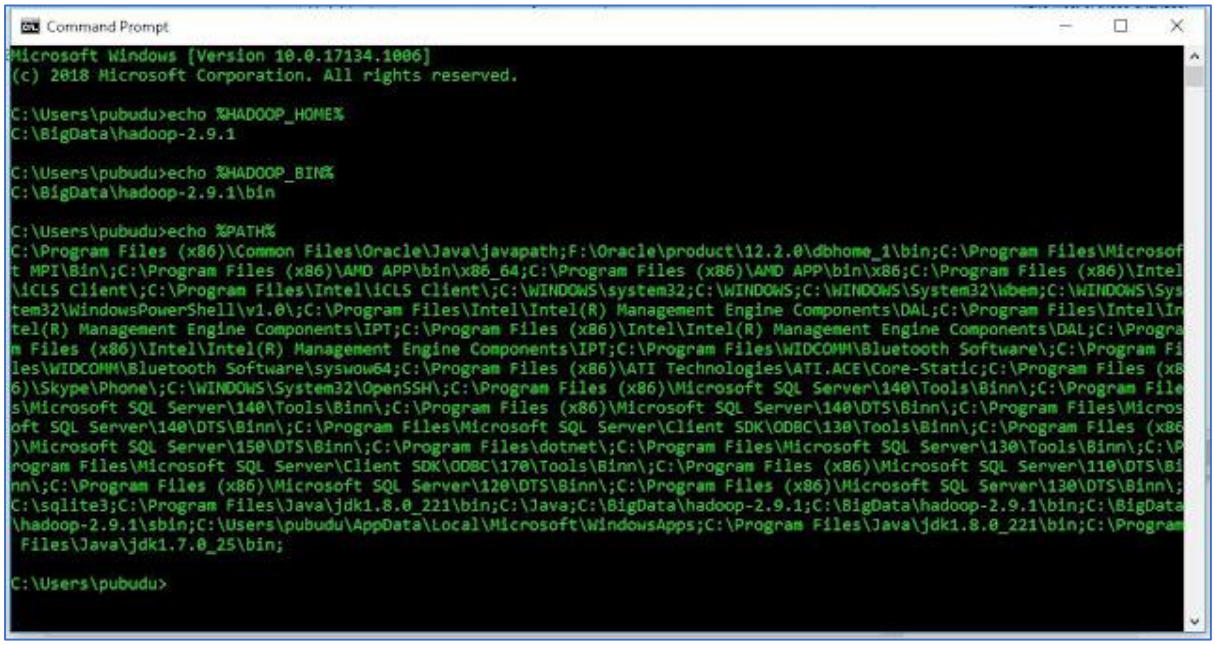


To validate the above setting, **open new cmd** and check the output.

echo %HADOOP_HOME%

echo %HADOOP_BIN%

echo %PATH%



To configure the Hadoop on windows we have to edit below mention files in the extracted location.

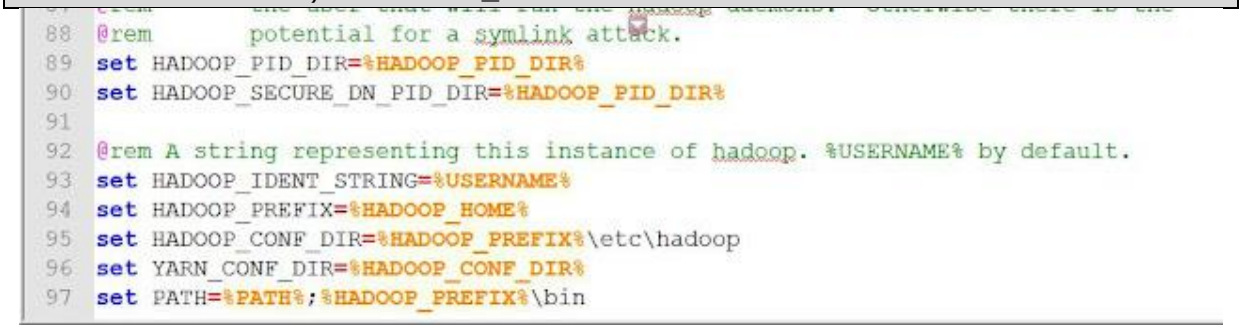
1. hadoop-env.cmd
2. core-site.xml
3. hdfs-site.xml
4. mapred-site.xml
5. yarn-site.xml

Edit hadoop-env.cmd

File location:-C:\BigData\hadoop-2.9.1\etc\hadoop\hadoop-env.cmd

Need to add:-

```
set HADOOP_PREFIX=%HADOOP_HOME%
set HADOOP_CONF_DIR=%HADOOP_PREFIX%\etc\hadoop
set YARN_CONF_DIR=%HADOOP_CONF_DIR%
set PATH=%PATH%;%HADOOP_PREFIX%\bin
```



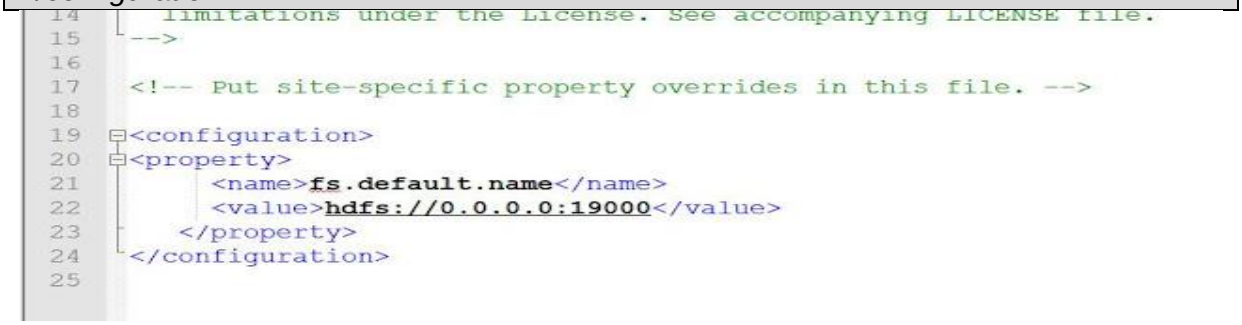
```
@rem       potential for a symlink attack.
set HADOOP_PID_DIR=%HADOOP_PID_DIR%
set HADOOP_SECURE_DN_PID_DIR=%HADOOP_PID_DIR%

@rem A string representing this instance of hadoop. %USERNAME% by default.
set HADOOP_IDENT_STRING=%USERNAME%
set HADOOP_PREFIX=%HADOOP_HOME%
set HADOOP_CONF_DIR=%HADOOP_PREFIX%\etc\hadoop
set YARN_CONF_DIR=%HADOOP_CONF_DIR%
set PATH=%PATH%;%HADOOP_PREFIX%\bin
```

Edit core-site.xml

File Location:- C:\BigData\hadoop-2.9.1\etc\hadoop\core-site.xml

Need to add:-content within <configuration> </configuration> tags.

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://0.0.0.0:19000</value>
  </property>
</configuration>
```



```
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
        <name>fs.default.name</name>
        <value>hdfs://0.0.0.0:19000</value>
    </property>
</configuration>
```

Edit hdfs-site.xml

File Location:- C:\BigData\hadoop-2.9.1\etc\hadoop\hdfs-site.xml.

Need to add;- below content within <configuration> </configuration> tags.

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>C:\BigData\hadoop-2.9.1\data\namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>C:\BigData\hadoop-2.9.1\data\datanode</value>
  </property>
</configuration>
```

Edit mapred-site.xml

File location:- Open C:\BigData\hadoop-2.9.1\etc\hadoop\mapred-site.xml

Need to add:- below content within <configuration> </configuration> tags. If you don't see mapred-site.xml then open mapred-site.xml.template file and rename it to mapred-site.xml

```
<configuration>
  <property>
    <name>mapreduce.job.user.name</name>
    <value>%USERNAME%</value>
  </property>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.apps.stagingDir</name>
    <value>/user/%USERNAME%/staging</value>
  </property>
  <property>
    <name>mapreduce.jobtracker.address</name>
```

```
    <value>local</value>
  </property>
</configuration>
```

Editing yarn-site.xml

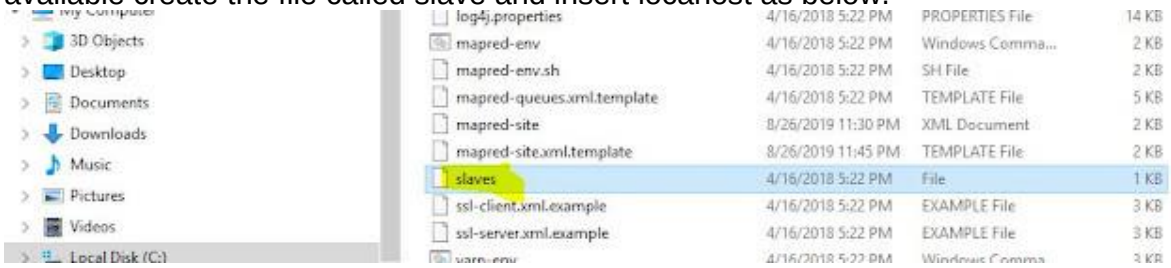
Right click on the file, select edit and paste the following content within <configuration> </configuration> tags.

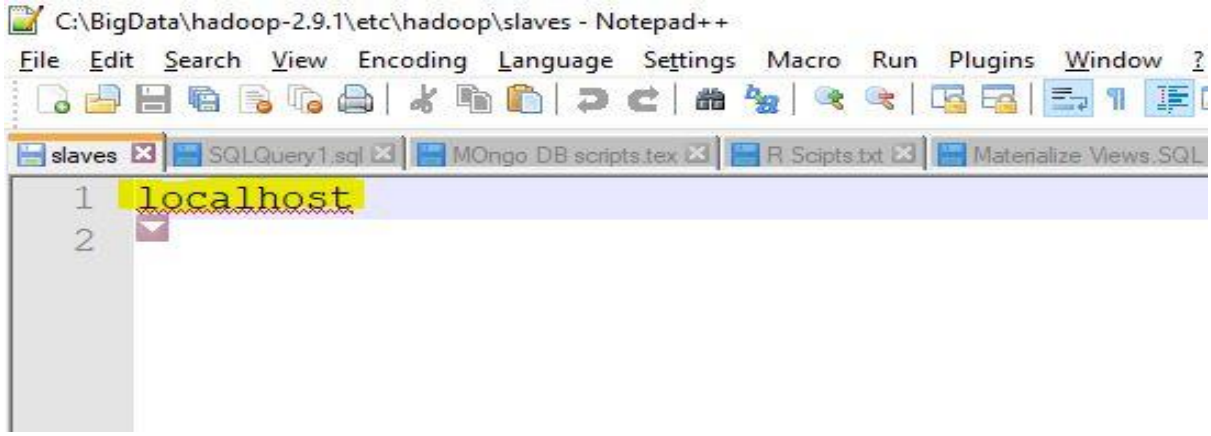
Note:- Below part already has the configuration tag, we need to copy only the part inside it.

```
<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>
```

Additional Configuration:-

Check if C:\BigData\hadoop-2.9.1\etc\hadoop\slaves file is present, if that file not available create the file called slave and insert locahost as below.

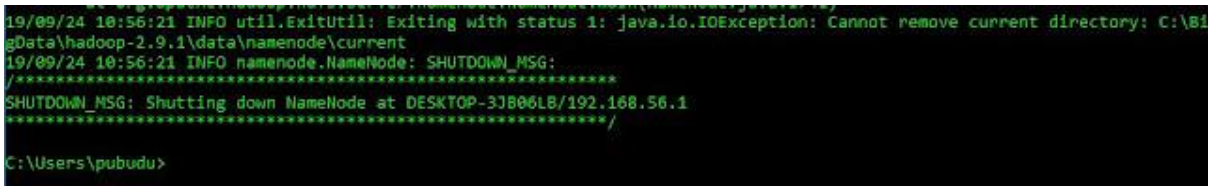




Node formatting

To format the node, **open the cmd** and execute the below command.

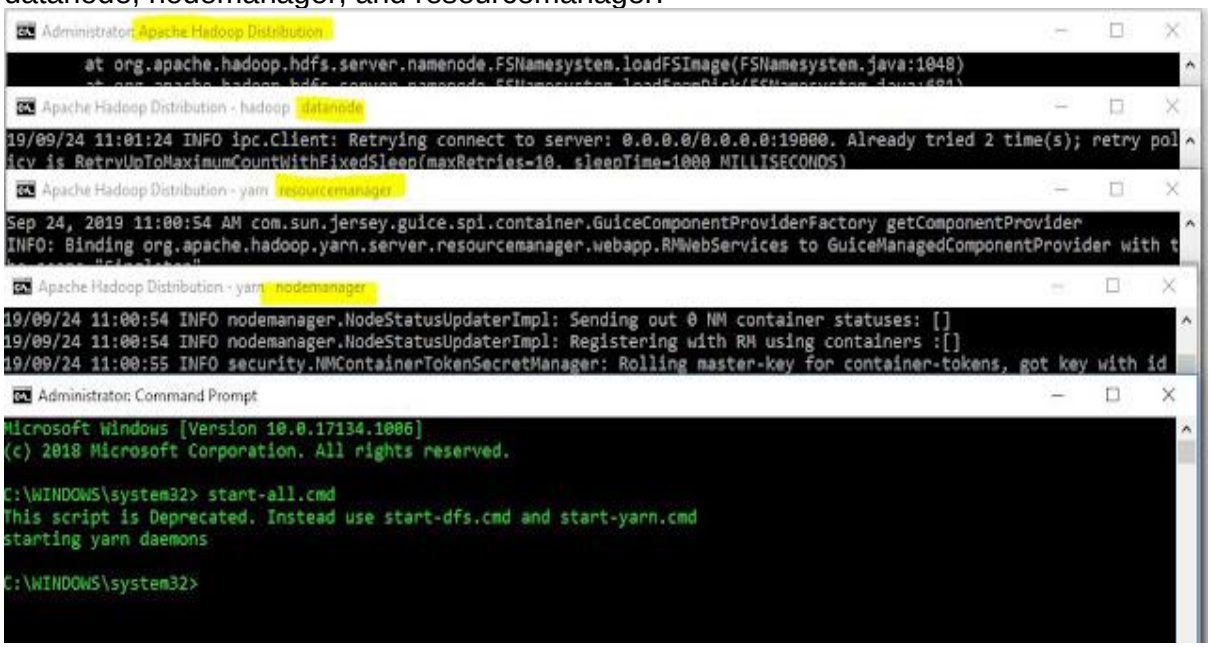
hadoop namenode -format



To enable the Hadoop open the **CMD as Administrator** and type below command

start-all.cmd

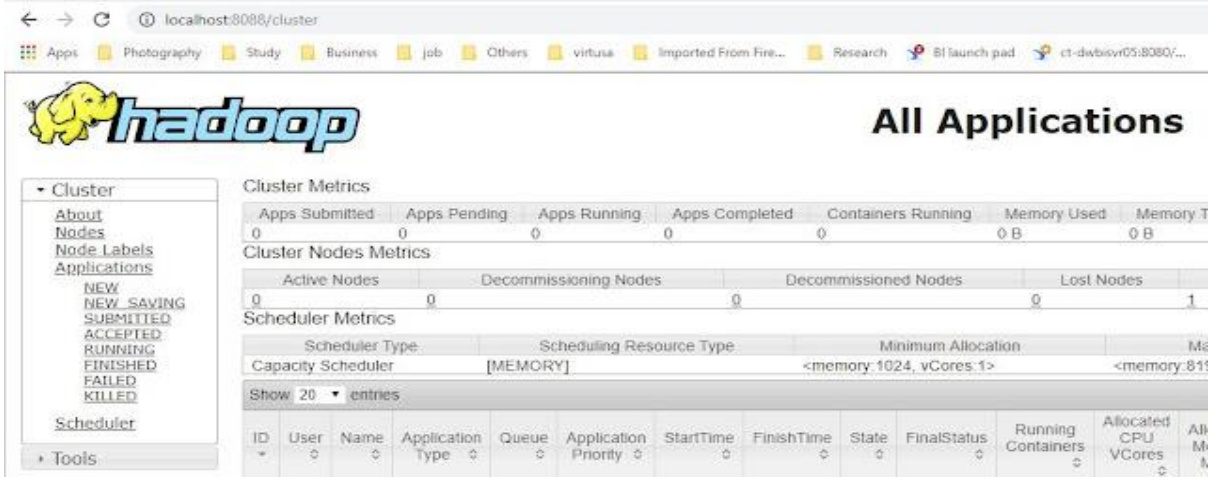
It will open 4 new windows cmd terminals for 4 daemon processes, namely namenode, datanode, nodemanager, and resourcemanager.



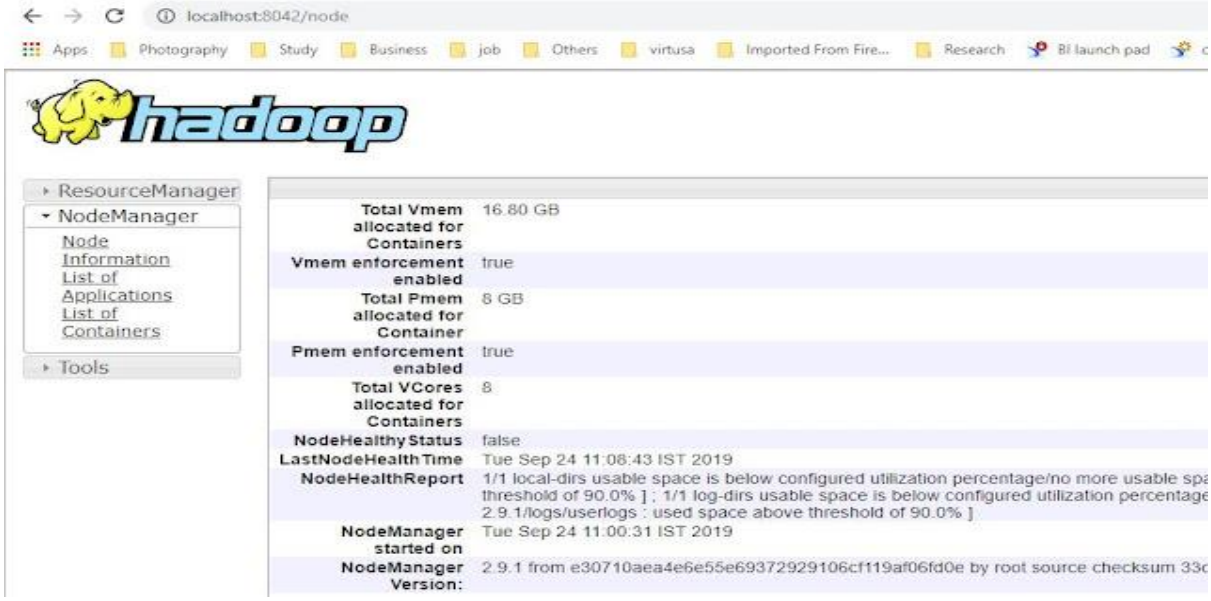
Then you have succesfully install the hadoop 2.9.1 on windows platform.

Now you can access all the Hadoop components via web urls.

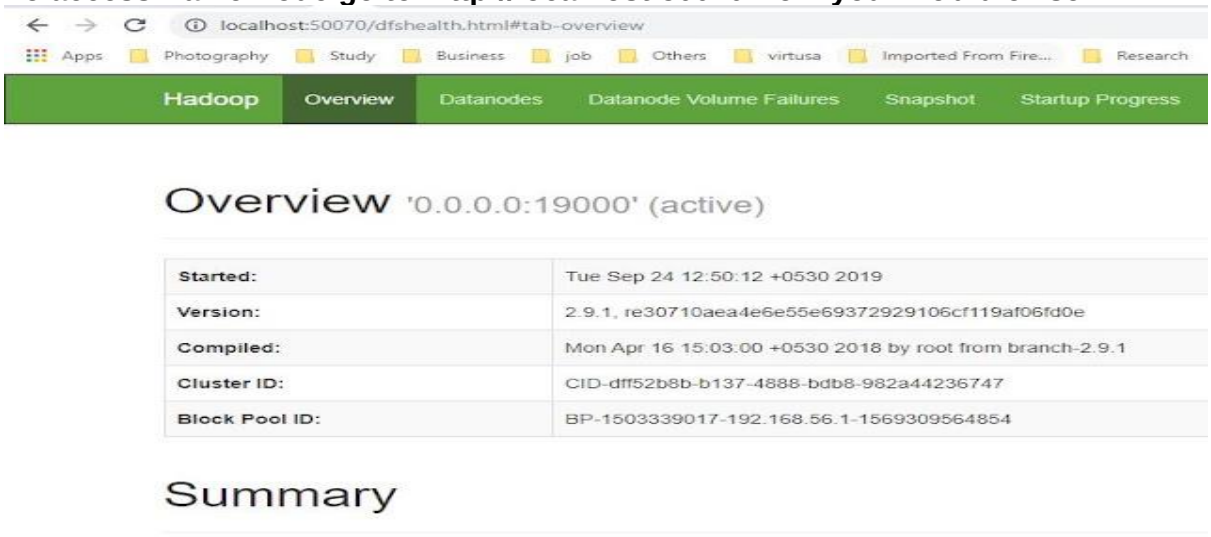
To access Resource Manager go to http://localhost:8088 from your web browser.



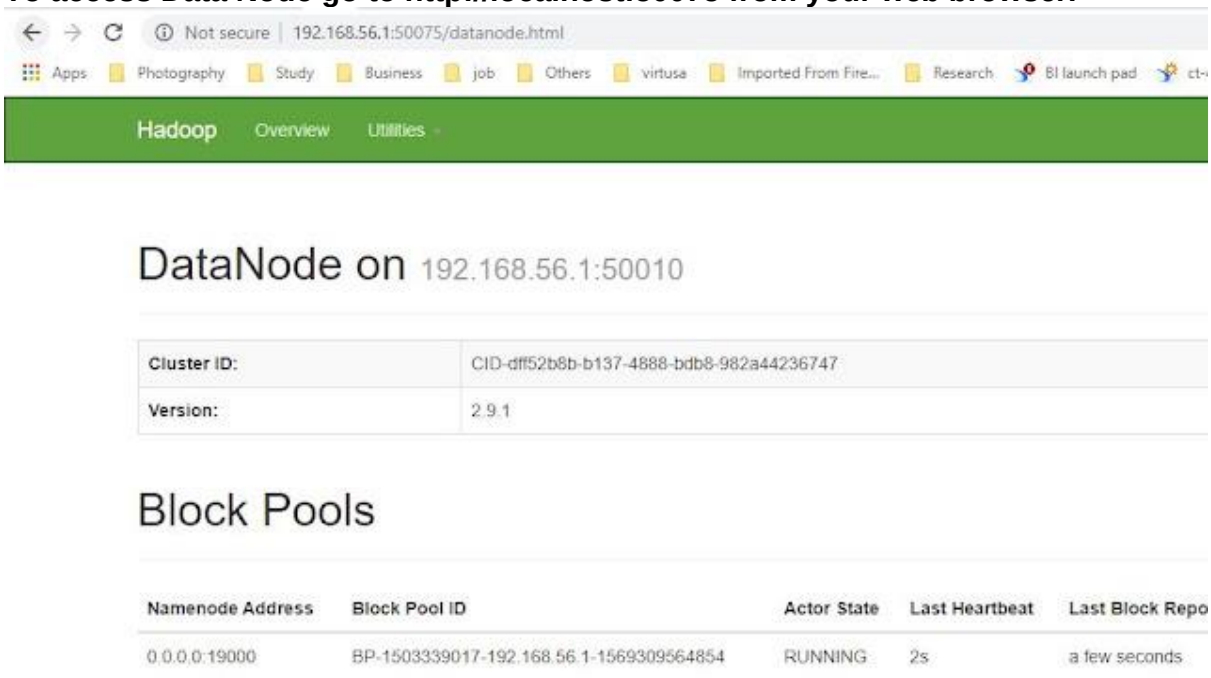
To access Node Manager go to http://localhost:8042 from your web browser.



To access Name Node go to http://localhost:50070 from your web browser.



To access Data Node go to http://localhost:50075 from your web browser.



Open a new Windows Command Prompt and run below commands.

C:\hadoop-2.9.1\hadoop-2.9.1>cd bin

ls: This command is used to list all the files. Use lsr for recursive approach. It is useful when we want a hierarchy of a folder.

C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /

```
Found 2 items
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:06 /sampledir
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 19:57 /test
```

mkdir: To create a directory. In Hadoop *dfs* there is no home directory by default.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -mkdir /user
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -mkdir /user/Lenovo
```

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:06 /sampledir
```

```
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 19:57 /test
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user
```

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -lsr /user
lsr: DEPRECATED: Please use 'ls -R' instead.
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user/Lenovo
```

touchz: It creates an empty file.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -touchz /user/myfile.txt
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls -R /user
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user/Lenovo
-rw-r--r--   1 LENOVO supergroup          0 2023-02-27 20:26 /user/myfile.txt
```

copyFromLocal (or) put: To copy files/folders from local file system to hdfs store. This is the most important command. Local filesystem means the files present on the OS.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -put C:/hadoop-2.9.1/hadoop-2.9.1/Sample.txt /user
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls -R /user
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user/Lenovo
-rw-r--r--   1 LENOVO supergroup         12 2023-02-27 20:30 /user/Sample.txt
-rw-r--r--   1 LENOVO supergroup          0 2023-02-27 20:26 /user/myfile.txt
```

cat: To print file contents.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -cat /user/Sample.txt
Hello Hadoop
```

copyToLocal (or) get: To copy files/folders from hdfs store to local file system.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -get /user/Sample.txt ../HadoopExamples
```

cp: This command is used to copy files within hdfs.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -mkdir /user_copied
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -cp /user /user_copied
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /user_copied
Found 1 items
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:48 /user_copied/user
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /user
Found 3 items
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user/Lenovo
-rw-r--r--   1 LENOVO supergroup         12 2023-02-27 20:30 /user/Sample.txt
-rw-r--r--   1 LENOVO supergroup          0 2023-02-27 20:26 /user/myfile.txt
```

mv: This command is used to move files within hdfs.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -mv /user/myfile.txt /user_copied
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /user
Found 2 items
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user/Lenovo
-rw-r--r--   1 LENOVO supergroup         12 2023-02-27 20:30 /user/Sample.txt
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /user_copied
Found 2 items
-rw-r--r--   1 LENOVO supergroup          0 2023-02-27 20:26 /user_copied/myfile.txt
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:48 /user_copied/user
```

rmr: This command deletes a file from HDFS *recursively*. It is very useful command when you want to delete a *non-empty directory*.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -rmr /user_copied
rmr: DEPRECATED: Please use '-rm -r' instead.
Deleted /user_copied
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /user_copied
ls: `/user_copied': No such file or directory
```

du: It will give the size of each file in directory.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -du /user
0   /user/Lenovo
12  /user/Sample.txt
```

dus: This command will give the total size of directory/file.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -dus /user
dus: DEPRECATED: Please use 'du -s' instead.
12  /user
```

stat: It will give the last modified time of directory or path. In short it will give stats of the directory or file.

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -stat /user
2023-02-27 15:20:27
```

setrep: This command is used to change the replication factor of a file/directory in HDFS. By default it is 3 for anything which is stored in HDFS (as set in hdfs *core-site.xml*).

```
C:\hadoop-2.9.1\hadoop-2.9.1\bin>hdfs dfs -ls /user
Found 2 items
drwxr-xr-x   - LENOVO supergroup          0 2023-02-27 20:23 /user/Lenovo
-rw-r--r--   6 LENOVO supergroup         12 2023-02-27 20:30 /user/Sample.txt
```