**Jaydeep Mahajan | CE066**

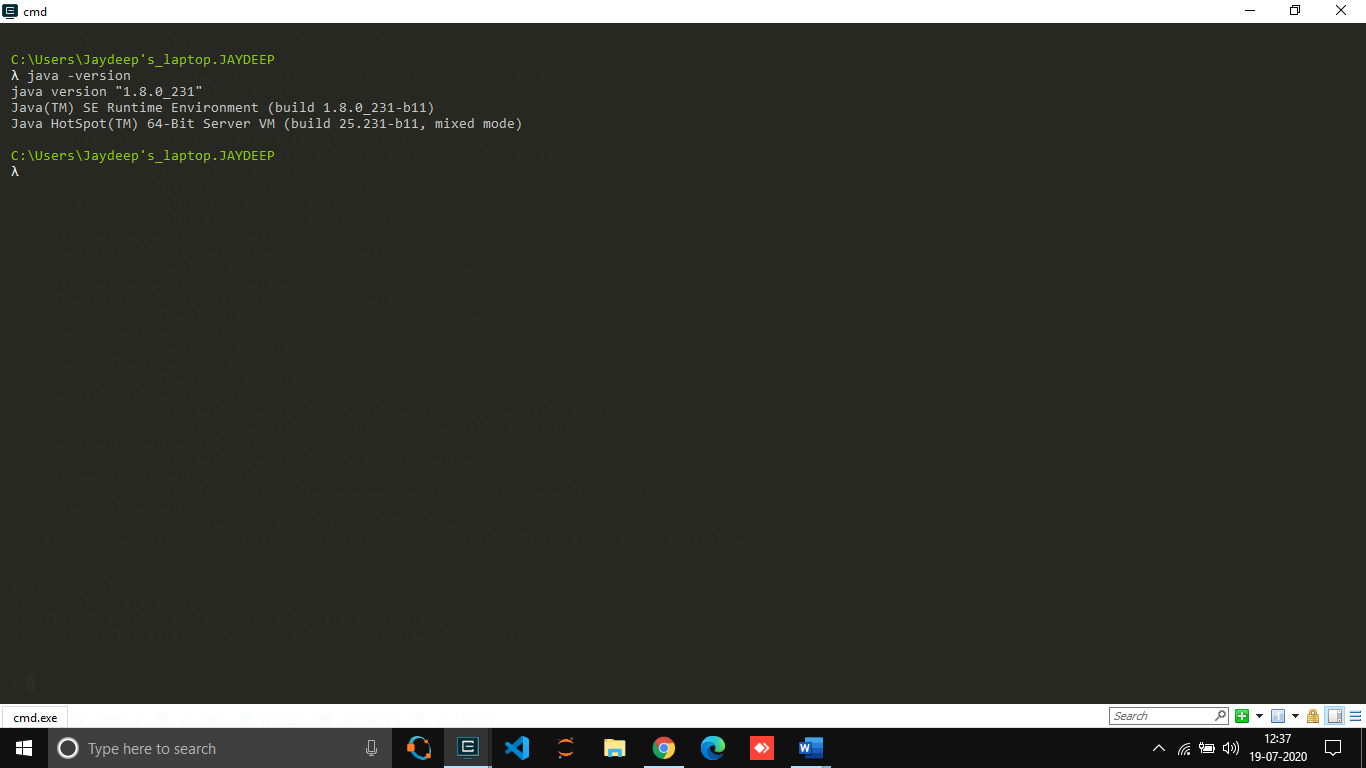
**BDA LAB 2**

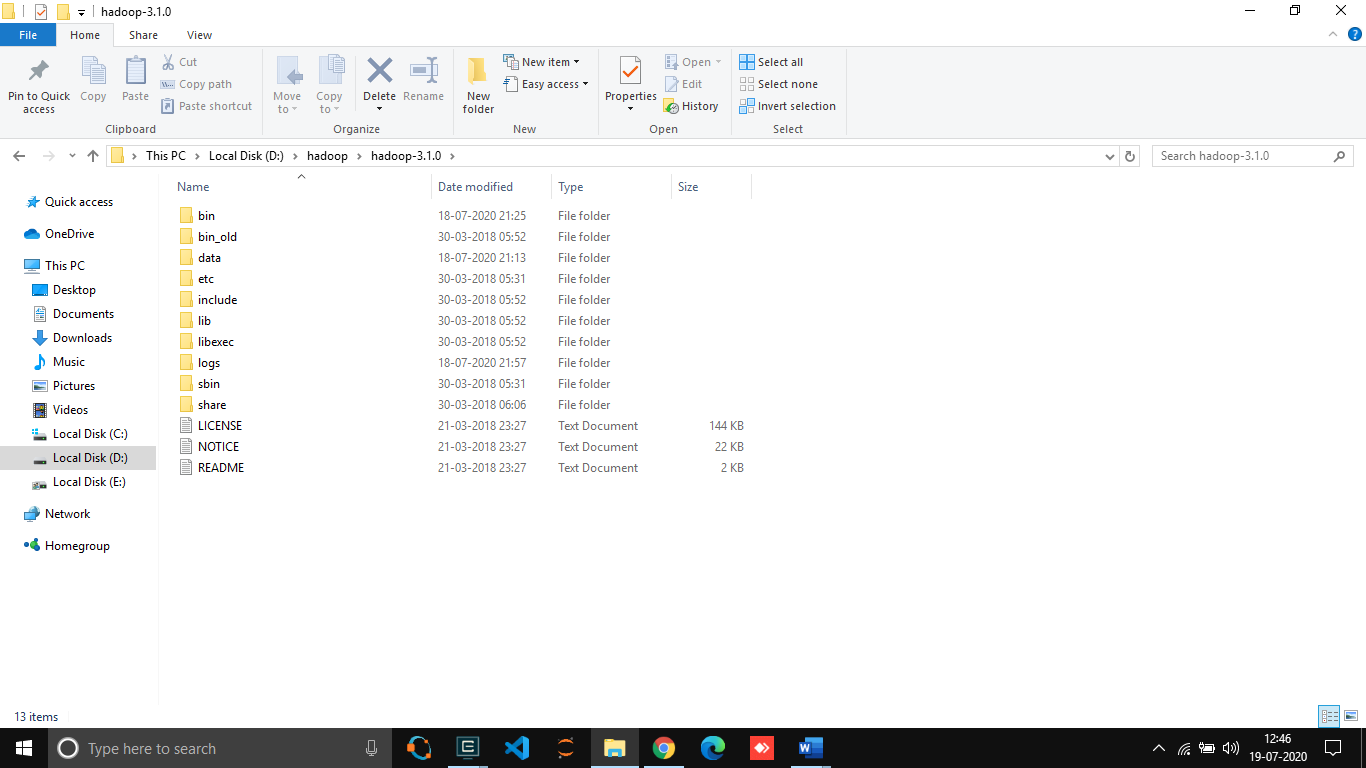
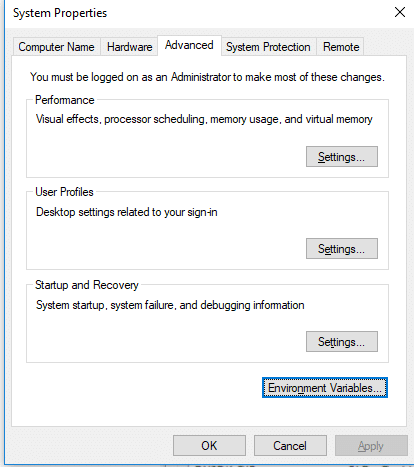
Exploring Hadoop

* What is Hadoop?

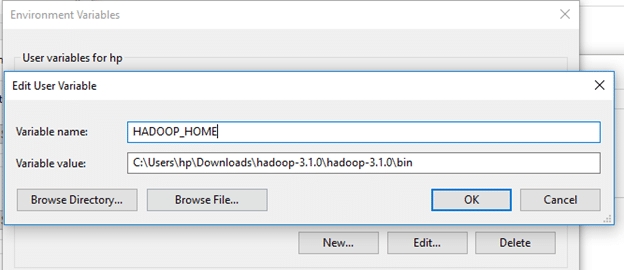
**Hadoop is an open source framework and is used to store process and analyse data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing.**

* Installation of Hadoop.
* **We will be installing single node pseudo-distributed Hadoop cluster on windows 10.**
* **To install Hadoop, you should have Java version 1.8 in your system if it is not present then install it.**
* **Check your java version through this command on command prompt.**
* **Java -version**

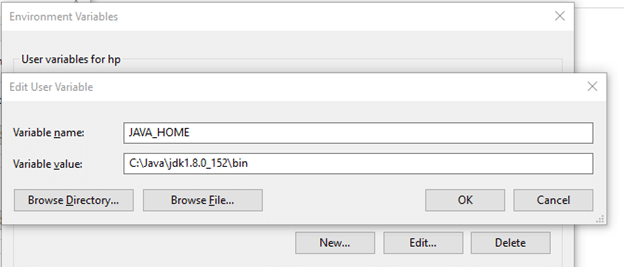
****

* After downloading java version 1.8, download Hadoop version 3.1 from this [link](https://archive.apache.org/dist/hadoop/common/hadoop-3.1.0/hadoop-3.1.0.tar.gz).
* Extract it to a folder.
* Setup System Environment Variables
* Open control panel to edit the system environment variable.

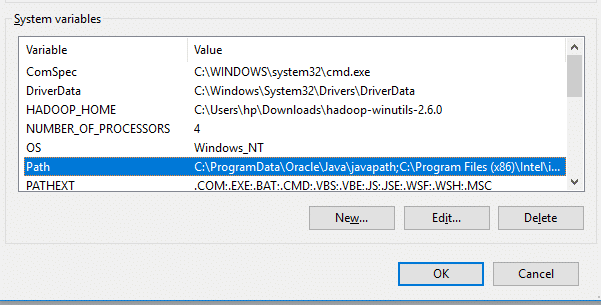
1. Create a new user variable. Put the Variable name as HADOOP\_HOME and Variable value as the path of the bin folder where you extracted Hadoop.

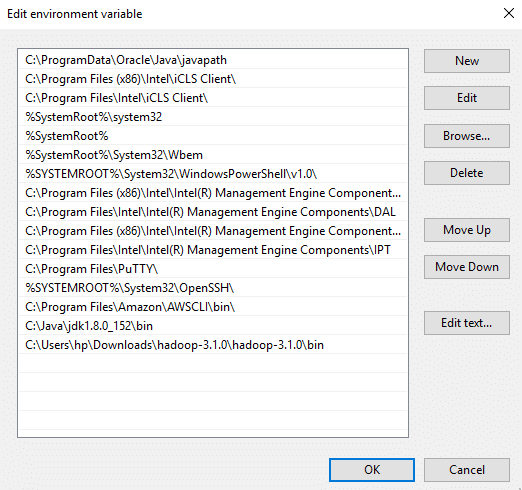


* Likewise, create a new user variable with variable name as JAVA\_HOME and variable value as the path of the bin folder in the Java directory.

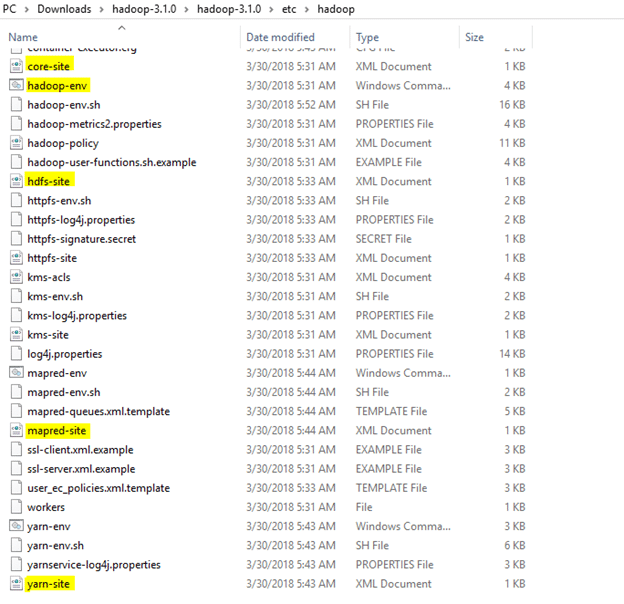


* Now we need to set Hadoop bin directory and Java bin directory path in system variable path. Edit Path in system variable.



* Click on New and add the bin directory path of Hadoop and Java in it.
* **Configurations**

Now we need to edit some files located in the Hadoop directory of the etc folder where we installed Hadoop. The files that need to be edited have been highlighted.



1. Edit the file core-site.xml in the Hadoop directory. Copy this xml property in the configuration in the file.

**<configuration>**

**<property>**

**<name>fs. defaultFS</name>**

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

**</property>**

**</configuration>**

1. Edit mapred-site.xml and copy this property in the configuration.

**<configuration>**

**<property>**

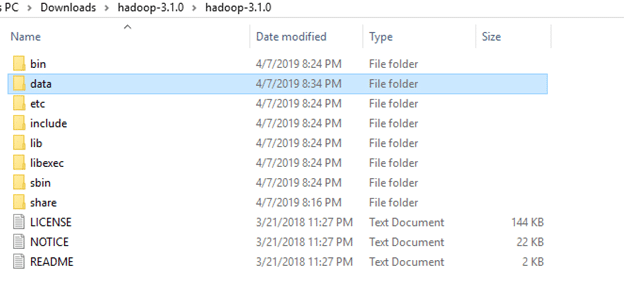
**<name>mapreduce.framework.name</name>**

**<value>yarn</value>**

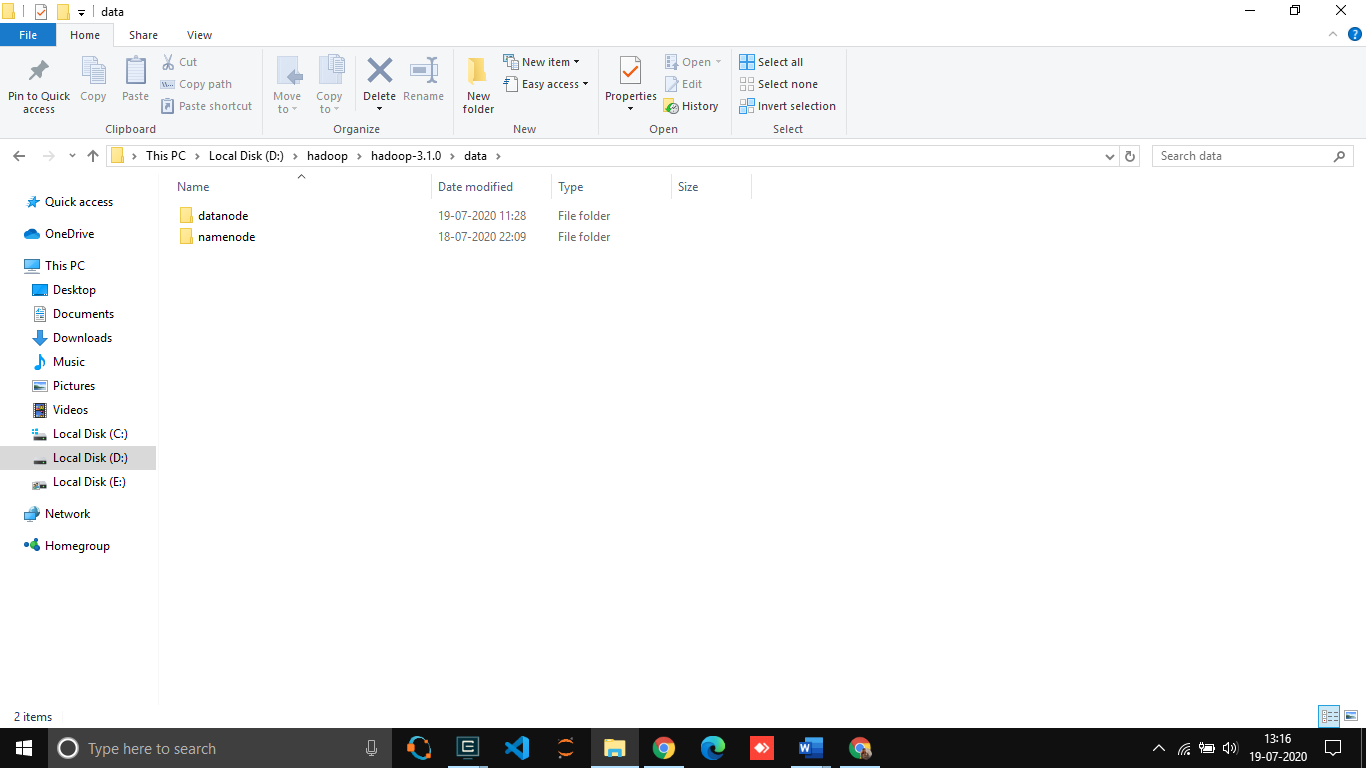
**</property>**

**</configuration>**

1. Create a folder ‘data’ in the Hadoop directory.



1. Create a folder with the name ‘datanode’ and a folder ‘namenode’ in this data directory.



1. Edit the file hdfs-site.xml and add below property in the configuration.

Note: The path of namenode and datanode across value would be the path of the datanode and namenode folders you just created.

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

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

<value> D:\hadoop\hadoop-3.1.0\data\namenode

</value>

</property>

<property>

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

<value> D:\hadoop\hadoop-3.1.0\data\datanode

</value>

</property>

</configuration>

1. Edit the file yarn-site.xml and add below property in the configuration.

<configuration>

<property>

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

<value>mapreduce\_shuffle</value>

</property>

<property>

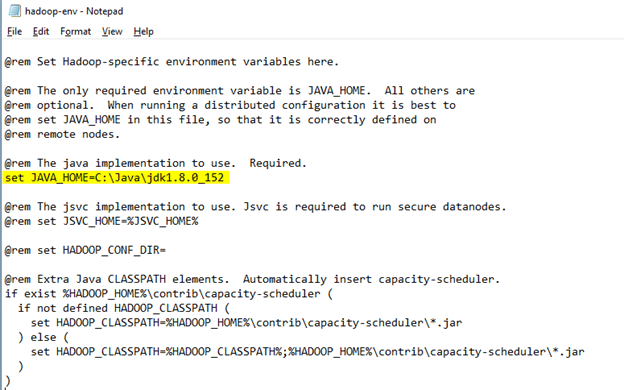
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

</configuration>

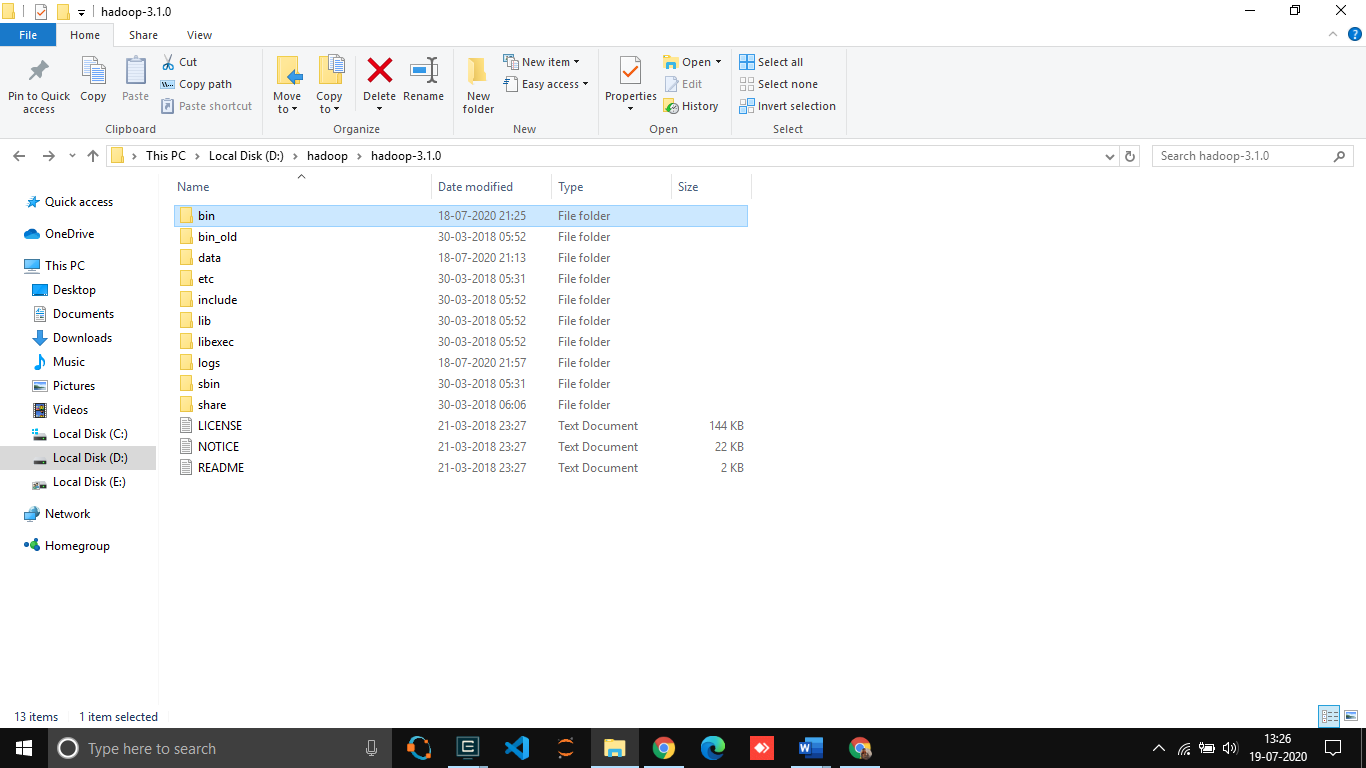
1. Edit hadoop-env.cmd and replace %JAVA\_HOME% with the path of the java folder where your jdk 1.8 is installed.



1. Hadoop needs windows OS specific files which does not come with default download of hadoop.
2. To include those files, replace the bin folder in hadoop directory with the bin folder provided in this github link.

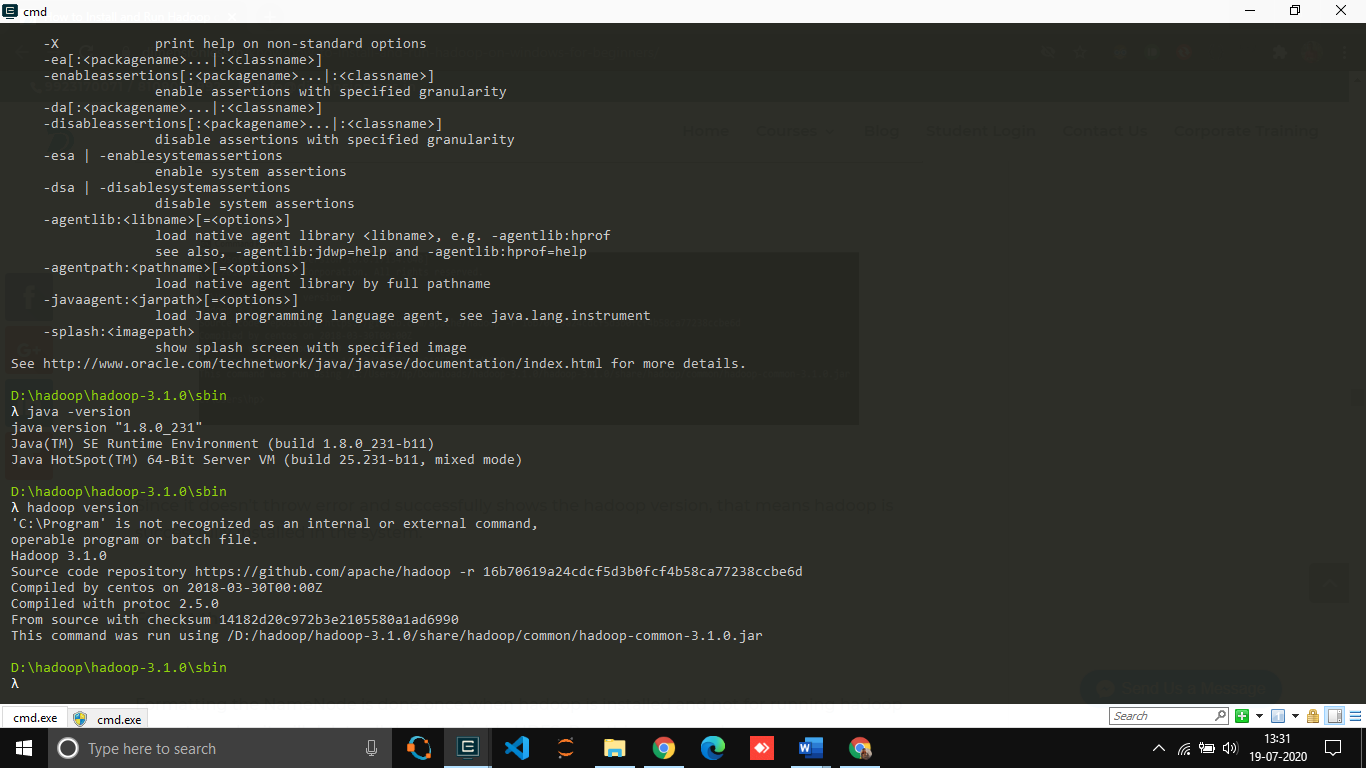
<https://github.com/s911415/apache-hadoop-3.1.0-winutils>

1. Download it as zip file. Extract it and copy the bin folder in it. If you want to save the old bin folder, rename it like bin\_old and paste the copied bin folder in that directory.



Check whether Hadoop is successfully installed by running this

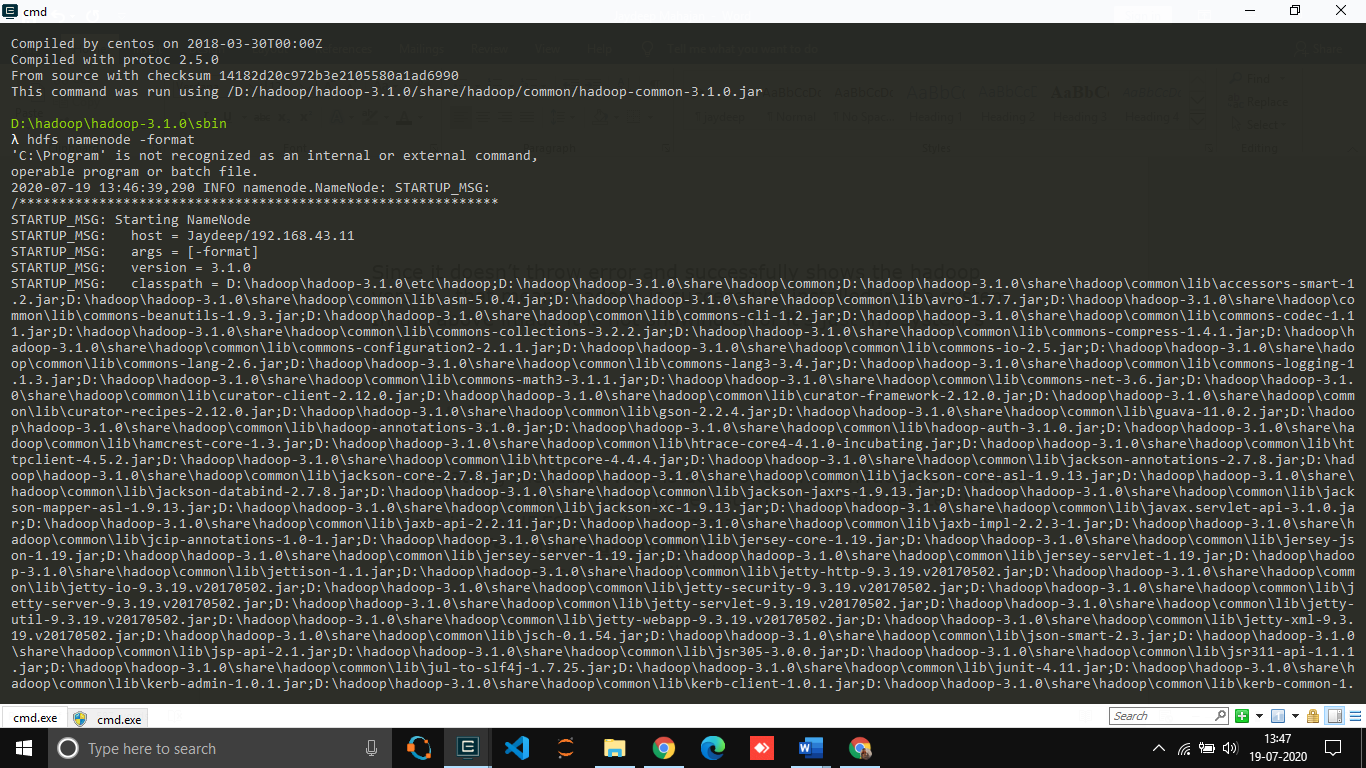
command on cmd(Run command prompt as a Administrator so it will not create any permission issue in future for any directory).

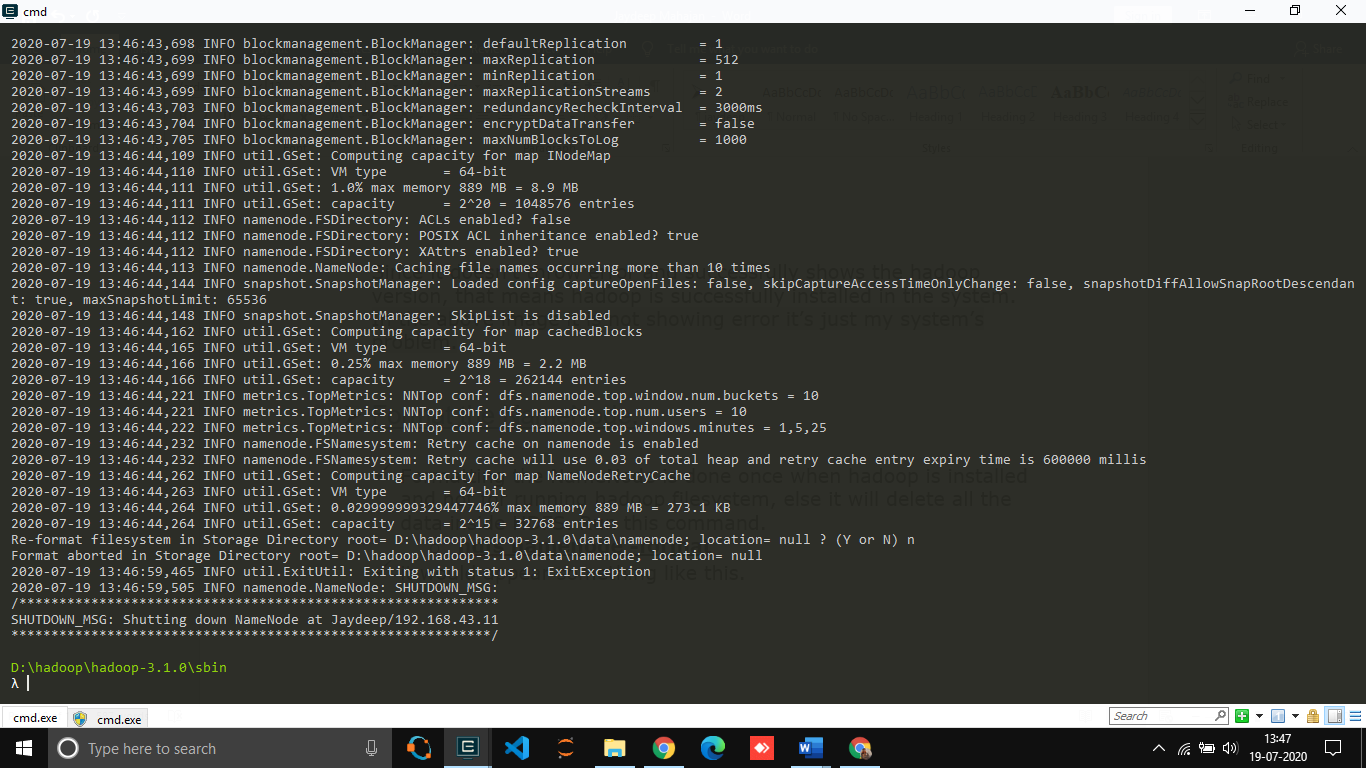
* + - hadoop -version

Since it doesn’t throw error and successfully shows the hadoop version, that means hadoop is successfully installed in the system.

In the above image it is not showing error it’s just my system’s problem.

* **Format the NameNode.**
* Formatting the NameNode is done once when hadoop is installed and not for running hadoop filesystem, else it will delete all the data inside HDFS. Run this command.
  + **hdfs namenode -format**
* it would appear something like this.



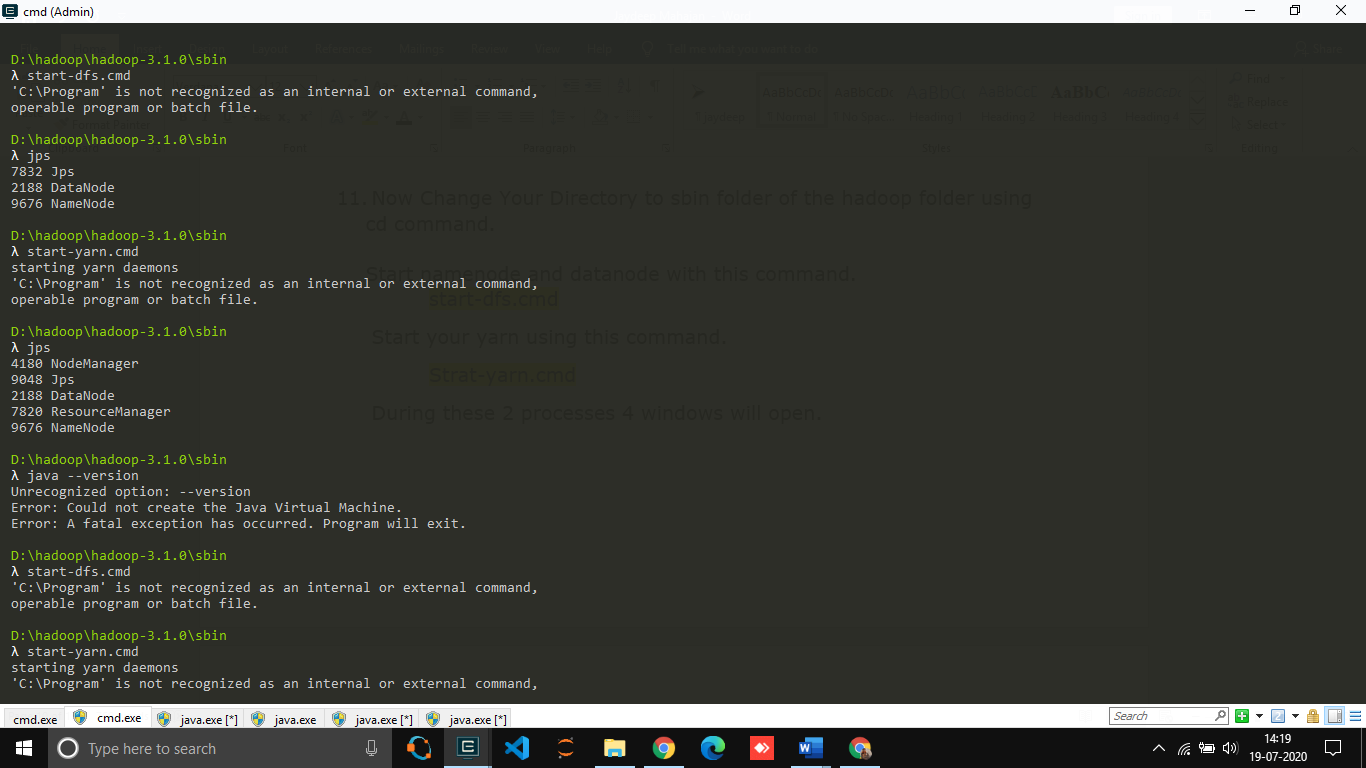


1. Now Change Your Directory to sbin folder of the hadoop folder using cd command.

Start namenode and datanode with this command.

start-dfs.cmd

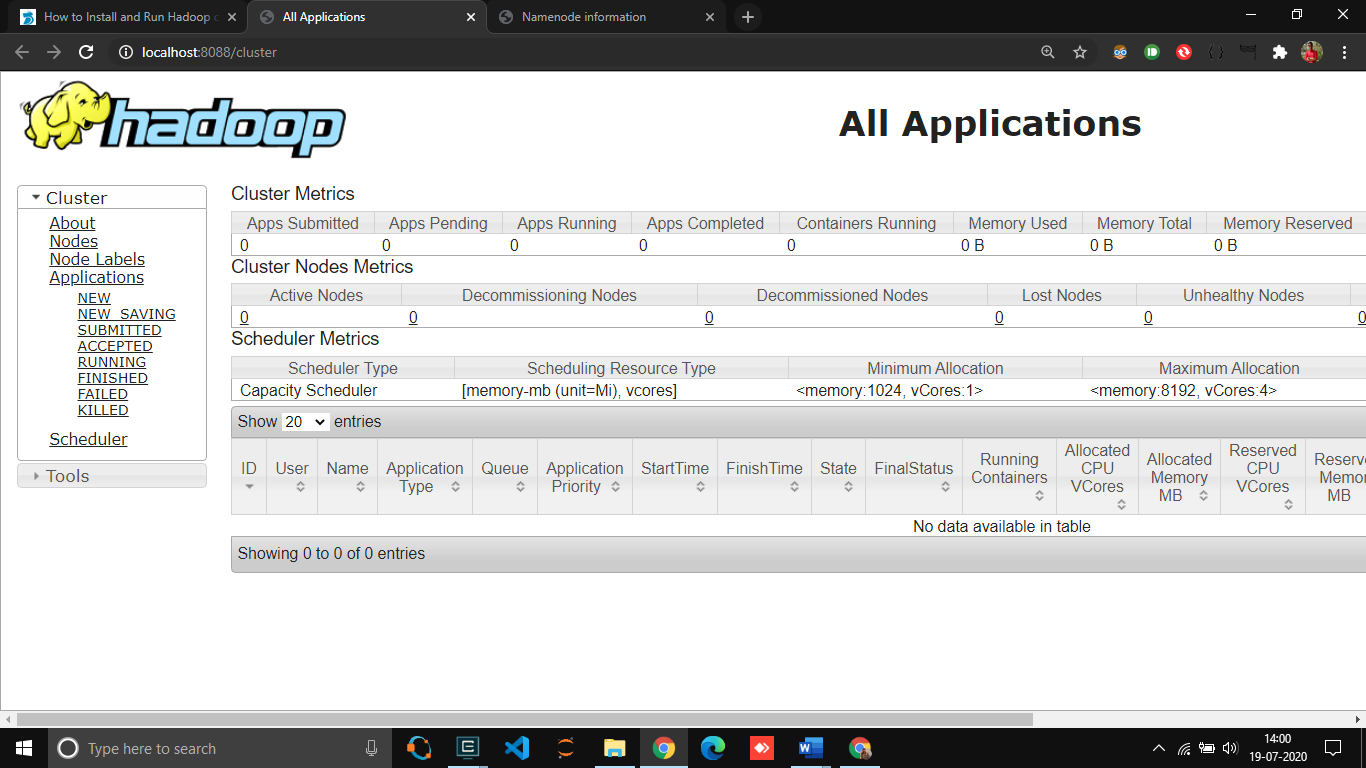
Start your yarn using this command.

 Strat-yarn.cmd

During these 2 processes 4 windows will open.

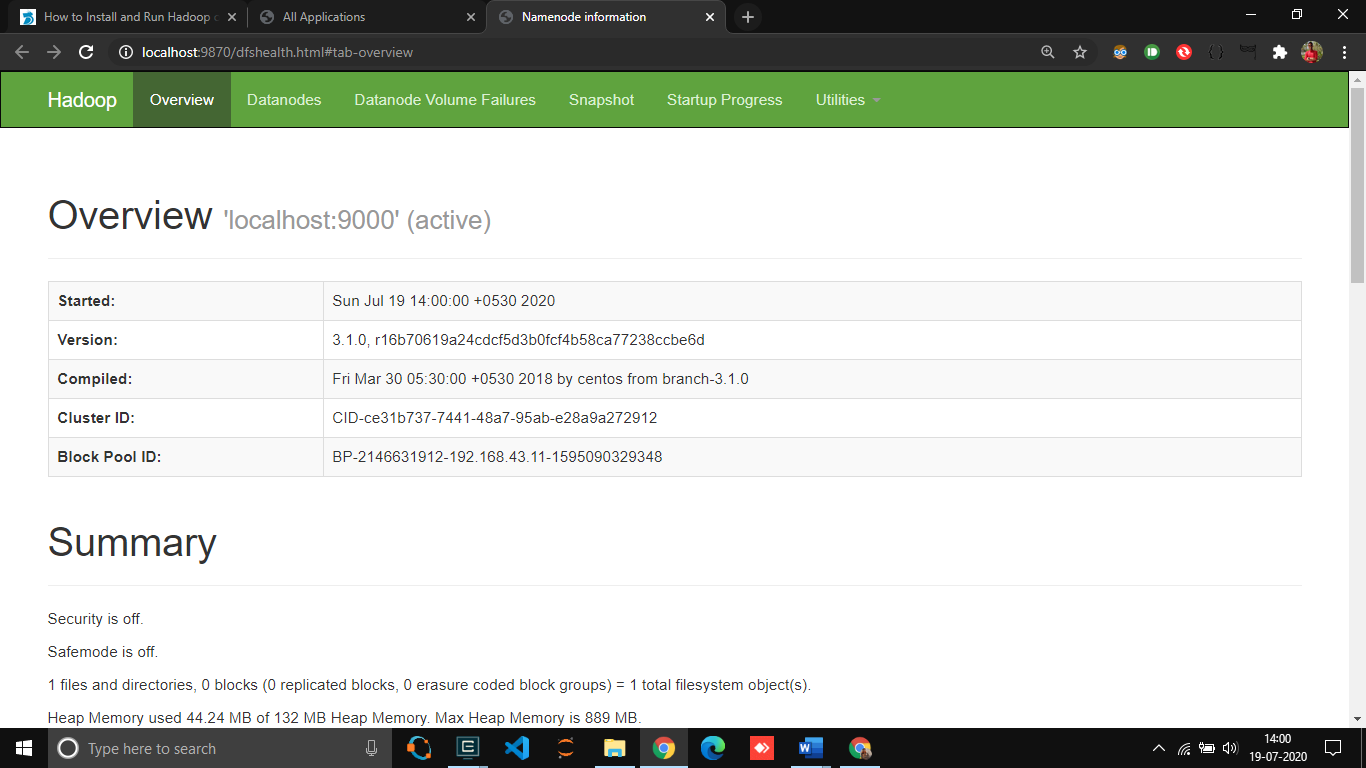
**Make sure all the 4 Apache Hadoop Distribution windows are up n running. If they are not running, you will see an error or a shutdown message. In that case, you need to debug the error.**

1. To access information about resource manager current jobs, successful and failed jobs, go to this link in browser-

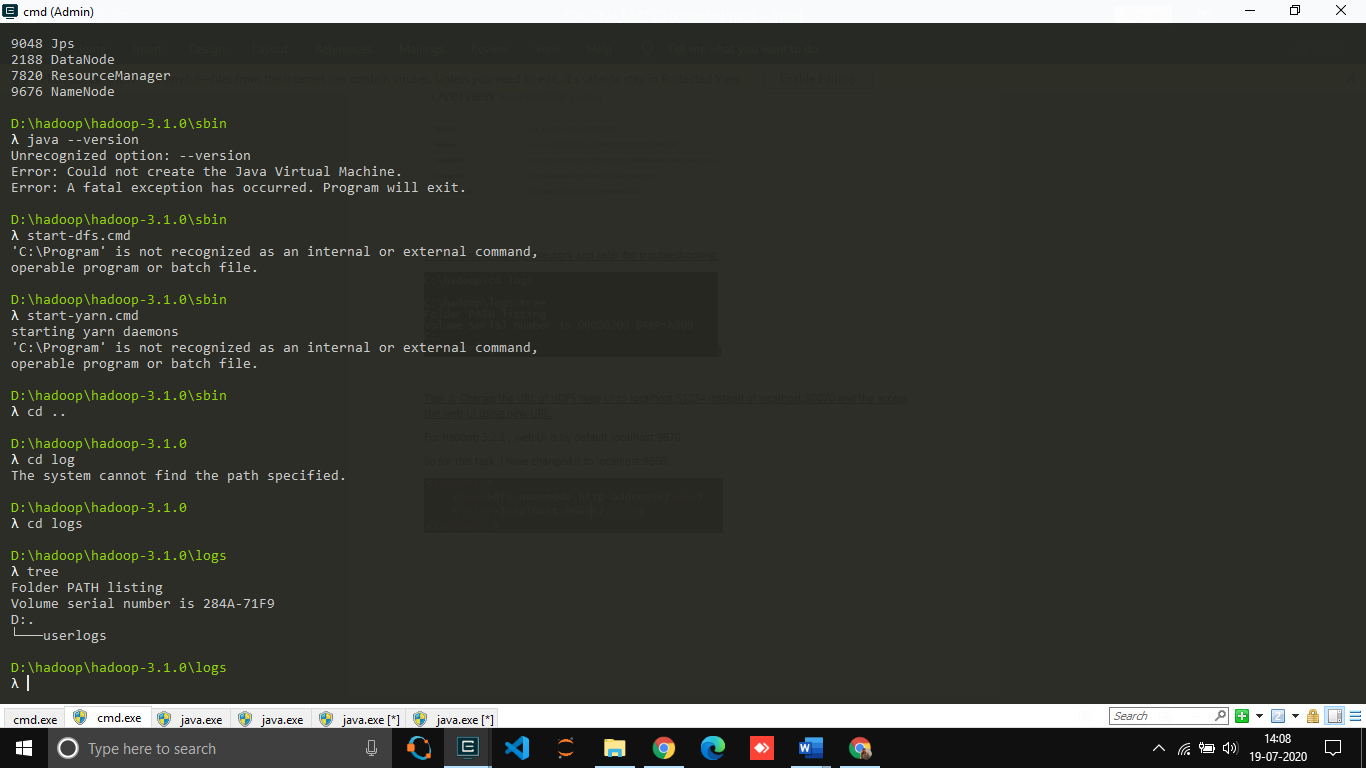
<http://localhost:8088/cluster>

To check the details about the hdfs (namenode and datanode),

Open this [link](http://localhost:9870/) on browser-

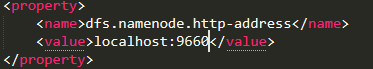


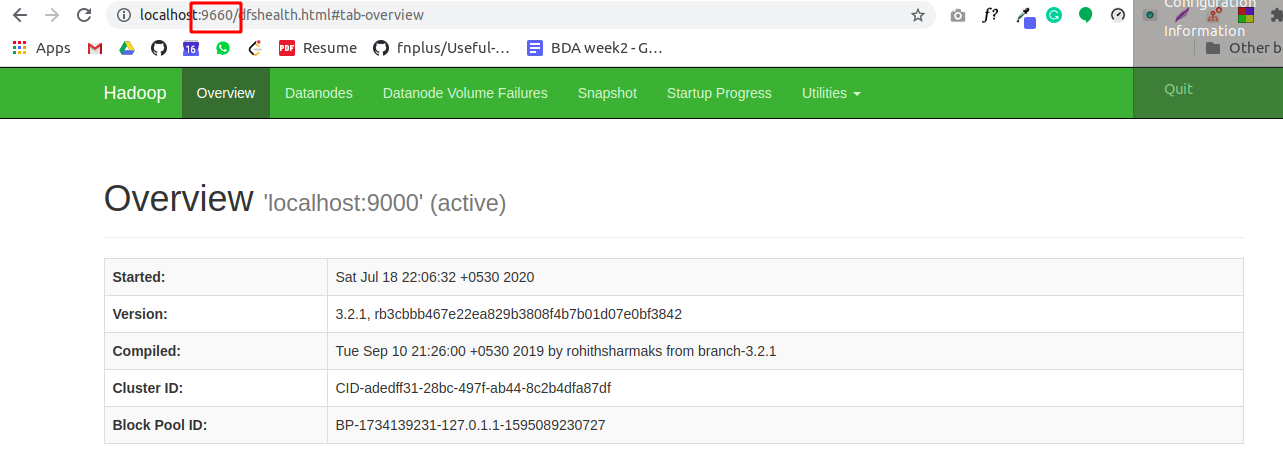
Task2: Locate the log directory and refer for troubleshooting.



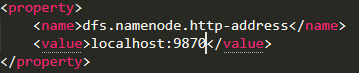
Task3: Change the URL of HDFS Web UI to localhost:51234 instead of localhost:50070 and the access the web UI using new URL.

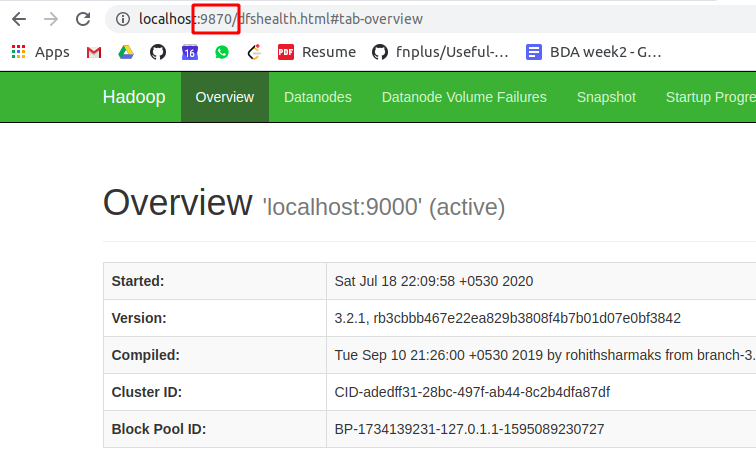
* For hadoop 3.1.0, web UI is by default localhost:9870.
* So, for this task, I have changed it to localhost:9660



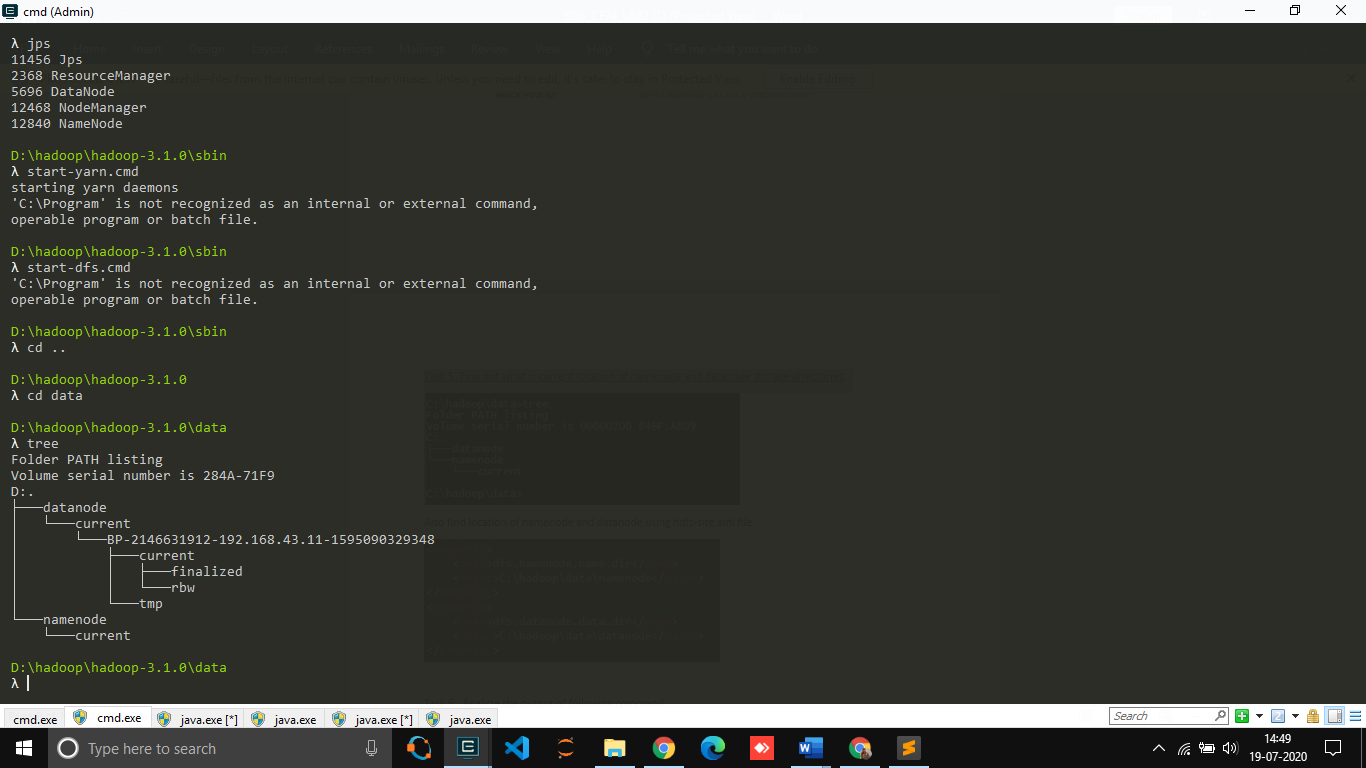


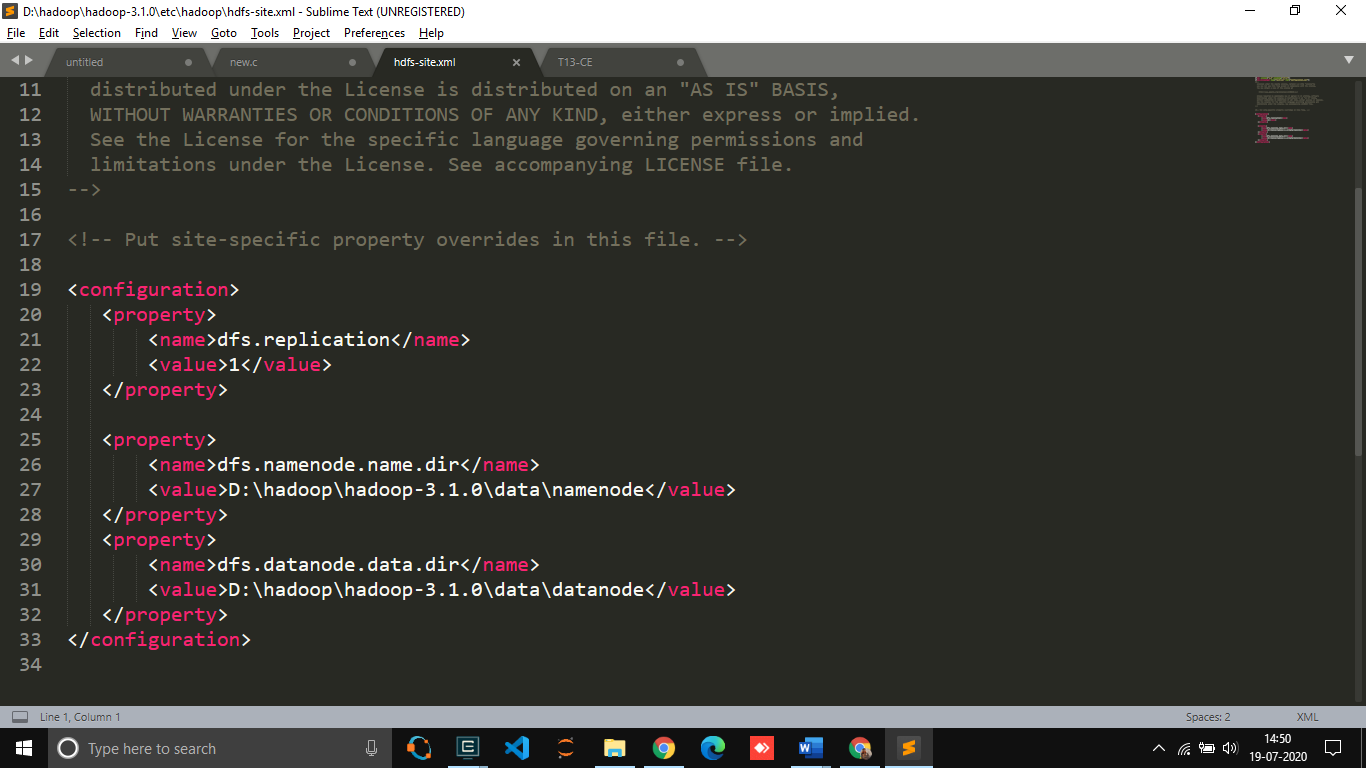
Task 4: Undo changes done in task1.





Task 5: Find out what is current location of namenode and datanode storage directories.



Also find location of namenode and datanode using hdfs-site.xml file.

Task 6: Analyse the output of following command:

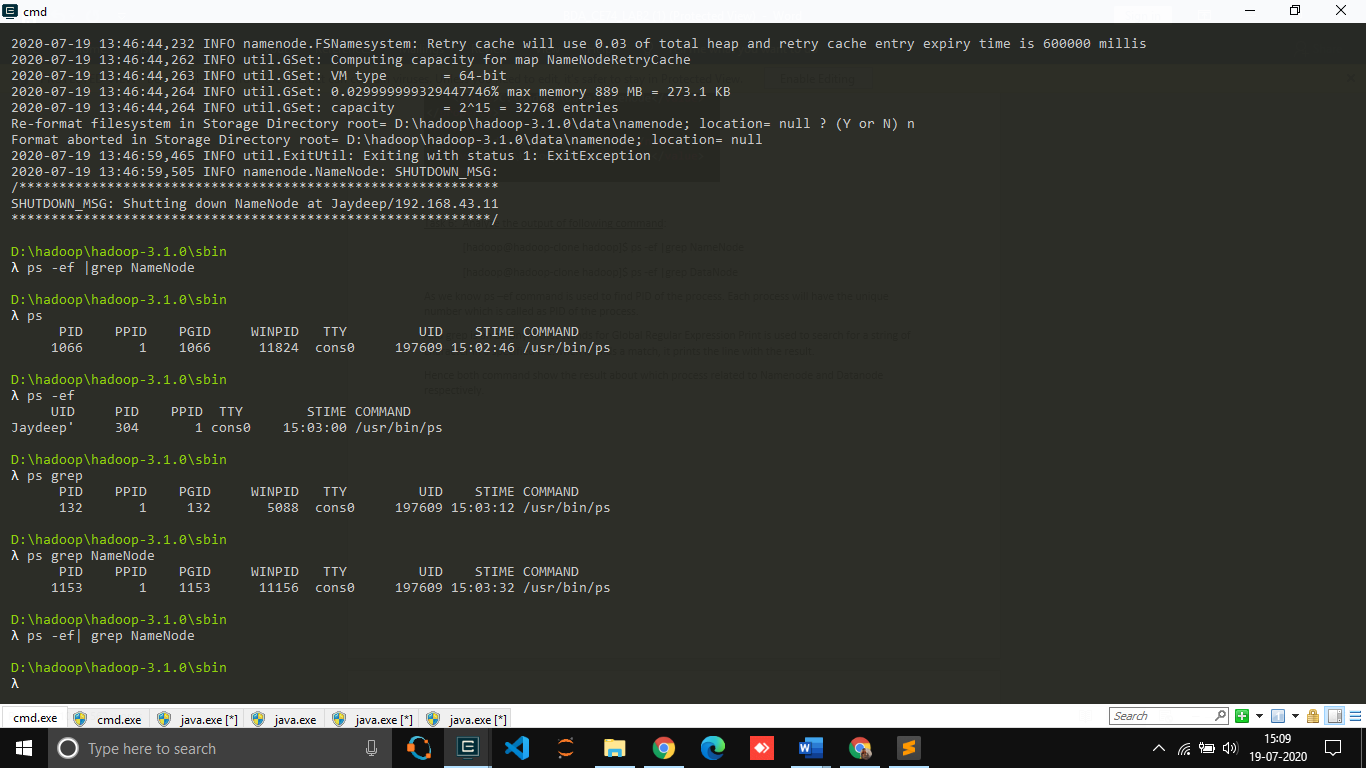
[hadoop@hadoop-clone hadoop] $ ps -ef |grep NameNode

[hadoop@hadoop-clone hadoop] $ ps -ef |grep DataNode

As we know the ps command enables you to check the status of active processes on a system, as well as display technical information about the processes.

With -ef option we can know the process id of process.

With grap option we can know what a particular process is doing.



In my opinion (as I Analyse the commands) the above 2 commands are alternative for jps command, to check whether Hadoop daemons are running or not. We can also check using Web UI.

References:

* <https://dimensionless.in/know-how-to-install-and-run-hadoop-on-windows-for-beginners/>
* <https://stackoverflow.com/questions/28513562/hadoop-resource-manager-not-starting>
* <https://stackoverflow.com/questions/15555965/how-to-check-if-hadoop-daemons-are-running>

**Jaydeep Mahajan | CE066**