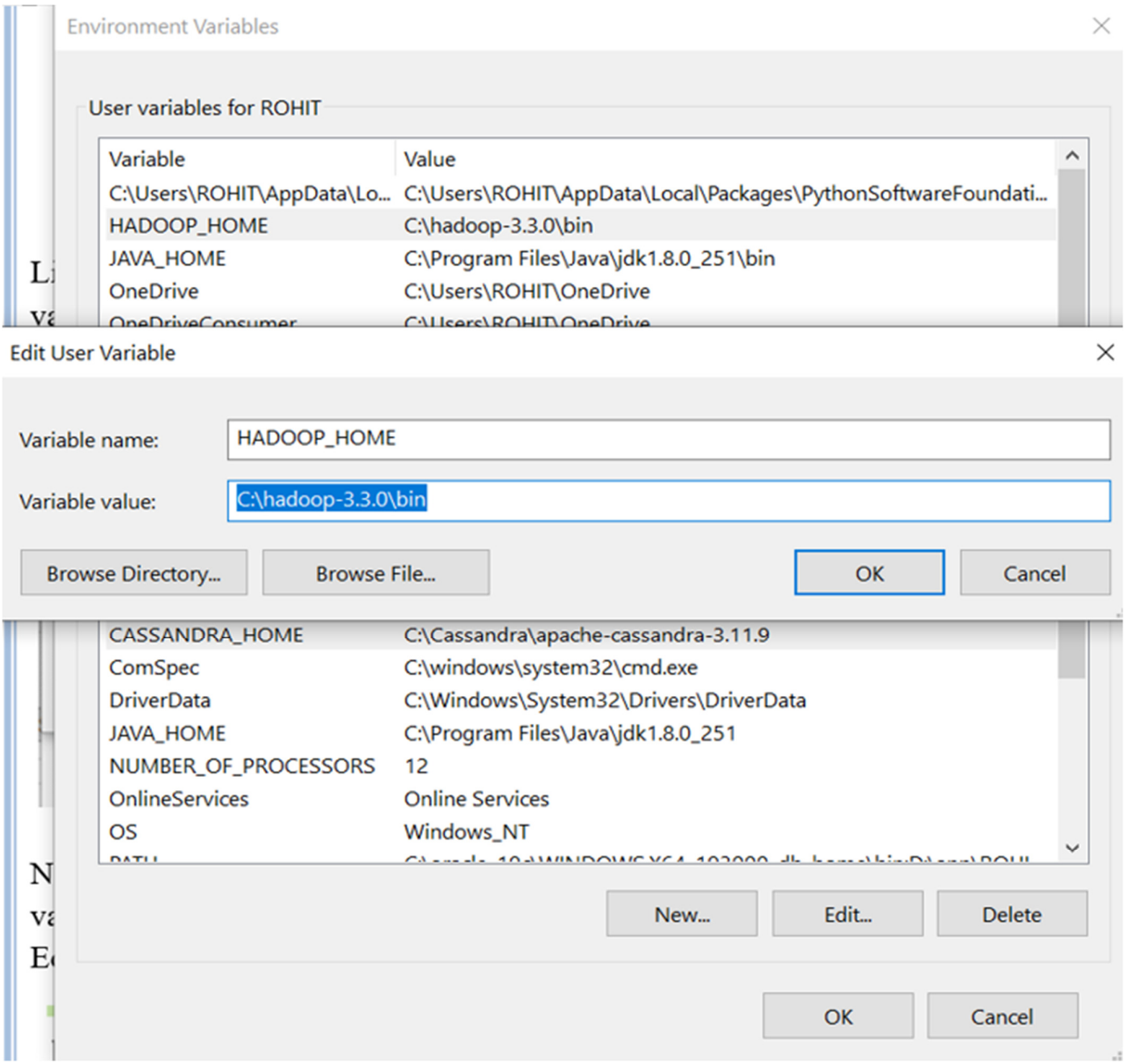
Practical:

Aim: Installation of Hadoop on Windows 10.

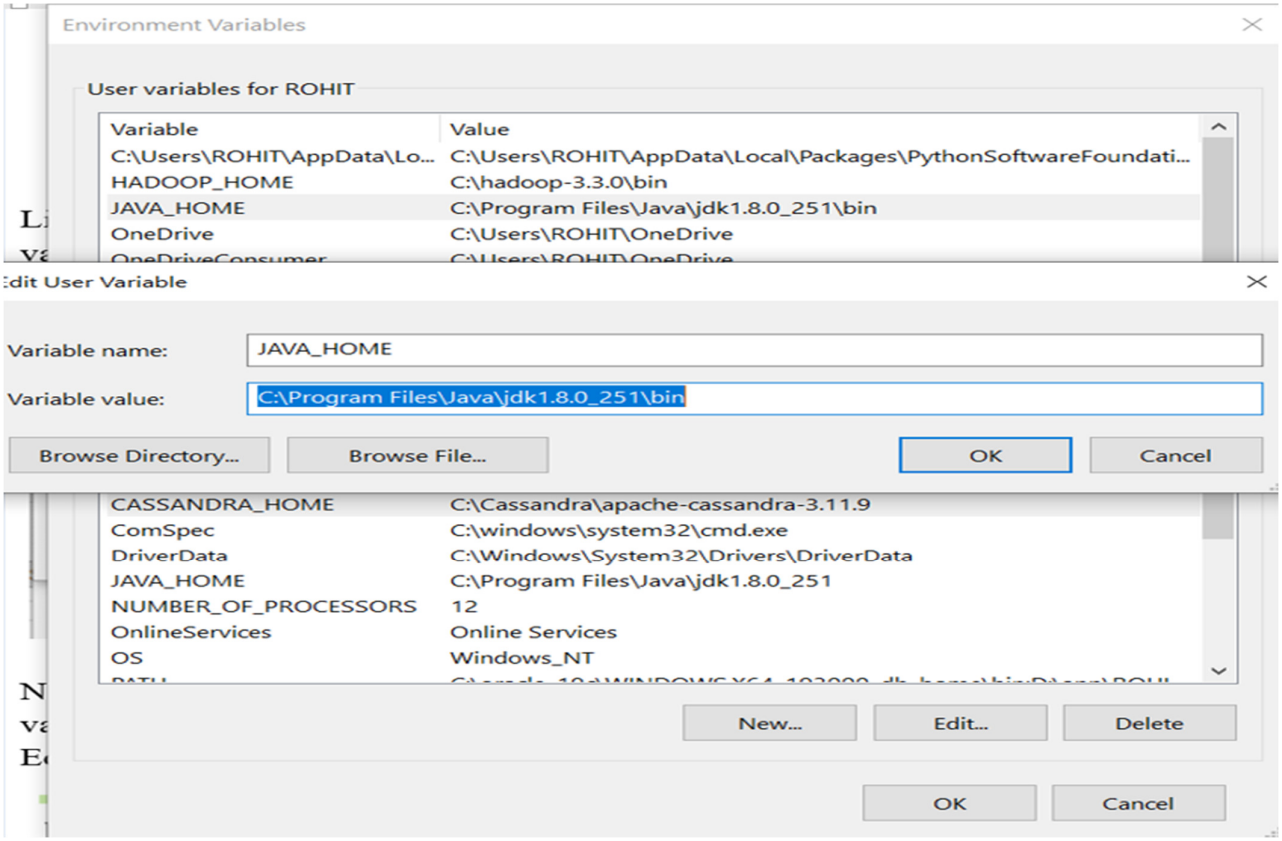
Writeup:

Aim: Hadoop Installation on Windows 10.

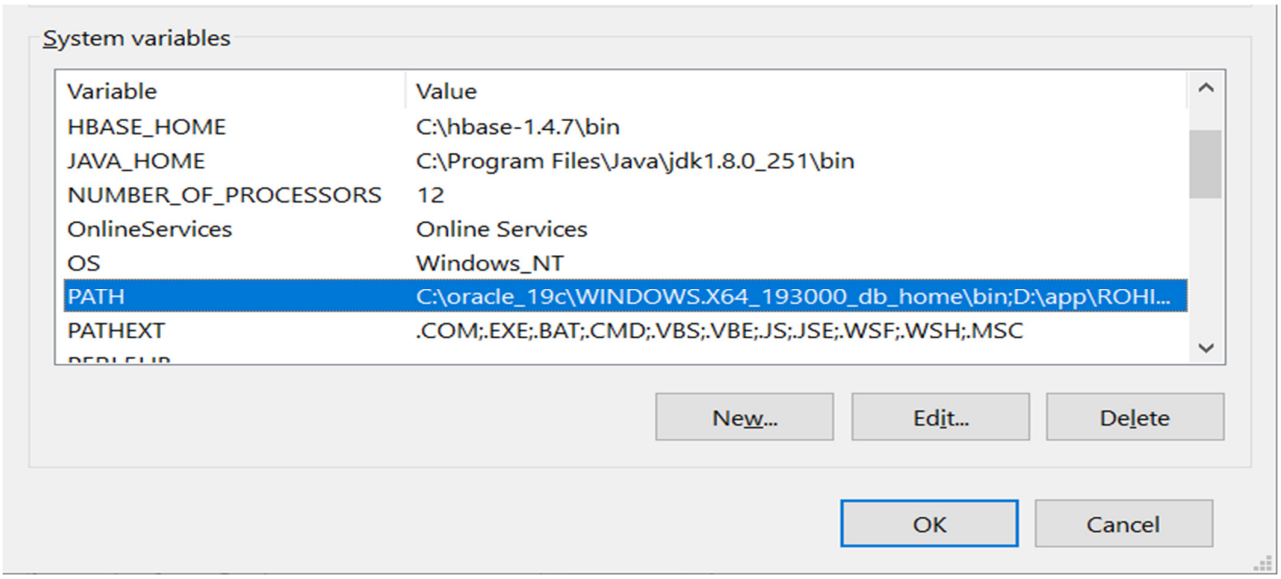
Prerequisite: To install Hadoop, you should have Java version 1.8 in your system.
Check your java version through this command on command prompt java -version
Create a new user variable. Put the Variable_name as “HADOOP_HOME” and Variable_value as the path of the bin folder where you extracted hadoop.



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.

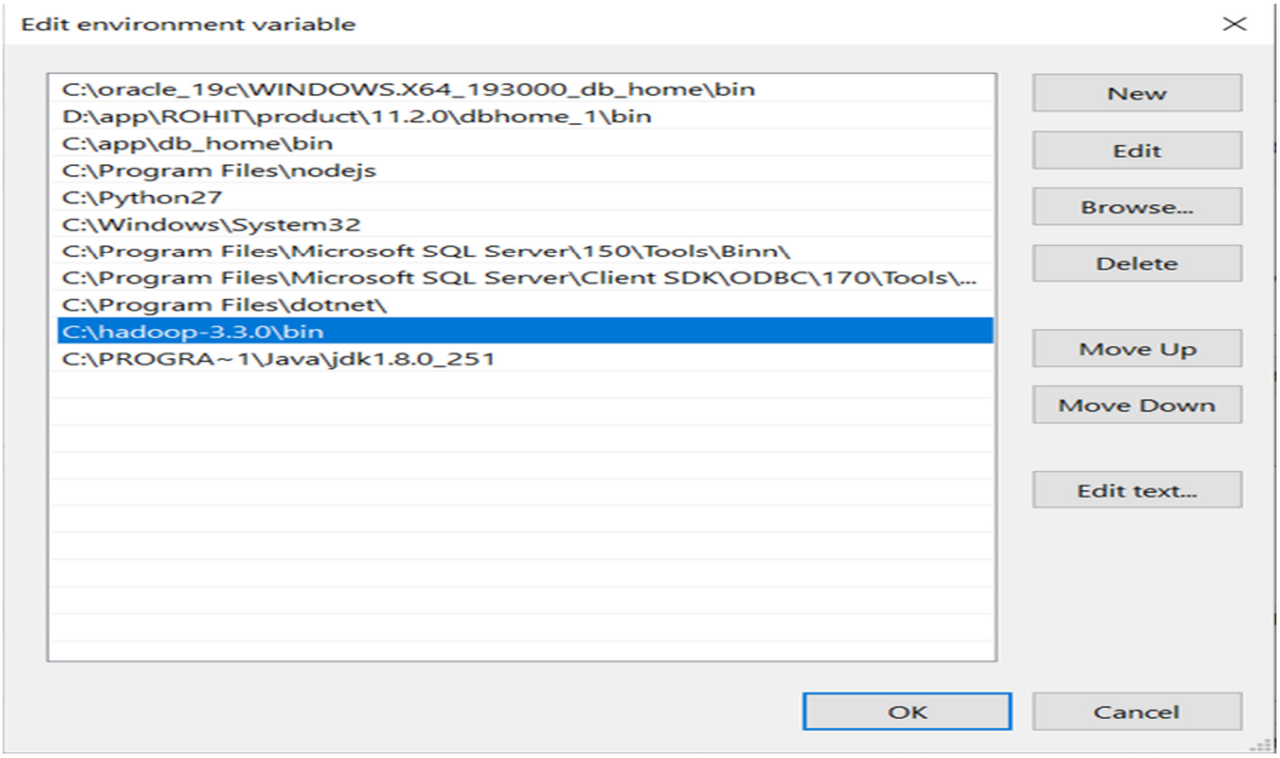


Now we need to set **Hadoop bin** directory and **Java bin** directory path in system variable path.
Edit Path in system variable.

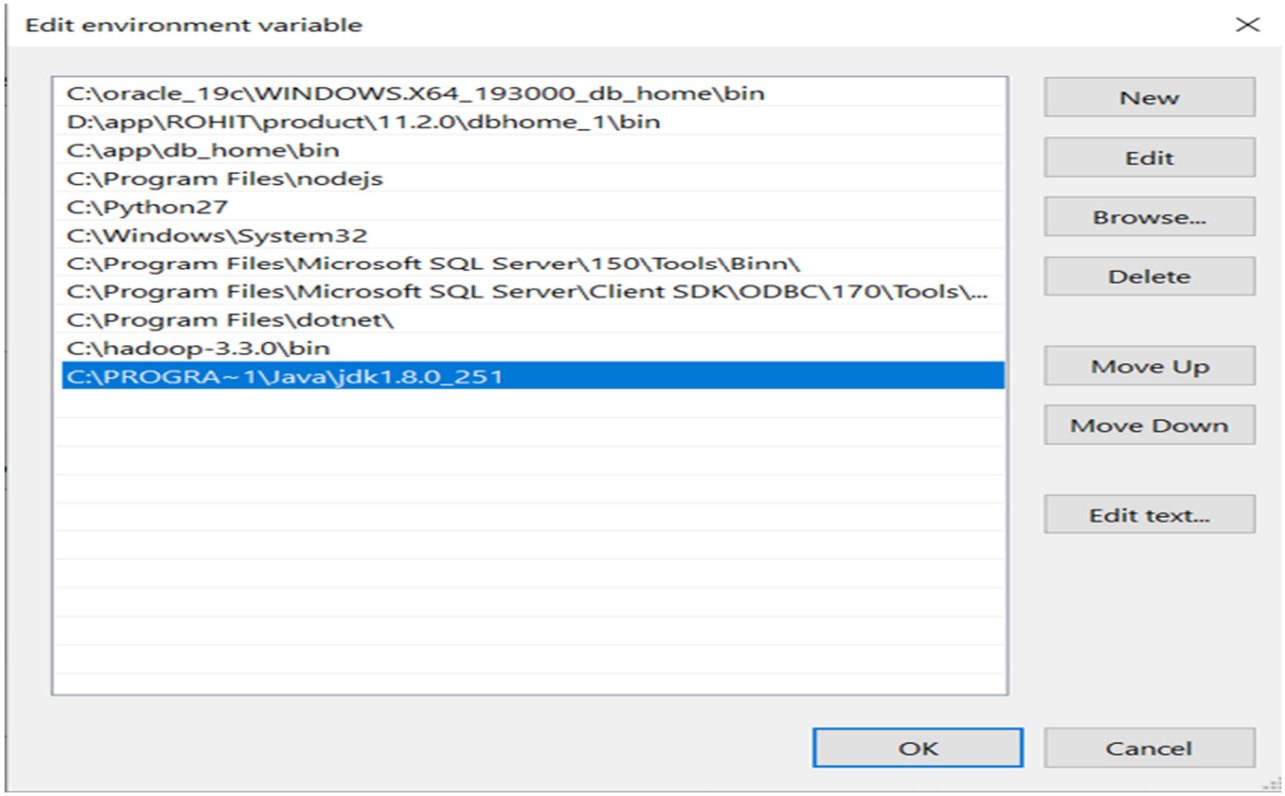


Click on New and add the bin directory path of Hadoop and Java in it.

Hadoop

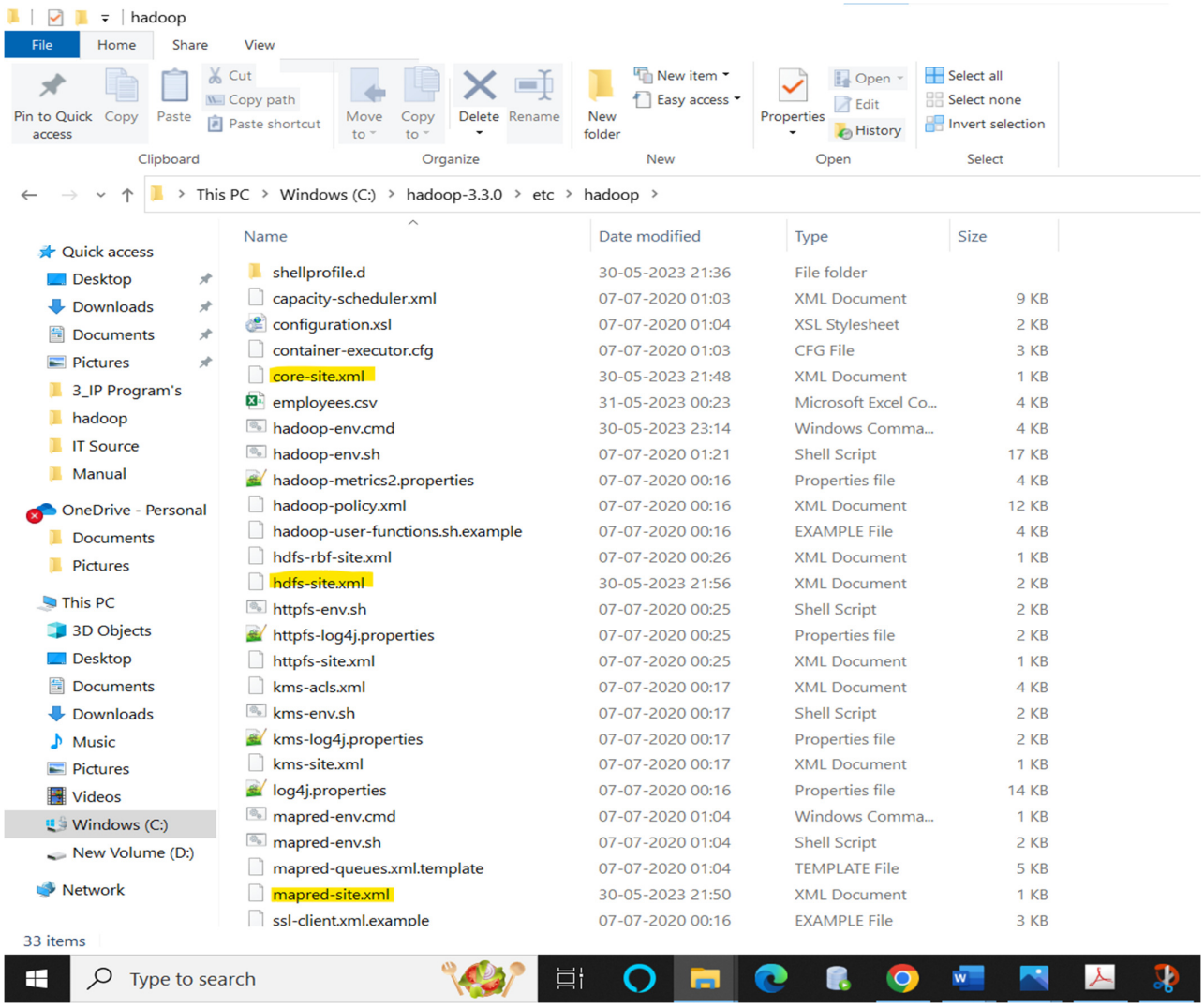


Java



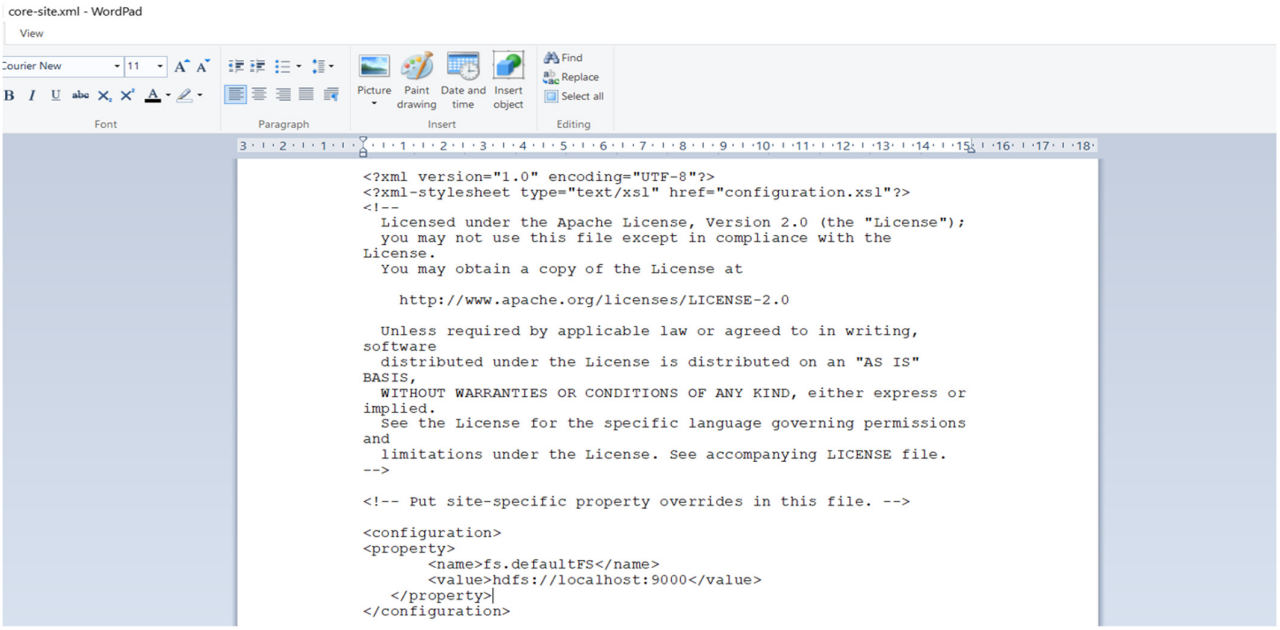
Configurations

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.



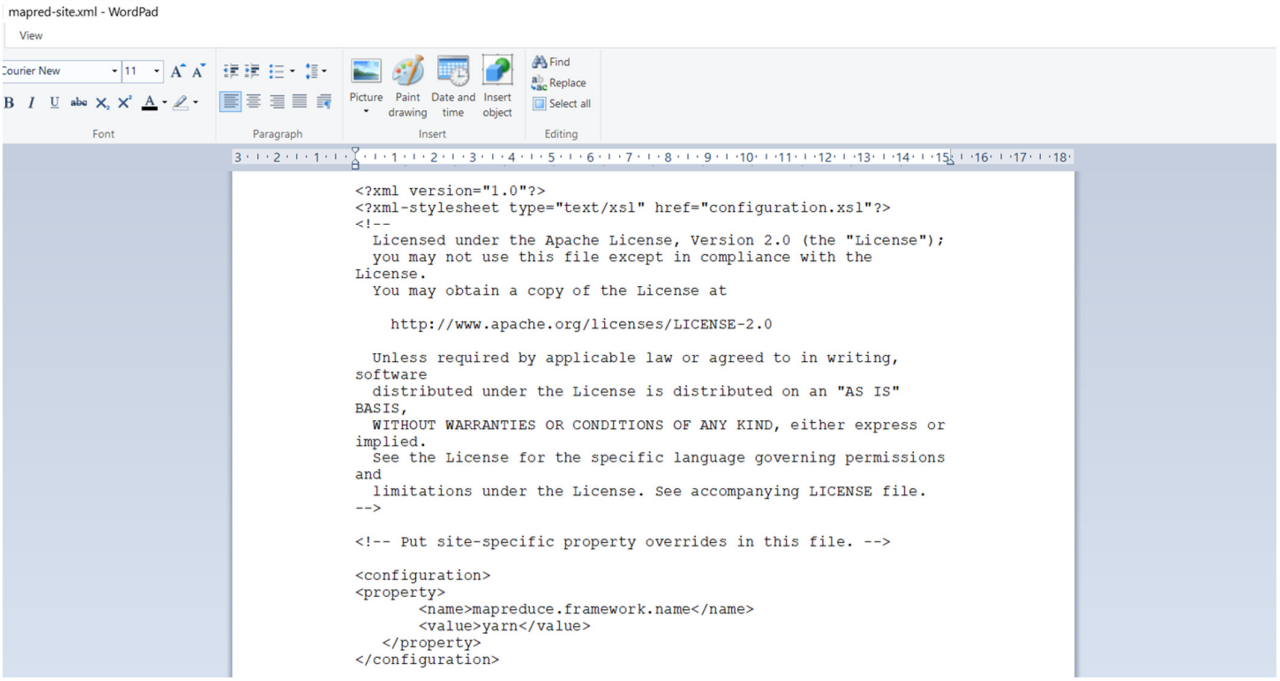
1. Edit the file **core-site.xml** in the hadoop directory. Copy this xml property in the configuration in the file.

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

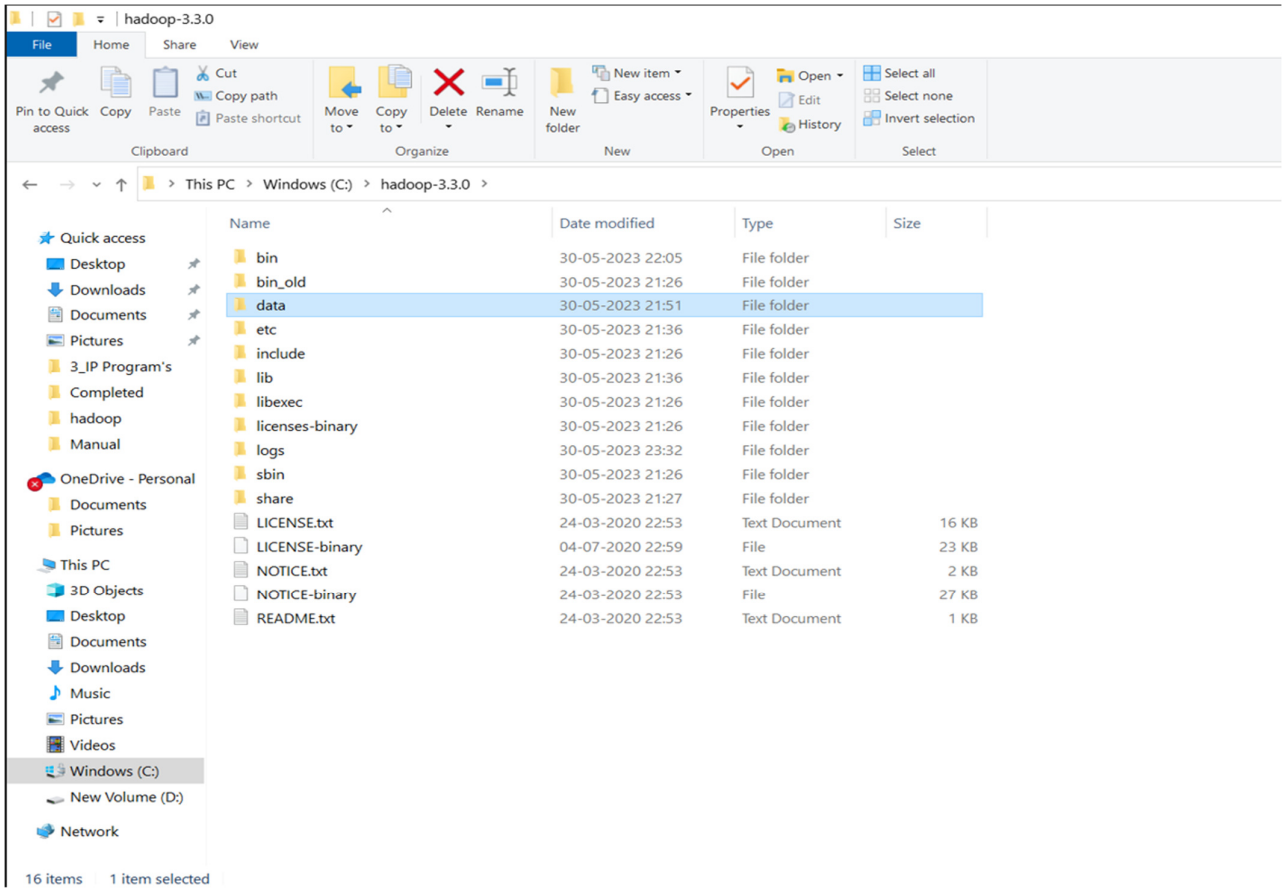


2. Edit **mapred-site.xml** and copy this property in the configuration.

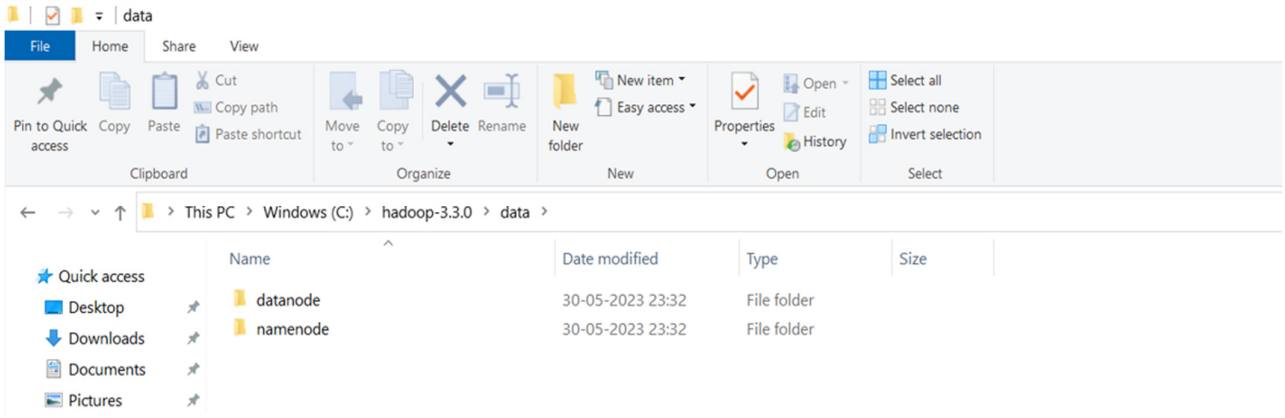
```
<configuration>
  <property>
     <name>mapreduce.framework.name</name>
     <value>yarn</value>
  </property>
</configuration>
```



3. Create a folder **"data"** in the hadoop directory.



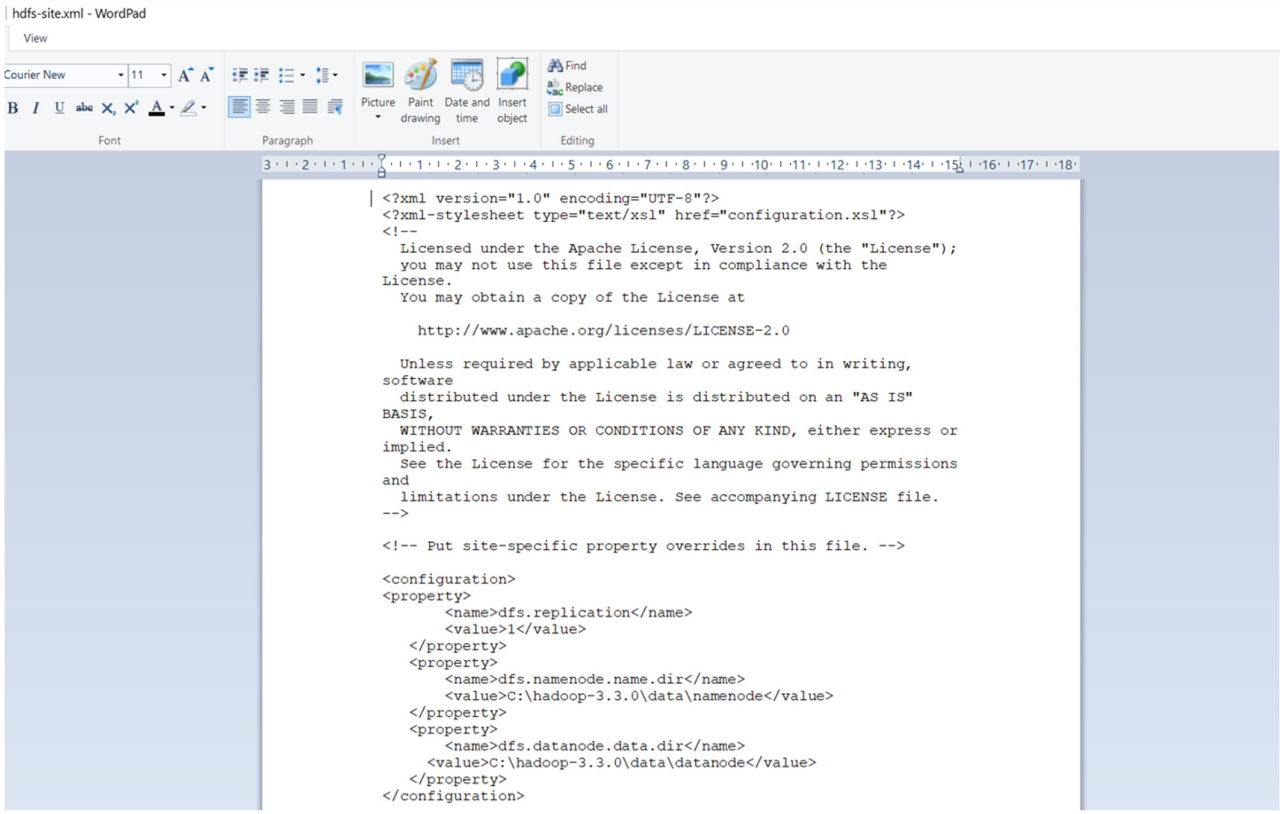
4. Create a folder with the name **"datanode"** and a folder **"namenode"** in this **data** directory



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

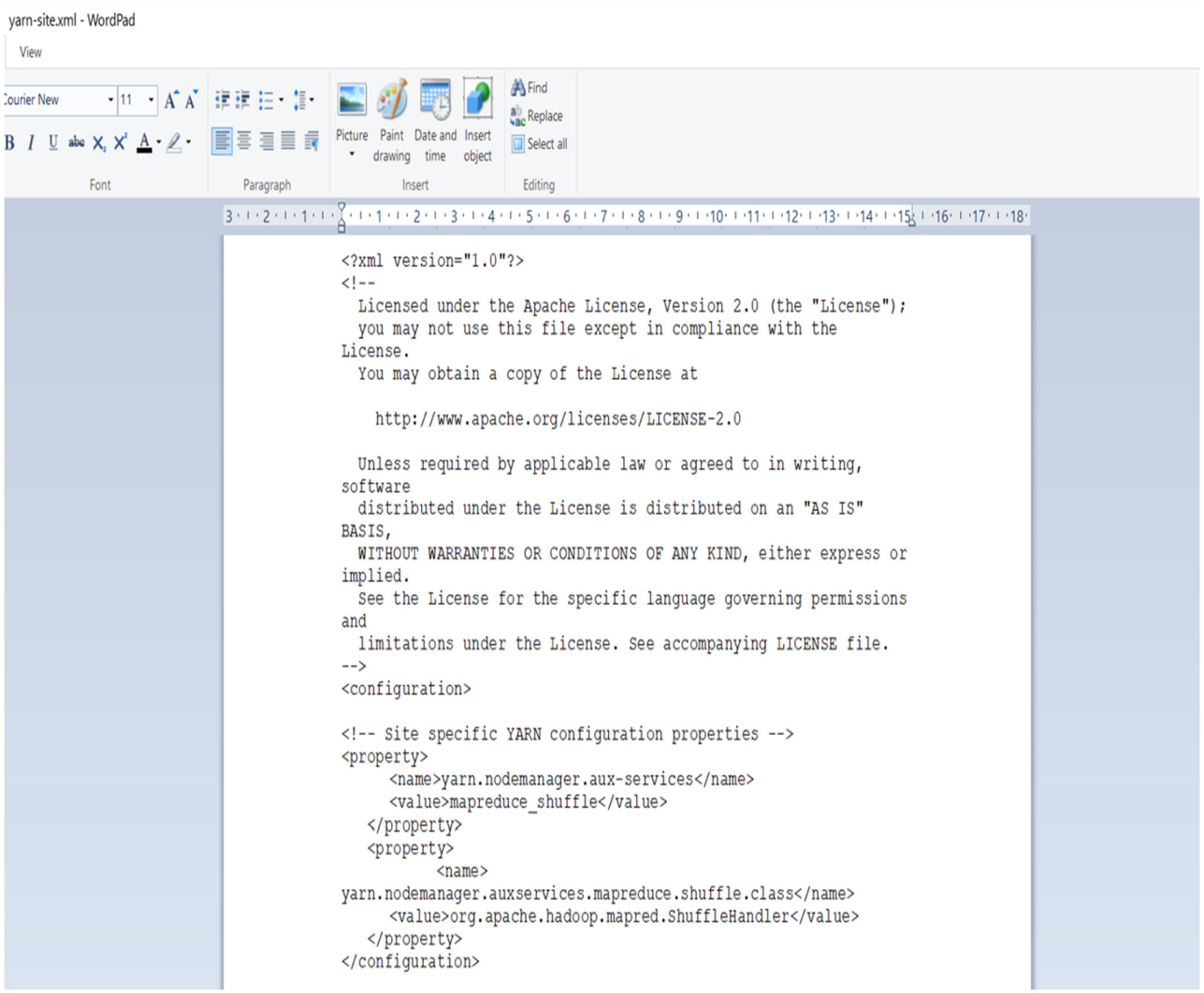
Note: The path of namenode and datanode across value would be the path of the datanode and namenode folders you just created.

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
   <value>C:\hadoop-3.3.0\data\namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
   <value>C:\hadoop-3.3.0\data\datanode</value>
  </property>
</configuration>
```

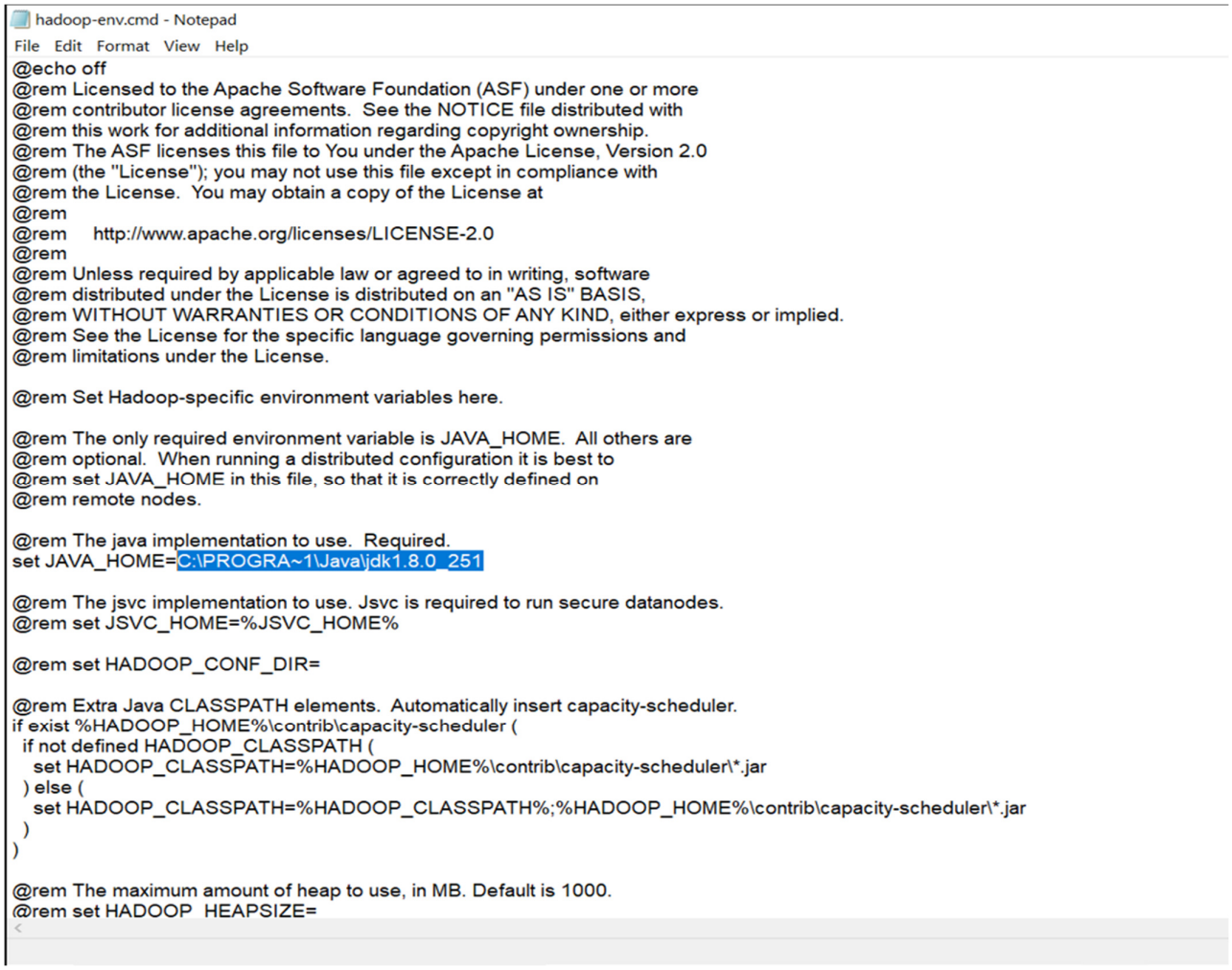


6. Edit the file **yarn-site.xml** and add below property in the configuration.

```
<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>
```



7. Edit **hadoop-env.cmd** and replace %JAVA_HOME% with the path of the java folder where your jdk 1.8 is installed.

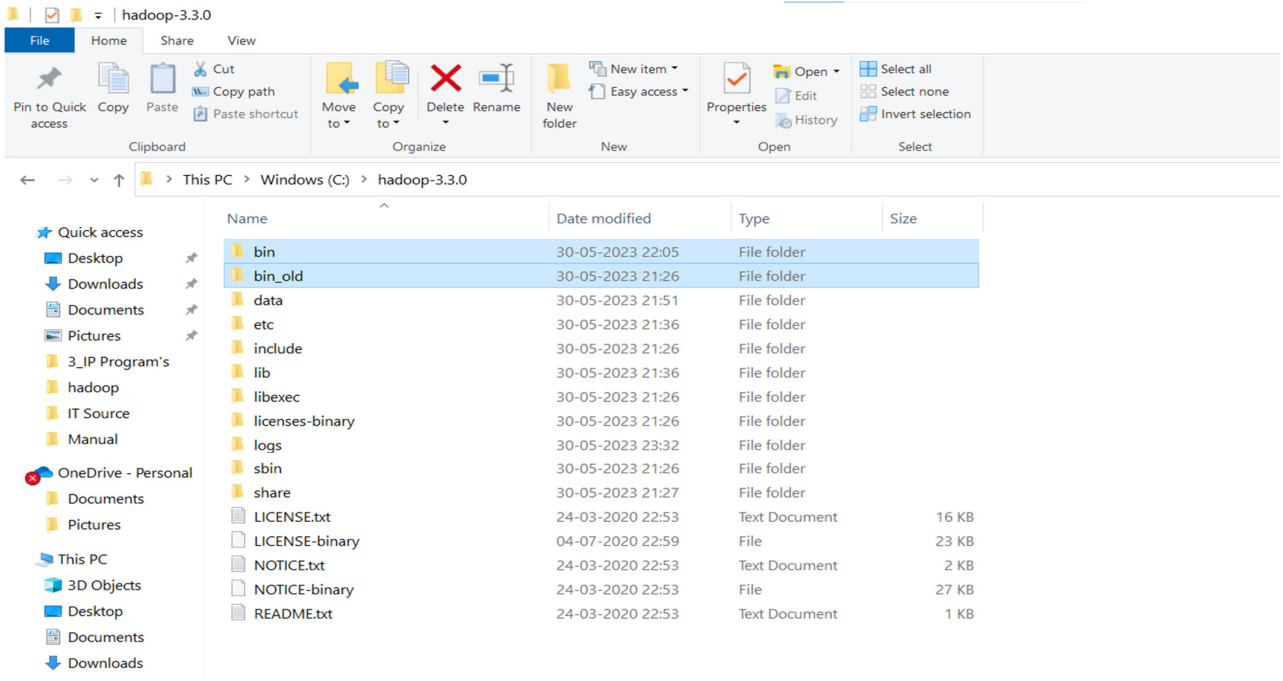


```
hadoop-env.cmd - Notepad
File  Edit  Format  View  Help
@echo off
@rem Licensed to the Apache Software Foundation (ASF) under one or more
@rem contributor license agreements.  See the NOTICE file distributed with
@rem this work for additional information regarding copyright ownership.
@rem The ASF licenses this file to You under the Apache License, Version 2.0
@rem (the "License"); you may not use this file except in compliance with
@rem the License.  You may obtain a copy of the License at
@rem
@rem     http://www.apache.org/licenses/LICENSE-2.0
@rem
@rem Unless required by applicable law or agreed to in writing, software
@rem distributed under the License is distributed on an "AS IS" BASIS,
@rem WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
@rem See the License for the specific language governing permissions and
@rem limitations under the License.

@rem Set Hadoop-specific environment variables here.

@rem The only required environment variable is JAVA_HOME.  All others are
@rem optional.  When running a distributed configuration it is best to
@rem set JAVA_HOME in this file, so that it is correctly defined on
@rem remote nodes.

@rem The java implementation to use.  Required.
set JAVA_HOME=C:\PROGRA~1\Java\jdk1.8.0_251

@rem The jsvc implementation to use. Jsvc is required to run secure datanodes.
@rem set JSVC_HOME=%JSVC_HOME%

@rem set HADOOP_CONF_DIR=

@rem Extra Java CLASSPATH elements.  Automatically insert capacity-scheduler.
if exist %HADOOP_HOME%\contrib\capacity-scheduler (
  if not defined HADOOP_CLASSPATH (
    set HADOOP_CLASSPATH=%HADOOP_HOME%\contrib\capacity-scheduler\*.jar
  ) else (
    set HADOOP_CLASSPATH=%HADOOP_CLASSPATH%;%HADOOP_HOME%\contrib\capacity-scheduler\*.jar
  )
)

@rem The maximum amount of heap to use, in MB. Default is 1000.
@rem set HADOOP_HEAPSIZE=
```

8. Hadoop needs windows OS specific files which does not come with default download of hadoop. 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

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-**hadoop –version**

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**

```
C:\Users\ROHIT>hdfs namenode -format
2023-05-30 23:29:47,088 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = LAPTOP-10MI48EB/192.168.43.154
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.0
STARTUP_MSG:   classpath = C:\hadoop-3.3.0\etc\hadoop;C:\hadoop-3.3.0\share\hadoop\common;C:\hadoop-3.3.0\share\hadoop\common\lib\accessors-smart-1.2.jar;C:\hadoop-3.3.0\share\hadoop\common
\lib\animal-sniffer-annotations-1.17.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\asm-5.0.4.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\audience-annotations-0.5.0.jar;C:\hadoop-3.3.0\share\h
adoop\common\lib\avro-1.7.7.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\checker-qual-2.5.2.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;C:\hadoop-3.3.0\share\hado
op\common\lib\commons-cli-1.2.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-codec-1.11.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-collections-3.2.2.jar;C:\hadoop-3.3.0\share\
hadoop\common\lib\commons-compress-1.19.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-configuration2-2.1.1.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-daemon-1.0.13.jar;C:\had
oop-3.3.0\share\hadoop\common\lib\commons-io-2.5.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-lang3-3.7.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-logging-1.1.3.jar;C:\hadoo
p-3.3.0\share\hadoop\common\lib\commons-math3-3.1.1.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-net-3.6.jar;C:\hadoop-3.3.0\share\hadoop\common\lib\commons-text-1.4.jar;C:\hadoop-3.
```

Now change the directory in cmd to **sbin** folder of hadoop directory with this command,

Start namenode and datanode with this command – **start-dfs.cmd**

Two more cmd windows will open for **NameNode** and **DataNode**

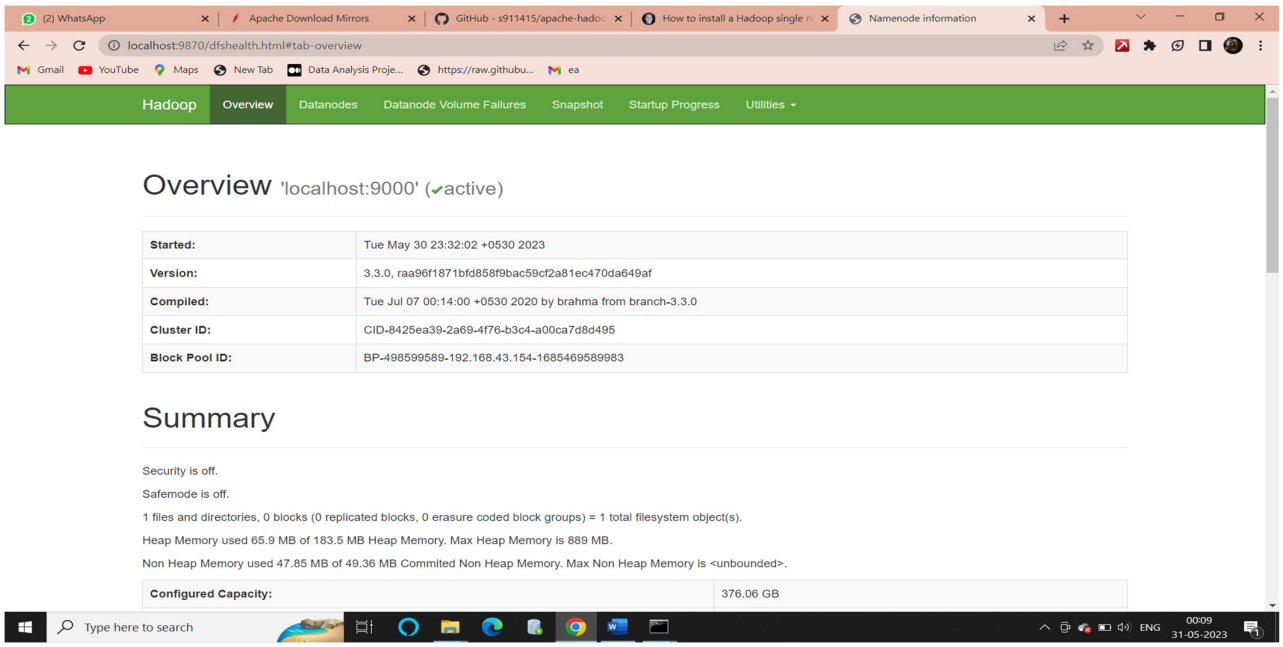
Now start yarn through this command - **start-yarn.cmd**

```
Command Prompt
Microsoft Windows [Version 10.0.19045.2965]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ROHIT>cd\

C:\>cd hadoop-3.3.0

C:\hadoop-3.3.0>cd sbin

C:\hadoop-3.3.0\sbin>start-dfs.cmd

C:\hadoop-3.3.0\sbin>start-yarn.cmd
starting yarn daemons

C:\hadoop-3.3.0\sbin>
```

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

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) - http://localhost:9870/



**********END**********

Aim: Exploring Hadoop Distributed File System (HDFS)

To implement the following file management tasks in Hadoop System (HDFS): Adding files and directories, Retrieving files, Deleting files

HDFS COMMANDS

Local File System and HDFS(Hadoop Distributed File System)

Local file system is the file system of your own computer. Say if you are using windows, windows operating system will be having it's own way of managing files and folders. Same applies for linux.

We have also seen Hadoop has got it's own way of storing files and that is called HDFS. We have also seen, the data nodes are independent computers which gets the instruction from Name Node for storing the files. So if you think a little, a Name Node is actually using the Hadoop Distributed File System where as the Data Nodes uses it's own file system (i.e linux).

HDFS Command line

The two commands that helps us to interact with the HDFS are 'hadoop fs' and 'hdfs dfs'. The only difference is 'hdfs dfs' helps us to deal only with the HDFS file system and using 'hadoop fs' we can work with other file systems as well.

1. Creating a directory in HDFS

The **'mkdir'** command is used to create a directory in HDFS.

Syntax :

hadoop fs -mkdir /<directory-name>

Example : hadoop fs -mkdir /mscit

A directory named 'mscit' is created under the root directory.

2. Copying files from local file system to HDFS

To copy the files from local file system to HDFS **'copyFromLocal'** command is used.

Syntax :

hadoop fs -copyFromLocal <local-file-path> <hdfs-file-path>

Example: hadoop fs -copyFromLocal C:\hadoop-3.3.0\etc\hadoop\employees.csv /mscit
The above command copies employee.csv from your local file system to the newly created directory 'mscit' in HDFS.

put command

Syntax: *hadoop fs -put <local-file-path> <hdfs-file-path>*

Example: hadoop fs -put C:\hadoop-3.3.0\etc\hadoop\employees.csv /mscit
The above command does the same thing. i.e. Copies employee.csv from your local file system to the newly created directory "mscit" in HDFS.

3. Copying files from HDFS to local file system

To copy the files from HDFS to local file system 'get' command is used.

Syntax: *hadoop fs -get <hdfs-file-path> <local-file-path>*

Example: hadoop fs -get \mscit\employee.csv C:\hadoop-3.3.0\etc\hadoop

```
C:\Users\ROHIT>hadoop fs -get \mscit\employees.csv C:\hadoop-3.3.0\etc\hadoop
get: `C:/hadoop-3.3.0/etc/hadoop/employees.csv': File exists
```

4. Viewing a file in HDFS

To view a file in HDFS 'cat' command is used.

Syntax: *hadoop fs –cat <filename>*

Example: hadoop fs -cat \mscit\employees.csv

```
C:\Users\ROHIT>hadoop fs -cat \mscit\employees.csv
EMPLOYEE_ID,FIRST_NAME,LAST_NAME,EMAIL,PHONE_NUMBER,HIRE_DATE,JOB_ID,SALARY,COMMISSION_PCT,MANAGER_ID,DEPARTMENT_ID
198,Donald,OConnell,DOCONNEL,650.507.9833,21-JUN-07,SH_CLERK,2600, - ,124,50
199,Douglas,Grant,DGRANT,650.507.9844,13-JAN-08,SH_CLERK,2600, - ,124,50
200,Jennifer,Whalen,JWHALEN,515.123.4444,17-SEP-03,AD_ASST,4400, - ,101,10
201,Michael,Hartstein,MHARTSTE,515.123.5555,17-FEB-04,MK_MAN,13000, - ,100,20
202,Pat,Fay,PFAY,603.123.6666,17-AUG-05,MK_REP,6000, - ,201,20
203,Susan,Mavris,SMAVRIS,515.123.7777,07-JUN-02,HR_REP,6500, - ,101,40
204,Hermann,Baer,HBAER,515.123.8888,07-JUN-02,PR_REP,10000, - ,101,70
205,Shelley,Higgins,SHIGGINS,515.123.8080,07-JUN-02,AC_MGR,12008, - ,101,110
206,William,Gietz,WGIETZ,515.123.8181,07-JUN-02,AC_ACCOUNT,8300, - ,205,110
100,Steven,King,SKING,515.123.4567,17-JUN-03,AD_PRES,24000, - , - ,90
101,Neena,Kochhar,NKOCHHAR,515.123.4568,21-SEP-05,AD_VP,17000, - ,100,90
102,Lex,De Haan,LDEHAAN,515.123.4569,13-JAN-01,AD_VP,17000, - ,100,90
103,Alexander,Hunold,AHUNOLD,590.423.4567,03-JAN-06,IT_PROG,9000, - ,102,60
104,Bruce,Ernst,BERNST,590.423.4568,21-MAY-07,IT_PROG,6000, - ,103,60
105,David,Austin,DAUSTIN,590.423.4569,25-JUN-05,IT_PROG,4800, - ,103,60
106,Valli,Pataballa,VPATABAL,590.423.4560,05-FEB-06,IT_PROG,4800, - ,103,60
107,Diana,Lorentz,DLORENTZ,590.423.5567,07-FEB-07,IT_PROG,4200, - ,103,60
108,Nancy,Greenberg,NGREENBE,515.124.4569,17-AUG-02,FI_MGR,12008, - ,101,100
109,Daniel,Faviet,DFAVIET,515.124.4169,16-AUG-02,FI_ACCOUNT,9000, - ,108,100
110,John,Chen,JCHEN,515.124.4269,28-SEP-05,FI_ACCOUNT,8200, - ,108,100
111,Ismael,Sciarra,ISCIARRA,515.124.4369,30-SEP-05,FI_ACCOUNT,7700, - ,108,100
112,Jose Manuel,Urman,JMURMAN,515.124.4469,07-MAR-06,FI_ACCOUNT,7800, - ,108,100
113,Luis,Popp,LPOPP,515.124.4567,07-DEC-07,FI_ACCOUNT,6900, - ,108,100
114,Den,Raphaely,DRAPHEAL,515.127.4561,07-DEC-02,PU_MAN,11000, - ,100,30
115,Alexander,Khoo,AKHOO,515.127.4562,18-MAY-03,PU_CLERK,3100, - ,114,30
116,Shelli,Baida,SBAIDA,515.127.4563,24-DEC-05,PU_CLERK,2900, - ,114,30
117,Sigal,Tobias,STOBIAS,515.127.4564,24-JUL-05,PU_CLERK,2800, - ,114,30
118,Guy,Himuro,GHIMURO,515.127.4565,15-NOV-06,PU_CLERK,2600, - ,114,30
119,Karen,Colmenares,KCOLMENA,515.127.4566,10-AUG-07,PU_CLERK,2500, - ,114,30
120,Matthew,Weiss,MWEISS,650.123.1234,18-JUL-04,ST_MAN,8000, - ,100,50
```

5. Display the contents of a directory in HDFS

To display the contents if a directory in HDFS 'ls' command is used.

Syntax: *hadoop fs –ls <filename>*

Example: hadoop fs -ls /mscit

```
C:\Users\ROHIT>hadoop fs -ls \mscit
Found 1 items
-rw-r--r--   1 ROHIT supergroup       3778 2023-05-31 00:32 /mscit/employees.csv
```

6. Deleting all files from a directory in HDFS

To delete files from HDFS 'rm' command is used.

Syntax: *hadoop fs –rm /<directory-name>/**

Example: hadoop fs -rm /mscit/*

```
C:\Users\ROHIT>hadoop fs -rm \mscit\*
Deleted /mscit/employees.csv

C:\Users\ROHIT>
```

**********END**********