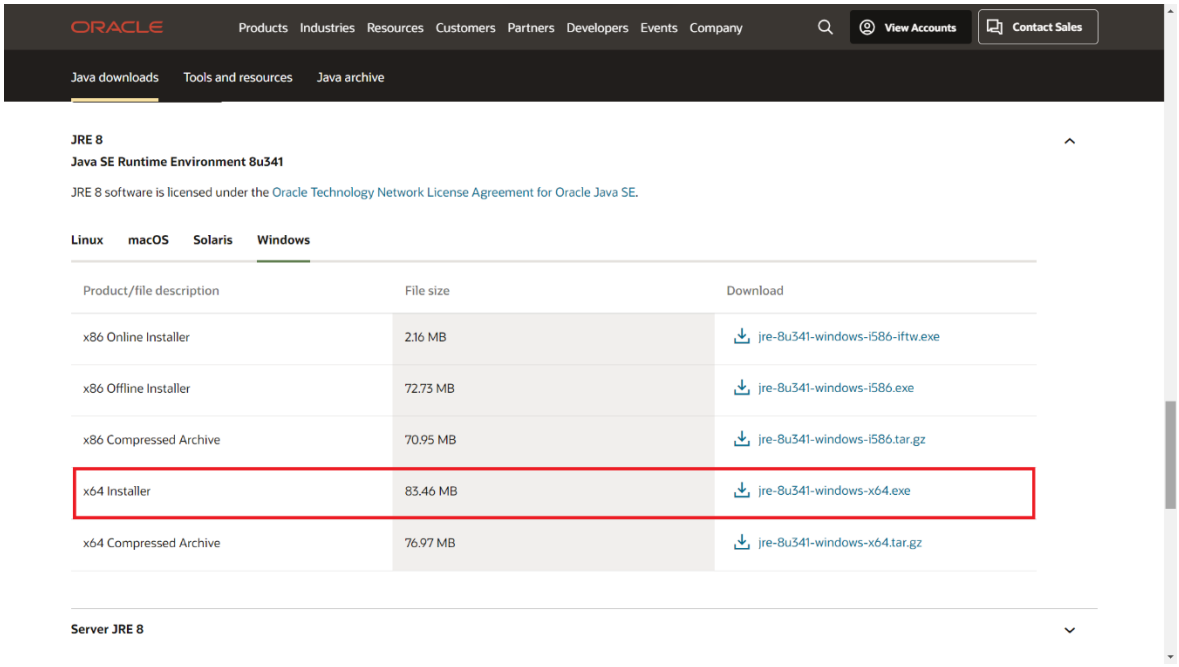
M2 Assignment 1 (Migration Process)

I. Java 8 Runtime Environment and Development Kit Installation

Before installing Hadoop into the local machine, Java 8 Runtime Environment and Development Kit had to be installed as these are required for Hadoop to work. To download JRE 8, follow this link and click "Download Java" to start the download: https://www.oracle.com/java/technologies/downloads/#jre8-windows.

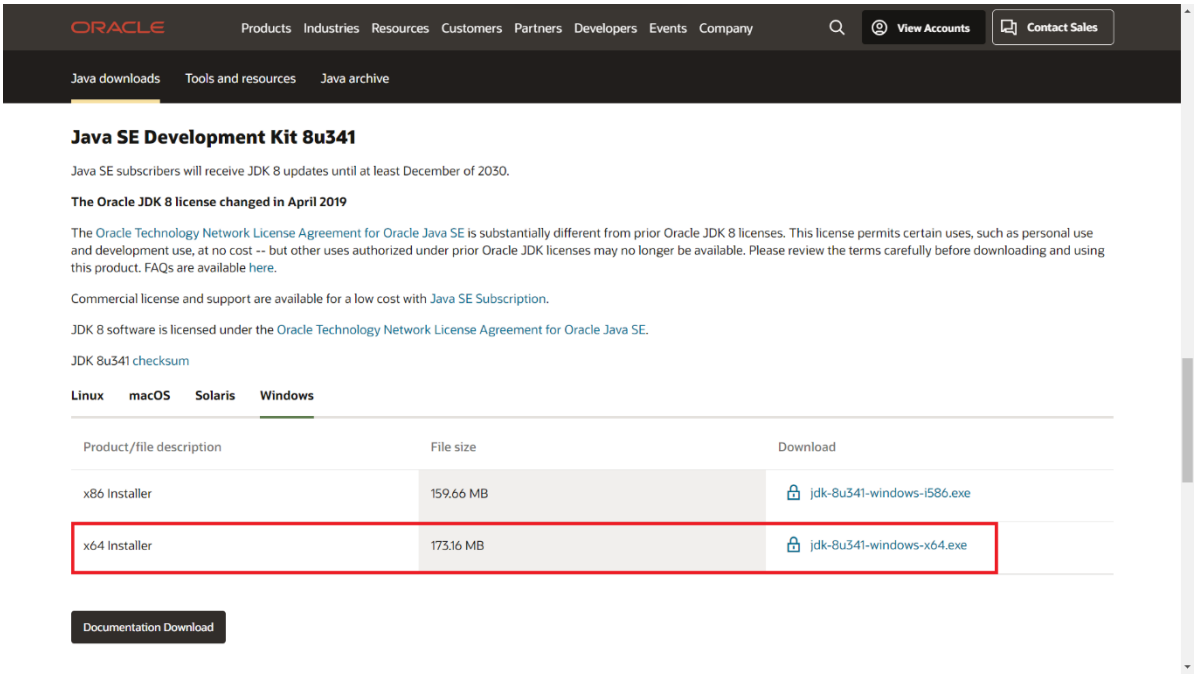


Then, save the file on the preferred directory, as shown below.

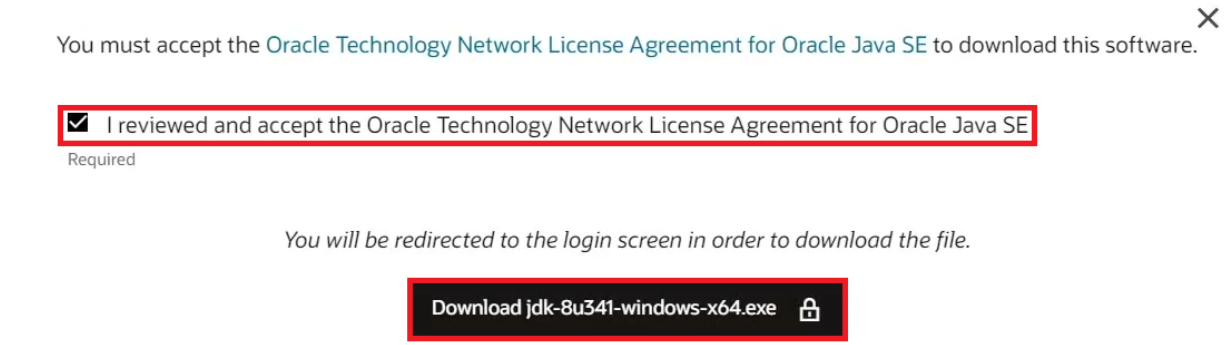
Once the installer has finished downloading, double click the application and follow the steps to complete the installation process.

Click “Install,” then wait for the installation to complete.

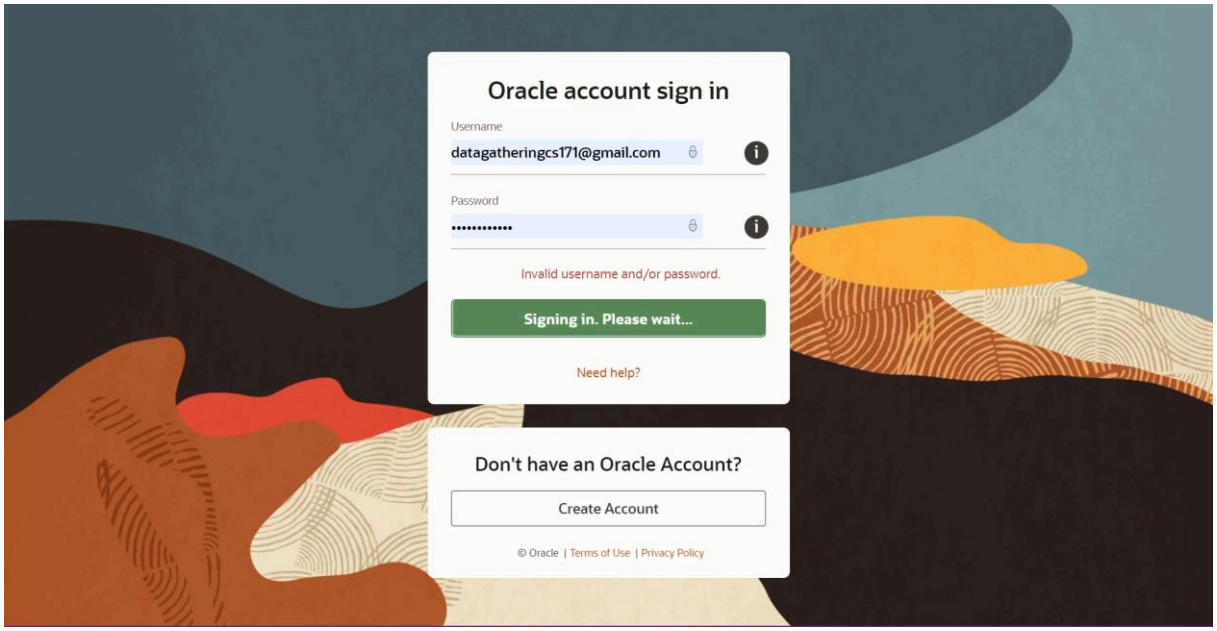
The next step after setting up the Java Runtime Environment is installing the Java SE Development Kit. To do this, follow this link and select the installer that is most appropriate for your local machine: https://www.oracle.com/java/technologies/downloads/#java8-windows. In this installation guide, the x64 installer was downloaded.



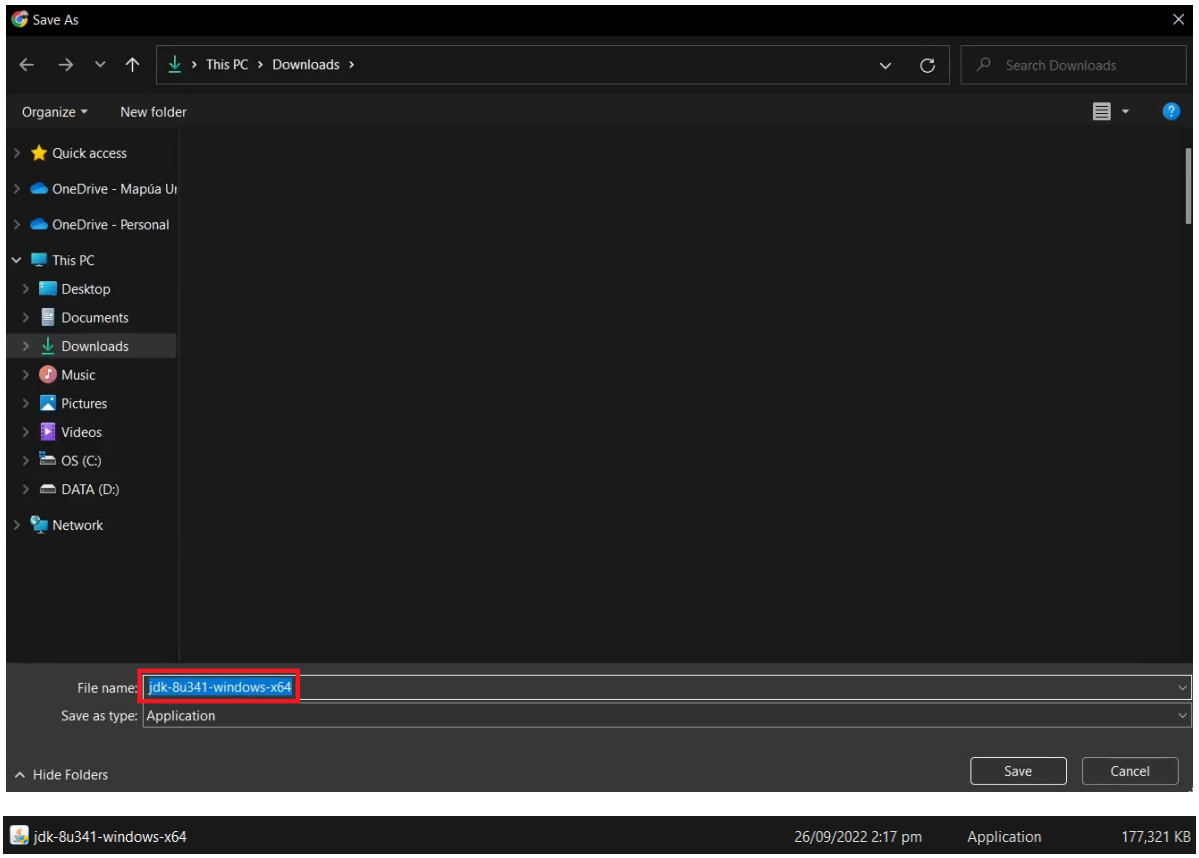
The following prompt should appear after selecting *jdk-8u341-windows-x64.exe*. Click the checkbox to agree with Oracle's terms and conditions and the download button to begin the download.



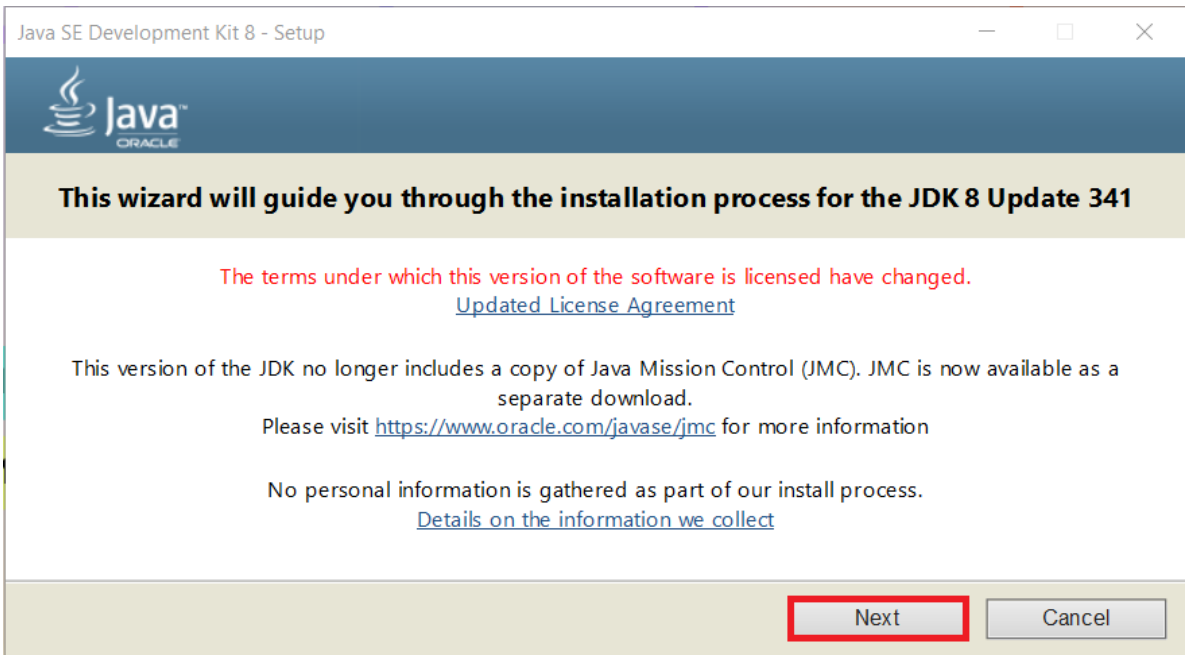
Clicking the download button will redirect the user to the Oracle account sign in window, where they will be prompted to enter a username and password to download the JDK installer. Since our group has previously created a joint Oracle account for situations like these, we used that account to download the JDK installer.



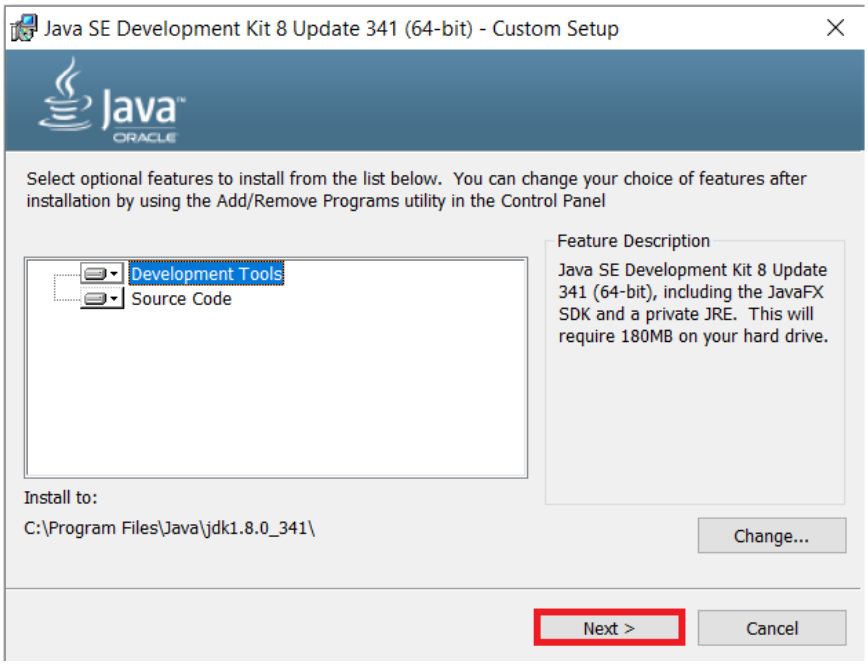
After signing in, the download should begin promptly, as indicated below.



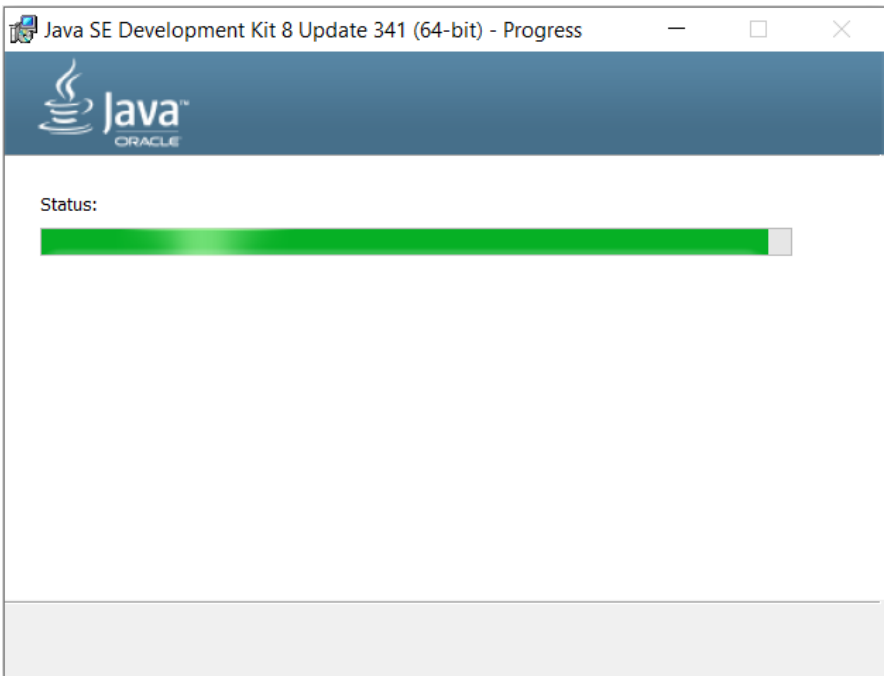
Double click on this downloaded executable file. Click ‘Next’ once the initial setup screen below is shown.



There will be a prompt to select and install optional features from a given list and change the installation folder for Java 8. Continue with the default settings and click on ‘Next’ to proceed.



We then wait for a couple of minutes for Java SE 8 to finish installing.



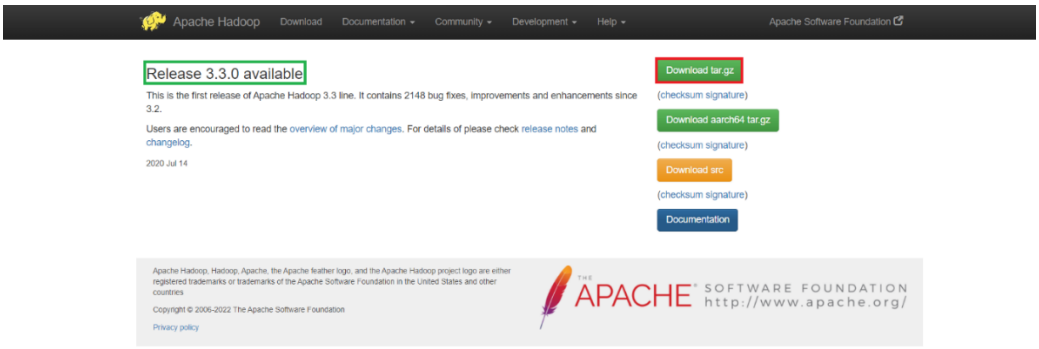
Once the screen below appears, Java SE 8 has been successfully installed on your computer. Click on ‘Close’ to end the setup wizard.

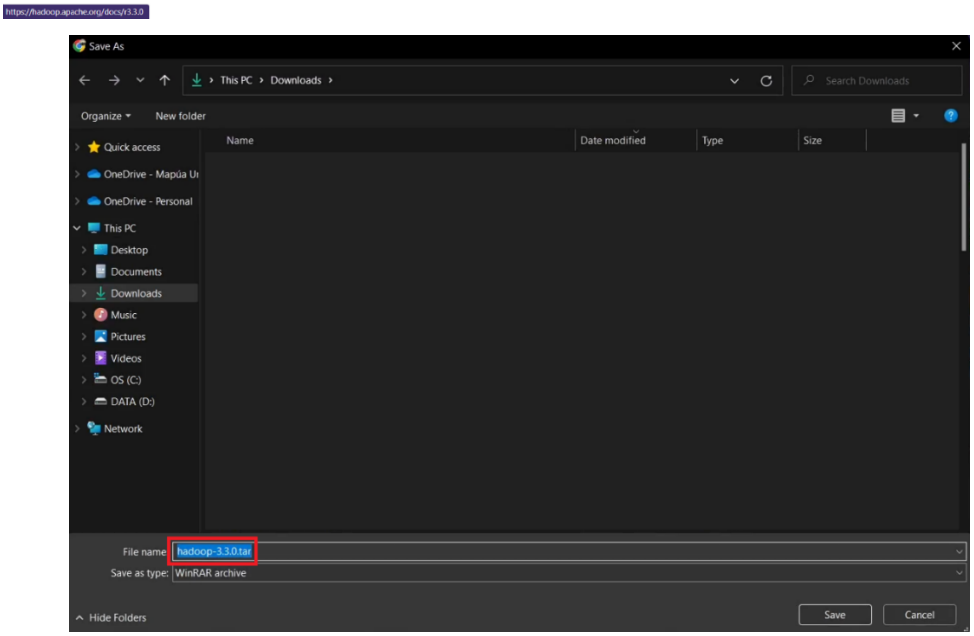
II. Hadoop Installation

a. Download Hadoop binaries (winutils)

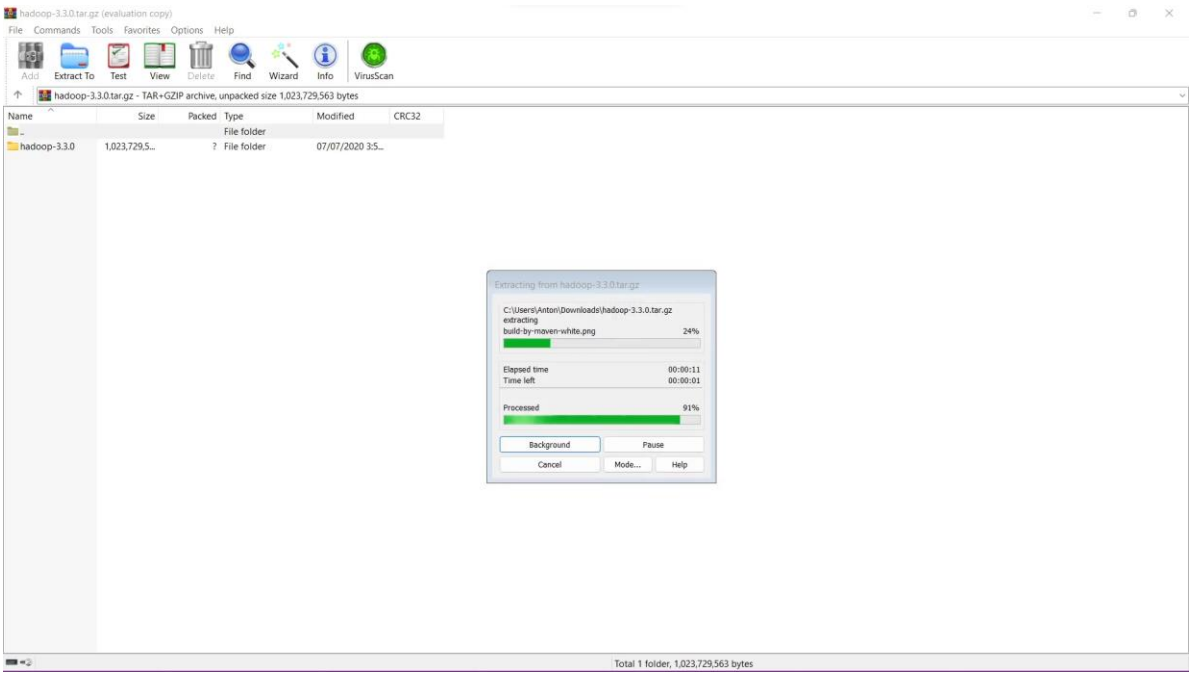
Before downloading the Hadoop binaries, the directory '*C:\hadoop-env*' was created. This folder is where all the necessary downloads in this installation guide will be placed.

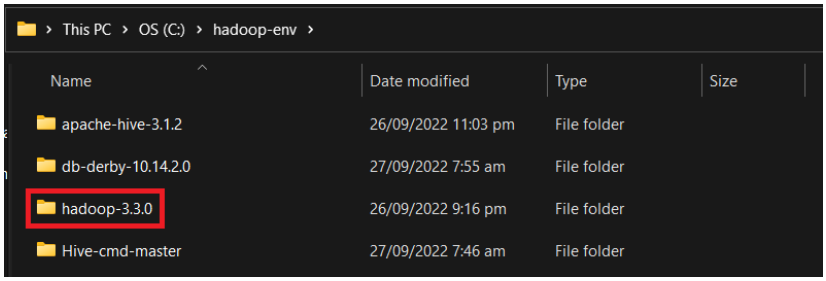
Now that the required JRE and JDK installations have been completed, Hadoop can already be downloaded by first clicking on the "Download tar.gz" button on this link: https://hadoop.apache.org/release/3.3.0.html. Clicking this button will download the zipped file into the user's specified directory.



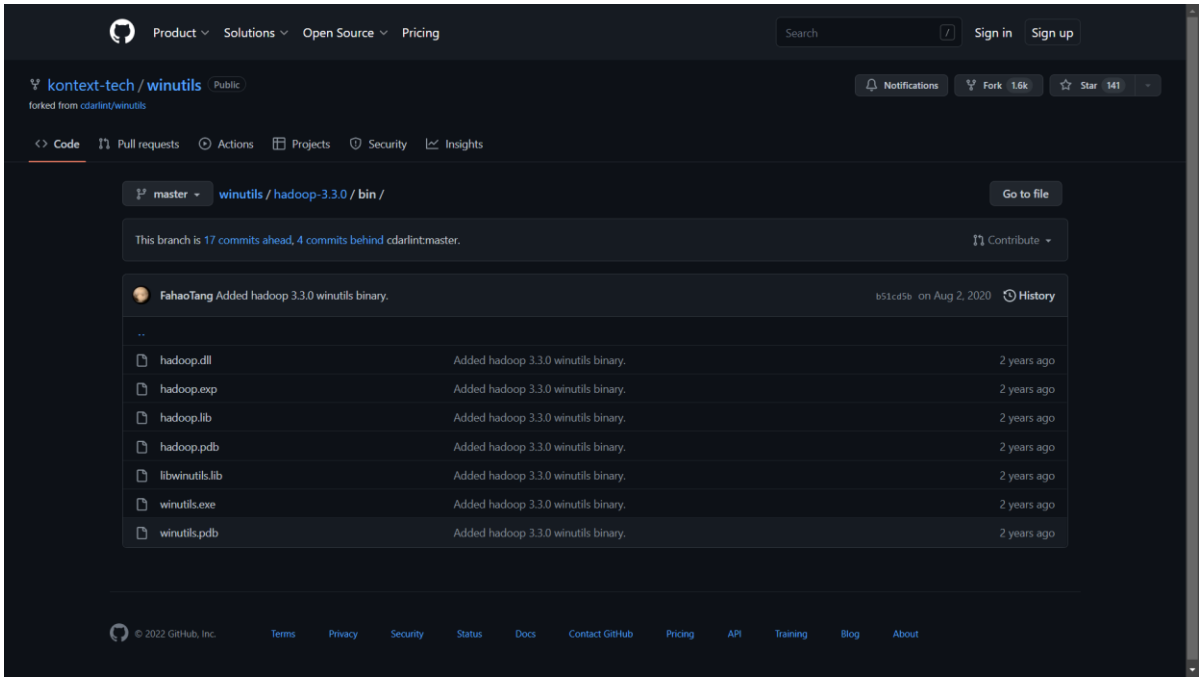


Once the file finishes downloading, unzip the file by extracting it using the WinRAR Archiver. (Other unzipping tools may also be used.) Then, move the unzipped file to the *hadoop-env* folder that was created prior to downloading the Hadoop package.

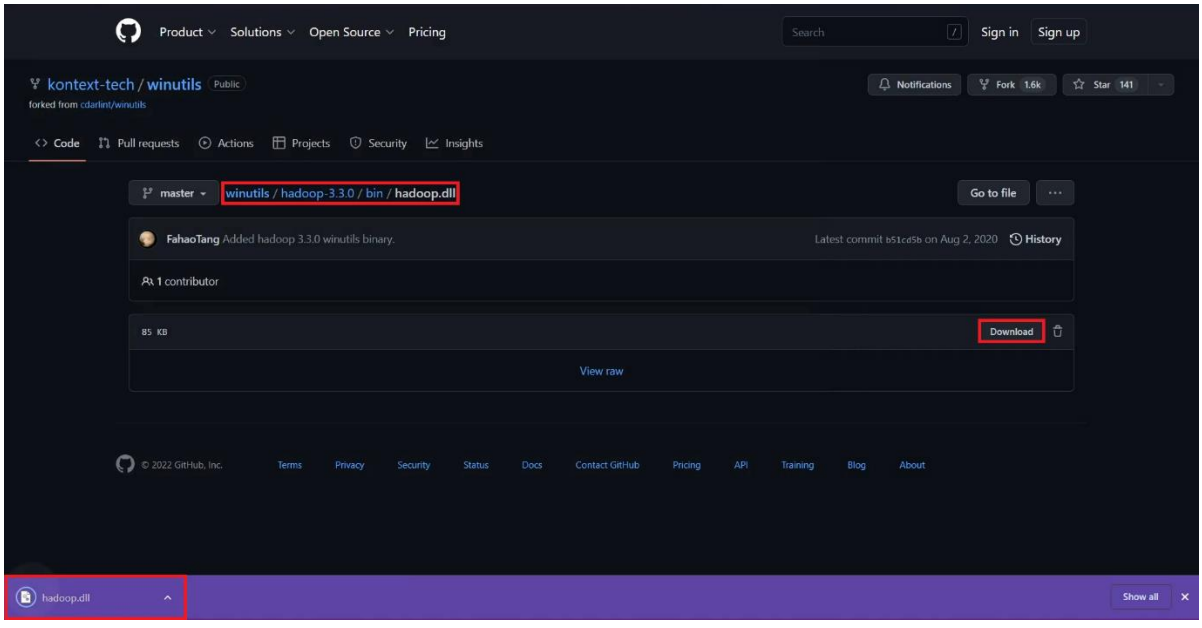




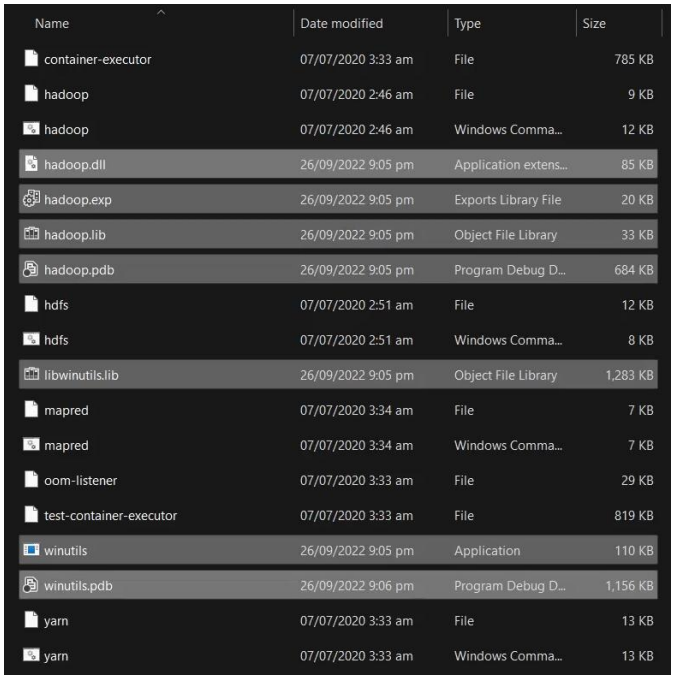
Afterwards, download Hadoop 3.3.0's binaries (winutils) by following this link and downloading all files inside the bin folder: https://github.com/kontext-tech/winutils/tree/master/hadoop-3.3.0/bin.



Download the contents of the bin folder by clicking on each file and selecting the "Download" button. Repeat this process for the other following files: *hadoop.exp*, *hadoop.lib*, *hadoop.pdb*, *libwinutils.lib*, *winutils.exe*, and *winutils.pdb*.



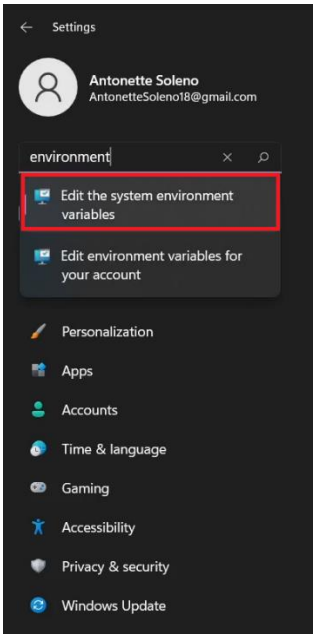
Once all files have been downloaded, copy these files into the 'hadoop-3.3.0\bin' directory, as indicated.

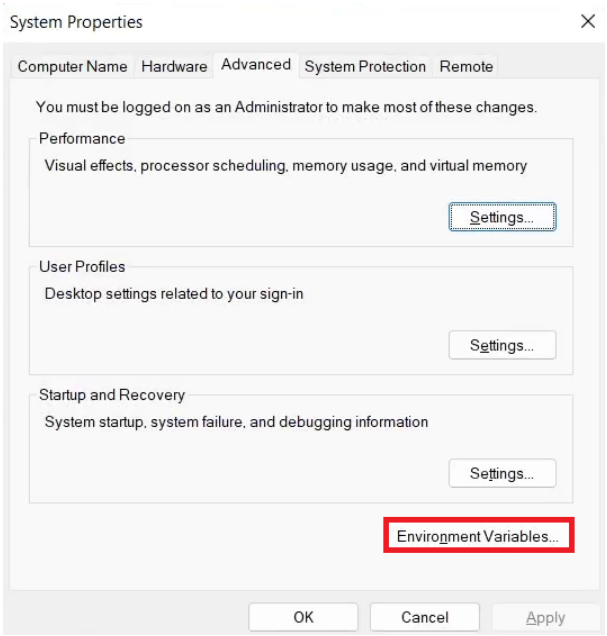


Name	Date modified	Type	Size
container-executor	07/07/2020 3:33 am	File	785 KB
hadoop	07/07/2020 2:46 am	File	9 KB
hadoop	07/07/2020 2:46 am	Windows Comma...	12 KB
hadoop.dll	26/09/2022 9:05 pm	Application extens...	85 KB
hadoop.exp	26/09/2022 9:05 pm	Exports Library File	20 KB
hadoop.lib	26/09/2022 9:05 pm	Object File Library	33 KB
hadoop.pdb	26/09/2022 9:05 pm	Program Debug D...	684 KB
hdfs	07/07/2020 2:51 am	File	12 KB
hdfs	07/07/2020 2:51 am	Windows Comma...	8 KB
libwinutils.lib	26/09/2022 9:05 pm	Object File Library	1,283 KB
mapred	07/07/2020 3:34 am	File	7 KB
mapred	07/07/2020 3:34 am	Windows Comma...	7 KB
oom-listener	07/07/2020 3:33 am	File	29 KB
test-container-executor	07/07/2020 3:33 am	File	819 KB
winutils	26/09/2022 9:05 pm	Application	110 KB
winutils.pdb	26/09/2022 9:06 pm	Program Debug D...	1,156 KB
yarn	07/07/2020 3:33 am	File	13 KB
yarn	07/07/2020 3:33 am	Windows Comma...	13 KB

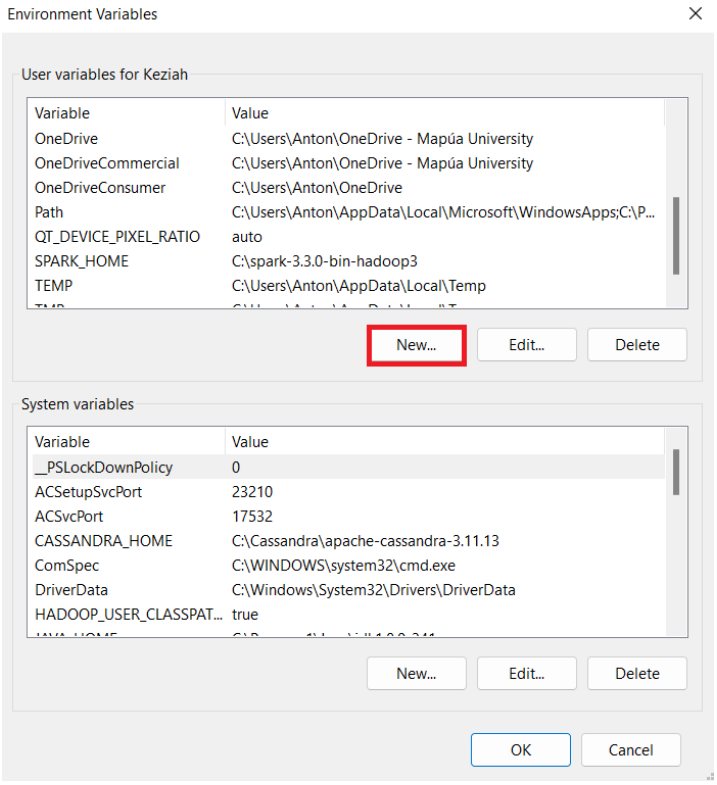
b. Set up environment variables for Java and Hadoop

After downloading Hadoop's package and binaries, the environment variables to define Hadoop and Java must now be configured. To do so, go to Settings and type "environment." When *Edit the system environment variables* appears, select it. The *System Properties* window should appear. Select *Environmental Variables…*.

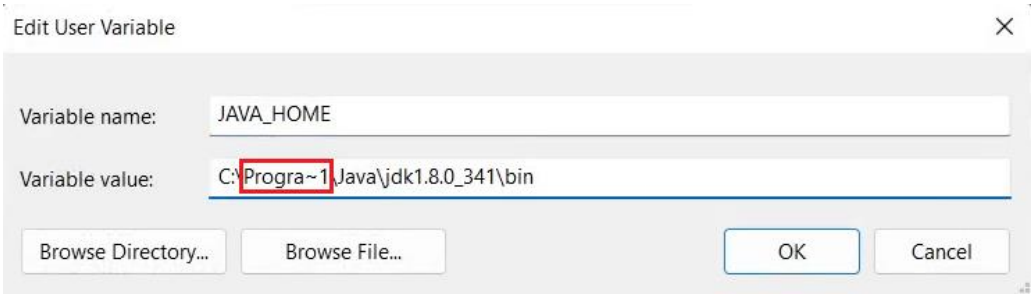


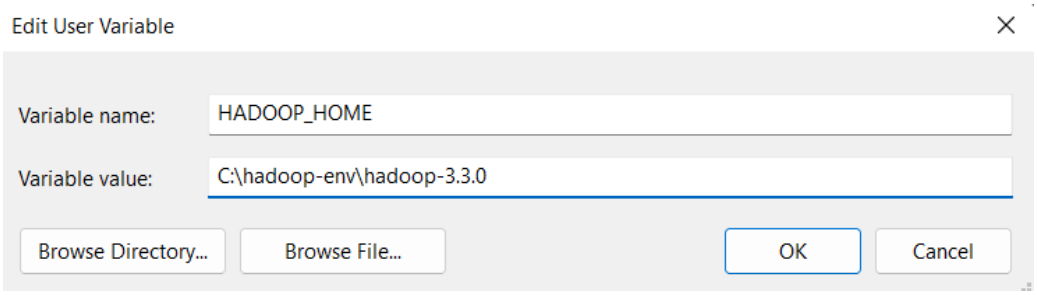


The *Environment Variables* window should appear after. Now, click "New…" under *User variables* and create two variables: JAVA_HOME and HADOOP_HOME. JAVA_HOME should contain the path for the JDK installation, and HADOOP_HOME the path for the Hadoop 3.3.0 installation. Click "OK" after each variable creation.

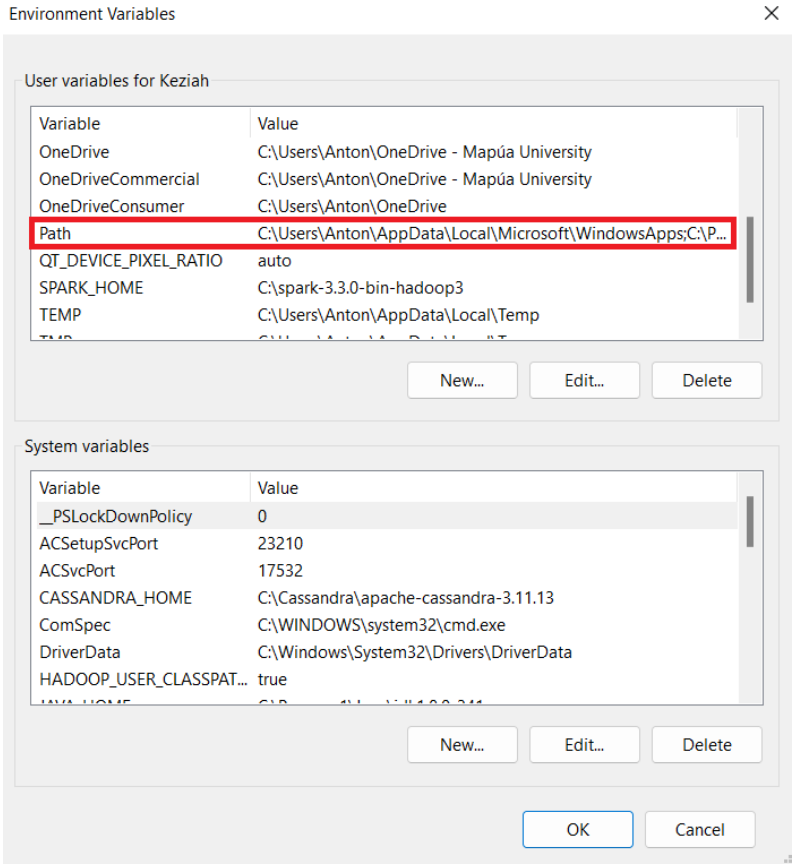


Note: The **Windows 8.3 path** was implemented while creating the path to JAVA_HOME to avoid future errors caused by white spaces in file naming.

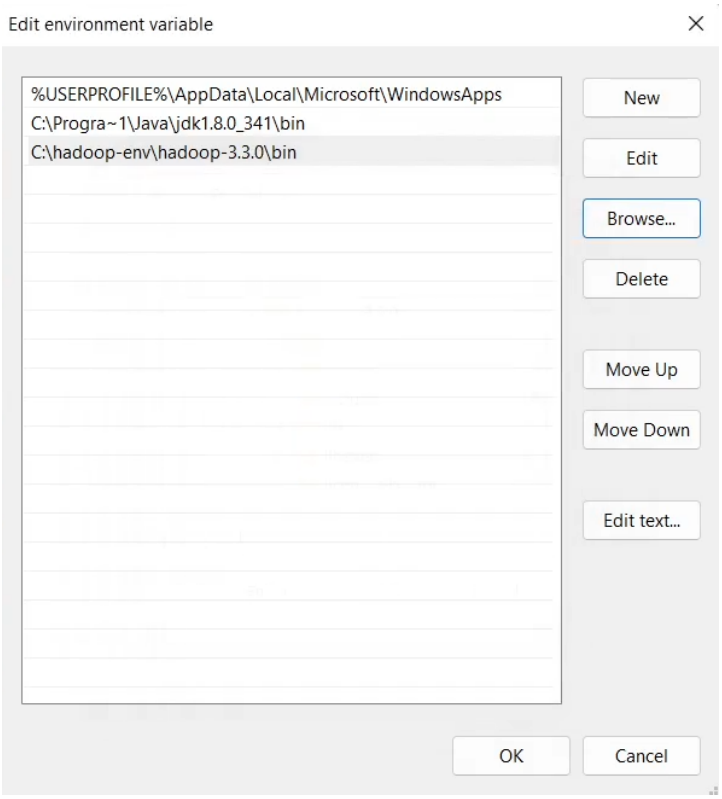




Under *User variables*, locate “Path” and double click.



Then, create two paths leading to each of the bin folders of JAVA_HOME and HADOOP_HOME under the *Edit environment variable* window, as indicated.



To verify that the configurations made were successful, run the following command on Windows Powershell to see the version of Java being used by Hadoop:

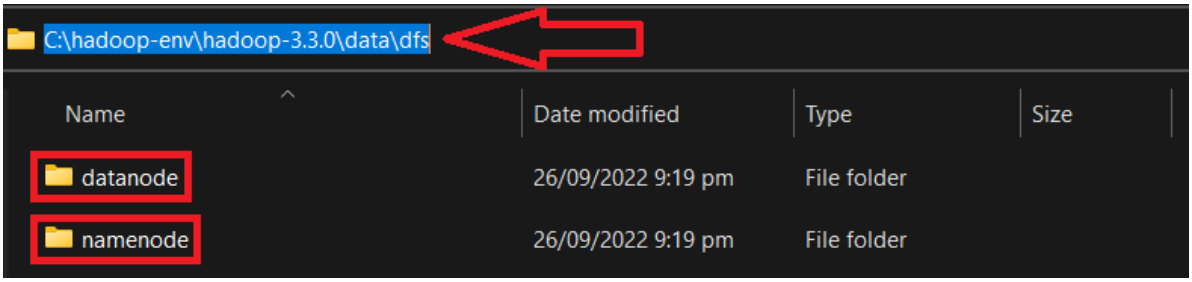
```
hadoop -version
```

The specific Java version for the SE Runtime Environment should appear, as indicated in the figure.

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\Anton> hadoop -version
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)
```

c. *Configure the Hadoop cluster*

Hadoop is built using a master-slave paradigm. Before altering the HDFS configuration file, a directory to store all master node (name node) data and another to store data (data node) must be created (Fadlallah, 2021). The following directories were created to store the master node into the *namenode* folder and data into the *datanode* folder.



1. Now that the directories have been created, the HDFS site file may now be configured by adding the following code into the <configuration></configuration> element located within its designated XML file. (*hdfs-site.xml*)

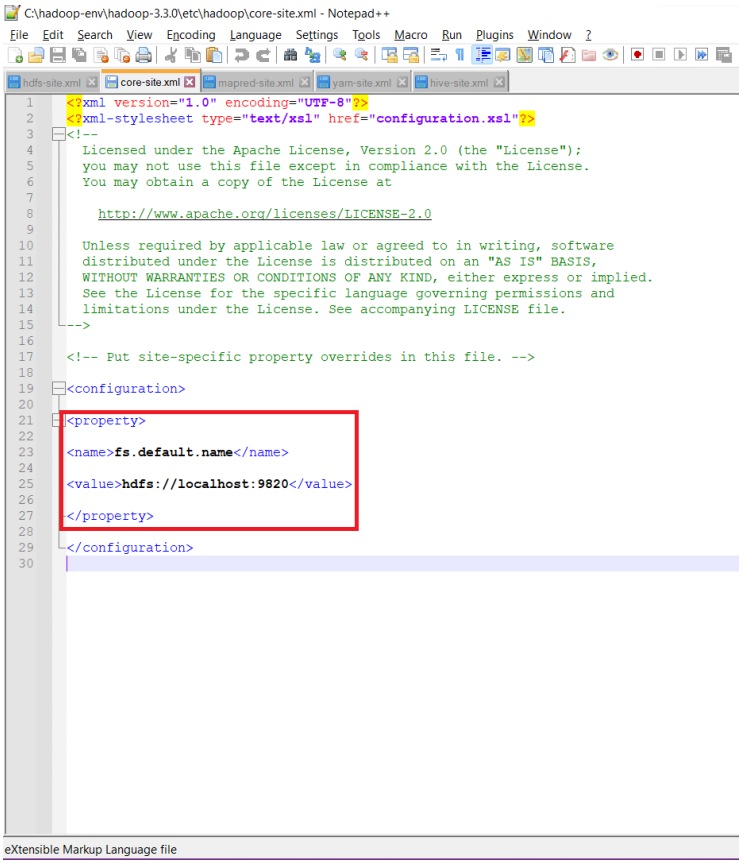
```
<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

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

<value>file:///C:/hadoop-env/hadoop-3.3.0/data/dfs/namenode</value>

</property>

<property>

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

<value>file:///C:/hadoop-env/hadoop-3.3.0/data/dfs/datanode</value>

</property>
```

The figure below depicts the modified version of *hdfs-site.xml*.

2. The next file to be modified is the Core site file. Paste the following code into the <configuration></configuration> element of *core-site.xml*.

```
<property>

<name>fs.default.name</name>

<value>hdfs://localhost:9820</value>

</property>
```

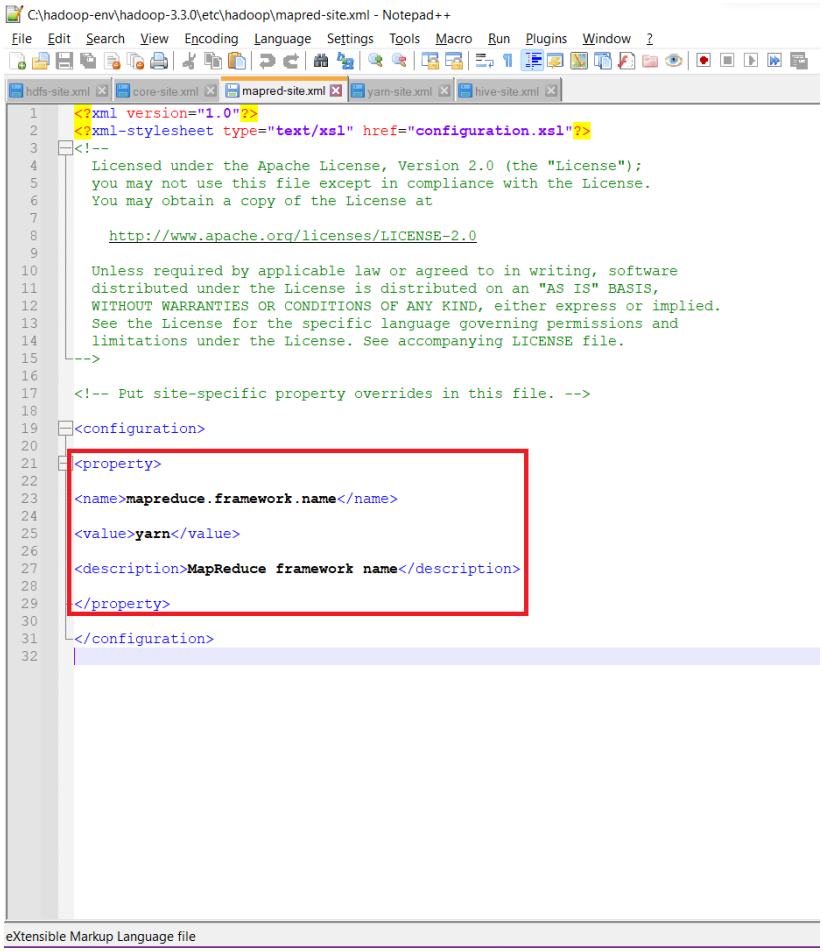
The figure below depicts the modified version of *core-site.xml*.



3. Afterwards, paste the following code into the <configuration></configuration> element of *mapred-site.xml*.

```
<property>

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

<value>yarn</value>

<description>MapReduce framework name</description>

</property>
```

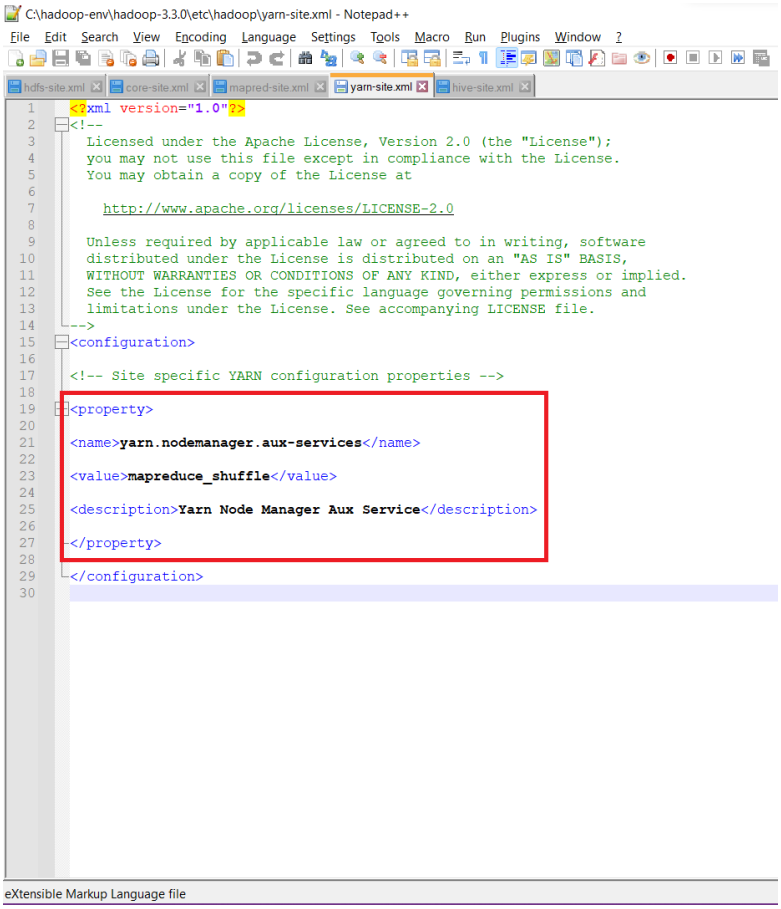
The figure below depicts the modified version of *mapred-site.xml*.



4. Lastly, paste the following code into the <configuration></configuration> element of *yarn-site.xml*.

```
<property>

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

<value>mapreduce_shuffle</value>

<description>Yarn Node Manager Aux Service</description>

</property>
```

The figure below depicts the modified version of *yarn-site.xml*.

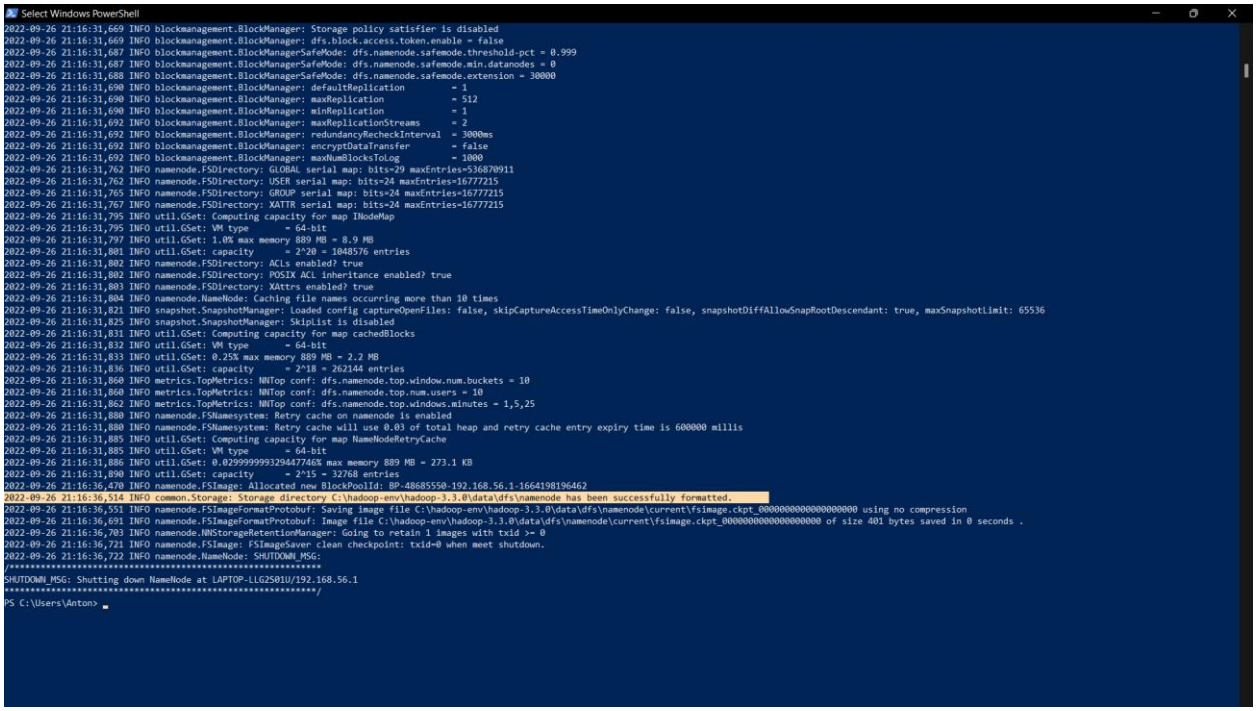


d. Format namenode

Once all configurations have finished, the name node may now be formatted by running the following command on Windows Powershell:

```
hdfs namenode -format
```

If all configurations were made correctly, the command should execute successfully by the declaration that the namenode has been successfully formatted.

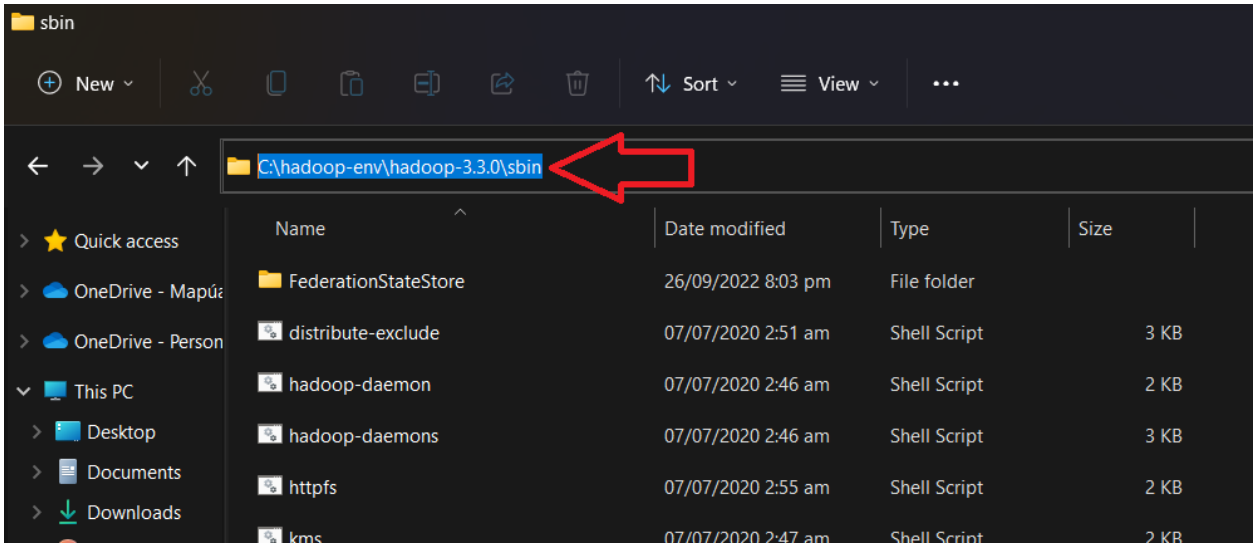


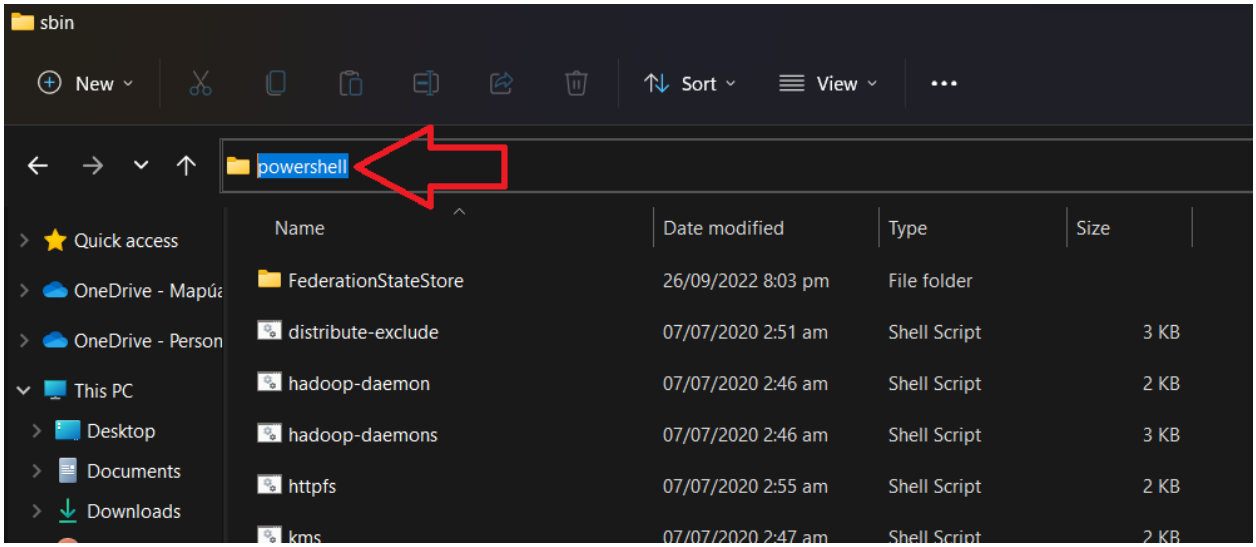
e. Start Hadoop services

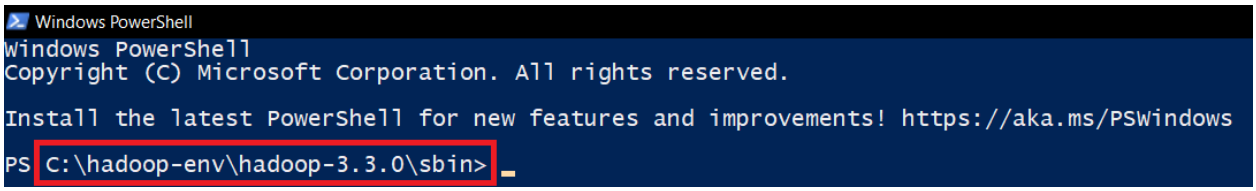
To start the Hadoop services, *open Windows PowerShell as Administrator on the directory C:\hadoop-env\hadoop-3.3.0\sbin and run the following command to open name node and data node in two separate command prompt terminals:

```
.\start-dfs.cmd
```

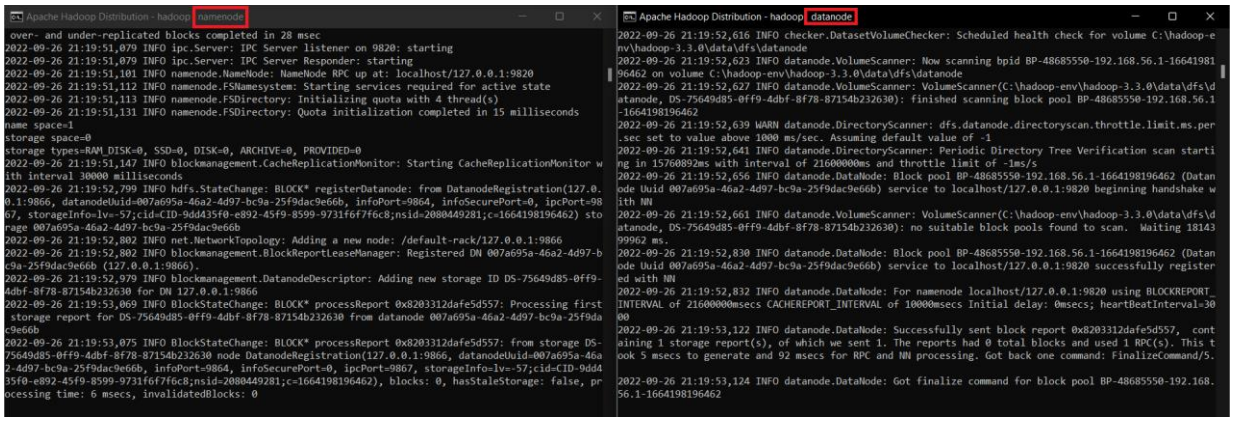
**To open PowerShell as Administrator on the specified directory, locate the sbin folder on the hadoop-3.3.0 folder and type* powershell *on the address bar. The PowerShell window running on the sbin folder should appear.*







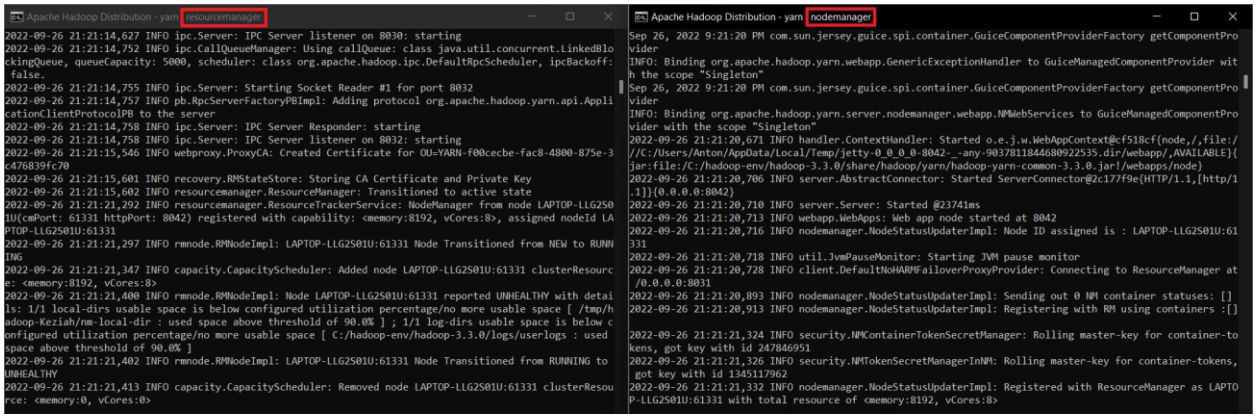
After running the start-dfs command, the two terminals, namenode and datenode, should appear.



The next Hadoop service to run is the Yarn service that can be begun by running the following command on Windows PowerShell:

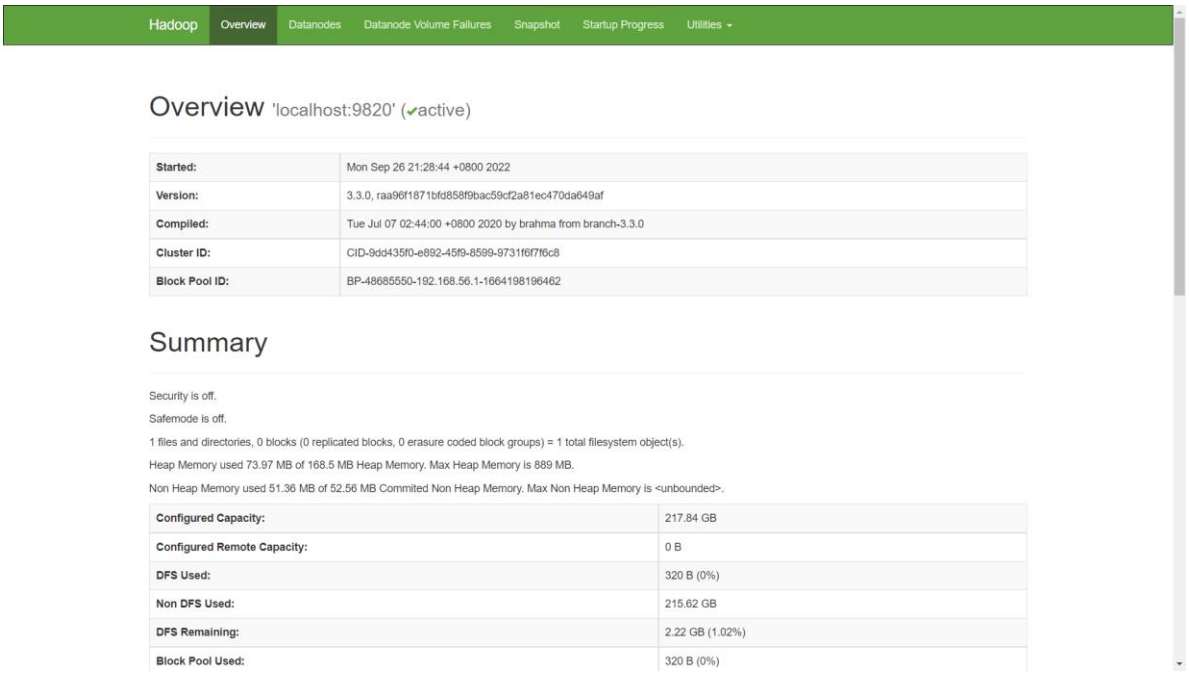
```
./start-yarn.cmd
```

After running the start-yarn command, the two terminals, resourcemanager and nodemanager, should appear.

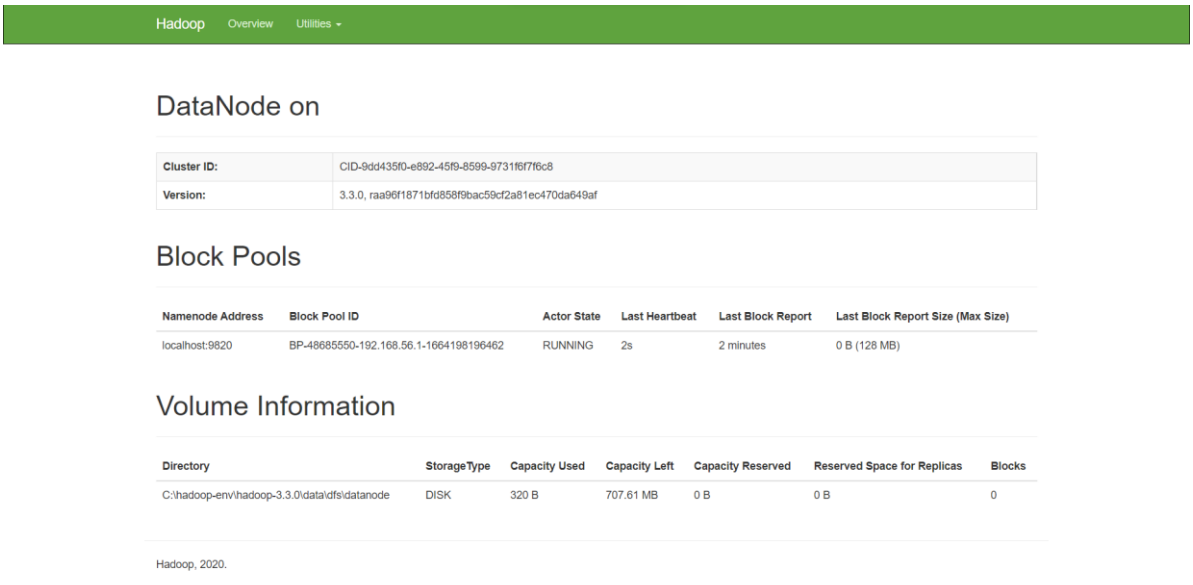


f. Hadoop Web UI

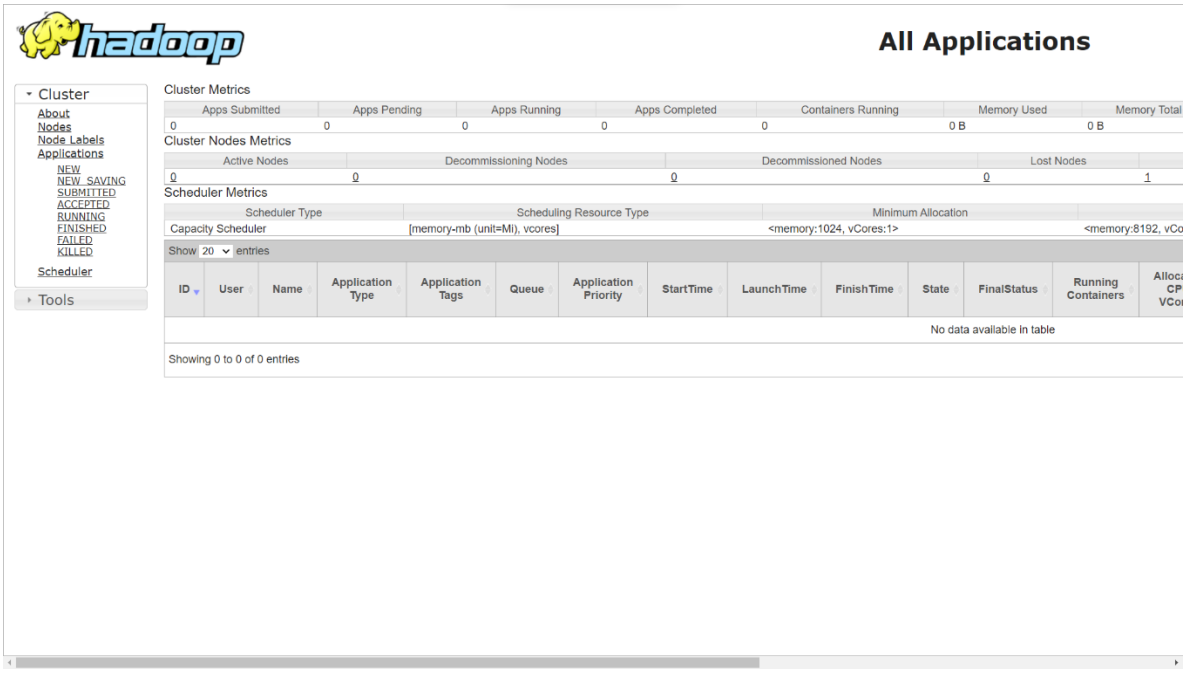
The following are the web user interfaces that could be used to access the name node, date node, and yarn services of Hadoop.



Name Node (web page)



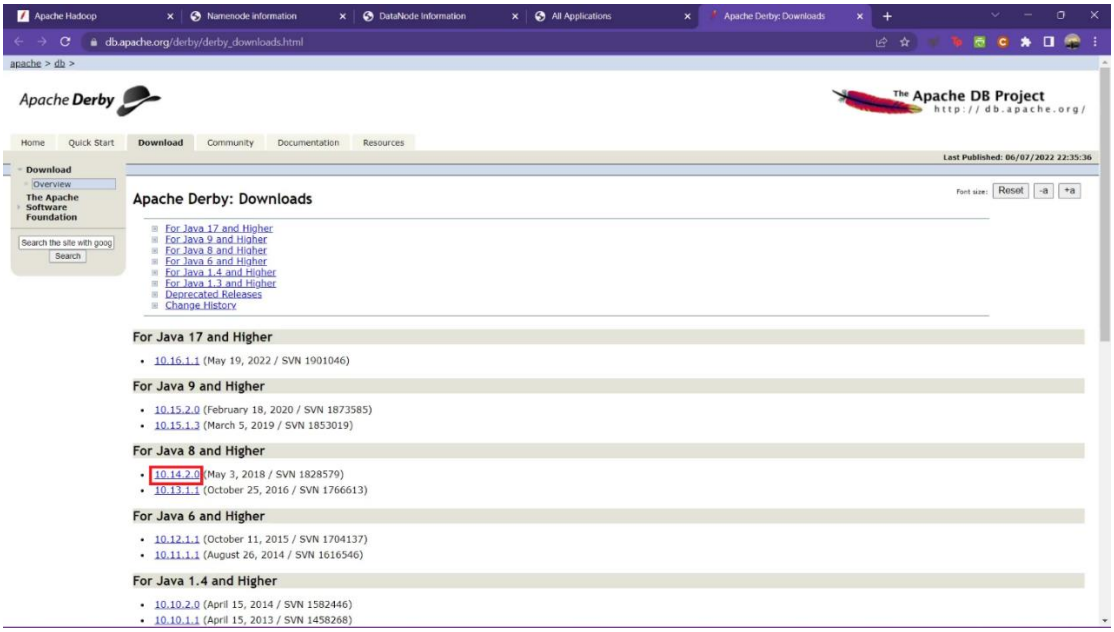
Data Node (web page)

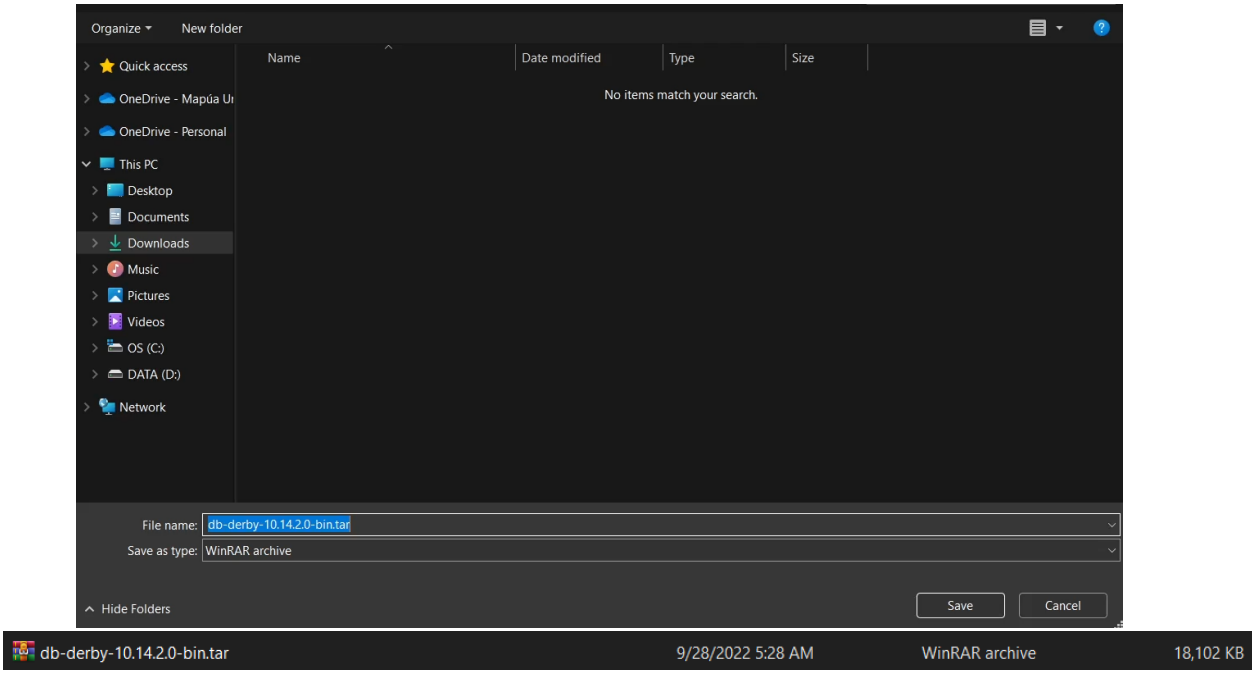


Yarn Service (web page)

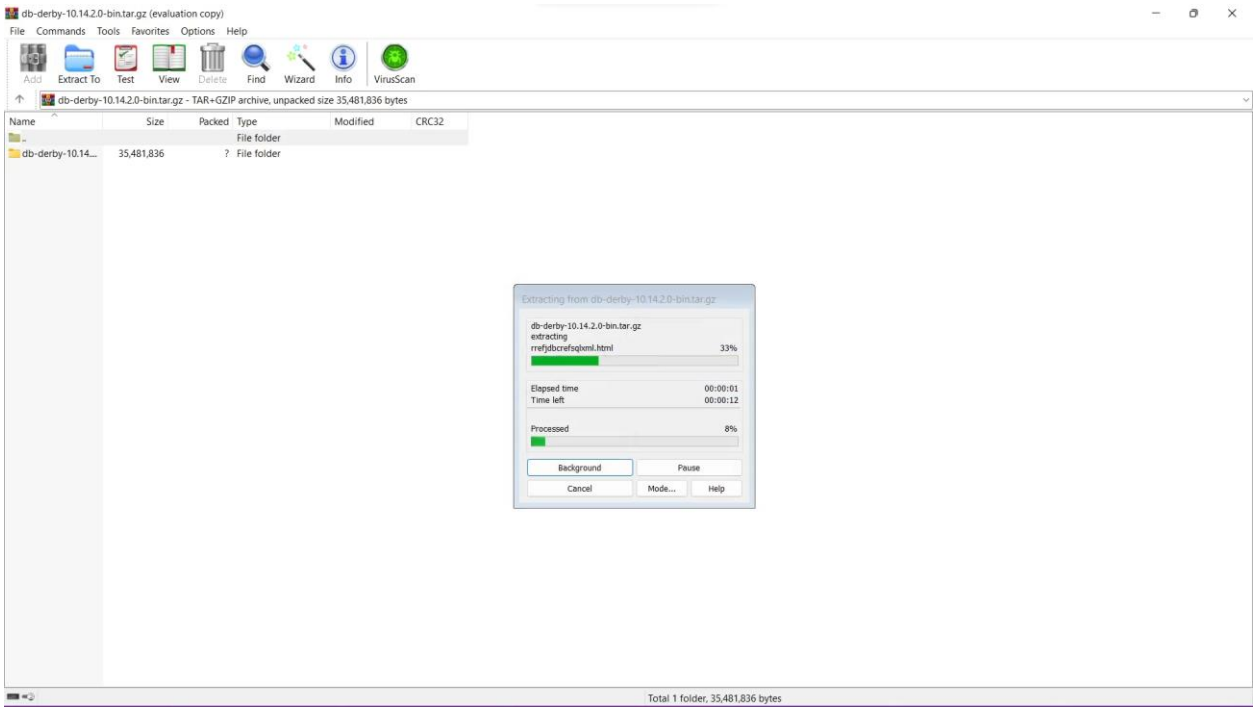
III. Apache Derby Download

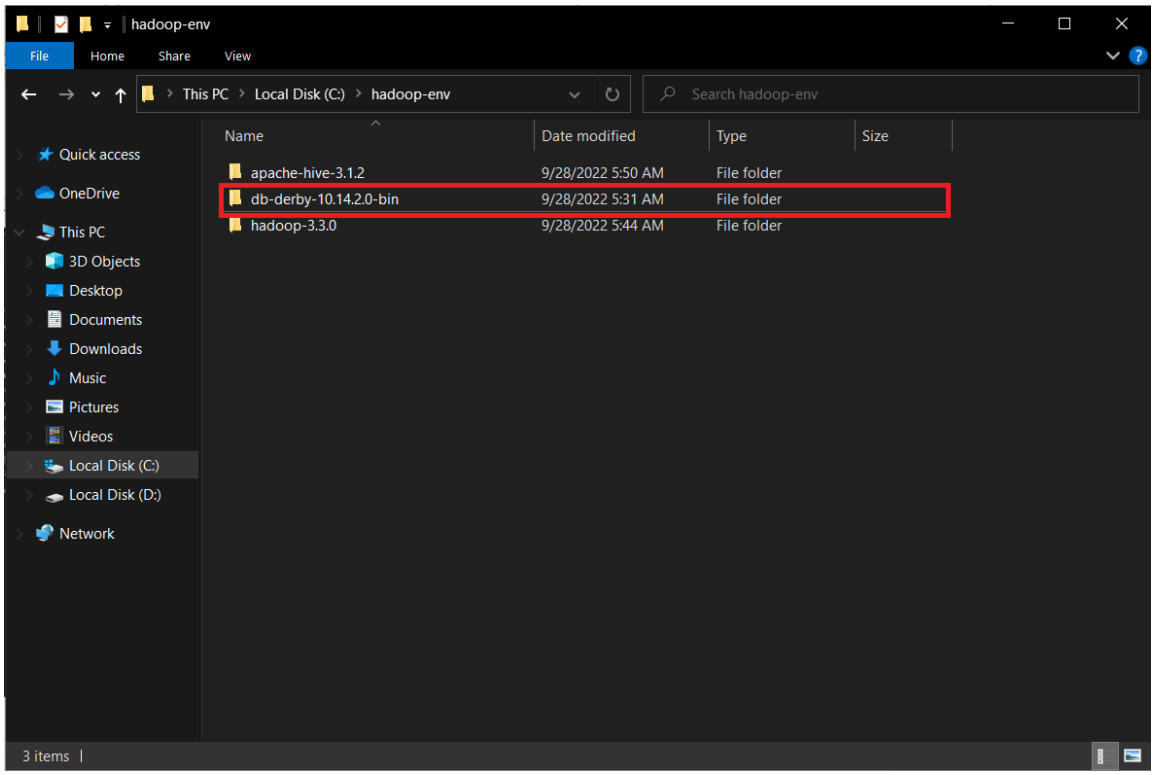
Another prerequisite of Apache Hive is Apache Derby, a relational database used to create Apache Hive's Metastore, where all Hive's metadata will be stored. Since Java 8 was the version previously installed for Hadoop, Apache Derby version 10.14.2.0 must be installed. This version can be directly downloaded from this link: https://downloads.apache.org//db/derby/db-derby-10.14.2.0/db-derby-10.14.2.0-bin.tar.gz. Alternatively, the download link may be accessed by following this link and selecting 10.4.2.0: https://db.apache.org/derby/derby_downloads.html.



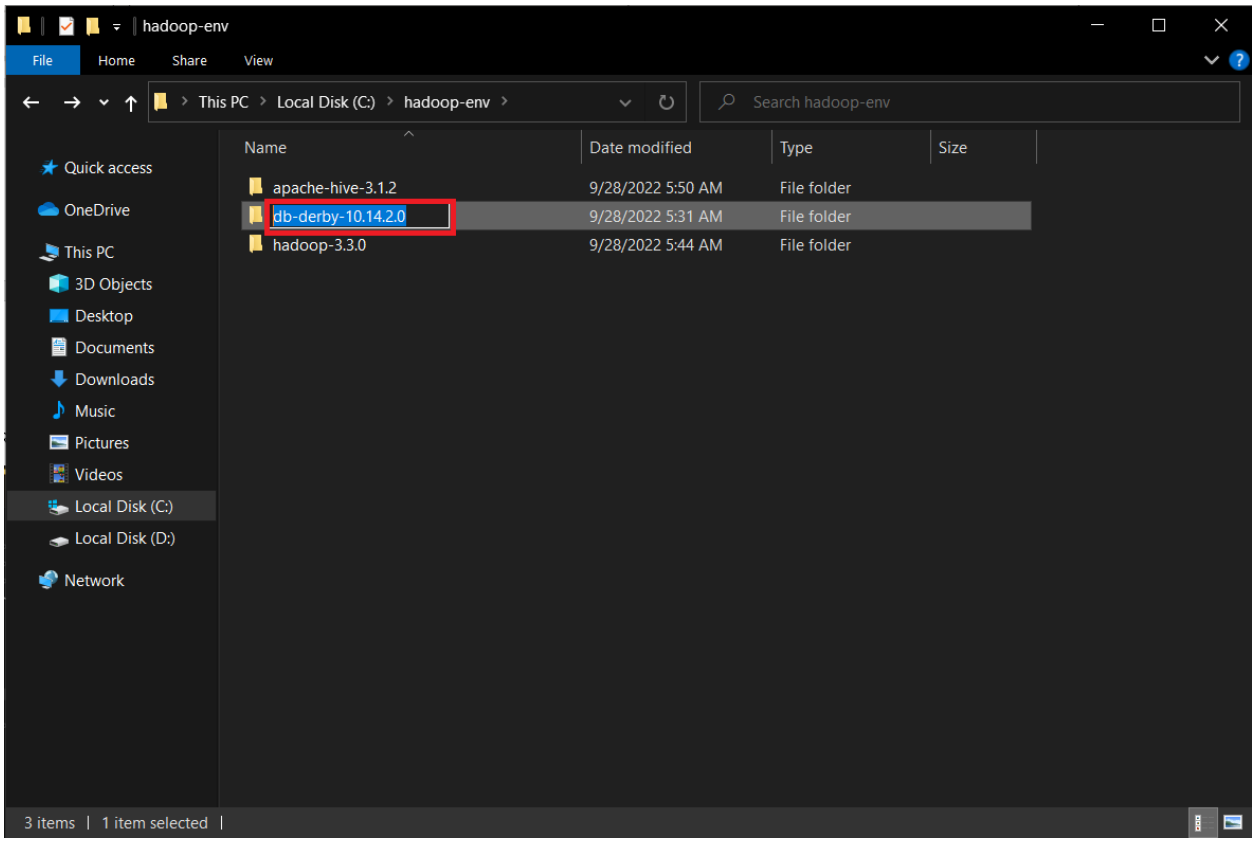


Locate the file in the downloads folder after it has downloaded. We must extract Derby using any zipping tool such as WinRAR Archiver. We extract the *db-derby-10.14.2.0-bin* file folder to the desired directory. Previously, we installed Hadoop within the "C:\hadoop-env" directory. Now, we will extract Derby into the same *hadoop-env* directory.

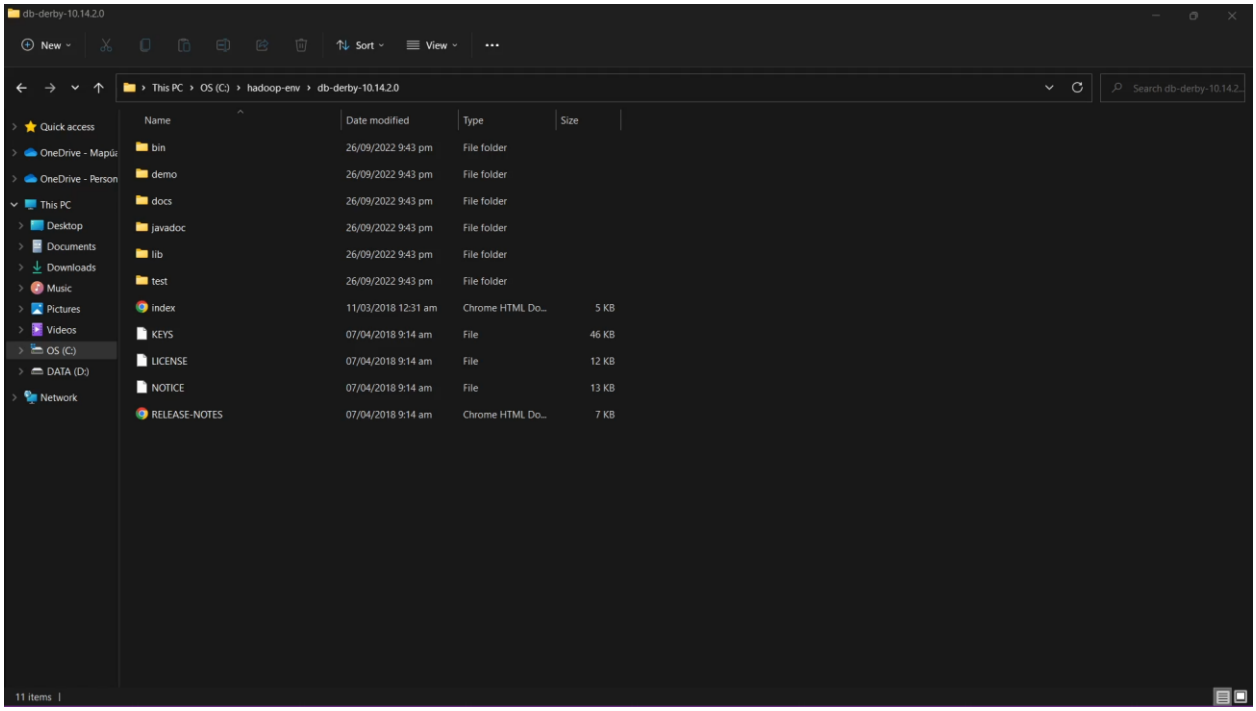




Rename this folder as *db-derby-10.14.2.0*.

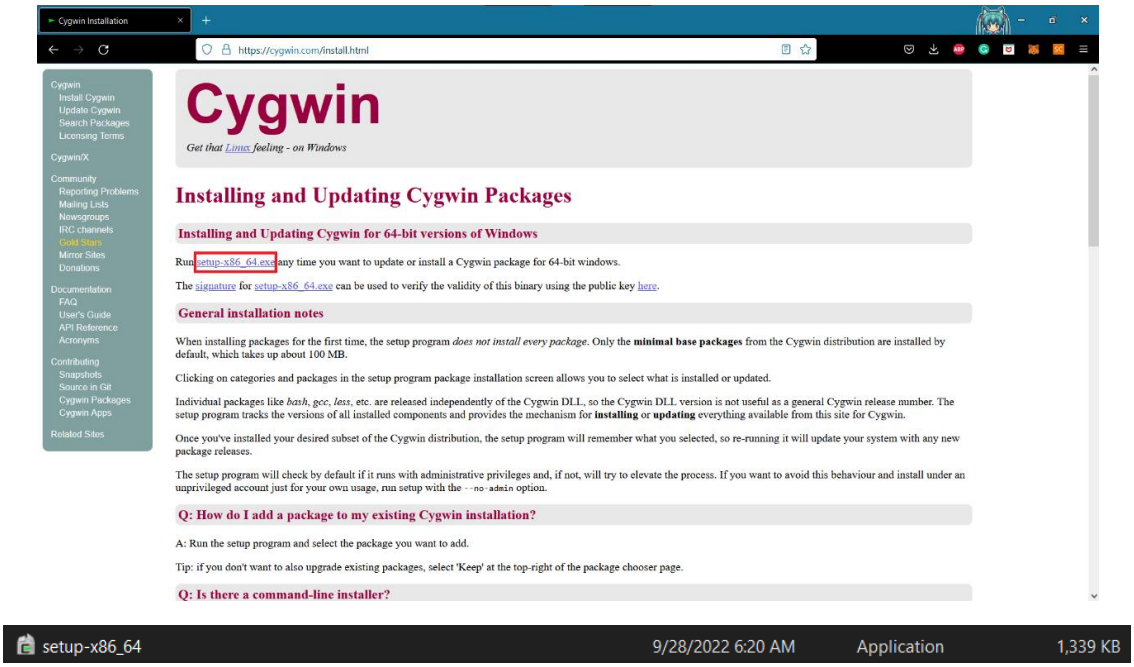


Apache Derby is now ready for Apache Hive.

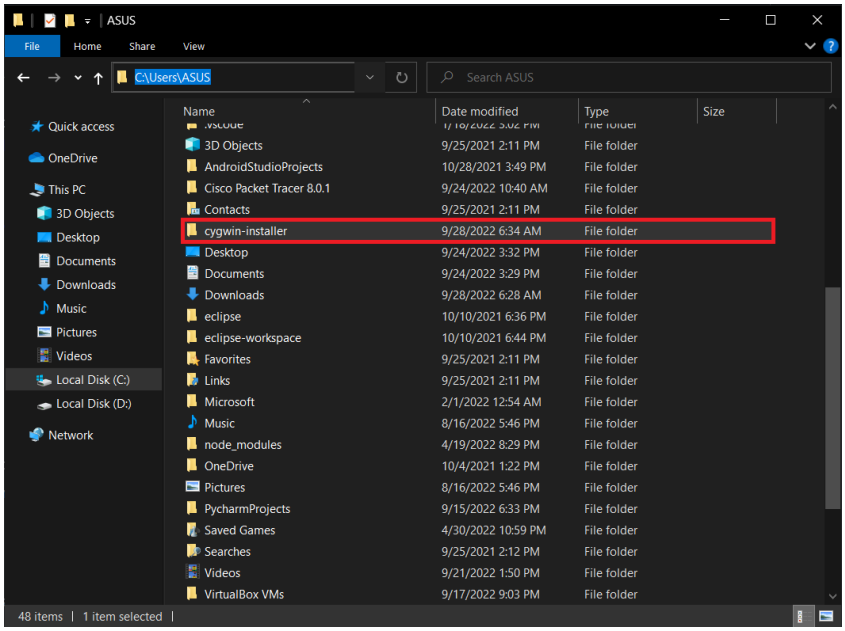


IV. Cygwin Installation

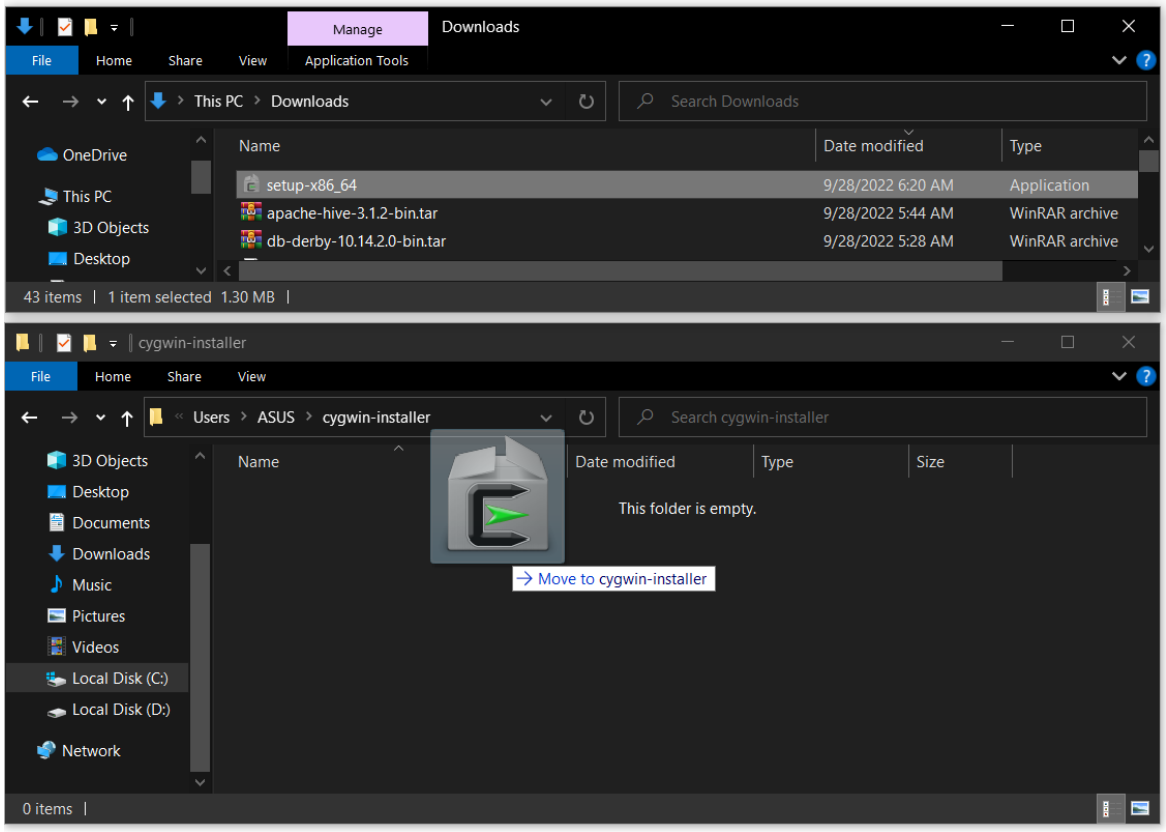
Unfortunately, there are some Hive tools that are compatible with Windows so we will need the Cygwin tool to run Linux commands. Head over to https://cygwin.com/install.html and click on *setup-x86_64.exe* to download the latest 64-bit version.



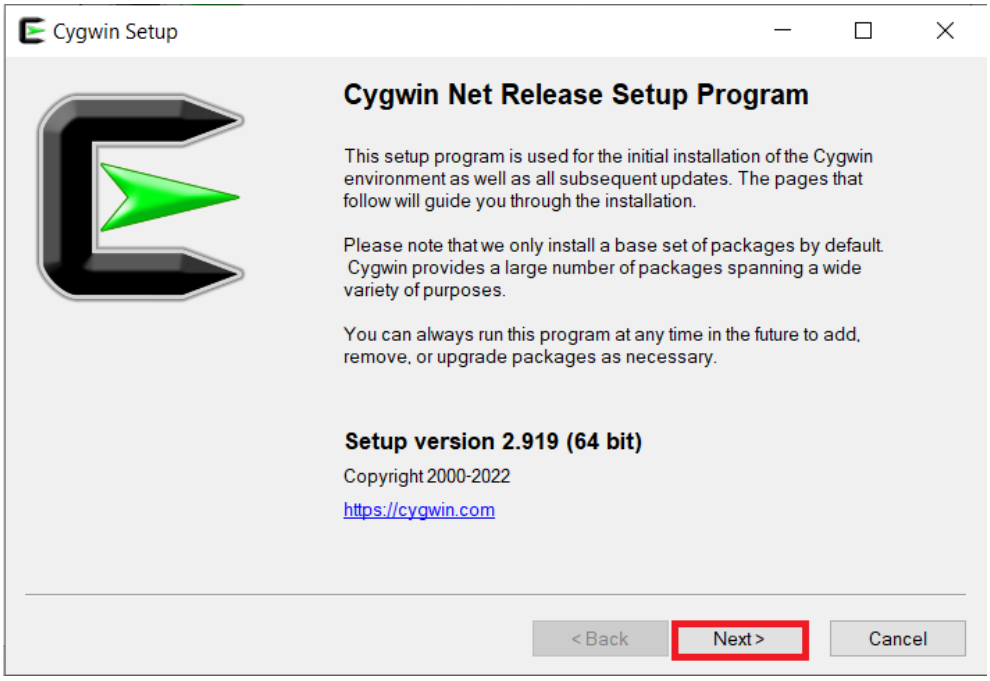
After downloading, open 'File Explorer' and go to your home directory (c:\Users\<username>) to create a new folder called 'cygwin-installer'.



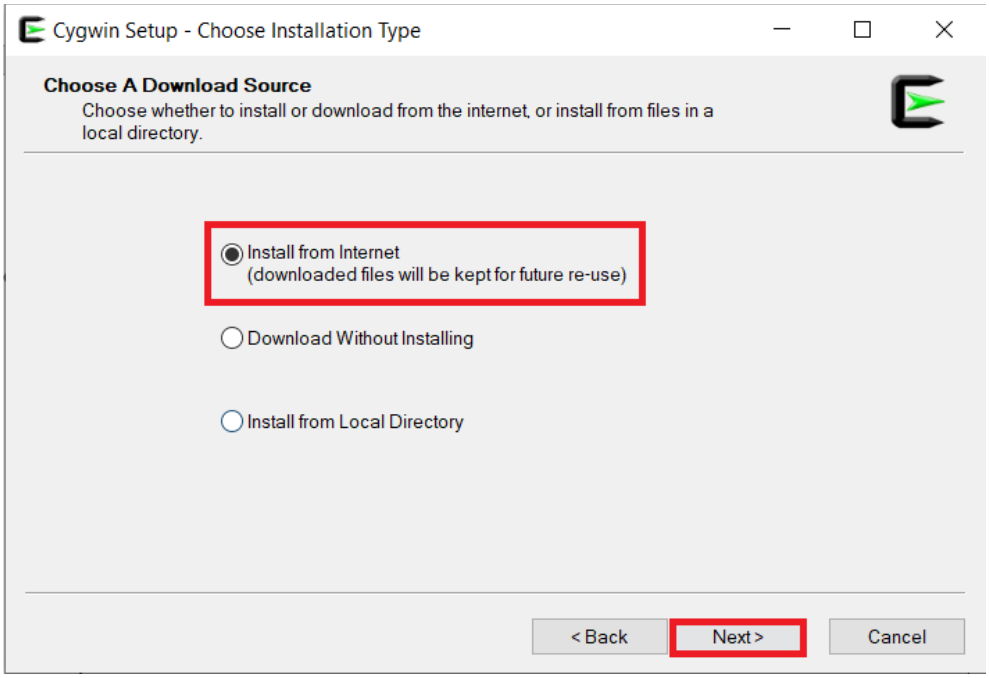
Next, move the Cygwin installer executable to the ‘cygwin-installer’ folder.



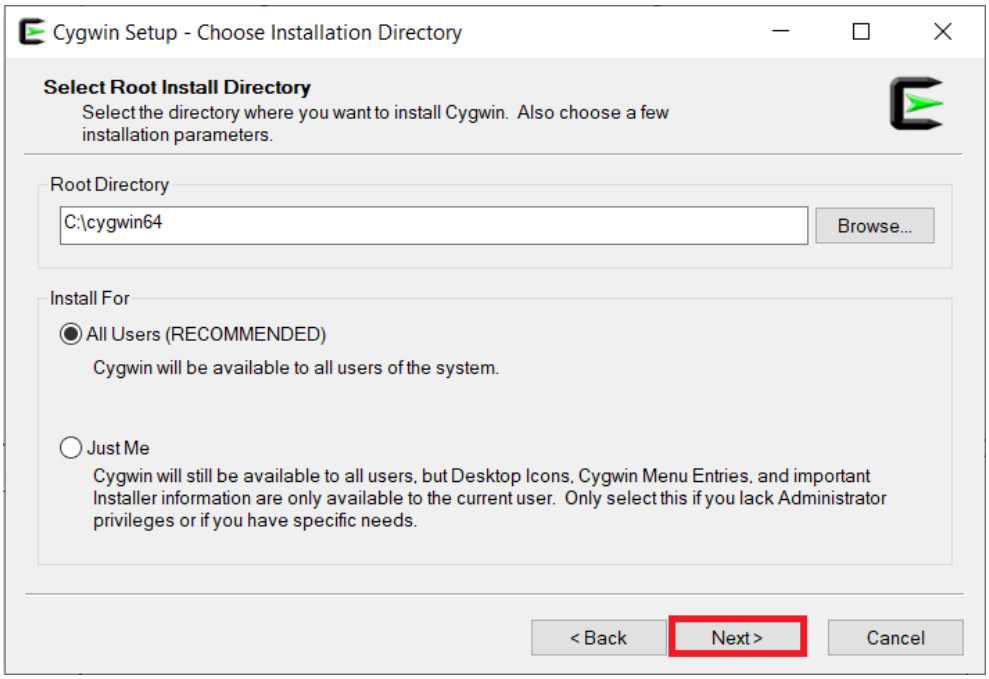
Inside the ‘cygwin-installer’ folder, we now run the cygwin installer executable. Once the setup program has started, click ‘Next’ to continue.



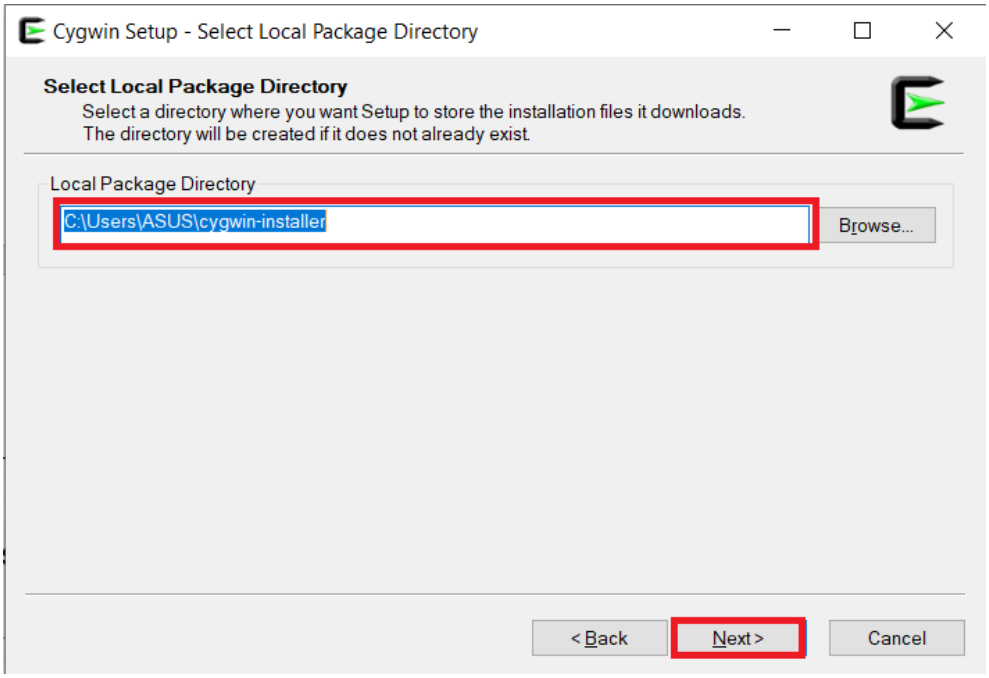
In the next section ‘Choose Installation Type’, we select Install from Internet. Click ‘Next’ to continue.



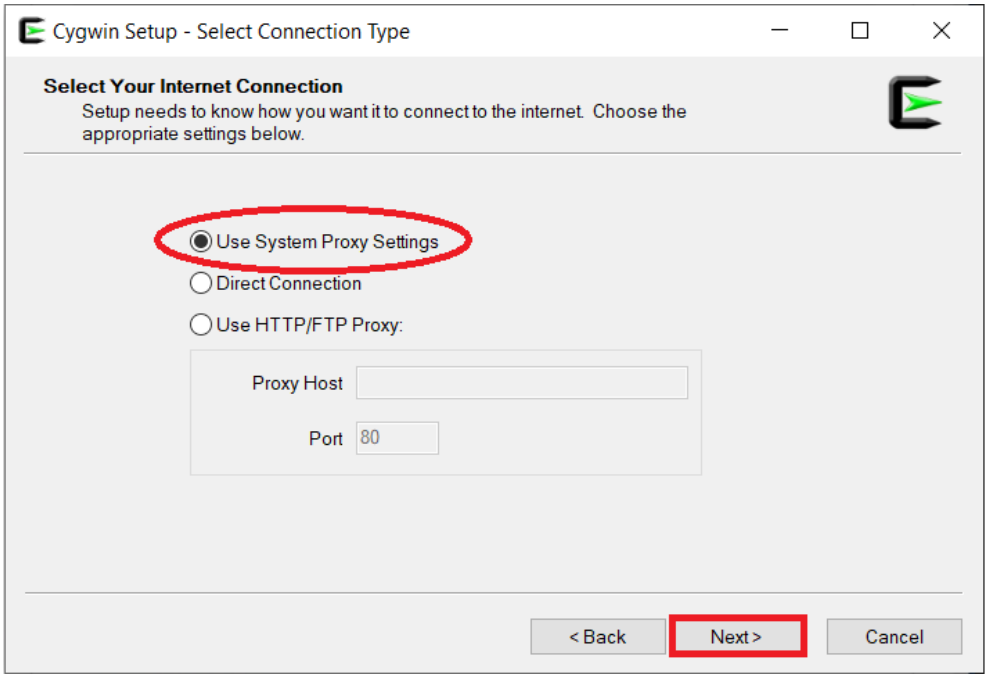
For the ‘Choose Installation Directory’ section, you can select a directory to where you want to install Cygwin. In this case, we continued with the default root directory ‘C:\cygwin64’. Next, for the ‘Install For’, keep the recommended option, ‘All Users’, selected unless you only want it to be installed to the current user. Click ‘Next’ to continue.



For this section, make sure that the ‘cygwin-installer’ folder is selected. Click ‘Next’ to continue with the setup.

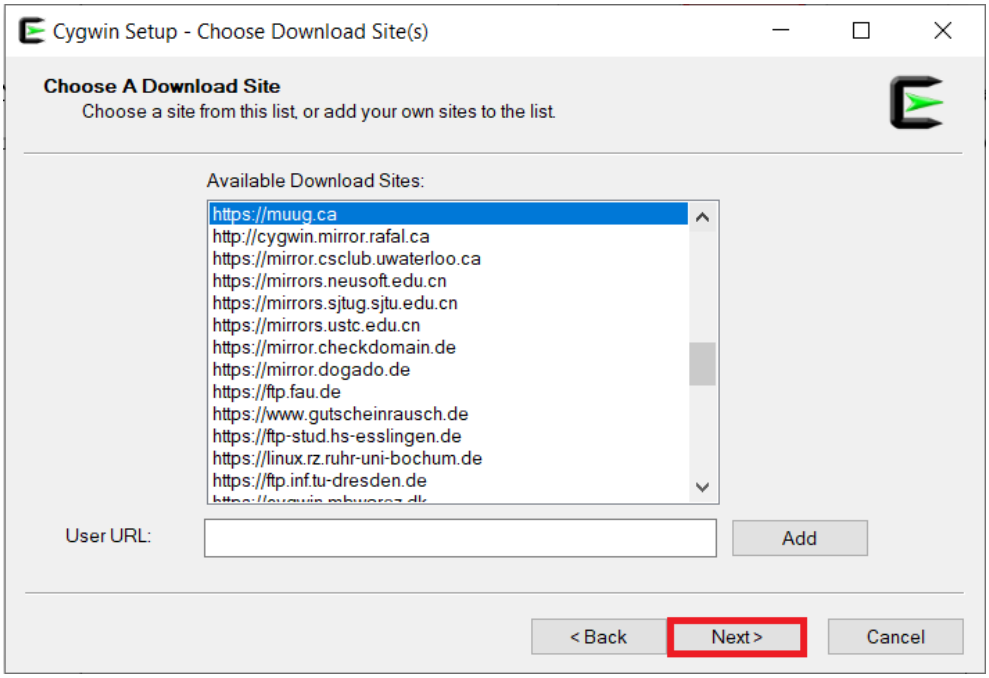


Now for the ‘Select Connection Type’, continue with the default setting ‘Use System Proxy Settings’ unless you want something to be specified here. Otherwise, click ‘Next’ to continue.

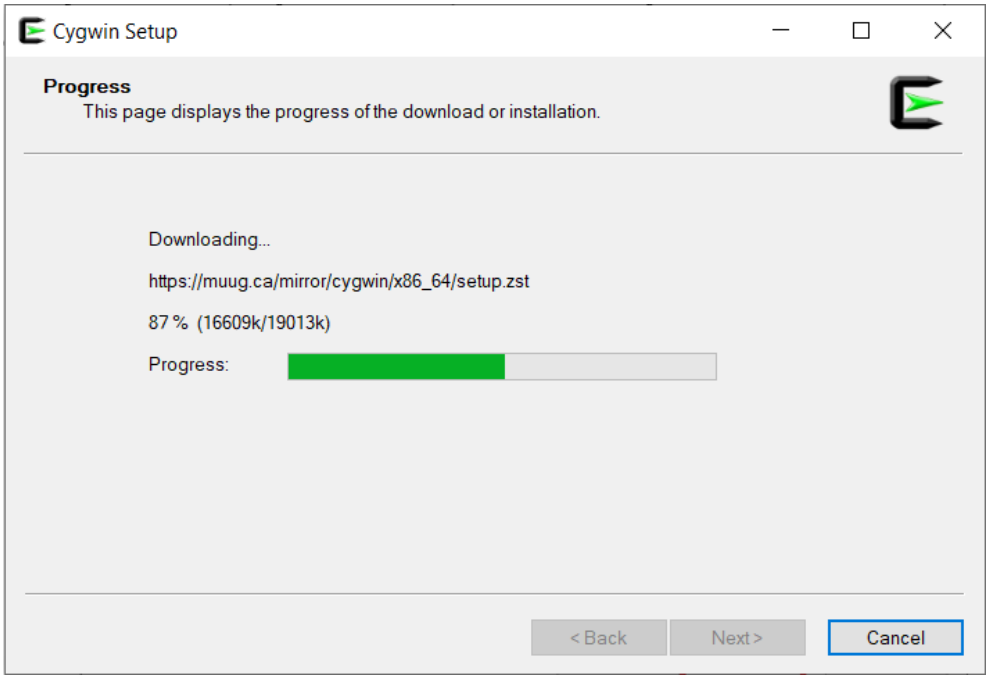


‘Choose Download Site(s)’ section will prompt you to choose a download site from a list from which you will download the Cygwin packages. It is better to select a mirror that is geographically

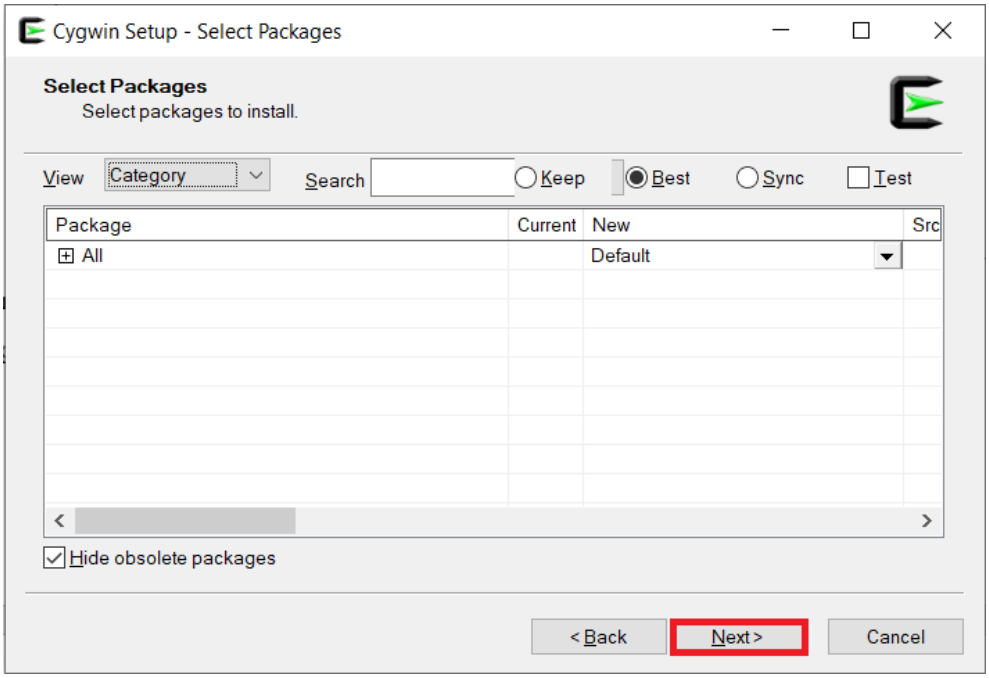
closer to you. However, if you are unsure, then choose any mirror. In this instance, we selected ‘https://muug.ca’. Click ‘Next’ to continue.



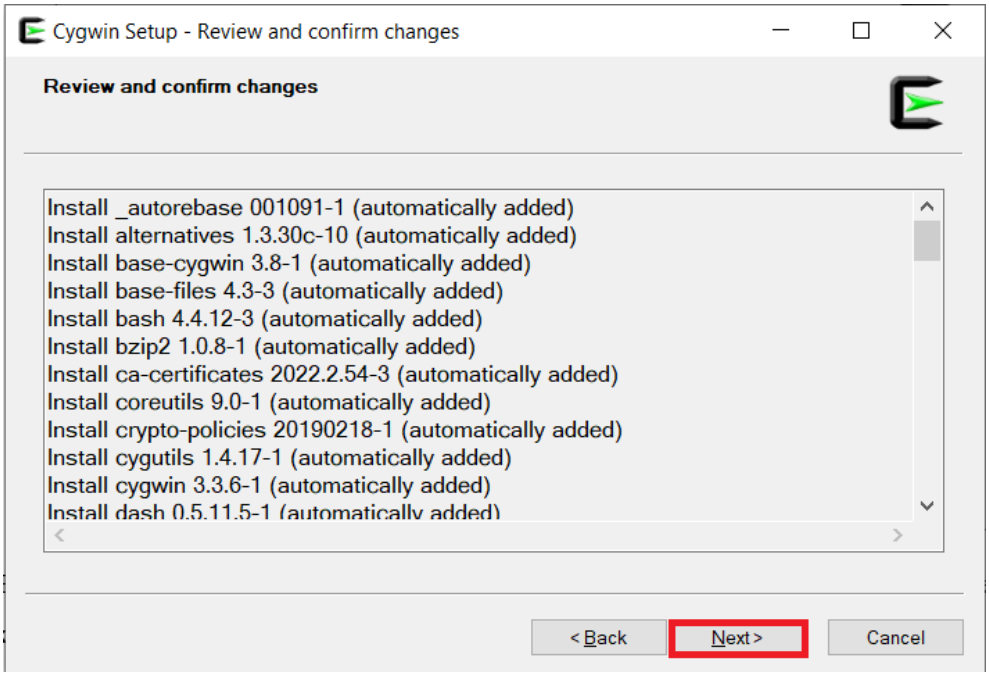
Now, we wait a few seconds for the download to finish.



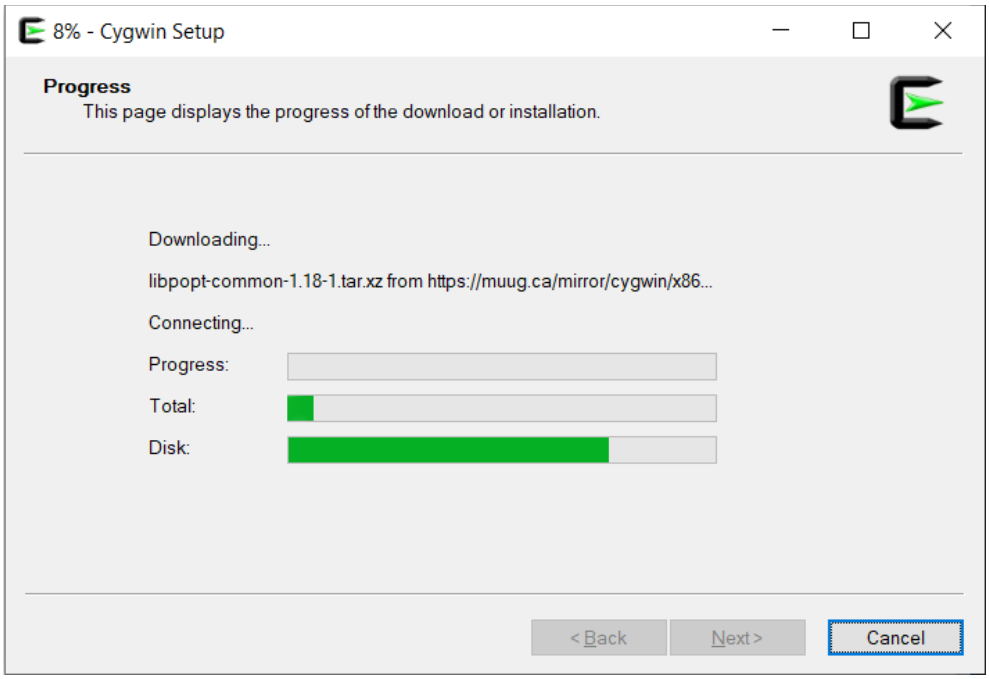
After it has downloaded, we are prompted to the ‘Select Packages’ section. For now, we will not select any packages, so click ‘Next’.



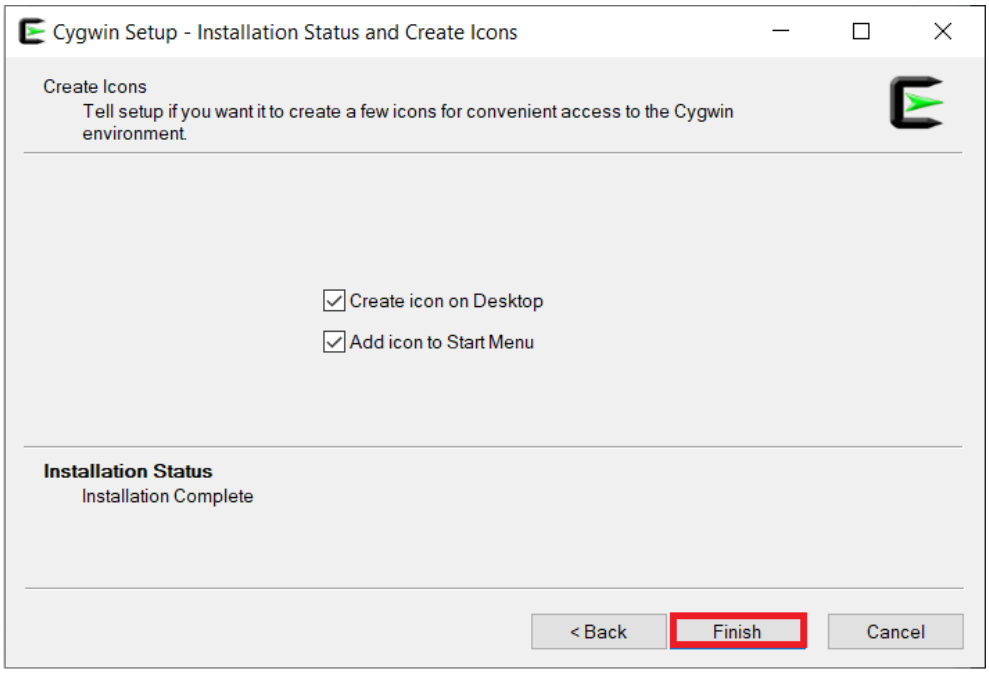
Before the final installation, we are prompted to review and confirm changes. Once done, click on ‘Next’ to start installing.



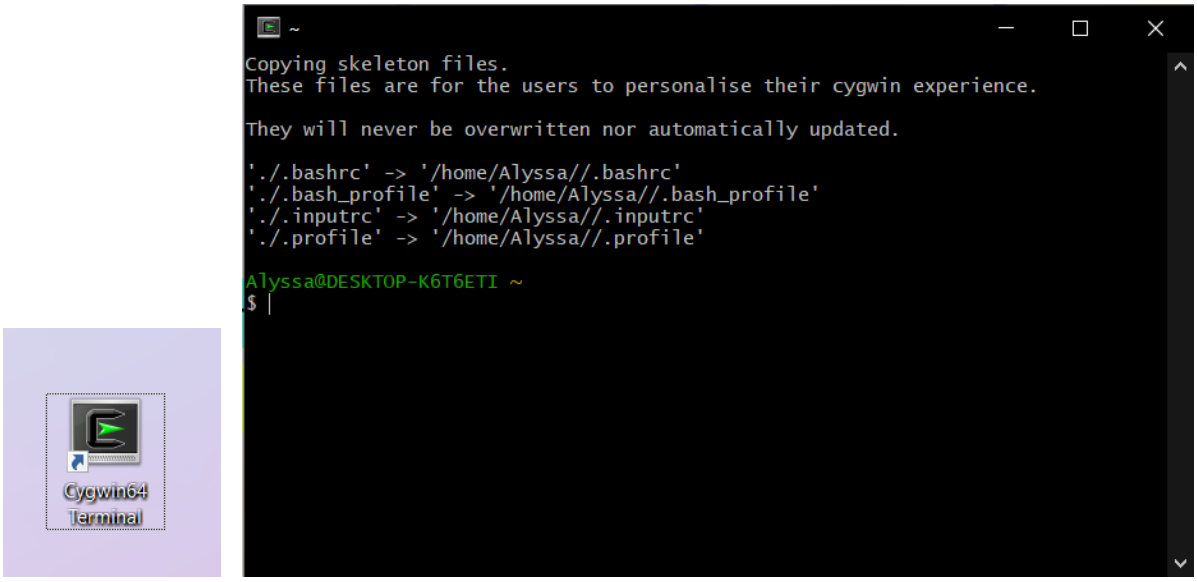
We now wait a few minutes for the Cygwin installation to finish.



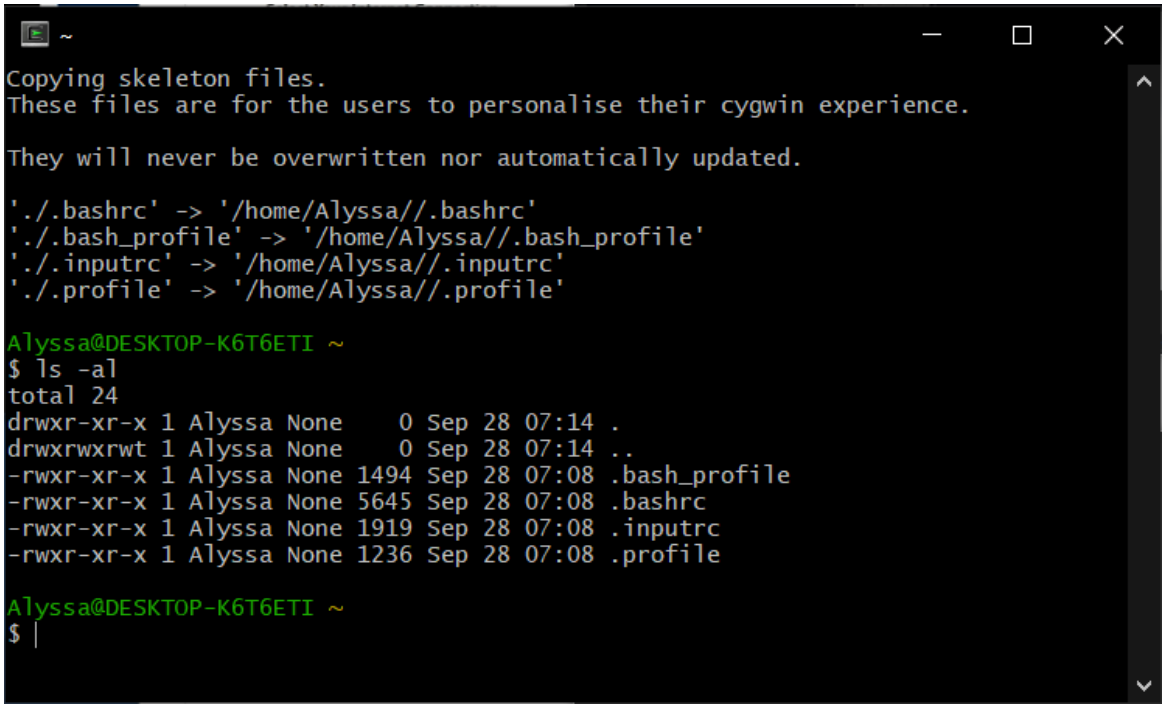
Once installation is finished, you may select ‘Create icon on Desktop’ and ‘Add icon to Start Menu’ if you want to. Click on ‘Finish’ to end the Cygwin setup.



To confirm that Cygwin is installed, you can double click on the ‘Cygwin Terminal’ icon on your desktop to run it. The Cygwin terminal should have launched.



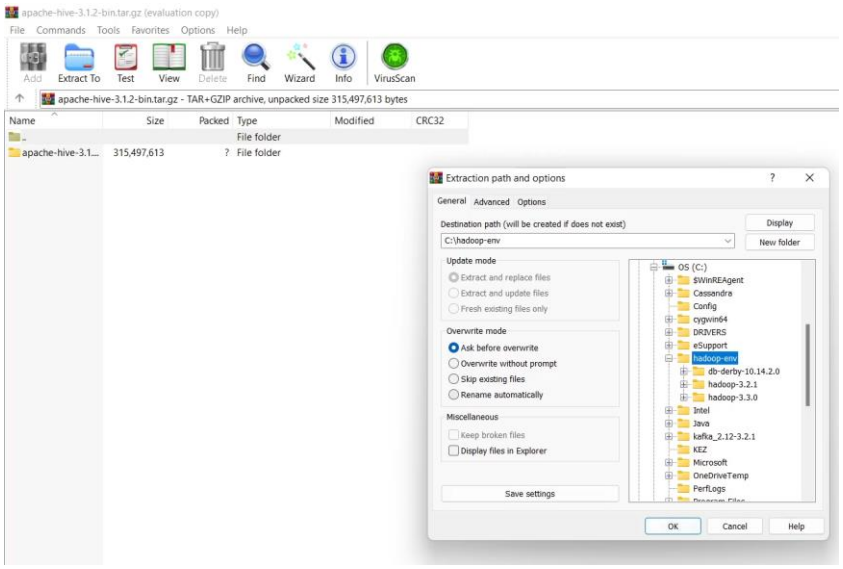
We input a command ‘ls -al’, which shows the content of that directory, to confirm the Cygwin is indeed running. Cygwin is now ready for Apache Hive.

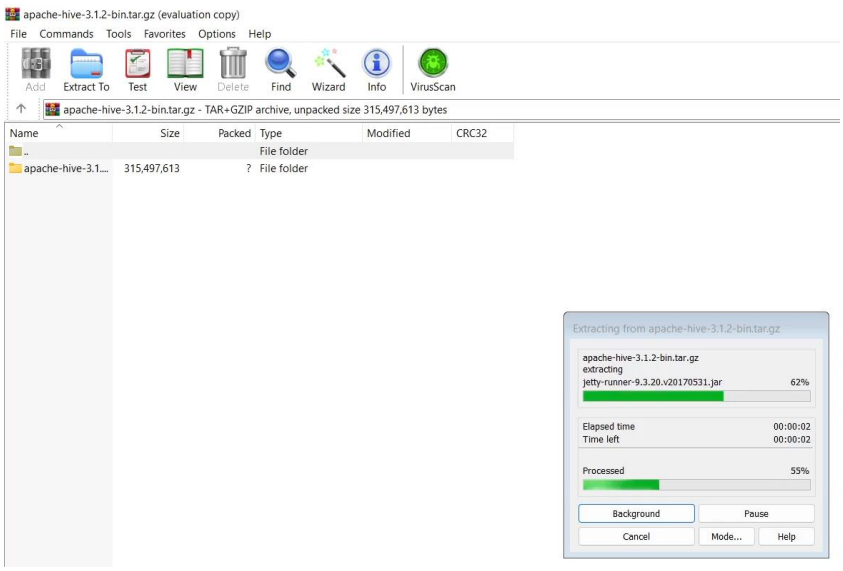


V. Hive Installation

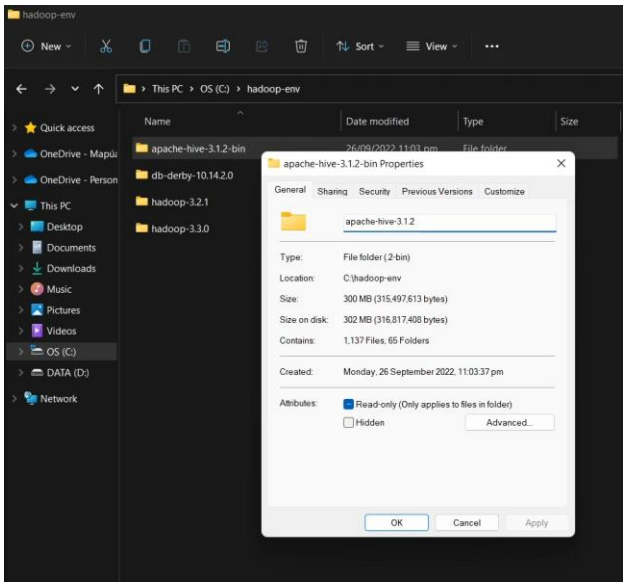
To download the Apache Hive binaries, click the following link and select *apache-hive-3.1.2-bin.tar.gz* to start the download: https://downloads.apache.org/hive/hive-3.1.2/

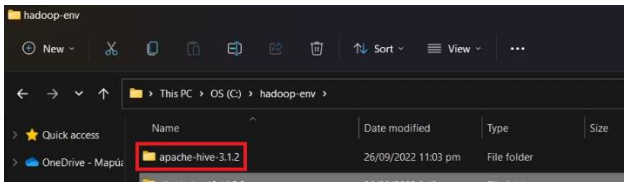
Once the package has finished downloading, open it using WinRAR Archiver and configure its extraction destination to be in the directory "C:\hadoop-env."



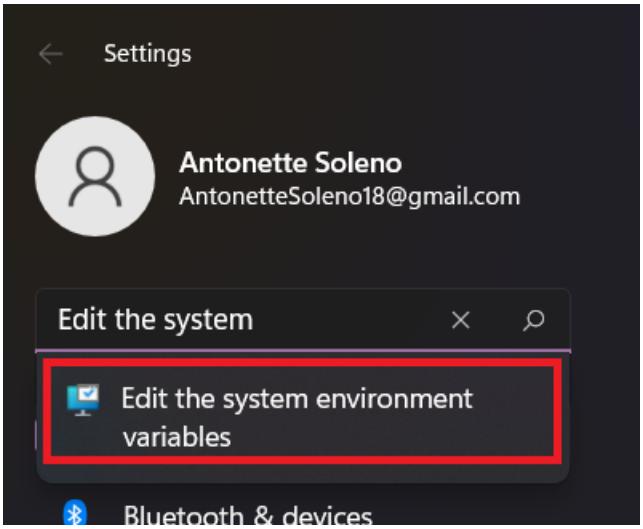


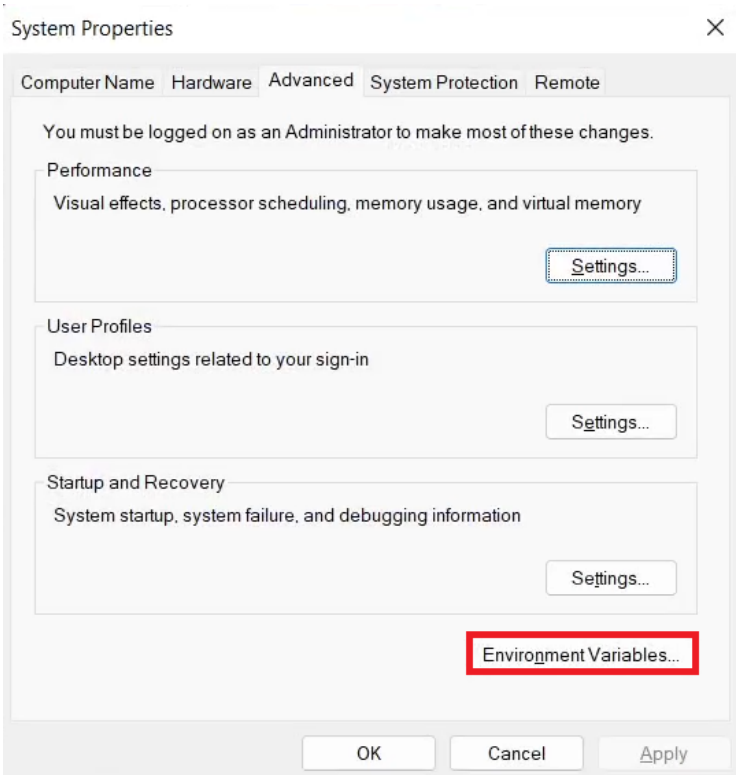
The extracted file should appear in the *hadoop-env* folder as *apache-hive-3.1.2-bin*. Rename this folder as *apache-hive-3.1.2*.



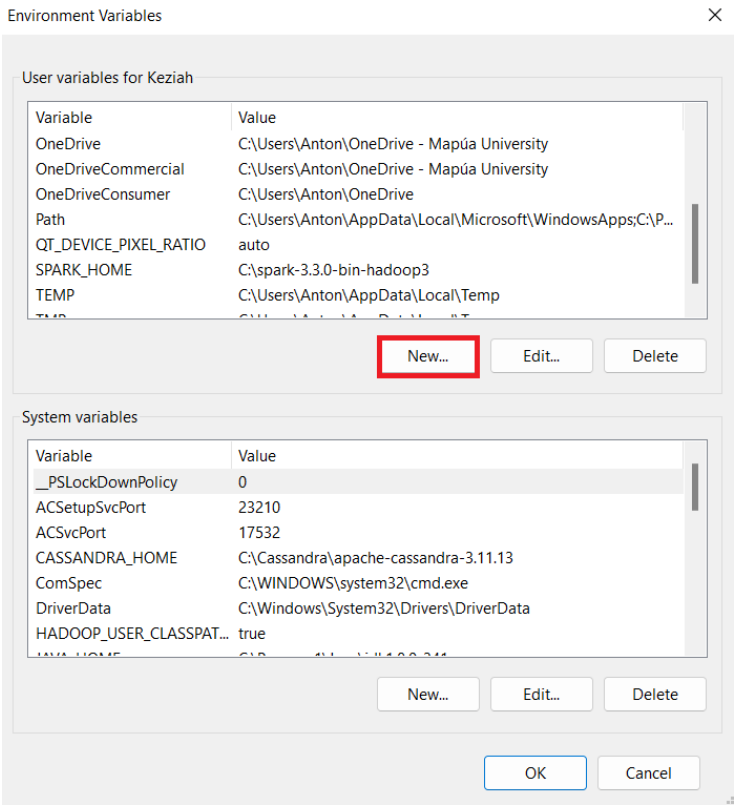


The next step is to set up environment variables for Hive. First, go to **Settings > "Edit the system environment variables"** to open the *System Properties* window. Click *Environment Variables…*.



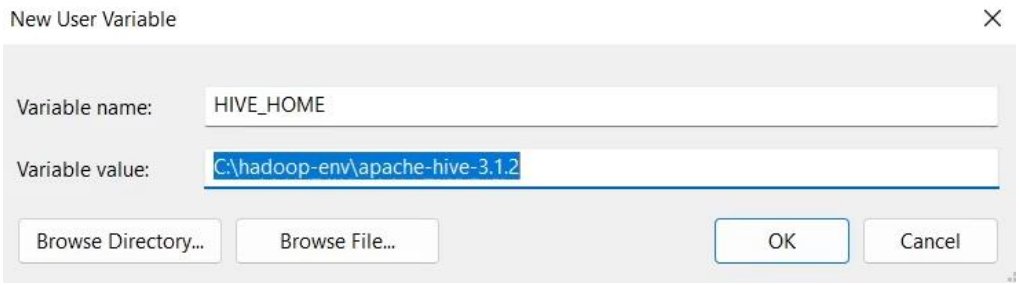


The *Environment Variables* window should appear shortly after. Click “New…” to create a new User variable.

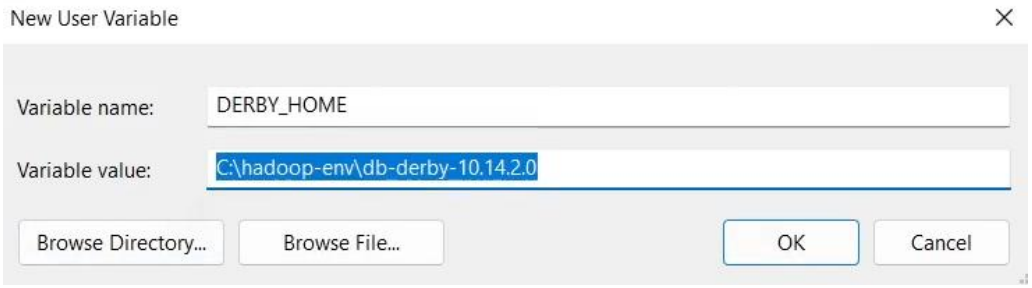


The user variables required in this installation guide are as follows:

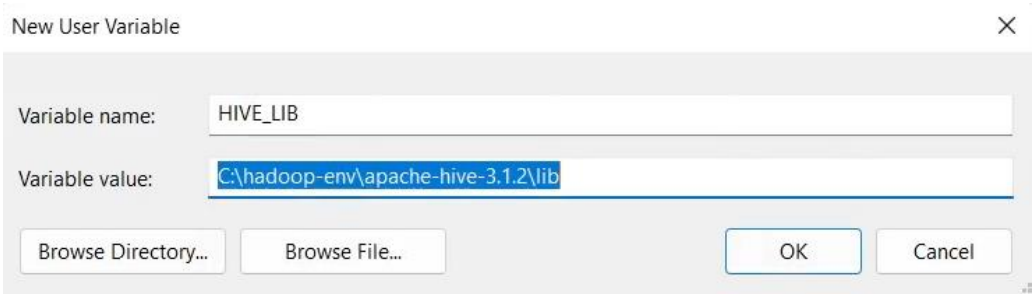
- **HIVE_HOME**: “C:\hadoop-env\apache-hive-3.1 .2”



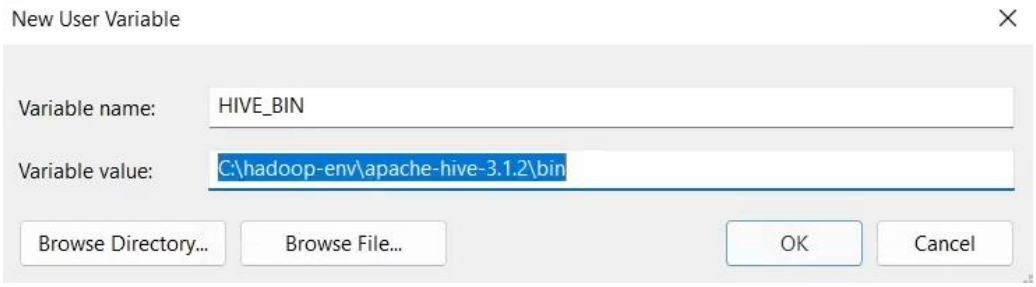
- **DERBY_HOME**: “C:\hadoop-env\db-derby-10.14.2.0”



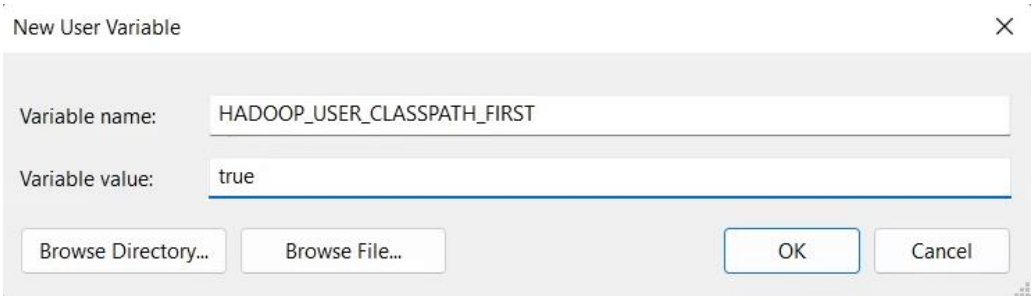
- **HIVE_LIB**: “C:\hadoop-env\apache-hive-3.1.2\lib”



- **HIVE_BIN**: “C:\hadoop-env\apache-hive-3.1.2\bin”

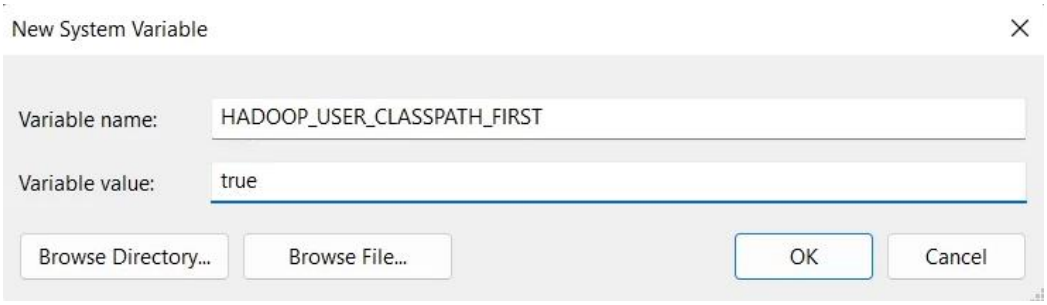


- **HADOOP_USER_CLASSPATH_FIRST**: “true”



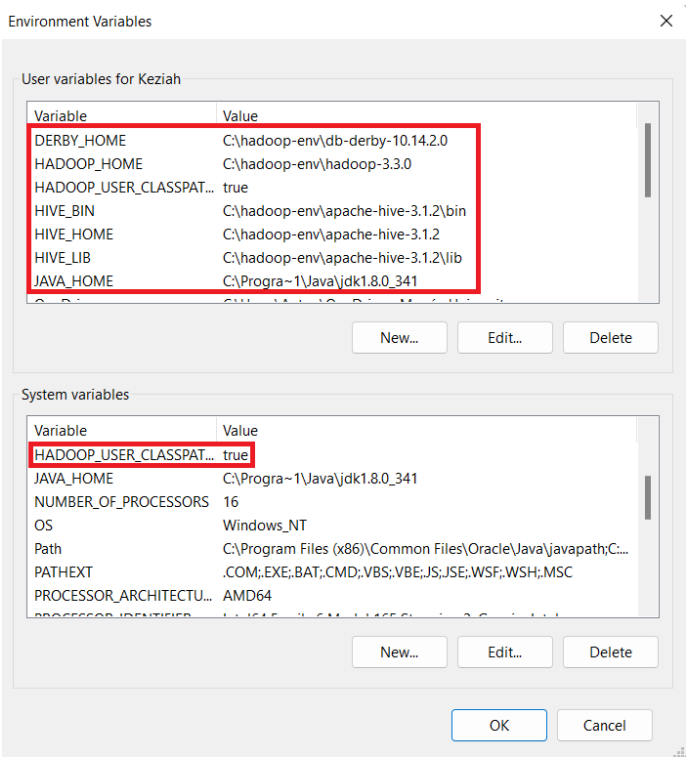
Additionally, the only system variable required is the following:

- **HADOOP_USER_CLASSPATH_FIRST**: “true”

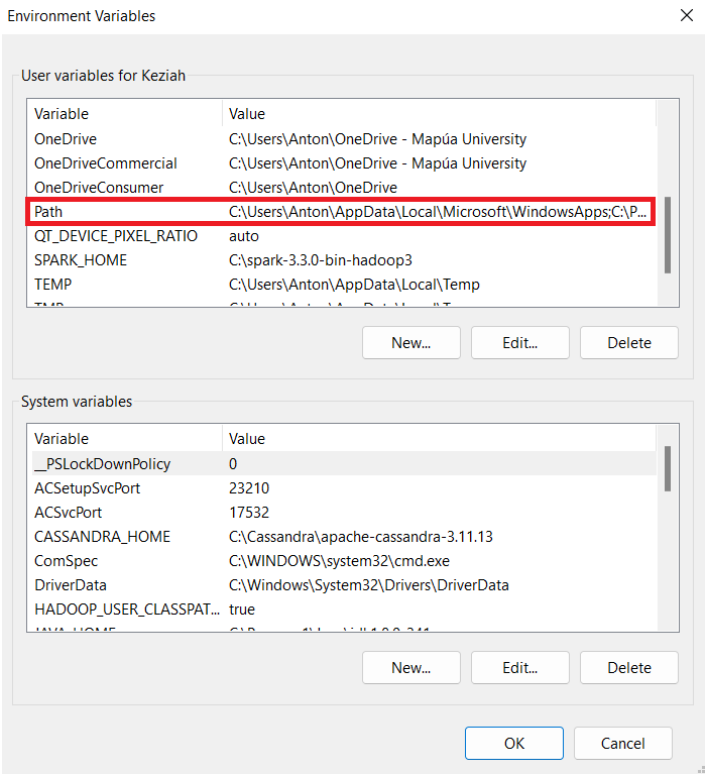


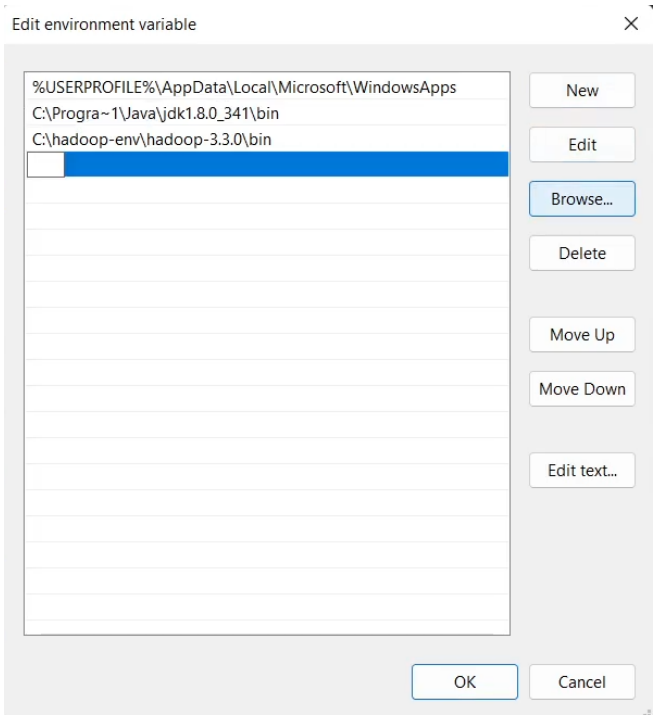
After every user or system variable creation, click “OK” to save the variable.

The *Environment variables* window should now look like this.

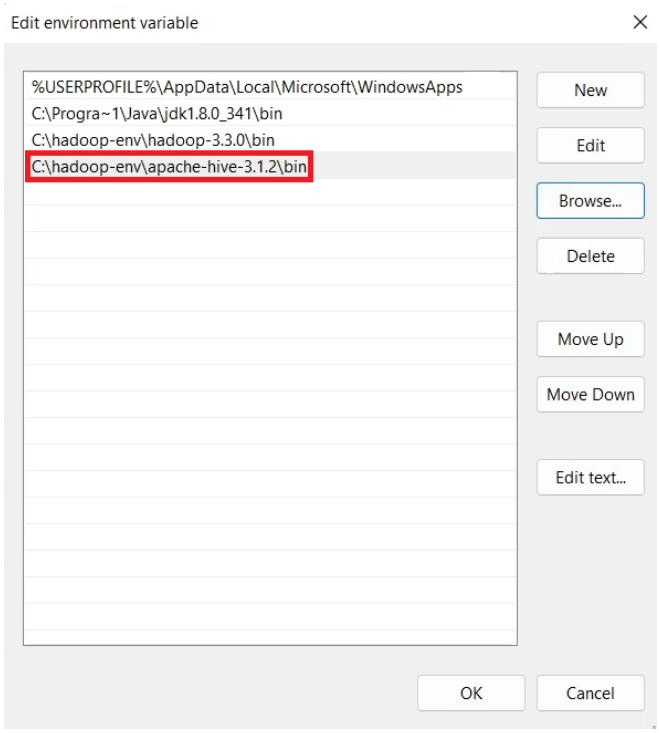


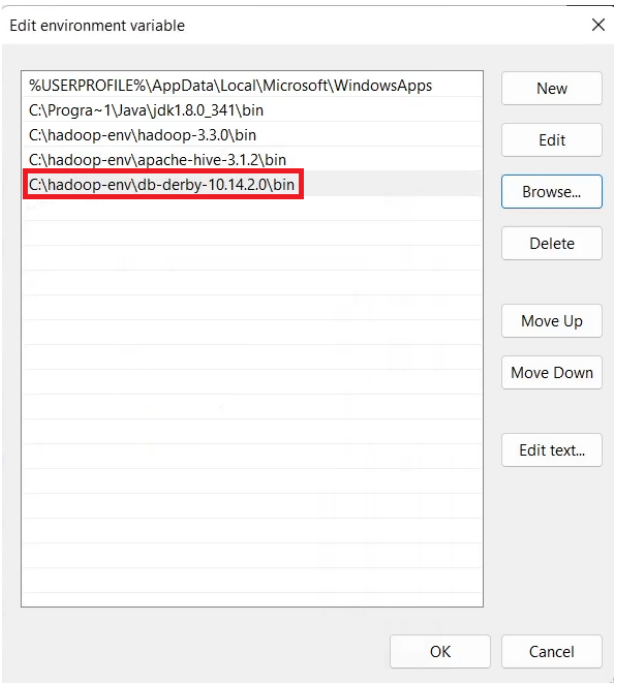
Once all required variables have been created, locate “Path” under *User variables* then double click to open the *Edit environment variable* window.





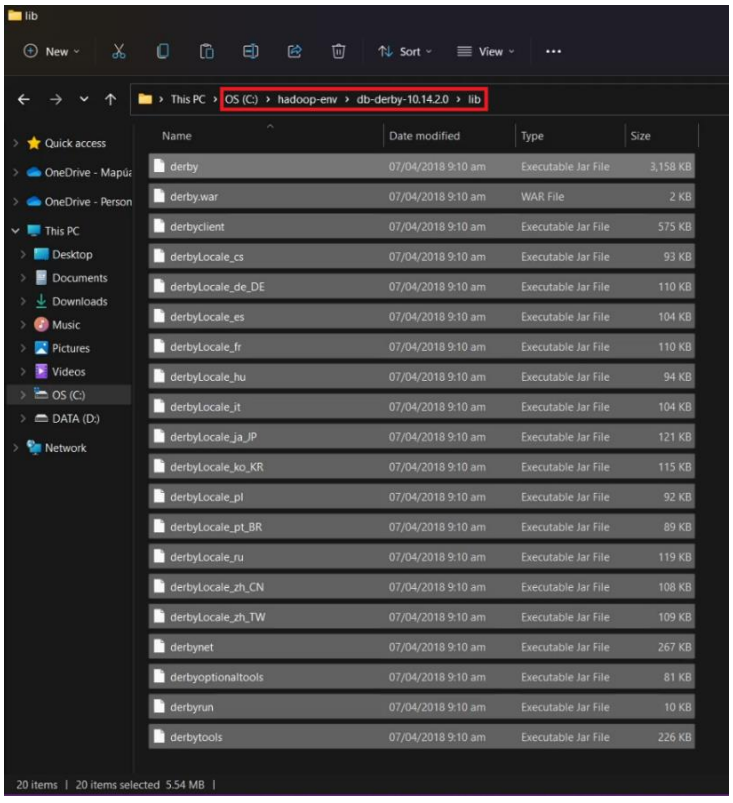
On the *Edit environment variable* window, add two paths leading to each bin folder of the Hive (apache-hive-3.1.2) and Derby (db-derby-10.14.2.0) folders.

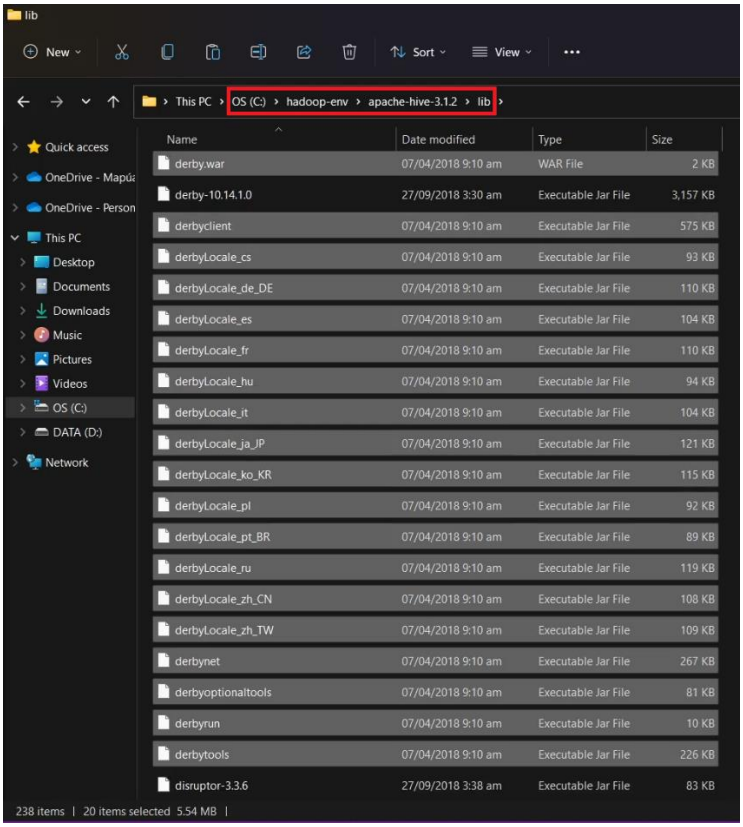




a. Configure Apache Hive

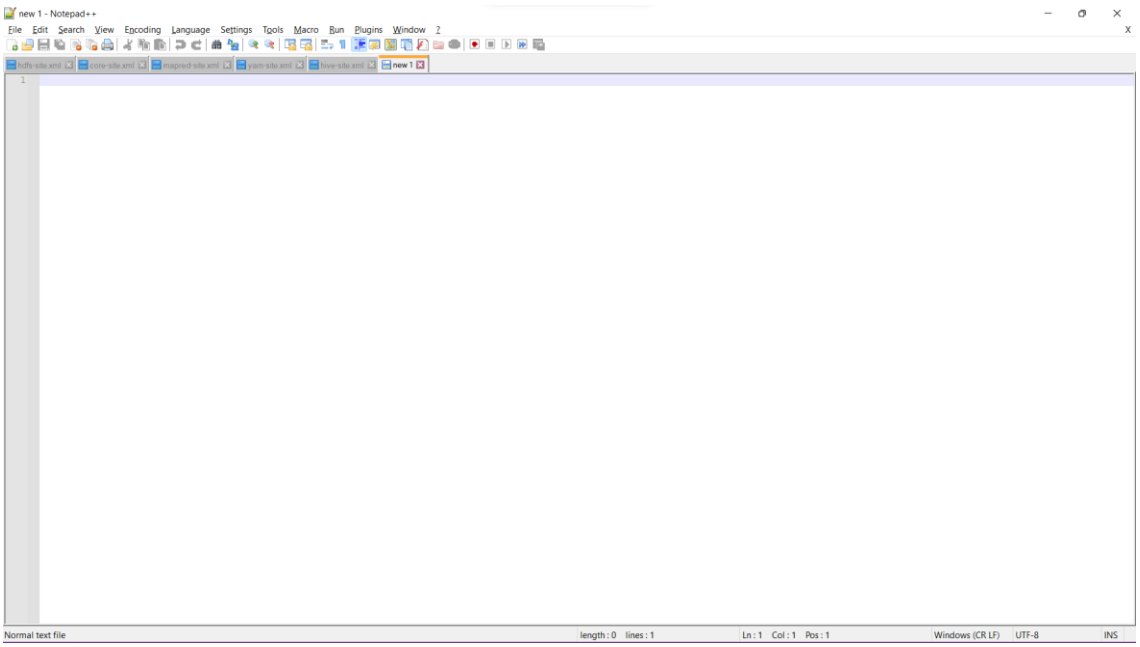
After setting the environment variables for Apache Hive and Derby, the next step is to copy all *.jar files from the Derby libraries directory (*C:\hadoop-env\db-derby-10.14.2.0\lib*) and paste them onto the Hive libraries directory (*C:\hadoop-env\apache-hive-3.1.2\lib*).





The next step is to create a new file entitled *hive-site.xml* in the Apache Hive configuration directory (*C:\hadoop-env\apache-hive-3.1.2\conf*).

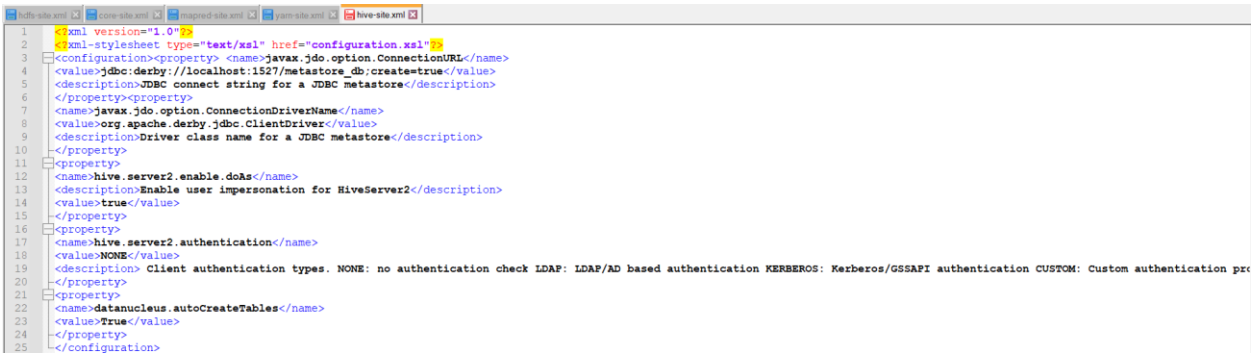
First, open Notepad++ on your local machine and click **File > New**. The content of the figure below should appear.



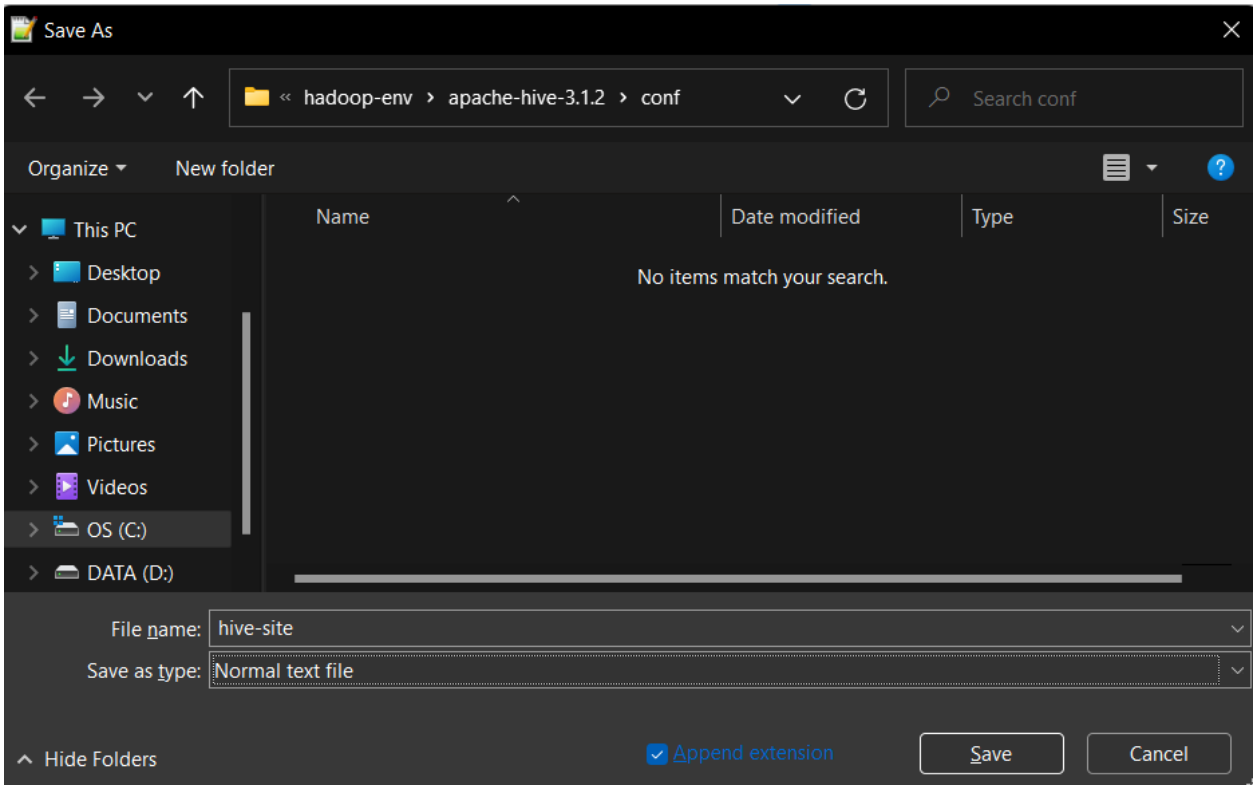
Then, copy the following XML code:

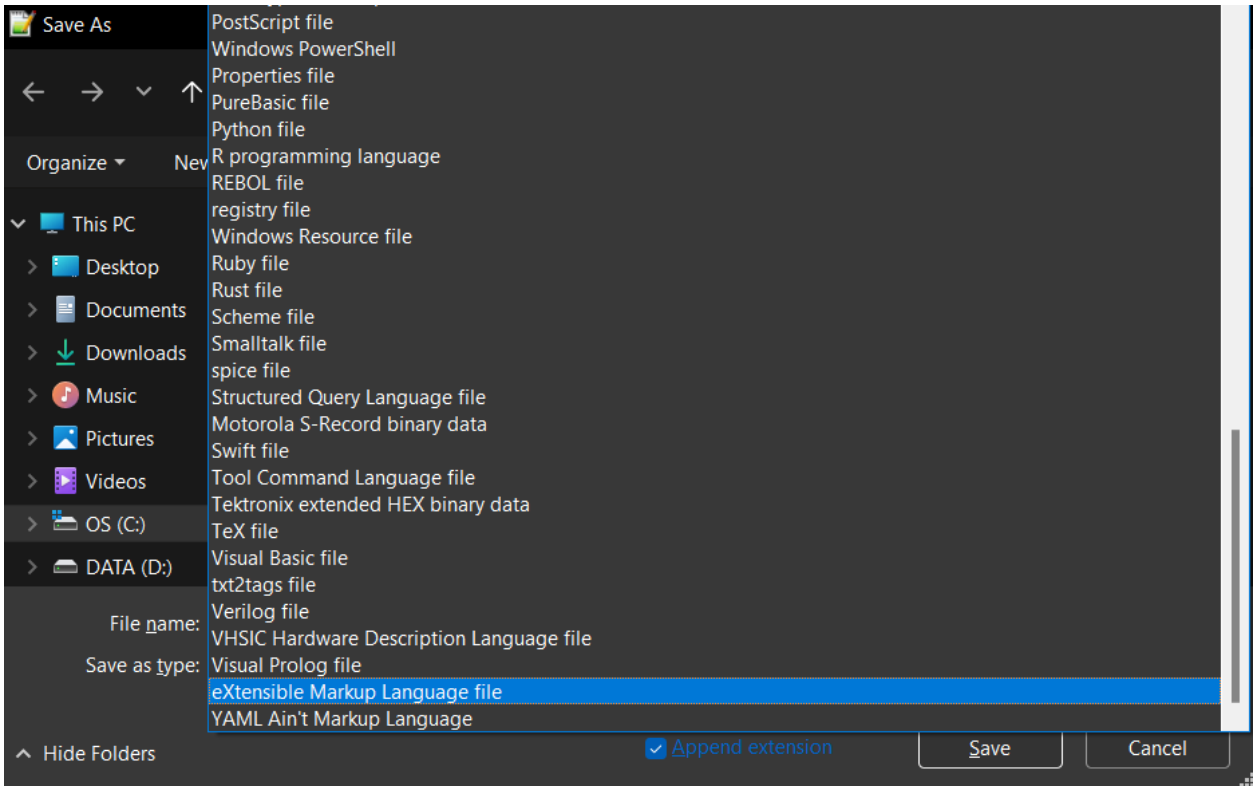
```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration><property> <name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:derby://localhost:1527/metastore_db;create=true</value>
<description>JDBC connect string for a JDBC metastore</description>
</property><property>
<name>javax.jdo.option.ConnectionDriverName</name>
<value>org.apache.derby.jdbc.ClientDriver</value>
<description>Driver class name for a JDBC metastore</description>
</property>
<property>
<name>hive.server2.enable.doAs</name>
<description>Enable user impersonation for HiveServer2</description>
<value>true</value>
</property>
<property>
<name>hive.server2.authentication</name>
<value>NONE</value>
<description> Client authentication types. NONE: no authentication check
LDAP: LDAP/AD based authentication KERBEROS: Kerberos/GSSAPI authentication
CUSTOM: Custom authentication provider (Use with property
hive.server2.custom.authentication.class) </description>
</property>
<property>
<name>datanucleus.autoCreateTables</name>
<value>True</value>
</property>
</configuration>
```

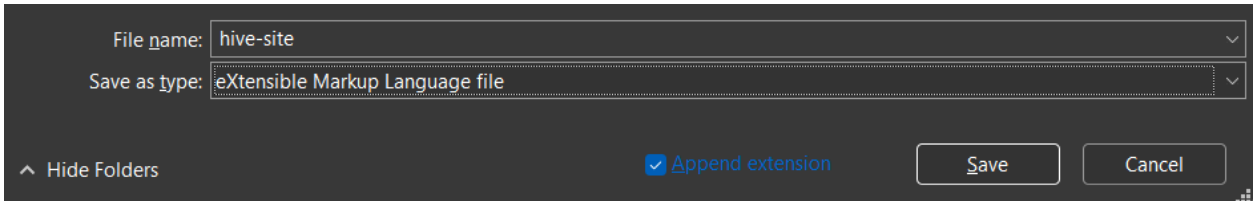
And paste it onto the newly created Notepad++ file. The file should look like the image below.



Next, click **File > Save as**. Set the file name as *hive-site*, and the file type as eXtensible Markup Language (XML) file, as indicated below.



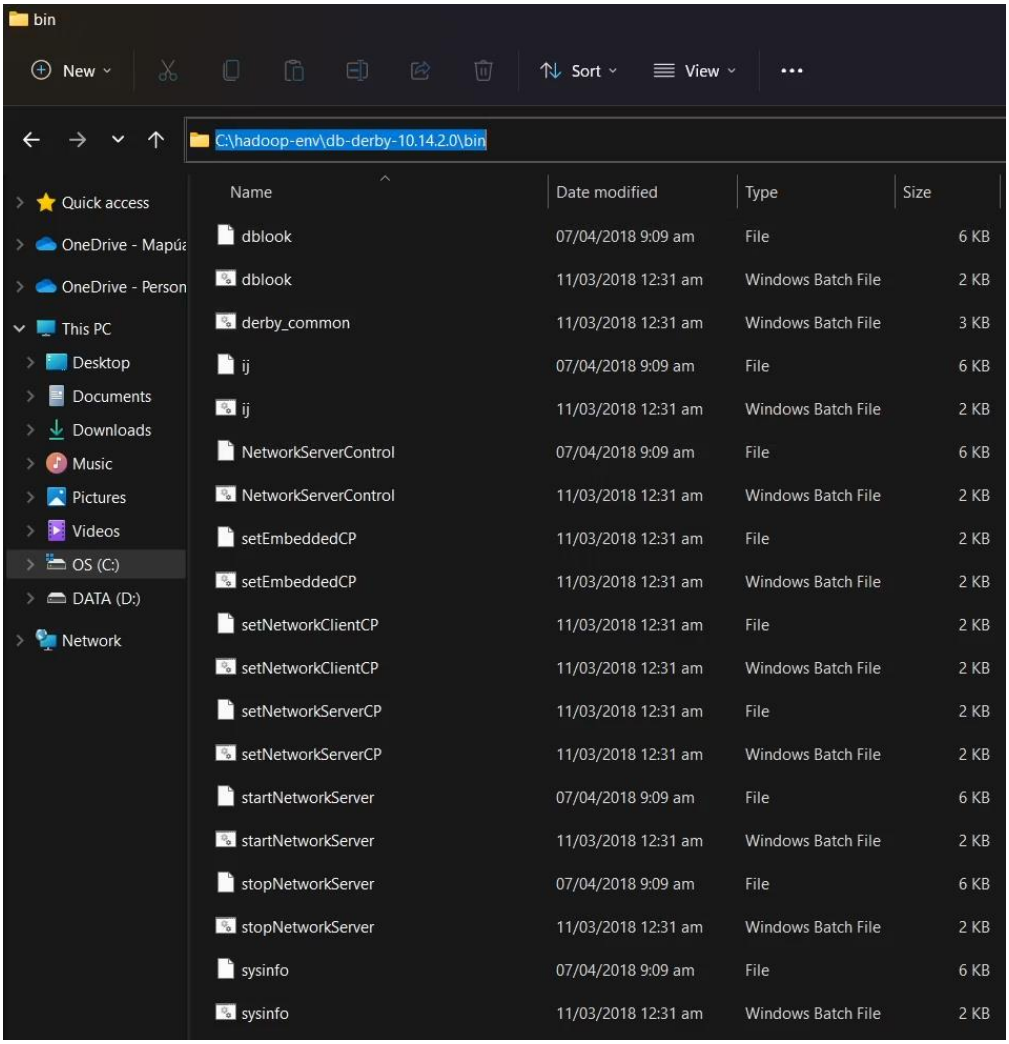


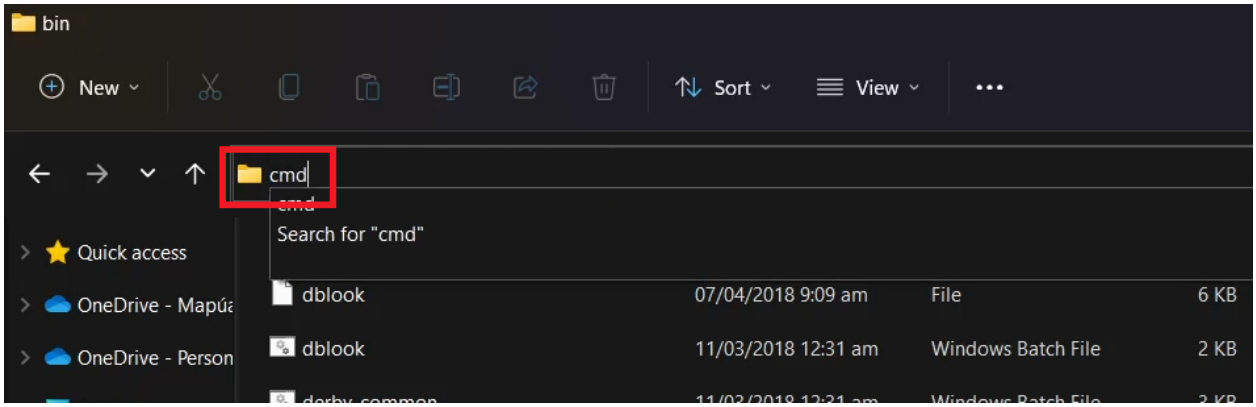


Now that the prerequisite configurations have been made, all the installed services may already be run.

To begin, run the Hadoop services by following the instructions under **II. Hadoop Installation** > *e. Start Hadoop services* found in this installation guide.

Then, start the Derby network server by first opening the command prompt terminal as administrator on the directory *C:\hadoop-env\db-derby-10.14.2.0\bin*,

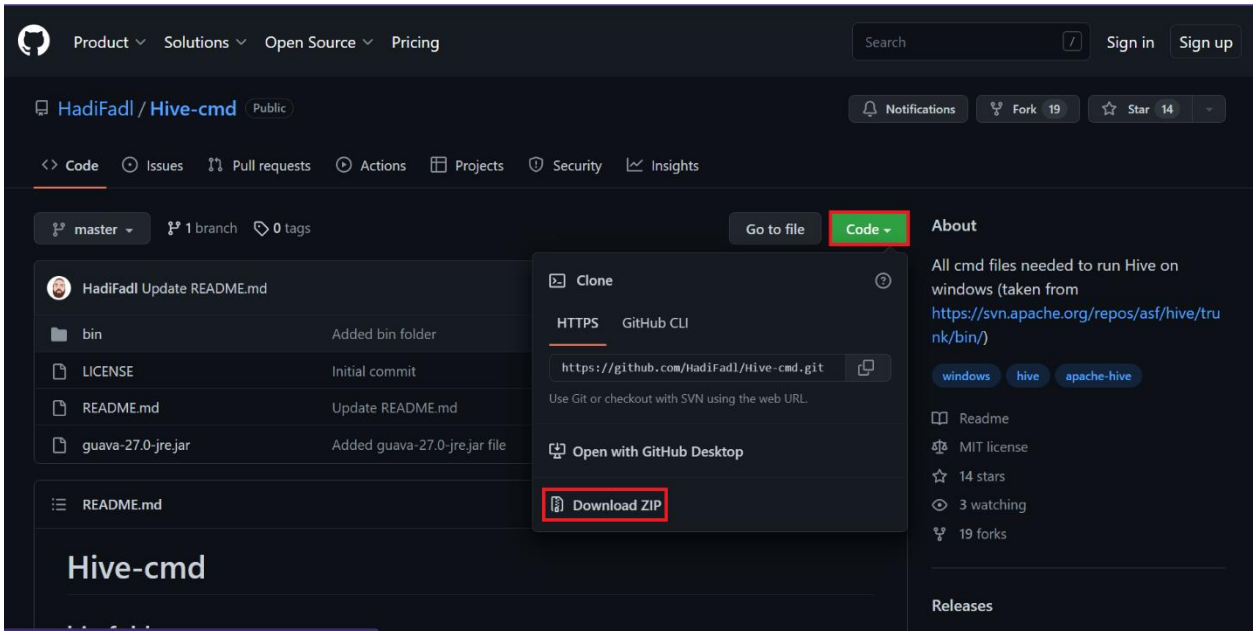




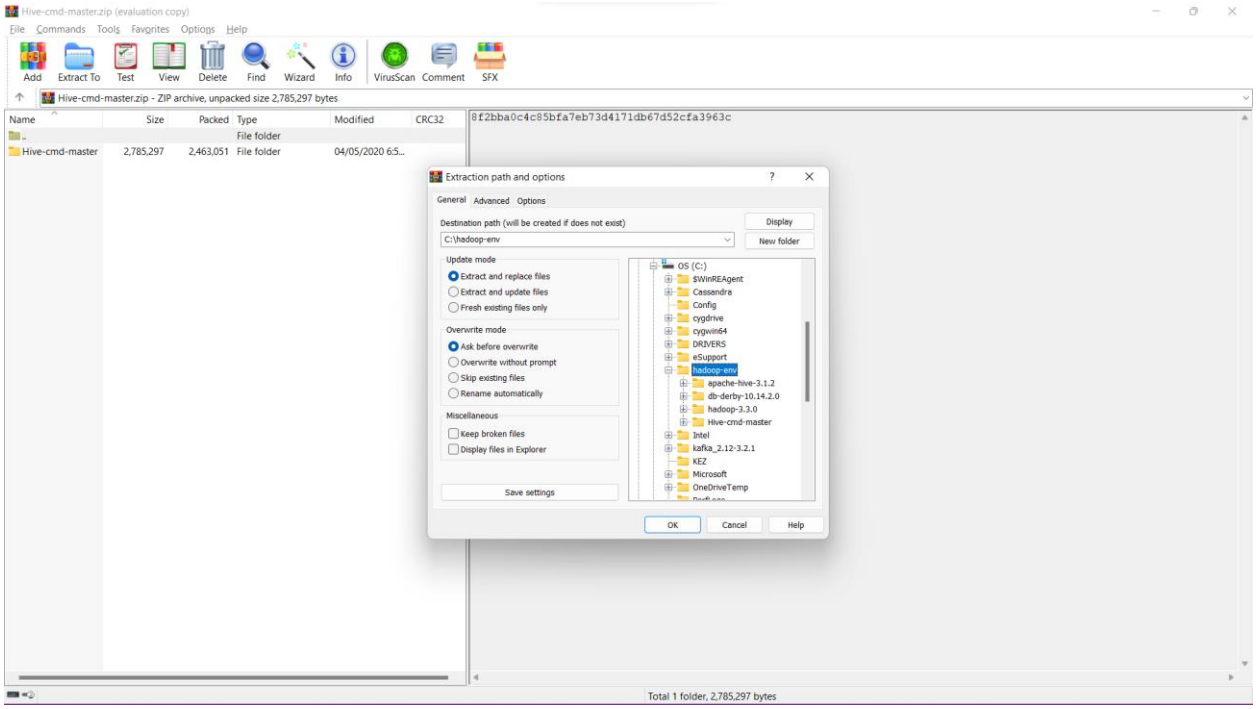
Now run the following command on the command prompt terminal.

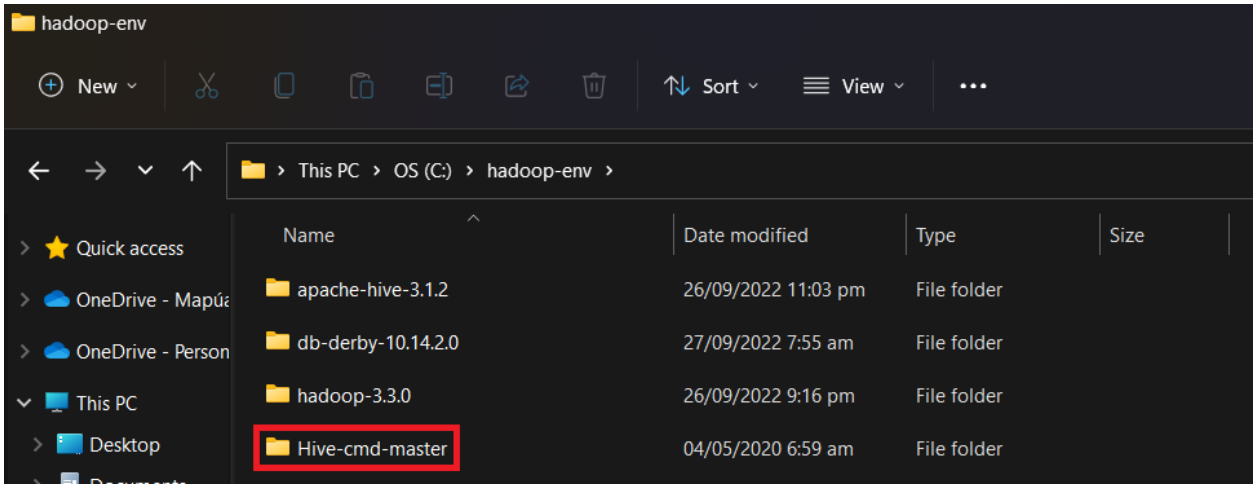
```
StartNetworkServer -h 0.0.0.0
```

Before starting Apache Hive, download the repository found in this link: https://github.com/HadiFadl/Hive-cmd by clicking **Code > Download ZIP**, as indicated below.

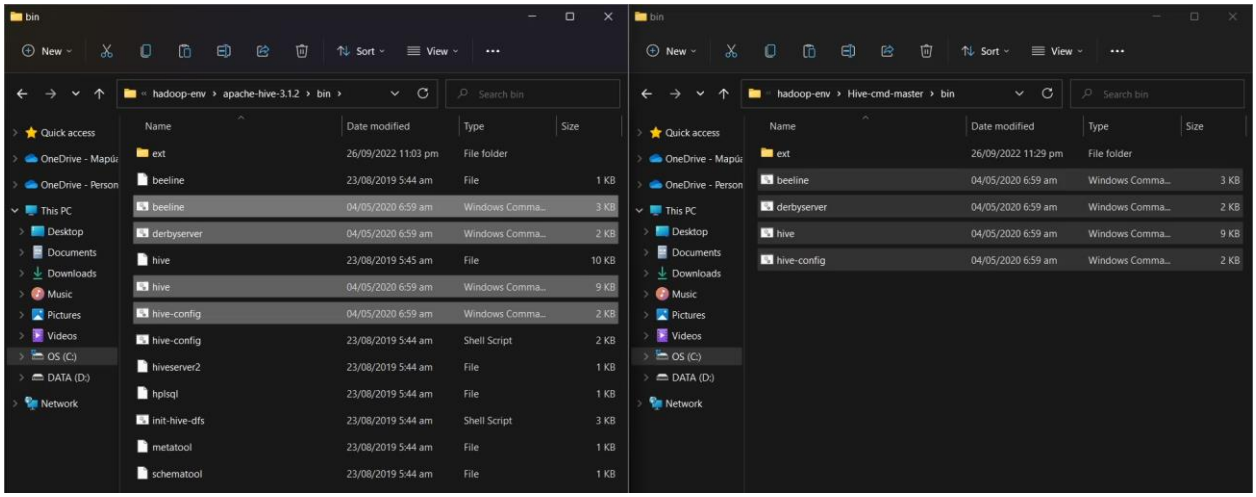


After the download has finished, unzip the file using WinRAR Archiver (or any other unzipping tool available) and save it in the hadoop-env folder.

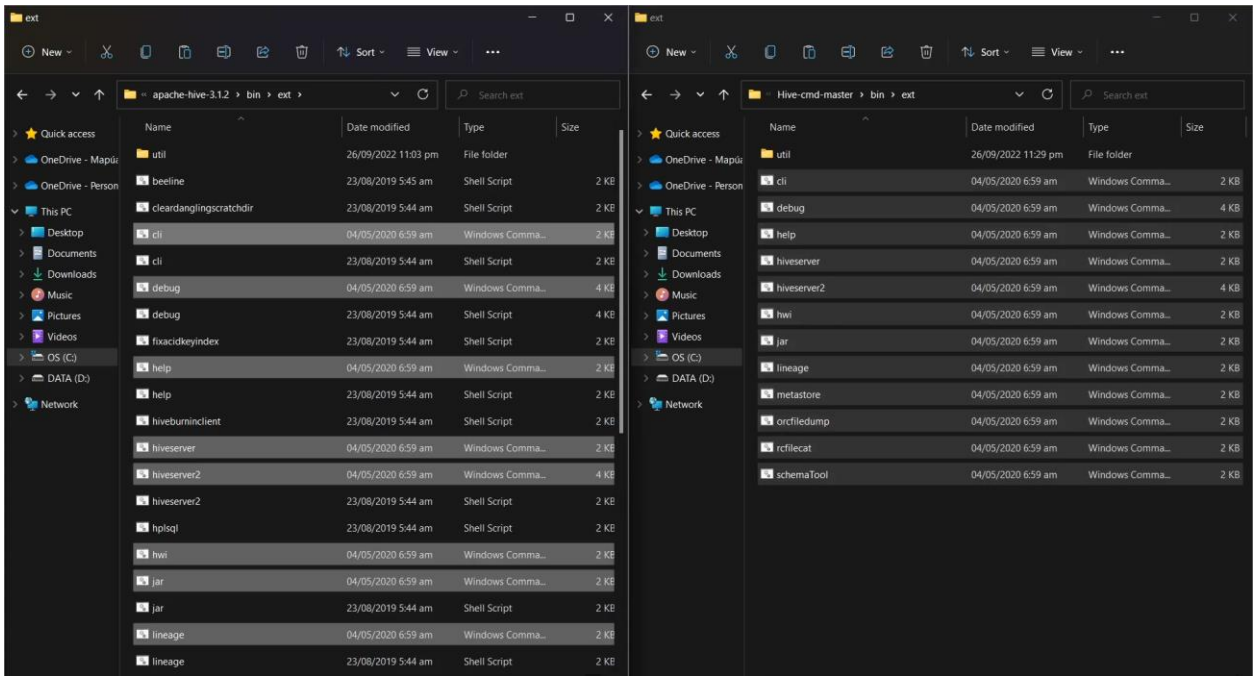




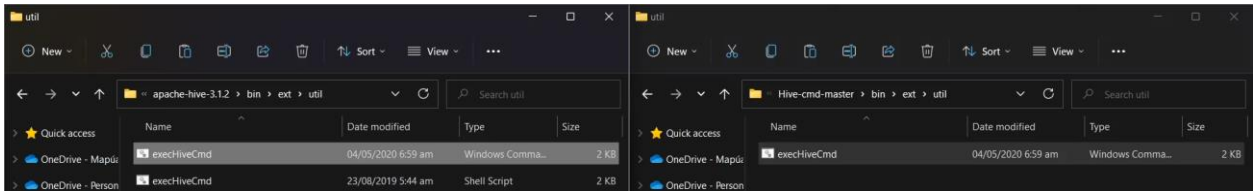
Next, go to the directory *C:\hadoop-env\Hive-cmd-master\bin*, then copy all *.cmd files and paste them onto the directory *C:\hadoop-env\apache-hive-3.1.2\bin*. Repeat the same process for folders *ext* and *util*.



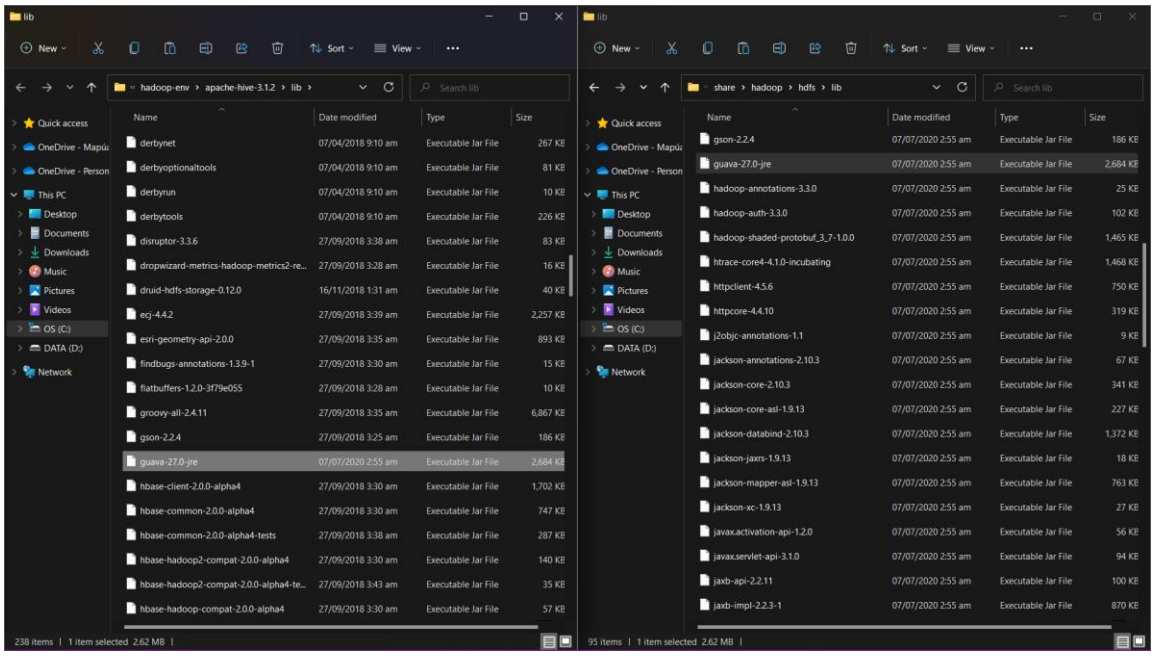
Go to the directory *C:\hadoop-env\Hive-cmd-master\bin\ext*, then copy all *.cmd files and paste them onto the directory *C:\hadoop-env\apache-hive-3.1.2\bin\ext*.



Go to the directory *C:\hadoop-env\Hive-cmd-master\bin\ext\util*, then copy all *.cmd files and paste them onto the directory *C:\hadoop-env\apache-hive-3.1.2\bin\ext\util*.

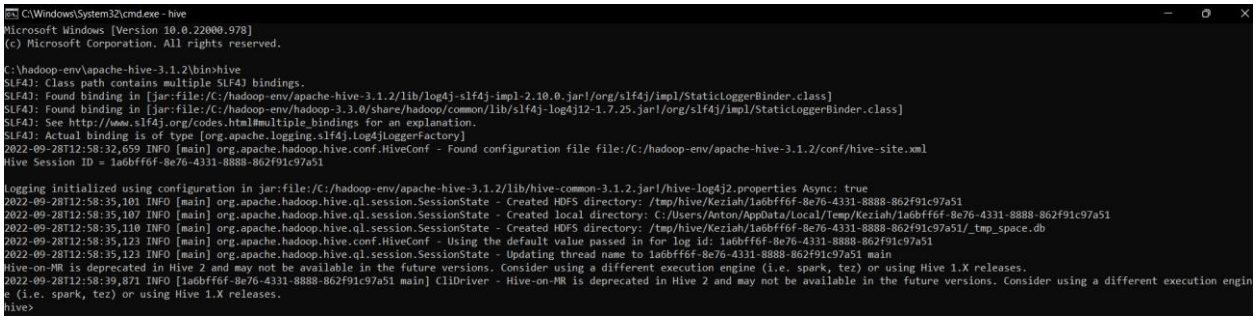


To bypass a Hive bug mentioned in this issue link: https://issues.apache.org/jira/browse/HIVE-22718, replace the *guava-19.0.jar* in directory *C:\hadoop-env\apache-hive-3.1.2\lib* with *guava-27.0-jre.jar* from the directory *C:\hadoop-env\hadoop-3.3.0\share\hadoop\hdfs\lib*.



You are now ready to run Apache Hive on your local machine. You can run the command prompt terminal as administrator on the directory *C:\hadoop-env\apache-hive-3.1.2\bin* and execute the following command to start Apache Hive:

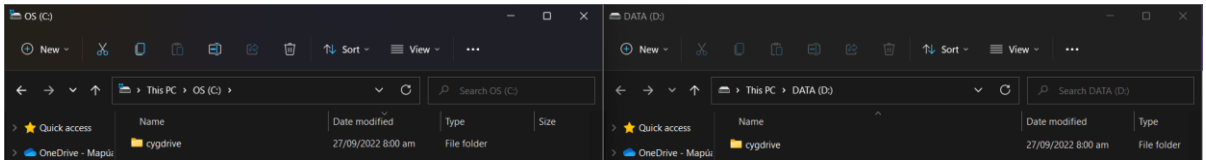
```
hive
```



b. Configure and using Hive2

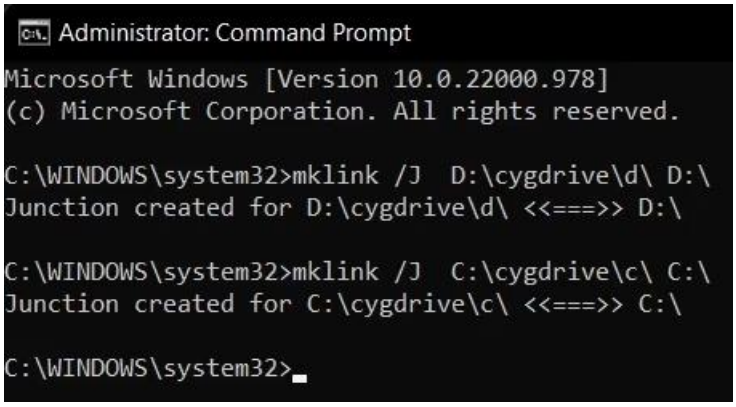
However, we may not be able to run any HiveQL commands. Therefore, Metastore must be initialized to run HiveQL commands using a schematool utility. Before starting Metastore, first close Apache Hive (Leave the other services open) to ensure that it does not interfere with the initialization. Since the schematool utility is not Windows-compatible, the Cygwin utility that was previously downloaded will be used to execute this Linux command from Windows. Before running the Cygwin utility, first create the following directories in your local machine:

- *C:\cygdrive*
- *D:\cygdrive*

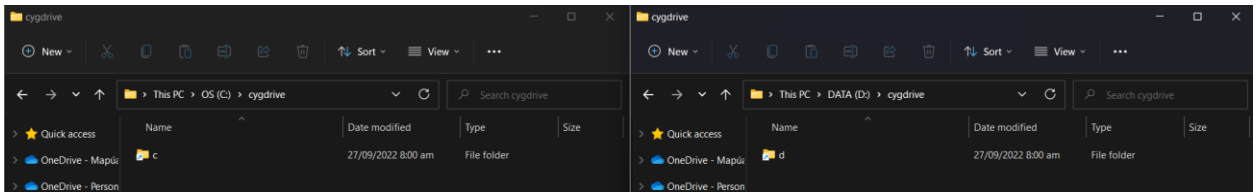


Then, run the command prompt terminal as administrator and execute the following commands:

```
mklink /J  D:\cygdrive\d\ D:\
mklink /J  C:\cygdrive\c\ C:\
```



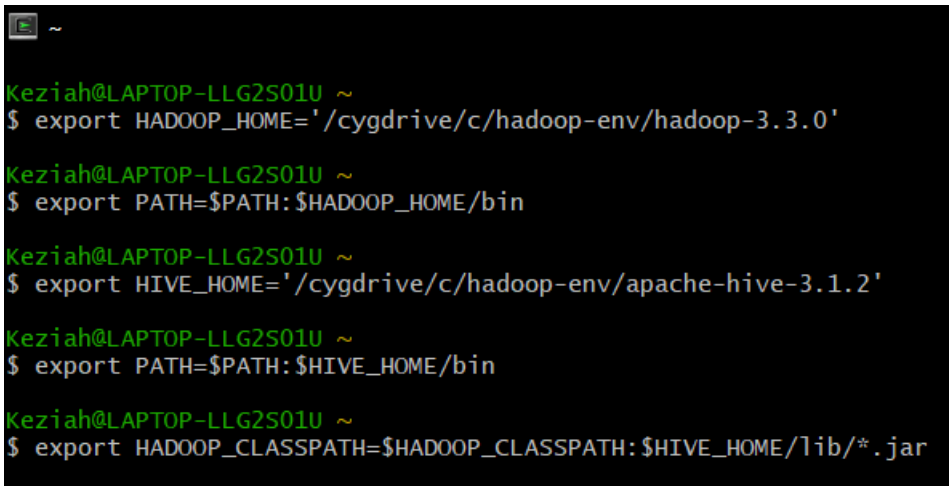
Running the above commands should result in the output below:



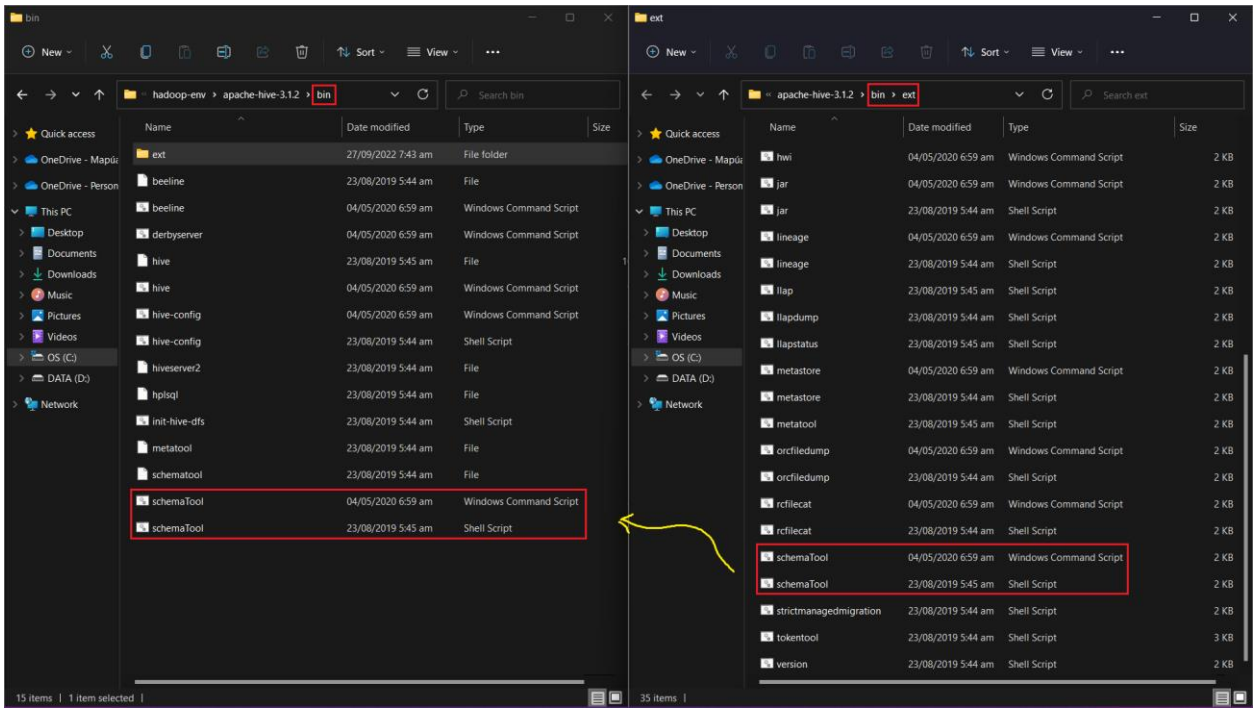
The symbolic links created above will be used to work with Cygwin as Java may produce issues in the long run.

To begin, open the Cygwin terminal and run the following commands:

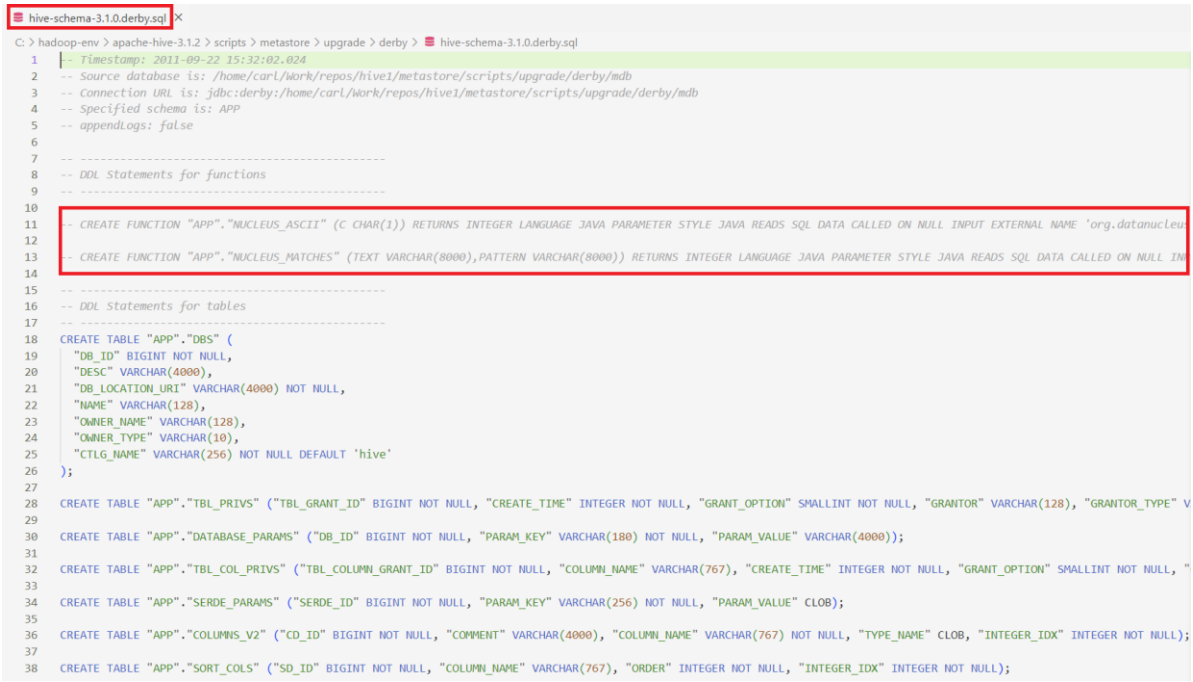
```
export HADOOP_HOME='/cygdrive/c/hadoop-env/hadoop-3.3.0'
export PATH=$PATH:$HADOOP_HOME/bin
export HIVE_HOME='/cygdrive/c/hadoop-env/apache-hive-3.1.2'
export PATH=$PATH:$HIVE_HOME/bin
export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$HIVE_HOME/lib/*.jar
```



Before running the schematool utility, ensure that **both schemaTool Windows Command Script and Shell Scripts** are copied from the *ext* folder and pasted into the *bin* folder of Apache Hive.

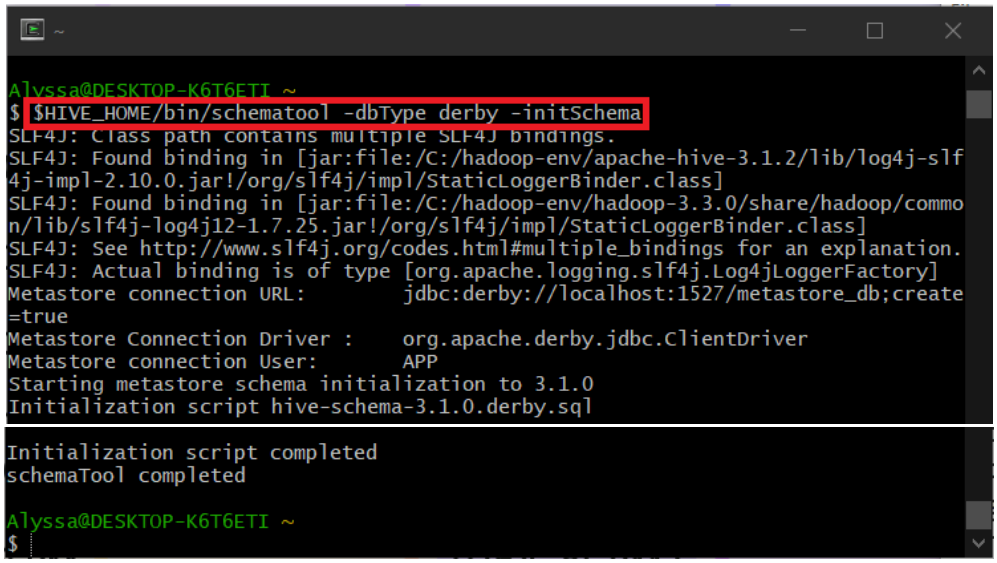


(Optional) Next, go to this directory: *C:\hadoop-env\apache-hive-3.1.2\scripts\metastore\upgrade\derby\hive-schema-3.1.0.derby* and comment out the function highlighted in the figure below. This will help with the error associated with initializing the Metastore.



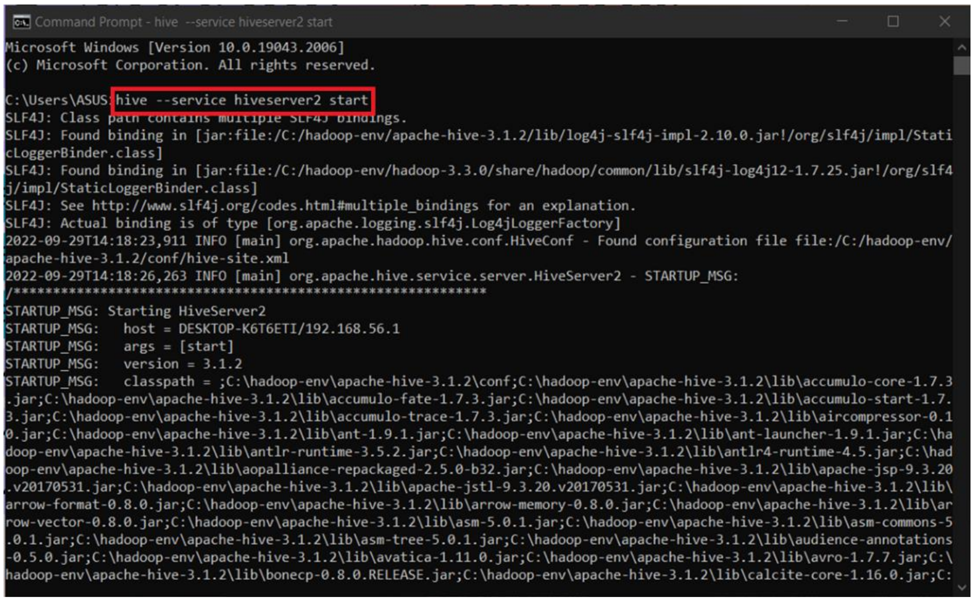
To initialize the metastore using the schematool utility, execute the following command on Cygwin. The output in the figure below it should appear, indicating that schematool utility initialization was successful.

```
$HIVE_HOME/bin/schematool -dbType derby -initSchema
```



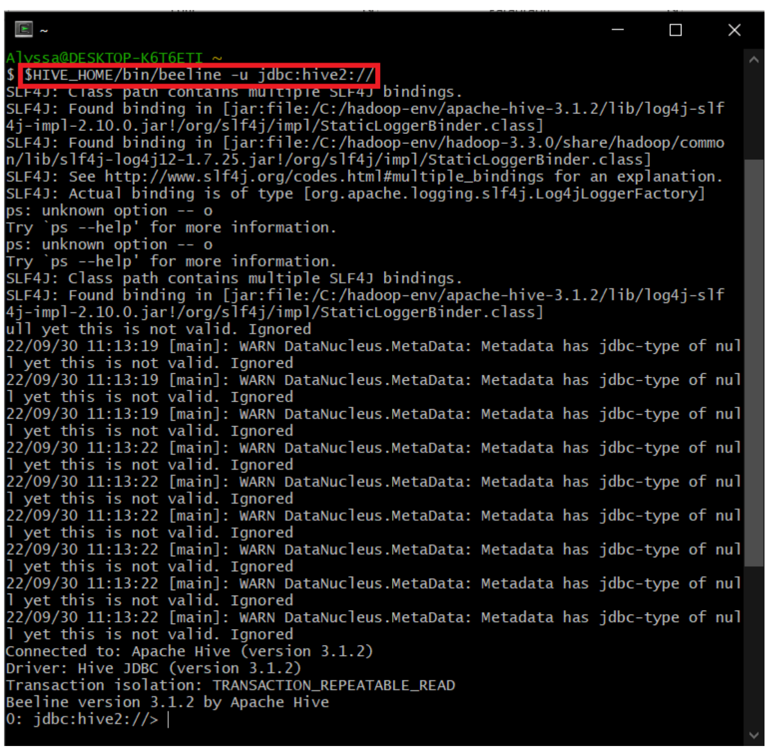
Afterwards, open a Windows Command Prompt terminal (or you may open another Cygwin terminal) and execute the following command and allow the service to run as a background process:

```
hive --service hiveserver2 start
```



Then, return to the Cygwin terminal, where the schematool utility was initialized, and execute the following code to run hive2:

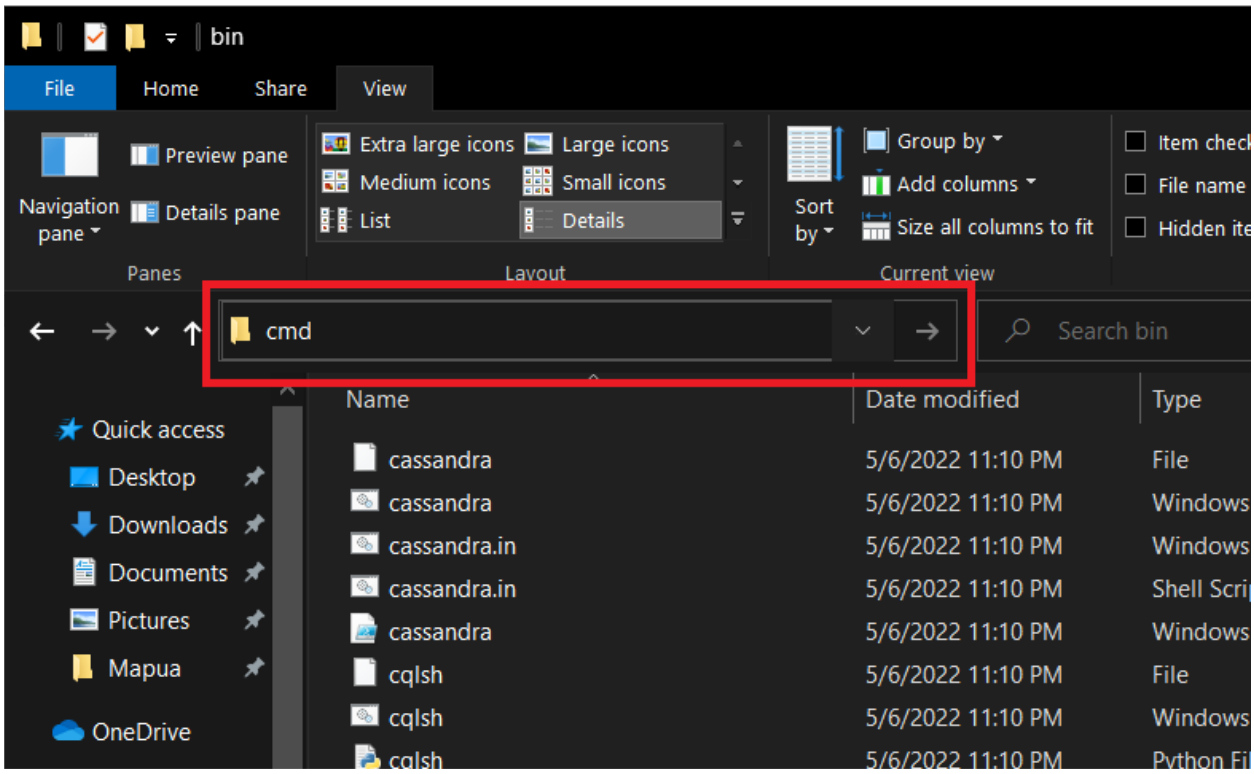
```
$HIVE_HOME/bin/beeline -u jdbc:hive2://
```



VI. Migration from Apache Cassandra to Hive Data Warehouse

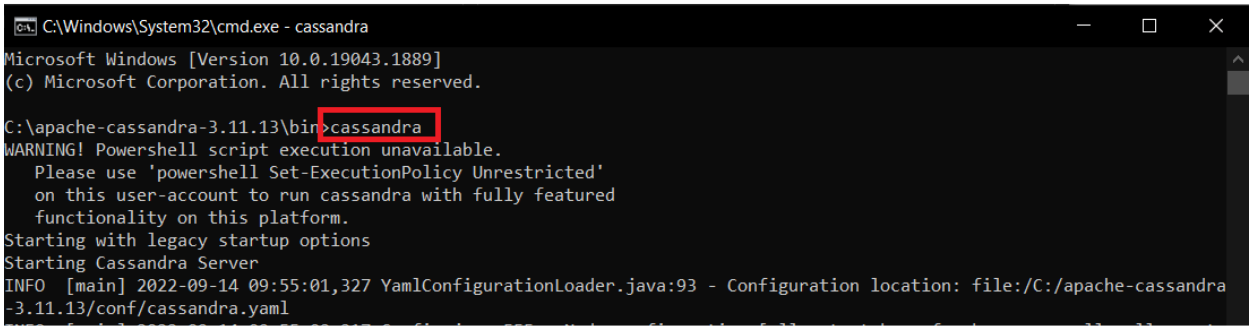
a. Initializing Cassandra and Hive

To begin the migration from Cassandra to Hive, start Cassandra by going to the directory *C:\apache-cassandra-3.11.13\bin* and typing 'cmd' on the address bar to open Windows Command Prompt terminal.



Once the terminal has opened, enter the following command to start Cassandra:

```
cassandra
```

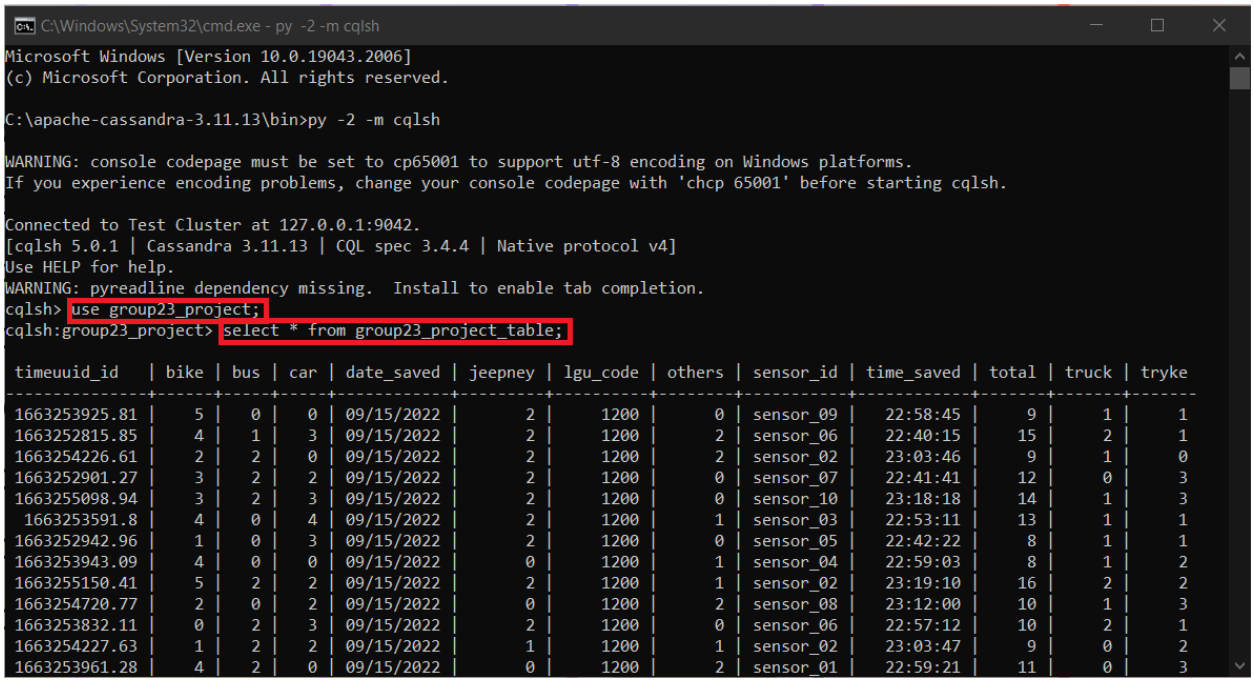


As the initial terminal is running, we can open a new command prompt terminal from the same directory and enter the following command to access Cassandra's database:

```
cqlsh
```

To check the data retrieved through the previous assignment, run the following commands to show all data from the previously made table in keyspace group23_project.

```
use group23_project;
select * from group23_project_table;
```



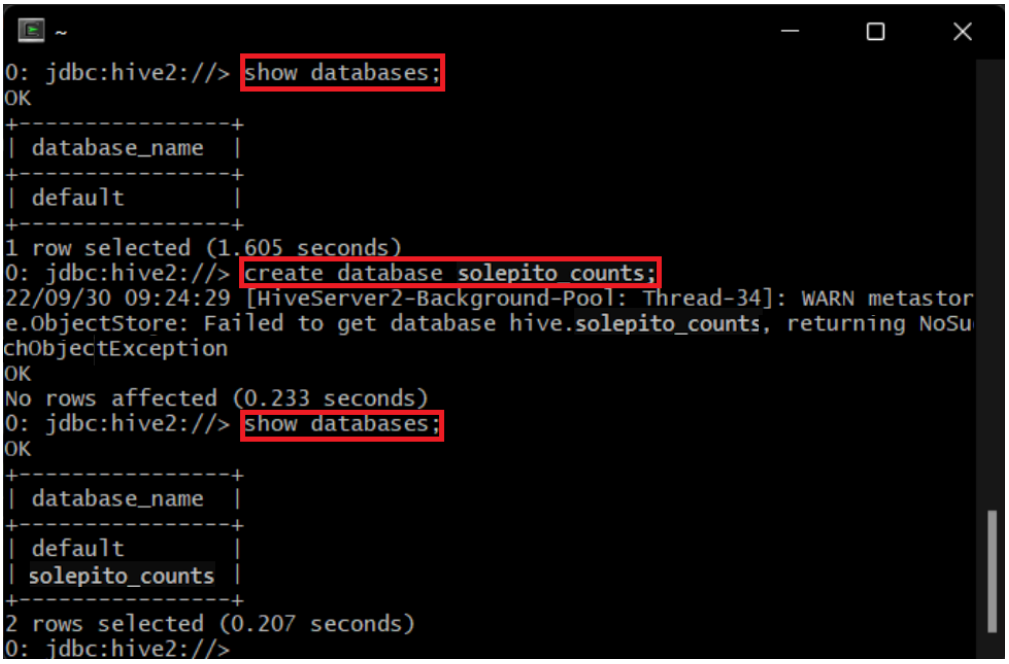
Since Hive has already been opened through the previous steps, the next step is to create the Python program that will migrate the data from the `group23_project_table` from Cassandra to a new table in Hive called `solepito_counts.sumry`.

First, we need to create a hive database, so go back to the Cygwin terminal where we ran hive2. We first ran the following command to show the available databases in hive2.

```
show databases;
```

We then created the database that would be used for the program using this command:

```
create database solepito_counts;
```



b. Creating the Python program to migrate data from Cassandra to Apache Hive

Now that both databases Cassandra and Hive have been initialized, a Python program to migrate data from Cassandra to Hive may be created. To begin, we imported the following libraries to access Hive and Cassandra using PySpark from Python:

```
import os
import sys
from pyspark.shell import spark
from pyspark.sql import SparkSession
from pyspark import SparkConf
from os.path import abspath
from datetime import datetime
```

Then, we need to set some arguments or configurations to make sure PySpark connects to our Cassandra node cluster.

```
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable
```

```
os.environ['PYSPARK_PYTHON'] = sys.executable
spark = SparkSession.builder.getOrCreate()
warehouse_location = abspath('spark-warehouse')
```

We also want to configure the connection between PySpark and Hive.

```
hive_conf = SparkConf()
hive_conf.set("spark.jars", "C:\hadoop-env\db-derby-
10.14.2.0\lib\derbyclient.jar")
```

Now, we need to make functions that will get the CCTV counts per minute and hour.

```
def extractHours(hours):
    return hours[:2]
def extractMinutes(mins):
    return mins[3:5]
def toDate(col):
    return datetime.strptime(col, '%m/%d/%Y')
```

After initializing the functions, we are going to need to get the data frame from the Cassandra table. The code below loads and returns the data frame from the Cassandra table and keyspace that was indicated earlier.

```
df_cass = spark.read \
    .format("org.apache.spark.sql.cassandra") \
    .options(table="group23_project_table", keyspace="group23_project") \
    .load()
```

Now we want to convert this data frame to the pandas data frame to easily modify and add the CCTV counts per hour and minute for the table.

```
df_pandas = df_cass.toPandas()
df_pandas['date_saved'] = df_pandas["date_saved"].apply(toDate)
df_minhours = df_pandas[["date_saved", "time_saved", "bike", "bus", "car",
"jeepney", "others", "truck", "tryke", "total"]]
df_minhours['hour'] = df_minhours['time_saved'].apply(extractHours)
df_minhours['min'] = df_minhours['time_saved'].apply(extractMinutes)
```

Then, we combine the columns for the hours and minutes to a new table where we count the total of each bike, bus, car, jeepney, truck, tryke, and other vehicles.

```
df_combine = df_minhours.groupby(['date_saved', 'hour', 'min']).agg(
    total=('total', sum),
    bike_total=('bike', sum),
    bus_total=("bus", sum),
    car_total=("car", sum),
    jeepney_total=("jeepney", sum),
    others_total=("others", sum),
```

```
    truck_total=("truck", sum),
    tryke_total=("tryke", sum)
).reset_index()
df_counts = spark.createDataFrame(df_combine)
df_counts.printSchema()
```

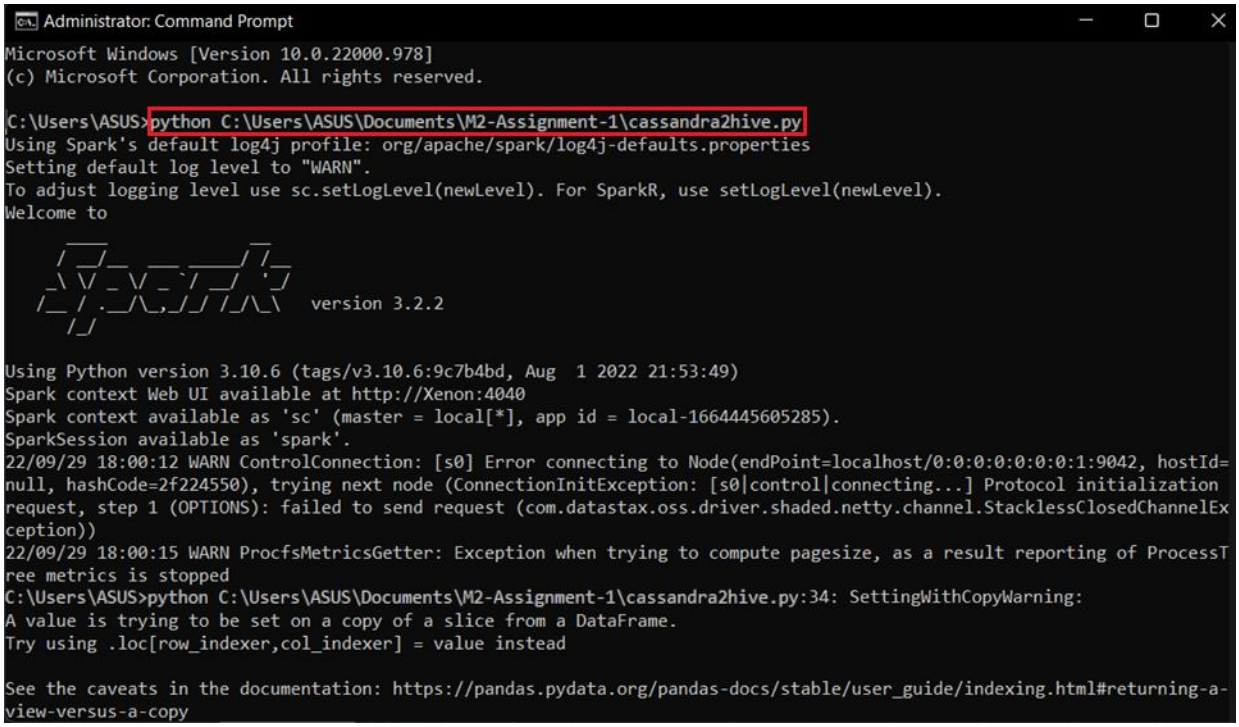
Finally, we save the data frame where everything has been combined into a table in Hive.

```
df_counts.write.mode("overwrite").saveAsTable("solepito_counts.sumry")
df_cass.write.mode("overwrite").saveAsTable("solepito_counts.cass")
```

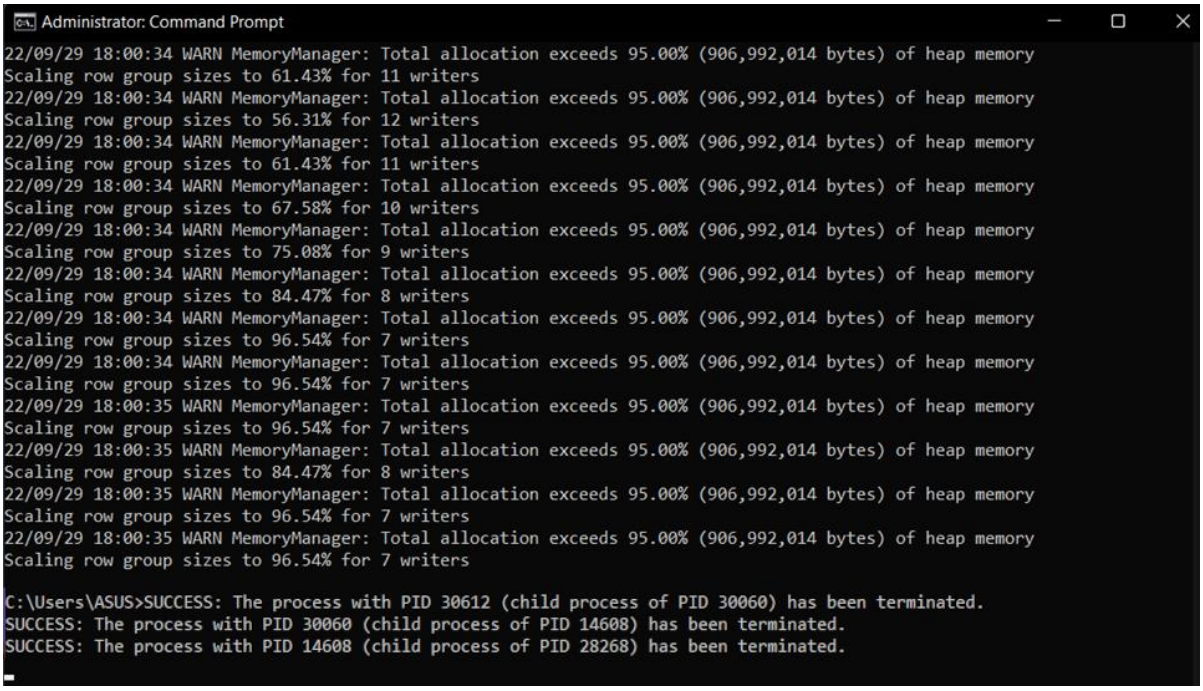
To run the program through Windows Command Prompt terminal, open a command prompt and enter the following command:

```
python C:\Users\ASUS\Documents\M2-Assignment-1\cassandra2hive.py
```

The directory from which the migration will occur is *C:\Users\ASUS\Documents\M2-Assignment-1\cassandra2hive.py*.

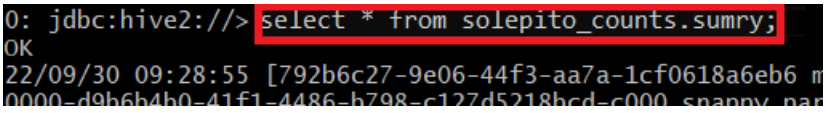


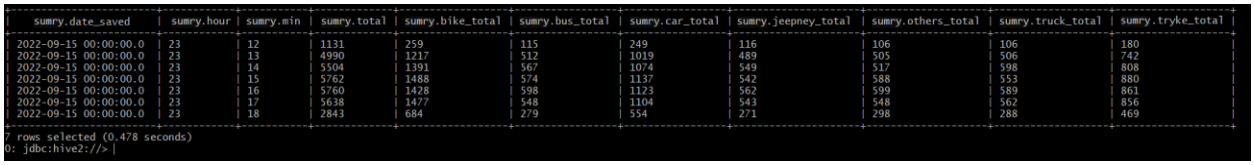
Since Spark had already been initialized at the beginning of the program, the migration process from Cassandra to Spark then to Hive will occur immediately after running the program, as indicated below.



Once the migration is complete, return to the Cygwin terminal, where Hive is already open, and run the following command to view the modified table with the migrated data from Cassandra:

```
select * from solepito_counts.sumry;
```





sumry.date_saved	sumry.hour	sumry.min	sumry.total	sumry.bike_total	sumry.bus_total	sumry.car_total	sumry.jeepney_total	sumry.others_total	sumry.truck_total	sumry.tryke_total
2022-09-15 00:00:00.0	23	12	1131	259	115	249	116	106	106	180
2022-09-15 00:00:00.0	23	13	4990	1217	512	1019	489	505	506	742
2022-09-15 00:00:00.0	23	14	5504	1391	567	1074	549	517	598	808
2022-09-15 00:00:00.0	23	15	5762	1488	574	1137	542	588	553	880
2022-09-15 00:00:00.0	23	16	5760	1428	598	1123	562	599	589	861
2022-09-15 00:00:00.0	23	17	5638	1477	548	1104	543	548	562	856
2022-09-15 00:00:00.0	23	18	2843	684	279	554	271	298	288	469

7 rows selected (0.478 seconds)
0: jdbc:hive2://>

REFERENCES

Apache Hive 3.0.0 Installation on Windows 10 Step by Step Guide. (2021, December 24). Hadoop, Hive & HBase. Retrieved September 27, 2022, from https://kontext.tech/article/291/apache-hive-300-installation-on-windows-10-step-by-step-guide

Fadlallah, H. (2021a, June 25). *Installing Hadoop 3.1.0 multi-node cluster on Ubuntu 16.04 Step by Step*. Medium. Retrieved September 27, 2022, from https://towardsdatascience.com/installing-hadoop-3-1-0-multi-node-cluster-on-ubuntu-16-04-step-by-step-8d1954b31505

Fadlallah, H. (2021b, December 14). *Installing Apache Hive 3.1.2 on Windows 10 - Towards Data Science*. Medium. Retrieved September 27, 2022, from https://towardsdatascience.com/installing-apache-hive-3-1-2-on-windows-10-70669ce79c79

Fadlallah, H. (2021c, December 14). *Installing Hadoop 3.2.1 Single node cluster on Windows 10*. Medium. Retrieved September 27, 2022, from https://towardsdatascience.com/installing-hadoop-3-2-1-single-node-cluster-on-windows-10-ac258dd48aef

Install Hadoop 3.2.1 on Windows 10 Step by Step Guide. (2022, September 6). Hadoop, Hive & HBase. Retrieved September 27, 2022, from https://kontext.tech/article/377/latest-hadoop-321-installation-on-windows-10-step-by-step-guide

Jarosciak, J. (2017, February 2). *How to install a hadoop single node cluster on windows 10*. Retrieved September 27, 2022, from https://www.joe0.com/2017/02/02/how-to-install-a-hadoop-single-node-cluster-on-windows-10/

[Jonathan MacDonald]. (2019, April 19). *How to install Cygwin on Windows 10 [Video]*. YouTube. Retrieved September 28, 2022, from https://www.youtube.com/watch?v=QonIPpKodCw

MacDonald, J. (2019, April 19). *How to install Cygwin on Windows 10* [Video]. YouTube. Retrieved September 27, 2022, from https://www.youtube.com/watch?v=QonIPpKodCw&feature=youtu.be

[mrunmayee kulkarni]. (2018, February 6). *EASY HIVE INSTALLATION ON WINDOWS* [Video]. YouTube. Retrieved September 27, 2022, from https://www.youtube.com/watch?v=npyRXkMhrgk

Stack Overflow - Where Developers Learn, Share, & Build Careers. (n.d.). Stack Overflow. Retrieved September 27, 2022, from https://stackoverflow.com/