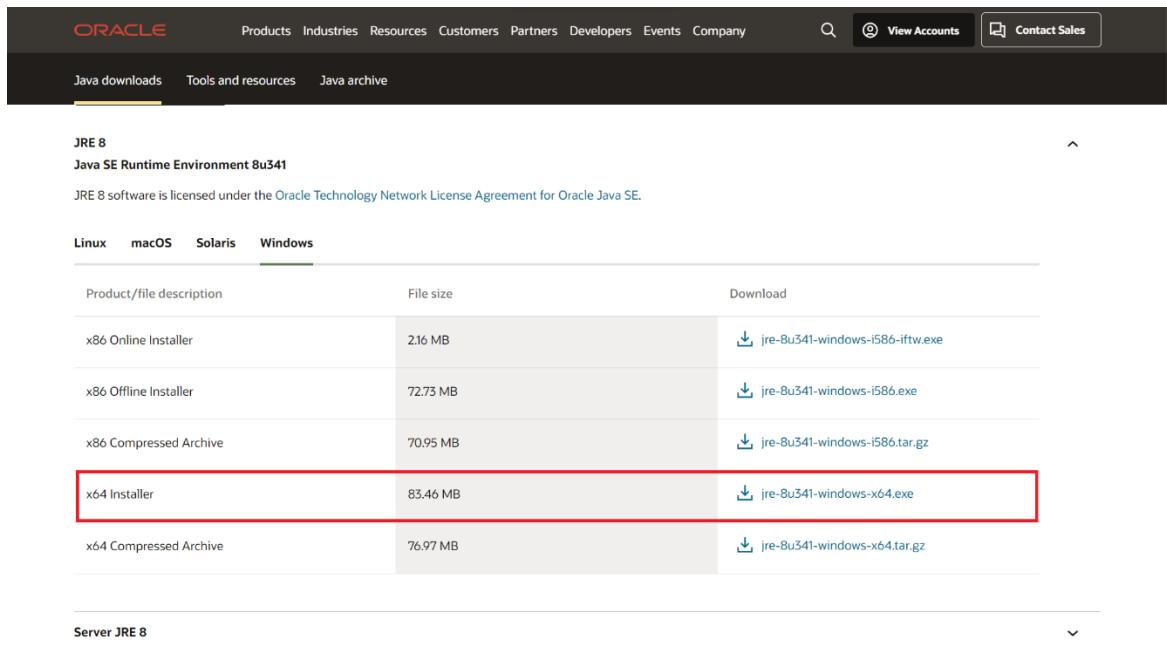


M2 Assignment 1 (Migration Process)

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>.



The screenshot shows the Oracle Java downloads page. The navigation bar includes links for Products, Industries, Resources, Customers, Partners, Developers, Events, Company, a search bar, and account-related buttons. Below the navigation bar, there are three main categories: Java downloads, Tools and resources, and Java archive. The Java downloads category is selected. Under the Java SE Runtime Environment 8u341 section, there are tabs for Linux, macOS, Solaris, and Windows. The Windows tab is selected. A table lists download options:

Product/file description	File size	Download
x86 Online Installer	2.16 MB	jre-8u341-windows-i586-iftw.exe
x86 Offline Installer	72.73 MB	jre-8u341-windows-i586.exe
x86 Compressed Archive	70.95 MB	jre-8u341-windows-i586.tar.gz
x64 Installer	83.46 MB	jre-8u341-windows-x64.exe
x64 Compressed Archive	76.97 MB	jre-8u341-windows-x64.tar.gz

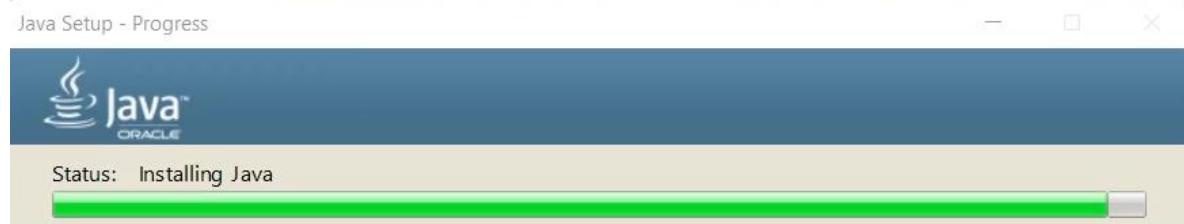
A red box highlights the x64 Installer row.

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.



#1 Development Platform

ORACLE



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.

A screenshot of the Oracle Java downloads page. The header includes the Oracle logo and navigation links for Products, Industries, Resources, Customers, Partners, Developers, Events, and Company. There are also "View Accounts" and "Contact Sales" buttons. The main menu has options for Java downloads, Tools and resources, and Java archive. The "Java downloads" option is highlighted. Below this, the "Java SE Development Kit 8u341" section is shown. It mentions that Java SE subscribers will receive JDK 8 updates until at least December of 2030. It notes that the Oracle JDK 8 license changed in April 2019. The Oracle Technology Network License Agreement for Oracle Java SE is described as substantially different from prior Oracle JDK 8 licenses. It also mentions that commercial license and support are available for a low cost with Java SE Subscription. The JDK 8 software is licensed under the Oracle Technology Network License Agreement for Oracle Java SE. A "JDK 8u341 checksum" link is provided. Below this, there are tabs for Linux, macOS, Solaris, and Windows, with "Windows" being the active tab. A table lists the download links for each version: "x86 Installer" (159.66 MB) and "x64 Installer" (173.16 MB). The "x64 Installer" row is highlighted with a red border. At the bottom, there is a "Documentation Download" button.

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.

You must accept the [Oracle Technology Network License Agreement for Oracle Java SE](#) to download this software. X

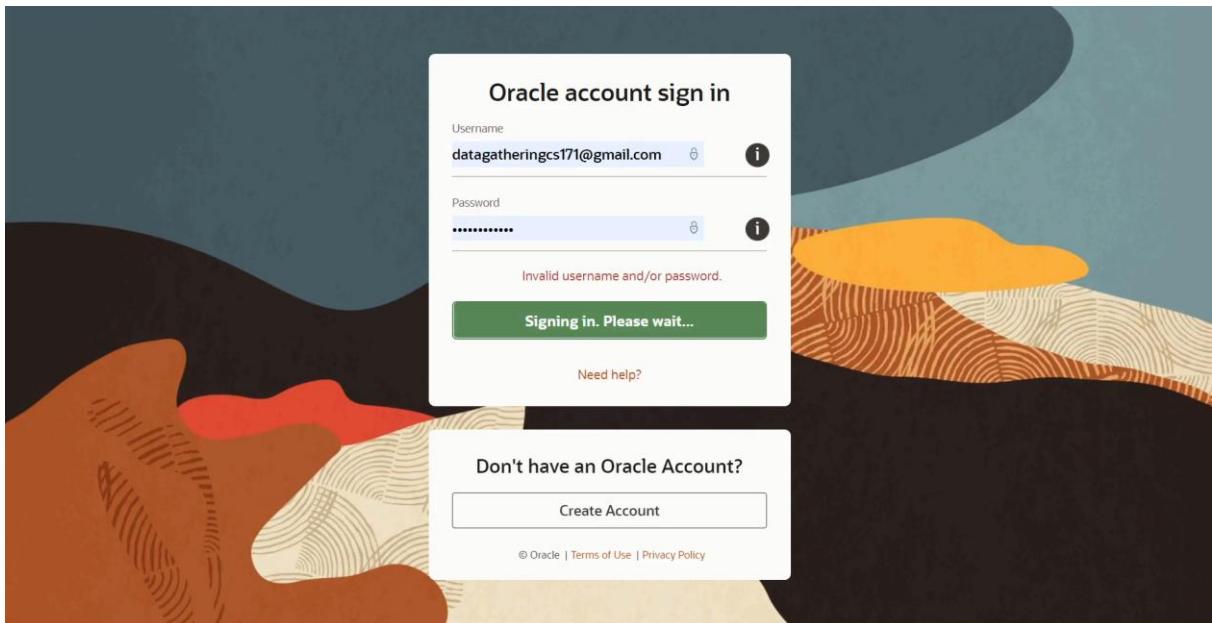
I reviewed and accept the Oracle Technology Network License Agreement for Oracle Java SE

Required

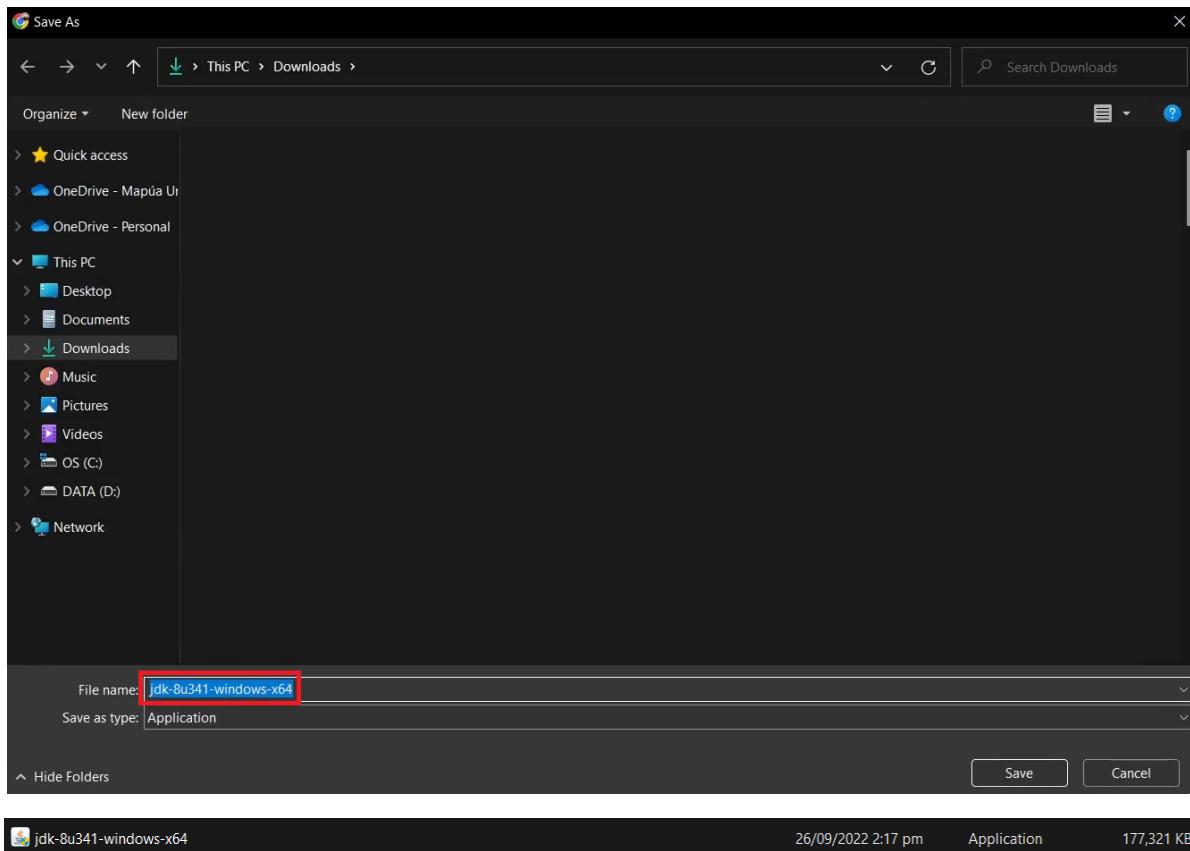
You will be redirected to the login screen in order to download the file.

[Download jdk-8u341-windows-x64.exe](#) 

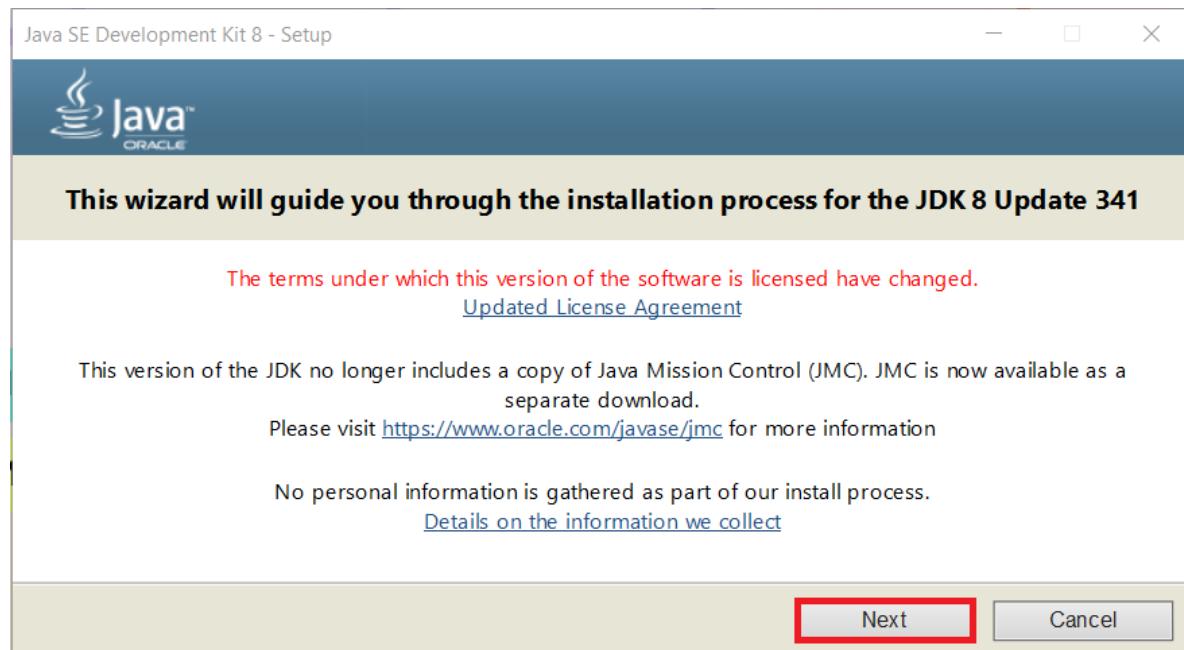
Clicking the download button will redirect the user to the Oracle account sign in window, where they will be prompted to enter a username and password to download the JDK installer. Since our group has previously created a joint Oracle account for situations like these, we used that account to download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



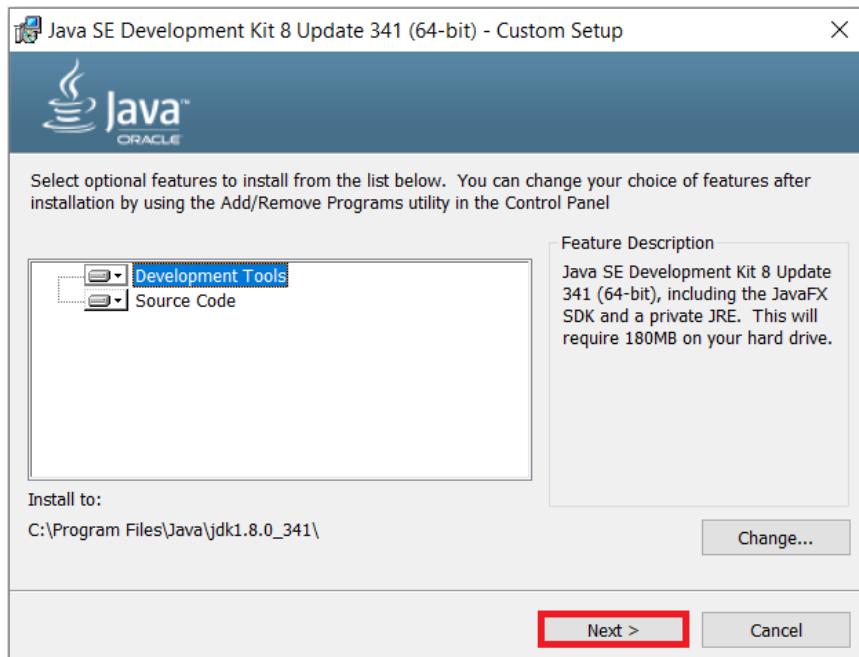
After signing in, the download should begin promptly, as indicated below.



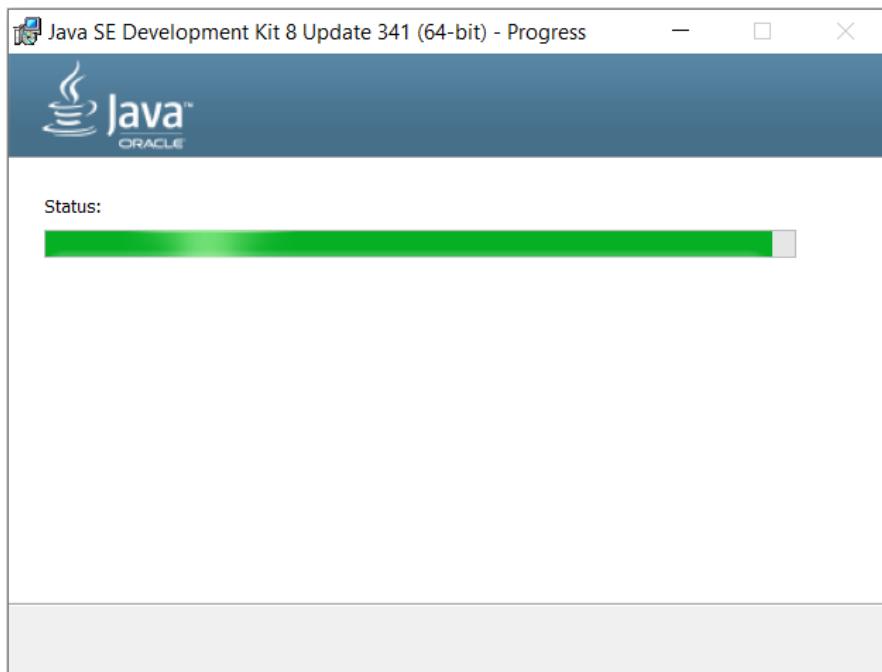
Double click on this downloaded executable file. Click 'Next' once the initial setup screen below is shown.



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.

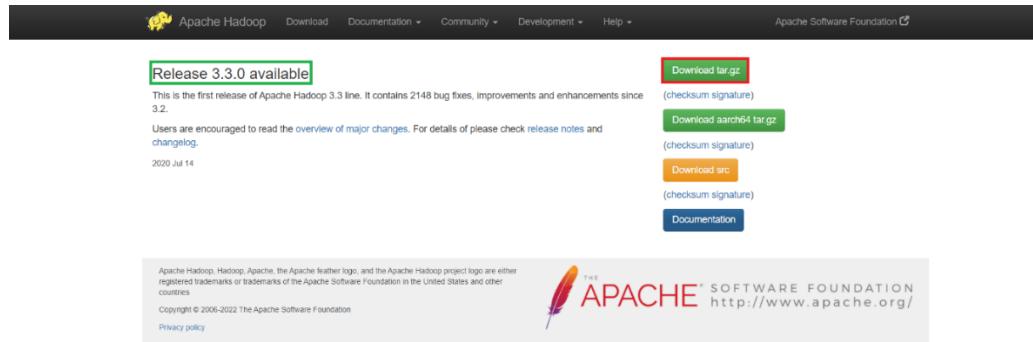


II. Hadoop Installation

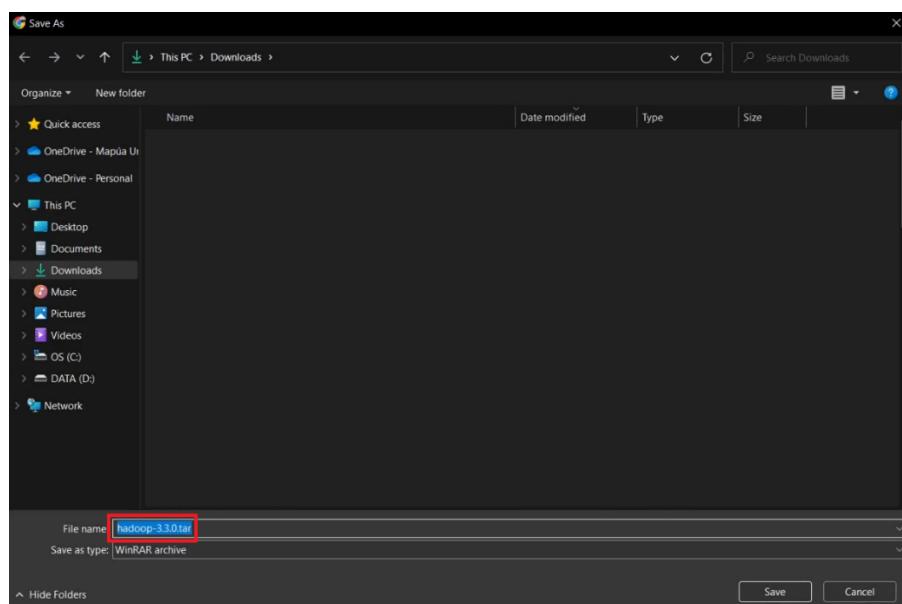
a. Download Hadoop binaries (winutils)

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.

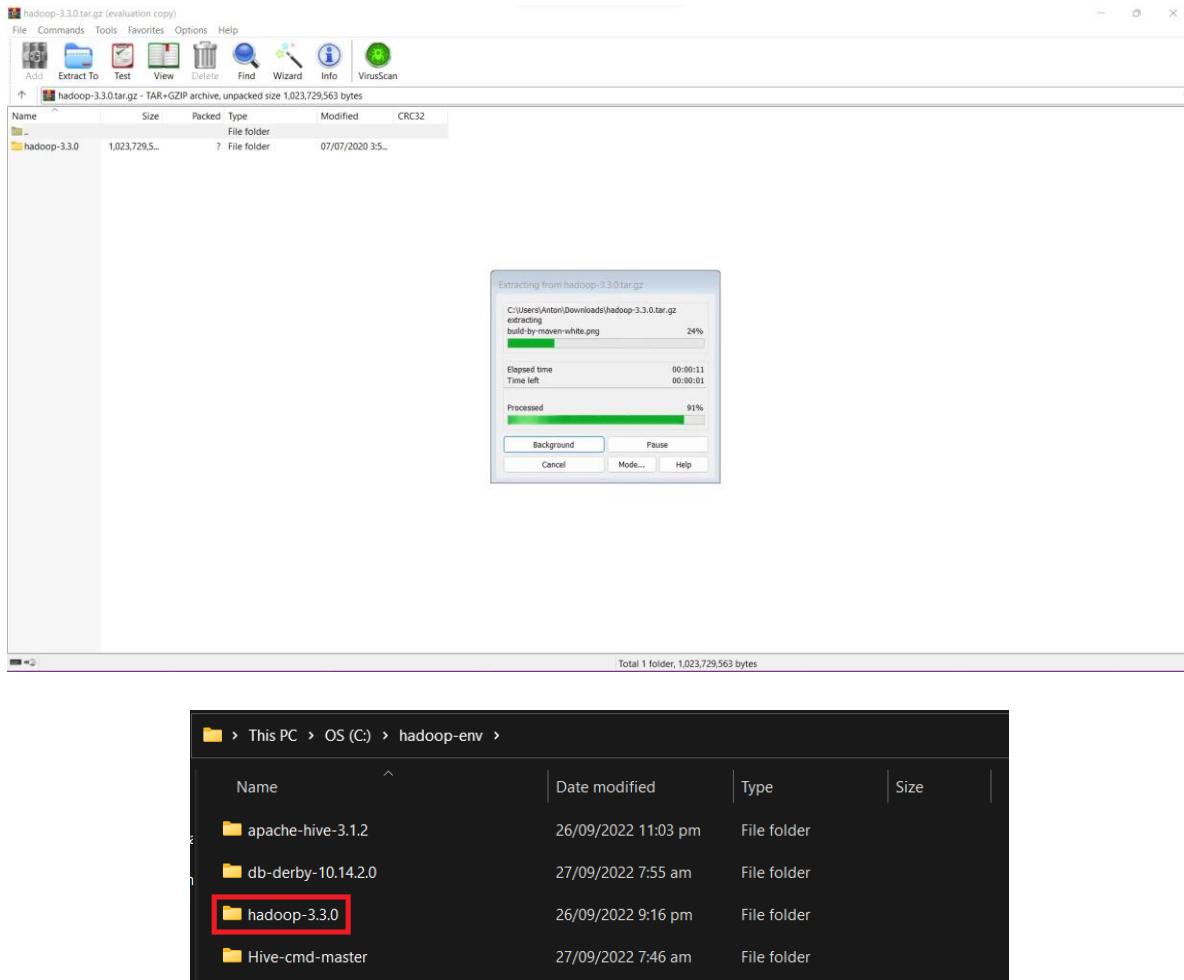
Now that the required JRE and JDK installations have been completed, Hadoop can already be downloaded by first clicking on the “Download tar.gz” button on this link: <https://hadoop.apache.org/release/3.3.0.html>. Clicking this button will download the zipped file into the user’s specified directory.



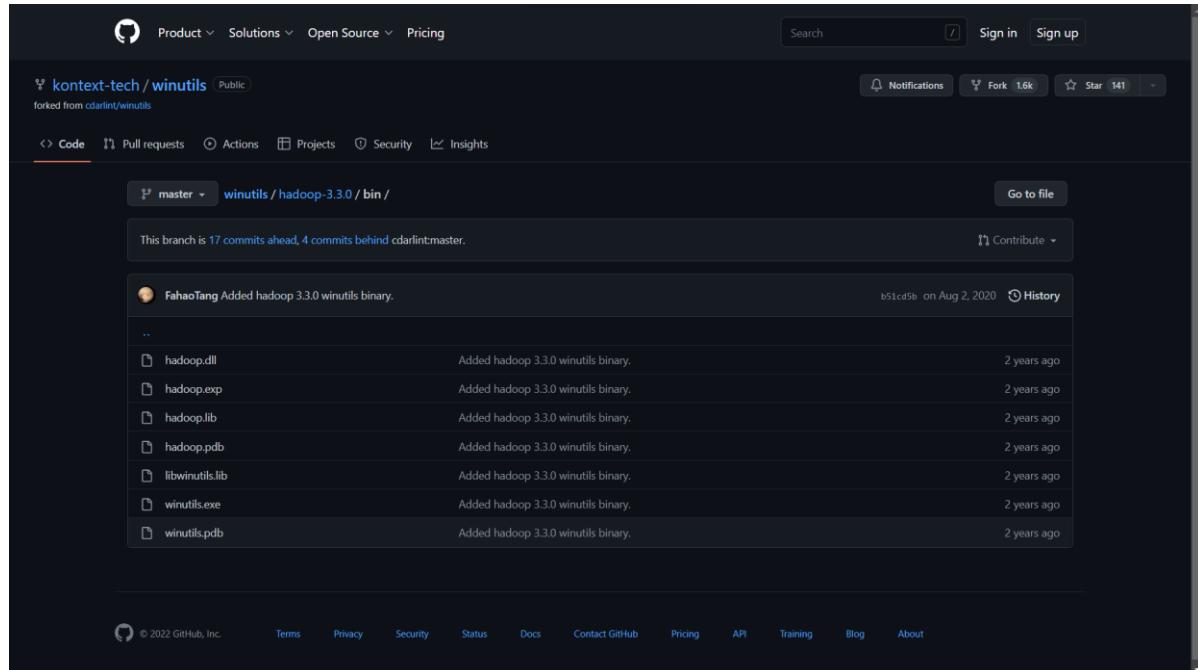
<https://hadoop.apache.org/docs/3.3.0>



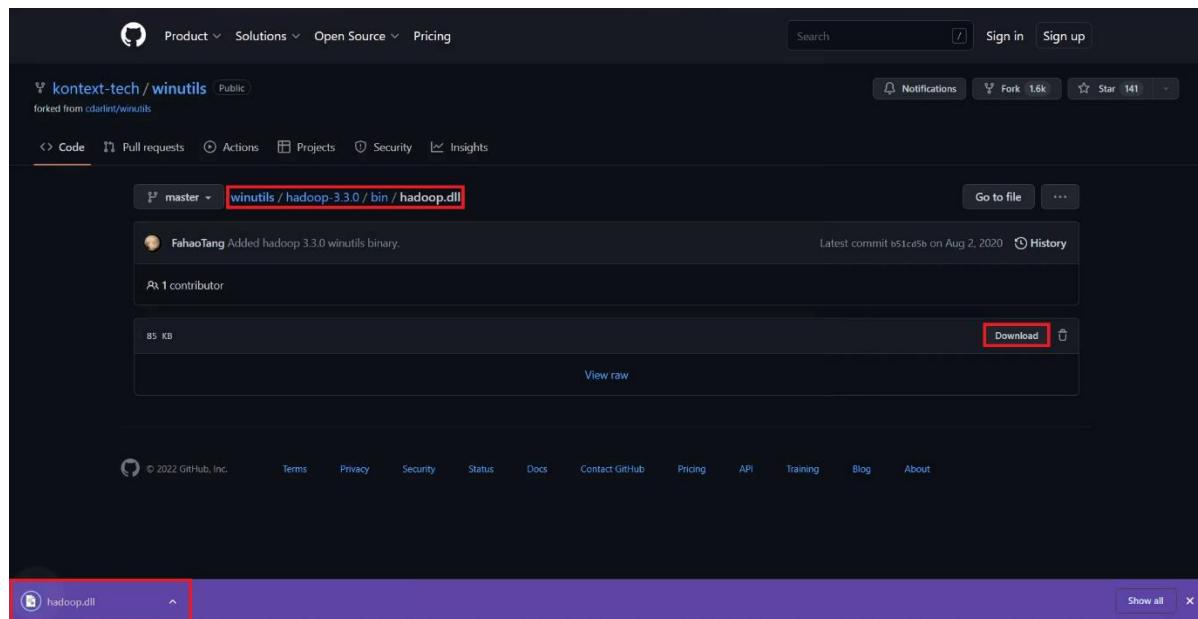
Once the file finishes downloading, unzip the file by extracting it using the WinRAR Archiver. (Other unzipping tools may also be used.) Then, move the unzipped file to the *hadoop-env* folder that was created prior to downloading the Hadoop package.



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*.

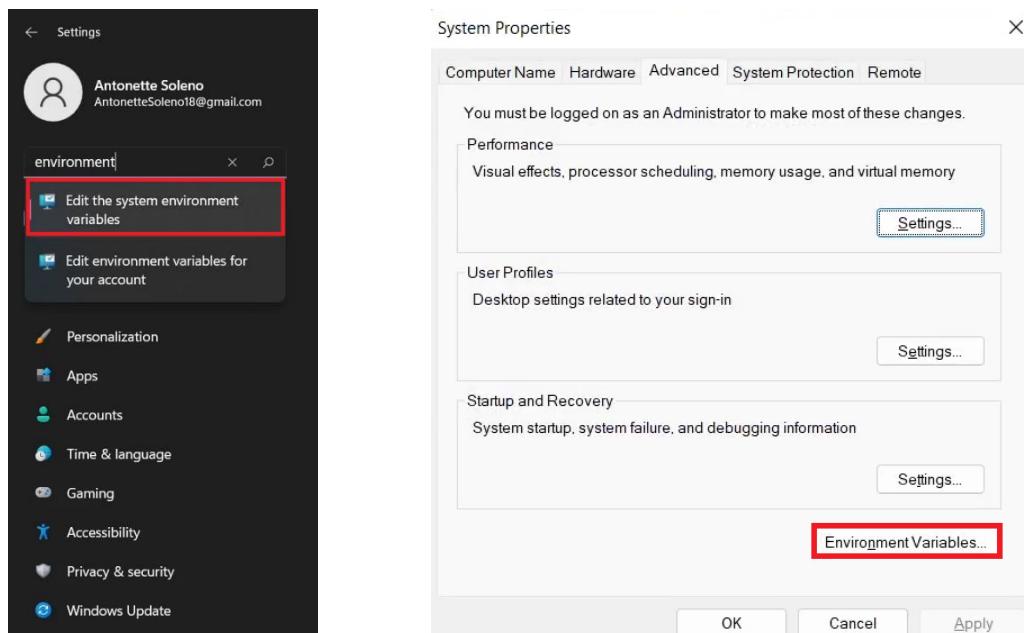


Once all files have been downloaded, copy these files into the ‘hadoop-3.3.0\bin’ directory, as indicated.

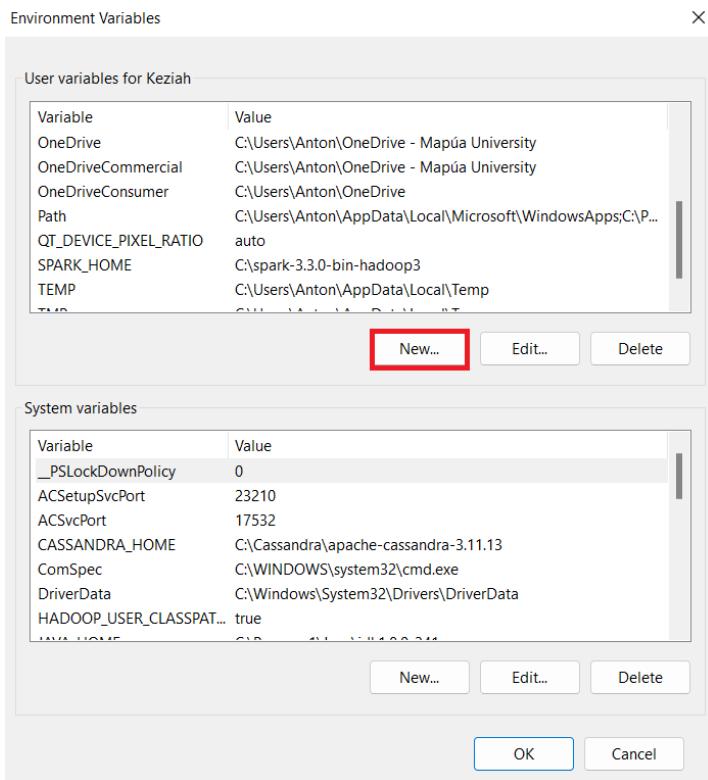
Name	Date modified	Type	Size
container-executor	07/07/2020 3:33 am	File	785 KB
hadoop	07/07/2020 2:46 am	File	9 KB
hadoop	07/07/2020 2:46 am	Windows Comma...	12 KB
hadoop.dll	26/09/2022 9:05 pm	Application extens...	85 KB
hadoop.exp	26/09/2022 9:05 pm	Exports Library File	20 KB
hadoop.lib	26/09/2022 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

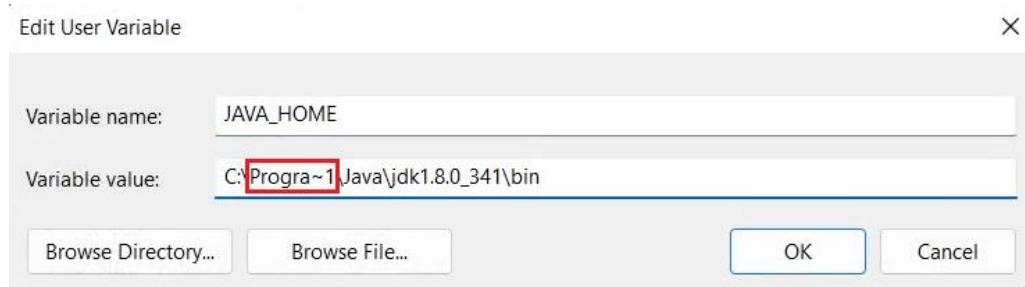
After downloading Hadoop’s package and binaries, the environment variables to define Hadoop and Java must now be configured. To do so, go to Settings and type “environment.” When *Edit the system environment variables* appears, select it. The *System Properties* window should appear. Select *Environmental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

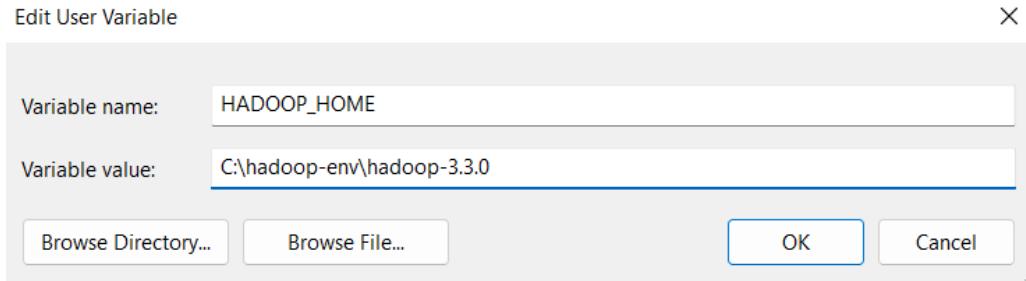


The *Environment Variables* window should appear after. Now, click “New...” under *User variables* and create two variables: JAVA_HOME and HADOOP_HOME. JAVA_HOME should contain the path for the JDK installation, and HADOOP_HOME the path for the Hadoop 3.3.0 installation. Click “OK” after each variable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

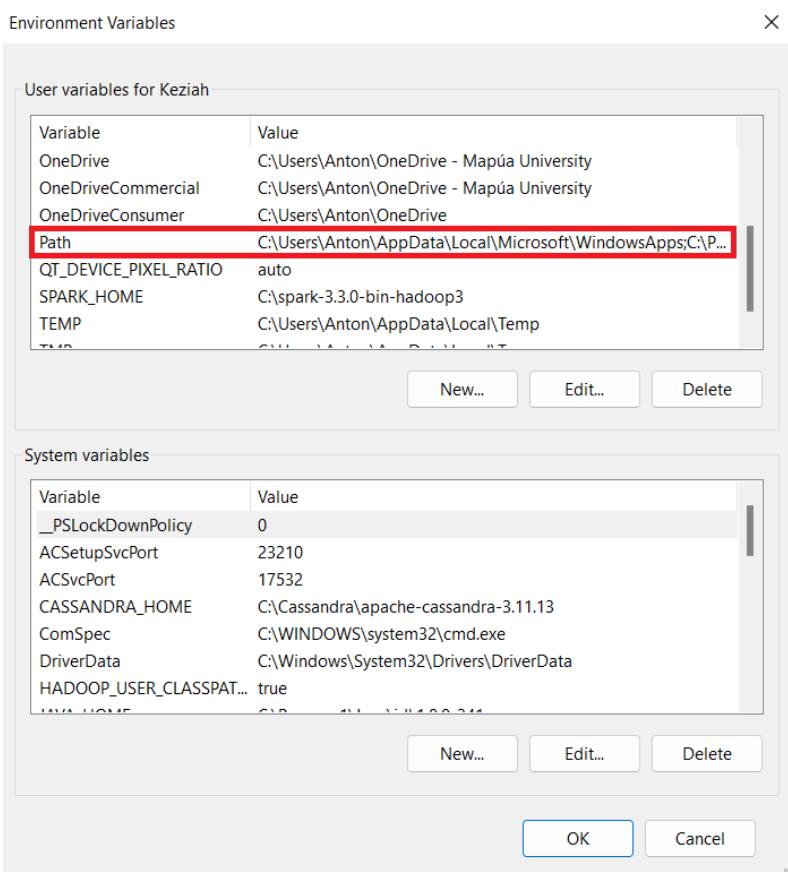


Note: The **Windows 8.3 path** was implemented while creating the path to JAVA_HOME to avoid future errors caused by white spaces in file naming.

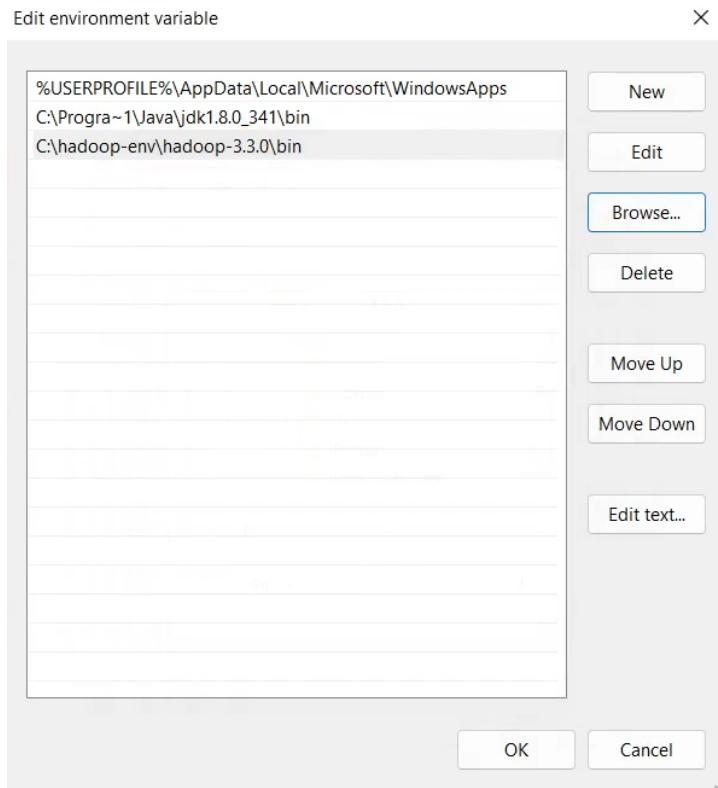




Under *User variables*, locate “Path” and double click.



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.

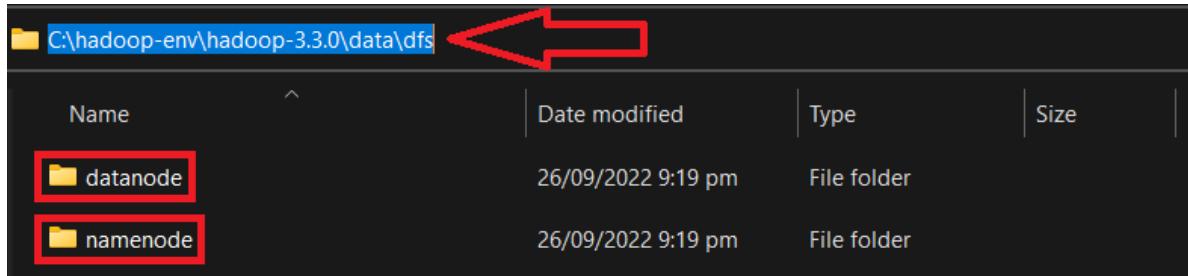
```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

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.



Name	Date modified	Type	Size
datanode	26/09/2022 9:19 pm	File folder	
namenode	26/09/2022 9:19 pm	File 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*.



```
1<?xml version="1.0" encoding="UTF-8"?>
2<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3<!!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15
16-->
17<!-- Put site-specific property overrides in this file. -->
18
19<configuration>
20<property>
21<name>dfs.replication</name>
22<value>1</value>
23</property>
24<property>
25<name>dfs.namenode.name.dir</name>
26<value>file:///C:/hadoop-env/hadoop-3.3.0/data/dfs/namenode</value>
27</property>
28<property>
29<name>dfs.datanode.data.dir</name>
30<value>file:///C:/hadoop-env/hadoop-3.3.0/data/dfs/datanode</value>
31</property>
32</configuration>
```

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*.

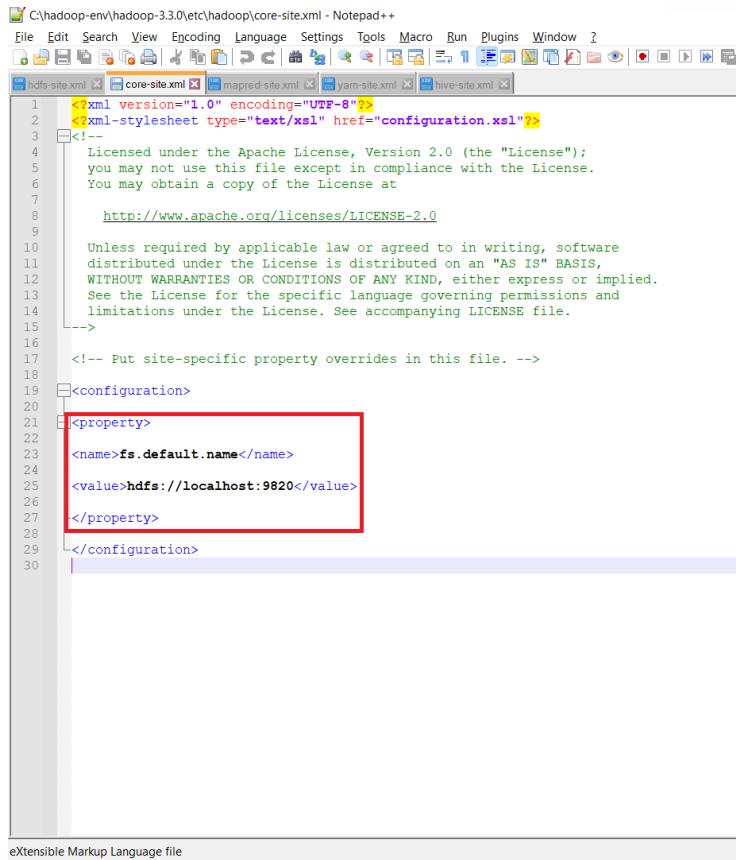
```
<property>

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

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

</property>
```

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



```
C:\hadoop-env\hadoop-3.3.0\etc\hadoop\core-site.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
hdfs-site.xml core-site.xml mapred-site.xml yarn-site.xml hive-site.xml

1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>fs.default.name</name>
22     <value>hdfs://localhost:9820</value>
23   </property>
24 </configuration>
25
26
27
28
29
30
```

eXtensible Markup Language file

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

```
<property>

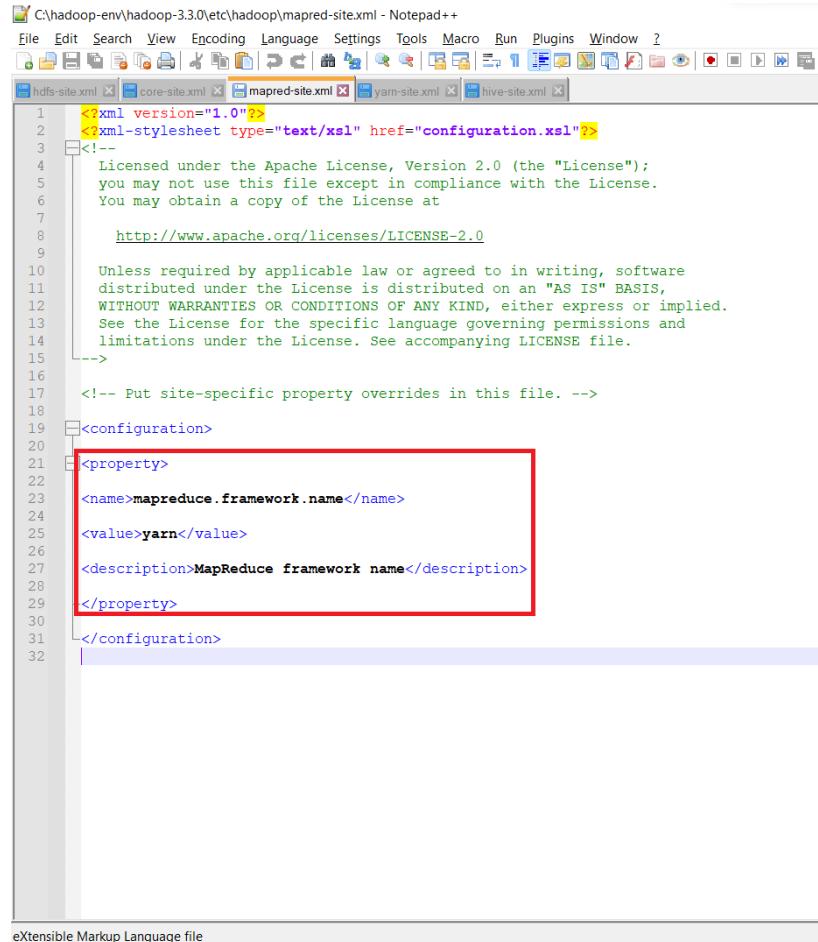
<name>mapreduce.framework.name</name>

<value>yarn</value>

<description>MapReduce framework name</description>

</property>
```

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



```
C:\hadoop-env\hadoop-3.3.0\etc\hadoop\mapred-site.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
hdbsite.xml core-site.xml mapred-site.xml yarn-site.xml hive-site.xml
1  <?xml version="1.0"?>
2  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3  <!--
4      Licensed under the Apache License, Version 2.0 (the "License");
5      you may not use this file except in compliance with the License.
6      You may obtain a copy of the License at
7
8          http://www.apache.org/licenses/LICENSE-2.0
9
10     Unless required by applicable law or agreed to in writing, software
11     distributed under the License is distributed on an "AS IS" BASIS,
12     WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13     See the License for the specific language governing permissions and
14     limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20
21 <property>
22
23 <name>mapreduce.framework.name</name>
24
25 <value>yarn</value>
26
27 <description>MapReduce framework name</description>
28
29 </property>
30
31 </configuration>
32
```

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

```
<property>

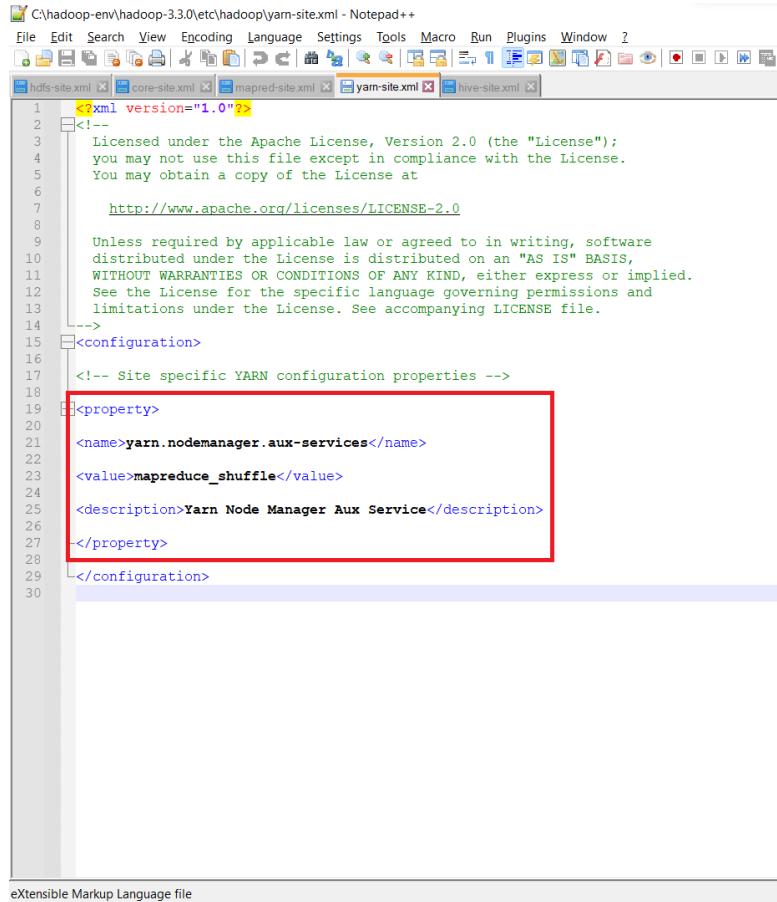
<name>yarn.nodemanager.aux-services</name>

<value>mapreduce_shuffle</value>

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

</property>
```

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



```
C:\hadoop-env\hadoop-3.3.0\etc\hadoop\yarn-site.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
hdfs-site.xml core-site.xml mapred-site.xml yarn-site.xml hive-site.xml
1 <?xml version="1.0"?>
2 <!--
3   Licensed under the Apache License, Version 2.0 (the "License");
4   you may not use this file except in compliance with the License.
5   You may obtain a copy of the License at
6
7     http://www.apache.org/licenses/LICENSE-2.0
8
9   Unless required by applicable law or agreed to in writing, software
10  distributed under the License is distributed on an "AS IS" BASIS,
11  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12  See the License for the specific language governing permissions and
13  limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16   <!-- Site specific YARN configuration properties -->
17 <property>
18   <name>yarn.nodemanager.aux-services</name>
19   <value>mapreduce_shuffle</value>
20   <description>Yarn Node Manager Aux Service</description>
21 </property>
22 </configuration>
23
24
25
26
27
28
29
30
```

eXtensible Markup Language file

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
```



```
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> 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:

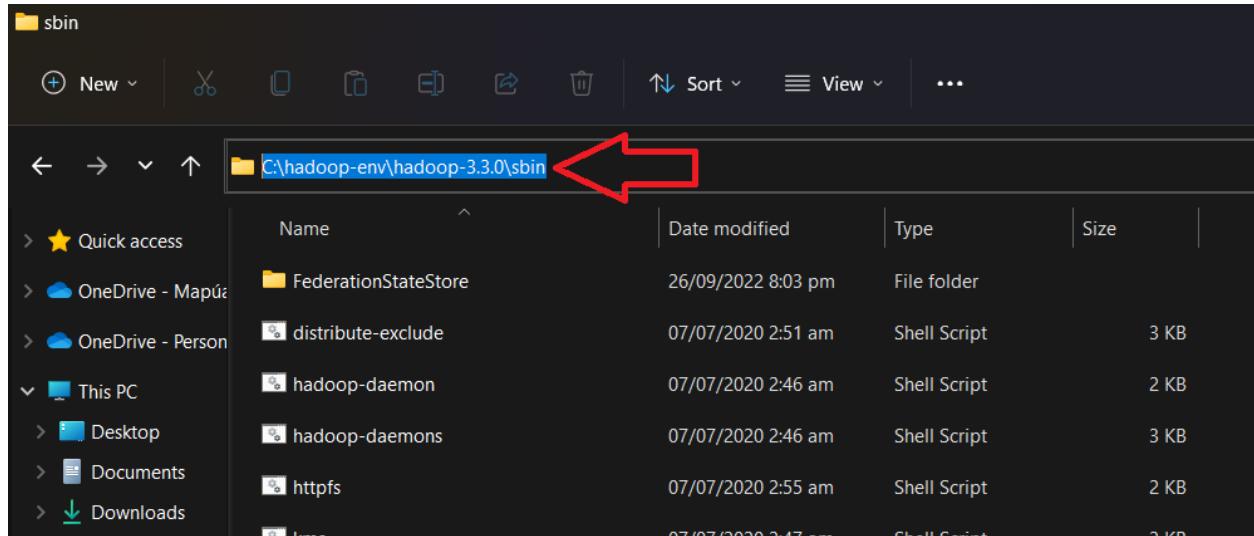
.\start-dfs.cmd

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

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

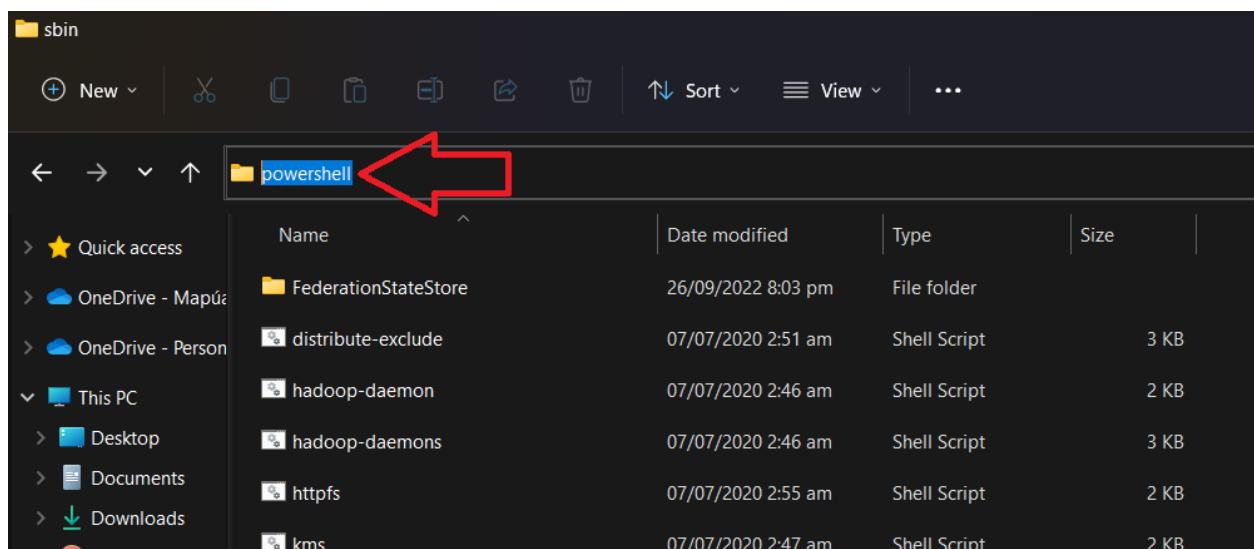
PS C:\hadoop-env\hadoop-3.3.0\sbin> .\start-dfs.cmd
```

*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.



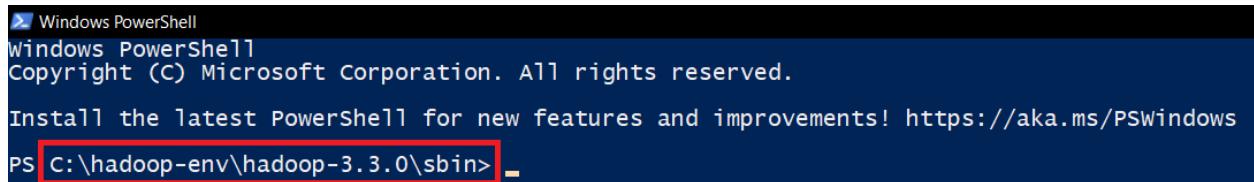
File Explorer showing the contents of the 'sbin' folder:

Name	Date modified	Type	Size
FederationStateStore	26/09/2022 8:03 pm	File folder	
distribute-exclude	07/07/2020 2:51 am	Shell Script	3 KB
hadoop-daemon	07/07/2020 2:46 am	Shell Script	2 KB
hadoop-daemons	07/07/2020 2:46 am	Shell Script	3 KB
httpfs	07/07/2020 2:55 am	Shell Script	2 KB
kms	07/07/2020 2:47 am	Shell Script	2 KB



File Explorer showing the contents of the 'sbin' folder:

Name	Date modified	Type	Size
FederationStateStore	26/09/2022 8:03 pm	File folder	
distribute-exclude	07/07/2020 2:51 am	Shell Script	3 KB
hadoop-daemon	07/07/2020 2:46 am	Shell Script	2 KB
hadoop-daemons	07/07/2020 2:46 am	Shell Script	3 KB
httpfs	07/07/2020 2:55 am	Shell Script	2 KB
kms	07/07/2020 2:47 am	Shell Script	2 KB



Windows PowerShell window:

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

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

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

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`

```
PS C:\hadoop-env\hadoop-3.3.0\sbin> ./start-yarn.cmd
starting yarn daemons
PS C:\hadoop-env\hadoop-3.3.0\sbin>
```

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, data node, and yarn services of Hadoop.

The screenshot shows the 'Overview' page of the Hadoop Web UI. At the top, there's a navigation bar with tabs: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The 'Overview' tab is active. Below the navigation bar, the title is 'Overview 'localhost:9820' (✓active)'. There are two tables of information:

Started:	Mon Sep 26 21:28:44 +0800 2022
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 02:44:00 +0800 2020 by brahma from branch-3.3.0
Cluster ID:	CID-9dd435f0-e892-45f9-8599-9731f6ff6c8
Block Pool ID:	BP-48685550-192.168.56.1-1664198196462

Configured Capacity:	217.84 GB
Configured Remote Capacity:	0 B
DFS Used:	320 B (0%)
Non DFS Used:	215.62 GB
DFS Remaining:	2.22 GB (1.02%)
Block Pool Used:	320 B (0%)

Name Node (web page)

The screenshot shows the 'DataNode on' page of the Hadoop Web UI. At the top, there's a navigation bar with tabs: Hadoop, Overview, and Utilities. The 'Overview' tab is active. Below the navigation bar, there's a table with two rows:

Cluster ID:	CID-9dd435f0-e892-45f9-8599-9731f6ff6c8
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9820	BP-48685550-192.168.56.1-1664198196462	RUNNING	2s	2 minutes	0 B (128 MB)

Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
C:\hadoop-env\hadoop-3.3.0\data\dfs\datanode	DISK	320 B	707.61 MB	0 B	0 B	0

Data Node (web page)



All Applications

Cluster Metrics							
Nodes	Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total
0	0	0	0	0	0 B	0 B	0 B
Cluster Nodes Metrics							
Active Nodes		Decommissioning Nodes		Decommissioned Nodes		Lost Nodes	
0	0	0	0	0	0	0	1
Scheduler Metrics							
Scheduler Type		Scheduling Resource Type		Minimum Allocation			
Capacity Scheduler		<memory:1024, vCores:1>		<memory:8192, vCo			
Show 20 entries							
ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime
LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CP	Allocated VCores	Allocated Memory
No data available in table							
Showing 0 to 0 of 0 entries							

Yarn Service (web page)

III. Apache Derby Download

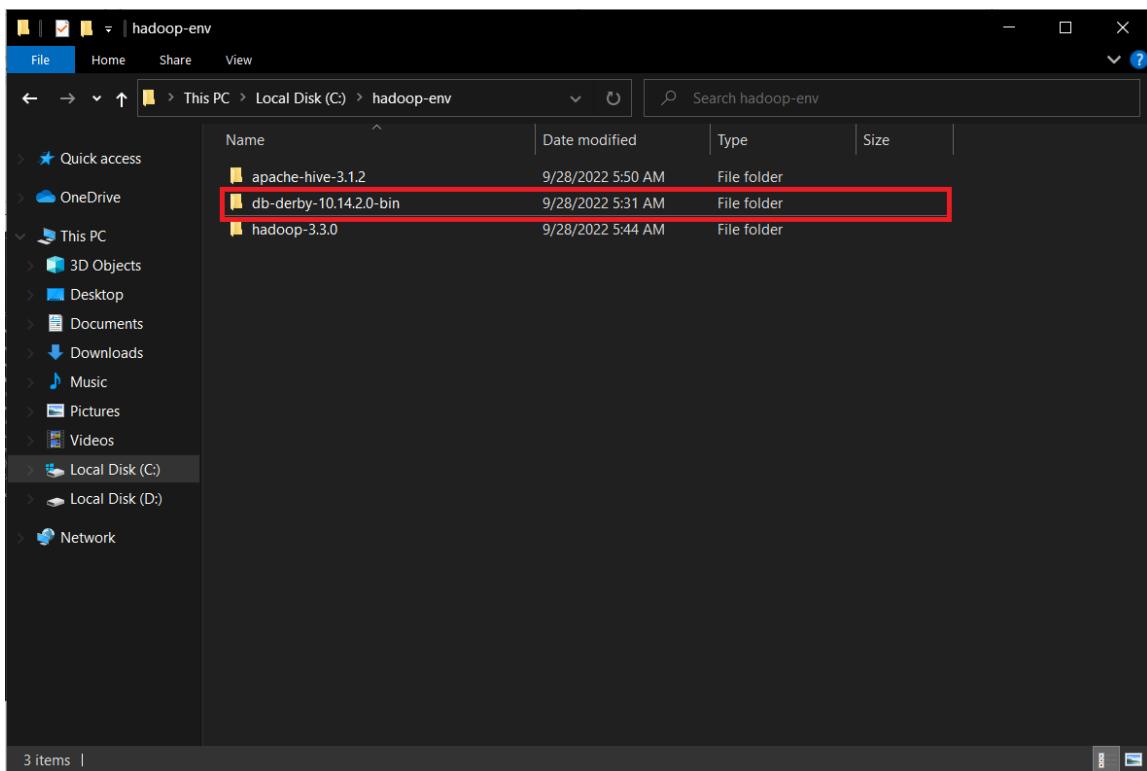
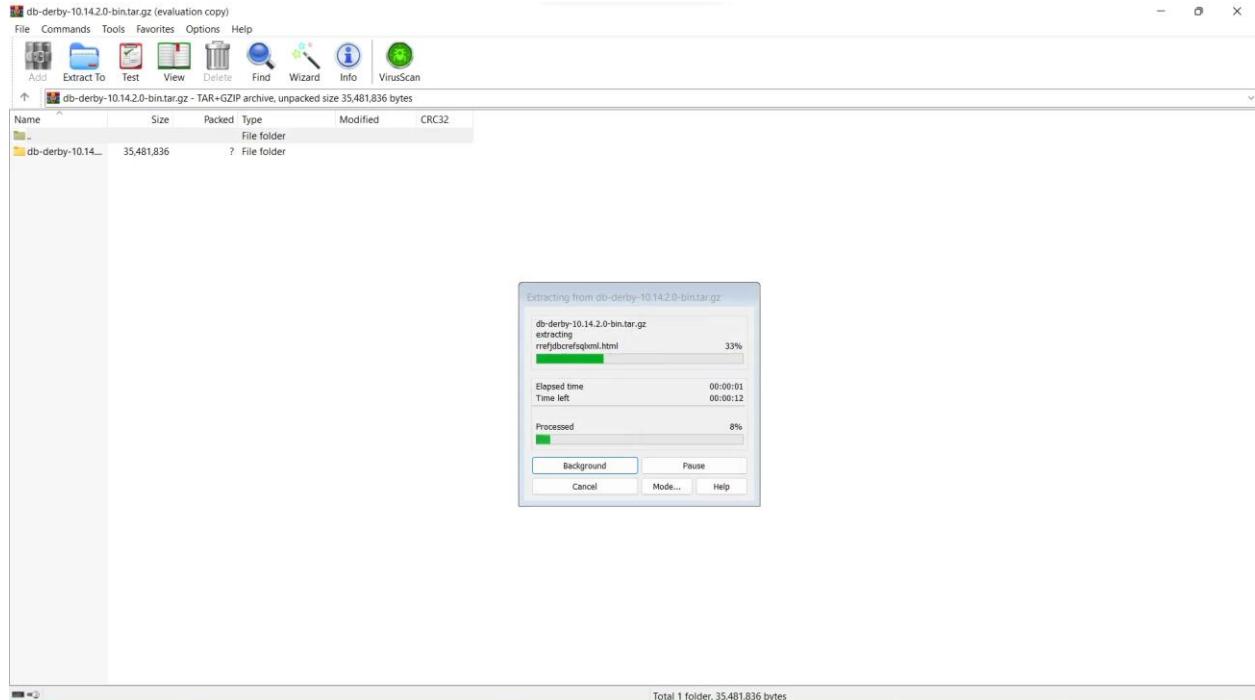
Another prerequisite of Apache Hive is Apache Derby, a relational database used to create Apache Hive's Metastore, where all Hive's metadata will be stored. Since Java 8 was the version previously installed for Hadoop, Apache Derby version 10.14.2.0 must be installed. This version can be directly downloaded 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.

The screenshot shows a Windows desktop environment. In the background, several application windows are open: 'Apache Hadoop', 'NameNode Information', 'DataNode Information', 'All Applications', and 'Apache Derby: Downloads'. The 'Apache Derby: Downloads' window is the active one, displaying the 'Apache Derby: Downloads' page. The page has a sidebar with links like 'Download Overview', 'The Apache Software Foundation', and a search bar. The main content area lists download links for different Java versions:

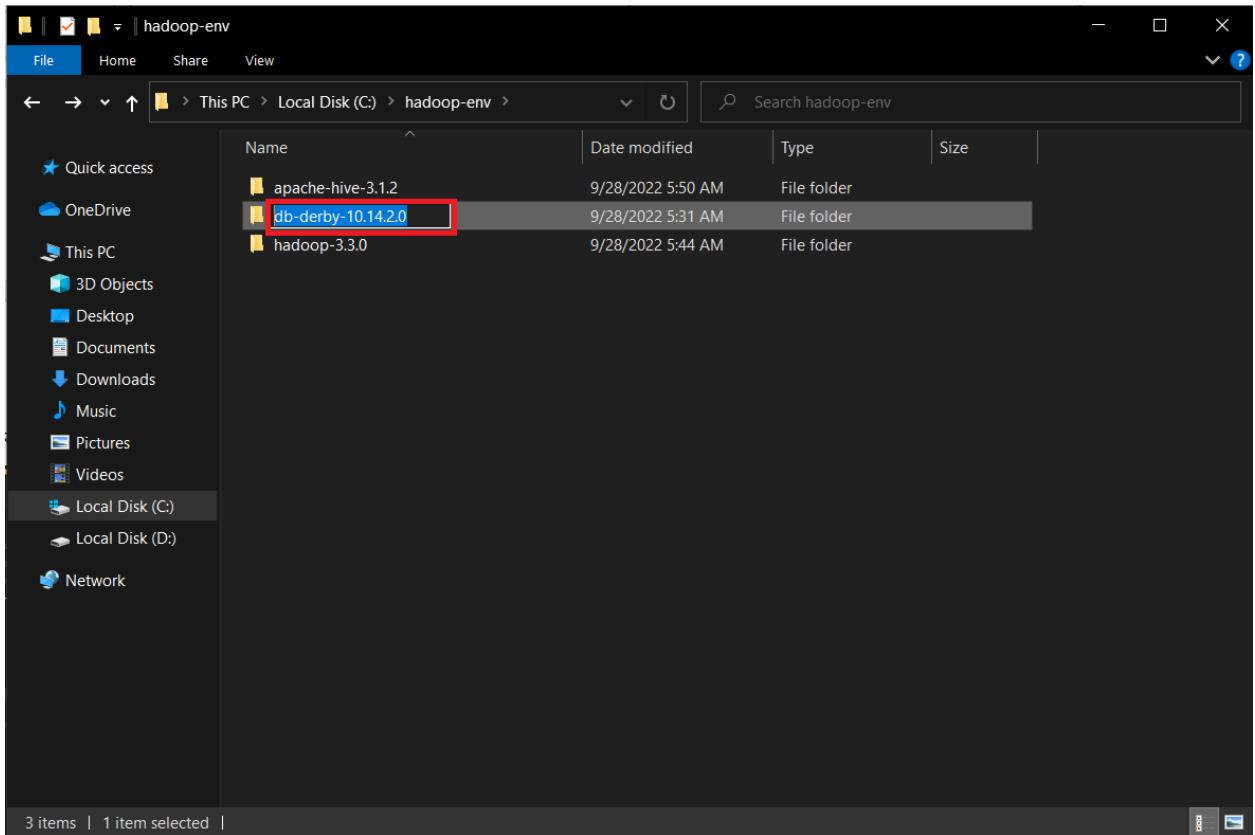
- For Java 17 and Higher:
 - [10.16.1.1](#) (May 19, 2022 / SVN 1901046)
- For Java 9 and Higher:
 - [10.15.2.0](#) (February 18, 2020 / SVN 1873585)
 - [10.15.1.3](#) (March 5, 2019 / SVN 1853019)
- For Java 8 and Higher:
 - [10.14.2.0](#) (May 3, 2018 / SVN 1828579)
 - [10.13.1.1](#) (October 25, 2016 / SVN 1766613)
- For Java 6 and Higher:
 - [10.12.1.1](#) (October 11, 2015 / SVN 1704137)
 - [10.11.1.1](#) (August 26, 2014 / SVN 1616546)
- For Java 1.4 and Higher:
 - [10.10.2.0](#) (April 15, 2014 / SVN 1582446)
 - [10.10.1.1](#) (April 15, 2013 / SVN 1458268)

In the foreground, a 'Save As' dialog box is open, showing the file path 'C:\Users\Mapúa\Downloads' and the file name 'db-derby-10.14.2.0-bin.tar'. The 'Save as type' dropdown is set to 'WinRAR archive'. At the bottom right of the dialog are 'Save' and 'Cancel' buttons.

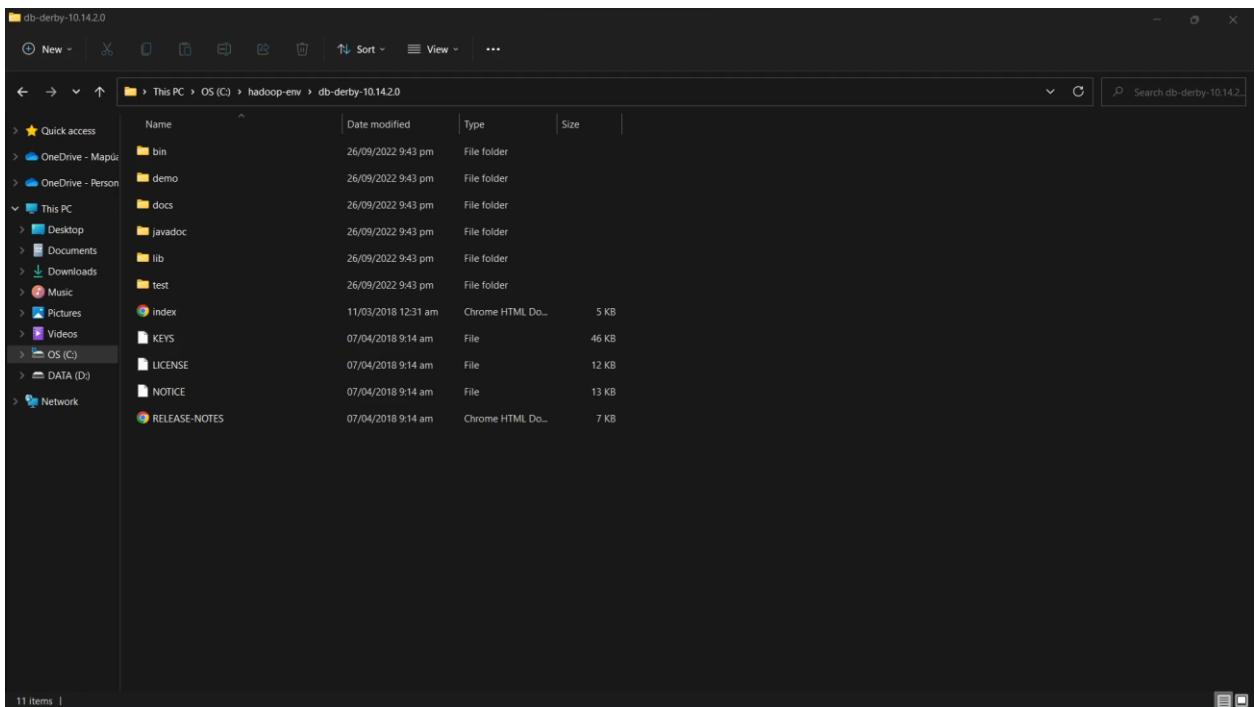
Locate the file in the downloads folder after it has downloaded. We must extract Derby using any zipping tool such as WinRAR Archiver. We extract the *db-derby-10.14.2.0-bin* file folder to the desired directory. Previously, we installed Hadoop within the "C:\hadoop-env" directory. Now, we will extract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

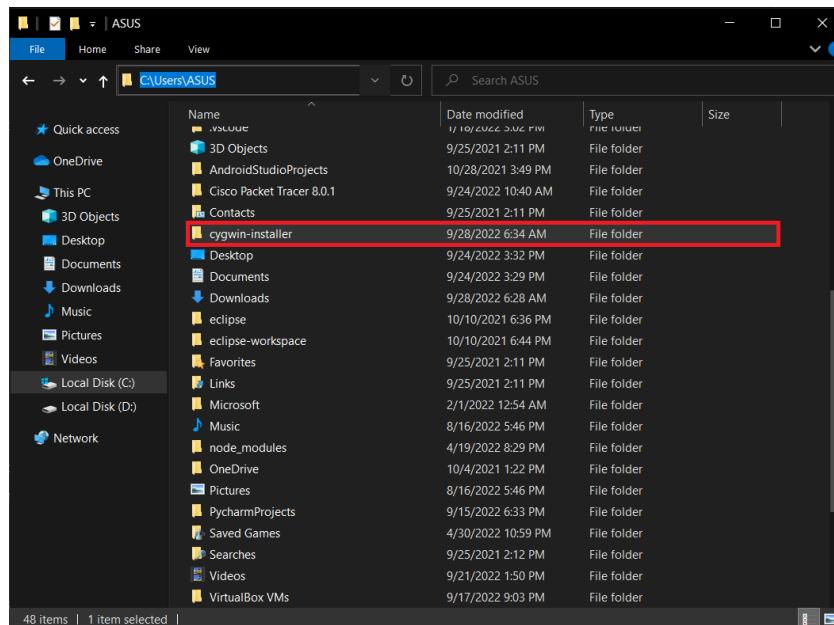


IV. Cygwin Installation

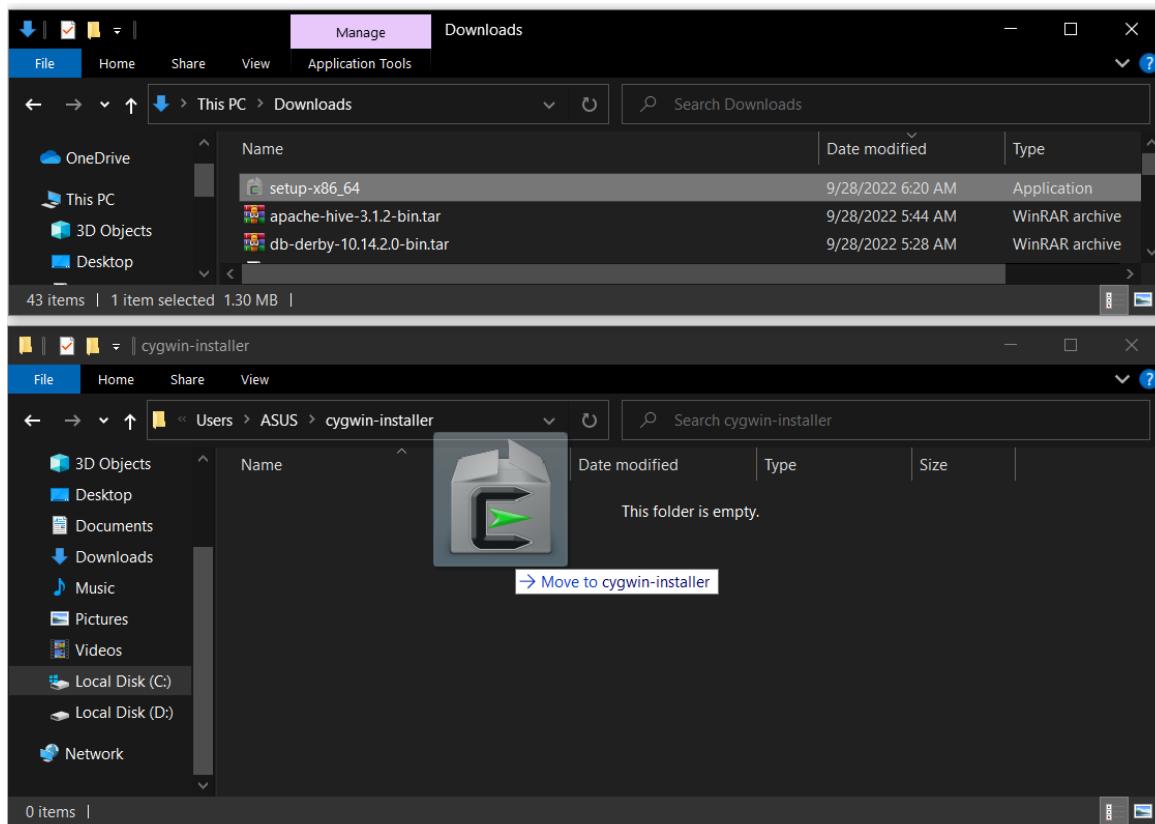
Unfortunately, there are some Hive tools that are compatible with Windows so we will need the Cygwin tool to run Linux commands. Head over to <https://cygwin.com/install.html> and click on *setup-x86_64.exe* to download the latest 64-bit version.

The screenshot shows a web browser window with the URL <https://cygwin.com/install.html>. The page title is "Cygwin". On the left, there's a sidebar with links like "Cygwin", "Install Cygwin", "Update Cygwin", "Search Packages", "Licensing Terms", "CygwinX", "Community", "Reporting Problems", "Mailing Lists", "Newsgroups", "IRC channels", "Gold Stars", "Mirror Sites", "Donations", "Documentation", "FAQ", "User's Guide", "API Reference", "Acronyms", "Contributing", "Snapshots", "Source in Git", "Cygwin Packages", "Cygwin Apps", and "Related Sites". The main content area has a heading "Installing and Updating Cygwin Packages". It includes sections for "Installing and Updating Cygwin for 64-bit versions of Windows", "General installation notes", and "Q: How do I add a package to my existing Cygwin installation?". Below these, there's a download link for "setup-x86_64.exe" with a file size of 1,339 KB and a timestamp of 9/28/2022 6:20 AM.

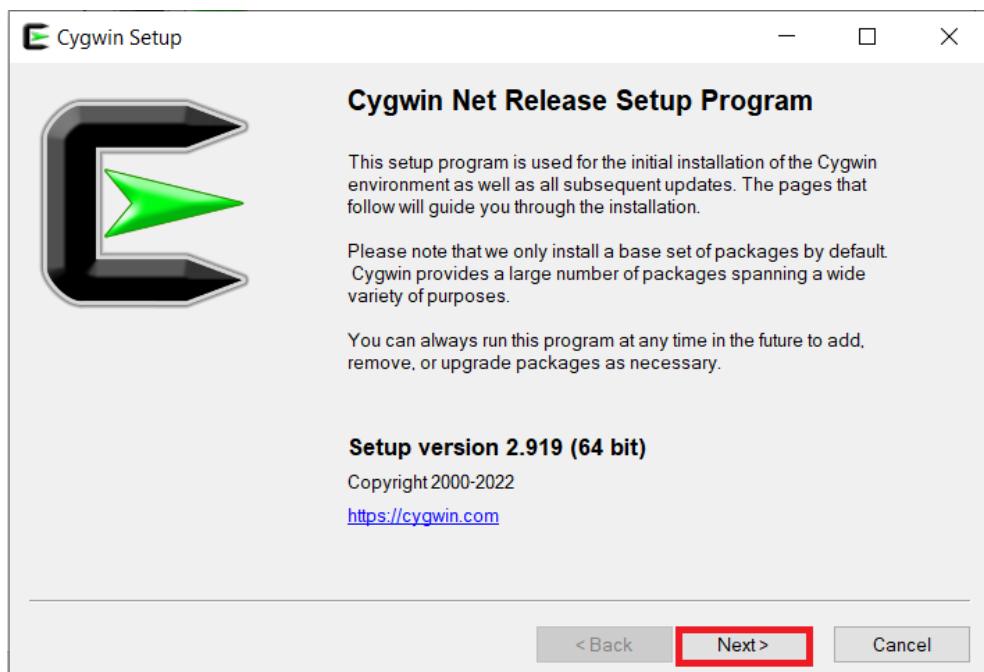
After downloading, open ‘File Explorer’ and go to your home directory (`c:\Users\<username>`) to create a new folder called ‘cygwin-installer’.



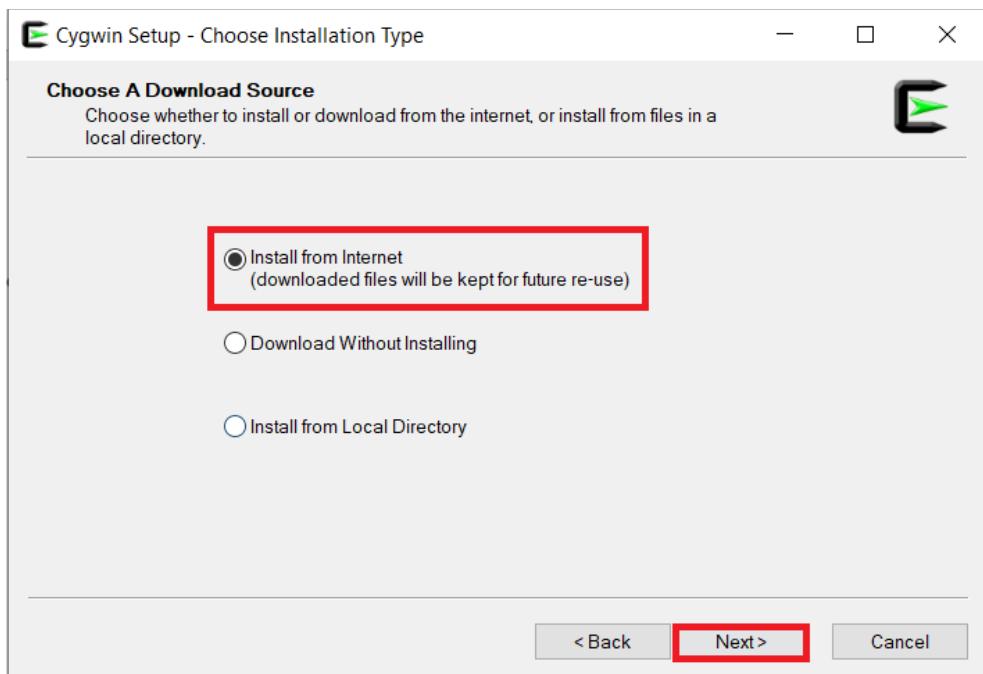
Next, move the Cygwin installer executable to the ‘cygwin-installer’ folder.



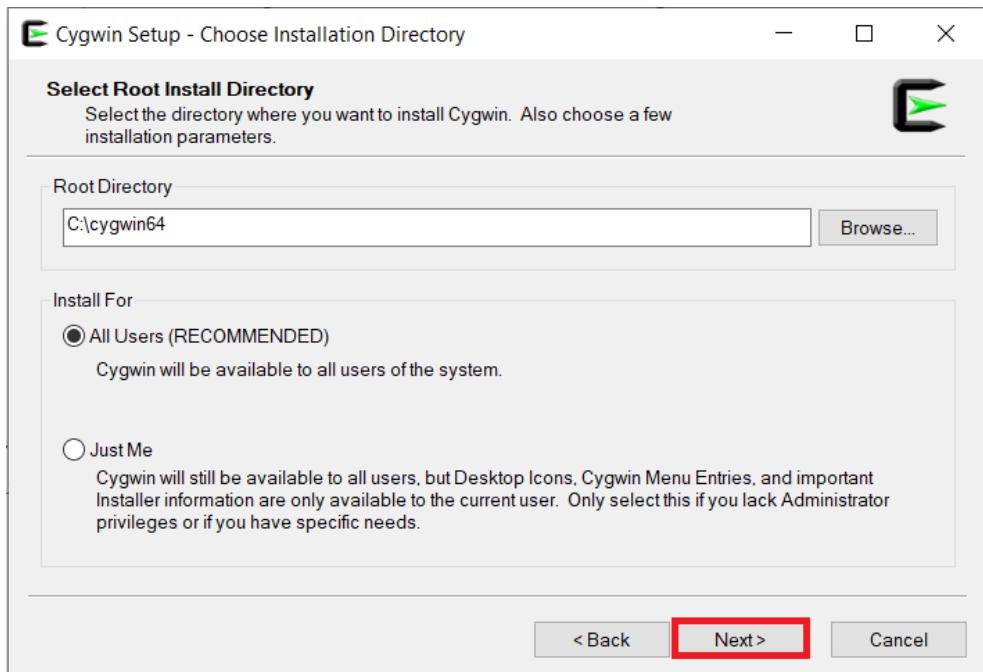
Inside the ‘cygwin-installer’ folder, we now run the cygwin installer executable. Once the setup program has started, click ‘Next’ to continue.



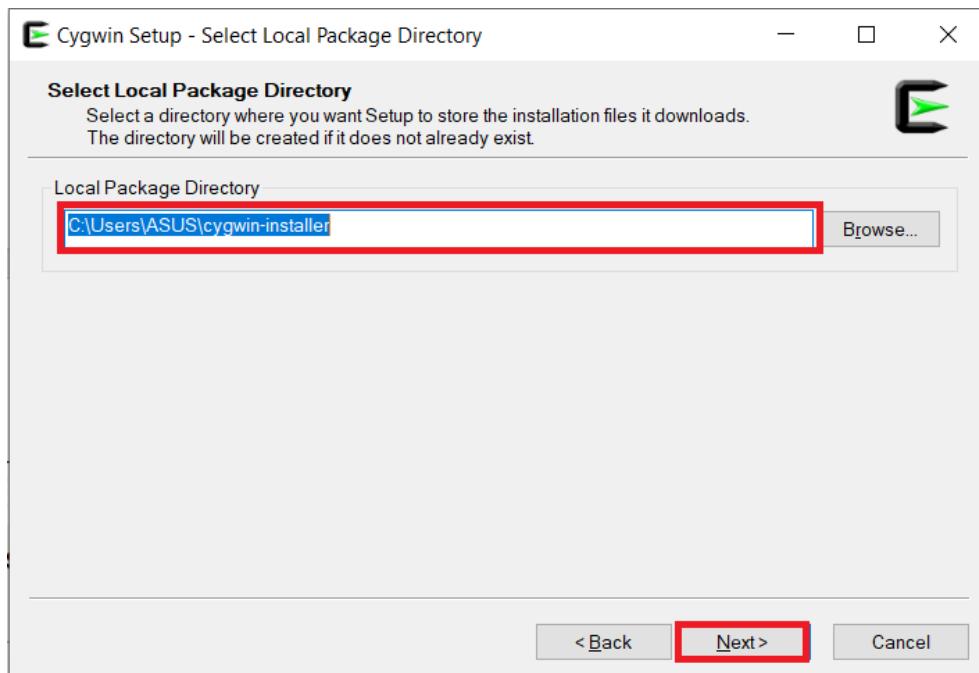
In the next section ‘Choose Installation Type’, we select Install from Internet. Click ‘Next’ to continue.



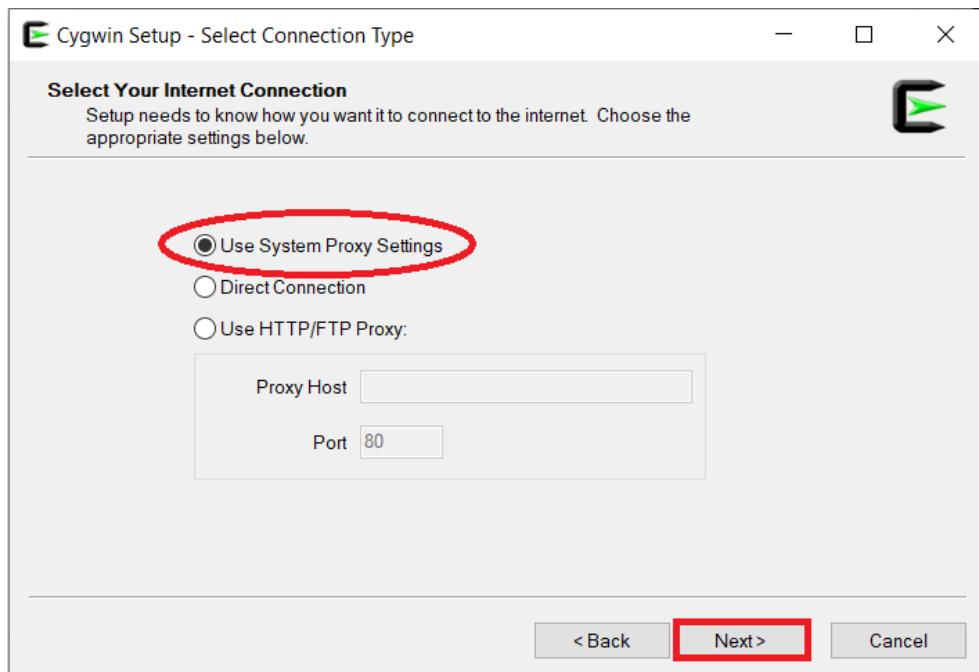
For the ‘Choose Installation Directory’ section, you can select a directory to where you want to install Cygwin. In this case, we continued with the default root directory ‘C:\cygwin64’. Next, for the ‘Install For’, keep the recommended option, ‘All Users’, selected unless you only want it to be installed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

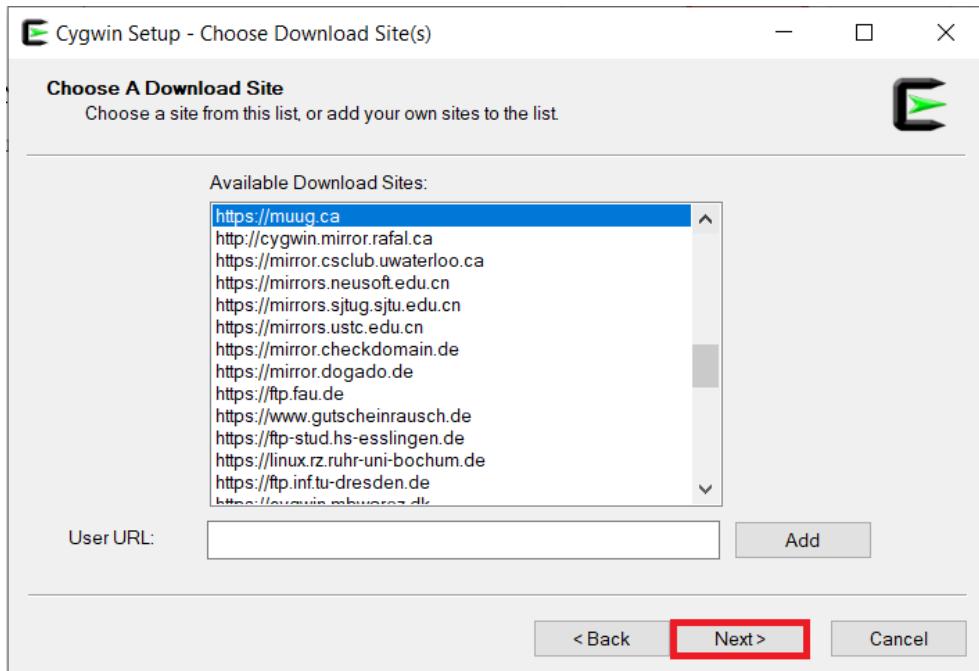


Now for the ‘Select Connection Type’, continue with the default setting ‘Use System Proxy Settings’ unless you want something to be specified here. Otherwise, click ‘Next’ to continue.

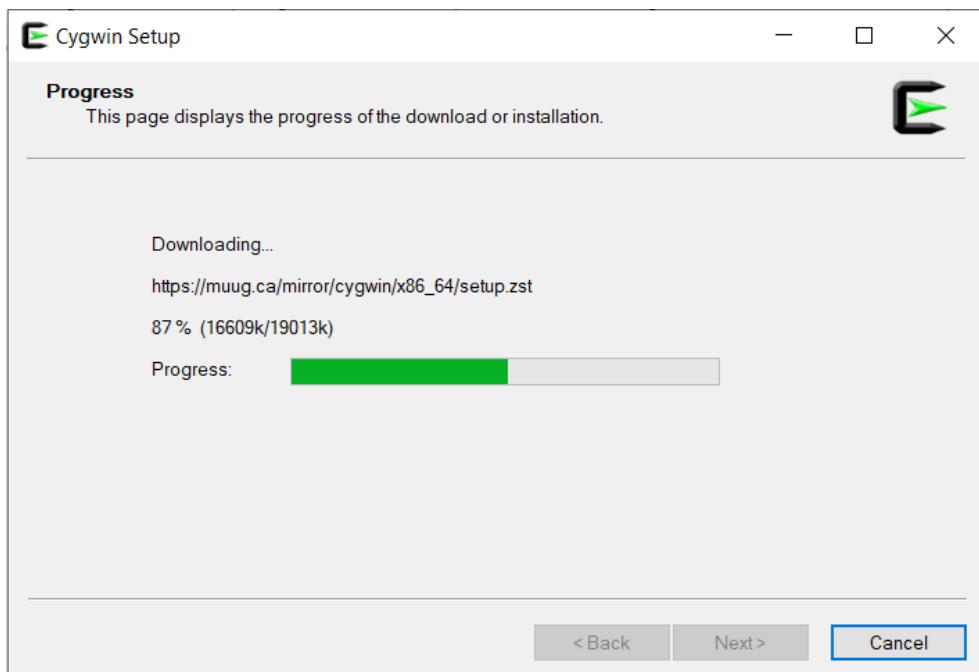


‘Choose Download Site(s)’ section will prompt you to choose a download site from a list from which you will download the Cygwin packages. It is better to select a mirror that is geographically

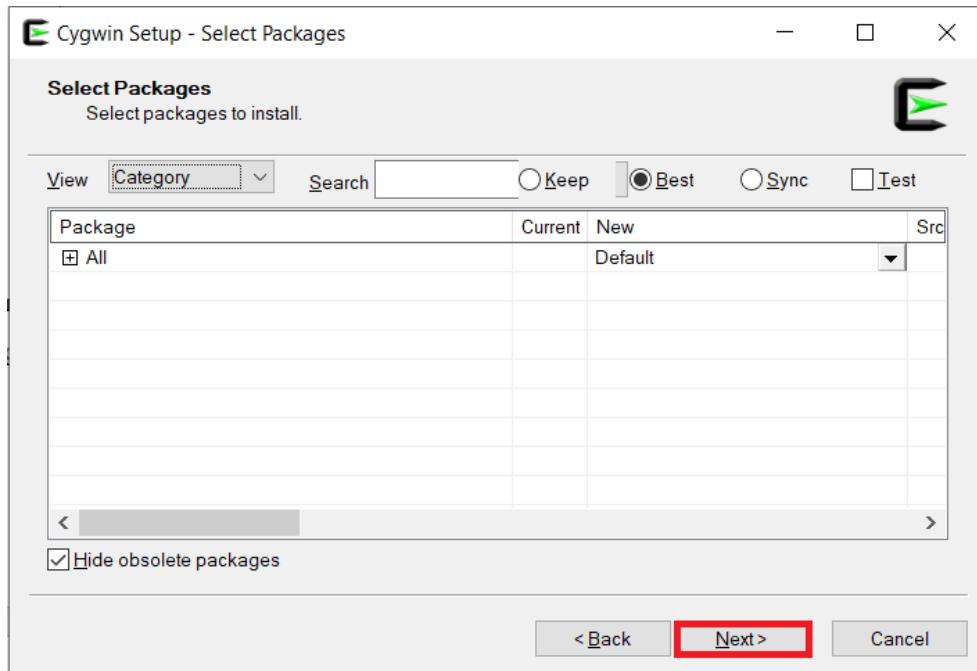
closer to you. However, if you are unsure, then choose any mirror. In this instance, we selected ‘<https://muug.ca>’. Click ‘Next’ to continue.



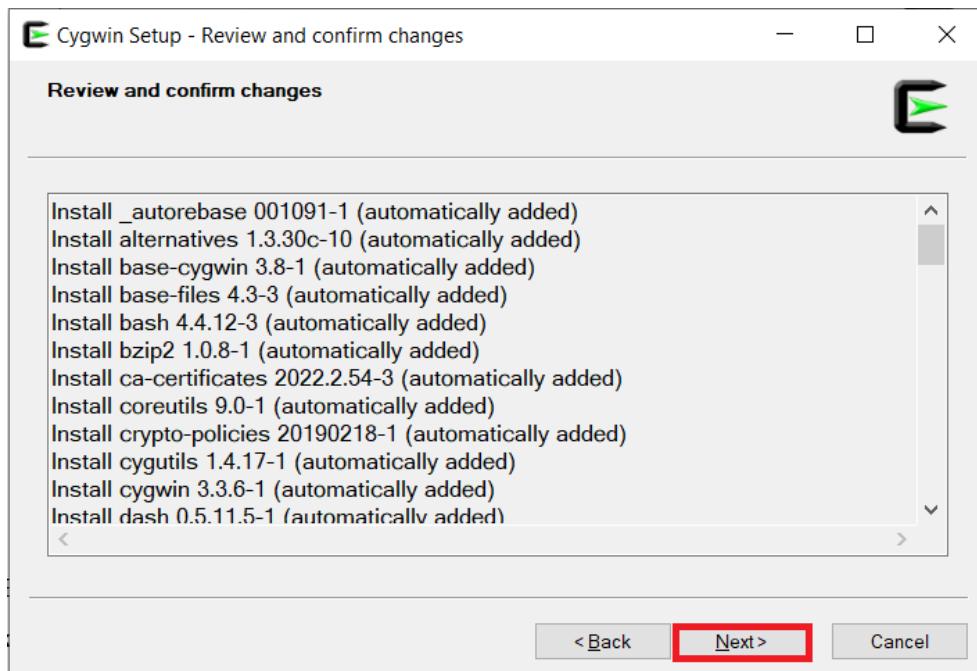
Now, we wait a few seconds for the download to finish.



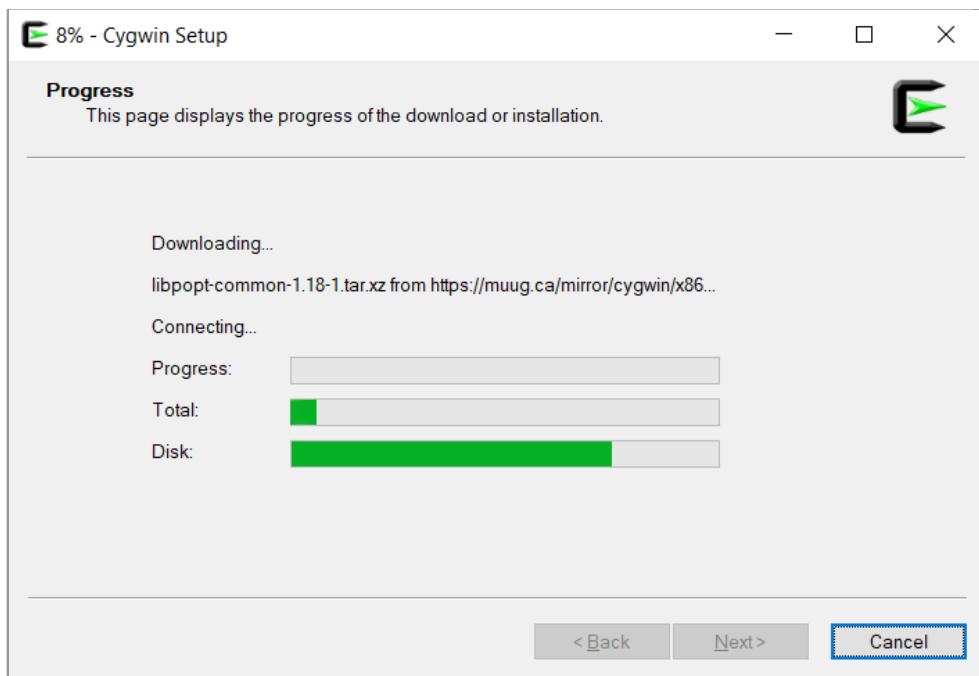
After it has downloaded, we are prompted to the ‘Select Packages’ section. For now, we will not select any packages, so click ‘Next’.



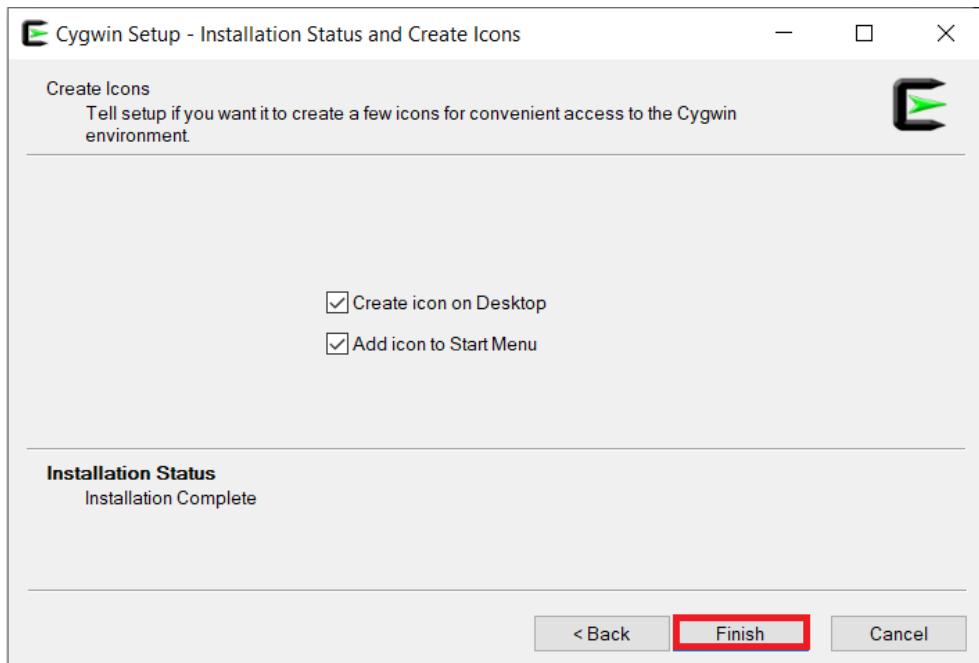
Before the final installation, we are prompted to review and confirm changes. Once done, click on ‘Next’ to start installing.



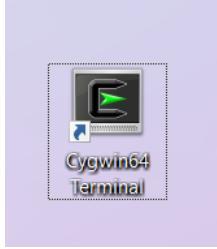
We now wait a few minutes for the Cygwin installation to finish.



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.



```
Copying skeleton files.
These files are for the users to personalise their cygwin experience.

They will never be overwritten nor automatically updated.

'./.bashrc' -> '/home/Alyssa//.bashrc'
'./.bash_profile' -> '/home/Alyssa//.bash_profile'
'./.inputrc' -> '/home/Alyssa//.inputrc'
'./.profile' -> '/home/Alyssa//.profile'

Alyssa@DESKTOP-K6T6ETI ~
$ |
```

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.

```
Copying skeleton files.
These files are for the users to personalise their cygwin experience.

They will never be overwritten nor automatically updated.

'./.bashrc' -> '/home/Alyssa//.bashrc'
'./.bash_profile' -> '/home/Alyssa//.bash_profile'
'./.inputrc' -> '/home/Alyssa//.inputrc'
'./.profile' -> '/home/Alyssa//.profile'

Alyssa@DESKTOP-K6T6ETI ~
$ ls -al
total 24
drwxr-xr-x 1 Alyssa None    0 Sep 28 07:14 .
drwxrwxrwt 1 Alyssa None    0 Sep 28 07:14 ..
-rw xr-xr-x 1 Alyssa None 1494 Sep 28 07:08 .bash_profile
-rw xr-xr-x 1 Alyssa None 5645 Sep 28 07:08 .bashrc
-rw xr-xr-x 1 Alyssa None 1919 Sep 28 07:08 .inputrc
-rw xr-xr-x 1 Alyssa None 1236 Sep 28 07:08 .profile

Alyssa@DESKTOP-K6T6ETI ~
$ |
```

V. Hive Installation

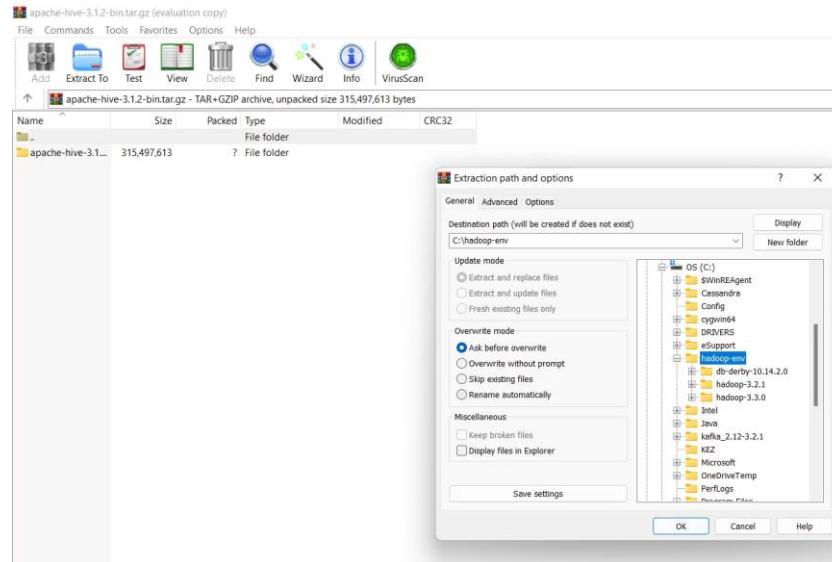
To download the Apache Hive binaries, click the following link and select *apache-hive-3.1.2-bin.tar.gz* to start the download: <https://downloads.apache.org/hive/hive-3.1.2/>

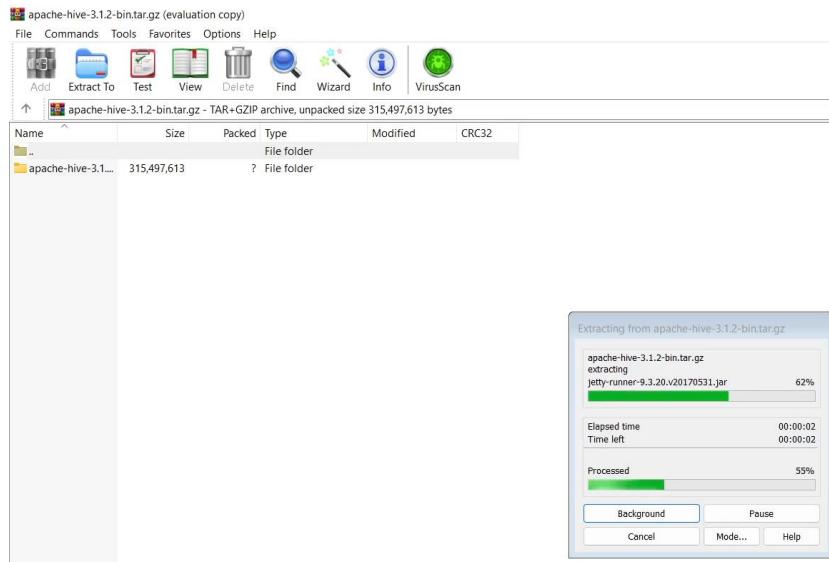
Index of /hive/hive-3.1.2

Name	Last modified	Size	Description
Parent Directory		-	
apache-hive-3.1.2-bin.tar.gz	2019-08-26 20:20	266M	
apache-hive-3.1.2-bin.tar.gz.asc	2019-08-26 20:20	833	
apache-hive-3.1.2-bin.tar.gz.sha256	2019-08-26 20:20	95	
apache-hive-3.1.2-src.tar.gz	2019-08-26 20:20	24M	
apache-hive-3.1.2-src.tar.gz.asc	2019-08-26 20:20	833	
apache-hive-3.1.2-src.tar.gz.sha256	2019-08-26 20:20	95	

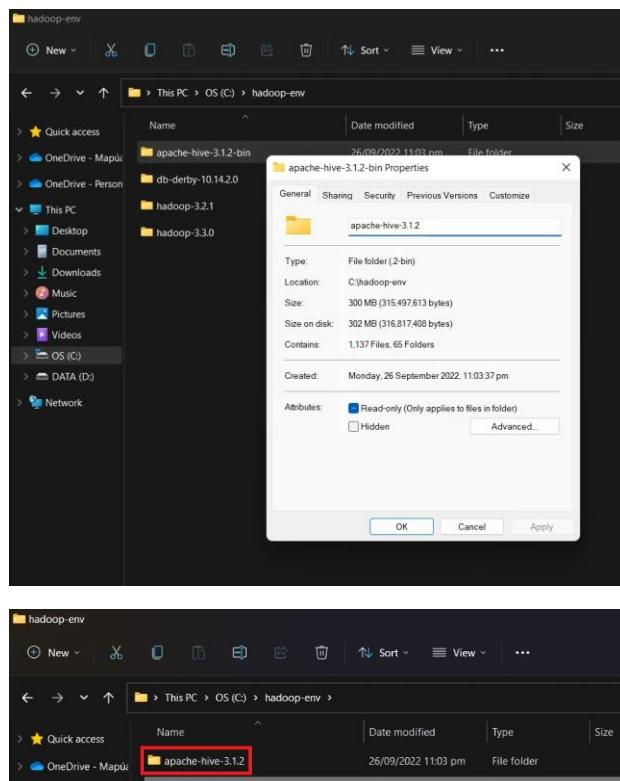


Once the package has finished downloading, open it using WinRAR Archiver and configure its extraction destination to be in the directory "C:\hadoop-env."

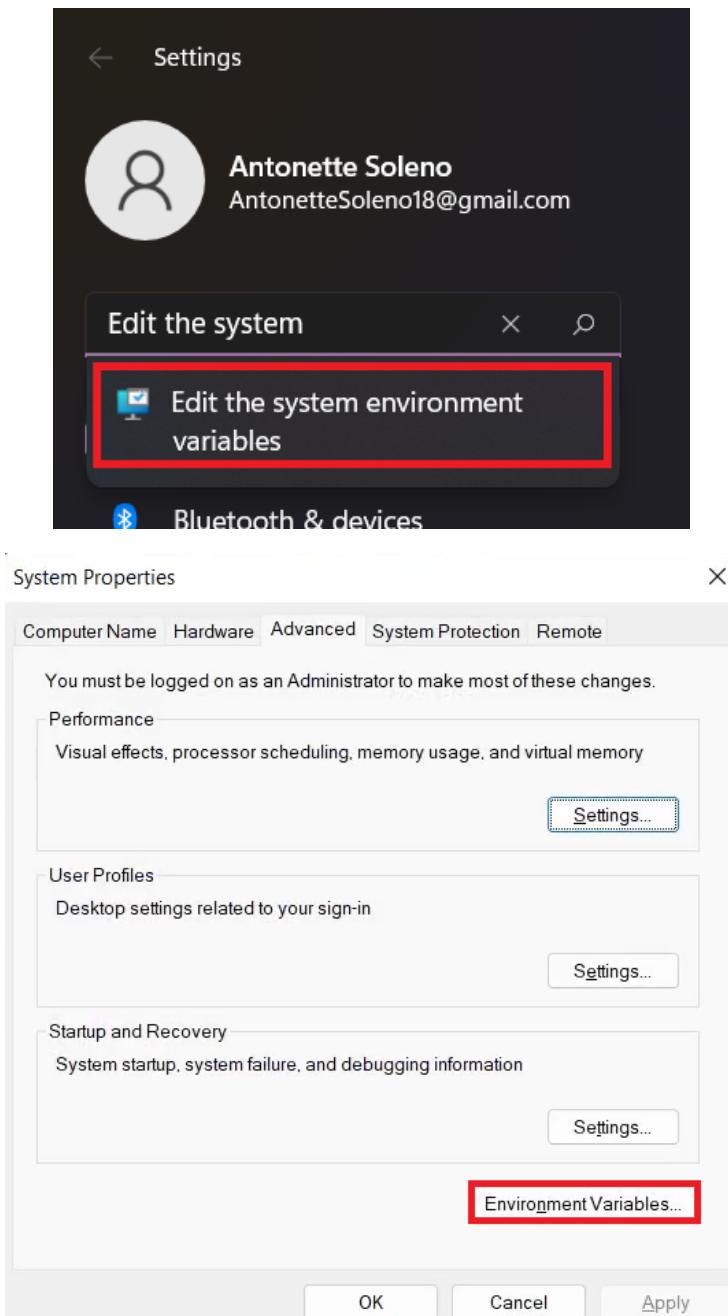




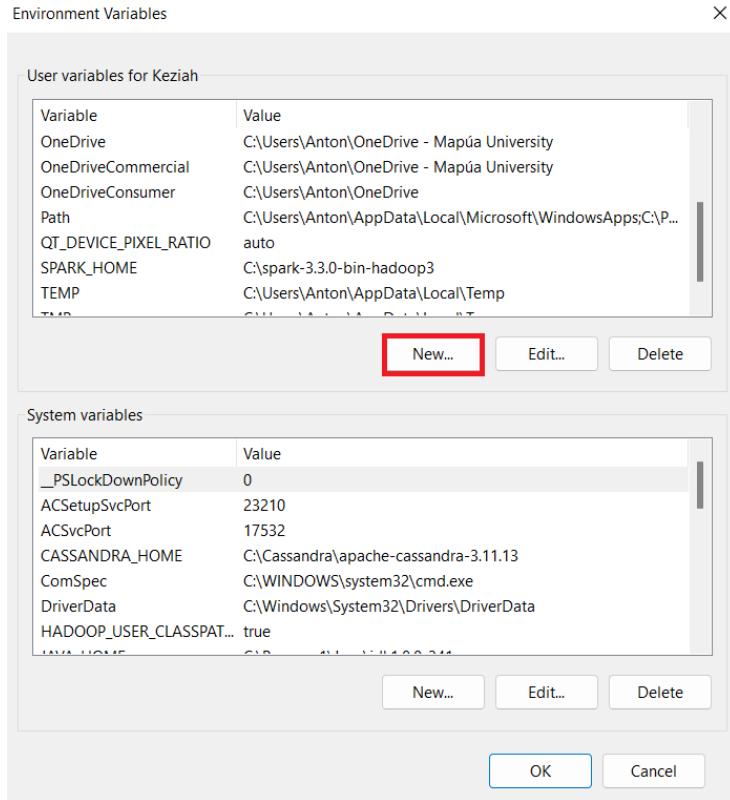
The extracted file should appear in the *hadoop-env* folder as *apache-hive-3.1.2-bin*. Rename this folder as *apache-hive-3.1.2*.



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.

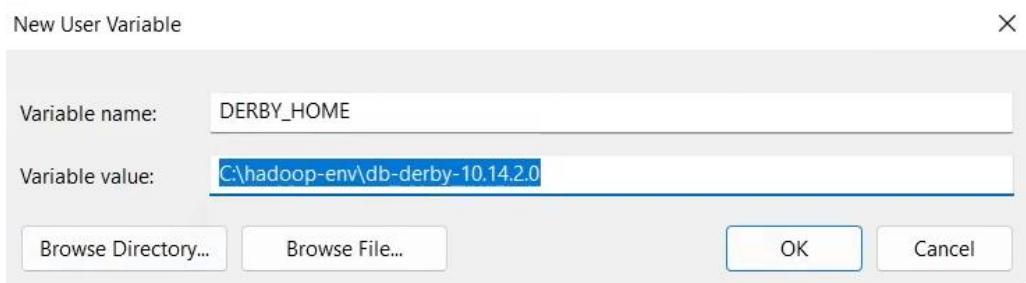


The user variables required in this installation guide are as follows:

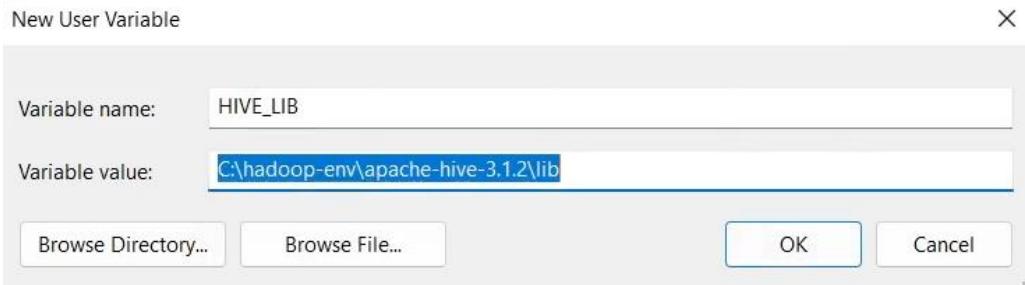
- **HIVE_HOME:** “C:\hadoop-env\apache-hive-3.1.2”



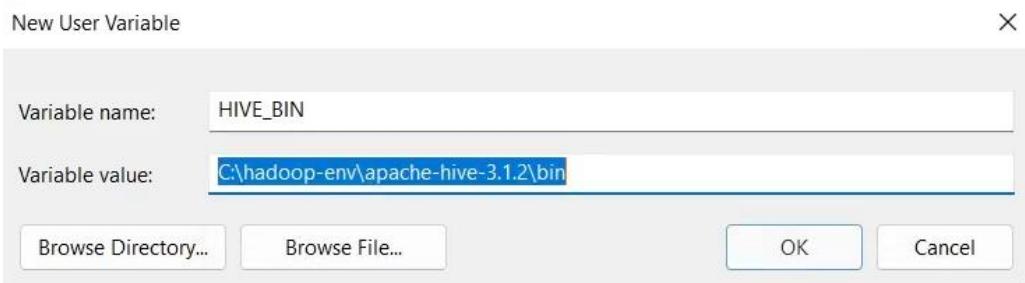
- **DERBY_HOME:** “C:\hadoop-env\db-derby-10.14.2.0”



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



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

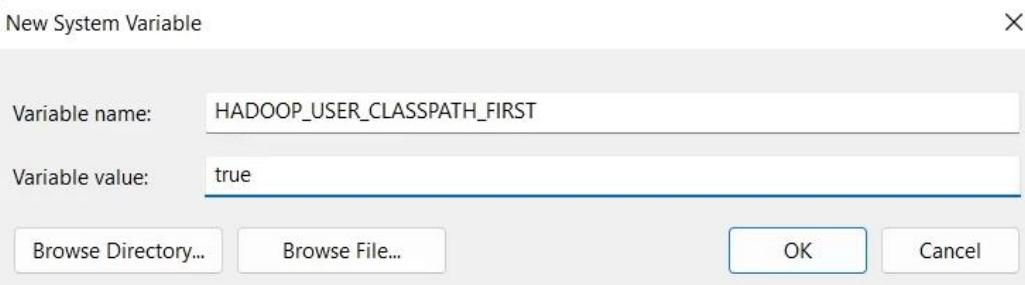


- **HADOOP_USER_CLASSPATH_FIRST:** “true”



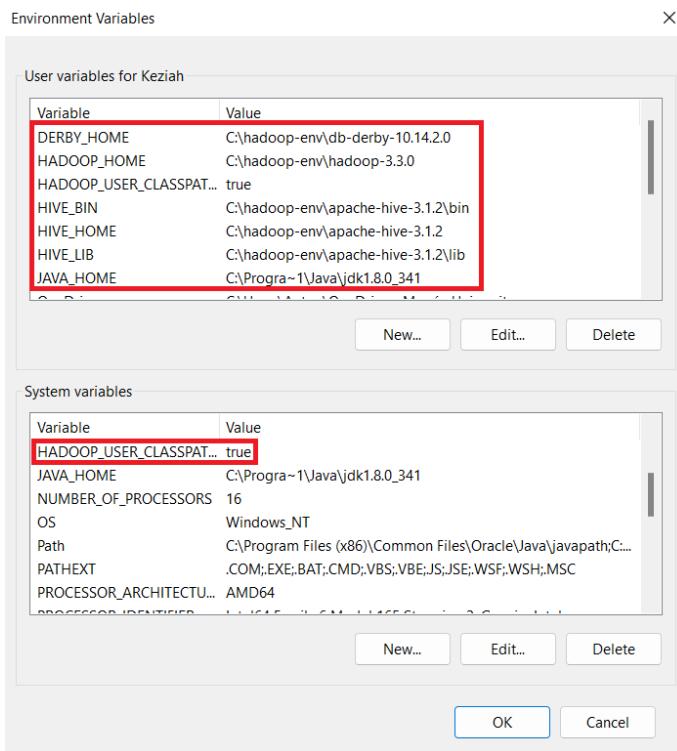
Additionally, the only system variable required is the following:

- **HADOOP_USER_CLASSPATH_FIRST:** “true”

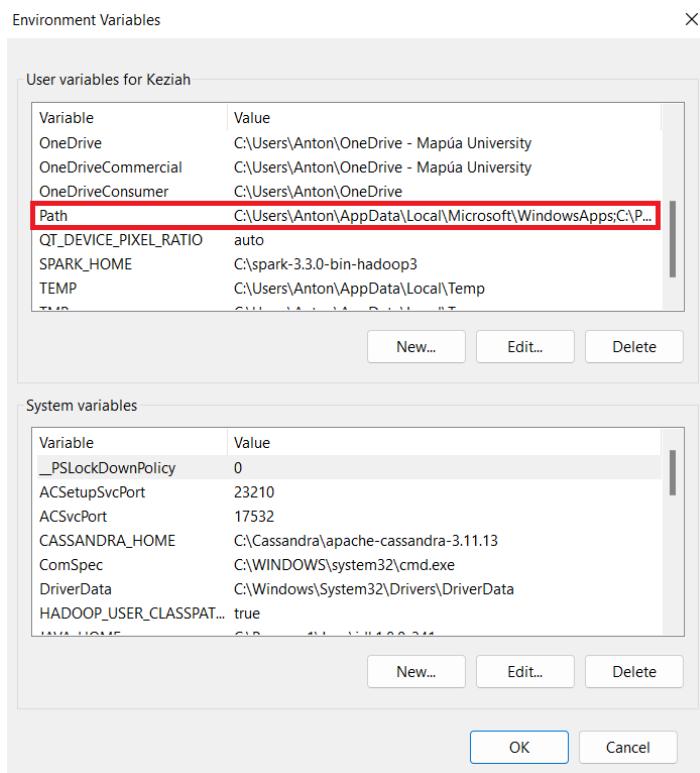


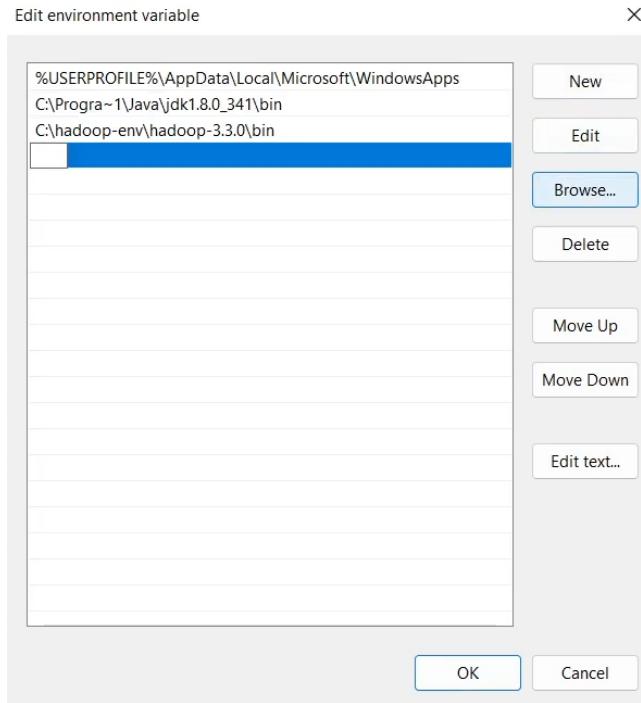
After every user or system variable creation, click “OK” to save the variable.

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

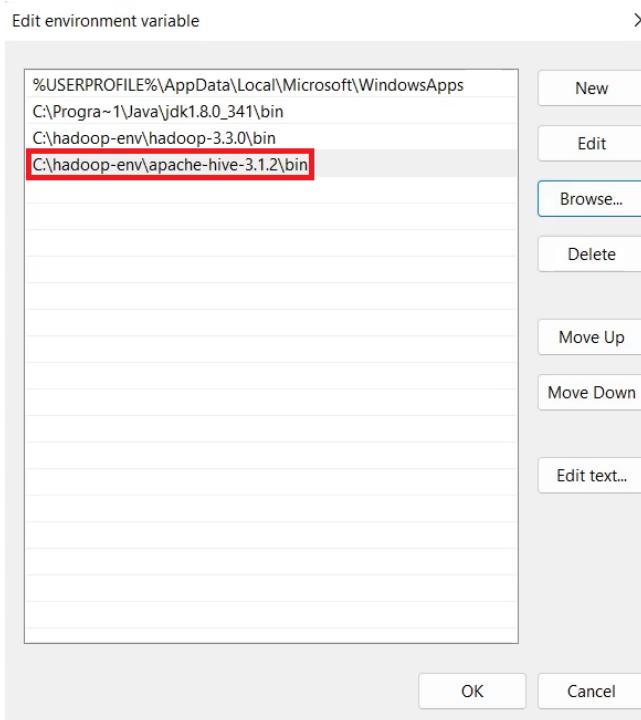


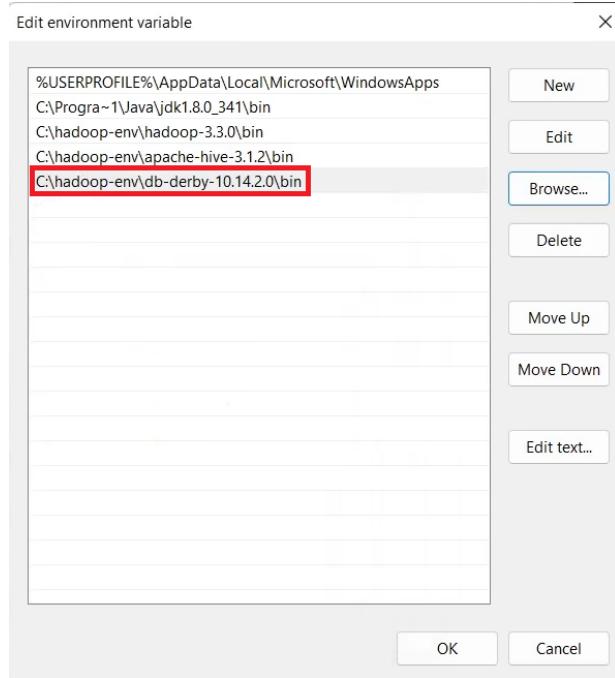
Once all required variables have been created, locate “Path” under *User variables* then double click to open the *Edit environment variable* window.





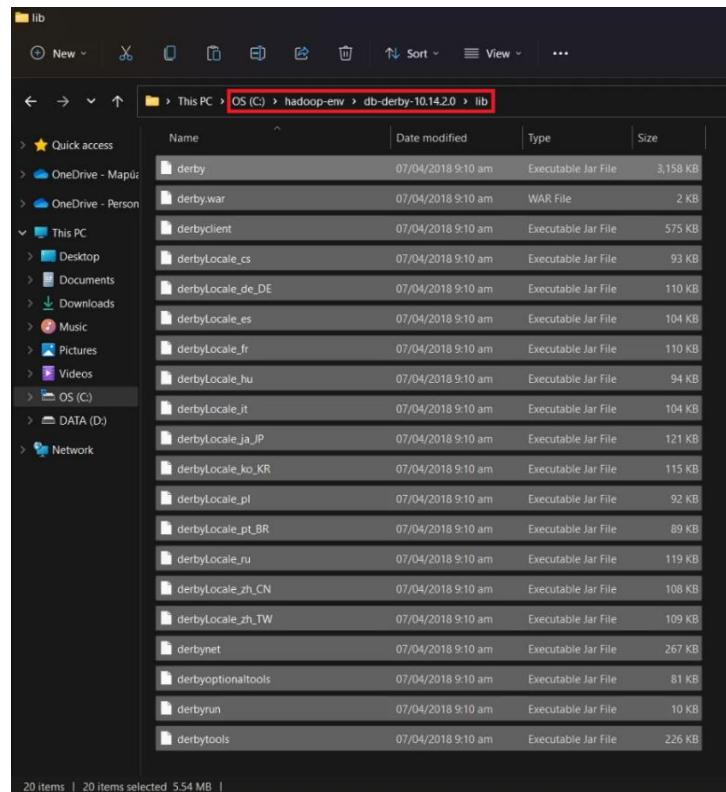
On the *Edit environment variable* window, add two paths leading to each bin folder of the Hive (apache-hive-3.1.2) and Derby (db-derby-10.14.2.0) folders.

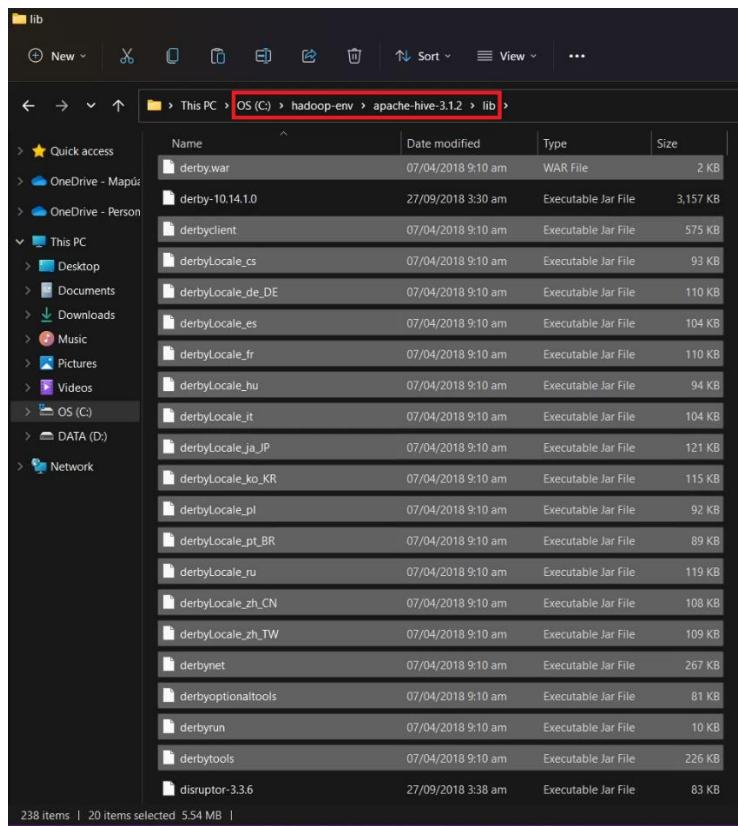




a. Configure Apache Hive

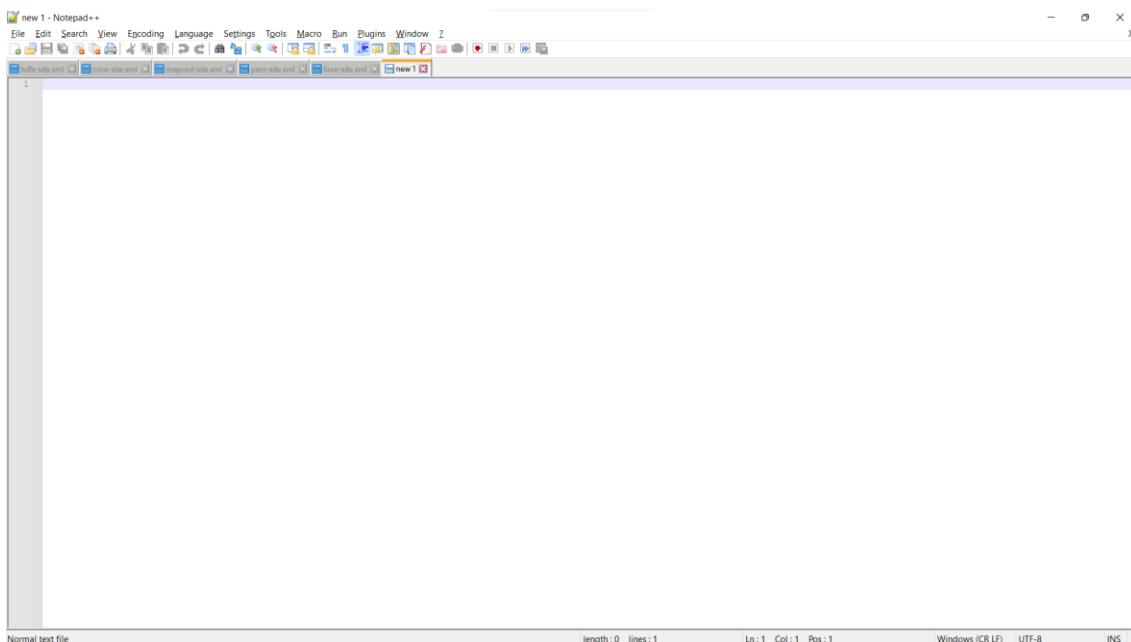
After setting the environment variables for Apache Hive and Derby, the next step is to copy all *.jar files from the Derby libraries directory (`C:\hadoop-env\db-derby-10.14.2.0\lib`) and paste them onto the Hive libraries directory (`C:\hadoop-env\apache-hive-3.1.2\lib`).





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>
```

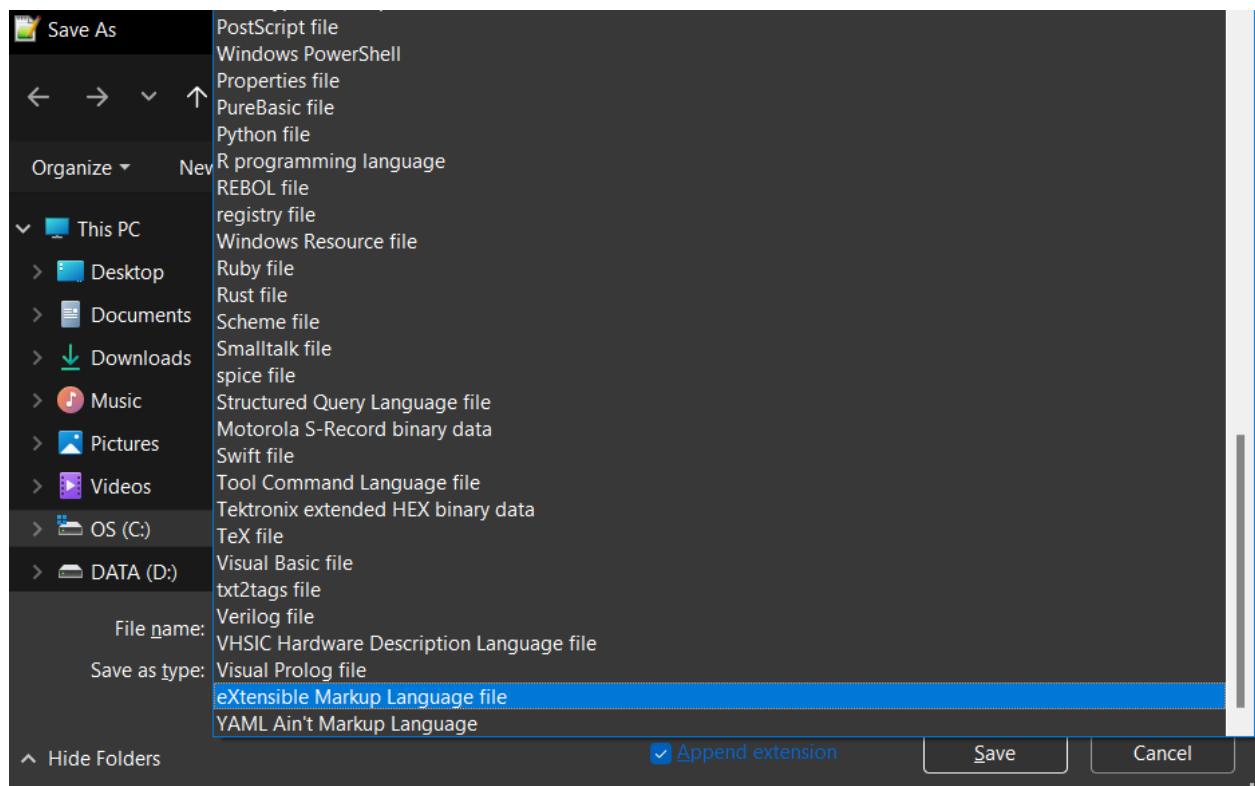
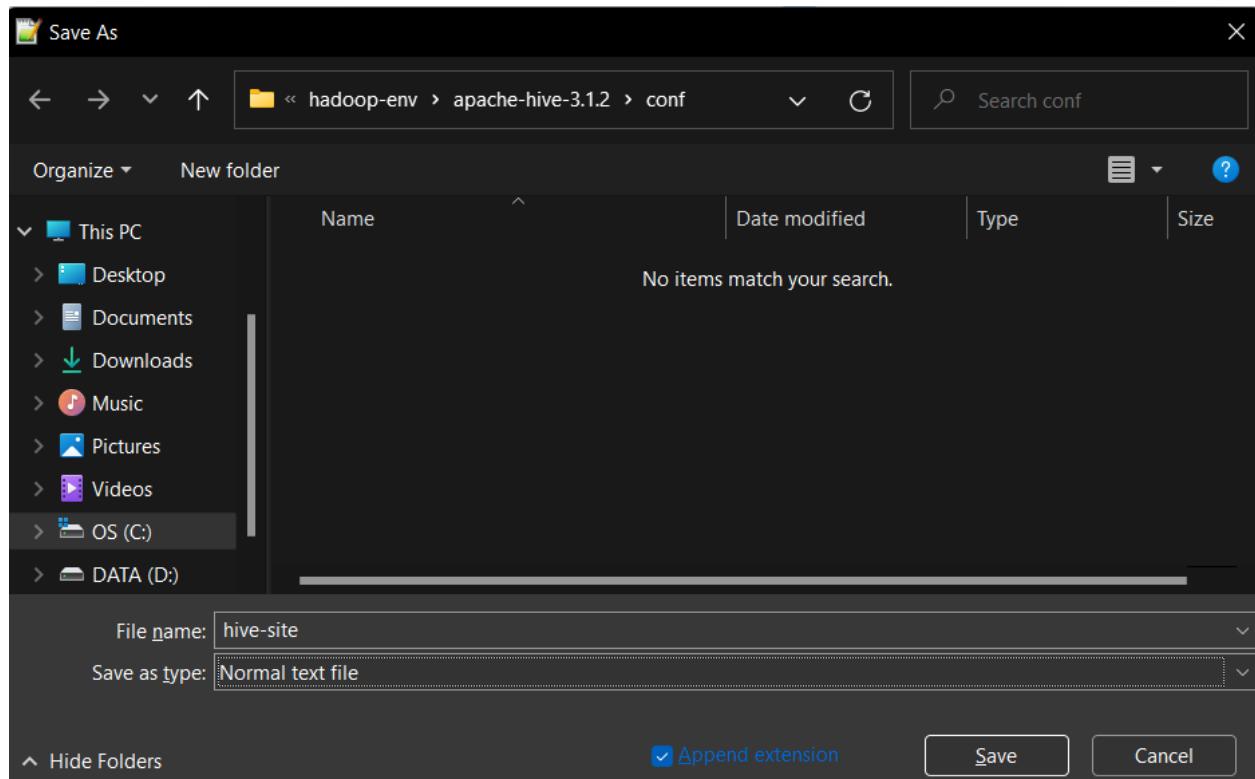
And paste it onto the newly created Notepad++ file. The file should look like the image below.



The screenshot shows a Notepad++ window with the title bar "hive-site.xml". The code area contains the XML configuration provided above. The code is syntax-highlighted, with tags in blue and values in black. Line numbers are visible on the left. The status bar at the bottom shows the file path "D:\Hive\hive-2.3.1\conf\hive-site.xml" and the word "File".

```
1 <?xml version="1.0"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <configuration><property> <name>javax.jdo.option.ConnectionURL</name>
4 <value>jdbc:derby://localhost:1527/metastore_db;create=true</value>
5 <description>JDBC connect string for a JDBC metastore</description>
6 </property>
7 <property>
8 <name>javax.jdo.option.ConnectionDriverName</name>
9 <value>org.apache.derby.jdbc.ClientDriver</value>
10 <description>Driver class name for a JDBC metastore</description>
11 </property>
12 <property>
13 <name>hive.server2.enable.doAs</name>
14 <description>Enable user impersonation for HiveServer2</description>
15 <value>true</value>
16 </property>
17 <property>
18 <name>hive.server2.authentication</name>
19 <value>NONE</value>
20 <description> Client authentication types. NONE: no authentication check
21 LDAP: LDAP/AD based authentication
22 KERBEROS: Kerberos/GSSAPI authentication
23 CUSTOM: Custom authentication provider (Use with property
24 hive.server2.custom.authentication.class) </description>
25 </property>
</configuration>
```

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.

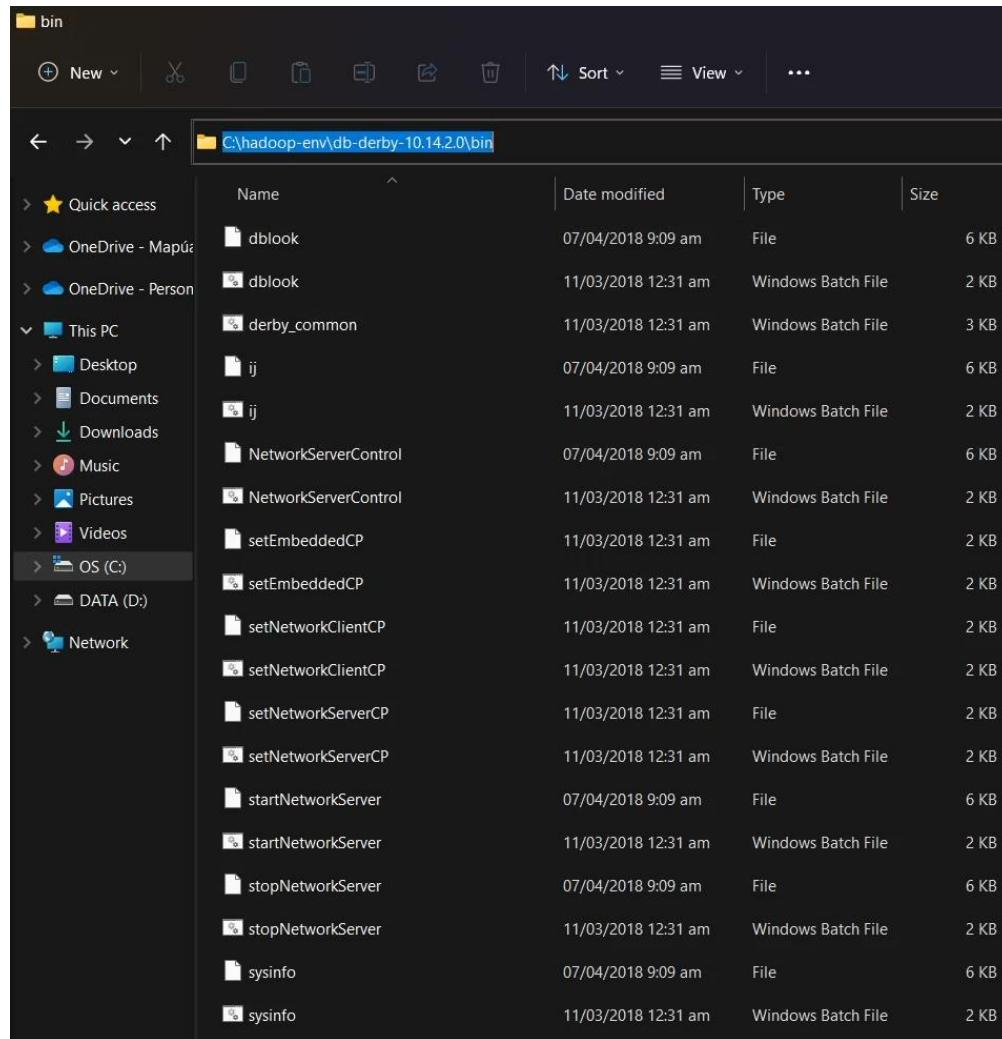


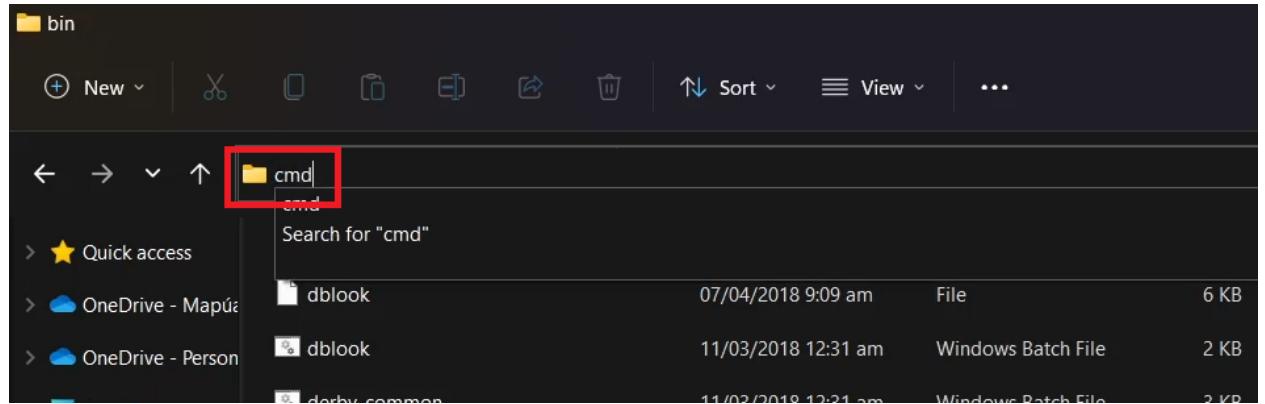


Now that the prerequisite configurations have been made, all the installed services may already be run.

To begin, run the Hadoop services by following the instructions under **II. Hadoop Installation > [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

```
C:\Windows\System32\cmd.exe - StartNetworkServer -h 0.0.0.0
Microsoft Windows [Version 10.0.22000.978]
(c) Microsoft Corporation. All rights reserved.

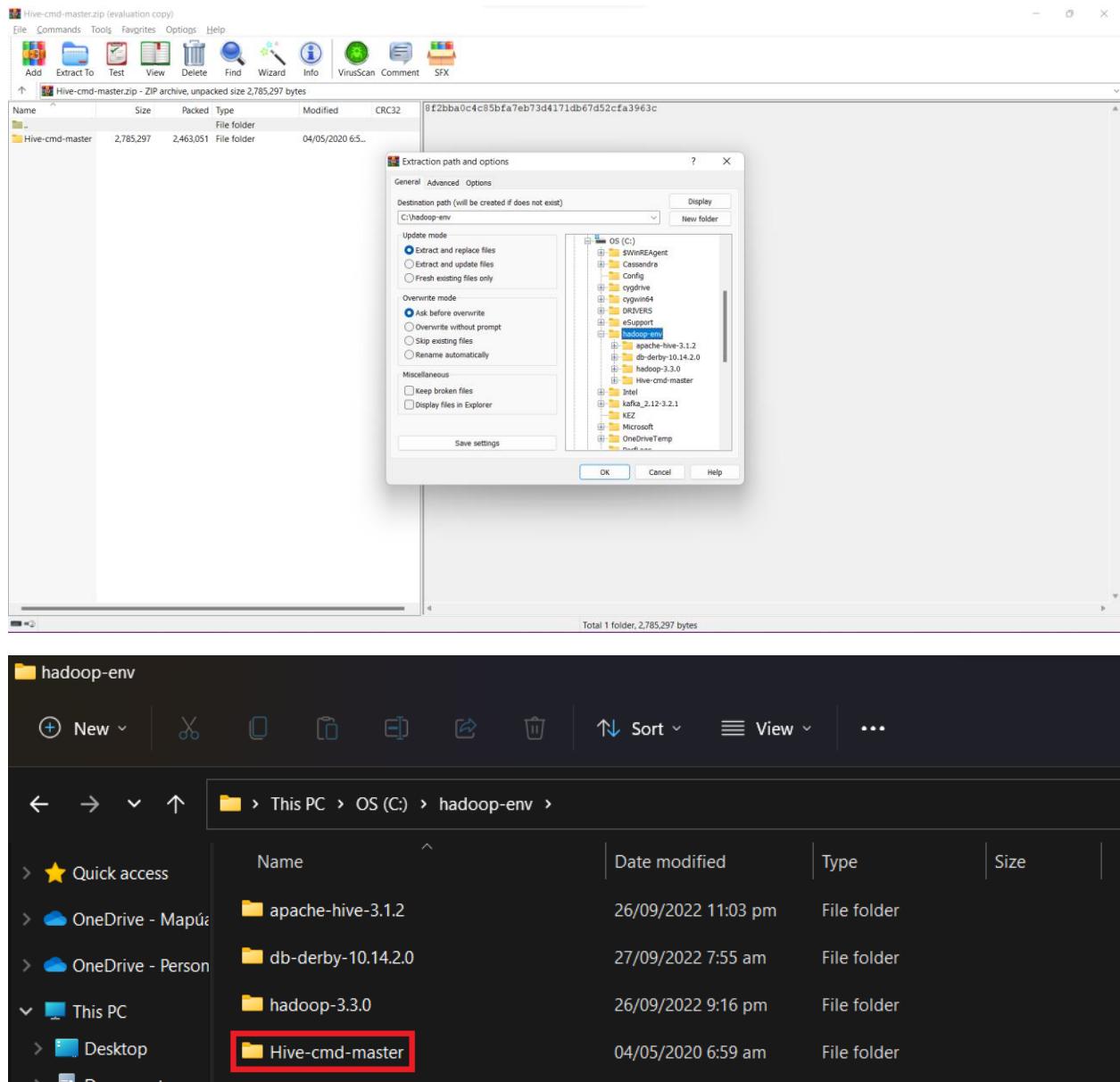
C:\hadoop-env\db-derby-10.14.2.0\bin>StartNetworkServer -h 0.0.0.0
Mon Sep 26 23:25:43 SGT 2022 : Security manager installed using the Basic server security policy.
Mon Sep 26 23:25:48 SGT 2022 : Apache Derby Network Server - 10.14.2.0 - (1828579) started and ready to accept connections on port 1527
```

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.

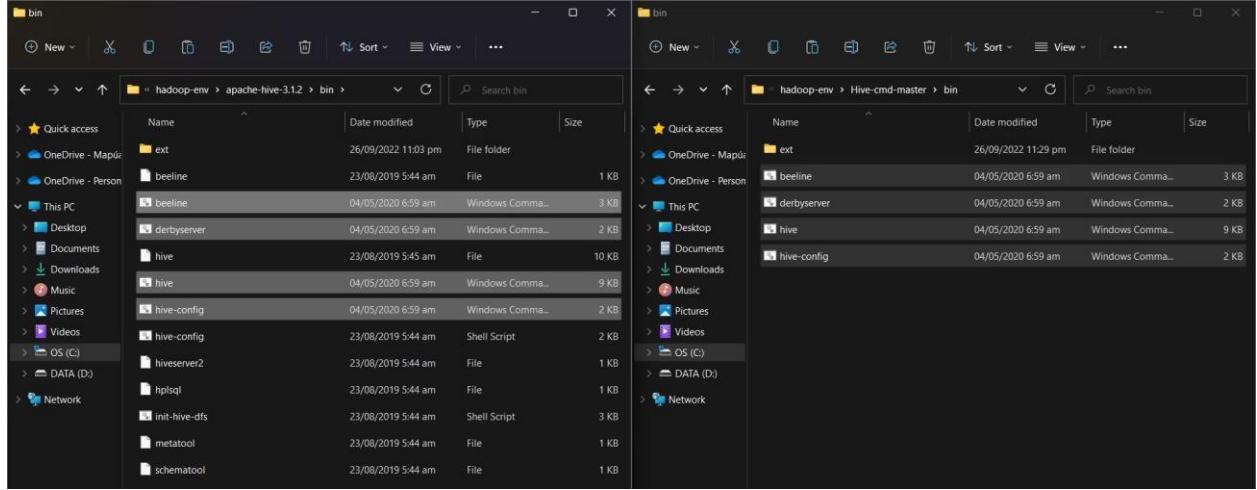
The screenshot shows a GitHub repository page for 'Hive-cmd'. The 'Code' button is highlighted with a red box. The 'Download ZIP' button is also highlighted with a red box. The repository details include:

- Owner: HadiFadl
- Name: Hive-cmd
- Branch: master
- Tags: 1 branch, 0 tags
- Clone options: HTTPS, GitHub CLI
- Repository URL: <https://github.com/HadiFadl/Hive-cmd.git>
- Actions: Notifications (19), Fork (19), Star (14)
- About: All cmd files needed to run Hive on windows (taken from <https://svn.apache.org/repos/asf/hive/trunk/bin/>)
- Tags: windows, hive, apache-hive
- Statistics: Readme, MIT license, 14 stars, 3 watching, 19 forks
- Releases: None

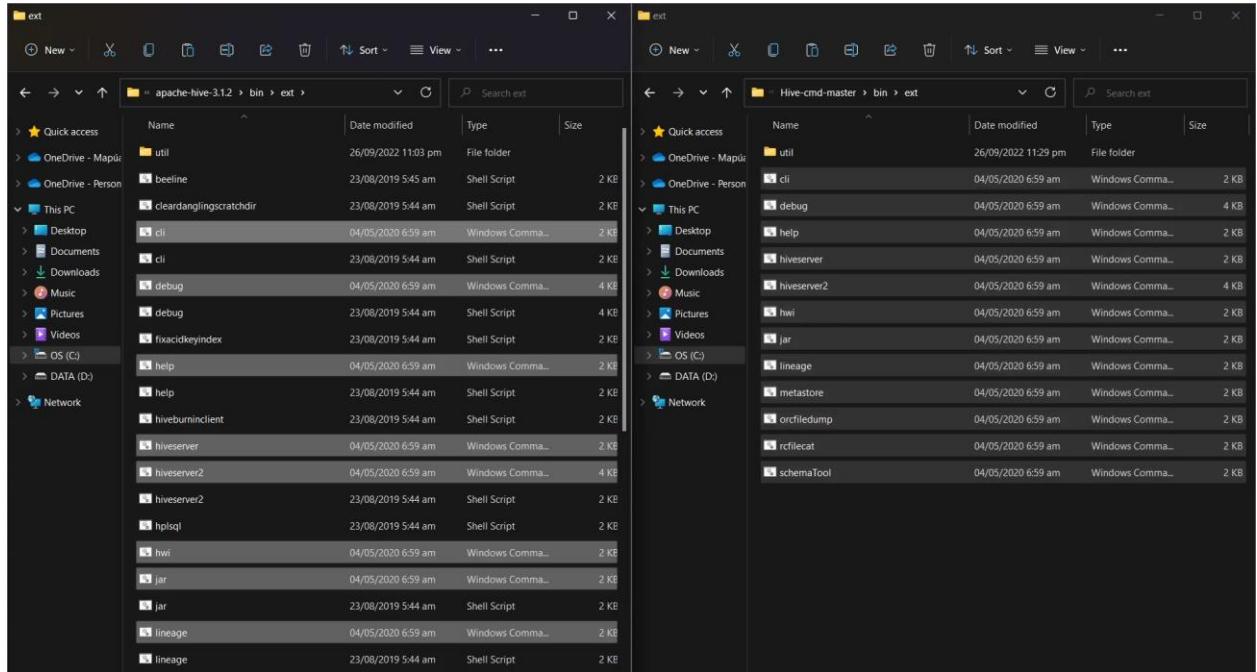
After the download has finished, unzip the file using WinRAR Archiver (or any other unzipping tool available) and save it in the hadoop-env folder.



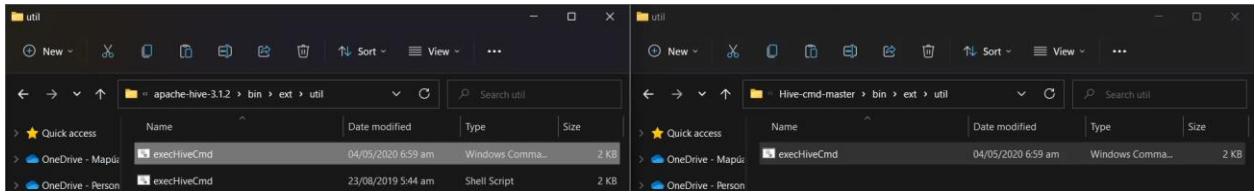
Next, go to the directory `C:\hadoop-env\Hive-cmd-master\bin`, then copy all *.cmd files and paste them onto the directory `C:\hadoop-env\apache-hive-3.1.2\bin`. Repeat the same process for folders `ext` and `util`.



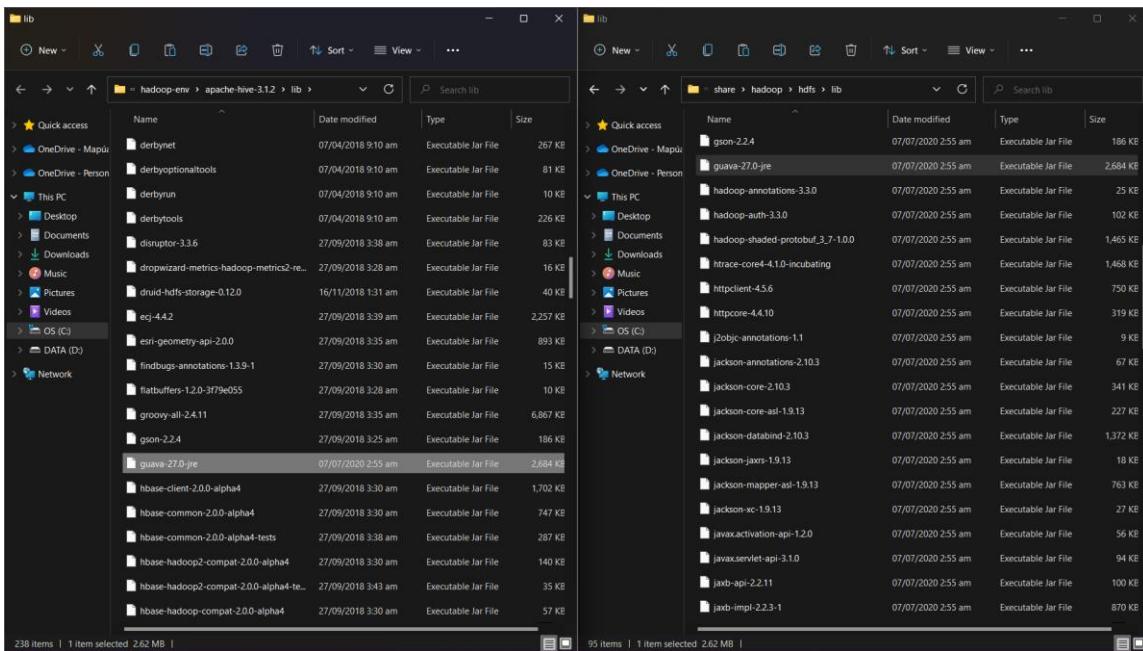
Go to the directory `C:\hadoop-env\Hive-cmd-master\bin\ext`, then copy all *.cmd files and paste them onto the directory `C:\hadoop-env\apache-hive-3.1.2\bin\ext`.



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`

```
C:\Windows\System32\cmd.exe - hive
Microsoft Windows [Version 10.0.22000.978]
(c) Microsoft Corporation. All rights reserved.

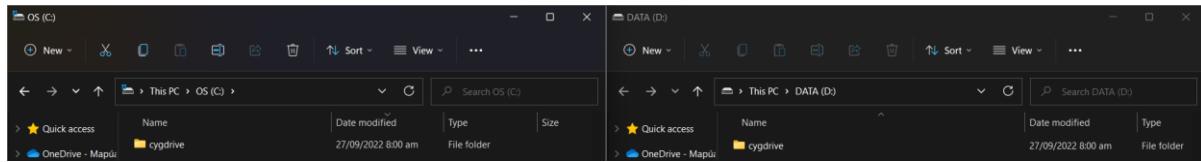
C:\hadoop-env\apache-hive-3.1.2\bin>hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop-env/apache-hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop-env/hadoop-3.3.0/share/hadoop/common/lib/slf4j-log4j12-1.7.20.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4JLoggerFactory]
2022-09-28T12:58:32,659 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Found configuration file file:/C:/hadoop-env/apache-hive-3.1.2/conf/hive-site.xml
Hive Session ID = 1ab6ffff-8e76-4331-8888-862f91c97a51

Logging initialized using configuration in [jar:file:/C:/hadoop-env/apache-hive-3.1.2/lib/hive-common-3.1.2.jar!/hive-log4j2.properties.Async: true
2022-09-28T12:58:35,101 INFO [main] org.apache.hadoop.hive ql.session.SessionState - Created HDFS directory: /tmp/hive/Keziah/1a6bffff-8e76-4331-8888-862f91c97a51
2022-09-28T12:58:35,107 INFO [main] org.apache.hadoop.hive ql.session.SessionState - Created local directory: C:/Users/Anton/AppData/Local/Temp/Keziah/1a6bffff-8e76-4331-8888-862f91c97a51
2022-09-28T12:58:35,110 INFO [main] org.apache.hadoop.hive ql.session.SessionState - Created local directory: /tmp/hive/Keziah/1a6bffff-8e76-4331-8888-862f91c97a51/_tmp_space.db
2022-09-28T12:58:35,123 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: 1ab6ffff-8e76-4331-8888-862f91c97a51
2022-09-28T12:58:35,123 INFO [main] org.apache.hadoop.hive ql.session.SessionState - Updating thread name to 1ab6ffff-8e76-4331-8888-862f91c97a51 main
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
2022-09-28T12:58:39,871 INFO [1ab6ffff-8e76-4331-8888-862f91c97a51 main] CliDriver - Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
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

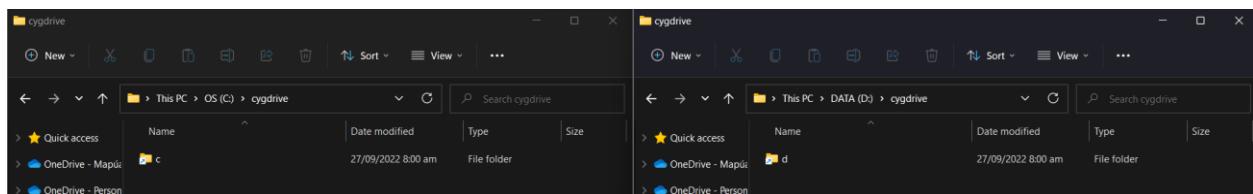


Then, run the command prompt terminal as administrator and execute the following commands:

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

A screenshot of an 'Administrator: Command Prompt' window. It shows the Windows version (10.0.22000.978) and copyright information. Two 'mklink /J' commands are run: one creating a junction point 'd' on drive D pointing to 'D:\cygdrive\d\' and another creating a junction point 'c' on drive C pointing to 'C:\cygdrive\c\'.

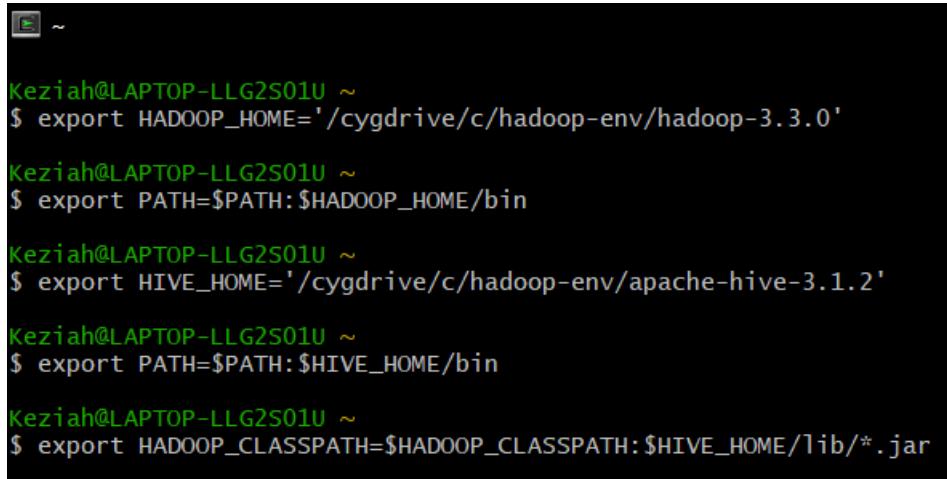
Running the above commands should result in the output below:



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

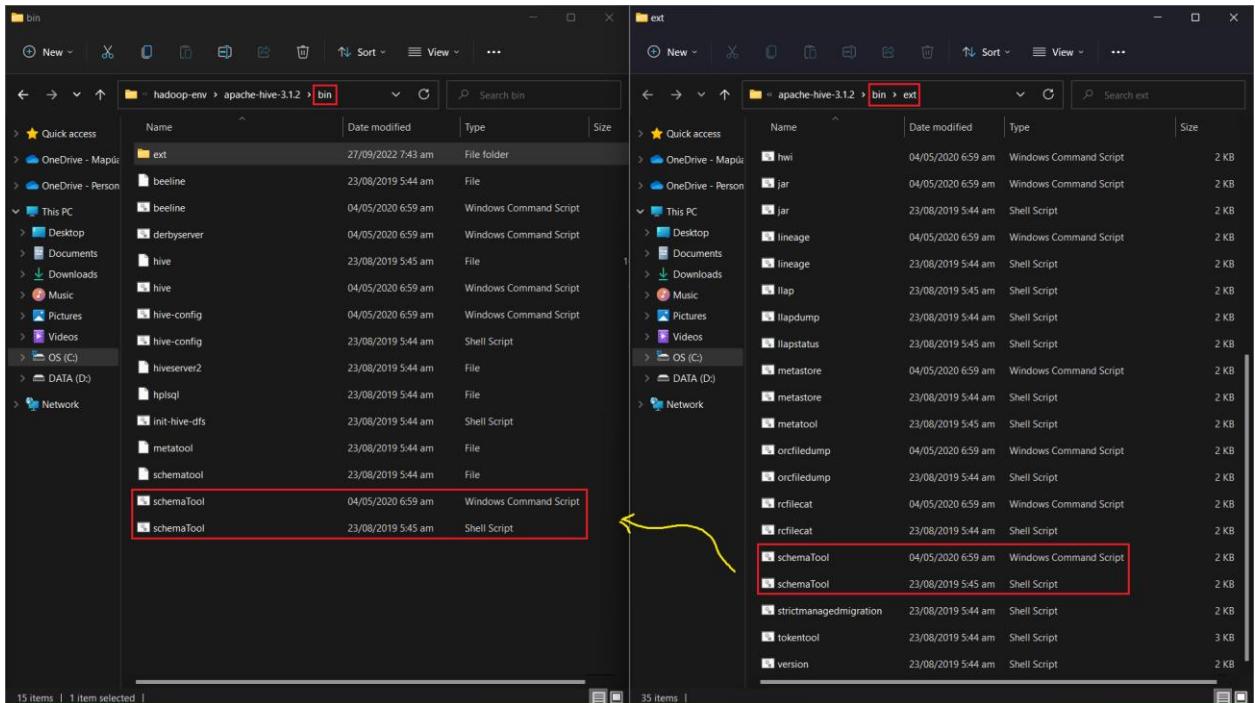
To begin, open the Cygwin terminal and run the following commands:

```
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
```

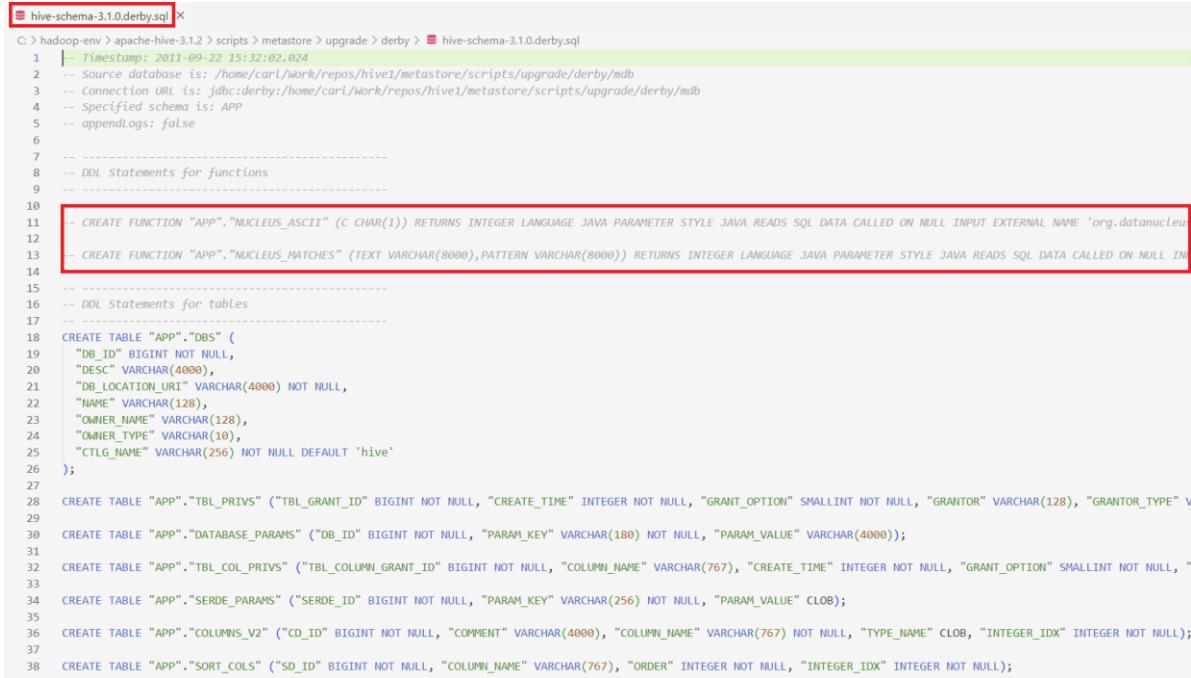


```
Keziah@LAPTOP-LLG2S01U ~  
$ export HADOOP_HOME='/cygdrive/c/hadoop-env/hadoop-3.3.0'  
Keziah@LAPTOP-LLG2S01U ~  
$ export PATH=$PATH:$HADOOP_HOME/bin  
Keziah@LAPTOP-LLG2S01U ~  
$ export HIVE_HOME='/cygdrive/c/hadoop-env/apache-hive-3.1.2'  
Keziah@LAPTOP-LLG2S01U ~  
$ export PATH=$PATH:$HIVE_HOME/bin  
Keziah@LAPTOP-LLG2S01U ~  
$ 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.



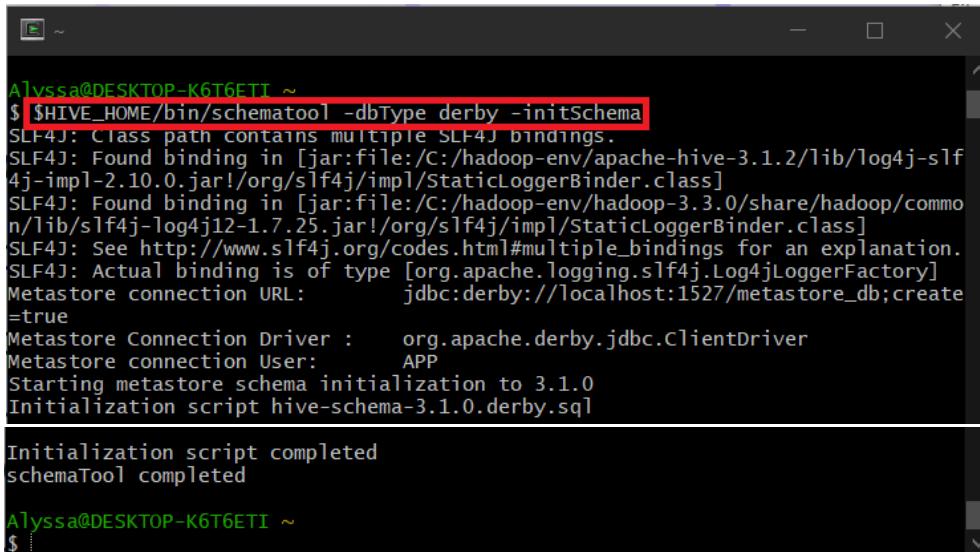
```

C:\> hadoop-env > apache-hive-3.1.2 > scripts > metastore > upgrade > derby > hive-schema-3.1.0.derby.sql
1   | Timestamp: 2011-09-22 15:32:02,024
2   -- Source database is: /home/carl/work/repos/hive1/metastore/scripts/upgrade/derby/mdb
3   -- Connection URL is: jdbc:derby:/home/carl/work/repos/hive1/metastore/scripts/upgrade/derby/mdb
4   -- Specified schema is: APP
5   -- appendLogs: false
6
7
8   -- DDL Statements for functions
9
10
11  -- CREATE FUNCTION "APP"."NUCLEUS_ASCII" (C CHAR(1)) RETURNS INTEGER LANGUAGE JAVA PARAMETER STYLE JAVA READS SQL DATA CALLED ON NULL INPUT EXTERNAL NAME 'org.datanucleus
12
13  -- CREATE FUNCTION "APP"."NUCLEUS_MATCHES" (TEXT VARCHAR(8000),PATTERN VARCHAR(8000)) RETURNS INTEGER LANGUAGE JAVA PARAMETER STYLE JAVA READS SQL DATA CALLED ON NULL IN
14
15
16   -- DDL Statements for tables
17
18  CREATE TABLE "APP"."DBS" (
19    "DB_ID" BIGINT NOT NULL,
20    "DESC" VARCHAR(4000),
21    "DB_LOCATION_URL" VARCHAR(4000) NOT NULL,
22    "NAME" VARCHAR(128),
23    "OWNER_NAME" VARCHAR(128),
24    "OWNER_TYPE" VARCHAR(10),
25    "CTLG_NAME" VARCHAR(256) NOT NULL DEFAULT 'hive'
26  );
27
28  CREATE TABLE "APP"."TBL_PRIVS" ("TBL_GRANT_ID" BIGINT NOT NULL, "CREATE_TIME" INTEGER NOT NULL, "GRANT_OPTION" SMALLINT NOT NULL, "GRANTOR" VARCHAR(128), "GRANTOR_TYPE" V
29
30  CREATE TABLE "APP"."DATABASE_PARAMS" ("DB_ID" BIGINT NOT NULL, "PARAM_KEY" VARCHAR(180) NOT NULL, "PARAM_VALUE" VARCHAR(4000));
31
32  CREATE TABLE "APP"."TBL_COL_PRIVS" ("TBL_COLUMN_GRANT_ID" BIGINT NOT NULL, "COLUMN_NAME" VARCHAR(767), "CREATE_TIME" INTEGER NOT NULL, "GRANT_OPTION" SMALLINT NOT NULL, "
33
34  CREATE TABLE "APP"."SERDE_PARAMS" ("SERDE_ID" BIGINT NOT NULL, "PARAM_KEY" VARCHAR(256) NOT NULL, "PARAM_VALUE" CLOB);
35
36  CREATE TABLE "APP"."COLUMNS_V2" ("CD_ID" BIGINT NOT NULL, "COMMENT" VARCHAR(4000), "COLUMN_NAME" VARCHAR(767) NOT NULL, "TYPE_NAME" CLOB, "INTEGER_IDX" INTEGER NOT NULL);
37
38  CREATE TABLE "APP"."SORT_COLS" ("SD_ID" BIGINT NOT NULL, "COLUMN_NAME" VARCHAR(767), "ORDER" INTEGER NOT NULL, "INTEGER_IDX" INTEGER NOT NULL);

```

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
```



```

Alyssa@DESKTOP-K6T6ETI ~
$ $HIVE_HOME/bin/schematool -dbType derby -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop-env/apache-hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop-env/hadoop-3.3.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Metastore connection URL:      jdbc:derby://localhost:1527/metastore_db;create=true
Metastore Connection Driver :  org.apache.derby.jdbc.ClientDriver
Metastore connection User:    APP
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.derby.sql

Initialization script completed
schemaTool completed

Alyssa@DESKTOP-K6T6ETI ~
$ ...

```

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:

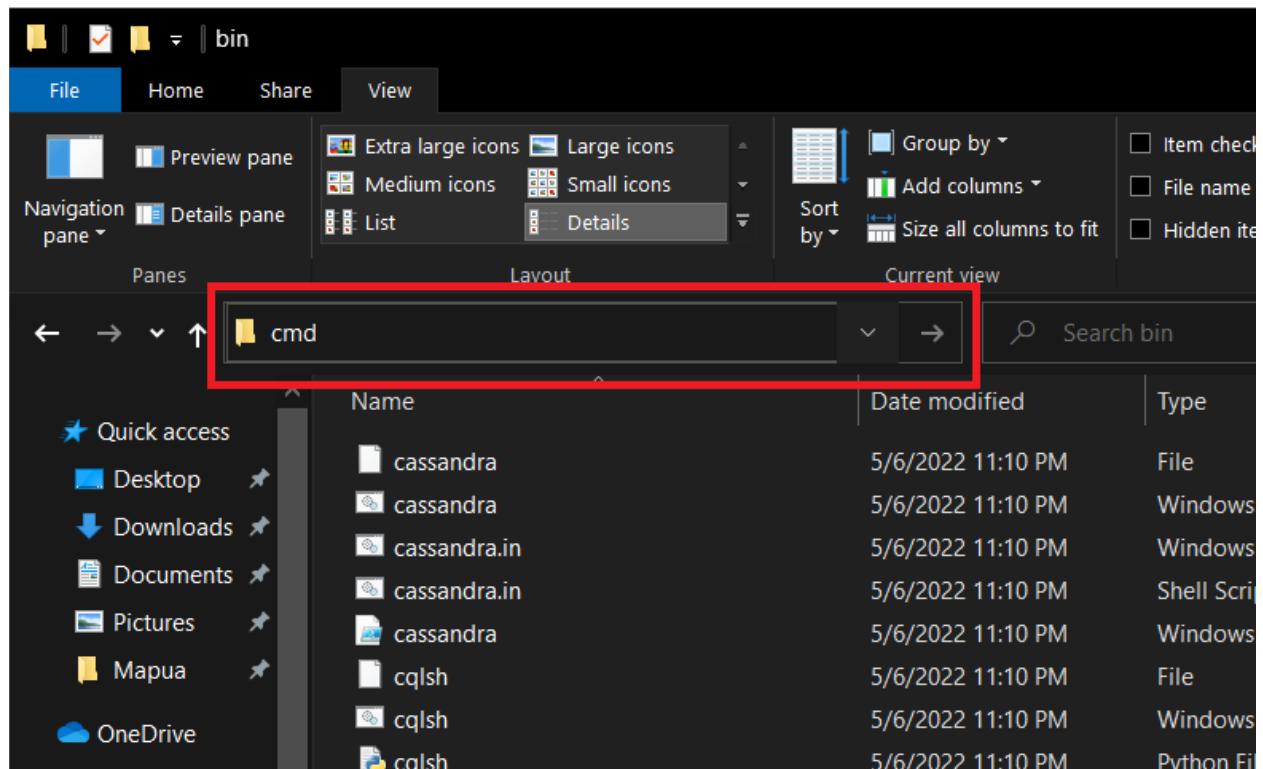
```
$HIVE_HOME/bin/beeline -u jdbc:hive2://
```

```
Allyssa@DESKTOP-K6T6FTT ~
$ $HIVE_HOME/bin/beeline -u jdbc:hive2://
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop-env/apache-hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop-env/hadoop-3.3.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
ps: unknown option -- o
Try 'ps --help' for more information.
ps: unknown option -- o
Try 'ps --help' for more information.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop-env/apache-hive-3.1.2/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
ull yet this is not valid. Ignored
22/09/30 11:13:19 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:19 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:19 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:22 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:22 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:22 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:22 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
22/09/30 11:13:22 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2 by Apache Hive
0: 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`

```
C:\Windows\System32\cmd.exe - cassandra
Microsoft Windows [Version 10.0.19043.1889]
(c) Microsoft Corporation. All rights reserved.

C:\apache-cassandra-3.11.13\bin>cassandra
WARNING! Powershell script execution unavailable.
Please use 'powershell Set-ExecutionPolicy Unrestricted'
on this user-account to run cassandra with fully featured
functionality on this platform.
Starting with legacy startup options
Starting Cassandra Server
INFO  [main] 2022-09-14 09:55:01,327 YamlConfigurationLoader.java:93 - Configuration location: file:/C:/apache-cassandra-3.11.13/conf/cassandra.yaml
```

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

```

C:\Windows\System32\cmd.exe - cqlsh
Microsoft Windows [Version 10.0.19043.1889]
(c) Microsoft Corporation. All rights reserved.

C:\apache-cassandra-3.11.13\bin>[cqlsh]

WARNING: console codepage must be set to cp65001 to support utf-8 encoding on Windows platforms.
If you experience encoding problems, change your console codepage with 'chcp 65001' before starting cqlsh.

Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.13 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing. Install to enable tab completion.
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; 
```

timeuuid	id	bike	bus	car	date_saved	jeepney	lgu_code	others	sensor_id	time_saved	total	truck	tryke
1663253925.81		5	0	0	09/15/2022	2	1200	0	sensor_09	22:58:45	9	1	1
1663252815.85		4	1	3	09/15/2022	2	1200	2	sensor_06	22:40:15	15	2	1
1663254226.61		2	2	0	09/15/2022	2	1200	2	sensor_02	23:03:46	9	1	0
1663252901.27		3	2	2	09/15/2022	2	1200	0	sensor_07	22:41:41	12	0	3
1663255098.94		3	2	3	09/15/2022	2	1200	0	sensor_10	23:18:18	14	1	3
1663253591.8		4	0	4	09/15/2022	2	1200	1	sensor_03	22:53:11	13	1	1
1663252942.96		1	0	3	09/15/2022	2	1200	0	sensor_05	22:42:22	8	1	1
1663253943.09		4	0	0	09/15/2022	0	1200	1	sensor_04	22:59:03	8	1	2
1663255150.41		5	2	2	09/15/2022	2	1200	1	sensor_02	23:19:10	16	2	2
1663254720.77		2	0	2	09/15/2022	0	1200	2	sensor_08	23:12:00	10	1	3
1663253832.11		0	2	3	09/15/2022	2	1200	0	sensor_06	22:57:12	10	2	1
1663254227.63		1	2	2	09/15/2022	1	1200	1	sensor_02	23:03:47	9	0	2
1663253961.28		4	2	0	09/15/2022	0	1200	2	sensor_01	22:59:21	11	0	3

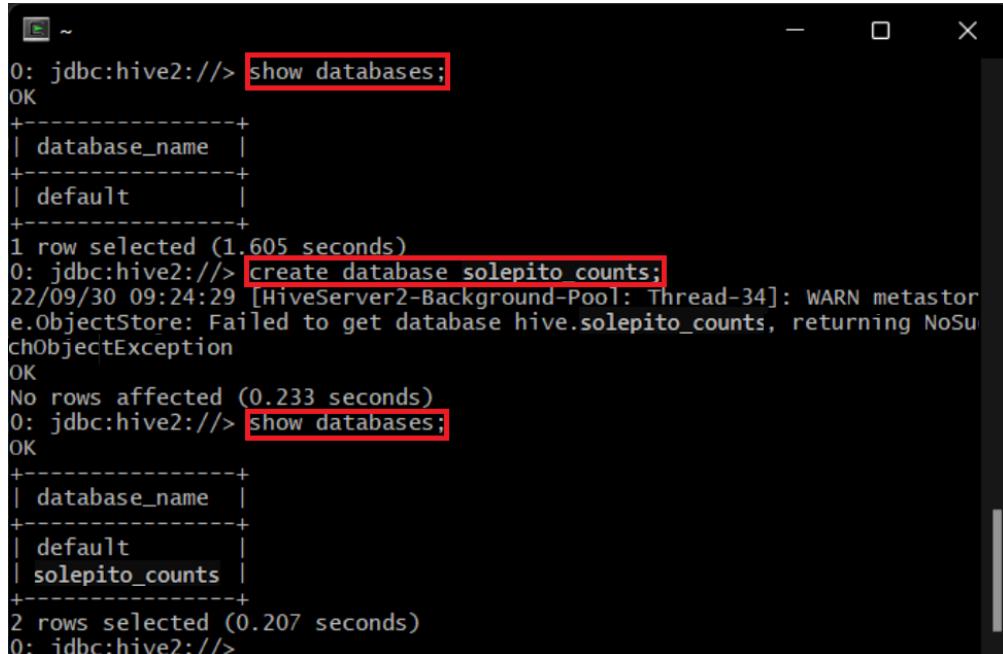
Since Hive has already been opened through the previous steps, the next step is to create the Python program that will migrate the data from the `group23_project_table` from Cassandra to a new table in Hive called `solepito_counts.sumry`.

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;
```



The screenshot shows a terminal window with the following session:

```
0: jdbc:hive2://> show databases;
OK
+-----+
| database_name |
+-----+
| default      |
+-----+
1 row selected (1.605 seconds)
0: jdbc:hive2://> create database solepito_counts;
22/09/30 09:24:29 [HiveServer2-Background-Pool: Thread-34]: WARN metastore.ObjectStore: Failed to get database hive.solepito_counts, returning NoSuchObjectException
OK
No rows affected (0.233 seconds)
0: jdbc:hive2://> show databases;
OK
+-----+
| database_name |
+-----+
| default      |
| solepito_counts |
+-----+
2 rows selected (0.207 seconds)
0: jdbc:hive2://>
```

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*.



The screenshot shows a Windows Command Prompt window titled "Administrator: Command Prompt". The window displays the output of a Python script named "cassandra2hive.py". The script uses Spark's default log4j profile and sets the log level to "WARN". It also prints the Python version (3.10.6) and the Spark context information. There are several warning messages related to connecting to a Node and sending requests. The script ends with a note about setting values on a copy of a DataFrame and a link to the pandas documentation.

```

Administrator: Command Prompt
Microsoft Windows [Version 10.0.22000.978]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ASUS>python C:\Users\ASUS\Documents\M2-Assignment-1\cassandra2hive.py
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

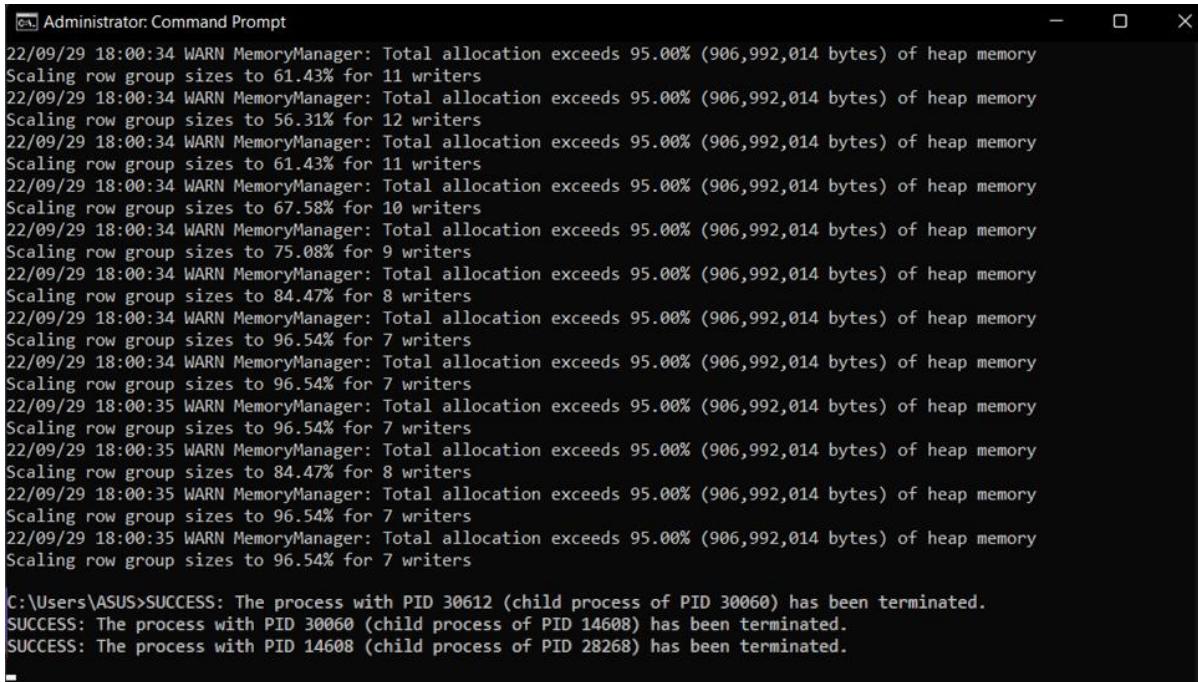
   / \ \ / \ / \ \ / \ / \
  / \ \ / \ / \ \ / \ / \ \ / \
 / \ \ / \ / \ \ / \ / \ \ / \
/ \ \ / \ / \ \ / \ / \ \ / \
version 3.2.2

Using Python version 3.10.6 (tags/v3.10.6:9c7b4bd, Aug 1 2022 21:53:49)
Spark context Web UI available at http://Xenon:4040
Spark context available as 'sc' (master = local[*], app id = local-1664445605285).
SparkSession available as 'spark'.
22/09/29 18:00:12 WARN ControlConnection: [s0] Error connecting to Node(endPoint=localhost/0:0:0:0:0:1:9042, hostId=null, hashCode=2f224550), trying next node (ConnectionInitException: [s0|control|connecting...] Protocol initialization request, step 1 (OPTIONS): failed to send request (com.datastax.oss.driver.shaded.netty.channel.StacklessClosedChannelException)
22/09/29 18:00:15 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped
C:\Users\ASUS>python C:\Users\ASUS\Documents\M2-Assignment-1\cassandra2hive.py:34: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```

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.



```

Administrator: Command Prompt
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 61.43% for 11 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 56.31% for 12 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 61.43% for 11 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 67.58% for 10 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 75.08% for 9 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 84.47% for 8 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 96.54% for 7 writers
22/09/29 18:00:34 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 96.54% for 7 writers
22/09/29 18:00:35 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 96.54% for 7 writers
22/09/29 18:00:35 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 84.47% for 8 writers
22/09/29 18:00:35 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 96.54% for 7 writers
22/09/29 18:00:35 WARN MemoryManager: Total allocation exceeds 95.00% (906,992,014 bytes) of heap memory
Scaling row group sizes to 96.54% for 7 writers

C:\Users\ASUS>SUCCESS: The process with PID 30612 (child process of PID 30060) has been terminated.
SUCCESS: The process with PID 30060 (child process of PID 14608) has been terminated.
SUCCESS: The process with PID 14608 (child process of PID 28268) has been terminated.

```

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;
```

```
0: jdbc:hive2://> select * from solepito_counts.sumry;
OK
22/09/30 09:28:55 [792b6c27-9e06-44f3-aa7a-1cf0618a6eb6 m
0000-d9b6b4b0-41f1-4486-b798-c127d5218bcd-c000_snappy.par
```

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	1133	259	1	249	116	106	106	180
2022-09-15 00:00:00.0	23	13	4990	1217	512	1019	189	505	507	142
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/>