Installing Hadoop 3.2.1 On Windows :

1.Prerequisites :

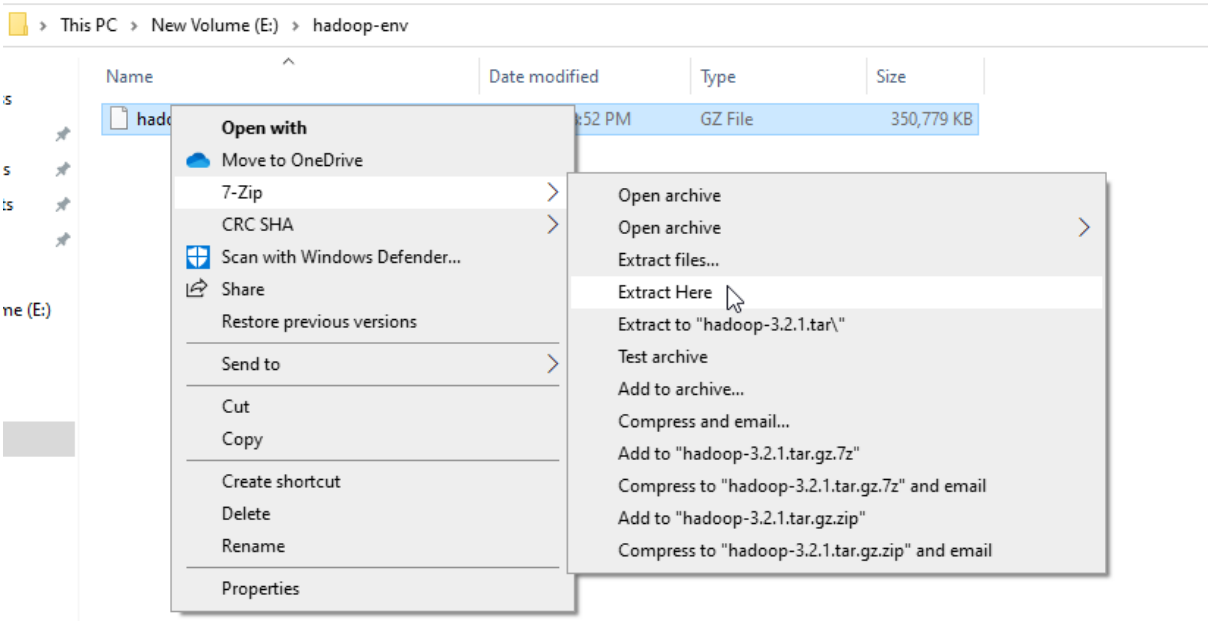
First, we need to make sure that the following prerequisites are installed:

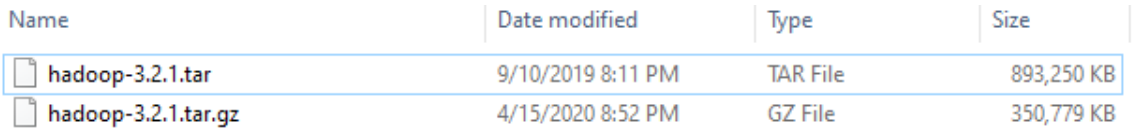
1. Java 8 runtime environment (JRE): Hadoop 3 requires a Java 8 installation. I prefer using the offline installer.
2. Java 8 development Kit (JDK)
3. To unzip downloaded Hadoop binaries, we should install 7zip.
4. I will create a folder "E:\hadoop-env" on my local machine to store downloaded files.

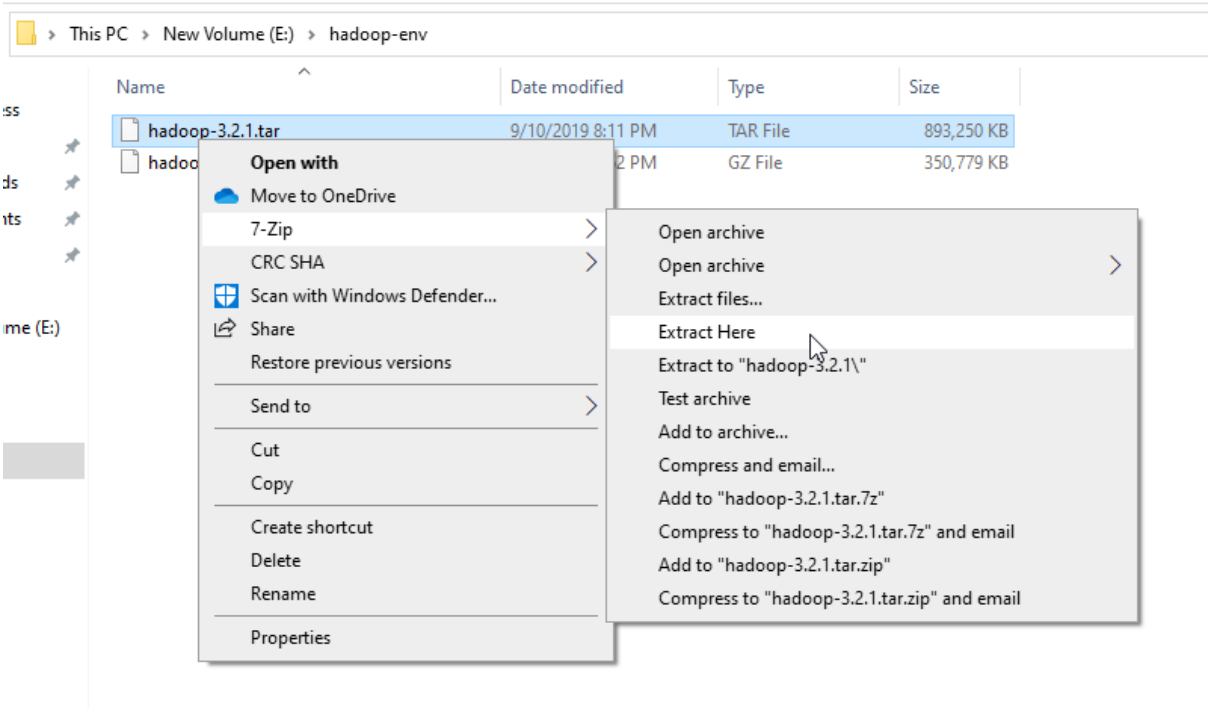
2. Download Hadoop binaries :

The first step is to download Hadoop binaries from the official website. The binary package size is about 342 MB.

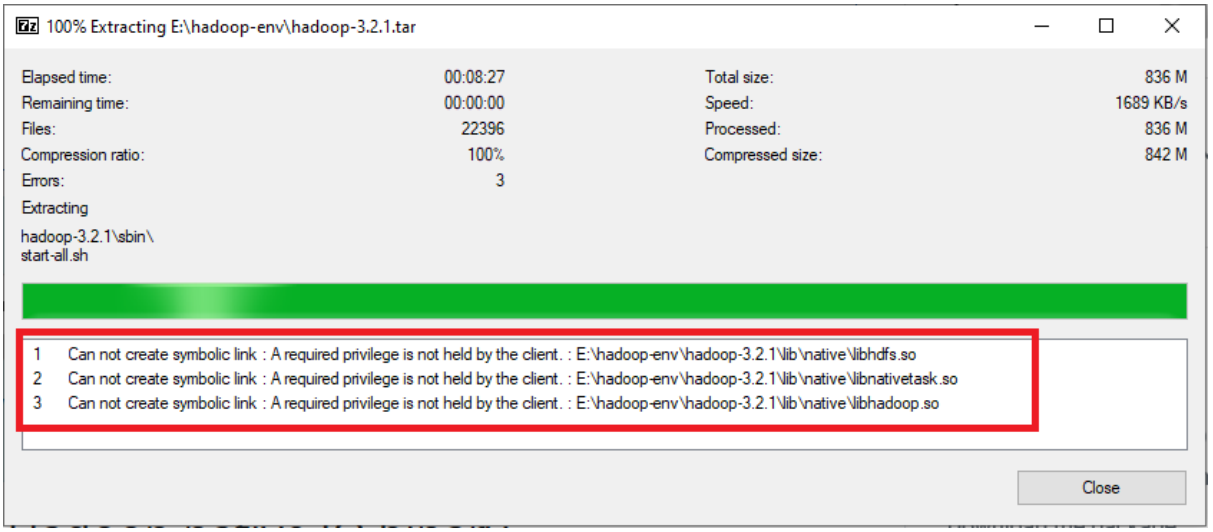
After finishing the file download, we should unpack the package using 7zip int two steps. First, we should extract the hadoop-3.2.1.tar.gz library, and then, we should unpack the extracted tar file:


This PC › New Volume (E:) › hadoop-env
Name
Date modified
Type
Size
GZ File
350,779 KB
Open with
Move to OneDrive
7-Zip
CRC SHA
Scan with Windows Defender...
Share
Restore previous versions
Send to
Cut
Copy
Create shortcut
Delete
Rename
Properties
Open archive
Open archive
Extract files...
Extract Here
Extract to "hadoop-3.2.1.tar\"
Test archive
Add to archive...
Compress and email...
Add to "hadoop-3.2.1.tar.gz.7z"
Compress to "hadoop-3.2.1.tar.gz.7z" and email
Add to "hadoop-3.2.1.tar.gz.zip"
Compress to "hadoop-3.2.1.tar.gz.zip" and email


Name
Date modified
Type
Size
hadoop-3.2.1.tar
9/10/2019 8:11 PM
TAR File
893,250 KB
hadoop-3.2.1.tar.gz
4/15/2020 8:52 PM
GZ File
350,779 KB


This PC › New Volume (E:) › hadoop-env
Name
Date modified
Type
Size
hadoop-3.2.1.tar
9/10/2019 8:11 PM
TAR File
893,250 KB
GZ File
350,779 KB
Open with
Move to OneDrive
7-Zip
CRC SHA
Scan with Windows Defender...
Share
Restore previous versions
Send to
Cut
Copy
Create shortcut
Delete
Rename
Properties
Open archive
Open archive
Extract files...
Extract Here
Extract to "hadoop-3.2.1\"
Test archive
Add to archive...
Compress and email...
Add to "hadoop-3.2.1.tar.7z"
Compress to "hadoop-3.2.1.tar.7z" and email
Add to "hadoop-3.2.1.tar.zip"
Compress to "hadoop-3.2.1.tar.zip" and email

The tar file extraction may take some minutes to finish. In the end, you may see some warnings about symbolic link creation. Just ignore these warnings since they are not related to windows.



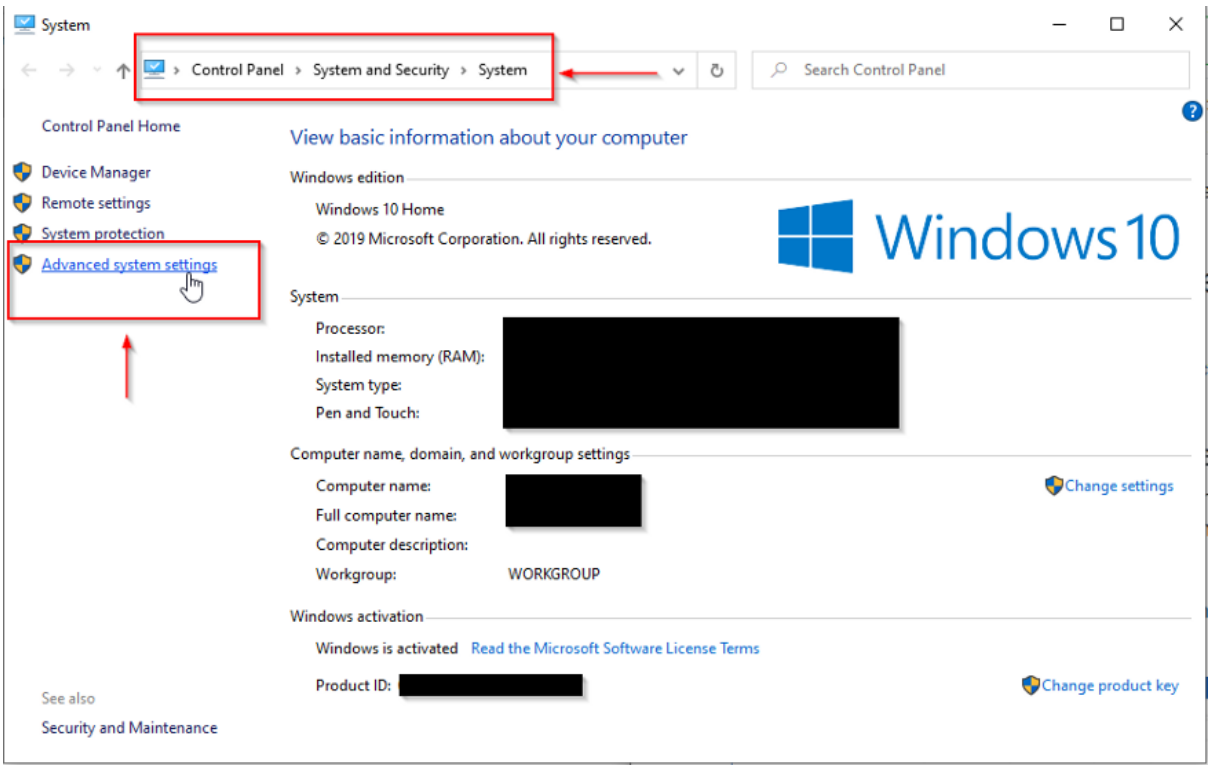
After unpacking the package, we should add the Hadoop native IO libraries, which can be found in the following GitHub repository: https://github.com/cdarlint/winutils.

Since we are installing Hadoop 3.2.1, we should download the files located in https://github.com/cdarlint/winutils/tree/master/hadoop-3.2.1/bin and copy them into the "hadoop-3.2.1\bin" directory.

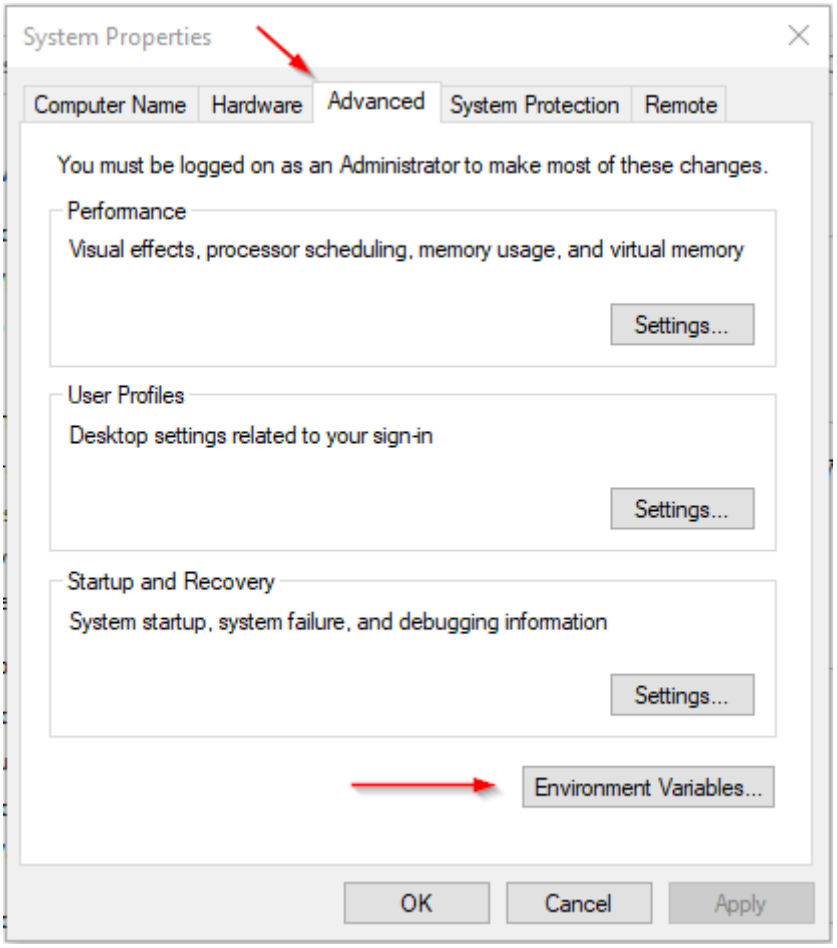
3. Setting up environment variables

After installing Hadoop and its prerequisites, we should configure the environment variables to define Hadoop and Java default paths.

To edit environment variables, go to Control Panel > System and Security > System (or right-click > properties on My Computer icon) and click on the "Advanced system settings" link.



When the “Advanced system settings” dialog appears, go to the “Advanced” tab and click on the “Environment variables” button located on the bottom of the dialog.



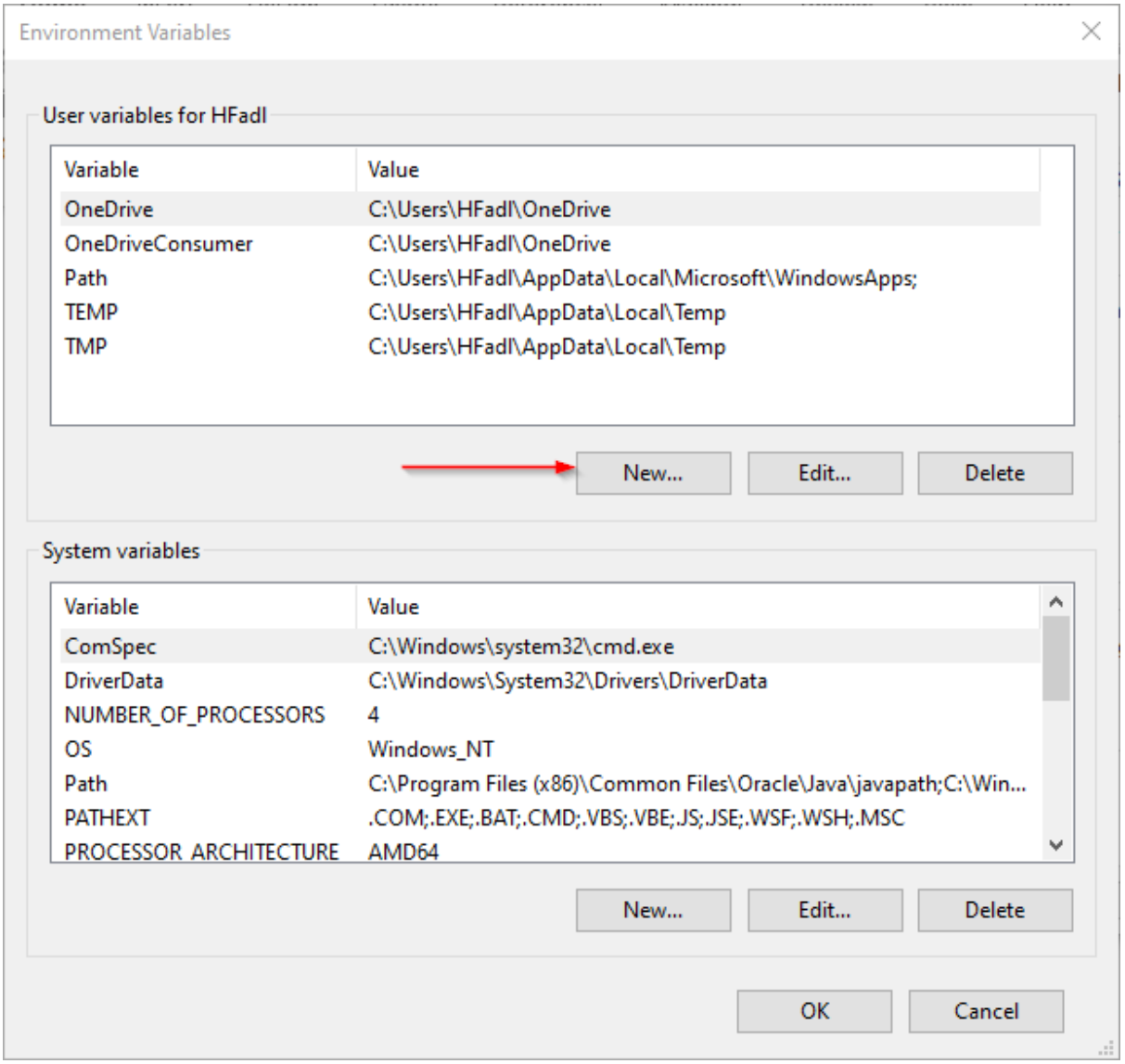
In the “Environment Variables” dialog, press the “New” button to add a new variable.

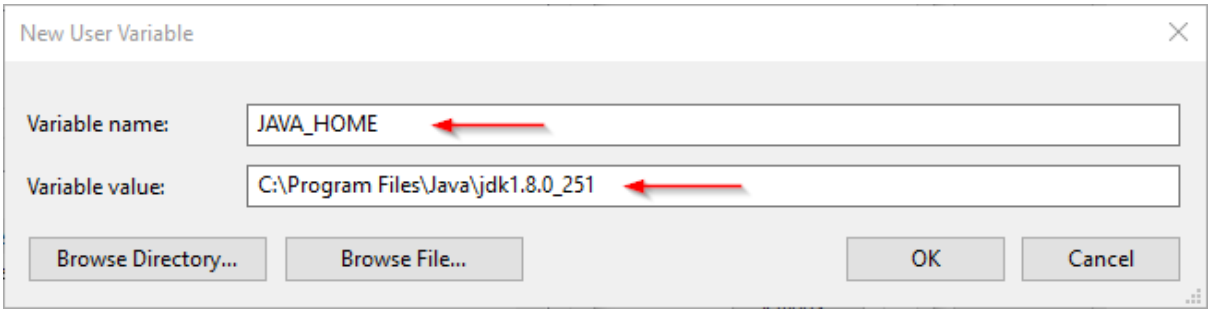
Note: In this guide, we will add user variables since we are configuring Hadoop for a single user. If you are looking to configure Hadoop for multiple users, you can define System variables instead.

There are two variables to define:

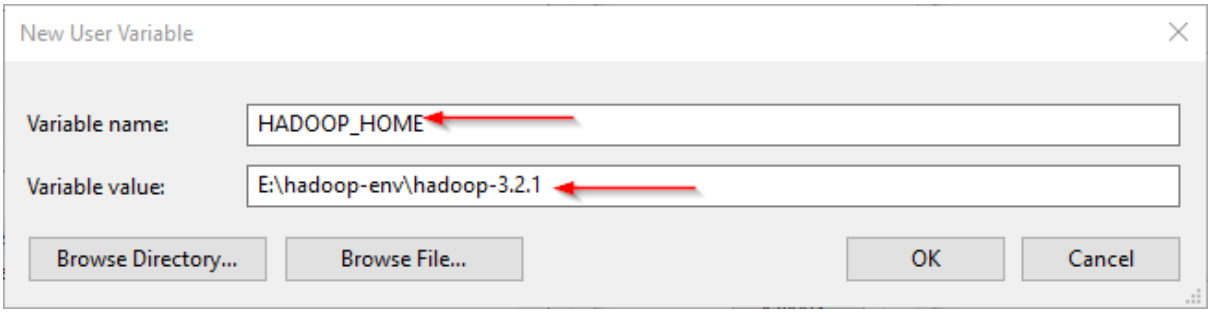
JAVA_HOME: JDK installation folder path

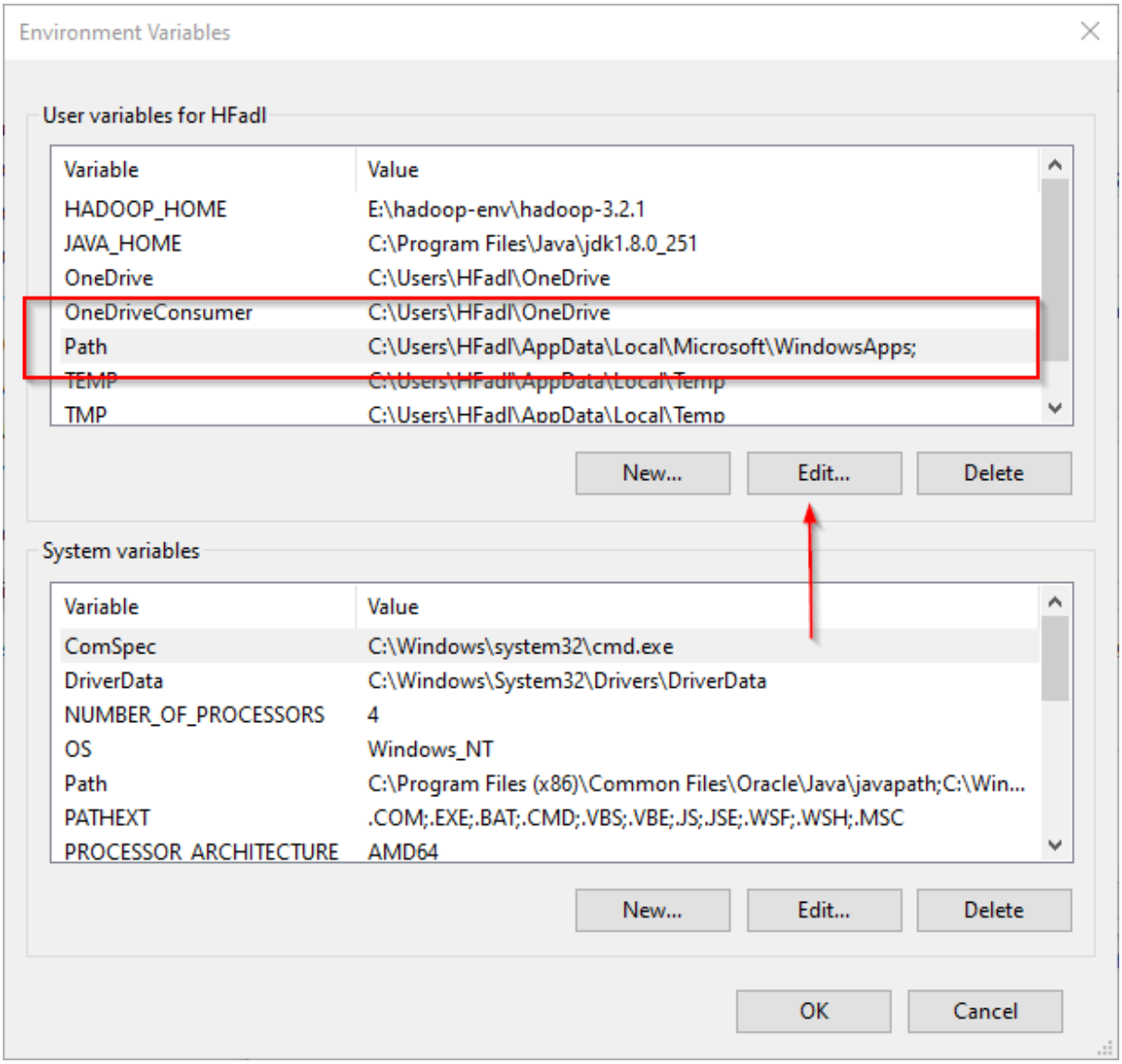
HADOOP_HOME: Hadoop installation folder path

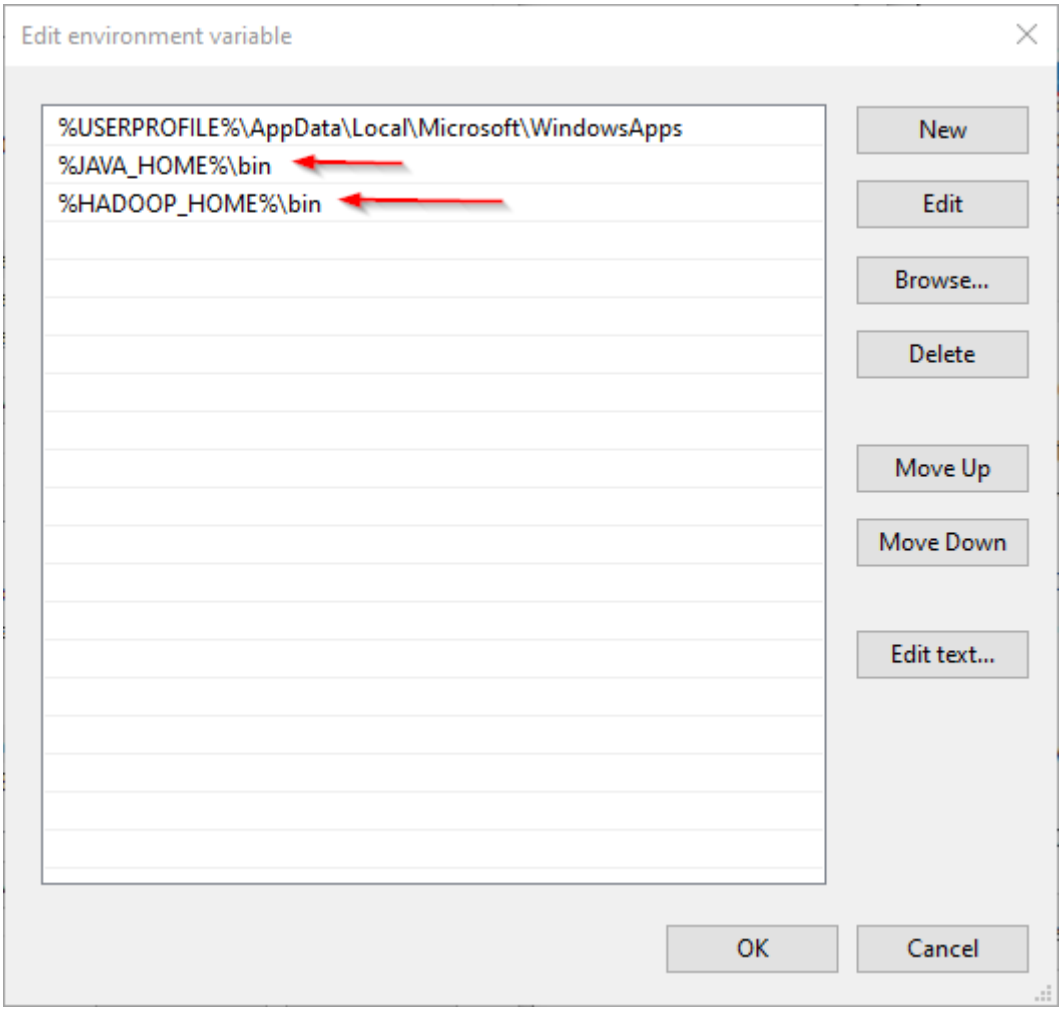




Now, we should edit the PATH variable to add the Java and Hadoop binaries paths as shown in the following screenshots.


New User Variable
Variable name:
HADOOP_HOME
Variable value:
E:\hadoop-env\hadoop-3.2.1
Browse Directory...
Browse File...
OK
Cancel


Environment Variables
User variables for HFadl
Variable
Value
HADOOP_HOME
E:\hadoop-env\hadoop-3.2.1
JAVA_HOME
C:\Program Files\Java\jdk1.8.0_251
OneDrive
C:\Users\HFadl\OneDrive
OneDriveConsumer
C:\Users\HFadl\OneDrive
Path
C:\Users\HFadl\AppData\Local\Microsoft\WindowsApps;
TEMP
C:\Users\HFadl\AppData\Local\Temp
TMP
C:\Users\HFadl\AppData\Local\Temp
New...
Edit...
Delete
System variables
Variable
Value
ComSpec
C:\Windows\system32\cmd.exe
DriverData
C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS
4
OS
Windows_NT
Path
C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Win...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE
AMD64
New...
Edit...
Delete
OK
Cancel



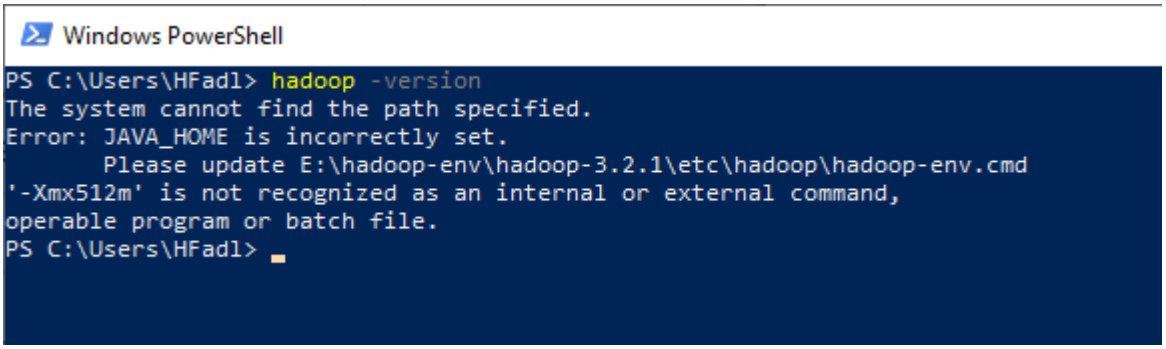
3.1. JAVA_HOME is incorrectly set error

Now, let's open PowerShell and try to run the following command:

```
hadoop -version
```

In this example, since the JAVA_HOME path contains spaces, I received the following error:

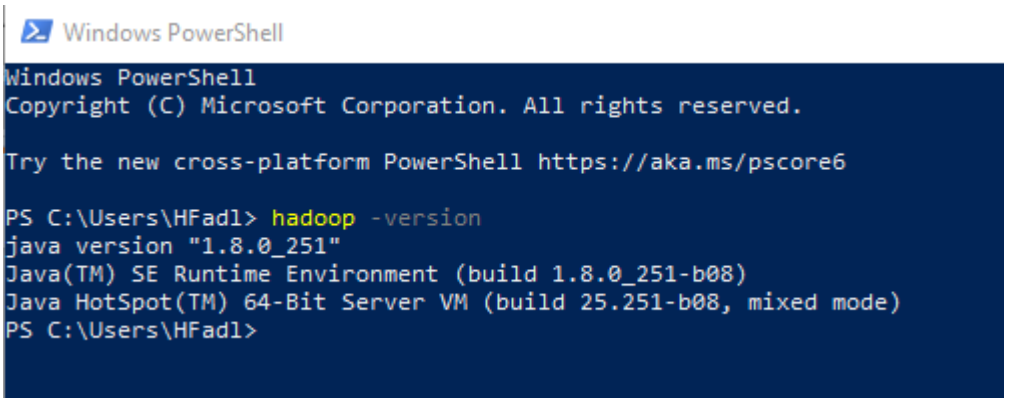
```
JAVA_HOME is incorrectly set
```



To solve this issue, we should use the windows 8.3 path instead. As an example:

- Use “Progra~1” instead of “Program Files”
- Use “Progra~2” instead of “Program Files(x86)”

After replacing “Program Files” with “Progra~1”, we closed and reopened PowerShell and tried the same command. As shown in the screenshot below, it runs without errors.



4. Configuring Hadoop cluster

There are four files we should alter to configure Hadoop cluster:

1. %HADOOP_HOME%\etc\hadoop\hdfs-site.xml
2. %HADOOP_HOME%\etc\hadoop\core-site.xml
3. %HADOOP_HOME%\etc\hadoop\mapred-site.xml
4. %HADOOP_HOME%\etc\hadoop\yarn-site.xml

4.1. HDFS site configuration

As we know, Hadoop is built using a master-slave paradigm. Before altering the HDFS configuration file, we should create a directory to store all master node (name node) data and another one to store data (data node). In this example, we created the following directories:

- E:\hadoop-env\hadoop-3.2.1\data\dfs\namenode
- E:\hadoop-env\hadoop-3.2.1\data\dfs\datanode

Now, let's open "hdfs-site.xml" file located in "%HADOOP_HOME%\etc\hadoop" directory, and we should add the following properties within the <configuration></configuration> element:

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:///E:/hadoop-env/hadoop-3.2.1/data/dfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:///E:/hadoop-env/hadoop-3.2.1/data/dfs/datanode</value>
</property>
```

4.2 Core site configuration

<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9820</value>
</property>

4.3 Map Reduce site configuration

<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
<description>MapReduce framework name</description>
</property>

4.4 Yarn site configuration

<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
<description>Yarn Node Manager Aux Service</description>
</property>