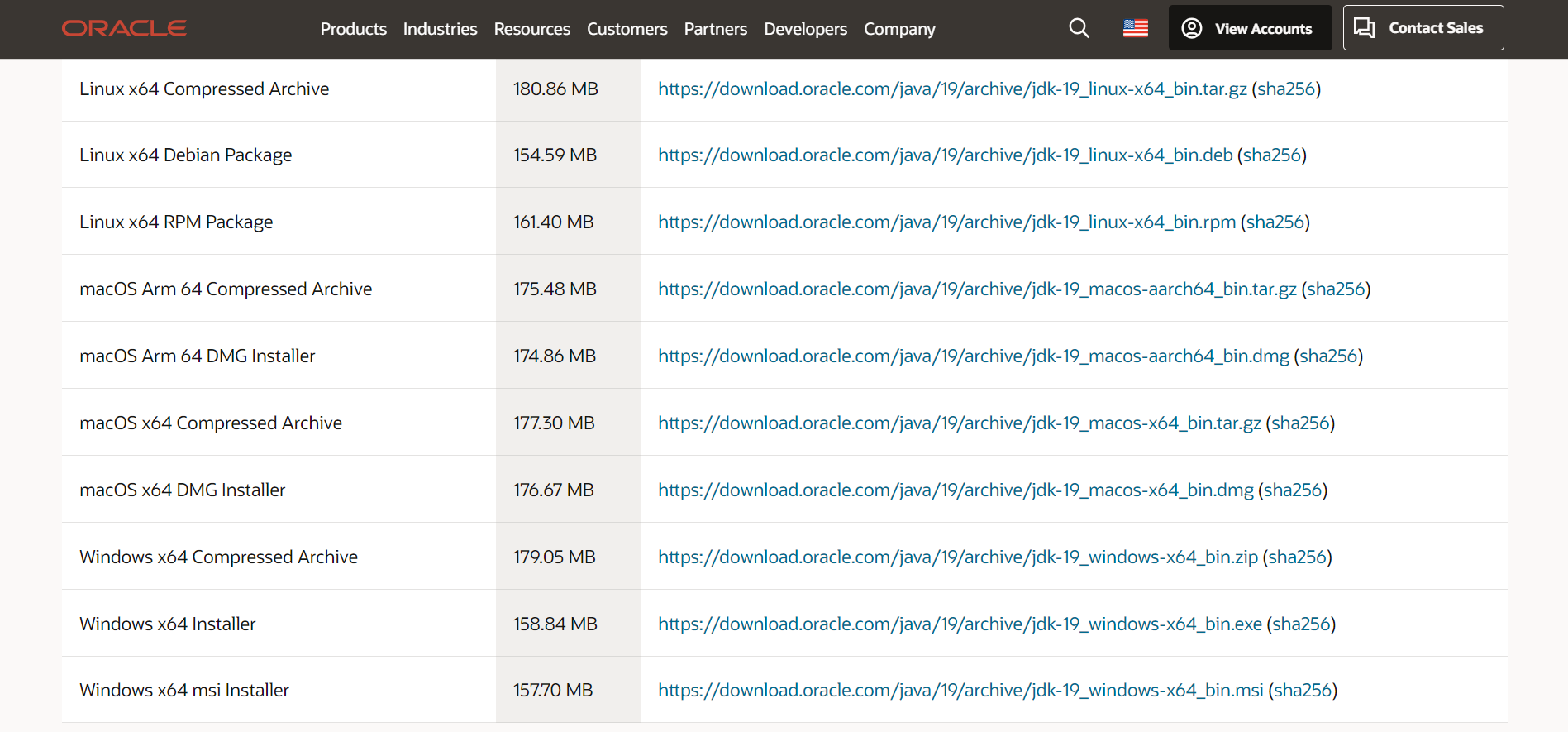
K. PRAVEEN KUMAR

ASSIGNMENT-1(PYSPARK)

**Installation of JDK(Java Development kit):**

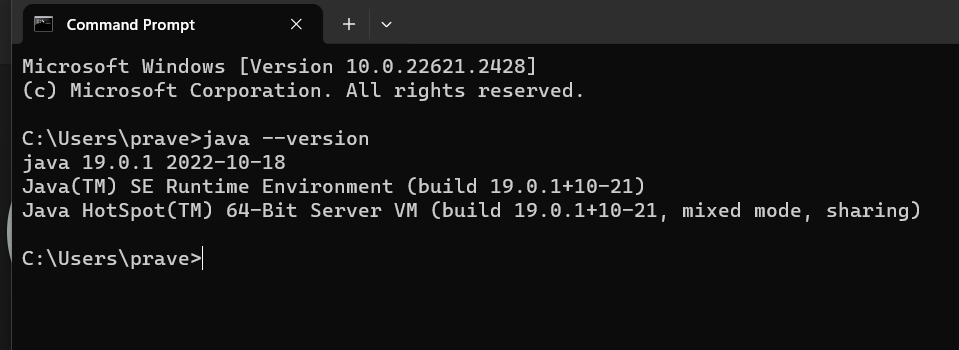
Visit the below link and download the jdk-19.exe file on your system.

<https://www.oracle.com/java/technologies/javase/jdk19-archive-downloads.html>

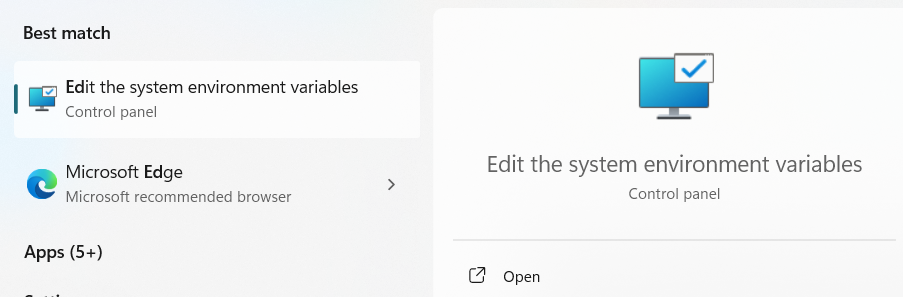


After downloading the file, open the file and the installation begins.

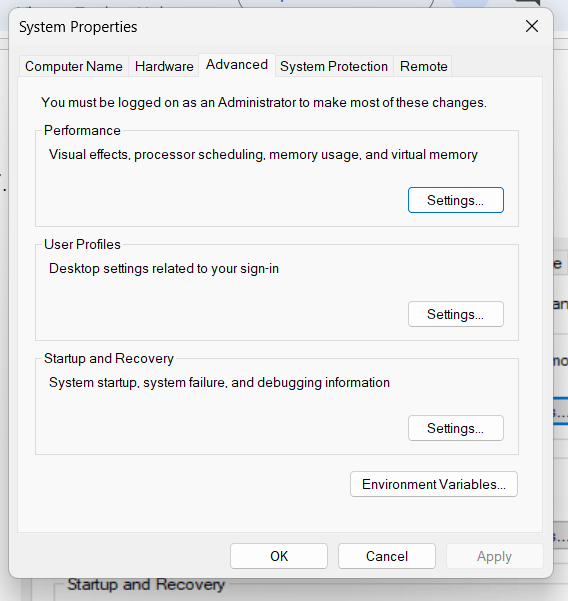
Go to the command prompt and type “java –version” to know the version and to check the java is installed or not.



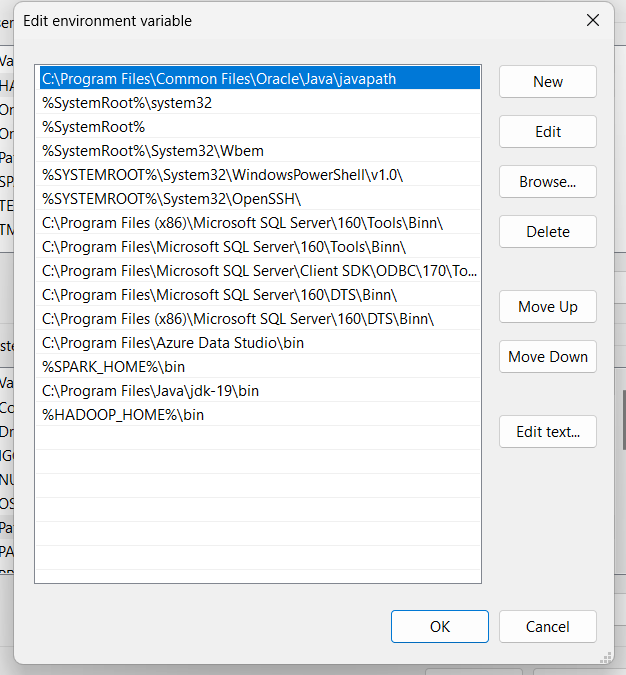
Go to the search bar and “Edit the environment variables”:



Click on the Environment Variables:



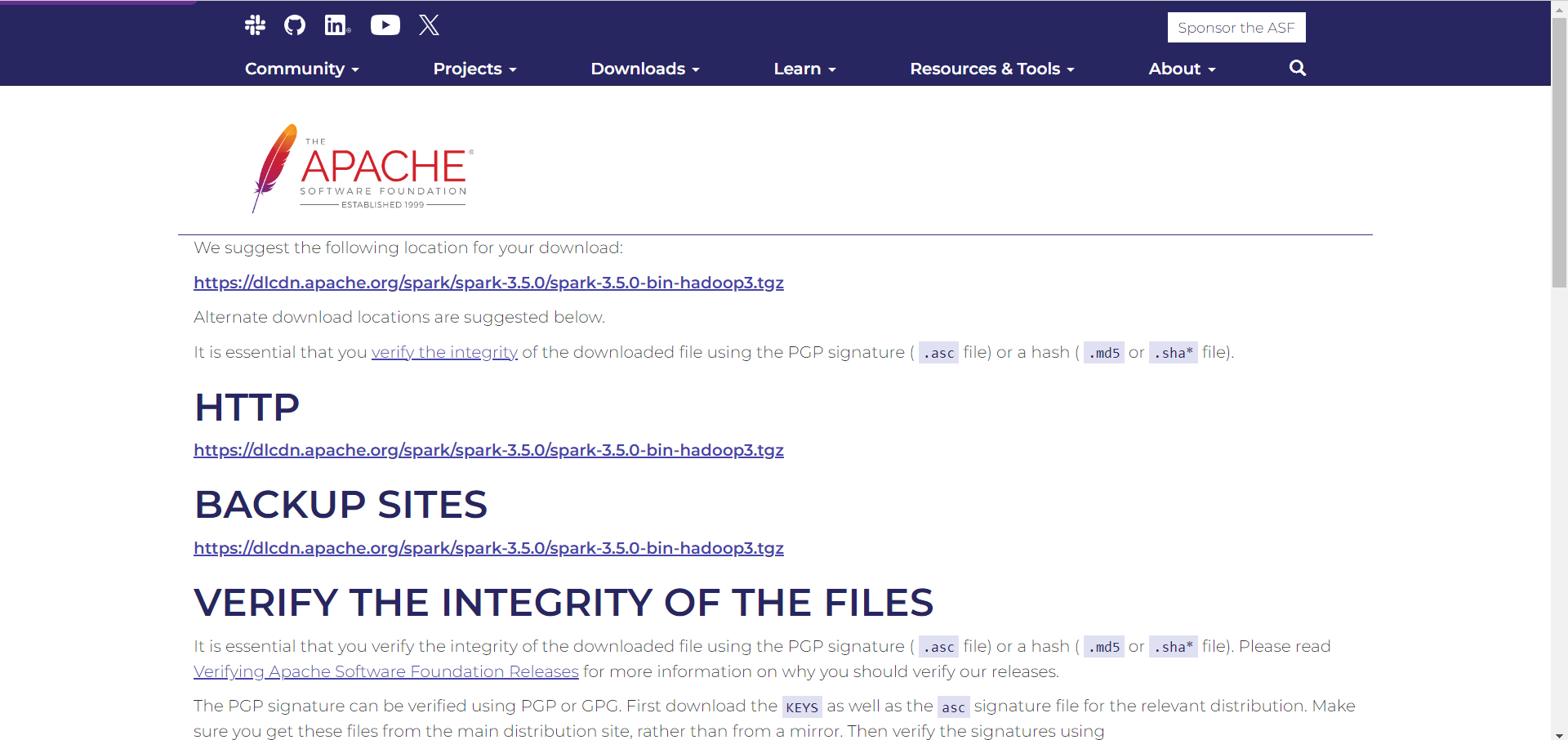
After opening the environment variables and then click on new and add the java path



**Installing Pyspark:**

Visit the below link and download the spark file with .tgz extension.

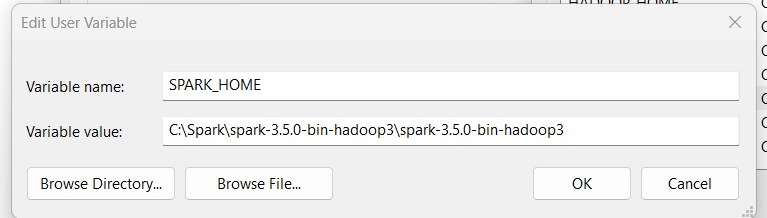
<https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz>

****

After installing the spark file, now go to C drive in local system and create a new folder called “Spark” and extract the downloaded zip file of spark in that folder.

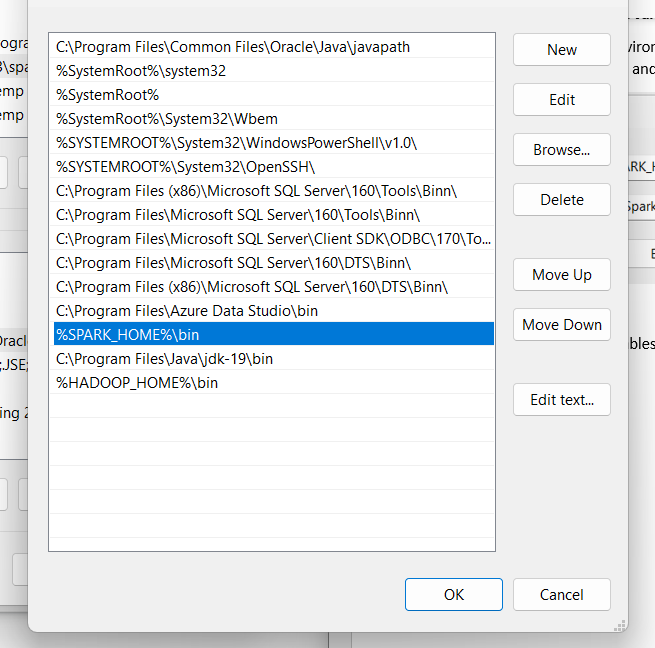
**Edit System Environment variables:**

Now open the system environment variables and click on ‘environment variables’ and then click ‘new’ in the user variable and give the “Variable name” and “variable value”:



Now click “OK”

Below in the system variables double click on “Path” and then click on “New” and add the path.



Click on “ok” and close the system environment variables.

Now open the command prompt and the give the command as “pip install Pyspark” and it will install the pyspark in your system.

**Downloading and setting up Winutils.exe:**

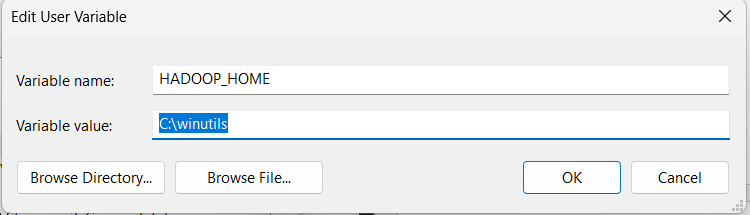
We need to download the winutils.exe file from the below github link.

<https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe>

After downloading the winutils.exe file, we need to create a folder “Winutils” inside the C drive and we need to create the folder named as “bin” inside the Wintuils folder. Now keep the downloaded file inside bin folder.

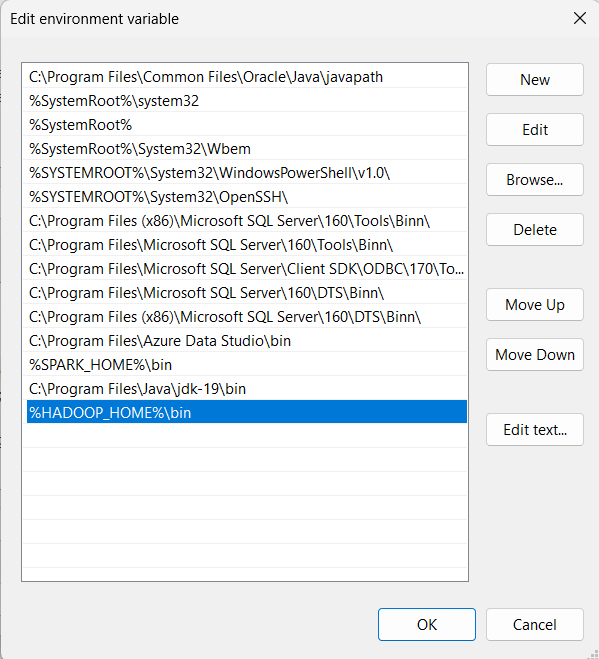
**Edit System Environment variables:**

Now open the system environment variables and click on ‘environment variables’ and then click ‘new’ in the user variable and give the “Variable name” and “variable value” as:



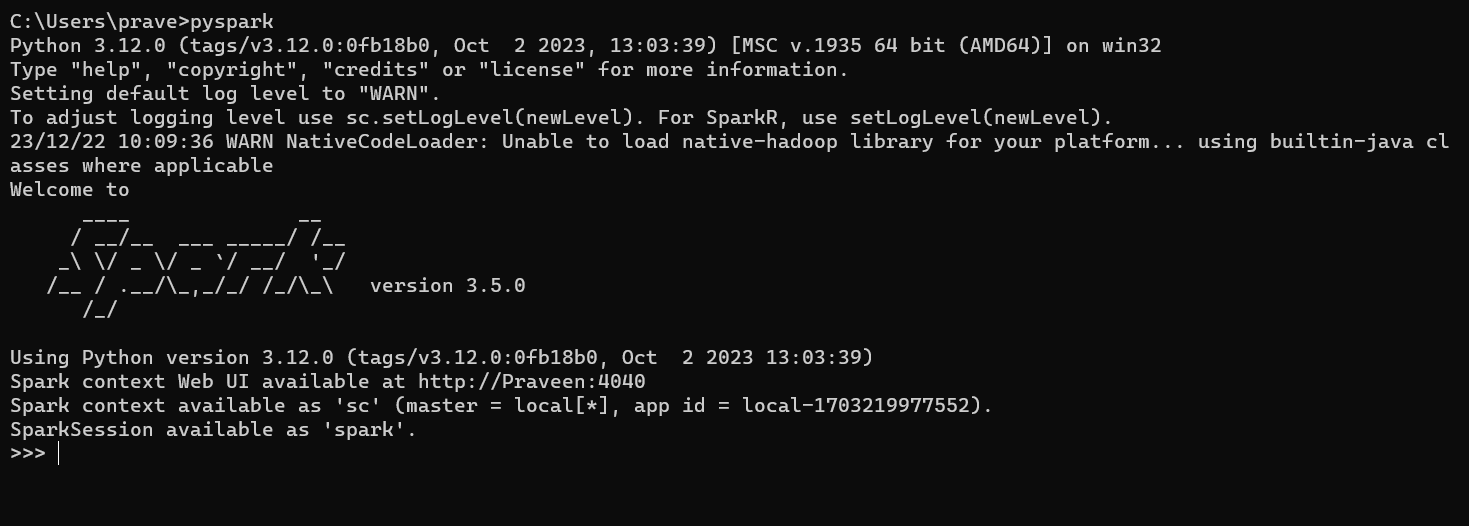
Now click “OK”

Below in the system variables double click on “Path” and then click on “New” and add the path.



Click on “ok” and close the system environment variables.

Open the command prompt and type “pyspark” to check whether it is installed or not.



Now the Spark is successfully installed and we need to check the webpage for this installed spark Is available or not by using the below link:

<http://localhost:4040/jobs/>

After opening the above link, we will get a webpage like this:

