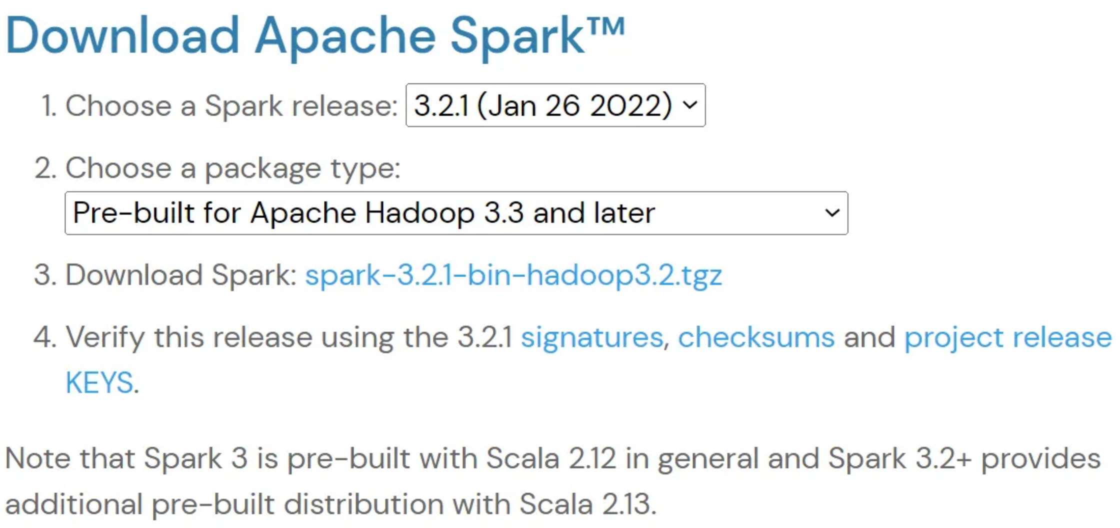
Name :- Nikhil Rajesh Bhanose
Student id:- 202318016

* Instructions for configuring PySpark in Windows.

* Required software language for installing Pyspark:
For the installation of PySpark we required following :

1. Jawa:-
   As Spark uses Java Virtual Machine internally, it has a dependency on JAVA. Install the latest version of the JAVA from here.
2. Python:-
   If you are going to work on a data science related project, I recommend you download Python and Jupyter Notebook together with the Anaconda Navigator.
   WE HAVE USE THE GOOGLE COLAB FOR THE ASSIGNMENT
3. PySpark



Download Apache Spark™

1. Choose a Spark release: 3.2.1 (Jan 26 2022)
2. Choose a package type: Pre-built for Apache Hadoop 3.3 and later
3. Download Spark: spark-3.2.1-bin-hadoop3.2.tgz
4. Verify this release using the 3.2.1 signatures, checksums and project release KEYS.

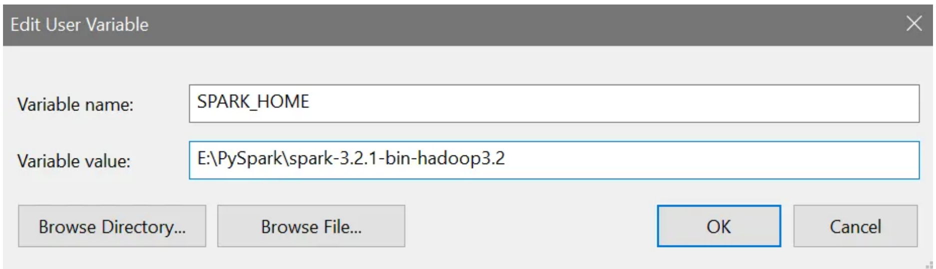
Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

4.Wintuils
In order to run Apache Spark locally, winutils.exe is required in the Windows Operating system**.** This is be*cause Spark needs elements of the Hadoop codebase called 'winutils' when it runs on non-windows clusters*. These windows utilities (winutils) help the management of the POSIX(Portable Operating System Interface) file system permissions that the HDFS (Hadoop Distributed File System) requires from the local (windows) file system.

Now installing PySpark , we need to set an environmental variables Environmental variables are the variables which are run through out the system during process
In order to set the environment variables

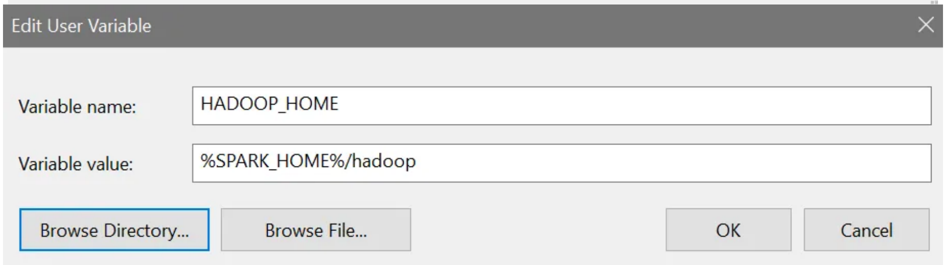
- Go to Windows search

- Type “**env**” —it will show the “edit environment variable for your account”, click on it
- Click on “**New**” for the user variables and add the following variable name and values (depending upon the location of the downloaded files



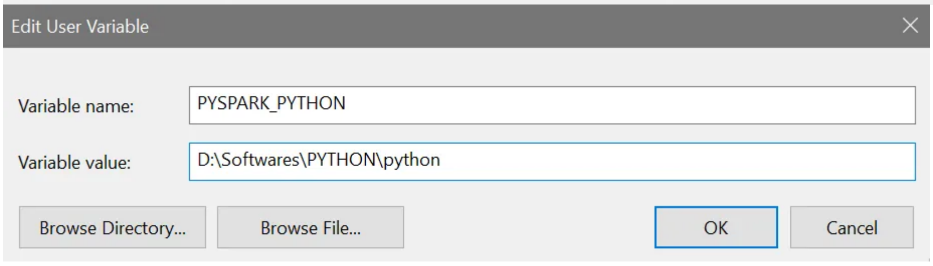
Variable name: SPARK_HOME

Variable value: E:\PySpark\spark-3.2.1-bin-hadoop3.2



Variable name: HADOOP_HOME

Variable value: %SPARK_HOME%/hadoop

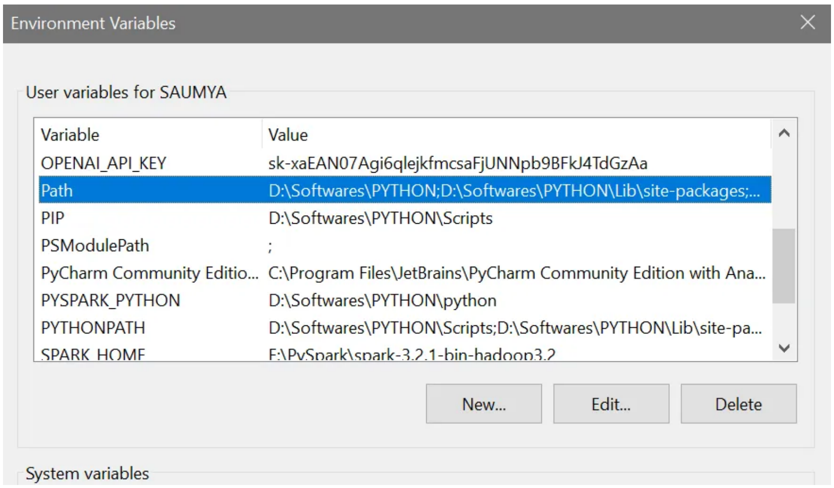


Variable name: PYSPARK_PYTHON

Variable value: D:\Softwares\PYTHON\python

Next, Update the PATH variable with the **\bin** folder address, containing the executable files of PySpark and Hadoop. This will help in executing Pyspark from the command prompt.

- Click on the “Path” variable



User variables for SAUMYA

Variable	Value
OPENAI_API_KEY	sk-xaEAN07Agi6qlejkfmcsaFjUNNpb9BFkJ4TdGzAa
Path	D:\Softwares\PYTHON;D:\Softwares\PYTHON\Lib\site-packages;...
PIP	D:\Softwares\PYTHON\Scripts
PSModulePath	;
PyCharm Community Editio...	C:\Program Files\JetBrains\PyCharm Community Edition with Ana...
PYSPARK_PYTHON	D:\Softwares\PYTHON\python
PYTHONPATH	D:\Softwares\PYTHON\Scripts;D:\Softwares\PYTHON\Lib\site-pa...
SPARK_HOME	E:\PySpark\spark-3.2.1-bin-hadoop3.2

Then add the following two values (we are using the previously defined Environment variables here)

%SPARK_HOME%\bin
%HADOOP_HOME%\bin

Here we install PySpark.
To run the PySpark go to command prompt
And type PySpark

*	Connecting the Spark and Python :-

As We used PySpark we can directly use the command in google cola .

Firstly install PySpark
Using command “pip install pyspark”
And We are good to go