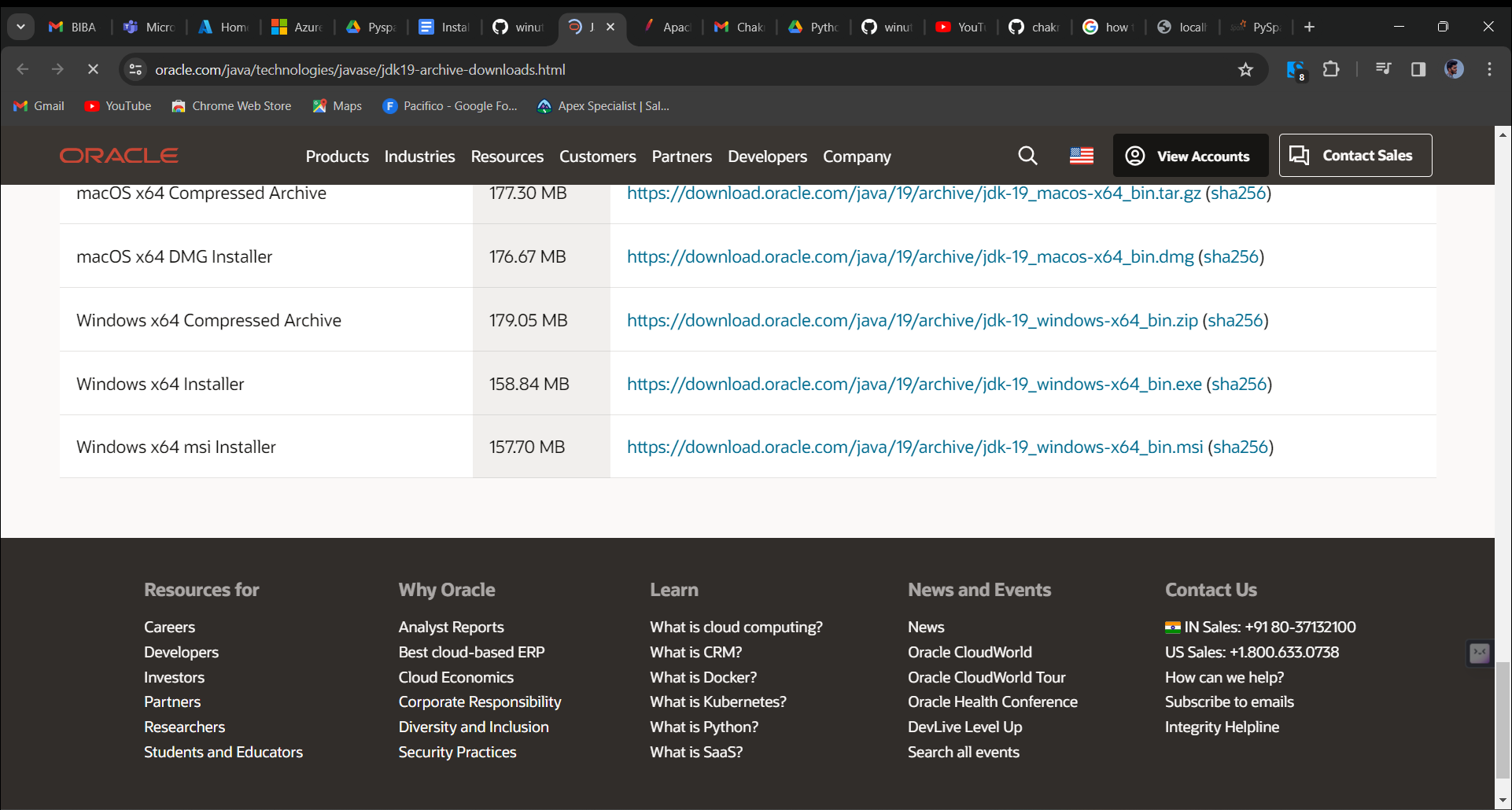
21/12/2023

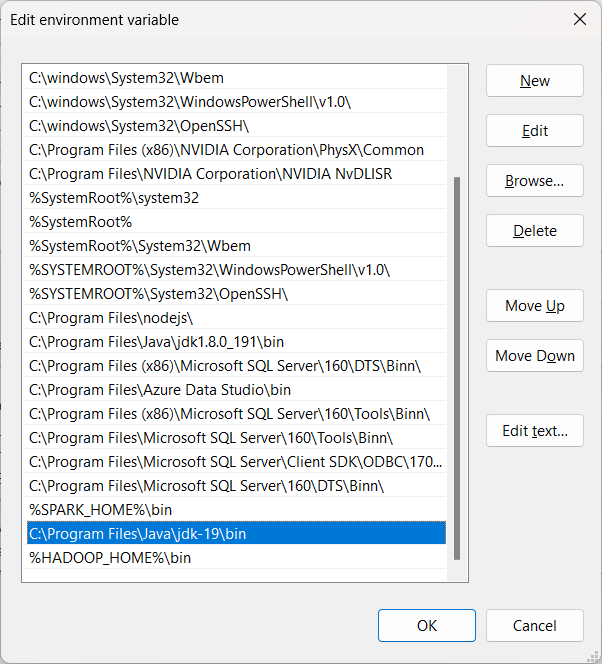
Chakradhar Bhogapurapu, BIBA

SPARK AND PYSPARK

Installation for JDK (Java Development Kit)



Download the jdk version 19 from this link. Now after installing, we get the version 19 jdk in program files. No we need the keep the jdk path in environment variables.

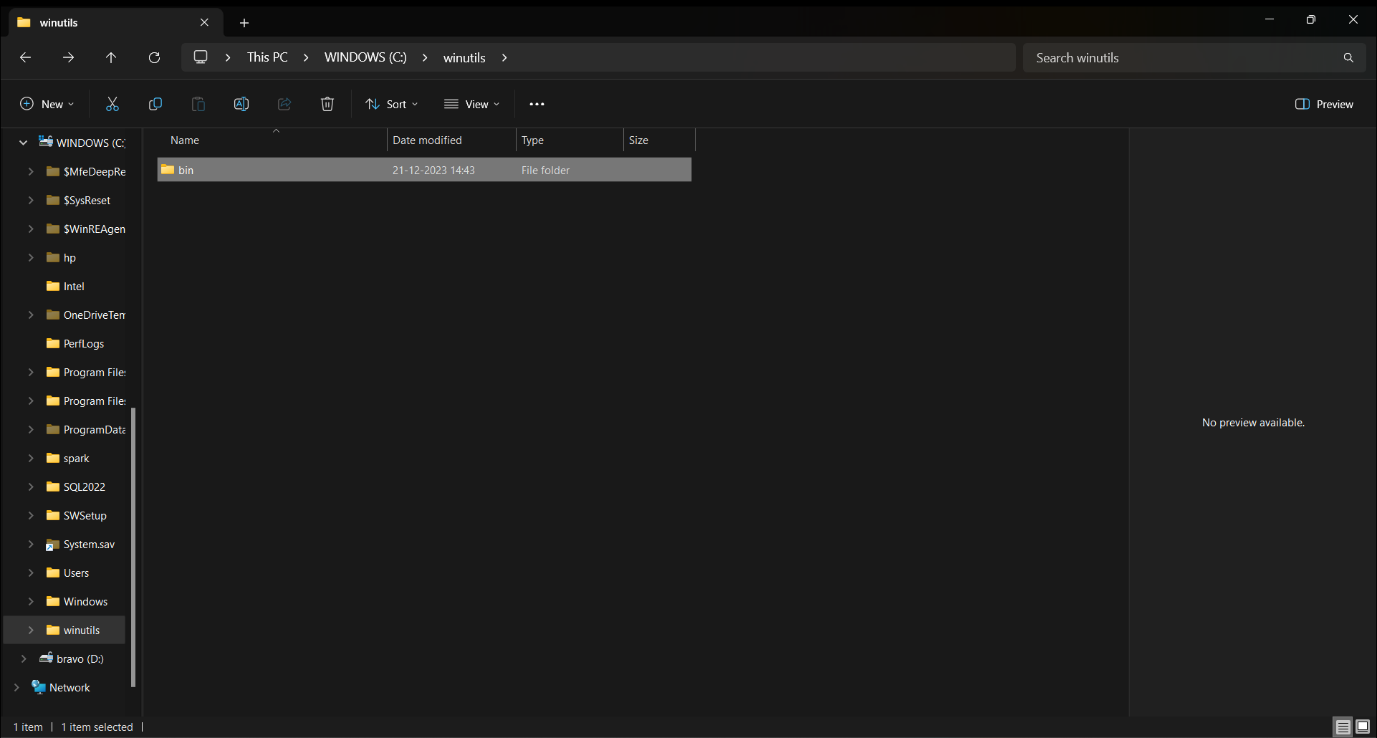


We need to give the path in the environment variables.

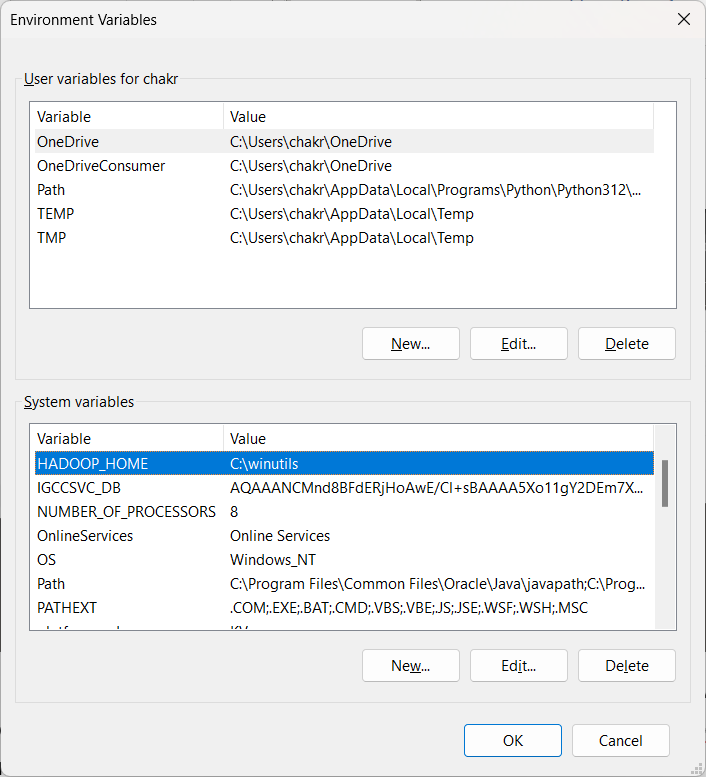
🡪winutils

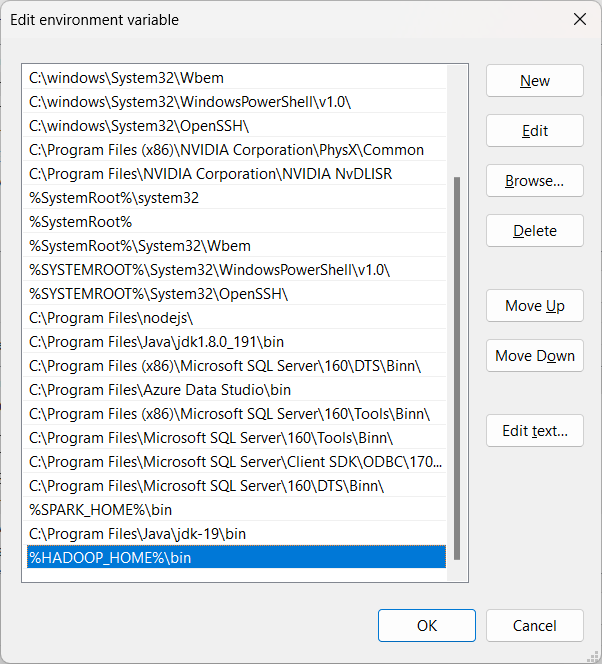
We need to download the winutils file from github link

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



After downloading the winutils file, we need to create file winutils file in C drive. And we need to create the folder named as “bin”. Now keep the downloaded file inside bin folder.





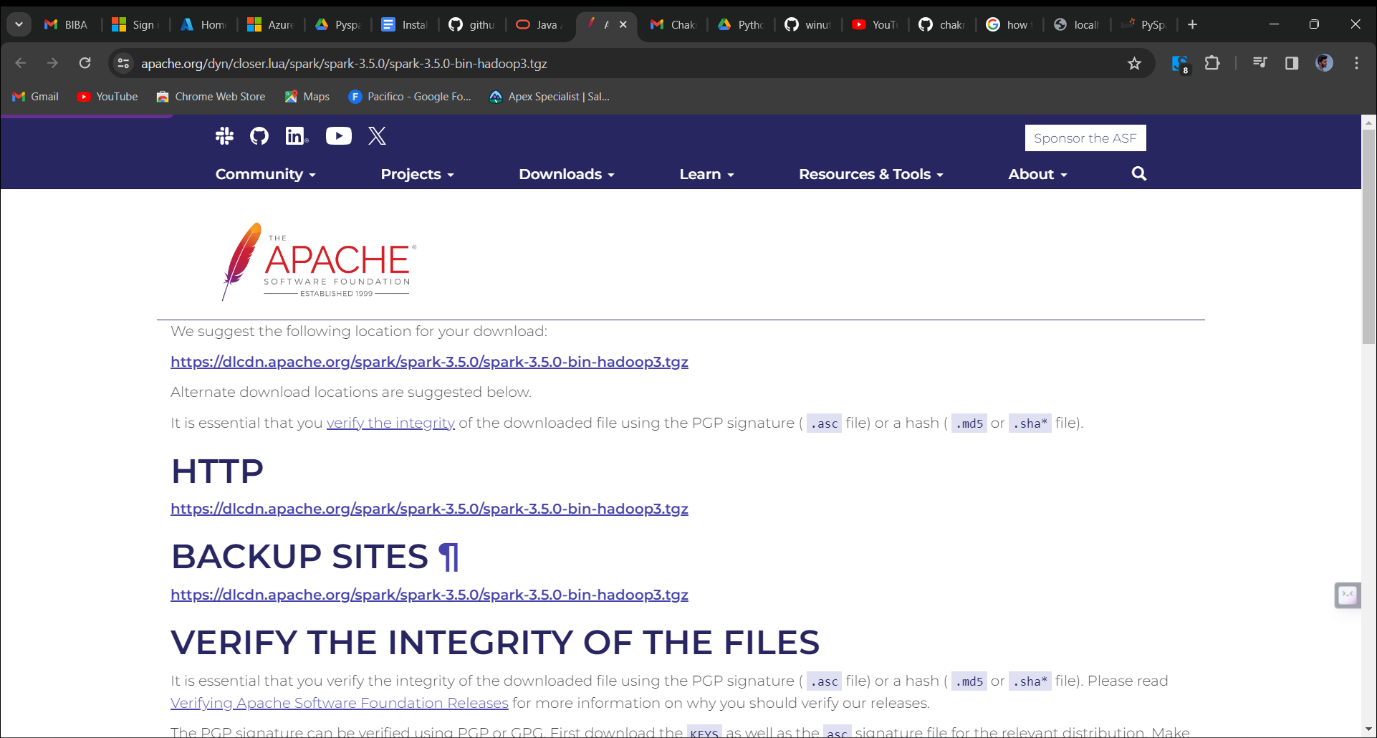
We need to give name as “HADOOP\_HOME” in environment variables and need to provide the path for winutils file.

🡪We need to install pyspark in command prompt by giving command as “pip install pyspark”

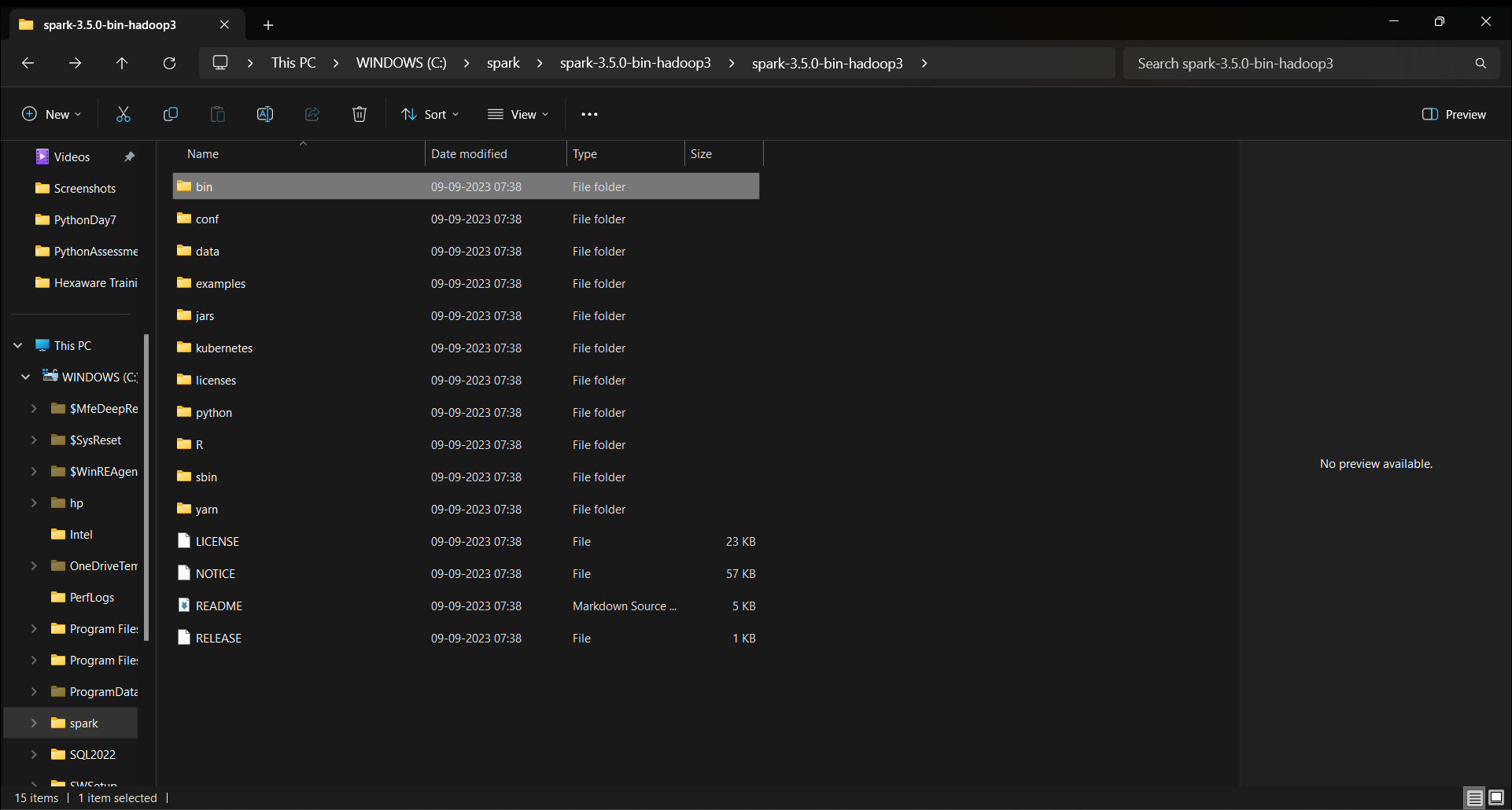
🡪Install pyspark

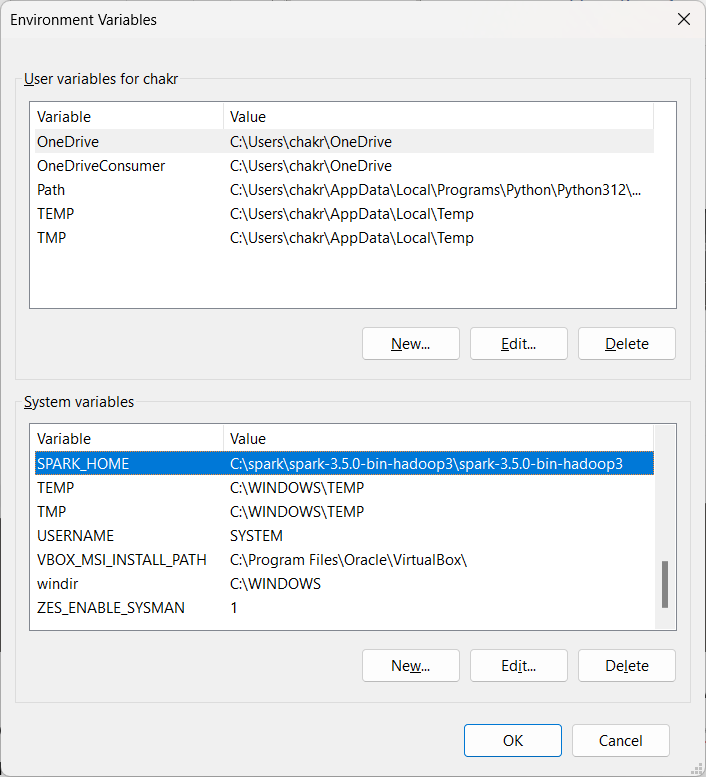
<https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz>

We need to download the spark file from the above link.



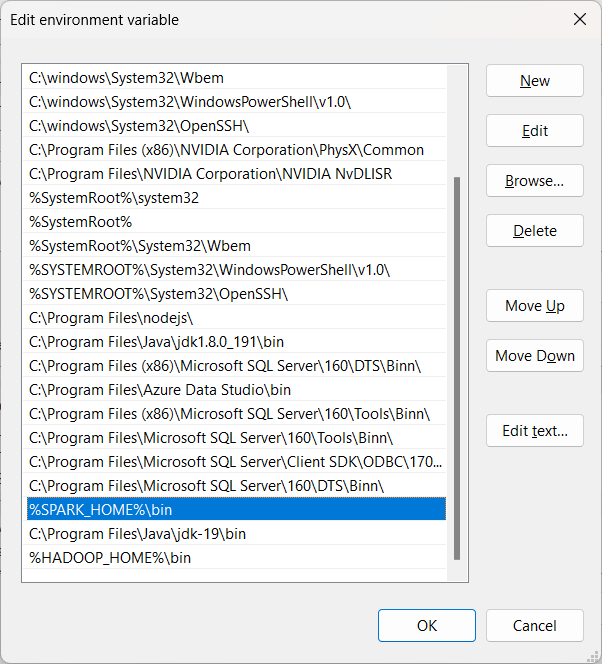
After downloading the zip file, we need to install the spark file.

We need to include the path in environment variables. 

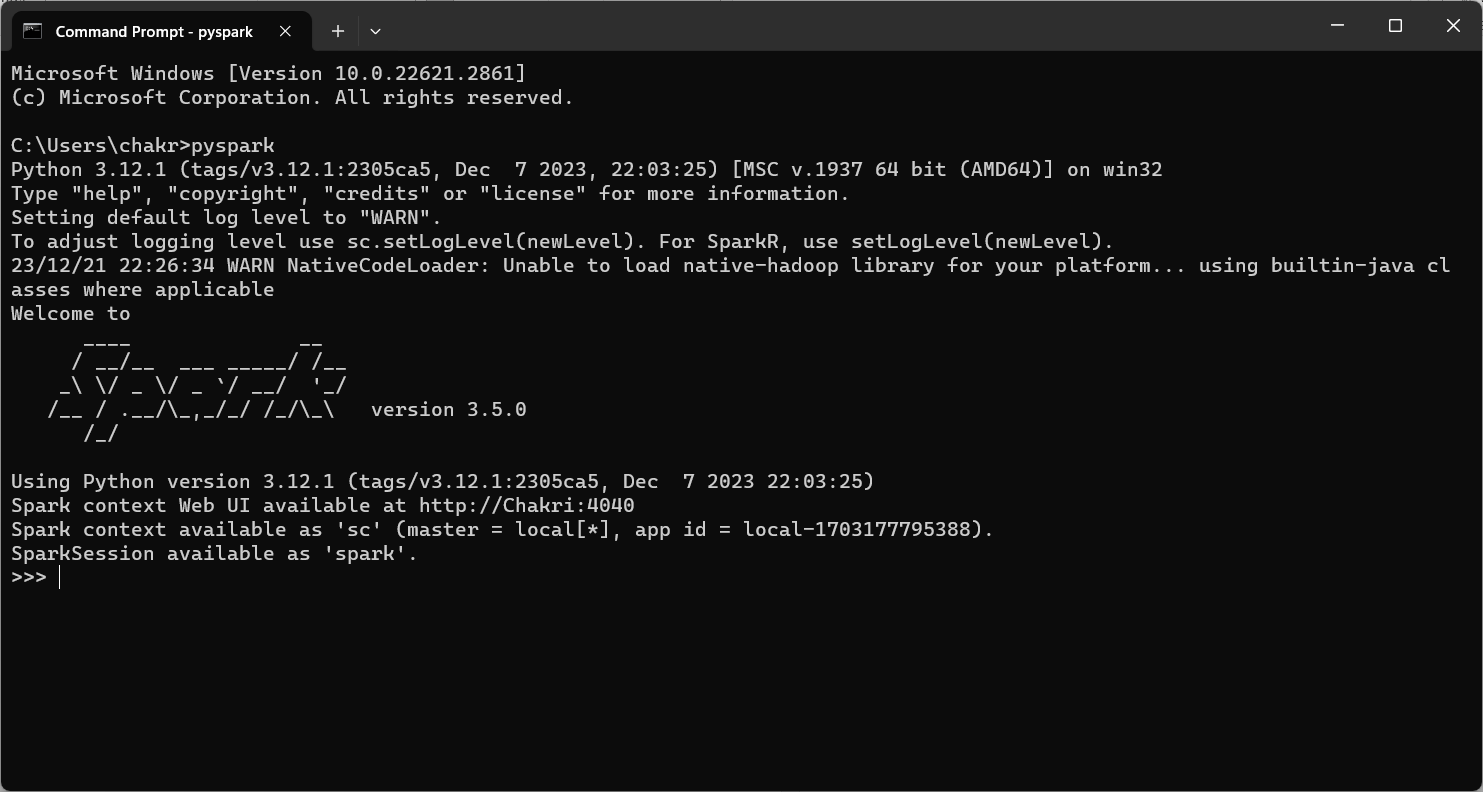


We need to name it as “SPARK\_HOME”.

In path, we need to include the path for spark file.



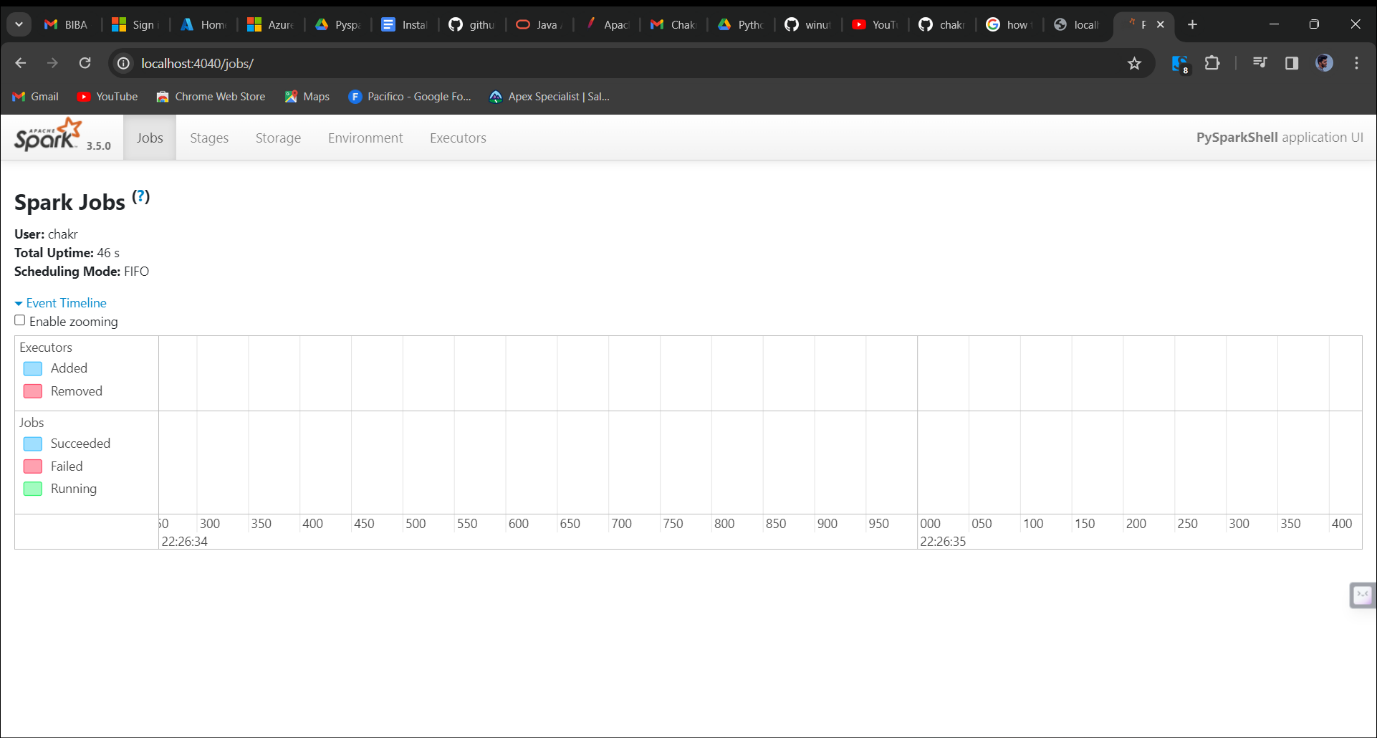
Now check in command prompt as pyspark



Now we need to click the below link

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

we need to get like the image below.



🡪Spark is designed because of to cover the loopholes of Hadoop. We cannot say that spark is an updated version of Hadoop. Spark is used to process the data very quickly.