ASSESSMENT - 16 | PySpark - 01

*“INSTALLING JDK, PYTHON, SPARK, HADOOP”*



Submitted By

VAIBHAV PATIDAR

IPS ACADEMY, INDORE (M.P.)

**Date :** 21-12-2023 | **Day 16**

**Week 3** | **Day - 04**

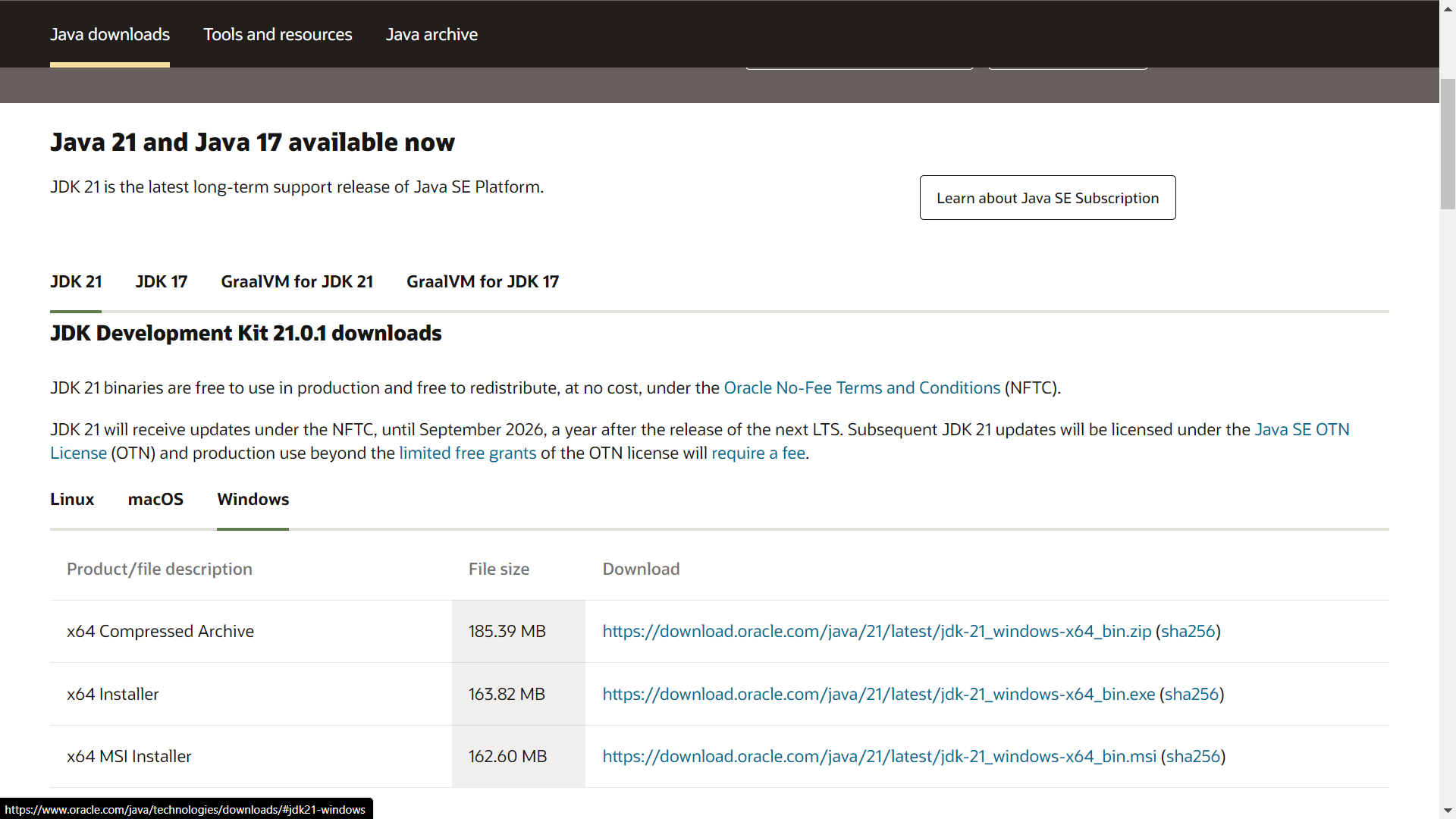
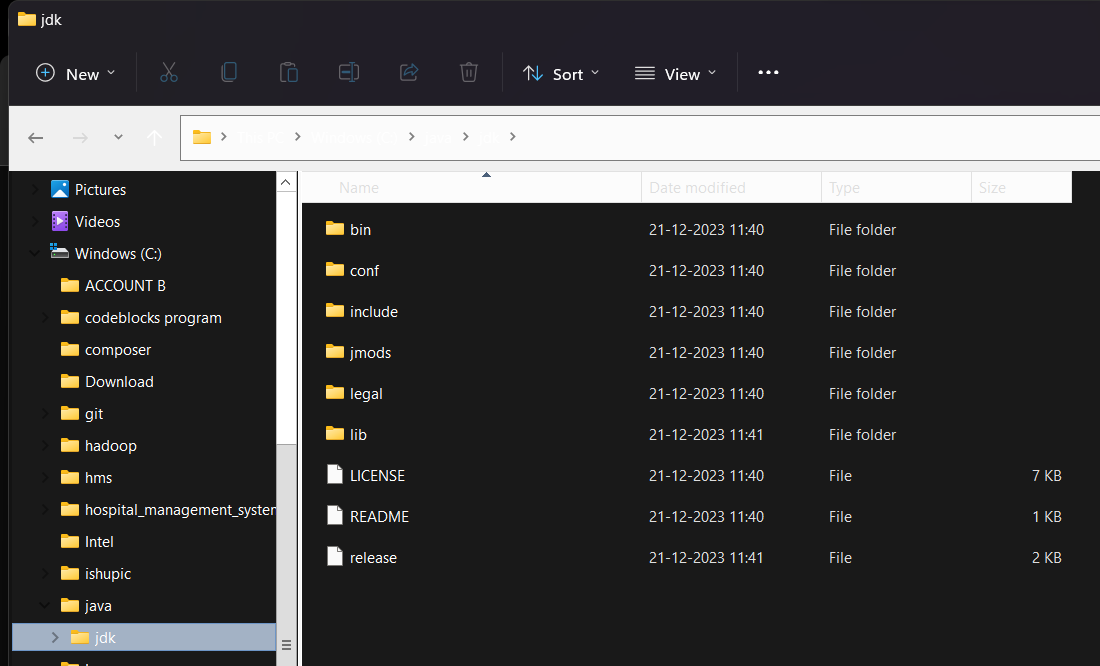
**→ Prerequisites for installing SPARK & PYSPARK :**

1. JDK with aversion above 8.
2. Latest Python version must be installed
3. Spark must be installed
4. Install Winutils.exe file
5. Edit the *“System Environment Variables” and “PATHs”.*
6. WebUI must be visible at machine localhost.

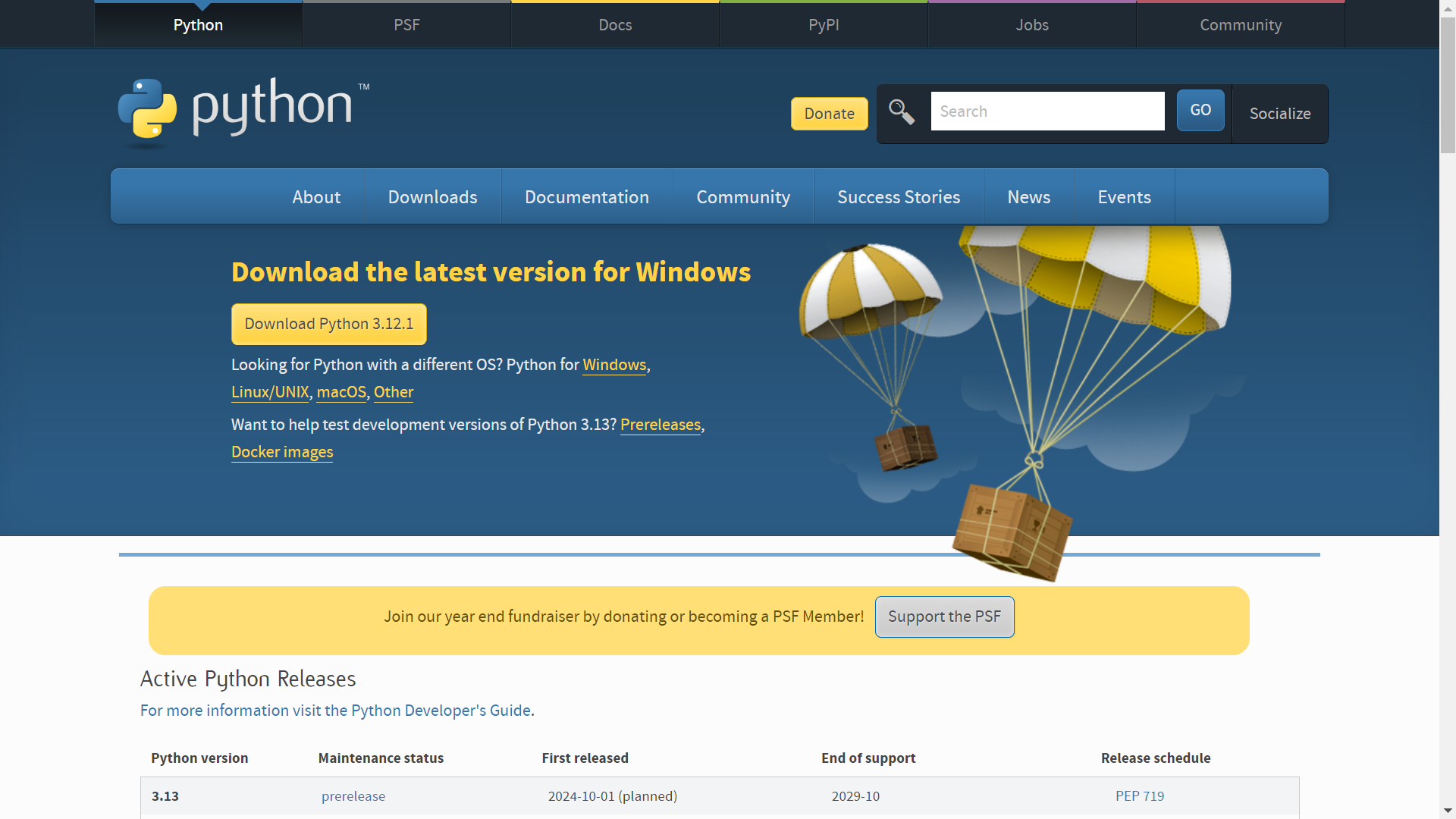
**→ WHAT IS JDK?**

* JDK (Java Development Kit) is essential for Java development.
* Includes Java compiler (javac) and Java Virtual Machine (JVM).
* Provides runtime libraries, APIs, and development tools.
* Enables Java application creation and execution.

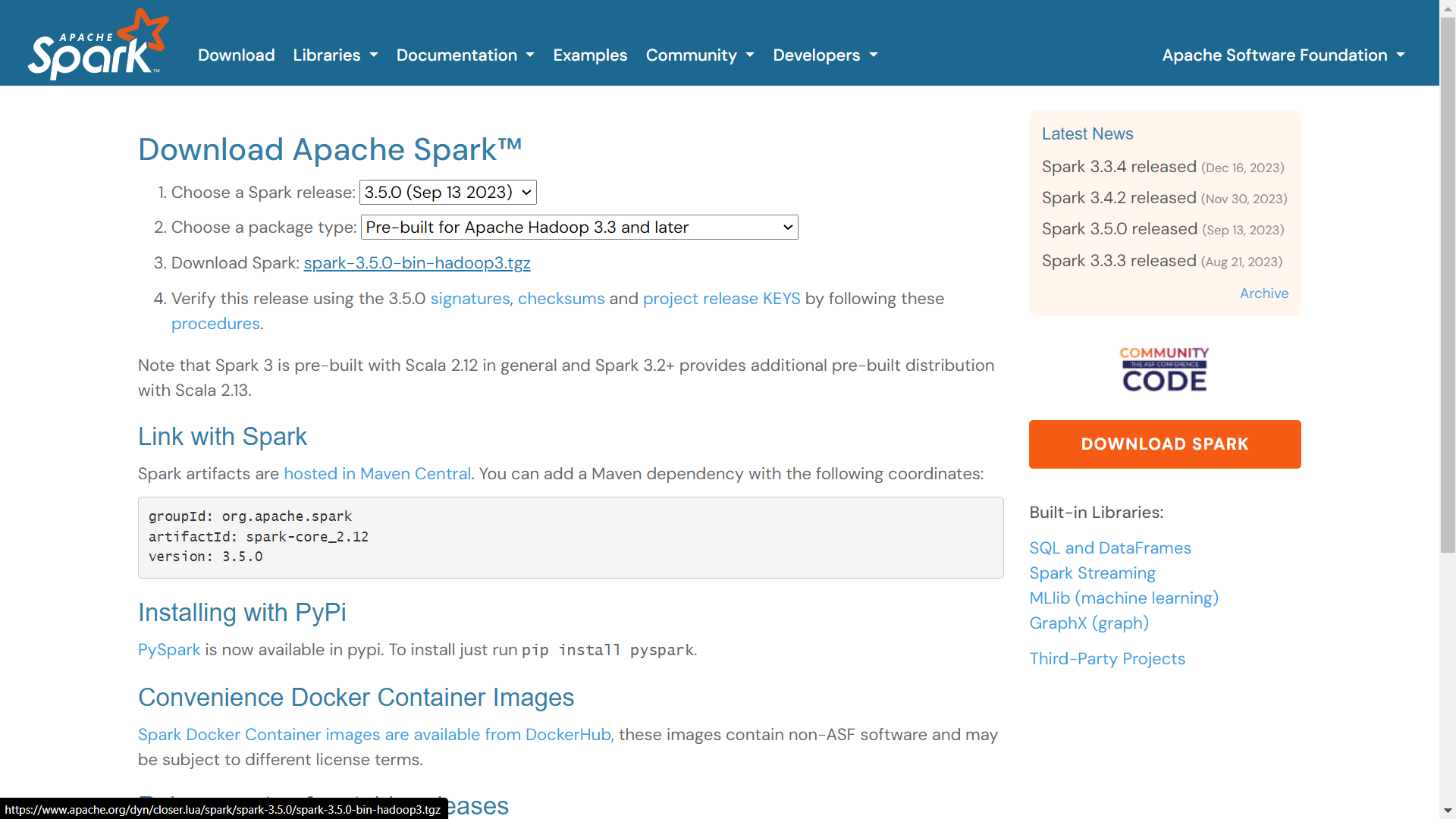
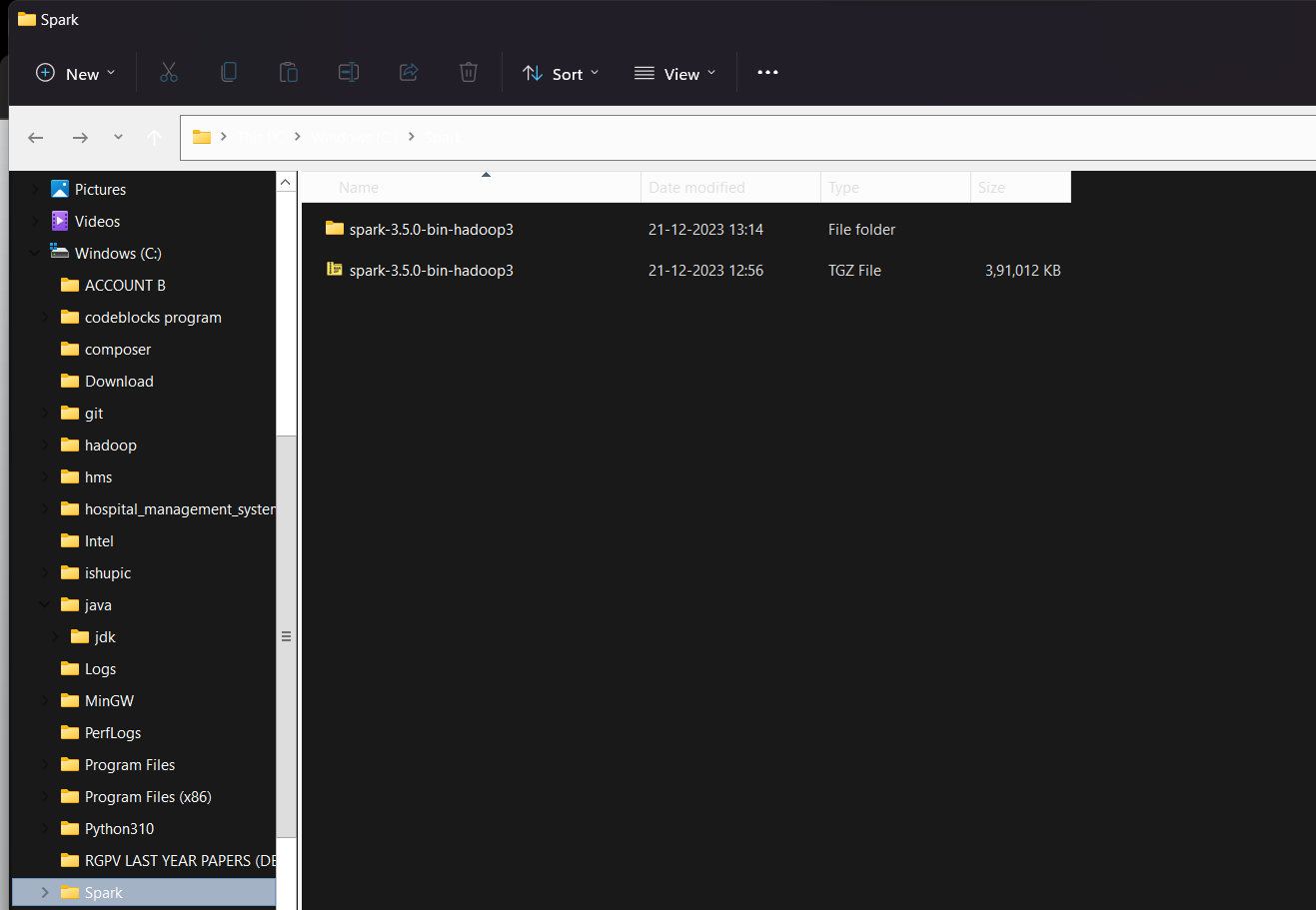
**→ Steps to install JDK :**

1. Go to google → search : Download JDK → Windows –< Download.
2. Or can visit to → <https://www.oracle.com/java/technologies/downloads/#jdk21-windows> → download .exe file
3. 
4. Then install JDK after downloading, keep in my mind while installing:
5. Create a folder named **“JAVA”** in WINDOWS(C) and a folder named **“JDK”** inside JAVA folder and install into it.
6. 

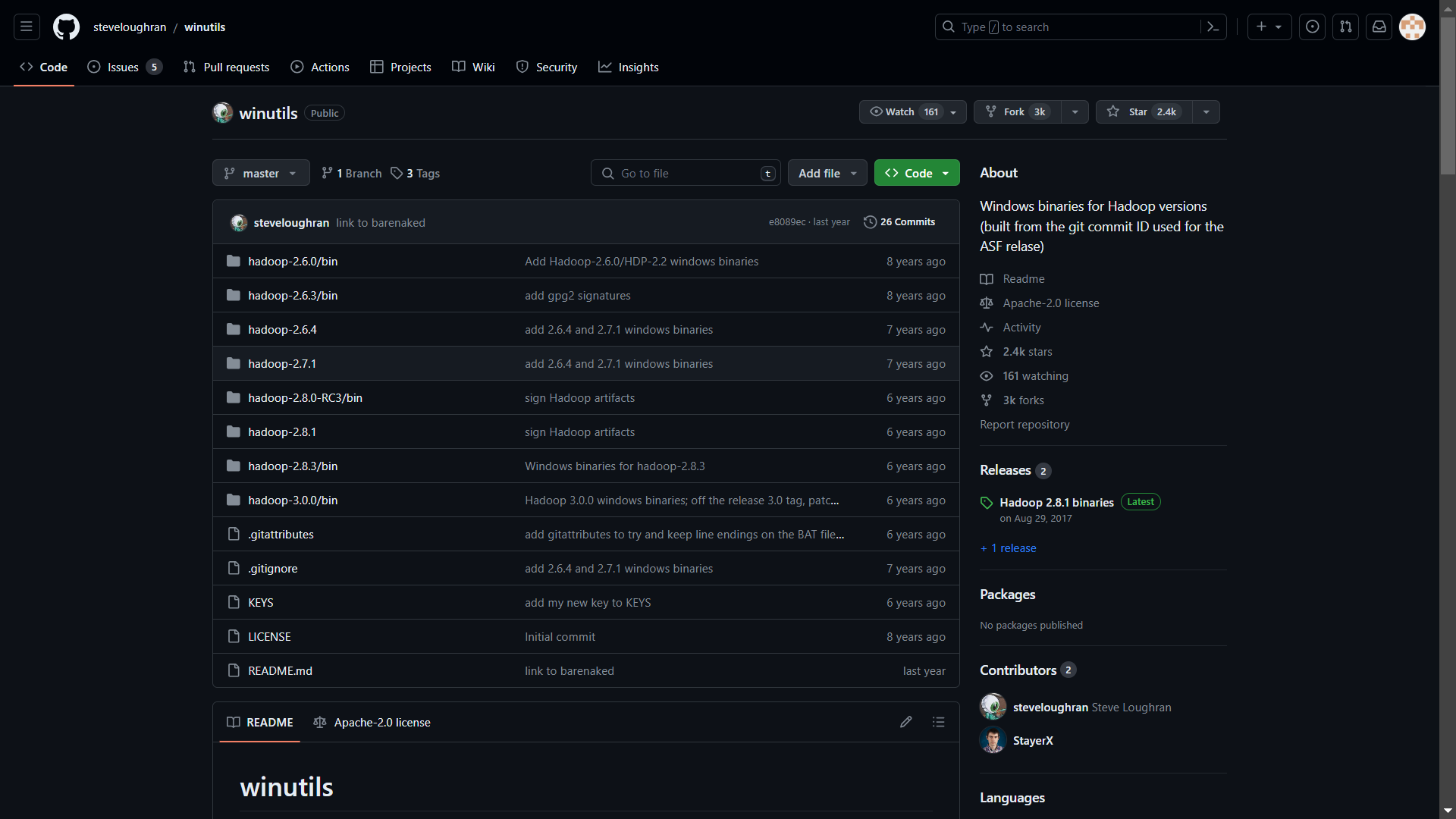
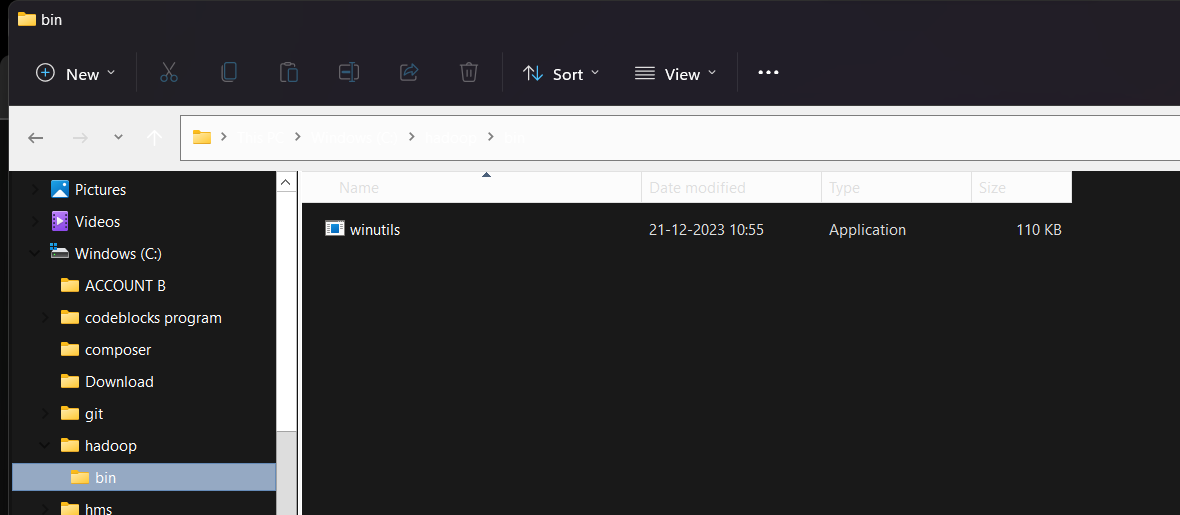
**→ Steps to install Python :**

1. Go to google search → Python Download or visit <https://www.python.org/downloads/>
2. Then download the latest version and install it.
3. 

**→ Steps to download Spark :**

1. Go o google and search → Download Spark or visit : <https://spark.apache.org/downloads.html>
2. THne download the latest version ex : 3.5.0
3. 
4. But, create a folder named **“SPARK”** in windows(c) and save the tar .tgz file there.
5. Now, click on the ***spark-3.5.0-bin-hadoop3.tgz*** file and extract it, until you get the .tar file in the spar folder.
6. 

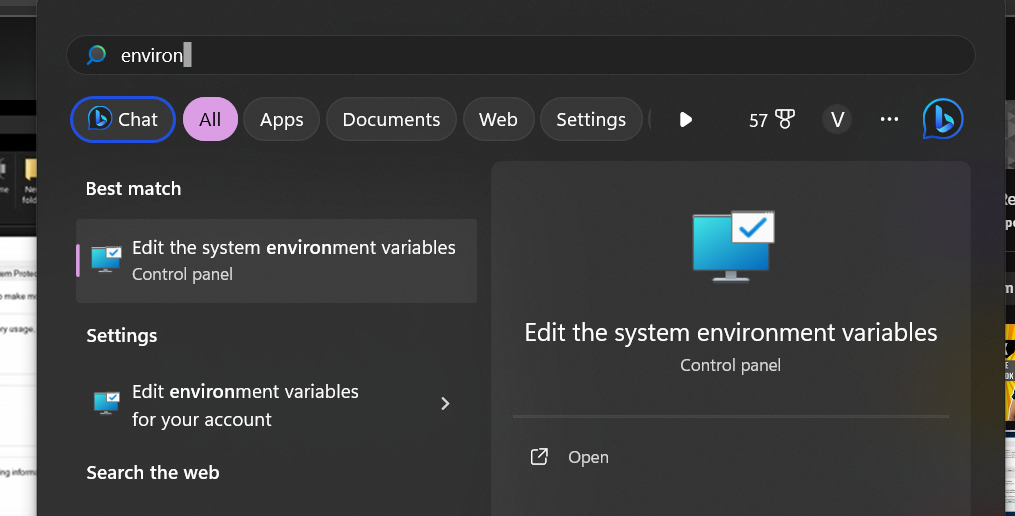
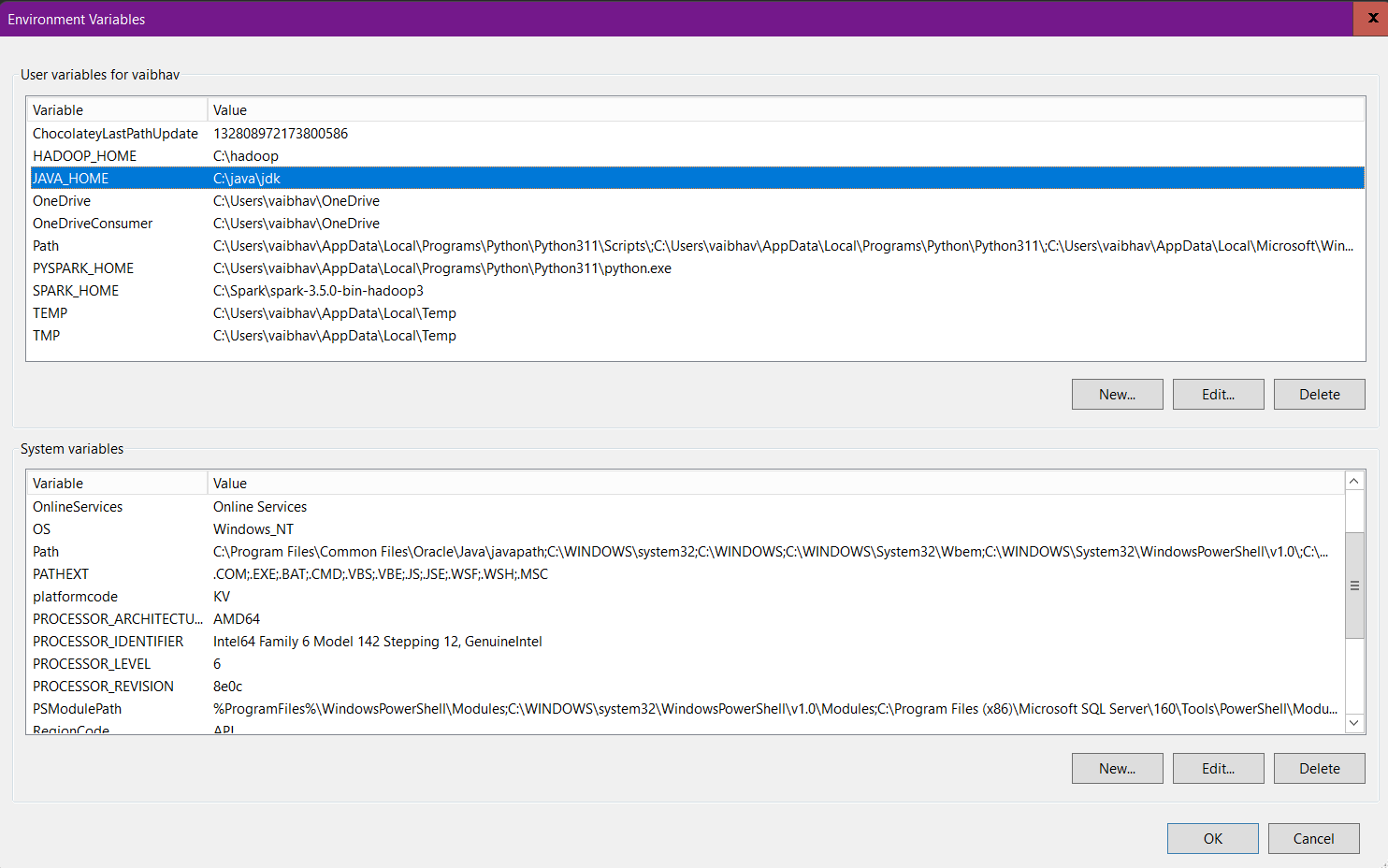
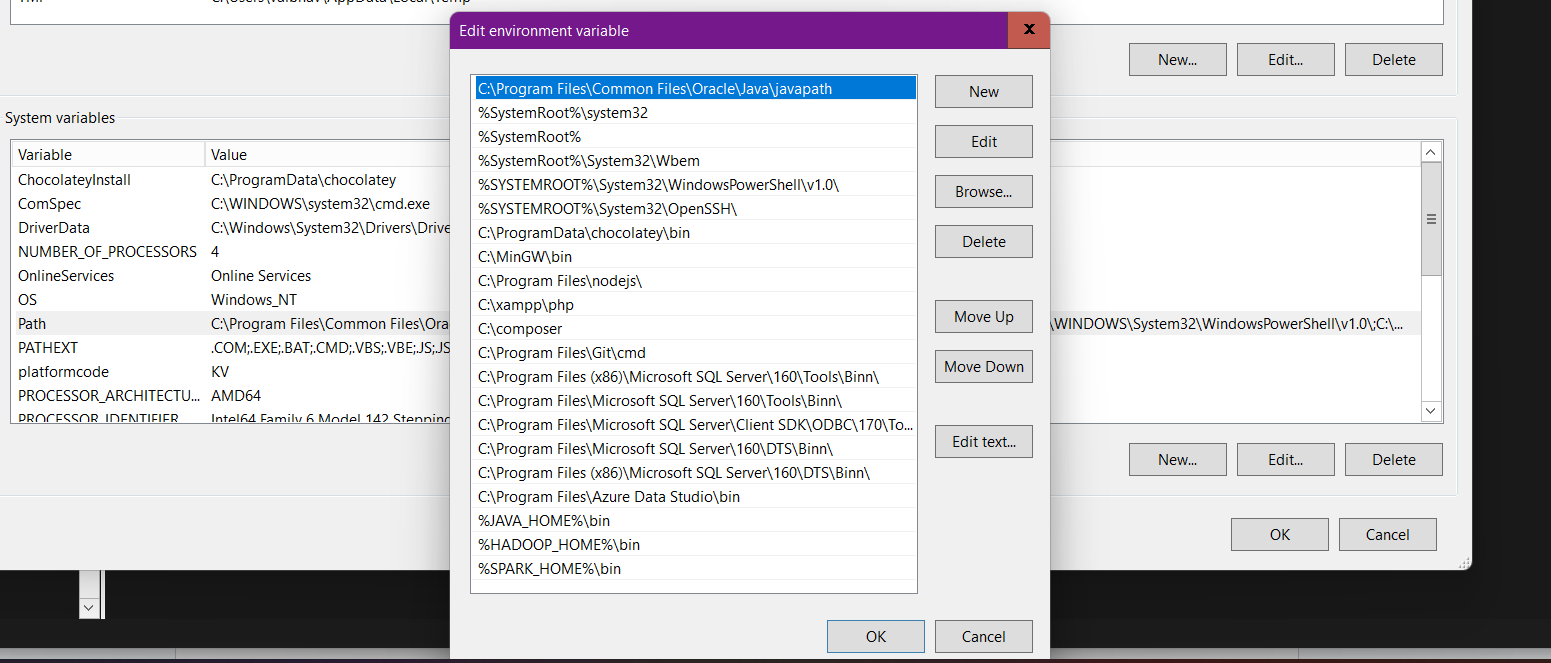
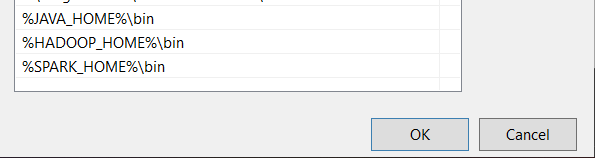
**→ Steps to install the Winutils.exe file :**

1. Click on the link : <https://github.com/steveloughran/winutils>
2. 
3. Click on Hadoop 2.7.1 → click bin → download Winutill.exe
4. After downloading, create a folder in Windows(c) named **“HADOOP”** and save the file here.
5. 

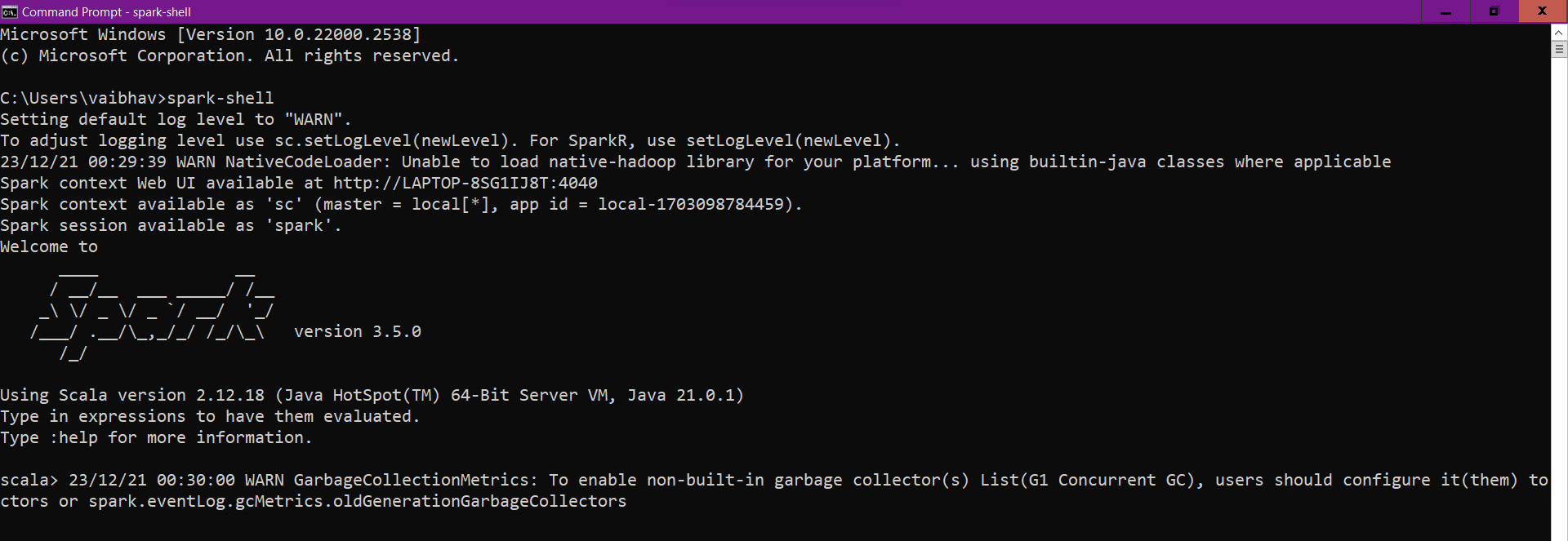
**→ Check Java and Python Installation :**

1. Open command prompt.
2. Type Java –Version and click enter
3. Type Python –version and click enter.

**→ Setup the Environment Variables and paths :**

1. Open the environment variable on your computer.
2. 
3. Then set the HOME for JAVA, SPARK, PYSPARK, & HADOOP as shown in image :
4. 
5. Then next step is to set the path for all 4 also. For this in “SYSTEM VARIABLE” section click on **PATH** and click **EDIT : **
6. Then click on **NEW** and add following three paths : 
7. Then click on **“OK” and “OK”**

**→ Check the installation of SPARK and PYSPARK :**

1. Open Command Prompt → type ***SPARK-SHELL*** and click enter.
2. If you see below image, ALL SET!!!
3. 
4. Now open the WEBUI on your system: <http://localhost:4040/jobs/>
5. 