MANDAPALLI HEMA

ASSESSMENT-16

**Pyspark = Python + Apache Spark**

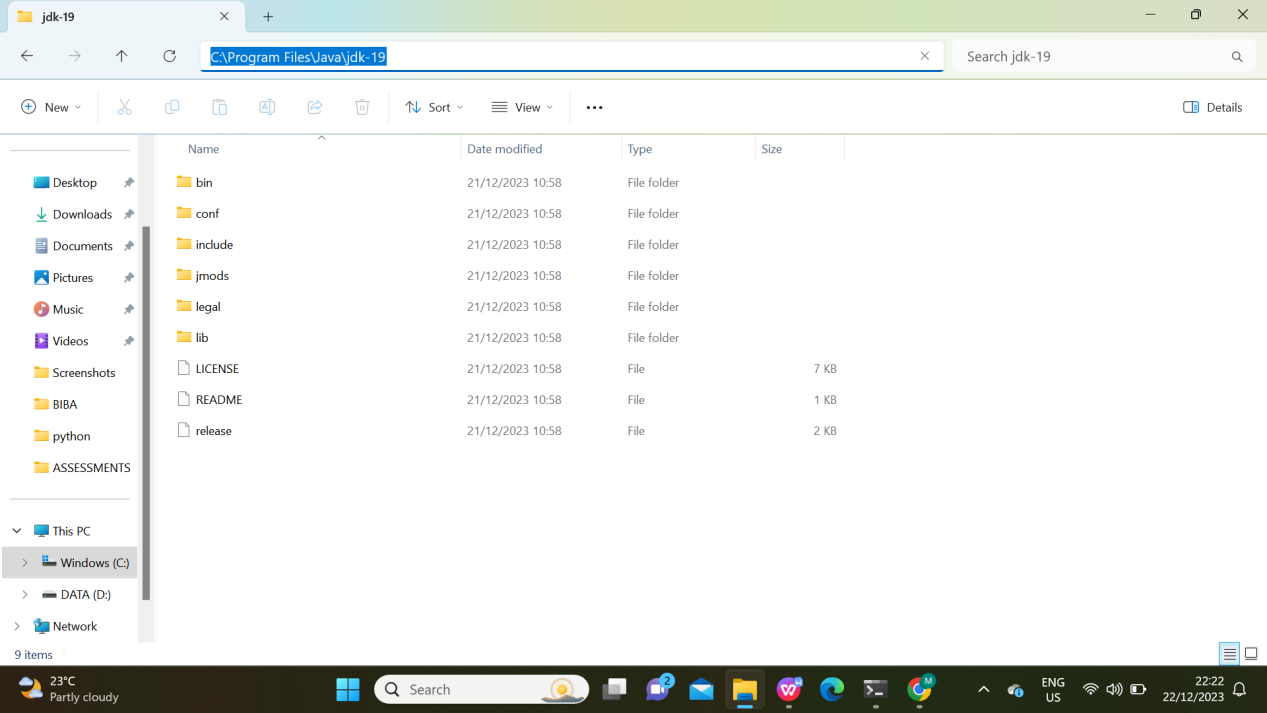
Apache Spark is a new and open-source framework used in the big data industry for real-time processing and batch processing. It supports different languages, like Python, Scala, Java, and R

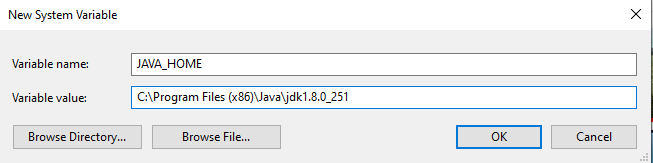
**Java installation**

1. Go to **Download Java JDK**.  
   Visit Oracle's website for the download of the Java Development Kit(JDK).
2. Open the installer file, and the download.

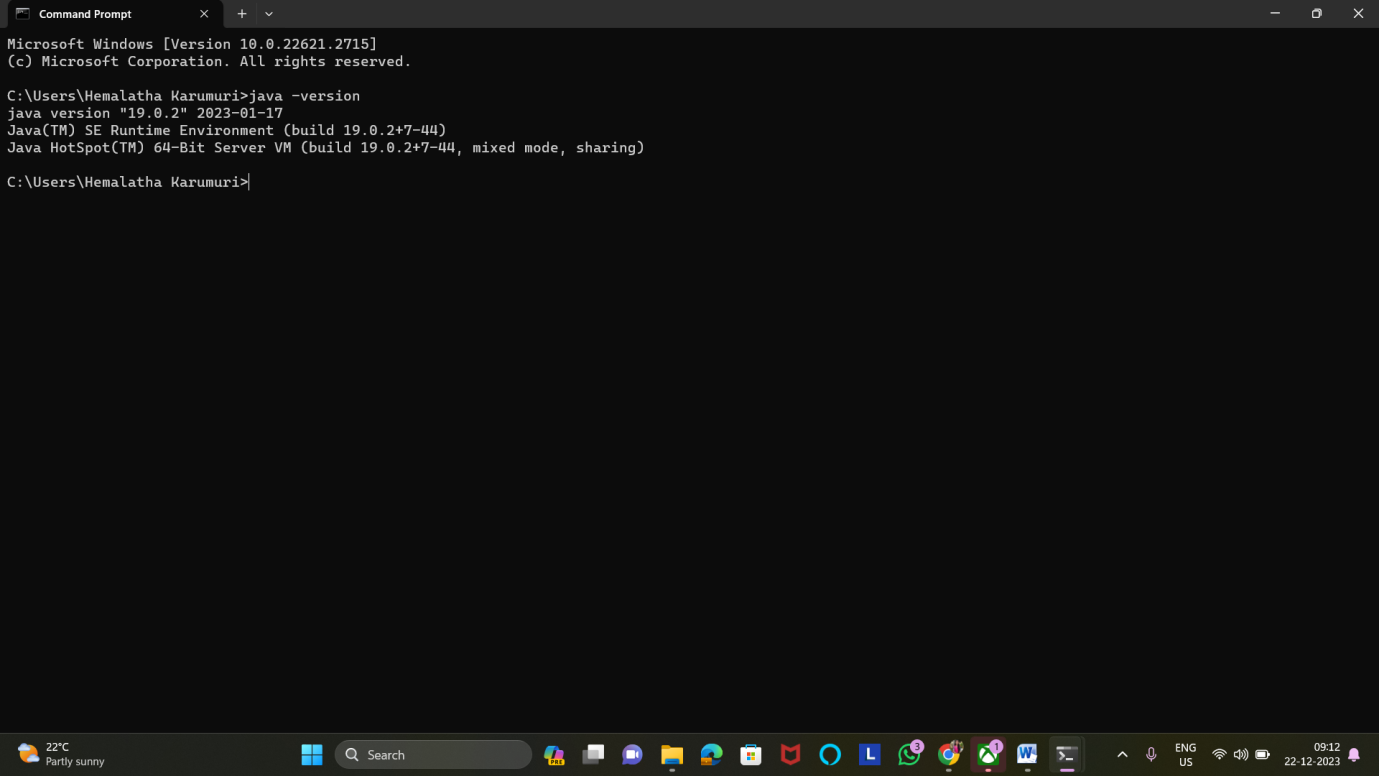


1. After installation we need to set up the jdk in to the environment variable.

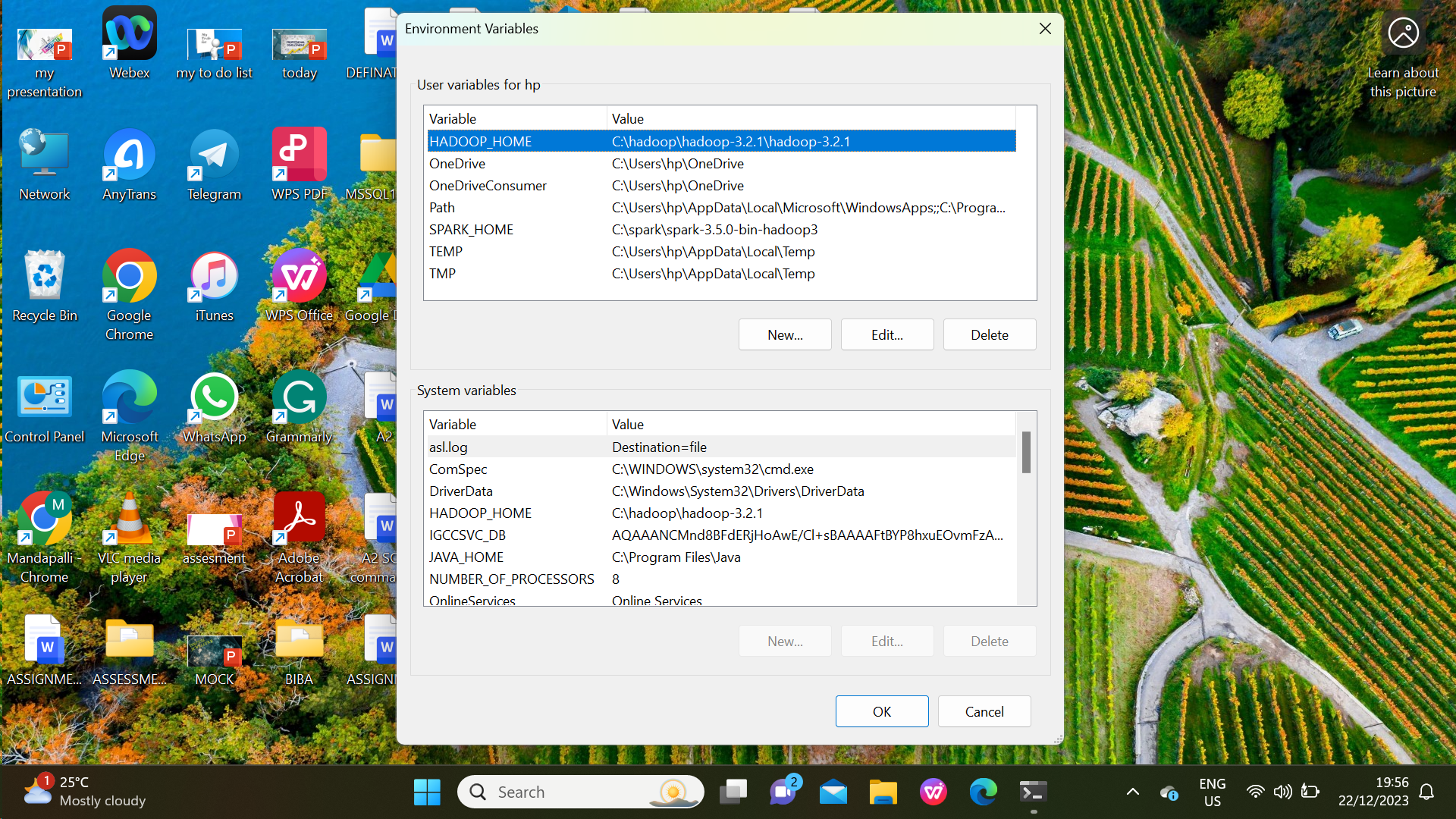




1. Go to "Command Prompt" and type "java -version" to know the version.

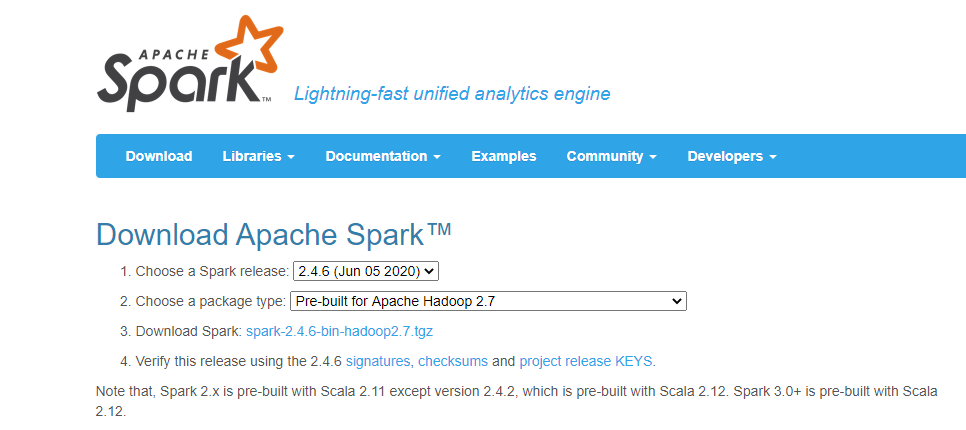


1. We can run java application as the set up was completed.



**Installing Pyspark**

1. Go to the **Spark homepage**.
2. Select the Spark release and package type as following and download the .tgz file.

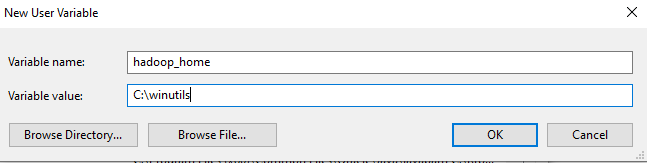


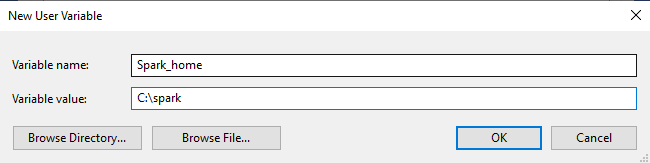
**Download and setup winutils.exe**

Go to **Winutils** choose your previously downloaded Hadoop version, then download the winutils.exe file by going inside 'bin'. The link to my Hadoop version is: **https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe**

Make a new folder called 'winutils' and inside of it create again a new folder called 'bin'.Then put the file recently download 'winutils' inside it.

Let's create a new environment where variable name as "hadoop\_home" and variable value to be the location of winutils, which is "C:\winutils" and click "OK".



1. For spark, also let's create a new environment where the variable name is "Spark\_home" and the variable value to be the location of spark, which is "C:\spark" and click "OK".  
   

Finally, double click the 'path' and change the following as done below where a new path is created "%Spark\_Home%\bin' is added

3.Check for pyspark in the command prompt it will displays like this