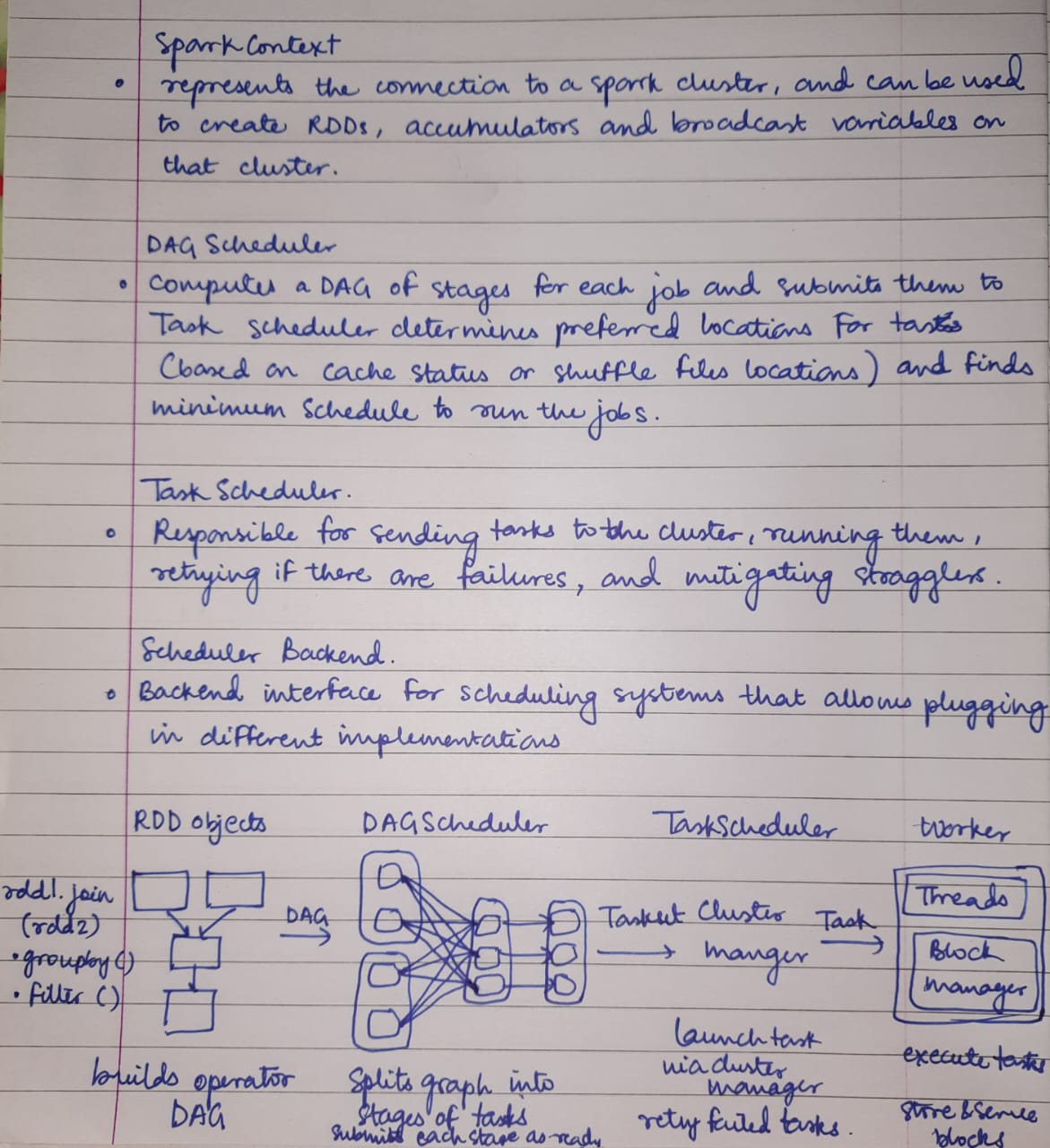
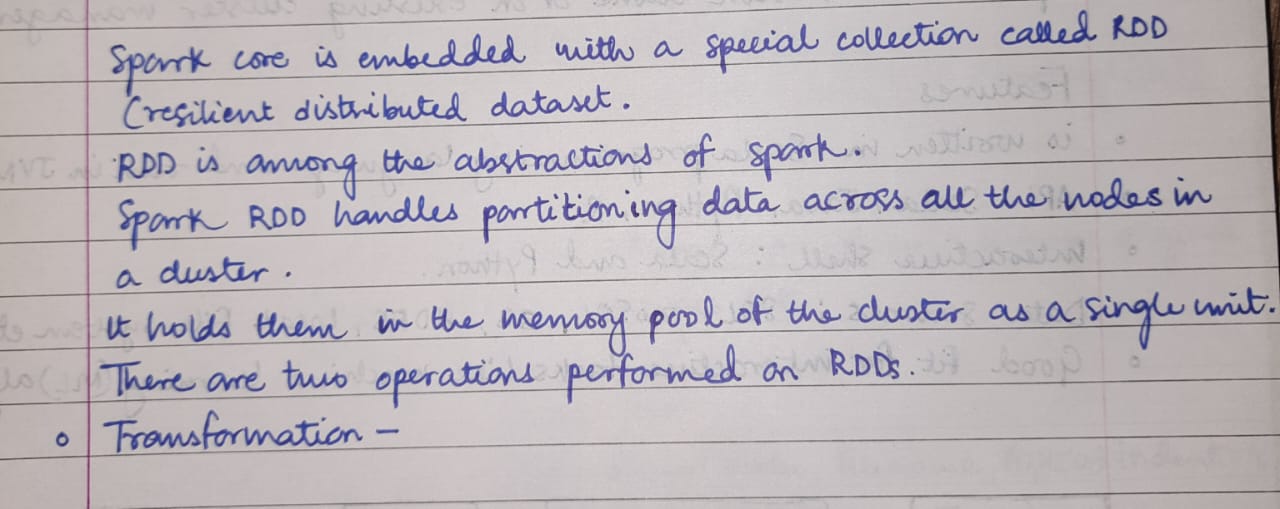
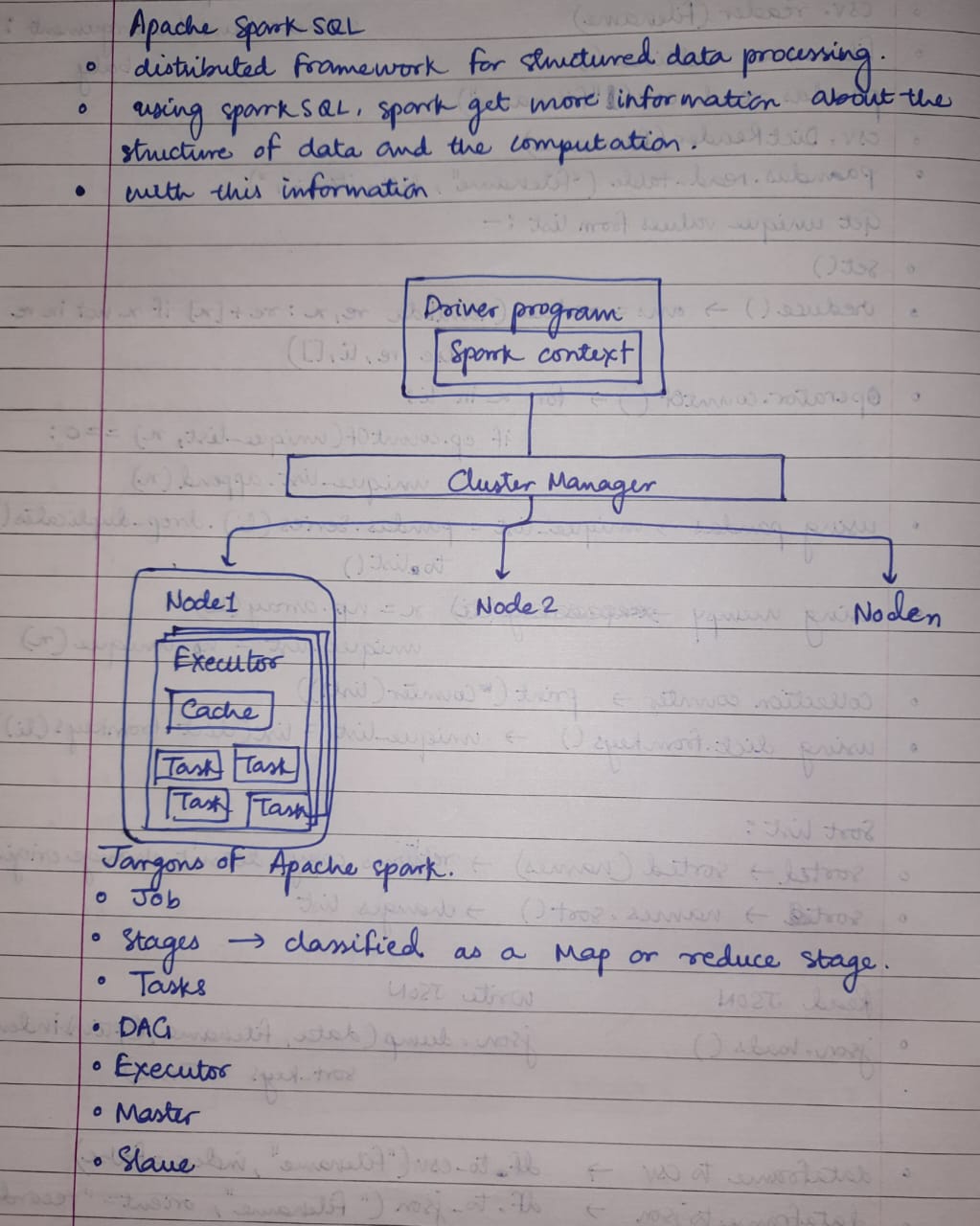
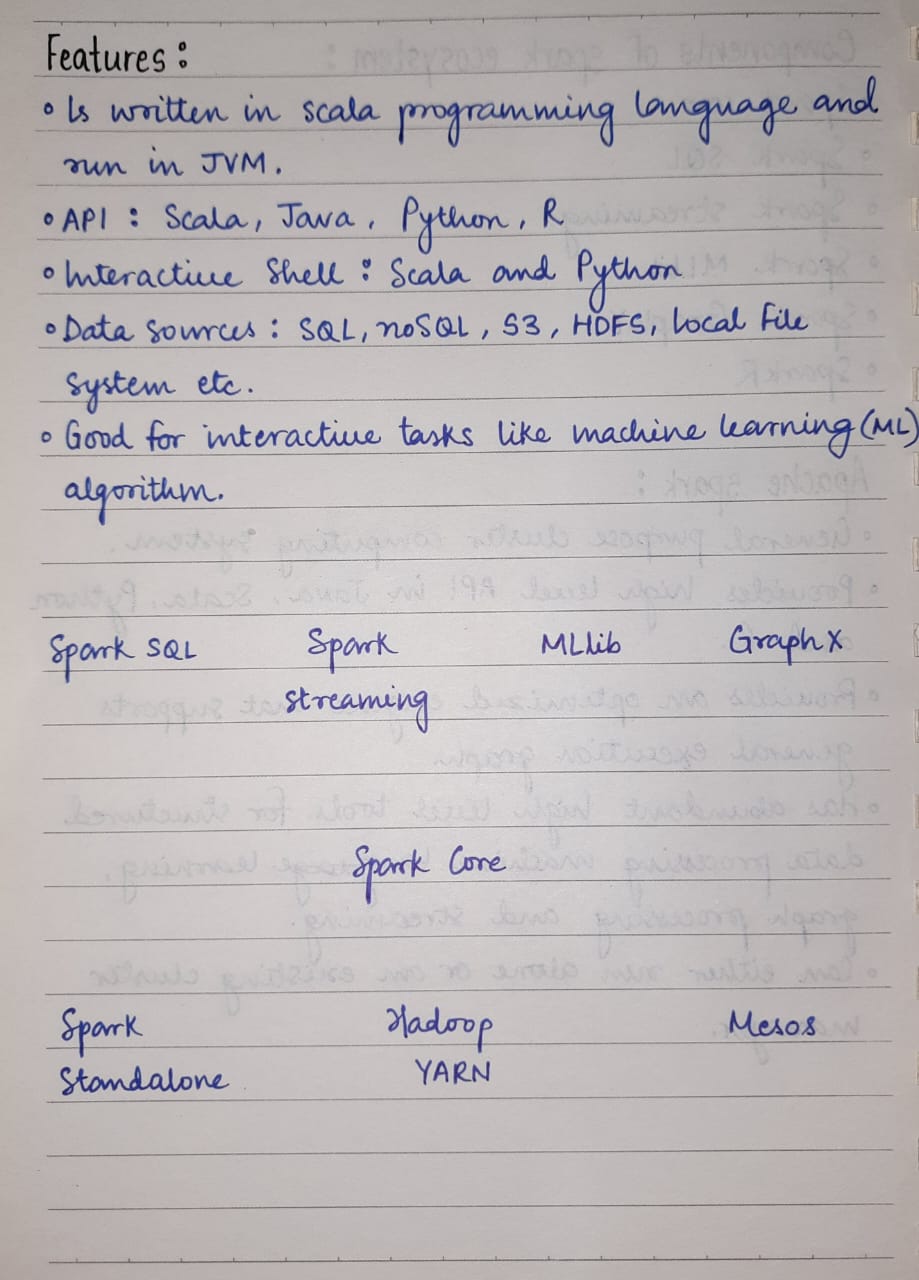
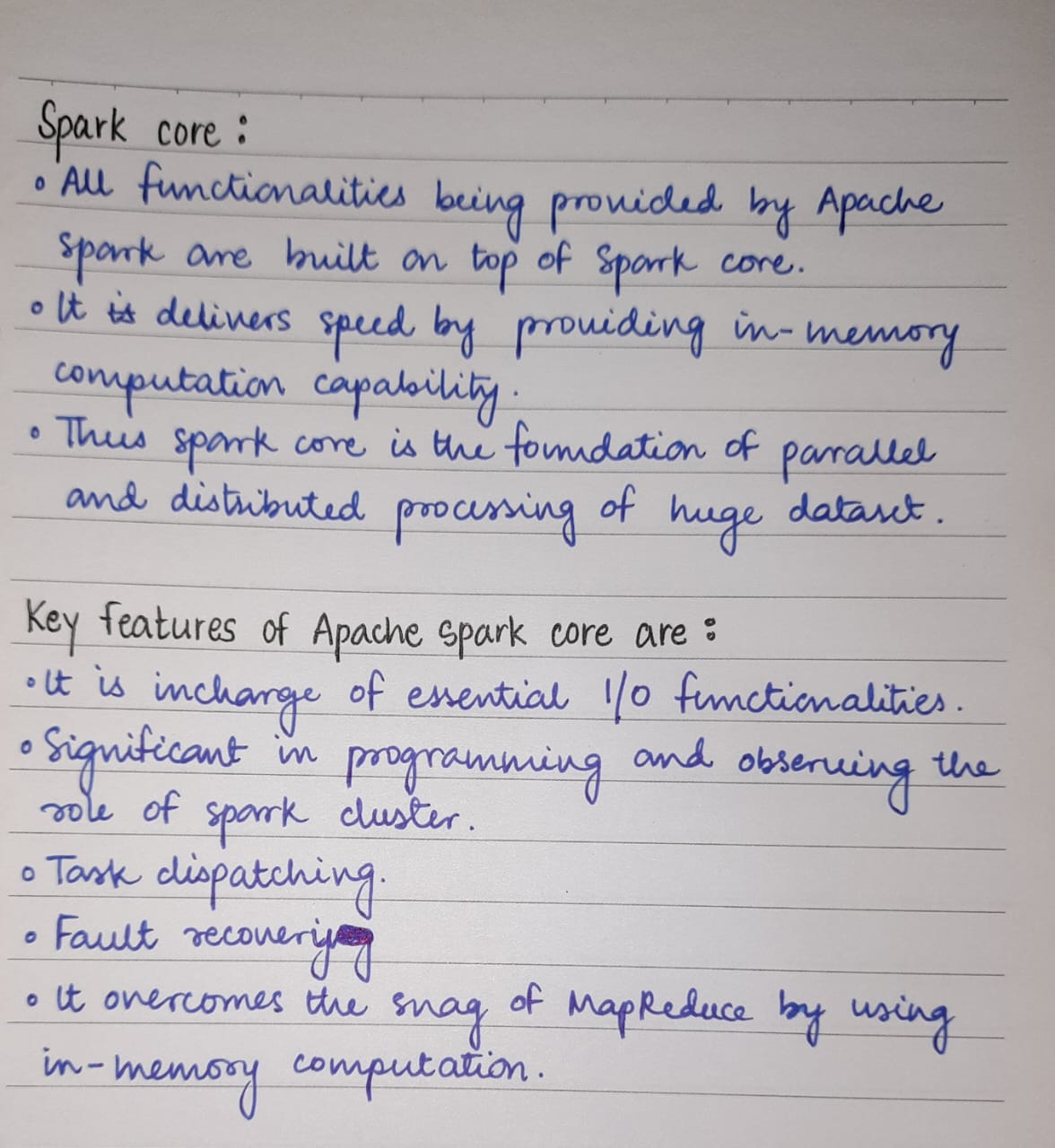
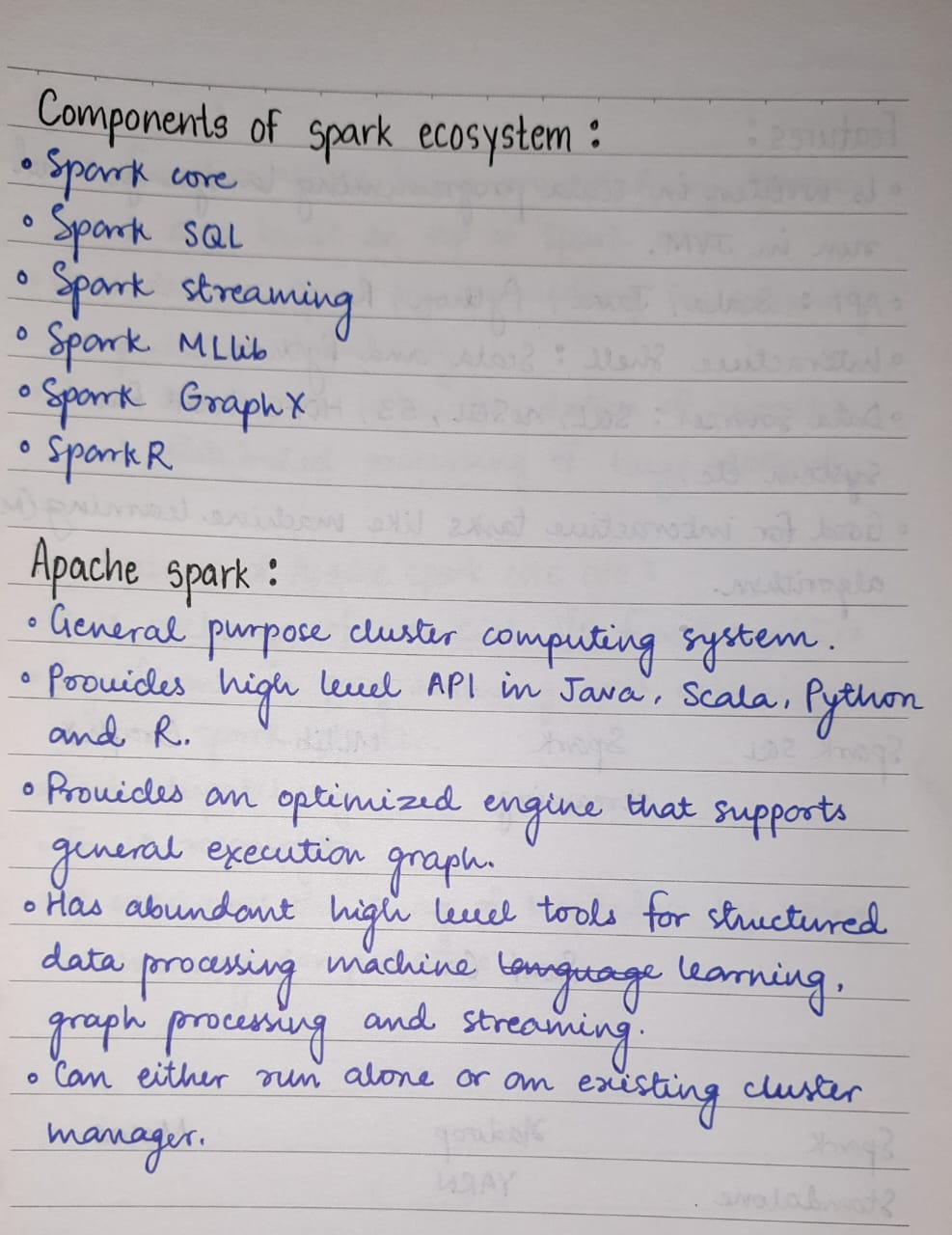
**VAISHALI BOKADIYA**

**PYSPARK DAY 1 ASSESSMENT**



**Pyspark installation:**

**Step1:** Go to <https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz> and download spark.

**Step2:** Go to <https://github.com/steveloughran/winutils/blob/master/hadoop-3.0.0/bin/winutils.exe> and download winutils.exe

**Step3:** Go to <https://www.oracle.com/java/technologies/downloads/> and download and install java development kit compatible for your pc.

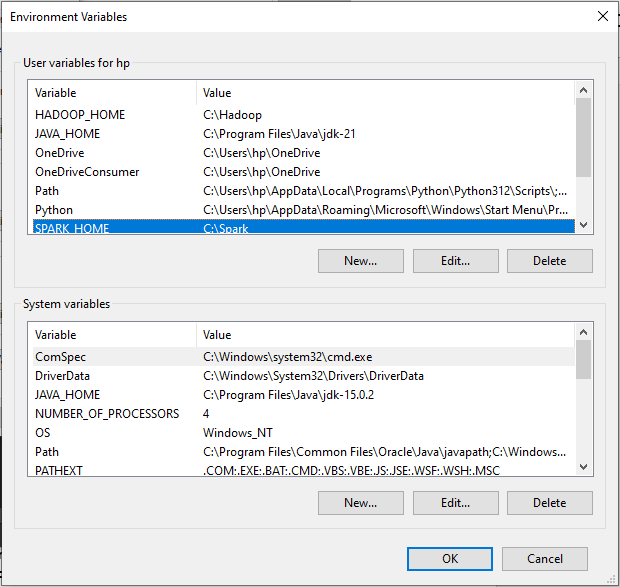
**Step4:** In C drive create a folder named “Spark” and extract all the files of spark-3.5.0-bin-hadoop3.tgz in this folder.

**Step4:** In C drive create a folder named “Hadoop” and then in this folder create another folder named bin. Inside bin copy the winutils.exe file.

**Step5:** Open system properties, go to environment variables and then add

The following variables:

|  |  |
| --- | --- |
| Variables | Values |
| HADOOP\_HOME | C:\Hadoop |
| SPARK\_HOME | C:\Spark |
| JAVA\_HOME | C:\Program Files\Java\jdk-21 |



**Step6:** In the user variables, go to the path variable, click on edit, and add the following new paths:

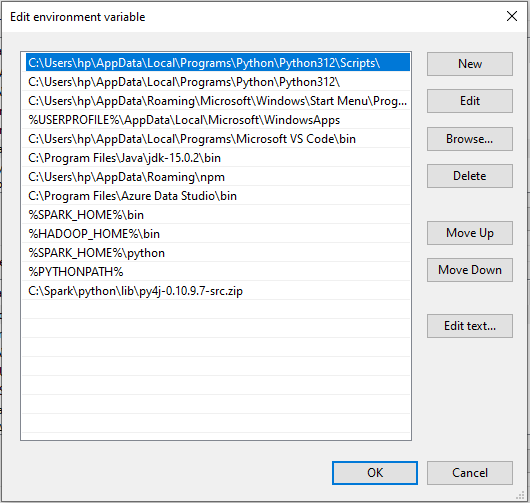
%SPARK\_HOME%\bin

%HADOOP\_HOME%\bin

%SPARK\_HOME%\python

%PYTHONPATH%

C:\Spark\python\lib\py4j-0.10.9.7-src.zip



**Step7:** Open Command Prompt and type “spark-submit –version”.

On successful installation the following window will appear.

