BUILDING AN APACHE SPARK LOCAL CLUSTER

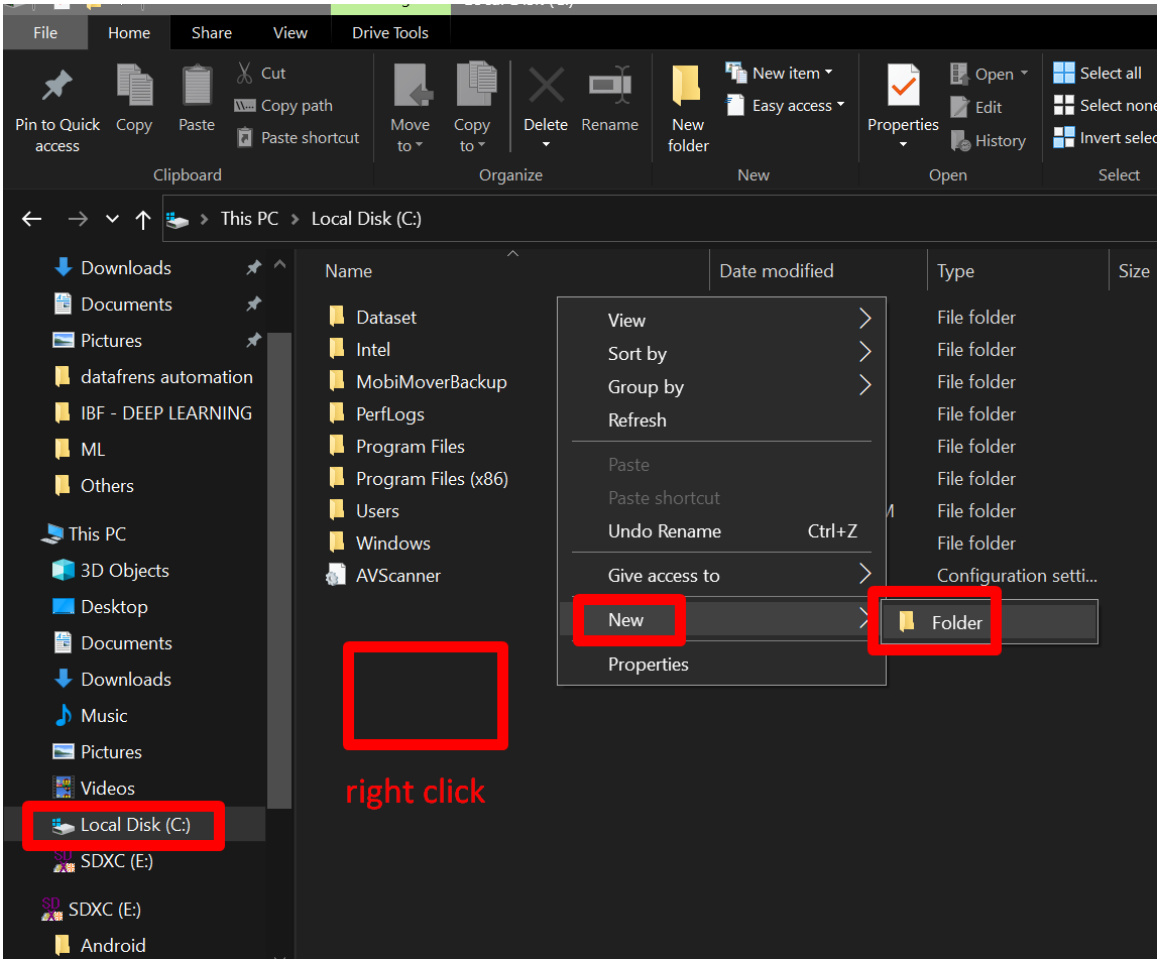
ON WINDOWS

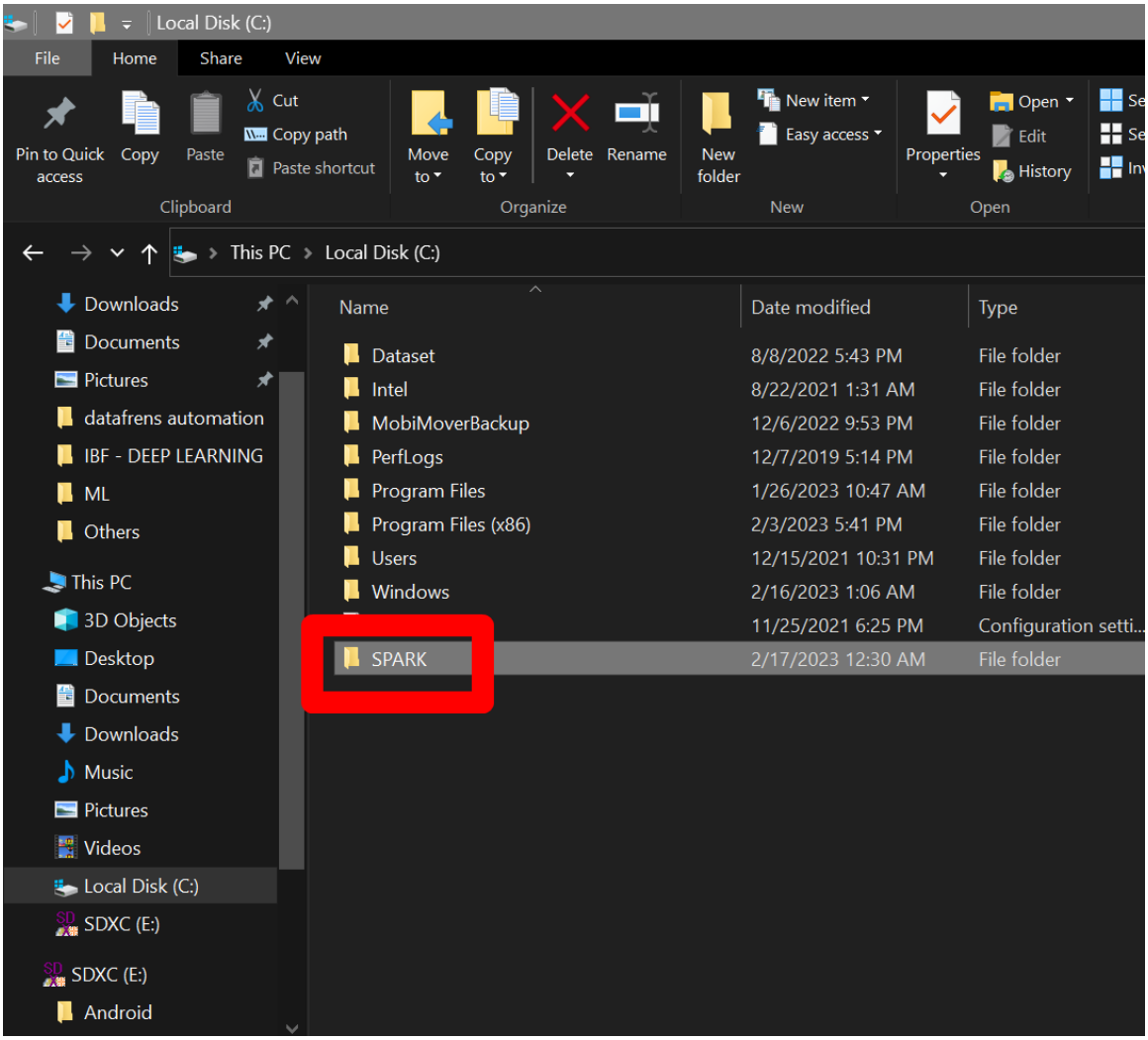
DR. ALVIN ANG

CONTENTS

I. INSTALLATION STEPS

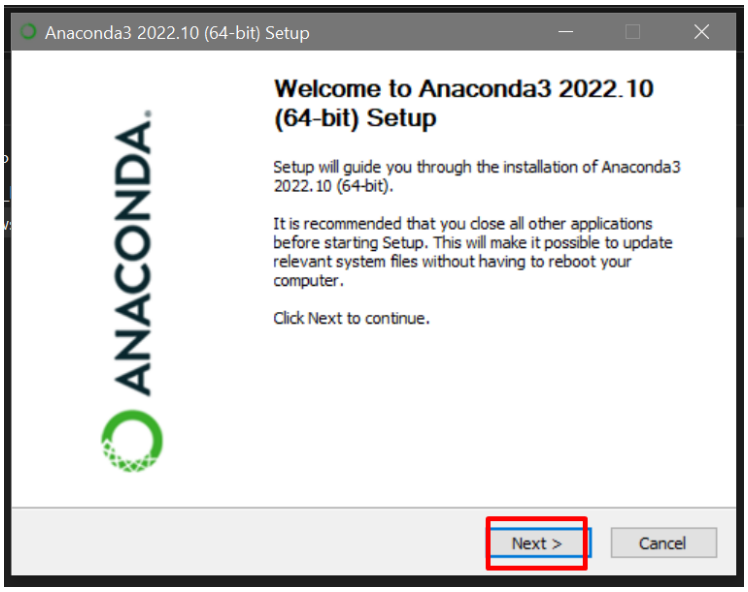
A. CREATE A NEW FOLDER CALLED SPARK IN C:\

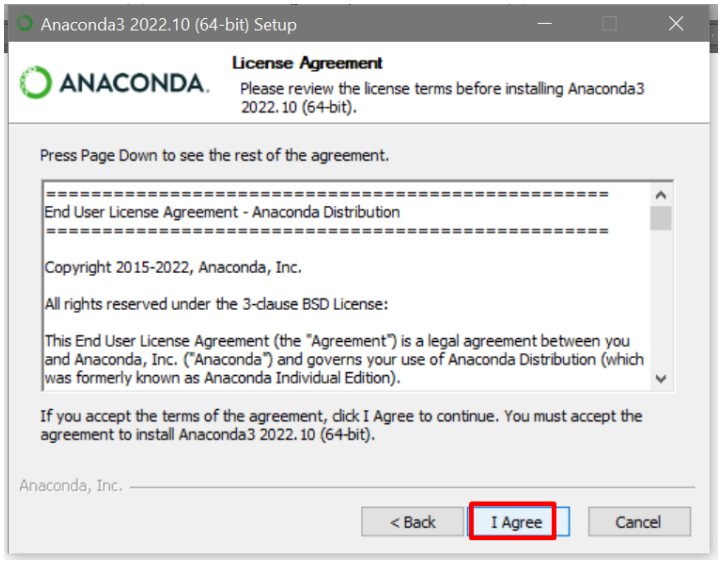


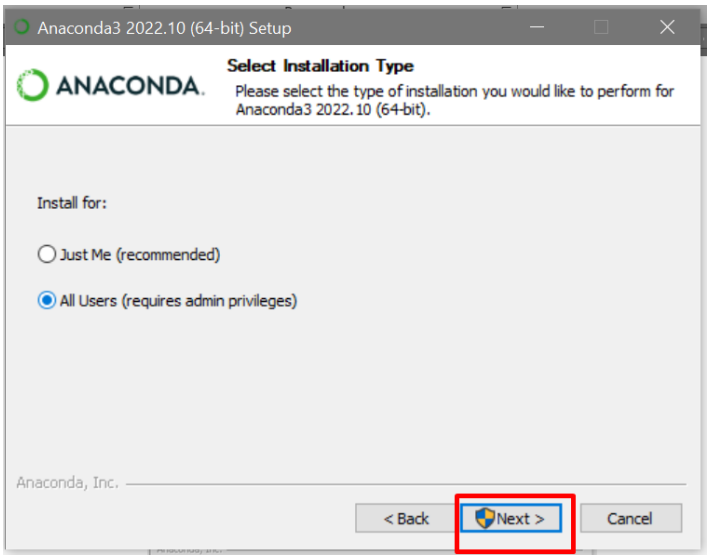


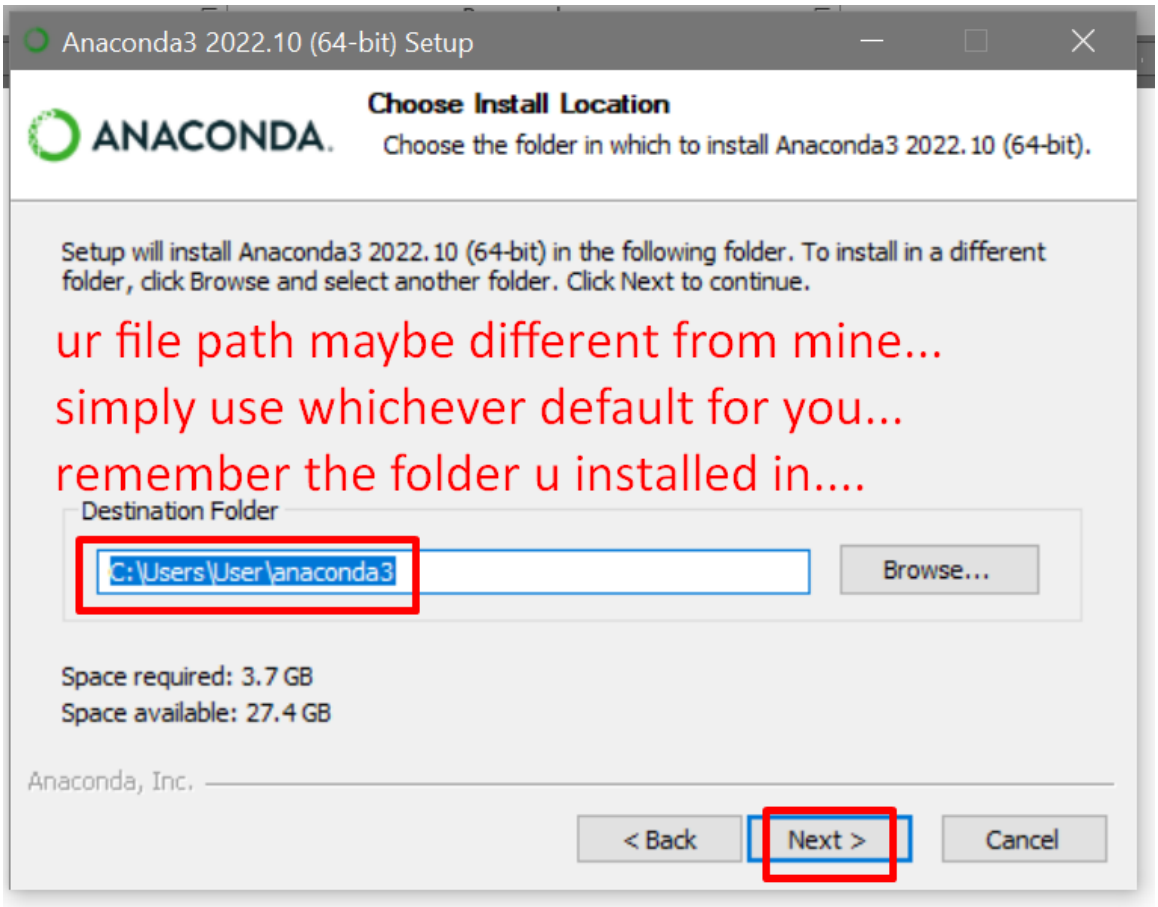
B. INSTALL ANACONDA

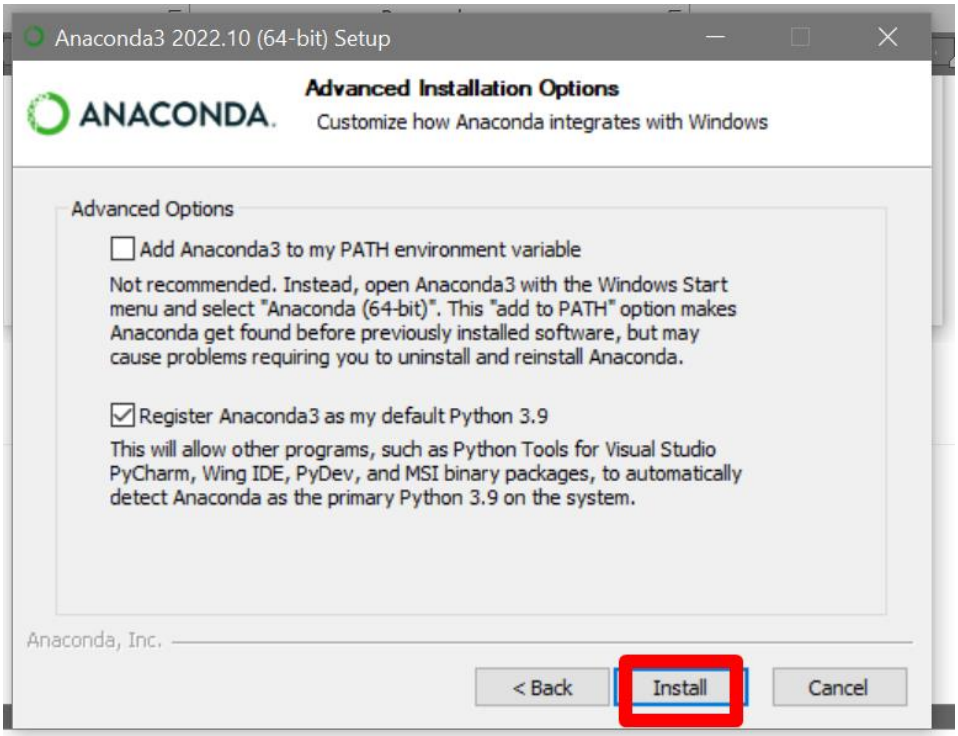
- https://repo.anaconda.com/archive/Anaconda3-2022.10-Windows-x86_64.exe

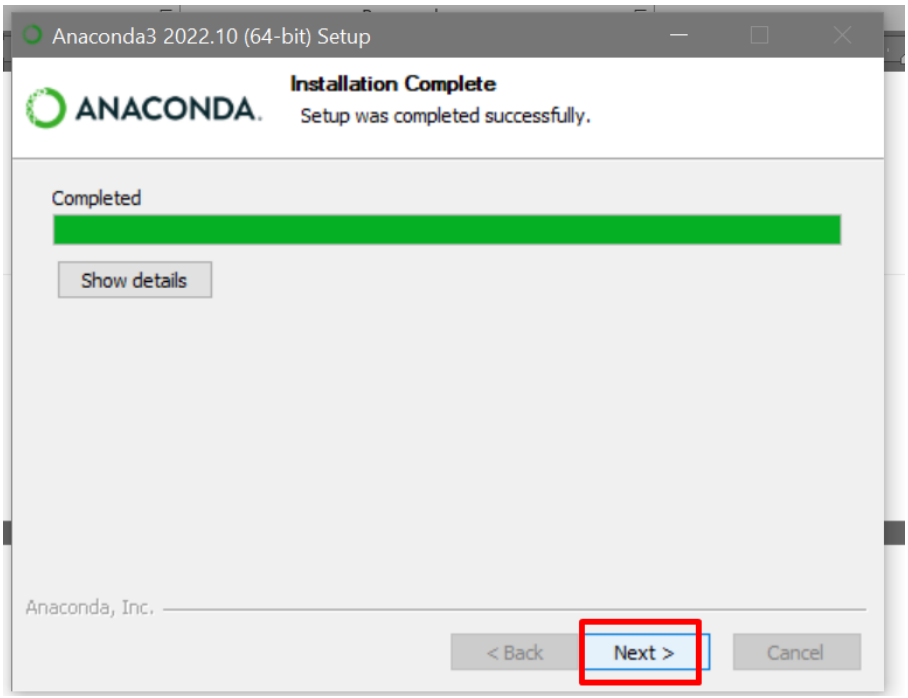


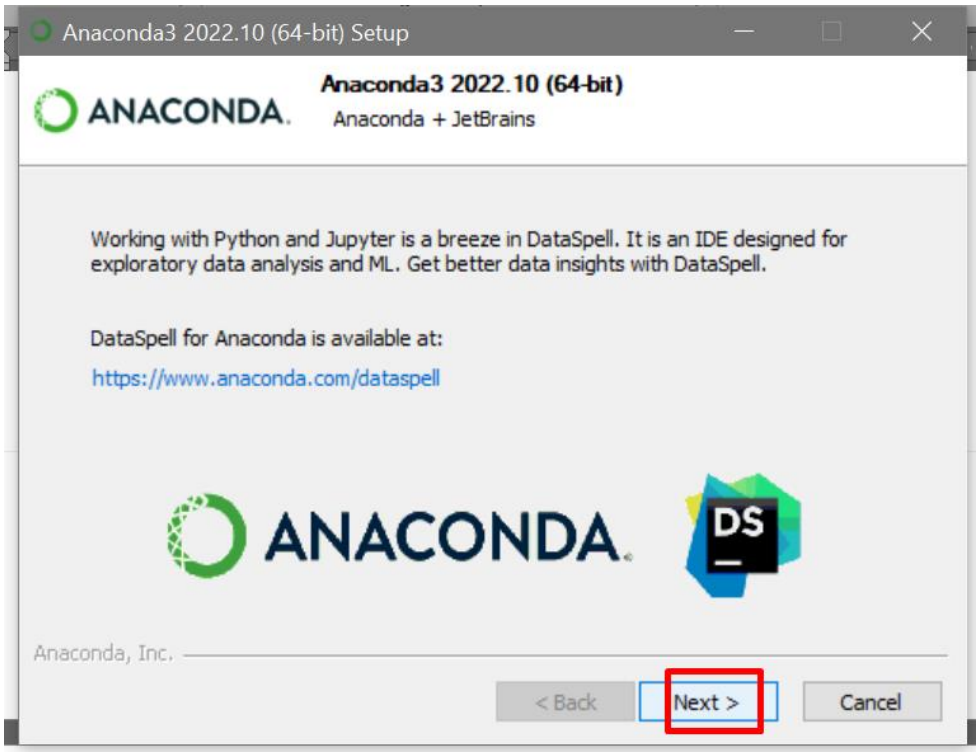


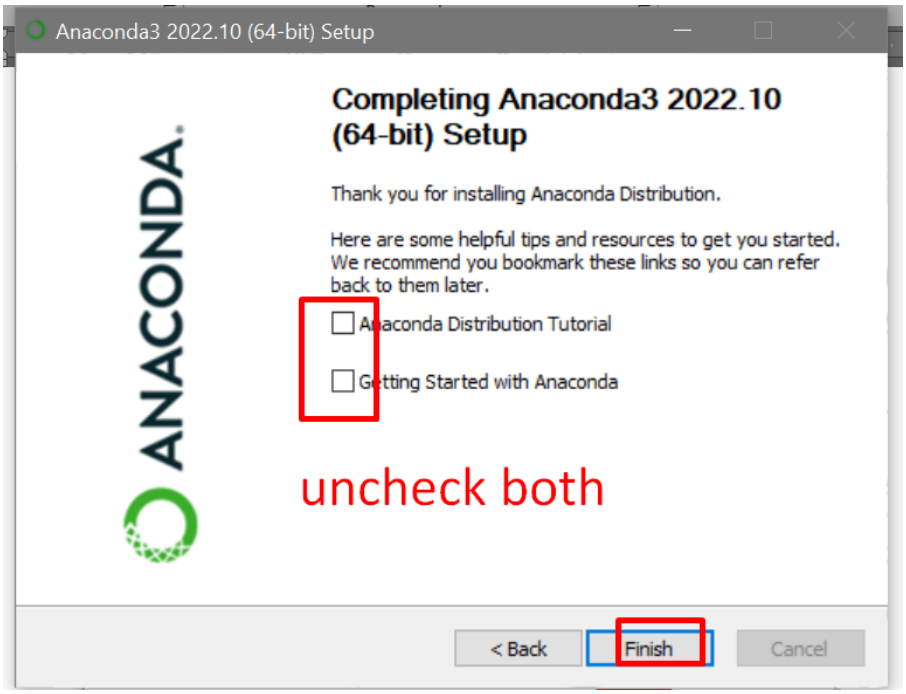


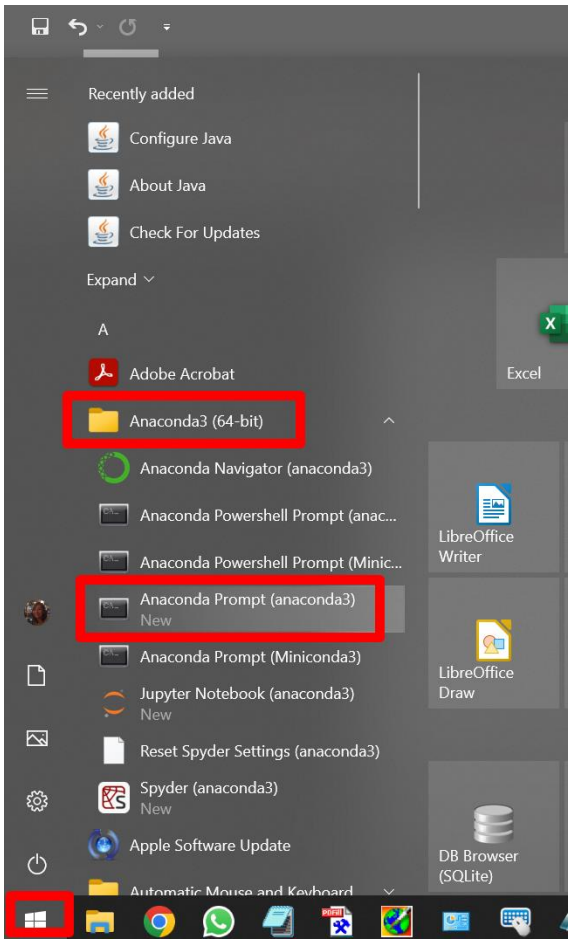


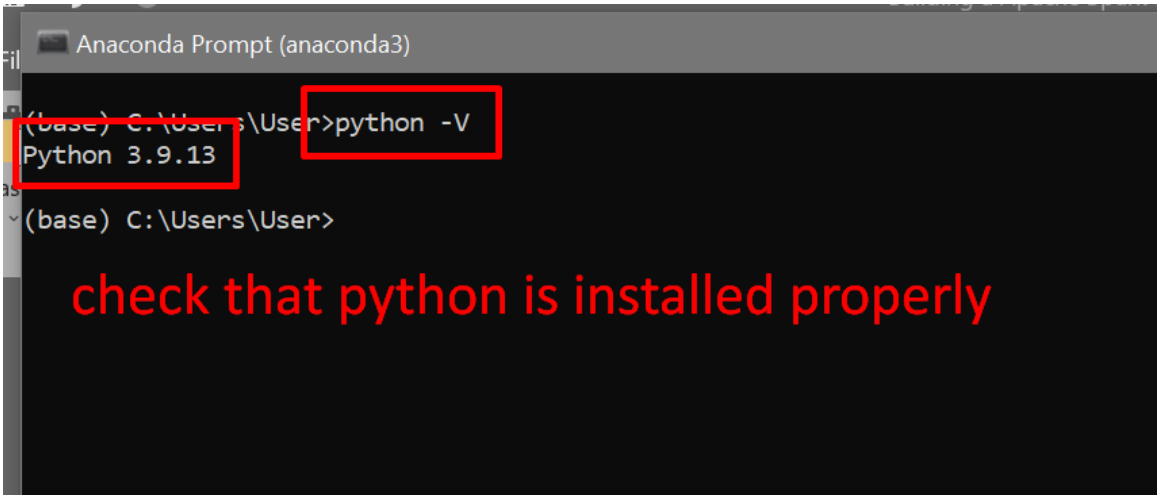






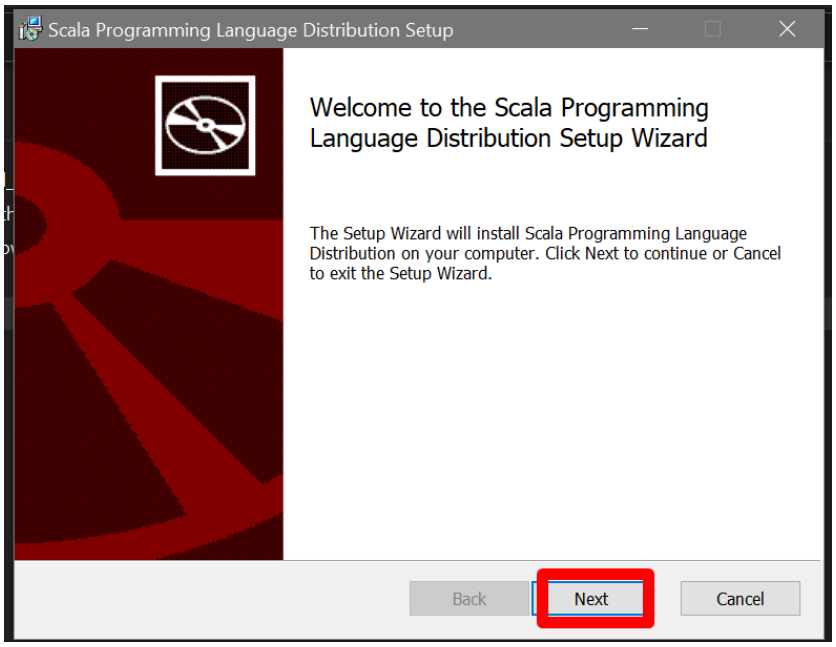


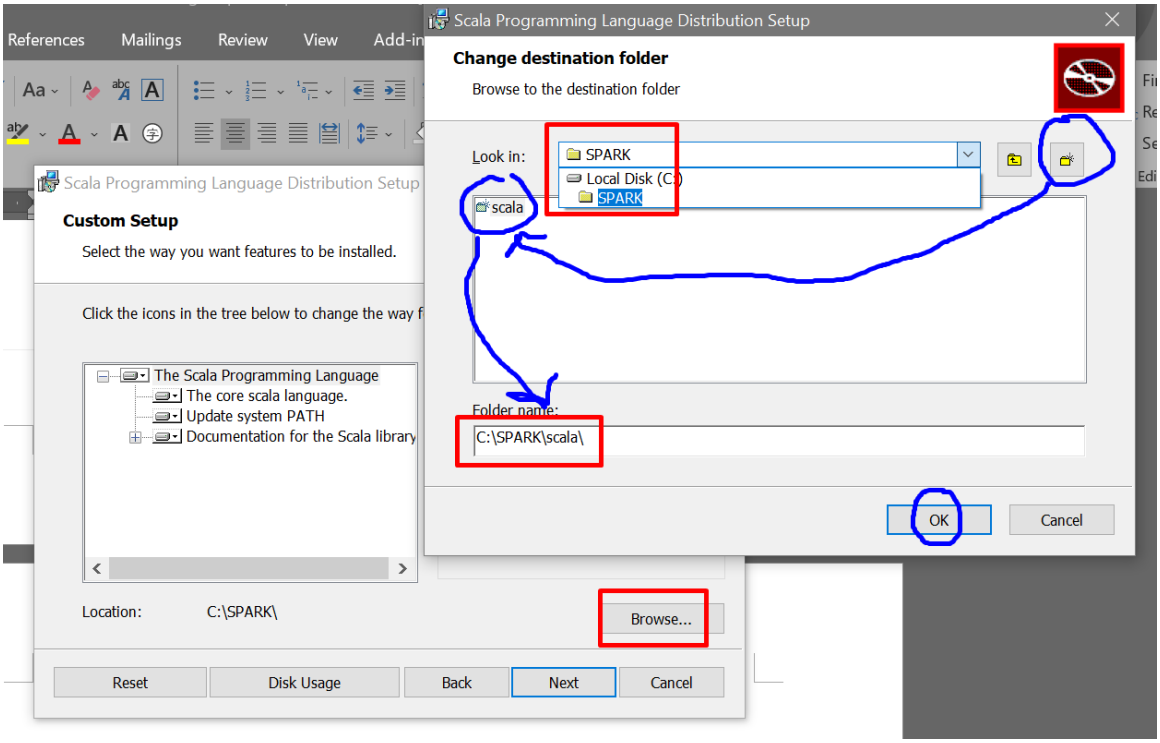


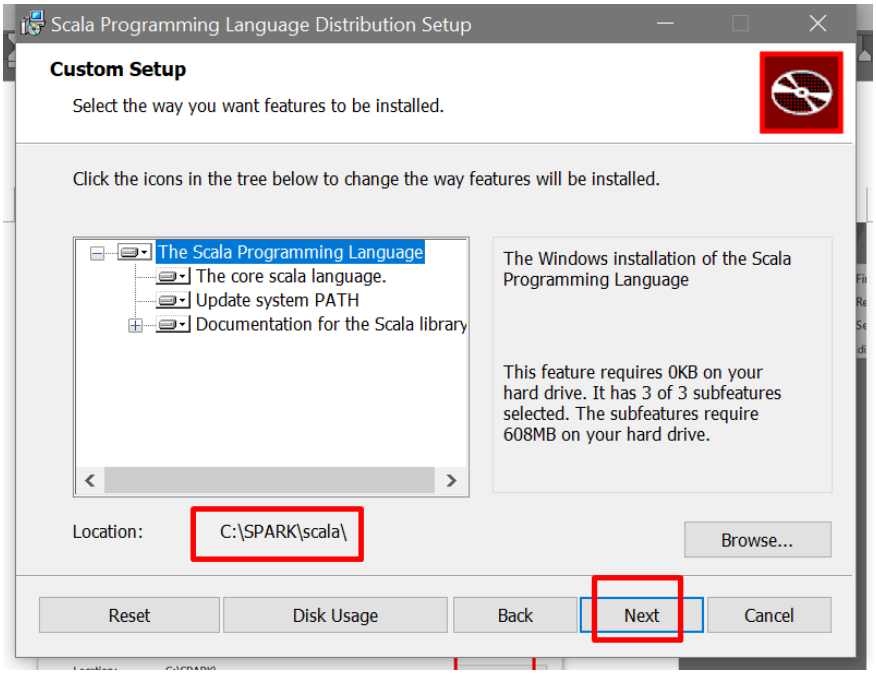


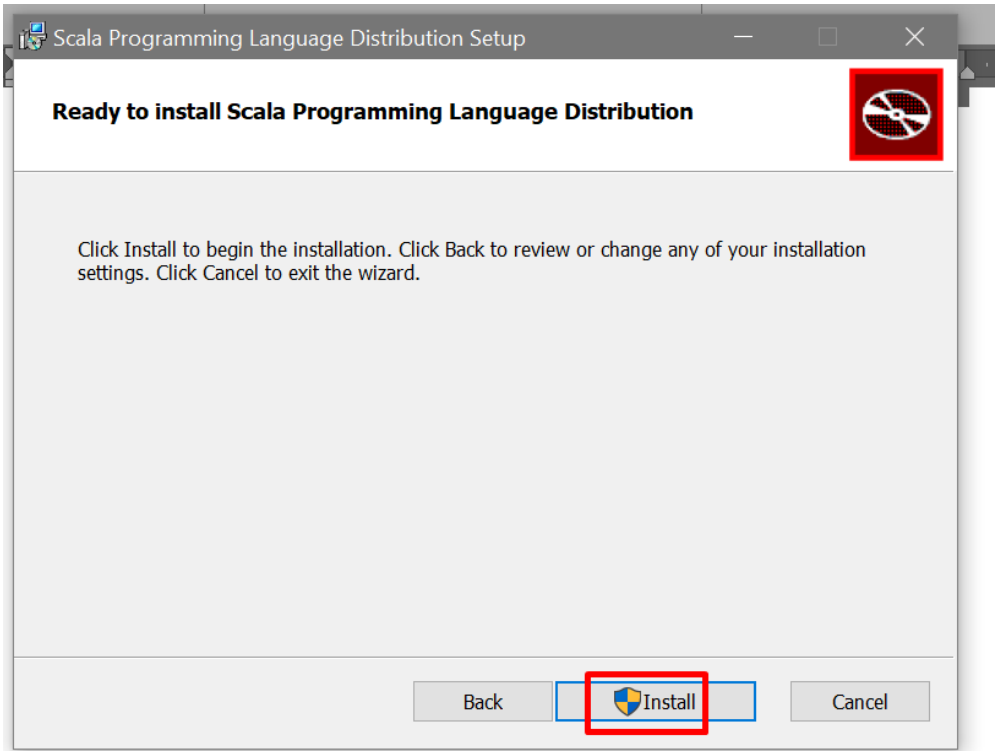
C. INSTALL SCALA

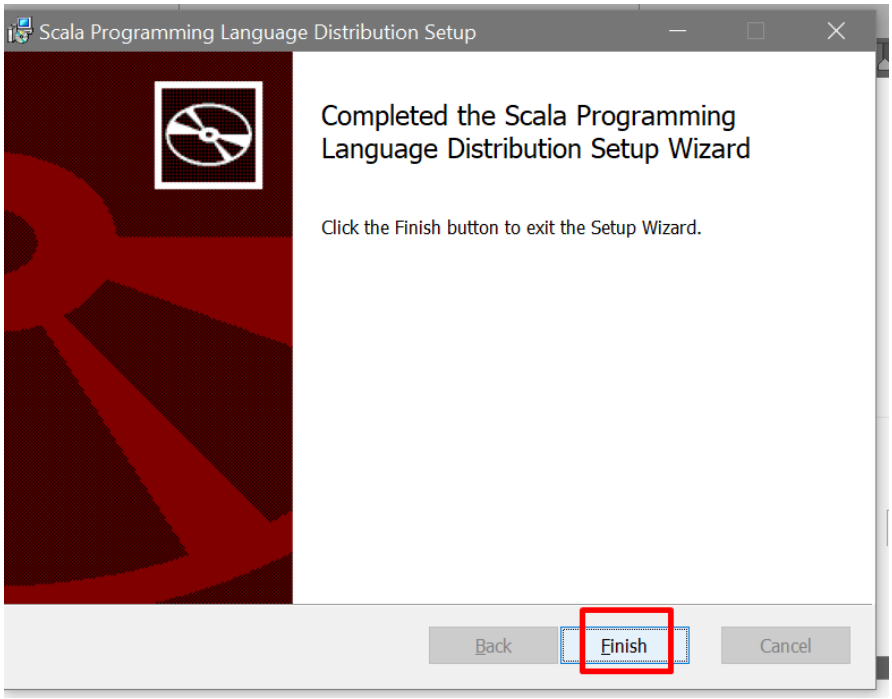
- https://downloads.lightbend.com/scala/2.12.4/scala-2.12.4.msi









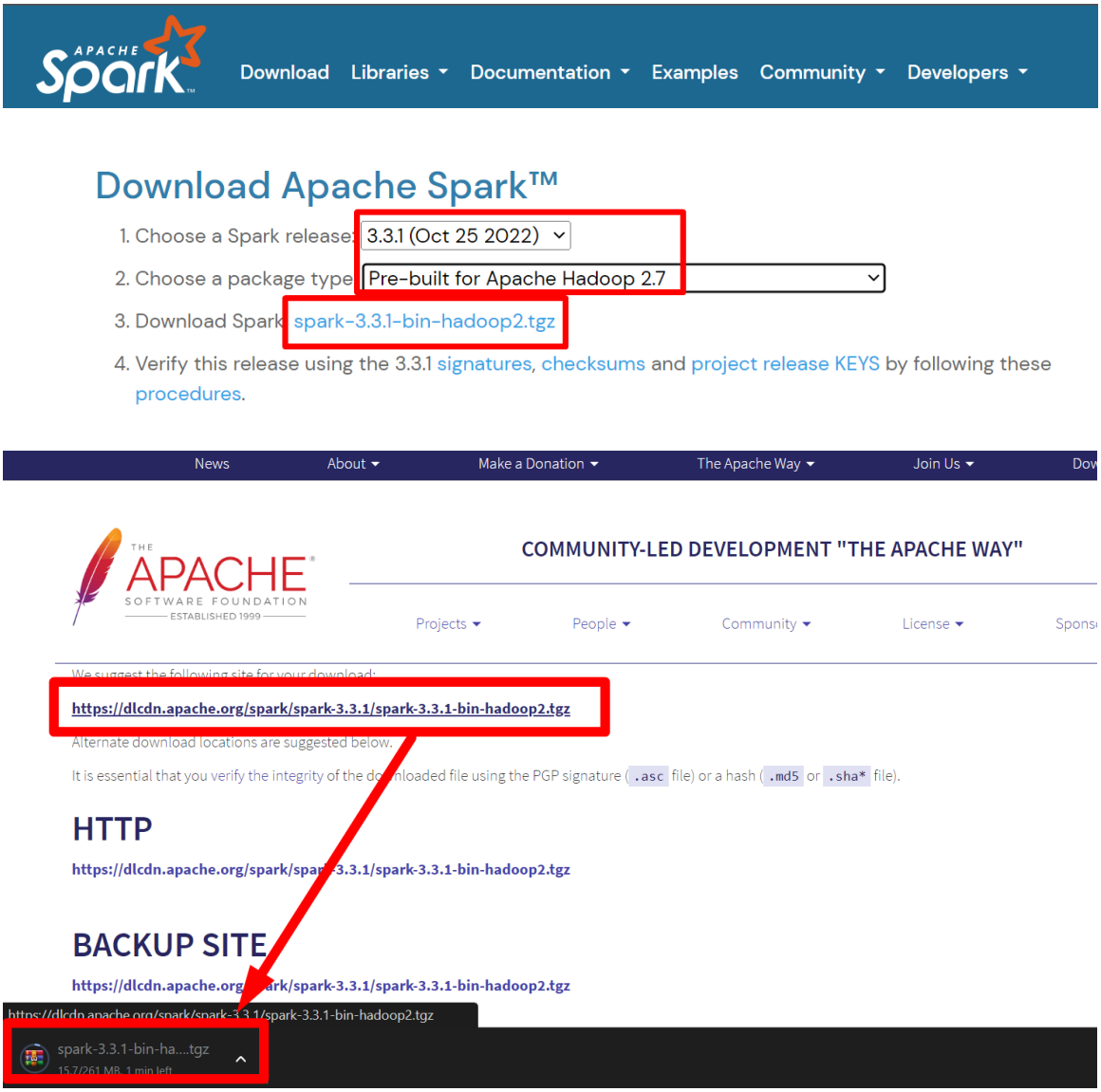


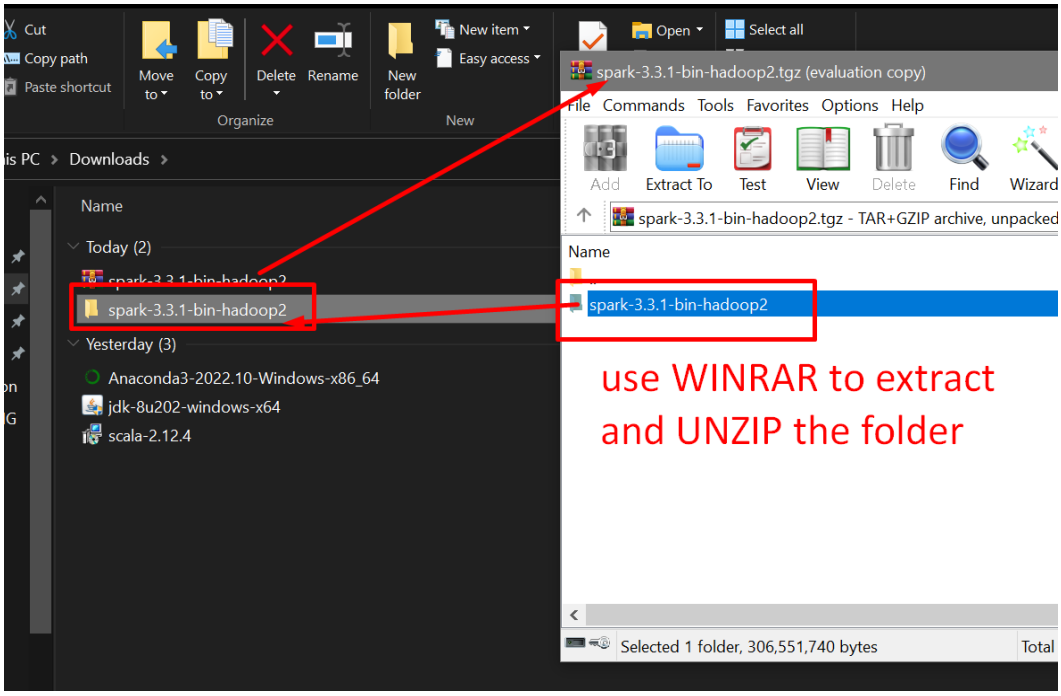
D. INSTALL APACHE SPARK

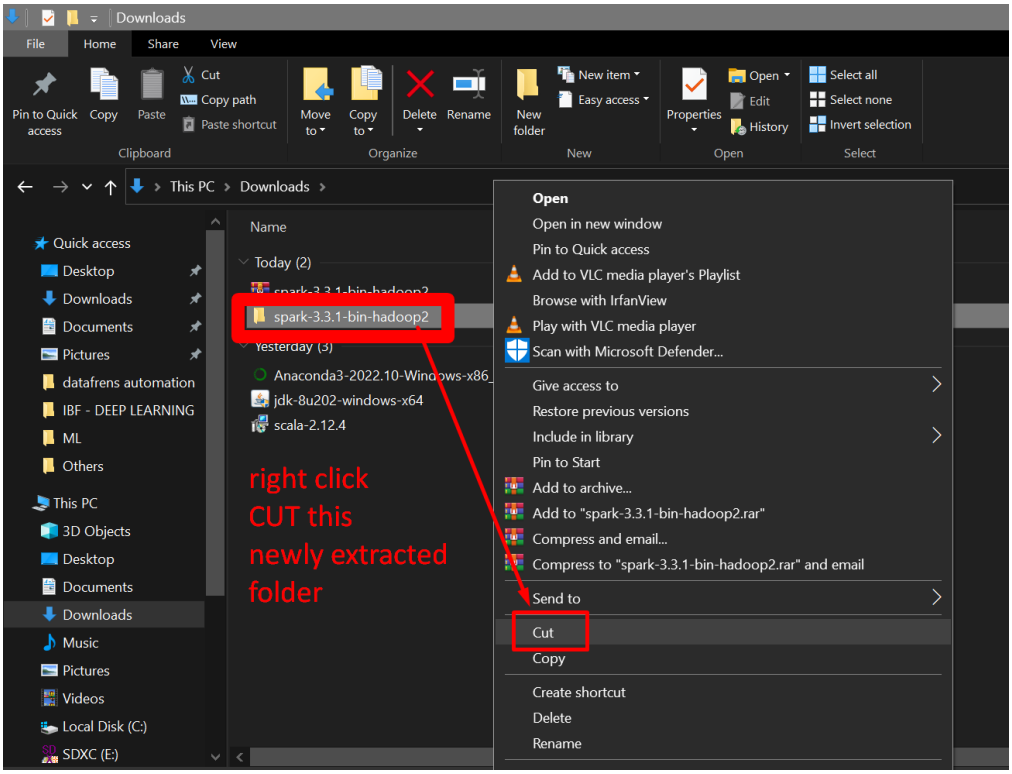
- https://www.apache.org/dyn/closer.lua/spark/spark-3.3.1/spark-3.3.1-bin-hadoop2.tgz

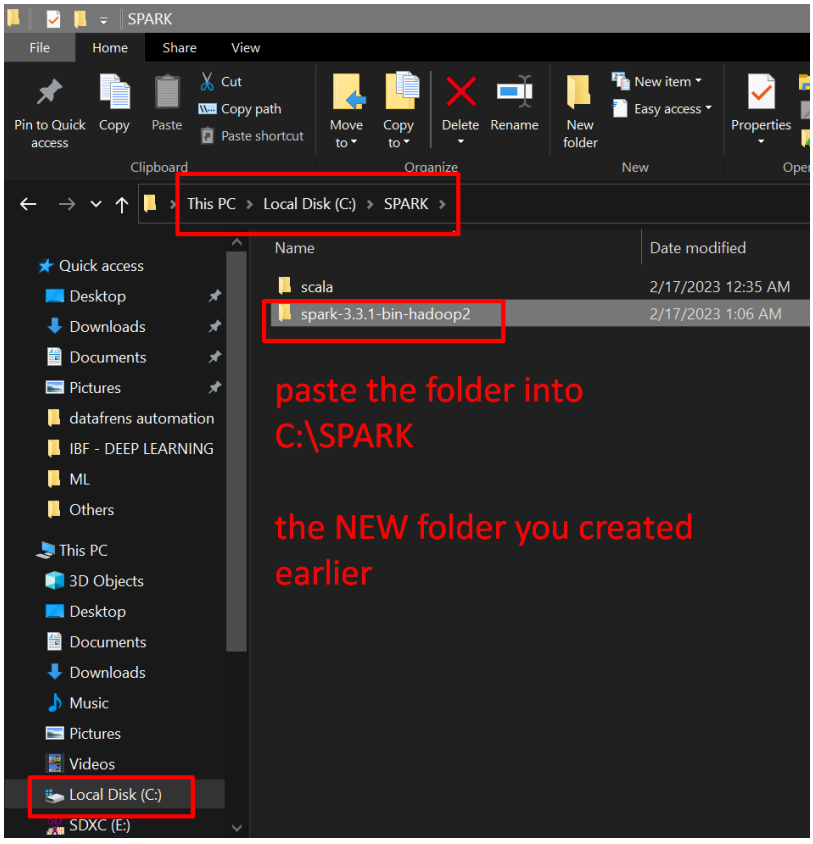
OR

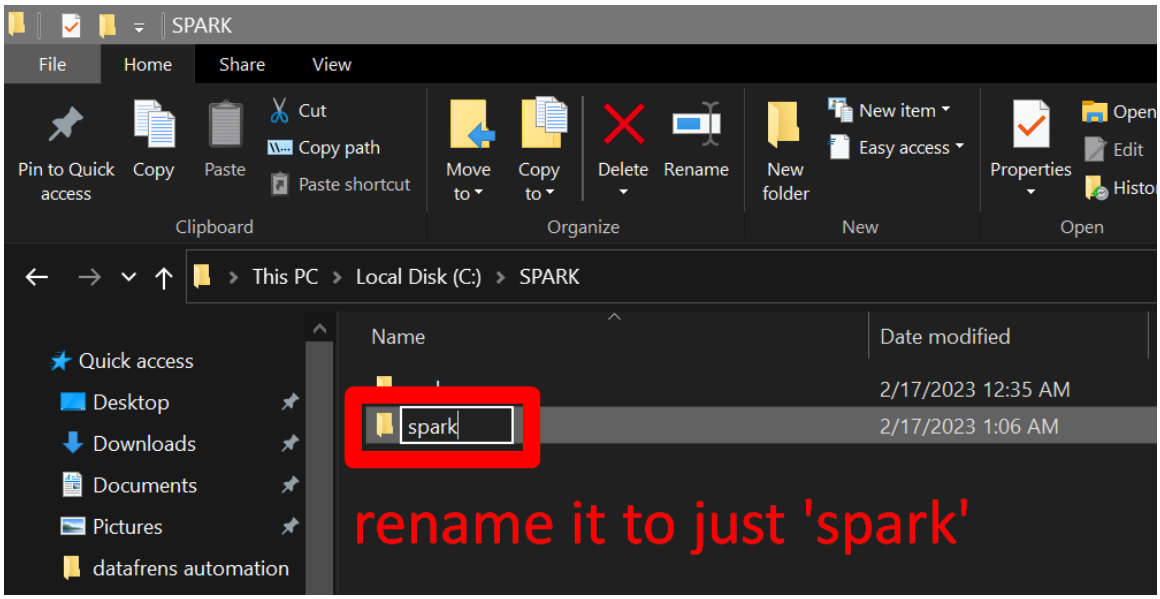
- https://spark.apache.org/downloads.html





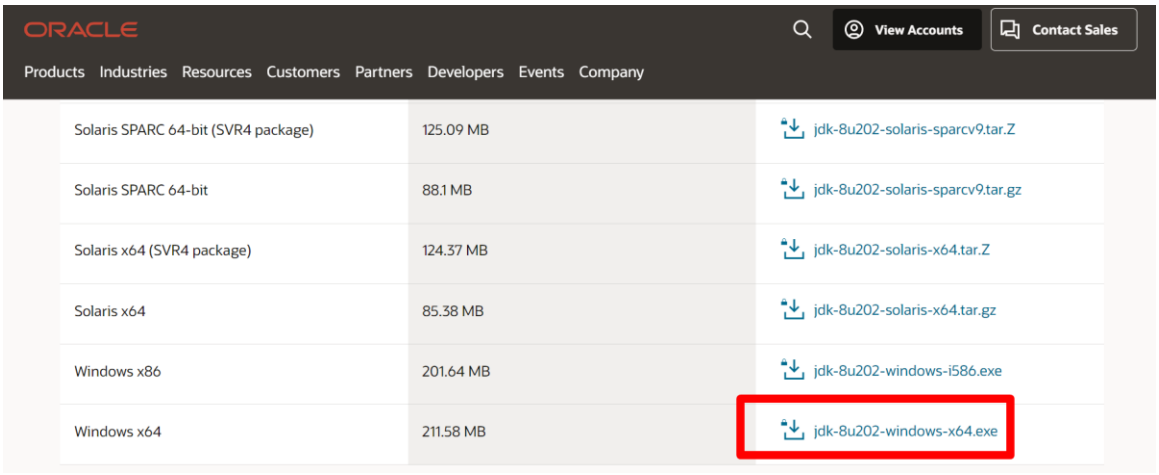


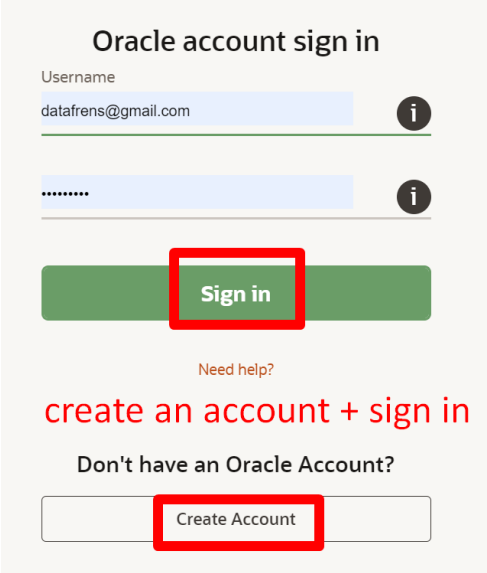




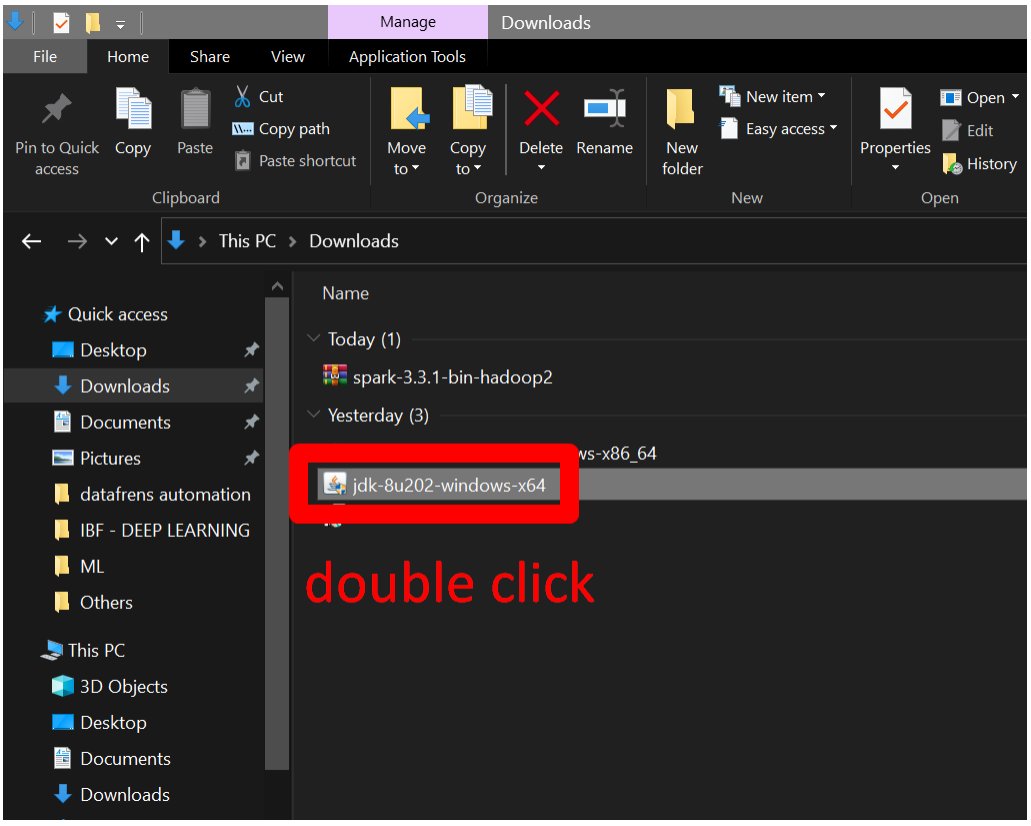
E. INSTALL JDK 8

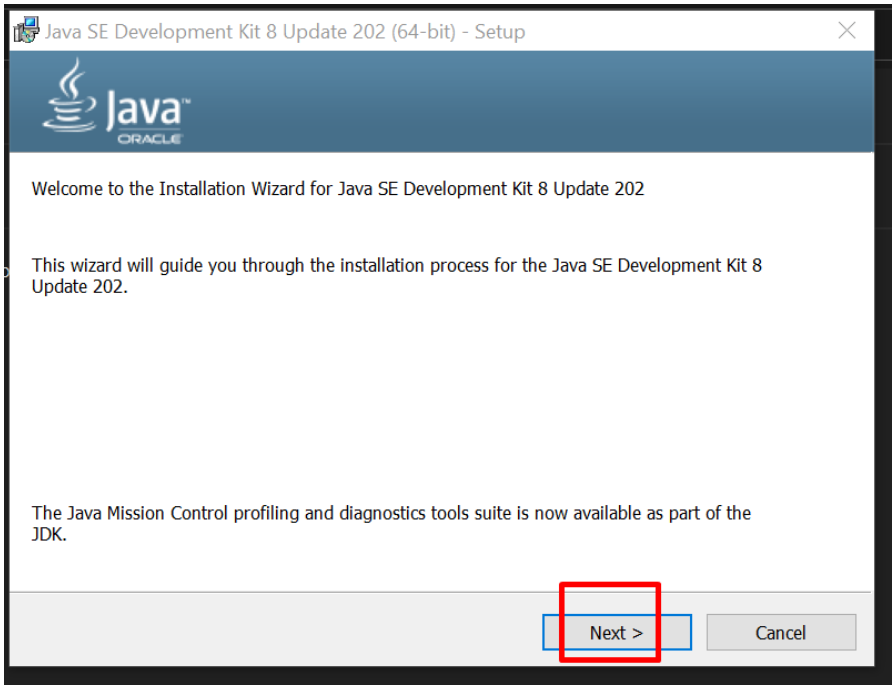
https://www.oracle.com/java/technologies/javase/javase8-archive-downloads.html#license-lightbox

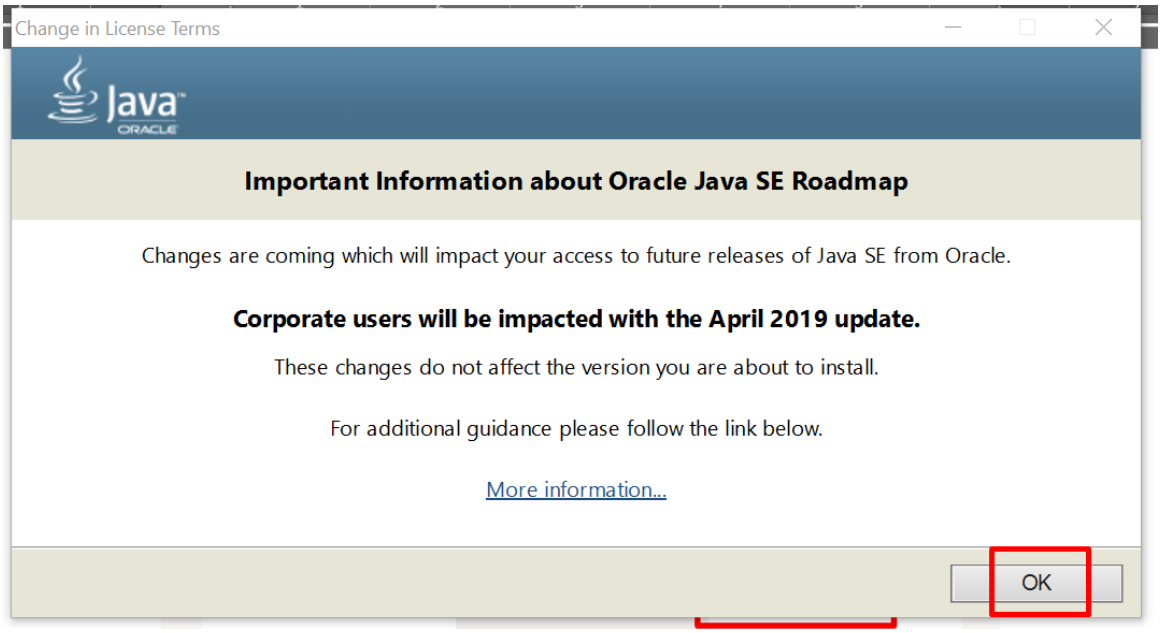


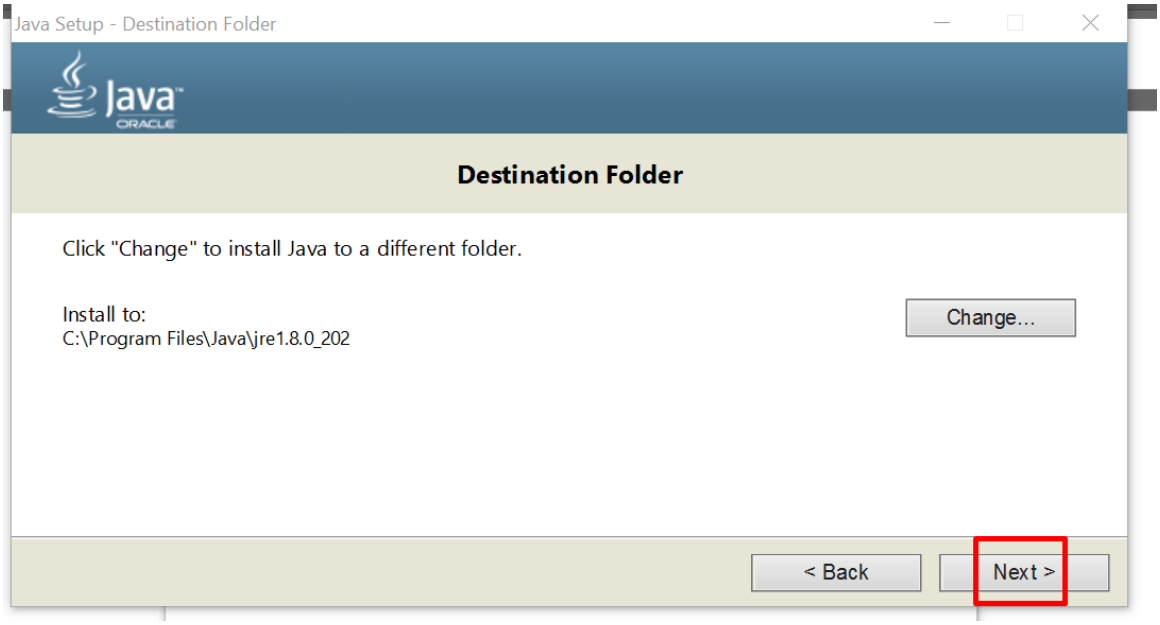


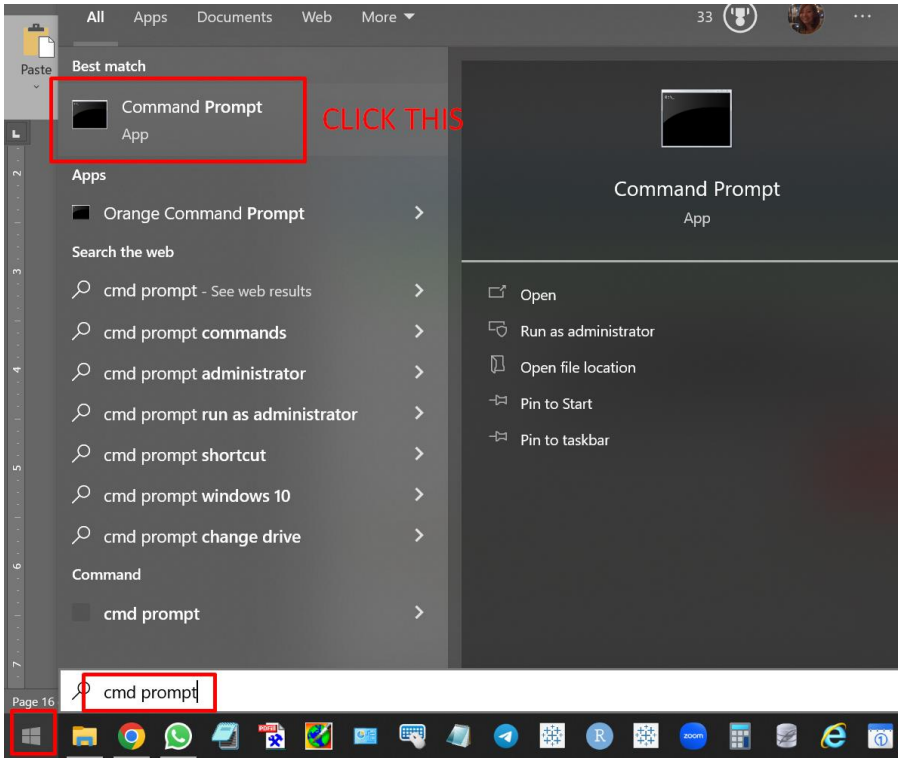
It will download a file after signing in…

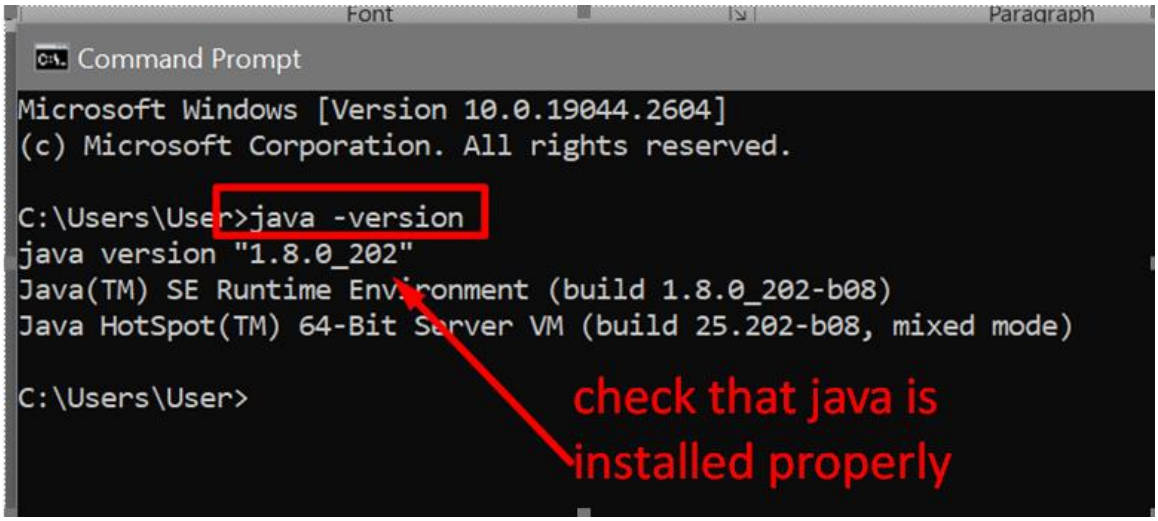










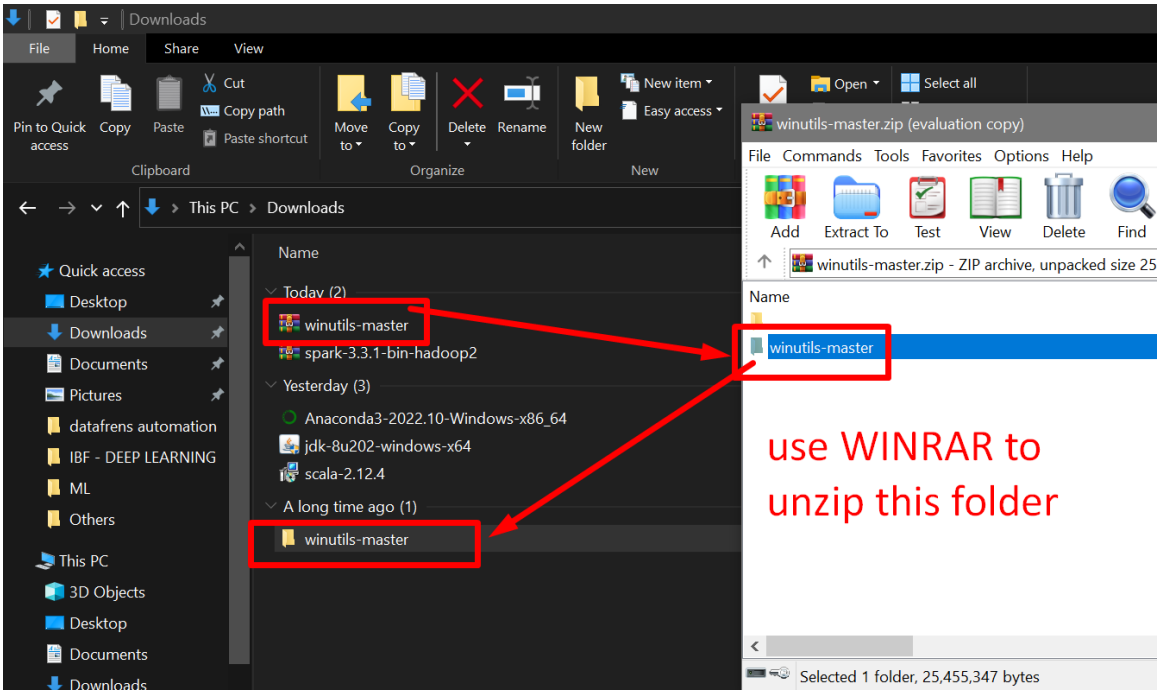


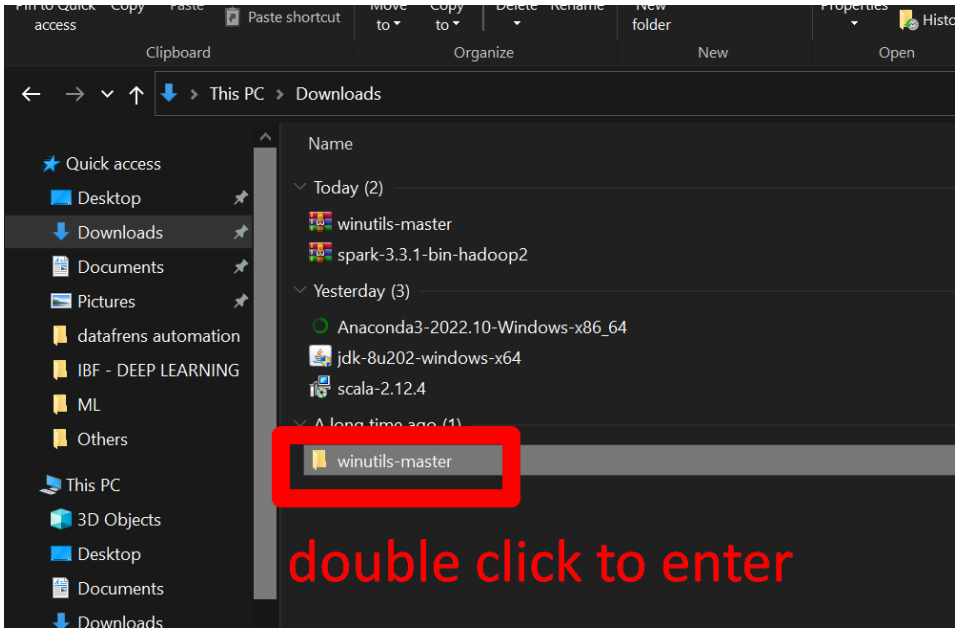
At the time of writing, I know that only JAVA 8 or 11 can work

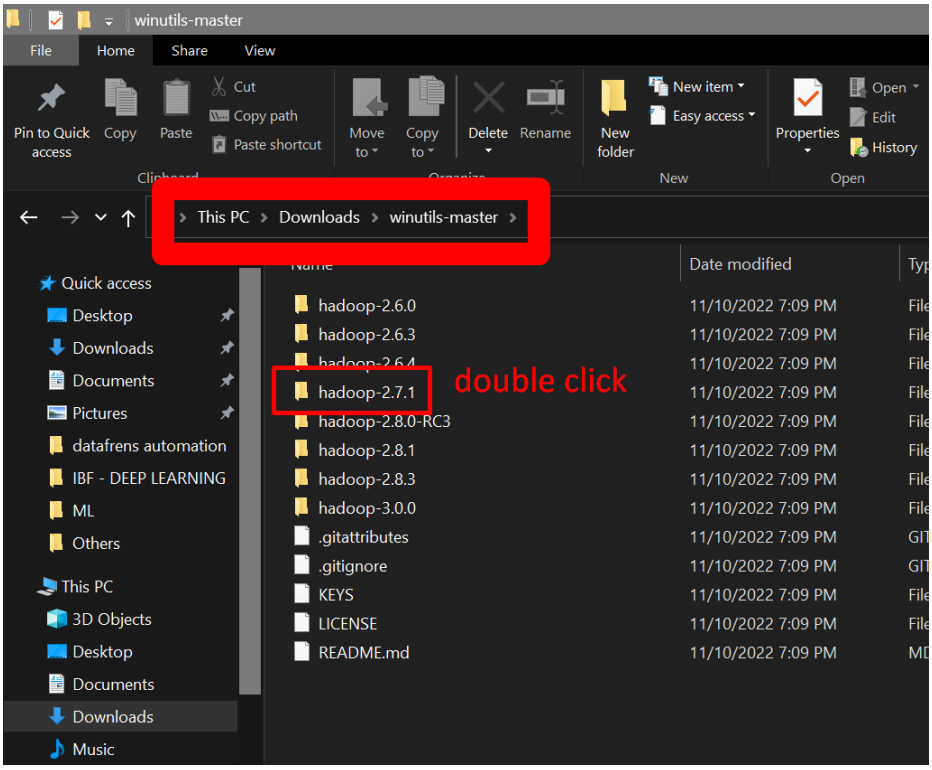
I read that JAVA 9 gives problems and should not be used.

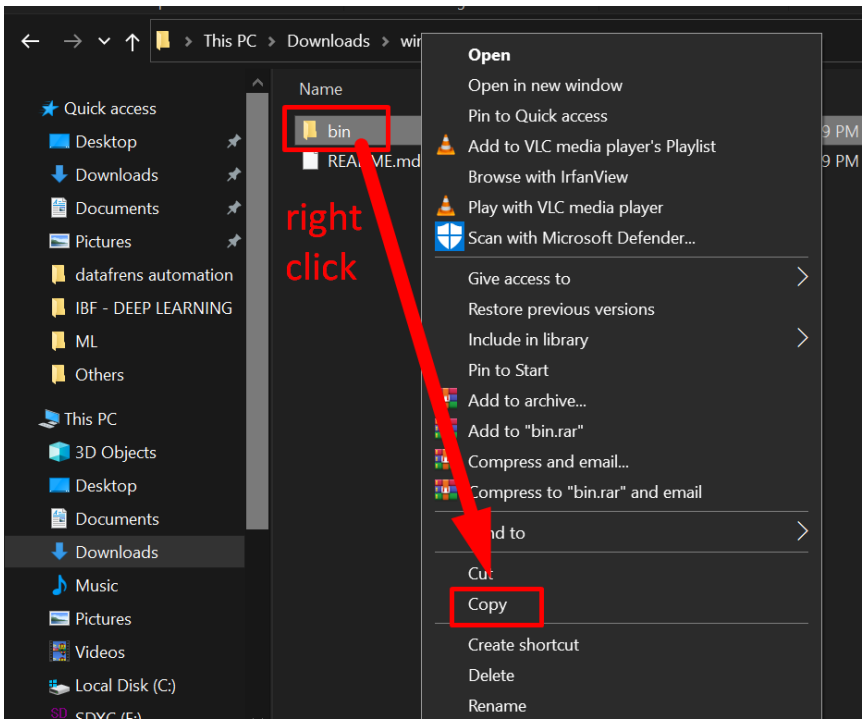
F. INSTALL HADOOP

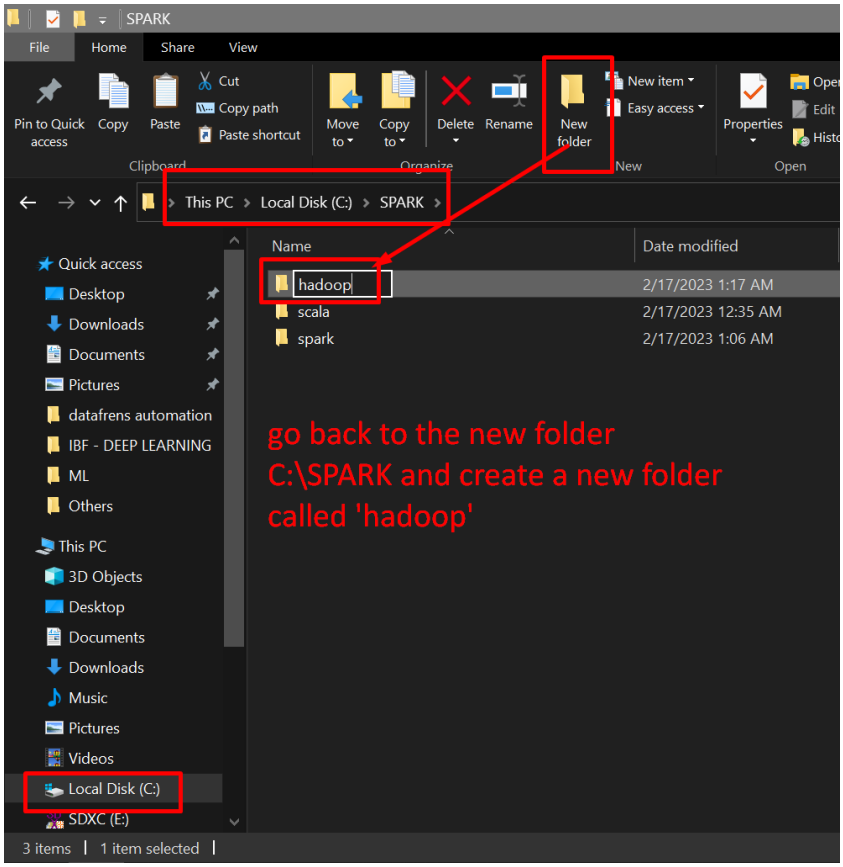
https://github.com/steveloughran/winutils/archive/refs/heads/master.zip

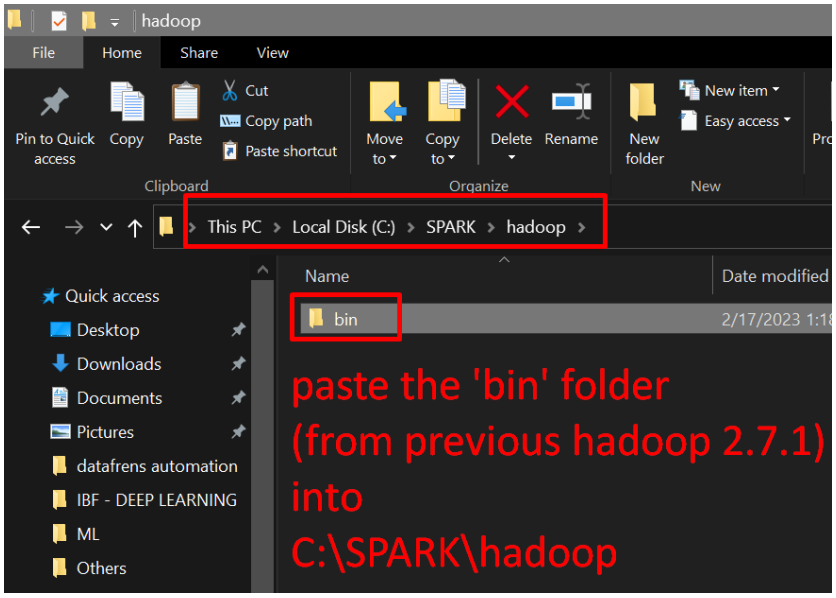








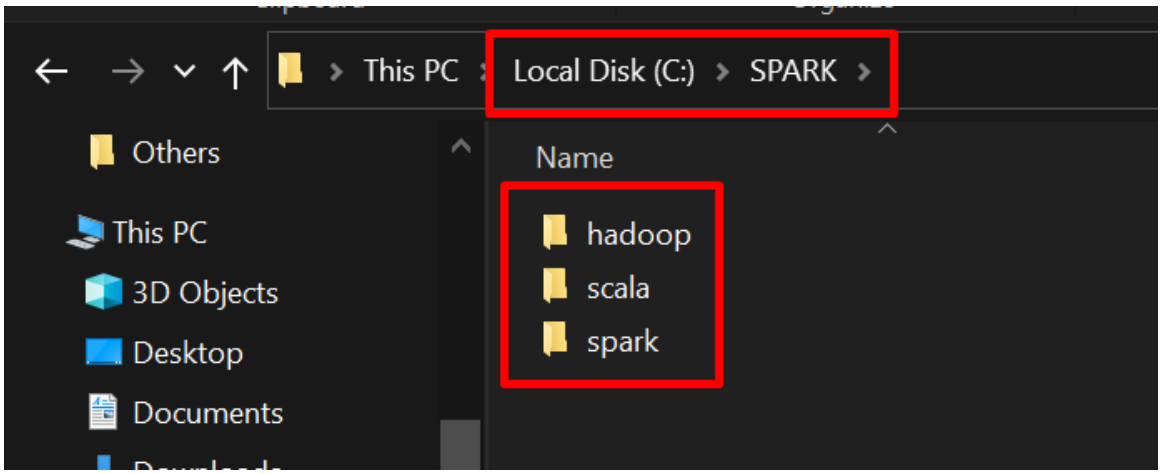


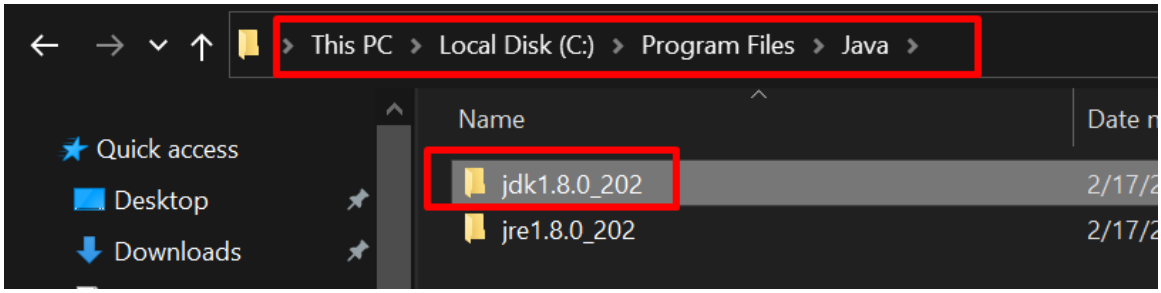


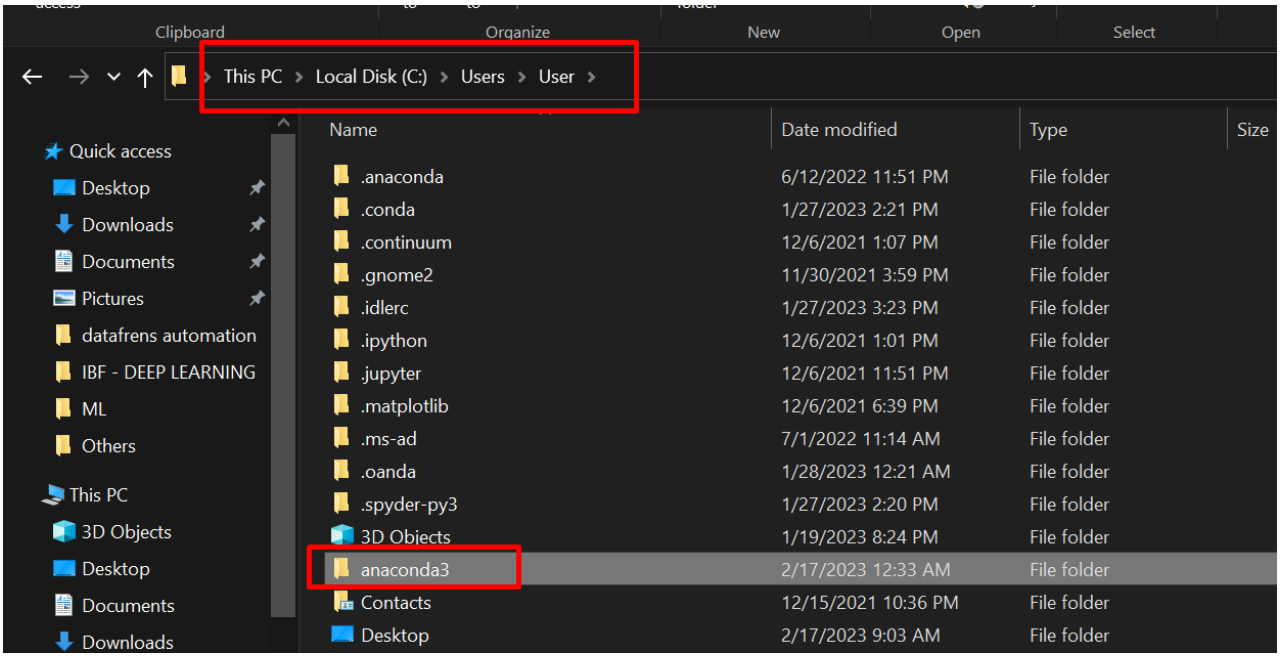
In a nutshell… we are trying to create all these…

. Environment Variables (windows)

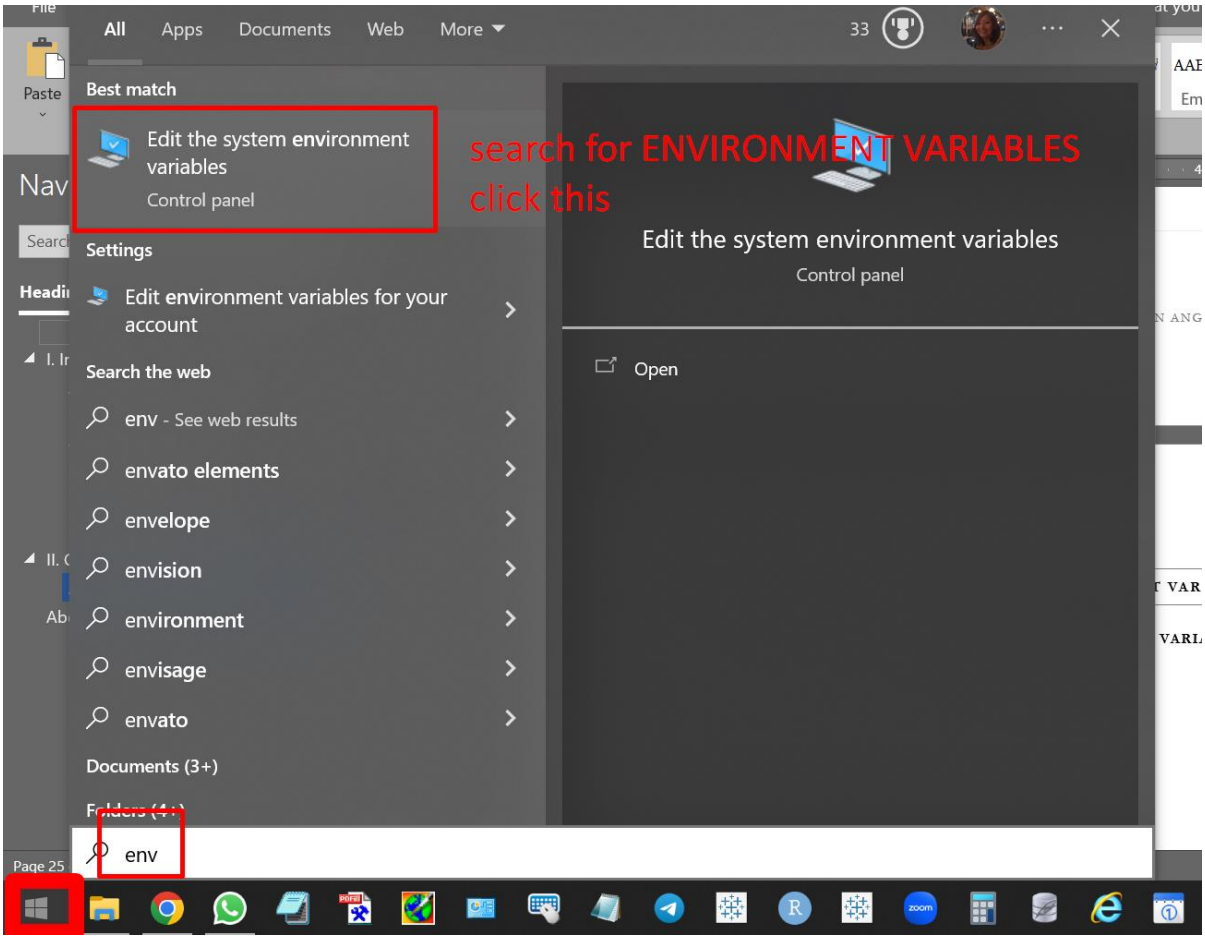
- HADOOP_HOME = C:\SPARK\hadoop
- JAVA_HOME = C:\Program Files\Java\jdk1.8.0_202
- SCALA_HOME = C:\SPARK\scala
- SPARK_HOME = C:\SPARK\spark
- PYSPARK_PYTHON = C:\Users\user\anaconda3\python.exe



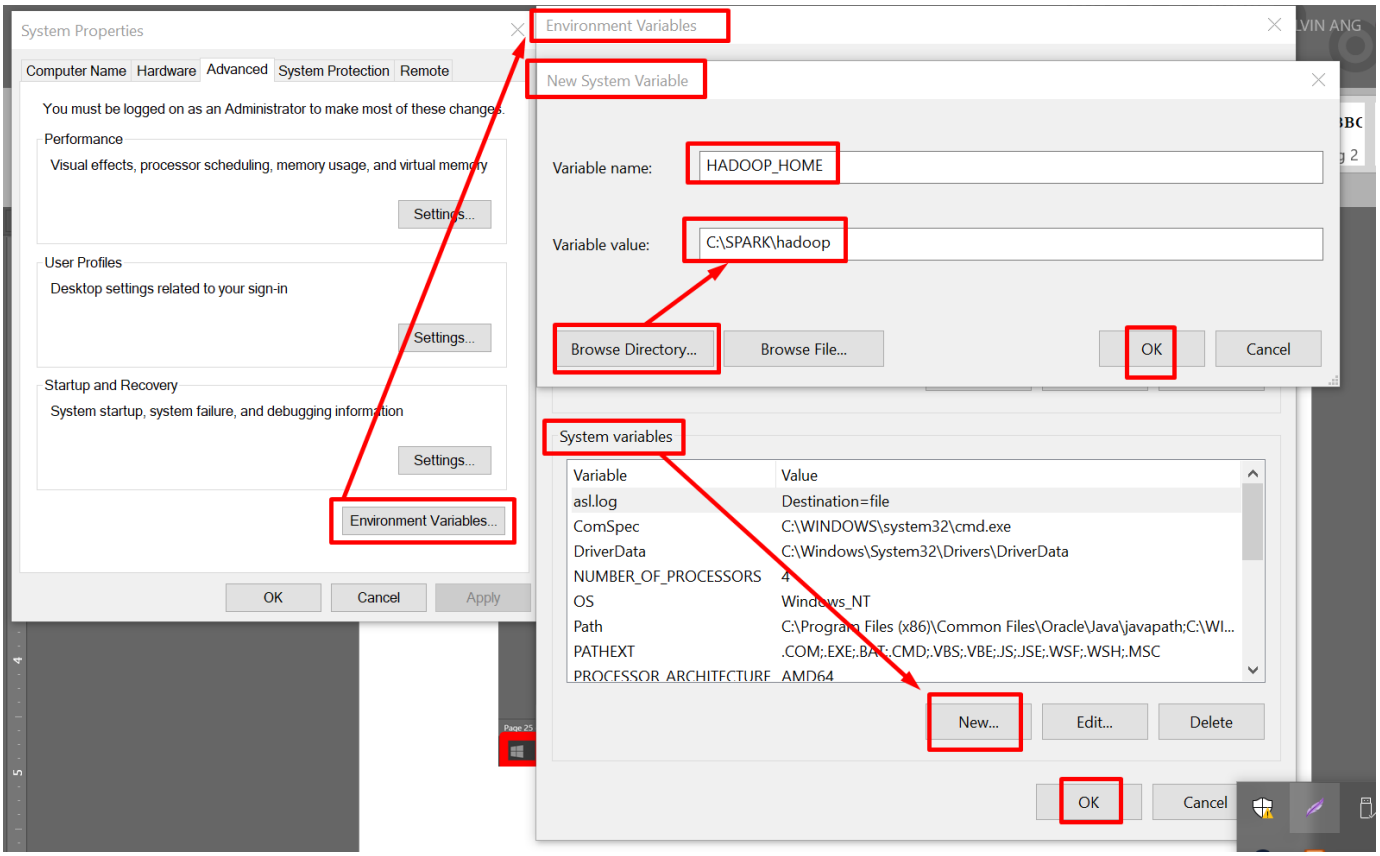




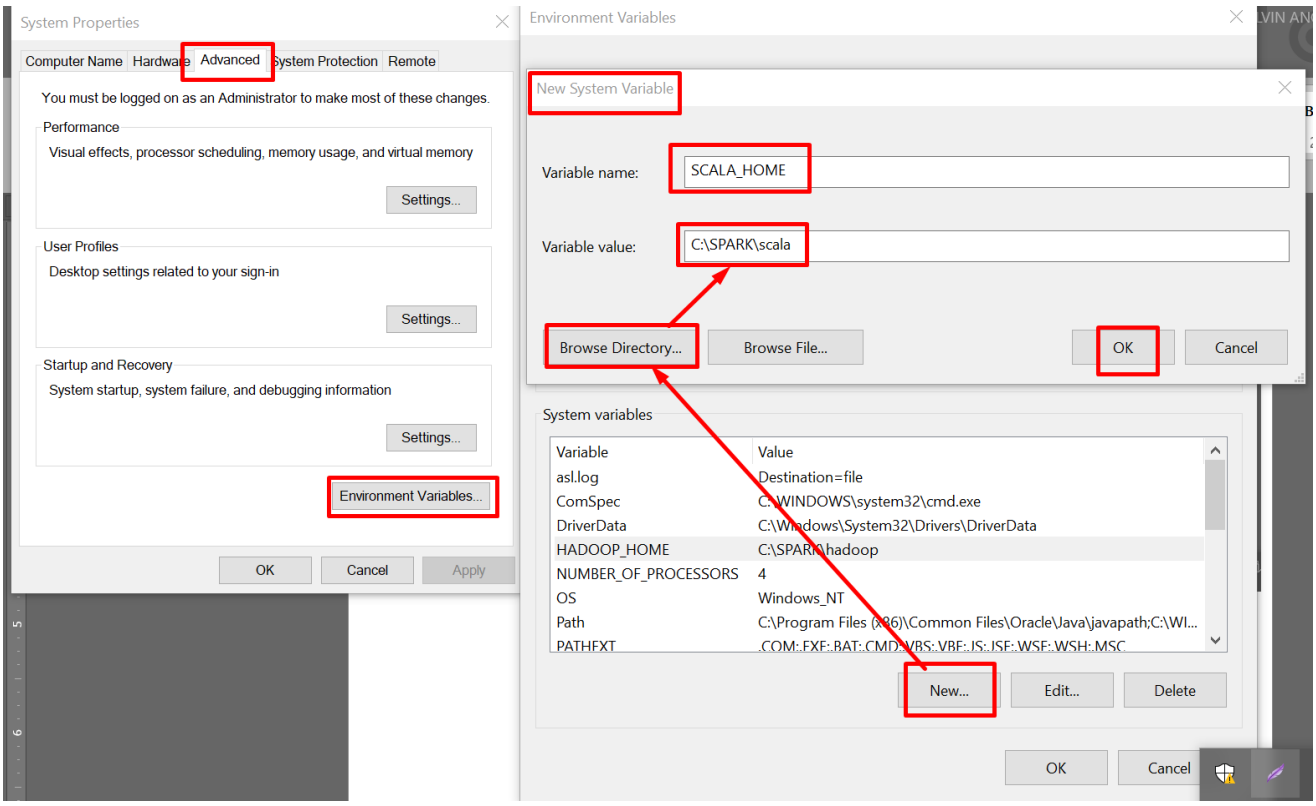
A. CREATING SYSTEM VARIABLES



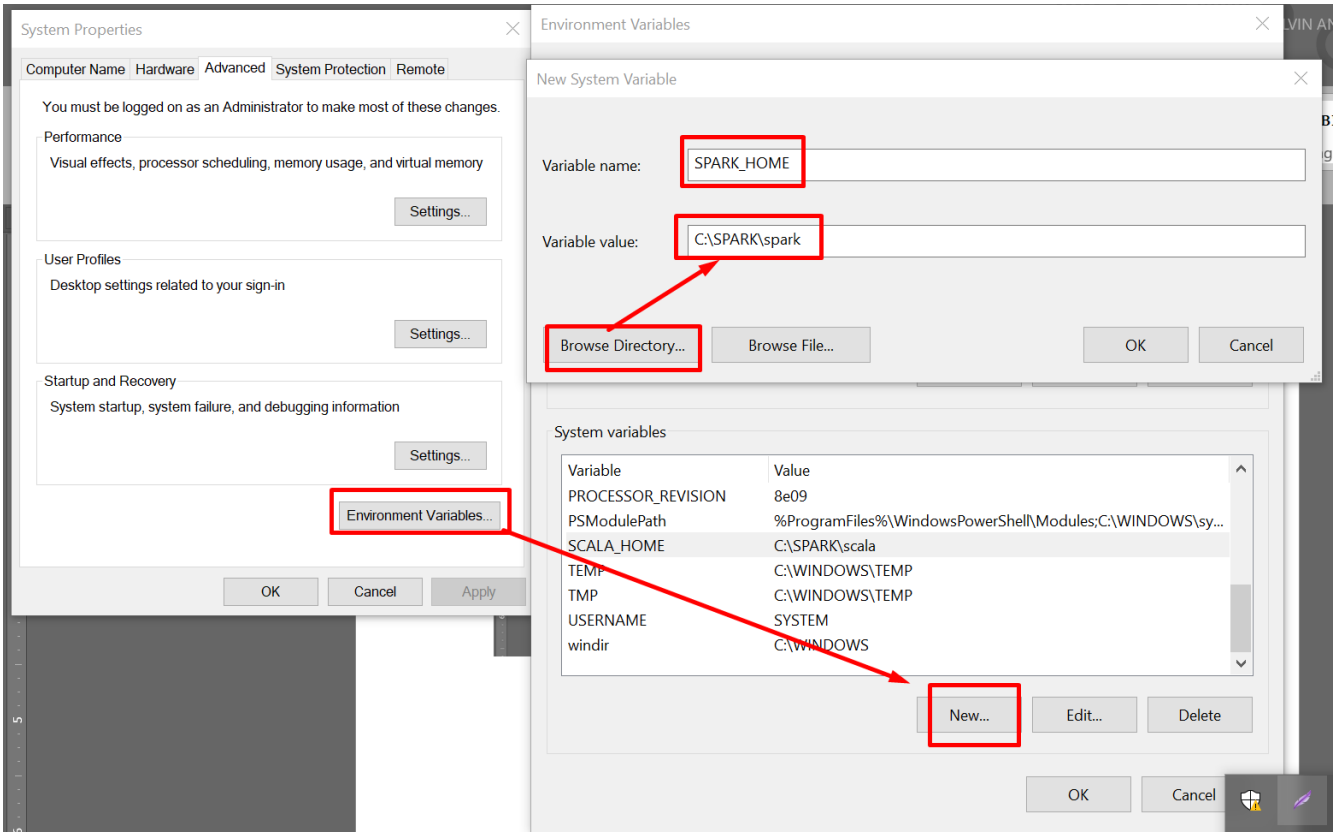
1. CREATE HADOOP_HOME VARIABLE



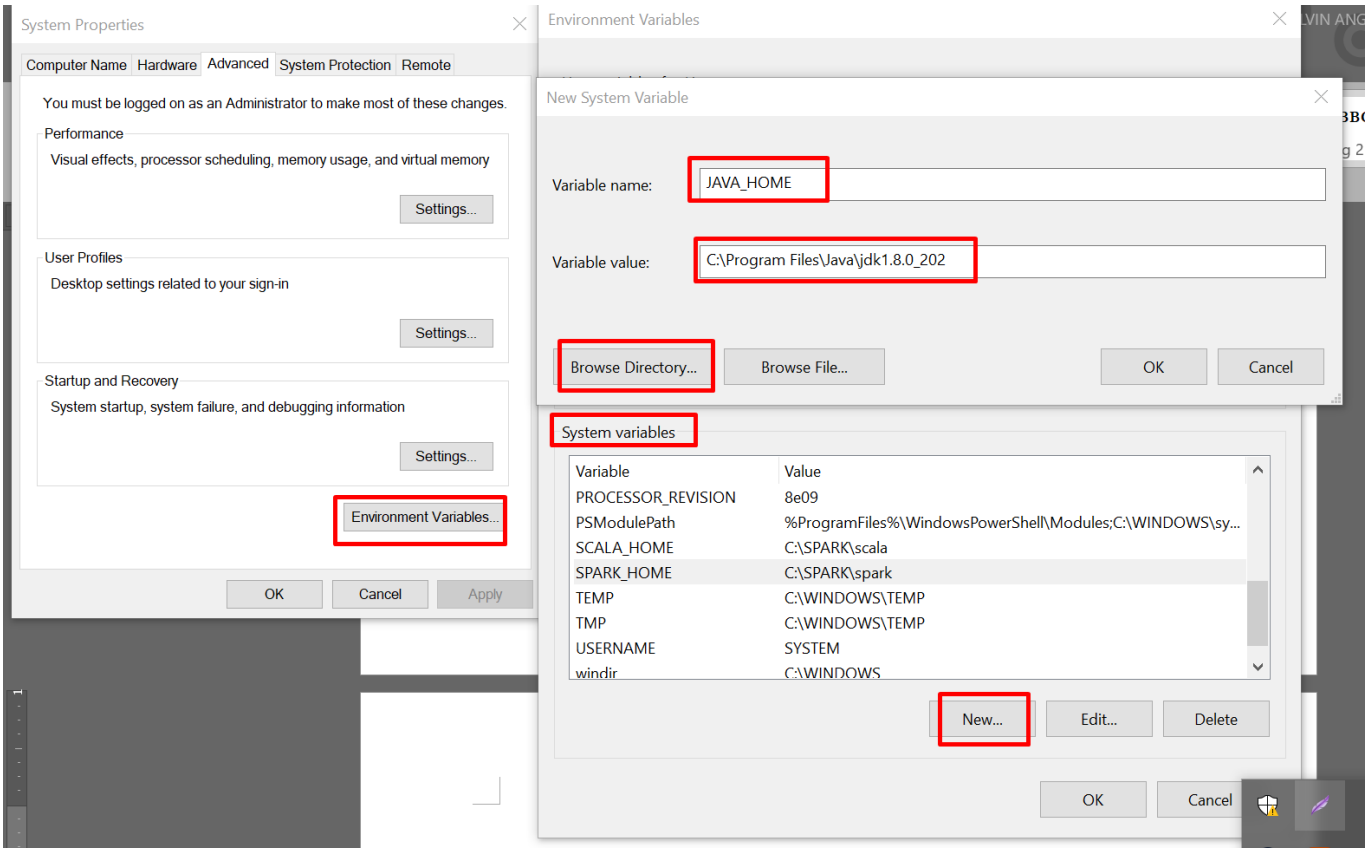
2. CREATE SCALA_HOME VARIABLE



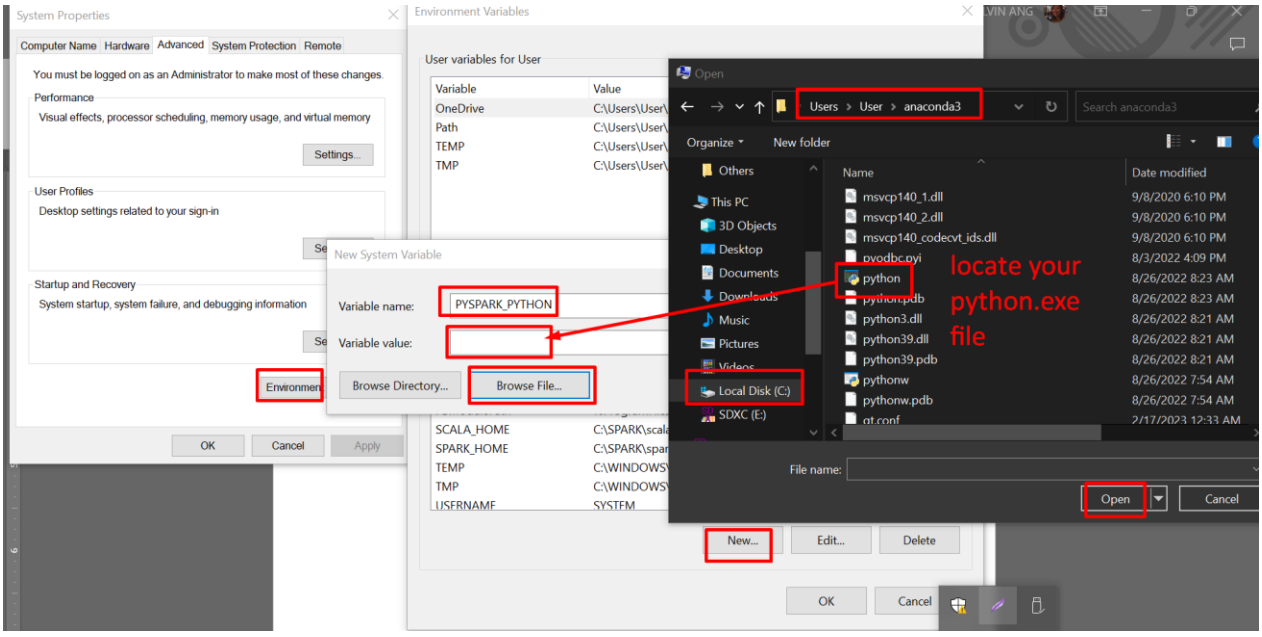
3. CREATE SPARK_HOME VARIABLE

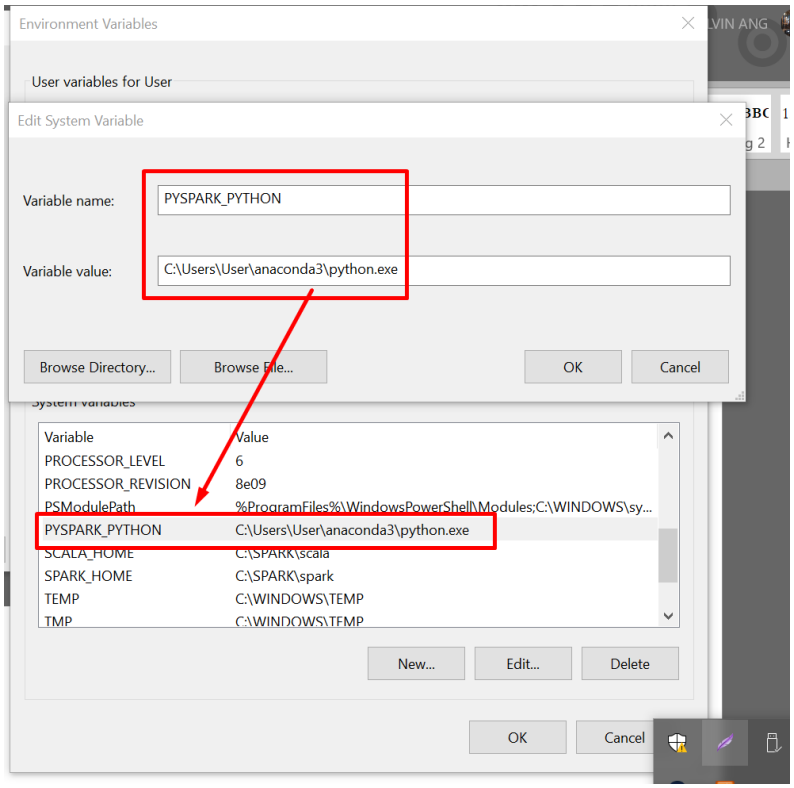


4. CREATE JAVA_HOME VARIABLE

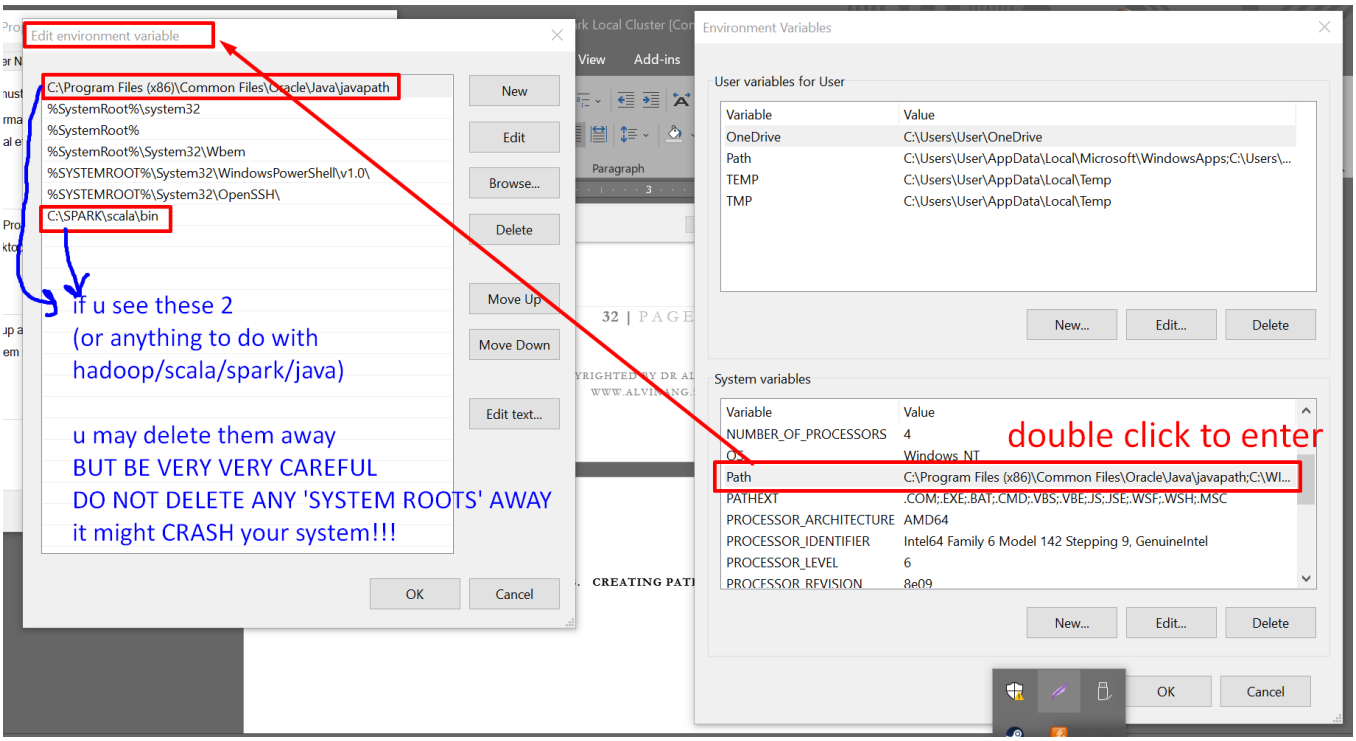


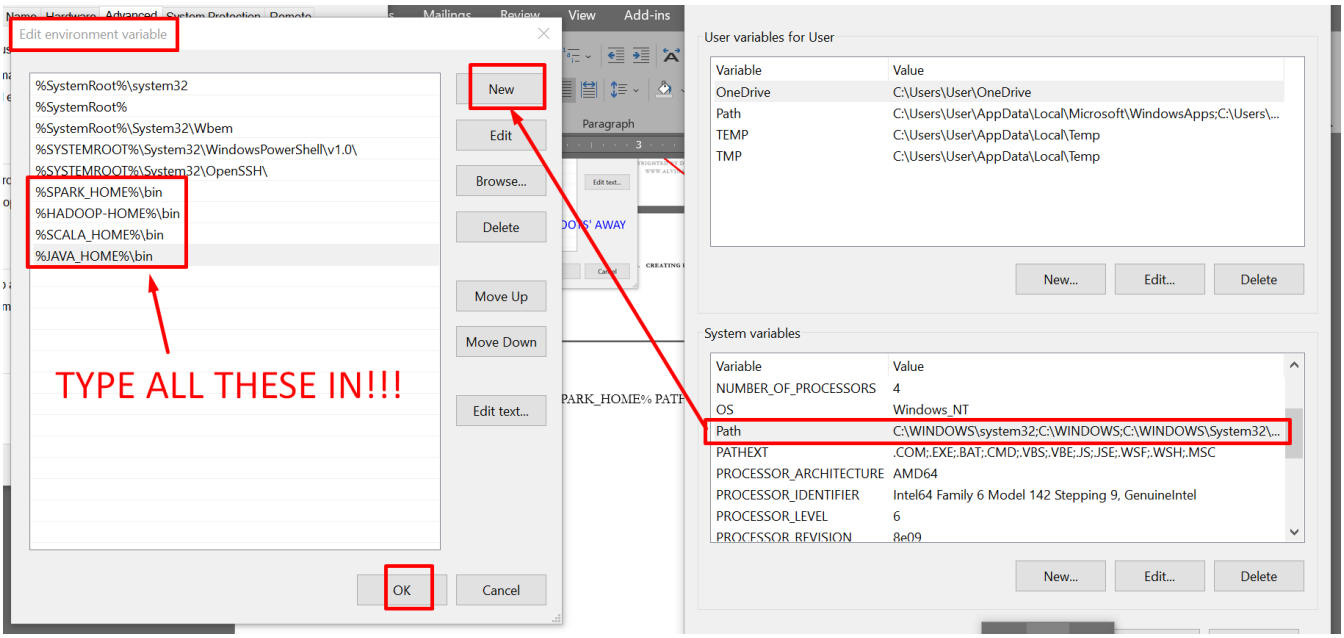
5. CREATE PYSPARK_PYTHON VARIABLE

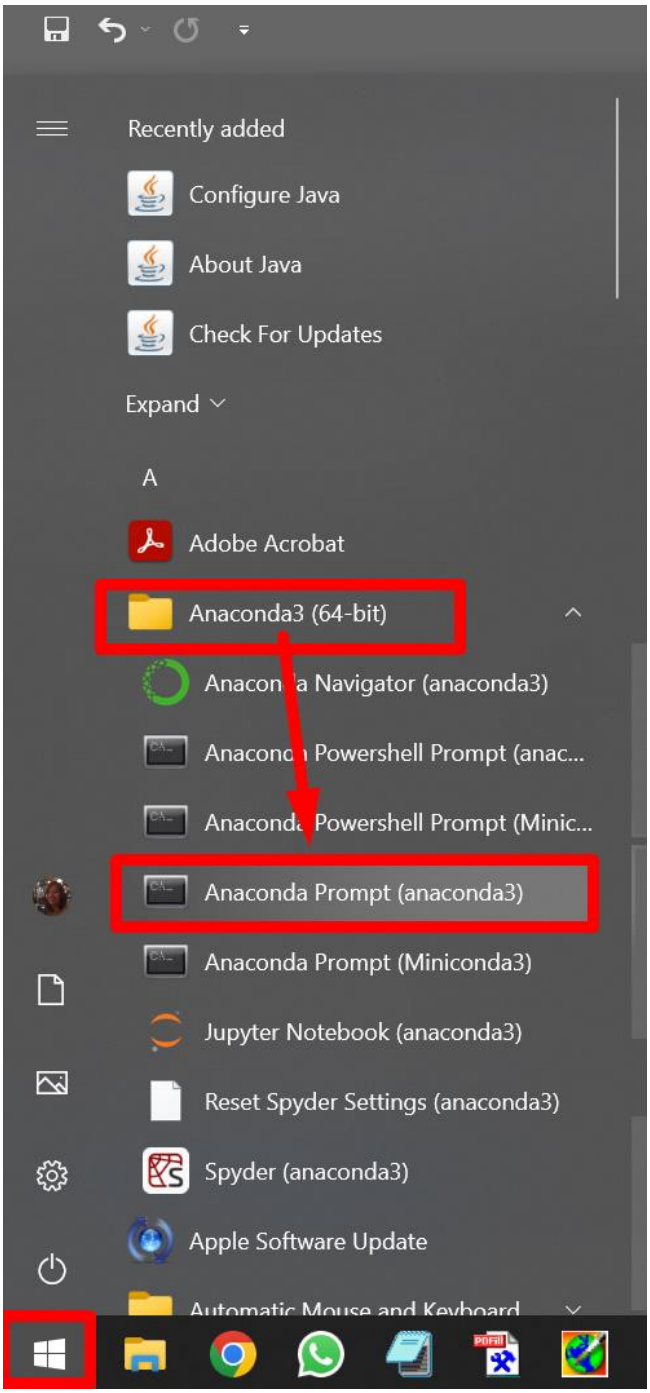


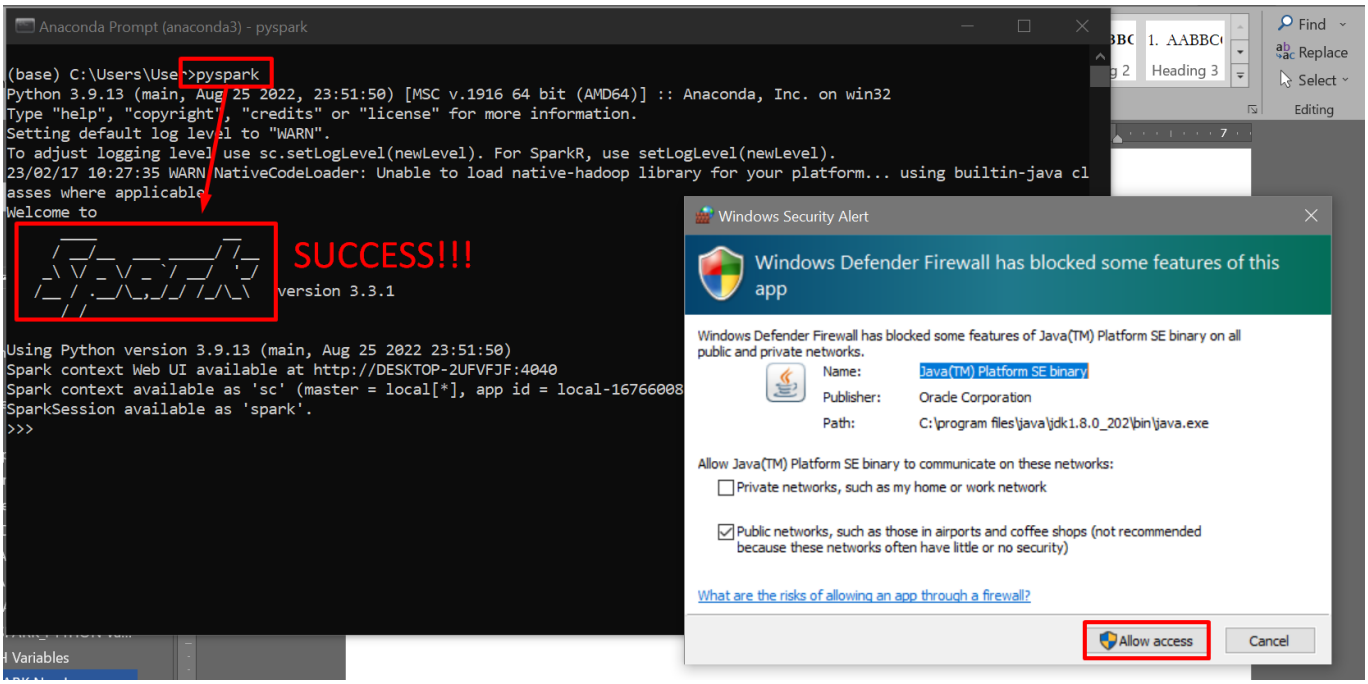


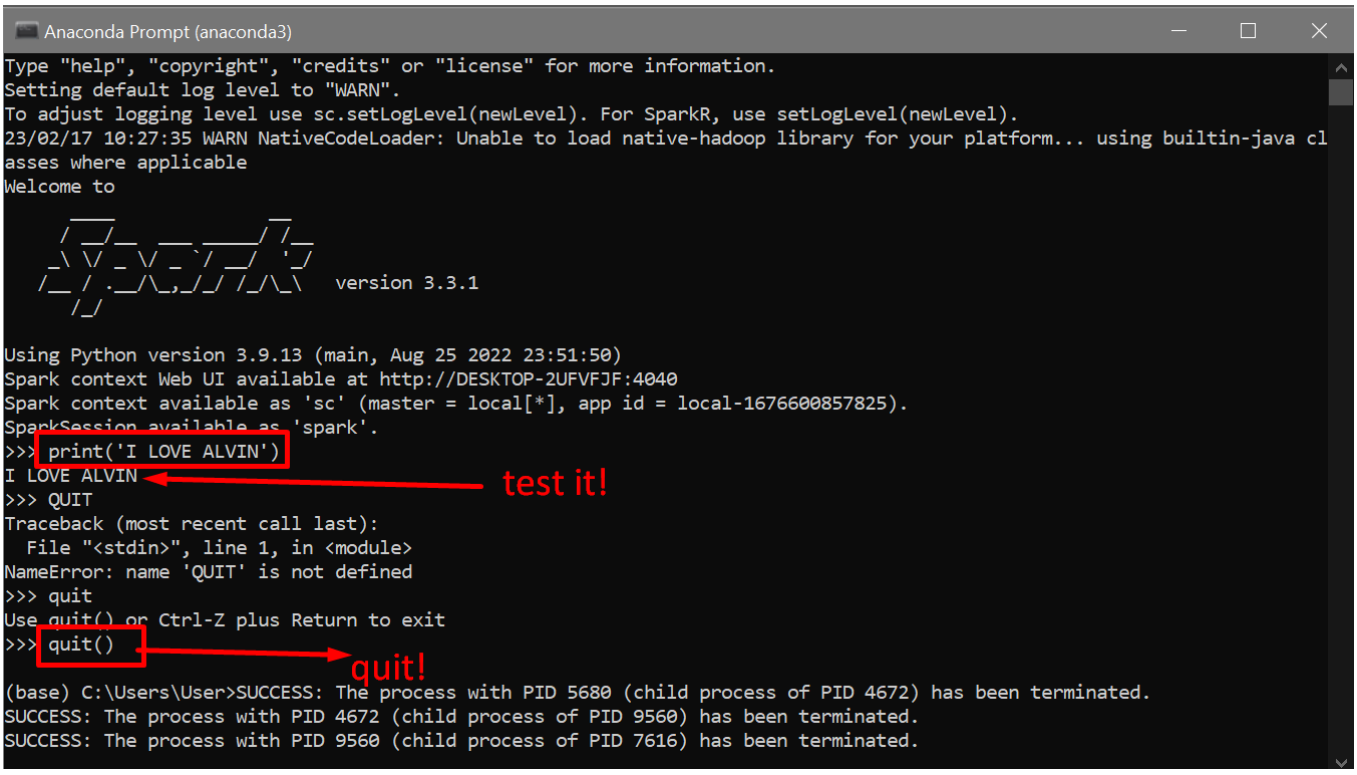
B. CREATING PATH VARIABLES

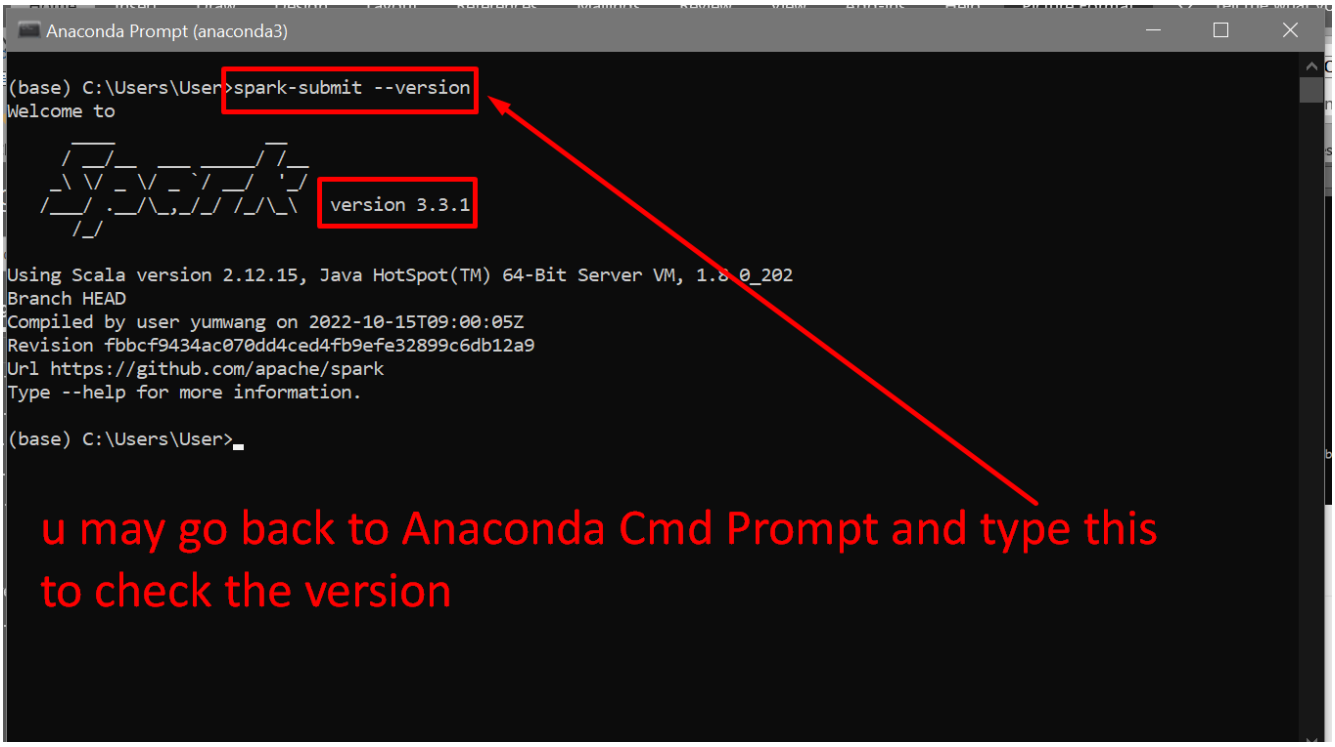


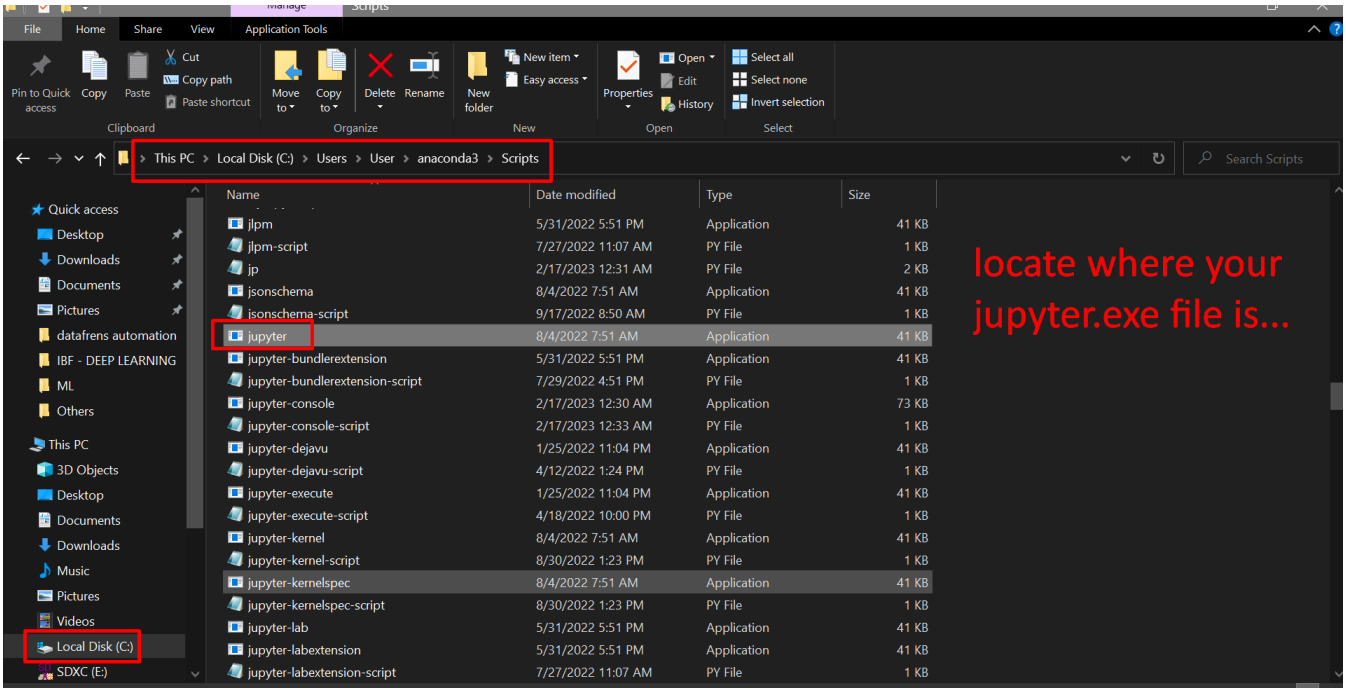


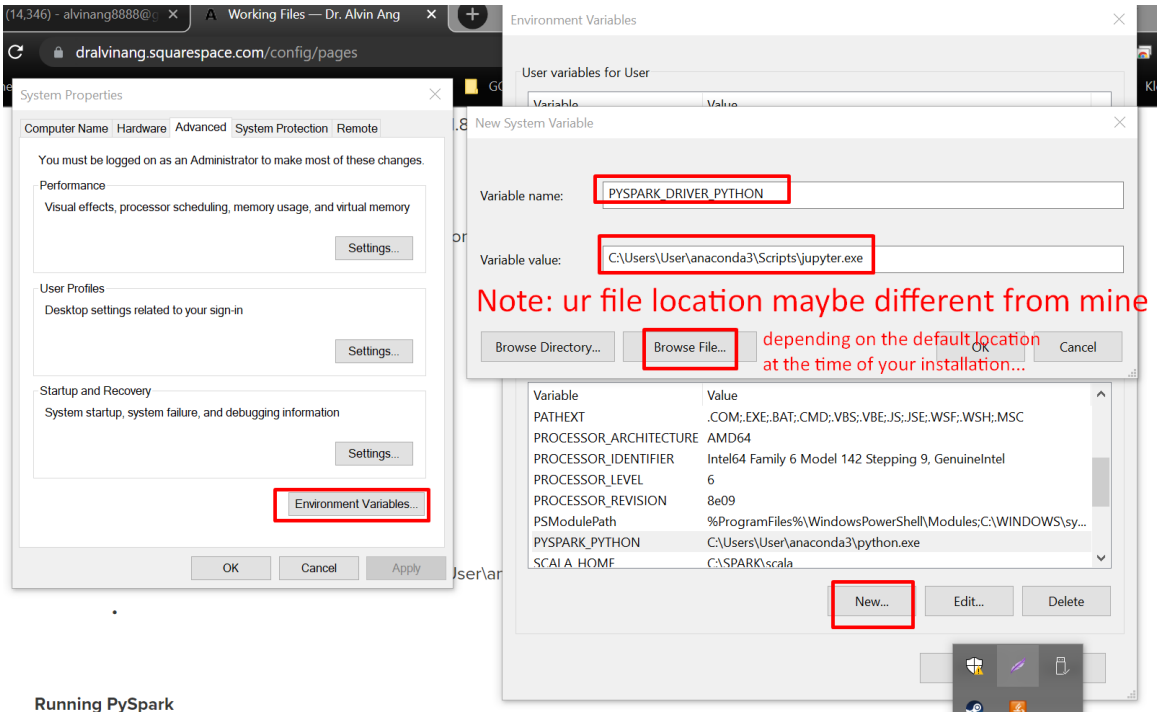




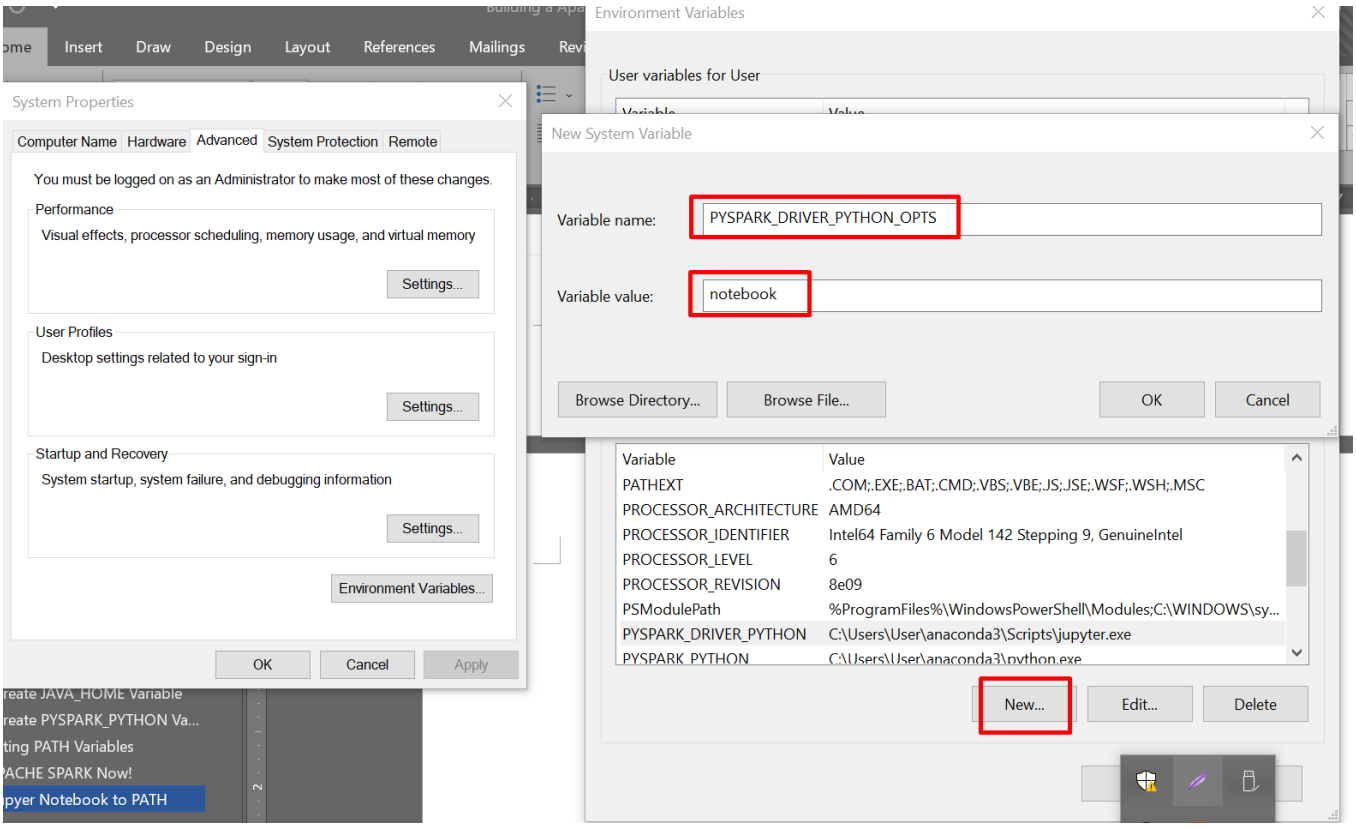






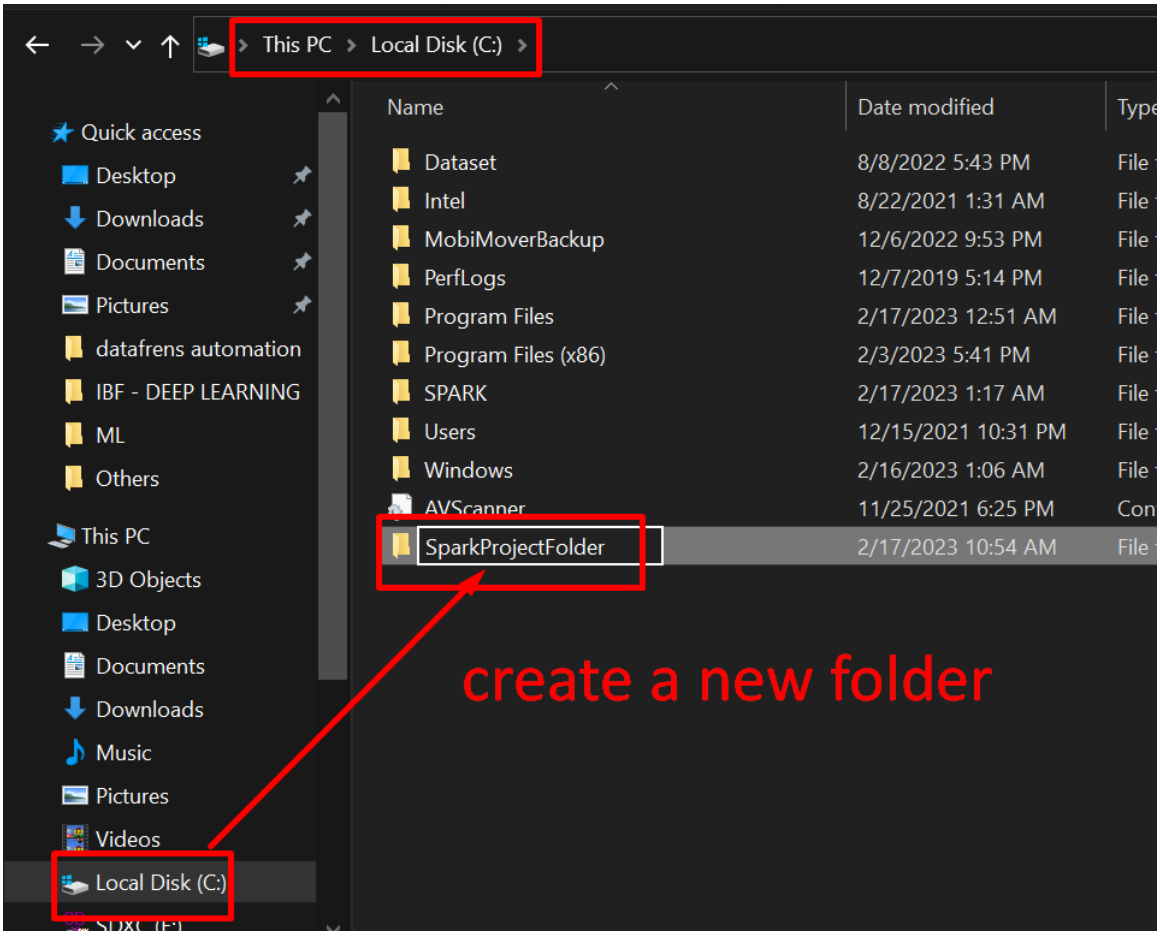


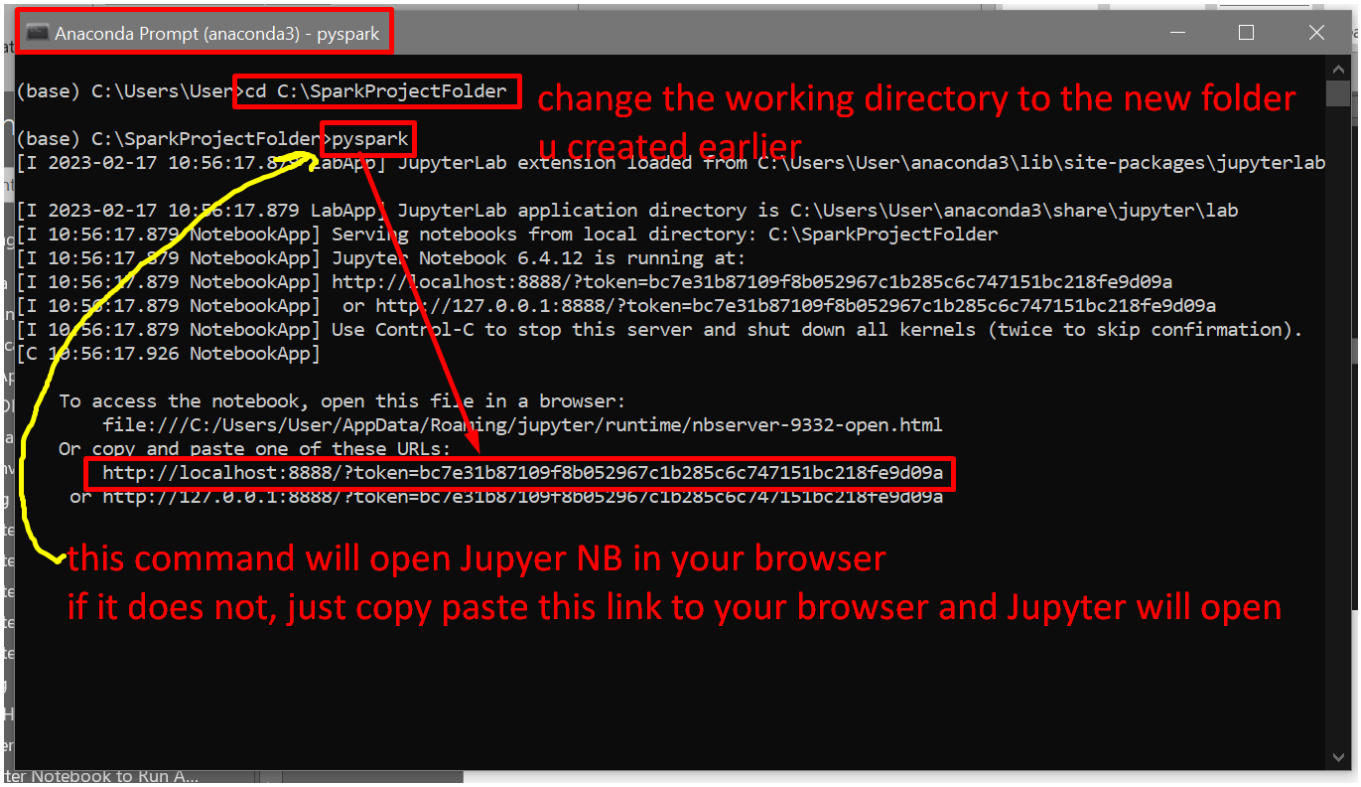
Running PySpark

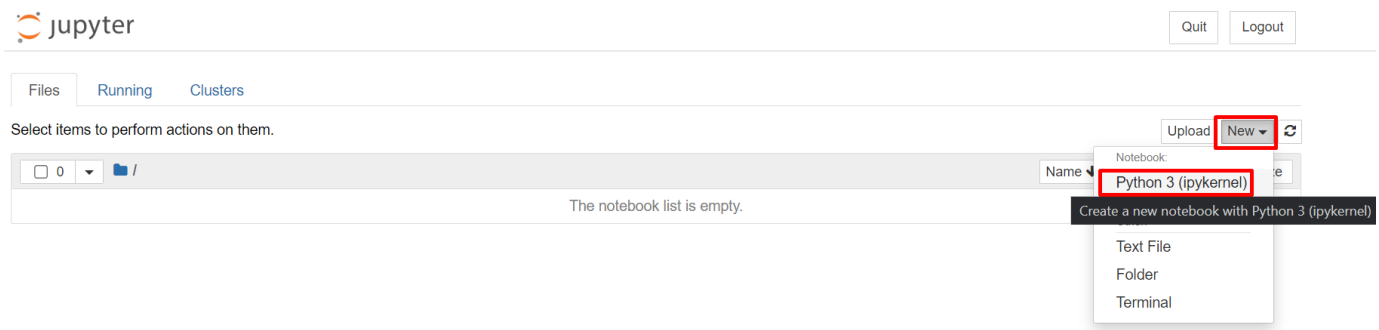


V. USING JUPYTER NOTEBOOK TO RUN APACHE SPARK

A. CREATE SPARK PROJECT FOLDER IN C:\

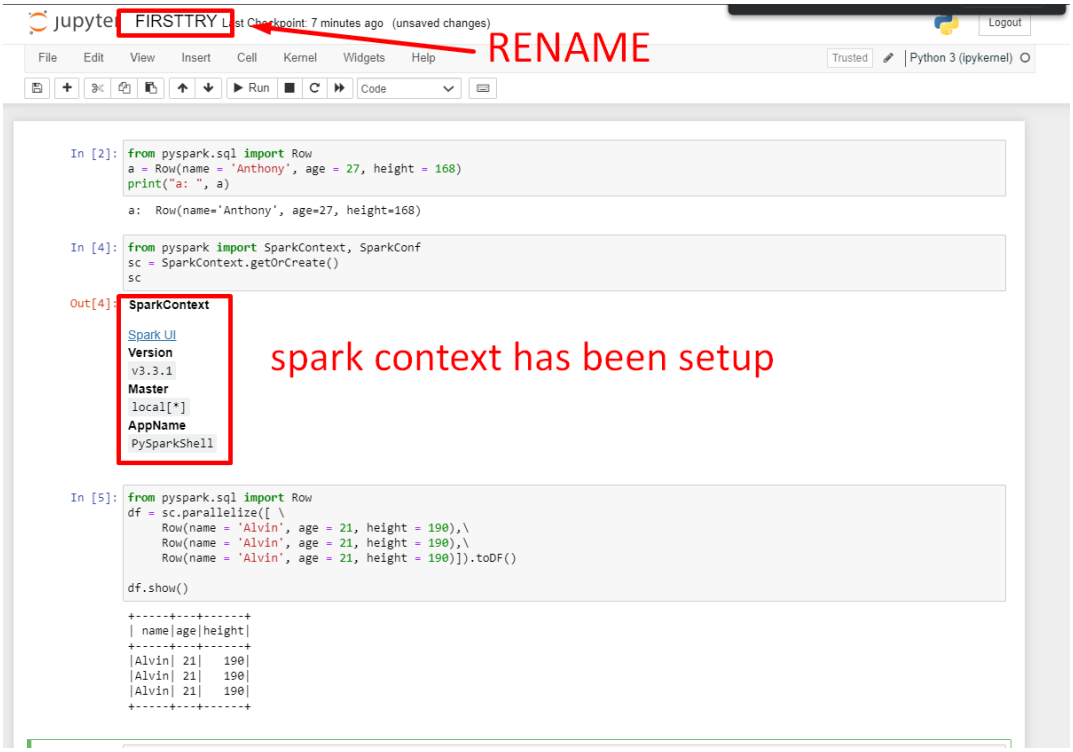






B. TEST CODE

```
from pyspark.sql import Row
a = Row(name = 'Anthony', age = 27, height = 168)
print("a: ", a)

#-------------------------------------------------------------------------------#

from pyspark import SparkContext, SparkConf
sc = SparkContext.getOrCreate()
sc

#-------------------------------------------------------------------------------#

from pyspark.sql import Row
df = sc.parallelize([ \
    Row(name = 'Alvin', age = 21, height = 190),\
    Row(name = 'Alvin', age = 21, height = 190),\
    Row(name = 'Alvin', age = 21, height = 190)]).toDF()

df.show()
```



VI. SETUP MASTER COMPUTER

A. STEP 1

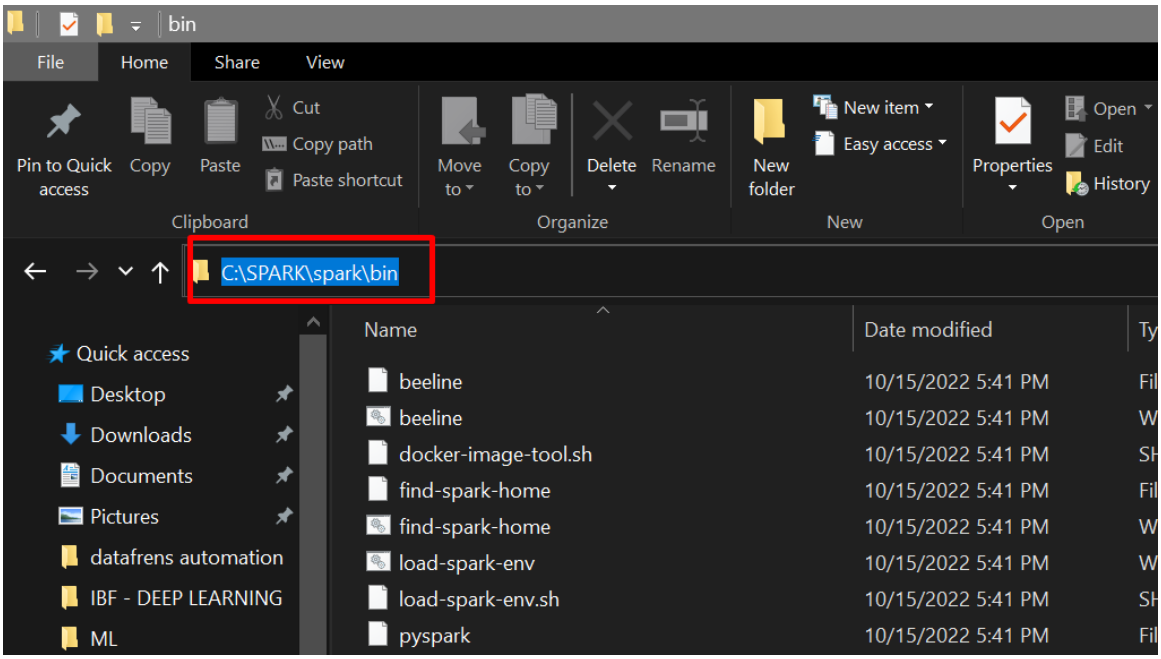
- Ensure that ALL Laptops have APACHE SPARK (and all Environment Variables) installed properly.

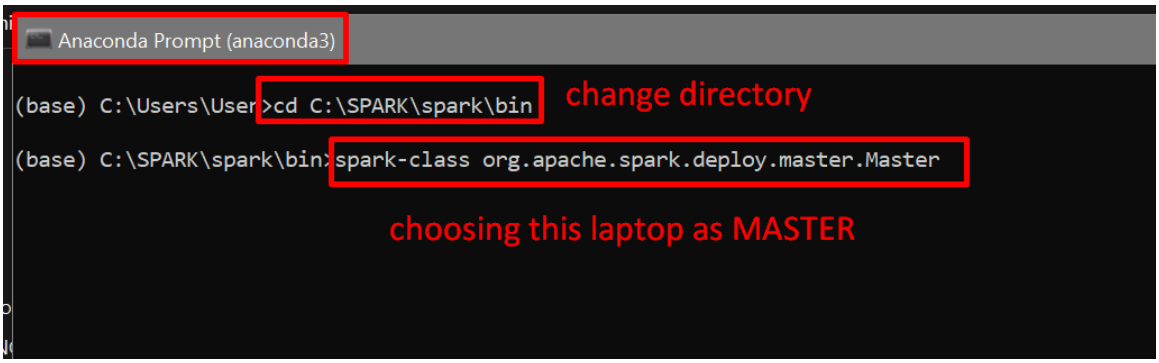
- In other words, ENSURE that you have followed through all previous steps in previous sections… including installing Java / Jupyter NB / Anaconda etc…. for ALL laptops.

B. STEP 2

- Decide which laptops you want to be MASTER… which ones you want to be SLAVE

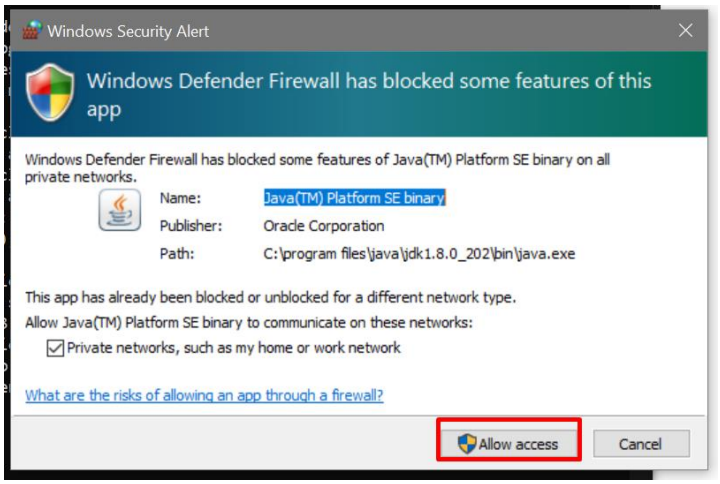
C. STEP 3: APPOINT YOUR MASTER LAPTOP

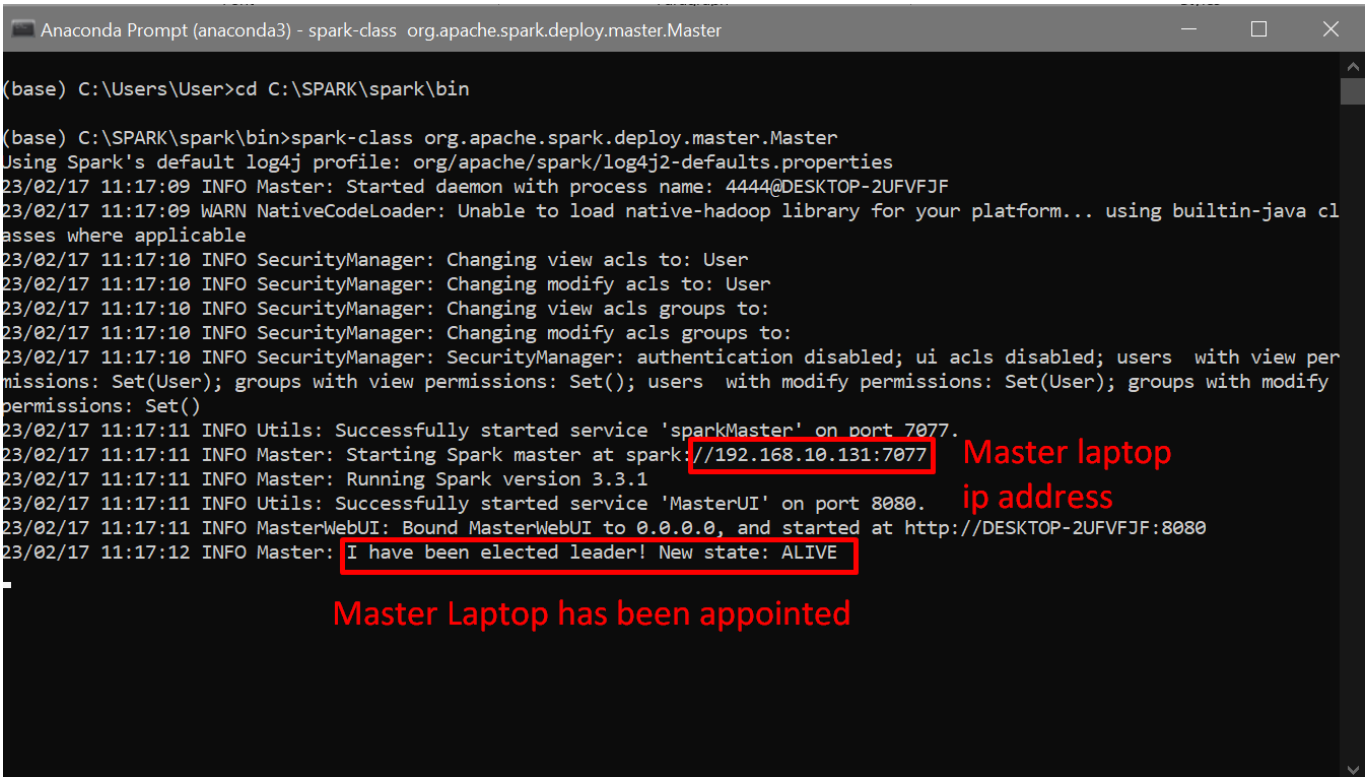


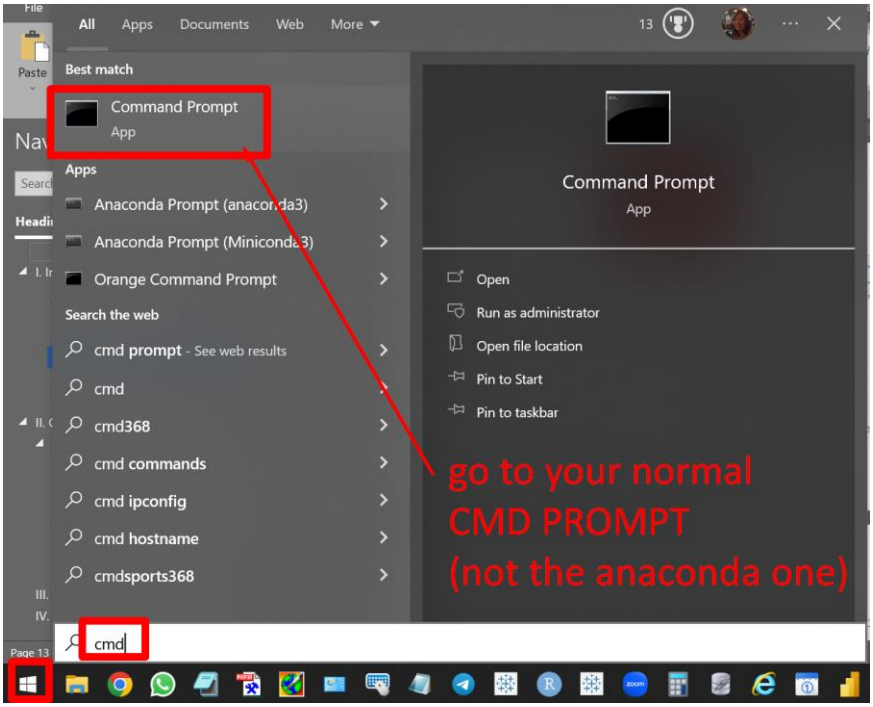


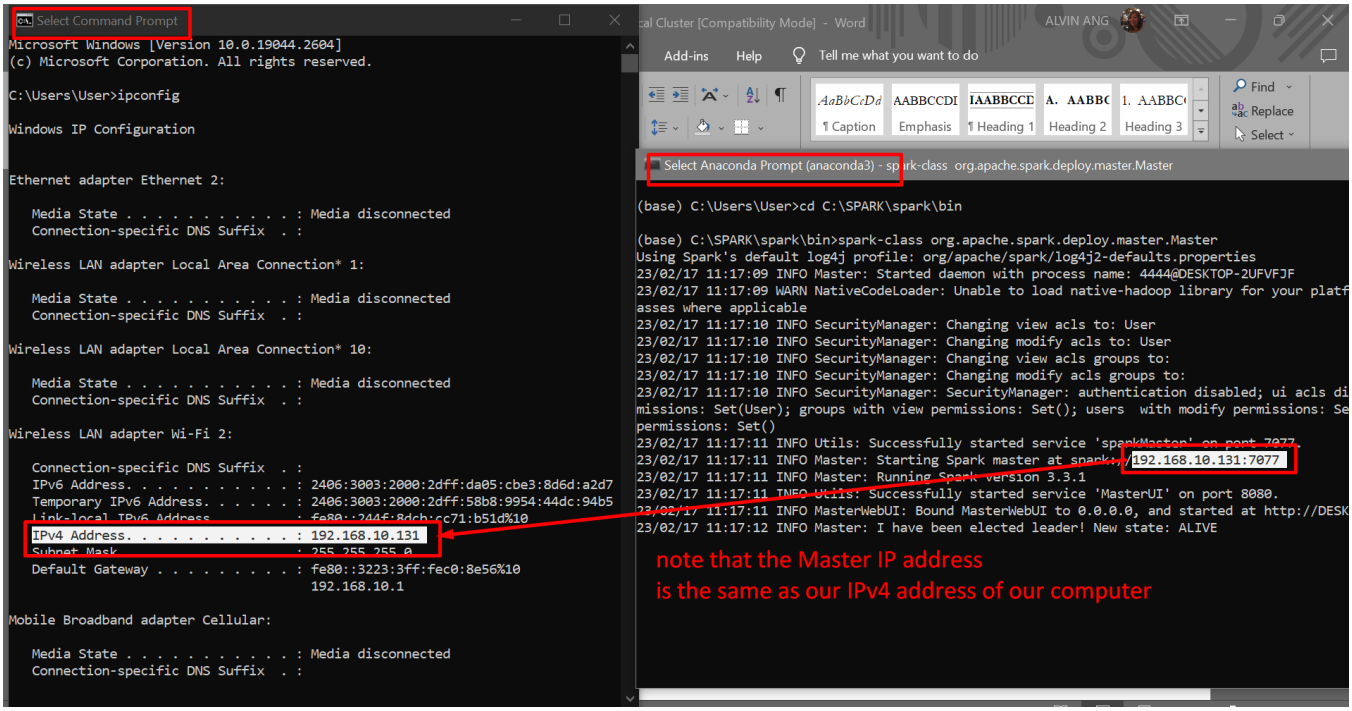
1. CODE

```
spark-class org.apache.spark.deploy.master.Master
```

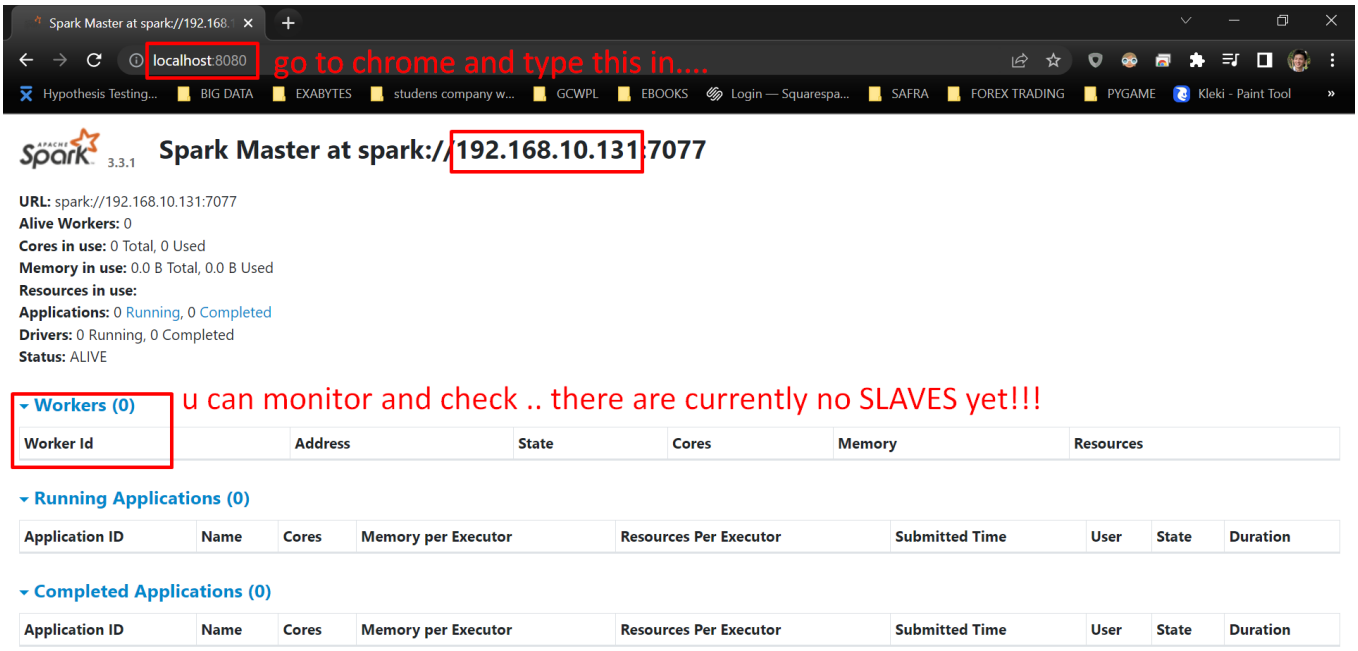




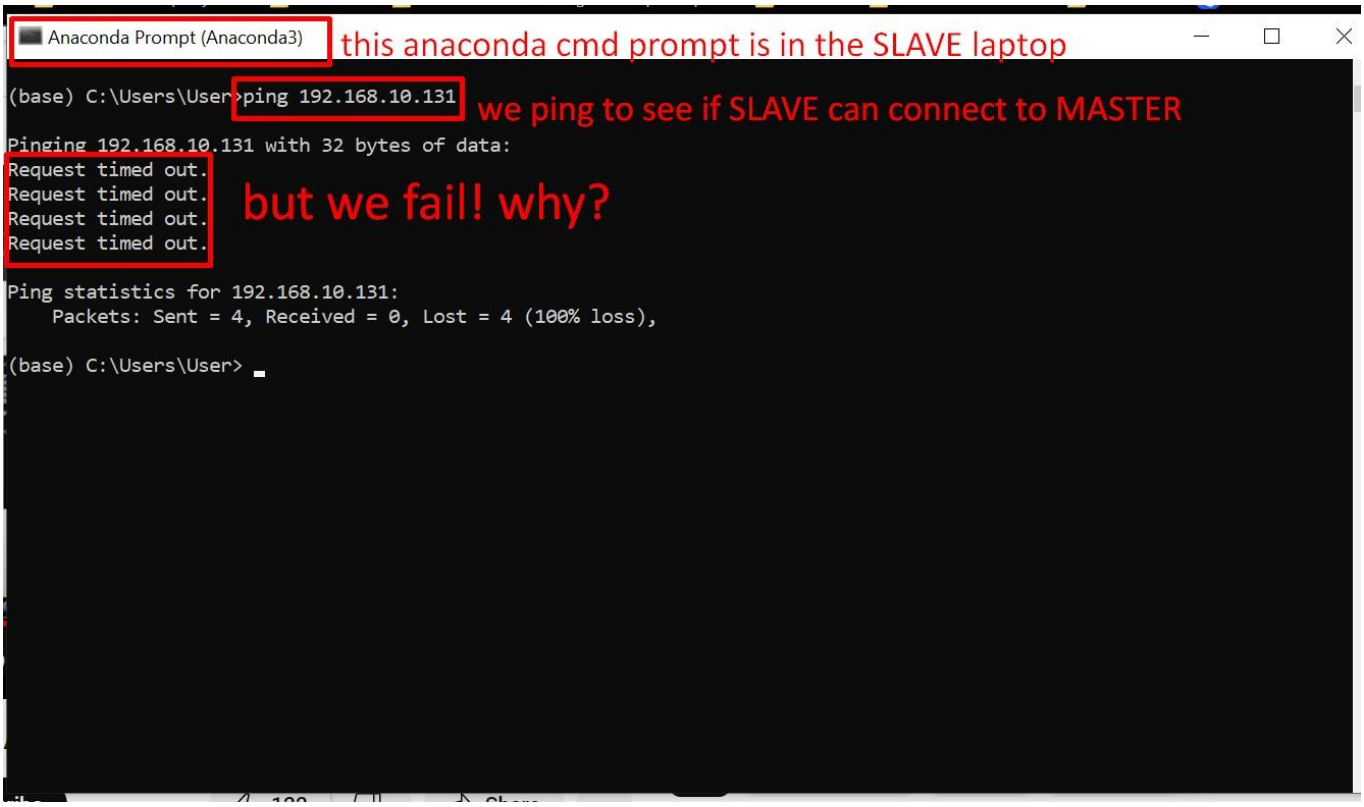




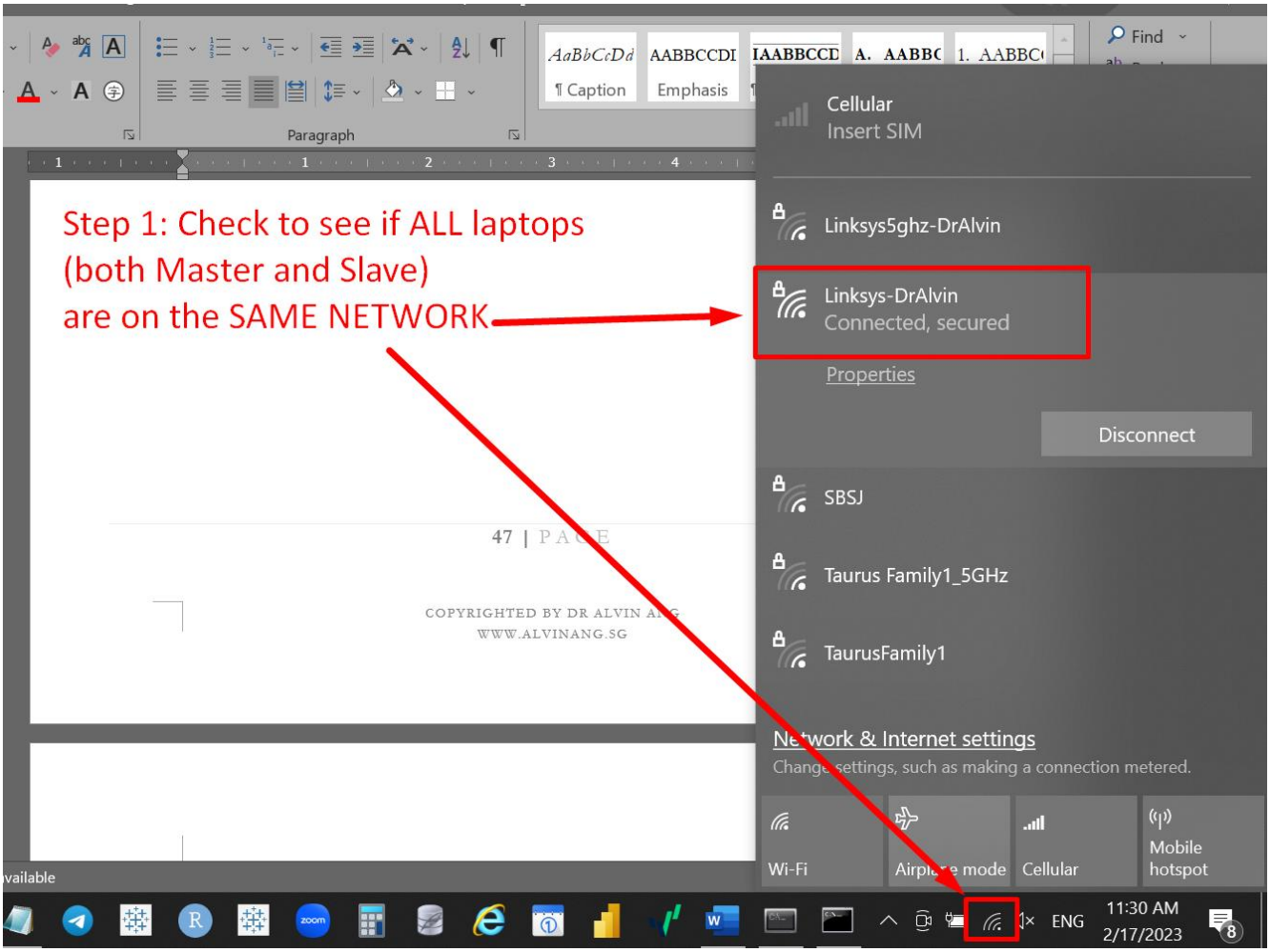
D. CHECK TO SEE IF THERE ARE ANY SLAVES....



Now…go to your SLAVE computer….

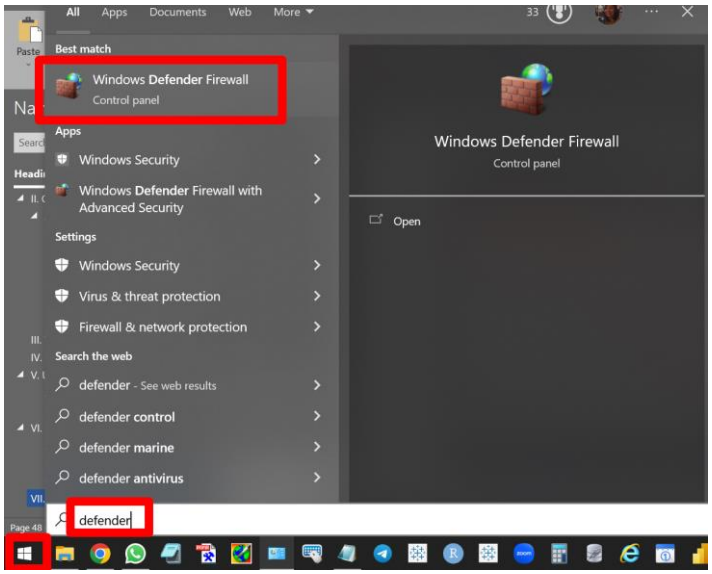


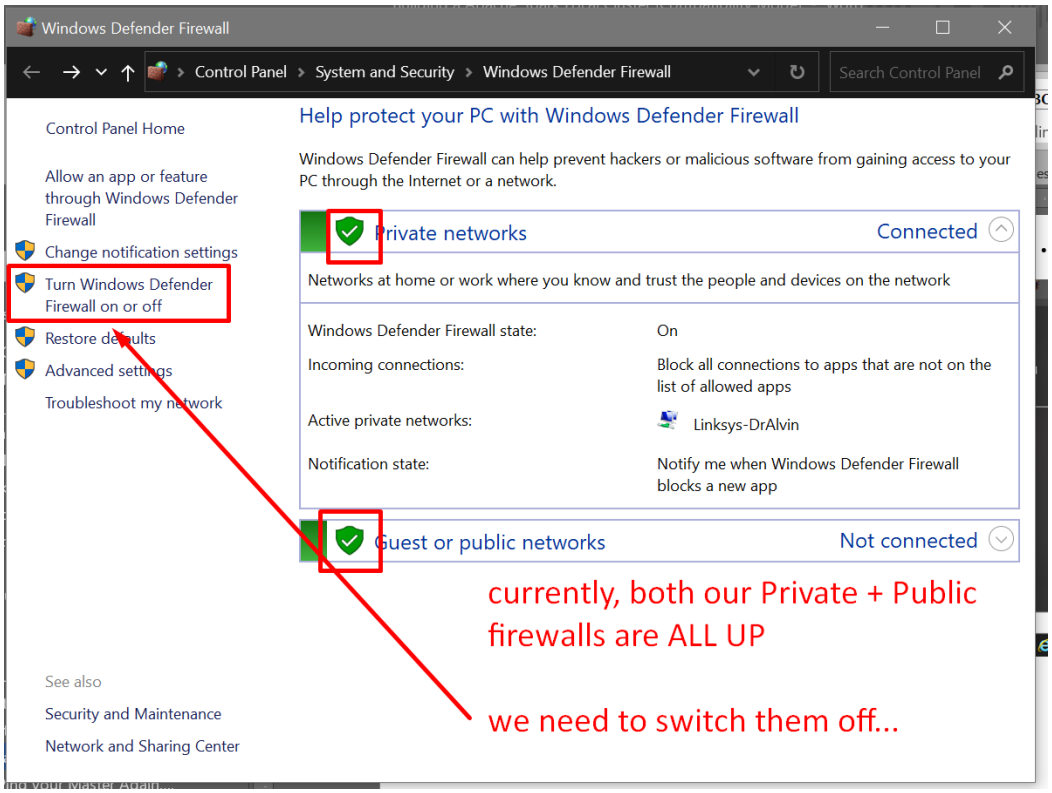
A. CHECK YOUR WIFI NETWORK (BOTH MASTER AND SLAVE)

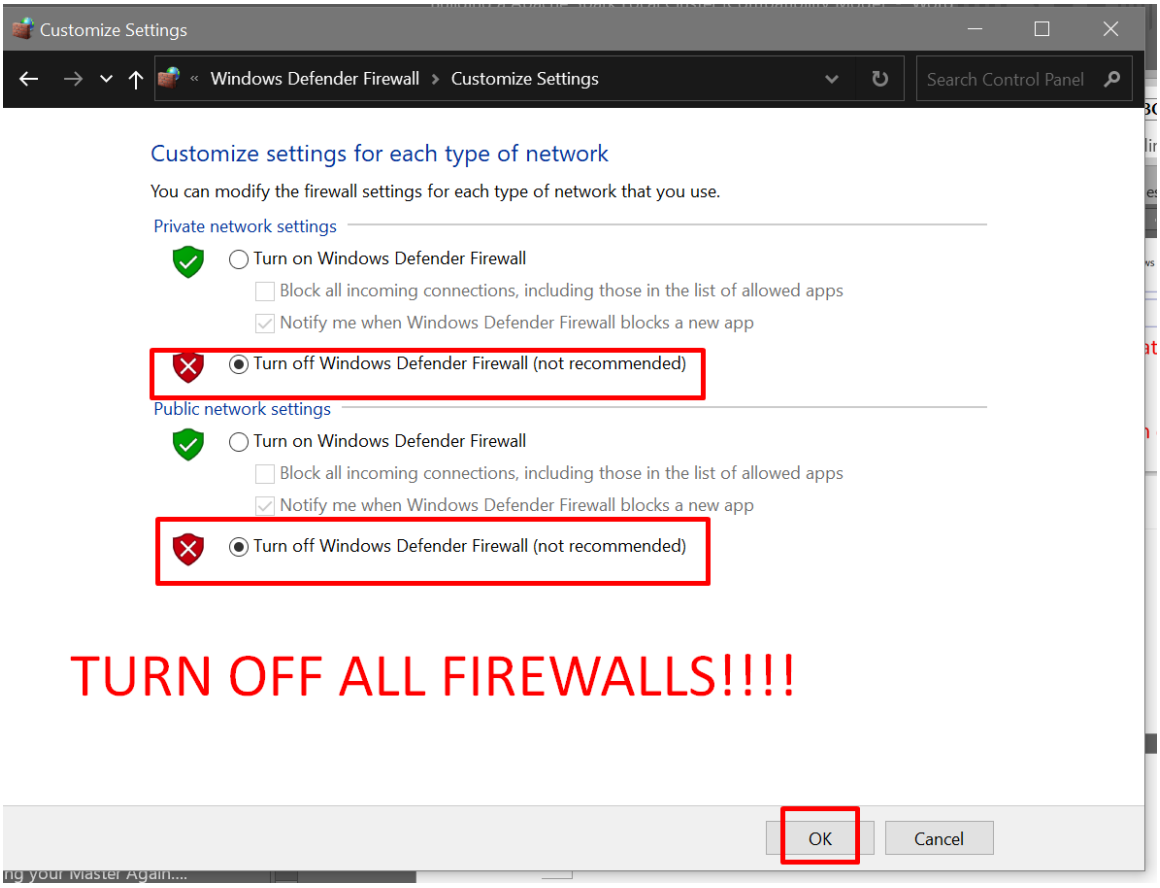


B. TURN OFF ALL YOUR FIREWALLS (MASTER ONLY)

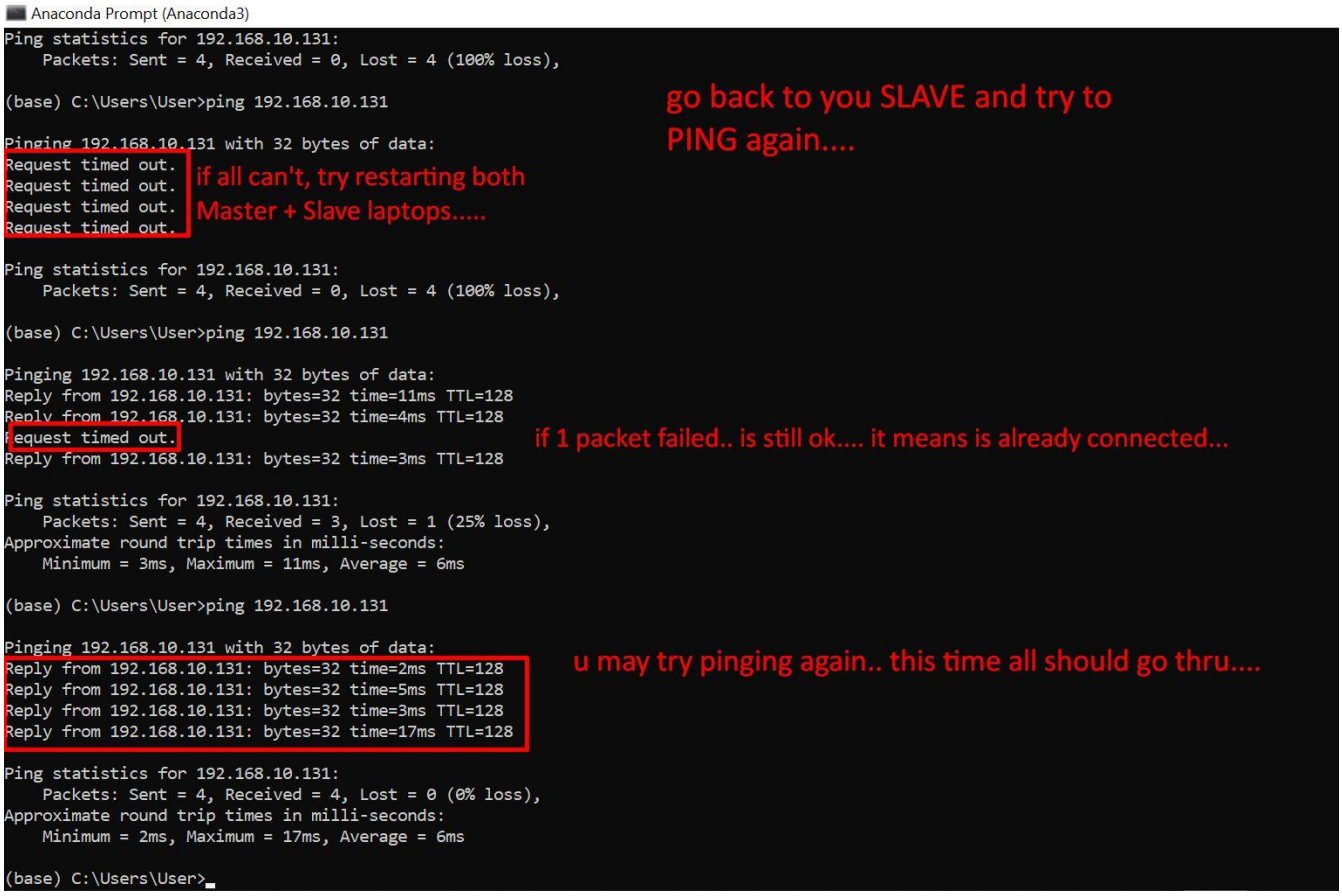
Now…go to your MASTER computer….



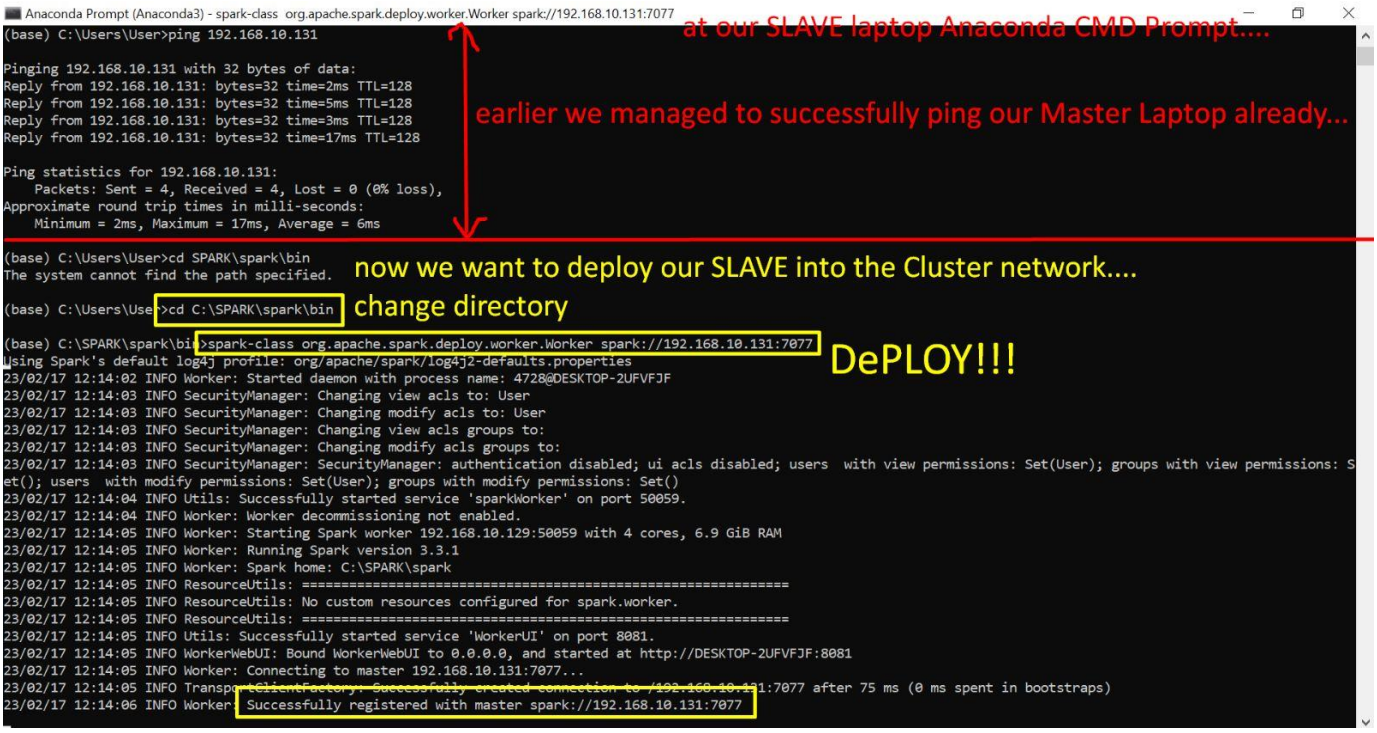




Now…go to your SLAVE computer….



Now…go to your SLAVE computer….



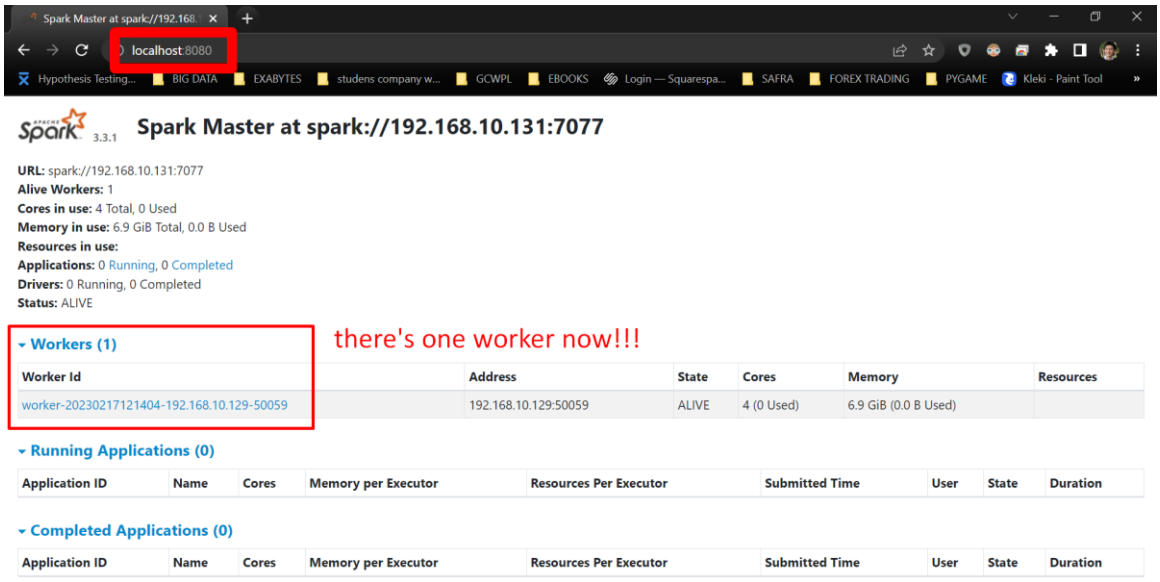
A. CODE

spark-class org.apache.spark.deploy.worker.Worker spark://192.168.10.131:7077

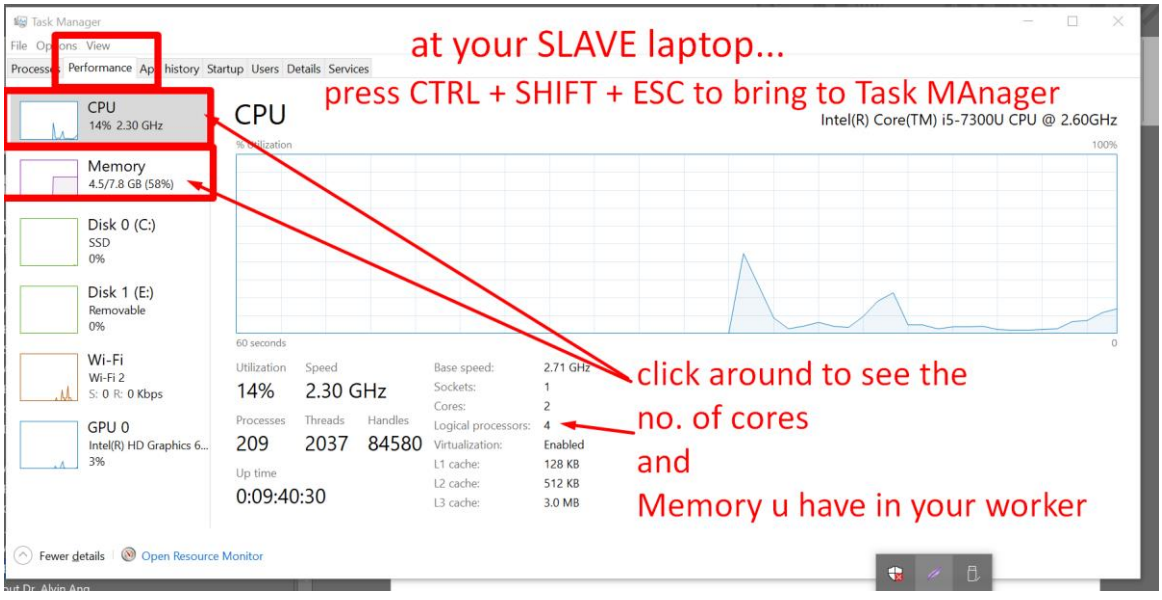
(your IP address will be different from mine…)

B. CHECK TO SEE IF THERE ARE ANY SLAVES….

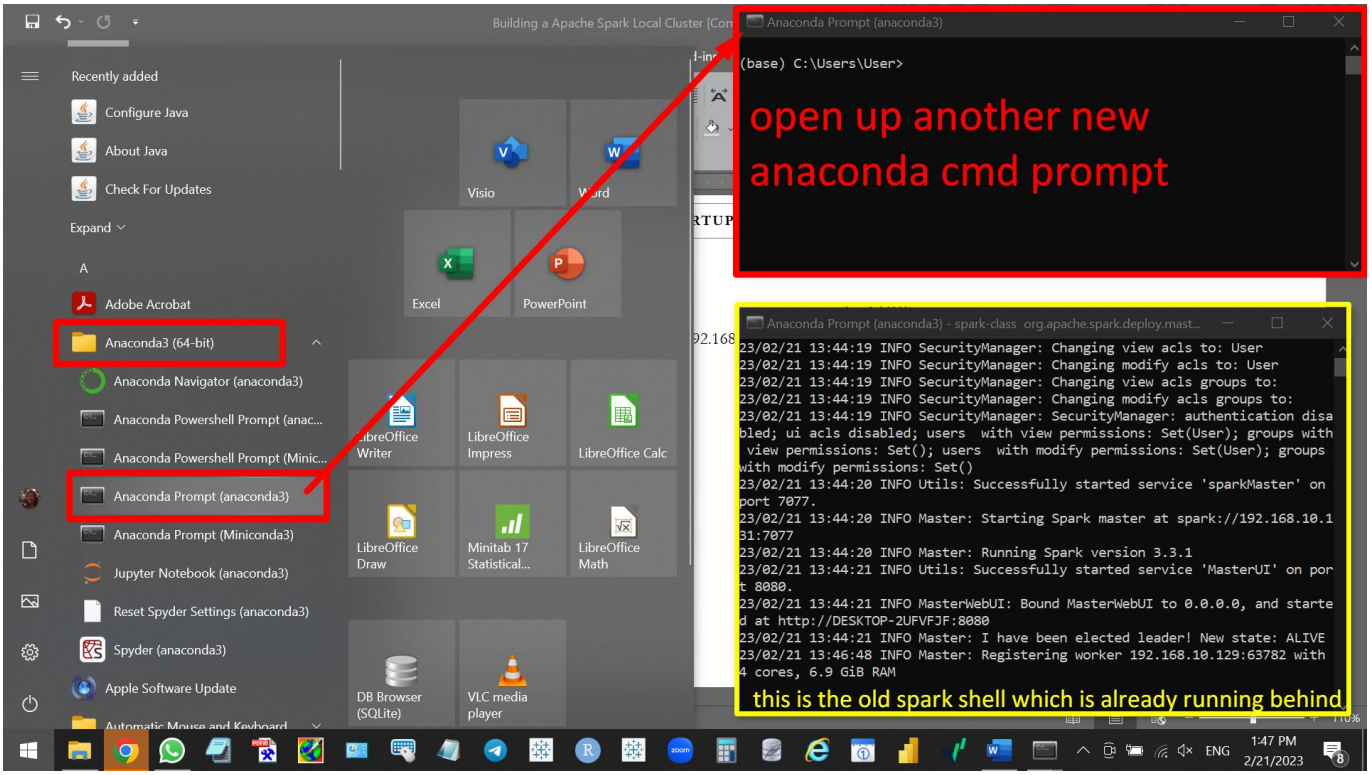
Now…go to your MASTER computer….

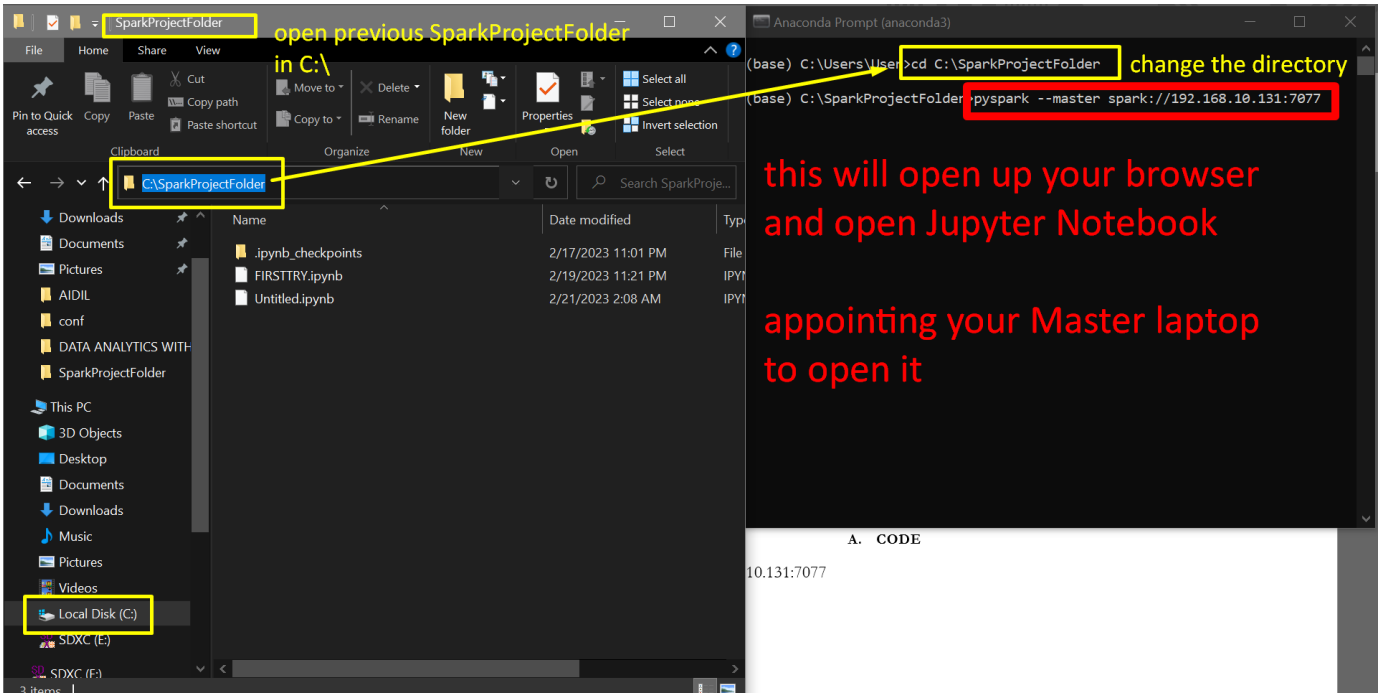


Now…go to your SLAVE computer….



Now…go to your MASTER computer….

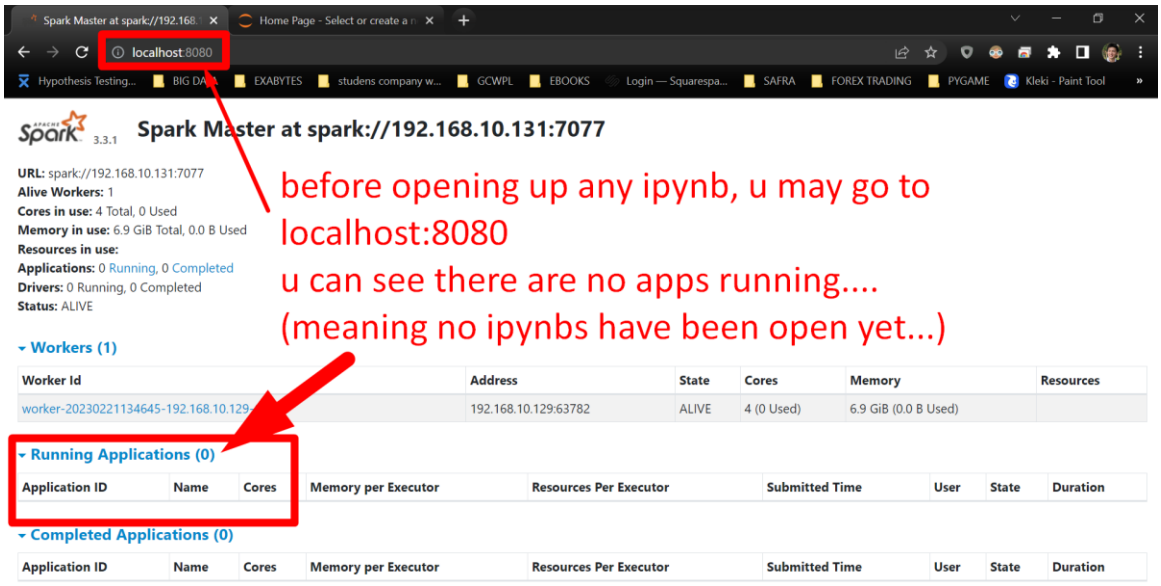




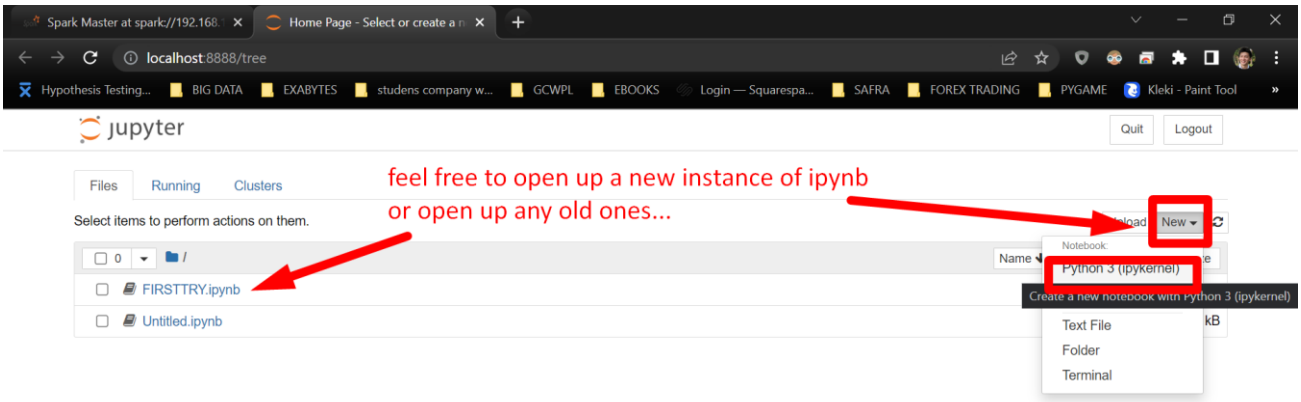
A. CODE

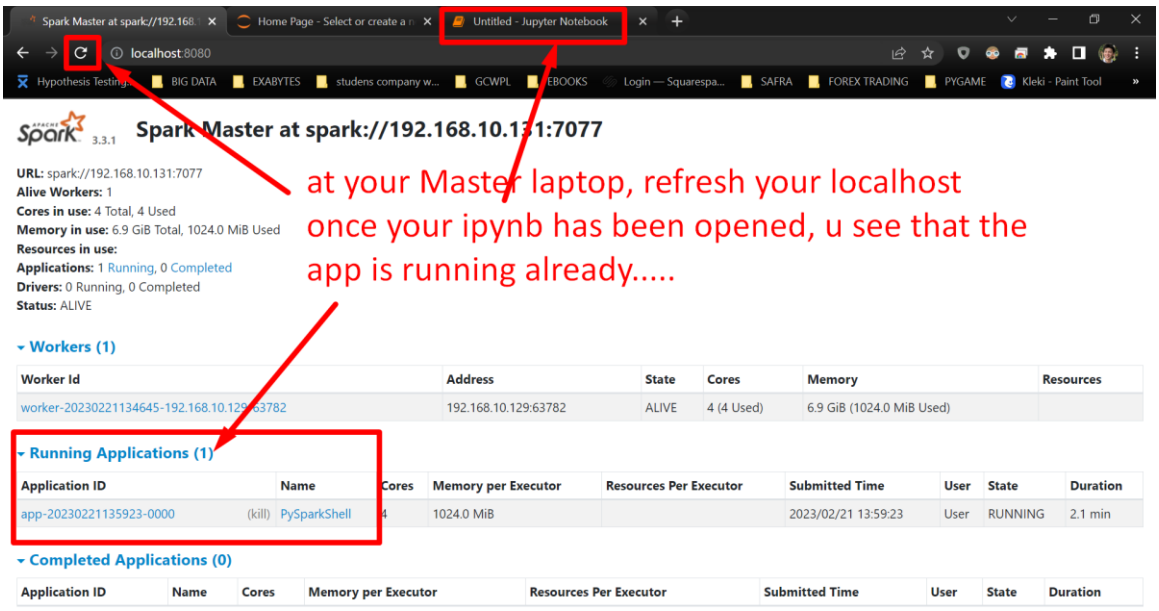
```
pyspark --master spark://192.168.10.131:7077
```

B. CHECK LOCALHOST

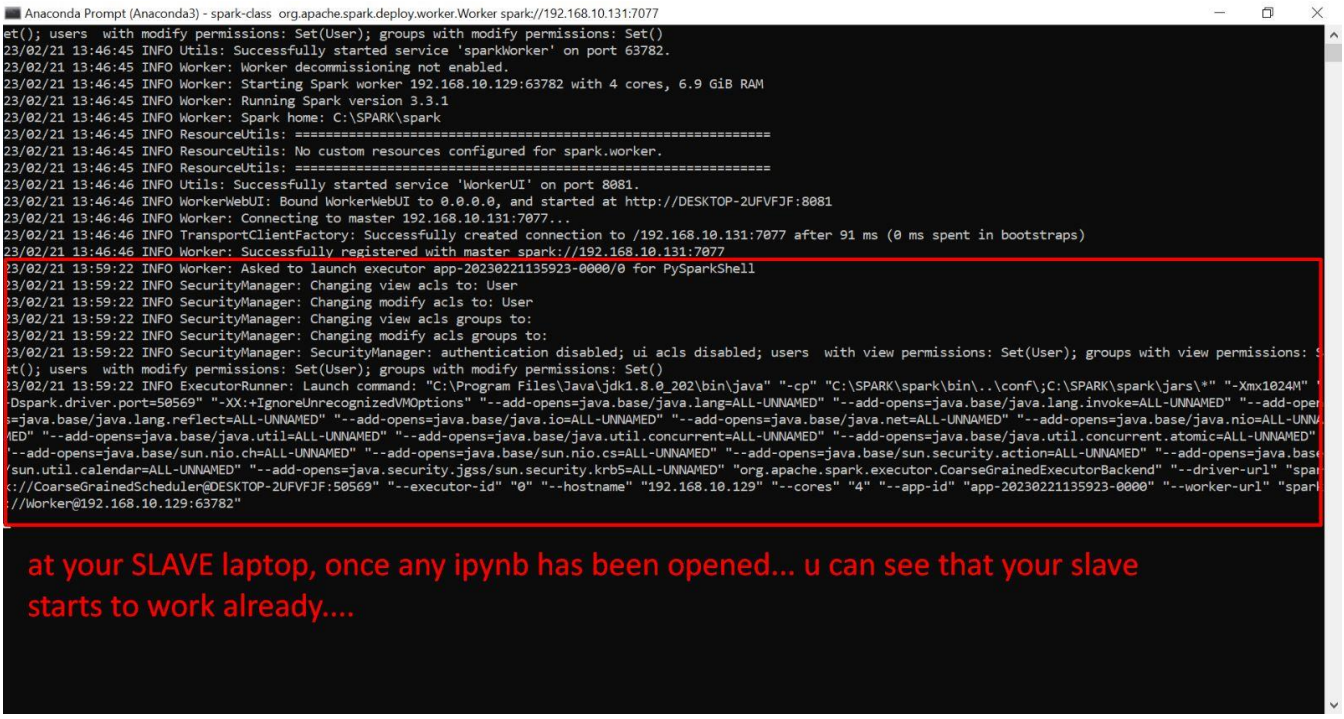


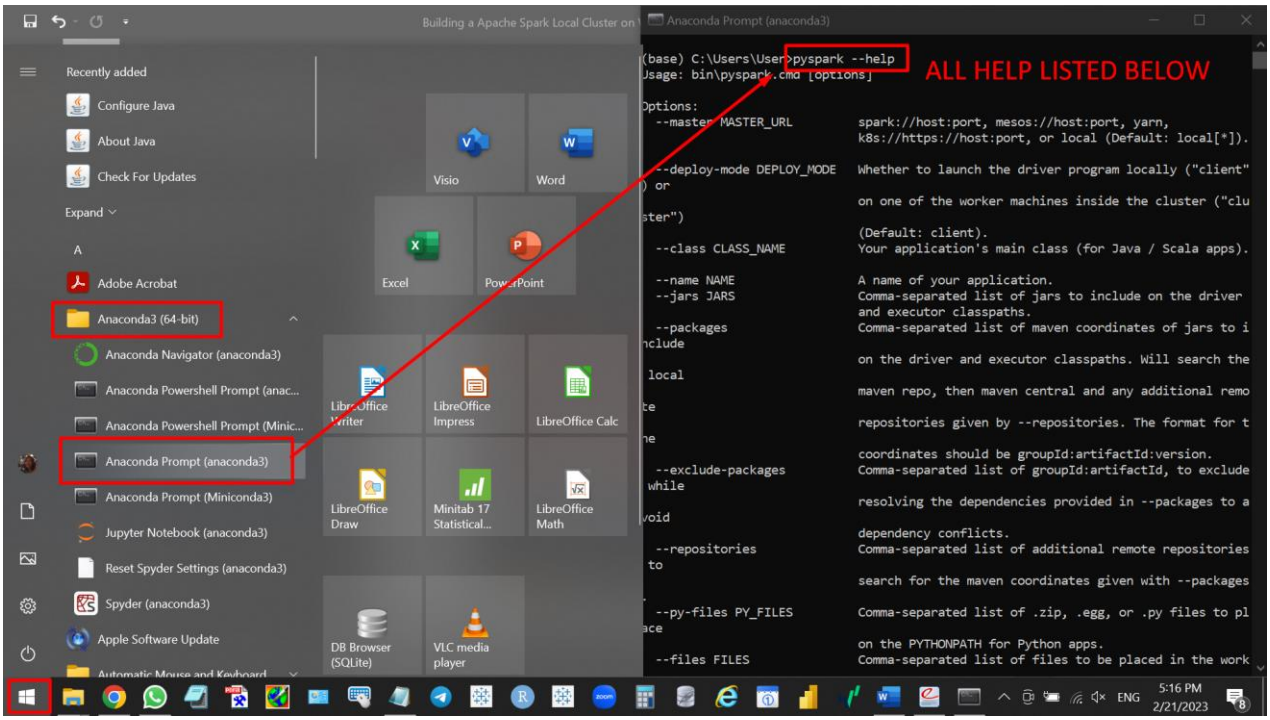
C. OPEN ANY IPYNB

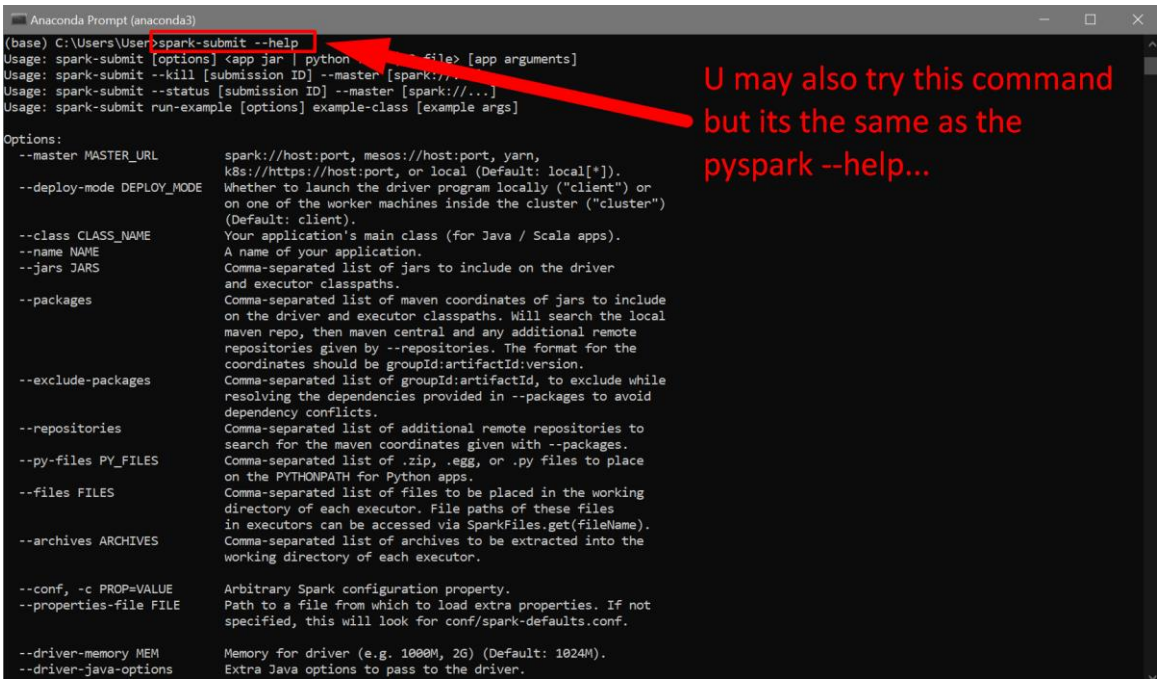




Now…go to your SLAVE computer….





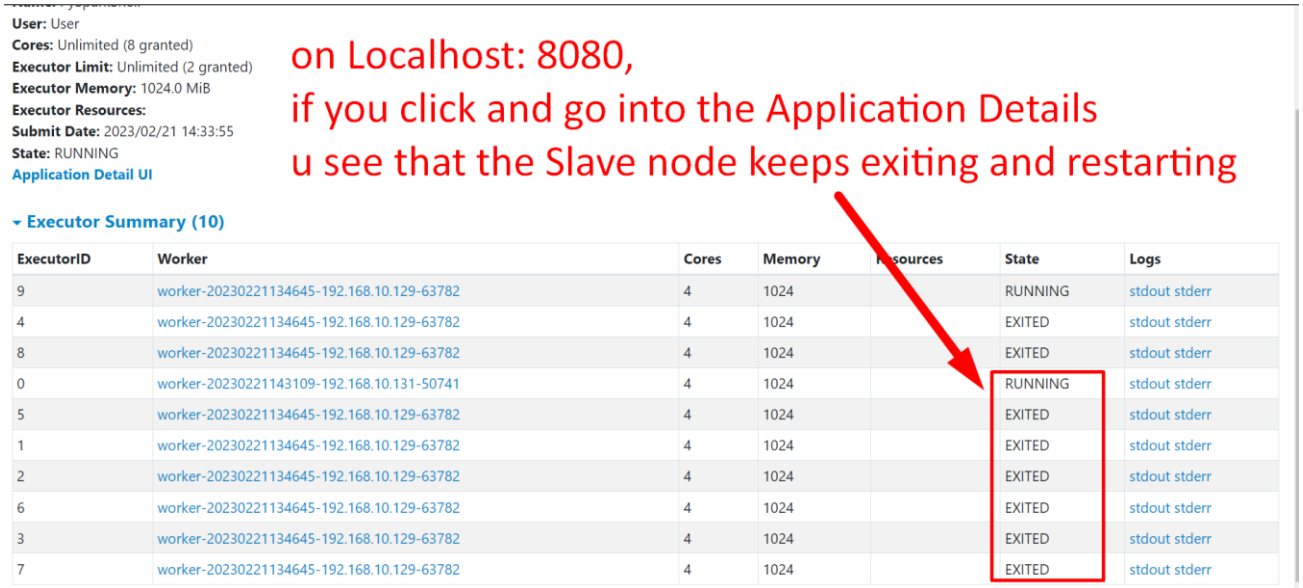


XI. DISCLAIMER

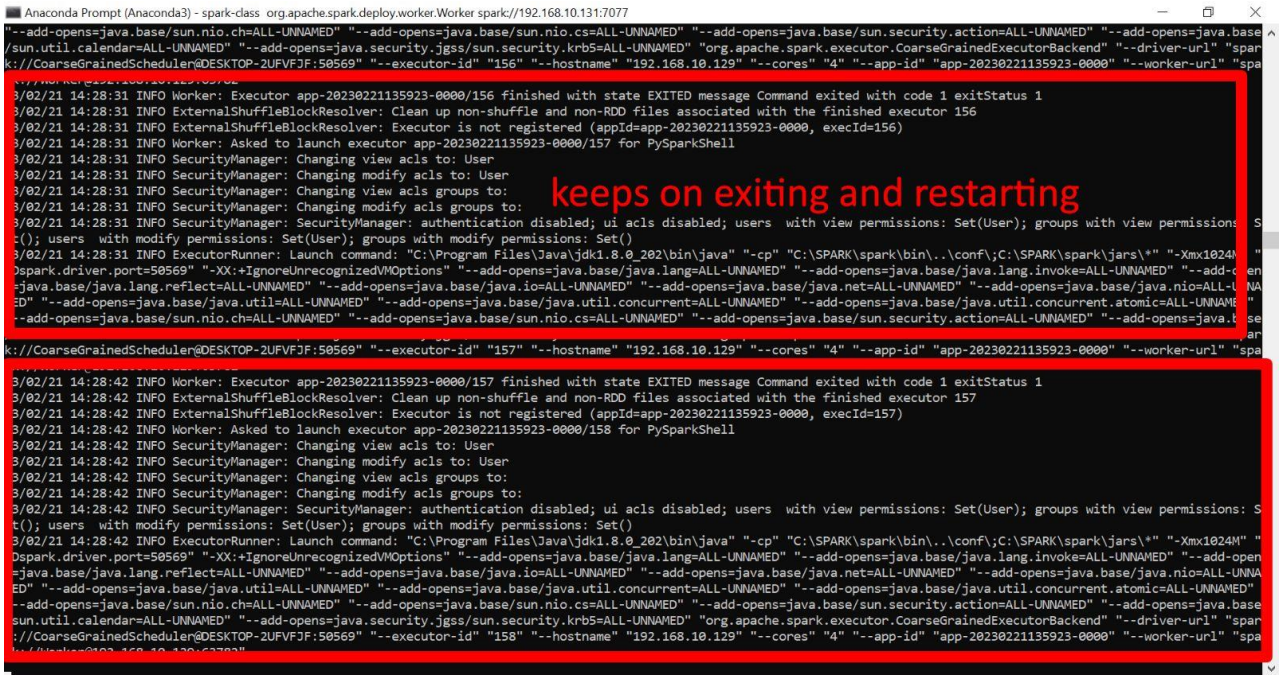
At the point of this writing, a key issue popped up while trying to do local Master / Slave configuration on Windows.

A. ISSUE: SLAVE APP KEEPS ON EXITING AND RESTARTING

At your Master laptop….



At your Slave laptop Cmd Prompt…



B. POSSIBLE SOLUTIONS

1. I can't find any solution to the above problem (even after a long time searching on the internet and trying and failing).

2. I have never tested live on real big data before …meaning, even the constant exiting and restarting could be due to "no real jobs running"….

3. You could test it on big data if you have….and maybe the problem will go away…

4. Spark was built mainly on Linux (Ubuntu)…. Thus, most codes on the internet use Ubuntu…

5. https://medium.com/codex/setup-a-spark-cluster-step-by-step-in-10-minutes-922c06f8e2b1

6. You might want to try the link above to install only on Ubuntu systems… which may work better…(else Windows systems don't seem to work)

7. But whatever installation we have learnt so far on Windows is the exact steps as Ubuntu OS (like creating Environment Variables).

8. However, you might need to have MULTIPLE Ubuntu laptops if you really want to crunch big data.. and install all of them with Spark….which is a lot of work…

9. A better solution will be to do Spark on the cloud…

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.