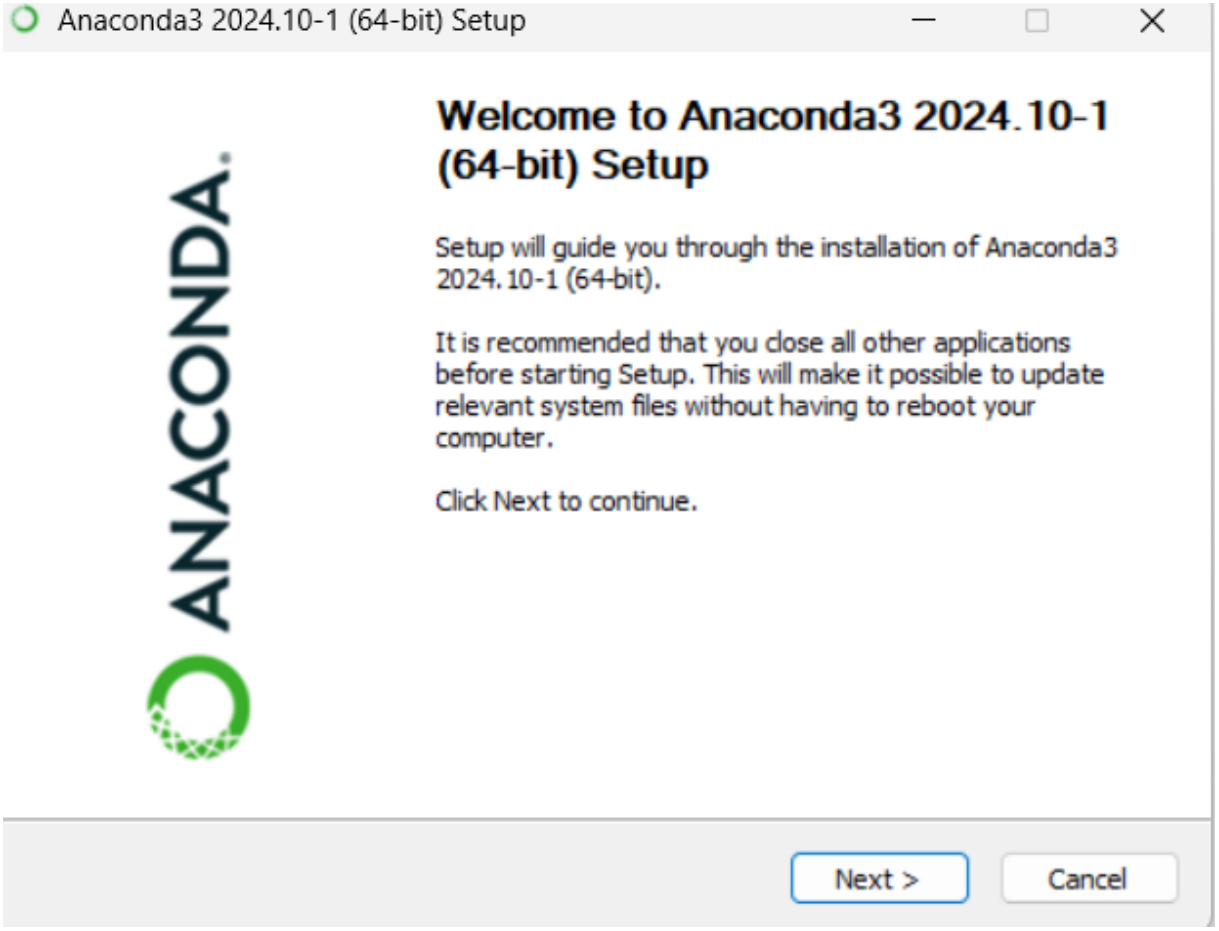
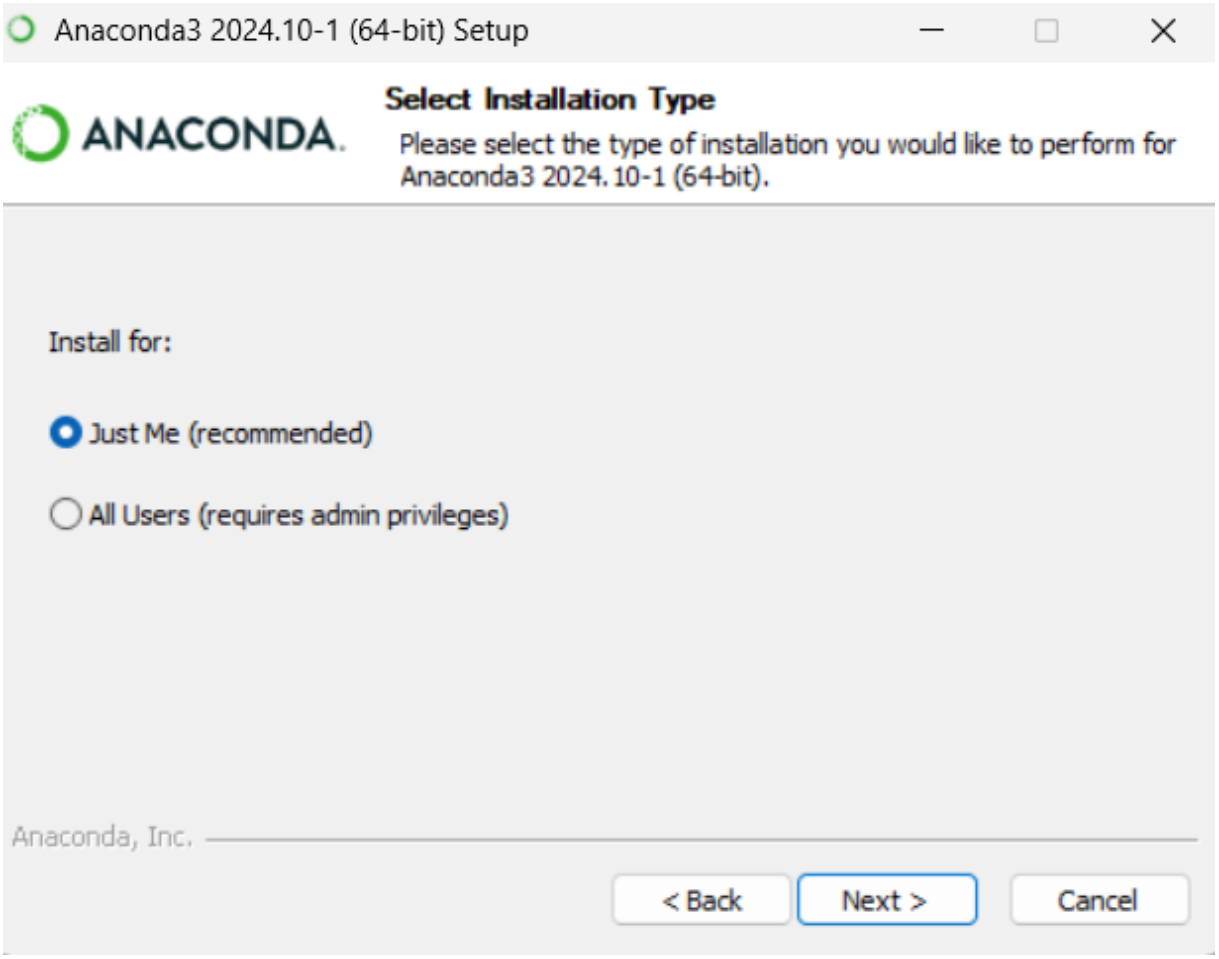
HƯỚNG DẪN TẢI SPARK TRÊN WINDOW

1. Tải IDM

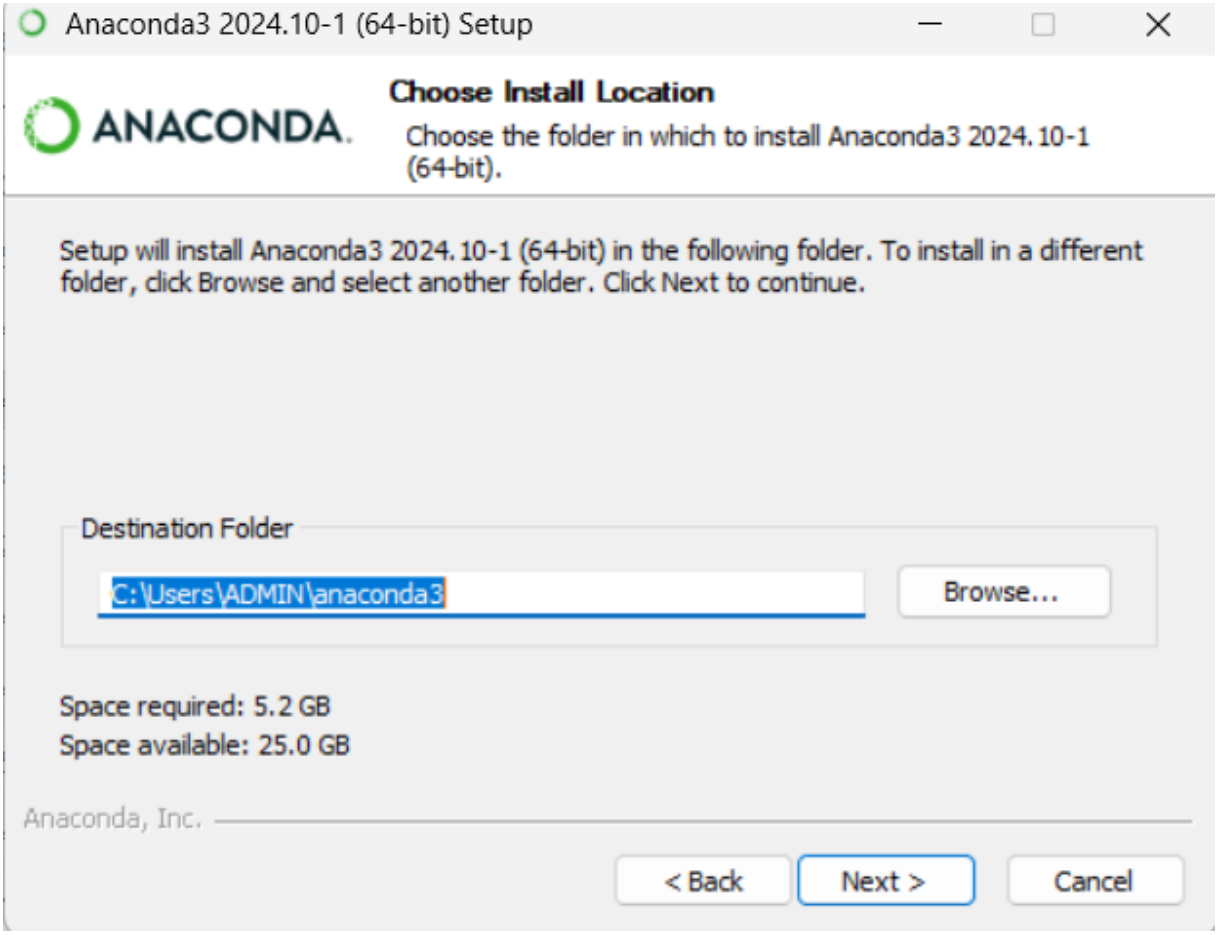
   https://khophanmem.vn/idm

2. Tải Pycharm Community
   https://www.jetbrains.com/pycharm/download
3. Tải anaconda
   Download Anaconda Distribution | Anaconda
   Chọn next

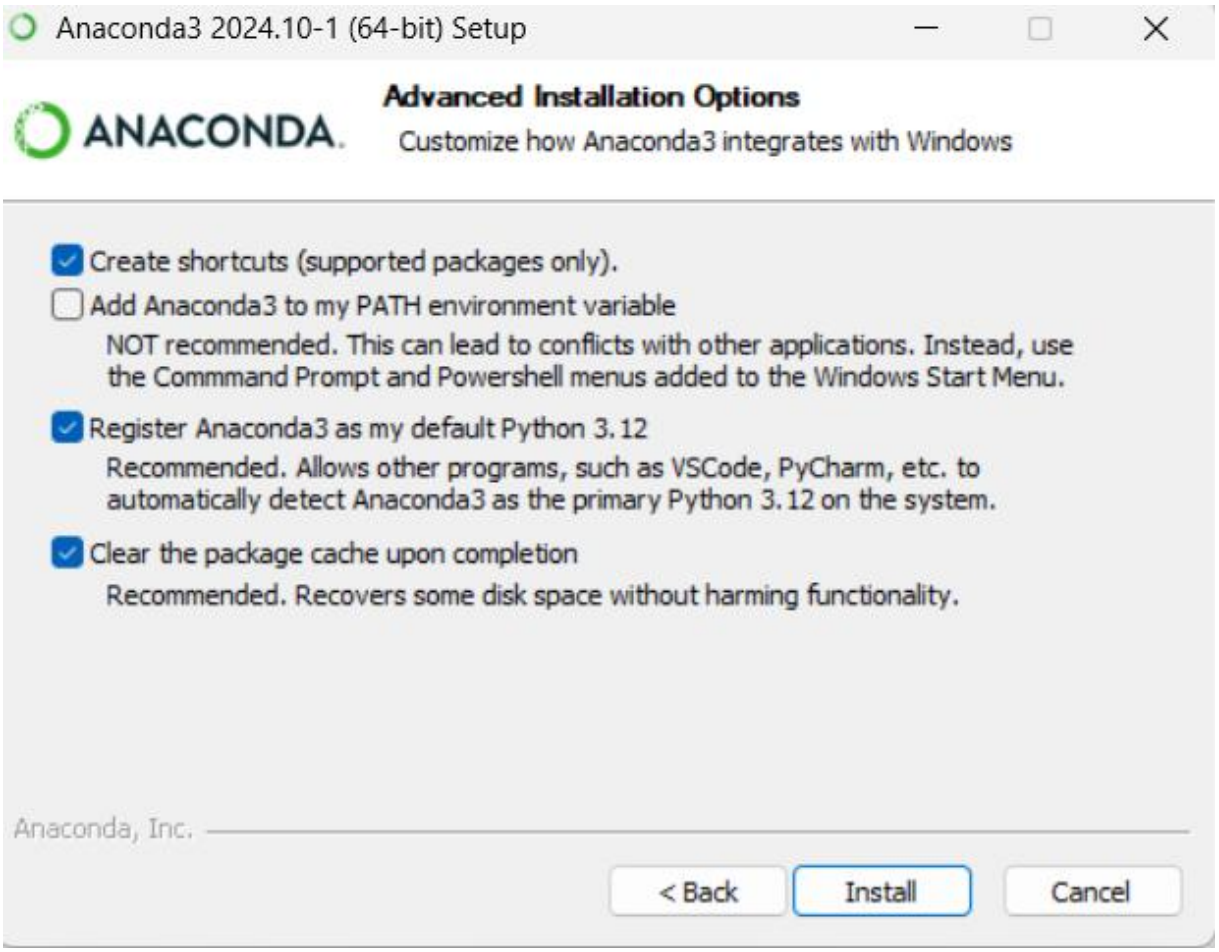



Anaconda3 2024.10-1 (64-bit) Setup
ANACONDA.
Select Installation Type
Please select the type of installation you would like to perform for Anaconda3 2024.10-1 (64-bit).
Install for:
Just Me (recommended)
All Users (requires admin privileges)
Anaconda, Inc.
< Back
Next >
Cancel

Next



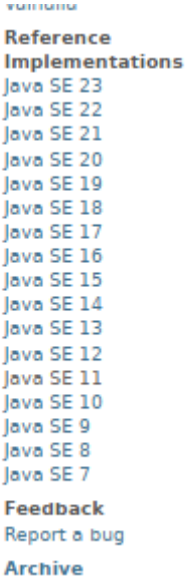
next(ghi nhớ là phải tải ổ c)



Nếu có ô cuối thì tích thêm và bấm install

4. Tải JAVA
   [JDK Builds from Oracle]
   Tải bản nào cx đc

Chọn JDK 23



Chọn java se 11

RI Binary (build 11.0.0.2+2) under the GNU General Public License version 2

- Oracle Linux 8.6/x64 Java Development Kit (sha256) 179 MB
- Windows 11/x64 Java Development Kit (sha256) 180 MB

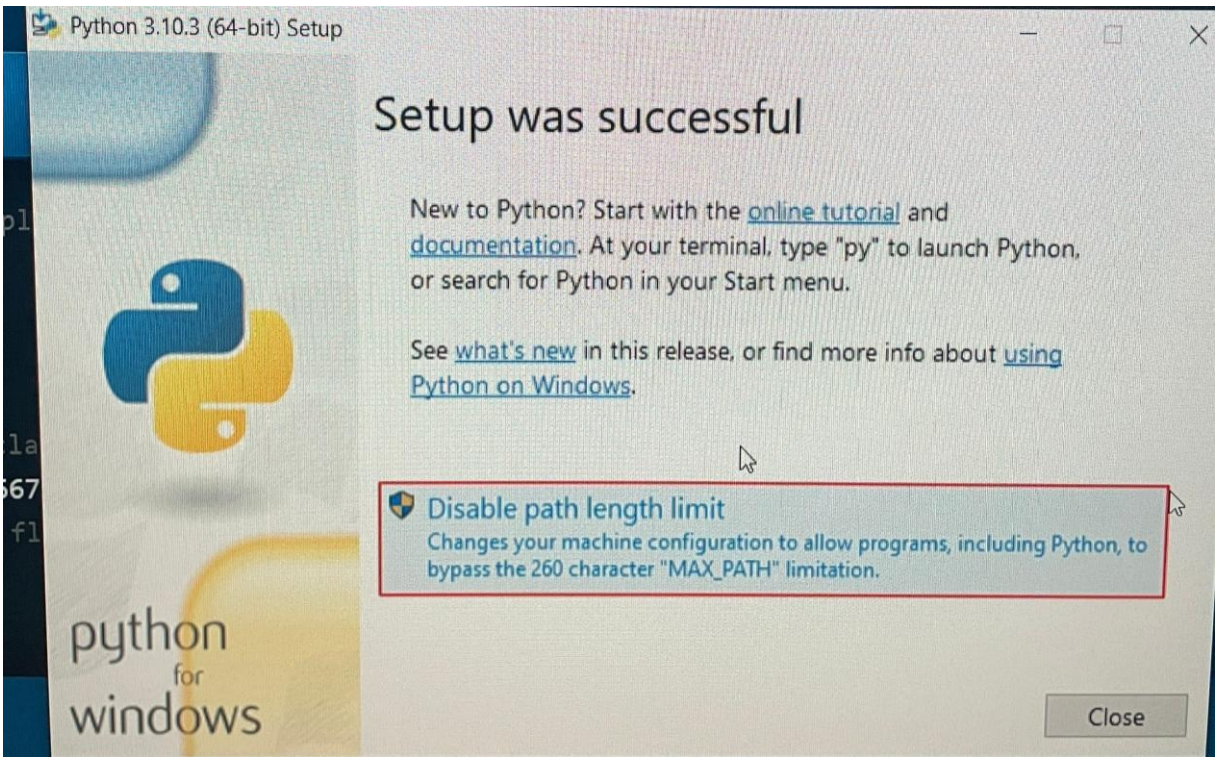
Chọn window

5. Tải python

[Python Release Python 3.10.4 | Python.org]

Nhớ tích cả 2

Nếu có cái này thì tích vào đấy



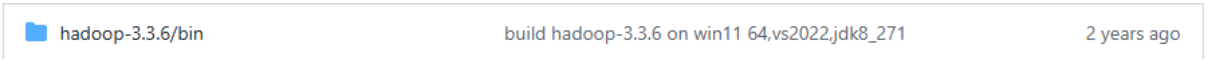
Vào cmd chạy python –version

```
C:\Users\ADMIN>python --version
Python 3.10.4
```

Hiện ra là đã cài xong

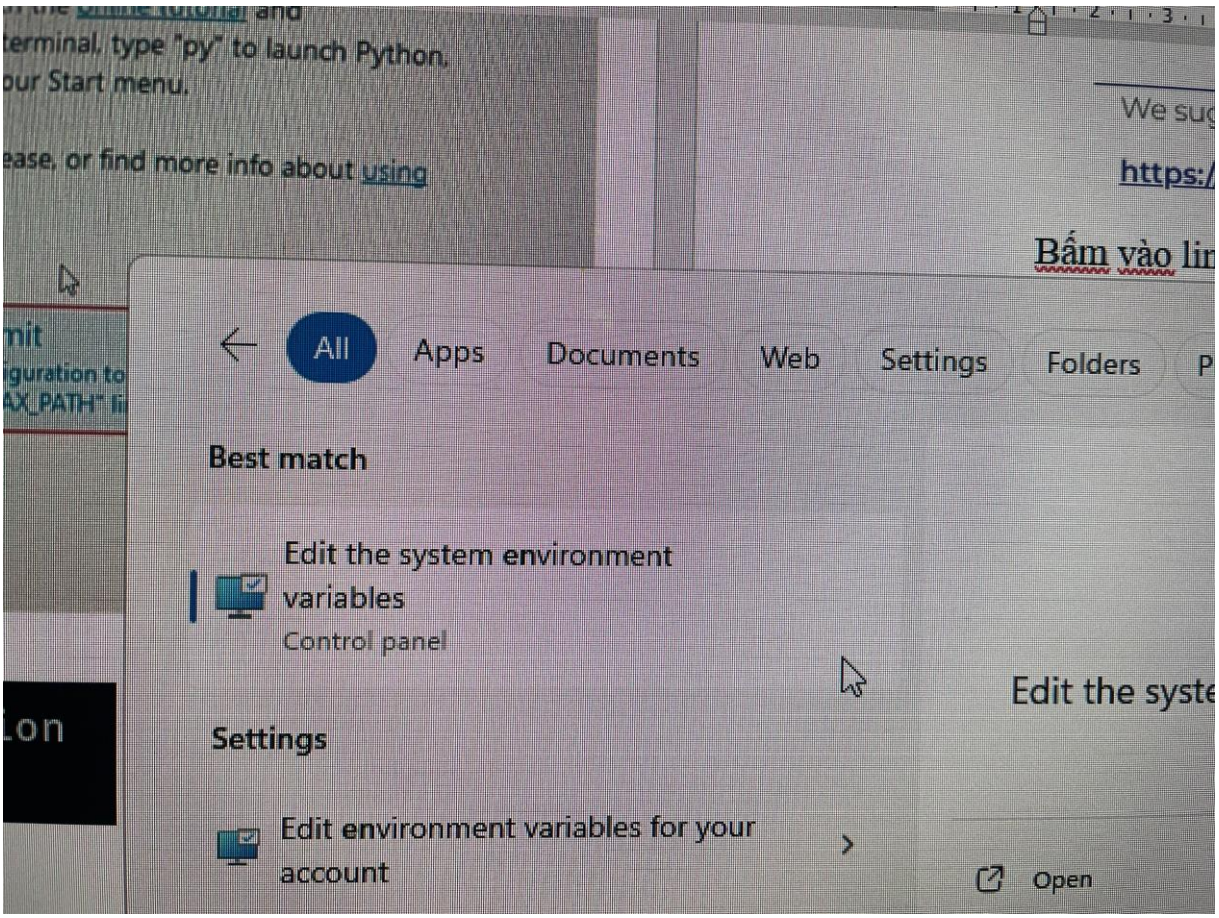
6. Tải Hadoop
   [cdarlint/winutils: winutils.exe hadoop.dll and hdfs.dll binaries for hadoop windows]

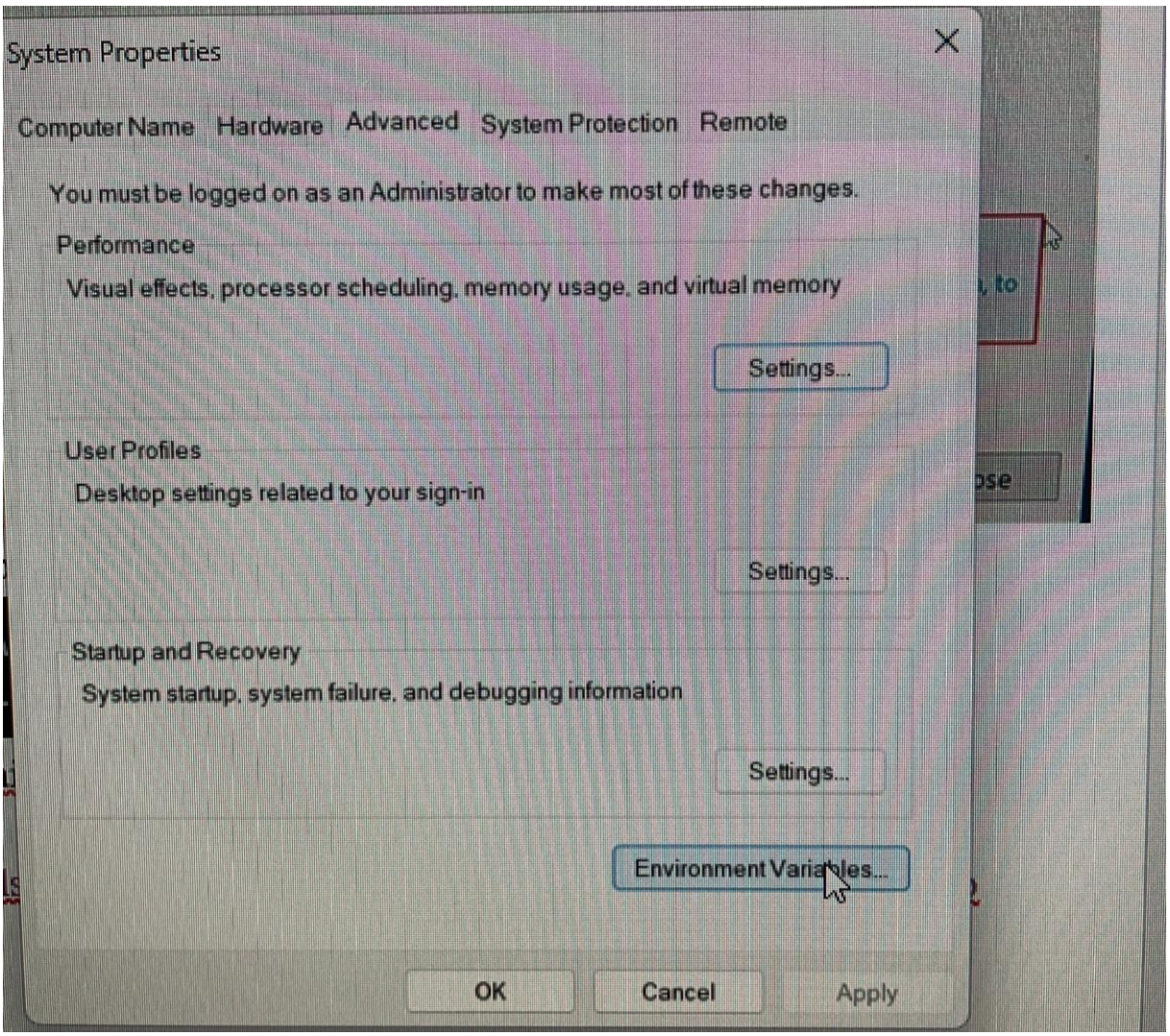
   chọn bản này



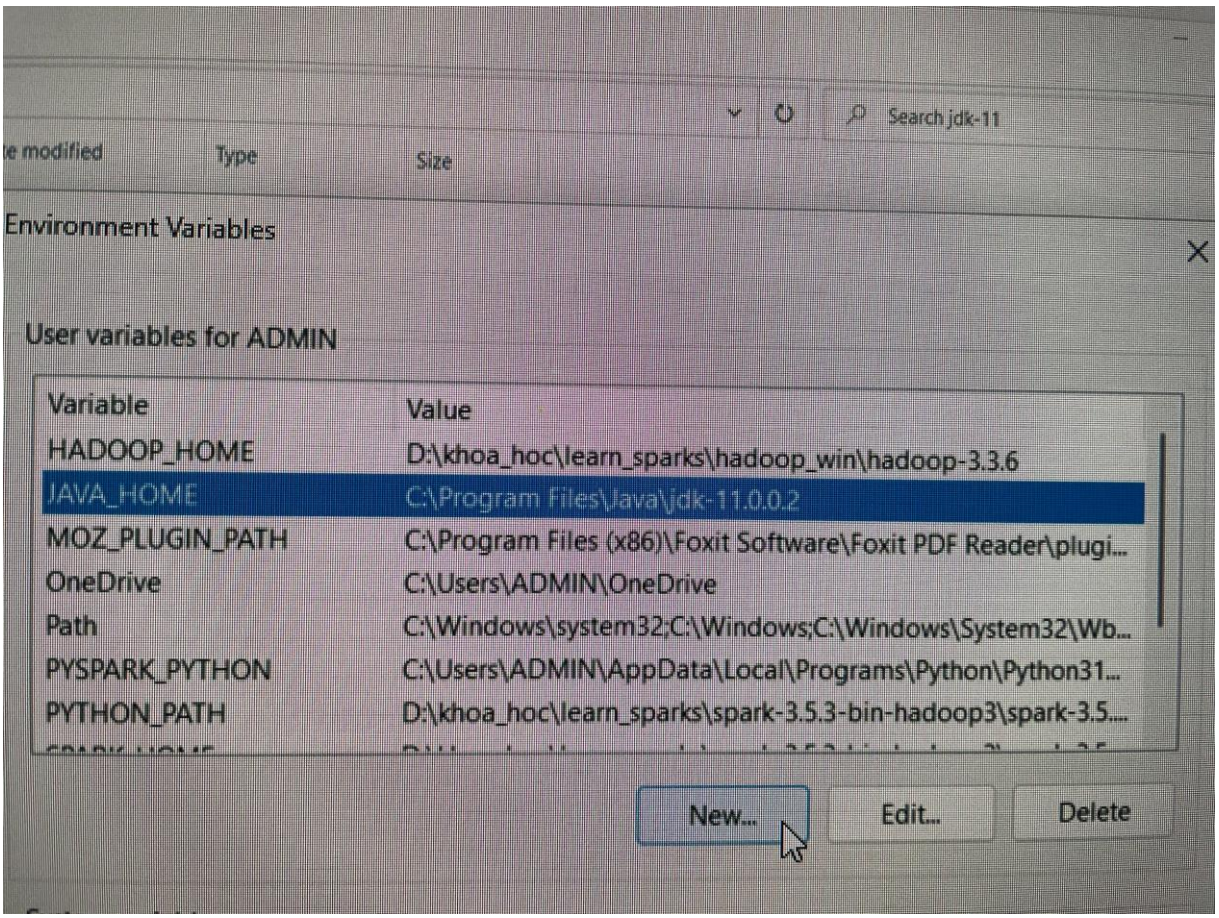
7. Tải apache spark
   Tải theo folder đã gửi

8. Tạo các biến enviroment

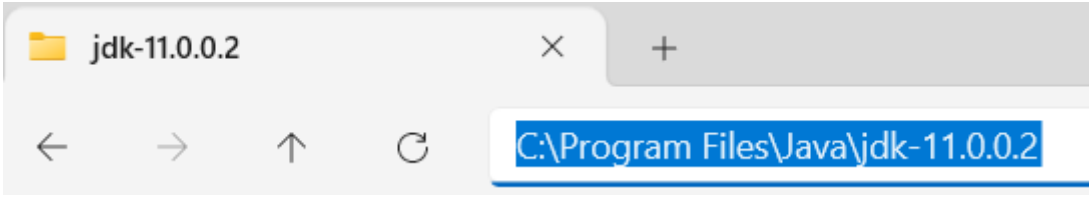

terminal, type "py" to launch Python.
our Start menu.
ease, or find more info about using
We sug
https:/
Bấm vào lin
All
Apps
Documents
Web
Settings
Folders
Best match
Edit the system environment variables
Control panel
Settings
Edit environment variables for your account
Open
Edit the syste



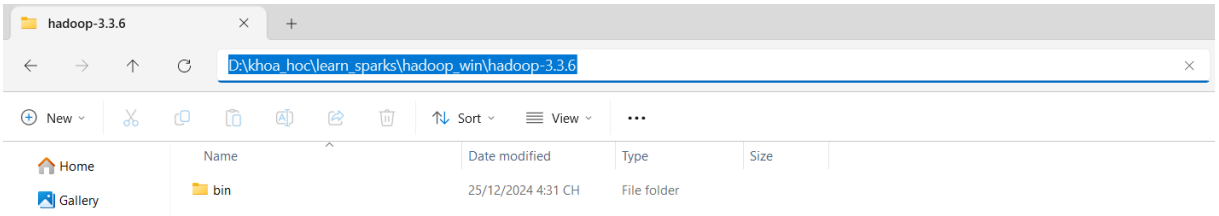
Với mỗi biến sau tạo 1 cái mới bấm new



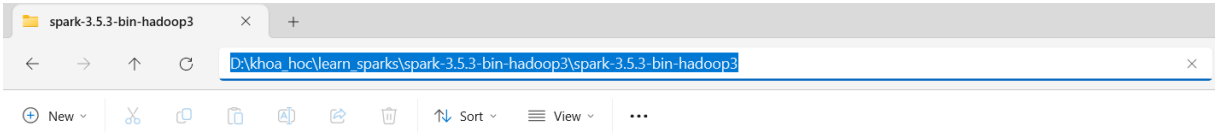
- JAVA_HOME: chọn nơi tải folder java ,ví dụ:



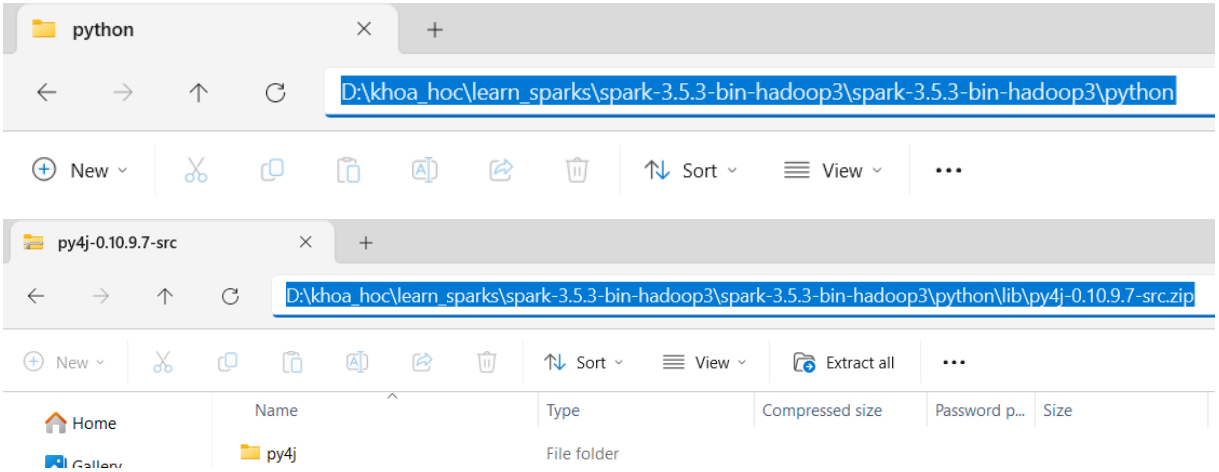
- HADOOP_HOME : folder Hadoop,ví dụ:



- SPARK_HOME:folder spark,ví dụ:



- PYTHON_PATH:trong thư mục spark chọn 2 địa chỉ



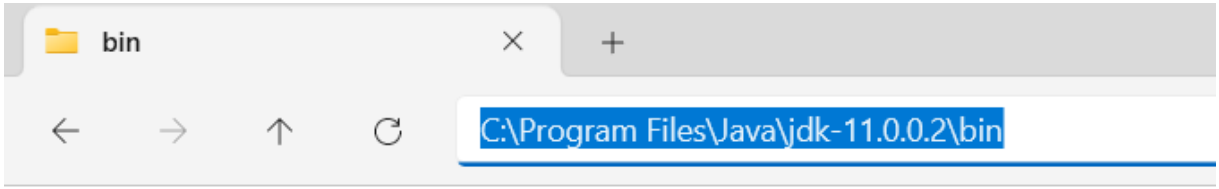
- PYSPARK_PYTHON:

Vào cmd gõ

```
C:\Users\ADMIN>where python
C:\Users\ADMIN\AppData\Local\Programs\Python\Python310\python.exe
```

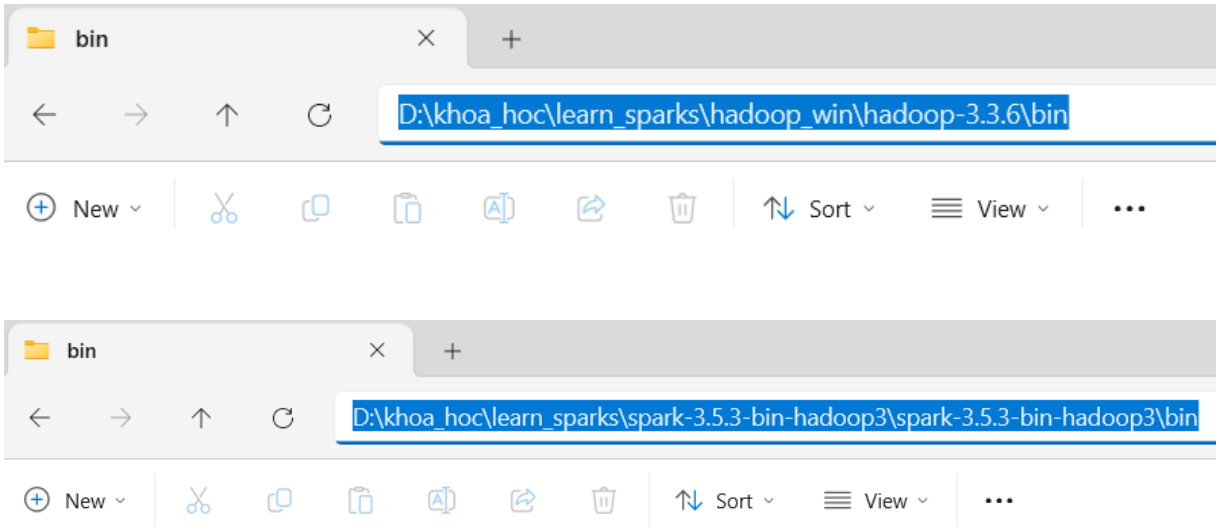
Copy đường dẫn và gán vào biến trên (new trong path)

- SET BIẾN PATH:bấm vào path và gán các địa chỉ sau



Vào cmd gõ java --version test có java chưa

Tiếp theo

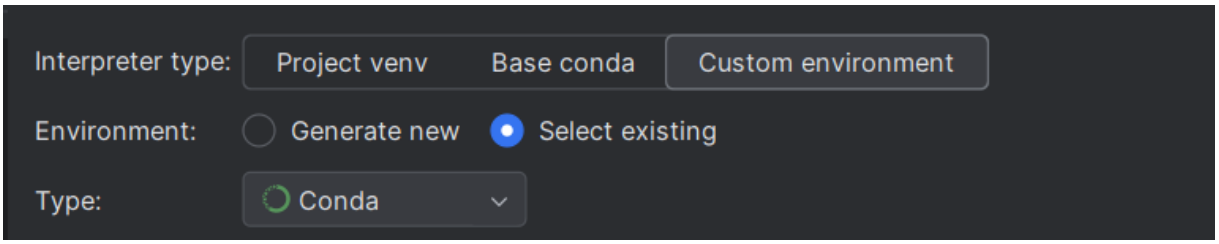


Vào cmd gõ pyspark xem đã chạy đc chưa nếu hiện logo spark là oke

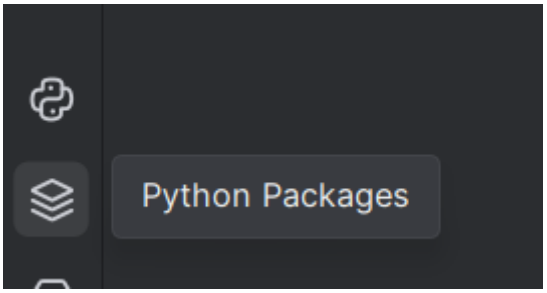
9. TEST CHƯƠNG TRÌNH

Vào pycharm

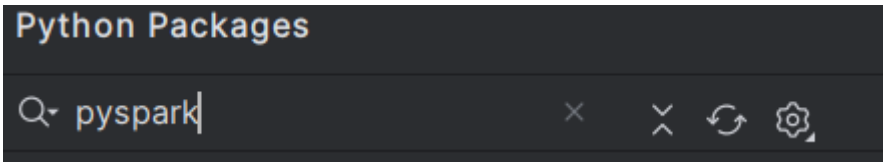
Chọn new project ở file



Vào package



Tìm pyspark và pytest trong conda



Pytest bấm install luôn nếu chưa cài

Pyspark thì chọn bản cùng với version spark đã cài

Lấy 1 file csv nào đó và copy địa chỉ

Tạo file python mới và copy code này

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

spark = SparkSession.builder \
    .master('local[*]') \
    .appName('spark df 2') \
    .getOrCreate()

input_df = spark.read \
    .format('csv') \
    .option('header','true') \
    .option('inferSchema','true') \
    .load('/FileStore/sample_data/orders_sh-1.csv')

input_df.show(5)
order_df = input_df.groupby('order_status') \
    .count()
order_df.show()
```

dán địa chỉ file csv vào load