


HƯỚNG DẪN CÀI ĐẶT PYSPARK TRÊN MÔI TRƯỜNG WINDOWS

1. CHUẨN BỊ

- [Python 3](#) (Nên sử dụng phiên bản 3.6.x, 3.7.x hoặc 3.8.x)
 - [JDK11](#)
 - [Spark](#):
 - Nếu sử dụng Python $\leq 3.7.x$: Sử dụng **spark-2.x.x** hoặc **spark-3.x.x**
 - Nếu sử dụng Python $\geq 3.8.x$: Sử dụng **spark-3.x.x**
- Ở hướng dẫn này sử dụng **Python 3.8.10** và **spark-3.3.0-bin-hadoop3.tgz**



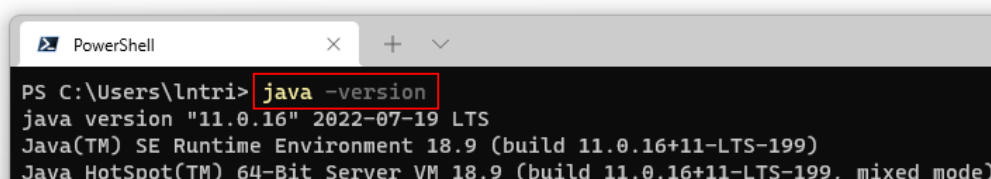
	spark-3.3.0-bin-hadoop3-scala2.13.tgz.sha512	2022-06-09 20:39	168	
	spark-3.3.0-bin-hadoop3.tgz	2022-06-09 20:39	285M	
	spark-3.3.0-bin-hadoop3.tgz.asc	2022-06-09 20:39	862	
	spark-3.3.0-bin-hadoop3.tgz.sha512	2022-06-09 20:39	158	
	spark-3.3.0-bin-without-hadoop.tgz	2022-06-09 20:39	201M	

2. CÀI ĐẶT JAVA 11

- Kiểm tra Java đã được cài đặt chưa?
- Mở cửa sổ **cmd** (Command prompt), gõ lệnh:

```
java -version
```

- Nếu xuất hiện thông tin sau có nghĩa là máy đã có Java:



```
PS C:\Users\lntri> java -version
java version "11.0.16" 2022-07-19 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.16+11-LTS-199)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.16+11-LTS-199, mixed mode)
```

- Nếu máy chưa cài đặt Java, download theo link ở mục 1 (Chuẩn bị):

Linux	macOS	Solaris	Windows
Product/file description	File size	Download	
x64 Installer	140.55 MB	jdk-11.0.16_windows-x64_bin.exe	
x64 Compressed Archive	158.30 MB	jdk-11.0.16_windows-x64_bin.zip	

- Sau khi đã kiểm tra đã có Java trên máy rồi, tiếp theo hãy xác định đường dẫn cài đặt.

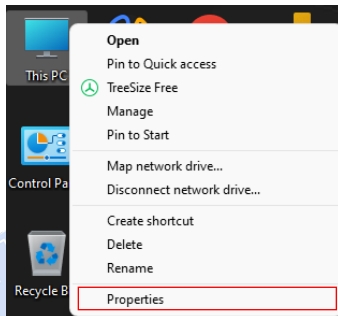
Nếu thao tác cài đặt mặc định thì đường dẫn sẽ có dạng như sau:

C:\Program Files\Java\jdk-11.0.16(*)

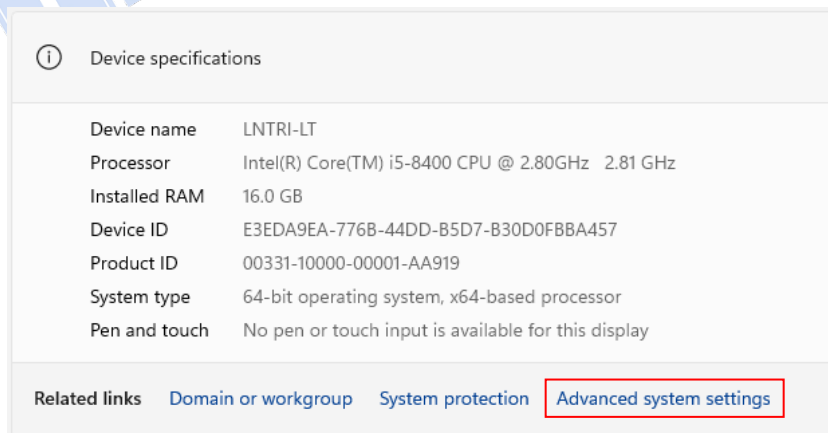
(*): Đây là phiên bản của JDK11, có thể sẽ khác tại từng thời điểm tải về

- Gán đường dẫn Java vào biến môi trường:

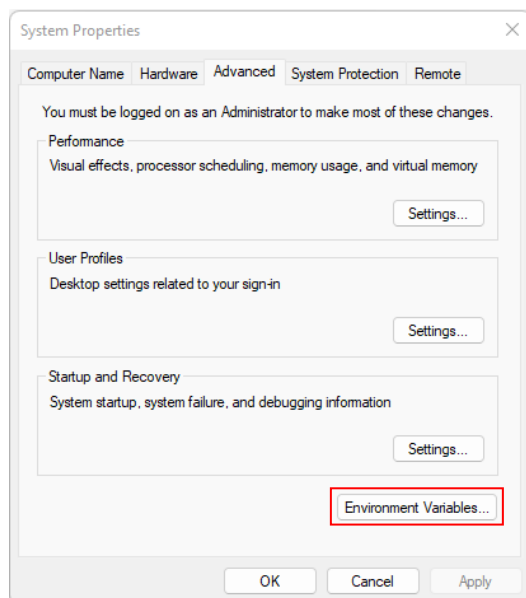
B1: Click phải vào **This PC** -> **Properties**



B2: Trong cửa sổ **Settings**, chọn **Advanced system settings**



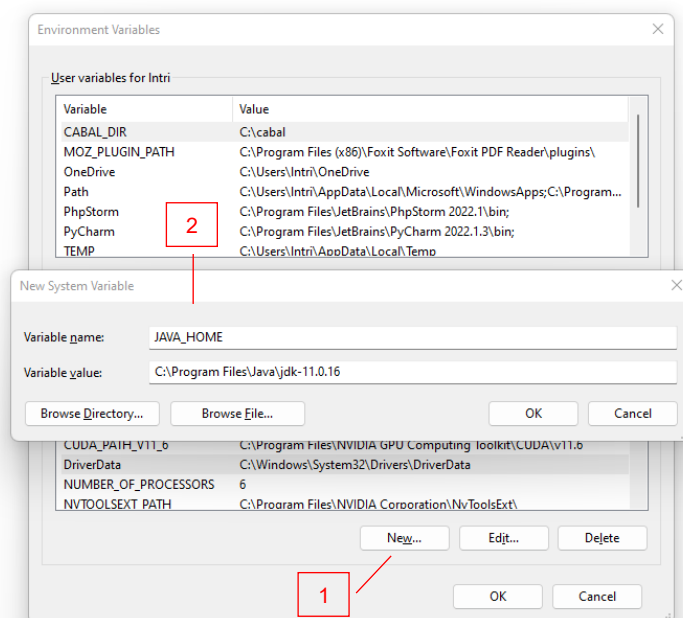
B3: Trong cửa sổ **System Properties**, chọn **Environment Variables...**



B4: Trong **Environment Variables**, chuyển đến **System variables**, thực hiện:

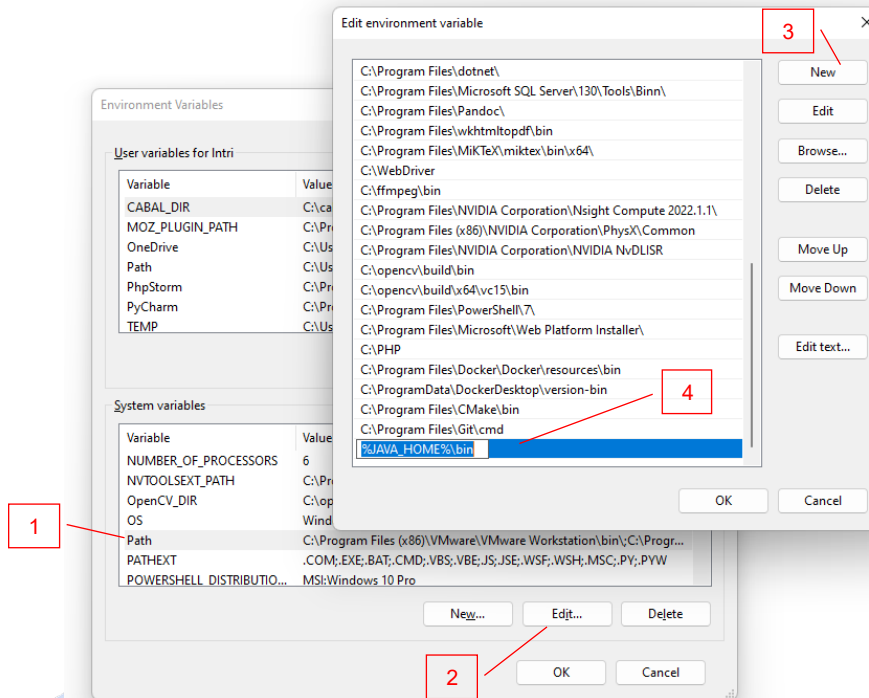
B4.1: Chọn **New...** để tạo mới biến môi trường tên **JAVA_HOME**

JAVA_HOME = C:\Program Files\Java\jdk-11.0.16



B4.2: Chọn biến môi trường **PATH** có sẵn (vẫn trong phần System variables), sau đó chọn **Edit...**, chọn **New**, thêm vào cú pháp như sau:

```
%JAVA_HOME%\bin
```



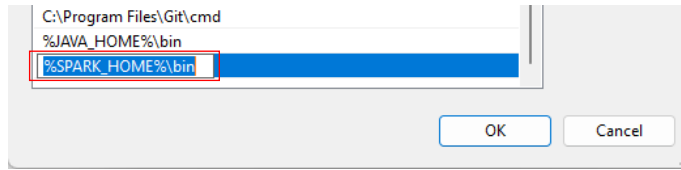
3. CÀI ĐẶT SPARK

- Giải nén file tại đường dẫn tùy chọn (ở hướng dẫn này là **C:\spark**)
- Cài đặt biến môi trường cho **Spark** và **Hadoop** (thao tác như **JAVA_HOME** phía trên)

```
SPARK_HOME = C:\spark\spark-3.3.0-bin-hadoop3
HADOOP_HOME = C:\spark\spark-3.3.0-bin-hadoop3
```

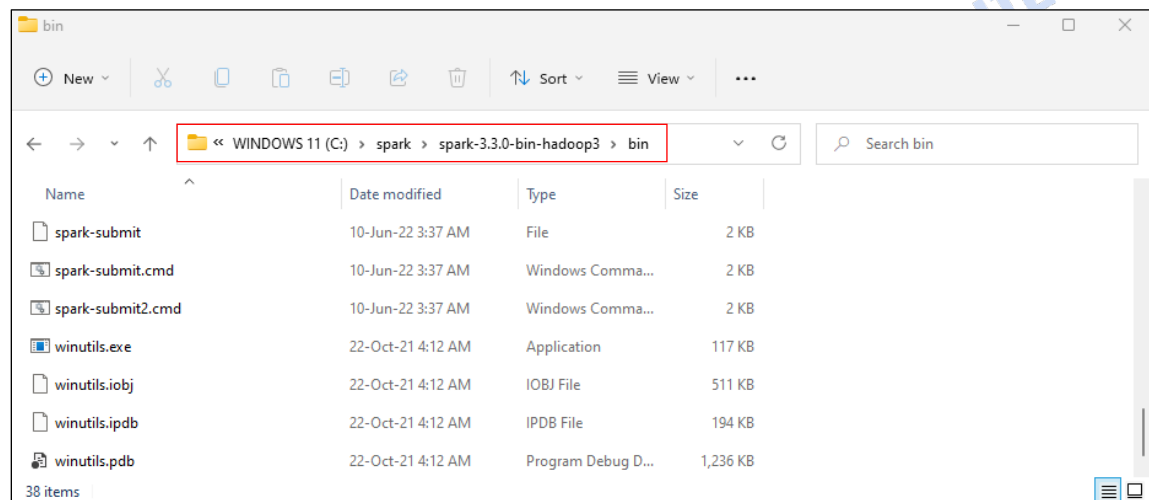
- Chọn biến môi trường **PATH**, chọn Edit..., chọn **New**, gán đường dẫn Spark vào (tương tự như %JAVA_HOME%\bin):

```
%SPARK_HOME%\bin
```



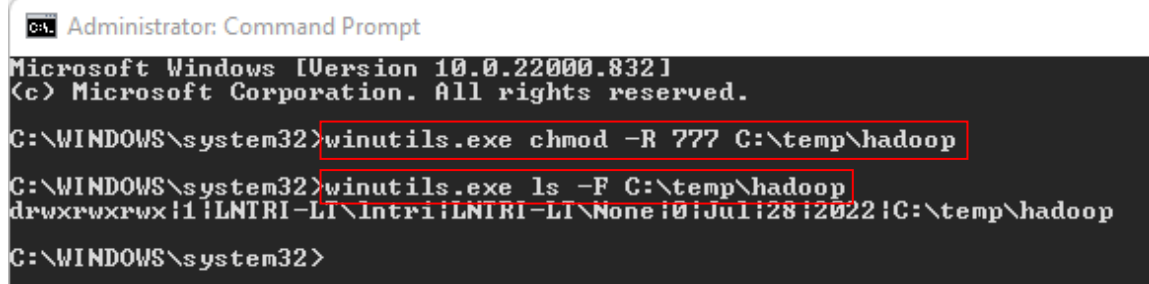
4. DOWNLOAD WINUTILS

- [Download](#) tập tin *winutils-hadoop3.zip*.
- Giải nén tập tin *winutils-hadoop3.zip* đã download, sau đó chép toàn bộ các tập tin trong thư mục *bin* vào thư mục *bin* của spark.



- Tạo thư mục tại 1 ổ đĩa bất kỳ (ở đây sử dụng tại ổ đĩa C):
C:\temp\hadoop
- Mở cửa sổ **cmd** với quyền **Administrator**, gõ lệnh:

```
winutils.exe chmod -R 777 C:\temp\hadoop
winutils.exe ls -F C:\temp\hadoop
```



- Mở cửa sổ **cmd**, gõ lệnh:

- Xuất hiện nội dung sau là đã cấu hình thành công:

```
PS C:\Users\lntri> pyspark
Python 3.8.10 (tags/v3.8.10:3d8993a, May 3 2021, 11:48:03) [MSC v.1928 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

      /---\
     /___\  _--_-----/_/_
    _\  \/_-_-_-_-_\/_/_/_/_/_
   /--/_/_-_-_-_-_-/_/_/_/_/_/_ version 3.3.0
    /_/

Using Python version 3.8.10 (tags/v3.8.10:3d8993a, May 3 2021 11:48:03)
Spark context Web UI available at http://host.docker.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1659423326053).
SparkSession available as 'spark'.
>>>
```

HẾT