

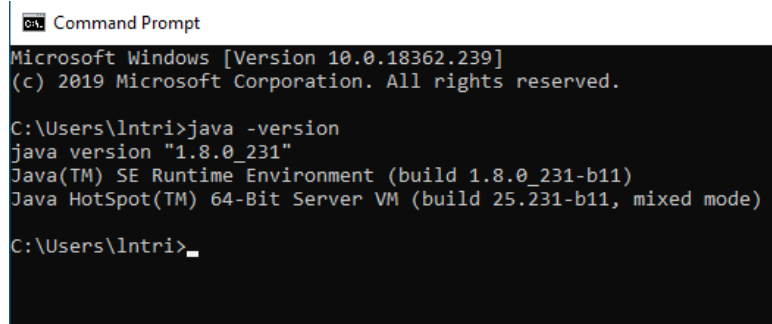
HƯỚNG DẪN CÀI ĐẶT PYSPARK TRÊN MÔI TRƯỜNG WINDOWS

1. CÀI ĐẶT JAVA 8

- Kiểm tra Java đã được cài đặt chưa?
 - Mở cửa sổ **cmd** (command prompt của windows), chạy lệnh sau:

```
java -version
```

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



```
Microsoft Windows [Version 10.0.18362.239]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\lntri>java -version
java version "1.8.0_231"
Java(TM) SE Runtime Environment (build 1.8.0_231-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.231-b11, mixed mode)

C:\Users\lntri>
```

- Nếu máy chưa cài đặt Java, download với từ khóa "**jdk**" (chọn JDK8) từ internet như hình sau (chọn phiên bản 32 bit hoặc 64 bit tùy thuộc vào hệ điều hành đang dùng):

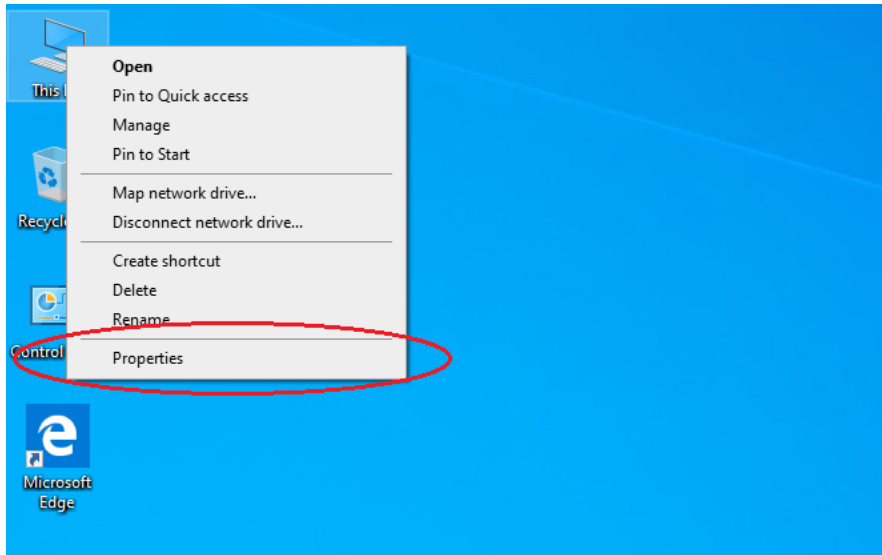
Java SE Development Kit 8u231		
You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.		
<input type="radio"/> Accept License Agreement <input checked="" type="radio"/> Decline License Agreement		
Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.9 MB	jdk-8u231-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.8 MB	jdk-8u231-linux-arm64-vfp-hflt.tar.gz
Linux x86	170.93 MB	jdk-8u231-linux-i586.rpm
Linux x86	185.75 MB	jdk-8u231-linux-i586.tar.gz
Linux x64	170.32 MB	jdk-8u231-linux-x64.rpm
Linux x64	185.16 MB	jdk-8u231-linux-x64.tar.gz
Mac OS X x64	253.4 MB	jdk-8u231-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	132.98 MB	jdk-8u231-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	94.16 MB	jdk-8u231-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.73 MB	jdk-8u231-solaris-x64.tar.Z
Solaris x64	91.96 MB	jdk-8u231-solaris-x64.tar.gz
Windows x86	200.22 MB	jdk-8u231-windows-i586.exe
Windows x64	210.18 MB	jdk-8u231-windows-x64.exe

Tiến hành cài đặt, đường dẫn sau khi cài đặt mặc định như sau:

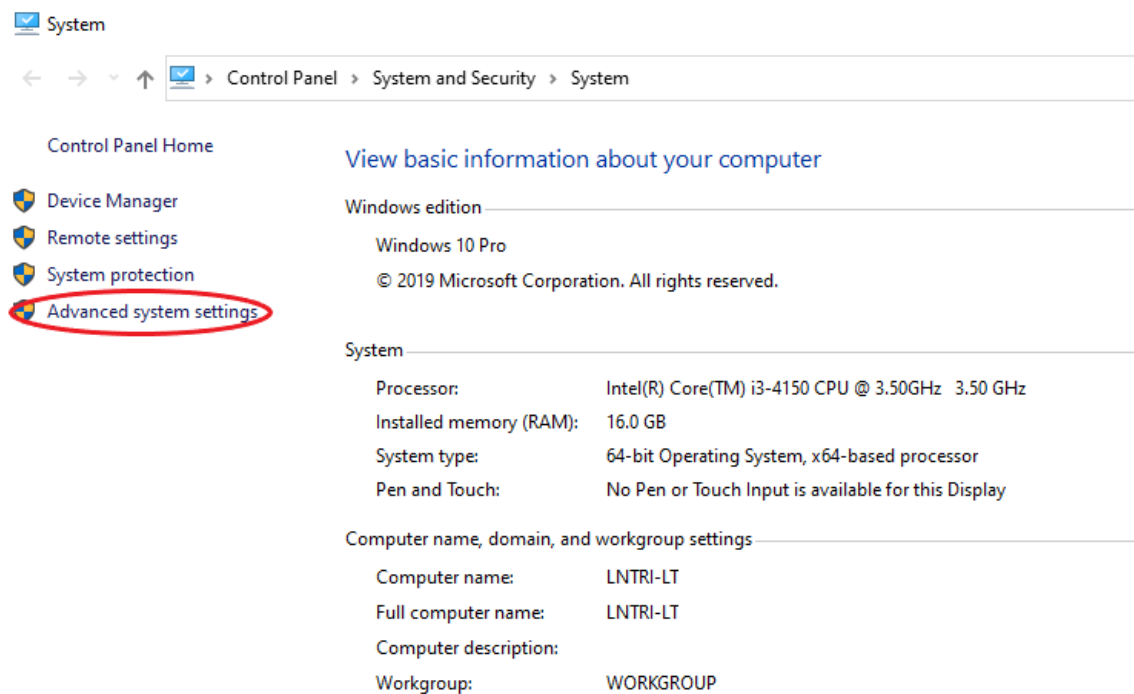
```
C:\ Program Files\Java\jdk1.8.0_231
```

- Thêm Java sau khi cài đặt vào biến môi trường

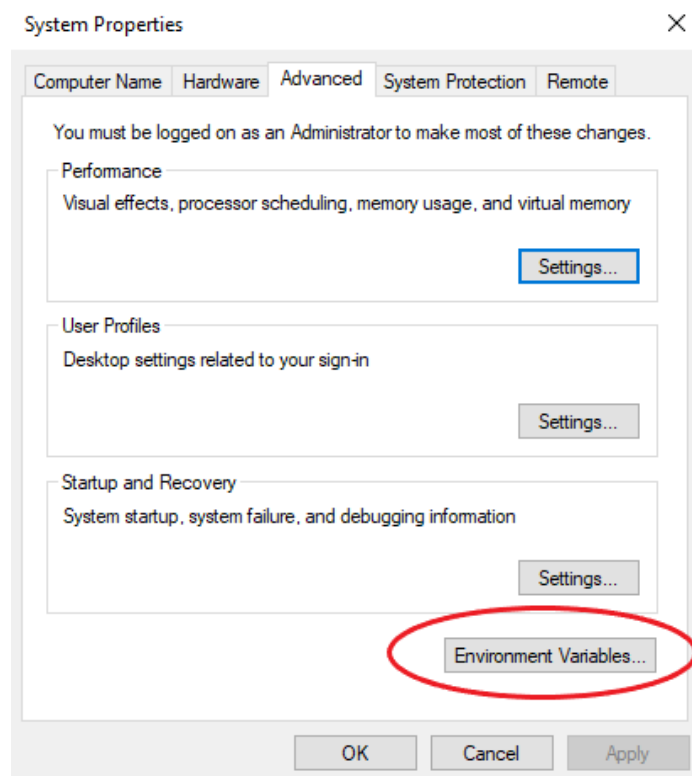
B1: Click phải vào **Computer** -> **Properties**



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



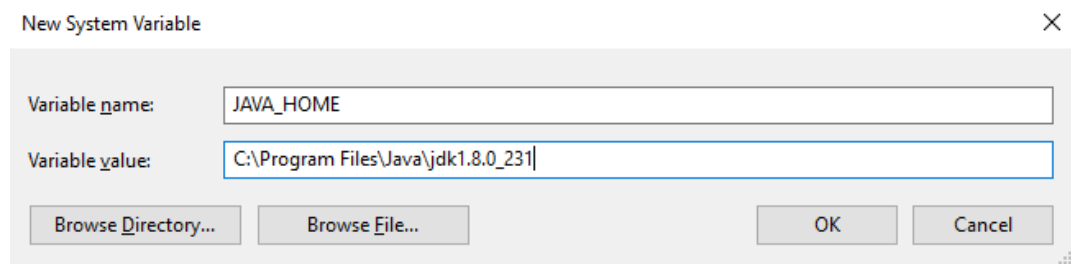
B3: Trong cửa sổ **System Properties**, chọn **Environment Variables...**



B4: Trong **Environment Variables**, chuyển đến **System variables**, ở bước này thực hiện 2 việc:

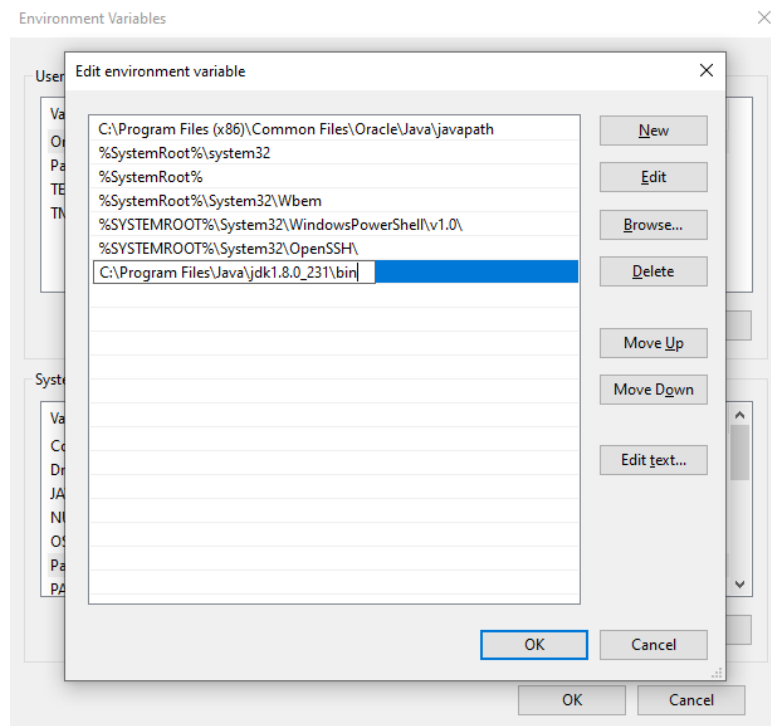
B4.1: Chọn **New...** để tạo mới biến môi trường tên JAVA_HOME

```
JAVA_HOME = C:\Program Files\Java\jdk1.8.0_231
```



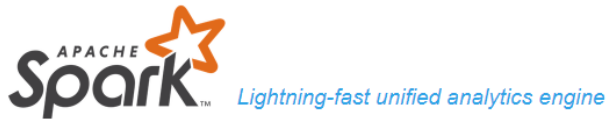
B4.2: Chọn biến môi trường **PATH** có sẵn (vẫn trong phần System variables), sau đó chọn **Edit...**, chọn **New**, thêm vào đường dẫn như sau:

```
C:\Program Files\Java\jdk1.8.0_231\bin
```



2. DOWNLOAD VÀ CÀI ĐẶT SPARK

- Truy cập vào đường dẫn: <http://spark.apache.org/downloads.html>
- Chọn phiên bản **Spark** và **Package type**, sau đó chọn Download file có đuôi mở rộng là **.tgz**



Download Apache Spark™

1. Choose a Spark release:
 2. Choose a package type:
 3. Download Spark: [spark-2.4.4-bin-hadoop2.7.tgz](#)
 4. Verify this release using the 2.4.4 [signatures](#), [checksums](#) and [project release KEYS](#).
- Note that, Spark is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12.

- 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-2.4.4-bin-hadoop2.7
HADOOP_HOME = C:\spark\spark-2.4.4-bin-hadoop2.7
```

- Chọn biến môi trường PATH, chọn Edit..., chọn New, gán đường dẫn Spark vào

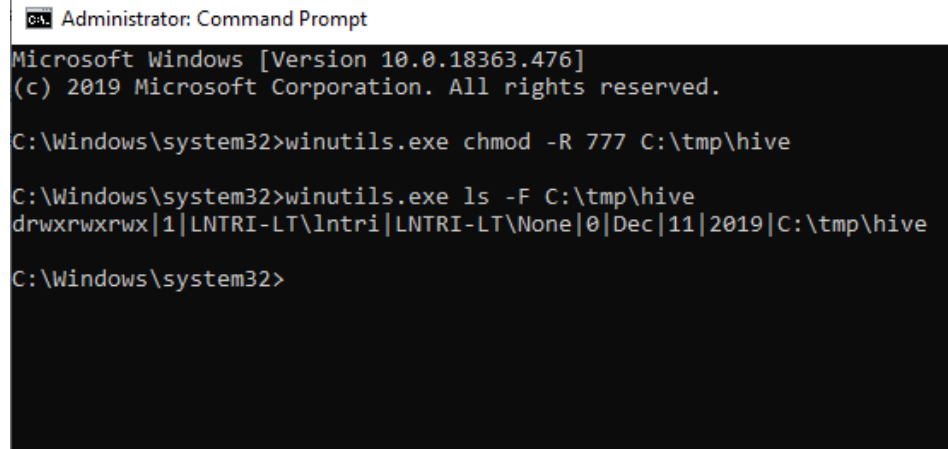
```
C:\spark\spark-2.4.4-bin-hadoop2.7\bin
```

3. DOWNLOAD VÀ CÀI ĐẶT WINUTILS.EXE

- Truy cập vào đường dẫn sau để download winutils.exe:
<https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe>
- Chép file winutils.exe vào thư mục bin của Spark (trong trường hợp này là **C:\spark\spark-2.4.4-bin-hadoop2.7\bin**)
- Tạo thư mục: **C:\tmp\hive**

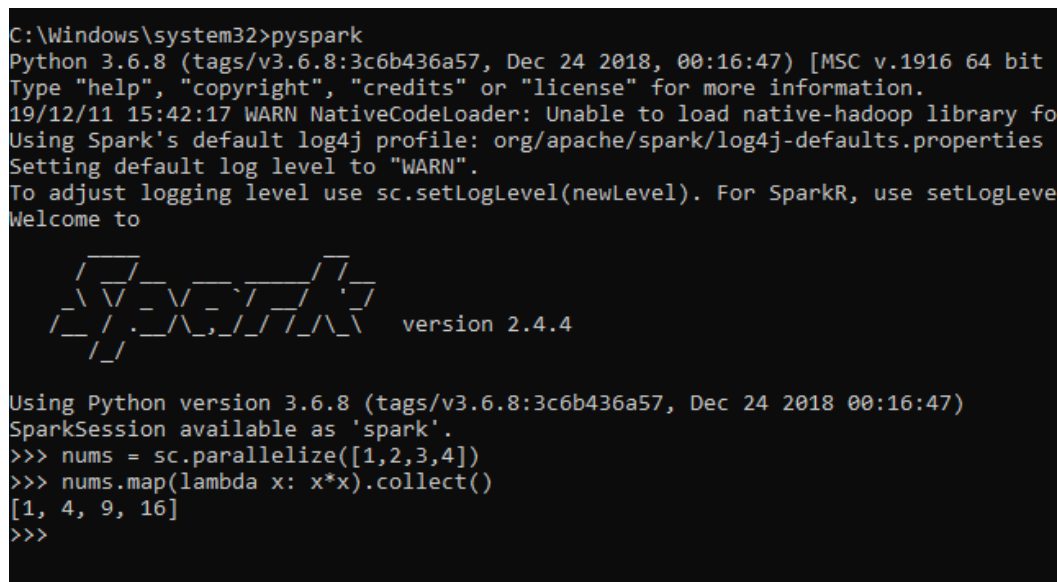
- Mở cửa sổ cmd với quyền Administrator (Run as administrator), chạy lệnh sau:

```
winutils.exe chmod -R 777 C:\tmp\hive  
winutils.exe ls -F C:\tmp\hive
```



A screenshot of a Windows Command Prompt window titled "Administrator: Command Prompt". The window shows the following text: "Microsoft Windows [Version 10.0.18363.476] (c) 2019 Microsoft Corporation. All rights reserved. C:\Windows\system32>winutils.exe chmod -R 777 C:\tmp\hive C:\Windows\system32>winutils.exe ls -F C:\tmp\hive drwxrwxrwx|1|LNTRI-LT\lntri|LNTRI-LT\None|0|Dec|11|2019|C:\tmp\hive C:\Windows\system32>".

4. KIỂM TRA CÀI ĐẶT PYSPARK



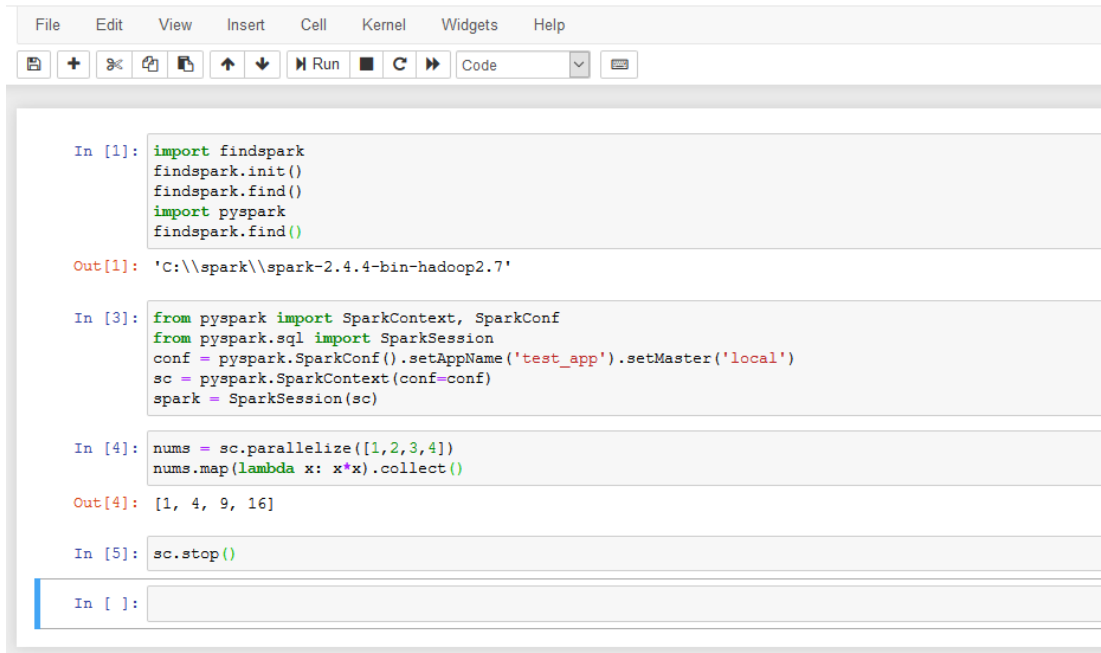
A screenshot of a Windows Command Prompt showing the installation and usage of PySpark. The text includes: "C:\Windows\system32>pyspark Python 3.6.8 (tags/v3.6.8:3c6b436a57, Dec 24 2018, 00:16:47) [MSC v.1916 64 bit Type 'help', 'copyright', 'credits' or 'license' for more information. 19/12/11 15:42:17 WARN NativeCodeLoader: Unable to load native-hadoop library for Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties Setting default log level to 'WARN'. To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(Welcome to version 2.4.4 Using Python version 3.6.8 (tags/v3.6.8:3c6b436a57, Dec 24 2018 00:16:47) SparkSession available as 'spark'. >>> nums = sc.parallelize([1,2,3,4]) >>> nums.map(lambda x: x*x).collect() [1, 4, 9, 16] >>>".

5. PYSPARK VỚI JUPYTER NOTEBOOK

- Cài đặt **findspark**:

```
pip install findspark
```

- Mở **jupyter notebook**, chạy các lệnh sau để kiểm tra:



```
File Edit View Insert Cell Kernel Widgets Help
[Icons] + < > [Icons] Run [Icons] Code [v] [Icon]

In [1]: import findspark
        findspark.init()
        findspark.find()
        import pyspark
        findspark.find()

Out[1]: 'C:\\spark\\spark-2.4.4-bin-hadoop2.7'

In [3]: from pyspark import SparkContext, SparkConf
        from pyspark.sql import SparkSession
        conf = pyspark.SparkConf().setAppName('test_app').setMaster('local')
        sc = pyspark.SparkContext(conf=conf)
        spark = SparkSession(sc)

In [4]: nums = sc.parallelize([1,2,3,4])
        nums.map(lambda x: x*x).collect()

Out[4]: [1, 4, 9, 16]

In [5]: sc.stop()

In [ ]:
```