

HƯỚNG DẪN CÀI ĐẶT PYSPARK TRÊN MÔI TRƯỜNG MACOS

1. CHUẨN BỊ

- [Python](#) (Nên sử dụng phiên bản 3.6 hoặc 3.7)
- [JDK8](#)
- [spark-2.4.7-bin-hadoop2.7.tgz](#)
- [scala-2.13.3.tgz](#)
- [sbt-1.4.0.tgz](#)

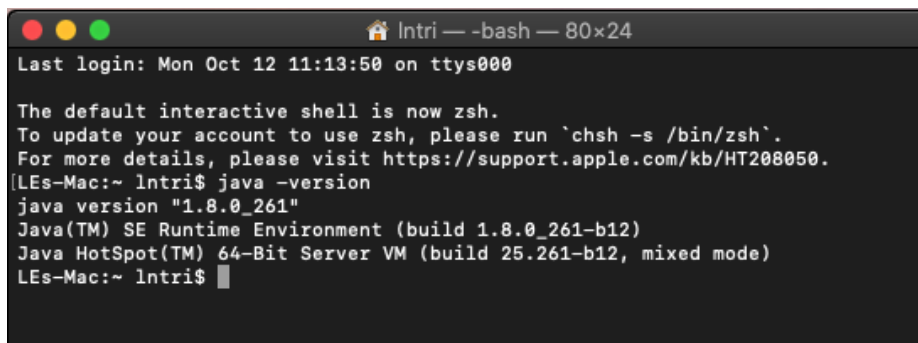
2. CÀI ĐẶT JAVA 8

- Kiểm tra Java đã được cài đặt chưa?

- *Mở cửa sổ **Terminal**, chạy lệnh sau:*

```
java -version
```

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



```
LEs-Mac:~ Intri$ java -version
java version "1.8.0_261"
Java(TM) SE Runtime Environment (build 1.8.0_261-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.261-b12, mixed mode)
LEs-Mac:~ Intri$
```

- Nếu máy chưa cài đặt Java, tìm kiếm với từ khóa "**jdk**" từ internet và chọn download như hình sau:

Linux x86 Compressed Archive	136.81 MB	jdk-8u261-linux-i586.tar.gz
Linux x64 RPM Package	121.53 MB	jdk-8u261-linux-x64.rpm
Linux x64 Compressed Archive	136.48 MB	jdk-8u261-linux-x64.tar.gz
macOS x64	203.94 MB	jdk-8u261-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	125.77 MB	jdk-8u261-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.72 MB	jdk-8u261-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	134.23 MB	jdk-8u261-solaris-x64.tar.Z
Solaris x64	92.47 MB	jdk-8u261-solaris-x64.tar.gz
Windows x86	154.52 MB	jdk-8u261-windows-i586.exe
Windows x64	166.28 MB	jdk-8u261-windows-x64.exe

- Sau khi download, thực hiện cài đặt JDK bình thường.

3. CẤU HÌNH

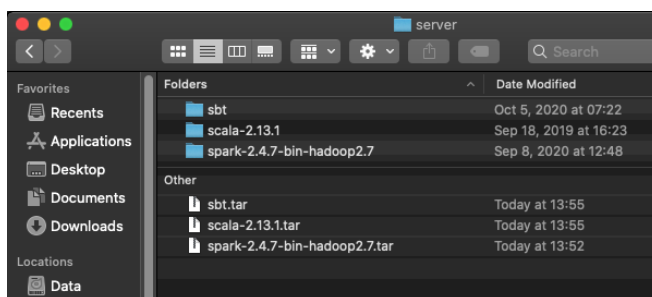
- Mở cửa sổ **Terminal**
- Trở về thư mục **Home** bằng cách gõ lệnh:

```
cd
```

- Tạo thư mục con (ở hướng dẫn này đặt tên là **server**) trong thư mục **Home**, gõ lệnh:

```
mkdir server
```

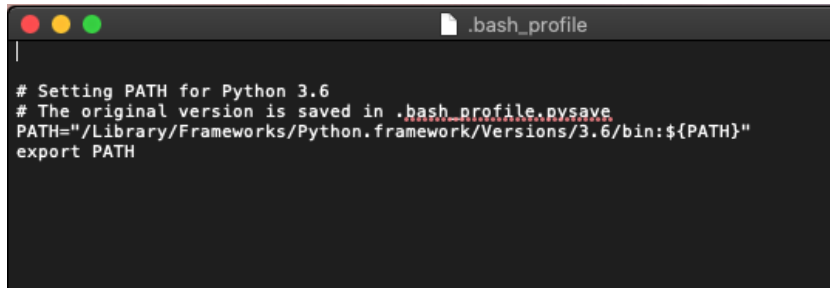
- Copy cả 3 tập tin **scala-2.13.3.tgz**, **sbt-1.4.0.tgz**, **spark-2.4.7-bin-hadoop2.7.tgz** đã download vào thư mục **server** vừa tạo. Sau đó thực hiện giải nén 3 tập tin này bằng ứng dụng **Archive Utility** có sẵn của HĐH macOS.



- Cấu hình biến môi trường: Tại thư mục **Home**, gõ lệnh:

```
touch -c .bash_profile  
open -a TextEdit .bash_profile
```

Xuất hiện cửa sổ:



- Sử dụng mẫu:

```
export  
JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_261.jdk/Contents/Home/  
  
export SPARK_HOME=/Users/intri/server/spark-2.4.7-bin-hadoop2.7  
export SBT_HOME=/Users/intri/server/sbt  
export SCALA_HOME=/Users/intri/server/scala-2.11.12  
  
export  
PATH=$JAVA_HOME/bin:$SBT_HOME/bin:$SBT_HOME/lib:$SCALA_HOME/bin  
:$SCALA_HOME/lib:$PATH  
  
export  
PATH=$JAVA_HOME/bin:$SPARK_HOME:$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH  
  
export PYSARK_PYTHON=python3
```

Lưu ý: Nội dung được đánh dấu màu **đỏ** phía trên là thông tin cần kiểm và cập nhật cho phù hợp với thông tin trên macOS đang dùng.

- Sau khi cấu hình, nội dung sẽ có dạng như sau:

```
export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_261.jdk/Contents/Home/
export SPARK_HOME=/Users/lntri/server/spark-2.4.7-bin-hadoop2.7
export SBT_HOME=/Users/lntri/server/sbt
export SCALA_HOME=/Users/lntri/server/scala-2.13.1
export PATH=$JAVA_HOME/bin:$SBT_HOME/bin:$SBT_HOME/lib:$SCALA_HOME/bin:$SCALA_HOME/lib:$PATH
export PATH=$JAVA_HOME/bin:$SPARK_HOME:$SPARK_HOME/bin:$SPARK_HOME/sbin:$PATH
export PYSARK_PYTHON=python3

# Setting PATH for Python 3.6
# The original version is saved in .bash_profile.pysave
PATH="/Library/Frameworks/Python.framework/Versions/3.6/bin:${PATH}"
export PATH
```

- Save và đóng file **.bash_profile**
- Reload file **.bash_profile** bằng cách gõ lệnh:

```
source ~/.bash_profile
```

4. KIỂM TRA PYSPARK

- Mở cửa sổ **Terminal**
- Gõ lệnh:

pyspark

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

```

Last login: Mon Oct 12 15:16:06 on ttys000

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
LEs-Mac:~ lntri$ pyspark
Python 3.6.8 (v3.6.8:3c6b436a57, Dec 24 2018, 02:04:31)
[GCC 4.2.1 Compatible Apple LLVM 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
20/10/12 17:47:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
in-java classes where applicable
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(newLevel).
Welcome to

      ____
     /___ \
    /   _ \
   /    /  \
  /_____/   \
 /         \
/_         _\
 \        /
  \      /
   \    /
    \__/\
       \

version 2.4.7

Using Python version 3.6.8 (v3.6.8:3c6b436a57, Dec 24 2018 02:04:31)
SparkSession available as 'spark'.
>>>

```

HẾT