SPARK INSTALLATION GUIDE

This is a step by step approach to install Spark on the EC2 instance with Amazon Linux as the OS. You must select the instance type as **t2.micro** and the region as **N. Virginia**.

Also, remember to stop the instance every time you take a break from the session.

Let's start with the steps now.

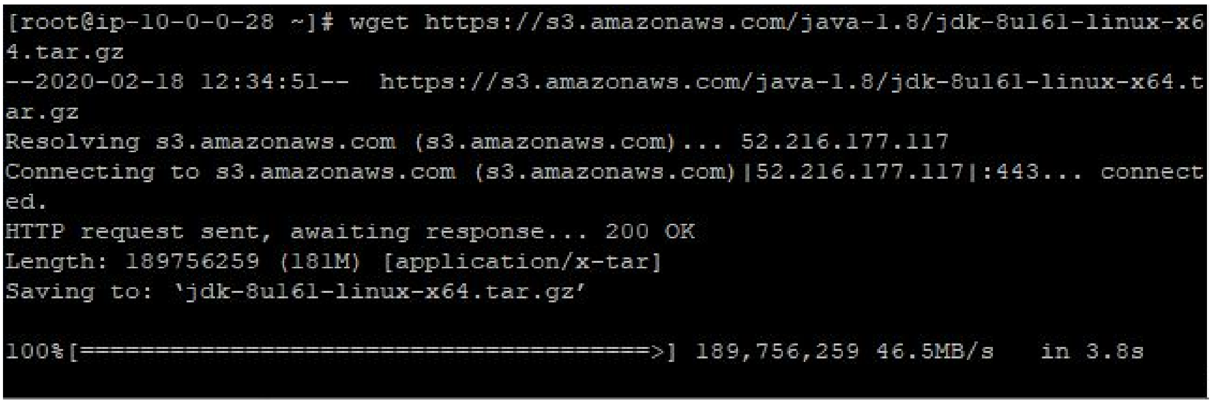
1. After the successful creation of EC2 instance, log in to the EC2 instance using PuTTy in Windows OS or through SSH in Mac/Linux. First, enter into the root directory by using the following command:

```
sudo -i
```

2. The first prerequisite for Spark is Java. For that, you will have to download the JDK 1.8.61 on the instance using the **wget** command provided below.

```
wget https://s3.amazonaws.com/java-1.8/jdk-8u161-linux-x64.tar.gz
```

The above command will give the following output:

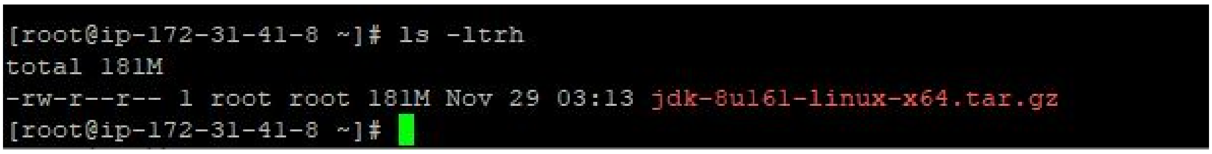


```
[root@ip-10-0-0-28 ~]# wget https://s3.amazonaws.com/java-1.8/jdk-8u161-linux-x6
4.tar.gz
--2020-02-18 12:34:51--  https://s3.amazonaws.com/java-1.8/jdk-8u161-linux-x64.t
ar.gz
Resolving s3.amazonaws.com (s3.amazonaws.com)... 52.216.177.117
Connecting to s3.amazonaws.com (s3.amazonaws.com)|52.216.177.117|:443... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 189756259 (181M) [application/x-tar]
Saving to: 'jdk-8u161-linux-x64.tar.gz'

100%[======================================>] 189,756,259 46.5MB/s   in 3.8s
```

3. To verify the downloaded file, you can list all the components using the below command:

```
ls -ltrh
```

This should show a file named **jdk-8u161-linux-x64.tar.gz** with a size **181MB** size. Refer to the image shown below.



```
[root@ip-172-31-41-8 ~]# ls -ltrh
total 181M
-rw-r--r-- 1 root root 181M Nov 29 03:13 jdk-8u161-linux-x64.tar.gz
[root@ip-172-31-41-8 ~]#
```

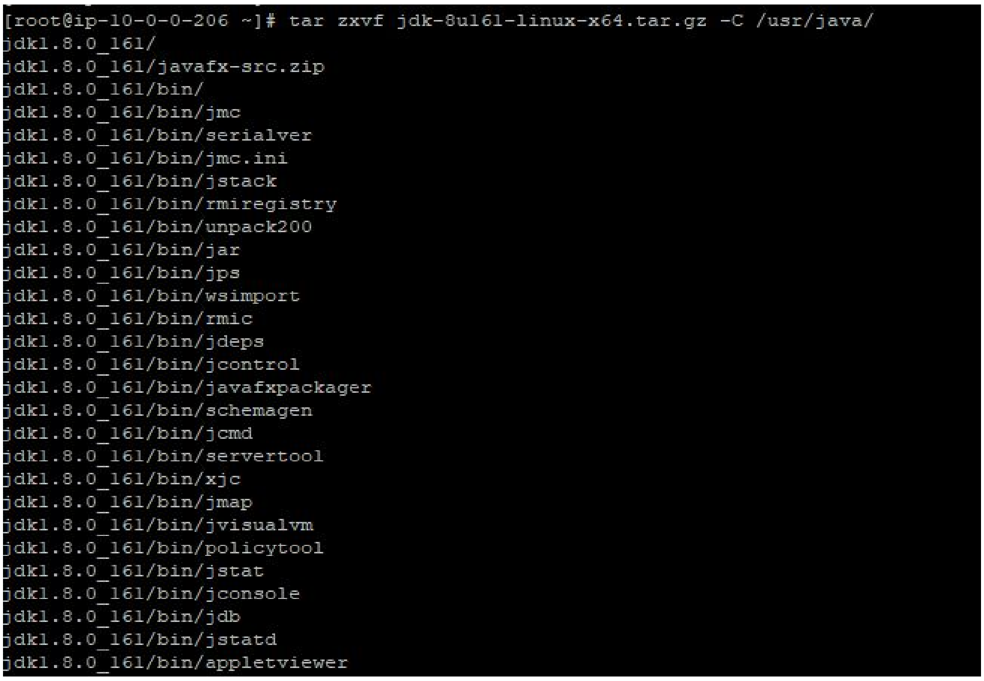
4. Now create a directory using the following command in the **root directory**:

```
mkdir /usr/java
```

5. Extract the tar file using the below command and installed into the location **/usr/java/**
   - Run the following command:

```
tar zxvf jdk-8u161-linux-x64.tar.gz -C /usr/java/
```

   - The above command will give the following output:



```
[root@ip-10-0-0-206 ~]# tar zxvf jdk-8u161-linux-x64.tar.gz -C /usr/java/
jdk1.8.0_161/
jdk1.8.0_161/javafx-src.zip
jdk1.8.0_161/bin/
jdk1.8.0_161/bin/jmc
jdk1.8.0_161/bin/serialver
jdk1.8.0_161/bin/jmc.ini
jdk1.8.0_161/bin/jstack
jdk1.8.0_161/bin/rmiregistry
jdk1.8.0_161/bin/unpack200
jdk1.8.0_161/bin/jar
jdk1.8.0_161/bin/jps
jdk1.8.0_161/bin/wsimport
jdk1.8.0_161/bin/rmic
jdk1.8.0_161/bin/jdeps
jdk1.8.0_161/bin/jcontrol
jdk1.8.0_161/bin/javafxpackager
jdk1.8.0_161/bin/schemagen
jdk1.8.0_161/bin/jcmd
jdk1.8.0_161/bin/servertool
jdk1.8.0_161/bin/xjc
jdk1.8.0_161/bin/jmap
jdk1.8.0_161/bin/jvisualvm
jdk1.8.0_161/bin/policytool
jdk1.8.0_161/bin/jstat
jdk1.8.0_161/bin/jconsole
jdk1.8.0_161/bin/jdb
jdk1.8.0_161/bin/jstatd
jdk1.8.0_161/bin/appletviewer
```

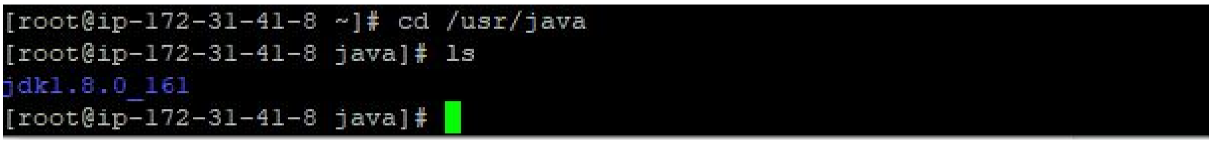
6. To verify the Java 1.8 installation.
   - Run the following command:

```
cd /usr/java
```

- Run the following command:

```
ls
```

- The directory named **jdk1.8.0_161** should be present here.

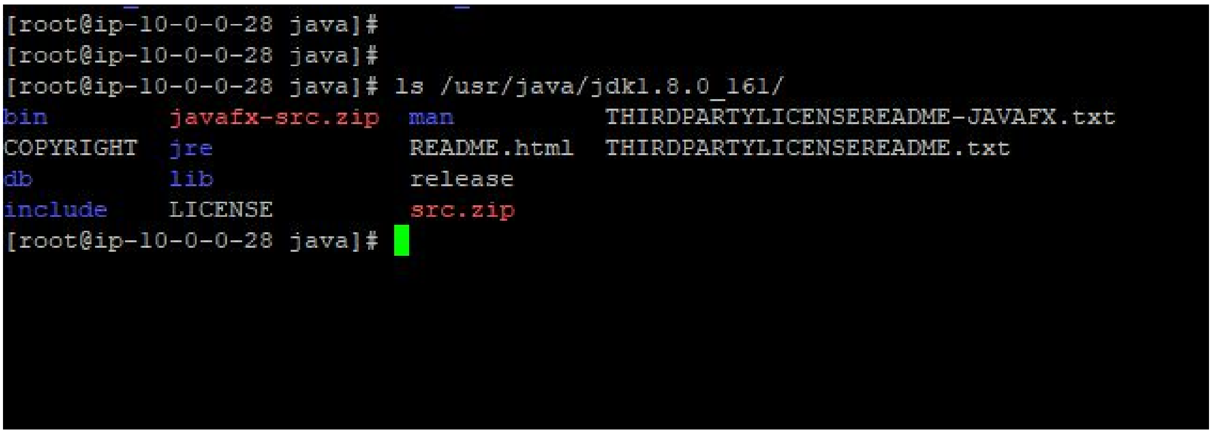


```
[root@ip-172-31-41-8 ~]# cd /usr/java
[root@ip-172-31-41-8 java]# ls
jdk1.8.0_161
[root@ip-172-31-41-8 java]#
```

7. Verify the java and JRE location
   - Run the following command:

```
ls /usr/java/jdk1.8.0_161/
```

   - The above command will give the following output:



```
[root@ip-10-0-0-28 java]#
[root@ip-10-0-0-28 java]#
[root@ip-10-0-0-28 java]# ls /usr/java/jdk1.8.0_161/
bin        javafx-src.zip  man          THIRDPARTYLICENSEREADME-JAVAFX.txt
COPYRIGHT  jre             README.html  THIRDPARTYLICENSEREADME.txt
db         lib             release
include    LICENSE         src.zip
[root@ip-10-0-0-28 java]#
```

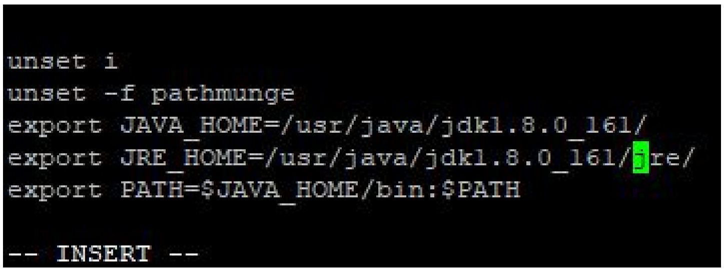
8. Once Java is installed, we need to change the home path of Java so that it can be accessed by all the users from all the folders. You can change the Java_home path for other users by editing to edit /etc/profile
   - Run the following command:

```
vi /etc/profile
```

- Scroll down till the end of the file and press **‘i’** to enter insert mode. **Paste the below three lines as shown in the below screenshot.**

```
export JAVA_HOME=/usr/java/jdk1.8.0_161/
export JRE_HOME=/usr/java/jdk1.8.0_161/jre/
export PATH=$JAVA_HOME/bin:$PATH
```

- It should look something like this:



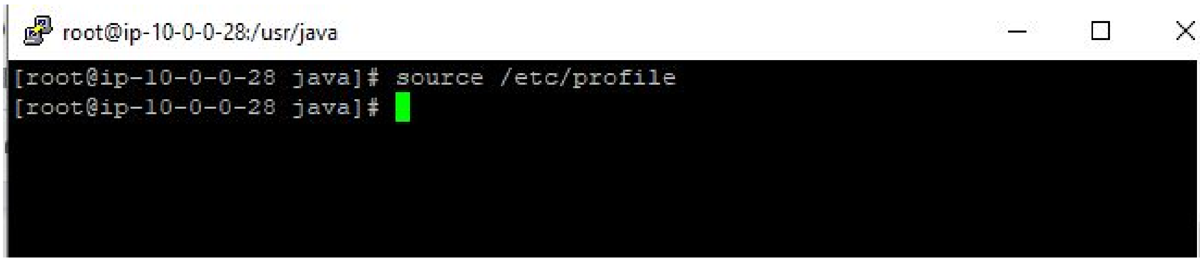
```
unset i
unset -f pathmunge
export JAVA_HOME=/usr/java/jdk1.8.0_161/
export JRE_HOME=/usr/java/jdk1.8.0_161/jre/
export PATH=$JAVA_HOME/bin:$PATH

-- INSERT --
```

- To save and exit the file:
  - Press **Esc**
  - Type**: wq!** and press Enter to save and exit from VI editor.

9. Now update the /etc/profile using the **source command** and check the new java version.

- Run the following command:

```
source /etc/profile
```



```
[root@ip-10-0-0-28 java]# source /etc/profile
[root@ip-10-0-0-28 java]#
```

10. Now change JAVA_HOME path in /etc/bashrc file

- Run the following command:

```
vi /etc/bashrc
```

- Scroll down till the end of the file. Press **‘i’** to enter the insert mode and paste the below three lines as shown in the below screenshot.

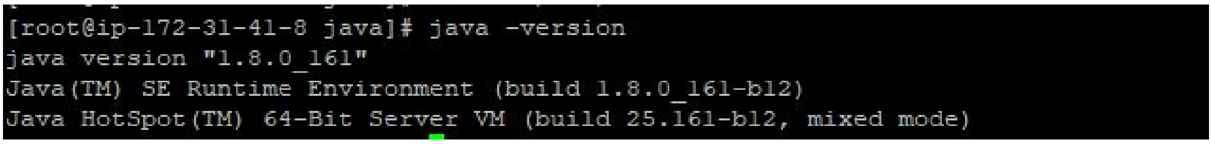
```
export JAVA_HOME=/usr/java/jdk1.8.0_161/
export JRE_HOME=/usr/java/jdk1.8.0_161/jre/
export PATH=$JAVA_HOME/bin:$PATH
```

Repeat the steps to exit the VI editor (**:wq!**)

11. After adding the path, the next step is source it using the following command:

```
source /etc/bashrc
```

- The above command will give the following output:



```
[root@ip-172-31-41-8 java]# java -version
java version "1.8.0_161"
Java(TM) SE Runtime Environment (build 1.8.0_161-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.161-b12, mixed mode)
```

- After sourcing it we will **exit the root directory** by simply typing **exit**.

```
exit
```

12. Now enter the following code in **ec2** directory and run the following command:

```
vi .bash_profile
```

- Scroll down till the export PATH command and just above it press **‘i’** to enter insert mode, **paste the below three lines as shown below.**

```
export JAVA_HOME=/usr/java/jdk1.8.0_161/
export JRE_HOME=/usr/java/jdk1.8.0_161/jre/
export PATH=$JAVA_HOME/bin:$PATH
```

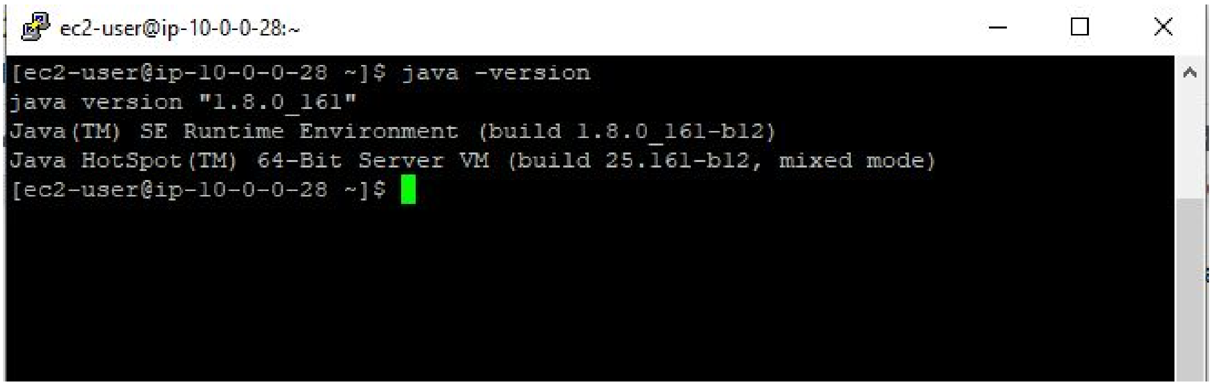
13. Save the bash_profile file and run the following command:

```
source .bash_profile
```

14. Now check the java version using the following command:

```
java -version
```

- The above command will give the following output:



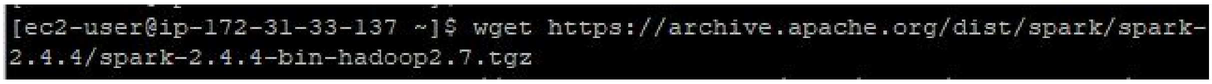
```
[ec2-user@ip-10-0-0-28 ~]$ java -version
java version "1.8.0_161"
Java(TM) SE Runtime Environment (build 1.8.0_161-b12)
Java HotSpot(TM) 64-Bit Server VM (build 25.161-b12, mixed mode)
[ec2-user@ip-10-0-0-28 ~]$
```

You have now installed the prerequisite for Spark - Java. After finishing the installation of Java, the next step is to download Spark.

Spark

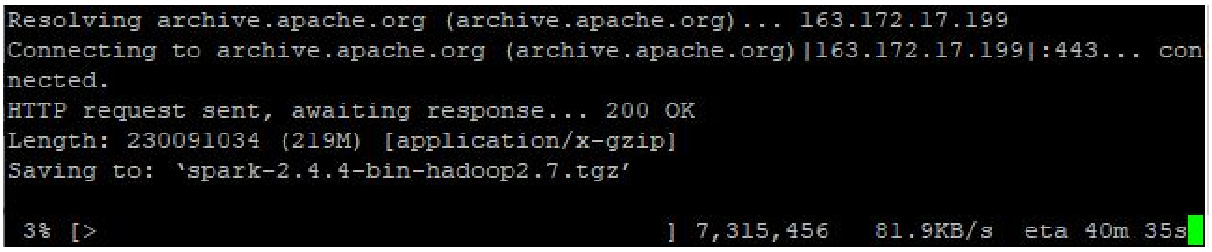
1. First, you will have to download the zip file containing Spark.

```
wget https://archive.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz
```



```
[ec2-user@ip-172-31-33-137 ~]$ wget https://archive.apache.org/dist/spark/spark-
2.4.4/spark-2.4.4-bin-hadoop2.7.tgz
```

- The above command will give the following output:

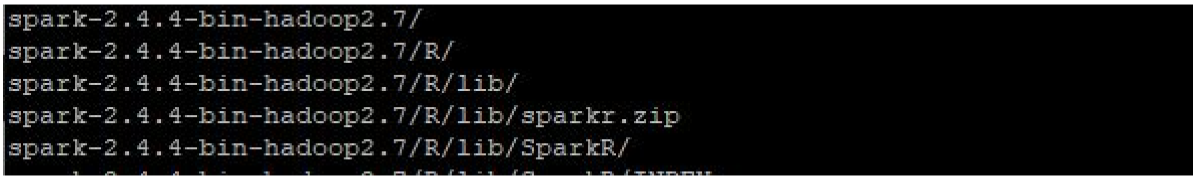


```
Resolving archive.apache.org (archive.apache.org)... 163.172.17.199
Connecting to archive.apache.org (archive.apache.org)|163.172.17.199|:443... con
nected.
HTTP request sent, awaiting response... 200 OK
Length: 230091034 (219M) [application/x-gzip]
Saving to: 'spark-2.4.4-bin-hadoop2.7.tgz'

 3% [>                                           ] 7,315,456   81.9KB/s   eta 40m 35s
```

2. After this step, use the following command in order to load all the Spark libraries:

```
sudo tar -zxvf spark-2.4.4-bin-hadoop2.7.tgz
```

- The above command will give the following output:



```
spark-2.4.4-bin-hadoop2.7/
spark-2.4.4-bin-hadoop2.7/R/
spark-2.4.4-bin-hadoop2.7/R/lib/
spark-2.4.4-bin-hadoop2.7/R/lib/sparkr.zip
spark-2.4.4-bin-hadoop2.7/R/lib/SparkR/
```

3. Now that Spark is installed, the files present in the spark bin can be seen using the following commands:

```
cd spark-2.4.4-bin-hadoop2.7/
```

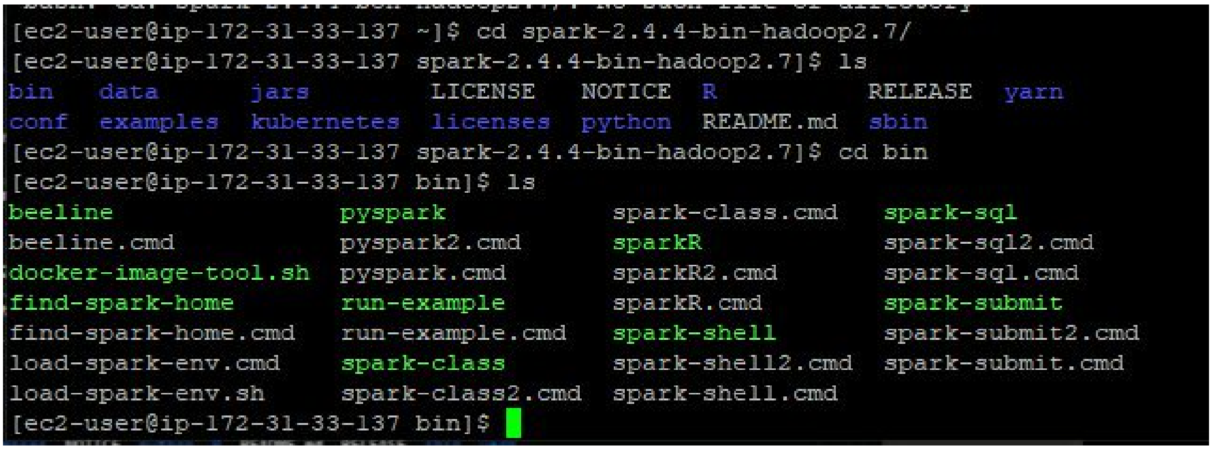
```
ls
```

```
cd bin
```

```
ls
```

```
cd
```

- The above command will give the following output:

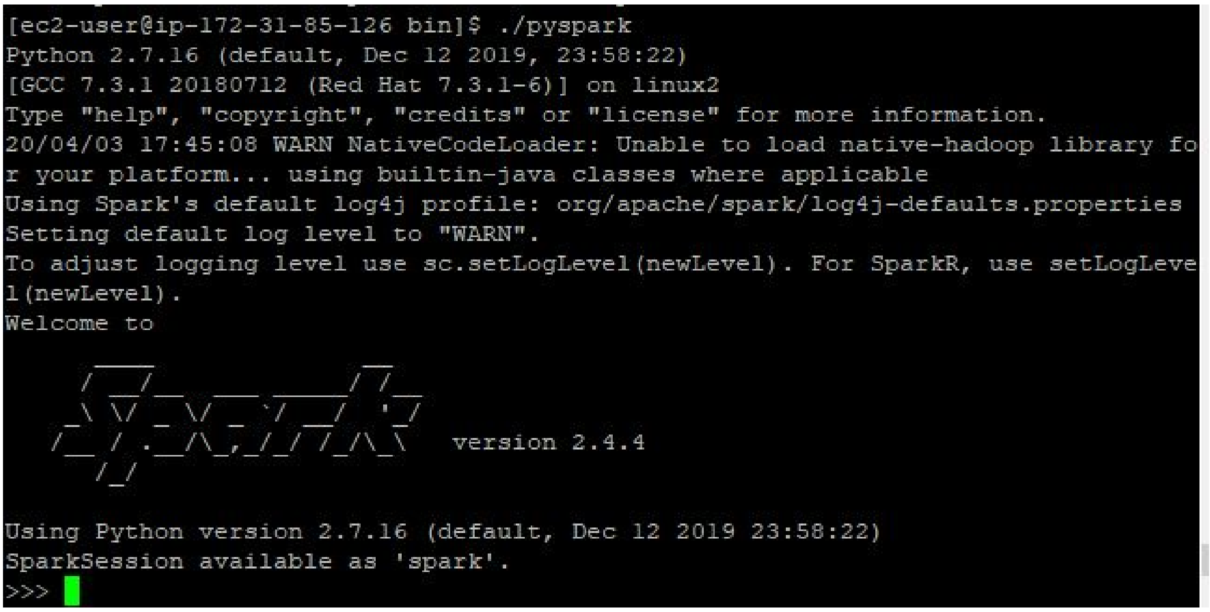


```
[ec2-user@ip-172-31-33-137 ~]$ cd spark-2.4.4-bin-hadoop2.7/
[ec2-user@ip-172-31-33-137 spark-2.4.4-bin-hadoop2.7]$ ls
bin   data      jars        LICENSE   NOTICE  R          RELEASE  yarn
conf  examples  kubernetes  licenses  python  README.md  sbin
[ec2-user@ip-172-31-33-137 spark-2.4.4-bin-hadoop2.7]$ cd bin
[ec2-user@ip-172-31-33-137 bin]$ ls
beeline               pyspark           spark-class.cmd   spark-sql
beeline.cmd           pyspark2.cmd      sparkR            spark-sql2.cmd
docker-image-tool.sh  pyspark.cmd       sparkR2.cmd       spark-sql.cmd
find-spark-home       run-example       sparkR.cmd        spark-submit
find-spark-home.cmd   run-example.cmd   spark-shell       spark-submit2.cmd
load-spark-env.cmd    spark-class       spark-shell2.cmd  spark-submit.cmd
load-spark-env.sh     spark-class2.cmd  spark-shell.cmd
[ec2-user@ip-172-31-33-137 bin]$
```

4. In order to enter into pyspark, use the following command:

```
bin/pyspark
```

- This will be the output of the above command ensuring that pyspark has been installed in your system:



```
[ec2-user@ip-172-31-85-126 bin]$ ./pyspark
Python 2.7.16 (default, Dec 12 2019, 23:58:22)
[GCC 7.3.1 20180712 (Red Hat 7.3.1-6)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
20/04/03 17:45:08 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-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 setLogLeve
l(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.4
      /_/

Using Python version 2.7.16 (default, Dec 12 2019 23:58:22)
SparkSession available as 'spark'.
>>>
```

5. The Spark object can be analysed using the following commands:

```
spark
```

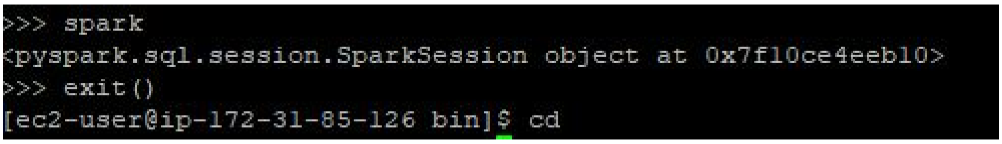
- To exit, use:

```
exit()
```

- In order to exit the bin directory:

```
cd
```

- The above command will give the following output:



```
>>> spark
<pyspark.sql.session.SparkSession object at 0x7f10ce4eeb10>
>>> exit()
[ec2-user@ip-172-31-85-126 bin]$ cd
```

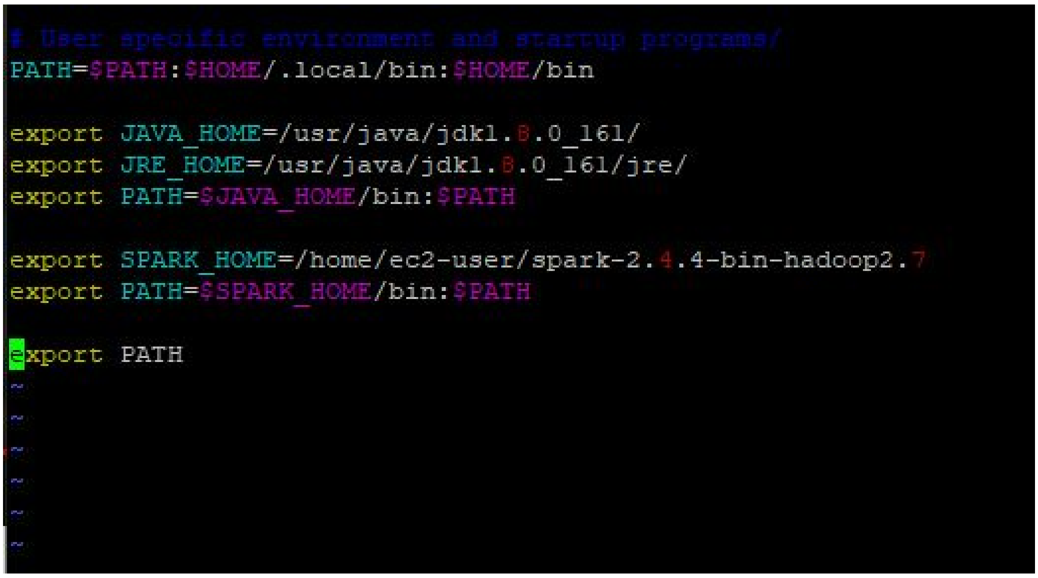
6. Now we can add spark home to our path, in order to make it easier and accessible:

```
vi .bash_profile
```

- Now we will be performing the following commands:

```
SPARK_HOME=/home/ec2-user/spark-2.4.4-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$PATH
```

- The Command Prompt will look like this:



```
# User specific environment and startup programs/
PATH=$PATH:$HOME/.local/bin:$HOME/bin

export JAVA_HOME=/usr/java/jdk1.8.0_161/
export JRE_HOME=/usr/java/jdk1.8.0_161/jre/
export PATH=$JAVA_HOME/bin:$PATH

export SPARK_HOME=/home/ec2-user/spark-2.4.4-bin-hadoop2.7
export PATH=$SPARK_HOME/bin:$PATH

export PATH
~
~
~
~
~
~
```

- Then exit using: wq! command

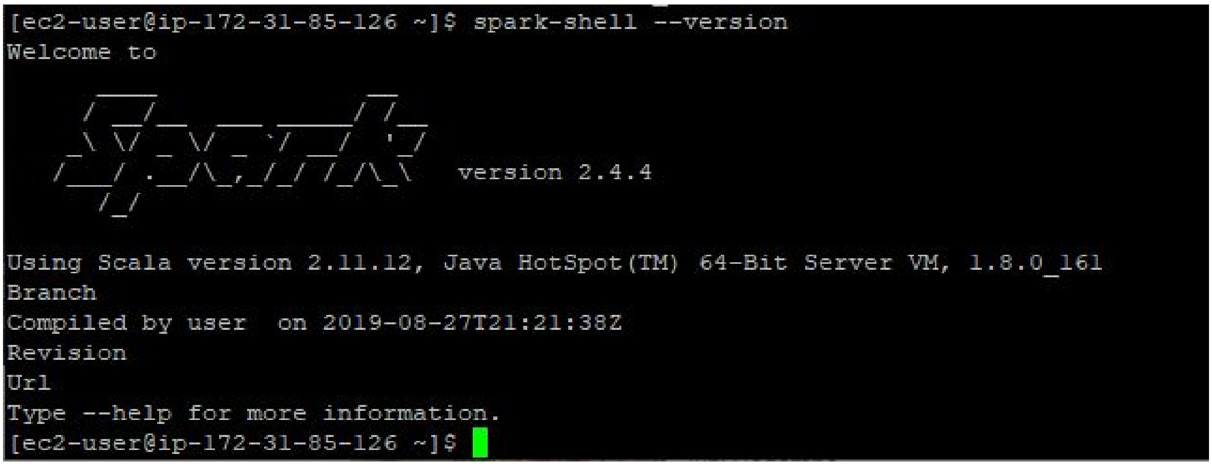
- Now source this file :

```
source .bash_profile
```

7. Now we can see that spark-shell has been installed.

```
spark-shell --version
```

- The above command will give the following output:



```
[ec2-user@ip-172-31-85-126 ~]$ spark-shell --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.4
      /_/

Using Scala version 2.11.12, Java HotSpot(TM) 64-Bit Server VM, 1.8.0_161
Branch
Compiled by user  on 2019-08-27T21:21:38Z
Revision
Url
Type --help for more information.
[ec2-user@ip-172-31-85-126 ~]$
```

Your instance is now ready with PySpark.

8. Once the work on pyspark is done, **exit** the shell using the following commands:

```
quit()
```

- Or exit using CTRL+D command