Installing Spark Pre-built for Hadoop version 2.7 and later

You will need to install Ubuntu and Java on Virtual Machine

1. Download VirtualBox from Oracle Website (32 bit)

https://www.virtualbox.org/wiki/Downloads

2. Install Oracle VirtualBox like any other software.

3. Download Ubuntu Desktop Image from the following URL (32 bit)

http://releases.ubuntu.com/16.04/

4. Use the instructions in the below link to setup Ubuntu in VirtuaBox

http://www.wikihow.com/Install-Ubuntu-on-VirtualBox

5. Now open Terminal and execute the following operations

a. ***Login as root user using the command – "su root"***

[To proceed to next step we need root password, so I followed below link to reset the password]

https://linuxconfig.org/how-to-reset-lost-root-password-on-ubuntu-16-04-xenial-xerus-linux

Now Login as root user using the command – "su root"

sudo add-apt-repository ppa:webupd8team/java

sudo apt-get update

sudo apt-get install oracle-java8-installer

sudo apt-get install default-jre

sudo apt-get install default-jdk

```
root@hp500-VirtualBox:/media/sf_Big_Data/spark-2.2.1-bin-hadoop2.7# java -version
openjdk version "1.8.0_151"
OpenJDK Runtime Environment (build 1.8.0_151-8u151-b12-0ubuntu0.16.04.2-b12)
OpenJDK Server VM (build 25.151-b12, mixed mode)
```

Check if Java is installed.

If above command doesn't install the JDK or throws connection refused error. That's because the oracle has moved the file to other destination than the one which the above command tries to access, hence the installation fails.

In that case follow the manual method to download and install the JDK/ JRE.

Download the JDK for the ubuntu 32 bit

The manual way

- Download the 32-bit or 64-bit Linux "compressed binary file" - it has a ".tar.gz" file extension.
- http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase7-521261.html
- Uncompress it

  `tar -xvf jdk-8-linux-i586.tar.gz` (32-bit)
  `tar -xvf jdk-8-linux-x64.tar.gz` (64-bit)

  The JDK 8 package is extracted into `./jdk1.8.0` directory. N.B.: Check carefully this folder name since Oracle seem to change this occasionally with each update.

- Now move the JDK 8 directory to `/usr/lib`
- `sudo mkdir -p /usr/lib/jvm`
- `sudo mv ./jdk1.8.0 /usr/lib/jvm/`

b. Set Java environment variable

Check where the Java files are installed or extracted and use that one for setting environment. Consider they are in "***/usr/lib/jvm"*** path then follow below instructions

vi /etc/environment

Paste the following line in that document

export JAVA_HOME=/usr/lib/jvm/

Press ESC key and type the following command to save the file in VI Editor
:wq!

vi /etc/profile

Paste the following line in that document

JAVA_HOME=/usr/lib/jvm/

PATH=$PATH:$HOME/bin:$JAVA_HOME/bin

export JAVA_HOME

export PATH

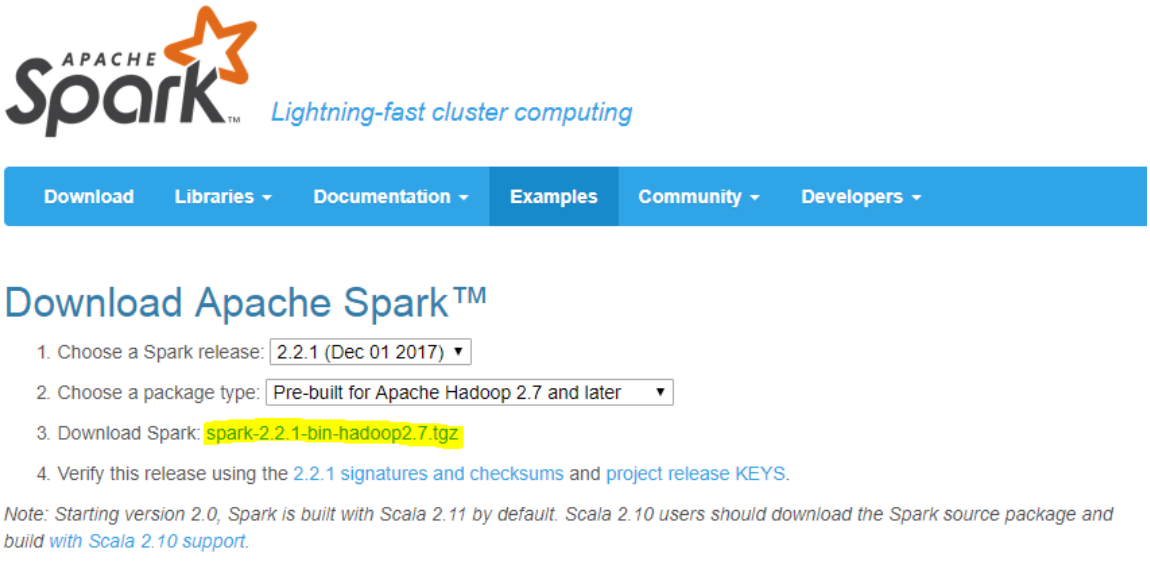
Press ESC key and type the following command to save the file in VI Editor
:wq!

6. Download the latest Spark version from

http://spark.apache.org/downloads.html

a. Select the Pre-built for Hadoop 2.7 and later

Download the .tgz file.



b. extract the file using the following command

tar zvfx spark-2.2.1-bin-hadoop2.7.tgz

c. get inside the extracted folder using the following command

cd spark-2.2.1-bin-hadoop2.7

d. Goto “sbin” folder and execute the following commands and check for **localhost:8080**

./stop-all.sh
./start-all.sh

If above command ran successfully, try *localhost:8080* in web browser it should open spark.

It was throwing the below error

localhost: ssh: connect to host localhost port 22: Connection refused

So I ran,

sudo apt-get install ssh

service ssh restart

After this the problem got resolved and the command ran successfully.

If the problem still persists then run below commands.

1. Remove SSH. Command:
```
sudo apt-get remove openssh-client openssh-server
```

2. Add SSH again. Command:
```
sudo apt-get install openssh-client openssh-server
```

Then the problem should be solved.

d. Get back to spark-2.2.1-bin-hadoop2.7 folder and execute following command
./bin/pyspark

7. To remove logging in pyspark, execute the following command.

a. Get into “conf” folder
cp log4j.properties.template log4j.properties

vi log4j.properties

Change
log4j.rootCategory = INFO, console
to

```
log4j.rootCategory = ERROR, console
```