

Instalación Pyspark en Linux

[Instalación Pyspark en Linux](#)

[Pasos previos](#)

[Instalación de spark](#)

[Instalación jupyter notebook](#)

1. Pasos previos

1.- Abrimos una shell:

2.- Instalamos openjdk-8-jdk de la siguiente forma:

sudo apt install openjdk-8-jdk

```
0-0ubuntu1 [47,8 kB]
Des:8 http://archive.ubuntu.com/ubuntu focal/main amd64 libpthread-stubs0-dev am
d64 0.4-1 [5.384 B]
Des:9 http://archive.ubuntu.com/ubuntu focal/main amd64 libsm-dev amd64 2:1.2.3-
1 [17,0 kB]
Des:10 http://archive.ubuntu.com/ubuntu focal/main amd64 libxau-dev amd64 1:1.0.
9-0ubuntu1 [9.552 B]
Des:11 http://archive.ubuntu.com/ubuntu focal/main amd64 libxdmcp-dev amd64 1:1.
1.3-0ubuntu1 [25,3 kB]
Des:12 http://archive.ubuntu.com/ubuntu focal/main amd64 xtrans-dev all 1.4.0-1
[68,9 kB]
Des:13 http://archive.ubuntu.com/ubuntu focal/main amd64 libxcb1-dev amd64 1.14-
2 [80,5 kB]
Des:14 http://archive.ubuntu.com/ubuntu focal-updates/main amd64 libx11-dev amd6
4 2:1.6.9-2ubuntu1.1 [649 kB]
Des:15 http://archive.ubuntu.com/ubuntu focal/main amd64 libxt-dev amd64 1:1.1.5
-1 [395 kB]
Des:16 http://archive.ubuntu.com/ubuntu focal-updates/universe amd64 openjdk-8-j
re-headless amd64 8u282-b08-0ubuntu1~20.04 [28,2 MB]
49% [16 openjdk-8-jre-headless 13,9 MB/28,2 MB 49%] 1.368 kB/s 18s
```

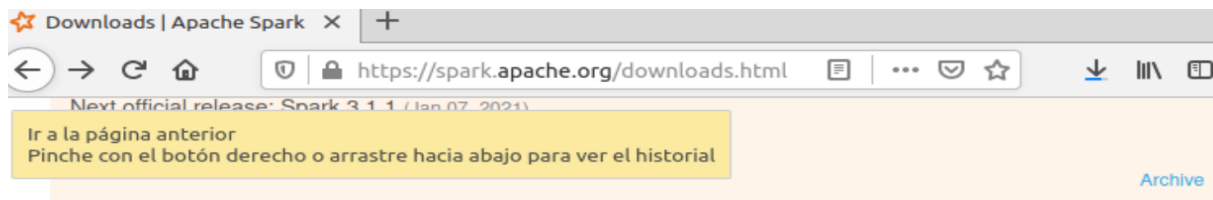
3.- Comprobamos que se ha instalado correctamente

java -version

```
fede@fede-VirtualBox:~$ java -version
openjdk version "11.0.9.1" 2020-11-04
OpenJDK Runtime Environment (build 11.0.9.1+1-Ubuntu-0ubuntu1.20.04)
OpenJDK 64-Bit Server VM (build 11.0.9.1+1-Ubuntu-0ubuntu1.20.04, mixed mode)
fede@fede-VirtualBox:~$
```

4.- Descarga Apache spark de la web oficial:

<https://spark.apache.org/downloads.html>



Download Apache Spark™

1. Choose a Spark release: **3.1.1 (Mar 02 2021)**
2. Choose a package type: **Pre-built for Apache Hadoop 2.7**
3. Download Spark: [spark-3.1.1-bin-hadoop2.7.tgz](#)
4. Verify this release using the 3.1.1 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.

5. Modificamos la variable de entorno

```
sudo vim /etc/environment
```

o

```
nano /etc/environment
```

5.b Si no tenemos vim instalado, podemos instalarlo

```
sudo apt install vim
```

Añadimos una nueva línea en el documento

```
JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64" (o la version que tengamos)
```

Salir y guardar

```
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
GNU nano 4.8 /etc/environment
PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin"
JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"
```

Ejecutamos:

```
source /etc/environment
```

Para que se ejecuten los cambios y comprobamos que ha ido bien ejecutando:

```
echo $JAVA_HOME
```

```
fede@fede-VirtualBox:~/Escritorio$ echo $JAVA_HOME
/usr/lib/jvm/java-8-openjdk-amd64
fede@fede-VirtualBox:~/Escritorio$
```

Una vez tenemos esto listo, vamos a instalar spark

2. Instalación de spark

1 Nos situamos en la carpeta donde hayamos descargado el comprimido

```
fede@fede-VirtualBox:~/Descargas$ ls
spark-3.1.1-bin-hadoop2.7.tgz
fede@fede-VirtualBox:~/Descargas$
```

Descomprimos:

```
sudo tar -zxvf nuestroArchivo.tgz
```

```
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-jakarta.activation-api.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-automaton.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-javax.transaction-api.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-jaxb-runtime.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-minlog.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-mustache.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-xmlenc.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-jline.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-istack-commons-runtime.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-py4j.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-vis-timeline.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-re2j.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-kryo.txt
spark-3.1.1-bin-hadoop2.7/licenses/LICENSE-cloudpickle.txt
fede@fede-VirtualBox:~/Descargas$
```

Abrimos

```
vim ~/.bashrc
```

y pegamos:

```
export SPARK_HOME=~/.Downloads/spark-3.1.1-bin-hadoop2.7
export PATH=$PATH:$SPARK_HOME/bin
export PATH=$PATH:~/anaconda3/bin
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
export PYSPARK_PYTHON=python3
export PATH=$PATH:$JAVA_HOME/jre/bin
```

```
export SPARK_HOME=~/.Downloads/NUESTRACARPETA
export PATH=$PATH:$SPARK_HOME/bin
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
#export PYSPARK_DRIVER_PYTHON="jupyter"
#export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
export PYSPARK_PYTHON=python3
export PATH=$PATH:$JAVA_HOME/jre/bin
```

```
Welcome to
      _ _ _ _ _
     / _ _ _ _ \   version 3.1.1
    / _ _ _ _ \
   / _ _ _ _ \

Using Python version 3.8.5 (default, Jul 28 2020 12:59:40)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1618391720)
SparkSession available as 'spark'.
>>>
```

Ya lo tendríamos

Ahora si queremos usar jupyter notebook

3. Instalación jupyter notebook

sudo apt install jupyter

o

Pip install jupyter

Debemos de descomentar las dos líneas:

```
#export PYSARK_DRIVER_PYTHON="jupyter"  
#export PYSARK_DRIVER_PYTHON_OPTS="notebook"
```

Y ya directamente cuando hagamos pyspark nos abrirá jupyter notebook y ya podremos trabajar ahí:

```
import pyspark  
from pyspark import SparkContext  
from pyspark.sql import SparkSession  
sc = SparkContext()
```

```
In [19]: import pyspark  
         from pyspark import SparkContext  
  
In [20]: from pyspark.sql import SparkSession  
  
In [24]: sc = SparkContext.getOrCreate();
```