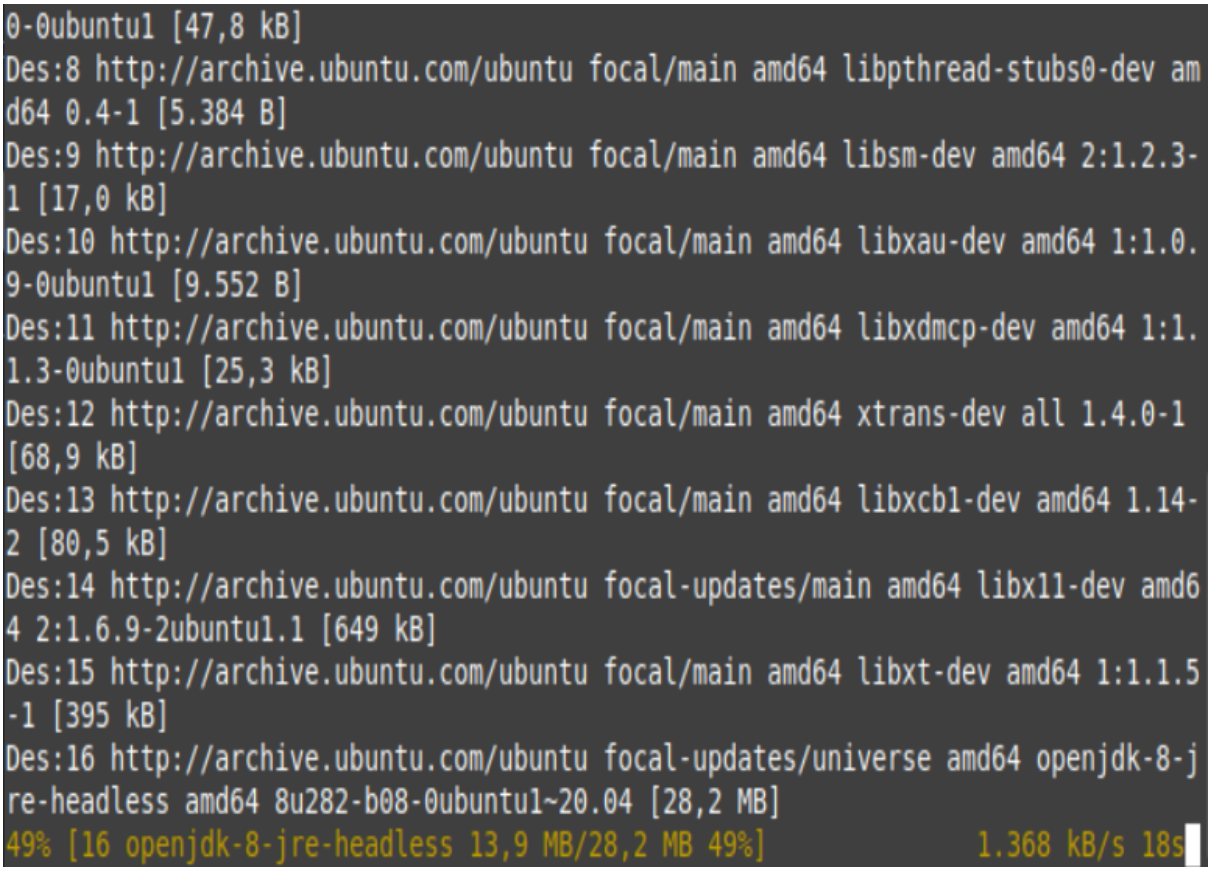
Instalación Pyspark en Linux

1. Pasos previos

1.- Abrimos una shell:

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

sudo apt install openjdk-8-jdk



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:

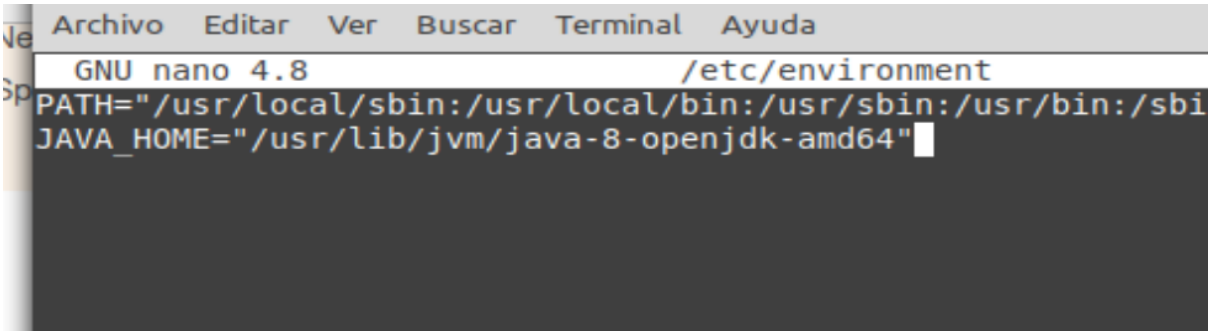
https://spark.apache.org/downloads.html

5. Modificamos la variable de entorno

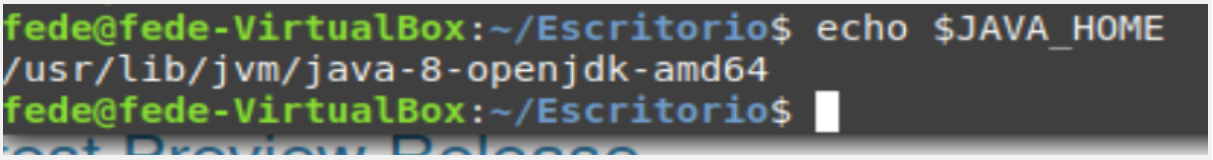
`sudo vim /etc/environment`
`o`
`nano /etc/environment`

5.b Si no tenemos vim instalado, podemos instalarlo
`sudo apt install vin`

Añadimos una nueva línea en el documento
`JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"(o la version que tengamos)`
Salir y guardar



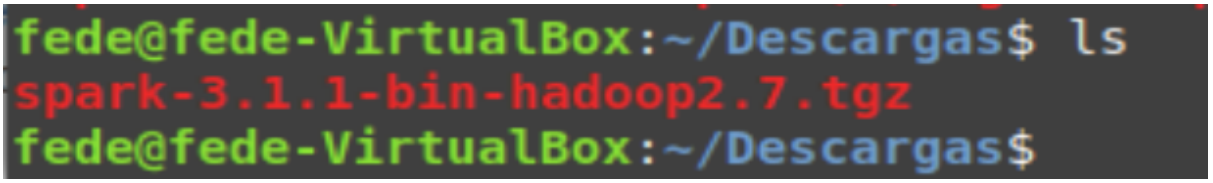
Ejecutamos:
`source /etc/environment`
Para que se ejecuten los cambios y comprobamos que ha ido bien ejecutando:
`echo $JAVA_HOME`



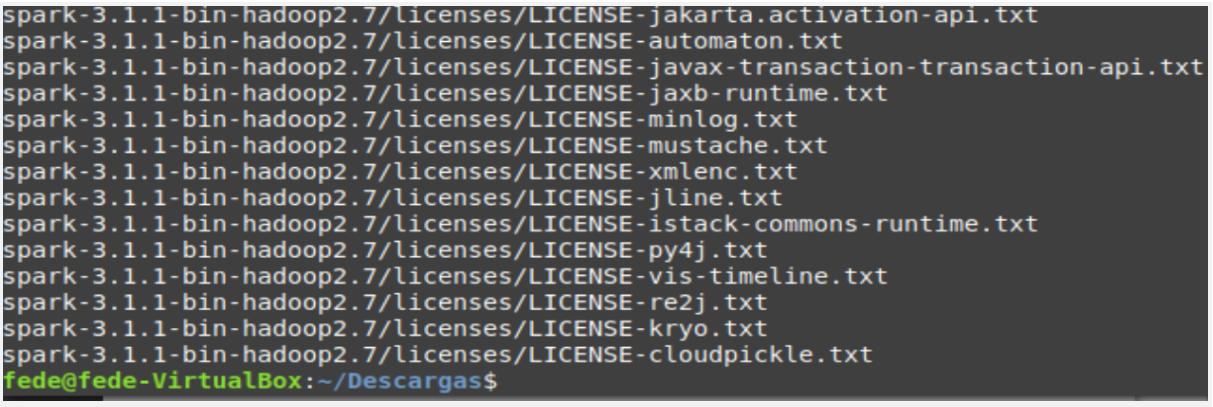
Una vez tenemos esto listo, vamos a instalar spark

2. Instalación de spark

1 Nos situamos en la carpeta donde hayamos descargado el comprimido



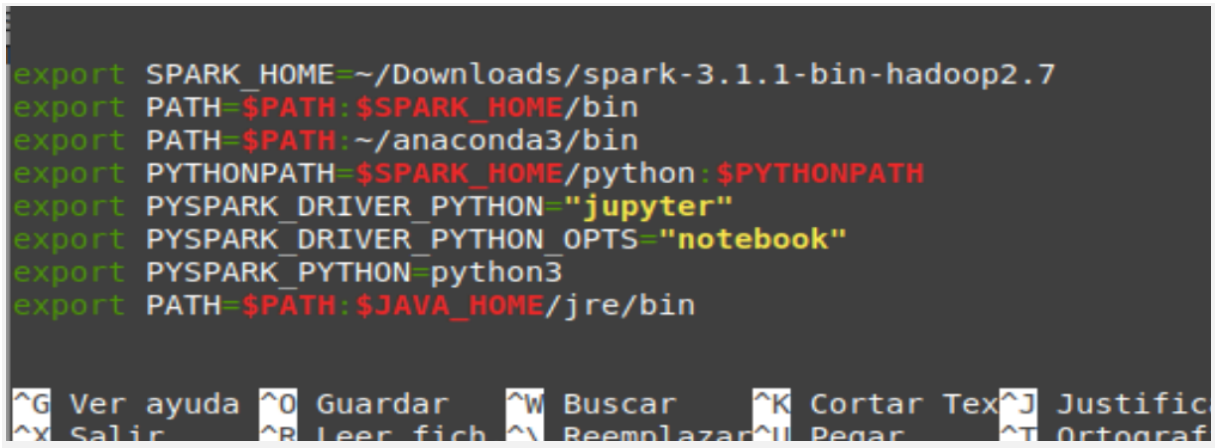
Descomprimimos:
`sudo tar -zxvf nuestroArchivo.tgz`



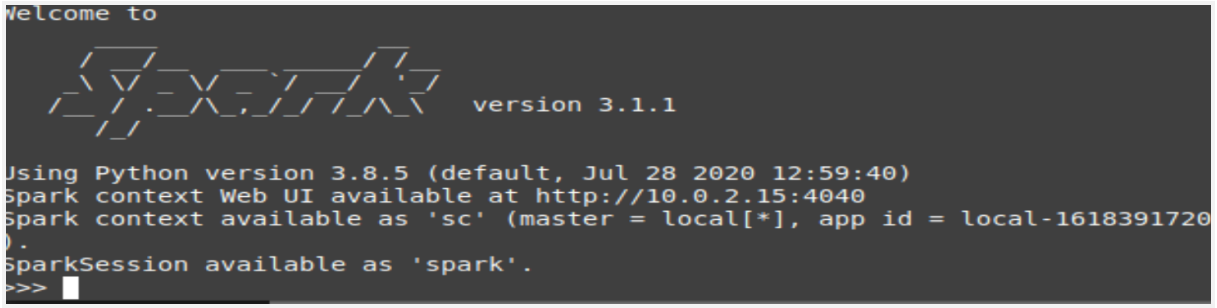
Abrimos

```
vim ~/.bashrc
```

y pegamos:




```
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
```



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 PYSPARK_DRIVER_PYTHON="jupyter"
#export PYSPARK_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();
```