

Fundamentos de manejo de datos con Python

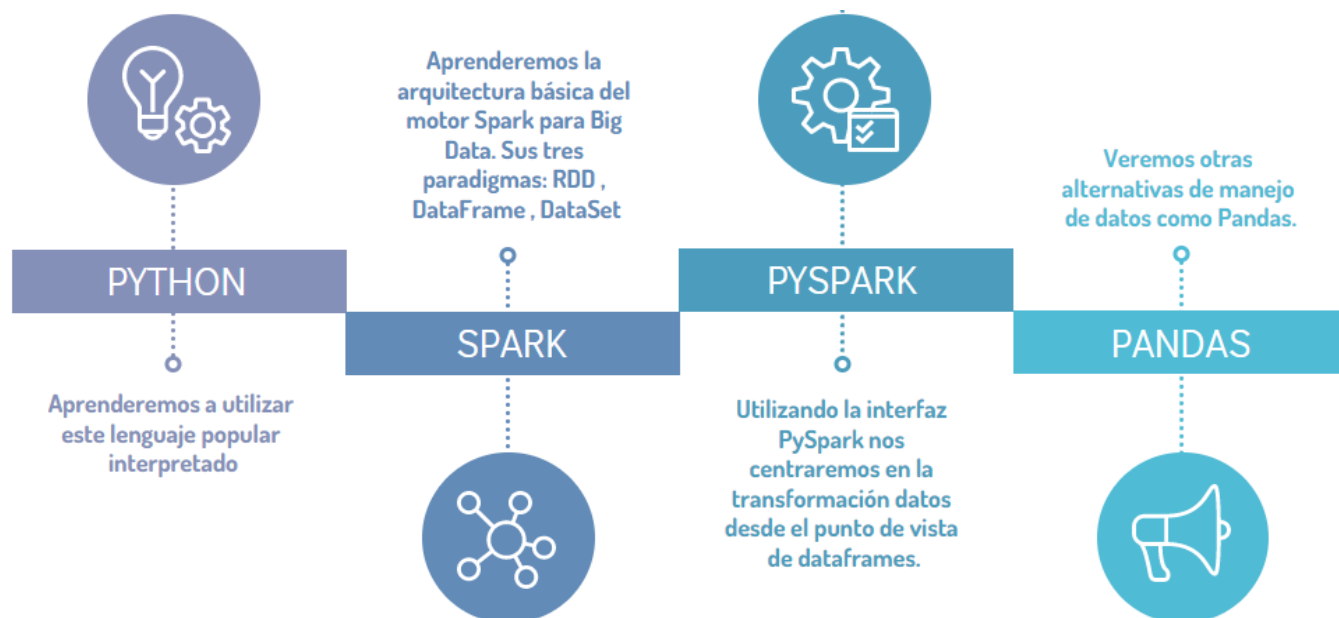
- Ambientación de herramientas de trabajo -

Linux

Autor: Marco Ycaza

Panorama general

Bienvenido al siguiente tutorial de fundamentos de manejo de datos , podrá observar a continuación la ruta propuesta de aprendizaje.



Pero para empezar con dicha ruta necesitamos tener ciertas herramientas.

A continuación listamos las siguientes:

N	Herramienta	Tipo	Versión recomendada	Usada por el autor
1	Java	lenguaje	8 en adelante	8_181
2	Python	lenguaje	3.5 en adelante	3.7.8
3	Spark	framework	2 en adelante	spark-2.4.4-bin-hadoop2.7

Sería recomendable que usted instale las mismas versiones con las cuales fue desarrollado este tutorial para evitar diferencias no contempladas en este tutorial.

Instalaciones de herramientas:

Instalación de Java:

Version 1:

```
sudo apt-get update

sudo apt-get install openjdk-8-jdk

java -version

/usr/lib/jvm/jdk1.8.0_version/bin/java
```

Version2:

Vamos a instalar el kit de desarrollo ya que este nos permitirá tener tanto el JRE (que incluye la virtual machine) y herramientas de compilación para desarrollo y testeo. De forma pura para correr Spark necesitamos solo el JRE pero conviene tener el **JDK** ⁽¹⁾ para posibles desarrollos.

Proceda a descargar la siguiente versión a través de:

[Java Archive Downloads - Java SE 8 \(oracle.com\)](#)

Precisamente en el enlace del apartado de "Java SE Development Kit" en el enlace de descarga nos llevara a una pagina de autenticación donde necesitaremos crear una cuenta oracle.

Java SE Development Kit 8u181

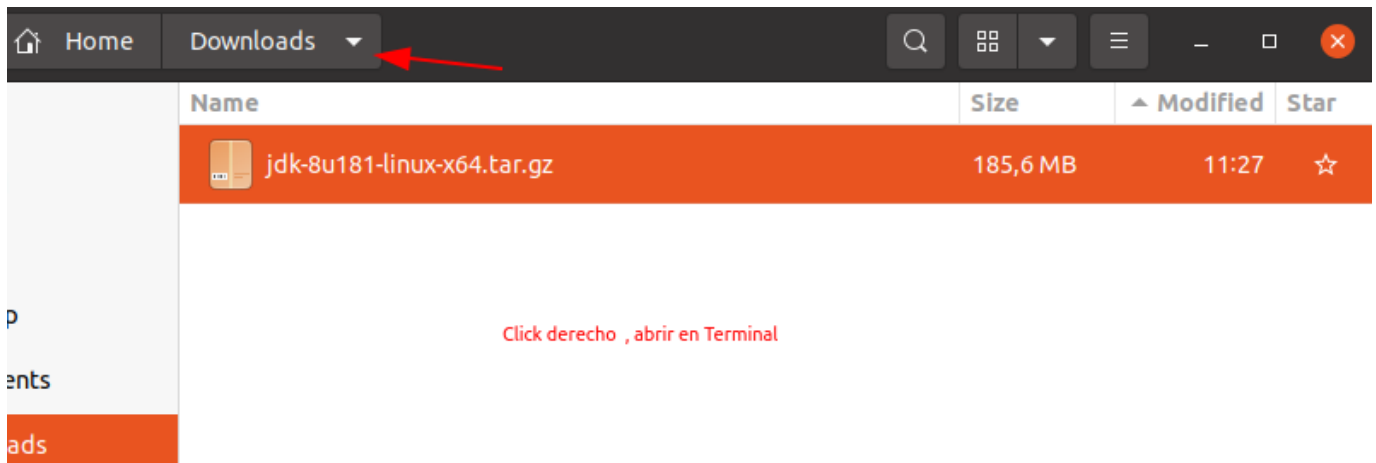
This software is licensed under the Oracle Binary Code License Agreement for Java SE Platform Products

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.95 MB	jdk-8u181-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.89 MB	jdk-8u181-linux-arm64-vfp-hflt.tar.gz
Linux x86	165.06 MB	jdk-8u181-linux-i586.rpm
Linux x86	179.87 MB	jdk-8u181-linux-i586.tar.gz

1. crear un folder llamado java dentro del directorio /usr/:

```
cd ~
sudo mkdir /usr/java
```

2. Abre una terminal en el folder donde esta ubicado tu archivo .tar.gz , en mi caso se encuentra en **Downloads**.



3. Luego movemos el archivo *****.tar.gz** hacia la carpeta **/usr/java**:

```
sudo mv jdk-8u181-linux-x64.tar.gz /usr/java
```

4. Nos movemos hacia esa carpeta y descomprimos la carpeta:

```
cd /usr/java
sudo tar zxvf jdk-8u181-linux-x64.tar.gz
```

(Nota) Puede encontrar mas información en : https://www.java.com/en/download/help/linux_install.html

5. Luego modificar el archivo **.bashrc** (4) que se encuentra en **cd \$HOME** con el editor de su preferencia y adicionar:

```
JAVA_HOME=/usr/java/jdk1.8.0_181/bin

PATH=$PATH:$JAVA_HOME
```

No olvidar actualizar el archivo .bashrc usando

```
source $HOME/.bashrc
```

Instalación de Python:

1. Corremos los siguientes comandos antes de proceder con la instalación!

```
sudo apt-get install zlib1g-dev

sudo apt install build-essential zlib1g-dev libncurses5-dev libgdbm-dev libnss3-dev libssl-dev
libreadline-dev libffi-dev wget

sudo apt install build-essential zlib1g-dev libncurses5-dev libgdbm-dev libnss3-dev libssl-dev
libreadline-dev libffi-dev wget

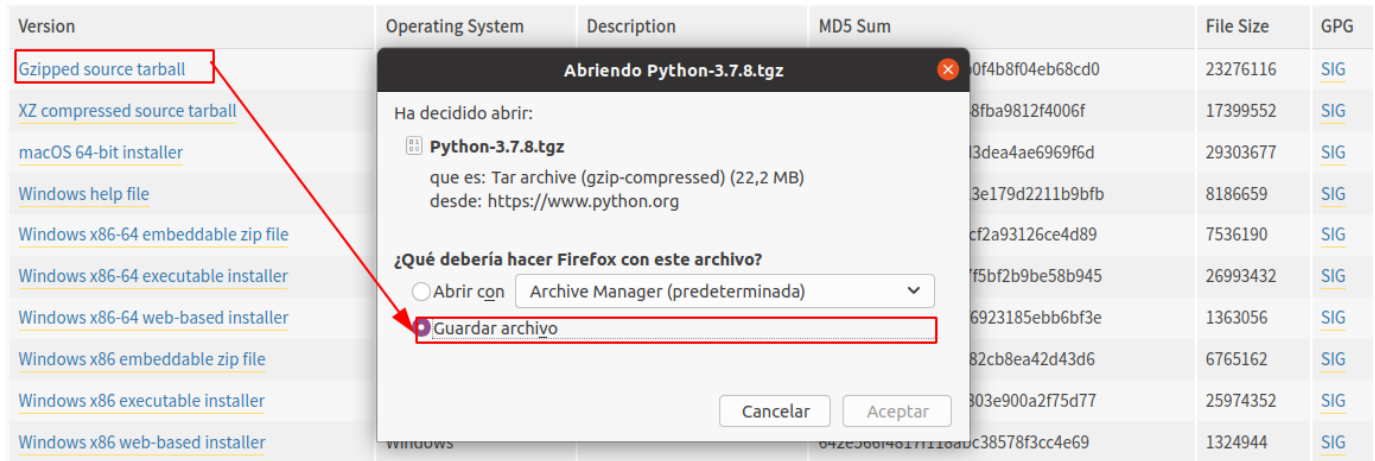
sudo apt-get install libsqlite3-dev
```

Luego:

Python es un lenguaje interpretado de simple sintaxis que nos permitirá usar el motor Spark de forma rápida. También será utilizado con Pandas al final del curso de fundamentos. El enlace de descarga es el siguiente:

[Python Release Python 3.7.8 | Python.org](https://www.python.org/downloads/release/python-378/)

2. En este caso descargaremos el tarball para linux.



3. Nos dirigimos a la carpeta Downloads y abrimos una terminal.

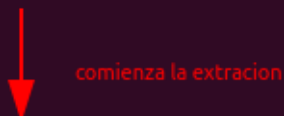
```
marco@marco:~/Downloads$ ls -lh
total 200M
drwxr-xr-x 7 uucp    143 4,0K Kun   7  2018 jdk1.8.0_181
-rw-rw-r-- 1 marco marco 178M Kup   13 11:27 jdk-8u181-linux-x64.tar.gz
-rw-rw-r-- 1 marco marco  23M Kup   13 12:17 Python-3.7.8.tgz
```

4. Luego ejecutamos (similar a la instalacion de java):

```
sudo mkdir /usr/python
sudo mv Python-3.7.8.tgz /usr/python
cd /usr/python
sudo tar zxvf Python-3.7.8.tgz
```

(ejemplo)

```
marco@marco:~/Downloads$ sudo mkdir /usr/python
[sudo] password for marco:
marco@marco:~/Downloads$ sudo mv Python-3.7.8.tgz /usr/python
marco@marco:~/Downloads$ cd /usr/python
marco@marco:/usr/python$ sudo tar zxvf Python-3.7.8.tgz
Python-3.7.8/
Python-3.7.8/Doc/
Python-3.7.8/Doc/c-api/
Python-3.7.8/Doc/c-api/sys.rst
Python-3.7.8/Doc/c-api/conversion.rst
```



5. Luego ejecutamos el archivo "configure"

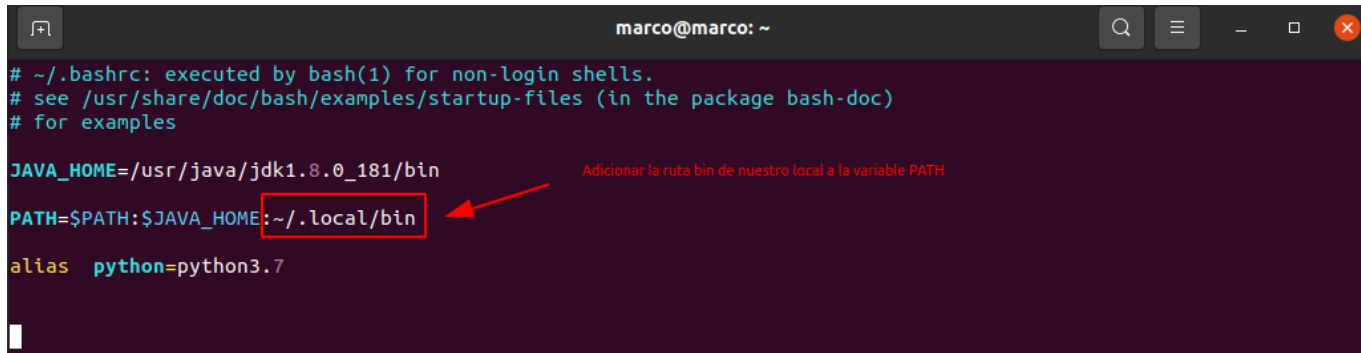
```
sudo ./configure
sudo make altinstall
```

6. En nuestro archivo .bashrc:

Adicionamos **~/local/bin** a nuestra variable **PATH** usando el separador :

Si tenemos otras versiones de python y queremos cambiar el comando "python" para que apunte a lo que hemos instalado creamos un alias en el archivo .bashrc usando **alias**

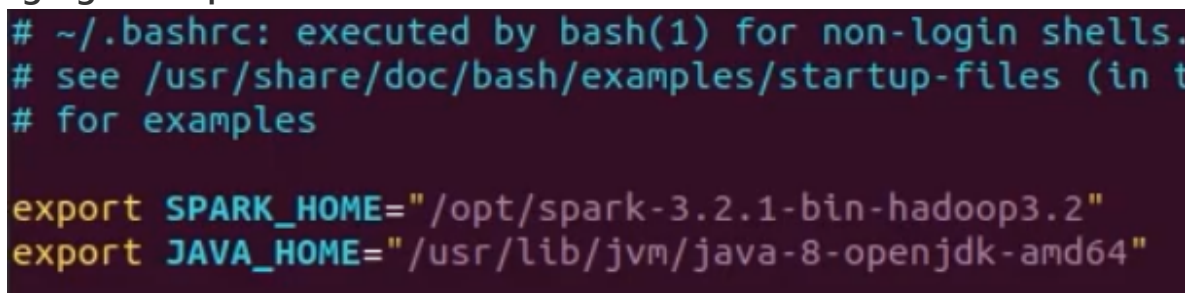
python=python3.7



```
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

JAVA_HOME=/usr/java/jdk1.8.0_181/bin
PATH=$PATH:$JAVA_HOME:~/local/bin
alias python=python3.7
```

Si usted ya cuenta con Python 3.8 y java instalado a través de open jdk , guiese de esta imagen. Notar que no es necesario agregar python porque por defecto viene agregado al path.



```
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

export SPARK_HOME="/opt/spark-3.2.1-bin-hadoop3.2"
export JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"
```

No olvidar actualizar el archivo .bashrc !!!

```
source $HOME/.bashrc
```

instalamos ciertos paquetes importantes como ipython y jupyter:

```
sudo -H pip3.7 install ipython
sudo -H pip3.7 install jupyter
```

Instalación de Spark con Hadoop:

Dentro de la página oficial de Apache Spark , vamos buscando el apartado de releases antiguos encontramos el siguiente recurso:

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

Download Apache Spark™

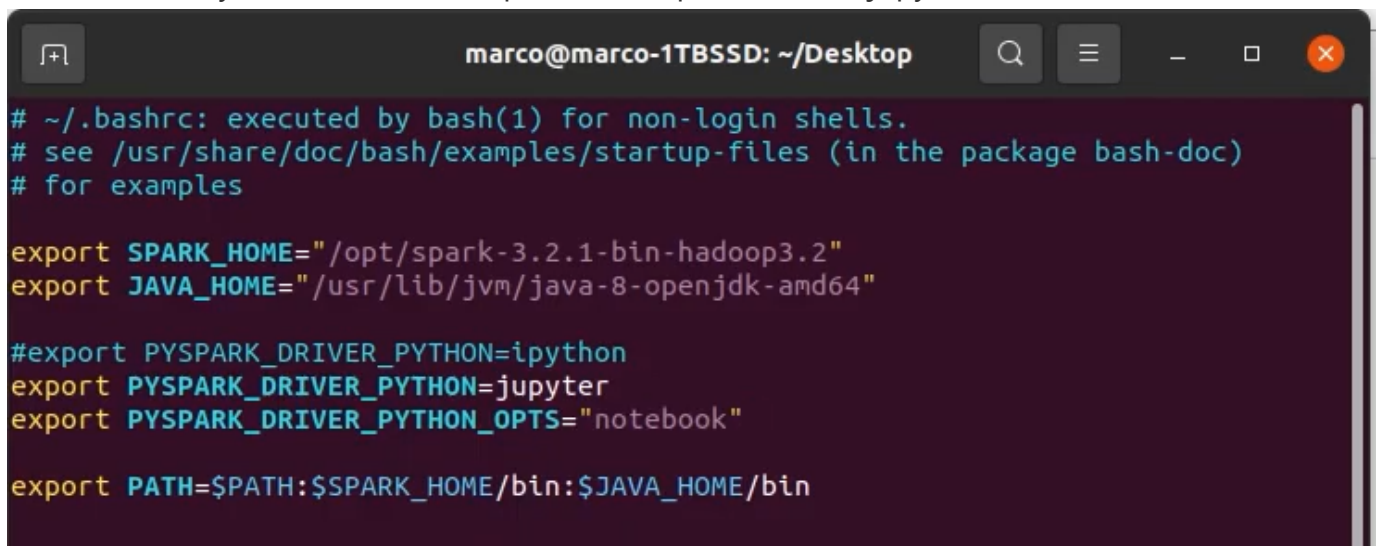
1. Choose a Spark release: **3.2.1 (Jan 26 2022)** ▾
2. Choose a package type: **Pre-built for Apache Hadoop 3.3 and later** ▾
3. Download Spark: **[spark-3.2.1-bin-hadoop3.2.tgz](#)** ← Descargamos la última versión
4. Verify this release using the 3.2.1 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that Spark 3 is pre-built with Scala 2.12 in general and Spark 3.2+ provides additional pre-built distribution with Scala 2.13.

Configuración de variables de entorno:

A continuación puede observar una línea comentada: "**export**

PYSPARK_DRIVER_PYTHON=ipython", si desea abrir spark usando esta interfaz, descomentarla y comentar las dos que vienen (que refieren a jupyter notebook).



```
marco@marco-1TBSSD: ~/Desktop
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

export SPARK_HOME="/opt/spark-3.2.1-bin-hadoop3.2"
export JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"

#export PYSPARK_DRIVER_PYTHON=ipython
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"

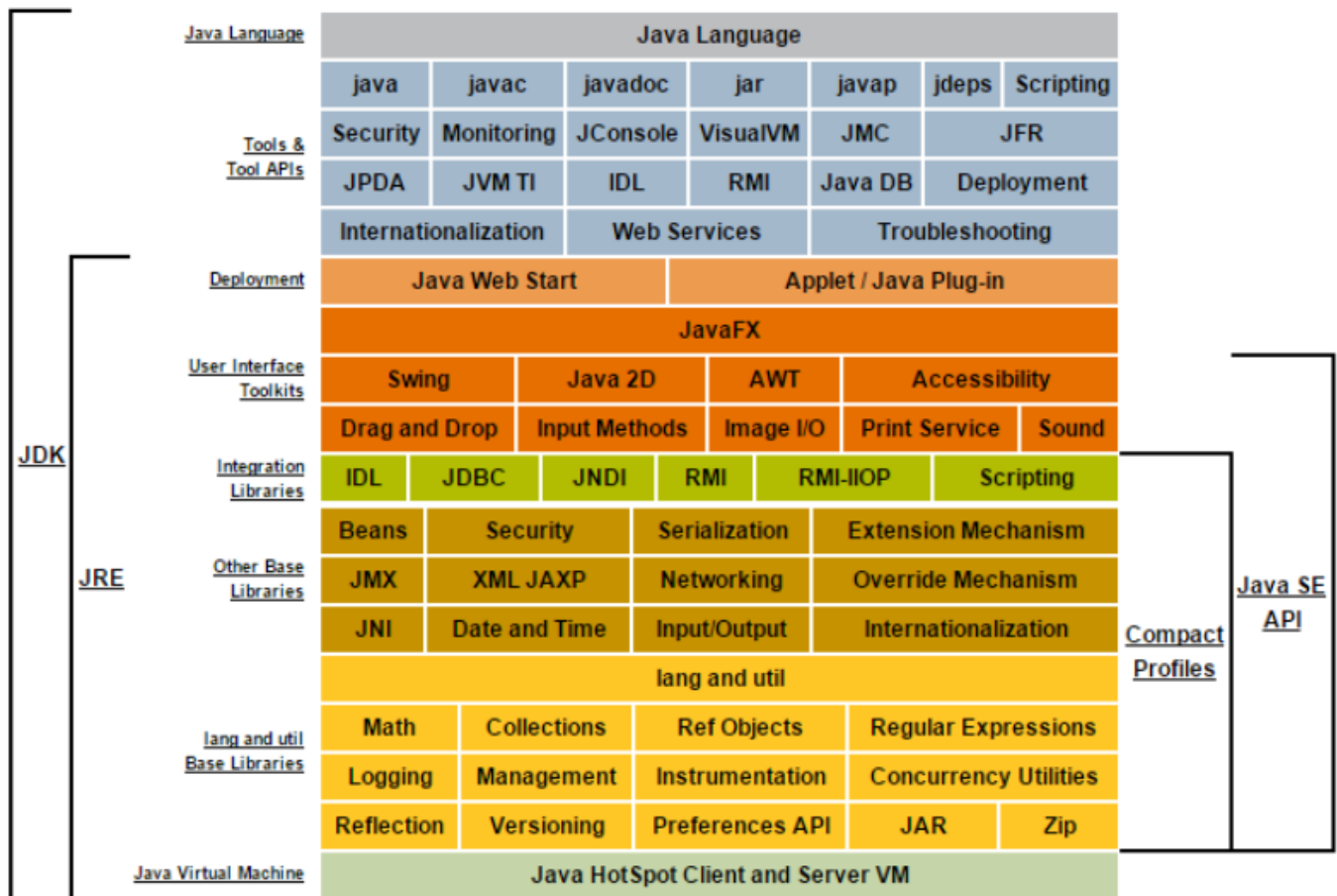
export PATH=$PATH:$SPARK_HOME/bin:$JAVA_HOME/bin
```

Si desea ayudarse con un video, tengo un [Video Tutorial](#) preparado para usted.

ANEXO

(1) Diagrama conceptual del JDK

Description of Java Conceptual Diagram



(3) Para poder usar pyspark en un kernel de python se tiene que tener acceso a la máquina virtual de java (JVM) para ello debemos definir la variable PYTHONPATH la cual nos permite adicionar directorios externos en donde se buscara módulos y paquetes. Justamente el módulo externo que queremos que python pueda acceder es Py4J. Para poder configurar correctamente esta variable revisemos la ruta %SPARK_HOME%/python/lib y veamos qué version de Py4J tenemos:

```
Directorio de C:\spark\spark-2.4.4-bin-hadoop2.7\python\lib
09/02/2022 10:37 a. m. <DIR> .
09/02/2022 10:37 a. m. <DIR> ..
29/12/2021 06:19 p. m. 42.437 py4j-0.10.7-src.zip
29/12/2021 06:19 p. m. 1.445 PY4J_LICENSE.txt
29/12/2021 06:19 p. m. 591.770 pyspark.zip
3 archivos 635.652 bytes
2 dirs 135.864.016.896 bytes libres
```

Debería salir 0.10.7 si es que has descargado spark 2.4.4 con hadoop 2.7

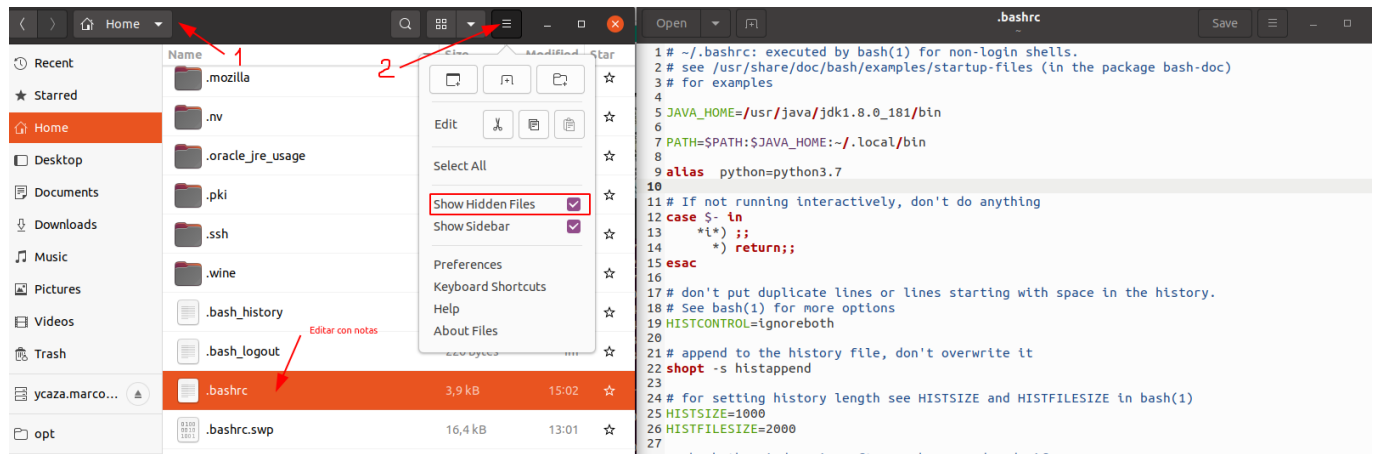
(4) Que es el archivo .bashrc ?

Los archivos BASHRC son archivos de intérprete de comandos. Este archivo se utiliza para establecer las preferencias de las solicitudes de comandos del usuario o las rutas comunes a los directorios o programas ejecutables.

Notar que este archivo se ejecuta automaticamente al invocar la shell. Notar el "." antes de bashrc que da a entender que este archivo debe ser ejecutado como comando (o conjunto de comandos).

Donde se encuentra el archivo?

Se encuentra en **HOME**



Al realizar cualquier modificación hay que abrir una terminal y cargar los cambios:
Esto lo hacemos con el comando **source** sobre el archivo bashrc.

