Repo: https://github.com/cpalacios01/spark-hadoop-local

Spark-Hadoop Local

Requisito:
jdk1.7.0_67 o superior (jdk1.8.0_361 por ej.)

Ingresar a: https://archive.apache.org/dist/spark/spark-3.3.2/
Descargar y descomprimir: **spark-3.3.2-bin-hadoop2.tgz**
Ejemplo en **C:\spark-3.3.2-bin-hadoop2**

En caso sea un entorno Windows
https://github.com/steveloughran/winutils
Obtener la versión apropiada de winutils y copiarlo en:
C:\spark-3.3.2-bin-hadoop2\bin

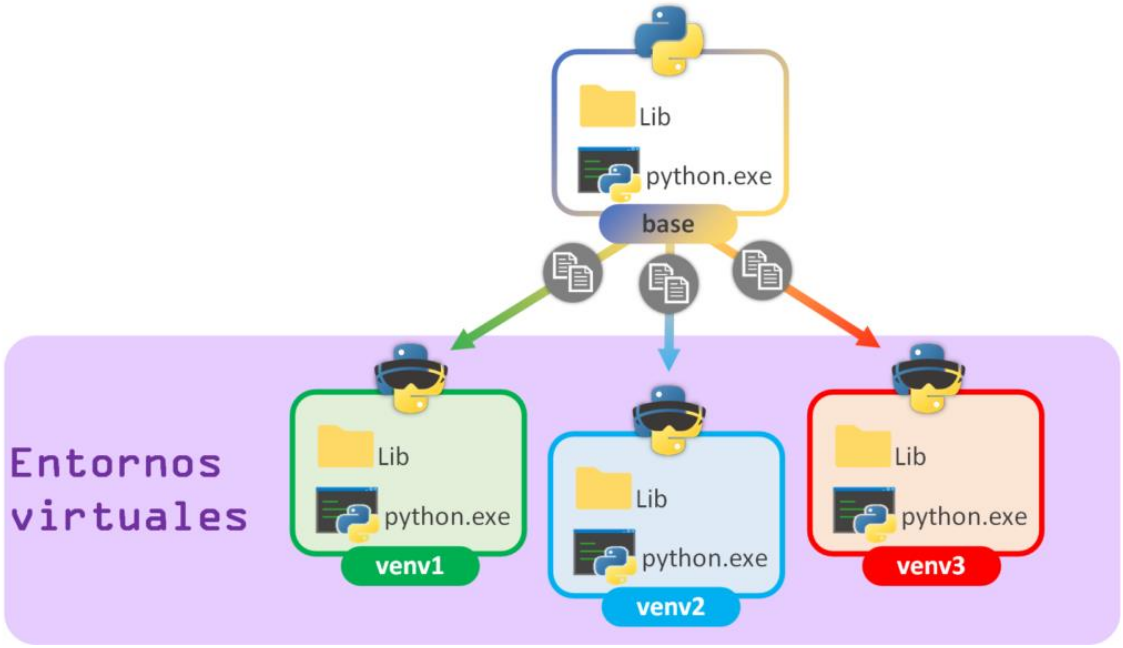
Añadir las variables de entorno
JAVA= C:\Program Files\Java\jdk1.8.0_361
PYSPARK_DRIVER_PYTHON= jupyter
PYSPARK_DRIVER_PYTHON_OPTS=notebook
SPARK_HOME= C:\spark-3.3.2-bin-hadoop2
HADOOP_HOME= C:\spark-3.3.2-bin-hadoop2

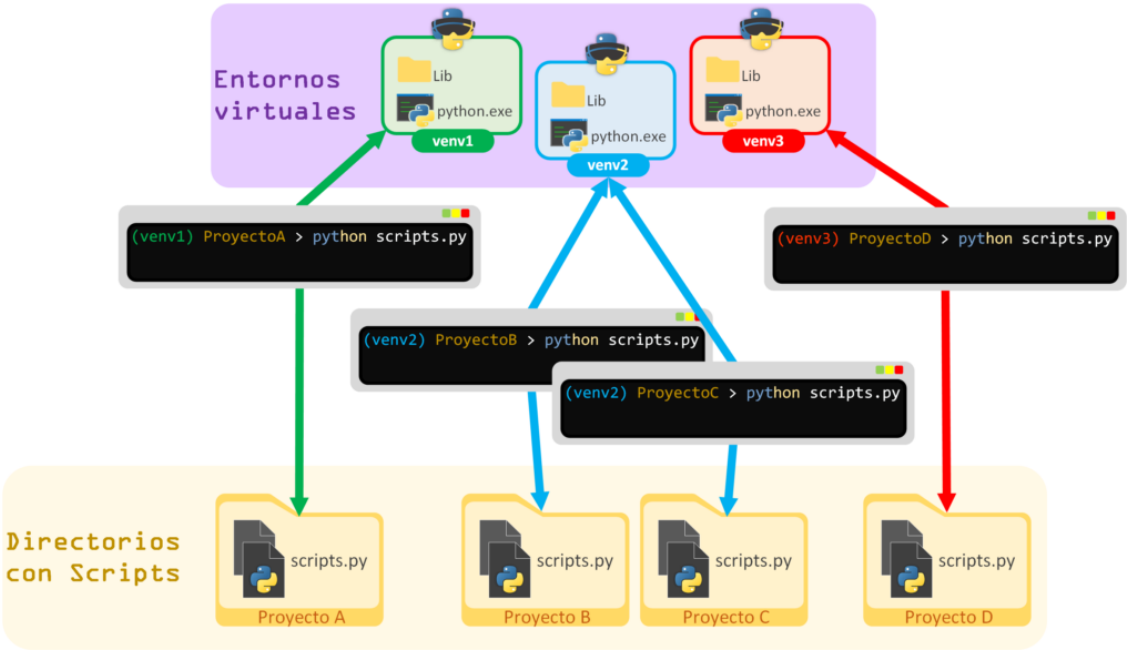
AGREGAR AL PATH
C:\spark-3.3.2-bin-hadoop2\bin

Anaconda-Conda

Asumiendo que tenemos Anaconda Instalado
Verificaciones varias

- Python
  ```
  python --version
  ```
- Pip (instalador estándar de librerías de Python)
  ```
  pip --version
  ```
- Jupyter Notebook
  ```
  jupyter notebook --version
  ```
- Conda (administrador de librerías, viene con Anaconda)
  ```
  conda --version
  ```





Ver listado de entornos existentes:

```
conda env list
```

Crear un entorno para BigData:

```
conda create -n <nombre de ambiente>
conda create -n <nombre de ambiente> python=x.x
```

Nota: la creación por consola, lo crea de forma vacía.
SE RECOMIENDA LA CREACIÓN DE ENTORNO CON ANACONDA, para que se incluyan las librerías base.

Activar entorno

```
conda activate BigData2023
```

```
C:\Users\ACER>conda activate BigData2023

(BigData2023) C:\Users\ACER>
```

Eliminar entorno (para cuando lo necesiten)

```
conda env remove -n nombre_entorno
conda remove --name nombre_entorno --all
```

Nota: Se debe estar fuera del entorno para poder eliminarlo

Ver lista de librerías dentro del entorno

```
conda list -n BigData2023
```

```
C:\Users\ACER>conda activate BigData2023

(BigData2023) C:\Users\ACER>conda list -n BigData2023
# packages in environment at C:\Users\ACER\anaconda3\envs\BigData2023:
#
# Name                    Version                   Build  Channel
bzip2                     1.0.8                he774522_0
ca-certificates           2023.01.10           haa95532_0
libffi                    3.4.2                hd77b12b_6
openssl                   1.1.1t               h2bbff1b_0
pip                       23.0.1          py310haa95532_0
python                    3.10.10              h966fe2a_2
setuptools                66.0.0          py310haa95532_0
sqlite                    3.41.2               h2bbff1b_0
tk                        8.6.12               h2bbff1b_0
tzdata                    2023c                h04d1e81_0
vc                        14.2                 h21ff451_1
vs2015_runtime            14.27.29016          h5e58377_2
wheel                     0.38.4          py310haa95532_0
xz                        5.2.10               h8cc25b3_1
zlib                      1.2.13               h8cc25b3_0

(BigData2023) C:\Users\ACER>
```

FindSpark

Instalar librería necesaria

```
python -m pip install findspark
```

```
(BigData2023) C:\Users\ACER>python -m pip install findspark
Collecting findspark
  Using cached findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Installing collected packages: findspark
Successfully installed findspark-2.0.1

(BigData2023) C:\Users\ACER>
```

Verificar lista librerías dentro del entorno

```
conda list -n BigData2023
```

```
(BigData2023) C:\Users\ACER>conda list -n BigData2023
# packages in environment at C:\Users\ACER\anaconda3\envs\BigData2023:
#
# Name                    Version                   Build  Channel
bzip2                     1.0.8                he774522_0
ca-certificates           2023.01.10           haa95532_0
findspark                 2.0.1                    pypi_0    pypi
libffi                    3.4.2                hd77b12b_6
openssl                   1.1.1t               h2bbff1b_0
pip                       23.0.1           py310haa95532_0
python                    3.10.10              h966fe2a_2
setuptools                66.0.0           py310haa95532_0
sqlite                    3.41.2               h2bbff1b_0
tk                        8.6.12               h2bbff1b_0
tzdata                    2023c                h04d1e81_0
vc                        14.2                 h21ff451_1
vs2015_runtime            14.27.29016          h5e58377_2
wheel                     0.38.4           py310haa95532_0
xz                        5.2.10               h8cc25b3_1
zlib                      1.2.13               h8cc25b3_0

(BigData2023) C:\Users\ACER>
```

Lanzar Jupyter Notebook

```
jupyter notebook
```

Crear un directorio de trabajo, descargar y descomprimir allí el archivo: base_datos_2022.rar

https://drive.google.com/drive/folders/1edB06b57ElfRDhWaaZKQVs9pCxJVoQr8?usp=share_link