

# Como instalar Jupyter Notebook en sus ambientes de Hadoop

Ejecutar el docker run pero abriendo el puerto 8889

```
docker run --name edvai_hadoop -p 8081:8081 -p 8080:8080 -p 8088:8088 -p 8889:8889 -p 9870:9870 -p 9868:9868 -p 9864:9864 -p 1527:1527 -p 10000:10000 -p 10002:10002 -p 8010:8010 -p 9093:9093 -p 2181:2182 -it fedepineyro/edvai_ubuntu:v6 /bin/bash -c "/home/hadoop/scripts/start-services.sh"
```

En edvai\_hadoop instalar jupyter:

```
pip install notebook
```

Luego crear en edvai\_hadoop el siguiente scripta en /home/hadoop/scripts:

```
cat > pyspark_jupyter.sh
```

```
#!/bin/bash

# Define env Python to use
export PYSARK_PYTHON=/usr/bin/python3

# Define IPython driver
export PYSARK_DRIVER_PYTHON='jupyter'

# Define Spark conf to use
export PYSARK_DRIVER_PYTHON_OPTS='notebook --ip=172.17.0.2 --port=8889'

/home/hadoop/spark/bin/pyspark
```

Cambiar los permisos del script:

```
chmod 777 /home/hadoop/scripts/pyspark_jupyter.sh
```

Ejecutar el script:

```
./home/hadoop/scripts/pyspark_jupyter.sh
```

Al ejecutar el script nos dara la ruta al cual debemos acceder:

```
[I 2024-05-12 10:29:20.189 ServerApp] Serving notebooks from local directory: /home/hadoop/scripts
[I 2024-05-12 10:29:20.189 ServerApp] Jupyter Server 2.14.0 is running at:
[I 2024-05-12 10:29:20.189 ServerApp] http://172.17.0.2:8889/tree?token=1ade7e5fc5b5522e7326e2c499fce838ce5f1885df55c1b6
[I 2024-05-12 10:29:20.190 ServerApp] http://127.0.0.1:8889/tree?token=1ade7e5fc5b5522e7326e2c499fce838ce5f1885df55c1b6
[I 2024-05-12 10:29:20.190 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[W 2024-05-12 10:29:20.196 ServerApp] No web browser found: Error('could not locate runnable browser').
[C 2024-05-12 10:29:20.197 ServerApp]

To access the server, open this file in a browser:
file:///home/hadoop/.local/share/jupyter/runtime/jpserver-9966-open.html
Or copy and paste one of these URLs:
http://172.17.0.2:8889/tree?token=1ade7e5fc5b5522e7326e2c499fce838ce5f1885df55c1b6
http://127.0.0.1:8889/tree?token=1ade7e5fc5b5522e7326e2c499fce838ce5f1885df55c1b6
[I 2024-05-12 10:29:20.215 ServerApp] Skipped non-installed server(s): bash-language-server, dockerfile-language-server-nodejs, javascript-language-server, julia-language-server, pyright, python-language-server, python-lsp-server, r-languageserver, sql-language-server, texlab, typescript-language-server
```

Ejemplo:

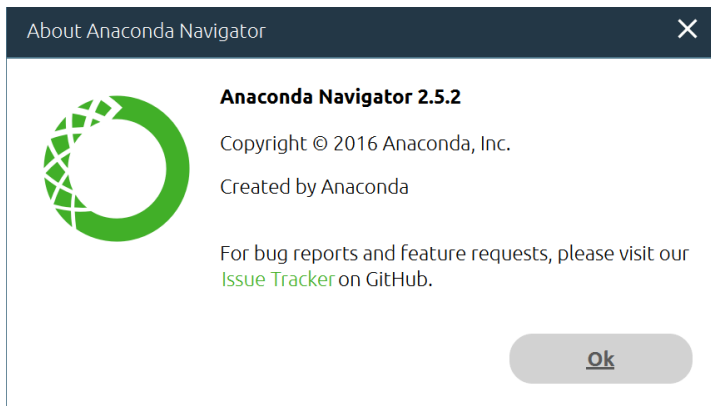
<http://127.0.0.1:8889/tree?token=1ade7e5fc5b5522e7326e2c499fce838ce5f1885df55c1b6>

Debemos ingresar siempre a la ruta que comienza con <http://127.0.0.1:8889/tree>.....

Podemos acceder con Ctrl y seleccionando esa ruta o la podemos copiar y pegar en el navegador.

Otra opción es siempre ingresar a: <http://127.0.0.1:8889/tree>

Debemos tener instalado Anaconda, en mi caso no actualicé el navegador:



Al abrir <http://127.0.0.1:8889/tree> veremos algo similar a esto:

127.0.0.1:8889/tree

jupyter

File View Settings Help

Files Running

Open Download Rename Duplicate Delete New Upload

/

Name	Last Modified	File Size
bin	2 years ago	
boot	4 years ago	
dev	7 minutes ago	
etc	7 minutes ago	
home	2 years ago	
lib	2 years ago	
lib32	2 years ago	
lib64	2 years ago	
libx32	2 years ago	
media	2 years ago	
mnt	2 years ago	
opt	2 years ago	
proc	7 minutes ago	
run	6 minutes ago	
sbin	2 years ago	
srv	2 years ago	
sys	7 minutes ago	
tmp	48 seconds ago	
usr	2 years ago	
var	2 years ago	

Debemos ir ingresando a través de las carpetas a /home/hadoop

Allí crearemos una carpeta llamada Notebooks, haciendo clic en nuevo y luego -> Nueva carpeta:

127.0.0.1:8889/tree/home/hadoop

jupyter

File View Settings Help

Files Running

Rename Delete

/ home / hadoop /

Name	Last Modified	File Size
airflow		
hadoop		
hadoopdata		
hive		

New Upload

Python 3 (ipykernel)

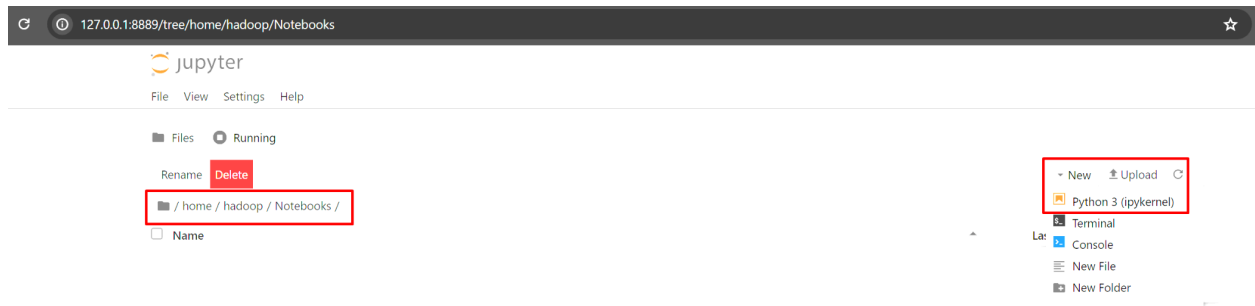
Terminal

Console

New File

New Folder

Una vez dentro de esa carpeta que hemos creado (/home/hadoop/notebooks) Debemos seleccionar New y luego Python 3



Al iniciar una nueva notebook debemos crear una SparkSession de la siguiente manera:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
    .master("spark://localhost:7077") \
    .getOrCreate()
```

Y luego continuamos trabajando en la notebook con Pyspark:

