Guía de ejecución (Proyecto Final)

Estudiante: Carlos Felipe Marin Sandi
Curso: Big Data

1. Clonar repositorio

Primero, clonamos el repositorio desde el docker Desktop con el siguiente comando:

```
git clone https://github.com/usuario/Proyecto_Final_CFMS.git
```

En caso de no poder correr la línea del código, descargar directamente desde el repositorio, y guardarlo en su directorio de preferencia.

2.Construir la imagen

Construimos una imagen Docker desde el Dockerfile del directorio actual (.) y la etiqueta como bigdata. (Tiene que estar dentro del directorio de la carpeta 'bigdataclass' para poder correr este comando de forma correcta):

```
docker build -t bigdata .
```

3. Iniciamos la imagen

Creamos y arrancamos el contenedor con la imagen ya construida con el siguiente comando para accederlo:

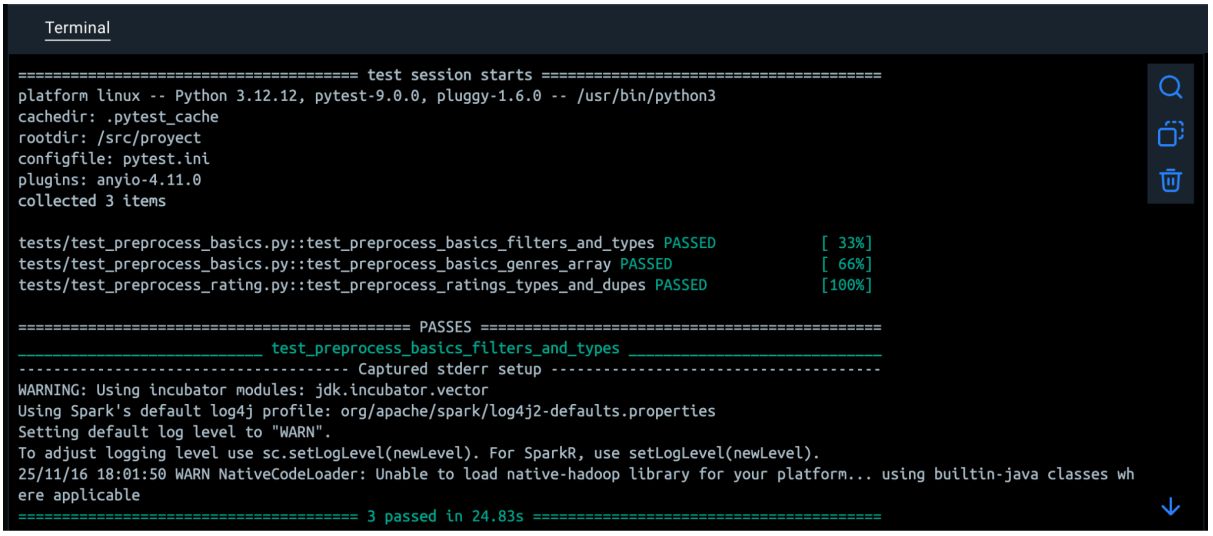
```
sudo docker run -p 8888:8888 -i -t bigdata /bin/bash
```

4. Corremos pruebas unitarias

Ya teniendo el contenedor construido, podemos correr las pruebas unitarias con el siguiente comando:

```
pytest -vv -rP
```

Se deben correr 3 pruebas unitarias para comprobar que el cargado de datos se hizo correctamente. Se debe obtener una resultado como la siguiente imagen:



Con tener estas 3 pruebas pasadas, sabemos que los datos fueron probados de forma correcta.

5. Correr datasets y modelos

Una vez corremos las pruebas unitarias, necesitamos correr los archivos de los datasets y de cargado de datos para poder visualizarlos y correrlos en nuestro Jupyter Notebook. Debes ejecutar estos comandos EN ESTE ORDEN:

python proyect/download_datasets.py

python proyect/src/etl/schemas.py

python proyect/src/etl/preprocess.py

6. Iniciamos Jupyter Notebook

Dentro del contenedor con la imagen ya construida, corremos jupyter notebook de una vez con el siguiente comando:

jupyter notebook --ip=0.0.0.0 --port=8888 --no-browser --allow-root

7. Corremos la base de datos

Ademas de tener el Jupyter abierto, creamos otra ventana en la terminal y corremos la base de datos de PostgreSQL con el siguiente comando:

docker run --name bigdata-db -e POSTGRES_PASSWORD=testPassword -p 5433:5432 -d postgres

Nota: Es muy importante tener el archivo postgresql-42.2.14.jar dentro de la misma carpeta donde se tiene el contenedor. Este archivo lo agarramos de la carpeta sparkml que venía desde el repositorio de bigdataclass en la clase 1.

Ya con estos pasos, dentro del Jupyter Notebook, puedes correr los comandos, y se realiza el cargado a la base de datos con el código correspondiente de forma satisfactoria.