

# Programa Ciencia de Datos – FUNDATEC

## Módulo 4 – Curso Big Data




















### Proyecto

Sebastián Porras

Octubre, 2022

**Instrucciones:** a continuación, se presentan indicaciones para crear un contenedor, ejecutar el programa principal

**Contenido de archivos en la carpeta comprimida:**

 .git	10
 Datos	10
 DB	10
 .gitignore	10
 build_image.sh	9
 conftest.py	10
 connect_db_from_spark_container.sh	9
 Datos.rar	9
 Dockerfile	9
 Entregable 1.docx	9
 Entregable 1.pdf	9
 functions.py	10
 load_jupyter_notebook.sh	9
 postgresql-42.2.14.jar	9
 Proyecto.ipynb	10
 proyecto.py	10
 README.md	9
 Run_image.sh	9
 test_unit_tests.py	10

- Archivo de datos Datos.rar
- Archivos para crear y configurar el contenedor: build\_image, Dockerfile, Run\_image
- Archivos .ipynb: Proyecto
- Archivos Postgres: Dentro de la carpeta DB
- Archivos .py: conftest, functions, proyecto, test\_unit\_tests

**Instrucciones para el código principal:**

Descargar y descomprimir el zip en un directorio local. Nombre esta carpeta 'Proyecto'.

Descomprimir el zip Datos.rar en un directorio local. Como en la imagen de arriba

Abra el command prompt del sistema, navegue los directorios y establezca a la carpeta 'Proyecto' como el directorio actual.

Construya la imagen utilizando el comando de docker que contiene el archivo 'build\_image.sh',  
<docker build --tag proyecto .> Alternativamente, también puede usar <bash build\_image.sh>

Luego, ejecutar la imagen con el comando presente en el archivo `'Run_image.sh'`, para crear el contenedor: `<docker run -p 8888:8888 -i -t proyecto /bin/bash>`

Para iniciar el contenedor de postgres, debemos irnos dentro de una terminal o command prompt del sistema a la carpeta `'DB'` y ejecutar la imagen con el comando presente en el archivo `'Run_image.sh'`, para crear el contenedor. Alternativamente, también puede usar `<docker run --name bigdata-db -e POSTGRES_PASSWORD=testPassword -p 5433:5432 -d postgres>`

Si se desea usar Jupyter Notebooks entonces:

1. Se debe ejecutar dentro del bash del contenedor `'proyecto'` el siguiente comando `<jupyter notebook --ip=0.0.0.0 --port=8888 --allow-root>`
2. Después de que se ejecute, en la terminal vamos a poder ver unas URL y debemos escoger una de esas para copiar y pegar en cualquier navegador para acceder al servidor de Jupyter.
3. Por último, en el servidor de Jupyter abrimos el archivo `Proyecto.ipynb`.

Si se desea usar python entonces:

1. Se debe ejecutar dentro del bash del contenedor `'proyecto'` el siguiente comando `<spark-submit --driver-class-path postgresql-42.2.14.jar --jars postgresql-42.2.14.jar proyecto.py>`

Si se quisiera revisar la tabla de los datos a usar para los modelos en postgres entonces debemos ejecutar el comando `<psql -h host.docker.internal -p 5433 -U postgres>` la contraseña es `testPassword`. Una vez conectado podemos ejecutar el comando `select * from datos;`

### **Instrucciones para las pruebas unitarias**

Finalmente, puede usar el comando `<pytest -s>` para ejecutar las pruebas unitarias definidas en el archivo `<test_unit_tests.py>` y visualizar su contenido.