Tarea #2 - Big Data

Estudiante: Ernesto Rivera

Instrucciones para correr la tarea

Una vez descomprimido el archivo tarea2_erivera.zip y abrir una consola (terminal) en la carpeta descomprimida:

1. Para crear la imagen
   a. `docker build -t bigdata:tarea2 .`
2. Si esta en **linux** o **mac**, puede modificar los archivos en caliente montando volúmenes:
   a. `docker run -v $(pwd)/src:/src -v $(pwd)/data:/data -i -t --name tarea2 bigdata:tarea2 bash`
3. Si está en **windows** tendrá que modificar los archivos volviendo a ejecutar el paso 1 y la siguiente instrucción:
   a. `docker run -i -t --name tarea2 bigdata:tarea2 bash`
4. Para correr el programa principal:
   a. `spark-submit programaestudiante.py`
5. Para correr las pruebas
   a. `pytest`
6. Para salir del contenedor
   a. `exit`
7. Para borrar el contendedor
   a. `docker container rm tarea2`

Archivos resultantes: pyspark genera carpetas distintas para cada archivo.

Si está en **Linux** o **mac** (paso 2.a de las instrucciones anteriores), los archivos quedan en **un solo archivo** por carpeta, **en el folder local**, que sigue los siguientes patrones:

- ./data/metricas.csv/part-00000-*.csv
- ./data/total_ingresos.csv/part-00000-*.csv
- ./data/total_viajes.csv/part-00000-*.csv

Si está en **Windows** (paso 3.a de las instrucciones anteriores), los archivos quedan en **un solo archivo** por carpeta, **dentro del contenedor**, que sigue los siguientes patrones:

- /data/metricas.csv/part-00000-*.csv
- /data/total_ingresos.csv/part-00000-*.csv
- /data/total_viajes.csv/part-00000-*.csv