Ingesta datos covid19

Objetivos Principal:

El principal objetivo de este proyecto es poder cargar, o realizar la ingesta de los datos relacionados al covid19 y que pone a disposición el gobierno de la Ciudad de México, esto con la finalidad de que los estudiantes, académicos o interesados en este tema, que requieran utilizar esta información no tengan que estar cargando la información de forma diaria y ya tenerla dentro de una base de datos lista para usarse.

Objetivo secundario

Poner a prueba los conocimientos aprendidos en el Bootcamp de Ingenieria de datos que impartió Códigofacilito, así como practicar algunas de las herramientas impartidas en este curso que no son de pago.

Sobre los datos

Los archivos utilizados para este proyecto son de libre acceso y fueron puestos a disposición por el gobierno de la Ciudad de México.

Estos datos son de primera fuente, pues fueron datos que se recopilaron durante la emergencia sanitaria ocurrida en el año 2020.

Arquitectura y Herramientas

Las herramientas que se utilizaron para realizar este proyecto son principalmente Python, Docker,PostgreSQL y DBeaver, por lo que el flujo del proyecto sería de la siguiente forma.

Describiendo el flujo:

En un principio es necesario identificar la pagina donde se encuentran los archivos, en este caso los archivos se encuentran en las siguientes URL:

https://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/datos_abiertos_covid19.zip

https://datosabiertos.salud.gob.mx/gobmx/salud/datos_abiertos/diccionario_datos_covid19.zip

una vez localizada nuestra información, se ejecutarán en el siguiente orden los scripts de Python:

- extract_data.py: Su principal función es descargar los archivos necesarios para la carga de la información
- modified_cat_file.py: Modificar los catálogos, con las respectivas descripciones y dejarlos todos en un mismo formato.
- load_cat.py: Una vez modificado los catálogos, se cargarán a nuestra BD
- load_estados.py: Carga la información de los catálogos Entidades y Municipios
- load_covid_data.py: Carga la información final del insumo principal, a nuestra BD

Así una vez ejecutado nuestros scripts de Python, la información estará almacena en nuestra BD.

En cuanto a nuestro almacenamiento se optó por utilizar Docker y PostgreSQL, esto con la finalidad de aprender más sobre el como usar estas herramientas, además una ventaja grande es que no instalamos programas en este caso pgAdmin4, el cual es bastante pesado.

Puntos para mejorar

Durante el desarrollo del proyecto, detectamos algunos puntos a mejorar principalmente en los scripts de Python, se pueden mejorar los scripts donde se realiza la carga de catálogos y más específicamente se puede mejorar el performance del script que realiza la cargar de los datos del insumo principal (COVID19MEXICO).

A pesar de que el gobierno de México ya no actualice de manera diaria los archivos como en los primeros años de la pandemia, podemos crear un orquestador en este caso Airflow para realizar la actualización de la información de manera mensual.