CODERFLEX

Data Engineering
Consigna de Tercer Pre entregable

CODERHOUSE

Script en un container de Docker y en un DAG de Apache Airflow

Para finalizar tu tercer pre entregable, te proponemos que el script el script de la 2da entrega corra en un container de Docker y esté embebido en un DAG de Airflow dentro del container.

Objetivos

- Crear un script liviano y funcional que pueda ser utilizado en cualquier Sistema operativo y por cualquier usuario.
- Dockerizar un script para hacerlo funcional en cualquier sistema operativo.

Requisitos

Este trabajo cuenta con una instancia que se debe **mostrar dentro de una misma presentación**:

1. Dockerfile

1. Dockerfile

Dockerfile y código con todo lo necesario para correr (si es necesario incluir un manual de instrucciones o pasos para correrlo), subido en repositorio de Github o en Google Drive.

El objetivo es entregar un Dockerfile donde sean incluídas las instrucciones para hacerlo correr de ser necesario. Considera los siguientes puntos:

- **El container debe ser lo más liviano posible:** Para que el script funcione sin problemas este debe ser ligero.
- **Cualquier usuario podría correr el container.**
- **El script debe estar listo para su ejecución.**
- **El Dockerfile debería utilizar un FROM de python y después instalar Airflow con pip.**
- **El Dockerfile debería tener un comando COPY para copiar el archivo a la carpeta de dag.**
- **El script debería estar dentro de una función que realice una llamada desde un PythonOperator**
- **El DAG debería correr de forma diaria**
- **El nombre del DAG y su descripción deberían ser fáciles y precisos de entender.**
- **Los parámetros elegidos fueron justificados de forma adecuada**

Este preentregable debería estar muy cercano ya al resultado del proyecto final.

2. Tabla en Amazon Redshift

Tabla creada en Redshift con los datos de muestra que hayan sido cargados mediante el script.

A su vez, la entrega involucra la creación de una versión inicial de la tabla donde los datos serán cargados posteriormente. Considera los siguientes puntos para su elaboración:

1. **Los datos deben ser extraidos y cargados con sus correspondientes tipos de datos en relación a la tabla creada en Redshift.**
2. **Todas las columnas deberían ser cargadas en la tabla.**
3. **Debe haber una clave primaria compuesta definida en la tabla o en el código**
4. **En caso de que se quiera insertar una fila con los mismos datos, debe ser reemplazada por los nuevos? Por ejemplo: la columna "fecha" y "ciudad" puede ser una clave primaria compuesta, ya que no deberían haber 2 datos diferentes para una misma ciudad en un mismo día.**

Recuerda que esto será la base para tu proyecto final.

Carga de datos en Amazon Redshift

Recomendaciones

- Utiliza la guía de actividades para poder tener mayor referencia de lo que se puede hacer para cumplir con este pre entregable.
- Consultar la documentación oficial relacionada con los temas de Apache Airflow. Investigar sobre Docker Compose para facilitar la tarea.
- La base de datos donde estará esta tabla no hace falta que viva en el container, sino que se tiene en cuenta que es un Redshift en la nube.

Ejemplos

Para guiarte, te compartimos el siguiente ejemplo:

- Proyecto final.

Criterios de evaluación

Para la evaluación de tu Proyecto Final, tendremos en cuenta los siguientes criterios de evaluación.