Data Modeling with Postgres

Introduction

This project is to do data modeling with Postgres and build an ETL pipeline using Python. The projects requires to define fact and dimension tables of star schema and transfer data from local folders to Postgres with Python and SQL.

Description of each file

Create_tables.py: It is to create databases and tables to get prepared for further modeling; sql_queries: SQL Queries that are used in Postgres database and in this project, python is going to make use of these queries to create tables and insert values; etl.py: A pipline that reads Json logs and metedata and populate data into tables;

How to Run the Script

I have put all sql queries in one file. So all you need to do is to run create_tables.py first and then run etl.py.

In jupyter notebook, execute "%run -i create_tables.py" first and then execute "%run -i etl.py"

Tricks of the project

1. There are duplicates records of dimensional tables. The solution here is to skip these columns when are inserted into the tables. In code, it should be "ON CONFLICT (artist_id) DO NOTHING".
2. It is a little bit special for user table. New user record could due to change of user behaviors. For example, user could switch to a paid user from a free one before. In this case, we need to update the field.
3. In python, spaces and tab should be clear. Four spaces equals a tab. The rule is very strict. In this project, I struggled for a while for this. While writing comments, I have 5 spaces in the beginning and no errors in the code, but still failed.

Database Schema

Database Schema