1_data_pipeline

Instruction

1. Open the parent directory of the project in VS code or your preferred code editors and open the **terminal**.

2. Execute **. start.sh** in the terminal to setup the docker environments.

3. Setup email alart

   i. Setup app passwords in https://security.google.com/settings/security/apppasswords

   ii. Setup smtp under **airflow.cfg** as below (The rests remind the same) smtp_host = smtp.gmail.com smtp_user = {your email} smtp_password = {password generated from app password} smtp_port = 587 smtp_mail_from = {your email}

4. Open web browser and go to **http://localhost:8080/**

   user: airflow

   password: airflow

5. Setup connection to postgres

   i. Click on Admin on the top bar

   ii. Click on Connection and add a new connection name **postgres**

   iii. Change the conn Type to **Postgres**

   iv. Input Host as **postgres**

   v. Input Schema as **postgres**

   vi. Input Login as **postgres**

   vii. Input Password as **postgres**

   viii. Input Port as **5432**

   ix. Save it

6. Setup postgres in localhost

   i. Install DBeaver

   ii. Create New Database Connection

   iii. Input Host as **locahost**

   iv. Input Database as **postgres**

v. Input Username as **postgres**

vi. Input Password as **postgres**

vii. Input Port as **15432**

viii. Save it

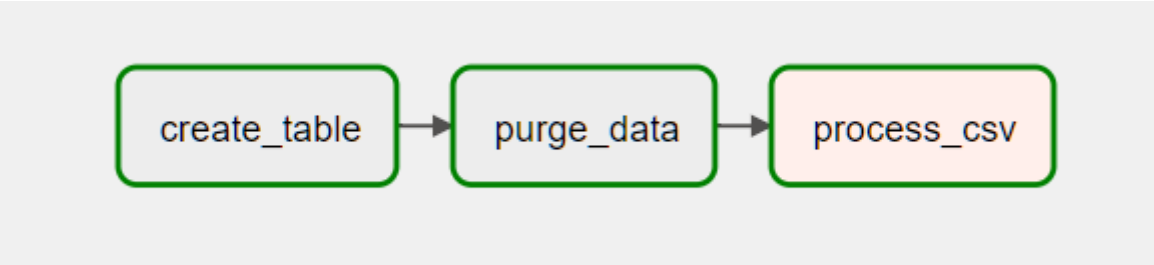
7. Switch on the DAG **1_data_pipeline** and let it run or create on play button on the actions bar to trigger the DAG manually
8. Read the results

   i Read processed data from results

   ii. Read processed data from DBeaver (as per step 6)

9. Close the project by executing **. stop.sh** in terminal

Methods explanation



1. Firstly, create the table in PostgreSQL if it does not exist.
2. Prior to reading the data, perform a purge to remove any existing data.
3. Process and write the datasets into the database within a for loop.
4. Provide an option to generate CSV output in the result folder with an absolute path, facilitating easy reading.

Note

1. Assuming that you have already set up a basic development environment and docker desktop on your workstation.
2. This setup is for showcase purpose, not for production deployment