Contents

[**SOFTWARES:** 3](#_Toc134657786)

[**INSTALLATION:** 3](#_Toc134657787)

[PostgreSQL: 3](#_Toc134657788)

[Ubuntu: 8](#_Toc134657789)

[Airflow: 9](#_Toc134657790)

[ETL Setup: 10](#_Toc134657791)

[EXECUTION: 17](#_Toc134657792)

[CONCLUSION: 17](#_Toc134657793)

[NOTE: 17](#_Toc134657794)

# **SOFTWARES:**

Total of three software’s are required.

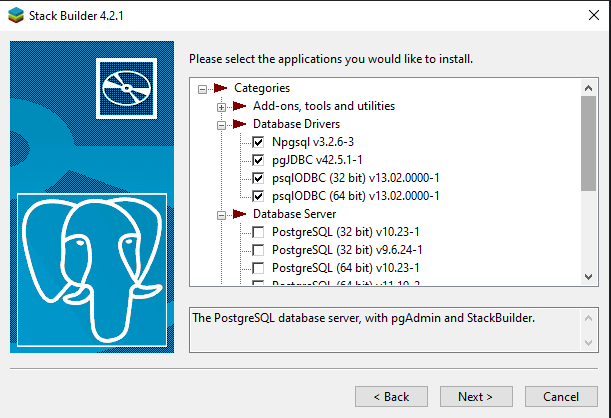
* **PostgreSQL**
* **Ubuntu(wsl)**
* **Airflow**

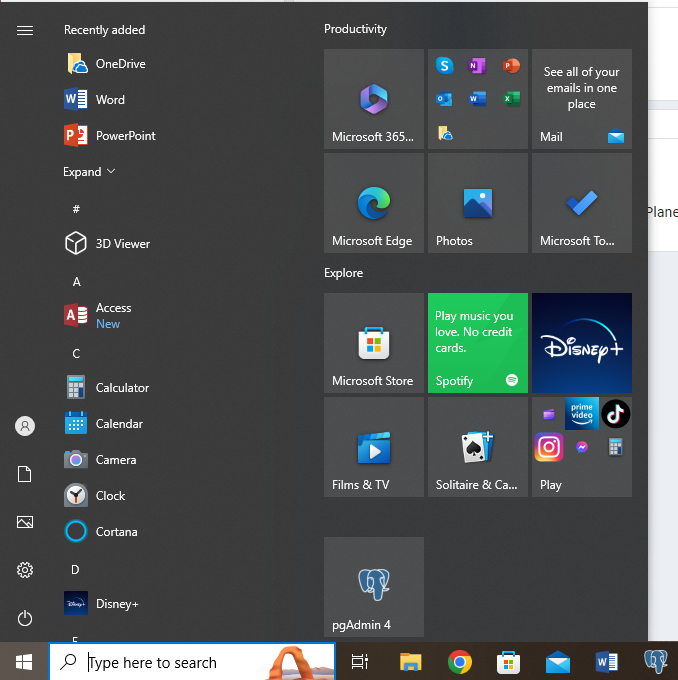
# **INSTALLATION:**

## PostgreSQL:

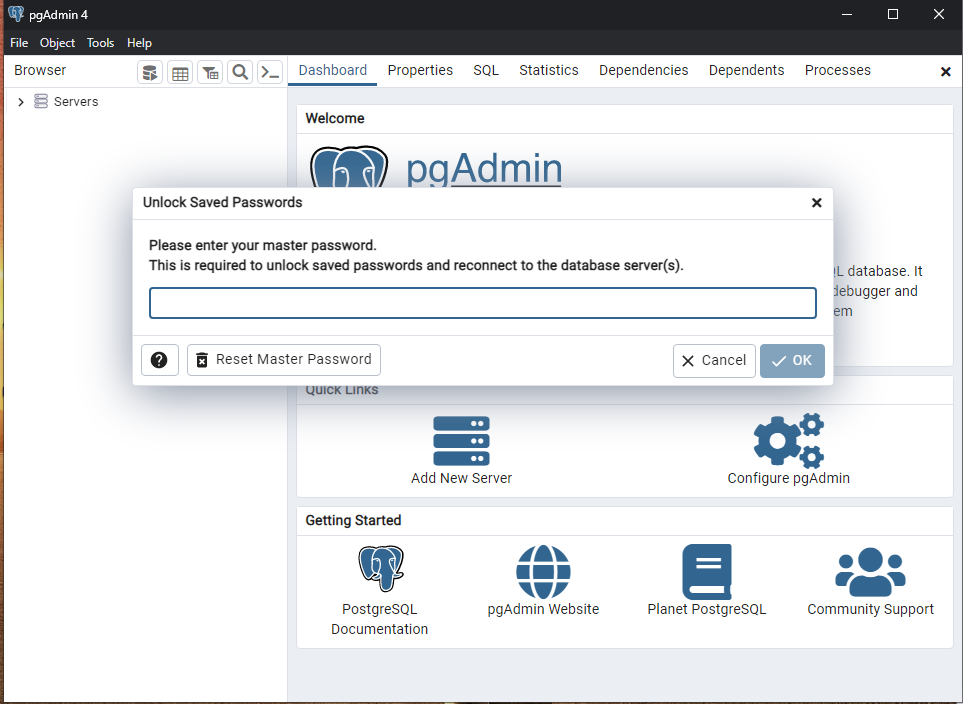
* Download the program from the following link and install it.

<https://www.postgresql.org/download/windows/>

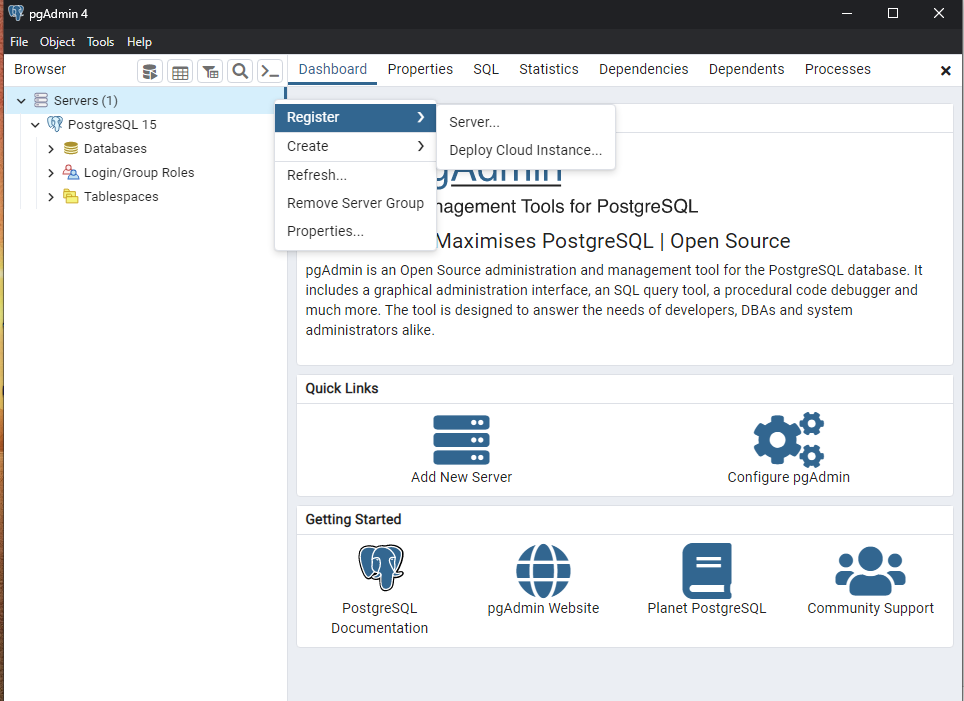
* Keep the settings default during the installation.
* Set the password for the PostgreSQL(Remember it).
* Check the fields for ODBC drivers and click next.
* After completing the installation launch the PgAdmin4 from the Start menu.



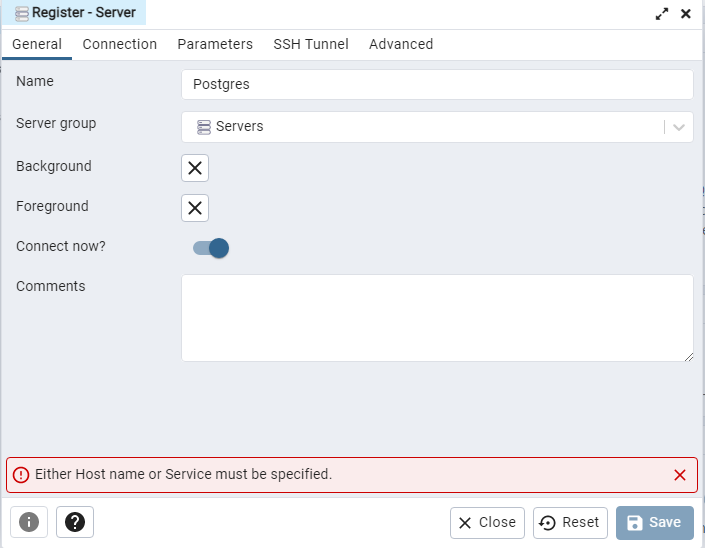
* Enter the password you set during the installation to connect to the database server.

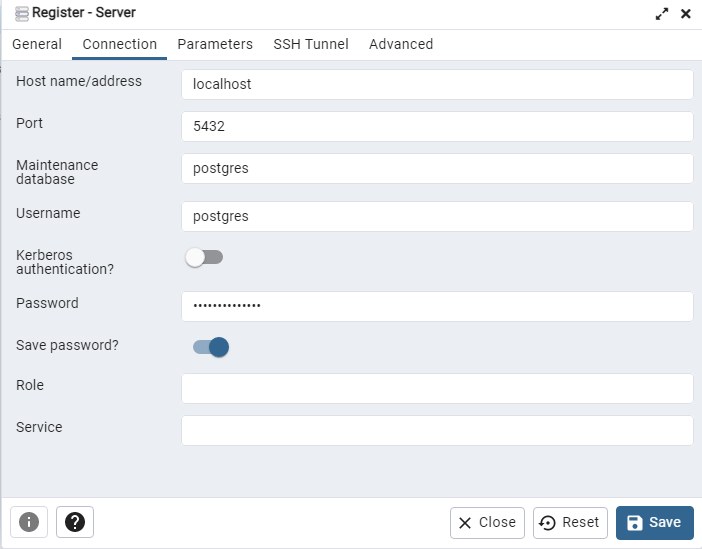


* Right click on the Servers and click on **Register**>>**server…**

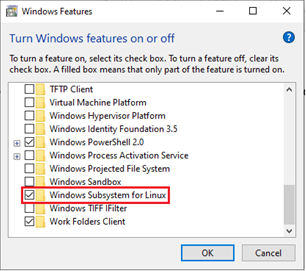


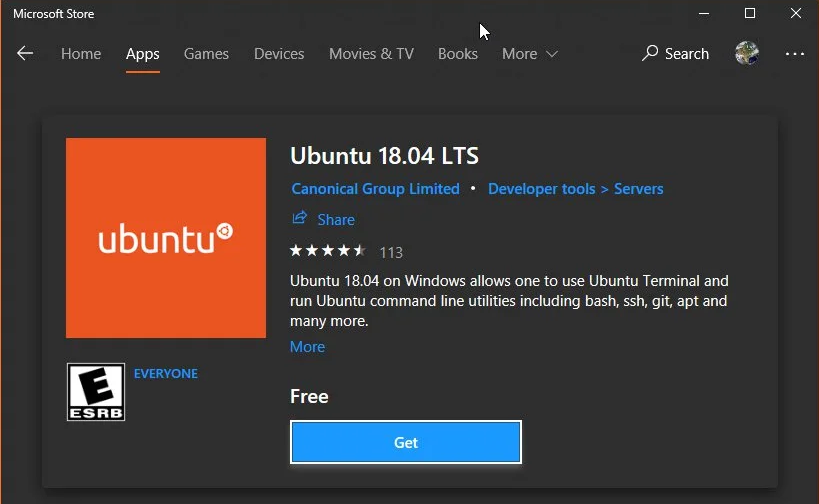
* Enter name and host name, and password (during installation one) in connection tab and then save.





## Ubuntu:

* From the Start menu >> Turn Windows features On or Off and check the box for Windows Subsystem for Linux then click ok and restart the system.  
   
* From the Start menu >> Microsoft Store >> Search of Ubuntu and install it.

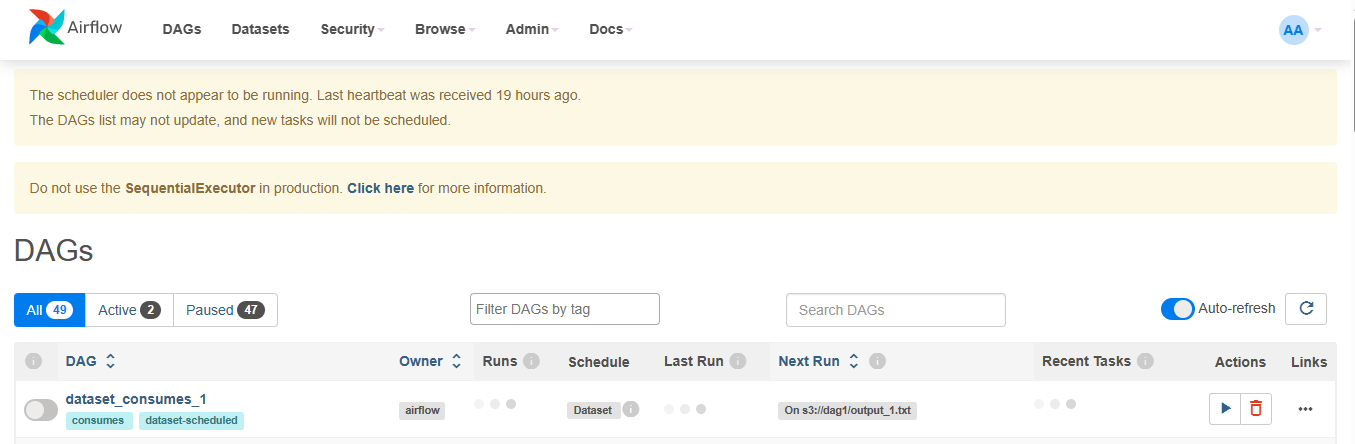


* Launch Ubuntu 18.04
* Enter a username. This will create a local user account and you will be automatically logged in to Ubuntu 18.04 as this user.
* Enter a password for the user and enter a second time to confirm.
* Run the following commands one by one:
  + sudo apt update && sudo apt upgrade
  + sudo apt-get install software-properties-common
  + sudo apt-add-repository universe
  + sudo apt-get install python3-pip

## Airflow:

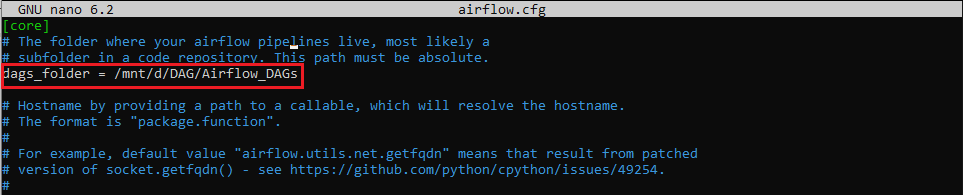
Run the following commands one by one in Ubuntu Terminal.

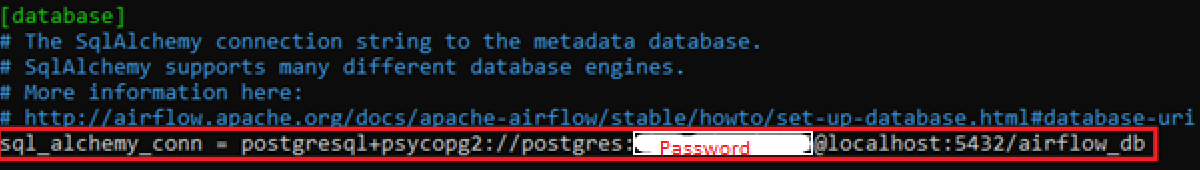
* *export SLUGIFY\_USES\_TEXT\_UNIDECODE=yes*
* *sudo pip3 install apache-airflow*
* *airflow db init*
* *airflow webserver -p 8080*
* *airflow users create --role Admin --username admin --email admin --firstname admin --lastname admin --password admin*
* Airflow webserver is ready open the address in a browser [’http://localhost:8080/](http://localhost:8080/)’.



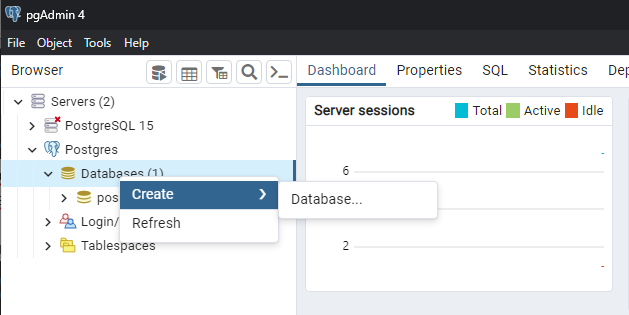
* Open the query tool in PgAdmin4 with connection to Postgres server and execute the following queries one by one.
  + CREATE DATABASE airflow\_db;
  + GRANT ALL PRIVILEGES ON DATABASE airflow\_db TO postgres;
  + GRANT ALL ON SCHEMA public TO postgres;

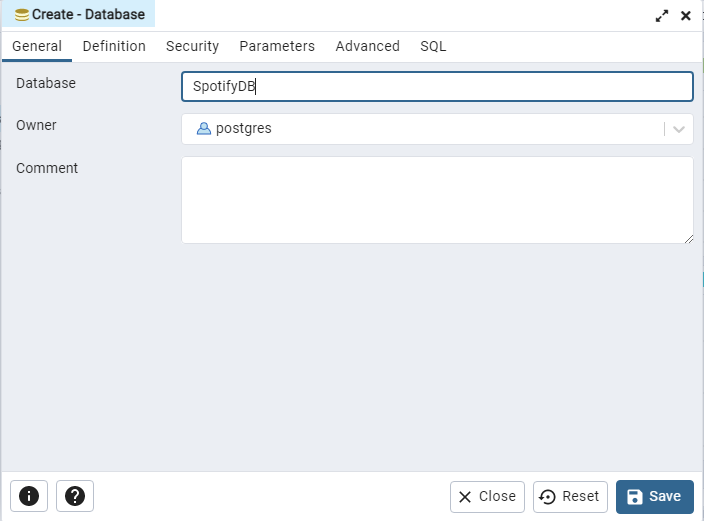
# ETL Setup:

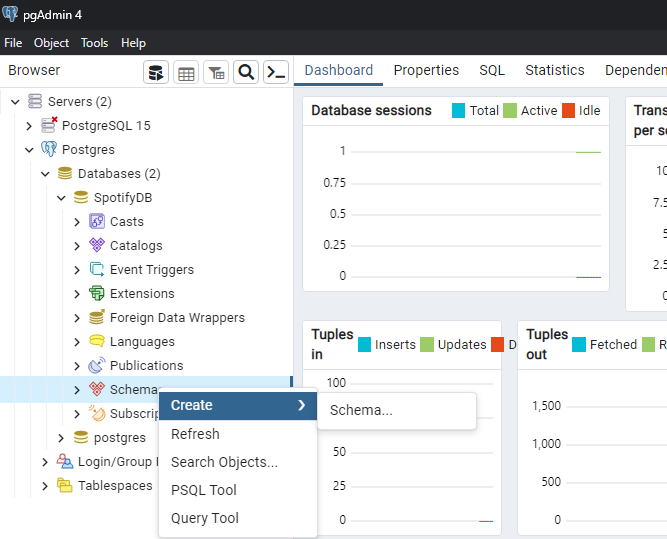
* Launch Ubuntu and in terminal run the following commands which download the libraries required to execute the python files.
  + *sudo pip3 install spotipy*
  + *Sudo pip3 install pandas*
  + *sudo pip3 install psycopg2-binary*
* Now run the following commands
  + *cd airflow*
  + nano airflow.cfg
* Change the dags\_folder in airflow.cfg where the python files are located.  
  
* Change the connection string in airflow.cfg to the
  + postgresql+psycopg2://user\_name:password @localhost:5432/airflow\_db
  + user\_name and password should be the same as in PostreSQL .

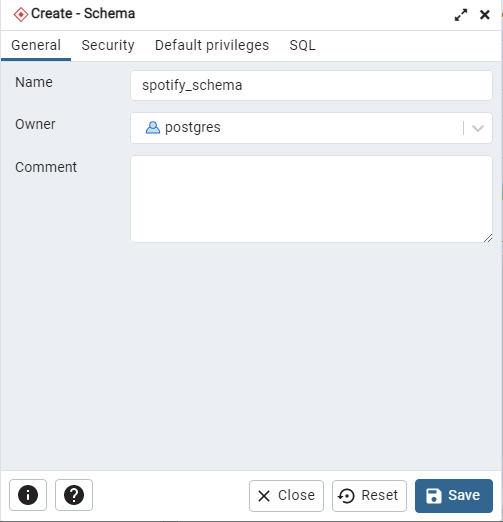


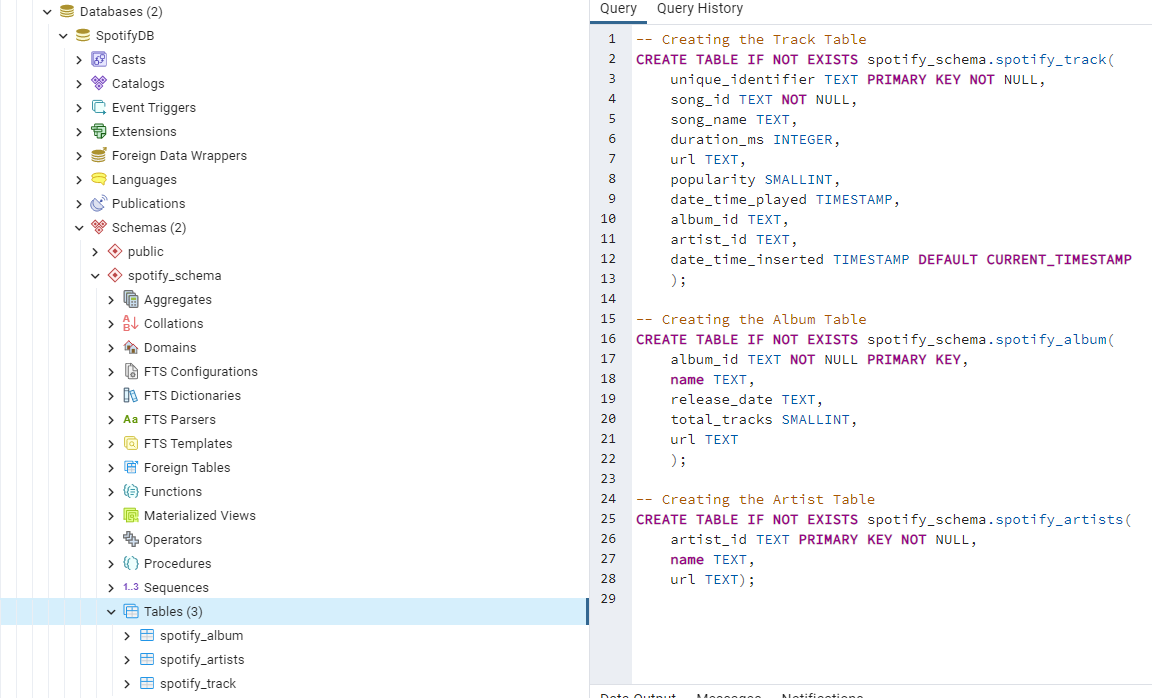
* Now create the “SpotifyDB” database in the postgres server and then click save.

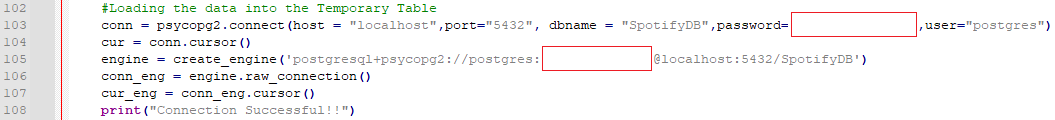


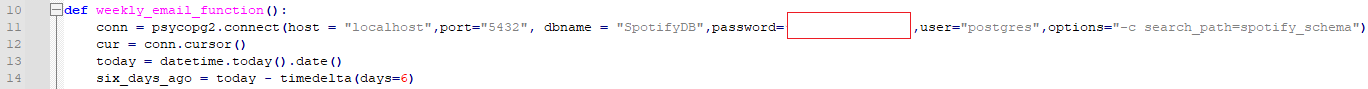


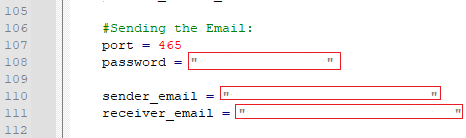
* Now run the PostgeSQL(pgAdmin4) and create “spotify\_schema”. Right click on Schemas >> Create >> Schema and the click save.  
   



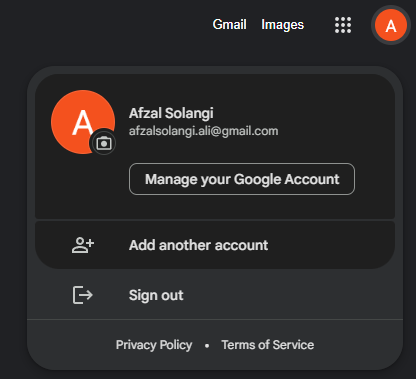
* Now right click on the created schema and then Query Tool.
* Now run the SQL queries provided in the **\DAG\SQL** folder from the file **Create\_TablesDB.txt.** 
* Now open the python files “ETL\_Spotify.py” and from the folder **\DAG\Airflow\_DAGs** using VSCode or Notepad++ and edit the connection for PostgreSQL, add your password in the red fields in the code.



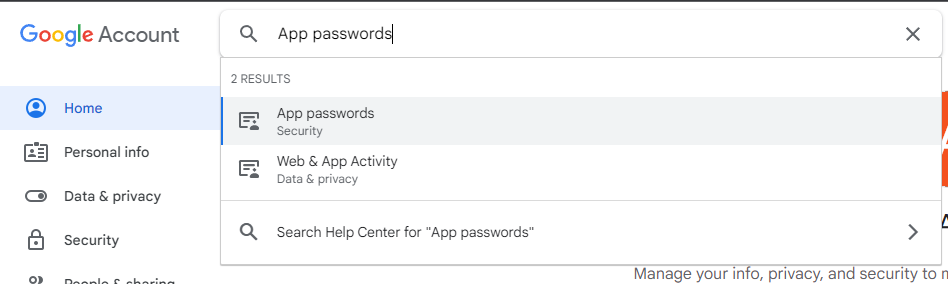
* Now open the python files “ETL\_Spotify.py” and from the folder **\DAG\Airflow\_DAGs** using VSCode or Notepad++ and edit the connection for PostgreSQL, add your password in the red fields in the code.
* In the same file, enter the sender email and receiver email.

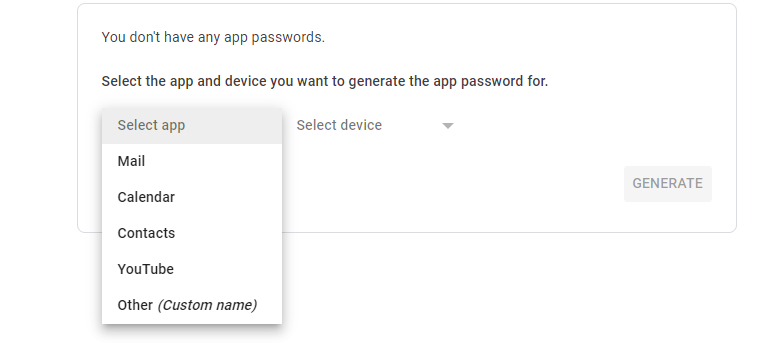


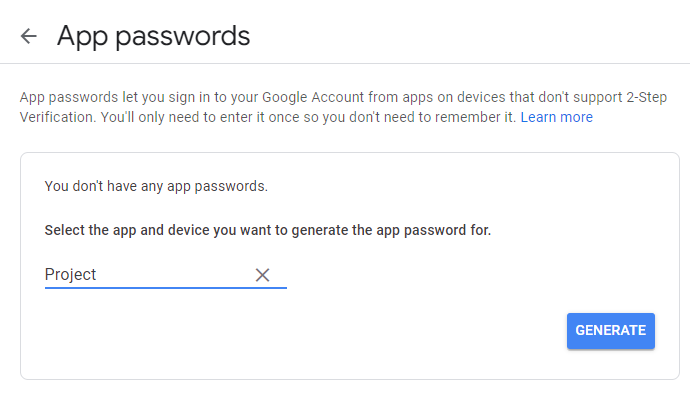
* For the password go to the, “manage your google account”.

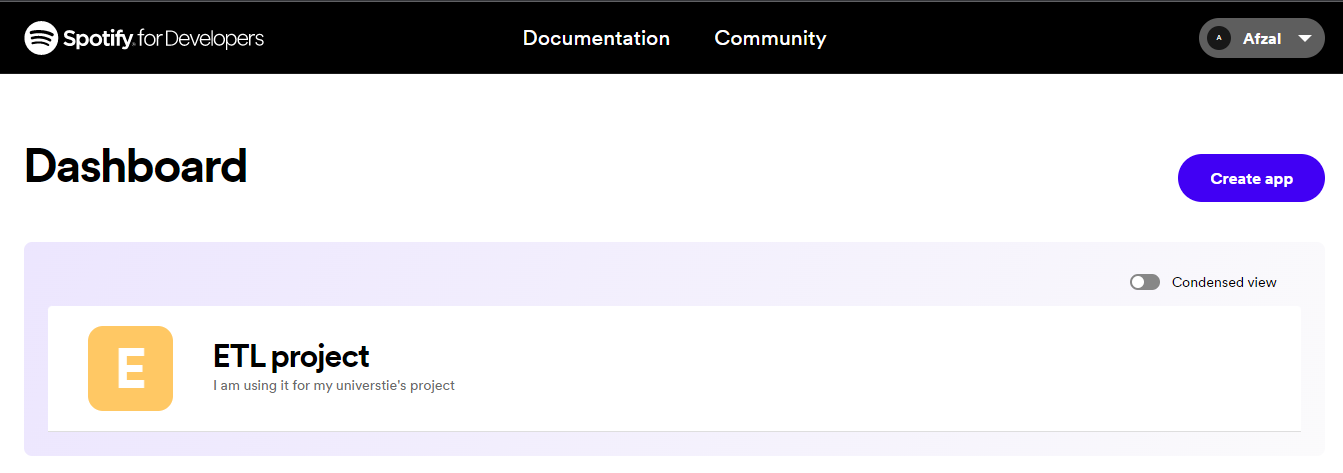


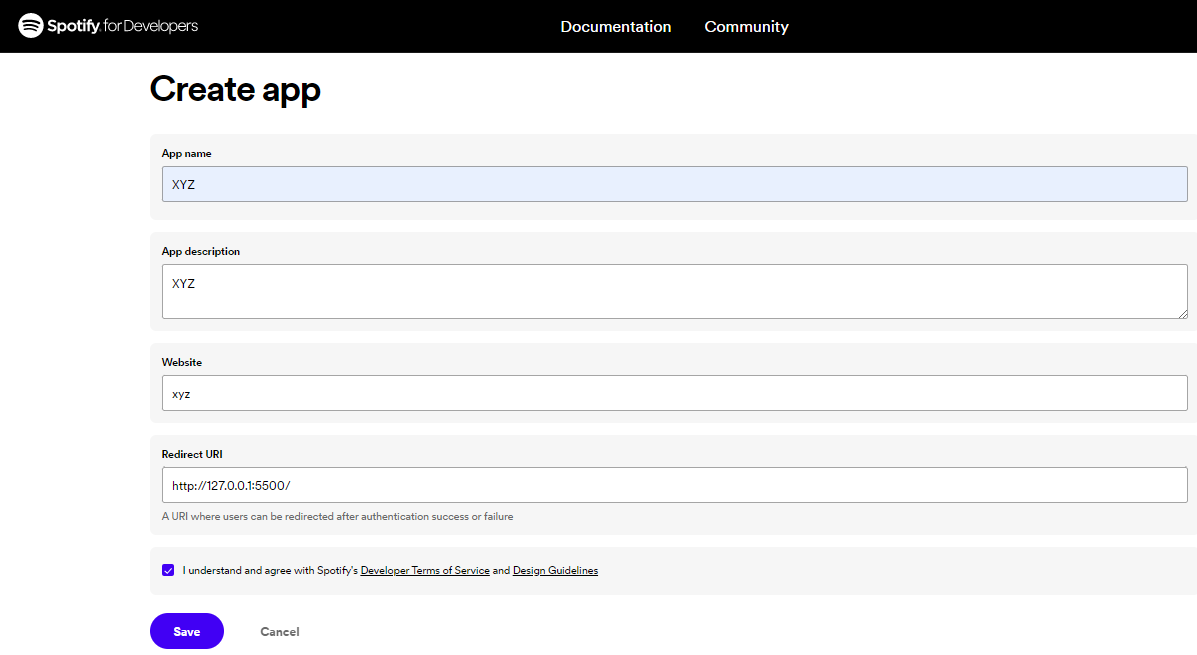
* Search for app passwords.

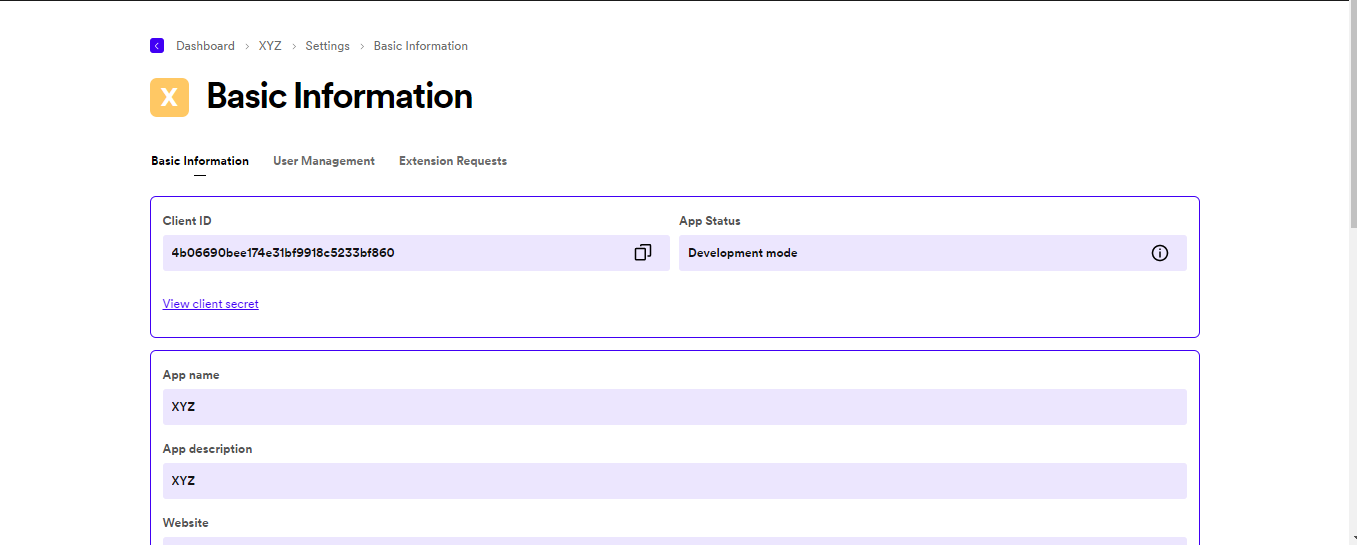


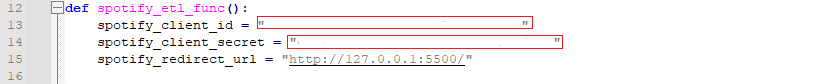
* In the “Select app” select Other from the drop down menu.  
  
* Enter the name for the app and then click on generate.



* Then copy the password and paste it into the password string in the “ETL\_Spotify.py” file.
* To connect to Spotify account follow the following steps.
* Go to the link <https://developer.spotify.com/dashboard>.
* Click on create app.  
  ****
* Fill the fields enter the redirect URI as ‘http://127.0.0.1:5500/’ and click save.

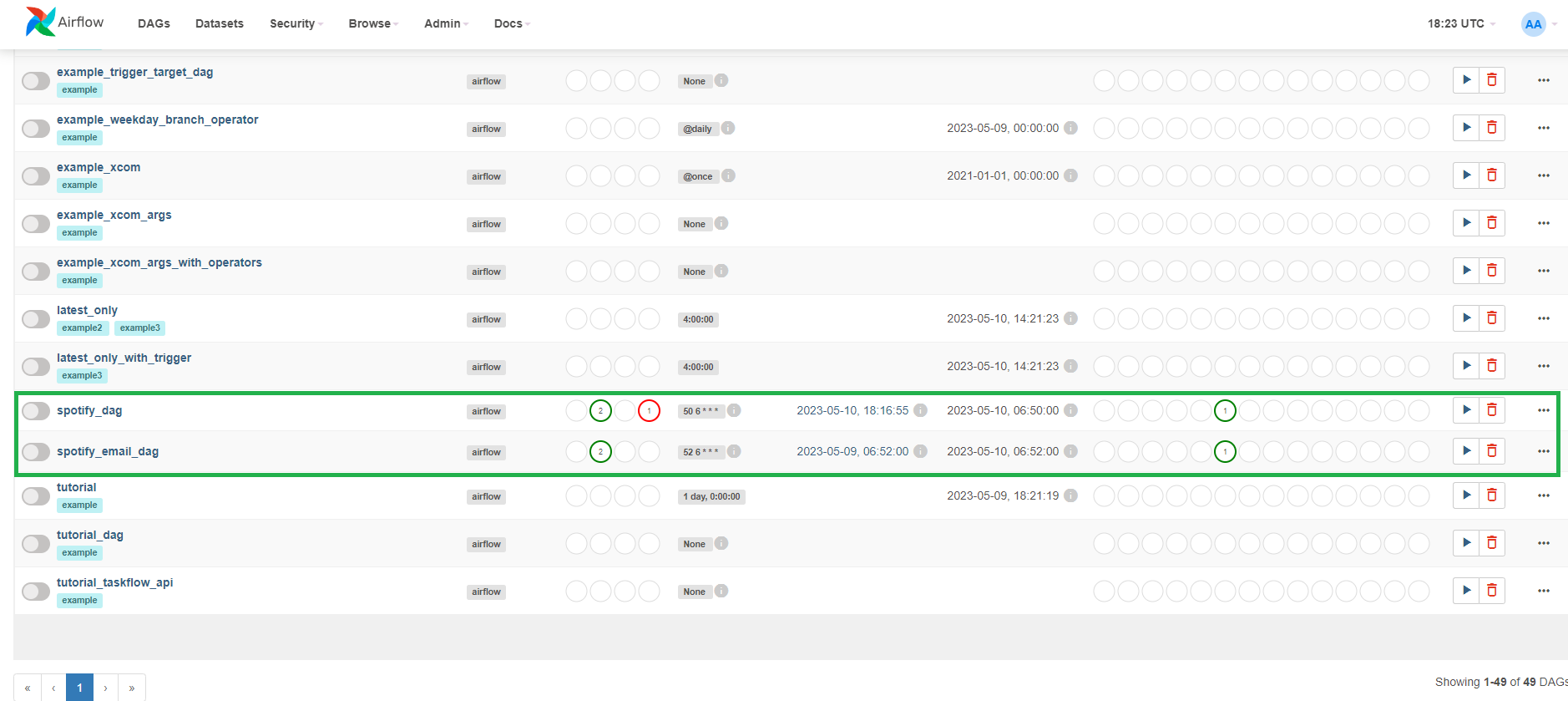


* Now click on settings and copy the client ID and client secret.  
  
* Now open the python file “ETL\_Spotify.py” and from the folder **\DAG\Airflow\_DAGs** and paste them here**.**



# EXECUTION:

* To schedule or run the ETLs follow the following steps.
* Run the PgAdmin4 and connect to the server “Postgres”.
* In ubuntu terminal run the following command:
  + airflow scheduler
* Open another ubuntu terminal and run the following command:
  + airflow webserver -p 8080
* Open another ubuntu terminal and run the following command:
  + airflow db init
* Now open airflow through link:
  + <http://localhost:8080/>



* Both DAG files now appear in the Airflow and will run at the scheduled time or it can also be manually triggered by clicking on the play button.

# CONCLUSION:

By following all the steps stated in this user manual the end user can easily deploy a fully schedulable ETL.

# NOTE:

In the DAG files the ETL schedule can be changed through the CRON expression. The reference website on how to use it is <https://crontab.cronhub.io/>