How to run it

1 - Clone the git repository: https://github.com/ferolifil/bees-case

2 - Run Docker. On PowerShell navigate to the root directory where git was cloned:

cd \<your_path>\bees-case\

3 - Run airflow container with command:

docker run -d -p 8080:8080 -v "$PWD/airflow/dags:/opt/airflow/dags/" --entrypoint=/bin/bash --name airflow apache/airflow:2.10.1-python3.8 -c '(airflow db init && airflow users create --username admin --password beescase --firstname Fernando --lastname Oliveira --role Admin --email admin@example.org); airflow webserver & airflow scheduler'

Connect to airflow container:

docker container exec -it airflow bash

Install required libraries:

pip install boto3 timedelta

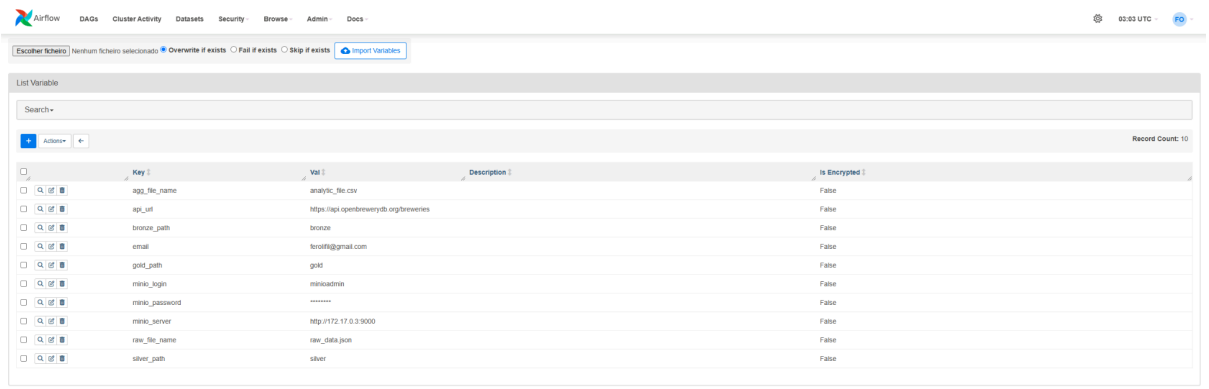
4 - Run Min.IO container with command:

docker run --name minio -d -p 9000:9000 -p 9001:9001 -v "$PWD/datalake:/data" minio/minio server /data --console-address ":9001”

5 - Connect to airflow UI and add airflow variables as example

Connect to: https://localhost:8080

login: admin
senha: beescase



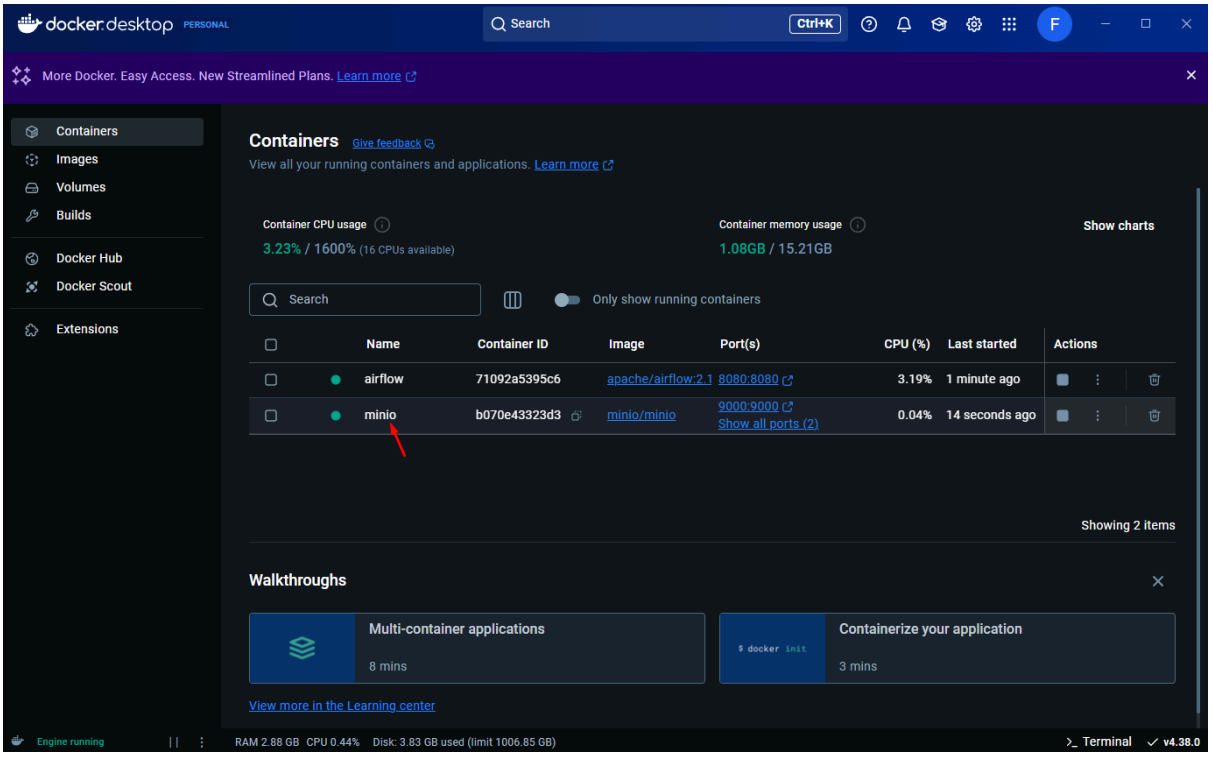
Admin >> Variables

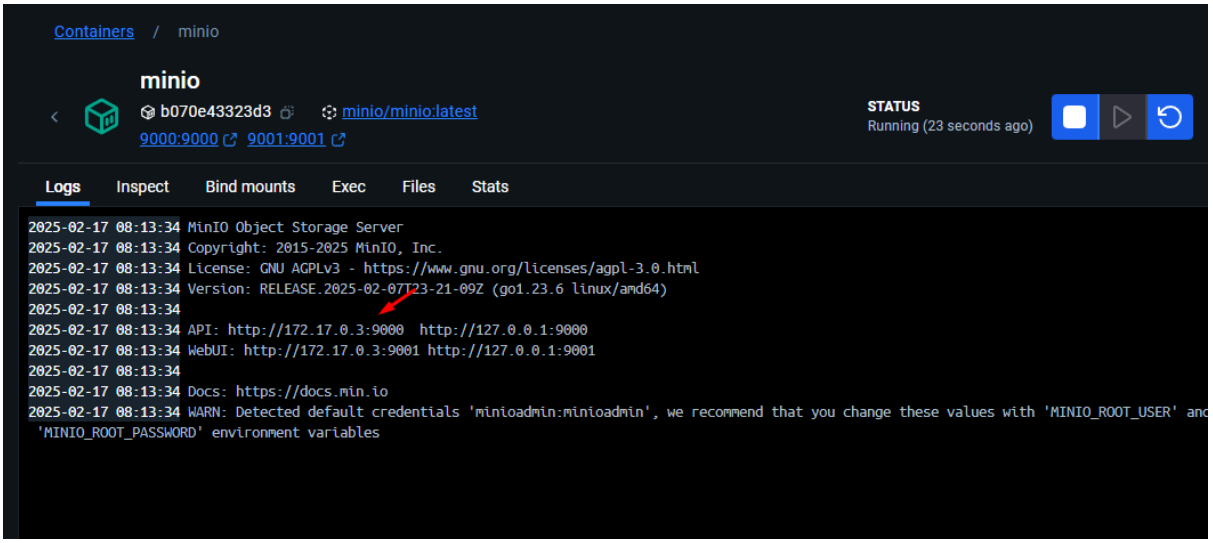
Variables List:

- agg_file_name / analytic_file.csv
- api_url / https://api.openbrewerydb.org/breweries
- bronze_path / bronze
- email / <your_email>
- gold_path / gold
- minio_login / minioadmin
- minio_password / minioadmin
- minio_server / <Min.IO server address>
- raw_file_name / raw_data.json
- silver_path / silver

5 - To find Min.IO server address:

Open Docker desktop:





To use Min.IO UI use:

login: minioadmin
password: minioadmin

6 - After finishing the configurations, enable DAGs in Airflow. Following the order:

api_etl >> parquet_etl >> agg_etl