AIS-data-pipeline

End-to-End Automated Data Pipeline: From Data Acquisition to Visualization with Dockerized Spark, HDFS, and Airflow, Postgres and metabase.

Pipeline Architecture

Overview

Objective:

- Create an automated system for data workflow from download to visualization.

Key Elements:

- Utilizes AIS data for maritime traffic analysis.
- Incorporates Docker, Apache Spark, HDFS, PostgreSQL, Metabase, and Airflow.
- Process Overview:
- Environment setup with Docker.
- Data download/storage automation.
- Daily task scheduling with Airflow.
- Data cleaning and destination analysis.

Outcome:

- Streamlined daily data processing for analysis readiness.

Recommendations:

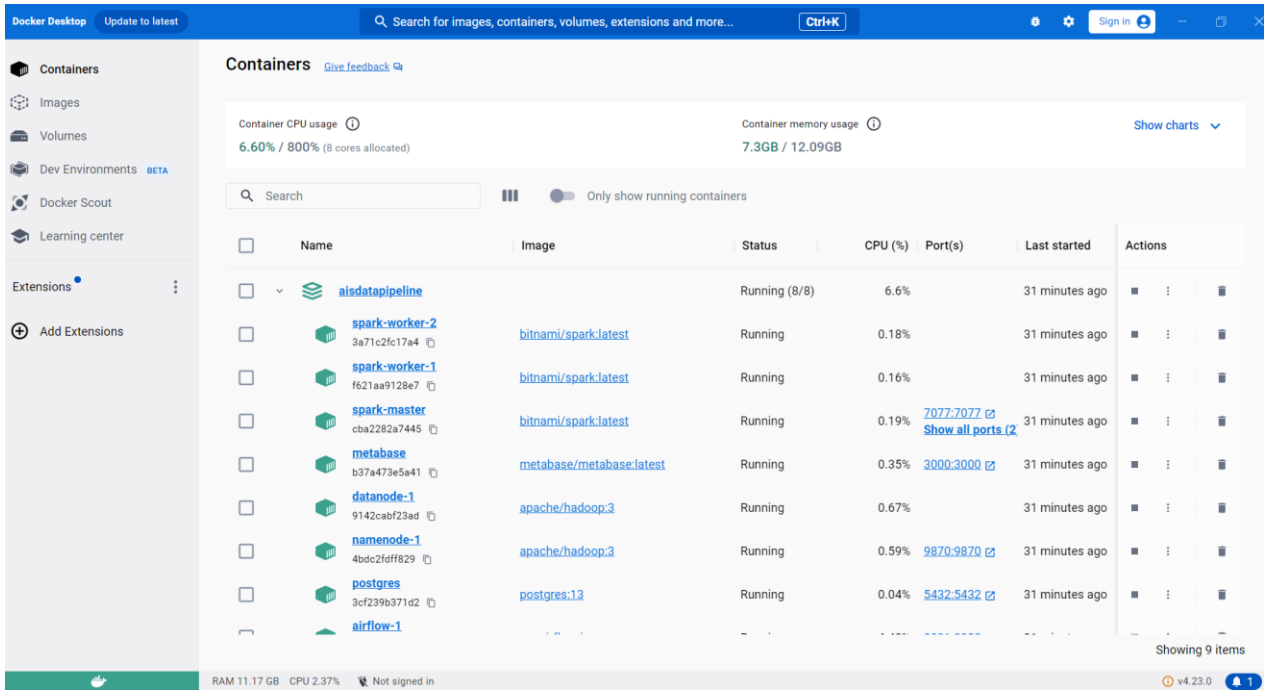
- Add error-handling.
- Visualize with PostgreSQL.
- Regular system monitoring.

Note:

- The code used get the data from the local directory to ease the process of testing the pipeline

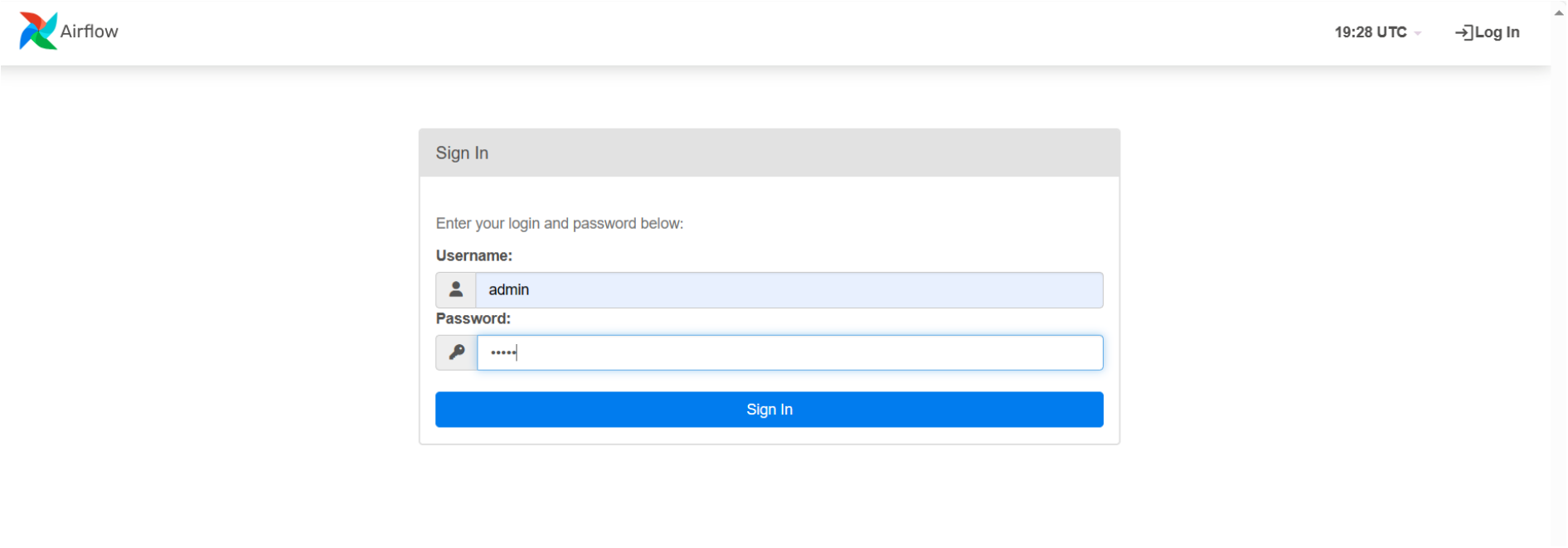
Running containers

- After building the image for Airflow using a python base image
- Run docker-compose up --scale spark-worker=2 -d to create two spark worker
- Containers should look like this in docker desktop



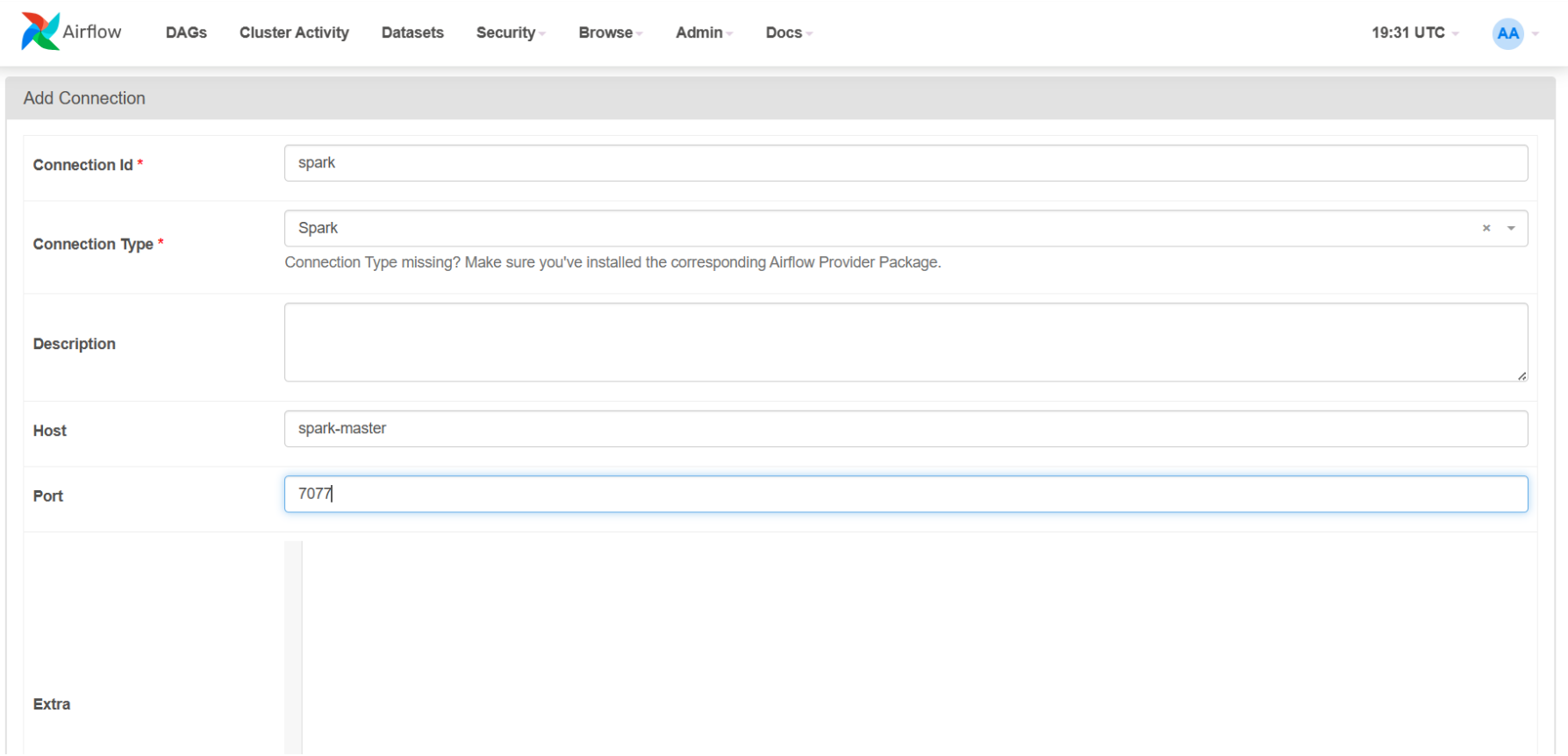
Airflow

- Log in on **http://localhost:8081/login/**



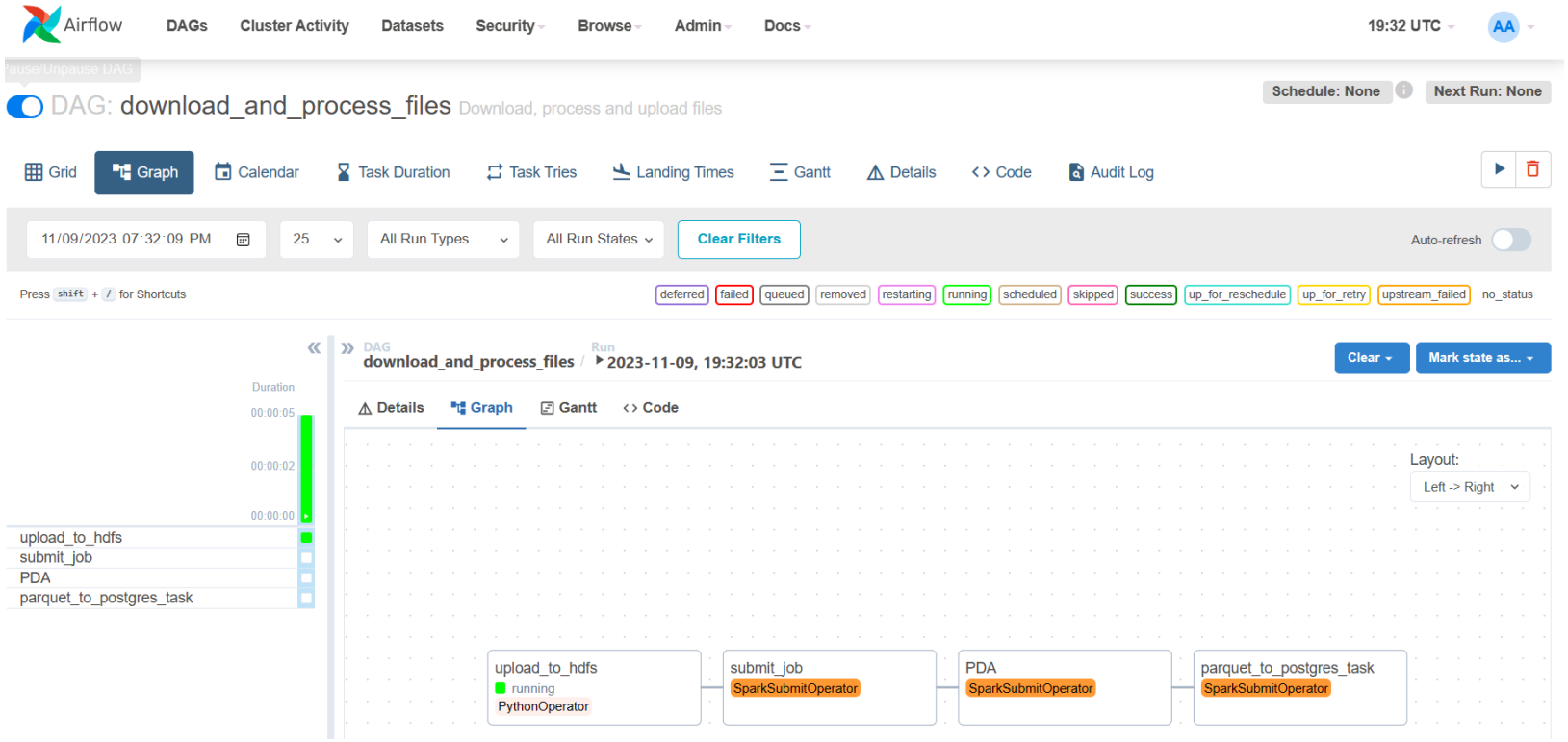
Airflow

- Create a spark connection on



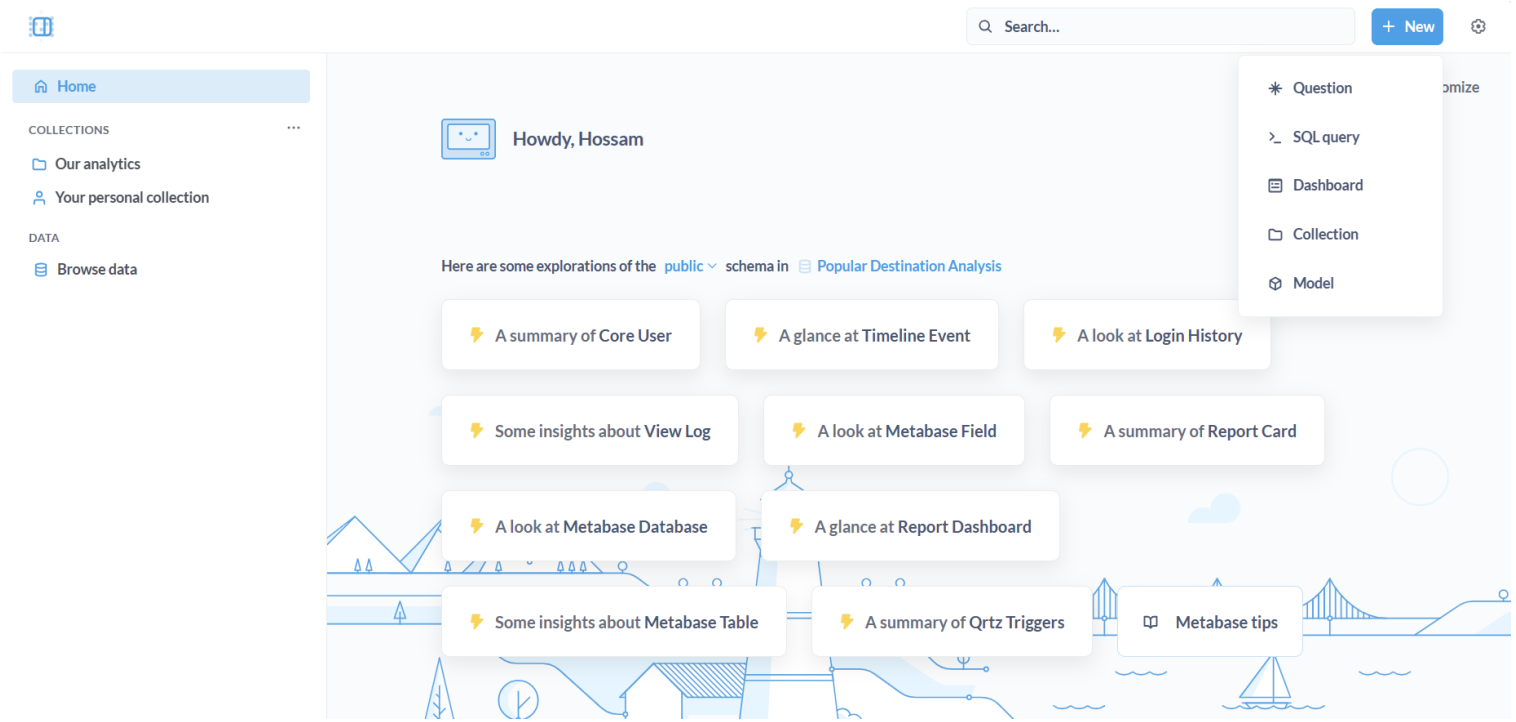
Airflow

- Run and monitor Dag



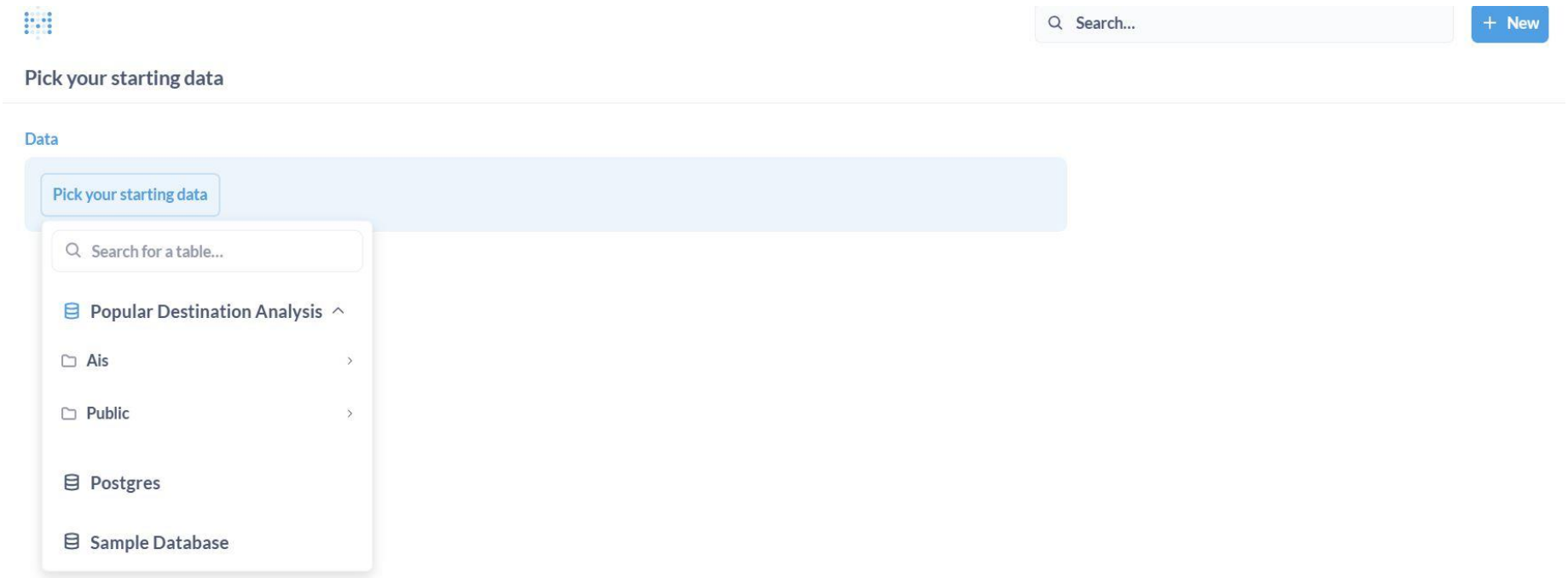
Metabase

- Once the dag is done go to http://localhost:3000/ to access metabase
- Sign up and add the postgres credentials
- Select New > SQL Query



Metabase

- Pick the data that was just added by the Airflow pipeline (Popular Destination Analysis)



Metabase

- Run a Query and visualize the top 10 destinations
- You can choose from a variety of charts to visualize data and create dashboards