

ALS-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:

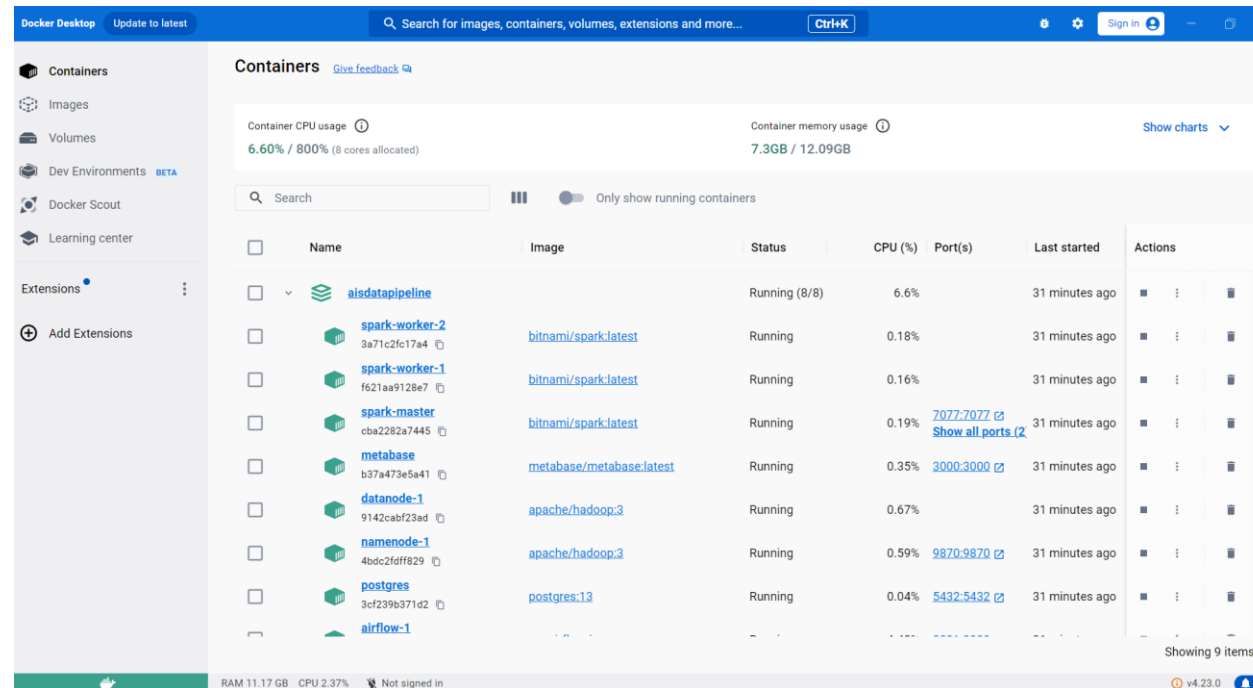
- Add error-handling.
- Visualize with PostgreSQL.
- Regular system monitoring.

Note:

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

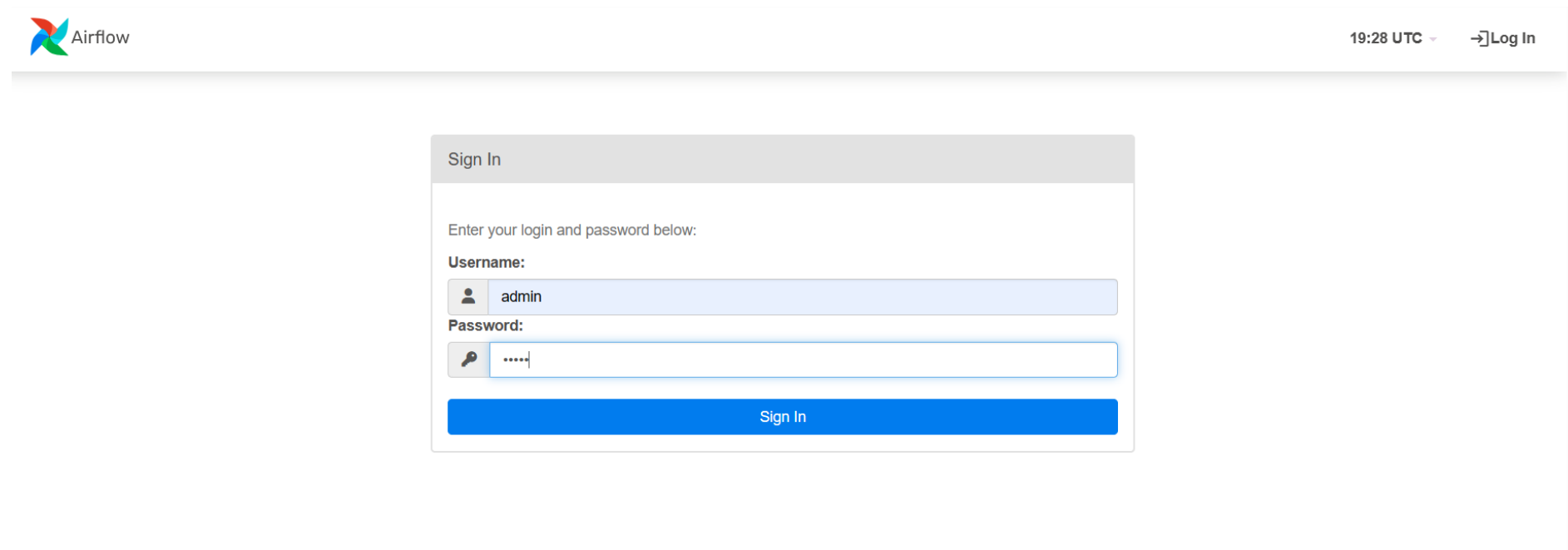
Running containers

- 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/`**



The image shows the Airflow web interface. At the top left is the Airflow logo. At the top right, it displays the time '19:28 UTC' and a 'Log In' link. The main content area features a 'Sign In' form. The form has a title 'Sign In' and a prompt 'Enter your login and password below:'. It contains two input fields: 'Username:' with the value 'admin' and 'Password:' with masked characters '.....'. A blue 'Sign In' button is at the bottom of the form.

Airflow

19:28 UTC → Log In

Sign In

Enter your login and password below:

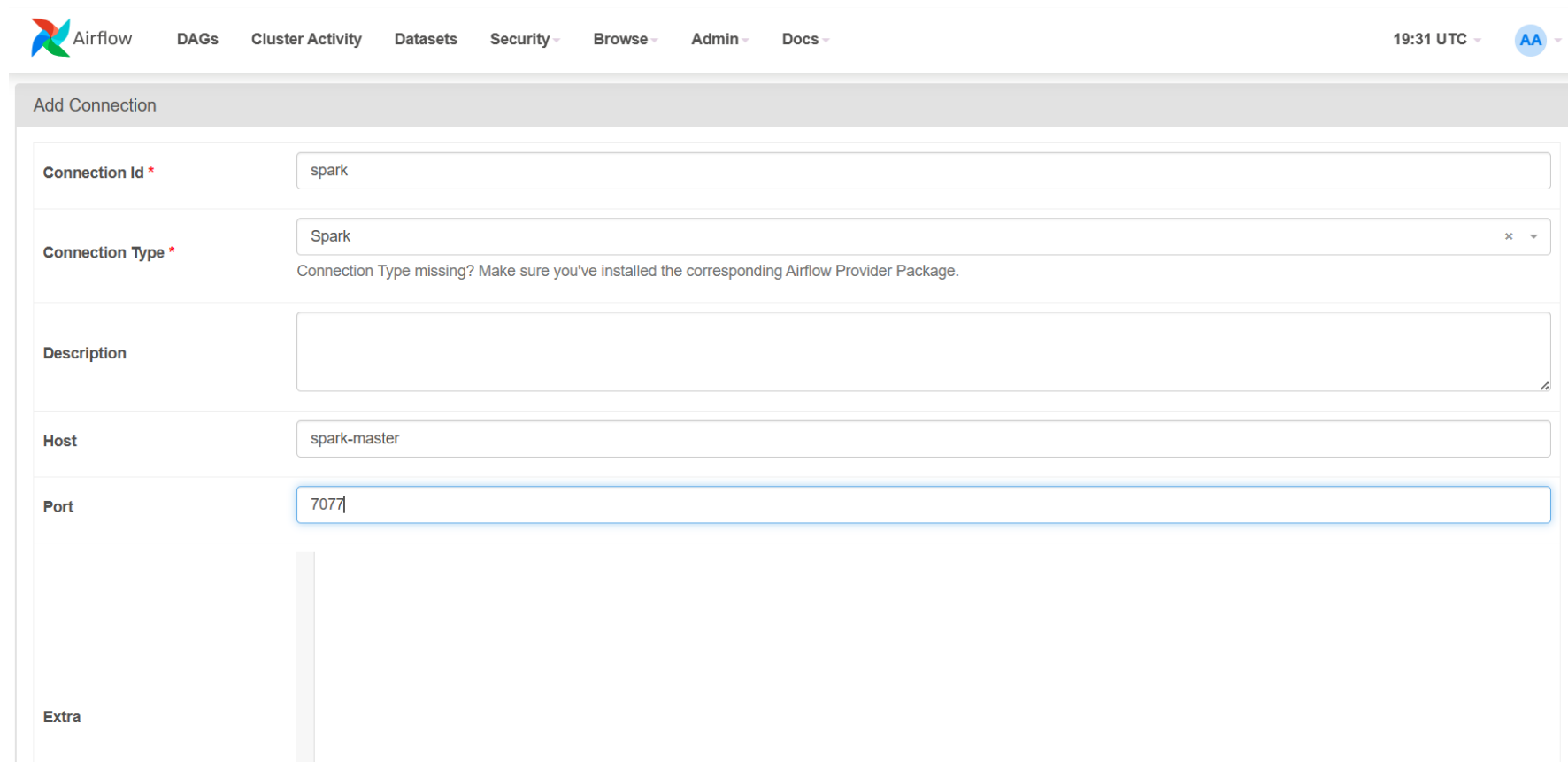
Username: admin

Password:

Sign In

Airflow

- Create a spark connection on



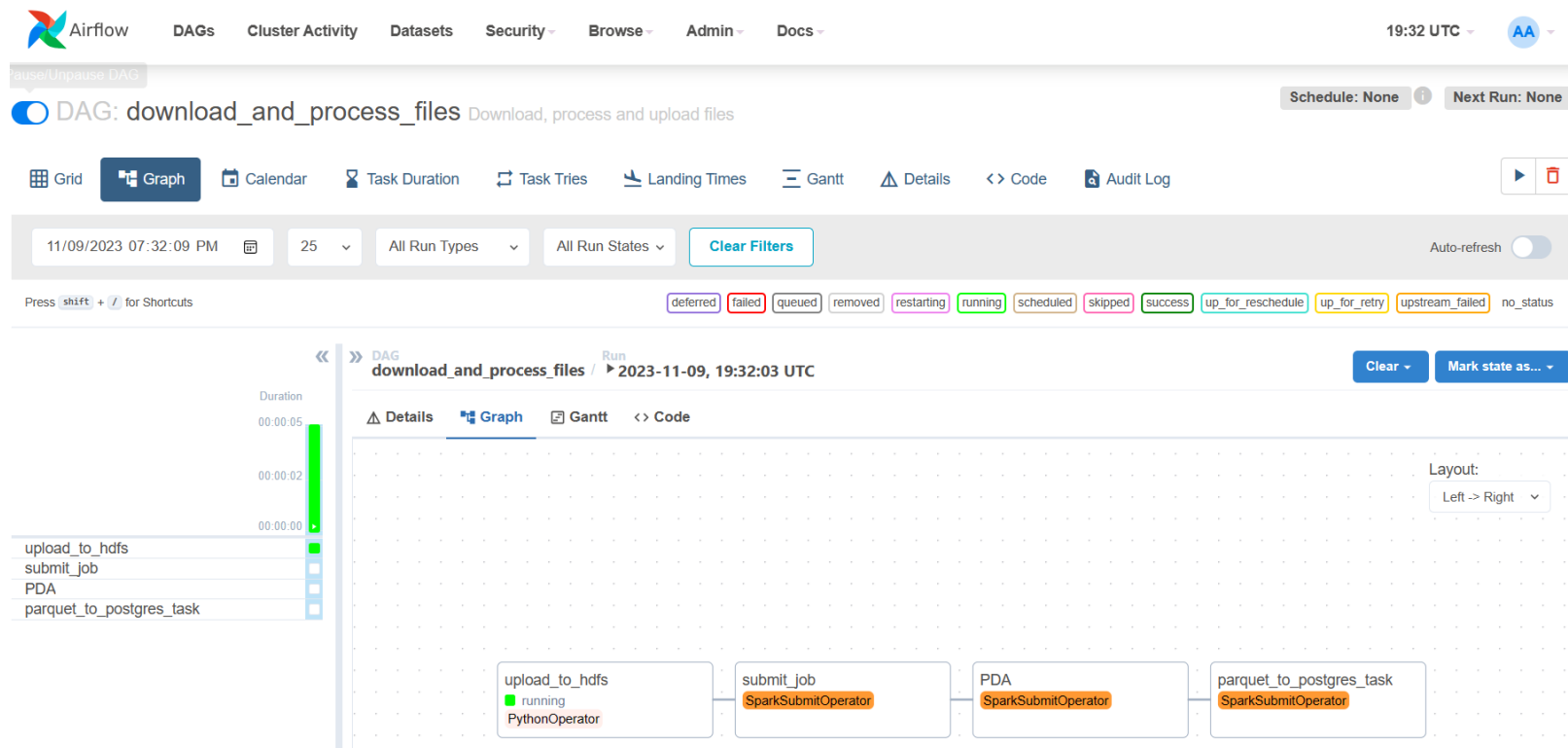
The screenshot shows the Airflow web interface with the 'Add Connection' form. The form is titled 'Add Connection' and contains the following fields:

- Connection Id ***: A text input field containing the value 'spark'.
- Connection Type ***: A dropdown menu with 'Spark' selected. Below the dropdown, a message reads: 'Connection Type missing? Make sure you've installed the corresponding Airflow Provider Package.'
- Description**: A large text area for adding a description.
- Host**: A text input field containing the value 'spark-master'.
- Port**: A text input field containing the value '7077'.
- Extra**: A section for additional configuration, currently empty.

The top navigation bar includes the Airflow logo, links to DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs, the current time (19:31 UTC), and a user profile icon labeled 'AA'.

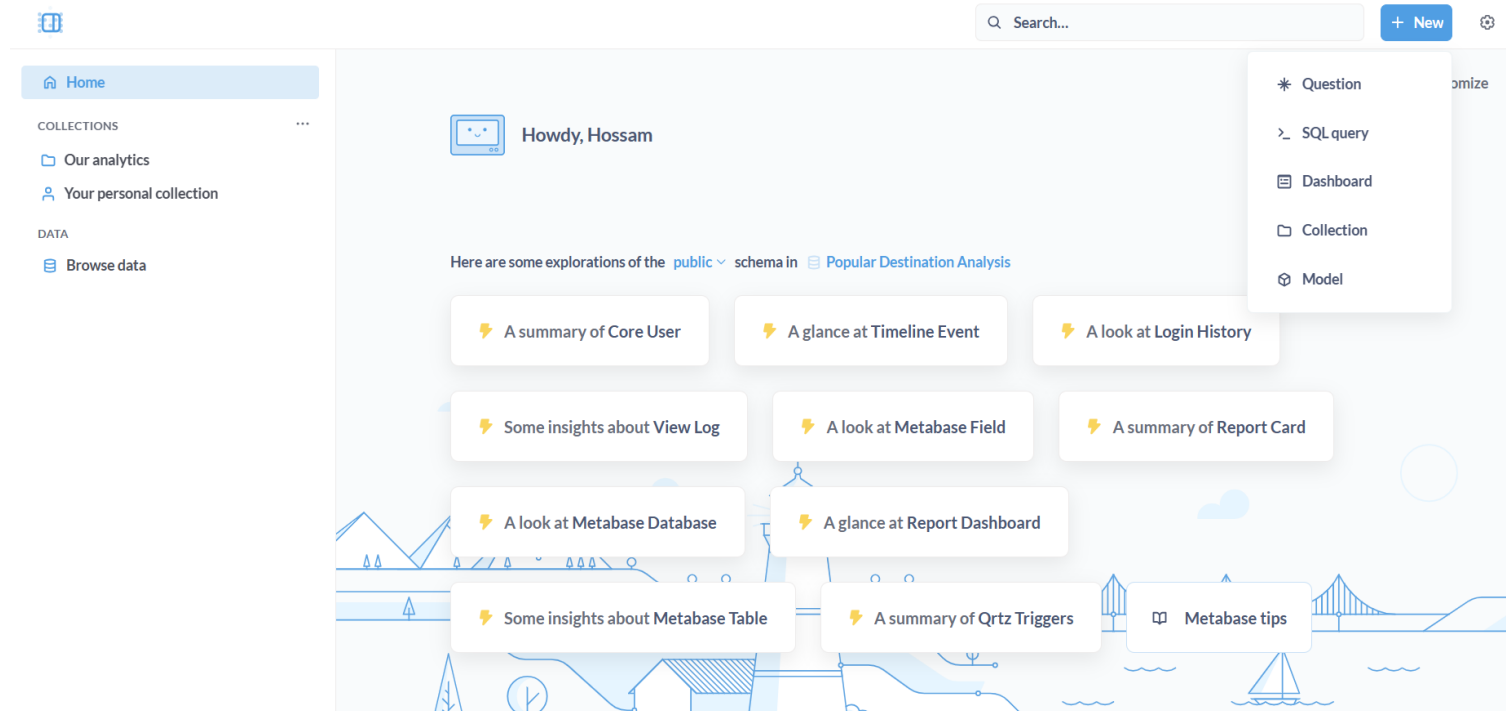
Airflow

- Run and monitor Dag



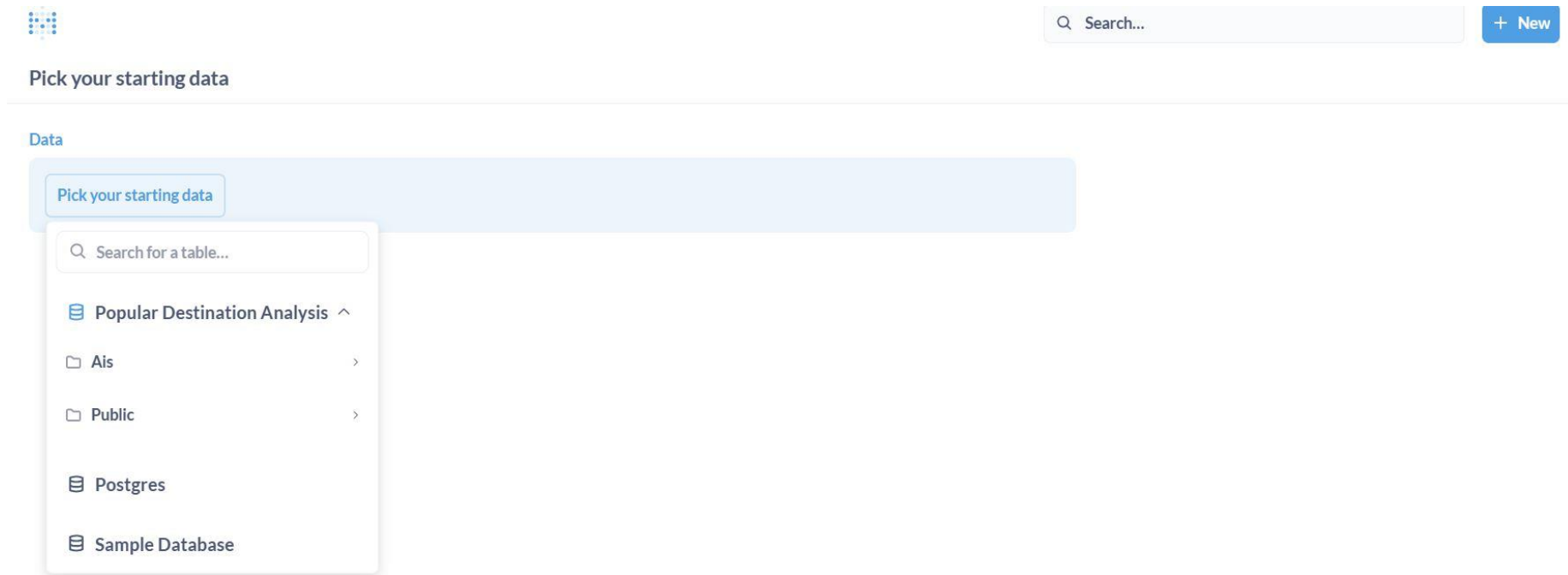
Metabase

- 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

The image shows the Metabase web interface. On the left is a sidebar with a 'New question' section containing icons for various chart types: Table, Bar, Line, Pie, Row, Area, Combo, Funnel, Detail, Map, Scatter, and Waterfall. Below this is an 'OTHER CHARTS' section with icons for Number, Pivot Table, Trend, Gauge, and Progress. A 'Done' button is at the bottom of the sidebar. The main area has a search bar, a '+ New' button, and a 'Save' link. The title 'Popular Destination Analysis' is followed by a dropdown arrow. Below the title is a SQL query editor with the text: `select * from ais.pda order by count desc limit 10`. To the right of the query are icons for a book, a close button, and a refresh button. Below the query editor is a table with two columns: 'destination' and 'Count'. The table contains six rows of data. At the bottom of the table are tabs for 'Visualization' and a settings icon. The bottom right corner shows 'Showing 10 rows' and icons for a refresh and a bell.

New question

Search... + New

Save

Popular Destination Analysis

```
select * from ais.pda order by count desc limit 10
```

destination	Count
Unknown	4,305,892
SKAGEN	205,562
VORDINGBORG	146,590
PLGDN	141,249
ESBJERG	131,899
FISHING GROUNDS	122,897

Visualization

Showing 10 rows