

Photo by [Ian Taylor](#) on [Unsplash](#)This member-only story is on us. [Upgrade](#) to access all of Medium.

★ Member-only story

Airflow, dbt, and Postgres

Tired of recreating docker images for PoC in data engineering? I created one for you.

Tomas Peluritis · [Follow](#)Published in [Geek Culture](#) · 4 min read · Nov 8, 2022

52



Introduction

Not sure how about you, but I like to do side projects quite a lot. Usually, they are to test different technologies/tools, how they interact, and if I could use them to some extent in my work.

The flow for me is to create some docker image or docker-compose, if it consists of interacting apps, to have an isolated and controlled environment. While I'm not an expert in docker, sometimes creating environments is more than 50% of the work of the whole PoC. I run a PoC, get some results, and rarely commit my code to GitHub. A couple of months pass and a new idea pops into my head; suddenly, I can't find where I've put my code. Almost every time I do this, I have the same issue.

To stop this from happening to me, I'll be creating multiple repositories for people to fork, and they can play around with different tools and their interactions as they will.

Airflow

My de-facto go-to scheduler. Using it since ~2019, starting as a simple dag creator to migrating with a team of people from Airflow 1.10.X to 2.1.X and applying best practices. You can read about it in my blog post while I was at HomeToGo:

Apache Airflow at HomeToGo

HomeToGo journey in having data flows orchestrated with Airflow. Pains and issues we encountered and how we tackled





I chose it because I'm the most familiar, but I will not limit myself later only to this, so there will be repositories with other orchestrators too!

The airflow image I create will be more straightforward without Redis and celery. I am doing two different machines, one as a web server, one as a scheduler, and of course, a back-end DB.

Dockerfile for Airflow

As you can see from Dockerfile, I'm using Airflow 2.4.2 with Python 3.9. I'm not going to lie; I chose this Airflow version to test their dataset scheduling later and familiarize myself with their new UI. I stopped checking other versions after 2.2.2.

The code itself is pretty straightforward. I am installing multiple packages, i.e., `libpq-dev` for interactions with Postgres DB, `git` for dbt, and requirements for python packages.

I've created two bash scripts to control the scheduler and webserver starting-up behavior.

scheduler.

Now before you go and judge me:

- I never knew well bash — all comments are more than welcome; glad to improve and learn better practices
- Hard-coding username, password, and IP — at the moment not sure how to do it another way, I guess I could have used some environment variables and added them all over the place, but I already was bashing my head here for more time than I wished

Webservers are way more straightforward:

- create admin user
- start web server

Since I've mounted the dags folder from the repository— put your dags there, and the scheduler will parse them.

dbt

At the moment my go-to tool for transformations. Installation is relatively easy — dbt-core package + dbt-DATABASE package.

Only adjustments to make sustainability I've written down about using separate profiles.YAML file for credentials, for PoC not to interfere with any prod things. I added as well a reminder on how it should look for Postgres.

Create your models in the transform/data_warehouse folder as you would for dbt projects.

Here's the way how to do that, so I gathered up the pieces from stack overflow and some other posts I've found in google.

It's pretty simple — create a networks object with relevant information and then use it in your services.

So each time I spin up my docker-compose, I'd always get my database on the same IP and port.

Running it

Running is also relatively easy if you're familiar with docker-compose.

```
docker-compose build
```

To build your images

```
docker-compose up
```

for running it.



Summary




In general fun and a pleasant experience making this all run. I learned a bit about networks and how to make at least some apps use static IP to mimic real life.

You can find this repository:

GitHub - TomasPel/airflow_dbt_postgres
When running some PoC or trying out how different tools plays
[Docker](#) [Data](#) [Engineering](#) [Development](#) [Database](#)

air/
dbt_postgres

 52 



Written by Tomas Peluritis

646 Followers · Writer for Geek Culture

Professional Data Wizard— Data Engineering/DWH/ETL/BI/Data Science.

Follow



More from Tomas Peluritis and Geek Culture



Tomas Peluritis

Is it the end for Apache Airflow?

Competition, pros and cons and some comparisons.

★ · 14 min read · May 31



349



8



...



Jacob Bennett in Geek Culture

The 5 paid subscriptions I actually use in 2023 as a software engineer

Tools I use that are cheaper than Netflix

★ · 4 min read · Mar 25



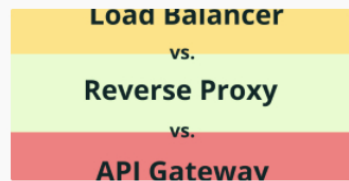
3.7K



37



...



Arslan Ahmad in Geek Culture

Load Balancer vs. Reverse Proxy vs. API Gateway

Understanding the Key Components for Efficient, Secure, and Scalable Web...

★ · 12 min read · May 17



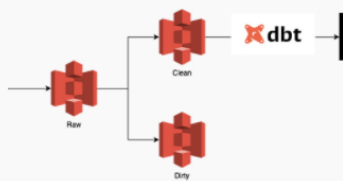
792



5



...



Tomas Peluritis in Geek Culture

Is it the end for Apache Airflow vol. 2

This is the continuation of my previous post/talk I did. In case you want to read it, yo...

★ · 8 min read · Jul 18



32



2



...

See all from Tomas Peluritis

See all from Geek Culture

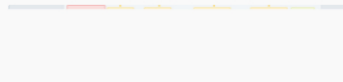
Recommended from Medium



Jacob Bennett in Level Up Coding

Use Git like a senior engineer

Git is a powerful tool that feels great to use when you know how to use it.



Love Shar... in ByteByteGo System Design Allian...

System Design Blueprint: The Ultimate Guide

Developing a robust, scalable, and efficient system can be daunting. However,...

★ · 4 min read · Nov 15, 2022

👍 8.2K 💬 84

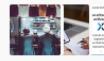


★ · 9 min read · Apr 20

👍 6.9K 💬 55



Lists



Leadership

35 stories · 85 saves



New_Reading_List

174 stories · 41 saves



It's never too late or early to start something

13 stories · 46 saves



Leadership upgrades

7 stories · 15 saves

👤 Ignacio de Gregorio

Microsoft Just Showed us the Future of ChatGPT with LongNet

Let's talk about Billions

★ · 8 min read · Jul 20

👍 2.5K 💬 35



👤 Dominik Polzer in Towards Data Science

All You Need to Know to Build Your First LLM App

A step-by-step tutorial to document loaders, embeddings, vector stores and prompt...

★ · 26 min read · Jun 22

👍 3.4K 💬 31



👤 Youssef Hosni in Level Up Coding

13 SQL Statements for 90% of Your Data Science Tasks

Structured Query Language (SQL) is a programming language designed for...

★ · 15 min read · Feb 27

👍 3.2K 💬 35



👤 The Coding Diaries in The Coding Diaries

Why Experienced Programmers Fail Coding Interviews

A friend of mine recently joined a FAANG company as an engineering manager, and...

★ · 5 min read · Nov 2, 2022

👍 5.8K 💬 119



See more recommendations

