

Data Engineering Zoomcamp FAQ

The purpose of this document is to capture frequently asked technical questions.

General course-related questions

When will the course start?

The exact day and hour of the course will be 16th Jan 2023 at 18h00. The course will start with the first “Office Hours” live.

Subscribe to course [public Google Calendar](#) (it works from Desktop only).

Yes, you can. Register for the course using this [link](#).

Don't forget to register in [DataTalks.Club's Slack](#) and join the [#course-data-engineering](#) channel.

What is the video/zoom link to the stream for the first “Office Hour”?

It should be posted in the announcements channel before it begins. Also, you will see it live on the DataTalksClub YouTube Channel.

I can't attend the “Office Hours” will it be recorded?

Yes! Every “Office Hours” will be recorded so you can attend whenever you want.

What can I do before the course starts?

You can start by installing and setup all the requirements:

- Google cloud account
- Git and GitHub
- Docker desktop with docker-compose
- Python 3 (installed with Anaconda)
- Google Cloud SDK
- Terraform

Is the 2023 cohort different from the 2022 cohort?

Yes. The main difference is the orchestration tool — we will use Prefect and not Airflow. And new homeworks 😊

Why are we using GCP and not other cloud providers?

Because everyone has a google account, GCP has a free trial period and gives \$300 in credits to new users. Also, we are working with BigQuery, which is a part of GCP.

Note that to sign up for a free GCP account, you need to have a valid credit card.

The GCP and other cloud providers are not available in some countries. Is it possible to provide a guide to installing a home lab?

You can do most of the course without a cloud. Almost everything we should (excluding BigQuery) can be run locally. We won't be able to provide guidelines for some things, but most of the materials are runnable without GCP.

For everything in the course, there's a local alternative. You could even do the whole course locally.

Why not AWS?

This course intends to focus on the data engineering processes and not the tools. Besides, the knowledge over one cloud is simply converted to another provider.

I want to use AWS. May I do that?

Yes, you can. Just remember to adapt all the information on the videos to AWS. Besides, the final capstone will be evaluated based on the task: Create a data pipeline! Develop a visualisation!

Besides the “Office Hour” which are the live zoom calls?

We will probably have some calls during the Capstone period to clear some questions but it will be announced in advance if that happens.

I don't want to watch the weekly videos or do homework. Can I still do the final capstone?

Yes :) You can do the final capstone and if you pass it, you will get a certificate.

Are we still using the NYC Trip data for January 2021 or are we using the 2022 data?

We will use the same data, as the project will essentially remain the same as last year's. The data is available [here](#)

Is the 2022 repo deleted?

No, but we moved the 2022 stuff [here](#)

Can I use Airflow instead of Prefect for my final project?

Yes, you can use any tool you want for your project.

Is it possible to use x tool instead of the one tool you use?

Yes, this applies if you want to use Airflow instead of Prefect, AWS or Snowflake instead of GCP products or Tableau instead of Metabase or Google data studio.

The course covers 2 alternative data stacks, one using GCP and one using local installation of everything. You can use either one of those, or use your tool of choice.

You do need to take in consideration that we can't support you if you choose to use a different stack, also you would need to explain the different choices of tool for the peer review of your capstone project.

How can we contribute to the course?

Star the repo! Share it with friends if you find it useful ❤️

Is the course [Windows/mac/Linux/...] friendly?

Yes! Linux is ideal but technically it should not matter. Students last year used all 3 OSes successfully

Any books or additional resources you recommend?

Yes to both! [check out this document](#)

Can I still join the course?

Yes, even if you don't submit the homeworks, you're still eligible for a at the end - as long as you successfully pass the project at the end.

Be aware, however, that there will be deadlines for turning in the final projects. So don't leave everything for the last minute.

For the homework PartB, which code are we to share, I have pasted the terminal output after running terraform apply to the form and a link to the terraform folder on my GitHub, is this okay?

Yes that is [okay](#)

I am trying to run postgres with volume mounting it ran success and the database server started but i could not see the files on my machine for persistency. I am using Ubuntu WSL this is the code I used to run docker:

```
docker run -it \  
  -e POSTGRES_USER="root" \  
  -e POSTGRES_PASSWORD="root" \  
  -e POSTGRES_DB="ny_taxi" \  
  -v $(pwd)/ny_taxi_postgres_data:/var/lib/postgres/data \  
  -p 5432:5432 \  
postgres:13
```

I had to do this to make it work. got it from the readme files on the repo

```
docker run -it \  
-e POSTGRES_USER="root" \  
-e POSTGRES_PASSWORD="root" \  
-e POSTGRES_DB="ny_taxi" \  
-v $(pwd)/ny_taxi_postgres_data:/var/lib/postgresql/data \  
-p 5432:5432 \  
postgres:13  
  
sudo chmod a+rwX ny_taxi_postgres_data
```

I'm using 2 email ids while working on the zoomcamp. While uploading the homework/projects etc are you mapping progress to the email id which is shared as an input in the form or the one used while being logged into Google(required to submit the form)?

You can use any email you want for the homeworks, it doesn't have to be the same as the one you used for signing up. Just make sure you use the same email for all the homeworks

While performing SQL query in python using pandas, I am facing the error : TypeError: __init__() got multiple values for argument 'schema'

Install an earlier version of sqlalchemy (1.4.46 release). SQLAlchemy version 2.0.0 was released recently and isn't compatible with pandasql. If you have the latest version of sqlalchemy (v2.0.0), just do:

- `pip uninstall sqlalchemy`
- `pip install sqlalchemy==1.4.46`

Alternative news source other than Slack

<https://t.me/dezoomcamp>

Week 1

SELECT * FROM zones_taxi WHERE Zone='Astoria Zone'; Error Column Zone doesn't exist

- For the HW1 I encountered this issue. The solution is

```
SELECT * FROM zones AS z WHERE z."Zone" = 'Astoria Zone';
```
- I think columns which start with uppercase need to go between "Column". I ran into a lot of issues like this and " " made it work out.
- Addition to the above point, for me, there is no 'Astoria Zone', only 'Astoria' is existing in the dataset.

```
SELECT * FROM zones AS z WHERE z."Zone" = 'Astoria';
```

SELECT Zone FROM taxi_zones Error Column Zone doesn't exist

- It is inconvenient to use quotation marks all the time, so it is better to put the data to the database all in lowercase, so in Pandas after

```
df = pd.read_csv('taxi+ zone lookup.csv')
```


Add the row:

```
df.columns = df.columns.str.lower()
```

Docker won't start or is stuck in settings (Windows 10 / 11)

- First off, make sure you're running the latest version of Docker for Windows, which you can download from [here](#). Sometimes using the menu to "Upgrade" doesn't work (which is another clear indicator for you to uninstall, and reinstall with the latest version)
- If docker is stuck on starting, first try to switch containers by right clicking the [docker symbol](#) from the running programs and switch the containers from windows to linux or vice versa
- **[Windows 10 / 11 Pro Edition]** The **Pro Edition** of Windows can run Docker either by using Hyper-V or WSL2 as its backend (Docker Engine)
 - In order to use **Hyper-V** as its back-end, you MUST have it enabled first, which you can do by following the tutorial: [Enable Hyper-V Option on Windows 10 / 11](#)
 - If you opt-in for **WSL2**, you can follow the same steps as detailed in the [tutorial here](#)

- **[Windows 10 / 11 Home Edition]** If you're running a **Home Edition**, you can still make it work with WSL2 (Windows Subsystem for Linux) by following the [tutorial here](#)

If even after making sure your WSL2 (or Hyper-V) is set up accordingly, Docker remains stuck, you can **try** the option to [Reset to Factory Defaults](#) or do a **fresh install**.

How to handle taxi data files, now that the files are available as *.csv.gz?

The pandas `read_csv` function can read csv.gz files directly. So no need to change anything in the script.

wget is not recognized as an internal or external command

“wget is not recognized as an internal or external command”, you need to install it.

On Ubuntu, run:

```
$ sudo apt-get install wget
```

On MacOS, the easiest way to install wget is to use [Brew](#):

```
$ brew install wget
```

On Windows, the easiest way to install wget is to use [Chocolatey](#):

```
$ choco install wget
```

Or you can download a binary (<https://gnuwin32.sourceforge.net/packages/wget.htm>) and put it to any location in your PATH (e.g. C:/tools/)

Also, you can following this step to install Wget on MS Windows

* Download the latest wget binary for windows from [eternallybored] (<https://eternallybored.org/misc/wget/>) (they are available as a zip with documentation, or just an exe)

* If you downloaded the zip, extract all (if windows built in zip utility gives an error, use [7-zip] (<https://7-zip.org/>)).

* Rename the file `wget64.exe` to `wget.exe` if necessary.

* Move wget.exe to your `Git\mingw64\bin\`.

Alternatively, you can use a Python wget library, but instead of simply using “wget” you’ll need to use

```
python -m wget
```

You need to install it with pip first:

```
pip install wget
```

Alternatively, you can just paste the file URL into your web browser and download the file normally that way. You’ll want to move the resulting file into your working directory.

Also recommended a look at the python library **requests** for the loading gz file <https://pypi.org/project/requests/>

**docker: Error response from daemon: invalid mode:
\\Program Files\\Git\\var\\lib\\postgresql\\data.**

Change the mounting path. Replace it with the following:

```
-v /e/zoomcamp/...:/var/lib/postgresql/data
```


(Optional) Should I run docker commands from the windows file system or from a file system of a Linux distribution in WSL?

It is recommended by the Docker docs to store all code in your default Linux distro to get the best out of file system performance (since Docker runs on WSL2 backend by default for Windows 10 Home / Windows 11 Home users).

More info in the [Docker Docs on Best Practises](#)

docker: Cannot connect to Docker daemon at unix:///var/run/docker.sock. Is the docker daemon running?

Make sure you're able to start the Docker daemon, and check the issue immediately down below:

docker: Error during connect: In the default daemon configuration on Windows, the docker client must be run with elevated privileges to connect.: Post: "http://%2F%2F.%2Fpipe%2Fdocker_engine/v1.24/containers/create" : open //./pipe/docker_engine: The system cannot find the file specified

As the official [Docker for Windows documentation](#) says, the Docker engine can either use the Hyper-V or WSL2 as its backend. However, a few constraints might apply

- **Windows 10 Pro / 11 Pro Users:**

In order to use **Hyper-V** as its back-end, you MUST have it enabled first, which you can do by following the tutorial: [Enable Hyper-V Option on Windows 10 / 11](#)

- **Windows 10 Home / 11 Home Users:**

On the other hand, Users of the 'Home' version do NOT have the option Hyper-V option enabled, which means, you can only get Docker up and running using the WSL2 (Windows Subsystem for Linux). Url

-

You can find the detailed instructions to do so here:

<https://pureinfotech.com/install-wsl-windows-11/>

In case, you run into another issue while trying to install WSL2

(**WslRegisterDistribution failed with error: 0x800701bc**), Make sure you update the WSL2 Linux Kernel, following the guidelines here:

<https://github.com/microsoft/WSL/issues/5393>

docker: Pull access denied for dbpage/pgadmin4, repository does not exist or may require 'docker login': denied: requested access to the resource is denied

Whenever a `docker pull` is performed (either manually or by `docker-compose up`), it attempts to fetch the given image name (**pgadmin4**, for the example above) from a repository (**dbpage**).

IF the repository is public, the fetch and download happens without any issue whatsoever.

For instance:

- `docker pull postgres:13`
- `docker pull dbpage/pgadmin4`

BE ADVISED:

The Docker Images we'll be using throughout the Data Engineering Zoomcamp are all public (except when or if explicitly said otherwise by the instructors or co-instructors).

Meaning: you are NOT required to perform a docker login to fetch them.

So if you get the message above saying *"docker login": denied: requested access to the resource is denied*. That is most likely due to a **typo** in your image name:

For instance:

```
$ docker pull dbpage/pgadmin4
```

- Will throw that exception telling you "repository does not exist or may require 'docker login'"

```
$ docker pull dbpage/pgadmin4
```

```
✓ base 07:45:46
```

```
Using default tag: latest
```

```
Error response from daemon: pull access denied for dbpage/pgadmin4, repository does not exist or  
may require 'docker login': denied: requested access to the resource is denied
```

- But that actually happened because the actual image is **dp**age/pgadmin4 and NOT **db**page/pgadmin4

How to fix it:

```
$ docker pull dpage/pgadmin4
```

```
$ docker pull dpage/pgadmin4
```

```
Using default tag: latest
```

```
latest: Pulling from dpage/pgadmin4
```

```
a9eaa45ef418: Already exists
```

```
942bbf3d7389: Pull complete
```

```
fbe23c71dc3b: Pull complete
```

```
7c1be9e99602: Pull complete
```

```
ccc31a15f27f: Pull complete
```

```
617b6e01309f: Pull complete
```

```
e6cfa0ba7132: Pull complete
```

```
9dd539b143fa: Pull complete
```

```
6f3ff58d53db: Pull complete
```

```
a79e40a556fb: Pull complete
```

```
b05884a10df3: Pull complete
```

```
3a39531f7518: Pull complete
```

```
0337d3baf297: Pull complete
```

```
c7a9de9c5d61: Pull complete
```

```
Digest: sha256:79b2d8da14e537129c28469035524a9be7cfe9107764cc96781a166c8374da1f
```

```
Status: Downloaded newer image for dpage/pgadmin4:latest
```

```
docker.io/dpage/pgadmin4:latest
```

EXTRA NOTES:

In the real world, occasionally, when you're working for a company or closed organisation, the Docker image you're trying to fetch might be under a private repo that your DockerHub Username was granted access to.

For which cases, you must first execute:

```
$ docker login
```

- Fill in the details of your username and password.
- And only then perform the **`docker pull`** against that private repository

connection failed: :1), port 5432 failed: could not receive data from server: Connection refused could not send SSL negotiation packet: Connection refused

Change

```
pgcli -h localhost -p 5432 -u root -d ny_taxi TO
```

```
pgcli -h 127.0.0.1 -p 5432 -u root -d ny_taxi
```

```
pgcli -h 127.0.0.1 -p 5432 -u root -d ny_taxi
```

Should we run pgcli inside another docker container?

In this section of the course, the 5432 port of postgres is mapped to your computer's 5432 port. Which means you can access the postgres database via pgcli directly from your computer.

So No, you don't need to run it inside another container. Your local system will do.

The input device is not a TTY (Docker run for Windows)

You may have this error:

```
$ docker run -it ubuntu bash
```

the input device is not a TTY. If you are using mintty, try prefixing the command with 'winpty'

Solution:

Use **winpty** before docker command ([source](#))

```
$ winpty docker run -it ubuntu bash
```

You also can make an alias:

```
echo "alias docker='winpty docker'" >> ~/.bashrc
```

OR

```
echo "alias docker='winpty docker'" >> ~/.bash_profile
```

Cannot pip install on Docker container (Windows)

You may have this error:

```
Retrying (Retry(total=4, connect=None, read=None, redirect=None, status=None)) after connection broken by 'NewConnectionError('<pip._vendor.u
```

```
rllib3.connection.HTTPSConnection object at 0x7efe331cf790>: Failed to establish a new connection: [Errno -3] Temporary failure in name resolution')':
```

```
/simple/pandas/
```

Possible solution might be:

```
$ winpty docker run -it --dns=8.8.8.8 --entrypoint=bash python:3.9
```

Setting up Docker on Mac

Check this article for details - [Setting up docker in macOS](#)

From researching it seems this method might be out of date, it seems that since docker changed their licensing model, the above is a bit hit and miss. What worked for me was to just go to the docker website and download their dmg. Haven't had an issue with that method.

1FATAL: password authentication failed for user "root" (You already have Postgres)

FATAL: password authentication failed for user "root"

observations: Below in bold do not forget the folder that was created
ny_taxi_postgres_data

This can happen if you already have Postgres installed on your computer. If it's the case, use a different port, e.g. 5431:

```
-p 5431:5432
```

And use it when connecting with pgcli:

```
pgcli -h localhost -p 5431 -U root -d ny_taxi
```

This will connect you to postgres.

If you want to debug: the following can help (on a MacOS)

To find out if something is blocking your port (on a MacOS):

- You can use the `lsof` command to find out which application is using a specific port on your local machine. ``lsof -i :5432``
- Or list the running postgres services on your local machine with `launchctl`
``launchctl list | grep postgres``

To unload the running service on your local machine (on a MacOS):

- unload the launch agent for the PostgreSQL service, which will stop the service and free up the port
``launchctl unload -w
~/Library/LaunchAgents/homebrew.mxcl.postgresql.plist``
- this one to start it again
``launchctl load -w
~/Library/LaunchAgents/homebrew.mxcl.postgresql.plist``

Changing port from 5432:5432 to 5431:5432 helped me to avoid this error.

OperationalError: (psycopg2.OperationalError) connection to server at "localhost" (:::1), port 5432 failed: FATAL: role "root" does not exist

Can happen when connecting via pgcli

```
pgcli -h localhost -p 5432 -U root -d ny_taxi
```

Or while uploading data via the connection in jupyter notebook

```
engine = create_engine('postgresql://root:root@localhost:5432/ny_taxi')
```

This can happen when Postgres is already installed on your computer. Changing the port can resolve that (e.g. from 5432 to 5431).

To check whether there even is a root user with the ability to login:

- Try: `docker exec -it <your_container_name> /bin/bash`
- And then run: `psql -h localhost -d ny_taxi -U root`

Also, you could change port from **5432:5432** to **5431:5432**

Other solution that worked:

Changing `POSTGRES_USER=root` to `PGUSER=postgres`

Based on this:

<https://stackoverflow.com/questions/60193781/postgres-with-docker-compose-gives-fatal-role-root-does-not-exist-error>

Also `docker compose down`, removing folder that had postgres volume, running `docker compose up` again.

Could not change permissions of directory "/var/lib/postgresql/data": Operation not permitted

```
$ docker run -it\
-e POSTGRES_USER="root" \
-e POSTGRES_PASSWORD="admin" \
-e POSTGRES_DB="ny_taxi" \
-v "/mnt/path/to/ny_taxi_postgres_data":"/var/lib/postgresql/data" \
-p 5432:5432 \
postgres:13
```

CCW

The files belonging to this database system will be owned by user "postgres".

This useThe database cluster will be initialized with locale "en_US.utf8".

The default databerrorase encoding has accordingly been set to "UTF8". xt search configuration will be set to "english".

Data page checksums are disabled.

fixing permissions on existing directory /var/lib/postgresql/data ...
initdb:

error: could not change permissions of directory
"/var/lib/postgresql/data": Operation not permittedvolume

One way to solve this issue is to create a local docker volume and map it to postgres data directory `/var/lib/postgresql/data`

```
$ docker volume create --name dtc_postgres_volume_local -d local
$ docker run -it\
-e POSTGRES_USER="root" \
-e POSTGRES_PASSWORD="root" \
-e POSTGRES_DB="ny_taxi" \
-v dtc_postgres_volume_local:/var/lib/postgresql/data \
-p 5432:5432 \
postgres:13
```

An alternate error could be:

```
initdb: error: directory "/var/lib/postgresql/data" exists but is not empty
If you want to create a new database system, either remove or empty
the directory "/var/lib/postgresql/data" or run initdb
with an argument other than "/var/lib/postgresql/data"
```

Which simply means that you already have a directory with that name and that it's populated with postgres files. You can either remove your local directory and run the command again, or restart the previous container that created it by starting the container with either its docker id or its randomly generated name.

Ideally, you want to be starting these with docker-compose.

invalid reference format: repository name must be lowercase (Mounting volumes with Docker on Windows)

Mapping volumes on Windows could be tricky. The way it was done in the course video doesn't work for everyone.

First, if you have spaces in the path, move your data to some folder without spaces. E.g. if your code is in "C:/Users/Alexey Grigorev/git/...", move it to "C:/git/..."

Try replacing the "-v" part with one of the following options:

- `-v /c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data`
- `-v //c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data`
- `-v /c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data`
- `-v //c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data`
- `--volume //driveletter/path/ny_taxi_postgres_data/:/var/lib/postgresql/data`

Try adding **winpty** before the whole command

```
winpty docker run -it
-e POSTGRES_USER="root"
-e POSTGRES_PASSWORD="root"
-e POSTGRES_DB="ny_taxi"
-v /c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data
-p 5432:5432
postgres:13
```

Try adding quotes:

- -v "/c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data"
- -v "//c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data"
- -v "/c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data"
- -v "//c:/some/path/ny_taxi_postgres_data:/var/lib/postgresql/data"
- -v "c:\some\path\ny_taxi_postgres_data":/var/lib/postgresql/data

Note: (Window) if it automatically create a folder call "ny_taxi_postgres_data;c" suggests you have problems with volume mapping, try delete both folder and replacing "-v" part with other options. for me "//c/" works instead of "/c/". and it will works by automatically create a correct folder call "ny_taxi_postgres_data".

A possible solution to this error would be to use

/"\$(pwd)"/ny_taxi_postgres_data:/var/lib/postgresql/data (with quotes' position varying as in the above list).

Important: note how the quotes are placed.

If none of these options work, you can use a volume name instead of the path:

- -v ny_taxi_postgres_data:/var/lib/postgresql/data

For Mac: You can wrap \$(pwd) with quotes like the highlighted.

```
docker run -it \
-e POSTGRES_USER="root" \
-e POSTGRES_PASSWORD="root" \
-e POSTGRES_DB="ny_taxi" \
-v "$(pwd)"/ny_taxi_postgres_data:/var/lib/postgresql/data \
-p 5432:5432 \
Postgres:13
```

```
docker run -it \
-e POSTGRES_USER="root" \
-e POSTGRES_PASSWORD="root" \
-e POSTGRES_DB="ny_taxi" \
-v "$(pwd)"/ny_taxi_postgres_data:/var/lib/postgresql/data \
-p 5432:5432 \
```

```
postgres:13
```

Source: <https://stackoverflow.com/questions/48522615/docker-error-invalid-reference-for-mat-repository-name-must-be-lowercase>

Persist/Save PGAdmin docker contents on GCP

So one common issue is when you run docker-compose on GCP, postgres won't persist its data to mentioned path for example:

```
services:
  ...
  pgadmin:
    ...
    Volumes:
      - "/pgadmin":"/var/lib/pgadmin:wr"
```

Might not work so in this use you can use Docker Volume to make it persist, by simply changing

```
services:
  ...
  pgadmin:
    ...
    Volumes:
      - pgadmin:/var/lib/pgadmin

volumes:
  pgadmin:
```

PermissionError: [Errno 13] Permission denied: '/some/path/.config/pgcli'

I get this error

```
pgcli -h localhost -p 5432 -U root -d ny_taxi
```

Traceback (most recent call last):

```
File "/opt/anaconda3/bin/pgcli", line 8, in <module>
    sys.exit(cli())
```

```
File "/opt/anaconda3/lib/python3.9/site-packages/click/core.py", line 1128,
in __call__
```

```

    return self.main(*args, **kwargs)
File "/opt/anaconda3/lib/python3.9/site-packages/click/core.py", line
1053, in main
    rv = self.invoke(ctx)
File "/opt/anaconda3/lib/python3.9/site-packages/click/core.py", line 1395,
in invoke
    return ctx.invoke(self.callback, **ctx.params)
File "/opt/anaconda3/lib/python3.9/site-packages/click/core.py", line 754,
in invoke
    return __callback(*args, **kwargs)
File "/opt/anaconda3/lib/python3.9/site-packages/pgcli/main.py", line 880,
in cli

    os.makedirs(config_dir)
File "/opt/anaconda3/lib/python3.9/os.py", line 225, in makedirspython
    mkdir(name, mode)PermissionError: [Errno 13] Permission denied:
'/Users/vray/.config/pgcli'

```

Make sure you install pgcli without sudo.

The recommended approach is to use conda/anaconda to make sure your system python is not affected.

If conda install gets stuck at "Solving environment" try these alternatives:

<https://stackoverflow.com/questions/63734508/stuck-at-solving-environment-on-anaconda>

pgcli error: no pq wrapper available.

```

ImportError: no pq wrapper available.
Attempts made:
- couldn't import psycopg 'c' implementation: No module named
'psycopg_c'
- couldn't import psycopg 'binary' implementation: No module
named 'psycopg_binary'
- couldn't import psycopg 'python' implementation: libpq library
not found

```

Solution:

First, make sure your Python is set to 3.9, at least.

And the reason for that is we have had cases of 'psycopg2-binary' failing to install because of an old version of Python (3.7.3).

0. You can check your current python version with:

```
$ python -V (the V must be capital)
```

1. Based on the previous output, if you've got a 3.9, skip to Step #2
Otherwise, you're better off with a new environment with 3.9

```
$ conda create -n de-zoomcamp python=3.9
```

```
$ conda activate de-zoomcamp
```

2. Next, you should be able to install the lib for postgres like this:

```
...
```

```
$ pip install psycopg2-binary
```

```
...
```

3. Finally, make sure you're also installing pgcli, but use conda for that:

```
...
```

```
$ conda install -c conda-forge pgcli
```

```
...
```

There, you should be good to go now!

ModuleNotFoundError: No module named 'psycopg2'

Issue:

```
In [14]: > engine = create_engine('postgresql://root:root@localhost:5431/ny_taxi')
-----
ModuleNotFoundError                                Traceback (most recent call last)
```

e...

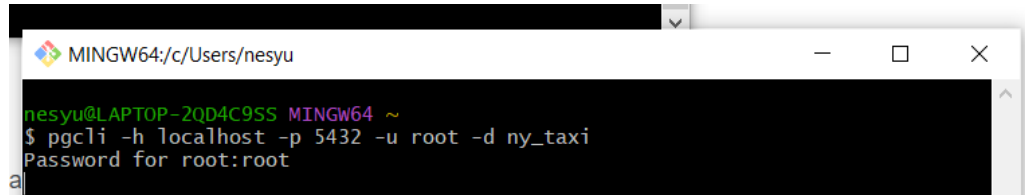
```
ModuleNotFoundError: No module named 'psycopg2'
```

Solution: pip install psycopg2-binary

*if you are still facing error with r psycopg2 and showing pg_config not found then you will have to install postgresql. in MAC it is **brew install postgresql**

pgcli stuck on password prompt

If your Bash prompt is stuck on the password command for postgres

A screenshot of a Windows terminal window titled 'MINGW64: c:/Users/nesyu'. The prompt is 'nesyu@LAPTOP-2QB4C9SS MINGW64 ~'. The command entered is '\$ pgcli -h localhost -p 5432 -u root -d ny_taxi'. Below the command, the text 'Password for root:root' is displayed, and the cursor is positioned at the end of the password, indicating it is stuck on the prompt.

Use winpty: `winpty pgcli -h localhost -p 5432 -u root -d ny_taxi`

Alternatively, try using Windows terminal or terminal in VS code.

Pgcli keeps rejecting my correct password

If you keep getting this error after adding the correct password:

connection failed: :1), port 5432 failed: FATAL: password authentication failed for user "user"

You should change your local port, the 5432 port on your system might have an already running container. So your docker command will be like this:

```
docker run -it\
  -e POSTGRES_USER="root" \
  -e POSTGRES_PASSWORD="root" \
  -e POSTGRES_DB="ny_taxi" \
  -v "/c/Users/user/Downloads/DataTalks
ZoomCamp/data-engineering-zoomcamp/week_1_basics_n_setup/2_docker_sql"
  \
  -p 5433:5432\
  Postgres:14
```

And your pgcli command will be:

Pgcli not loading on Git Bash

If Pgcli keeps freezing after you enter your password, I just used windows powershell instead and ran the same commands and it worked.

pgcli: command not found

Problem: If you have already installed pgcli but bash doesn't recognize pgcli

- On Git bash: bash: pgcli: command not found
- On Windows Terminal: pgcli: The term 'pgcli' is not recognized...

Solution: Try adding a Python path

C:\Users\...\AppData\Roaming\Python\Python39\Scripts to Windows PATH

For details:

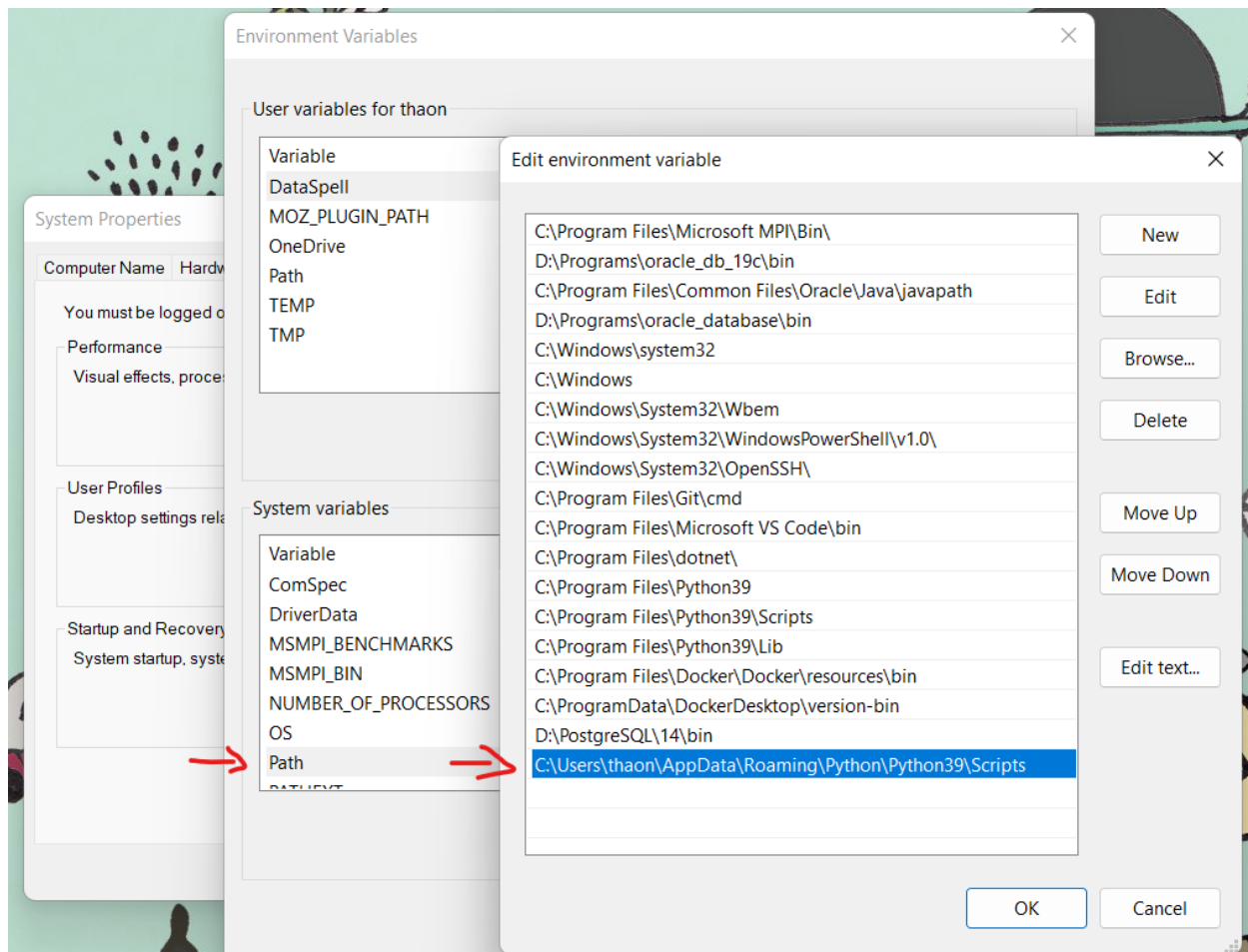
1. Get the location: `pip list -v`
2. Copy `C:\Users\...\AppData\Roaming\Python\Python39\site-packages`
3. 3. Replace site-packages with Scripts:
`C:\Users\...\AppData\Roaming\Python\Python39\Scripts`

It can also be that you have Python installed elsewhere.

For me it was under `c:\python310\lib\site-packages`

So I had to add c:\python310\lib\Scripts to PATH, as shown below.

Put the above path in "Path" (or "PATH") in System Variables



Reference: <https://stackoverflow.com/a/68233660>

OperationalError: (psycopg2.OperationalError) connection to server at "localhost" (:::1), port 5432 failed: FATAL: database "ny_taxi" does not exist

```
~\anaconda3\lib\site-packages\psycopg2\__init__.py in connect(dsn, connection_factory,
cursor_factory, **kwargs)
120
121     dsn = _ext.make_dsn(dsn, **kwargs)
--> 122     conn = _connect(dsn, connection_factory=connection_factory, **kwargsync)
123     if cursor_factory is not None:
124         conn.cursor_factory = cursor_factory
```

OperationalError: (psycopg2.OperationalError) connection to server at "localhost" (:::1), port 5432 failed: FATAL: database "ny_taxi" does not exist

Make sure postgres is running. You can check that by running ``docker ps``

✓ Solution: If you have postgres software installed on your computer before now, build your instance on a different port like 8080 instead of 5432

curl: (6) Could not resolve host: output.csv

Solution (for mac users): `os.system(f"curl {url} --output {csv_name}")`

Jupyter Notebook not opening with error in git bash.

ImportError: DLL load failed while importing _sqlite3: The specified module could not be found. ModuleNotFoundError: No module named 'pysqlite2'

The issue seems to arise from the missing of sqlite3.dll in path ".\Anaconda\DLLs\".

✓ I solved it by simply copying that .dll file from \Anaconda3\Library\bin and put it under the path mentioned above. (if you are using anaconda)

docker build error: error checking context: 'can't stat '/home/user/repos/data-engineering/week_1_basics_n_setup/2_docker_sql/ny_taxi_postgres_data'.

This error appeared when running the command: `docker build -t taxi_ingest:v001 .`

When feeding the database with the data the user id of the directory `ny_taxi_postgres_data` was changed to 999, so my user couldn't access it when running the above command. Even though this is not the problem here it helped to raise the error due to the permission issue.

Since at this point we only need the files `Dockerfile` and `ingest_data.py`, to fix this error one can run the `docker build` command on a different directory (having only these two files).

A more complete explanation can be found here:

<https://stackoverflow.com/questions/41286028/docker-build-error-checking-context-cant-stat-c-users-username-appdata>

Yellow Taxi Trip Records downloading error, Error 403 or XML error webpage

When you try to download the 2021 data from [TLC website](#), you get this error:

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<Error>
  <Code>AccessDenied</Code>
  <Message>Access Denied</Message>
  <RequestId>KA0D58E64XX4W/CDC</RequestId>
  <HostId>7okNtNIhIKLrvGjzCLzdfm+LeGXWDRJ0UNmSUCJLBArWmzFfzKicPVRxf40X0Rb4ToMXvs6mu4s</HostId>
</Error>
```

If you click on the link, and ERROR 403: Forbidden on the terminal.

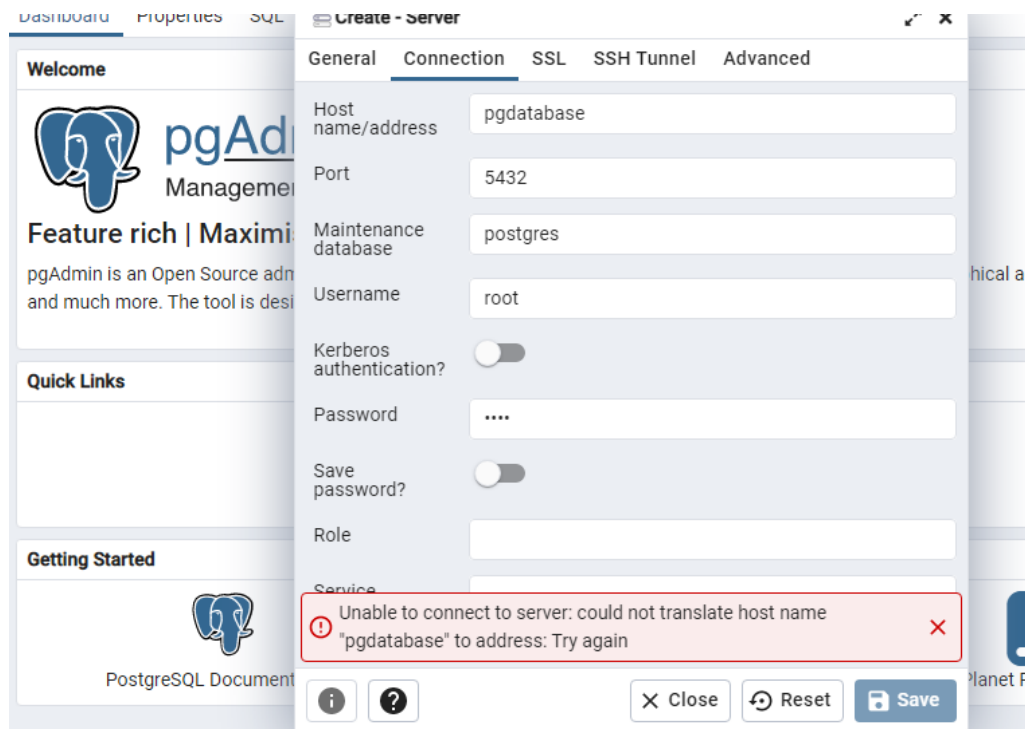
We have a backup, so use it instead: <https://github.com/DataTalksClub/nyc-tlc-data>

Update for Jan 2023, the correct link would be

https://github.com/DataTalksClub/nyc-tlc-data/releases/download/yellow/yellow_tripdata_2021-01.csv.gz

Docker-compose - Error translating host name to address

Couldn't translate host name to address



Make sure postgres database is running.

Use the command to start containers in detached mode: `docker-compose up -d`

```
(data-engineering-zoomcamp) hw % docker compose up -d
[+] Running 2/2
  # Container pg-admin      Started
0.6s
  # Container pg-database   Started
```

To view the containers use: `docker ps.`

```
(data-engineering-zoomcamp) hw % docker ps
```

| CONTAINER ID | IMAGE | COMMAND NAMES | CREATED | STATUS |
|--------------|----------------|---|----------------|---------------|
| faf05090972e | postgres:13 | "docker-entrypoint.s..." pg-database | 39 seconds ago | Up 37 seconds |
| 6344dcecd58f | dpage/pgadmin4 | "/entrypoint.sh" pg-admin | 39 seconds ago | Up 37 seconds |

```
443/tcp, 0.0.0.0:8080->80/tcp
hw
```

To view logs for a container: `docker logs <containerid>`

```
(data-engineering-zoomcamp) hw % docker logs faf05090972e
```

PostgreSQL Database directory appears to contain a database; Skipping initialization

```
2022-01-25 05:58:45.948 UTC [1] LOG:  starting PostgreSQL 13.5 (Debian 13.5-1.pgdg110+1) on
aarch64-unknown-linux-gnu, compiled by gcc (Debian 10.2.1-6) 10.2.1 20210110, 64-bit
2022-01-25 05:58:45.948 UTC [1] LOG:  listening on IPv4 address "0.0.0.0", port 5432
2022-01-25 05:58:45.948 UTC [1] LOG:  listening on IPv6 address ":::", port 5432
2022-01-25 05:58:45.954 UTC [1] LOG:  listening on Unix socket
"/var/run/postgresql/.s.PGSQL.5432"
2022-01-25 05:58:45.984 UTC [28] LOG:  database system was interrupted; last known up at
2022-01-24 17:48:35 UTC
2022-01-25 05:58:48.581 UTC [28] LOG:  database system was not properly shut down; automatic
recovery in
progress
2022-01-25 05:58:48.602 UTC [28] LOG:  redo starts at 0/872A5910
2022-01-25 05:59:33.726 UTC [28] LOG:  invalid record length at 0/98A3C160: wanted 24, got 0
2022-01-25 05:59:33.726 UTC [28] LOG:  redo done at 0/98A3C128
2022-01-25 05:59:48.051 UTC [1] LOG:  database system is ready to accept connections
```

If docker ps doesn't show pgdatabase running, run: `docker ps -a`

This should show all containers, either running or stopped.

Get the container id for pgdatabase-1, and run

Docker-Compose Up - Issues

After executing `docker-compose up` - if you lose database data and are unable to successfully execute your Ingestion script (to re-populate your database) but receive the following error:

```
sqlalchemy.exc.OperationalError: (psycopg2.OperationalError) could not translate host name "pg-database" to address: Name or service not known
```

Docker compose is creating its own default network since it is no longer specified in a docker execution command or file. Docker Compose will emit to logs the new network name. See the logs after executing `docker compose up` to find the network name and change the network name argument in your Ingestion script.

Docker-compose up -d, issue/ Hostname does not resolve

It returns --> `Error response from daemon: network 66ae65944d643fdeb89bd0329f1409dec2c9e12248052f5f4c4be7d1bdc6a3 not found`

Try:

`docker ps -a` to see all the stopped&running containers

`docker rm -f $(docker ps -aq)` to nuke all the containers

Try: `docker-compose up -d` again

On localhost:8080 server → `Unable to connect to server: could not translate host name 'pg-database' to address: Name does not resolve`

Try: new host name, best without “ - ” e.g. pgdatabase

And on `docker-compose.yml`, should `specify docker network & specify the same network in both containers`

```
services:
  pgdatabase:
    image: postgres:13
    environment:
      - POSTGRES_USER=root
      - POSTGRES_PASSWORD=root
```

```

    - POSTGRES_DB=ny_taxi
volumes:
    - "/ny_taxi_postgres_data:/var/lib/postgresql/data:rw"
ports:
    - "5431:5432"
networks:
    - pg-network
pgadmin:
    image: dpage/pgadmin4
    environment:
        - PGADMIN_DEFAULT_EMAIL=admin@admin.com
        - PGADMIN_DEFAULT_PASSWORD=root
    ports:
        - "8080:80"
    networks:
        - pg-network
networks:
    pg-network:
        name: pg-network

```

pgAdmin - Create server dialog does not appear

pgAdmin has a new version. Create server dialog may not appear. Try using register->server instead.

Docker issue - ERRO[0000] error waiting for container: context canceled

You might have installed docker via snap. Run “sudo snap status docker” to verify. If you have “error: unknown command "status", see 'snap help'.” as a response than reinstall docker and install via the [official website](https://docs.docker.com/engine/install/)

Cd Terraform issue - Error acquiring the state lock

<https://github.com/hashicorp/terraform/issues/14513>

Terraform issue - Error 403 : Access denied

| Error: googleapi: Error 403: Access denied., forbidden

Your `$GOOGLE_APPLICATION_CREDENTIALS` might not be pointing to the correct file

run = `export GOOGLE_APPLICATION_CREDENTIALS=~/.gc/YOUR_JSON.json`

And then = `gcloud auth activate-service-account --key-file`

`$GOOGLE_APPLICATION_CREDENTIALS`

Terraform: Do I need to make another service account for terraform before I get the keys (.json file)?

One service account is enough for all the services/resources you'll use in this course. After you get the file with your credentials and set your environment variable, you should be good to go.

Ingestion with Jupyter notebook - missing 100000 records

If you follow the video [1.2.2 - Ingesting NY Taxi Data to Postgres](#) and you execute all the same steps as Alexey does, you will ingest all the data (~1.3 million rows) into the table `yellow_taxi_data` as expected.

However, if you try to run the whole script in the Jupyter notebook for a second time from top to bottom, you will be missing the first chunk of 100000 records. This is because there is a call to the iterator before the while loop that puts the data in the table. The while loop therefore starts by ingesting the second chunk, not the first.

✅ **Solution:** remove the cell “`df=next(df_iter)`” that appears higher up in the notebook than the while loop. The first time `next(df_iter)` is called should be *within* the while loop.

📌 **Note:** As this notebook is just used as a way to test the code, it was not intended to be run top to bottom, and the logic is tidied up in a later step when it is instead inserted into a `.py` file for the pipeline

Python - Iteration csv without error

```
_iter = pd.read_csv(csv_name, iterator=True, chunksize=100000, on_bad_lines='warn',
low_memory=False)

df = next(_iter)
df.lpep_dropoff_datetime = pd.to_datetime(df.lpep_dropoff_datetime)
df.lpep_pickup_datetime = pd.to_datetime(df.lpep_pickup_datetime)

engine = create_engine(f'postgresql://{user}:{password}@{host}:{port}/{db}')

df.to_sql(name=table, con=engine, if_exists='replace')

for chunk in _iter:
    t_start = time()
    tpep_pickup_datetime
    chunk.lpep_dropoff_datetime = pd.to_datetime(chunk.lpep_dropoff_datetime)
    chunk.lpep_pickup_datetime = pd.to_datetime(chunk.lpep_pickup_datetime)
    # print(chunk.head())
    chunk.to_sql(name=table, con=engine, if_exists='append')

    t_end = time()
```

```
print(f"Inserted another chunk, took {t_end - t_start} seconds")
```

GCP - Trying to initialize gcloud sdk:

It asked me to create a project. This should be done from the cloud console. So maybe we don't need this FAQ.

```
WARNING: Project creation failed: HttpError accessing
<https://cloudresourcemanager.googleapis.com/v1/projects?alt=json>: response:
<{'vtpep_pickup_datetime': 'Origin, X-Origin, Referer', 'content-type':
'application/json; charset=UTF-8', 'content-encoding': 'gzip', 'date': 'Mon,
24 Jan 2022 19:29:12 GMT', 'server': 'ESF', 'cache-control': 'private',
'x-xss-protection': '0', 'x-frame-options': 'SAMEORIGIN',
'x-content-type-options': 'nosniff', 'server-timing': 'gfet4t7; dur=189',
'alt-svc': 'h3=":443"; ma=2592000,h3-29=":443"; ma=2592000,h3-Q050=":443";
ma=2592000,h3-Q046=":443"; ma=2592000,h3-Q043=":443"; ma=2592000,quic=":443";
ma=2592000; v="46,43"', 'transfer-encoding': 'chunked', 'status': 409}>,
content <{

  "error": {

    "code": 409,

    "message": "Requested entity alreadytpep_pickup_datetime exists",

    "status": "ALREADY_EXISTS"

  }

}
```

}From Stackoverflow:

<https://stackoverflow.com/questions/52561383/gcloud-cli-cannot-create-project-the-project-id-you-specified-is-already-in-us?rq=1>

Project IDs are unique across all projects. That means if *any* user *ever* had a project with that ID, you cannot use it. testproject is pretty common, so it's not surprising it's already taken.

GCP - The project to be billed is associated with an absent billing account

If you receive the error: “Error 403: The project to be billed is associated with an absent billing account., accountDisabled” It is most likely because you did not enter **YOUR** project ID. The snip below is from video 1.3.2.

```
+ enabled = true
}
}

Plan: 2 to add, 0 to change, 0 to destroy.

Note: You didn't use the -out option to save this plan, so Terraform can't guarantee
"terraform apply" now.
sejalvaidya@Sejals-MBP terraform % terraform apply
var.project
  Your GCP Project ID

Enter a value: global-maxim-338113
```

The value you enter here will be unique to each student. You can find this value on your GCP Dashboard when you login.

Ashish Agrawal

Another possibility is that you have not linked your billing account to your current project

GCP - OR-CBAT-15 ERROR Google cloud free trial account

GCP Account Suspension Inquiry

If Google refuses your credit/debit card, try another - I've got an issue with Kaspi (Kazakhstan) but it worked with TBC (Georgia).

Unfortunately, there's small hope that support will help.

It seems that Pyypl web-card should work too.



Hello Razvodov,

Thank you for your patience with this case.

Please be advised that I have received the response of our specialized team regarding the ticket that we have submitted for review.

Based on the feedback, unfortunately, the returned error indicates that the information provided through the form could not be verified, and therefore, we cannot activate your Cloud Billing account.

We apologize for the inconvenience.

Sincerely,

Mark Angelo
Google Cloud Support

GCP VM - mkdir: cannot create directory '.ssh': Permission denied

I am trying to create a directory but it won't let me do it

```
User1@DESKTOP-PD6UM8A MINGW64 /
```

```
$ mkdir .ssh
```

```
mkdir: cannot create directory '.ssh': Permission denied
```

You should do it in your home directory. Should be your home (~)

Local. But it seems you're trying to do it in the root folder (/). Should be your home (~)

[Link to Video 1.4.1](#)

GCP VM - Error while saving the file in VM via VS Code

```
Failed to save '<file>': Unable to write file  
'vscode-remote://ssh-remote+de-zoomcamp/home/<user>/data_engineering_course/week_2/airflow/dags/<file>' (NoPermissions (FileSystemError):  
Error: EACCES: permission denied, open  
'/home/<user>/data_engineering_course/week_2/airflow/dags/<file>')
```

You need to change the owner of the files you are trying to edit via VS Code. You can run the following command to change the ownership.

```
sudo chown -R <user> <path to your directory>
```


GCP VM - VM connection request timeout

Question: I connected to my VM perfectly fine last week (ssh) but when I tried again this week, the connection request keeps timing out.

✓ Answer: Start your VM. Once the VM is running, copy its External IP and paste that into your config file within the ~/.ssh folder.

```
cd ~/.ssh  
code config ← this opens the config file in VSCode
```

GCP VM - connect to host port 22 no route to host

(reference:

<https://serverfault.com/questions/953290/google-compute-engine-ssh-connect-to-host-ip-port-22-operation-timed-out>)

1. Go to edit your VM.
2. Go to section Automation
3. Add Startup script

...

```
#!/bin/bash  
sudo ufw allow ssh  
...
```

4. Stop and Start VM.

Windows Google Cloud SDK install issue:

for windows if you having trouble install SDK try follow these steps on the link, if you getting this error:

These credentials will be used by any library that requests Application Default Credentials (ADC).

WARNING:

Cannot find a quota project to add to ADC. You might receive a "quota exceeded" or "API not enabled" error. Run `$ gcloud auth application-default set-quota-project` to add a quota project.

For me:

- I reinstalled the sdk using `unzip file "install.bat"`,
- aftDell Precision M4800er successfully checking `gcloud version`,

- run `gcloud init` to set up project before
- you run `gcloud auth application-default login`

https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/week_1_basics_n_setup/1_terraform_gcp/windows.md

Docker build error checking context: can't stat '/home/fhrzn/Projects/..../ny_taxi_postgres_data'

Found the issue in the PopOS linux. It happened because our user didn't have authorization rights to the host folder (which also caused folder seems empty, but it didn't!).

✅ Solution:

Just add permission for everyone to the corresponding folder

```
sudo chmod -R 777 <path_to_folder>
```

Example:

```
sudo chmod -R 777 ny_taxi_postgres_data/
```

Docker failed to solve with frontend dockerfile.v0: failed to read dockerfile: error from sender: open ny_taxi_postgres_data: permission denied.

This happens on Ubuntu/Linux systems when trying to run the command to build the Docker container again.

```
$ docker build -t taxi_ingest:v001 .
```

A folder is created to host the Docker files. When the build command is executed again to rebuild the pipeline or create a new one the error is raised as there are no permissions on this new folder. Grant permissions by running this command;

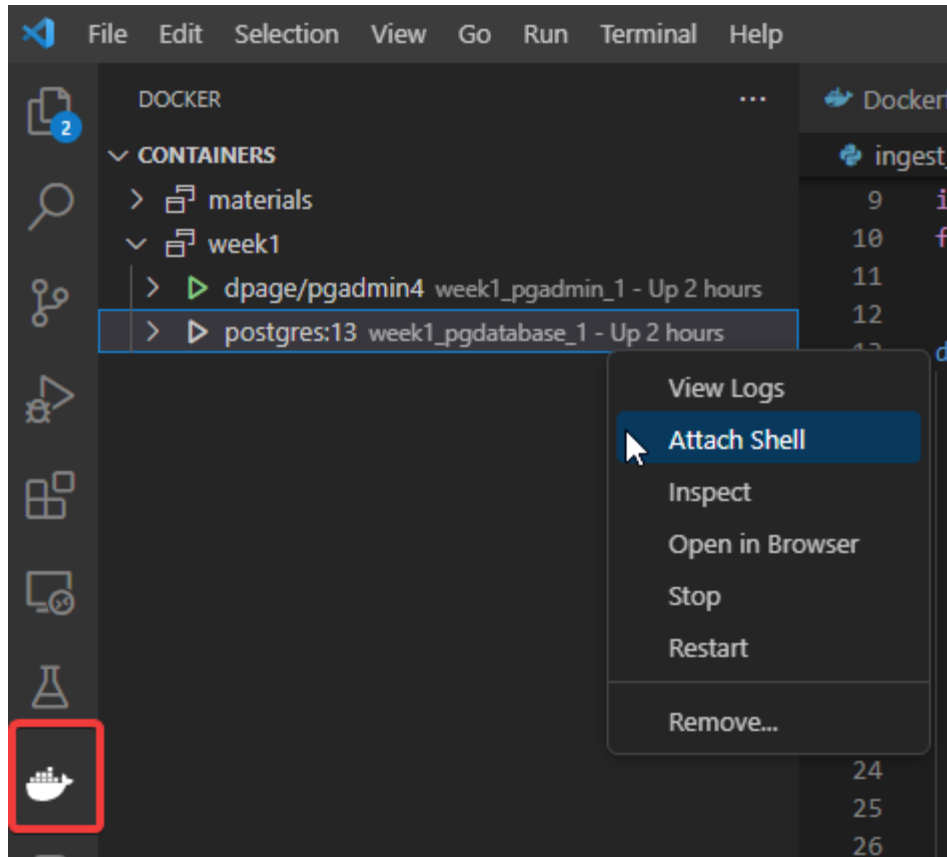
```
$ sudo chmod -R 755 ny_taxi_postgres_data
```

Or use 777 if you still see problems. 755 grants write access to only the owner.

Manage your Docker Infra from VS Code

It's very easy to manage your docker container, images, network and compose projects from VS Code.

Just [install the official extension](#) and launch it from the left side icon.



It will work even if your Docker runs on WSL2, as VS Code can easily connect with your Linux.

Is it necessary to use a GCP VM? When is it useful?

The reason this video about the GCP VM exists is because many students had problems configuring their env. You can use your own env if it works for you.

And advantage of using your own environment is that if you are working in a Github repo where you can commit, you will be able to commit the changes that you do. In the VM the repo is cloned via HTTPS so it is not possible to direct commit, even if you are the owner of the repo.

Docker compose still not available after changing .bashrc

This is happen to me after following 1.4.1 video where we are installing docker compose in our Google Cloud VM. In my case, the docker-compose file downloaded from github

named `docker-compose-linux-x86_64` while it is more convenient to use `docker-compose` command instead. So just change the `docker-compose-linux-x86_64` into `docker-compose`.

Error getting credentials after running docker-compose up -d

Installing pass via 'sudo apt install pass' helped to solve the issue. More about this can be found here: <https://github.com/moby/buildkit/issues/1078>

Docker compose: docker-compose up -d gives an error dial unix /var/run/docker.sock: connect: permission denied:

This happens if you did not create the docker group and added your user. Follow these steps from the

link: <https://github.com/sindresorhus/guides/blob/main/docker-without-sudo.md>

And then press `ctrl+D` to log-out and log-in again. pgAdmin: Maintain state so that it remembers your previous connection

If you are tired of having to setup your database connection each time that you fire up the containers, all you have to do is create a volume for pgAdmin:

In your `docker-compose.yml` file, enter the following into your *pgAdmin* declaration:

```
volumes:
  - type: volume
    source: pgadmin_data
    target: /var/lib/pgadmin
```

Also add the following to the end of the file:

```
volumes:
  Pgadmin_data:
```

Where can I find the Terraform 1.1.3 Linux (AMD 64)?

Here: https://releases.hashicorp.com/terraform/1.1.3/terraform_1.1.3_linux_amd64.zip

I can't find the yellow taxi data. Where is it?

Here:

https://github.com/DataTalksClub/nyc-tlc-data/releases/download/yellow/yellow_tripdata_2021-01.csv.gz

Note: Make sure to [unzip the “gz” file](#) (no, the “unzip” command won’t work for this.)

```
"gzip -d file.gz"
```

Where can I find the “ny-ride.json” file?

The ny-rides.json is your private file in Google Cloud Platform (GCP).

And here’s the way to find it:

GCP -> Select project with your instance -> IAM & Admin -> Service Accounts Keys tab
-> add key, JSON as key type, then click create

Note: Once you go into Service Accounts Keys tab, click the email, then you can see the “KEYS” tab where you can add key as a JSON as its key type

**[In this lecture](#), Alexey deleted his instance in Google Cloud.
Do I have to do it.**

Nope. Do not delete your instance in Google Cloud platform. Otherwise, you have to do this twice for the week 1 readings.

Couldn’t run python ingest_data.py → execute_values() got an unexpected keyword argument ‘fetch’

When tried to run ingest_data.py faced the issue [argument-fetch-error-in-pandas-to-sql](#)
Fixed the problem with

```
$ pip3 install psycopg2-binary==2.8.6 and then run
```

```
$ python3 ingest_data.py \  
--user=root \  
--password=root \  

```

Couldn’t convert python notebook to python script

✓ Solution: `winpty python -m nbconvert [name of notebook] -to python`

Data Dictionary for NY Taxi data?

Yellow Trips:


https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Green Trips:

https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf


No module failure because of pycopg2 library

Fixed the problem with [this blogpost: "No Module Named pycopg2"](#)

- First i did: `pip3 install pycopg2-binary`
- And then: `pip3 install pycopg2-binary --upgrade`
-  And it works for me

PostgreSQL "Column does not exist" but it actually does (Pycopg2 error in MacBook Pro M2)

In the join queries, if we mention the column name directly or enclosed in single quotes it'll throw an error says "column does not exist".

 Solution: But if we enclose the column names in double quotes then it will work

Docker ingestion when using docker-compose could not translate host name

Typical error: sqlalchemy.exc.OperationalError: (psycopg2.OperationalError) could not translate host name "pgdatabase" to address: Name or service not known

When running `docker-compose up -d` see which network is created and use this for the ingestions script instead of pg-network and see the name of the database to use instead of pgdatabase

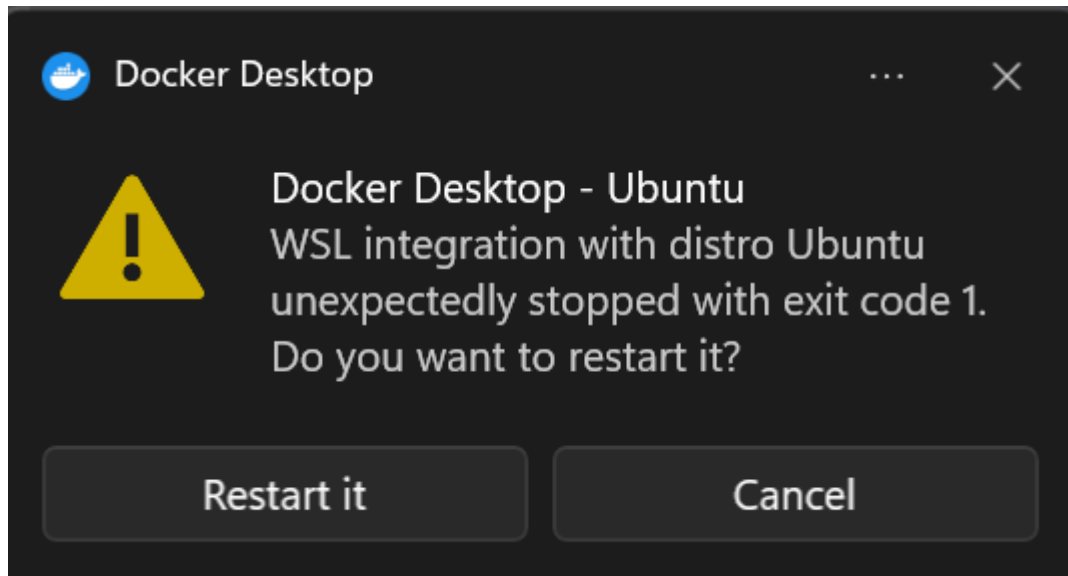
E.g.:

- pg-network becomes 2docker_default
- Pgdatabase becomes 2docker-pgdatabase-1

What is my Docker network name (solution for mac) ?

Get the network name via: \$ [docker network ls](#).

Docker on Windows WSL: WSL integration with distro Ubuntu unexpectedly stopped with exit code 1.



Up restarting the same issue appears. Happens out of the blue on windows.

Solution 1: Fixing DNS Issue (credit: [reddit](#)) this worked for me personally

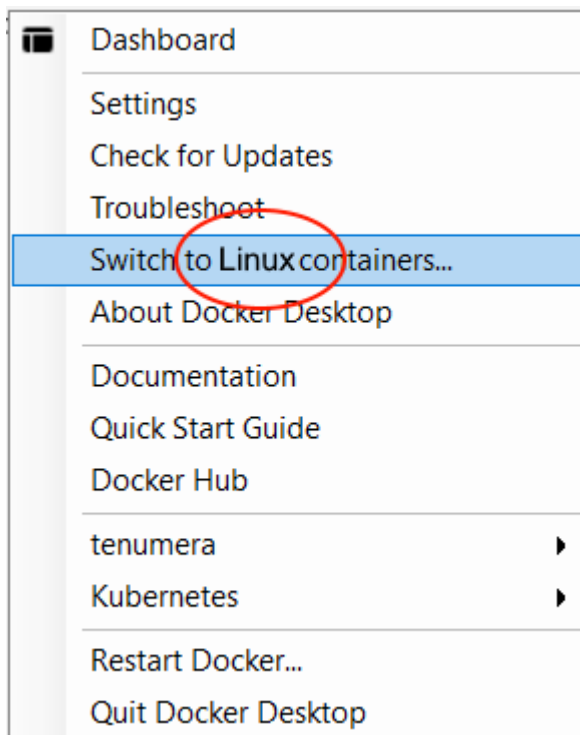
```
reg add "HKLM\System\CurrentControlSet\Services\Dnscache" /v "Start" /t REG_DWORD /d "4" /f
```

Restart your computer and then enable it with the following

```
reg add "HKLM\System\CurrentControlSet\Services\Dnscache" /v "Start" /t REG_DWORD /d "2" /f
```

Restart your OS again. It should work.

Solution 2: right click on running Docker icon (next to clock) and chose "Switch to Linux containers"



Docker: Error response from daemon: Conflict. The container name "pg-database" is already in use by container "xxx". You have to remove (or rename) that container to be able to reuse that name.

Sometimes, when you try to restart a docker image configured with a network name, the above message appears. In this case, use the following command with the appropriate container name:

>>> If the container is running state, use `docker stop <container_name>`

>>> then, `docker rm pg-database`

Or use `docker start` instead of `docker run` in order to restart the docker image without removing it.

Backslash as an escape character in Git Bash for Windows

For those who wish to use the backslash as an escape character in Git Bash for Windows (as Alexey normally does), type in the terminal: `bash.escapeChar=\\` (no need to include in .bashrc)

PGCLI is case sensitive use “Quotations” around columns with capital letters

PULocationID will not be recognized but “PULocationID” will be. This is because unquoted identifiers are case insensitive. [See docs](#).

Pandas can read *.csv.gzip

When a CSV file is compressed using Gzip, it is saved with a ".csv.gz" file extension. This file type is also known as a Gzip compressed CSV file. When you want to read a Gzip compressed CSV file using Pandas, you can use the `read_csv()` function, which is specifically designed to read CSV files. The `read_csv()` function accepts several parameters, including a file path or a file-like object. To read a Gzip compressed CSV file, you can pass the file path of the ".csv.gz" file as an argument to the `read_csv()` function.

Here is an example of how to read a Gzip compressed CSV file using Pandas:

```
import pandas as pd

df = pd.read_csv('path/to/file.csv.gz', compression='gzip')
```

If you prefer to keep the uncompressed csv (easier preview in vscode and similar), gzip files can be unzipped using gunzip (but not unzip). On a Ubuntu local or virtual machine, you may need to apt-get install gunzip first.

How to iterate through and ingest parquet file

Contrary to pandas's `read_csv` method there's no such easy way to iterate through and set chunksize for parquet files. We can use PyArrow (Apache Arrow Python bindings) to resolve that.

```
import pyarrow.parquet as pq
```

```
output_name =
"https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2021-01.parquet"
```

```
parquet_file = pq.ParquetFile(output_name)
parquet_size = parquet_file.metadata.num_rows
```

```
engine=
create_engine(f'postgresql://{user}:{password}@{host}:{port}/{db}')
```

```

table_name="yellow_taxi_schema"

# Clear table if exists
pq.read_table(output_name).to_pandas().head(n=0).to_sql(name=table_name,
con=engine, if_exists='replace')

# default (and max) batch size
index = 65536

for i in parquet_file.iter_batches(use_threads=True):
    t_start = time()
    print(f'Ingesting {index} out of {parquet_size} rows ({index /
parquet_size:.0%})')
    i.to_pandas().to_sql(name=table_name, con=engine,
if_exists='append')
    index += 65536
    t_end = time()
    print(f'\t- it took %.1f seconds' % (t_end - t_start))

```

How do I use Git / GitHub for this course?

After you create a GitHub account, you should clone the course repo to your local machine using the process outlined in this video:

<https://www.youtube.com/watch?v=CKcqniGu3tA>

Having this local repository on your computer will make it easy for you to access the instructors' code and make pull requests (if you want to add your own notes or make changes to the course content).

You will probably also create your own repositories that host your notes, versions of your file, to do this. Here is a great tutorial that shows you how to do this:

<https://www.atlassian.com/git/tutorials/setting-up-a-repository>

Remember to ignore large database, .csv, and .gz files, as well as other files that should not be saved to a repository. Use .gitignore for this:

<https://www.atlassian.com/git/tutorials/saving-changes/gitignore> NEVER store passwords or keys in a git repo (even if that repo is set to private).

This is also a great resource: <https://dangitgit.com/>

Terraform initialized in an empty directory! The directory has no Terraform configuration files. You may begin working with Terraform immediately by creating Terraform configuration files.

You get this error because I run the command `terraform init` outside the working directory, and this is wrong. You need first to navigate to the working directory terraform that contains terraform configuration files, and then run the command.

Which docker-compose binary to use for WSL?

To figure out which docker-compose you need to download from <https://github.com/docker/compose/releases> you can check your system with these commands:

- `uname -s` -> return Linux most likely
- `uname -m` -> return "flavor"

Or try this command -

```
sudo curl -L "https://github.com/docker/compose/releases/download/1.29.2/docker-compose-$(uname -s)-$(uname -m)" -o /usr/local/bin/docker-compose
```

ERROR: Could not find a version that satisfies the requirement psycopg2 (from versions: none). No matching distribution found for psycopg2

May appear because of several reasons:

1. Pip installing a wrong package. The package name should be **psycopg2-binary**
2. Using pip instead of **pip3**. So, try `pip3 install psycopg2-binary`

Port forwarding from GCP VM to Ubuntu local machine without using Visual Studio Code nor any client.

You can easily forward the ports of pgAdmin, postgres and Jupyter Notebook using the built-in tools in Ubuntu and without any additional client:

1. First, in the VM machine, launch `docker-compose up -d` and `jupyter notebook` in the correct folder.
2. From the local machine, execute: `ssh -i ~/.ssh/gcp -L 5432:localhost:5432 username@external_ip_of_vm`
3. Execute the same command but with ports 8080 and 8888.

4. Now you can access pgAdmin on local machine in browser typing `localhost:8080`
5. For Jupyter Notebook, type `localhost:8888` in the browser of your local machine. If you have problems with the credentials, it is possible that you have to copy the link with the access token provided in the logs of the terminal of the VM machine when you launched the `jupyter notebook` command.

Cannot install docker on MacOS/Windows 11 VM running on top of Linux (due to Nested virtualization).

Run this command before starting your VM:

- On Intel CPU:
`modprobe -r kvm_intel`
`modprobe kvm_intel nested=1`
- On AMD CPU:
`modprobe -r kvm_amd`
`modprobe kvm_amd nested=1`

WSL Error - “ Insufficient system resources exist to complete the requested service. ”

Cause:

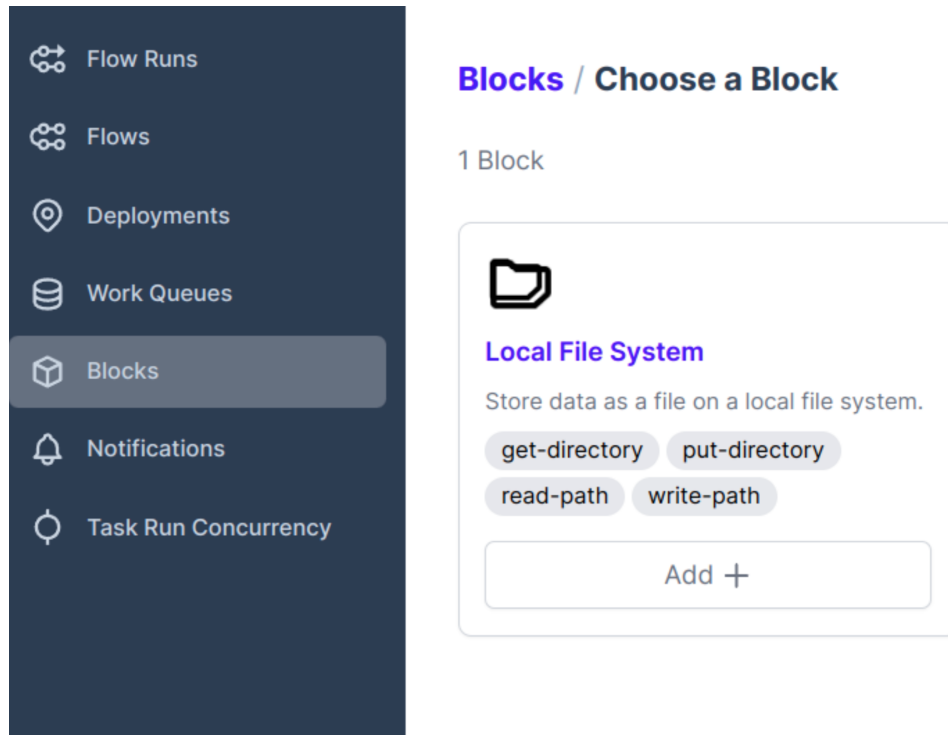
It happens because the apps are not updated. To be specific, search for any pending updates for Windows Terminal, WSL and Windows Security updates.

Solution

- for updating Windows terminal which worked for me:
 1. Go to Microsoft Store.
 2. Go to the library of apps installed in your system.
 3. Search for Windows terminal.
 4. Update the app and restart your system to see the changes.
- For updating the Windows security updates:
 1. Go to Windows updates and check if there are any pending updates from Windows, especially security updates.
 2. Do restart your system once the updates are downloaded and installed successfully.

Week 2

Prefect Blocks are not showing even after installing the requirements.txt as shown in the video. How to fix:



With the Prefect Orion up & running, execute:

- `prefect orion database reset`
- `prefect block register -m prefect_sqlalchemy`
- `prefect block register -m prefect_gcp`

Why and for what do we need to work with Conda environments exactly? What does this environment give us in terms of ability?

Python and many other languages that advise on a virtual env suffer from a major flaw called dependency hell. In a nutshell, that means:

It's common sense that different projects have different requirements, right? That could mean different python versions, different libraries, frameworks, etc

So imagine you have a single Python installation (the one that comes with the operating system, for example), and you're invited to work in a Project A and in a Project B

Project A Dependencies:

- A1
- A2
- A3

Project B Dependencies:

- B1
- B2
- B3

Now, a dependency, at the end of the day, also make use of other dependencies.

(These "implicit" dependencies of the libraries/frameworks you are using in your project are called: transient dependencies)

Now, with that in mind:

A1 depends on X version 1.0

Meaning "X", v1.0, is a transient dependency of A1

B1 depends on X version 2.0

Meaning "X", v2.0, is a transient dependency of B1

When you are working on the same "python environment", the rule that applies is:

Always keep the most updated version of a library. So when you ask your system to install the dependency "A1", it will also bring "X" at version 1.0

But then you switch to another project, called B, which also depends on X - but at version 2.0

That means that this Python environment will upgrade "X" from v1 to v2 to make it work for B1. But the thing is: when you hop between major versions of a library, things often break.

Meaning: B1 will work just fine, because its transient dependency (X v2.0) is satisfied.

But A1 may stop working if v2.0 is incompatible with its previous version (v1.0).

The scenario I just said is quite simple, but in the real world, you're going to be working with different versions of Python and other sets of constraints. But instead we had many ppl in here with Python at 3.7 unable to install the project dependencies for psycopg2, and once they updated to Python 3.9, everything started to work fine.

But what if you DO have another project that is running in production on Python 3.7 and starts bugging out ? The best you can do to reproduce/replicate said environment is to make yourself an equivalent environment (w/ 3.7 instead of 3.9)... and the list goes on and on and on

So long story short,

It is considered a best practice to prevent catastrophic project issues like the ones I listed above, not because "it's a best practice because it's a best practice".

Besides, you NEVER EVER want to mess up with Python environment that comes with your Operating System (macOS / Linux) -> many many system tools today use Python, and if you break it the one that comes bundled with the OS, you're sure gonna have a lot of headache (as in: unexpected behaviors) to put the pieces back together

Hence why, once again, you use virtualenvs to provide you with isolation. Not only between your projects, but also, between the projects and the underlying infrastructure from the OS

WHICH is why you don't use virtualenvs for containers.

Did you notice that the Dockerfile you're using already comes with Python and we didn't actually have to install conda in there?

That's because the containers are not only meant to be disposable if/when they break, but they also run in an isolated workspace of their own (but this is new and entire different discussion)

Why does Jeff have the same line twice in video 2.2.3 and 2.2.4, etl_web_to_gcs.py and parameterized_flow.py?

```
df["tpep_pickup_datetime"] =  
pd.to_datetime(df["tpep_pickup_datetime"])
```

Second one should be:

```
df["tpep_dropoff_datetime"] =  
pd.to_datetime(df["tpep_dropoff_datetime"])
```

Repo is updated. Thank you jralduaveuthey and Valentine Zaretsky for catching!

Why does Jeff use a default parameter assignment in etl_web_to_gcs.py instead of a type hint around time 12:40 in video 2.2.3 and in 2.2.4?

That's a typo. Should be:

```
@task(log_prints=True)  
def clean(df: pd.DataFrame) -> pd.DataFrame:
```

not

```
@task(log_prints=True)
def clean(df=pd.DataFrame) -> pd.DataFrame:
```

Repo is updated. Thank you Valentine Zaretsky for catching!

I have the following error:
raise RuntimeError(

RuntimeError: Unable to load 'de-zoomcamp-gcs' of block type None due to failed validation. To load without validation, try loading again with `validate=False`.

Probable cause of error:

Copied json key for credentials incorrectly or the service account doesn't have the necessary permissions

How to fix it:

Create a new Service Account with the permissions necessary, copy the json key, paste it into the credentials block and finally execute the pipeline again.

This should be fixed!

With Windows, Prefect-gcp 0.2.3 converted / slashes in a path to \ in the to_path statement

Bug with prefect-gcp 0.2.3 on Windows only. Couldn't upload the file into a folder as in the video.

✅ **SOLUTION:** Use **prefect-gcp 0.2.4** You can specify the new version in *requirements.txt* before installing or **pip install -U prefect-gcp** to upgrade in an existing environment.

Then use **path = Path(path).as_posix()** before the upload command.

(TimeoutError)requests.exceptions.ConnectionError: ('Connection aborted.', timeout('The write operation timed out'))

I was hitting the following error in the `gcs_block.upload_from_path` function

The solution for me was to set the `timeout` parameter of the function to 120 (seconds).
`gcs_block.upload_from_path(from_path=path, to_path=path, timeout=120)`

The default timeout is 60 seconds. Timeout may vary depending on your internet speed.

You could also opt to work from a VM.

ValueError: Path


opt/prefect/C:\Users\user\.prefect\storage/5eeca69056a042a284e87ea46f757188 does not exist.

Error when you run the command `prefect deployment run etl-parent-flow/docker-flow -p "months=[1,2]"`

It looks like you ran the flow before with caching on. Now when it tries to find the cached location, it's outside Docker and can't be accessed. If using Prefect 2.7.8 you can refresh the cache and it should work.

Alternatively, you can set the `cache_key_expiration` to a short period of time - say a minute - and rerun outside Docker and then in Docker, it won't try to find the cached result.

See more instructions in the [docs](#).

1.  Deleted the part that use the cache from the fetch task and the flow runs, i'll provide further details if i find why it was trying to pull from local cache

```
# @task(retries=3, cache_key_fn=task_input_hash,
cache_expiration=timedelta(days=1))
@task(retries=3)
def fetch(dataset_url: str) -> pd.DataFrame:
```

If using Docker with a function that says to look for a cached run, but you have cache stored locally outside Docker from a previous run, Docker can't access the cached file, so it throws an error.

2. Also, make sure the Prefect Block you have for the GCS creds is having the service account JSON data directly instead of the path to the service account JSON file

3. made sure I removed the CACHE settings on the @task and I re-ran every command (building docker image, pushing, updating deployment) works!!!

For more info, you can also refer this slack thread for the same error -

<https://datatalks-club.slack.com/archives/C01FABYF2RG/p1674928505612379>

ERROR | Task run 'write_local' - Encountered exception during execution

Remember, you have to create the folders where you keep the files from the repository.


ERROR | Flow run 'xxxxxx' - Finished in state Failed('Flow run encountered an exception.

google.api_core.exceptions.Forbidden: 403 GET: Access Denied: Table xxxxx: Permission bigquery.tables.get denied on table xxxxxx (or it may not exist).\n')

If you reuse the block with the service account which you created before to connect to Cloud Storage Bucket you have to add permissions of BigQuery Administrator.

Attempt to run deployment etl-parent-flow/docker-flow

prefect.exceptions.ScriptError: Script at 'parameterized_flow.py' encountered an exception: FileNotFoundError(2, 'No such file or directory')

Please help with answer here  -> place parameterized_flow.py in flows folder

Trying to build the docker image throws this error:

=> ERROR [5/5] RUN mkdir /opt/prefect/data/yellow 0.3s

—

> [5/5] RUN mkdir /opt/prefect/data/yellow: #9 0.299 mkdir: cannot create directory '/opt/prefect/data/yellow': No such file or directory

—— executor failed running [/bin/sh -c mkdir /opt/prefect/data/yellow]: exit code: 1

Check dockerfile and add -p to mkdir command:

```
RUN mkdir -p /opt/prefect/data/yellow
```

Thank you! Fixed in repo now.

OSError: Cannot save file into a non-existent directory: '/data/yellow'

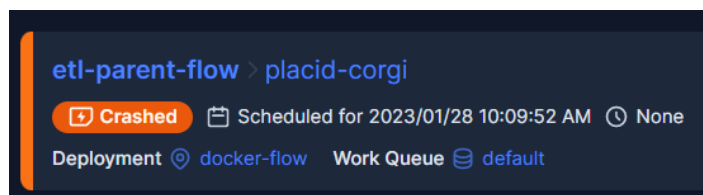
Change the path to go back 2 directories or however many you need so it goes to the proper path

```
path = Path(f"../../data/{color}/{dataset_file}.parquet")
```

etl-parent-flow/docker-flow Crashed with ConnectionRefusedError: [Errno 111] Connect call failed ('127.0.0.1', 4200)

This error occur when you run your prefect on WSL 2 on Window 10. The error looks like this:

UI



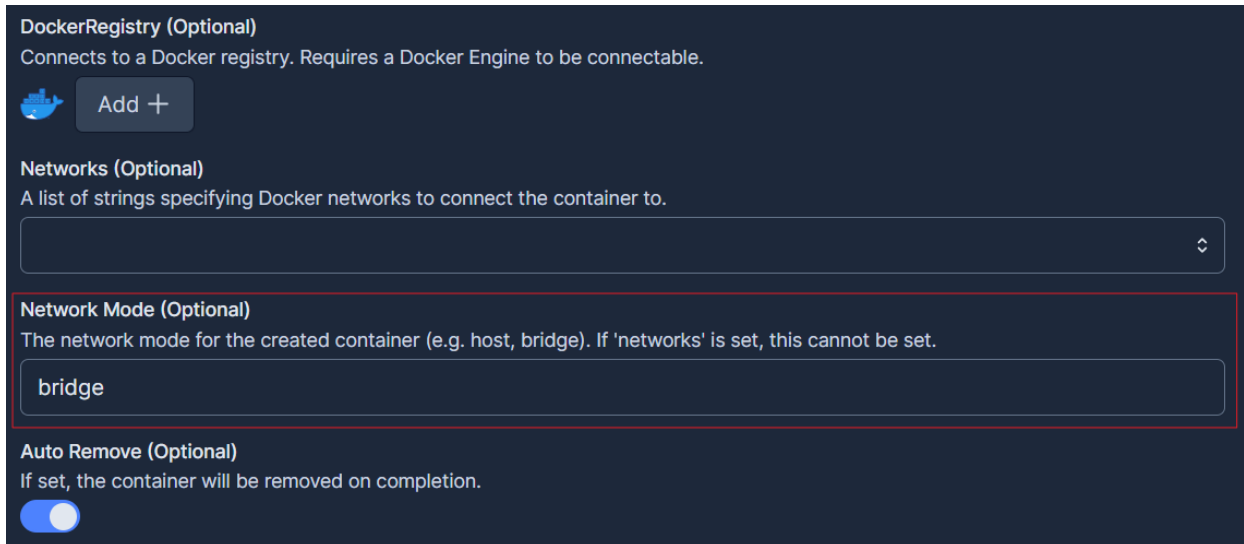
Prefect Agent Log

```
httpx.ConnectError: All connection attempts failed
10:10:16.423 | INFO      | prefect.infrastructure.docker-container -
Docker container 'placid-corgi' has status 'removing'
10:10:16.437 | INFO      | prefect.infrastructure.docker-container -
Docker container 'placid-corgi' has status 'removing'
10:10:16.502 | INFO      | prefect.agent - Reported flow run
'24feeb23-6eb9-4523-b330-19c365bd68fc' as crashed: Flow run
infrastructure exited with non-zero status code 1.
```

The error above is because we run Prefect locally on our machine at localhost:4200, when we run docker without specifying their network, docker call the localhost:4200 inside the container but not the localhost:4200 on our machine.

To solve this, you only need to specify **Network Mode** to **bridge**.

Config from UI



The screenshot shows the Prefect Docker container configuration interface. It includes sections for 'DockerRegistry (Optional)', 'Networks (Optional)', 'Network Mode (Optional)', and 'Auto Remove (Optional)'. The 'Network Mode (Optional)' section is highlighted with a red border and contains a text input field with the value 'bridge'.

Config from Code (make_docker_block.py)

```
from prefect.infrastructure.docker import DockerContainer
```

```
# alternative to creating DockerContainer block in the UI
```

```
docker_block = DockerContainer(  
    image="discdiver/prefect:zoom", # insert your image here  
    image_pull_policy="ALWAYS",  
    auto_remove=True,  
    network_mode="bridge"  
)
```

```
docker_block.save("zoom", overwrite=True)
```

Why are we writing the .csv locally in etl_web_to_gcs.py?

Just to do some transformation. It is not necessary.

ERROR | Flow could not be retrieved from deployment

When using the GitHub block, you have to run the prefect deployment build command in the same local folder as the root folder in the GitHub repository.

In the build command, you have to provide the path to the python file:
path/to/my_file.py:my_flow

--path is used as the upload path, which doesn't apply for GitHub repository-based storage.

Why cant I use same path as Jeff does which is “`gcs_path = f'data/{color}/{color}_tripdata_{year}-{month:02}.parquet'`” instead I need to have “`gcs_path = f'../../data/{color}/{color}_tripdata_{year}-{month:02}.parquet'`” in order to have same structure as in video but then I get the same path in GCS? I am using Linux machine to follow the Zoomcamp.

Problem with prefect with macOS M1 (arm) with poetry the package manager, I got the below message:

I got this problem while tried to run ``prefect orion start`` to start the prefect ui on my laptop

```
“ ValueError: the greenlet library is required to use this function.  
No module named 'greenlet' “
```

For the people who got the same problem with poetry I solve this issue by add ``greenlet`` to ``pyproject.toml``

So run this command:
`poetry add greenlet`

“Localhost (127.0.0.1), port 5433 failed: Connection refused.”

Instructor using port **5433** for week2. If you're using previous containers you should set port **5432** and user/pass **root/root**

Why when I used again terraform to create the infrastructure I got an error?

The error:

```
Error: googleapi: Error 403: Access denied., forbidden
```

```
|
```

and

```
| Error: Error creating Dataset: googleapi: Error 403: Request had  
insufficient authentication scopes.
```

For this solution make sure to run:

```
echo $GOOGLE_APPLICATION_CREDENTIALS
```

```
echo $?
```

If you get 1 then....

Solution:

You have to set again the `GOOGLE_APPLICATION_CREDENTIALS` as Alexey did in the environment set-up video in week1:

```
export
```

```
GOOGLE_APPLICATION_CREDENTIALS="<path/to/your/service-account-authkeys  
>.json"
```

Client error '404 Not Found'

Error Message:

```
httpx.HTTPStatusError: Client error '404 Not Found' for url
```

```
'http://ephemeral-orion/api/flow_runs/6aed1011-1599-4fba-afad-b8d999cf  
9073'
```

For more information check: <https://httpstatuses.com/404>

Solution:

[reference](#)

[method-1]

Input the below command on local PC.

```
prefect config set PREFECT_API_URL=http://127.0.0.1:4200/api
```

[method-2]

Add the below lines in the Dockerfile

```
ENV PREFECT_API_URL=http://127.0.0.1:4200/api
```

```
RUN prefect config set PREFECT_API_URL="${PREFECT_API_URL}"
```

(if you need, you can set the Env in Prefect Orion, like below. And you can change the value depending on your demand, such as connection to Prefect Cloud)

Week 3

Docker-compose takes infinitely long to install zip unzip packages for linux, which are required to unpack datasets

A:

1 solution) Add `-Y` flag, so that apt-get automatically agrees to install additional packages

2) Use python ZipFile package, which is included in all modern python distributions

If you're having problems loading the FHV_2021 data from the github repo into GCS and then into BQ (input file not of type parquet), you need to do two things. First, append the URL Template link with '?raw=true' like so:

```
URL_TEMPLATE = URL_PREFIX + "/fhv_tripdata_{  
execution_date.strftime('%Y-%m') }.parquet?raw=true"
```

Second, update make sure the URL_PREFIX is set to the following value:

```
URL_PREFIX =  
"https://github.com/alexeygrigorev/datasets/blob/master/nyc-tlc/fhv"
```

It is critical that you use this link with the keyword blob. If your link has 'tree' here, replace it. Everything else can stay the same, including the curl -sSLf command.

I am having problems with columns datatype while running DBT/BigQuery

R: If you don't define the column format while converting from csv to parquet Python will "choose" based on the first rows.

✓ **Solution:** Defined the schema while running web_to_gcp.py pipeline.

Sebastian adapted the script:

https://github.com/sebastian2296/data-engineering-zoomcamp/blob/main/week_4_analytics_engineering/web_to_gcs.py

Same ERROR - When running dbt run for fact_trips.sql, the task failed with error:

"Parquet column 'ehail_fee' has type DOUBLE which does not match the target cpp_type INT64"

Reason: Parquet files have their own schema. Some parquet files for green data have records with decimals in ehail_fee column.

There are some possible fixes:

Drop ehail_fee column since it is not really used. For instance when creating a partitioned table from the external table in BigQuery

```
SELECT * EXCEPT (ehail_fee) FROM...
```

Modify stg_green_tripdata.sql model using this line cast(0 as numeric) as ehail_fee.

Modify Airflow dag to make the conversion and avoid the error.

```
pv.read_csv(src_file,  
convert_options=pv.ConvertOptions(column_types = {'ehail_fee':  
'float64'}))
```

Problem with prefect with macOS M1 (arm)

In [video 2.2.2](#) at 15:16 when running the command `python ingest_data_flow.py`, I got a big error message who said this:

```
ValueError: the greenlet library is required to use this function....  
... is an incompatible architecture (have 'x86_64', need 'arm64'))
```

My computer is a MacBook Pro M1 and miniconda arm64 installed.

I searched for the arm version of greenlet but couldn't find it, so I found the following solution instead.

I used the instructions from this site [How to Manage Conda Environments on an Apple Silicon M1 Mac](#) to create a conda x86 environment.

So instead of running the command:

```
conda create -n zoom python=3.9
```

Instead, I ran the command:

```
create_x86_conda_environment myenv_x86 python=3.9
```

This solution works! And I can now continue to prefect on my arm computer!

Week 4

When running your first dbt model, if it fails with an error: 404 Not found: Dataset was not found in location US

R: Go to BigQuery, and check the location of BOTH

1. The source dataset (trips_data_all), and
2. The schema you're trying to write to (name should be <first initial><last name>)

Likely, your source data will be in your region, but the write location will be a multi-regional location (US in this example). Delete these datasets, and recreate them with your specified region and the correct naming format.

Alternatively, instead of removing datasets, you can specify the single-region location you are using. E.g. instead of 'location: US', specify the region, so 'location: US-east1'. See [this Github comment](#) for more detail. Additionally please see [this post of Sandy](#)

In **DBT cloud** you can actually specify the location using the following steps:

1. **Go** to your profile page (top right drop-down --> profile)

2. Then **go** to under Credentials --> Analytics (you may have customised this name)
3. **Click** on Bigquery >
4. **Hit** Edit
5. **Update** your location, you may need to re-upload your service account JSON to re-fetch your private key, and **save**.

When executing dbt run after fact_trips.sql has been created, the task failed with error:

R: "Access Denied: BigQuery BigQuery: Permission denied while globbing file pattern."

1. Fixed by adding the Viewer role to the service account in use in BigQuery.
2. Add the related roles to the service account in use in GCS.

Why do my Fact_trips only contain a few days of data?

Make sure you use:

```
dbt run --var 'is_test_run: false' or dbt build --var 'is_test_run: false'
```

 (watch out for formatted text from this document: re-type the single quotes)

BigQuery returns an error when I try to run the dm_monthly_zone_revenue.sql model.

R: After the second SELECT, change this line:

```
date_trunc('month', pickup_datetime) as revenue_month,
```

To this line:

```
date_trunc(pickup_datetime, month) as revenue_month,
```

Make sure that "month" isn't surrounded by quotes!

Error thrown by format_to_parquet_task when converting fhv_tripdata_2020-01.csv using Airflow

R: This conversion is needed for the question 3 of homework, in order to process files for fhv data. The error is:

```
pyarrow.lib.ArrowInvalid: CSV parse error: Expected 7 columns, got 1: B02765
```

Cause: Some random line breaks in this particular file.

Fixed by opening a bash in the container executing the dag and manually running the following command that deletes all \n not preceded by \r.

```
perl -i -pe 's/(?<!\r)\n/\1/g' fhv_tripdata_2020-01.csv
```

After that, clear the failed task in Airflow to force re-execution.

Why do we need the Staging dataset?

Vic created three different datasets in the videos.. dbt_<name> was used for development and you used a production dataset for the production environment. What was the use for the staging dataset?

R: Staging, as the name suggests, is like an intermediate between the raw datasets and the fact and dim tables, which are the finished product, so to speak. You'll notice that the datasets in staging are materialised as views and not tables.

Vic didn't use it for the project, you just need to create production and dbt_name + trips_data_all that you had already.

My main branch on dbt suddenly changed to read-only... How do I change it back while working on DBT, the branch of the project

R: Since you are on the main branch, it doesn't allow you to change. Just create a new branch to keep going

DBT Docs Served but Not Accessible via Browser

Try removing the “network: host” line in docker-compose.

Week 5

RuntimeError: Java gateway process exited before sending its port number

After installing all including pyspark (and it is successfully imported), but then running this script on the jupyter notebook

```
import pyspark
from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder \
    .master("local[*]") \
    .appName('test') \
    .getOrCreate()
```

```
df = spark.read \
    .option("header", "true") \
    .csv('taxi+_zone_lookup.csv')
```

```
df.show()
```

it gives the error:

RuntimeError: Java gateway process exited before sending its port number

✅ The solution (for me) was:

- `pip install findspark` on the command line and then
- Add

```
import findspark
findspark.init()
```

to the top of the script.

Another possible solution is:

- Check that pyspark is pointing to the correct location.
- Run `pyspark.__file__`. It should be `list /home/<your user name>/spark/spark-3.0.3-bin-hadoop3.2/python/pyspark/__init__.py` if you followed the videos.
- If it is pointing to your python site-packages remove the pyspark directory there and check that you have added the correct exports to your `.bashrc` file and that there are not any other exports which might supersede the ones provided in the course content.

Module Not Found Error in Jupyter Notebook .

Even after installing pyspark correctly on linux machine (VM) as per course instructions, faced a module not found error in jupyter notebook .

The solution which worked for me (use following in jupyter notebook) :

```
!pip install findspark
```

```
Import findspark
```

```
findspark.init()
```

Thereafter , import pyspark and create spark context as usual

None of the solutions above worked for me till I ran `!pip3 install pyspark` instead `!pip install pyspark`.

ModuleNotFoundError: No module named 'py4j' while executing `import pyspark`

Make sure that the version under ``${SPARK_HOME}/python/lib/`` matches the filename of py4j or you will encounter `ModuleNotFoundError: No module named 'py4j'`` while executing ``import pyspark``.

For instance, if the file under ``${SPARK_HOME}/python/lib/`` was ``py4j-0.10.9.3-src.zip``.

Then the `export PYTHONPATH` statement above should be changed to ``export PYTHONPATH="${SPARK_HOME}/python/lib/py4j-0.10.9.3-src.zip:$PYTHONPATH "`` appropriately.

Additionally, you can check for the version of 'py4j' of the spark you're using from [here](#) and update as mentioned above.

Exception: Jupyter command `jupyter-notebook` not found.

Even after we have exported our paths correctly you may find that even though Jupyter is installed you might not have Jupyter Notebook for one reason or another. Full instructions are found [here](#) (for my walkthrough) or [here](#) (where I got the original instructions from) but are included below. These instructions include setting up a virtual environment (handy if you are on your own machine doing this and not a VM):

Full steps:

1. Update and upgrade packages:
 - a. `sudo apt update && sudo apt -y upgrade`
2. Install Python:
 - a. `sudo apt install python3-pip python3-dev`
3. Install Python virtualenv:
 - a. `sudo -H pip3 install --upgrade pip`
 - b. `sudo -H pip3 install virtualenv`
4. Create a Python Virtual Environment:
 - a. `mkdir notebook`
 - b. `cd notebook`
 - c. `virtualenv jupyterenv`
 - d. `source jupyterenv/bin/activate`
5. Install Jupyter Notebook:

a. `pip install jupyter`

6. Run Jupyter Notebook:

a. `jupyter notebook`

Error java.io.FileNotFoundException

Code executed:

```
df = spark.read.parquet(pq_path)
```

```
... some operations on df ...
```

```
df.write.parquet(pq_path, mode="overwrite")
```

```
java.io.FileNotFoundException: File  
file:/home/xxx/code/data/pq/fhvhv/2021/02/part-00021-523f9ad5-14af-4332-9434-bdcb0831  
f2b7-c000.snappy.parquet does not exist
```

The problem is that Sparks performs lazy transformations, so the actual action that trigger the job is `df.write`, which does delete the parquet files that is trying to read (`mode="overwrite"`)

✅ Solution: Write to a different directory

```
df.write.parquet(pq_path_temp, mode="overwrite")
```

Which type of SQL is used in Spark? Postgres? MySQL? SQL Server?

Actually Spark SQL is one independent “type” of SQL - Spark SQL.

The several SQL providers are very similar:

```
SELECT [attributes]
```

```
FROM [table]
```

```
WHERE [filter]
```

GROUP BY [grouping attributes]
HAVING [filtering the groups]
ORDER BY [attribute to order]
(INNER/FULL/LEFT/RIGHT) JOIN [table2]
ON [attributes table joining table2] (...)

What differs the most between several SQL providers are built-in functions.

For Built-in Spark SQL function check this link:

<https://spark.apache.org/docs/latest/api/sql/index.html>

Extra information on SPARK SQL :

<https://databricks.com/glossary/what-is-spark-sql#:~:text=Spark%20SQL%20is%20a%20Spark,on%20existing%20deployments%20and%20data.>

The spark viewer on localhost:4040 was not showing the current run

✓ Solution: I had two notebooks running, and the one I wanted to look at had opened a port on localhost:4041.

If port is in use, then Spark uses next. It can be even 4044. You can run

```
spark.sparkContext.uiWebViewUrl
```

and result will be some like

```
'http://172.19.10.61:4041'
```

java.lang.NoSuchMethodError: sun.nio.ch.DirectBuffer.cleaner()Lsun/misc/Cleaner Error during repartition call (conda pyspark installation)

✓ Solution: replace Java Developer Kit 11 with Java Developer Kit 8.

RuntimeError: Java gateway process exited before sending its port number

Shows java_home is not set on the notebook log

<https://sparkbyexamples.com/pyspark/pyspark-exception-java-gateway-process-exited-before-sending-the-driver-its-port-number/>

Spark fails when reading from BigQuery and using `.show()` on `SELECT` queries

✅ I got it working using `gcs-connector-hadoop3-2.2.5-shaded.jar` and Spark 3.1

I also added the `google_credentials.json` and `.p12` to auth with gcs. These files are downloadable from GCP Service account.

To create the SparkSession:

```
spark = SparkSession.builder.master('local[*]') \
    .appName('spark-read-from-bigquery') \
    .config('BigQueryProjectId', 'razor-project-xxxxxxx') \
    .config('BigQueryDatasetLocation', 'de_final_data') \
    .config('parentProject', 'razor-project-xxxxxxx') \
    .config("google.cloud.auth.service.account.enable", "true") \
    .config("credentialsFile", "google_credentials.json") \
    .config("GcpJsonKeyFile", "google_credentials.json") \
    .config("spark.driver.memory", "4g") \
    .config("spark.executor.memory", "2g") \
    .config("spark.memory.offHeap.enabled", True) \
    .config("spark.memory.offHeap.size", "5g") \
    .config('google.cloud.auth.service.account.json.keyfile',
"google_credentials.json") \
    .config("fs.gs.project.id", "razor-project-xxxxxxx") \
    .config("fs.gs.impl",
"com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystem") \
    .config("fs.AbstractFileSystem.gs.impl",
"com.google.cloud.hadoop.fs.gcs.GoogleHadoopFS") \
    .getOrCreate()
```

Spark BigQuery connector Automatic configuration

While creating a SparkSession using the config `spark.jars.packages` as `com.google.cloud.spark:spark-bigquery-with-dependencies_2.12:0.23.2`

```
spark =  
SparkSession.builder.master('local').appName('bq').config("spark.jars.package  
s",  
"com.google.cloud.spark:spark-bigquery-with-dependencies_2.12:0.23.2").getOrCreate()  
create()
```

automatically downloads the required dependency jars and configures the connector, removing the need to manage this dependency. More details available [here](#)

Spark Cloud Storage connector

[Link to Slack Thread](#)

has anyone figured out how to read from GCP data lake instead of downloading all the taxi data again?

There's a few extra steps to go into reading from GCS with PySpark

1.) IMPORTANT: Download the Cloud Storage connector for Hadoop here:
<https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-storage#clusters>

As the name implies, this .jar file is what essentially connects PySpark with your GCS

2.) Move the .jar file to your Spark file directory. I installed Spark using homebrew on my MacOS machine and I had to create a /jars directory under
"/opt/homebrew/Cellar/apache-spark/3.2.1/ (where my spark dir is located)

3.) In your Python script, there are a few extra classes you'll have to import:

```
import pyspark  
from pyspark.sql import SparkSession  
from pyspark.conf import SparkConf  
from pyspark.context import SparkContext
```

4.) You must set up your configurations before building your SparkSession. Here's my code snippet:

```

conf = SparkConf() \
    .setMaster('local[*]') \
    .setAppName('test') \
    .set("spark.jars",
"/opt/homebrew/Cellar/apache-spark/3.2.1/jars/gcs-connector-hadoop3-latest.jar") \
    .set("spark.hadoop.google.cloud.auth.service.account.enable",
"true") \

.set("spark.hadoop.google.cloud.auth.service.account.json.keyfile",
"path/to/google_credentials.json")

sc = SparkContext(conf=conf)

sc._jsc.hadoopConfiguration().set("fs.AbstractFileSystem.gs.impl",
"com.google.cloud.hadoop.fs.gcs.GoogleHadoopFS")
sc._jsc.hadoopConfiguration().set("fs.gs.impl",
"com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystem")
sc._jsc.hadoopConfiguration().set("fs.gs.auth.service.account.json.keyfile", "path/to/google_credentials.json")
sc._jsc.hadoopConfiguration().set("fs.gs.auth.service.account.enable",
"true")

```

5.) Once you run that, build your SparkSession with the new parameters we'd just instantiated in the previous step:

```

spark = SparkSession.builder \
    .config(conf=sc.getConf()) \
    .getOrCreate()

```

6.) Finally, you're able to read your files straight from GCS!

```

df_green = spark.read.parquet("gs://{BUCKET}/green/202*/*")

```

How can I read a small number of rows from the parquet file directly?

```

from pyarrow.parquet import ParquetFile
pf = ParquetFile('fhvhv_tripdata_2021-01.parquet')
#pyarrow builds tables, not dataframes

```

```
tbl_small = next(pf.iter_batches(batch_size = 1000))
#this function converts the table to a dataframe of manageable size
df = tbl_small.to_pandas()
```

Alternatively without PyArrow:

```
df = spark.read.parquet('fhvhv_tripdata_2021-01.parquet')
df1 = df.sort('DOLocationID').limit(1000)
pdf = df1.select("*").toPandas()
```

DataType error when creating Spark DataFrame with a specified schema?

Probably you'll encounter this if you followed the video '5.3.1 - First Look at Spark/PySpark' and used the parquet file from the TLC website (csv was used in the video).

When defining the schema, the PULocation and DOLocationID are defined as IntegerType. This will cause an error because the Parquet file is INT64 and you'll get an error like:

```
Parquet column cannot be converted in file [...] Column [...] Expected: int, Found: INT64
```

Change the schema definition from IntegerType to LongType and it should work

Week 6

Could not start docker image “control-center” from the docker-compose.yaml file.

On Mac OSX 12.2.1 (Monterey) I could not start the kafka control center. I opened Docker Desktop and saw docker images still running from week 4, which I did not see when I typed “docker ps.” I deleted them in docker desktop and then had no problem starting up the kafka environment.

Module “kafka” not found when trying to run producer.py

Solution from Alexey: create a virtual environment and run requirements.txt and the python files in that environment.

To create a virtual env and install packages (run only once)

```
python -m venv env
source env/bin/activate
pip install -r requirements.txt
```

To activate it (you'll need to run it every time you need the virtual env):

```
source env/bin/activate
```

To deactivate it:

```
deactivate
```

This works on MacOS, Linux and Windows - but for Windows the path is slightly different (it's env/Scripts/activate)

Also the virtual environment should be created only to run the python file. Docker images should first all be up and running.

Error importing cimpl dll when running avro examples

ImportError: DLL load failed while importing cimpl: The specified module could not be found

... you may have to load librdkafka-5d2e2910.dll in the code. Add this before importing avro:

```
from ctypes import CDLL
CDLL("C:\\Users\\YOUR_USER_NAME\\anaconda3\\envs\\dtcde\\Lib\\site-packages\\confluent_kafka.libs\\librdkafka-5d2e2910.dll")
```

It seems that the error may occur depending on the OS and python version installed.

ALTERNATIVE:

ImportError: DLL load failed while importing cimpl

✅ SOLUTION: `$env:CONDA_DLL_SEARCH_MODIFICATION_ENABLE=1` in Powershell.

You need to set this DLL manually in Conda Env.

Source:

<https://github.com/confluentinc/confluent-kafka-python/issues/1186?page=2>

ModuleNotFoundError: No module named 'avro'

✓ SOLUTION: `pip install confluent-kafka[avro]`.

For some reason, Conda also doesn't include this when installing confluent-kafka via pip.

More sources on Anaconda and confluent-kafka issues:

- <https://github.com/confluentinc/confluent-kafka-python/issues/590>
- <https://github.com/confluentinc/confluent-kafka-python/issues/1221>
- <https://stackoverflow.com/questions/69085157/cannot-import-producer-from-confluent-kafka>

Error while running python3 stream.py worker

If you get an error while running the command `python3 stream.py worker`

Run `pip uninstall kafka-python`

Then run `pip install kafka-python==1.4.6`

Negsignal:SIGKILL while converting dta files to parquet format

Got this error because the docker container memory was exhausted. The dta file was upto 800MB but my docker container does not have enough memory to handle that.

Solution was to load the file in chunks with Pandas, then create multiple parquet files for each dat file I was processing. This worked smoothly and the issue was resolved.

Project

Does anyone know nice and relatively large datasets?

See a list of datasets here:

https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/week_7_project/datasets.md

Spark Streaming - How do I read from multiple topics in the same Spark Session

Initiate a Spark Session

```
spark = (SparkSession
        .builder
        .appName(app_name)
        .master(master=master)
        .getOrCreate())
```

```
spark.streams.resetTerminated()
```

```
query1 = spark
        .readStream
        ...
        .load()
```

```
query2 = spark
        .readStream
        ...
        .load()
```

```
query3 = spark
        .readStream
        ...
        .load()
```

```
query1.start()
```

```
query2.start()  
query3.start()
```

```
spark.streams.awaitAnyTermination() #waits for any one of the query to  
receive kill signal or error failure. This is asynchronous
```

```
# On the contrary query3.start().awaitTermination() is a blocking  
call. Works well when we are reading only from one topic.
```

Orchestrating dbt with Airflow

The trial dbt account provides access to dbt API. Job will still be needed to be added manually. Airflow will run the job using a python operator calling the API. You will need to provide api key, job id, etc. (be careful not committing it to Github).

Detailed explanation here: <https://docs.getdbt.com/blog/dbt-airflow-spiritual-alignment>

Source code example here:

https://github.com/sungchun12/airflow-toolkit/blob/95d40ac76122de337e1b1cdc8eed35ba1c3051ed/dags/examples/dbt_cloud_example.py

Orchestrating DataProc with Airflow

https://airflow.apache.org/docs/apache-airflow-providers-google/stable/_api/airflow/providers/google/cloud/operators/dataproc/index.html

https://airflow.apache.org/docs/apache-airflow-providers-google/stable/_modules/airflow/providers/google/cloud/operators/dataproc.html

Give the following roles to you service account:

- DataProc Administrator
- Service Account User (explanation [here](#))

Use DataprocSubmitPySparkJobOperator, DataprocDeleteClusterOperator and DataprocCreateClusterOperator.

When using DataprocSubmitPySparkJobOperator, do not forget to add:


```
dataproc_jars =  
["gs://spark-lib/bigquery/spark-bigquery-with-dependencies_2.12-0.24.0  
.jar"]
```

Because DataProc does not already have the BigQuery Connector.

2022 - Week 2 (Airflow)

Airflow - I've got this error:

**google.auth.exceptions.DefaultCredentialsError: File
/.google/credentials/google_credentials.json was not found.**


Change the path of the *google_credentials* mounting in the docker-compose file to an absolute one. For example in Ubuntu,

Instead of this: **/.google/credentials:/google_credentials:ro**

Use this: **/home/<username>/.google/credentials:/google_credentials**

I got the error below when I was running `download_dataset_task`:

```
*** Log file does not exist:  
/opt/airflow/logs/taxi_zone_dag/download_dataset_task/2022-02-02T09:39:17.124318+00:00/6.log  
  
*** Fetching from:  
http://:8793/log/taxi_zone_dag/download_dataset_task/2022-02-02T09:39:17.124318+00:00/6.log  
  
*** Failed to fetch log file from worker. Request URL missing either an 'http://' or  
'https://' protocol.
```

 I resolved it by running:

```
docker-compose down -v --rmi all --remove-orphans
```

After that, remove the following line from my codes:

```
from datetime import time
```

And then, restart docker-compose again:

```
docker-compose up
```

Installing python libraries in airflow

Under this section of the docker-compose.yaml file, find the

```
_PIP_ADDITIONAL_REQUIREMENTS:  
  build:  
    context: .  
    dockerfile: ./Dockerfile  
  environment:
```

```
_PIP_ADDITIONAL_REQUIREMENTS:${_PIP_ADDITIONAL_REQUIREMENTS:-}
```

E.g

```
_PIP_ADDITIONAL_REQUIREMENTS:${_PIP_ADDITIONAL_REQUIREMENTS:- pyspark}
```

See documentation:

<https://airflow.apache.org/docs/docker-stack/entrypoint.html#installing-additional-requirements>

Airflow won't update the DAG / It keeps returning errors even though I supposedly installed additional Python libraries

Make sure that you update your Airflow image to a more recent one. Inside your Dockerfile, modify the *FROM apache/airflow:2.2.3* to any of the more recent images available in the official Airflow Docker repository, available at

<https://hub.docker.com/r/apache/airflow/tags>

Airflow web login issue on docker:

I was unable to log onto my linux instance of airflow with the web password until I modified the config file in docker_compose.yaml from:

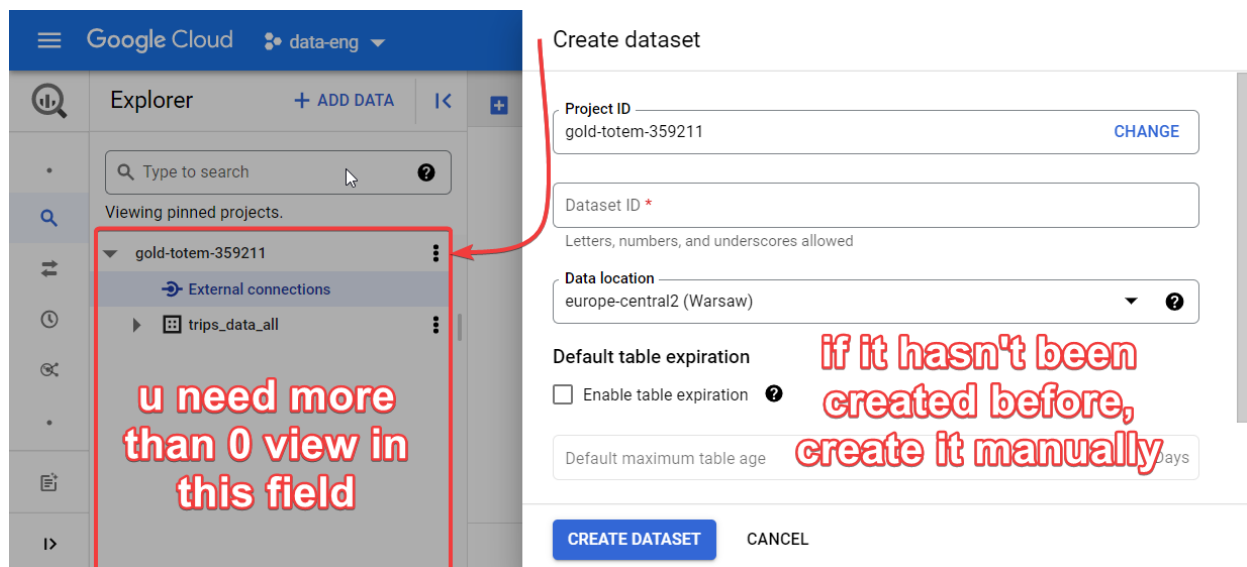
```
_AIRFLOW_WWW_USER_CREATE: 'true'
_AIRFLOW_WWW_USER_USERNAME: ${_AIRFLOW_WWW_USER_USERNAME:-airflow}
_AIRFLOW_WWW_USER_PASSWORD: ${_AIRFLOW_WWW_USER_PASSWORD:-airflow}
```

to :

```
_AIRFLOW_WWW_USER_CREATE=True
_AIRFLOW_WWW_USER_USERNAME=airflow
_AIRFLOW_WWW_USER_PASSWORD=airflow
```

google.api_core.exceptions.NotFound:

If you stuck with this problem - check this -
<https://github.com/mozilla/bigquery-etl/issues/1409>



P.S. These entities must be created by your terraform main.tf file

GCP credentials json file cannot be found when running a DAG

I got this error when running a DAG which needs to authenticate connection to the GCP:

```
File "/home/airflow/.local/lib/python3.7/site-packages/google/auth/_default.py", line 108,
```

```
in load_credentials_from_filechown
    "File {} was not found.".format(filename)
google.auth.exceptions.DefaultCredentialsError: File
/.google/credentials/google_credentials.json was not found.
```

Make sure first that you put the credentials file into the directory. Then proceed.

How did I solve it? Those are the changes I made on the docker-compose.yaml:

```
volumes:
  ~/.google/credentials/<credentials_file_name>.json:/.google/credential
ls/google_credentials.json:ro
```

```
environment:
  AIRFLOW_CONN_GOOGLE_CLOUD_DEFAULT:
    'google-cloud-platform:///key_path=%2F.google%2Fcredentialsgoogle_cre
    dentials.json&scope=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-p
    latform&project=<project_id>&num_retries=5'
```

In this stage, you might get another error, associated with directory permissions. Just grant a permission to the directory via this expression (run it on the terminal):

```
chmod 774 ~/.google/credentials/<filename.json>
```

Postgres is failing with 'could not open relation mapping file "global/pg_filenode.map" '

See [this link](#)

Assigning the unprivileged Postgres user as the owner of the Postgres data directory

```
sudo chown -R postgres /usr/local/var/postgres
```

Running the ingestion file using python

File "/usr/lib/python3/dist-packages/psycopg2/__init__.py", line 122, in connect

```
conn = _connect(dsn, connection_factory=connection_factory, **kwasync)
psycopg2.OperationalError: could not translate host name "pg-database" to address: Name or service not known
```

This is due to not being able to connect to the container port. In this case, we can point to the localhost port.

Command:

```
python3 ingest_data.py \
    --user=root \
    --password=root \
    --host=localhost \
    --port=5432 \
    --db=ny_taxi \
    --table_name=green_taxi_trips \

--url="https://github.com/DataTalksClub/nyc-tlc-data/releases/download/green/green_tripdata_2019-01.csv.gz"
```

Connecting to postgres in docker from local machine (windows) using pgcli via port 5432

Returns : Password Authentcation Failed for user <username>

Cause: I had pgAdmin installed in my windows PC and it was already using port 5432.

Fix: delete the container and rerun it. This time change the port binding from 5432:5432 to 5433:5432 . This will ask enable docker binds it's port 5432 to your PC's port 5433 since your PC's port 5432 is already in use.

Blocker: DE Zoomcamp 2.2.6 at around 11:11: denied: requested access to the resource is denied

✅ Solution:

It is assumed that you have an active Docker account. If not sign-up [here](#).

Make sure you use below command, before pushing image to Docker hub
`docker login`

```
docker build -t docker image DOCKER_USERNAME/IMAGE_NAME:VERSION
docker build -t docker image timapple/prefect_cloud:v000
docker push DOCKER_USERNAME/IMAGE_NAME:VERSION
docker push timapple/prefect_cloud:v000
```

URL: In case the above does not work. Kindly read [this thread](#).

Blocker: NotADirectoryError: [Errno 20] Not a directory: '/opt/prefect/flows'

Code block:

```
# base Docker image that we will build on
FROM prefecthq/prefect:2.7.7-python3.9

# Copy the list of env packages needed. Don't forget what's on the
list
COPY docker_env_req.txt .

# Setup the env packages requirement
RUN pip install -r docker_env_req.txt --trusted-host pypi.python.org
--no-cache-dir

# Copy the flow code
COPY parametric_web_to_gcs.py /opt/prefect/flows
COPY data /opt/prefect/data
```

✓Solution:

From Jeff Hale: Looks like a Docker copy error. Maybe you can't copy a file to a directory that is being created in the same statement. So change the copy command from the .py file to the name of the folder. For example, copy the local flows folder into the /opt/prefect/flows folder.

Change code to:

```
COPY parametric_web_to_gcs.py /opt/prefect/flows/week_2_workflow_orchestration
```

Code

But got an error:

```
Script at 'parametric_web_to_gcs.py' encountered an exception:
FileNotFoundError(2, 'No such file or directory')
```

Then I revert to

```
COPY parametric_web_to_gcs.py /opt/prefect/flows
```

Build the image and push it again then deploy. And it works! The mystery of the universe

URL:

https://datatalks-club.slack.com/archives/C01FABYF2RG/p1674964665817089?thread_ts=1674570724.323079&cid=C01FABYF2RG

Running pip command in Dockerfile

THE ERROR:

```
WARNING: Retrying (Retry(total=4, connect=None, read=None, redirect=None, status=None)) after connection broken by
'NewConnectionError('<pip._vendor.urllib3.connection.HTTPSConnection object
at 0x7fc88c50a940>: Failed to establish a new connection: [Errno -3]
Temporary failure in name resolution')': /simple/pip/
```

✓THE SOLUTION:

Add the nameserver line below into /etc/resolv.conf

```
- sudo nano /etc/resolv.conf
```

```
'''
nameserver 8.8.8.8
'''
```

Unable to find block document named zoom-gcs for block type gcs-bucket when running \$ python3 etl_web_to_gcs.py

You have not set up the gcs-block yet. Follow this [excellent summary](#) by Padhila, Section: [DE Zoomcamp 2.2.3 - ETL with GCP & Prefect](#) up until Step 8. Note: If you've replaced your GCP credentials you can generate new one via: GCP > MENU > IAM & Admin > Service Accounts > "User" > Keys > Add Key > Create New Key > JSON

Question 4. Github Storage Block

pydantic.error_wrappers.ValidationError: 1 validation error for Deployment infrastructure.

Infrastructure block must have 'run-infrastructure' capabilities. (type=value_error)

-sb, --storage-block TEXT

The [storage block](#) to use, in block-type/block-name or block-type/block-name/path format. Note that the appropriate library supporting the storage filesystem must be installed.

| Block name | Block class name | Block type for a slug |
|--------------------|------------------|-----------------------|
| Azure | Azure | azure |
| Docker Container | DockerContainer | docker-container |
| GitHub | GitHub | github |
| GCS | GCS | gcs |
| Kubernetes Job | KubernetesJob | kubernetes-job |
| Process | Process | process |
| Remote File System | RemoteFileSystem | remote-file-system |
| S3 | S3 | s3 |
| SMB | SMB | smb |
| GitLab Repository | GitLabRepository | gitlab-repository |

prefect deployment build path/to/etl_web_to_gcs.py:etl_web_to_gcs -n "github-deployment" -sb github/github-zoomcamp

Using infrastructure

You may create customized infrastructure blocks through the Prefect UI or Prefect Cloud [Blocks](#) page or create them in code and save them to the API using the blocks `.save()` method.

Once created, there are two distinct ways to use infrastructure in a deployment:

- Starting with Prefect defaults — this is what happens when you pass the `-i` or `--infra` flag and provide a type when building deployment files.
- Pre-configure infrastructure settings as blocks and base your deployment infrastructure on those settings — by passing `-ib` or `--infra-block` and a block slug when building deployment files.

For example, when creating your deployment files, the supported Prefect infrastructure types are:

- `process`
- `docker-container`
- `kubernetes-job`
- `ecs-task`
- `cloud-run-job`
- `container-instance-job`

```
$ prefect deployment build ./my_flow.py:my_flow -n my-flow-deployment -t test -i docker-container
Found flow 'my-flow'
Successfully uploaded 2 files to s3://bucket-full-of-sunshine
Deployment YAML created at './Users/terry/test/flows/infra/deployment.yaml'
```