

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: <https://github.com/DataTalksClub/nyc-tlc-data>

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, if you don't submit the homeworks, you're still eligible for a certificate at the end - as long as you successfully pass the project at the end. So you can take the course at your own pace.

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

Week 1

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

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

On Ubuntu, run

```
sudo apt-get 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/)

On MacOS, the easiest way to install wget is to use Brew:

```
brew install wget
```

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
```

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
```

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).

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 dpage/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 **dpage/pgadmin4** and NOT **dbpage/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
f8e23c71dc3b: 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 0.0.0.0 -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.

pgcliThe 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``

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
```

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 permitted
```

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="admin" \
  -e POSTGRES_DB="ny_taxi" \
  -v dtc_postgres_volume_local:/var/lib/postgresql/data \
  -p 5432:5432 \
  postgres:13
```

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"

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
```

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

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 makedirs
    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.

Error:

```

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:

(in git bash) pip install psycopg2-binary

ModuleNotFoundError: No module named 'psycopg2'

Issue:

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

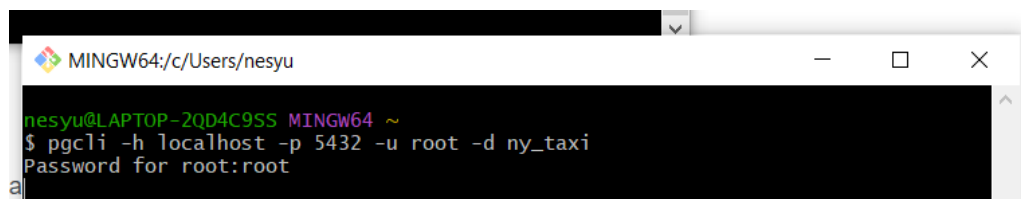
e...

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

Solution: pip install psycopg2-binary

pgcli stuck on password prompt

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



Use winpty:

winpty pgcli -h localhost -p 5432 -u root -d ny_taxi

•

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

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. Replace site-packages with Scripts:

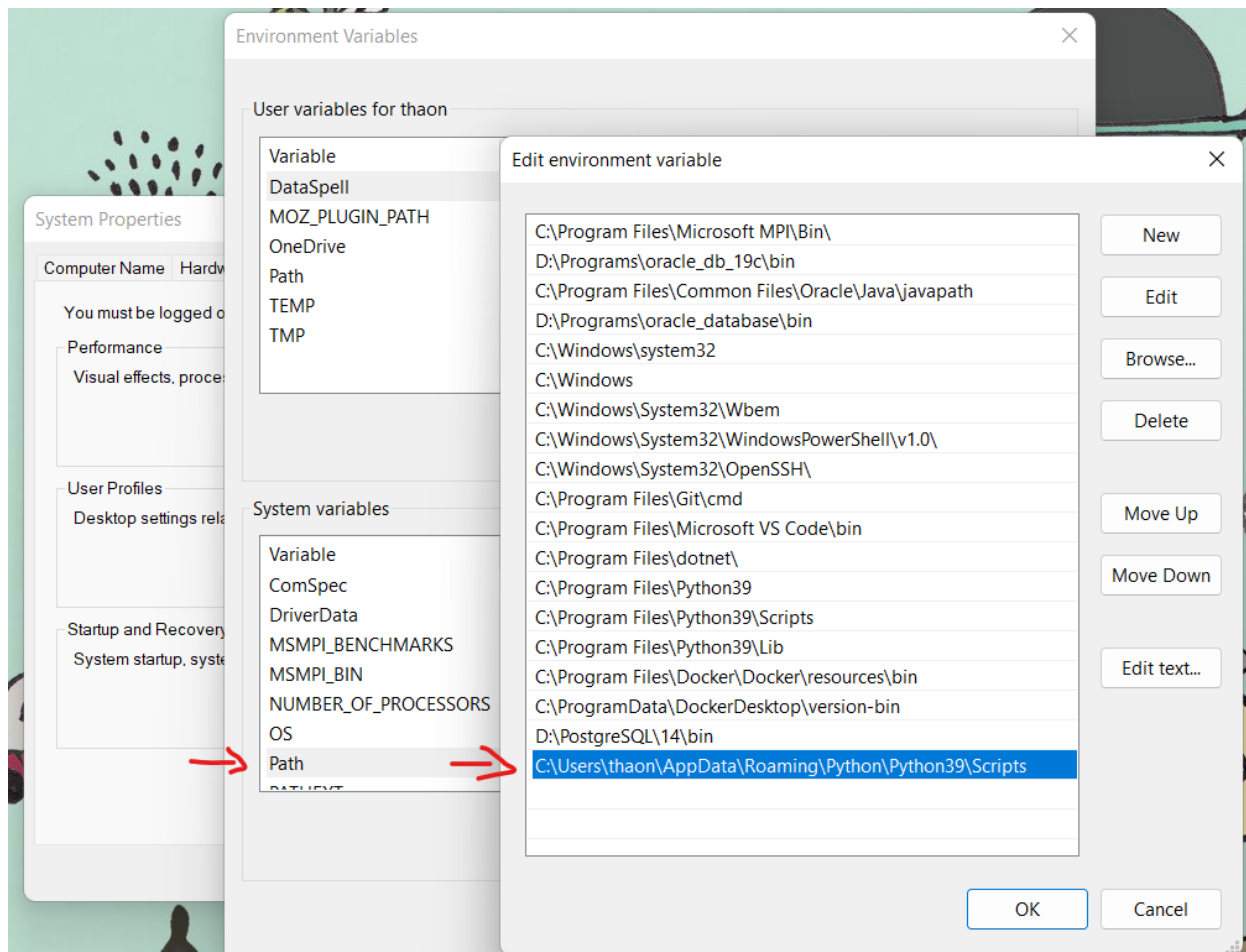
C:\Users\...\AppData\Roaming\Python\Python39\Scripts

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: [tps://stackoverflow.com/a/68233660](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" encoding="UTF-8" ?>
<Error>
  <Code>AccessDenied</Code>
  <Message>Access Denied</Message>
  <RequestId>KA0B59E64XX4WCDC</RequestId>
  <HostId>7okNtNIh1KLrv9jzClzDfm+leGXWDRj0UNmSUC3LBArWmzFfzKicPVRxf40XORb4ToMXvs6mu4s</HostId>
</Error>
```

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

We have a backup, so use it instead:

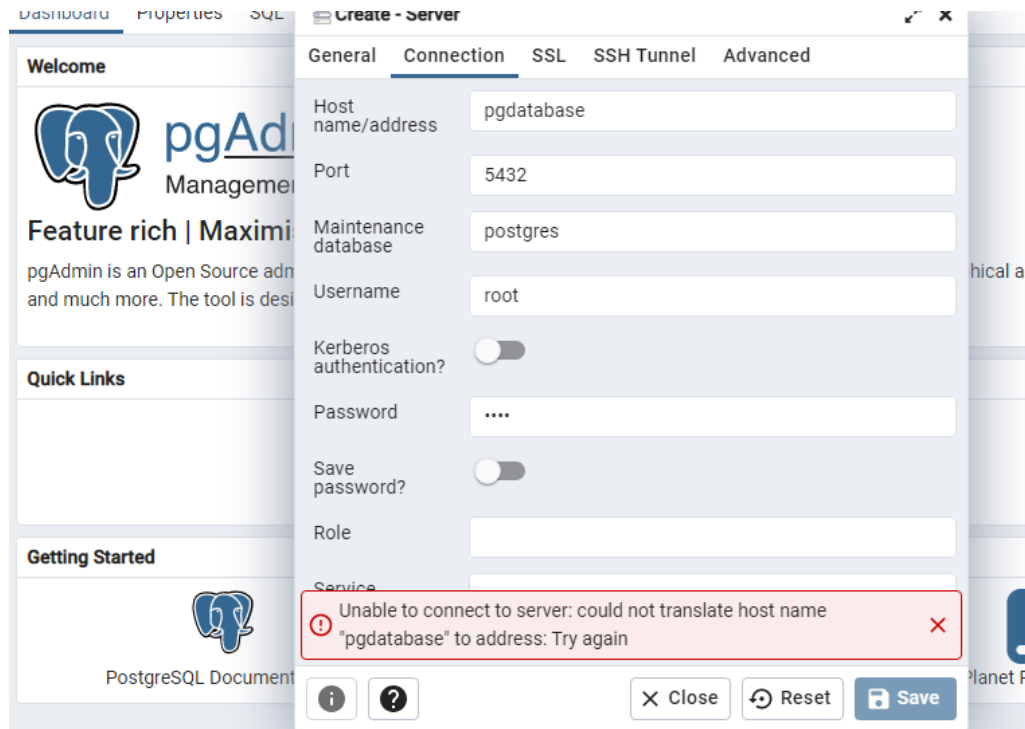
<https://github.com/DataTalksClub/nyc-tlc-data>

So 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 PORTS	IMAGE	COMMAND NAMES	CREATED	STATUS
faf05090972e 0.0.0.0:5432->5432/tcp	postgres:13	"docker-entrypoint.s..." pg-database	39 seconds ago	Up 37 seconds
6344dcecd58f 443/tcp, 0.0.0.0:8080->80/tcp hw	dpage/pgadmin4	"/entrypoint.sh" pg-admin	39 seconds ago	Up 37 seconds

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 logs <container_id>`

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 - ERROR[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](#)

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

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()

    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:
<{'vary': '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 already exists",
  "status": "ALREADY_EXISTS"
}
```

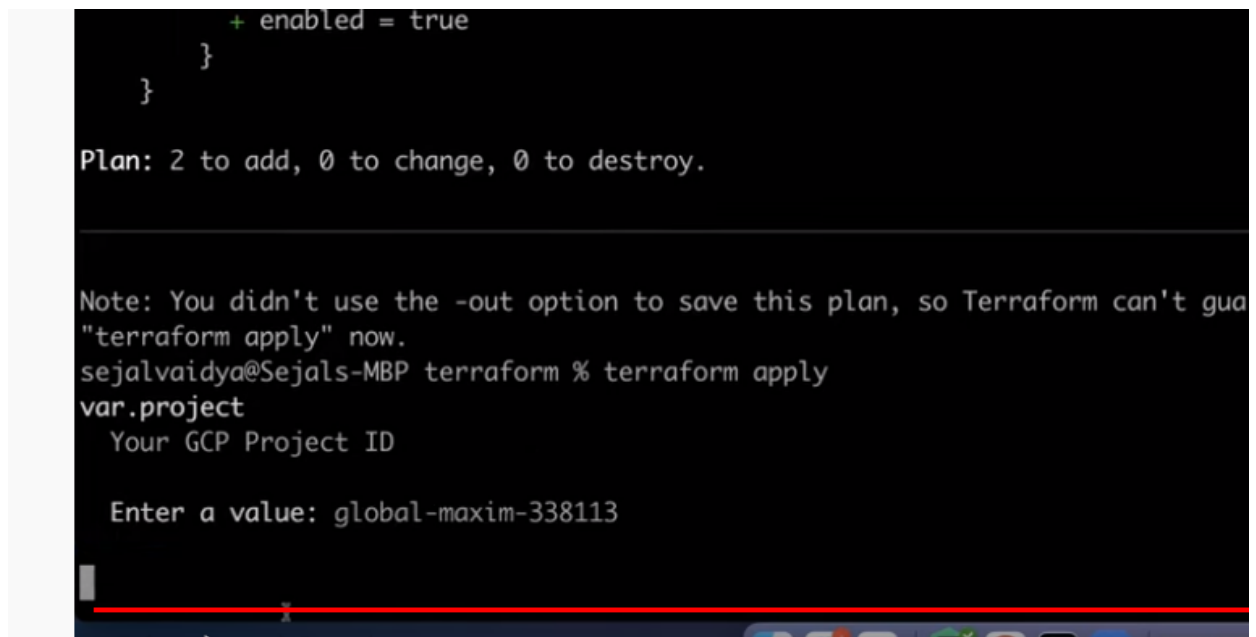
From Stackoverflow:

<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.

A terminal window with a dark background. At the top, there is a snippet of JSON configuration:

```
+ enabled = true
}
}
```

 Below this, the Terraform plan output is shown:

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

 A horizontal line separates the plan from the apply command. The apply command is:

```
sejalvaidya@Sejals-MBP terraform % terraform apply
```

 The prompt `var.project` is shown, followed by the instruction `Your GCP Project ID`. The user has entered the value `global-maxim-338113`.

```
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 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 (~)

Video 1.4.1

https://www.youtube.com/watch?v=ae-CV2KfoN0&list=PL3MmuxUbc_hJed7dXYoJw8DoCuVHhGEQb

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_c
ourse/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 `_dag` 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.

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/
```

Is 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`.

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

This is 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 cntr+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.yaml` 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.)

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.

Week 2

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.

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).

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'}))
```

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

<https://github.com/dbt-labs/dbt-bigquery/issues/19#issuecomment-635545315>

Additionally please see this post of Sandy:

<https://learningdataengineering540969211.wordpress.com/dbt-cloud-and-bigquery-an-easy-way-to-try-and-resolve-location-issues/>

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 dataset 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 updated 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.uiWebUrl`

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.packages",
"com.google.cloud.spark:spark-bigquery-with-dependencies_2.12:0.23.2").getOrCreate()

```

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

https://datatalks-club.slack.com/archives/C01FABYF2RG/p1646013709648279?thread_ts=1646008578.136059&cid=C01FABYF2RG

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 Alexei: 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://githubhot.com/repo/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

Can I still submit the final 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 `DataprocsSubmitPySparkJobOperator`, `DataprocsDeleteClusterOperator` and `DataprocsCreateClusterOperator`.

When using `DataprocsSubmitPySparkJobOperator`, 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)

2022:

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}
```

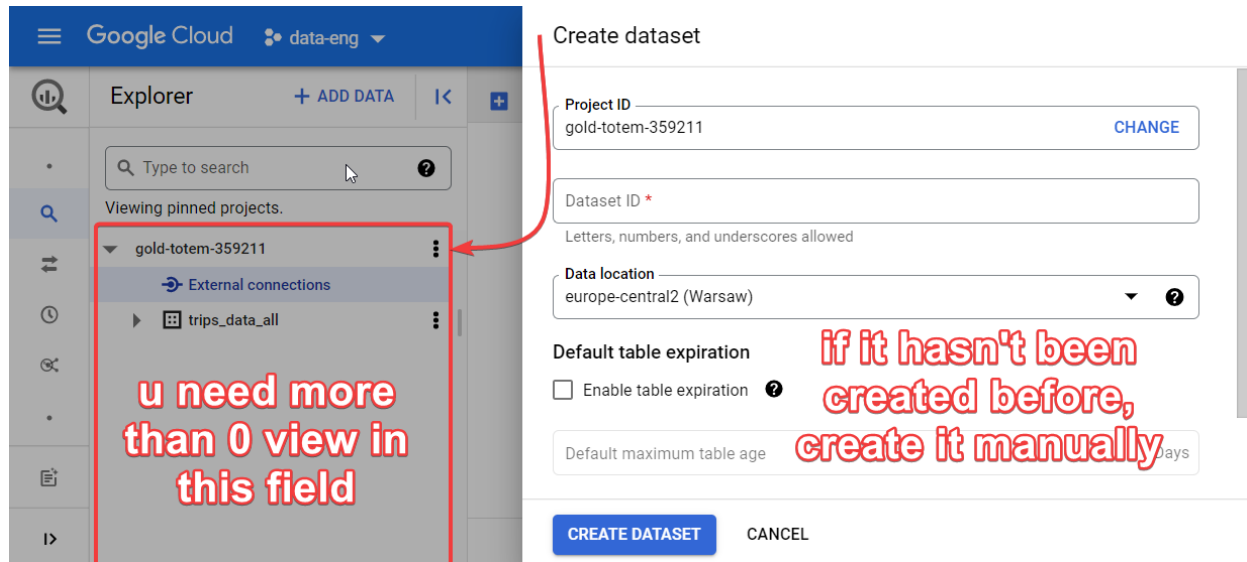
to :

```
_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_file
    "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/credentia
```

```
ls/google_credentials.json:ro
```

```
environment:  
AIRFLOW_CONN_GOOGLE_CLOUD_DEFAULT:  
'google-cloud-platform://?key_path=%2F.google%2Fcredentials%2Fgoogle_credentials.json&scope=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-platform&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" '

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

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