Data Engineering Zoomcamp FAQ

Data Engineering Zoomcamp FAQ

The purpose of this document is to capture Frequently asked technical questions

Editing guidelines:

- When adding a new FAQ entry, make sure the question is “Heading 2”
- Feel free to improve if you see something is off
- **Don’t change the formatting in the Data document or add any visual “improvements” (make a copy for yourself first if you need to do it for whatever reason)**
- **Don’t change the pages format (it should be “pageless”)**
- Add name and date for reference, if possible

General course-related questions

Course - When does the course start?

The next cohort starts January 13th 2025. More info at DTC.

- Register before the course starts using this link.
- Joint the course Telegram channel with announcements.
- Don’t forget to register in DataTalks.Club's Slack and join the channel.

Course - What are the prerequisites for this course?

See DE zoomcamp 2025 pre-course Q&A

To get the most out of this course, you should have:

- Basic coding experience
- Familiarity with SQL
- Experience with Python (helpful but not required)

No prior data engineering experience is necessary. See Readme on GitHub

Course - Can I still join the course after the start date?

Yes, even if you don't register, you're still eligible to submit the homework.

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

Course - I have registered for the Data Engineering Bootcamp. When can I expect to receive the confirmation email?

You don't need it. You're accepted. You can also just start learning and submitting homework without registering. It is not checked against any registered list. Registration is just to gauge interest *before* the start date.

Course - What can I do before the course starts?

Start by installing and setting up all the dependencies and requirements:

- Google cloud account
- Google Cloud SDK
- Python 3 (installed with Anaconda)
- Terraform
- Git

Look over the prerequisites and syllabus to see if you are comfortable with these subjects.

Course - how many Zoomcamps in a year?

There are multiple Zoomcamps in a year, as of 2025. More info at DTC Article.

However, they are five separate courses, estimated to be during these months:

1. Data-Engineering (Jan - Apr)
2. Stock Market Analytics (Apr - May)
3. MLOps (May - Aug)
4. LLM (June - Sep)
5. Machine Learning (Sep - Jan)

There's only one Data-Engineering Zoomcamp “live” cohort per year, for the certification. Same as for the other Zoomcamps.

They follow pretty much the same schedule for each cohort per zoomcamp. For Data-Engineering it is (generally) from Jan-Apr of the year. If you’re not interested in the Certificate, you can take any zoom camps at any time, at your own pace, out of sync with any “live” cohort.

Course - Is the current cohort going to be different from the previous cohort?

For the 2025 edition we are using Kestra (see Demo) instead of MageAI (Module 2). Lookout for new videos. See Playlist

For the 2024 edition we used Mage AI instead of Prefect and **re-recorded the terraform videos**, For 2023, we used Prefect instead of Airflow. See Playlists on YouTube and cohorts folder in Github repo.

Course - Can I follow the course after it finishes?

Yes, we will keep all the materials after the course finishes, so you can follow the course at your own pace after it finishes.

You can also continue looking at the homeworks and continue preparing for the next cohort. I guess you can also start working on your final capstone project.

Course - Can I get support if I take the course in the self-paced mode?

Yes, the slack channel remains open and you can ask questions there. But always search the channel first and second, check the FAQ (this document), most likely all your questions are already answered here.

You can also tag the bot @ZoomcampQABot to help you conduct the search, but don't rely on its answers 100%, it is pretty good though.

Course - Which playlist on YouTube should I refer to?

All the main videos are stored in the Main "DATA ENGINEERING ZOOMCAMP" playlist (no year specified). The Github repository has also been updated (if not create a pull request) to show each video with a thumbnail, that would bring you directly to the same playlist below.

Below is the MAIN PLAYLIST'. And then you refer to the year specific playlist for additional videos for that year like for office hours videos etc. Also find this playlist pinned to the slack channel.

- Data Engineering Zoomcamp
- Data Engineering Zoomcamp 2022
- Data Engineering Zoomcamp 2023
- Data Engineering Bootcamp 2024
- Data Engineering Bootcamp 2025
- DE Zoomcamp 2025 (Module 2 Kestra)

Course - How many hours per week am I expected to spend on this course?

It depends on your background and previous experience with modules. It is expected to require about 5 - 15 hours per week. [source1] [source2]

You can also calculate it yourself using this data and then update this answer.

Office Hours - What is the video/zoom link to the stream for the “Office Hour” or workshop sessions?

The zoom link is only published to instructors/presenters/TAs.

Students participate via Youtube Live and submit questions to Slido (link would be pinned in the chat when Alexey goes Live). The video URL should be posted in the announcements channel on Telegram & Slack and is in google calendar before it begins. Also, you will see it live on the DataTalksClub YouTube Channel.

Don’t post your questions in chat as it would be off-screen before the instructors/moderators have a chance to answer it if the room is very active.

Office Hours - I can’t attend the “Office hours” / workshop, will it be recorded?

Yes! Every “Office Hours” will be recorded and available a few minutes after the live session is over; so you can view (or rewatch) whenever you want.

Course Management Platform for Homeworks, Project and Certificate

Edit Course Profile.

The display name listed on the leaderboard is an auto-generated randomized name. You can edit it to be a nickname, or your real name, if you prefer. Your entry on the Leaderboard is the one highlighted in light green.

The Certificate name should be your actual name that you want to appear on your certificate after completing the course.

The "Display on Leaderboard" option indicates whether you want your name to be listed on the course leaderboard.

Certificate - Do I need to do the homeworks to get the certificate?

No, as long as you do the peer-reviewed capstone projects in time then you can get the certificate. You do not need to do the homeworks if you join late for example.

Certificate - Can I follow the course in a self-paced mode and get a certificate?

No, you can only get a certificate if you finish the course with a “live” cohort. We don't award certificates for the self-paced mode. The reason is you need to peer-review capstone(s) after submitting a project. You can only peer-review projects at the time the course is running.

Homework - What are homework and project deadlines?

2025 deadlines will be announced on https://courses.datatalks.club/de-zoomcamp-2025/ and in Google Calendar

You can find the 2024 deadlines here: https://docs.google.com/spreadsheets/d/e/2PACX-1vQACMLuutV5rvXg5qICuJGL-yZqIV0FBD84CxPdC5eZHf8TfzB-CJT_3Mo7U7oGVTXmSihPgQxuuoku/pubhtml

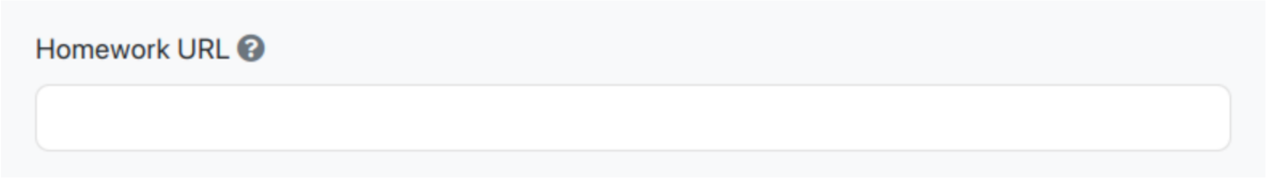
Also, take note of Announcements from **@Au-Tomator** for any extensions or other news. Or, the form may also show the updated deadline, if Instructor(s) has updated it.

Homework - Are late submissions of homework allowed?

No, late submissions are not allowed. But if the form is still not closed and it’s after the due date, you can still submit the homework. Confirm your submission by the date-timestamp on the Course page. Make sure you are logged in.

Older news:[source1] [source2]

Homework - What is the homework URL in the homework link?



Homework URL

Answer: In short, it’s your repository on github, gitlab, bitbucket, etc

In long, your repository or any other location you have your code where a reasonable person would look at it and think yes, you went through the week and exercises. Think of it like a portfolio you could present to an employer.

Homework and Leaderboard - what is the system for points in the course management platform?

After you submit your homework it will be graded based on the amount of questions in a particular homework. You can see how many points you have right on the page of the homework up top. Additionally in the leaderboard you will find the sum of all points you’ve earned - points for Homeworks, FAQs and Learning in Public. If homework is clear,(https://datatalks-club.slack.com/archives/C01FABYF2RG/p1706846846359379? others work as follows:

- You get maximum 1 point for the FAQ Contribution in the respective week

For each learning in a public link you get one point, so you can get maximum 7 points.

Check this Video: https://www.loom.com/share/710e3297487b409d94df0e8da1c984ce

Leaderboard - I am not on the leaderboard / how do I know which one I am on the leaderboard?

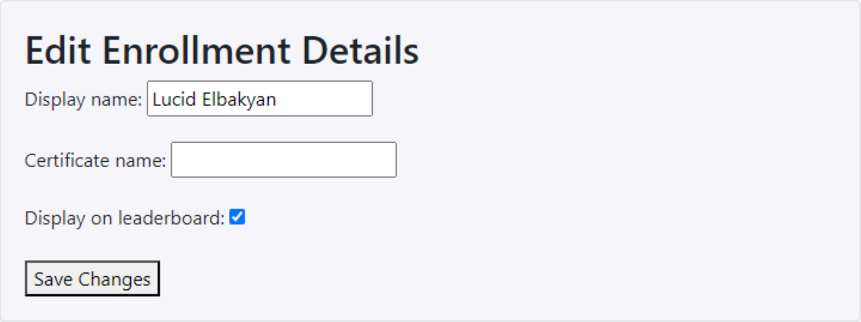
When you set up your account you are automatically assigned a random name such as “Lucid Elbakyan” for example. If you want to see what your Display name is.

Go to your profile: →

2025: https://courses.datatalks.club/de-zoomcamp-2025/enrollment

2024: https://courses.datatalks.club/de-zoomcamp-2024/enrollment

Log in -> your display name is here, you can also change it should you wish. Make sure your Certificate name is correct, this name will later be printed on your certificate!!!



Edit Enrollment Details

Display name: Lucid Elbakyan

Certificate name:

Display on leaderboard: ☑

Environment - Is Python 3.9 still the recommended version to use in 2024?

Yes, for simplicity (of troubleshooting against the recorded videos) and stability. [source]

But Python 3.10 and 3.11 should work fine.

Environment - Should I use my local machine, GCP, or GitHub Codespaces for my environment?

You can set it up on your laptop or PC if you prefer to work locally from your laptop or PC.

You might face some challenges, especially for Windows users.

If you prefer to work on the local machine, you may start with the week 1 Introduction to Docker and follow through.

However, if you prefer to set up a virtual machine, you may start with these first:

1. Using GitHub Codespaces

2. Setting up the environment on a cloud VM codespace

I decided to work on a virtual machine because I have different laptops & PCs for my home & office, so I can work on this boot camp virtually anywhere.

Environment - Is GitHub codespaces an alternative to using cli/git bash to ingest the data and create a docker file?

GitHub Codespaces offers you computing Linux resources with many pre-installed tools (Docker, Docker Compose, Python).

You can also open any GitHub repository in a GitHub Codespace.

Environment - Do we really have to use GitHub codespaces? I already have PostgreSQL & Docker installed.

It's up to you which platform and environment you use for the course.

Github codespaces or GCP VM are just possible options, but you can do the entire course from your laptop.

Environment - Do I need both GitHub Codespaces and GCP?

Choose the approach that aligns the most with your idea for the end project

One of those should suffice. However, BigQuery, which is part of GCP, will be used, so learning that is probably a better option. Or you can set up a local environment for most of this course.

Environment - Could not establish connection to "MyServerName": Got bad result from install script

This happens when attempting to connect to a GCP VM using VSCode on a Windows machine. Changing registry value in registry editor

1. To open Run command window, you can either:

(1-1) Use the shortcut keys: 'Windows + R', or

(1-2) Right Click "Start", and click "Run" to open.

2. Registry Values Located in Registry Editor, to open it: Type 'regedit' in the Run command window, and then press Enter.' 3. Now you can change the registry values

"Autorun" in "HKEY_CURRENT_USER\Software\Microsoft\Command Processor" from "if exists" to a blank.

Alternatively, You can simplify the solution by deleting the fingerprint saved within the `known_hosts` file. In Windows, this file is placed at `C:\Users\<your_user_name>\.ssh\known_host`

Environment - Why are we using GCP and not other cloud providers?

For uniformity.

You can use other cloud platforms, since you get every service that's been provided by GCP in Azure and AWS, you're not restricted to GCP, you can use other cloud platforms like AWS if you're comfortable with AWS or others.

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 must have a valid credit card.

Should I pay for cloud services?

No, if you use and take advantage of their free trial.

Environment - The GCP and other cloud providers are unavailable 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 use (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. HW3 needed BigQuery.

Environment - Can DE Zoomcamp course be completed using only the GCP Sandbox option, or is the Free Trial required at any point?

Google Cloud Platform (GCP) provides two free trial options: the Free Trial and the Sandbox. Note that users can switch from Sandbox to Free Trial anytime by adding billing details. The reverse is true at anytime as well. You can switch from the GCP Free Trial to the Sandbox option. To do this, you'll need to **disable billing** on your project. Once billing is disabled, your project will revert to the Sandbox mode, allowing you to use the limited free resources without a billing account.

However, completing the course using the GCP Sandbox option is not possible because the Sandbox has limited features compared to the full Free Trial with $300 credit. The course will involve using services that are not available in the Sandbox environment. The FAQ indicates that while the course may start locally, it will eventually transition to using VMs, GCS Bucket and other paid services on on GCP, which would require the full capabilities provided by the $300 credit option. Additionally, the course emphasizes the use of BigQuery, which is a key component of GCP, and the Sandbox may not support all necessary functionalities for working with it effectively. Therefore, it's recommended to utilize the full Free Trial with billing details to ensure access to all required features for the course.

Environment - 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!

The problem would be when you need help. You'd need to rely on fellow coursemates who also use AWS (or have experience using it before), which might be in smaller numbers than those learning the course with GCP.

See the de-course-aws channel on slack

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

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.

See Google Calendar

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 to the cohort 2022 folder on github (here)

Can I use Airflow instead for my final project?

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

Is it possible to use tool "X" instead of the one tool you use in the course?

Yes, this applies if you want to use Airflow or Prefect instead of Mage, 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 one of them or use your tool of choice.

Should you consider it instead of the one tool you use? 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 ❣️

Create a PR if you see you can improve the text or the structure of the repository.

Update this FAQ.

Environment - Is the course [Windows/macOS/Linux/...] friendly?

Yes! Linux is ideal but technically it should not matter. Students in the 2024 cohort used all 3 OSes successfully.

Environment - Roadblock for Windows users in modules with *.sh (shell scripts).

Later modules (module-05 & RisingWave workshop) use shell scripts in *.sh files and most Windows users not using WSL would hit a wall and cannot continue, even in git bash or MINGW64. This is why WSL environment setup is recommended from the start.

Any books or additional resources you recommend?

Yes to both! check out this document: https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/awesome-data-engineering.md

Project - What is Project Attempt #1 and Project Attempt #2 exactly?

You will have two attempts for a project. If the first project deadline is over and you're late or you submit the project and fail the first attempt, you have another chance to submit the project with the second attempt.

How to troubleshoot issues

The first step is to try to solve the issue on your own. Get used to solving problems and reading documentation. This will be a real life skill you need when employed. `[ctrl+f]` is your friend, use it! It is a universal shortcut and works in all apps/browsers.

1. What does the error say? There will often be a description of the error or instructions on what is needed or even how to fix it. I have even seen a link to the solution. Does it reference a specific line of your code?

2. Restart app or server/pc. In

3. Google it, use ChatGPT, Bing AI etc.

    a. It is going to be rare that you are the first to have the problem, someone out there has posted the fly issue and likely the solution.

    b. Search using: `<technology> <problem statement>`. Example: `pgcli error column c.relhasoids does not exist.`

    c. There are often different solutions for the same problem due to variation in environments.

4. Check the tech’s documentation. Use its search if available or use the browsers search function.
5. Try uninstall (this may remove the bad actor) and reinstall of application or reimplementation of action. Remember to restart the server/pc for reinstalls.
    a. Sometimes reinstalling fails to resolve the issue but works if you uninstall first.
6. Post your question to Stackoverflow. Read the Stackoverflow guide on posting good questions.
    a. https://stackoverflow.com/help/how-to-ask
    b. This will be your real life. Ask an expert in the future (in addition to coworkers).
7. Ask in Slack
    a. Before asking a question,
        i. Check Pins 📌 in channel (where the shortcut to the repo and this FAQ is located)
        ii. Use the slack app’s search function
        iii. check the FAQ (this document), use search `[ctrl+f]`
        iv. Use the bot **@ZoomcampQABot** to do the search for you
    b. When asking a question, include as much information as possible:
        i. What are you coding on? What OS?
        ii. What command did you run, which video did you follow? Etc etc
        iii. What error did you get? Does it have a line number to the “offending” code and have you check it for typos?
        iv. What have you tried that did not work? This answer is crucial as without it, helpers would ask you to do the suggestions in the error log first. Or just read this FAQ document.
    c. DO NOT use screenshots, especially don’t take pictures from a phone.
    d. DO NOT tag instructors, it may discourage others from helping you. Copy and paste errors; if it’s long, just post it in a reply to your thread.
        i. Use ``` for formatting your code.

    e. Use the same thread for the conversation (that means reply to your own thread).

        i. DO NOT create multiple posts to discuss the issue.

        ii. You may create a new post if the issue reemerges down the road. Describe what has changed in the environment.

    f. Provide additional information in the same thread of the steps you have taken for resolution.

8. Take a break and come back later. You will be amazed at how often you figure out the solution after letting your brain rest. Get some fresh air, workout, play a video game, watch a tv show, whatever allows your brain to not think about it for a little while or even until the next day.

9. Remember technology issues in real life sometimes take days or even weeks to resolve.

10. If somebody helped you with your problem and it's not in the FAQ, please add it there. It will help other students.

How to ask questions

When the troubleshooting guide above does not help resolve it and you need another pair of eyeballs to spot mistakes. When asking a question, include as much information as possible:

1. What are you coding on? What OS?

2. What command did you run, which video did you follow? Etc etc

3. What error did you get? Does it have a line number to the “offending” code and have you check it for typos?

4. What have you tried that did not work? This answer is crucial as without it, helpers would ask you to do the suggestions in the error log first. Or just read this FAQ document.

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: Git for Everybody: How to Clone a Repository from GitHub

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: How to Create a Git Repository | Atlassian Git Tutorial

Remember to ignore large database, .csv, and .gz files, and other files that should not be saved to a repository. Use `.gitignore` for this: .gitignore file - ignoring files in Git | Atlassian Git Tutorial

NEVER stores passwords or keys in a git repo (even if that repo is set to private). Put files containing sensitive information (`.env`, `secret.json` etc.) in your `.gitignore`.

This is also a great resource: Dangit, Git!?!

VS Code: Tab using spaces

Error: Makefile:2: *** missing separator. Stop.

Solution: Tabs in documents should be converted to Tab instead of spaces. Follow this stack.

Opening an HTML file with a Windows browser from Linux running on WSL

If you're running Linux on Windows Subsystem for Linux (WSL) 2, you can open HTML files from the guest (Linux) with whatever Internet Browser you have installed on the host (Windows). Just install wslu and open the page with `wslview <file>`, for example:

```
wslview index.html
```

You can customise which browser to use by setting the `BROWSER` environment variable first. For example:

```
export BROWSER='/mnt/c/Program Files/Firefox/firefox.exe'
```

Set up Chrome Remote Desktop for Linux on Compute Engine

This tutorial shows you how to set up the Chrome Remote Desktop service on a Debian Linux virtual machine (VM) instance on Compute Engine. Chrome Remote Desktop allows you to remotely access applications with a graphical user interface.

Certificate - generating, receiving after projects graded

Q: When will it be sent out / released?

Q: How do I get my certificate after project(s) have been reviewed and graded?

A: There'll be an announcement in Telegram and the course channel for

(1) checking that your proper full name is how you want displayed on the Certificate (see Editing course profile on the Course Management webpage), and

(2) when the grading is completed.

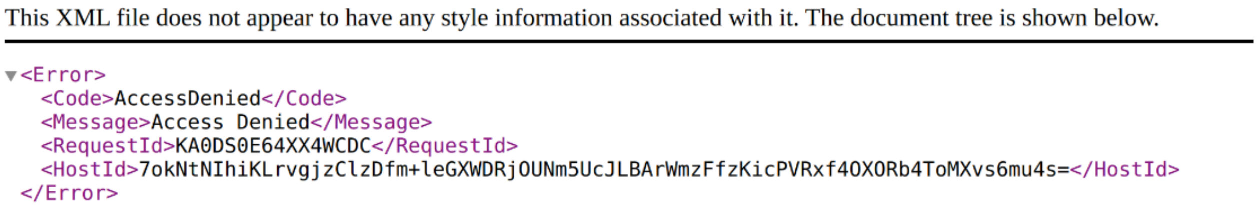
After the second announcement, please follow instructions in https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/certificates.md on how to generate the Certificate document yourself.

Module 1: Docker and Terraform

Taxi Data - Yellow Taxi Trip Records downloading error, Error no or XML error webpage

When you try to download the 2021 data from TLC website, you get this error:

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



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

```
▼<Error>
   <Code>AccessDenied</Code>
   <Message>Access Denied</Message>
   <RequestId>KA0DS0E64XX4WCDC</RequestId>
   <HostId>7okNtNIhiKLrvgjzClzDfm+leGXWDRjOUNm5UcJLBArWmzFfzKicPVRxf40X0Rb4ToMXvs6mu4s=</HostId>
 </Error>
```

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

So the link should be 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.)

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

In this video, we store the data file as `“output.csv”`. The data file won’t store correctly if the file extension is csv.gz instead of csv. One alternative is to replace `csv_name = “output.cs -v”` with the file name given at the end of the URL. Notice that the URL for the yellow taxi data is: https://github.com/DataTalksClub/nyc-tlc-data/releases/download/yellow/yellow_tripdata_2021-01.csv.gz where the highlighted part is the name of the file. We can parse this file name from the URL and use it as `csv_name`. That is, we can replace `csv_name = “output.csv”` with `csv_name = url.split(“/”)[-1]` . Then when we use `csv_name` to using `pd.read_csv`, there won’t be an issue even though the file name really has the extension csv.gz instead of csv since the pandas `read_csv` function can read csv.gz files directly.

Taxi Data - Data Dictionary for NY Taxi data?

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

Green Trips: Data Dictionary - LPEP Trip Records May 1, 2018

Taxi Data - Unzip Parquet file

You can unzip this downloaded parquet file, in the command line. The result is a csv file which can be imported with pandas using the pd.read_csv() shown in the videos.

‘’’gunzip green_tripdata_2019-09.csv.gz’’’

SOLUTION TO USING PARQUET FILES DIRECTLY IN PYTHON SCRIPT
ingest_data.py

In the def main(params) add this line

parquet_name= 'output.parquet'

Then edit the code which downloads the files

os.system(f"wget {url} -O {parquet_name}")

Convert the download .parquet file to csv and rename as csv_name to keep it relevant to the rest of the code df = pd.read_parquet(parquet_name)

df.to_csv(csv_name, index=False)

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.

“No such file or directory: 'output.csv.gz'”, may also caused by wget not recognized

.
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

wget - ERROR: cannot verify <website> certificate (MacOS)

Firstly, make sure that you add "!" before wget if you're running your command in a Jupyter Notebook or CLI. Then, you can check one of this 2 things (from CLI):

1. Using the Python library `wget` you installed with pip, try `python -m wget <url>`
2. Write the usual command and add `--no-check-certificate` at the end. So it should be:

   ```
   !wget <website_url> --no-check-certificate
   ```

https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2021-01.csv

Git Bash - 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)

GitHub Codespaces - How to store secrets

Instruction on how to store secrets that will be avialable in GitHub Codespaces.
Managing your account-specific secrets for GitHub Codespaces - GitHub Docs

Github Codespaces - Running pgadmin in docker

With the default instructions and running pgadmin in docker you may receive a blank screen after logging in to the pgadmin console. To resolve this, add the following two environment variables to your pgadmin config to allow it to work with codespace's reverse proxy:

```
PGADMIN_CONFIG_PROXY_X_HOST_COUNT: 1
PGADMIN_CONFIG_PROXY_X_PREFIX_COUNT: 1
```

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:

And don't forget to update the wsl in powershell the command is wsl –update

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 credentials(Windows Subsystem for Linux). Url

You can find the detailed instructions to do so here: rt
ghttps://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 - docker pull dbpage

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'

```
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
```

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

Docker - "permission denied" error when creating a PostgreSQL Docker with a mounted volume on macOS M1

Issue Description:

When attempting to run a Docker command similar to the one below:

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 \mount

 postgres:13

You encounter the error message:

docker: Error response from daemon: error while creating mount source path '/path/to/ny_taxi_postgres_data': chown /path/to/ny_taxi_postgres_data: permission denied.

Solution:

1- Stop Rancher Desktop:

> If you are using Rancher Desktop and face this issue, stop Rancher Desktop to resolve compatibility problems.

2- Install Docker Desktop:

Install Docker Desktop, ensuring that it is properly configured and has the required permissions.

2-Retry Docker Command:

Run the Docker command again after switching to Docker Desktop. This step resolves compatibility issues on some systems.

Note: The issue occurred because Rancher Desktop was in use. Switching to Docker Desktop resolves compatibility problems and allows for the successful creation of PostgreSQL containers with mounted volumes for the New York Taxi Database on macOS M1.

Docker - can’t delete local folder that mounted to docker volume

When I runned command to create postgre in docker container it created folder on my local machine to mount it to volume inside container. It has write and read protection and owned by user 999, so I could not delete it by simply drag to trash. My obsidian could not started due to access error, so I had to change placement of this folder and delete old folder by this command:

```
sudo rm -r -f docker_test/
```

- where `rm` - remove, `-r` - recursively, `-f` - force, `docker_test/` - folder.

Docker - 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

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

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.**

Docker - 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'
```

error:

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
```

Docker - 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
```

Docker - ny_taxi_postgres_data is empty

Even after properly running the docker script the **folder is empty** in the vs code then try this (**For Windows**)

```
 winpty docker run -it \
  -e POSTGRES_USER="root" \
  -e POSTGRES_PASSWORD="root" \
  -e POSTGRES_DB="ny_taxi" \
  -v "C:\Users\abhin\dataengg\DE_Project_git_connected\DE_OLD\week1_set_up\docker_sql/ny_taxi_postgres_data:/var/lib/postgresql/data" \
  -p 5432:5432 \
  postgres:13
```

Here **quoting the absolute path in the -v parameter** is solving the issue and all the files are visible in the Vs-code ny_taxi folder as shown in the video.

**Note: Check he example for the direction of the / **

**Another possible solution for windows, make sure to finish the folder path with a forward slash / :

```
 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 - 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.

brew install conflict with docker desktop and command line tools. You need to install docker desktop first and then the command line tools.
[Issue](https://github.com/Homebrew/brew/issues/16309)

```
brew install –cask docker
```

```
brew install docker docker-compose
```

Docker - 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 use The database cluster will be initialized with locale
"en_US.utf8".
The default database 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: f

error: could not change permissions of directory
"/var/lib/postgresql/data": Operation not permitted  volume
```

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

The input `dtc_postgres_volume_local` must match in both commands below

```
$ 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
```

To verify the above command works in (WSL2 Ubuntu 22.04, verified 2024-Jan), go to the Docker Desktop app and look under **Volumes -** `dtc_postgres_volume_local` would be listed there. The folder `ny_taxi_postgres_data` would however be empty, since we used an alternative config.

```
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
```

Docker - 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 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:1
```

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 creates a folder called "**`ny_taxi_postgres_data;C`**" suggests you have problems with volume mapping, try deleting both folders and replacing "`-v`" part with other options. For me "***//c/***" works instead of "/c/". And it will work by automatically creating a correct folder called "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).

Yes for windows use the command it works perfectly fine

- `-v /"$(pwd)"/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
```

```
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-format-repository-name-must-be-lowercase

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

Change the mounting path. Replace it with one of following:

- `-v /e/zoomcamp/...:/var/lib/postgresql/data`
- `-v /c:/.../ny_taxi_postgres_data:/var/lib/postgresql/data\` (leading slash in front of c:)

Docker - Error response from daemon: error while creating buildmount source path '/run/desktop/mnt/host/c/<your path>': mkdir /run/desktop/mnt/host/c: file exists

When you run this command second time

```
docker run -it \

  -e POSTGRES_USER="root" \

  -e POSTGRES_PASSWORD="root" \

  -e POSTGRES_DB="ny_taxi" \

  -v <your path>:/var/lib/postgresql/data \

  -p 5432:5432 \

  postgres:13
```

The error message above could happen. That means you should not mount on the second run. This command helped me:

When you run this command second time

```
docker run -it \

  -e POSTGRES_USER="root" \

  -e POSTGRES_PASSWORD="root" \

  -e POSTGRES_DB="ny_taxi" \

  -p 5432:5432 \

  postgres:13
```

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

You can fix the problem by changing the permission of the directory on ubuntu with following command:

sudo chown -R $USER dir_path

On windows follow the link: **https://thegeekpage.com/take-ownership-of-a-file-folder-through-command-prompt-in-windows-10/**

Added by Kenan Arslanbay

Docker - 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 deinstall docker and install via the official website

Bind for 0.0.0.0:5432 failed: port is a

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 comtionmand;

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

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

Docker - Docker network name

Get the network name via: $ docker network ls.

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.

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

Typical error:`n.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

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

terraformRun this command before starting your VM:

- On Intel CPU:

  ```
  modprobe -r kvm_intel
  ```

  ```
  modprobe kvm_intel nested=1
  ```

- On AMD CPU:

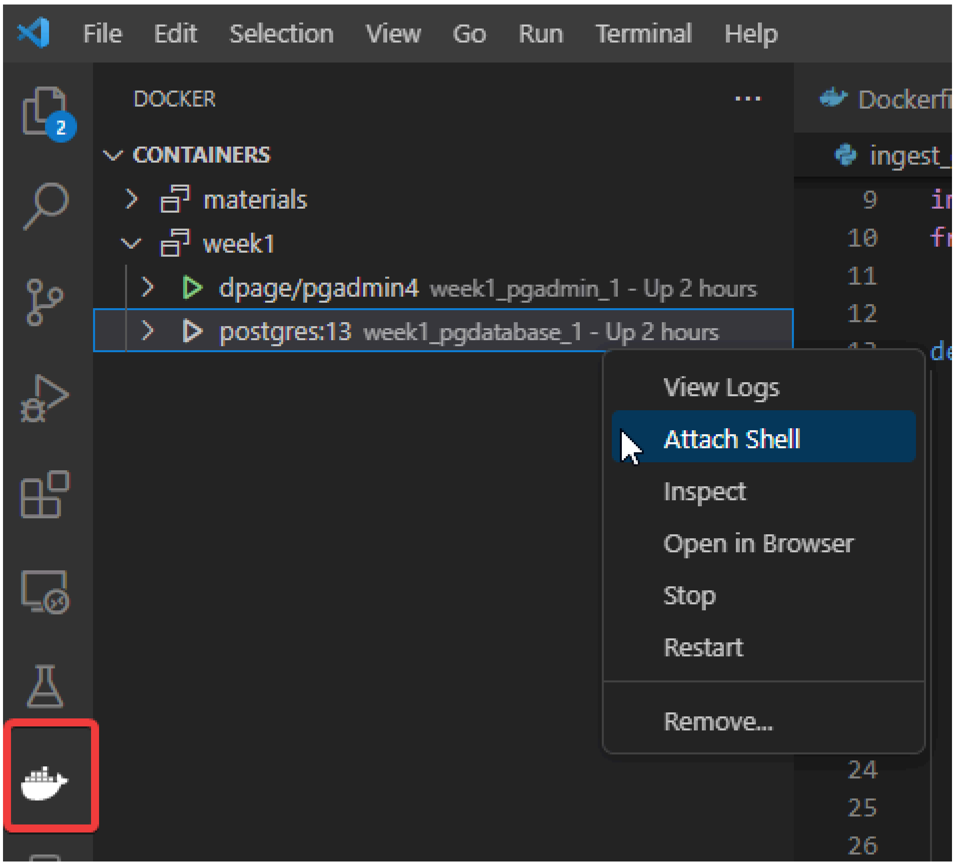
  ```
  modprobe -r kvm_amd
  ```

  ```
  modprobe kvm_amd nested=1
  ```

Docker - Connecting 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.

Docker - How to stop a container?

Use the following command:

```
$ docker stop <container_id>
```

Docker - PostgreSQL Database directory appears to contain a database. Database system is shut down

When you see this in logs, your container with postgres is not accepting any requests, so if you attempt to connect, you'll get this error:

connection failed: server closed the connection unexpectedly

This probably means the server terminated abnormally before or while processing the request.

In this case, you need to delete the directory with data (the one you map to the container with the -v flag) and restart the container.

Solution 2:

If your data is critical, you may be able to reset the write-ahead lock from within the docker container (see here)

```
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 \
  --network pg-network \
  postgres:13 \
  /bin/bash -c 'gosu postgres pg_resetwal /var/lib/postgresql/data'
```

Docker not installable on Ubuntu

On some versions of Ubuntu, snap command can be used to install Docker.

sudo snap install docker

Docker-Compose - mounting error

```
error: could not change permissions of directory
"/var/lib/postgresql/data": Operation not permitted  volume
```

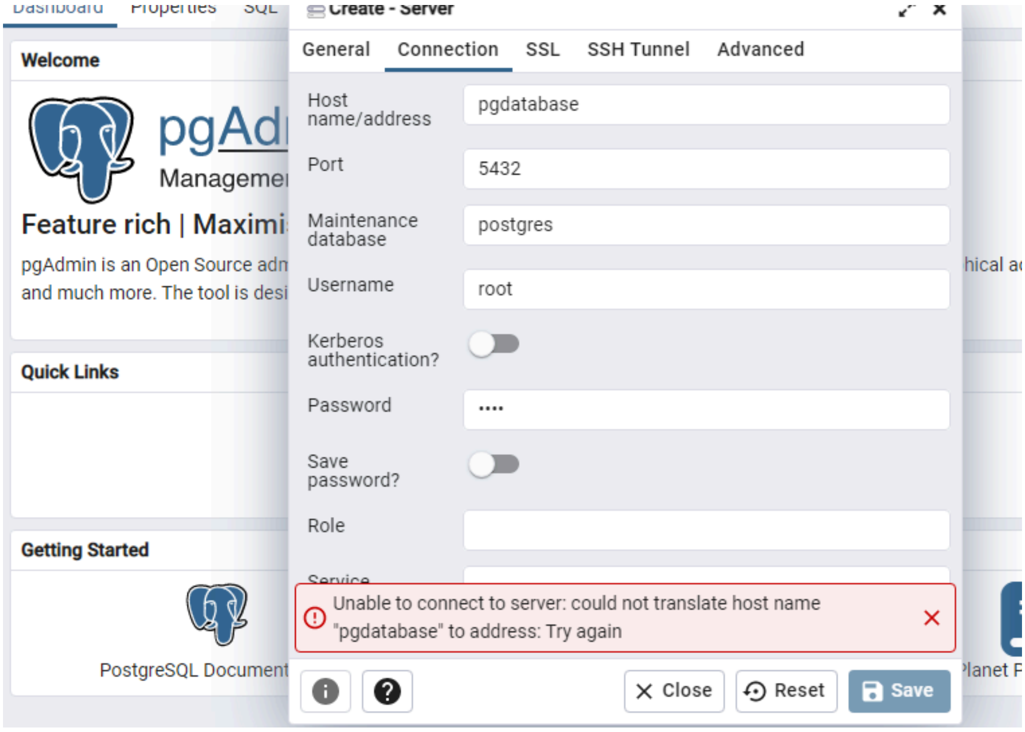
if you have used the prev answer (just before this) and have created a local docker volume, then you need to tell the compose file about the named volume:

```
volumes:
dtc_postgres_volume_local:  # Define the named volume here

# services mentioned in the compose file auto become part of the same network!
services:
your remaining code here . . .
```

- now use docker volume inspect dtc_postgres_volume_local to see the location by checking the value of Mountpoint
- In my case, after i ran docker compose up the mounting dir created was named ‘docker_sql_dtc_postgres_volume_local’ whereas it should have used the already existing ‘dtc_postgres_volume_local’
- All i did to fix this is that I renamed the existing ‘dtc_postgres_volume_local’ to ‘docker_sql_dtc_postgres_volume_local’ and removed the newly created one (just be careful when doing this)
- run docker compose up again and check if the table is there or not!

Docker-Compose - Error translating host name to address

Couldn’t translate host name to address



Host name/address: pgdatabase

Port: 5432

Maintenance database: postgres

Username: root

Kerberos authentication?

Password: ••••

Save password?

Role:

Service

Unable to connect to server: could not translate host name "pgdatabase" to address: Try again

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                  CREATED          STATUS
PORTS                                                               NAMES

faf05090972e   postgres:13       "docker-entrypoint.s…"   39 seconds ago   Up 37
seconds   0.0.0.0:5432->5432/tcp          pg-database
```

```
6344dcecd58f   dpage/pgadmin4   "/entrypoint.sh"         39 seconds ago   Up 37
seconds   443/tcp, 0.0.0.0:8080->80/tcp   pg-admin
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 - Data retention (could not translate host name "pg-database" to address: Name or service not known)

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 /data_pgadmin:/var/lib/pgadmin"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.

If problems persist with pgcli, we can use HeidiSQL

Krishna Anand

Docker-Compose - Hostname does not resolve

It returns --> `Error response from daemon: network 66ae65944d643fdebbc89bd0329f1409dec2c9e12248052f5f4c4be7d1bdc6a3 not found`

Try:

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

d to nuke all the containers

Try: `docker-compose up -d` again ports

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
```

Docker-Compose + PgAdmin – no database in PgAdmin

When you login into PgAdmin and see empty database, the solution below can help:

When you run

```
docker-compose up
```

and at the same time

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

with

```
docker run -it \
  --network=pg-network \ ← <---- NETWORK NAME IS THE SAME AS THAT CREATED BY DOCKER COMPOSE
  taxi_ingest:v001 \
    --user=postgres \
    --password=postgres \
    --host=db \
    --port=5432 \
    --db=ny_taxi \
    --table_name=green_tripdata \
    --url=${URL}
```

It’s important to use the same `--network` which states in the file docker-compose.yaml (`networks`, as mentioned above). OR The file docker-compose.yaml might not specify a network, as in the example below.

```
services:
  db:
    container_name: postgres
    image: postgres:17-alpine
    environment:
      ...
    ports:
      - '5433:5432'
    volumes:
      - ...
  pgadmin:
    container_name: pgadmin
    image: dpage/pgadmin4:latest
    environment:
      ...
    ports:
      - "8080:80"
    volumes:
      - ...
volumes:
  vol-pgdata:
    name: vol-pgdata
  vol-pgadmin_data:
    name: vol-pgadmin_data
```

In this case, the network name is generated automatically: The name of the directory containing the `docker-compose.yaml` file in lowercase + `_default`.

You can find the network's name during running docker-compose up

```
pg-database Pulling
pg-database Pulled
Network week_1_default  Creating <-- THIS ONE
Network week_1_default  Created
```

Docker-Compose - Persist PGAdmin docker contents on GCP

So one common issue is when you run docker-compose on GCP, postgres won't persist it's 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:
```

Docker engine stopped_failed to fetch extensions

- The docker will keep on crashing continuously
- Not working after restart

docker engine stopped

And failed to fetch extensions pop ups will on screen non-stop

Solution :

1. Try checking if latest version of docker is installed / Try updating the docker
2. If Problem still persist then final solution is to reinstall docker
3. (Just have to fetch images again else no issues)

Docker-Compose - Persist PGAdmin configuration

As per the lessons,

Persisting pgAdmin configuration (i.e. server name) is done by adding a “volumes” section:

services:

 pgdatabase:

[...]

pgadmin:

   image: dpage/pgadmin4

```
  environment:
    - PGADMIN_DEFAULT_EMAIL=admin@admin.com
    - PGADMIN_DEFAULT_PASSWORD=root
  volumes:
    - "./pgAdmin_data:/var/lib/pgadmin/sessions:rw"
  ports:
    - "8080:80"
```

In the example above, ”pgAdmin_data” is a folder on the host machine, and “/var/lib/pgadmin/sessions” is the session settings folder in the pgAdmin container.

Before running docker-compose up on the YAML file, we also need to give the pgAdmin container access to write to the “pgAdmin_data” folder. The container runs with a username called “5050” and user group “5050”. The bash command to give access over the mounted volume is:

```
sudo chown -R 5050:5050 pgAdmin_data
```

Docker-Compose - 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:

guides/docker-without-sudo.md at main · sindresorhus/guides · GitHub

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.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:ls

```
volumes:
  Pgadmin_data:
```

Docker-Compose - 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`.

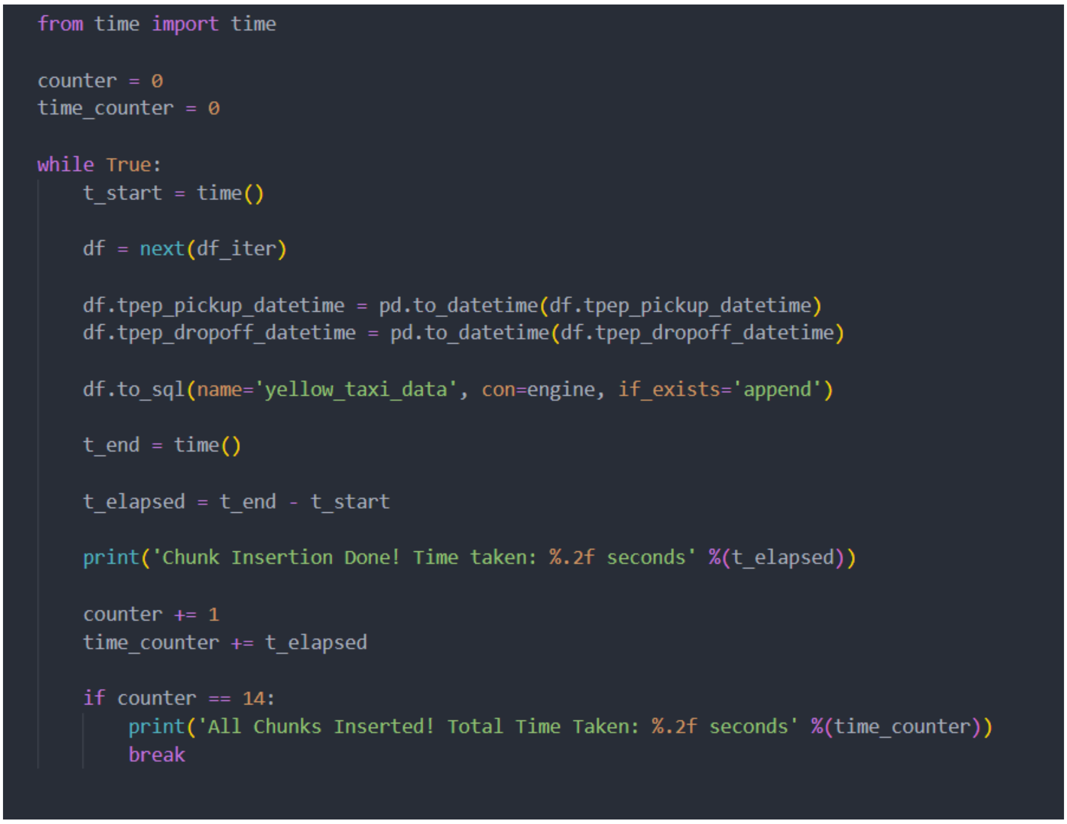
Docker-Compose - Error getting credentials after running docker-compose up -d

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 - Errors pertaining to docker-compose.yml and pgadmin setup

For everyone who's having problem with Docker compose, getting the data in postgres and similar issues, please take care of the following:

- create a new volume on docker (either using the command line or docker desktop app)
- make the following changes to your docker-compose.yml file (see attachment)
- set low_memory=false when importing the csv file (df = pd.read_csv('yellow_tripdata_2021-01.csv', nrows=1000, low_memory=False))
- use the below function (in the upload-data.ipynb) for better tracking of your ingestion process (see attachment)



```
from time import time

counter = 0
time_counter = 0

while True:
    t_start = time()

    df = next(df_iter)

    df.tpep_pickup_datetime = pd.to_datetime(df.tpep_pickup_datetime)
    df.tpep_dropoff_datetime = pd.to_datetime(df.tpep_dropoff_datetime)

    df.to_sql(name='yellow_taxi_data', con=engine, if_exists='append')

    t_end = time()

    t_elapsed = t_end - t_start

    print('Chunk Insertion Done! Time taken: %.2f seconds' %(t_elapsed))

    counter += 1
    time_counter += t_elapsed

    if counter == 14:
        print('All Chunks Inserted! Total Time Taken: %.2f seconds' %(time_counter))
        break
```

- Order of execution:
    - ○ (1) open terminal in 2_docker_sql folder and run docker compose up
    - ○ (2) ensure no other containers are running except the one you just executed (pgadmin and pgdatabase)
    - ○ (3) open jupyter notebook and begin the data ingestion
    - ○ (4) open pgadmin and set up a server (make sure you use the same configurations as your docker-compose.yml file like the same name (pgdatabase), port, databasename (ny_taxi) etc.

Docker Compose up -d error getting credentials - err: exec: "docker-credential-desktop": executable file not found in %PATH%, out: ``

Locate config.json file for docker (check your home directory; Users/username/.docker).

Modify credsStore to credStore

Save and re-run

Docker-Compose - 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
```

Docker-Compose - Error undefined volume in Windows/WSL

If you wrote the docker-compose.yaml file exactly like the video, you might run into an error like this:dev

```
service "pgdatabase" refers to undefined volume
dtc_postgres_volume_local: invalid compose project
```

In order to make it work, you need to include the volume in your docker-compose file. Just add the following:

```
volumes:

  dtc_postgres_volume_local:
```

Docker-Compose - cannot execute binary file: Exec format error

This error means the docker-compose executable can’t be opened in current OS. Make sure the file you download from github matches your system environment.

As of 2025/1/17, docker-compose (v2.32.4) docker-compose-linux-aarch64 does not work, try v2.32.3 docker-compose-linux-x86_64

Docker-Compose - Postgres container fails to launch with exit code (1) when attempting to compose

This happens due to the Postgres database not being initialized before running docker-compose up -d. There are other potential ways around it (thread) but you can simply initialize the database first and the compose will work afterward.

```
docker run -it \

-e POSTGRES_USER="root" \

-e POSTGRES_PASSWORD="root" \

-e POSTGRES_DB="ny_taxi" \

-v $(pwd)/ny_taxi_data:/var/lib/postgresql/data \

-p 5432:5432 \

--network=pg-network \

--name=pg_database \

postgres:13
```

WSL Docker directory permissions error

Error: initdb: error: could not change permissions of directory

Issue: WSL and Windows do not manage permissions in the same way causing conflict if using the Windows file system rather than the WSL file system.

Solution: Use Docker volumes.

> **Why:** Volume is used for storage of persistent data and not for use of transferring files. A local volume is unnecessary.

Benefit: This resolves permission issues and allows for better management of volumes.

NOTE: the ‘user:’ is not necessary if using docker volumes, but is if using local drive.

</> docker-compose.yaml

```
services:
  postgres:
    image: postgres:15-alpine
    container_name: postgres
    user: "0:0"
    environment:
      - POSTGRES_USER=postgres
      - POSTGRES_PASSWORD=postgres
      - POSTGRES_DB=ny_taxi
    volumes:
      - "pg-data:/var/lib/postgresql/data"
    ports:
      - "5432:5432"
    networks:
      - pg-network

  pgadmin:
    image: dpage/pgadmin4
    container_name: pgadmin
    user: "${UID}:${GID}"
    environment:
      - PGADMIN_DEFAULT_EMAIL=email@some-site.com
      - PGADMIN_DEFAULT_PASSWORD=pgadmin
    volumes:
      - "pg-admin:/var/lib/pgadmin"
    ports:
      - "8080:80"
    networks:
      - pg-network


networks:
  pg-network:
    name: pg-network


volumesta:
    name: ingest_pgdata
  pg-admin:
    name: ingest_pgadmin:
  pg-da
```

WSL - 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

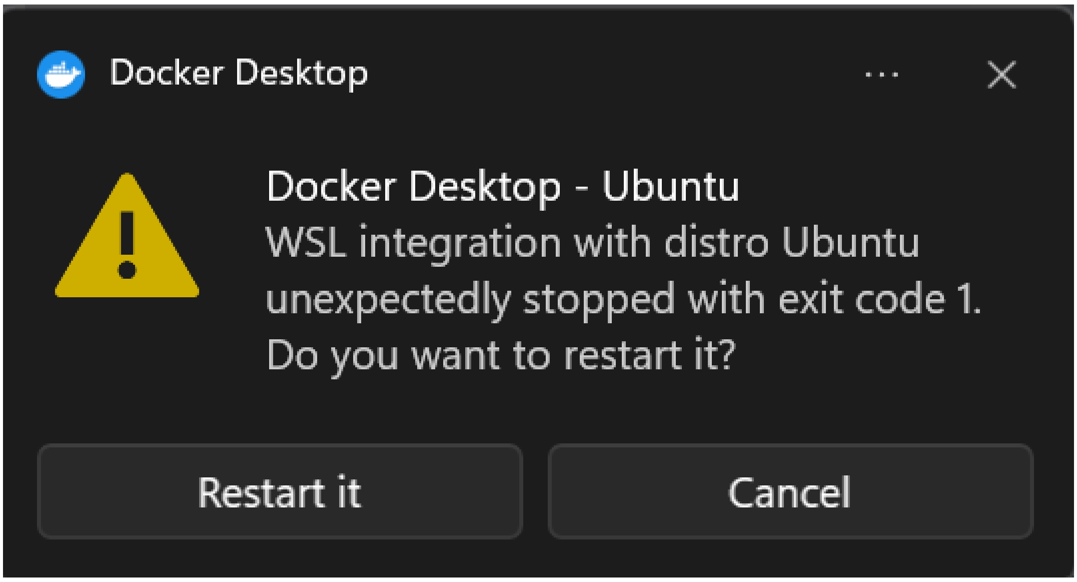
- for updating Windows terminal which worked for me:

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.unexpectedly

WSL - WSL integration with distro Ubuntu unexpectedly stopped with exit code 1.



Docker Desktop - Ubuntu
WSL integration with distro Ubuntu unexpectedly stopped with exit code 1. Do you want to restart it?

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" n

bash: conda: command not found

Database is uninitialized and superuser password is not specified.

Database is uninitialized and superuser password is not specified.

WSL - Permissions too open at Windows

Issue when trying to run the GPC VM through SSH through WSL2, probably because WSL2 isn’t looking for .ssh keys in the correct folder. My case I was trying to run this command in the terminal and getting an error

```
PC:/mnt/c/Users/User/.ssh$ ssh -i gpc [username]@[my external IP]
```

You can try to use sudo before the command

```
Sudo .ssh$ ssh -i gpc [username]@[my external IP]
```

You can also try to cd to your folder and change the permissions for the private key SSH file.

```
chmod 600 gpc
```

If that doesn’t work, create a .ssh folder in the home diretory of WSL2 and copy the content of windows .ssh folder to that new folder.

```
cd ~
```

```
mkdir .ssh
```

```
cp -r /mnt/c/Users/YourUsername/.ssh/* ~/.ssh/
```

You might need to adjust the permissions of the files and folders in the .ssh directory.

WSL - Could not resolve host name

Such as the issue above, WSL2 may not be referencing the correct .ssh/config path from Windows. You can create a config file at the home directory of WSL2.

```
cd ~
```

```
mkdir .ssh
```

Create a config file in this new .ssh/ folder referencing this folder:

```
  HostName [GPC VM external IP]
  User [username]
```

```
  IdentityFile ~/.ssh/[private key]
```

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

Change TO Socket

```
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
```

PGCLI - should we run pgcli inside another docker container?

In this section of the course, the 5432 port of pgsql 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.

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

For a more visual and detailed explanation, feel free to check the video 1.4.2 - Port Mapping and Networks in Docker

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`wi ``

- Or list the running postgres services on your local machine with launchctl

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.

PGCLI - 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/sitYe-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'
```

Solution 1:

This error indicates that your user doesn’t have the necessary permissions to access or modify the specified directory or file (/some/path/.config/pgcli).

This can happen in the context of Docker when privileges were assigned to root and not to the user you have.

For example, if a process inside the container creates the file as root, your user might not have write permissions to that file on the host.

To resolve this:

- **Check file permissions** on the directory /some/path/.config/pgcli and ensure that your user has read/write access. You can do this with the command:

```
ls -l /some/path/.config/pgcli
```

- **Change ownership/permissions** of the file or directory so that your user has the necessary permissions. For example, to grant your user read/write permissions, use:

```
sudo chown -R user_name /Users/user_name/.config
```

  - The **sudo** stands for Super User DO
  - The **chown** means change owner
  - **-R** is doing so recursively
  - **User_name** is the name you gave to your PC (e.g. vray)

Solution 2:

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 - no pq wrapper available.

```
ImportError: no pq wrapper available.

Attempts made:

- couldn't import \dt

opg '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 better off with a new environment with 3.9**

```
$ conda create --name 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

$ pip install psycopg_binary
```

3. If above steps do not work, try:

```
$ pip install --upgrade pgcli
```

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

```
$ pgcli -h localhost -U root -d ny_taxisudo
```

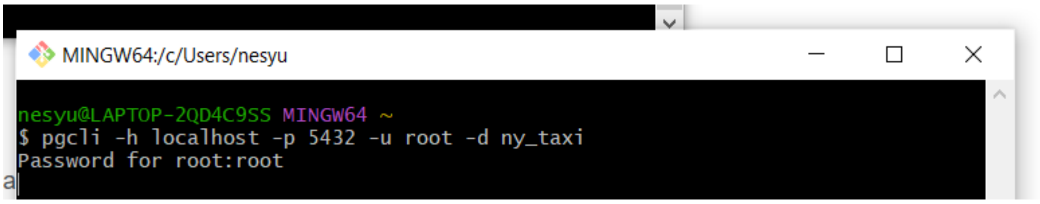
There, you should be good to go now!

Another solution:

Run this

PGCLI - stuck on password prompt

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



```
nesyu@LAPTOP-2QD4C9SS MINGW64 ~
$ pgcli -h localhost -p 5432 -u root -d ny_taxi
Password for root:root
```

Use winpty:

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

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

PGCLI -connection failed: FATAL: password authentication failed for user "root"

The error above was faced continually despite inputting the correct password

Solution

Option 1: Stop the PostgreSQL service on Windows

Option 2 (using WSL): Completely uninstall Protgres 12 from Windows and install postgresql-client on WSL (sudo apt install postgresql-client-common postgresql-client libpq-dev)

Option 3: Change the port of the docker container

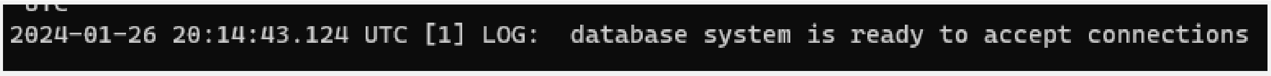
Option 4: `NEW SOLUTION: 27/01/2024`

PGCLI -connection failed: FATAL: password authentication failed for user "root"

If you’ve got the error above, it’s probably because you were just like me, closed the connection to the Postgres:13 image in the previous step of the tutorial, which is

```
docker run -it \
        -e POSTGRES_USER=root \

        -e POSTGRES_PASSWORD=root \

        -e POSTGRES_DB=ny_taxi \

        -v
d:/git/data-engineering-zoomcamp/week_1/docker_sql/ny_taxi_postgres_data:/var/lib/postgresql/data \

        -p 5432:5432 \

        postgres:13
```

So keep the database connected and you will be able to implement all the next steps of the tutorial.



```
2024-01-26 20:14:43.124 UTC [1] LOG:  database system is ready to accept connections
```

Option 5: Change the Port for Docker PostgreSQL
After running the command: pgcli -h localhost -p 5432 -u root -d ny_taxi User get the enter password prompt and despite using the correct one, the error persist. This is provably due to user having installed Postgres in local machine. The easiest solution to this port conflict between host and container is by Changing the Port for Docker PostgreSQL: You can configure your Docker PostgreSQL container to use a different port. This way, it won't conflict with the PostgreSQL instance running on your local machine. When running the PostgreSQL container, map it to a different port on your host machine. E.g.:

```
 docker run -it \\
  -e POSTGRES_USER="root" \\
  -e POSTGRES_PASSWORD="root" \\
  -e POSTGRES_DB="ny_taxi" \\
  -v c:/workspace/de-zoomcamp/1_intro_to_data_engineering/docker_sql/ny_taxi_postgres_data:/var/lib/postgresql/data \\
  -p 5433:5432 \\
  Postgres:13
```

- 5433 refers to the port on the host machine.
- 5432 refers to the port inside the Docker Postgres container.

PGCLI - 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:

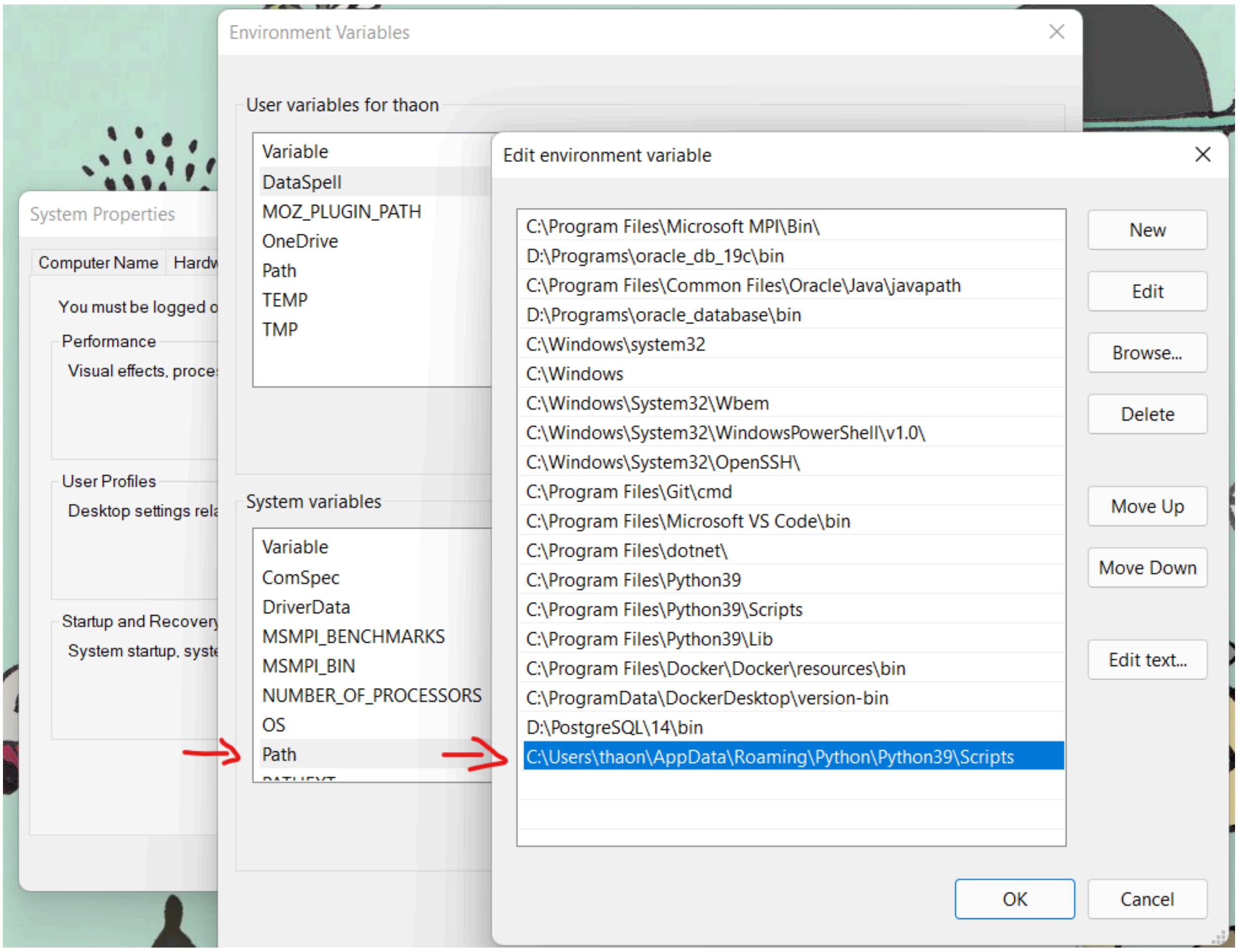
1. Get the location: `pip list -v`
2. Copy
   `C:\Users\...\AppData\Roaming\Python\Python39\site-packages`
3. 3. Replace site-packages with Scripts:
   `C:\Users\...\AppData\Roaming\Python\Python39\Scripts`

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



```
C:\Program Files\Microsoft MPI\Bin\
D:\Programs\oracle_db_19c\bin
C:\Program Files\Common Files\Oracle\Java\javapath
D:\Programs\oracle_database\bin
C:\Windows\system32
C:\Windows
C:\Windows\System32\Wbem
C:\Windows\System32\WindowsPowerShell\v1.0\
C:\Windows\System32\OpenSSH\
C:\Program Files\Git\cmd
C:\Program Files\Microsoft VS Code\bin
C:\Program Files\dotnet\
C:\Program Files\Python39
C:\Program Files\Python39\Scripts
C:\Program Files\Python39\Lib
C:\Program Files\Docker\Docker\resources\bin
C:\ProgramData\DockerDesktop\version-bin
D:\PostgreSQL\14\bin
C:\Users\thaon\AppData\Roaming\Python\Python39\Scripts
```

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

PGCLI - running in a Docker container

In case running `pgcli` locally causes issues or you do not want to install it locally you can use it running in a Docker container instead.

Below the usage with values used in the videos of the course for:

- network name (docker network)
- postgres related variables for pgcli
  - Hostname
  - Username
  - Port
  - Database name

```
$ docker run -it --rm --network pg-network ai2ys/dockerized-pgcli:4.0.1

175dd47cda07:/# pgcli -h pg-database -U root -p 5432 -d ny_taxi

Password for root:

Server: PostgreSQL 16.1 (Debian 16.1-1.pgdg120+1)

Version: 4.0.1

Home: http://pgcli.com

root@pg-database:ny_taxi> \dt

+--------+------------------+-------+-------+

| Schema | Name             | Type  | Owner |

|--------+------------------+-------+-------|

| public | yellow_taxi_data | table | root  |

+--------+------------------+-------+-------+

SELECT 1

Time: 0.009s

root@pg-database:ny_taxi>
```

RRPGCLI - case sensitive use “Quotations” around columns with capital letters

PULocationID will not be recognized but “PULocationID” will be. This is because unquoted "Localidentifiers are case insensitive. See docs.

PGCLI - error column c.relhasoids does not exist

When using the command `` `\d <database name>` `` you get the error column `` `c.relhasoids does not exist` ``.

Resolution:

1. Uninstall pgcli
2. Reinstall pgclidatabase "ny_taxi" does not exist
3. Restart pc

Postgres - bind: address already in use

1.2.2 Postgres commandline for docker

1. Various errors when first pasting docker run command - make sure there is only 1 space before “\” and only a newline after “\”
2. Error - posgres post is already in use. This seems to happen every time i try to start the docker postgres container.

```
$ docker run -it   -e POSTGRES_USER="root"   -e POSTGRES_PASSWORD="root"   -e
POSTGRES_DB="ny_taxi"   -v
./ny_taxi_postgres_data:/var/lib/postgresql/data:rw   -p  5432:5432
postgres:13

docker: Error response from daemon: driver failed programming external
connectivity on endpoint jolly_chatterjee
(9e21c5bf0aa3dcc711185bc6fb1dc7b2722fc568fa47655dab98ab55ff8c23f2): failed to
bind port 0.0.0.0:5432/tcp: Error starting userland proxy: listen tcp4
0.0.0.0:5432: bind: address already in use.

$ sudo lsof -i :5432
COMMAND   PID     USER   FD   TYPE DEVICE SIZE/OFF NODE NAME
postgres 3082 postgres  3u  IPv4  31345   0t0  TCP localhost:postgresql
(LISTEN)

$ sudo service postgresql stop
```

Option 1: Figure out what service is using the port (`sudo lsof -i :5432)` and stop that service: `sudo service postgresql stop`.

Option 2: more long term.
I actually eventually ended up mapping to a different port, because this happened every time I restarted my VM. So I would map <local 5433: container 5432> in the docker file or docker compose file. Since i am using a VM, I also need to make sure that port 5433 is forwarded.

PGCLI - After installing PGCLI and checking with pgcli --help we get the error: ImportError: no pq wrapper available

The error persists because the psycopg library cannot find the required libpq library. Ensure the required PostgreSQL client library is installed:

```
sudo apt install libpq-dev
```

Rebuild psycopg

```
pip uninstall psycopg psycopg_binary psycopg_c -y
pip install psycopg --no-binary psycopg
```

The issue should be resolved by now. However, even after these steps you get the error:
ModuleNotFoundError: No module named 'psycopg2'
Then run the following:

```
pip install psycopg2-binary
```

Postgres - OperationalError: (psycopg2.OperationalError) connection to server at "localhost" (::1), port 5432 failed: FATAL: password authentication failed for user "root"

This happens while uploading data via the connection in jupyter notebook

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

The port 5432 was taken by another postgres. You could already have installed Postgres in the past at the same port, so when you are trying to connect it does not reach docker, but the old Postgres installation instead. We are not connecting to the port in docker, but to the port on our machine. Substitute 5431 or whatever port you mapped to for port 5432. Another option is to remove the old Postgres installation if it is useless.

Also if this error is still persistent , kindly check if you have a service in windows running postgres. Stopping that service will resolve the issue

Postgres - connection failed: connection to server at "127.0.0.1", port 5432 failed: FATAL: password authentication failed for user "root"

check that the port was properly forwarded. If 5432 is being used, kill the process:

```
sudo lsof -i :5432
```

sudo kill -9 PID

Windows users:

Found that my issue was related to PostgresSQL running locally on my machine and that pgAdmin4 was using my 5432 port.

To stop this process:

1. Press Win + R to open the Run dialog.
2. Type services.msc and press Enter.
3. In the Services window, scroll down and look for a service with a name like PostgreSQL, postgresql-x64-13, or similar (the exact name depends on your PostgreSQL version).
4. Right-click the PostgreSQL service and select Stop.

Postgres - 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).

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

Other solution that worked:

Changing `` `POSTGRES_USER=juroot` `` to `` `PGUSER=postgres` ``

Based on this: postgres with docker compose gives FATAL: role "root" does not exist error - Stack Overflow

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

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

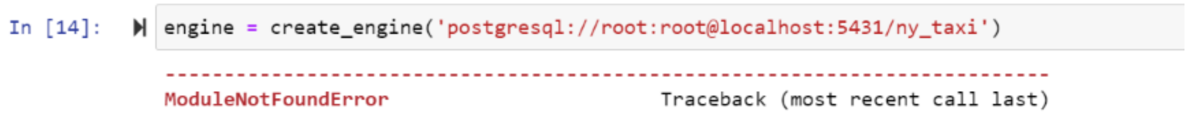
```
~\anaconda3\lib\site-packages\psycopg2\__init__.py in connect(dsn, connection_factory,
cursor_factory, **kwargs)
    120
    121     dsn = _ext.make_dsn(dsn, **kwargs)
--> 122     conn = _connect(dsn, connection_factory=connection_factory, **kwasync)
    123     if cursor_factory is not None:
  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

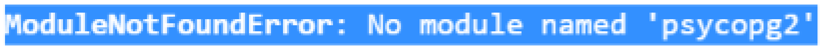
Postgres - ModuleNotFoundError: No module named 'psycopg2'

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 already have it, you might need to update it:

pip install psycopg2-binary --upgrade

Other methods, if the above fails:

- if you are getting the " ModuleNotFoundError: No module named 'psycopg2' " error even after the above installation, then try updating conda using the command conda update -n base -c defaults conda. Or if you are using pip, then try updating it before installing the psycopg packages i.e
    - First uninstall the psycopg package
    - Then update conda or pip
    - Then install psycopg again using pip.
- if you are still facing error with r pcycopg2 and showing pg_config not found then you will have to install postgresql. in MAC it is **brew install postgresql**

Postgres - "Column does not exist" but it actually does (Pyscopg2 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

pgAdmin - Create server dialog does not appear

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

pgAdmin - Blank/white screen after login (browser)

Using GitHub Codespaces in the browser resulted in a blank screen after the login to pgAdmin (running in a Docker container). The terminal of the pgAdmin container was showing the following error message:

```
CSRFError: 400 Bad Request: The referrer does not match the host.
```

Solution #1:

As recommended in the following issue https://github.com/pgadmin-org/pgadmin4/issues/5432 setting the following environment variable solved it.

```
PGADMIN_CONFIG_WTF_CSRF_ENABLED="False"
```

Modified “docker run” command

```
docker run --rm -it \

    -e PGADMIN_DEFAULT_EMAIL="admin@admin.com" \

    -e PGADMIN_DEFAULT_PASSWORD="root" \

    -e PGADMIN_CONFIG_WTF_CSRF_ENABLED="False" \

    -p "8080:80" \

    --name pgadmin \

    --network=pg-network \

    dpage/pgadmin4:8.2
```

Solution #2:

Using the local installed VSCode to display GitHub Codespaces.

When using GitHub Codespaces in the locally installed VSCode (opening a Codespace or creating/starting one) this issue did not occur.

pgAdmin - Can not access/open the PgAdmin address via browser

I am using a Mac Pro device and connect to the GCP Compute Engine via Remote SSH - VSCode. But when I trying to run the PgAdmin container via docker run or docker compose command, I am failed to access the pgAdmin address via my browser. I have switched to another browser, but still can not access the pgAdmin address. So I modified a little bit the configuration from the previous DE Zoomcamp repository like below and can access the pgAdmin address:

Solution #1:

Modified “docker run” command

```
docker run --rm -it \
    -e PGADMIN_DEFAULT_EMAIL="admin@admin.com" \
    -e PGADMIN_DEFAULT_PASSWORD="pgadmin" \
    -e PGADMIN_CONFIG_WTF_CSRF_ENABLED="False" \
    -e PGADMIN_LISTEN_ADDRESS=0.0.0.0 \
    -e PGADMIN_LISTEN_PORT=5050 \
    -p 5050:5050 \
    --network=de-zoomcamp-network \
    --name pgadmin-container \
    --link postgres-container \
    -t dpage/pgadmin4
```

Solution #2:

Modified docker-compose.yaml configuration (via “docker compose up” command)

```
  pgadmin:
    image: dpage/pgadmin4
    container_name: pgadmin-conntainer
    environment:
      - PGADMIN_DEFAULT_EMAIL=admin@admin.com
      - PGADMIN_DEFAULT_PASSWORD=pgadmin
      - PGADMIN_CONFIG_WTF_CSRF_ENABLED=False
      - PGADMIN_LISTEN_ADDRESS=0.0.0.0
      - PGADMIN_LISTEN_PORT=5050
    volumes:
      - "./pgadmin_data:/var/lib/pgadmin/data"
    ports:
      - "5050:5050"
    networks:
      - de-zoomcamp-network
    depends_on:
      - postgres-conntainer
```

pgAdmin - How to Persist pgAdmin Configurations

Question: How can I keep pgAdmin settings after restarting the container?

Answer: Create a directory, map it to /var/lib/pgadmin, and fix permissions:

Create the directory for pgAdmin data:
# mkdir -p /path/to/pgadmin-data

Assign ownership to pgAdmin's user (ID 5050):
# sudo chown -R 5050:5050 /path/to/pgadmin-data

# sudo chmod -R 755 /path/to/pgadmin-data

pgAdmin - Unable to connect to server: [Errno -3] Try again

This error occurs in connecting pgAdmin with Docker Postgres. In tutorial, in the pgAdmin server creation under Connection > Host name/address: pg-database is given and resulted in the above mentioned error when saved.

Solution 1:

- Verify that both containers are connected to `pg-network` : `docker network inspect pg-network`
- If Docker Postgres container is not connected, then connect it to `pg-network`: `docker network connect pg-network postgresContainer_name`
- Retry connection, and if error persist, instead of using `pg-database` under `Connection > Host name/address: pg-database`, **Try using IP Address:** Use the IP address of the `postgresContainer_name` container e.g.(172.19.0.3) in the pgAdmin configuration instead of the container name or pg-database.

Python - ModuleNotFoundError: No module named 'pysqlite2'

```
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)

Python - 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 w(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

iPython - Pandas parsing dates with ‘read_csv’

Pandas can interpret “string” column values as “datetime” directly when reading the CSV file using “pd.read_csv” using the parameter “parse_dates”, which for example can contain a list of column names or column indices. Then the conversion afterwards is not required anymore.

pandas.read_csv — pandas 2.1.4 documentation (pydata.org)

Example from week 1

```
import pandas as pd
df = pd.read_csv(
    'yellow_tripdata_2021-01.csv',
    nrows=100,
    parse_dates=['tpep_pickup_datetime', 'tpep_dropoff_datetime'])
df.info()
```

which will output

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 18 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   VendorID               100 non-null    int64
 1   tpep_pickup_datetime   100 non-null    datetime64[ns]
 2   tpep_dropoff_datetime  100 non-null    datetime64[ns]
 3   passenger_count        100 non-null    int64
 4   trip_distance          100 non-null    float64
 5   RatecodeID             100 non-null    int64
 6   store_and_fwd_flag     100 non-null    object
 7   PULocationID           100 non-null    int64
 8   DOLocationID           100 non-null    int64
 9   payment_type           100 non-null    int64
 10  fare_amount            100 non-null    float64
 11  extra                  100 non-null    float64
 12  mta_tax                100 non-null    float64
 13  tip_amount             100 non-null    float64
 14  tolls_amount           100 non-null    float64
 15  improvement_surcharge  100 non-null    float64
 16  total_amount           100 non-null    float64
 17  congestion_surcharge   100 non-null    float64
dtypes: datetime64[ns](2), float64(9), int64(6), object(1)
memory usage: 14.2+ KB
```

Python - Python cant ingest data from the github link provided using curl

os.system(f"curl -LO {url} -o {csv_name}")

Python - 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:

```
df = pd.read_csv('file.csv.gz'
                 , compression='gzip'
                 , low_memory=False
     )
```

*Python - How to iterate through an*d ingest parquet file

Contrary to panda’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))
```

Python - SQLAlchemy - ImportError: cannot import name 'TypeAliasType' from 'typing_extensions'.

Error raised during the jupyter notebook’s cell execution:

```
from sqlalchemy import create_engine.
```

Solution: Version of Python module “typing_extensions” >= 4.6.0. Can be updated by Conda or pip.

Python - SQLALchemy - TypeError 'module' object is not callable

create_engine('postgresql://root:root@localhost:5432/ny_taxi') I get the error "TypeError: 'module' object is not callable"

Solution:

```
conn_string = "postgresql+psycopg://root:root@localhost:5432/ny_taxi"
engine = create_engine(conn_string)
```

Python - SQLAlchemy - ModuleNotFoundError: No module named 'psycopg2'.

Error raised during the jupyter notebook’s cell execution:

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

Solution: Need to install Python module “psycopg2”. Can be installed by Conda or pip.

Python - SQLAlchemy - NoSuchModuleError: Can't load plugin: sqlalchemy.dialects:postgresql.psycopg

Error raised during the jupyter notebook’s cell execution:

```
conn_string = "postgresql+psycopg://root:root@localhost:5432/ny_taxi"

engine = create_engine(conn_string)
```

Solution: We had a scenario of a virtualenv (created by Pycharm) being run on top of another virtual env (on conda). Solution was:

to get rid of the .venv

create a brand new virtualenv with conda conda create -n pyingest python=3.12

install the required dependencies pip install pandas sqlalchemy psycopg2-binary jupyterlab

And re-execute the code.

For psycopg2, the connection string should be:

```
postgresql+psycopg2://{db_user}:{db_password}@{db_host}:{db_port}/{db_name}
```

Reference - Kayla Tinker 1/14/25

Python - SQLAlchemy - read_sql_query() throws "'OptionEngine' object has no attribute 'execute'"

First, check SQLAlchemy and Pandas version. Make sure they are both up-to-date. Upgrade them using pip/conda if needed.

Then, try to wrap the query using text:

```
from sqlalchemy import text

query = text("""SELECT * FROM tbl""") df = pd.read_sql_query(query, conn)
```

GCP - Static vs Ephemeral IP / Setting up static IP for VM

I had my contig file set up from the first instance of my VM setup, but once I shut the VM down and restarted it later, the config no longer worked. This was because the IP address of my VM had changed, so my config was out of date. I didn’t want to change my config file every time so I wondered if there was a solution – there is!

You can make a static IP address. The default is ephemeral, which changes every time you start/stop. This way, you can keep the same ip address in your config file every time you start/stop the VM.

Set up a static IP in VPC Network > IP addresses. Make sure you attach it to your VM instance to avoid extra fees. You are only charged for a static IP if it is not assigned to a specific virtual machine. There is also pretty good documentation for this on gcp.

GCP - Unable to add Google Cloud SDK PATH to Windows

Unable to add Google Cloud SDK PATH to Windows

Windows error: The installer is unable to automatically update your system PATH. Please add C:\tools\google-cloud-sdk\bin

if you are constantly getting this feedback. Might be that you needed to add Gitbash to your Windows path:

One way of doing that is to use conda: ‘If you are not already using it

Download the Anaconda Navigator

Make sure to check the box (add conda to the path when installing navigator: although not recommended do it anyway)

You might also need to install git bash if you are not already using it(or you might need to uninstall it to reinstall it properly)

Make sure to check the following boxes while you install Gitbash

- Add a GitBash to Windows Terminal
- Use Git and optional Unix tools from the command prompt

Now open up git bash and type **conda init bash** This should modify your bash profile

Additionally, you might want to use Gitbash as your default terminal.

Open your Windows terminal and go to settings, on the default profile change Windows power shell to git bash

GCP - Project creation failed: HttpError accessing … Requested entity alreadytpep_pickup_datetime exists

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_datetimeary': '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

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

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

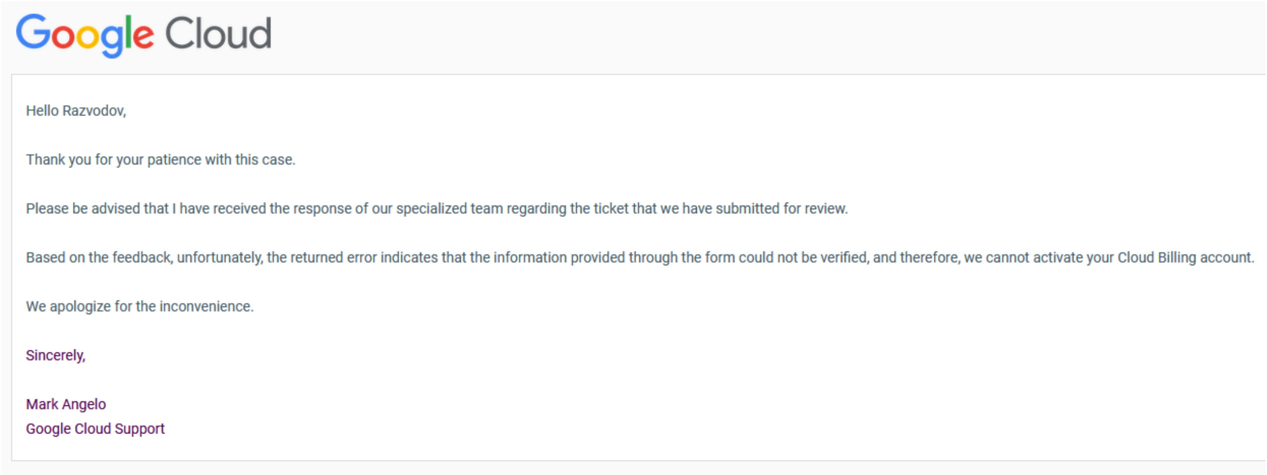
GCP Account Suspension Inquiry

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.

ny-rides.json



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 - Where can I find the “ny-rides.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

GCP - “Failed to load” when accessing Compute Engine’s metadata section (e.g., to add a SSH key)

You likely didn’t enable the Compute Engine API.

GCP - Do I need to delete my instance in Google Cloud?

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.

GCP - ssh public key error - multiple users / usernames

Initially, I could not ssh into my VM from my windows laptop. I thought at first it was because I did not follow along exactly with the tutorial. Instead of generating ssh key using the MINGW/git bash with the linux style command, I did it in command-prompt using the windows style command. I kept getting a public key error.

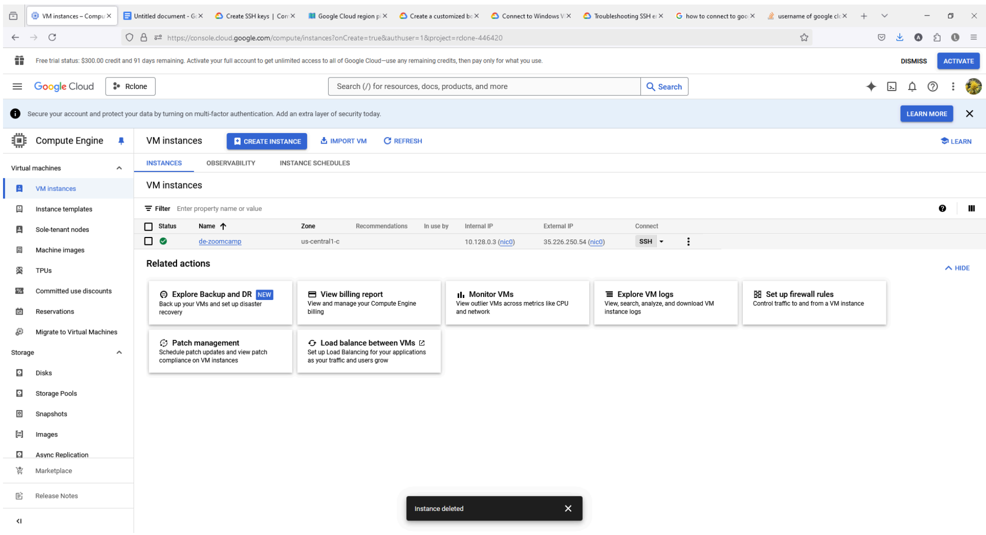
Permanent solution:

It turns out it wasn’t an issue with the keygen at all! It was silly, as with most “bugs.” I had given my ssh key a different username than what showed in my VM (my google account username). So I had been trying to log in with googleacctuser@[ipaddr] instead of mySSHuser@[ipaddr]. I figured this out by retracing my steps to double check that I had set up an ssh key in GCP console, where it showed the user and ssh key. I quickly changed the username to the correct one (googleacctuser) in my config file and it works!

Now, the catch is that I've created two users! I made all the installations, permissions granting, etc. on googleacctuser and it's not accessible from liv. So there's a couple avenues I could take, but since I set up googleacctuser and I don't need mySSHuser, I'm just going to change the username at the end of the ssh key to mySSHuser from mySSHuser on local (open up public gcp ssh file in texteditor), and re-paste that into the GCP console. Then update the config file and use mySSHuser to log in.

Then delete mySSHuser account in the VM terminal just to keep things clean. (i skipped this because i am now a bit attached :))

Temporary solution: Before i figured out my issue, I took a shortcut by ssh'ing into the VM in the browser (see screenshot), which actually worked nicely for a while. But eventually I wanted to use VScode.



https://console.cloud.google.com/compute/instances?onCreate=true&authuser=1&project=rclone-446420

Free trial status: $300.00 credit and 91 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Search (/) for resources, docs, products, and more

Secure your account and protect your data by turning on multi-factor authentication. Add an extra layer of security today.

VM instances

Enter property name or value

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
✓	de-zoomcamp	us-central1-c			10.128.0.3 (nic0)	35.226.250.54 (nic0)	SSH

Related actions

Explore Backup and DR — Back up your VMs and set up disaster recovery

View billing report — View and manage your Compute Engine billing

Monitor VMs — View outlier VMs across metrics like CPU and network

Explore VM logs — View, search, analyze, and download VM instance logs

Set up firewall rules — Control traffic to and from a VM instance

Patch management — Schedule patch updates and view patch compliance on VM instances

Load balance between VMs — Set up Load Balancing for your applications as your traffic and users grow

Instance deleted

GCP Virtual Machine (VM) Size, Slow, Clean Up

If you are progressing through the course and find that your VM is starting to become slow you can run the following commands to inspect and detect areas where you can improve this.

NB: What size VM should I start with? I started with 30GB but this wasn't enough, I had to restart the project with a 60GB machine so I'd recommend choosing the 60GB version.

Commands to inspect the health of your VM:

System Resource Usage:

- `top` or `htop`: Shows real-time information about system resource usage, including CPU, memory, and processes.
- `free -h`: Displays information about system memory usage and availability.
- `df -h`: Shows disk space usage of file systems.
- `du -h <directory>`: Displays disk usage of a specific directory.

Running Processes:

- `ps aux`: Lists all running processes along with detailed information.

Network:

- `ifconfig` or `ip addr show`: Shows network interface configuration.
- `netstat -tuln`: Displays active network connections and listening ports.

Hardware Information:

- `lscpu`: Displays CPU information.
- `lsblk`: Lists block devices (disks and partitions).
- `lshw`: Lists hardware configuration.

User and Permissions:

- `who`: Shows who is logged on and their activities.
- `w`: Displays information about currently logged-in users and their processes.

Package Management:

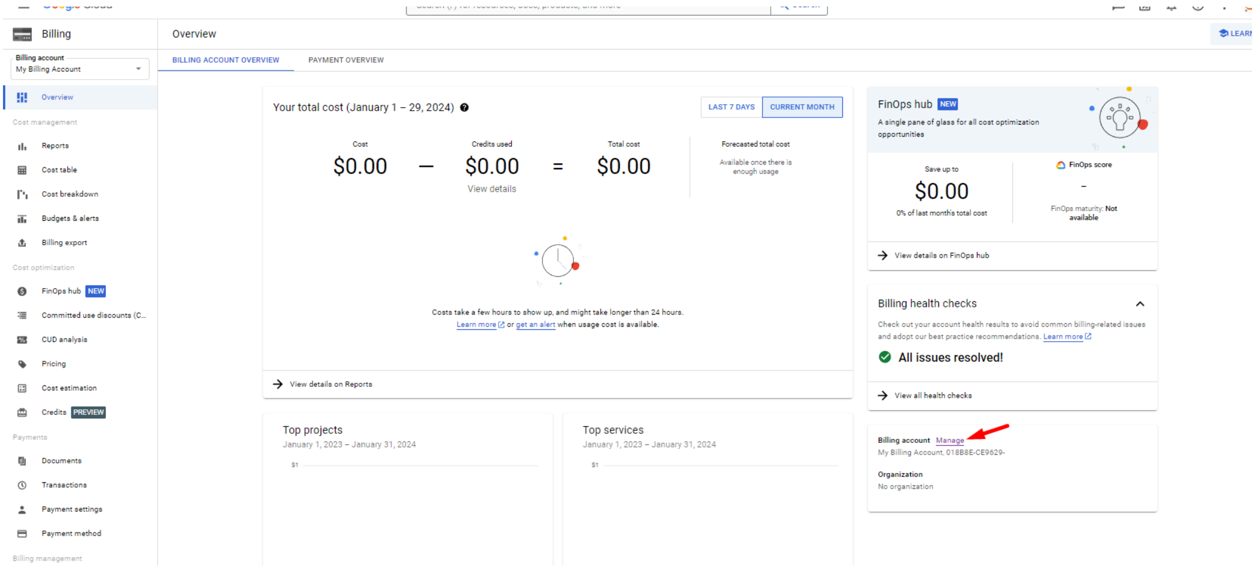
- `apt list --installed`: Lists installed packages (for Ubuntu and Debian-based systems)

Billing account has not been enabled for this project. But you've done it indeed!

if you've got the error

```
Error: Error updating Dataset
"projects/<your-project-id>/datasets/demo_dataset": googleapi: Error
403: Billing has not been enabled for this project. Enable billing at
https://console.cloud.google.com/billing. The default table expiration
time must be less than 60 days, billingNotEnabled
```

but you’ve set your billing account indeed, then try to disable billing for the project and enable it again. It worked for ME!



Your total cost (January 1 – 29, 2024)

Cost		Credits used		Total cost	Forecasted total cost
$0.00	—	$0.00	=	$0.00	Available once there is enough usage
		View details			

Costs take a few hours to show up, and might take longer than 24 hours. Learn more or get an alert when usage cost is available.

FinOps hub NEW
A single pane of glass for all cost optimization opportunities

Save up to $0.00
0% of last month's total cost

FinOps score -
FinOps maturity: Not available

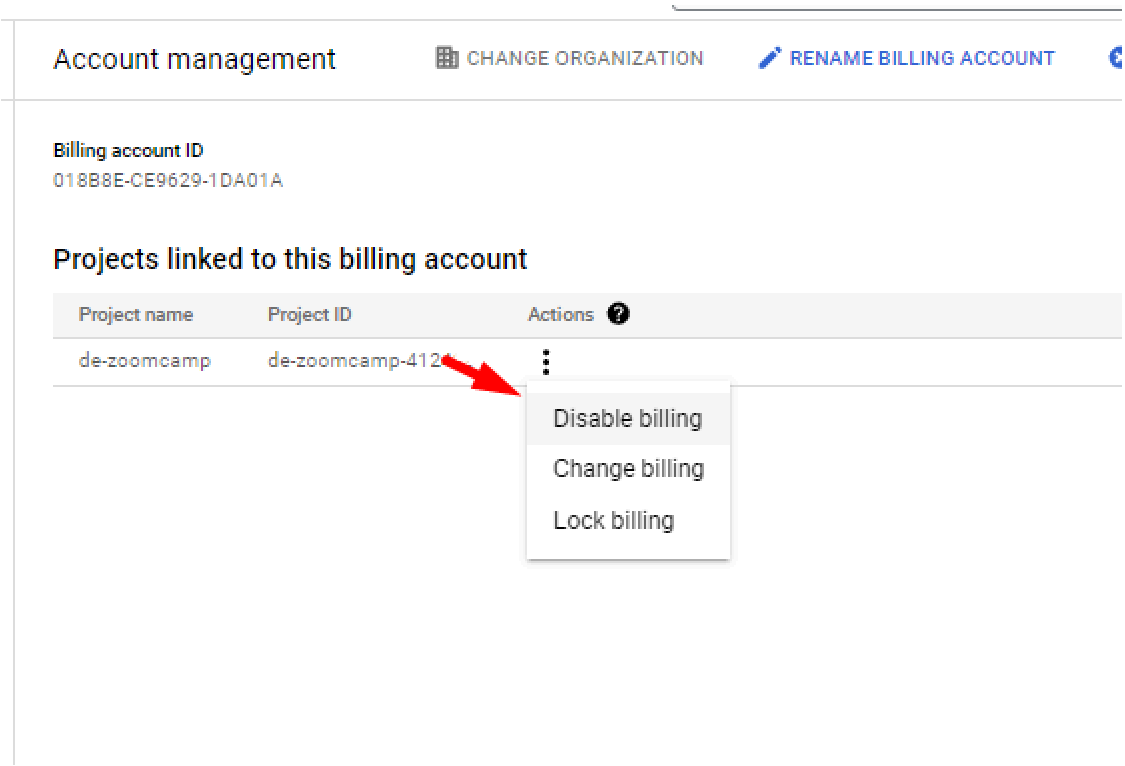
Billing health checks
Check out your account health results to avoid common billing-related issues and adopt our best practice recommendations. Learn more
All issues resolved!

Top projects
January 1, 2023 – January 31, 2024
$1

Top services
January 1, 2023 – January 31, 2024
$1

Billing account Manage
My Billing Account, 018B8E-CE9629-

Organization
No organization



Account management

Billing account ID
018B8E-CE9629-1DA01A

Projects linked to this billing account

Project name	Project ID	Actions
de-zoomcamp	de-zoomcamp-412	⋮

GCP - Windows Google Cloud SDK install issue:gcp

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”,`
- after 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

GCP VM - I cannot get my Virtual Machine to start because GCP has no resources.

1. Click on your VM
2. Create an image of your VM
3. On the page of the image, tell GCP to create a new VM instance via the image
4. On the settings page, change the location

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

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

And the 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 directly commit, even if you are the owner of the repo.

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.

ssh

```
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)Go to edit your VM.

1. Go to section Automation

2. Add Startup script
   ```
   #!/bin/bash
   sudo ufw allow ssh
   ```

4. Stop and Start VM.

GCP VM - Port forwarding from GCP without using VS Code

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.

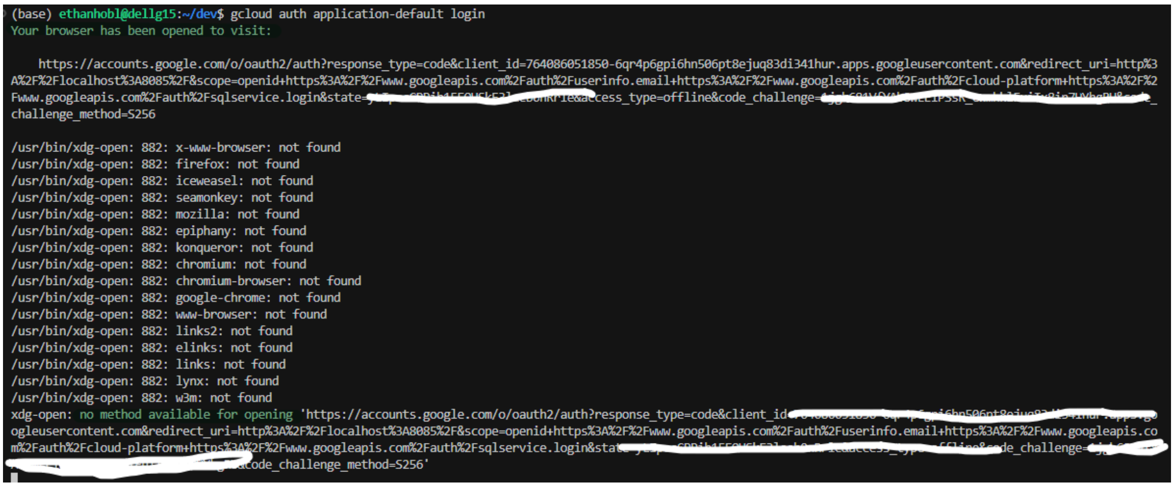
4. Now you can access pgAdmin on local machine in browser typing `localhost:8080`

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.

6. To forward both pgAdmin and postgres use, ssh -i ~/.ssh/gcp -L 5432:localhost:5432 -L 8080:localhost:8080 modito@35.197.218.128

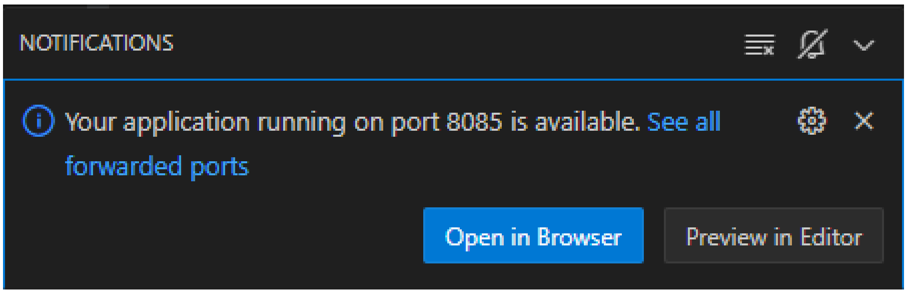
GCP gcloud + MS VS Code - gcloud auth hangs

- If you are using MS VS Code and running gcloud in WSL2, when you first try to login to gcp via the gcloud cli `gcloud auth application-default login`, you will see a message like this, and nothing will happen



```
(base) ethanhobl@dellg15:~/dev$ gcloud auth application-default login
Your browser has been opened to visit:

    https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=764086051850-6qr4p6gpi6hn506pt8ejuq83di341hur.apps.googleusercontent.com&redirect_uri=http%3
A%2F%2Flocalhost%3A8085%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fcloud-platform+https%3A%2F%2
Fwww.googleapis.com%2Fauth%2Fsqlservice.login&state=[redacted]&access_type=offline&code_challenge=[redacted]
challenge_method=S256

/usr/bin/xdg-open: 882: x-www-browser: not found
/usr/bin/xdg-open: 882: firefox: not found
/usr/bin/xdg-open: 882: iceweasel: not found
/usr/bin/xdg-open: 882: seamonkey: not found
/usr/bin/xdg-open: 882: mozilla: not found
/usr/bin/xdg-open: 882: epiphany: not found
/usr/bin/xdg-open: 882: konqueror: not found
/usr/bin/xdg-open: 882: chromium: not found
/usr/bin/xdg-open: 882: chromium-browser: not found
/usr/bin/xdg-open: 882: google-chrome: not found
/usr/bin/xdg-open: 882: www-browser: not found
/usr/bin/xdg-open: 882: links2: not found
/usr/bin/xdg-open: 882: elinks: not found
/usr/bin/xdg-open: 882: links: not found
/usr/bin/xdg-open: 882: lynx: not found
/usr/bin/xdg-open: 882: w3m: not found
xdg-open: no method available for opening 'https://accounts.google.com/o/oauth2/auth?response_type=code&client_id[redacted]
ogleusercontent.com&redirect_uri=http%3A%2F%2Flocalhost%3A8085%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.co
m%2Fauth%2Fcloud-platform+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login&stat[redacted]de_challenge=[redacted]
[redacted]code_challenge_method=S256'
```

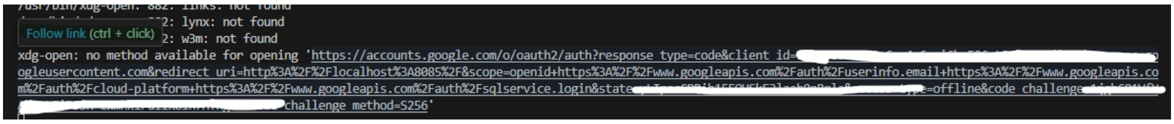
- And there might be a prompt to ask if you want to open it via browser, if you click on it, it will open up a page with error message



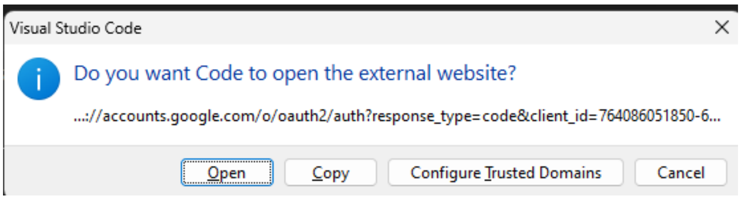
NOTIFICATIONS

Your application running on port 8085 is available. See all forwarded ports

- Solution : you should instead hover on the long link, and ctrl + click the long link



```
/usr/bin/xdg-open: 882: links: not found
Follow link (ctrl + click) 2: lynx: not found
                           2: w3m: not found
xdg-open: no method available for opening 'https://accounts.google.com/o/oauth2/auth?response_type=code&client_id=[redacted]o
ogleusercontent.com&redirect_uri=http%3A%2F%2Flocalhost%3A8085%2F&scope=openid+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%2Fwww.googleapis.co
m%2Fauth%2Fcloud-platform+https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fsqlservice.login&state[redacted]e=offline&code_challenge[redacted]
[redacted]challenge_method=S256'
```

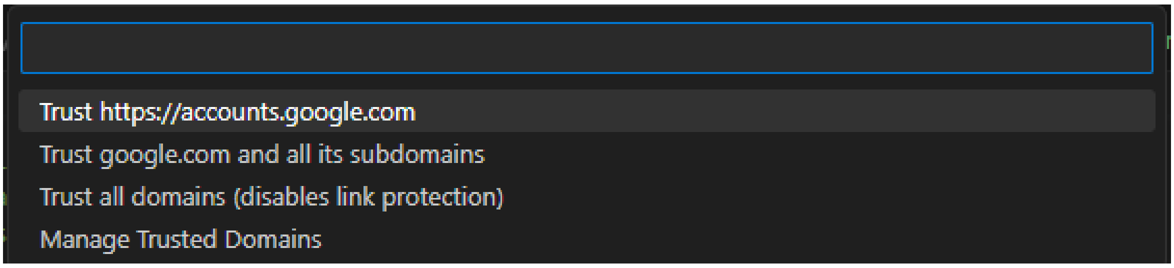


Visual Studio Code

Do you want Code to open the external website?

...://accounts.google.com/o/oauth2/auth?response_type=code&client_id=764086051850-6...

Click configure Trusted Domains here



Trust https://accounts.google.com
Trust google.com and all its subdomains
Trust all domains (disables link protection)
Manage Trusted Domains

Popup will appear, pick first or second entry

- Next time you gcloud auth, the login page should popup via default browser without issues

Terraform - Error: Failed to query available provider packages │ Could not retrieve the list of available versions for provider hashicorp/google: could not query │ provider registry for registry.terrafogorm.io/hashicorp/google: the request failed after 2 attempts, │ please try again later

It is an internet connectivity error, terraform is somehow not able to access the online registry. Check your VPN/Firewall settings (or just clear cookies or restart your network). Try terraform init again after this, it should work.

Terraform - Error:Post "https://storage.googleapis.com/storage/v1/b?alt=json&prettyPrint=false&project=coherent-ascent-379901": oauth2: cannot fetch token: Post "https://oauth2.googleapis.com/token": dial tcp 172.217.163.42:443: i/o timeout

The issue was with the network. Google is not accessible in my country, I am using a VPN. And The terminal program does not automatically follow the system proxy and requires separate proxy configuration settings.I opened a Enhanced Mode in Clash, which is a VPN app, and 'terraform apply' works! So if you encounter the same issue, you can ask help for your vpn provider.

Terraform - Install for WSL

https://techcommunity.microsoft.com/t5/azure-developer-community-blog/configuring-terraform-on-windows-10-linux-sub-system/ba-p/393845

Terraform - Error acquiring the state lock

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

Terraform - Error 400 Bad Request. Invalid JWT Token on WSL.

When running

```
terraform apply
```

on wsl2 I've got this error:

```
│ Error: Post
"https://storage.googleapis.com/storage/v1/b?alt=json&prettyPrint=false&project=<your-project-id>": oauth2: cannot fetch token: 400 Bad Request

│ Response: {"error":"invalid_grant","error_description":"Invalid JWT: Token must be a short-lived token (60 minutes) and in a reasonable timeframe. Check your iat and exp values in the JWT claim."}
```

It happens because there may be time desync on your machine which affects computing JWT

To fix this, run the command

```
sudo hwclock -s
```

which fixes your system time.

Terraform - 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.

Terraform - 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

Terraform - 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.g

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 that contains terraform configuration files, and then run the command.

Terraform - Error creating Dataset: googleapi: Error 403: Request had insufficient authentication scopes

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 $?
```

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
```

stoTerraform - Error creating Bucket: googleapi: Error 403: Permission denied to access ‘storage.buckets.create’

The error:

```
Error: googleapi: Error 403: terraform-trans-campus@trans-campus-410115.iam.gserviceaccount.com does not have storage.buckets.create access to the Google Cloud
```

```
project. Permission 'storage.buckets.create' denied on resource
(or it may not exist)., forbidden
```

The solution:

You have to declare the project name as your *Project ID*, and not your *Project name,* available on GCP console Dashboard.

Terraform google provider requires credentials.

To ensure the sensitivity of the credentials file, I had to spend lot of time to input that as a file.

```
provider "google" {
  project     = var.projectId
  credentials = file("${var.gcpkey}")
  #region      = var.region
  zone = var.zone
}
```

Terraform Teardown of BigQuery Dataset

When running `terraform destroy`, the following error can occur:

```
Do you really want to destroy all resources?
  Terraform will destroy all your managed infrastructure, as shown
above.
  There is no undo. Only 'yes' will be accepted to confirm.
```

~~~
  Enter a value: yes


google_bigquery_dataset.homework_dataset: Destroying...
[id=projects/terraform-demo-449214/datasets/homework_dataset]

╷

│ Error: Error when reading or editing Dataset: googleapi: Error
400: Dataset terraform-demo-449214:homework_dataset is still in
use, resourceInUse
```
~~~

This is because the dataset is still in use by a table. To delete the dataset, we need to set the `delete_contents_on_destroy` property to `true` in the `main.tf` file.

SQL - 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’;
  ```

SQL - 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()

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

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

SSH Error: ssh: Could not resolve hostname linux: Name or service not known

To resolve this, ensure that your config file is in C/User/Username/.ssh/config

'pip' is not recognized as an internal or external command, operable program or batch file.

If you use Anaconda (recommended for the course), it comes with pip, so the issues is probably that the anaconda’s Python is not on the PATH.

Adding it to the PATH is different for each operation system.

For Linux and MacOS:

1. Open a terminal.
2. Find the path to your Anaconda installation. This is typically `~/anaconda3` or `~/opt/anaconda3`.
3. Add Anaconda to your PATH with the command: `export PATH="/path/to/anaconda3/bin:$PATH"`.
4. To make this change permanent, add the command to your `.bashrc` (Linux) or `.bash_profile` (MacOS) file.

On Windows, python and pip are in different locations (python is in the anaconda root, and pip is in Scripts). With GitBash:

1. Locate your Anaconda installation. The default path is usually `C:\Users\[YourUsername]\Anaconda3`.

2. Determine the correct path format for Git Bash. Paths in Git Bash follow the Unix-style, so convert the Windows path to a Unix-style path. For example, `C:\Users\[YourUsername]\Anaconda3` becomes `/c/Users/[YourUsername]/Anaconda3`.
3. Add Anaconda to your PATH with the command: `export PATH="/c/Users/[YourUsername]/Anaconda3/:/c/Users/[YourUsername]/Anaconda3/Scripts/$PATH"`.
4. To make this change permanent, add the command to your `.bashrc` file in your home directory.
5. Refresh your environment with the command: `source ~/.bashrc`.

For Windows (without Git Bash):

1. Right-click on 'This PC' or 'My Computer' and select 'Properties'.
2. Click on 'Advanced system settings'.
3. In the System Properties window, click on 'Environment Variables'.
4. In the Environment Variables window, select the 'Path' variable in the 'System variables' section and click 'Edit'.
5. In the Edit Environment Variable window, click 'New' and add the path to your Anaconda installation (typically `C:\Users\[YourUsername]\Anaconda3` and C:\Users\[YourUsername]\Anaconda3\Scripts`).
6. Click 'OK' in all windows to apply the changes.

After adding Anaconda to the PATH, you should be able to use `pip` from the command line. Remember to restart your terminal (or command prompt in Windows) to apply these changes.

Error: error starting userland proxy: listen tcp4 0.0.0.0:8080: bind: address already in use

Resolution: You need to stop the services which is using the port.

Run the following:

```
sudo kill -9 `sudo lsof -t -i:<port>`
```

<port> being 8080 in this case. This will free up the port for use.

~ Abhijit Chakraborty

Error: error response from daemon: cannot stop container: 1afaf8f7d52277318b71eef8f7a7f238c777045e769dd832426219d6c4b8dfb4: permission denied

Resolution: In my case, I had to stop docker and restart the service to get it running properly

Use the following command:

```
sudo systemctl restart docker.socket docker.service
```

~ Abhijit Chakraborty

Error: docker build Error checking context: 'can't stat '<path-to-file>'

Resolution: This happens due to insufficient permission for docker to access a certain file within the directory which hosts the Dockerfile.

1. You can create a .dockerignore file and add the directory/file which you want Dockerfile to ignore while build.

2. If the above does not work, then put the dockerfile and corresponding script, `  1.py` in our case to a subfolder. and run `docker build ...`

from inside the new folder.

~ Abhijit Chakraborty

Docker-Compose - it is illegal to have any blank spaces between the environment argument in docker-compose.yml

The following ways of configuring it will not work:

- PGADMIN_DEFAULT_EMAIL = admin@admin.com
- PGADMIN_DEFAULT_PASSWORD = root

- PGADMIN_DEFAULT_EMAIL=admin@admin.com
- PGADMIN_DEFAULT_PASSWORD=root

Anaconda to PIP

To get a pip-friendly requirements.txt file file from Anaconda use

`conda install pip` then `pip list –format=freeze > requirements.txt`.

`conda list -d > requirements.txt` will not work and `pip freeze > requirements.txt` may give odd pathing.

Jupyter - Install, open Jupyter and convert Jupyter notebook to Python script

Install and open Jupyter Notebook

```
pip install jupyter
```

```
python3 -m notebook
```

Notebook convert

```
pip install nbconvert --upgrade
```

```
Python3 -m jupyter nbconvert --to=script upload-data.ipynb
```

Alternative way to convert Jupyter notebook to Python script (via jupytext)

If you keep getting errors with nbconvert after: jupyter nbconvert --to script <your_notebook.ipynb>

you could try to convert your Jupyter notebook via another tool called jupytext

Jupytext is another excellent tool for converting Jupyter Notebooks to Python scripts, which works very similar to nbconvert

Install jupytext

```
pip install jupytext
```

Convert your Notebook to a Python script

```
jupytext --to py <your_notebook.ipynb>
```

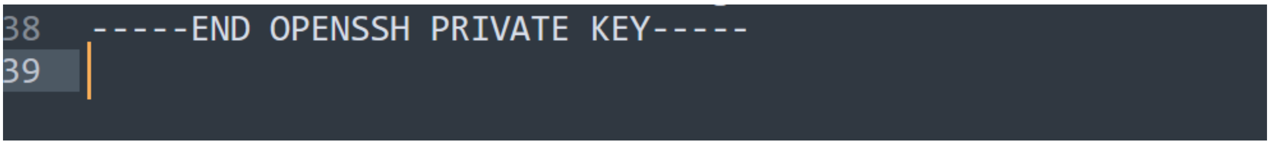
SSH error in VS Code - “Could not establish connection to "de-zoomcamp": Permission denied (publickey).”

If you are using Windows, try copying the .ssh folder from the Linux file path to Windows. In the config file, use

IdentityFile C:\Users\<username>\.ssh\gcp

Instead of IdentityFile ~/.ssh/gcp

Another reason: The private key in its file at the local path C:\Users\<username>\.ssh\gcp needs an extra line in the end:



```
-----END OPENSSH PRIVATE KEY-----

```

Module 2: Workflow Orchestration

Where are the FAQ questions from the previous cohorts for the orchestration module?

[Prefect](#) [Airflow](#) [Mage](#)

How do I launch Kestra?

Start docker in linux with docker run --pull=always --rm -it -p 8080:8080 --user=root \

```
  -v /var/run/docker.sock:/var/run/docker.sock \
  -v /tmp:/tmp kestra/kestra:latest server local
```

Once run you can login to dashboard at localhost:8080
For windows instructions see the Kestra github here https://github.com/kestra-io/kestra

Here sample docker-compose for kestra

```
services:
  kestra:
    build: .
    image: kestra/kestra:latest
    container_name: kestra
    user: "0:0"
    environment:
      DOCKER_HOST: tcp://host.docker.internal:2375  # for Windows
      KESTRA_CONFIGURATION: |
        kestra:
          repository:
            type: h2
          queue:
            type: memory
          storage:
            type: local
            local:
              basePath: /app/storage
          tasks:
            tmp-dir:
              path: /app/tmp
          plugins:
            repositories:
              - id: central
                type: maven
                url: https://repo.maven.apache.org/maven2
            definitions:
              - io.kestra.plugin.core:core:latest
              - io.kestra.plugin.scripts:python:1.3.4
              - io.kestra.plugin.http:http:latest
      KESTRA_TASKS_TMP_DIR_PATH: /app/tmp
    ports:
      - "8080:8080"
    volumes:
      - //var/run/docker.sock:/var/run/docker.sock  # Windows path
      - /yourpath/.dbt:/app/.dbt
      - /yourpath/kestra/plugins:/app/plugins
      - /yourpath/kestra/workflows:/app/workflows
      - /yourpath/kestra/storage:/app/storage
      - /yourpath//kestra/tmp:/app/tmp
      - /yourpath//dbt_prj:/app/workflows/dbt_project
      - /yourpath//my-creds.json:/app/.dbt/my-creds.json
    command: server standalone
```

docker: Error response from daemon: mkdir C:\Program Files\Git\var: Access is denied.

Description:

Running the command below in Bash with Docker running and WSL2 installed. Even running Bash as admin won’t work

```:

$ docker run --pull=always --rm -it -p 8080:8080 --user=root -v

/var/run/docker.sock:/var/run/docker.sock -v /tmp:/tmp kestra/kestra:latest server local

latest: Pulling from kestra/kestra

Digest:
sha256:af02a309ccbb52c23ad1f1551a1a6db8cf0523cf7aac7c7eb878d7925bc85a62

Status: Image is up to date for kestra/kestra:latest

docker: Error response from daemon: mkdir C:\\Program Files\\Git\\var: Access is denied.

See 'docker run --help'.
```

The error mentioned above will appear and localhost wont shows the Kestra UI, the solution is to run Command Prompt as admin with the following command:

```

docker run --pull=always --rm -it -p 8080:8080 --user=root ^

   -v "/var/run/docker.sock:/var/run/docker.sock" ^

   -v "C:/Temp:/tmp" kestra/kestra:latest server local
```

This works flawlessly and localhost shows Kestra UI as usual.

Error when running Kestra flow connecting to postgres.

Error: org.postgresql.util.psqlexception the connection attempt failed due to this config on kestra flow -> jdbc:postgresql://host.docker.internal:5432/postgres-zoomcamp

Solution: Just replace host.docker.internal for the name of the service for postgres in docker compose.

—---

I also encountered a similar error as above, slightly different error message:

```
org.postgresql.util.PSQLException: The connection attempt failed. 2025-01-29
22:52:22.281 green_create_table The connection attempt failed. host.docker.internal
```

I could download my dataset by executing my flow, but when i wanted to ingest it to the pg database, the connection to pg failed.

The main issue was that the pg database url is different for linux than the url in the tutorial. Namely, instead of host.docker.internal, linux users will use the service or container name for postgres, which for me was just postgres.

```
url: jdbc:postgresql://postgres:5432/kestra
```

Voila. Also, make sure to double check your pg database name. Mine was kestra in the docker compose file, whereas in the tutorial they had named it postgres-zoomcamp.

Adding a pgadmin service with volume mounting to the docker-compose:

I encountered an error where the localhost url for pgadmin would just hang up (i chose localhost:8080 for my pgadmin, and made kestra localhost:8090, personal preference).

The associated error was:

```
| [2025-01-30 02:38:49 +0000] [91] [INFO] Worker exiting (pid: 91)
2_kestra-pgadmin-1 | ERROR : Failed to create the directory
/var/lib/pgadmin/sessions: 2_kestra-pgadmin-1 | [Errno 13] Permission denied:
'/var/lib/pgadmin/sessions' 2_kestra-pgadmin-1 | HINT : Create the directory
/var/lib/pgadmin/sessions, ensure it is writeable by 2_kestra-pgadmin-1 |
'pgadmin', and try again, or, create a config_local.py file
2_kestra-pgadmin-1 | and override the SESSION_DB_PATH setting per
2_kestra-pgadmin-1 |
https://www.pgadmin.org/docs/pgadmin4/8.14/config_py.html 2_kestra-pgadmin-1
| [2025-01-30 02:38:50 +0000] [1] [ERROR] Worker (pid:91) exited with code 1
2_kestra-pgadmin-1 | [2025-01-30 02:38:50 +0000] [1] [ERROR] Worker (pid:91)
exited with code 1. 2_kestra-pgadmin-1 | [2025-01-30 02:38:50 +0000] [92]
[INFO] Booting worker with pid: 92
```

And the resolution involved changing the ownership of my local directory to the user “5050” which is pgadmin. Unlike postgres, pgadmin requires you to give it permission. Apparently the postgres user inside the docker container creates the postgres volume/dir, so it has permission`s already. This is a good source: https://stackoverflow.com/questions/64781245/permission-denied-var-lib-pgadmin-sessions-in-dockerG

Running out of storage when using kestra with postgres on GCP VM

Running out of storage while trying to backfill. I realized my GCP VM only has 30GB of storage and I was eating it up! Couple things I did/would suggest:

1. Clean up your GCP VM drive. You can use this command to see what is taking up the most space: $ sudo du -sh *
2. (~1gb) For me, Anaconda installer was taking up lots of space - you can delete that immediately because I already installed anaconda. I don't need the installer anymore.
   Rm -rf <anacondainstaller_fpath>
3. (~3gb) Anaconda also takes up lots of space. You can't delete it all if you want to run python, but you can clean it up significantly. I don't care much about libs, etc. because I can build them in a docker container! Command is $ conda clean --all -y
4. You can clean up your kestra files with a purge flow. Here is the generic one: https://kestra.io/docs/administrator-guide/purge
   a. I personally wanted to do it immediately, not at end of month, so I made end date just now and got rid of the trigger block. You can also specify if you want to removed FAILED state executions, but I chose not to: `endDate: "{{ now() }}"`
5. You can clean up your pg database by manually deleting tables in pgadmin. Or possibly set up a workflow for it in kestra, but it was easy enough to manually delete.

How can Kestra access service account credential?

Do not directly add the content of service account credential json in Kestra script, especially if we are pushing to Github. Follow the instruction to add the service account as a secret Configure Google Service Account.

When we need to use it in Kestra, we can pull it through `{{ secret('GCP_SERVICE_ACCOUNT') }}`

In the pluginDefaults.

Storage Bucket Permission Denied Error when running the gcp_setup flow

When following the youtube lesson and then running the gcp_setup flow, I get the following error:

```
2025-02-03 08:12:17.991create_gcs_bucket2i78v7qCFw9Q7rKzR424iM
zoomcamp@kestra-sandbox-449806.iam.gserviceaccount.com does not have
storage.buckets.get access to the Google Cloud Storage bucket.
Permission 'storage.buckets.get' denied on resource (or it may not
exist).
2025-02-03 08:12:17.991create_gcs_bucket2i78v7qCFw9Q7rKzR424iM
403 Forbidden
```

```
 GET
https://storage.googleapis.com/storage/v1/b/kestra-de-zoomcamp-bucket?
projection=full
 {
 "code" : 403,
 "errors" : [ {
 "domain" : "global",
 "message" : "zoomcamp@kestra-sandbox-449806.iam.gserviceaccount.com
does not have storage.buckets.get access to the Google Cloud Storage
bucket. Permission 'storage.buckets.get' denied on resource (or it may
not exist).",
 "reason" : "forbidden"
 } ],
 "message" : "zoomcamp@kestra-sandbox-449806.iam.gserviceaccount.com
does not have storage.buckets.get access to the Google Cloud Storage
bucket. Permission 'storage.buckets.get' denied on resource (or it may
not exist)."
 }
```

I tried manually creating the bucket in the GCP console, but this showed me that the bucket already existed. So I came up with another name for the bucket and it worked.

The GCP bucket name has to be unique globally across all buckets, even if those are not your buckets, because the bucket will be accessible by URL.

Invalid dataset ID Error Error when running the gcp_setup flow

When following the youtube lesson and then running the gcp_setup flow, it works until the create_bq_dataset task, where I got the following error:

```
2025-02-03 08:44:12.162Invalid dataset ID "de-zoomcamp". Dataset IDs
must be alphanumeric (plus underscores) and must be at most 1024
characters long.
2025-02-03 08:44:12.162400 Bad Request
 POST
https://bigquery.googleapis.com/bigquery/v2/projects/kestra-sandbox-44
9806/datasets?prettyPrint=false
 {
```

```
 "code": 400,
 "errors": [
 {
 "domain": "global",
 "message": "Invalid dataset ID \"de-zoomcamp\". Dataset IDs must be
alphanumeric (plus underscores) and must be at most 1024 characters
long.",
 "reason": "invalid"
 }
 ],
 "message": "Invalid dataset ID \"de-zoomcamp\". Dataset IDs must be
alphanumeric (plus underscores) and must be at most 1024 characters
long.",
 "status": "INVALID_ARGUMENT"
```

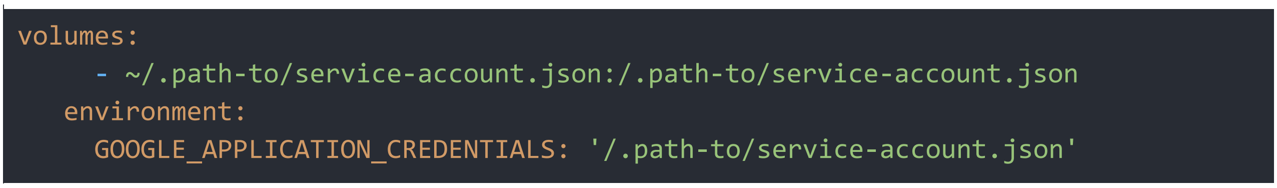
While not very apparent from the error message, we are not suppose to use a dash in the dataset name, so I changed the dataset name to “de_zoomcamp” and it worked.

How do I properly authenticate a Google Cloud Service Account in Kestra?

Several authentication methods are available;
These are some of the most straightforward approaches.

Method 1:

Update your `docker-compose.yml` file as follows:



```
volumes:
     - ~/.path-to/service-account.json:/.path-to/service-account.json
   environment:
     GOOGLE_APPLICATION_CREDENTIALS: '/.path-to/service-account.json'
```

Method 2:

Step 1: Store the Service Account as a Secret

a. Run this command, specifying the **correct path** to your `service-account.json` file and `.env_encoded`:

```
echo SECRET_GCP_SERVICE_ACCOUNT=$(cat /path/to/service-account.json |
base64 -w 0) >> /path/to/.env_encoded
```

b. Modify `docker-compose.yml` to include the encoded secrets:

```
kestra:
  env_file: /path/to/.env_encoded
```

Step 2: Configure Kestra Plugin Defaults

This ensures all GCP tasks use the secret automatically:

```
pluginDefaults:
  - type: io.kestra.plugin.gcp
    values:
      serviceAccount: "{{ secret('GCP_SERVICE_ACCOUNT') }}"
```

Step 3: Verify it’s working in a testing GCP workflow

```
namespace: testing-credentials

tasks:
  - id: create_gcs_bucket
    type: io.kestra.plugin.gcp.gcs.CreateBucket
    ifExists: SKIP
    storage class: REGIONAL
    name: “testing-cred-bucket” # "{{ kv('GCP_BUCKET_NAME') }}"
```

Additional - QA

Question: How do I update the Service Account key?

Answer: Generate a new key, re-run the Base64 command, and restart Kestra.

Question: Why use secrets instead of embedding the JSON key in the task?

Answer: Secrets prevent credential exposure and make workflows easier to manage.

Question: Can I apply this method to other GCP tasks?

Answer: Yes, all GCP plugins will automatically inherit the secret.

Should I include my .env_encoded file in my .gitignore?

⚠️ Yes, you should definitely include the .env_encoded file in your .gitignore file. Here's why:

- Security: The .env_encoded file contains sensitive information, namely the base64 encoded version of your GCP Service Account key. Even though it's encoded, it's not secure to share this in a public repository as anyone can decode it back to the original JSON.
- Best Practices: It's a common practice to not commit environment files or any files containing secrets to version control systems like Git. This prevents accidental exposure of sensitive data.

⚠️ ***How to do it:***

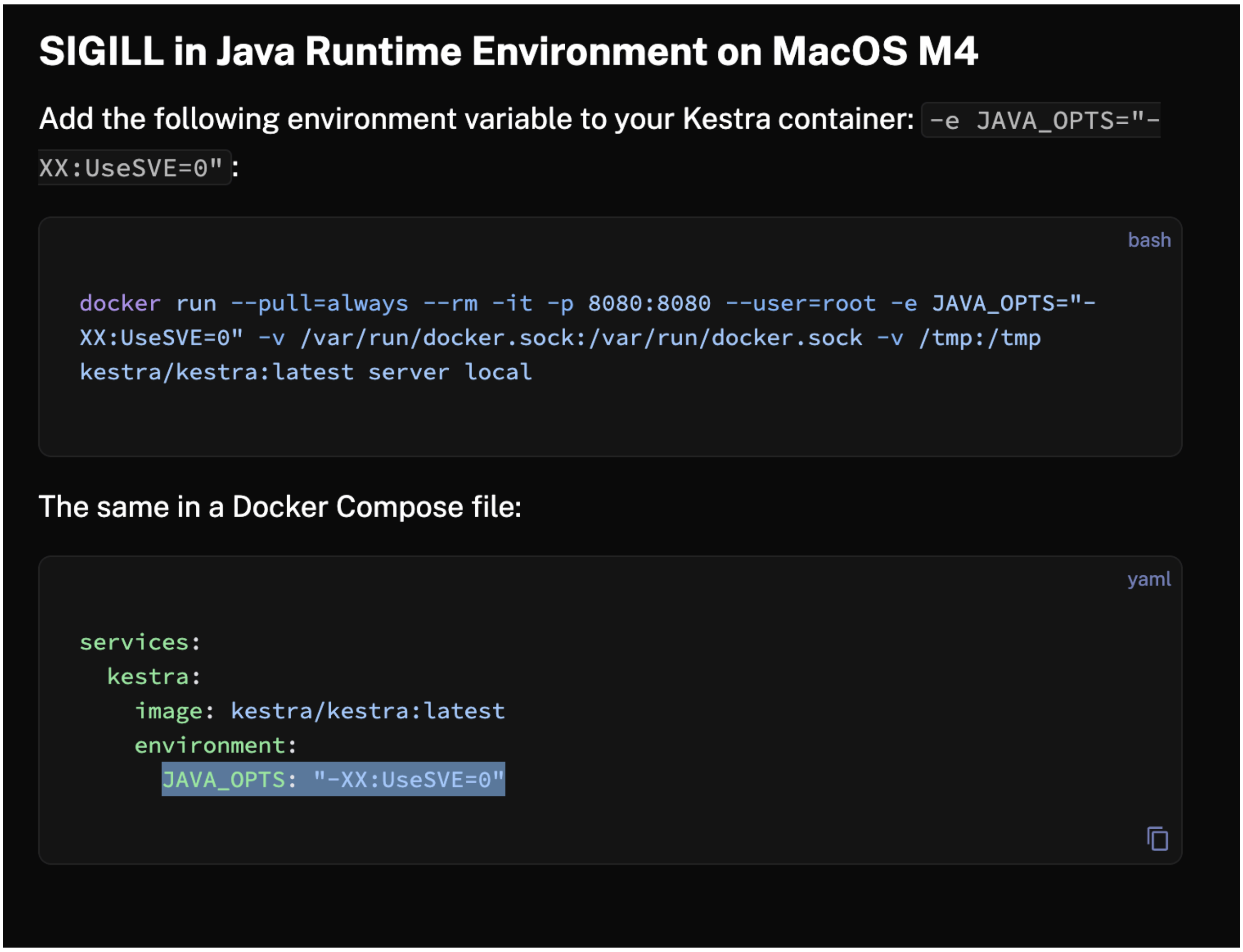
```
# Add this line to your .gitignore file
.env_encoded
```

⚠️ ***More on Security:***

Base64 encoding is easily reversible. Base64 is an encoding scheme, not an encryption method. It's designed to encode binary data into ASCII characters that can be safely transmitted over systems that are designed to deal with text. Here's why it's not secure for protecting sensitive information:

- **Reversibility:** Base64 encoding simply translates binary data into a text string using a specific set of 64 characters. Decoding it back to the original data is straightforward and doesn't require any secret key or password.
- **Public Availability of Tools:** Numerous online tools, software libraries, and command-line utilities exist that can decode base64 with just a few clicks or commands.
- **No Security:** Since base64 encoding does not change or hide the actual content of the data, anyone with access to the encoded string can decode it back to the original data.

Question: Getting SIGILL in JRE when running latest kestra image on Mac M4 MacOS 15.2/3



SIGILL in Java Runtime Environment on MacOS M4

Add the following environment variable to your Kestra container: `-e JAVA_OPTS="-XX:UseSVE=0"`:

```bash
docker run --pull=always --rm -it -p 8080:8080 --user=root -e JAVA_OPTS="-XX:UseSVE=0" -v /var/run/docker.sock:/var/run/docker.sock -v /tmp:/tmp kestra/kestra:latest server local
```

The same in a Docker Compose file:

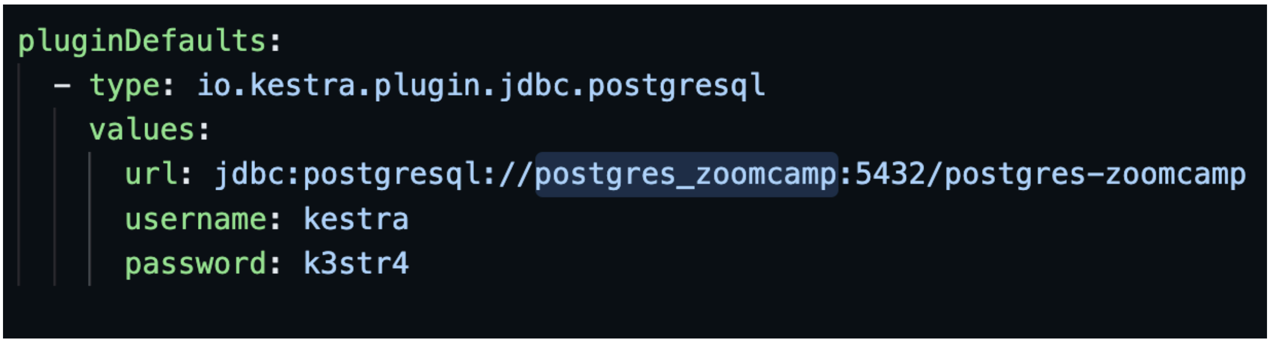
```yaml
services:
  kestra:
    image: kestra/kestra:latest
    environment:
      JAVA_OPTS: "-XX:UseSVE=0"
```

taskid: yellow_create_table The connection attempt failed. Host.docker.internal

If you're using Linux, you might encounter Connection Refused errors when connecting to the Postgres DB from within Kestra. This is because host.docker.internal works differently on Linux.

Using the modified Docker Compose file in 02-workflow-orchestration readme troubleshooting tips **Docker Compose Example**, you can run both Kestra and its dedicated Postgres DB, as well as the Postgres DB for the exercises all together. You can access it within Kestra by referring to the container name postgres_zoomcamp instead of host.docker.internal in pluginDefaults.

The pluginDefaults exist in both 2_postgres_taxi_scheduled.yaml, 02_postgres_taxi.yaml, please modify as shown below.



```yaml
pluginDefaults:
  - type: io.kestra.plugin.jdbc.postgresql
    values:
      url: jdbc:postgresql://postgres_zoomcamp:5432/postgres-zoomcamp
      username: kestra
      password: k3str4
```

Fix: Add extra_hosts for host.docker.internal on Linux

This update corrects the Docker Compose configuration to resolve the error when using the alias `host.docker.internal` on Linux systems. Since this alias does not resolve natively on Linux, the following entry was added to the affected container:

```yaml
kestra:

   image: kestra/kestra:latest

   pull_policy: always

   user: "root"

   command: server standalone

   volumes:...

   environment:...

   ports:...

   depends_on:...

   extra_hosts:

     - "host.docker.internal:host-gateway"
```

```yaml
extra_hosts:

  - "host.docker.internal:host-gateway"
```

With this change, containers that need to access host services via `host.docker.internal` will be able to do so correctly. For inter-container communication within the same network, it is recommended to use the service name directly.

Fix: Add extra_hosts for taskRunner in the dbt-build

Adds the extraHosts configuration to the taskRunner in the dbt-build task to resolve the issue with host.docker.internal not being recognized on Linux.

```
    taskRunner:
      type: io.kestra.plugin.scripts.runner.docker.Docker
      extraHosts:
          - "host.docker.internal:host-gateway"
```

Kestra: Don’t forget to set GCP_CREDS variable

If you plan on using Kestra with Google Cloud Platform, make sure you setup the GCP_CREDS that’s gonna be used in the flows that has “gcp” on its name.

To set it, go to Namespaces, and then select “zoomcamp” if you are using the same examples used in the lessons. Then in the “KV Store” tab create the new key as GCP_CREDS and set the type to JSON and paste the content of the .json file with credentials for the service account created.

Kestra: Backfill showing getting executed but not getting results or showing up in executions:

It seems to be a bug. Current fix is to remove the timezone from triggers in the script. More on this bug is here.

Module 3: Data Warehousing

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

GCS Bucket - error when writing data from web to GCS:

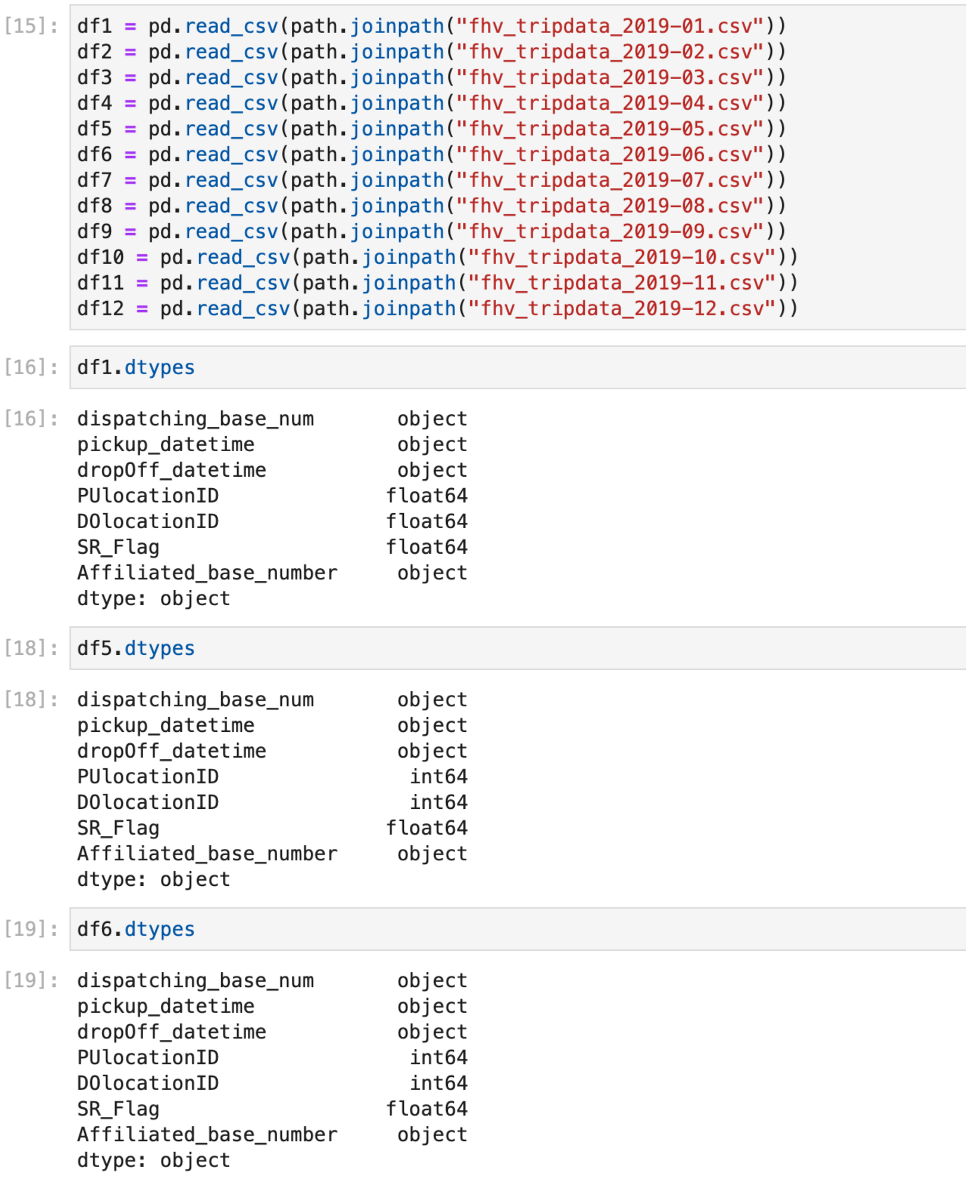
Make sure to use **Nullable** dataTypes, such as **Int64** when appliable.

GCS Bucket - te table: Error while reading data, error message: Parquet column 'XYZ' has type INT which does not match the target cpp_type DOUBLE. File: gs://path/to/some/blob.parquet

Ultimately, when trying to ingest data into a BigQuery table, all files within a given directory must have the same schema.

When dealing for example with the FHV Datasets from 2019, however (see image below), one can see that the files for '2019-05', and 2019-06, have the columns "PUlocationID" and "DOlocationID" as Integers, while for the period of '2019-01' through '2019-04', the same column is defined as FLOAT.parquet

So while importing these files as parquet to BigQuery, the first one will be used to define the schema of the table, while all files following that will be used to append data on the existing table. Which means, they must all follow the very same schema of the file that created the table.



```
[15]: df1 = pd.read_csv(path.joinpath("fhv_tripdata_2019-01.csv"))
      df2 = pd.read_csv(path.joinpath("fhv_tripdata_2019-02.csv"))
      df3 = pd.read_csv(path.joinpath("fhv_tripdata_2019-03.csv"))
      df4 = pd.read_csv(path.joinpath("fhv_tripdata_2019-04.csv"))
      df5 = pd.read_csv(path.joinpath("fhv_tripdata_2019-05.csv"))
      df6 = pd.read_csv(path.joinpath("fhv_tripdata_2019-06.csv"))
      df7 = pd.read_csv(path.joinpath("fhv_tripdata_2019-07.csv"))
      df8 = pd.read_csv(path.joinpath("fhv_tripdata_2019-08.csv"))
      df9 = pd.read_csv(path.joinpath("fhv_tripdata_2019-09.csv"))
      df10 = pd.read_csv(path.joinpath("fhv_tripdata_2019-10.csv"))
      df11 = pd.read_csv(path.joinpath("fhv_tripdata_2019-11.csv"))
      df12 = pd.read_csv(path.joinpath("fhv_tripdata_2019-12.csv"))

[16]: df1.dtypes

[16]: dispatching_base_num       object
      pickup_datetime            object
      dropOff_datetime           object
      PUlocationID              float64
      DOlocationID              float64
      SR_Flag                   float64
      Affiliated_base_number     object
      dtype: object

[18]: df5.dtypes

[18]: dispatching_base_num       object
      pickup_datetime            object
      dropOff_datetime           object
      PUlocationID                int64
      DOlocationID                int64
      SR_Flag                   float64
      Affiliated_base_number     object
      dtype: object

[19]: df6.dtypes

[19]: dispatching_base_num       object
      pickup_datetime            object
      dropOff_datetime           object
      PUlocationID                int64
      DOlocationID                int64
      SR_Flag                   float64
      Affiliated_base_number     object
      dtype: object
```

So, in order to prevent errors like that, make sure to enforce the data types for the columns on the DataFrame before you serialize/upload them to BigQuery. Like this:

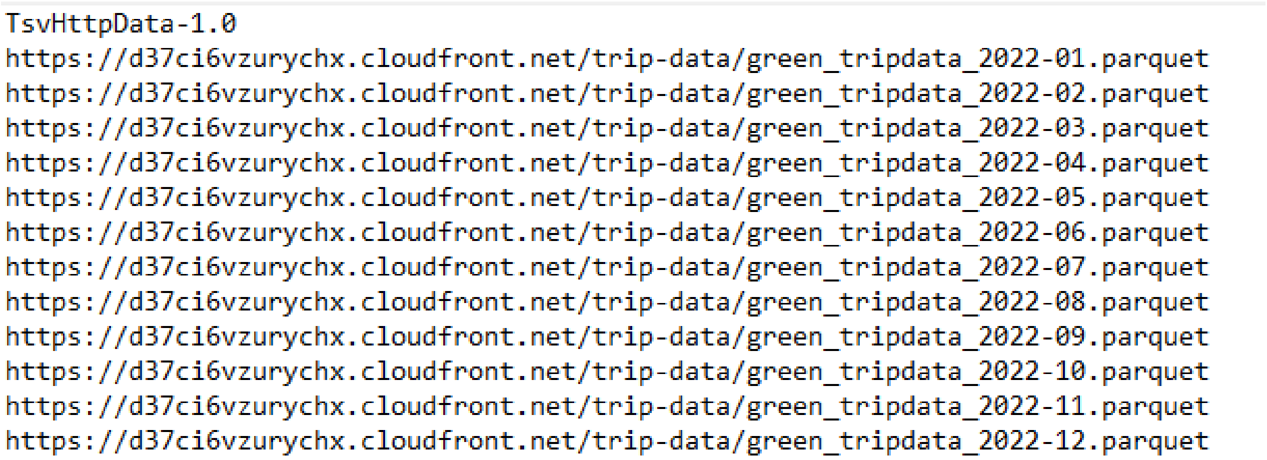
```
pd.read_csv("path_or_url").astype({
     "col1_name": "datatype",
     "col2_name": "datatype",
     ...
     "colN_name": "datatype"
})
```

GCS Bucket - Fix Error when importing FHV data to GCS

If you receive the error gzip.BadGzipFile: Not a gzipped file (b'\n\n'), this is because you have specified the wrong URL to the FHV dataset. Make sure to use

https://github.com/DataTalksClub/nyc-tlc-data/releases/download/fhv/{dataset_file}.csv.gz
Emphasising the ‘/releases/download’ part of the URL.

GCS Bucket - Load Data From URL list in to GCP Bucket



```
TsvHttpData-1.0
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-01.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-02.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-03.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-04.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-05.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-06.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-07.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-08.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-09.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-10.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-11.parquet
https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2022-12.parquet
```

Krishna Anand

GCS Bucket - I query my dataset and get a Bad character (ASCII 0) error?

- Check the Schema
- You might have a wrong formatting
- Try to upload the CSV.GZ files without formatting or going through pandas via wget
- See this Slack conversation for helpful tips

GCP BQ - “bq: command not found”

Run the following command to check if “BigQuery Command Line Tool” is installed or not:
`gcloud components list`

You can also use `bq.cmd` instead of `bq` to make it work.

GCP BQ - Caution in using bigquery:no

Use big queries carefully,

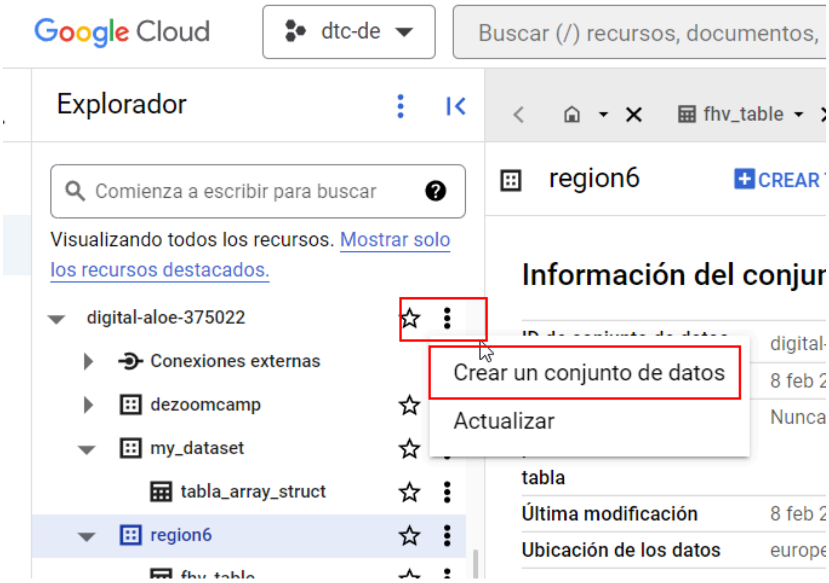
I created by bigquery dataset on an account where my free trial was exhausted, and got a bill of $80.

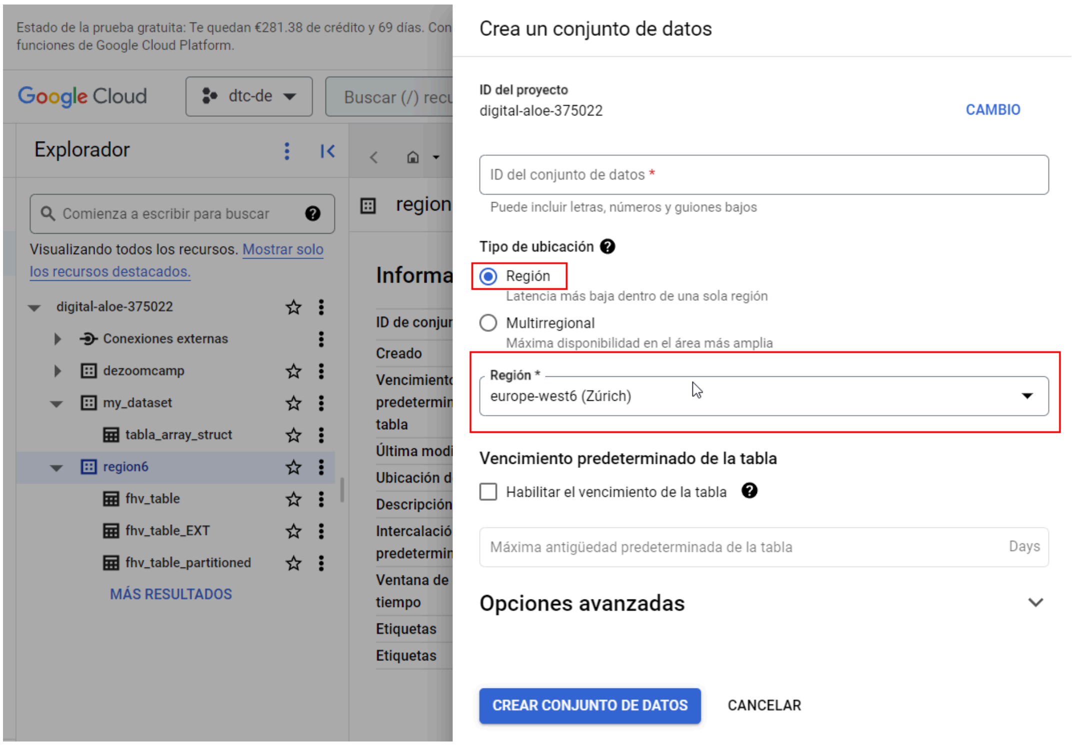
Use big query in free credits and destroy all the datasets after creation.

Check your Billing daily! Especially if you’ve spinned up a VM.

GCP BQ - Cannot read and write in different locations: source: EU, destination: US - Loading data from GCS into BigQuery (different Region):

Be careful when you create your resources on GCP, all of them have to share the same Region in order to allow load data from GCS Bucket to BigQuery. If you forgot it when you created them, you can create a new dataset on BigQuery using the same Region which you used on your GCS Bucket.





Estado de la prueba gratuita: Te quedan €281.38 de crédito y 69 días. Con
funciones de Google Cloud Platform.

Visualizando todos los recursos. Mostrar solo los recursos destacados.

- digital-aloe-375022
  - Conexiones externas
  - dezoomcamp
  - my_dataset
    - tabla_array_struct
  - region6
    - fhv_table
    - fhv_table_EXT
    - fhv_table_partitioned
  - MÁS RESULTADOS

Crea un conjunto de datos

ID del proyecto
digital-aloe-375022 CAMBIO

ID del conjunto de datos *
Puede incluir letras, números y guiones bajos

Tipo de ubicación

● Región
Latencia más baja dentro de una sola región

○ Multirregional
Máxima disponibilidad en el área más amplia

Región *
europe-west6 (Zúrich)

Vencimiento predeterminado de la tabla

□ Habilitar el vencimiento de la tabla

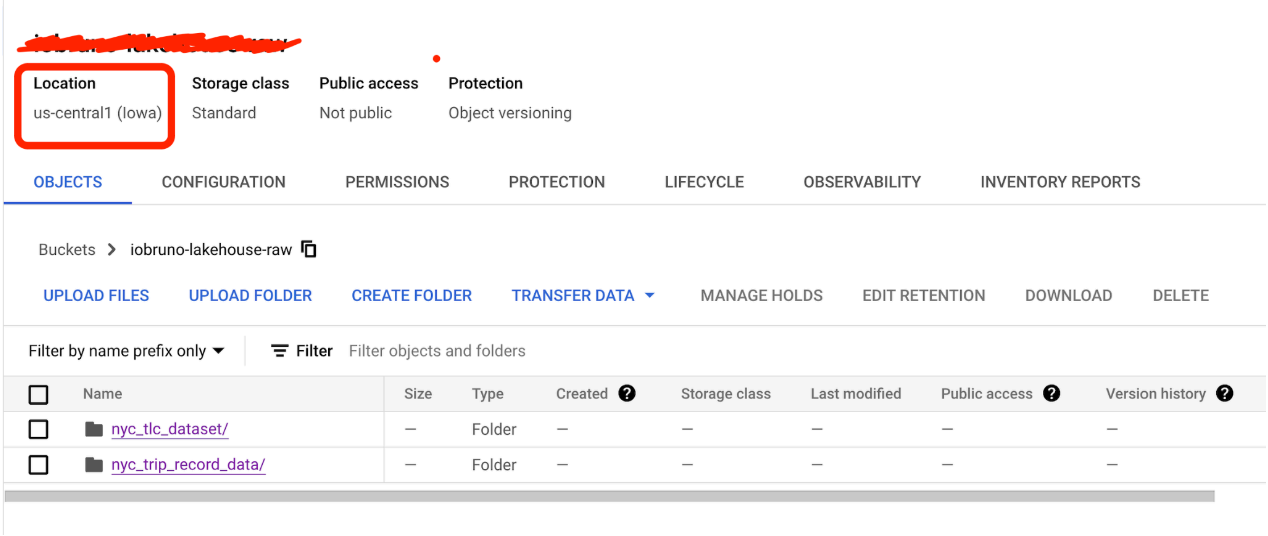
Máxima antigüedad predeterminada de la tabla Days

Opciones avanzadas

This means that your GCS Bucket and the BigQuery dataset are placed in different regions. You have to create a new dataset inside BigQuery in the same region with your GCS bucket and store the data in the newly created dataset.

GCP BQ - Cannot read and write in different locations: source: <REGION_HERE>, destination: <ANOTHER_REGION_HERE>

Make sure to create the BigQuery dataset in the very same location that you've created the GCS Bucket. For instance, **if your GCS Bucket was created in** `us-central1`**, then BigQuery dataset must be created in the same region** (us-central1, in this example)

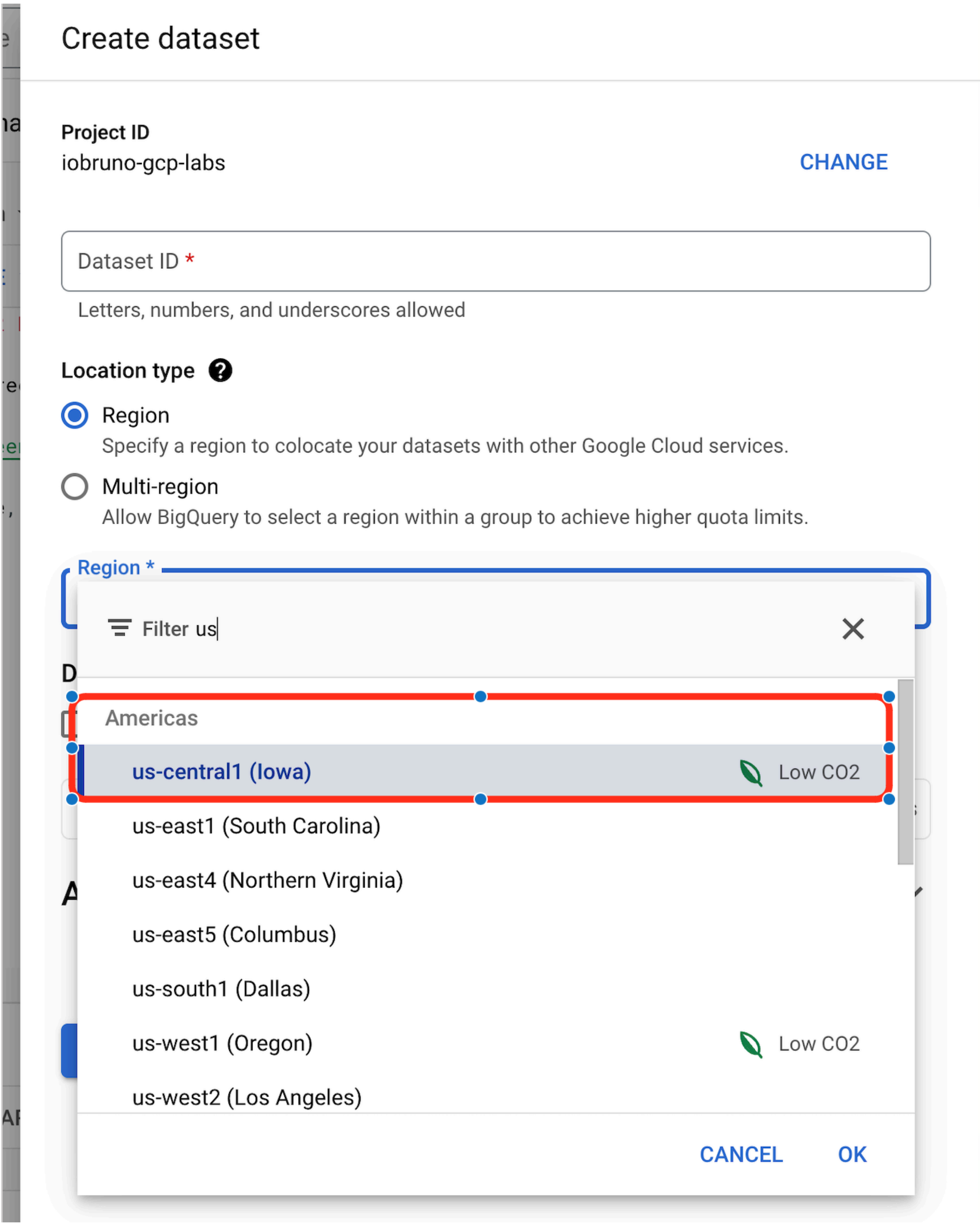


[redacted]

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	Object versioning

Buckets > iobruno-lakehouse-raw

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history
□	nyc_tlc_dataset/	—	Folder	—	—	—	—	—
□	nyc_trip_record_data/	—	Folder	—	—	—	—	—



Create dataset

Project ID
iobruno-gcp-labs

CHANGE

Dataset ID *

Letters, numbers, and underscores allowed

Location type

● Region
Specify a region to colocate your datasets with other Google Cloud services.

○ Multi-region
Allow BigQuery to select a region within a group to achieve higher quota limits.

Region *

Filter us

Americas

us-central1 (Iowa) Low CO2

us-east1 (South Carolina)

us-east4 (Northern Virginia)

us-east5 (Columbus)

us-south1 (Dallas)

us-west1 (Oregon) Low CO2

us-west2 (Los Angeles)

CANCEL OK

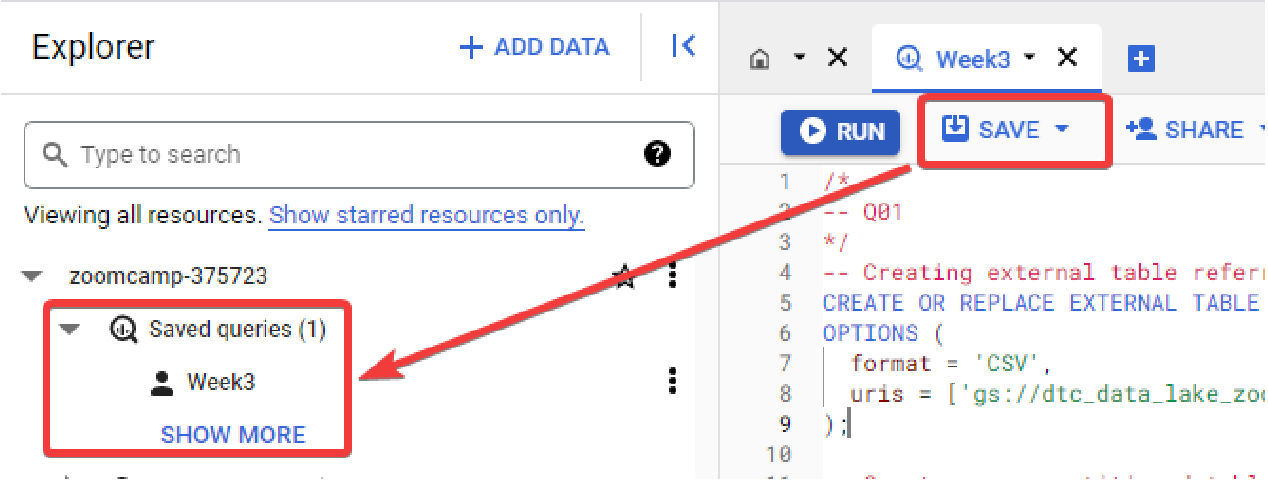
GCP BQ - Remember to save your queries

By the way, this isn’t a problem/solution, but a useful hint:

- Please, remember to save your progress in BigQuery SQL Editor.

- I was almost finishing the homework, when my Chrome Tab froze and I had to reload it. Then I lost my entire SQL script.

- Save your script from time to time. Just click on the button at the top bar. Your saved file will be available on the left panel.



```
/*
-- Q01
*/
-- Creating external table refer
CREATE OR REPLACE EXTERNAL TABLE
OPTIONS (
  format = 'CSV',
  uris = ['gs://dtc_data_lake_zo
);
```

Alternatively, you can copy paste your queries into an .sql file in your preferred editor (Notepad++, VS Code, etc.). Using the .sql extension will provide convenient color formatting.

GCP BQ - Can I use BigQuery for real-time analytics in this project?

Ans : While real-time analytics might not be explicitly mentioned, BigQuery has real-time data streaming capabilities, allowing for potential integration in future project iterations.

GCP BQ - Unable to load data from external tables into a materialized table in BigQuery due to an invalid timestamp error that are added while appending data to the file in Google Cloud Storage

```
could not parse 'pickup_datetime' as timestamp for field
pickup_datetime (position 2)
```

This error is caused by invalid data in the timestamp column. A way to identify the problem is to define the schema from the external table using string datatype. This enables the queries to work at which point we can filter out the invalid rows from the import to the materialised table and insert the fields with the timestamp data type.

GCP BQ - Error Message in BigQuery: annotated as a valid Timestamp, please annotate it as TimestampType(MICROS) or TimestampType(MILLIS)

Background:

- `pd.read_parquet`
- `pd.to_datetime`
- `pq.write_to_dataset`

Reference:

- https://stackoverflow.com/questions/48314880/are-parquet-file-created-with-pyarrow-vs-pyspark-compatible
- https://stackoverflow.com/questions/57798479/editing-parquet-files-with-python-causes-errors-to-datetime-format
- https://www.reddit.com/r/bigquery/comments/16aoq0u/parquet_timestamp_to_bq_coming_across_as_int/?share_id=YXqCs5Jl6hQcw-kg6-VgF&utm_content=1&utm_medium=ios_app&utm_name=ioscss&utm_source=share&utm_term=1

Solution:

Add `use_deprecated_int96_timestamps=True` to `pq.write_to_dataset` function, like below

```
pq.write_to_dataset(
        table,
        root_path=root_path,
        filesystem=gcs,
        use_deprecated_int96_timestamps=True
# Write timestamps to INT96 Parquet format
)
```

GCP BQ - Datetime columns in Parquet files created from Pandas show up as integer columns in BigQuery

Solution:

If you’re using Mage, in the last Data Exporter that writes to Google Cloud Storage use PyArrow to generate the Parquet file with the correct logical type for the datetime columns, otherwise they won't be converted to timestamp when loaded by BigQuery later on.

```
import pyarrow as pa
import pyarrow.parquet as pq
import os

if 'data_exporter' not in globals():
    from mage_ai.data_preparation.decorators import data_exporter

# Replace with the location of your service account key JSON file.
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] =
'/home/src/personal-gcp.json'

bucket_name = "<YOUR_BUCKET_NAME>"
object_key = 'nyc_taxi_data_2022.parquet'
where = f'{bucket_name}/{object_key}'

@data_exporter
def export_data(data, *args, **kwargs):
    table = pa.Table.from_pandas(data, preserve_index=False)
    gcs = pa.fs.GcsFileSystem()

    pq.write_table(
        table,
        where,

        # Convert integer columns in Epoch milliseconds
        # to Timestamp columns in microseconds ('us') so
        # they can be loaded into BigQuery with the right
        # data type
        coerce_timestamps='us',

        filesystem=gcs
    )
```

Solution 2:

If you're using Mage, in the last Data Exporter that writes to Google Cloud Storage, provide PyArrow with explicit schema to generate the Parquet file with the correct logical type for the datetime columns, otherwise they won't be converted to timestamp when loaded by BigQuery later on.

```
schema = pa.schema([
    ('vendor_id', pa.int64()),
    ('lpep_pickup_datetime', pa.timestamp('ns')),
    ('lpep_dropoff_datetime', pa.timestamp('ns')),
    ('store_and_fwd_flag', pa.string()),
    ('ratecode_id', pa.int64()),
    ('pu_location_id', pa.int64()),
    ('do_location_id', pa.int64()),
    ('passenger_count', pa.int64()),
    ('trip_distance', pa.float64()),
    ('fare_amount', pa.float64()),
    ('extra', pa.float64()),
    ('mta_tax', pa.float64()),
    ('tip_amount', pa.float64()),
    ('tolls_amount', pa.float64()),
    ('improvement_surcharge', pa.float64()),
    ('total_amount', pa.float64()),
    ('payment_type', pa.int64()),
    ('trip_type', pa.int64()),
    ('congestion_surcharge', pa.float64()),
    ('lpep_pickup_month', pa.int64())
])

table = pa.Table.from_pandas(data, schema=schema)
```

GCP BQ - Create External Table using Python

Reference:

https://cloud.google.com/bigquery/docs/external-data-cloud-storage

Solution:

```
from google.cloud import bigquery

    # Set table_id to the ID of the table to create
    table_id = f"{project_id}.{dataset_name}.{table_name}"

    # Construct a BigQuery client object
    client = bigquery.Client()

    # Set the external source format of your table
    external_source_format = "PARQUET"

    # Set the source_uris to point to your data in Google Cloud
    source_uris = [ f'gs://{bucket_name}/{object_key}/*']

    # Create ExternalConfig object with external source format
    external_config =
bigquery.ExternalConfig(external_source_format)
```

```
    # Set source_uris that point to your data in Google Cloud
    external_config.source_uris = source_uris
    external_config.autodetect = True

    table = bigquery.Table(table_id)
    # Set the external data configuration of the table
    table.external_data_configuration = external_config

    table = client.create_table(table)  # Make an API request.

    print(f'Created table with external source: {table_id}')
    print(f'Format:
{table.external_data_configuration.source_format}')
```

GCP BQ - Check BigQuery Table Exist And Delete

Reference:

https://stackoverflow.com/questions/60941726/can-bigquery-api-overwrite-existing-table-view-with-create-table-tables-inser

Solution:

Combine with “Create External Table using Python”, use it before “client.create_table” function.

```
def tableExists(tableID, client):
    """
    Check if a table already exists using the tableID.
    return : (Boolean)
    """
    try:
        table = client.get_table(tableID)
        return True
    except Exception as e: # NotFound:
        return False
```

GCP BQ - Error: Missing close double quote (") character

To avoid this error you can upload data from Google Cloud Storage to BigQuery through BigQuery Cloud Shell using the command:

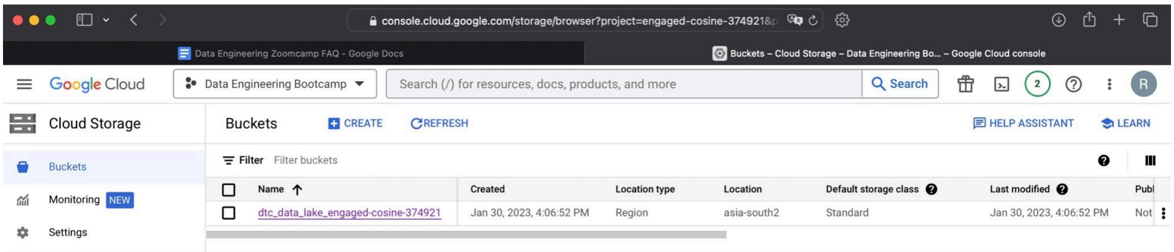
```
$ bq load  --autodetect --allow_quoted_newlines --source_format=CSV
dataset_name.table_name
"gs://dtc-data-lake-bucketname/fhv/fhv_tripdata_2019-*.csv.gz"
```

GCP BQ - Cannot read and write in different locations: source: asia-south2, destination: US

Solution: This problem arises if your gcs and bigquery storage is in different regions.

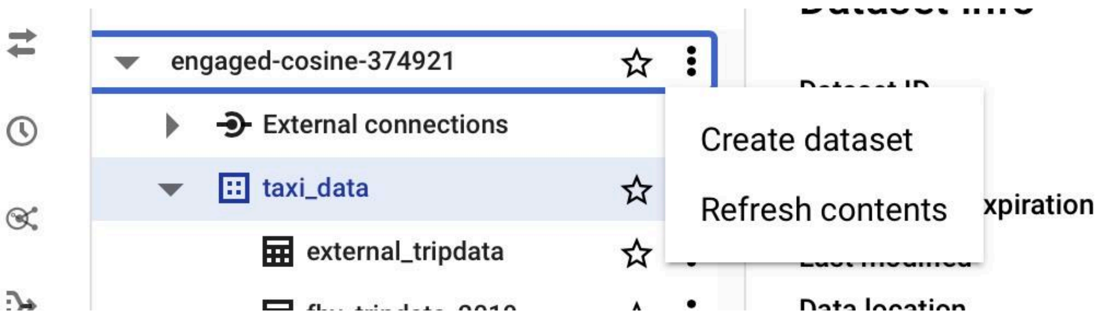
One potential way to solve it:

1. Go to your google cloud bucket and check the region in field named “Location”

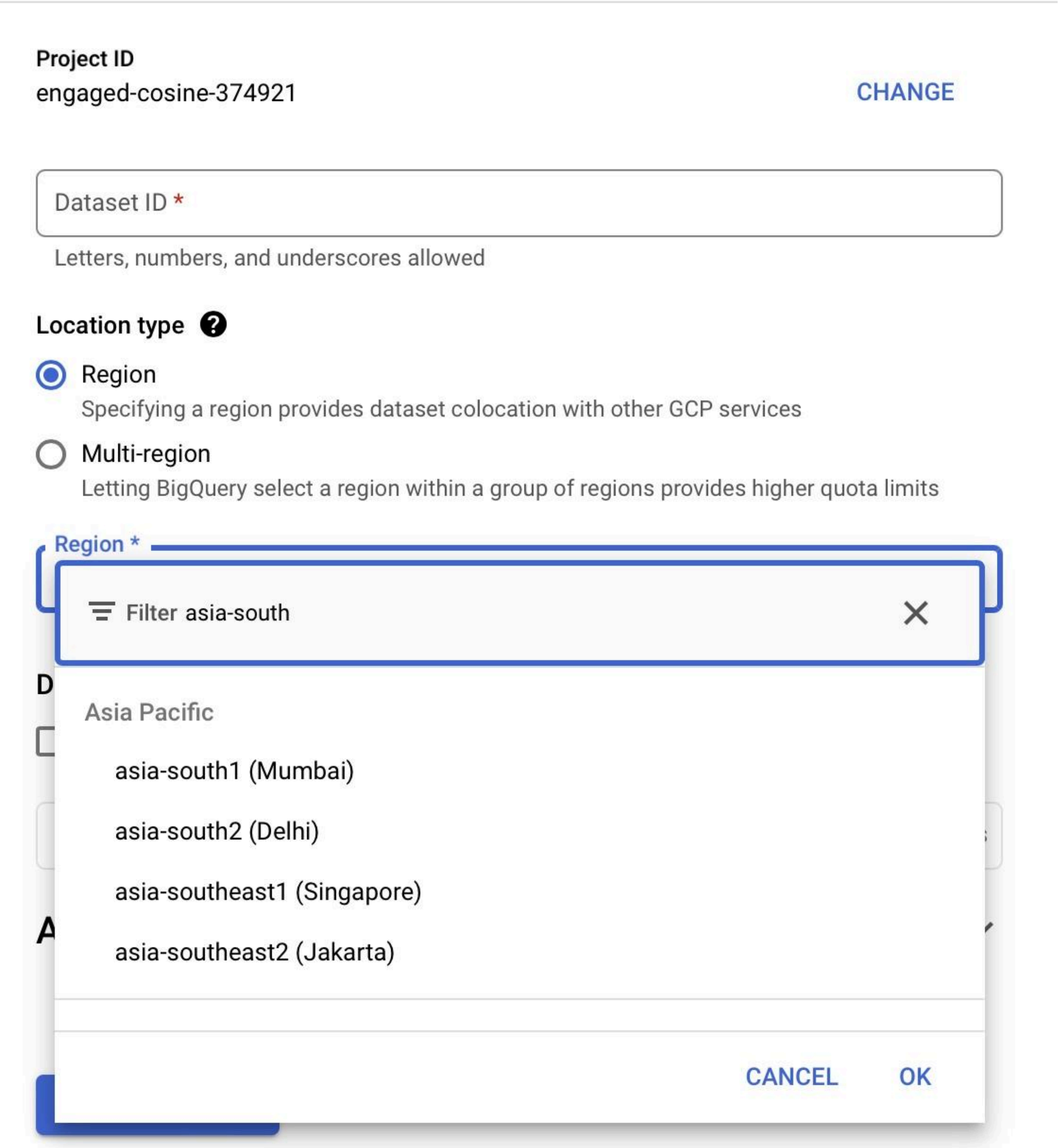

console.cloud.google.com/storage/browser?project=engaged-cosine-374921&p
Data Engineering Zoomcamp FAQ - Google Docs
Buckets – Cloud Storage – Data Engineering Bo... – Google Cloud console
Google Cloud
Data Engineering Bootcamp
Search (/) for resources, docs, products, and more
Search
Cloud Storage
Buckets
CREATE
REFRESH
HELP ASSISTANT
LEARN
Buckets
Monitoring NEW
Settings
Filter Filter buckets

Name	Created	Location type	Location	Default storage class	Last modified	Publ
dtc_data_lake_engaged-cosine-374921	Jan 30, 2023, 4:06:52 PM	Region	asia-south2	Standard	Jan 30, 2023, 4:06:52 PM	Not

2. Now in bigquery, click on three dot icon near your project name and select create dataset.



3. In region filed choose the same regions as you saw in your google cloud bucket



Project ID
engaged-cosine-374921

Dataset ID *
Letters, numbers, and underscores allowed

Location type

● Region
Specifying a region provides dataset colocation with other GCP services

○ Multi-region
Letting BigQuery select a region within a group of regions provides higher quota limits

Region *
Filter asia-south

Asia Pacific
- asia-south1 (Mumbai)
- asia-south2 (Delhi)
- asia-southeast1 (Singapore)
- asia-southeast2 (Jakarta)

GCP BQ - Tip: Using Cloud Function to read csv.gz files from github directly to BigQuery in Google Cloud:

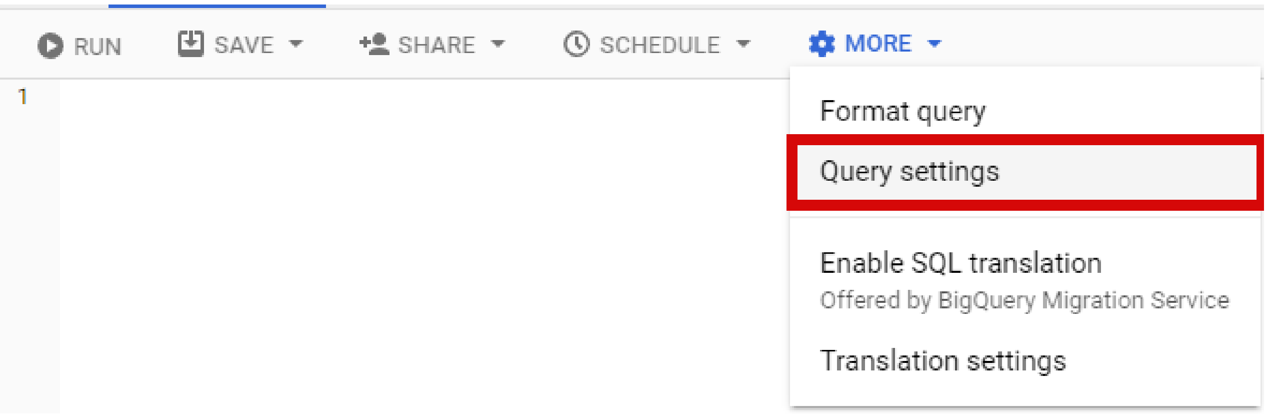
There are multiple benefits of using Cloud Functions to automate tasks in Google Cloud.

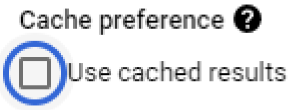
Use below Cloud Function python script to load files directly to BigQuery. Use your project id, dataset id & table id as defined by you.

```
import tempfile
import requests
import logging
from google.cloud import bigquery

def hello_world(request):

    # table_id = <project_id.dataset_id.table_id>
    table_id = 'de-zoomcap-project.dezoomcamp.fhv-2019'

    # Create a new BigQuery client
    client = bigquery.Client()


    for month in range(4, 13):
        # Define the schema for the data in the CSV.gz files
        url =
'https://github.com/DataTalksClub/nyc-tlc-data/releases/download/f
hv/fhv_tripdata_2019-{:02d}.csv.gz'.format(month)

        # Download the CSV.gz file from Github
        response = requests.get(url)

        # Create new table if loading first month data else append
        write_disposition_string = "WRITE_APPEND" if month > 1
else "WRITE_TRUNCATE"

        # Defining LoadJobConfig with schema of table to prevent
it from changing with every table
        job_config = bigquery.LoadJobConfig(
                schema=[
                    bigquery.SchemaField("dispatching_base_num",
"STRING"),
                    bigquery.SchemaField("pickup_datetime",
"TIMESTAMP"),
                    bigquery.SchemaField("dropOff_datetime",
"TIMESTAMP"),
                    bigquery.SchemaField("PUlocationID",
"STRING"),
                    bigquery.SchemaField("DOlocationID",
"STRING"),
                    bigquery.SchemaField("SR_Flag", "STRING"),
                    bigquery.SchemaField("Affiliated_base_number",
"STRING"),
                ],
                    skip_leading_rows=1,
                    write_disposition=write_disposition_string,
                    autodetect=True,
                    source_format="CSV",
                )
        # Load the data into BigQuery
        # Create a temporary file to prevent the exception-
AttributeError: 'bytes' object has no attribute 'tell'"
        with tempfile.NamedTemporaryFile() as f:
            f.write(response.content)
            f.seek(0)
            job = client.load_table_from_file(
                f,
                table_id,
                location="US",
                job_config=job_config,
            )
            job.result()
            logging.info("Data for month %d successfully loaded
into table %s.", month, table_id)
    return 'Data loaded into table {}.'.format(table_id)
```

GCP BQ - When querying two different tables external and materialized you get the same result when count(distinct(*))

You need to uncheck cache preferences in query settings




Cache preference

☐ Use cached results

GCP BQ - How to handle type error from big query and parquet data?

Problem: When you inject data into GCS using Pandas, there is a chance that some dataset has missing values on DOlocationID and PUlocationID. Pandas by default will cast these columns as float data type, causing inconsistent data type between parquet in GCS and schema defined in big query. You will see something like this:

error: Error while reading table: trips_data_all.external_fhv_tripdata, error message: Parquet column 'DOlocationID' has type INT64 which does not match the target cpp_type DOUBLE.

Solution:

- Fix the data type issue in data pipeline
- Before injecting data into GCS, use astype and Int64 (which is different from int64 and accept both missing value and integer exist in the column) to cast the columns.

Something like:

```
    df["PUlocationID"] = df.PUlocationID.astype("Int64")

    df["DOlocationID"] = df.DOlocationID.astype("Int64")

NOTE: It is best to define the data type of all the columns in the
Transformation section of the ETL pipeline before loading to BigQuery
```

GCP BQ - Invalid project ID . Project IDs must contain 6-63 lowercase letters, digits, or dashes. Some project

Problem occurs when misplacing content after fro``m clause in BigQuery SQLs.
Check to remove any extra apaces or any other symbols, keep in lowercases, digits and dashes only

GCP BQ - Does BigQuery support multiple columns partition?

No. Based on the documentation for Bigquery, it does not support more than 1 column to be partitioned.

[source]

GCP BQ - DATE() Error in BigQuery

Error Message:

```
PARTITION BY expression must be DATE(<timestamp_column>),
DATE(<datetime_column>), DATETIME_TRUNC(<datetime_column>,
DAY/HOUR/MONTH/YEAR), a DATE column,
TIMESTAMP_TRUNC(<timestamp_column>, DAY/HOUR/MONTH/YEAR),
DATE_TRUNC(<date_column>, MONTH/YEAR), or
RANGE_BUCKET(<int64_column>, GENERATE_ARRAY(<int64_value>,
<int64_value>[, <int64_value>]))
```

Solution:

Convert the column to datetime first.

```
df["pickup_datetime"] = pd.to_datetime(df["pickup_datetime"])
df["dropOff_datetime"] = pd.to_datetime(df["dropOff_datetime"])
```

GCP BQ - When trying to cluster by DATE(tpep_pickup_datetime) it gives an error: Entries in the CLUSTER BY clause must be column names

No need to convert as you can cluster by a TIMESTAMP column directly in BigQuery. BigQuery supports clustering on TIMESTAMP, DATE, DATETIME, STRING, INT64, and BOOL types.

clustering sorts data based on the timestamp to optimize queries with filters like WHERE tpep_pickup_datetime BETWEEN ..., rather than creating discrete partitions.

If your goal is to improve performance for time-based queries, combining partitioning by DATE(event_time) and clustering by tpep_pickup_datetime is a good approach.

GCP BQ - Native tables vs External tables in BigQuery?

Native tables are tables where the data is stored in BigQuery. External tables store the data outside BigQuery, with BigQuery storing metadata about that external table.

External tables: They are not stored directly in big query tables but pulled in from a data lake such as Google Cloud Storage or S3.

Materialized table: Copy of this external table. Now the data is stored in the bigquery table and consumes the space.

Resources:

- https://cloud.google.com/bigquery/docs/external-tables
- https://cloud.google.com/bigquery/docs/tables-intro

Why does my partitioned table in BigQuery show as non-partitioned even though BigQuery says it's partitioned?

If your partitioned table in BigQuery shows as non-partitioned, it may be due to a delay in updating the table's details in the UI. The table is likely partitioned, but it may not show the updated information immediately.

Here’s what you can do:

1. Refresh your BigQuery UI:
   If you're already inspecting the table in the BigQuery UI, try refreshing the page after a few minutes to ensure the table details are updated correctly.
2. Open a new tab:
   Alternatively, try opening a new tab in BigQuery and inspect the table details again. This can sometimes help to load the most up-to-date information.
3. Be patient:
   In some cases, there might be a slight delay in reflecting changes, but the table is very likely partitioned.

GCP BQ ML - Unable to run command (shown in video) to export ML model from BQ to GCS

Issue: Tried running command to export ML model from BQ to GCS from Week 3

```
bq --project_id taxi-rides-ny extract -m nytaxi.tip_model
gs://taxi_ml_model/tip_model
```

It is failing on following error:

```
BigQuery error in extract operation: Error processing job Not found: Dataset was
not found in location US
```

I verified the BQ data set and gcs bucket are in the same region- us-west1. Not sure how it gets location US. I couldn’t find the solution yet.

Solution: Please enter correct project_id and gcs_bucket folder address. My gcs_bucket folder address is

```
gs://dtc_data_lake_optimum-airfoil-376815/tip_model
```

Dim_zones.sql Dataset was not found in location US When Running fact_trips.sql

To solve this error mention the location = US when creating the dim_zones table

```
{{ config(
    materialized='table',
    location='US'
) }}
```

Just Update this part to solve the issue and run the dim_zones again and then run the fact_trips

GCP BQ ML - Export ML model to make predictions does not work for MacBook with Apple M1 chip (arm architecture).

Solution: proceed with setting up serving_dir on your computer as in the extract_model.md file. Then instead of

```
docker pull tensorflow/serving
```

use

```
docker pull emacski/tensorflow-serving
```

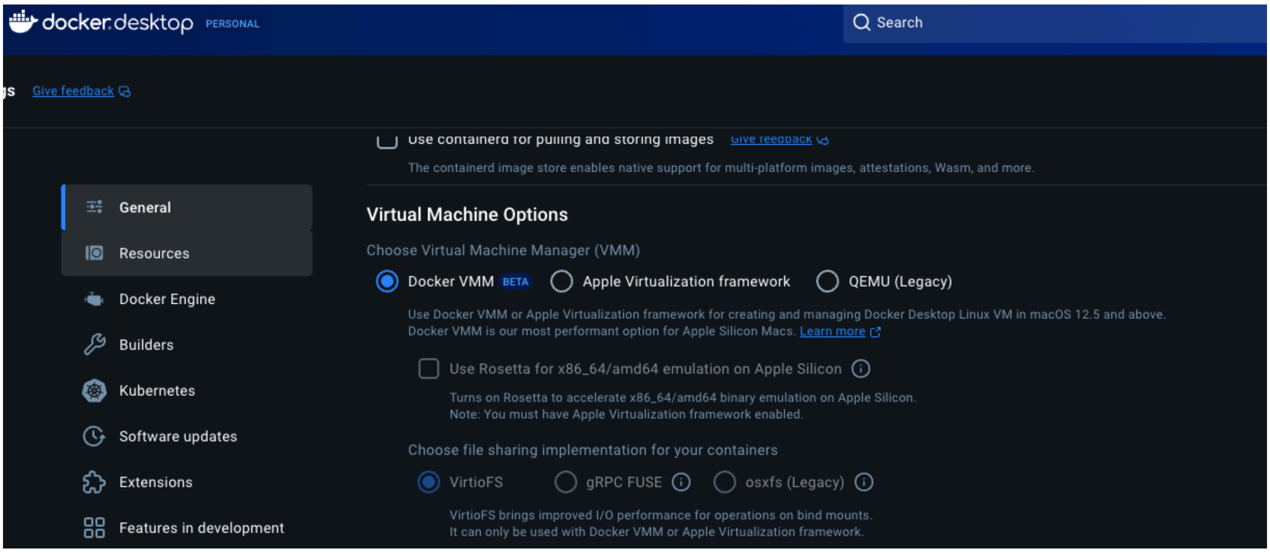
Then

```
docker run -p 8500:8500 -p 8501:8501 --mount
type=bind,source=`pwd`/serving_dir/tip_model,target=/models/tip_model
-e MODEL_NAME=tip_model -t emacski/tensorflow-serving
```

Then run the curl command as written, and you should get a prediction.

Or new since Oct 2024:

Beta release of Docker VMM - the more performant alternative to Apple Virtualization Framework on macOS (requires Apple Silicon and macOS 12.5 or later).
https://docs.docker.com/desktop/features/vmm/



□ Use containerd for pulling and storing images Give feedback

The containerd image store enables native support for multi-platform images, attestations, Wasm, and more.

Virtual Machine Options

Choose Virtual Machine Manager (VMM)

● Docker VMM BETA ○ Apple Virtualization framework ○ QEMU (Legacy)

Use Docker VMM or Apple Virtualization framework for creating and managing Docker Desktop Linux VM in macOS 12.5 and above.
Docker VMM is our most performant option for Apple Silicon Macs. Learn more

□ Use Rosetta for x86_64/amd64 emulation on Apple Silicon

Turns on Rosetta to accelerate x86_64/amd64 binary emulation on Apple Silicon.
Note: You must have Apple Virtualization framework enabled.

Choose file sharing implementation for your containers

● VirtioFS ○ gRPC FUSE ○ osxfs (Legacy)

VirtioFS brings improved I/O performance for operations on bind mounts.
It can only be used with Docker VMM or Apple Virtualization framework.

VMs - What do I do if my VM runs out of space?

- Try deleting data you've saved to your VM locally during ETLs
- Kill processes related to deleted files
- Download ncdu and look for large files (pay particular attention to files related to Prefect)
- If you delete any files related to Prefect, eliminate caching from your flow code

GCP BQ - External and regular table

External Table (data remains in GCS bucket)

Regular Table (data is copied into BigQuery storage)

Example of creating external table:

```
 CREATE OR REPLACE EXTERNAL TABLE `your_project.your_dataset.tablenamel`

OPTIONS (

  format = 'PARQUET',

  uris = ['gs://your-bucket-name/yellow_tripdata_2024-*.parquet']
);
```

Example of creating regular table from extermal table

```
CREATE OR REPLACE TABLE `your_project.your_dataset.tablename`

AS

SELECT * FROM `your_project.your_dataset.yellow_taxi_external`;
```

Or directly load data form GCS into a regular BigQuery table without creating an external table using:

```
CREATE OR REPLACE TABLE `your_project.your_dataset.yellow_taxi_table`

OPTIONS (

  format = 'PARQUET'

) AS

SELECT * FROM `your_project.your_dataset.external_table_placeholder`

FROM EXTERNAL_QUERY(

  'your_project.region-us.gcs_external',

  'SELECT * FROM `gs://your-bucket-name/yellow_tripdata_2024-*.parquet`'

);
```

Can BigQuery work with parquet files directly?

Yes, you can load your Parquet files directly into your GCP (Google Cloud Platform) Bucket first, then via BigQuery, you can create an external table of these Parquet files with a query statement like this:

```
CREATE OR REPLACE EXTERNAL TABLE
`module-3-data-warehouse.taxi_data.external_yellow_tripdata_2024`
OPTIONS (
  format = 'PARQUET',
  uris = ['gs://module3-dez/yellow_tripdata_2024-*.parquet']
);
```

Make sure to adjust the sql statement to your own situation and directories.
The * symbol can be used as a wildcard, which you will need to target Parquet files of all the months of 2024.

Homework - What does it mean “Stop with loading the files into a bucket.' Stop with loading the files into a bucket?”

Ans: What they mean is that they don't want you to do anything more than that. You should load the files into the bucket and create an external table based on those files (but nothing like cleaning the data and putting it in parquet format)

Homework - Reading parquets from nyc.gov directly into pandas returns Out of bounds error

If for whatever reason you try to read parquets directly from nyc.gov’s cloudfront into pandas, you might run into this error:

pyarrow.lib.ArrowInvalid: Casting from timestamp[us] to timestamp[ns] would result in out of bounds

Cause:

1. there is one errant data record where the dropOff_datetime was set to year 3019 instead of 2019.
2. pandas uses “timestamp[ns]” (as noted above), and int64 only allows a ~580 year range, centered on 2000. See `pd.Timestamp.max` and `pd.Timestamp.min`
3. This becomes out of bounds when pandas tries to read it because 3019 > 2300 (approx value of pd.Timestamp.Max

Fix:

1. Use pyarrow to read it:
   ```
   import pyarrow.parquet as pq df =
   pq.read_table('fhv_tripdata_2019-02.parquet').to_pandas(safe=
   False)
   ```
   However this results in weird timestamps for the offending record
2. Read the datetime columns separately using pq.read_table

```
table = pq.read_table(‘taxi.parquet’)
datetimes = [‘list of datetime column names’]
df_dts = pd.DataFrame()
    for col in datetimes:
        df_dts[col] = pd.to_datetime(table .column(col),
errors='coerce')
```

The `errors=’coerce’` parameter will convert the out of bounds timestamps into either the max or the min

3. Use parquet.compute.filter to remove the offending rows

```
import pyarrow.compute as pc
table = pq.read_table("‘taxi.parquet")
df = table.filter(
    pc.less_equal(table["dropOff_datetime"],
pa.scalar(pd.Timestamp.max))
).to_pandas()
```

Homework - Uploading files to GCS via GUI

This can help avoid schema issues in the homework.
Download files locally and use the ‘upload files’ button in GCS at the desired path. You can upload many files at once. You can also choose to upload a folder.

Homework - Qn 5: The partitioned/clustered table isn’t giving me the prediction I expected

Ans: Take a careful look at the format of the dates in the question.

Homework - Qn 6: Did anyone get an exact match for one of the options given in Module 3 homework Q6?

Many people aren’t getting an exact match, but are very close to one of the options. As per **Alexey said to choose the closest option**.

Python - invalid start byte Error Message

```
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xa0 in
position 41721: invalid start byte
```

Solution:

Step 1: When reading the data from the web into the pandas dataframe mention the encoding as follows:

```
pd.read_csv(dataset_url, low_memory=False, encoding='latin1')
```

Step 2: When writing the dataframe from the local system to GCS as a csv mention the encoding as follows:

```
df.to_csv(path_on_gsc, compression="gzip", encoding='utf-8')
```

Alternative: use `pd.read_parquet(url)`

Python - Generators in python

A generator is a function in python that returns an iterator using the yield keyword.

A generator is a special type of iterable, similar to a list or a tuple, but with a crucial difference. Instead of creating and storing all the values in memory at once, a generator generates values on-the-fly as you iterate over it. This makes generators memory-efficient, particularly when dealing with large datasets.

Python - Easiest way to read multiple files at the same time?

The read_parquet function supports a list of files as an argument. The list of files will be merged into a single result table.

Python - These won't work. You need to make sure you use Int64:

Incorrect:

```
df['DOlocationID'] = pd.to_numeric(df['DOlocationID'], downcast=integer) or
```

```
df['DOlocationID'] = df['DOlocationID'].astype(int)
```

Correct:

```
df['DOlocationID'] = df['DOlocationID'].astype('Int64')
```

Warning when run load_yellow_data python script

```
RuntimeWarning: As the c extension couldn't be imported,
google-crc32c is using a pure python implementation that is
significantly slower. If possible, please configure a c build
environment and compile extention
warnings.warn(_SLOW_CRC32C_WARNING, RuntimeWarning)

Failed to upload ./yellow_tripdata_2024-01.parquet to GCS:
Timeout of 120.0s exceeded, last exception: ('Connection
aborted.', timeout('The write operation timed out'))

Failed to upload ./yellow_tripdata_2024-03.parquet to GCS:
Timeout of 120.0s exceeded, last exception: ('Connection
aborted.', timeout('The write operation timed out'))
```

Im facing two separate issues in my script:

1. google-crc32c Warning: The Google Cloud Storage library is using a slow Python implementation instead of the optimized C version.
2. Upload Timeout Error: Your file uploads are timing out after 120 seconds.

✅ Solution: Install the C-optimized google-crc32c

pip install --upgrade google-crc32c

2. Fix Google Cloud Storage Upload Timeout

✅ Solution 1: Increase Timeout

blob.upload_from_filename(file_path, timeout=300) # Set timeout to 5 minutes

Module 4: analytics engineering with dbt

dbt cloud Developer

Please be aware that the demos are done using **dbt cloud Developer** licensing. Although Team license is available to you upon creation of dbt cloud account for 14 days, **the interface won't fully match the demo-ed experience.**

DBT-Config ERROR on CLOUD IDE: No dbt_project.yml found at expected path

(Lower left Corner after setting all connections to BQ and Github)

```
14:48:39 Running dbt...

14:48:39 Encountered an error:

Runtime Error

  No dbt_project.yml found at expected path /usr/src/develop/user-70471823426120/environment-70471823422561/repository-70471823410839/dbt_project.yml

  Verify that each entry within packages.yml (and their transitive dependencies) contains a file named dbt_project.yml
```

Solution: Initialize a project through UI.

Importing git repo of an existing dbt project:

Please read through these details for doing it: https://docs.getdbt.com/docs/cloud/git/import-a-project-by-git-url

DBT Cloud production error: prod dataset not available in location EU

Problem: I am trying to deploy my DBT models to production, using DBT Cloud. The data should live in BigQuery. The dataset location is EU. However, when I am running the model in production, a prod dataset is being create in BigQuery with a location US and the dbt invoke build is failing giving me "ERROR 404: porject.dataset:prod not available in location EU". I tried different ways to fix this. I am not sure if there is a more simple solution then creating my project or buckets in location US. Hope anyone can help here.

Note: Everything is working fine in development mode, the issue is just happening when scheduling and running job in production

Solution: I created the prod dataset manually in BQ and specified EU, then I ran the job.

How do I solve the Dbt Cloud error: prod was not found in location?

You might get this error while trying to run dbt in production aftering following the instructions in the video ‘DE Zoomcamp 4.4.1 - Deployment Using dbt Cloud (Alternative A’):

Database Error in model stg_yellow_tripdata (models/staging/stg_yellow_tripdata.sql)
Not found: Dataset module-4-analytics-eng:prod was not found in location europe-west10

This error is easily solved. There are two solutions to solve this issue:

Solution #1: Matching the dataset's data location with the source dataset

Set your ‘prod’ dataset's data location to match the data location of your ‘trips_data_all’ dataset's data location (in BigQuery). Running dbt in production works for the instructor, because her ‘ prod’ is in the same region as her source data. Since your ‘trips_data_all’ is in europe-west10 (or anything else besides US), your prod needs to be there too; not US (which is a default setting when dbt creates a dataset for you in BigQuery).

Solution #2: Changing the dataset to <development dataset>

Go into your dbt production environment settings:

1. Go to: Deploy / Environments / Production (your production environment) / Settings
2. Now look at the Deployment credentials. There is an input field here called Dataset. The input of ‘prod’ is likely in here.
3. Replace ‘prod’ with the name of the Dataset that you worked with while in development (before moving to Production). This is the Dataset name inside your BigQuery where you successfully ran ‘dbt debug’ and ‘dbt build’ with.
4. After saving, you are ready to rerun your Job!

Setup - No development environment

Error: `This project does not have a development environment configured. Please create a development environment and configure your development credentials to use the dbt IDE.`

The error itself tells us how to solve this issue, the guide is here. And from videos @1:42 and also slack chat

Setup - Connecting dbt Cloud with BigQuery Error

```
Runtime Error

dbt was unable to connect to the specified database.

  The database returned the following error:
```

```
 >Database Error

Access Denied: Project <project_name>: User does not have
bigquery.jobs.create permission in project <project_name>.

Check your database credentials and try again. For more information,
visit:

https://docs.getdbt.com/docs/configure-your-profile
```

Steps to resolve error in Google Cloud:

1. Navigate to **IAM & Admin** and select **IAM**

2. Click **Grant Access** if your newly created dbt service account isn't listed

3. In ***New principals*** field, add your service account

4. Select a **Role** and search for **BigQuery Job User** to add

5. Go back to *dbt cloud project setup* and Test your connection

6. ***Note***: Also add **BigQuery Data Owner**, **Storage Object Admin**, & **Storage Admin** to prevent permission issues later in the course

Setup - Failed to clone repository.

Error: `Failed to clone repository.`

```
git clone
git@github.com:DataTalksClub/data-engineering-zoomcamp.git
/usr/src/develop/…
Cloning into '/usr/src/develop/...
Warning: Permanently added 'github.com,140.82.114.4' (ECDSA) to
the list of known hosts.
git@github.com: Permission denied (publickey).
fatal: Could not read from remote repository.
```

Issue: You don’t have permissions to write to `DataTalksClub/data-engineering-zoomcamp.git`

Solution 1: Clone the repository and use this forked repo, which contains your github username. Then, proceed to specify the path, as in:

```
[your github username]/data-engineering-zoomcamp.git
```

Solution 2: create a fresh repo for dbt-lessons. We’d need to do branching and PRs in this lesson, so it might be a good idea to also not mess up your whole other repo. Then you don’t have to create a subfolder for the **dbt** project files

Solution 3: Use https link

Errors when I start the server in dbt cloud: Failed to start server. Permission denied (publickey)

Failed to start server. Permission denied (publickey). fatal: Could not read from remote repository. Please make sure you have the correct access rights and the repository exists.

Use the deploy keys in dbt repo details to create a public key in your repo, the issue will be solved.

Steps in details:

1. **Find dbt Cloud’s SSH Key**
   - In dbt Cloud, go to **Settings > Account Settings > SSH Keys**
   - Copy the **public SSH key** displayed there.
2. **Add It to GitHub**
   - Go to **GitHub > Settings > SSH and GPG Keys**
   - Click **"New SSH Key"**, name it "dbt Cloud", and paste the key.
   - Click **"Add SSH Key"**.
3. **Try Restarting dbt Cloud**

dbt job - Triggered by pull requests is disabled prerequisites when I try to create a new Continuous Integration job in dbt cloud.

Solution:

Check if you're on the Developer Plan. As per the prerequisites, you'll need to be enrolled in the Team Plan or Enterprise Plan to set up a CI Job in dbt Cloud.

So If you're on the Developer Plan, you'll need to upgrade to utilise CI Jobs.

Note from another user: I'm in the Team Plan (trial period) but the option is still disabled. What worked for me instead was this. It works for the Developer (free) plan.

Setup - `Your IDE session was unable to start. Please contact support.`

Issue: If the DBT cloud IDE loading indefinitely then giving you this error

Solution: check the dbt_cloud_setup.md file and make a SSH Key and use gitclone to import repo into dbt project, copy and paste deploy key back in your repo setting.

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

Issue: 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

Need a quick change to make the file work with gz files, added the following lines (and don't forget to delete the file at the end of each iteration of the loop to avoid any problem of disk space)

```
file_name_gz = f"{service}_tripdata_{year}-{month}.csv.gz"

open(file_name_gz, 'wb').write(r.content)

os.system(f"gzip -d {file_name_gz}")

os.system(f"rm {file_name_init}.*")
```

“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_feel 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'}))
```

Same type of ERROR - parquet files with different data types - Fix it with pandas

Here is another possibility that could be interesting:

You can specify the dtypes when importing the file from csv to a dataframe with pandas

pd.from_csv(..., dtype=type_dict)

One obstacle is that the regular int64 pandas use (I think this is from the numpy library) does not accept null values (NaN, not a number). But you can use the pandas Int64 instead, notice capital ‘I’. The type_dict is a python dictionary mapping the column names to the dtypes.

Sources:

https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

Nullable integer data type — pandas 1.5.3 documentation

Ingestion: When attempting to use the provided quick script to load trip data into GCS, you receive error Access Denied from the S3 bucket

If the provided URL isn’t working for you (https://nyc-tlc.s3.amazonaws.com/trip+data/):

We can use the GitHub CLI to easily download the needed trip data from https://github.com/DataTalksClub/nyc-tlc-data, and manually upload to a GCS bucket.

Instructions on how to download the CLI here: https://github.com/cli/cli

Commands to use:

```
gh auth login

gh release list -R DataTalksClub/nyc-tlc-data

gh release download yellow -R DataTalksClub/nyc-tlc-data

gh release download green -R DataTalksClub/nyc-tlc-data

etc.
```

Now you can upload the files to a GCS bucket using the GUI.

Hack to load yellow and green trip data for 2019 and 2020

I initially followed data-engineering-zoomcamp/03-data-warehouse/extras/web_to_gcs.py at main · DataTalksClub/data-engineering-bootcamp (github.com)

But it was taking forever for the yellow trip data and when I tried to download and upload the parquet files directly to GCS, that works fine but when creating the Bigquery table, there was a schema inconsistency issue

Then I found another hack shared in the slack which was suggested by Victoria.

[Optional] Hack for loading data to BigQuery for Week 4 - YouTube

Please watch until the end as there is few schema changes required to be done

GCP VM - All of sudden ssh stopped working for my VM after my last restart

One common cause experienced is lack of space after running prefect several times. When running prefect, check the folder ‘.prefect/storage’ and delete the logs now and then to avoid the problem.

GCP FREE TRIAL ACCOUNT ERROR

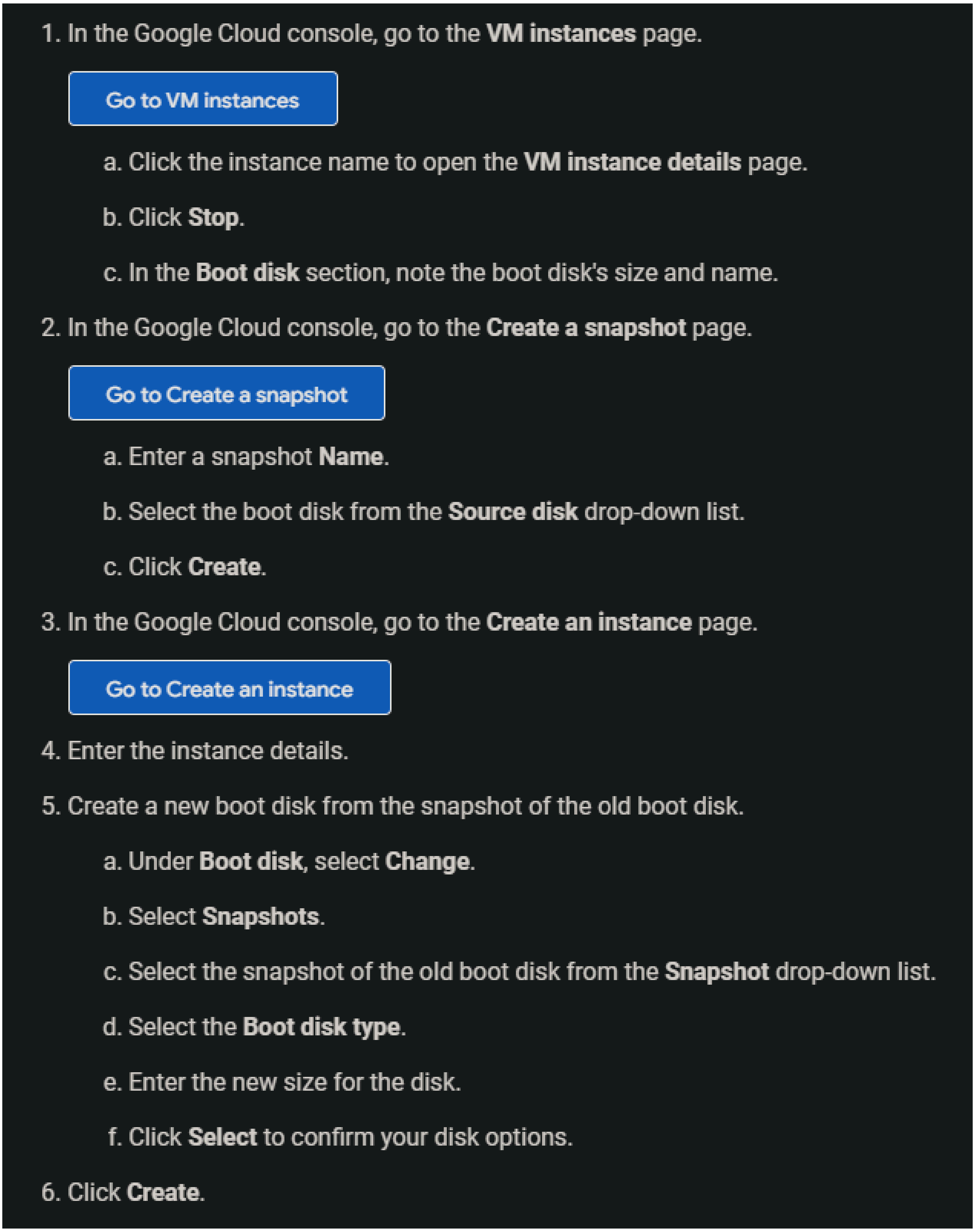
If you're encountering an error when trying to create a GCP free trial account, the issue isn't related to country restrictions, credit/debit card problems, or IP issues, it's a random problem with no clear logical reason behind it. Here's a simple workaround that worked for me:

I asked a few friends in my country to try signing up for the free trial using their Gmail accounts and their debit/credit cards. One of them was able to successfully create the account, and I'm temporarily using their Gmail to access the trial.

If you're still running into the issue, this method could help you bypass the problem!

GCP VM - If you have lost SSH access to your machine due to lack of space. Permission denied (publickey)

You can try to do this steps:



1. In the Google Cloud console, go to the **VM instances** page.

   Go to VM instances

   a. Click the instance name to open the **VM instance details** page.

   b. Click **Stop**.

   c. In the **Boot disk** section, note the boot disk's size and name.

2. In the Google Cloud console, go to the **Create a snapshot** page.

   Go to Create a snapshot

   a. Enter a snapshot **Name**.

   b. Select the boot disk from the **Source disk** drop-down list.

   c. Click **Create**.

3. In the Google Cloud console, go to the **Create an instance** page.

   Go to Create an instance

4. Enter the instance details.

5. Create a new boot disk from the snapshot of the old boot disk.

   a. Under **Boot disk**, select **Change**.

   b. Select **Snapshots**.

   c. Select the snapshot of the old boot disk from the **Snapshot** drop-down list.

   d. Select the **Boot disk type**.

   e. Enter the new size for the disk.

   f. Click **Select** to confirm your disk options.

6. Click **Create**.

DBT - When running your first dbt model, if it fails with an error:

- 404 Not found: Dataset was not found in location US
- 404 Not found: Dataset eighth-zenith-372015:trip_data_all was not found in location us-west1

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 dbt_<first initial><last name> (if you didn't change the default settings at the end when setting up your project))

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. **GPo** 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. (NOTE:** be sure to exactly copy the region BigQuery specifies your dataset is in.**)**

DBT - When executing dbt run after installing dbt-utils latest version i.e., 1.0.0 warning has generated

Error: `dbt_utils.surrogate_key` has been replaced by `dbt_utils.generate_surrogate_key`

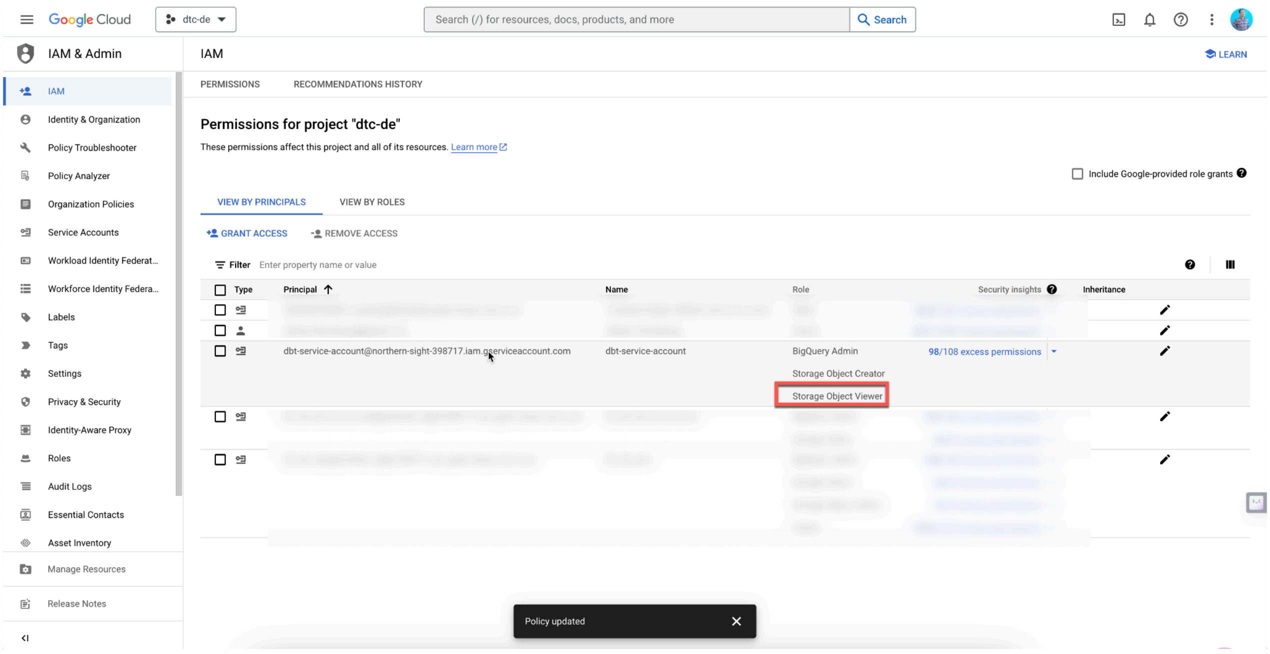
Fix:

Replace `dbt_utils.surrogate_key` with `dbt_utils.generate_surrogate_key` in stg_green_tripdata.sql

When executing dbt run after fact_trips.sql has been created, the task failed with error: “Access Denied: BigQuery BigQuery: Permission denied while globbing file pattern.”

1. Fixed by adding the Storage Object Viewer role to the service account in use in BigQuery.

2. Add the related roles to the service account in use in GCS.



Permissions for project "dtc-de"

These permissions affect this project and all of its resources. Learn more

□ Include Google-provided role grants

□ Type	Principal ↑	Name	Role	Security insights	Inheritance
□	[redacted]	[redacted]	[redacted]	[redacted]	
□	[redacted]	[redacted]	[redacted]	[redacted]	
□	dbt-service-account@northern-sight-398717.iam.gserviceaccount.com	dbt-service-account	BigQuery Admin Storage Object Creator Storage Object Viewer	98/108 excess permissions	
□	[redacted]	[redacted]	[redacted]	[redacted]	
□	[redacted]	[redacted]	[redacted]	[redacted]	

Policy updated ✕

When You are getting error dbt_utils not found

You need to create packages.yml file in main project directory and add packages' meta data:

```
packages:

  - package: dbt-labs/dbt_utils

        version: 0.8.0
```

After creating file run:

```
dbt deps
```

And hit enter.

Lineage is currently unavailable. Check that your project does not contain compilation errors or contact support if this error persists.

Ensure you properly format your yml file. Check the build logs if the run was completed successfully. You can expand the command history console (where you type the `--vars '{'is_test_run': 'false'}'`) and click on any stage's logs to expand and read errors messages or warnings.

Build - 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). If that does not work, use `--vars '{'is_test_run': 'false'}'` with each phrase separately quoted.

Build - Why do my fact_trips only contain one month of data?

Check if you specified `if_exists` argument correctly when writing data from GCS to BigQuery. When I wrote my automated flow for each month of the years 2019 and 2020 for green and yellow data I had specified `if_exists="replace"` while I was experimenting with the flow setup. Once you want to run the flow for all months in 2019 and 2020 make sure to set `if_exists="append"`

- `if_exists="replace"` will replace the whole table with only the month data that you are writing into BigQuery in that one iteration -> you end up with only one month in BigQuery (the last one you inserted)
- `if_exists="append"` will append the new monthly data -> you end up with data from all months

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!

DBT - Warning: dbt_utils.surrogate_key has been replaced by dbt_utils.generate_surrogate_key. The new macro treats null values(...)To restore the behaviour of the original macro,

That means the surrogate_key has been deprecated, and it indicates you should replace it with the new method `generate_surrogate_key`

Replace:

```
{{ dbt_utils.surrogate_key([
    field_a,
    field_b,
    field_c,
    …,
    field_z
]) }}
```

For this instead:

```
{{ dbt_utils.generate_surrogate_key([
    field_a,
    field_b,
    field_c,
    …,
    field_z
]) }}
```

add a global variable in dbt_project.yml(...)

```
Warning: `dbt_utils.surrogate_key` has been replaced by


`dbt_utils.generate_surrogate_key`. The new macro treats null values differently to empty
strings. To restore the behaviour of the original macro, add a global variable in
dbt_project.yml called `surrogate_key_treat_nulls_as_empty_strings` to your dbt_project.yml file
with a value of True. The taxi_rides_ny.stg_yellow_tripdata model triggered this warning.
```

I changed location in dbt, but dbt run still gives me an error

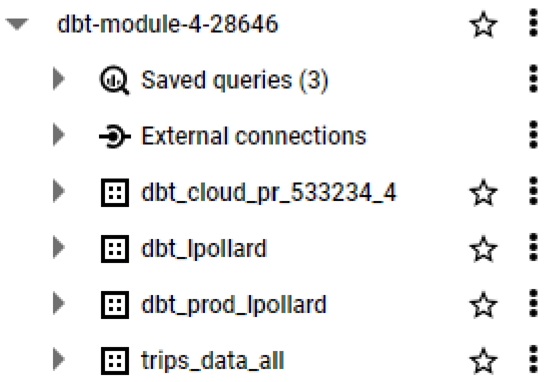
Remove the dataset from BigQuery which was created by dbt and run dbt run again so that it will recreate the dataset in BigQuery with the correct location

DBT - I ran dbt run without specifying variable which gave me a table of 100 rows. I ran again with the variable value specified but my table still has 100 rows in BQ.

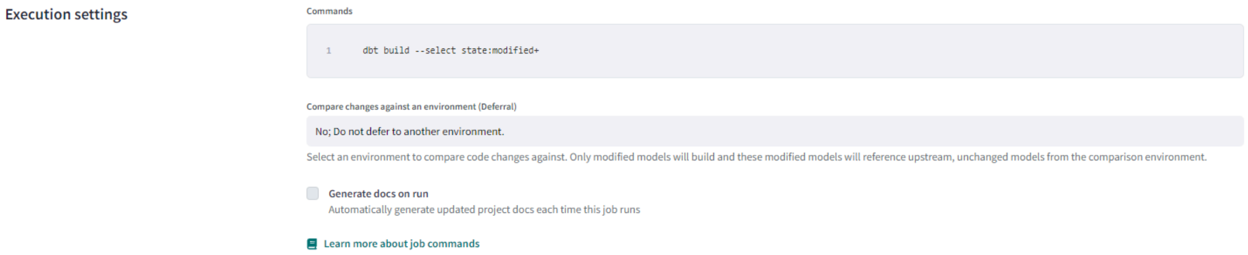
Remove the dataset from BigQuery created by dbt and run again (with test disabled) to ensure the dataset created has all the rows.

DBT - Why am I getting a new dataset after running my CI/CD Job? / What is this new dbt dataset in BigQuery?

Answer: *when you create the CI/CD job, under 'Compare Changes against an environment (Deferral) make sure that you select ' No; do not defer to another environment' - otherwise dbt won't merge your dev models into production models; it will create a new environment called 'dbt_cloud_pr_number of pull request'*



- dbt-module-4-28646
  - Saved queries (3)
  - External connections
  - dbt_cloud_pr_533234_4
  - dbt_lpollard
  - dbt_prod_lpollard
  - trips_data_all



Execution settings

Commands

```
dbt build --select state:modified+
```

Compare changes against an environment (Deferral)

No; Do not defer to another environment.

Select an environment to compare code changes against. Only modified models will build and these modified models will reference upstream, unchanged models from the comparison environment.

□ Generate docs on run

Automatically generate updated project docs each time this job runs

Learn more about job commands

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.

DBT - Docs Served but Not Accessible via Browser

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

BigQuery adapter: 404 Not found: Dataset was not found in location europe-west6

1. Go to Account settings >> Project >> Analytics >> Click on your connection >> go all the way down to Location and type in the GCP location just as displayed in GCP (e.g. europe-west6). You might need to reupload your GCP key.
2. Delete your dataset in GBQ
3. Rebuild project: dbt build
4. Newly built dataset should be in the correct location

Dbt+git - Main branch is “read-only”

Create a new branch to edit. More on this can be found here in the dbt docs.

Dbt+git - It appears that I can't edit the files because I'm in read-only mode. Does anyone know how I can change that?

Create a new branch for development, then you can merge it to the main branch

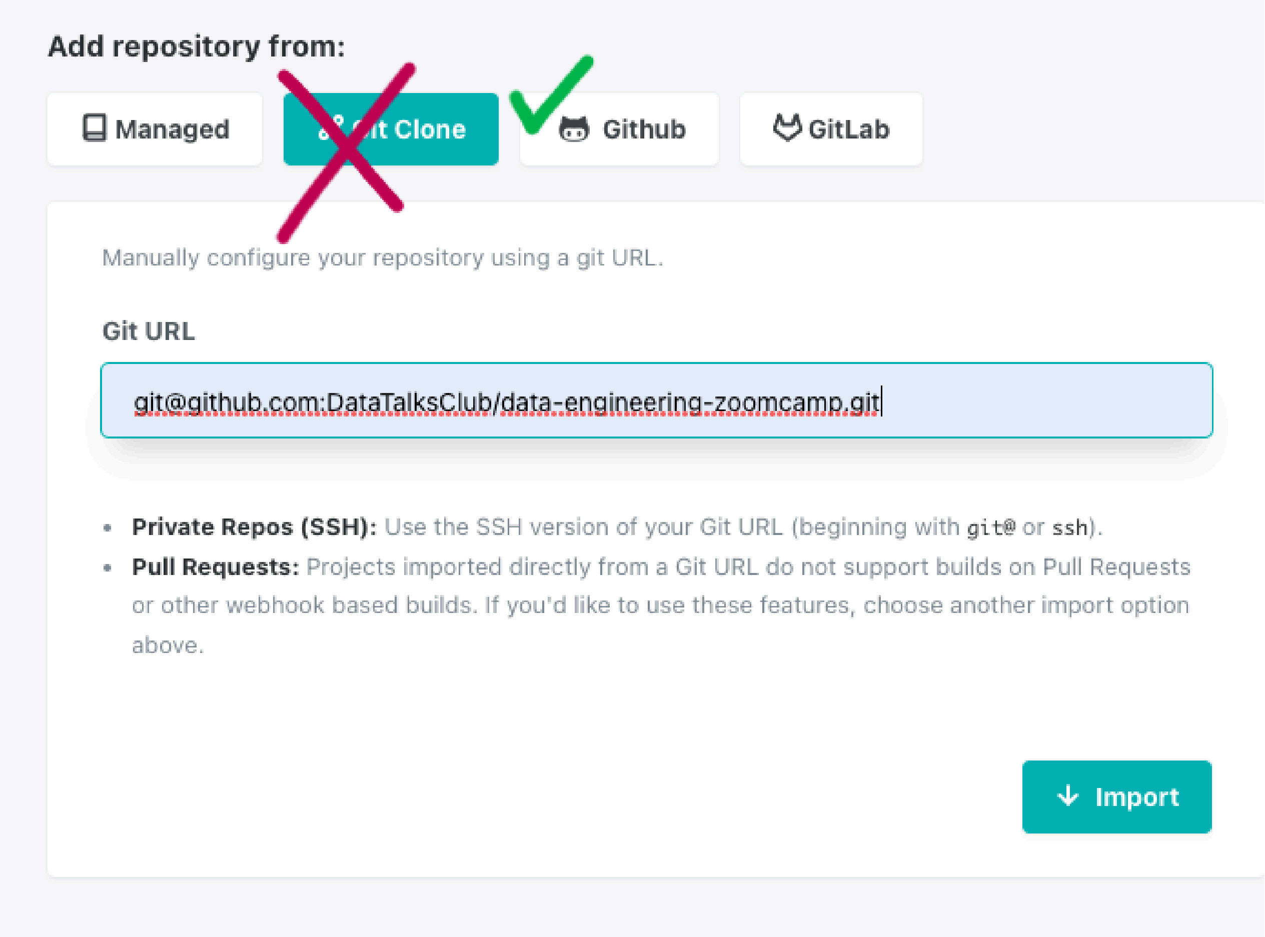
Create a new branch and switch to this branch. It allows you to make changes. Then you can commit and push the changes to the “main” branch.

Dbt deploy + Git CI - cannot create CI checks job for deployment to Production. See more discussion in slack chat

Error:

```
Triggered by pull requests

This feature is only available for dbt repositories connected through dbt Cloud's native
integration with Github, Gitlab, or Azure DevOps
```

Solution: Contrary to the guide on DTC repo, don’t use the **Git Clone** option. Use the **Github** one instead. Step-by-step guide to UN-LINK **Git Clone** and RE-LINK with **Github** in the next entry below



Add repository from:

Manually configure your repository using a git URL.

Git URL

git@github.com:DataTalksClub/data-engineering-zoomcamp.git

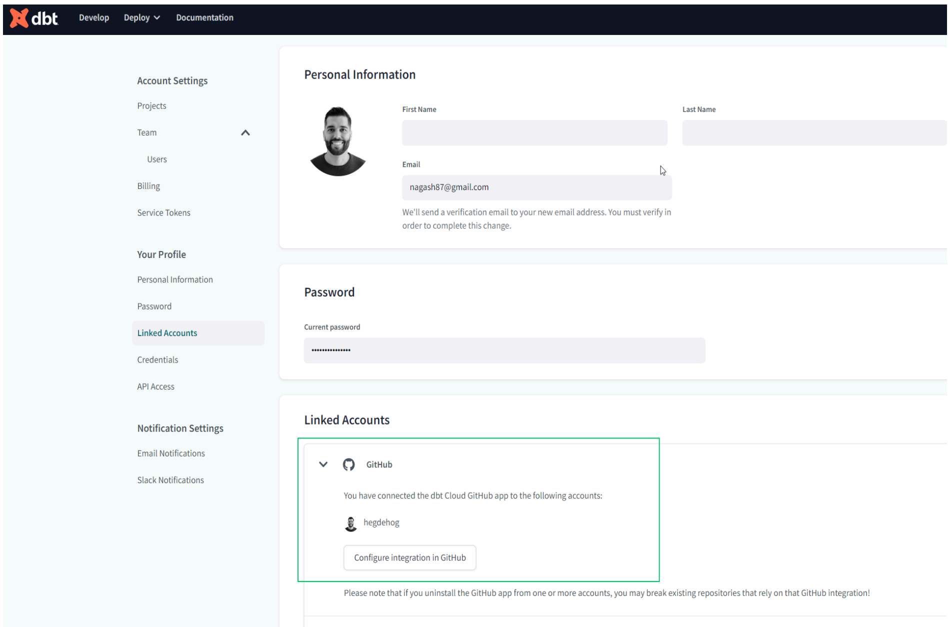
- **Private Repos (SSH):** Use the SSH version of your Git URL (beginning with `git@` or `ssh`).
- **Pull Requests:** Projects imported directly from a Git URL do not support builds on Pull Requests or other webhook based builds. If you'd like to use these features, choose another import option above.

Dbt deploy + Git CI - Unable to configure Continuous Integration (CI) with Github

If you're trying to configure CI with Github and on the job's options you can't see **Run on Pull Requests?** on triggers, you have to reconnect with Github using native connection instead clone by SSH. Follow these steps:

1. On **Profile Settings > Linked Accounts** connect your Github account with dbt project allowing the permissions asked. More info at https://docs.getdbt.com/docs/collaborate/git/connect-gith

2.



Personal Information

First Name

Last Name

Email: nagash87@gmail.com

We'll send a verification email to your new email address. You must verify in order to complete this change.

Password

Current password: ••••••••••••••

Linked Accounts

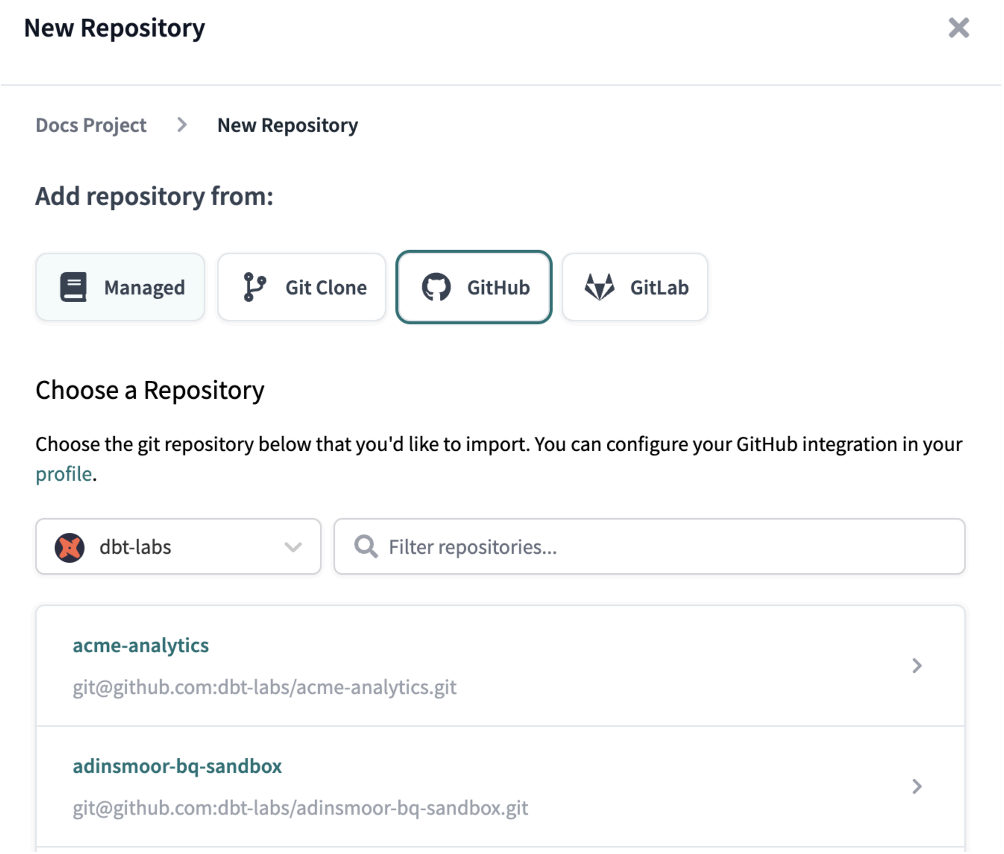
You have connected the dbt Cloud GitHub app to the following accounts:

hegdehog

Please note that if you uninstall the GitHub app from one or more accounts, you may break existing repositories that rely on that GitHub integration!

3. Disconnect your current Github's configuration from ***Account Settings > Projects (analytics) > Github connection.*** At the bottom left appears the button *Disconnect,* press it.

4. Once we have confirmed the change, we can configure it again. This time, choose *Github* and it will appear in all repositories which you have allowed to work with dbt. Select your repository and it's ready.



New Repository

Docs Project > New Repository

Add repository from:

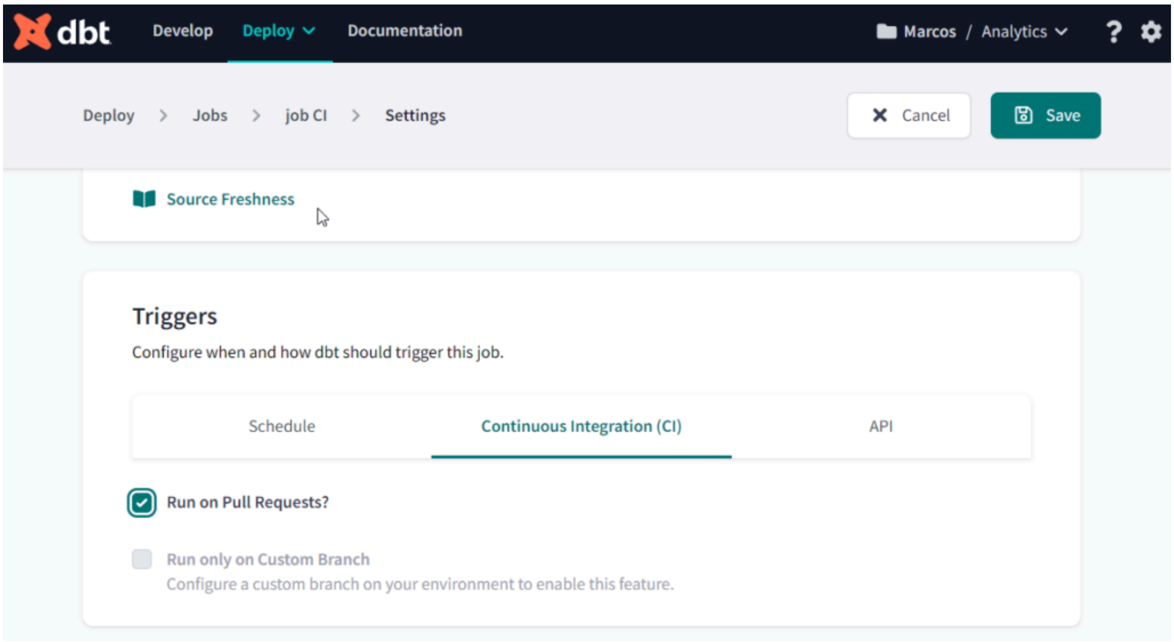
Choose a Repository

Choose the git repository below that you'd like to import. You can configure your GitHub integration in your profile.

acme-analytics
git@github.com:dbt-labs/acme-analytics.git

adinsmoor-bq-sandbox
git@github.com:dbt-labs/adinsmoor-bq-sandbox.git

5. Go to the **Deploy > job** configuration's page and go down until ***Triggers*** and now you can see the option *Run on Pull Requests*:



Triggers

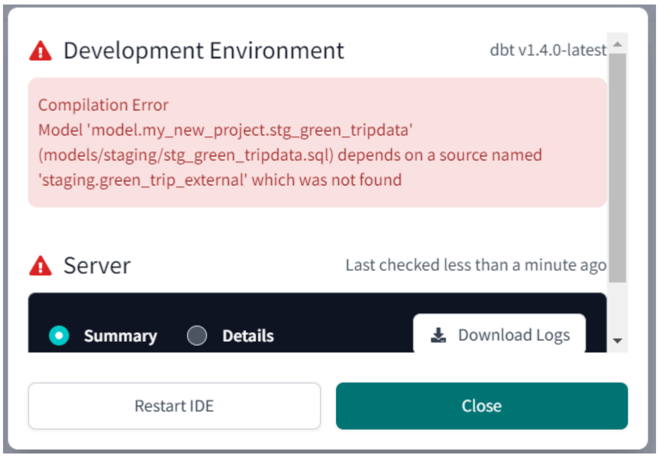
Configure when and how dbt should trigger this job.

☑ Run on Pull Requests?

□ Run only on Custom Branch
Configure a custom branch on your environment to enable this feature.

Compilation Error (Model 'model.my_new_project.stg_green_tripdata' (models/staging/stg_green_tripdata.sql) depends on a source named 'staging.green_trip_external' which was not found)

If you're following video DE Zoomcamp 4.3.1 - Building the First DBT Models, you may have encountered an issue at 14:25 where the Lineage graph isn't displayed and a Compilation Error occurs, as shown in the attached image. Don't worry - a quick fix for this is to simply **save your schema.yml** file. Once you've done this, you should be able to view your Lineage graph without any further issues.



Compilation Error
Model 'model.my_new_project.stg_green_tripdata' (models/staging/stg_green_tripdata.sql) depends on a source named 'staging.green_trip_external' which was not found

Compilation Error in test accepted_values_stg_green_tripdata_Payment_type__False___var_payment_type_values_ (models/staging/schema.yml) 'NoneType' object is not iterable

```
 > in macro test_accepted_values (tests/generic/builtin.sql)

 > called by test
accepted_values_stg_green_tripdata_Payment_type__False___var_payme
nt_type_values_ (models/staging/schema.yml)
```

Remember that you have to add to dbt_project.yml the vars:

```
vars:

  payment_type_values: [1, 2, 3, 4, 5, 6]
```

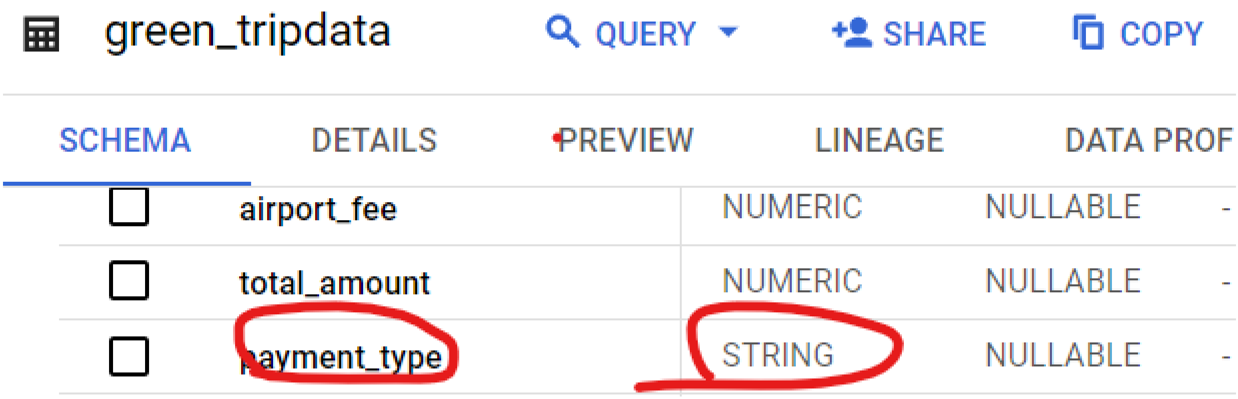
dbt macro errors with get_payment_type_description(payment_type)

You will face this issue if you copied and pasted the exact macro directly from data-engineering-zoomcamp repo.

```
BigQuery adapter: Retry attempt 1 of 1 after error: BadRequest('No
matching signature for operator CASE for argument types: STRING, INT64,
STRING, INT64, STRING, INT64, STRING, INT64, STRING, INT64, STRING,
INT64, STRING, NULL at [35:5]; reason: invalidQuery, location: query,
message: No matching signature for operator CASE for argument types:
STRING, INT64, STRING, INT64, STRING, INT64, STRING, INT64, STRING,
INT64, STRING, INT64, STRING, NULL at [35:5]')
```

What you'd have to do is to change the data type of the numbers (1, 2, 3 etc.) to text by inserting '', as the initial 'payment_type' data type should be string (Note: I extracted and loaded the green trips data using Google BQ Marketplace)

```
 {#

    This macro returns the description of the payment_type

#}


{% macro get_payment_type_description(payment_type) -%}


    case {{ payment_type }}

        when '1' then 'Credit card'
        when '2' then 'Cash'
        when '3' then 'No charge'
        when '4' then 'Dispute'
        when '5' then 'Unknown'
        when '6' then 'Voided trip'
    end

{%- endmacro %}
```



□	airport_fee	NUMERIC	NULLABLE	-
□	total_amount	NUMERIC	NULLABLE	-
□	payment_type	STRING	NULLABLE	-

Troubleshooting in dbt:

The dbt error log contains a link to BigQuery. When you follow it you will see your query and the problematic line will be highlighted.

DBT - Why changing the target schema to “marts” actually creates a schema named “dbt_marts” instead?

It is a default behaviour of dbt to append custom schema to initial schema. To override this behaviour simply create a macro named “generate_schema_name.sql”:

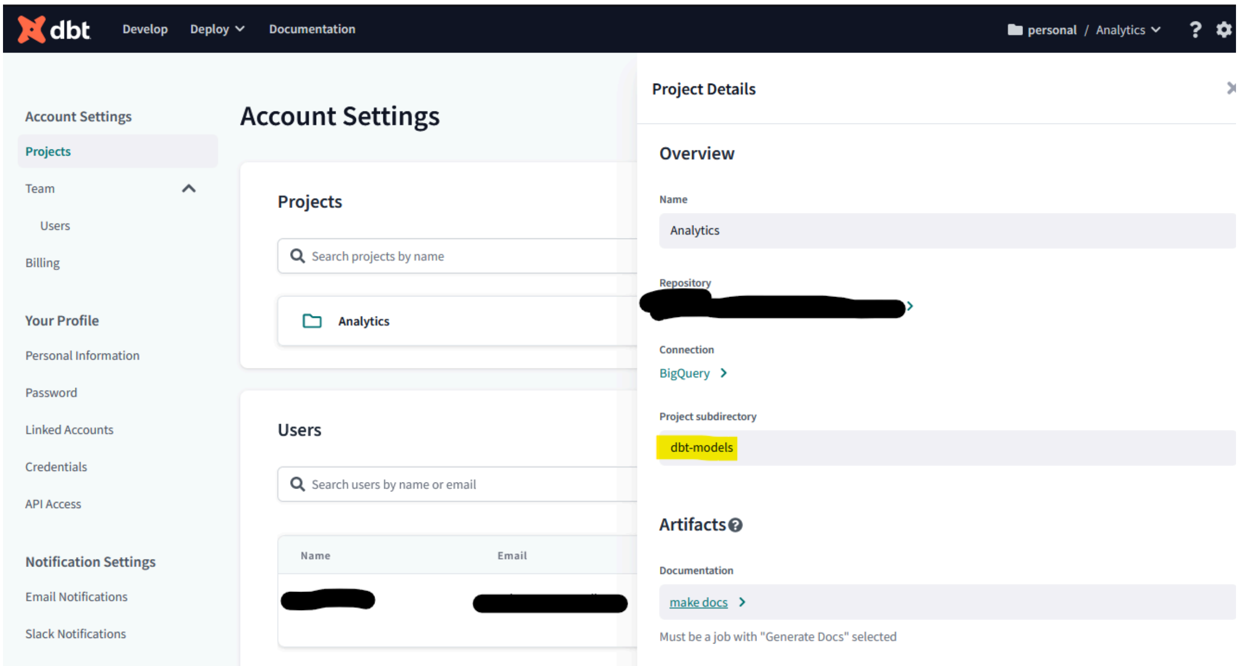
```
{% macro generate_schema_name(custom_schema_name, node) -%}
    {%- set default_schema = target.schema -%}
    {%- if custom_schema_name is none -%}
```

```
        {{ default_schema }}
    {%- else -%}
        {{ custom_schema_name | trim }}
    {%- endif -%}
{%- endmacro %}
```

Now you can override default custom schema in “dbt_project.yml”:

How to set subdirectory of the github repository as the dbt project root

There is a project setting which allows you to set `Project subdirectory` in dbt cloud:



Name	Email
[redacted]	[redacted]

Project Details

Overview

Name
Analytics

Repository
[redacted]

Connection
BigQuery

Project subdirectory
dbt-models

Artifacts

Documentation
make docs

Must be a job with "Generate Docs" selected

Compilation Error : Model 'model.XXX' (models/<model_path>/XXX.sql) depends on a source named '<a table name>' which was not found

Remember that you should modify accordingly your .sql models, to read from existing table names in BigQuery/postgres db

Example: `select * from {{ source('staging',<your table name in the database>') }}`

Compilation Error : Model '<model_name>' (<model_path>) depends on a node named '<seed_name>' which was not found   (Production Environment)

Make sure that you create a pull request from your Development branch to the Production branch (`main` by default). After that, check in your 'seeds' folder if the seed file is inside it. Another thing to check is your `.gitignore` file. Make sure that the .csv extension is not included.

When executing dbt run after using fhv_tripdata as an external table: **you** get "Access Denied: BigQuery BigQuery: Permission denied"

1. Go to your dbt cloud service account

1. Adding the [Storage Object Admin,Storage Admin] role in addition tco BigQuery Admin.

How to automatically infer the column data type (pandas missing value issues)?

Problem: when injecting data to bigquery, you may face the type error. This is because pandas by default will parse integer columns with missing value as float type.

Solution:

- One way to solve this problem is to specify/ cast data type Int64 during the data transformation stage.
- However, you may be lazy to type all the int columns. If that is the case, you can simply use `convert_dtypes` to infer the data type

```
    # Make pandas to infer correct data type (as pandas parse int
with missing as float)

    df.fillna(-999999, inplace=True)ingesting
    df = df.convert_dtypes()

    df = df.replace(-999999, None)
```

When loading github repo raise exception that ‘taxi_zone_lookup’ not found

Seed files loaded from directory with name ‘seed’, that’s why you should rename dir with name ‘data’ to ‘seed’

‘taxi_zone_lookup’ not found

Check the .gitignore file and make sure you don’t have *.csv in it

Dbt error 404 was not found in location

My specific error:
Runtime Error in rpc request (from remote system.sql) 404 Not found: Table dtc-de-0315:trips_data_all.green_tripdata_partitioned was not found in location europe-west6 Location: europe-west6 Job ID: 168ee9bd-07cd-4ca4-9ee0-4f6b0f33897c

Make sure all of your datasets have the correct region and not a generalised region:
Europe-west6 as opposed to EU

Match this in dbt settings:
dbt -> projects -> optional settings -> manually set location to match

Data type errors when ingesting with parquet files

The easiest way to avoid these errors is by ingesting the relevant data in a .csv.gz file type. Then, do:

CREATE OR REPLACE EXTERNAL TABLE `dtc-de.trips_data_all.fhv_tripdata`

OPTIONS (

format = 'CSV',

uris = ['gs://dtc_data_lake_dtc-de-updated/data/fhv_all/fhv_tripdata_2019-*.csv.gz']

```
);
```

As an example. You should no longer have any data type issues for week 4.

Inconsistent number of rows when re-running fact_trips model

This is due to the way the deduplication is done in the two staging files.

Solution: add `order by` in the `partition by` part of both staging files. Keep adding columns to order by until the number of rows in the fact_trips table is consistent when re-running the fact_trips model.

Explanation (a bit convoluted, feel free to clarify, correct etc.)

We partition by vendor id and pickup_datetime and choose the first row (rn=1) from all these partitions. These partitions are not ordered, so every time we run this, the first row might be a different one. Since the first row is different between runs, it might or might not contain an unknown borough. Then, in the fact_trips model we will discard a different number of rows when we discard all values with an unknown borough.

Data Type Error when running fact table

If you encounter data type error on trip_type column, it may due to some nan values that isn't null in bigquery.

Solution: try casting it to FLOAT datatype instead of NUMERIC

CREATE TABLE has columns with duplicate name locationid.

This error could result if you are using some select * query without mentioning the name of table for ex:

```
with dim_zones as (
    select * from `engaged-cosine-374921`.`dbt_victoria_mola`.`dim_zones`
    where borough != 'Unknown'
),
```

fhv as (

select * from `engaged-cosine-374921`.`dbt_victoria_mola`.`stg_fhv_tripdata`

)

select * from fhv

inner join dim_zones as pickup_zone

on fhv.PUlocationID = pickup_zone.locationid

inner join dim_zones as dropoff_zone

on fhv.DOlocationID = dropoff_zone.locationid

);

To resolve just replace use : **select fhv.* from fhv**

Bad int64 value: 0.0 error

Some ehail fees are null and casting them to integer gives Bad int64 value: 0.0 error,

Solution:

Using safe_cast returns NULL instead of throwing an error. So use safe_cast from dbt_utils function in the jinja code for casting into integer as follows:

```
 {{ dbt_utils.safe_cast('ehail_fee',
api.Column.translate_type("integer"))}} as ehail_fee,
```

Can also just use safe_cast(ehail_fee as integer) without relying on dbt_utils.

Bad int64 value: 2.0/1.0 error

You might encounter this when building the fact_trips.sql model. The issue may be with the **payment_type_description** field.

Using safe_cast as above, would cause the entire field to become null. A better approach is to drop the offending decimal place, then cast to integer.

```
cast(replace({{ payment_type }},'.0','') as integer)
```

Bad int64 value: 1.0 error (again)

I found that there are more columns causing the bad INT64: ratecodeid and trip_type on Green_tripdata table.
You can use the queries below to address them:

```
CAST(

        REGEXP_REPLACE(CAST(rate_code AS STRING), r'\.0', '') AS INT64

    ) AS ratecodeid,

CAST(

        CASE

            WHEN REGEXP_CONTAINS(CAST(trip_type AS STRING), r'\.\d+') THEN NULL

            ELSE CAST(trip_type AS INT64)

        END AS INT64

    ) AS trip_type,
```

DBT - Error on building fact_trips.sql: Parquet column 'ehail_fee' has type DOUBLE which does not match the target cpp_type INT64. File: gs://<gcs bucket>/<table>/green_taxi_2019-01.parquet")

The two solution above don’t work for me - I used the line below in `stg_green_trips.sql` to replace the original ehail_fee line:

`{{ dbt.safe_cast('ehail_fee',  api.Column.translate_type("numeric"))}} as ehail_fee,`

The - vars argument must be a YAML dictionary, but was of type str

Remember to add a space between the variable and the value. Otherwise, it won't be interpreted as a dictionary.

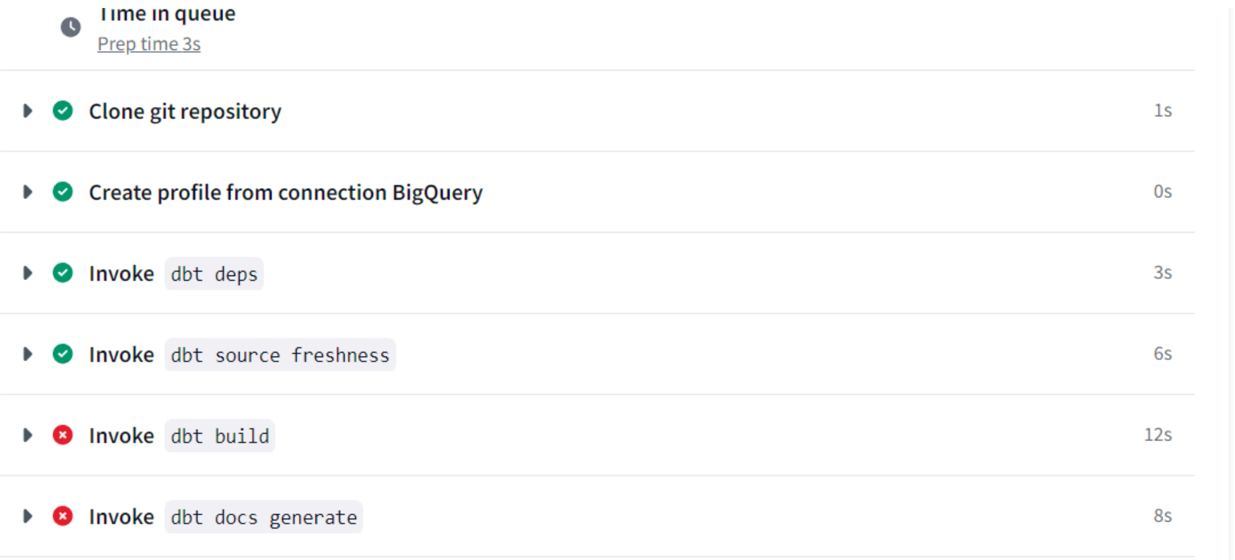
It should be:

dbt run --var 'is_test_run: false'

Not able to change Environment Type as it is greyed out and inaccessible

You don't need to change the environment type. If you are following the videos, you are creating a Production Deployment, so the only available option is the correct one.'

Access Denied: Table yellow_tripdata: User does not have permission to query table yellow_tripdata, or perhaps it does not exist in location US.



Time in queue
Prep time 3s

✓ Clone git repository 1s

✓ Create profile from connection BigQuery 0s

✓ Invoke `dbt deps` 3s

✓ Invoke `dbt source freshness` 6s

✗ Invoke `dbt build` 12s

✗ Invoke `dbt docs generate` 8s

```
Database Error in model stg_yellow_tripdata (models/staging/stg_yellow_tripdata.sql)

  Access Denied: Table taxi-rides-ny-339813-412521:trips_data_all.yellow_tripdata: User does not
have permission to query table taxi-rides-ny-339813-412521:trips_data_all.yellow_tripdata, or
perhaps it does not exist in location US.

  compiled Code at target/run/taxi_rides_ny/models/staging/stg_yellow_tripdata.sql
```

In my case, I was set up in a different branch, so always check the branch you are working on. Change the 04-analytics-engineering/taxi_rides_ny/models/staging/**schema.yml** file in the

```
sources:

  - name: staging

    database: your_database_name
```

If this error will continue when running dbt job, As for changing the branch for your job, you can use the ‘Custom Branch’ settings in your dbt Cloud environment. This allows you to run your job on a different branch than the default one (usually main). To do this, you need to:

Go to an environment and select Settings to edit it

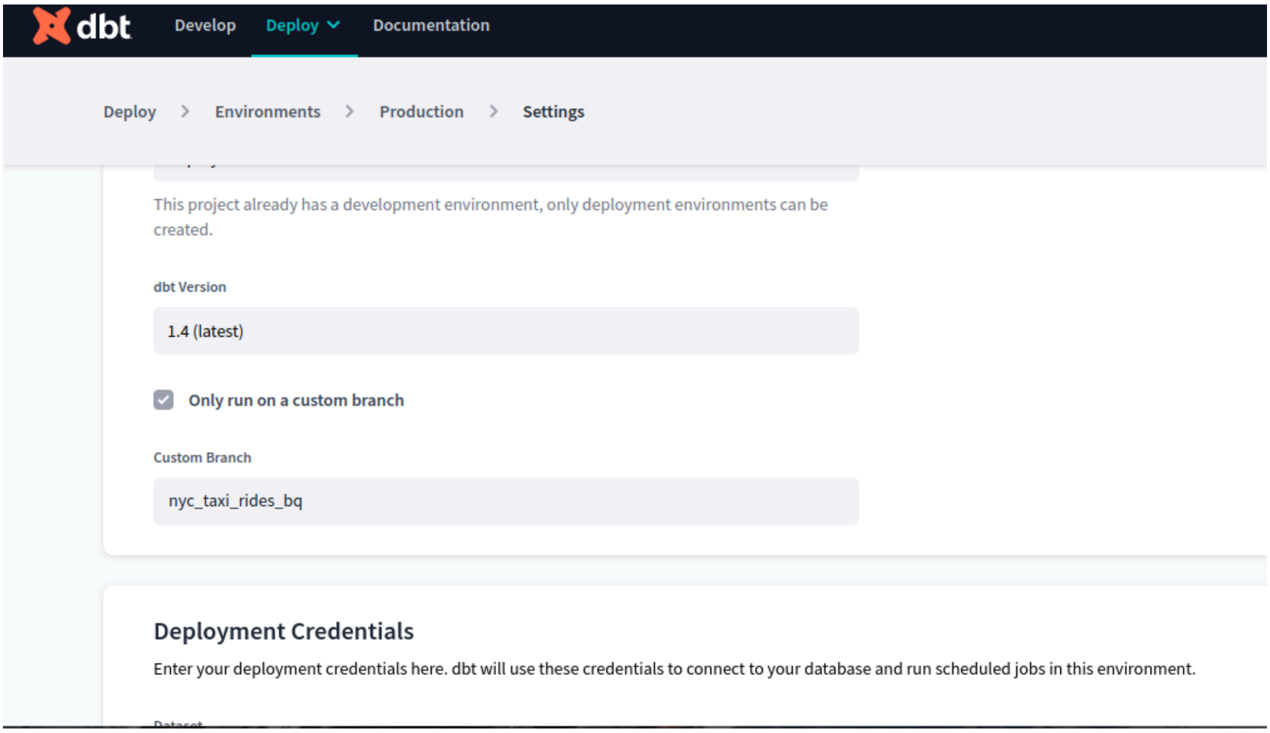
Select Only run on a custom branch in General settings

Enter the name of your custom branch (e.g. HW)

Click Save

Could not parse the dbt project. please check that the repository contains a valid dbt project

Running the Environment on the master branch causes this error, you must activate “Only run on a custom branch” checkbox and specify the branch you are working when Environment is setup.



This project already has a development environment, only deployment environments can be created.

dbt Version: 1.4 (latest)

☑ Only run on a custom branch

Custom Branch: nyc_taxi_rides_bq

Deployment Credentials

Enter your deployment credentials here. dbt will use these credentials to connect to your database and run scheduled jobs in this environment.

Made change to your modelling files and commit the your development branch, but Job still runs on old file?

Change to main branch, make a pull request from the development branch.
Note: this will take you to github.
Approve the merging and rerun you job, it would work as planned now

Setup - I’ve set Github and Bigquery to dbt successfully. Why nothing showed in my Develop tab?

Before you can develop some data model on dbt, you should create development environment and set some parameter on it. After the model being developed, we should also create deployment environment to create and run some jobs.

BigQuery returns an error when i try to run ‘dbt run’:

My taxi data was loaded into gcs with etl_web_to_gcs.py script that converts csv data into parquet. Then I placed raw data trips into external tables and when I executed dbt run I got an error message: Parquet column 'passenger_count' has type INT64 which does not match the target cpp_type DOUBLE. It is because several columns in files have different formats of data.

When I added df[col] = df[col].astype('Int64') transformation to the columns: passenger_count, payment_type, RatecodeID, VendorID, trip_type it went ok. Several people also faced this error and more about it you can read on the slack channel.

DBT - Running dbt run --models stg_green_tripdata --var 'is_test_run: false' is not returning anything:

Use the syntax below instead if the code in the tutorial is not working.

```
dbt run --select stg_green_tripdata --vars '{"is_test_run": false}'
```

DBT - Error: No module named 'pytz' while setting up dbt with docker

Following dbt with BigQuery on Docker readme.md, after `docker-compose build` and `docker-compose run dbt-bq-dtc init`, encountered error `ModuleNotFoundError: No module named 'pytz'`

Solution:

Add `**RUN python -m pip install --no-cache pytz**` in the **Dockerfile** under `FROM --platform=$build_for python:3.9.9-slim-bullseye as base`

VS Code: NoPermissions (FileSystemError): Error: EACCES: permission denied (linux)

If you have problems editing *dbt_project.yml* when using Docker after 'docker-compose run dbt-bq-dtc init', to change profile '`taxi_rides_ny`' to `'bq-dbt-workshop'`, just run:

sudo chown -R username path

DBT - Internal Error: Profile should not be None if loading is completed

When running `dbt debug`, change the directory to the newly created subdirectory (e.g: the newly created `taxi_rides_ny` directory, which contains the dbt project).

Google Cloud BigQuery Location Problems

When running a query on BigQuery sometimes could appear a this table is not on the specified location error.

For this problem there is not a straightforward solution, you need to dig a little, but the problem could be one of these:

- Check the locations of your bucket, datasets and tables. Make sure they are all on the same one.
- Change the query settings to the location you are in: on the query window select more -> query settings -> select the location
- Check if all the paths you are using in your query to your tables are correct: you can click on the table -> details -> and copy the path.

DBT Deploy - This dbt Cloud run was cancelled because a valid dbt project was not found.

1. This happens because we have moved the dbt project to another directory on our repo.
2. Or might be that you’re on a different branch than is expected to be merged from / to.

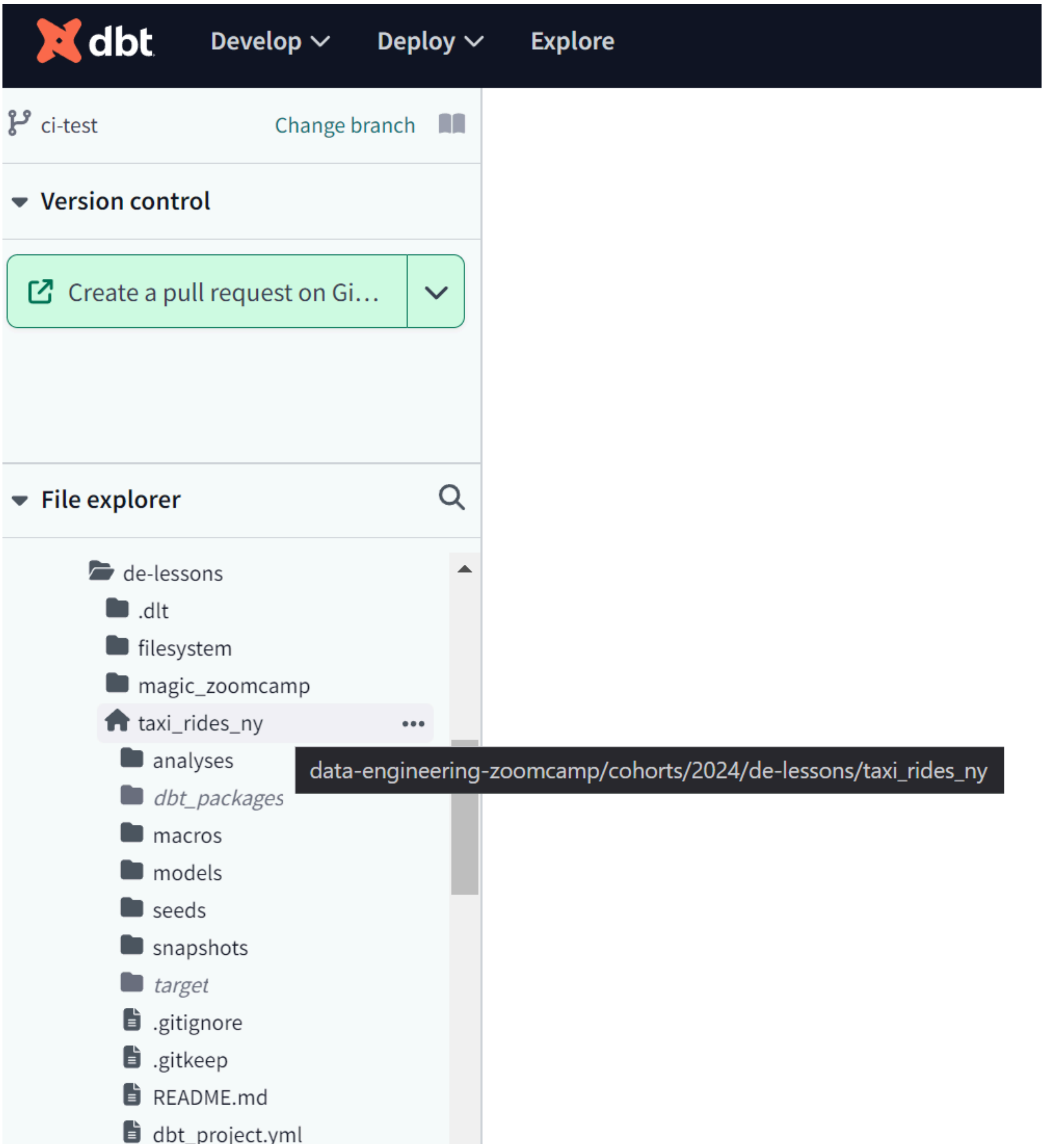
Solution:

Go to the projects window on dbt cloud -> settings -> edit -> and add directory (the extra path to the dbt project)

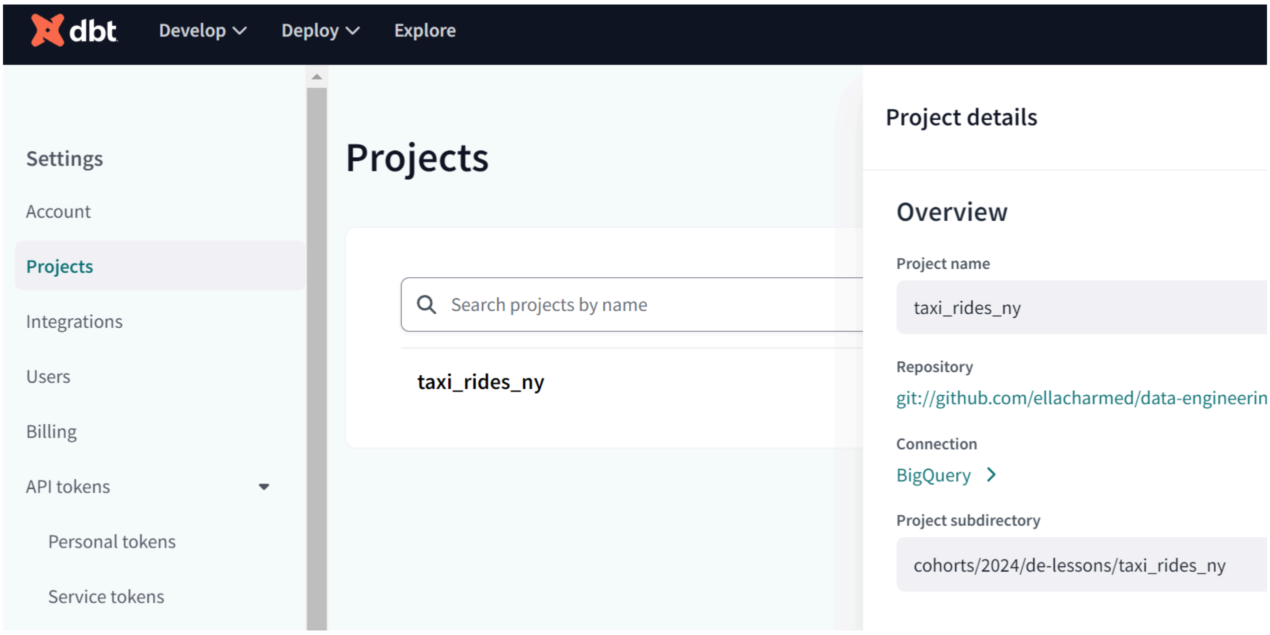
For example:

/week5/taxi_rides_ny

Make sure your file explorer path and this Project settings path matches and there’s no files waiting to be committed to github if you’re running the job to deploy to PROD.



data-engineering-zoomcamp/cohorts/2024/de-lessons/taxi_rides_ny



Projects

taxi_rides_ny

Project details

Overview

Project name

taxi_rides_ny

Repository

git://github.com/ellacharmed/data-engineerin

Connection

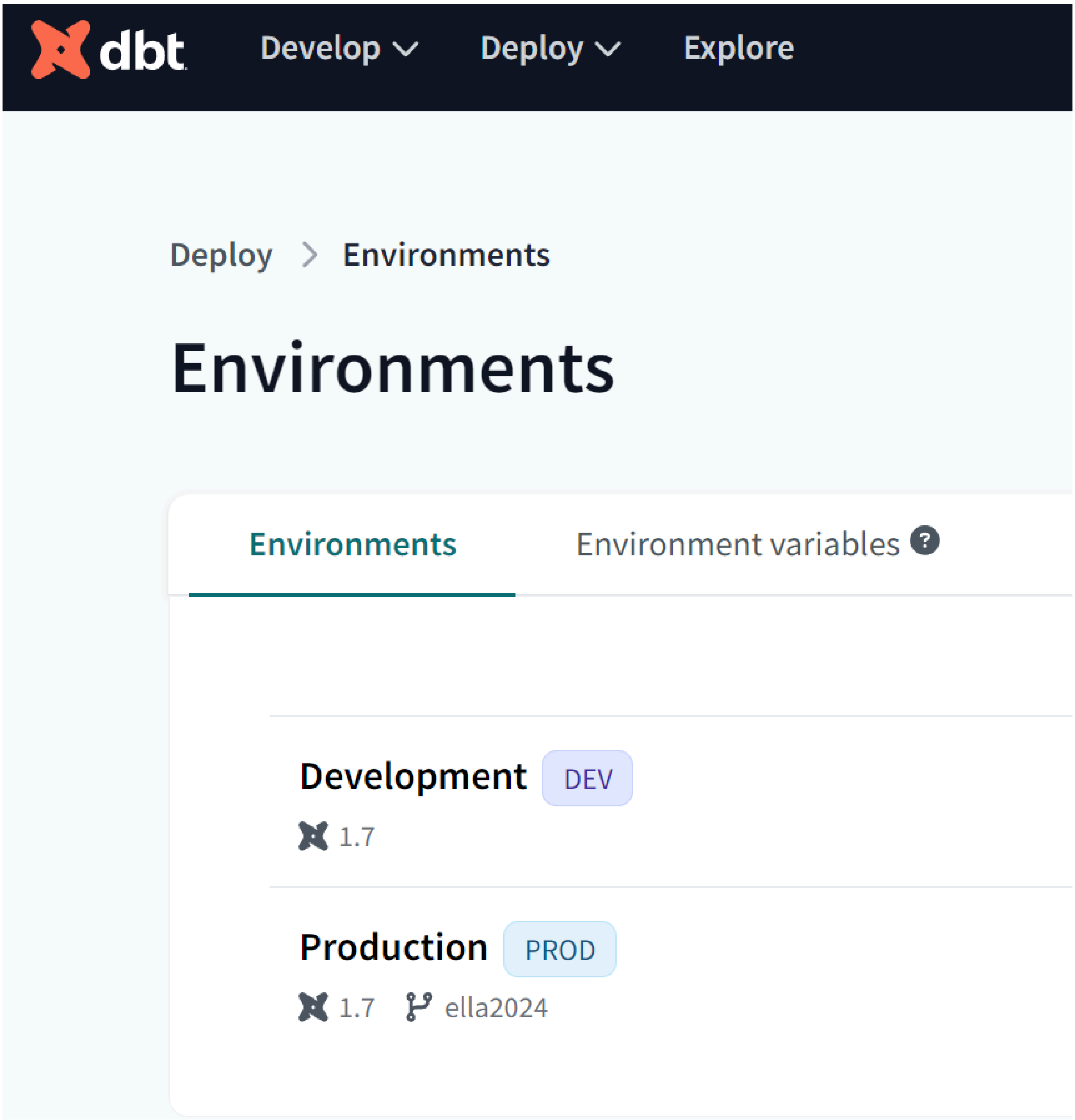
BigQuery

Project subdirectory

cohorts/2024/de-lessons/taxi_rides_ny

And that you had setup the PROD environment to check in the `main` branch, or whichever you specified.

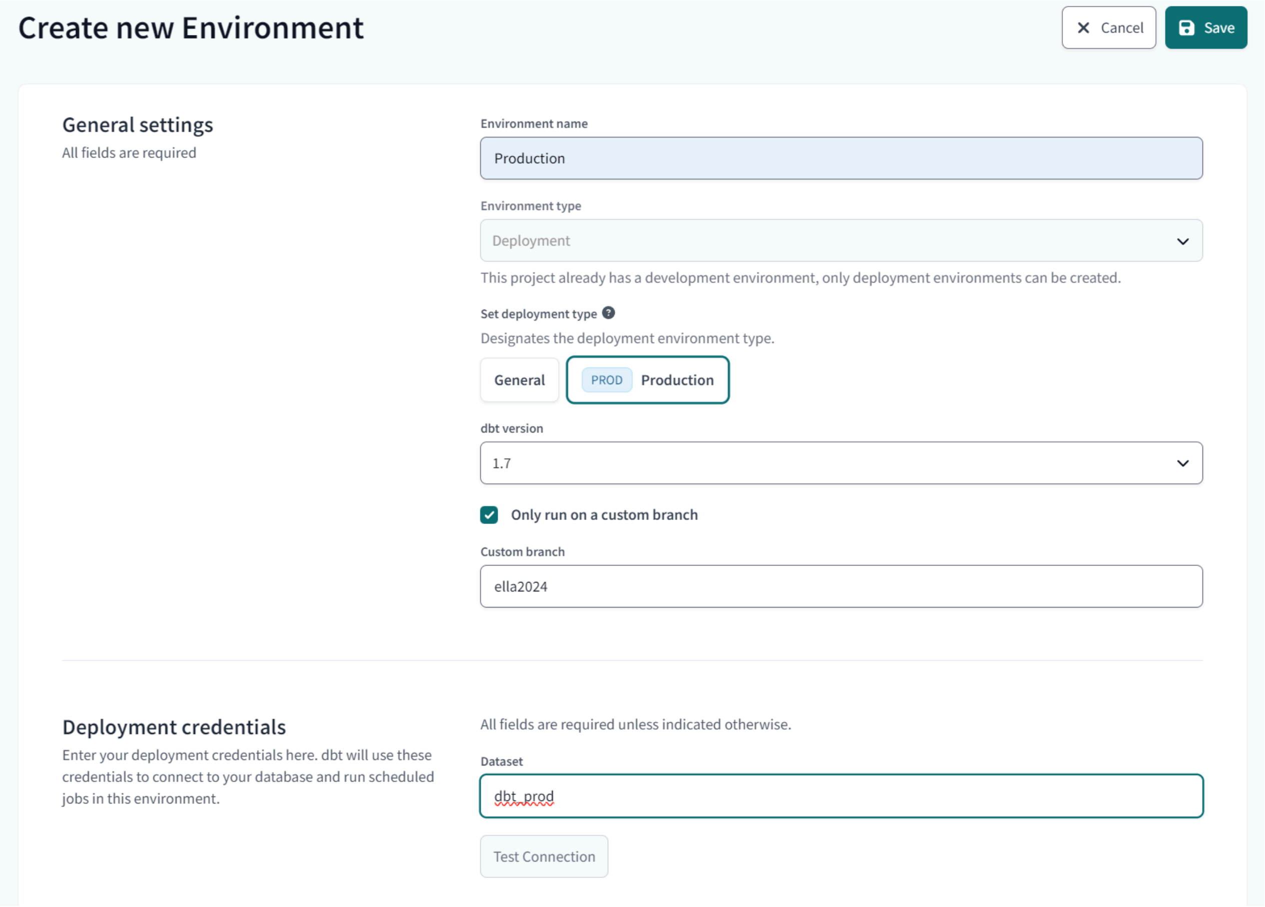
In the picture below, I had set it to `ella2024` to be checked as “production-ready” by the “freshness” check mark at the PROD environment settings. So each time I merge a branch from something else into `ella2024` and then trigger the PR, the CI check job would kick-in. But we still do need to Merge and close the PR manually, I believe, that part is not automated.



Development DEV
1.7

Production PROD
1.7 ella2024

You set up the PROD custom branch (if not default `main`) in the Environment setup screen.


Create new Environment

General settings

All fields are required

Environment name: Production

Environment type: Deployment

This project already has a development environment, only deployment environments can be created.

Set deployment type

Designates the deployment environment type.

dbt version: 1.7

☑ Only run on a custom branch

Custom branch: ella2024

Deployment credentials

Enter your deployment credentials here. dbt will use these credentials to connect to your database and run scheduled jobs in this environment.

All fields are required unless indicated otherwise.

Dataset: dbt_prod

DBT Deploy + CI - Location Problems on BigQuery

When you are creating the pull request and running the CI, dbt is creating a new schema on BIgQuery. By default that new schema will be created on 'US' location, if you have your dataset, schemas and tables on 'EU' that will generate an error and the pull request will not be accepted. To change that location to 'EU' on the connection to BigQuery from dbt we need to add the location 'EU' on the connection optional settings:

Dbt -> project -> settings -> connection BIgQuery -> OPtional Settings -> Location -> EU

DBT Deploy - Error When trying to run the dbt project on Prod

When running trying to run the dbt project on prod there is some things you need to do and check on your own:

- First Make the pull request and Merge the branch into the main.
- Make sure you have the latest version, if you made changes to the repo in another place.
- Check if the dbt_project.yml file is accessible to the project, if not check this solution (Dbt: This dbt Cloud run was cancelled because a valid dbt project was not found.).
- Check if the name you gave to the dataset on BigQuery is the same you put on the dataset spot on the production environment created on dbt cloud.

DBT - Error: “404 Not found: Dataset <dataset_name>:<dbt_schema_name> was not found in location EU” after building from stg_green_tripdata.sql

In the step in this video (DE Zoomcamp 4.3.1 - Build the First dbt Models), after creating `stg_green_tripdata.sql` and clicking `build`, I encountered an error saying dataset not found in location EU. The default location for dbt Bigquery is the US, so when generating the new Bigquery schema for dbt, unless specified, the schema locates in the US.

Solution:
Turns out I forgot to specify **Location** to be `**EU**` when adding connection details.

Develop -> Configure Cloud CLI -> Projects -> taxi_rides_ny -> (connection) Bigquery -> Edit -> Location (Optional) -> type `EU` -> Save

Homework - Ingesting FHV_20?? data

Issue: If you’re having problems loading the FHV_20?? 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. ‘

Ingesting FHV : alternative with kestra

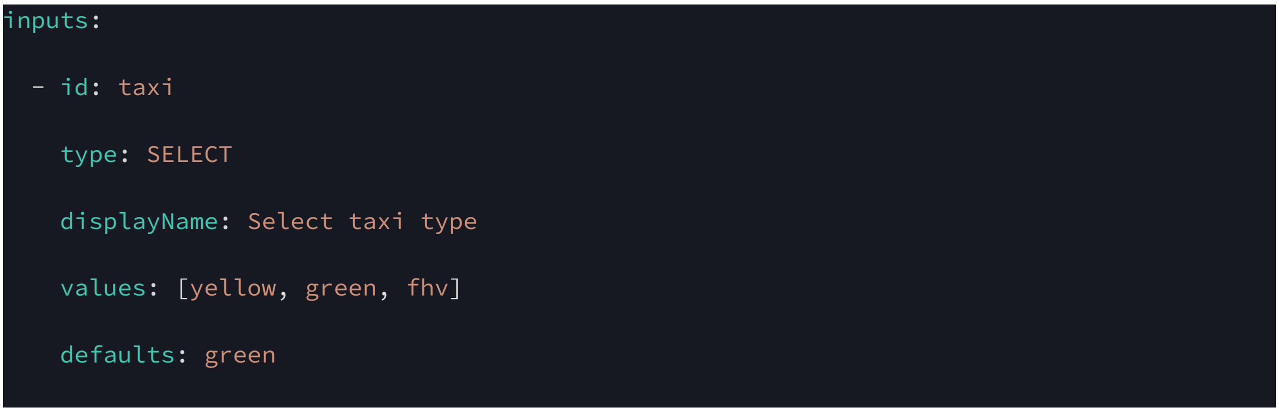
Add this task based on the previous ones :

```
- id: if_fhv_taxi
    type: io.kestra.plugin.core.flow.If
    condition: "{{inputs.taxi == 'fhv'}}"
    then:
      - id: bq_fhv_tripdata
        type: io.kestra.plugin.gcp.bigquery.Query
        sql: |
          CREATE TABLE IF NOT EXISTS `{{kv('GCP_PROJECT_ID')}}.{{kv('GCP_DATASET')}}.fhv_tripdata`
          (
              unique_row_id BYTES OPTIONS (description = 'A unique identifier for the trip, generated by hashing key trip attributes.'),
              filename STRING OPTIONS (description = 'The source filename from which the trip data was loaded.'),
              dispatching_base_num STRING,
              pickup_datetime TIMESTAMP,
              dropoff_datetime TIMESTAMP,
              PUlocationID NUMERIC,
              DOlocationID NUMERIC,
              SR_Flag STRING,
              Affiliated_base_number STRING


          )
          PARTITION BY DATE(pickup_datetime);
      - id: bq_fhv_table_ext
        type: io.kestra.plugin.gcp.bigquery.Query
        sql: |
          CREATE OR REPLACE EXTERNAL TABLE
`{{kv('GCP_PROJECT_ID')}}.{{render(vars.table)}}_ext`
          (
              dispatching_base_num STRING,
              pickup_datetime TIMESTAMP,
              dropoff_datetime TIMESTAMP,
              PUlocationID NUMERIC,
              DOlocationID NUMERIC,
              SR_Flag STRING,
              Affiliated_base_number STRING
          )
          OPTIONS (
              format = 'CSV',
              uris = ['{{render(vars.gcs_file)}}'],
              skip_leading_rows = 1,
              ignore_unknown_values = TRUE
          );

      - id: bq_fhv_table_tmp
        type: io.kestra.plugin.gcp.bigquery.Query
        sql: |
          CREATE OR REPLACE TABLE `{{kv('GCP_PROJECT_ID')}}.{{render(vars.table)}}`
          AS
          SELECT
            MD5(CONCAT(
              COALESCE(CAST(pickup_datetime AS STRING), ""),
              COALESCE(CAST(dropoff_datetime AS STRING), ""),
              COALESCE(CAST(PUlocationID AS STRING), ""),
              COALESCE(CAST(DOLocationID AS STRING), "")
            )) AS unique_row_id,
            "{{render(vars.file)}}" AS filename,
            *
          FROM `{{kv('GCP_PROJECT_ID')}}.{{render(vars.table)}}_ext`;


      - id: bq_fhv_merge
        type: io.kestra.plugin.gcp.bigquery.Query
        sql: |
          MERGE INTO `{{kv('GCP_PROJECT_ID')}}.{{kv('GCP_DATASET')}}.fhv_tripdata` T
          USING `{{kv('GCP_PROJECT_ID')}}.{{render(vars.table)}}` S
          ON T.unique_row_id = S.unique_row_id
          WHEN NOT MATCHED THEN
            INSERT (unique_row_id, filename, dispatching_base_num, pickup_datetime, dropoff_datetime, PUlocationID, DOlocationID, SR_Flag, Affiliated_base_number)
            VALUES (S.unique_row_id, S.filename, S.dispatching_base_num, S.pickup_datetime, S.dropoff_datetime, S.PUlocationID, S.DOlocationID, S.SR_Flag, S.Affiliated_base_number);
```

Add a trigger too :

```
- id: fhv_schedule
    type: io.kestra.plugin.core.trigger.Schedule
    cron: "0 11 1 * *"
    inputs:
      taxi: fhv
```

And modify inputs :



```
inputs:

  - id: taxi

    type: SELECT

    displayName: Select taxi type

    values: [yellow, green, fhv]

    defaults: green
```

Homework - Ingesting NYC TLC Data

I found out that the easies way to upload datasets form github for the homework is utilising this script git_csv_to_gcs.py. Thank you Lidia!!
It is similar to a script that Alexey provided us in
03-data-warehouse/extras/**web_to_gcs.py**

How to set environment variable easily for any credentials

If you have to securely put your credentials for a project and, probably, push it to a git repository then the best option is to use an environment variable
For example for **web_to_gcs.py** or **git_csv_to_gcs.py** we have to set these variables:
GOOGLE_APPLICATION_CREDENTIALS
GCP_GCS_BUCKET
The easises option to do it is to use .env (dotenv).
Install it and add a few lines of code that inject these variables for your project

```
pip install python-dotenv

from dotenv import load_dotenv
import os

# Load environment variables from .env file
load_dotenv()

# Now you can access environment variables like GCP_GCS_BUCKET and
GOOGLE_APPLICATION_CREDENTIALS
credentials_path = os.getenv("GOOGLE_APPLICATION_CREDENTIALS")
BUCKET = os.environ.get("GCP_GCS_BUCKET")
```

Invalid date types after Ingesting FHV data through CSV files: Could not parse 'pickup_datetime' as a timestamp

If you uploaded manually the fvh 2019 csv files, you may face errors regarding date types. Try to create an the external table in bigquery but define the pickup_datetime and dropoff_datetime to be strings

```
CREATE OR REPLACE EXTERNAL TABLE `gcp_project.trips_data_all.fhv_tripdata`  (
    dispatching_base_num STRING,
    pickup_datetime STRING,
    dropoff_datetime STRING,
    PUlocationID STRING,
    DOlocationID STRING,
    SR_Flag STRING,
    Affiliated_base_number STRING
)
OPTIONS (
    format = 'csv',
    uris = ['gs://bucket/*.csv']
);
```

Then when creating the fhv core model in dbt, use TIMESTAMP(CAST(()) to ensure it first parses as a string and then convert it to timestamp.

```
with fhv_tripdata as (
    select * from {{ ref('stg_fhv_tripdata') }}
),
dim_zones as (
    select * from {{ ref('dim_zones') }}
    where borough != 'Unknown'
)
select fhv_tripdata.dispatching_base_num,
    TIMESTAMP(CAST(fhv_tripdata.pickup_datetime AS STRING)) AS pickup_datetime,
    TIMESTAMP(CAST(fhv_tripdata.dropoff_datetime AS STRING)) AS dropoff_datetime,
```

Invalid data types after Ingesting FHV data through parquet files: Could not parse SR_Flag as Float64,Couldn’t parse datetime column as timestamp,couldn’t handle NULL values in PULocationID,DOLocationID

If you uploaded manually the fvh 2019 parquet files manually after downloading from `https://d37ci6vzurychx.cloudfront.net/trip-data/fhv_tripdata_2019-*.parquet` you may face errors regarding date types while loading the data in a landing table (say fhv_tripdata). Try to create an the external table with the schema defines as following and load each month in a loop.

```
-----Correct load with schema defination----will not throw error----------------------
CREATE OR REPLACE EXTERNAL TABLE
`dw-bigquery-week-3.trips_data_all.external_tlc_fhv_trips_2019` (
    dispatching_base_num STRING,
    pickup_datetime TIMESTAMP,
    dropoff_datetime TIMESTAMP,
    PUlocationID FLOAT64,
    DOlocationID FLOAT64,
    SR_Flag FLOAT64,
    Affiliated_base_number STRING
)
OPTIONS (
  format = 'PARQUET',
  uris = ['gs://project id/fhv_2019_8.parquet']
);
Can Also USE  uris = ['gs://project id/fhv_2019_*.parquet'] (THIS WILL remove the need
for the loop and can be done for all month in single RUN )
```

– THANKYOU FOR THIS –

Join Error on LocationID “Unable to find common supertype for templated argument”

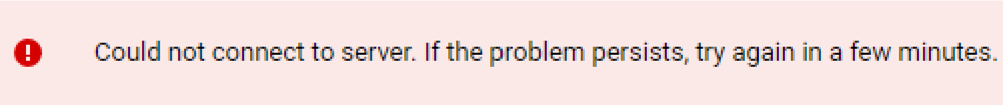
```
No matching signature for operator = for argument types: STRING, INT64
    Signature: T1 = T1
      Unable to find common supertype for templated argument
```

Make sure the LocationID field is in the same type. If it is in string format in one table, we can use the following code in dbt to convert it to integer:

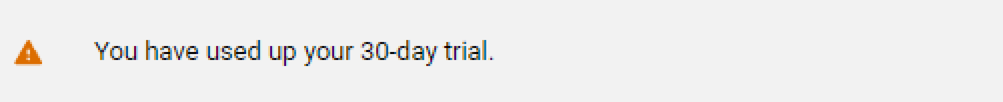
```
{{ dbt.safe_cast("PULocationID", api.Column.translate_type("integer")) }} as
pickup_locationid
```

Google Looker Studio - you have used up your 30-day trial

When accessing Looker Studio through the Google Cloud Project console, you may be prompted to subscribe to the Pro version and receive the following errors:



Could not connect to server. If the problem persists, try again in a few minutes.



You have used up your 30-day trial.

Instead, navigate to https://lookerstudio.google.com/navigation/reporting which will take you to the free version.

How does dbt handle dependencies between models?

Ans: Dbt provides a mechanism called "ref" to manage dependencies between models. By referencing other models using the "ref" keyword in SQL, dbt automatically understands the dependencies and ensures the correct execution order.

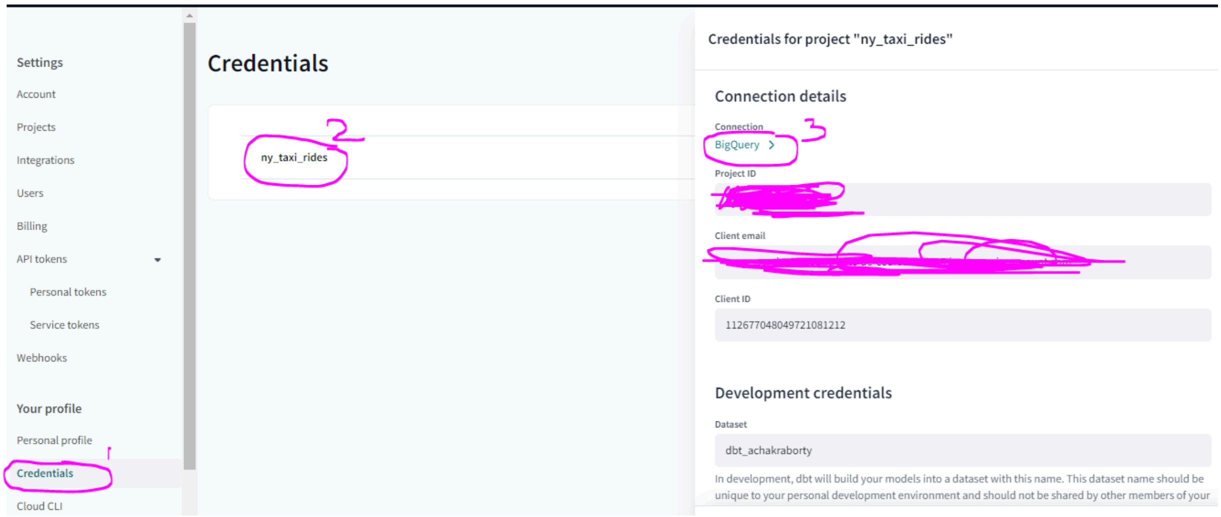
Loading FHV Data goes into slumber using Mage?

Try loading the data using jupyter notebooks in a local environment. There might be bandwidth issues with Mage.

Load the data into a pandas dataframe using the urls, make necessary transformations, upload the gcp bucket / alternatively download the parquet/csv files locally and then upload to GCP manually.

Region Mismatch in DBT and BigQuery

If you are using the datasets copied into BigQuery from BigQuery public datasets, the region will be set as US by default and hence it is much easier to set your dbt profile location as US while transforming the tables and views.
You can change the location as follows:



Credentials for project "ny_taxi_rides"

Connection details

Connection: BigQuery >

Project ID: [redacted]

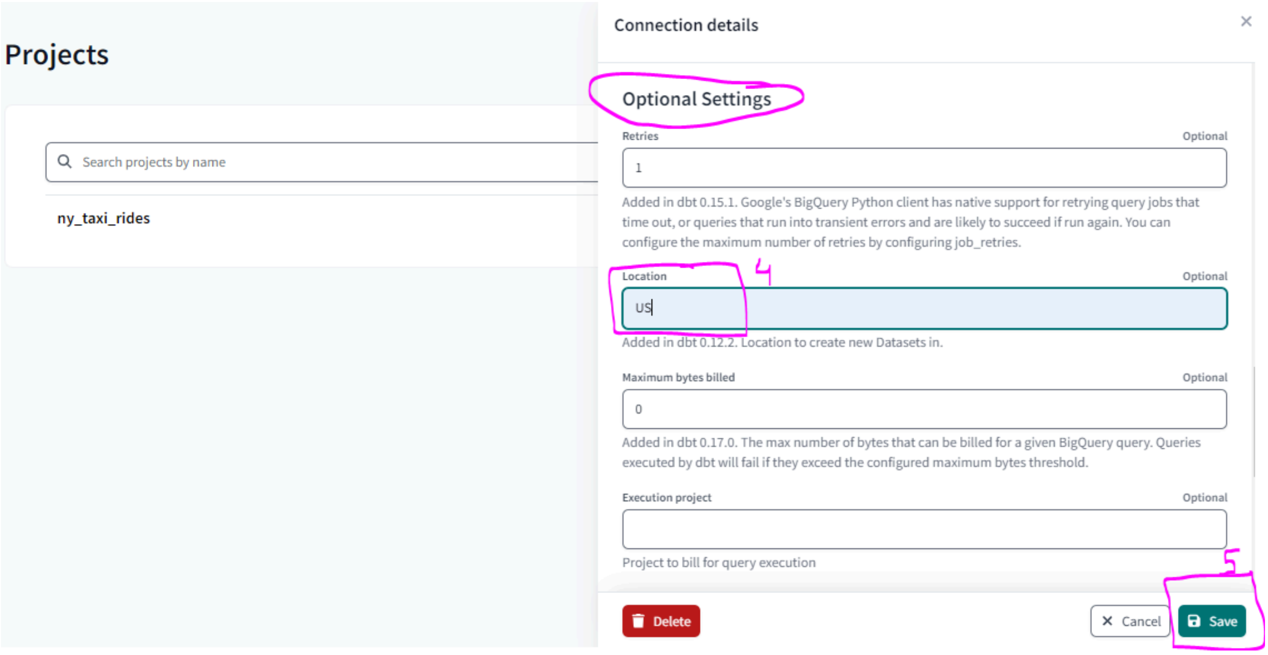
Client email: [redacted]

Client ID: 112677048049721081212

Development credentials

Dataset: dbt_achakraborty

In development, dbt will build your models into a dataset with this name. This dataset name should be unique to your personal development environment and should not be shared by other members of your



Connection details

Optional Settings

Retries — Optional: 1

Added in dbt 0.15.1. Google's BigQuery Python client has native support for retrying query jobs that time out, or queries that run into transient errors and are likely to succeed if run again. You can configure the maximum number of retries by configuring job_retries.

4 Location — Optional: us

Added in dbt 0.12.2. Location to create new Datasets in.

Maximum bytes billed — Optional: 0

Added in dbt 0.17.0. The max number of bytes that can be billed for a given BigQuery query. Queries executed by dbt will fail if they exceed the configured maximum bytes threshold.

Execution project — Optional:

Project to bill for query execution

5 Save

What is the fastest way to upload taxi data to dbt-postgres?

Use the PostgreSQL COPY FROM feature that is compatible with csv files

First create the table like (as an example):

```
CREATE TABLE taxis (
…
);
```

And then use copy functionality (as an example):

```
COPY taxis FROM PROGRAM
'url'
WITH (
 FORMAT csv,
 HEADER true,
```

ENCODING utf8

);

```
COPY table_name [ ( column_name [, ...] ) ]
FROM { 'filename' | PROGRAM 'command' | STDIN }
[ [ WITH ] ( option [, ...] ) ]
[ WHERE condition ]
```

dbt - Where should we create `profiles.yml` ?

For local environment i.e. dbt-core, the profile configuration is valid for all projects. Note: dbt Cloud doesn't require it.

The ~/.dbt/profiles.yml file should be located in your user's home directory. On Windows, this would typically be:

C:\Users\<YourUsername>\.dbt\profiles.yml

Replace <YourUsername> with your actual Windows username. This file is used by dbt to store connection profiles for different projects.

Here's how you can create the profiles.yml file in the appropriate directory:

1. Open File Explorer and navigate to C:\Users\<YourUsername>\.
2. Create a new folder named .dbt if it doesn't already exist.
3. Inside the .dbt folder, create a new file named profiles.yml.

Usage example can be found here.

dbt - Are there UI for dbt Core like dbt Cloud?

- Second only to dbt Cloud functionality: https://github.com/AltimateAI/vscode-dbt-power-user Sign up for the community plan for free usage at Altimate and add the API into your VS Code extension.
- VSCode Snippets Package for dbt and Jinja functions in SQL, YAML, and Markdown: https://github.com/bastienboutonnet/vscode-dbt
- For monitoring purposes: https://github.com/elementary-data/elementary Read more here.

When configuring the profiles.yml file for dbt-postgres with jinja templates with environment variables, I'm getting "Credentials in profile "PROFILE_NAME", target: 'dev', invalid: '5432'is not of type 'integer'

```
dbt_postgres_analytics:
  outputs:
    dev:
      type: postgres
      host:   "{{ env_var('DBT_POSTGRES_HOST', 'localhost') }}"
      port:   "{{ env_var('DBT_POSTGRES_PORT', 5432) }}"
      dbname: "{{ env_var('DBT_POSTGRES_DATABASE') }}"
      schema: "{{ env_var('DBT_POSTGRES_TARGET_SCHEMA') }}"
      user:   "{{ env_var('DBT_POSTGRES_USER') }}"
      pass:   "{{ env_var('DBT_POSTGRES_PASSWORD') }}"
      threads: 4
```

Update the line:

```
port:   "{{ env_var('DBT_POSTGRES_PORT', 5432) }}"
```

With:

```
port:   "{{ env_var('DBT_POSTGRES_PORT', 5432) | as_number }}"
```

DBT - The database is correct but I get Error with Incorrect Schema in Models

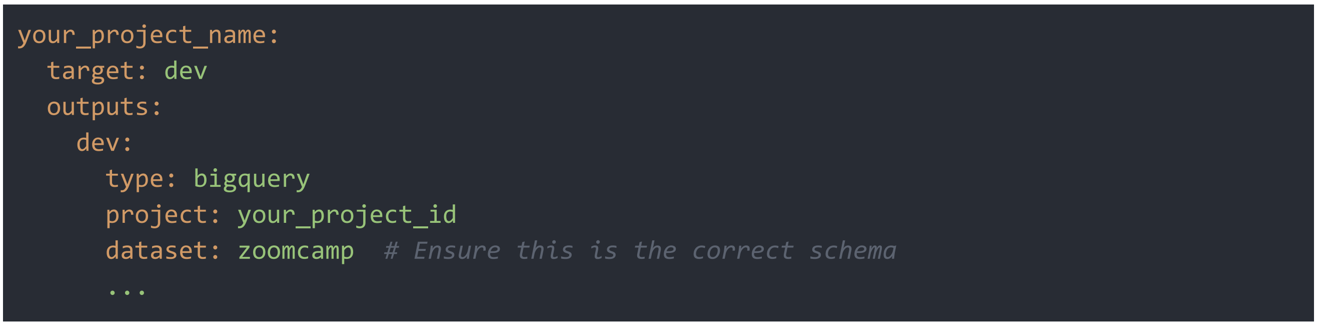
What to do if your dbt model fails with an error similar to:

```
Database Error in model <model_name> Not found: Dataset <dataset_name>
was not found in location <location_id>
```

1. DBT-CORE

- **Check** `profiles.yml`:
    - Ensure your `profiles.yml` file is correctly configured with the correct schema and database under your target. This file is typically located in `~/.dbt/`.

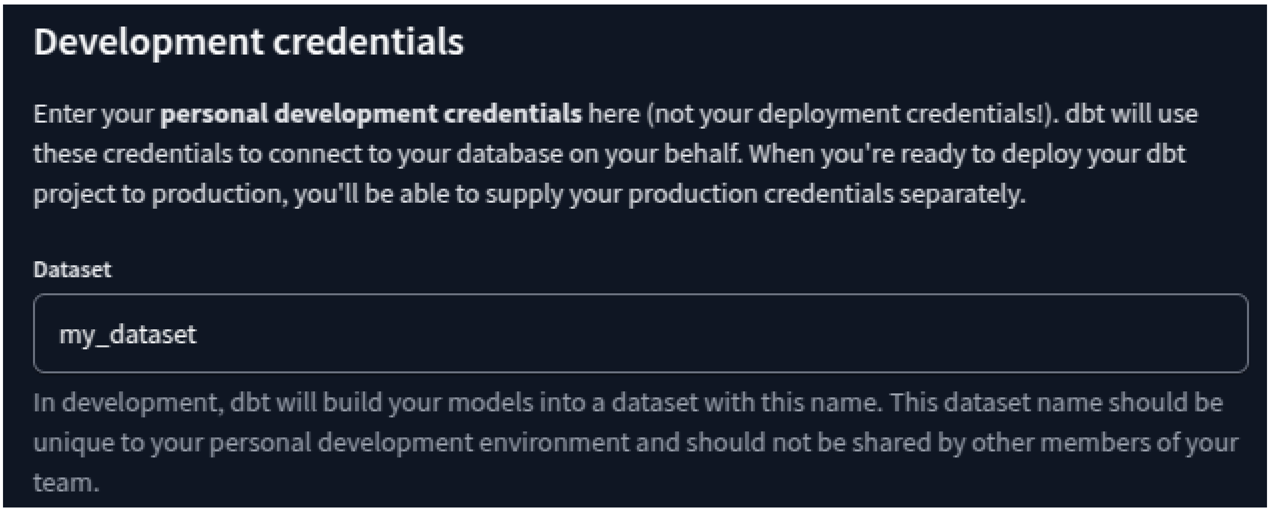
      Example configuration:



```
your_project_name:
  target: dev
  outputs:
    dev:
      type: bigquery
      project: your_project_id
      dataset: zoomcamp  # Ensure this is the correct schema
      ...
```

2. DBT-CLOUD-IDE

- **Check Credentials in dbt Cloud UI:**
    - Navigate to the Credentials section in the dbt Cloud project settings.
    - Ensure the correct database and schema are set (e.g., ‘my_dataset’).



Development credentials

Enter your **personal development credentials** here (not your deployment credentials!). dbt will use these credentials to connect to your database on your behalf. When you're ready to deploy your dbt project to production, you'll be able to supply your production credentials separately.

Dataset

my_dataset

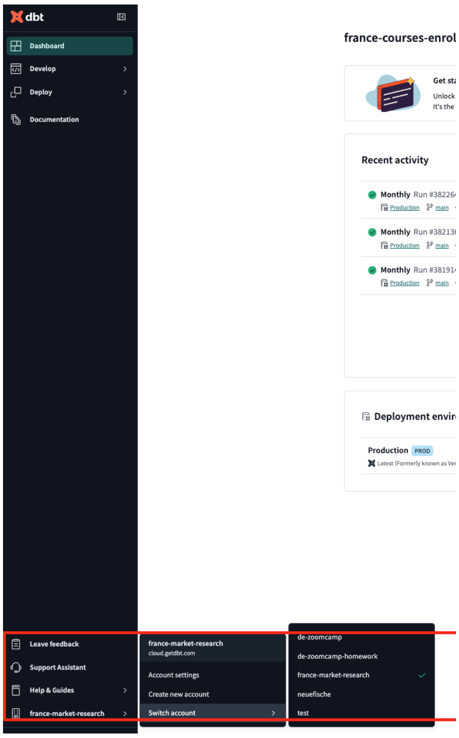
In development, dbt will build your models into a dataset with this name. This dataset name should be unique to your personal development environment and should not be shared by other members of your team.

- **Verify Environment Settings:**
    - Double-check that you are working in the correct environment (dev, prod, etc.), as dbt Cloud allows different settings for different environments.
- **No Need for** `profiles.yml`**:**

- ○ In dbt Cloud, you don’t need to configure `profiles.yml` manually. All connection settings are handled via the UI.

DBT allows only 1 project in free developer version.

Yes, DBT allows only 1 project under one account. But you can create multiple accounts as shown below:



Documentation or book sign not shown even after doing dbt docs generate.

In the free version, it does not show the docs when models are run in development environment. Create a production job and tick generate docs section. Execute it and it will generate the documentation.

Module 5: pyspark

Setting up Java and Spark (with PySpark) on Linux (Alternative option using SDKMAN)

1. Install SDKMAN:

   ```
   curl -s "https://get.sdkman.io" | bash
   source "$HOME/.sdkman/bin/sdkman-init.sh"
   ```

2. Using SDKMAN, install Java 11 and Spark 3.3.2:

   ```
   sdk install java 11.0.22-tem
   sdk install spark 3.3.2
   ```

   Open a new terminal or run the following in the same shell:

   ```
   source "$HOME/.sdkman/bin/sdkman-init.sh"
   ```

3. Verify the locations and versions of Java and Spark that were installed:

   ```
   echo $JAVA_HOME
   java -version
   echo $SPARK_HOME
   spark-submit --version
   ```

PySpark - Setting Spark up in Google Colab

If you're seriously struggling to set things up "locally" (here locally meaning non/partly-managed environment like own laptop, a VM or Codespaces) you can use the following guide to use Spark in Google Colab:

https://medium.com/gitconnected/launch-spark-on-google-colab-and-connect-to-sparkui-342cad19b304

Starter notebook:

https://github.com/aaalexlit/medium_articles/blob/main/Spark_in_Colab.ipynb

It's advisable to spend some time setting things up locally rather than jumping right into this solution.

Spark-shell: unable to load native-hadoop library for platform - Windows

If after installing Java (either jdk or openjdk), Hadoop and Spark, and setting the corresponding environment variables you find the following error when spark-shell is run at CMD:

```
java.lang.IllegalAccessError: class
org.apache.spark.storage.StorageUtils$ (in unnamed module
@0x3c947bc5) cannot access class sun.nio.ch.DirectBuffer (in
module java.base) because module java.base does not export
sun.nio.ch to unnamed
module @0x3c947bc5
```

Solution: Java 17 or 19 is not supported by Spark. Spark 3.x: requires Java 8/11/16. Install Java 11 from the website provided in the windows.md setup file.

PySpark - Python was not found; run without arguments to install from the Microsoft Store, or disable this shortcut from Settings > Manage App Execution Aliases.

I found this error while executing the user defined function in Spark (crazy_stuff_udf). I am working on Windows and using conda. After following the setup instructions, I found that the PYSPARK_PYTHON environment variable was not set correctly, given that conda has different python paths for each environment.

Solution:

- `pip install findspark` on the command line inside proper environment
- Add to the top of the script

  ```
  import findspark

  findspark.init()
  ```

PySpark - TypeError: code() argument 13 must be str, not int , while executing `import pyspark` (Windows/ Spark 3.0.3 - Python 3.11)

This is because Python 3.11 has some inconsistencies with such an old version of Spark. The solution is a downgrade in the Python version. Python 3.9 using a conda environment takes care of it. Or install newer PySpark >= 3.5.1 works for me (Ella) [source].

Import pyspark - Error: No Module named 'pyspark'

Ensure that your `PYTHONPATH` is set correctly to include the PySpark library. You can check if PySpark is pointing to the correct location by running:

import pyspark

print(pyspark.__file__)

It should point to the location where PySpark is installed (e.g., `/home/<your username>/spark/spark-3.x.x-bin-hadoop3.x/python/pyspark/__init__.py`)

Cannot find Spark jobs UI at localhost

This is because current port is in use, Spark UI will run on a different port. You can check which port Spark is using by running this command:

spark.sparkContext.uiWebUrl

If it indicates a different port, you should access that specific port instead. Additionally, ensure that there are no other notebooks or processes that might be using the same port. Clean up unused resources to avoid port conflicts.

Java+Spark - Easy setup with miniconda env (worked on MacOS)

If anyone is a Pythonista or becoming one (which you will essentially be one along this journey), and desires to have all python dependencies under same virtual environment (e.g. conda) as done with prefect and previous exercises, simply follow these steps

1. Install OpenJDK 11,

   a. on MacOS: `$ brew install java11`
   b. Add `export PATH="/opt/homebrew/opt/openjdk@11/bin:$PATH"` to `~/.bashrc` or `~/zshrc`
2. Activate working environment (by pipenv / poetry / conda)
3. Run `$ pip install pyspark`
4. Work with exercises as normal

All default commands of spark will be also available at shell session under activated enviroment.

Hope this can help!

P.s. you won't need findspark to firstly initialize.

Py4J - Py4JJavaError: An error occurred while calling (...) java.net.ConnectException: Connection refused: no further information;

If you're getting `Py4JavaError` with a generic root cause, such as the described above (Connection refused: no further information). You're most likely using incompatible versions of the JDK or Python with Spark.

As of the current latest Spark version (3.5.0), it supports JDK 8 / 11 / 17. All of which can be easily installed with SDKMan! on macOS or Linux environments

```
$ sdk install java 17.0.10-librca
$ sdk install spark 3.5.0
$ sdk install hadoop 3.3.5py4j
```

As PySpark 3.5.0 supports Python 3.8+ make sure you're setting up your virtualenv with either 3.8 / 3.9 / 3.10 / 3.11 (Most importantly avoid using 3.12 for now as not all libs in the data-science/engineering ecosystem are fully package for that)

```
$ conda create -n ENV_NAME python=3.11
$ conda activate ENV_NAME
$ pip install pyspark==3.5.0
```

This setup makes installing `findspark` and the likes of it unnecessary. Happy coding.

Py4J - `Py4JJavaError:` `An error occurred while calling o54.parquet. Or any kind of` **`Py4JJavaError that show up after run df.write.parquet('zones')(On window)`**

This assume you already correctly set up the PATH in the nano ~/.bashrc

Here my

```
export JAVA_HOME="/c/tools/jdk-11.0.21"
export PATH="${JAVA_HOME}/bin:${PATH}"

export HADOOP_HOME="/c/tools/hadoop-3.2.0"
export PATH="${HADOOP_HOME}/bin:${PATH}"

export SPARK_HOME="/c/tools/spark-3.3.2-bin-hadoop3"
export PATH="${SPARK_HOME}/bin:${PATH}"

export PYTHONPATH="${SPARK_HOME}/python/:$PYTHONPATH"
export PYTHONPATH="${SPARK_HOME}spark-3.5.1-bin-hadoop3py4j-0.10.9.5-src.zip:$PYTHONPATH"
```

You also need to add environment variables correctly which paths to java jdk, spark and hadoop through

Go to Stephenlaye2/winutils3.3.0: winutils.exe hadoop.dll and hdfs.dll binaries for hadoop windows (github.com), download the right winutils for hadoop-3.2.0. Then create a new folder,bin and put every thing in side to make a /c/tools/hadoop-3.2.0/bin(You might not need to do this, but after testing it without the /bin I could not make it to work)

Then follow the solution in this video: How To Resolve Issue with Writing DataFrame to Local File | winutils | msvcp100.dll (youtube.com)

Remember to restart IDE and computer, After the error `An error occurred while calling o54.parquet. is` fixed but new errors like `o31.parquet.` `Or o35.parquet.` appear.

Spark - Installation Error Code 1603

Issue: Spark installation on Windows completed but failed to run.

This is a common Windows Installer error code indicating that there was a fatal error during installation. It often occurs due to issues like insufficient permissions, conflicts with other software, or problems with the installer package.

Step to solve the issue:

Installing Chocolatey

Chocolatey is a package manager for Windows, which makes it easy to install, update, and manage software.

Installation Steps

1. **Open PowerShell as an Administrator**

   - Press `Win + X` and select `Windows PowerShell (Admin)` or search for `PowerShell`, right-click, and select `Run as administrator`.

2. **Run the following command to install Chocolatey**

```
 Set-ExecutionPolicy Bypass -Scope Process -Force;
[System.Net.ServicePointManager]::SecurityProtocol =
[System.Net.ServicePointManager]::SecurityProtocol -bor 3072;
iex ((New-Object
System.Net.WebClient).DownloadString('<https://community.chocolatey.org/install.ps1>'))
```

3. **Verify the installation**

   - Close and reopen PowerShell as an administrator and run:

```
choco -v
```

  - You should see the Chocolatey version number indicating that it has been installed successfully.

Command for Global Acceptance

To globally accept all licenses for all packages installed using Chocolatey, run the following command:

```
choco feature enable -n allowGlobalConfirmation
```

This command configures Chocolatey to automatically accept license agreements for all packages, streamlining the installation process and avoiding prompts for each package.

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 you .bashrc file and that there are not any other exports which might supersede the ones provided in the course content.

To add to the solution above, if the errors persist in regards to setting the correct path for spark, an alternative solution for permanent path setting solve the error is to set environment variables on system and user environment variables following this tutorial:

- Once everything is installed, skip to 7:14 to set up environment variables. This allows for the environment variables to be set permanently.

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 contex<<t as usual

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

Filter based on conditions based on multiple columns

```
from pyspark.sql.functions import col

new_final.filter((new_final.a_zone=="Murray Hill") &
(new_final.b_zone=="Midwood")).show()
```

Krishna Anand

Py4JJavaError - ModuleNotFoundError: No module named 'py4j'` while executing `import pyspark`

You need to look for the Py4J file and note the version of the filename. Once you know the version, you can update the export command accordingly, this is how you check yours:
` ls ${SPARK_HOME}/python/lib/ ` and then you add it in the export command, mine was:

```
export
PYTHONPATH=”${SPARK_HOME}/python/lib/Py4J-0.10.9.5-src.zip:${PYTHO
NPATH}”
```

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.

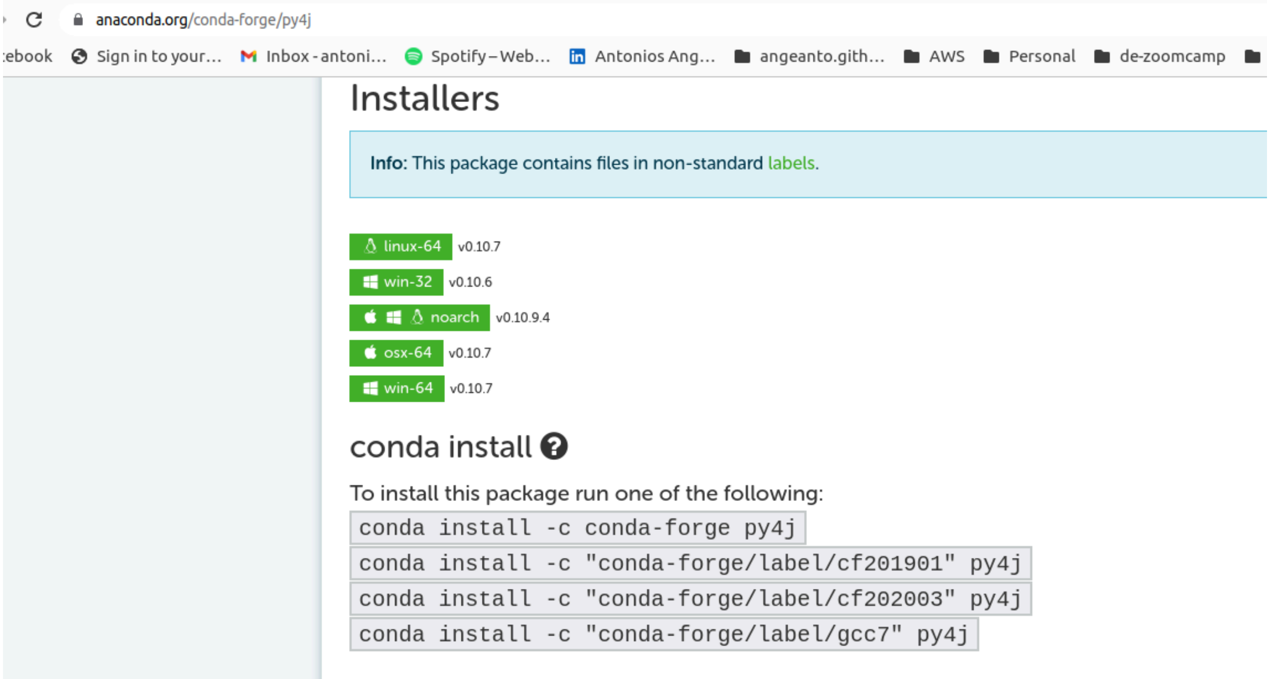
~ Abhijit Chakraborty: Sometimes, even with adding the correct version of py4j might not solve the problem. Simply run `pip install py4j` and problem should be resolved.

Py4J Error - ModuleNotFoundError: No module named 'py4j' (Solve with latest version)

If below does not work, then download the latest available py4j version with

```
conda install -c conda-forge py4j
```

Take care of the latest version number in the website to replace appropriately.



Installers

Info: This package contains files in non-standard labels.

linux-64	v0.10.7
win-32	v0.10.6
noarch	v0.10.9.4
osx-64	v0.10.7
win-64	v0.10.7

conda install

To install this package run one of the following:

```
conda install -c conda-forge py4j
conda install -c "conda-forge/label/cf201901" py4j
conda install -c "conda-forge/label/cf202003" py4j
conda install -c "conda-forge/label/gcc7" py4j
```

Now add

```
export PYTHONPATH="${SPARK_HOME}/python/:$PYTHONPATH"
```

```
export
PYTHONPATH="${SPARK_HOME}/python/lib/py4j-0.10.9.7-src.zip:$PYTHONPATH"
```

in your .bashrc file.

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 Noteboopgak 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

Following 5.2.1, I am getting an error - Head:cannot open ‘taxi+_zone_lookup.csv’ for reading: No such file or directory

The latest filename is just ‘taxi_zone_lookup.sv’ so it should work after removing the ‘+’ now.

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-bdcb0831f2b7-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 directorydf

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

Hadoop - FileNotFoundException: Hadoop bin directory does not exist , when trying to write (Windows)

You need to create the Hadoop `/bin` directory manually and add the downloaded files in there, since the shell script provided for Windows installation just puts them in `/c/tools/hadoop-3.2.0/` .

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 a port is in use, then Spark uses the next available port number. It can be even 4044. Clean up after yourself when a port does not work or a container does not run.

You can run `spark.sparkContext.uiWebUrl`

and result will be some like
'http://172.19.10.61:4041'

Java - 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.

Java - 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/

https://twitter.com/drkrishnaanand/status/1765423415878463839

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

✅I got it working using `` `gcs-connector-hadoop-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.j
ars.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 : 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-hadoop
3-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.enab
le", "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!

```
start-slave.sh: command not found
```

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

Remove white spaces from column names in Pyspark

```
df_finalx=df_finalw.select([col(x).alias(x.replace(" ","")) for x in df_finalw.columns])
```

Krishna Anand

AttributeError: 'DataFrame' object has no attribute 'iteritems'

This error comes up on the Spark video 5.3.1 - First Look at Spark/PySpark,

because as at the creation of the video, 2021 data was the most recent which utilised csv files but as at now its parquet.

So when you run the command `spark.createDataFrame(df1_pandas).show(),`

You get the Attribute error. This is caused by the pandas version 2.0.0 which seems incompatible with Spark 3.3.2, so to fix it you have to downgrade pandas to 1.5.3 using the command **pip install -U pandas==1.5.3**

Another option is adding the following after importing pandas, if one does not want to downgrade pandas version (source) :

pd.DataFrame.iteritems = pd.DataFrame.items

Note that this problem is solved with Spark versions from 3.4.1

AttributeError: 'DataFrame' object has no attribute 'iteritems'

Another alternative is to install pandas 2.0.1 (it worked well as at the time of writing this), and it is compatible with Pyspark 3.5.1. Make sure to add or edit your environment variable like this:

```
export SPARK_HOME="${HOME}/spark/spark-3.5.1-bin-hadoop3"
 export PATH="${SPARK_HOME}/bin:${PATH}"
```

Spark Standalone Mode on Windows

- Open a CMD terminal in administrator mode
- cd %SPARK_HOME%
- Start a master node: `bin\spark-class org.apache.spark.deploy.master.Master`

- Start a worker node: `bin\spark-class org.apache.spark.deploy.worker.Worker spark://<master_ip>:<port> --host <IP_ADDR>`
-
- `bin/spark-class org.apache.spark.deploy.worker.Worker` spark://localhost:7077 `--host <IP_ADDR>`
    - `spark://<master_ip>:<port>:` copy the address from the previous command, in my case it was spark://localhost:7077
    - Use `--host <IP_ADDR>` if you want to run the worker on a different machine. For now leave it empty.
- Now you can access Spark UI through localhost:8080

Export PYTHONPATH command in linux is temporary

You can either type the export command every time you run a new session, add it to the .bashrc/ which you can find in /home or run this command at the beginning of your homebook:

```
import findspark

findspark.init()
```

Compression Error: zcat output is gibberish, seems like still compressed

In the code along from Video 5.3.3 Alexey downloads the CSV files from the NYT website and gzips them in their bash script. If we now (2023) follow along but download the data from the GH course Repo, it will already be zippes as csv.gz files. Therefore we zip it again if we follow the code from the video exactly. This then leads to gibberish outcome when we then try to cat the contents or count the lines with zcat, because the file is zipped twitch and zcat only unzips it once.

✅solution: do not gzip the files downloaded from the course repo. Just wget them and save them as they are as csv.gz files. Then the zcat command and the showSchema command will also work

```
URL="${URL_PREFIX}/${TAXI_TYPE}/${TAXI_TYPE}_tripdata_${YEAR}-${FMONTH}.csv.gz"
   LOCAL_PREFIX="data/raw/${TAXI_TYPE}/${YEAR}/${FMONTH}"
   LOCAL_FILE="${TAXI_TYPE}_tripdata_${YEAR}_${FMONTH}.csv.gz"
   LOCAL_PATH="${LOCAL_PREFIX}/${LOCAL_FILE}"

   echo "downloading ${URL} to ${LOCAL_PATH}"
   mkdir -p ${LOCAL_PREFIX}
   wget ${URL} -O ${LOCAL_PATH}

   echo "compressing ${LOCAL_PATH}"
   # gzip ${LOCAL_PATH} <- uncomment this line
```

PicklingError: Could not serialise object: IndexError: tuple index out of range.

Occurred while running : `spark.createDataFrame(df_pandas).show()`

This error is usually due to the python version, since spark till date of 2 march 2023 doesn't support python 3.11, try creating a new env with python version 3.8 and then run this command.

On the virtual machine, you can create a conda environment (here called myenv) with python 3.10 installed:

```
conda create -n myenv python=3.10 anaconda
```

Then you must run `conda activate myenv` to run python 3.10. Otherwise you'll still be running version 3.11. You can deactivate by typing `conda deactivate`.

Connecting from local Spark to GCS - Spark does not find my google credentials as shown in the video?

Make sure you have your credentials of your GCP in your VM under the location defined in the script.

Spark docker-compose setup

To run spark in docker setup

1. Build bitnami spark docker

a. clone bitnami repo using command

```
git clone https://github.com/bitnami/containers.git
```

(tested on commit 9cef8b892d29c04f8a271a644341c8222790c992)

b. edit file `bitnami/spark/3.3/debian-11/Dockerfile` and update java and spark version as following

```
        "python-3.10.10-2-linux-${OS_ARCH}-debian-11" \
        "java-17.0.5-8-3-linux-${OS_ARCH}-debian-11" \
```

reference: https://github.com/bitnami/containers/issues/13409

c. build docker image by navigating to above directory and running docker build command

navigate `cd bitnami/spark/3.3/debian-11/`

build command `docker build -t spark:3.3-java-17 .`

2. run docker compose

using following file

```yaml docker-compose.yml
version: '2'

services:
  spark:
    image: spark:3.3-java-17
    environment:
    - SPARK_MODE=master
    - SPARK_RPC_AUTHENTICATION_ENABLED=no
    - SPARK_RPC_ENCRYPTION_ENABLED=no
    - SPARK_LOCAL_STORAGE_ENCRYPTION_ENABLED=no
```

```
        - SPARK_SSL_ENABLED=no
        volumes:
        - "./:/home/jovyan/work:rw"
        ports:
        - '8080:8080'
        - '7077:7077'
  spark-worker:
        image: spark:3.3-java-17
        environment:
        - SPARK_MODE=worker
        - SPARK_MASTER_URL=spark://spark:7077
        - SPARK_WORKER_MEMORY=1G
        - SPARK_WORKER_CORES=1
        - SPARK_RPC_AUTHENTICATION_ENABLED=no
        - SPARK_RPC_ENCRYPTION_ENABLED=no
        - SPARK_LOCAL_STORAGE_ENCRYPTION_ENABLED=no
        - SPARK_SSL_ENABLED=no
        volumes:
        - "./:/home/jovyan/work:rw"
        ports:
        - '8081:8081'
  spark-nb:
        image: jupyter/pyspark-notebook:java-17.0.5
        environment:
        - SPARK_MASTER_URL=spark://spark:7077
        volumes:
        - "./:/home/jovyan/work:rw"
```

~~~
        ports:
        - '8888:8888'
        - '4040:4040'
```
~~~

run command to deploy docker compose

```
docker-compose up
```

Access jupyter notebook using link logged in docker compose logs

Spark master url is `spark://spark:7077`

How do you read data stored in gcs on pandas with your local computer?

To do this
```
pip install gcsfs,
```

Thereafter copy the uri path to the file and use
```
df = pandas.read_csc(gs://path)
```

TypeError when using spark.createDataFrame function on a pandas df

Error:

```
spark.createDataFrame(df_pandas).schema
TypeError: field Affiliated_base_number: Can not merge type <class 'pyspark.sql.types.StringType'> and <class 'pyspark.sql.types.DoubleType'>
```

Solution:

`Affiliated_base_number` is a mix of letters and numbers (you can check this with a preview of the table), so it cannot be set to *DoubleType* (only for double-precision numbers). The suitable type would be *StringType*. Spark `inferSchema` is more accurate than Pandas infer type method in this case. You can set it to `true` while reading the csv, so you don’t have to take out any data from your dataset. Something like this can help:

```
df = spark.read \
    .options(
    header = "true", \
    inferSchema = "true", \
          ) \
    .csv('path/to/your/csv/file/')
```

Solution B:

It's because some rows in the affiliated_base_number are null and therefore it is assigned the datatype String and this cannot be converted to type Double. So if you really want to convert this pandas df to a pyspark df only take the rows from the pandas df that are not null in the 'Affiliated_base_number' column. Then you will be able to apply the pyspark function createDataFrame.

```
# Only take rows that have no null values
pandas_df= pandas_df[pandas_df.notnull().all(1)]
```

MemoryManager: Total allocation exceeds 95.00% (1,020,054,720 bytes) of heap memory

Default executor memory is 1gb. This error appeared when working with the homework dataset.

```
Error: MemoryManager: Total allocation exceeds 95.00%
(1,020,054,720 bytes) of heap memory
Scaling row group sizes to 95.00% for 8 writers
```

Solution:

Increase the memory of the executor when creating the Spark session like this:

```
spark = SparkSession.builder \
      .master("local[*]") \
      .appName('test') \
      .config("spark.executor.memory", "4g") \
      .config("spark.driver.memory", "4g") \
      .getOrCreate()
```

Remember to restart the Jupyter session (ie. close the Spark session) or the config won’t take effect.

How to spark standalone cluster is run on windows OS

Change the working directory to the spark directory:

if you have setup up your SPARK_HOME variable, use the following;

```
cd %SPARK_HOME%
```

if not, use the following;

```
cd <path to spark installation>
```

Creating a Local Spark Cluster

To start Spark Master:

```
bin\spark-class org.apache.spark.deploy.master.Master --host localhost
```

Starting up a cluster:

```
bin\spark-class org.apache.spark.deploy.worker.Worker spark://localhost:7077 --host localhost
```

Env variables set in ~/.bashrc are not loaded to Jupyter in VS Code

I added PYTHONPATH, JAVA_HOME and SPARK_HOME to `~/.bashrc`, `import pyspark` worked ok in iPython in terminal, but couldn’t be found in .ipynb opened in VS Code

After adding new lines to `~/.bashrc`, need to **restart** the shell to activate the new lines, do either

- `source ~/.bashrc`
- `exec bash`

Instead of configuring paths in `~/.bashrc`, I created .env file in the root of my workspace:

```
JAVA_HOME="${HOME}/app/java/jdk-11.0.2"
PATH="${JAVA_HOME}/bin:${PATH}"
SPARK_HOME="${HOME}/app/spark/spark-3.3.2-bin-hadoop3"
PATH="${SPARK_HOME}/bin:${PATH}"
PYTHONPATH="${SPARK_HOME}/python/:$PYTHONPATH"
PYTHONPATH="${SPARK_HOME}/python/lib/py4j-0.10.9.5-src.zip:$PYTHON
PATH"
PYTHONPATH="${SPARK_HOME}/python/lib/pyspark.zip:$PYTHONPATH"
```

hadoop “wc -l” is giving a different result then shown in the video

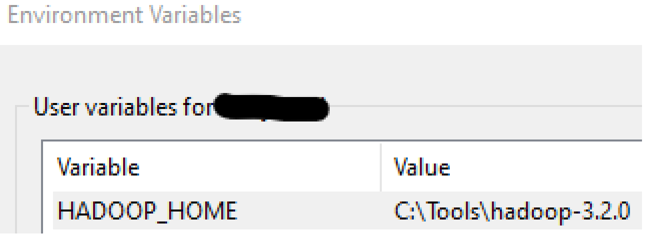
If you are doing `wc -l fhvhv_tripdata_2021-01.csv.gz` with the gzip file as the file argument, you will get a different result, obviously! Since the file is compressed.

Unzip the file and then do `wc -l fhvhv_tripdata_2021-01.csv` to get the right results.

Hadoop - Exception in thread "main" java.lang.UnsatisfiedLinkError: org.apache.hadoop.io.nativeio.NativeIO$Windows.access0(Ljava/lang/String;I)Z

If you are seeing this (or similar) error when attempting to write to parquet, it is likely an issue with your path variables.

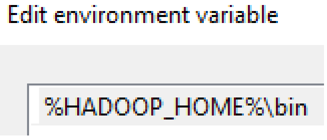
For Windows, create a new User Variable “HADOOP_HOME” that points to your Hadoop directory. Then add “%HADOOP_HOME%\bin” to the PATH variable.



Environment Variables

User variables for [redacted]

Variable	Value
HADOOP_HOME	C:\Tools\hadoop-3.2.0



Edit environment variable

%HADOOP_HOME%\bin

Additional tips can be found here: https://stackoverflow.com/questions/41851066/exception-in-thread-main-java-lang-unsatisfiedlinkerror-org-apache-hadoop-io

Java.io.IOException. Cannot run program “C:\hadoop\bin\winutils.exe”. CreateProcess error=216, This version of 1% is not compatible with the version of Windows you are using.

Change the hadoop version to 3.0.1.Replace all the files in the local hadoop bin folder with the files in this repo: winutils/hadoop-3.0.1/bin at master · cdarlint/winutils (github.com)

If this does not work try to change other versions found in this repository.

For more information please see this link: This version of %1 is not compatible with the version of Windows you're running · Issue #20 · cdarlint/winutils (github.com)

Dataproc - ERROR: (gcloud.dataproc.jobs.submit.pyspark) The required property [project] is not currently set. It can be set on a per-command basis by re-running your command with the [--project] flag.

Fix is to set the flag like the error states. Get your project ID from your dashboard and set it like so:

```
gcloud dataproc jobs submit pyspark \
     --cluster=my_cluster \
     --region=us-central1 \
     --project=my-dtc-project-1010101 \
     gs://my-dtc-bucket-id/code/06_spark_sql.py
     -- \
          …
```

Run Local Cluster Spark in Windows 10 with CMD

1. Go to `%SPARK_HOME%\bin`
2. Run `spark-class org.apache.spark.deploy.master.Master` to run the master. This will give you a URL of the form `spark://ip:port`
3. Run `spark-class org.apache.spark.deploy.worker.Worker spark://ip:port` to run the worker. Make sure you use the URL you obtained in step 2.
4. Create a new Jupyter notebook:

```
spark = SparkSession.builder \
    .master("spark://{ip}:7077") \
    .appName('test') \
    .getOrCreate()
```

5. Check on Spark UI the master, worker and app.

lServiceException: 401 Anonymous caller does not have storage.objects.list access to the Google Cloud Storage bucket. Permission 'storage.objects.list' denied on resource (or it may not exist).

This occurs because you are not logged in “gcloud auth login” and maybe the project id is not settled. Then type in a terminal:

```
gcloud auth login
```

This will open a tab in the browser, accept the terms, after that close the tab if you want. Then set the project is like:

```
gcloud config set project <YOUR PROJECT_ID>
```

Then you can run the command to upload the pq dir to a GCS Bucket:

```
gsutil -m cp -r pq/ <YOUR URI from gsutil>/pq
```

py4j.protocol.Py4JJavaError GCP

When submit a job, it might throw an error about Java in log panel within Dataproc. I changed the Versioning Control when I created a cluster, so it means that I delete the cluster and created a new one, and instead of choosing Debian-Hadoop-Spark, I switch to Ubuntu 20.02-Hadoop3.3-Spark3.3 for Versioning Control feature, the main reason to choose this is because I have the same Ubuntu version in mi laptop, I tried to find documentation to sustent this but unfortunately I couldn't nevertheless it works for me.

Repartition the Dataframe to 6 partitions using df.repartition(6) - got 8 partitions instead

Use both repartition and coalesce, like so:

```
df = df.repartition(6)
df = df.coalesce(6)
df.write.parquet('fhv/2019/10', mode='overwrite')
```

Jupyter Notebook or SparkUI not loading properly at localhost after port forwarding from VS code?

Possible solution - Try to forward the port using ssh cli instead of vs code.

Run > “`ssh -L <local port>:<VM host/ip>:<VM port> <ssh hostname>`”

ssh hostname is the name you specified in the ~/.ssh/config file

In case of Jupyter Notebook run

“`ssh -L 8888:localhost:8888 gcp-vm`”

from your local machine’s cli.

NOTE: If you logout from the session, the connection would break. Also while creating the spark session notice the block's log because sometimes it fails to run at 4040 and then switches to 4041.

~Abhijit Chakrborty: If you are having trouble accessing localhost ports from GCP VM consider adding the forwarding instructions to .ssh/config file as following:

```
Host <hostname>

    Hostname <external-gcp-ip>

    User xxxx

    IdentityFile yyyy

    LocalForward 8888 localhost:8888

    LocalForward 8080 localhost:8080

    LocalForward 5432 localhost:5432
    LocalForward 4040 localhost:4040
```

This should automatically forward all ports and will enable accessing localhost ports.

Installing Java 11 on codespaces

~ Abhijit Chakraborty

`sdk list java` to check for available java sdk versions.

`sdk install java 11.0.22-amzn` as java-11.0.22-amzn was available for my codespace.

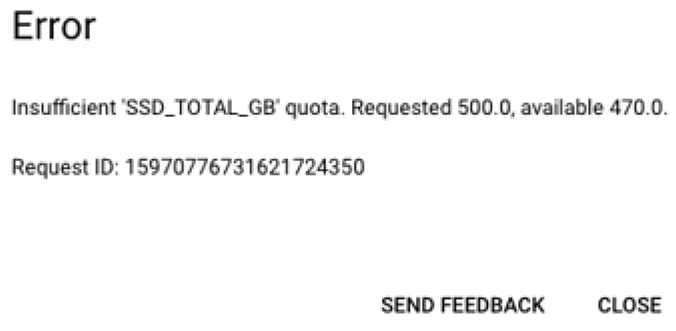
click on Y if prompted to change the default java version.

Check for java version using `java -version `.

If working fine great, else `sdk default java 11.0.22-amzn` or whatever version you have installed.

Error: Insufficient 'SSD_TOTAL_GB' quota. Requested 500.0, available 470.0.

Sometimes while creating a dataproc cluster on GCP, the following error is encountered.



Error

Insufficient 'SSD_TOTAL_GB' quota. Requested 500.0, available 470.0.

Request ID: 15970776731621724350

Solution: As mentioned here, sometimes there might not be enough resources in the given region to allocate the request. Usually, gets freed up in a bit and one can create a cluster. – abhirup ghosh

Solution 2: Changing the type of boot-disk from PD-Balanced to PD-Standard, in terraform, helped solve the problem.- Sundara Kumar Padmanabhan

Homework - how to convert the time difference of two timestamps to hours

Pyspark converts the difference of two TimestampType values to Python's native datetime.timedelta object. The timedelta object only stores the duration in terms of days, seconds, and microseconds. Each of the three units of time must be manually converted into hours in order to express the total duration between the two timestamps using only hours.

Another way for achieving this is using the **datediff** (sql function). It receives this parameters

- Upper Date: the closest date you have. For example dropoff_datetime
- Lower Date: the farthest date you have. For example pickup_datetime

And the result is returned in terms of days, so you could multiply the result for 24 in order to get the hours.

PicklingError: Could not serialize object: IndexError: tuple index out of range

This version combination worked for me:

PySpark = 3.3.2
Pandas = 1.5.3

If it still has an error,

Py4JJavaError: An error occurred while calling o180.showString. : org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 6.0 failed 1 times, most recent failure: Lost task 0.0 in stage 6.0 (TID 6) (host.docker.internal executor driver): org.apache.spark.SparkException: Python worker failed to connect back.

Run this before SparkSession

```
import os
import sys
os.environ['PYSPARK_PYTHON'] = sys.executable
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable
```

RuntimeError: Python in worker has different version 3.11 than that in driver 3.10, PySpark cannot run with different minor versions. Please check environment variables PYSPARK_PYTHON and PYSPARK_DRIVER_PYTHON are correctly set.

```
import os
import sys
os.environ['PYSPARK_PYTHON'] = sys.executable
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable
```

Dataproc Pricing: https://cloud.google.com/dataproc/pricing#on_gke_pricing

Dataproc Qn: Is it essential to have a VM on GCP for running Dataproc and submitting jobs ?

Ans: No, you can submit a job to DataProc from your local computer by installing gsutil (https://cloud.google.com/storage/docs/gsutil_install) and configuring it. Then, you can execute the following command from your local computer.

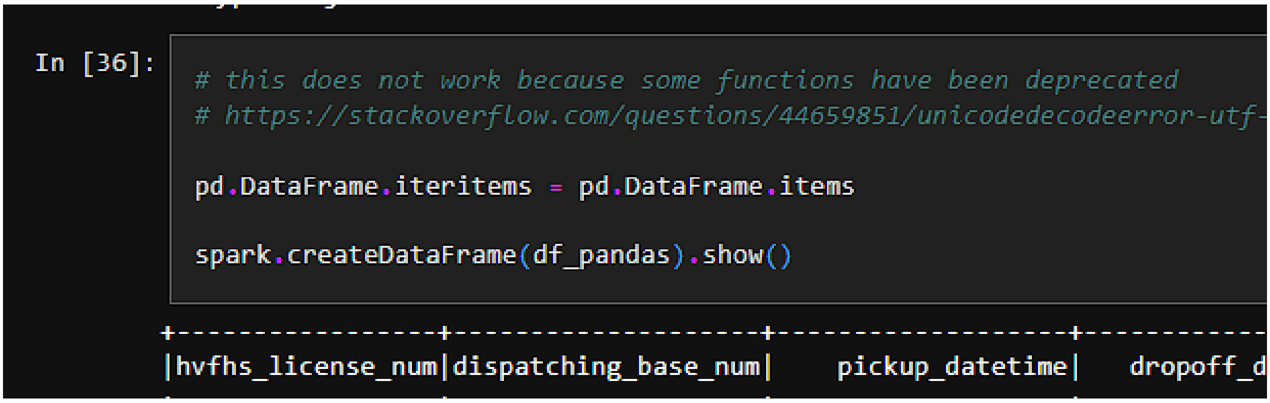
```
gcloud dataproc jobs submit pyspark \
    --cluster=de-zoomcamp-cluster \
    --region=europe-west6 \
    gs://dtc_data_lake_de-zoomcamp-nytaxi/code/06_spark_sql.py \
    -- \

--input_green=gs://dtc_data_lake_de-zoomcamp-nytaxi/pq/green/2020/
*/ \

--input_yellow=gs://dtc_data_lake_de-zoomcamp-nytaxi/pq/yellow/202
0/*/ \
      --output=gs://dtc_data_lake_de-zoomcamp-nytaxi/report-2020
(edited)
```

In module 5.3.1, trying to run spark.createDataFrame(df_pandas).show() returns error

AttributeError: 'DataFrame' object has no attribute 'iteritems' this is because the method inside the pyspark refers to a package that has been already deprecated

(https://stackoverflow.com/questions/76404811/attributeerror-dataframe-object-has-no-attribute-iteritems)

You can do this code below, which is mentioned in the stackoverflow link above:



In [36]:

```
# this does not work because some functions have been deprecated
# https://stackoverflow.com/questions/44659851/unicodedecodeerror-utf-

pd.DataFrame.iteritems = pd.DataFrame.items

spark.createDataFrame(df_pandas).show()
```

```
+-----------------+-------------------+-------------------+-----------
|hvfhs_license_num|dispatching_base_num|     pickup_datetime|    dropoff_d
```

Another work around here is to create a conda enviroment to donwgrade python’s version to 3.8 and pandas to 1.5.3

```
conda create -n pyspark_env python=3.8 pandas=1.5.3 jupyter

conda activate pyspark_env
```

Cannot create a cluster: Insufficient 'SSD_TOTAL_GB' quota. Requested 500.0, available 250.0.

A: The master and worker nodes are allocated a maximum of 250 GB of memory combined. In the configuration section, adhere to the following specifications:

Master Node:

Machine type: n2-standard-2

Primary disk size: 85 GB

Worker Node:

Number of worker nodes: 2

Machine type: n2-standard-2

Primary disk size: 80 GB

You can allocate up to 82.5 GB memory for worker nodes, keeping in mind that the total memory allocated across all nodes cannot exceed 250 GB.

Setting JAVA_HOME with Homebrew on Apple Silicon

The MacOS setup instruction (https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/05-batch/setup/macos.md#installing-java) for setting the ***JAVA_HOME*** environment variable is for Intel-based Macs which have a default install location at `/usr/local/`. If you have an Apple Silicon mac, you will have to set ***JAVA_HOME*** to `/opt/homebrew/`, specifically in your `.bashrc` or `.zshrc`:

```
export JAVA_HOME="/opt/homebrew/opt/openjdk/bin"
```

```
export PATH="$JAVA_HOME:$PATH"
```

Confirm that your path was correctly set by running the command: `which java`

You should expect to see the output:

```
/opt/homebrew/opt/openjdk/bin/java
```

```
Check java version with the next command:
```

```
Java -version
```

Reference: https://docs.brew.sh/Installation

Subnetwork 'default' does not support Private Google Access which is required for Dataproc clusters when 'internal_ip_only' is set to 'true'. Enable Private Google Access on subnetwork 'default' or set 'internal_ip_only' to 'false'.

Search for VPC in Google Cloud Console

Navigate to the second tab “SUBNETS IN CURRENT PROJECT”

Look for the region/location where your dataproc will be located and click on it

Click the edit button and toggle on the button for “Private Google Access”

Save changes.

Spark is working, however, nothing appears in the Spark UI (e.g., .show())?

Since we used multiple notebooks during the course, it's possible that there are more than one Spark session active. It’s highly likely that you are observing the incorrect one. Follow these steps to troubleshoot:

- Spark uses port **4040** by default, but if more than one session is active, it will try to use the next available port (e.g., **4041**).
- Ensure you're viewing the correct **Spark Web UI** for the application where your jobs are running.
- Verify the **current application session address**:
  - Eg: Using `spark.sparkContext.uiWebUrl` command in your session.
  - Expected output: `http://your.application.session.address.internal:4041`
  - Indicating **port** `4041`
- If using a VM, make sure you forward the identified port to access the web ui on the `localhost:<port>`.

Module 6: streaming with kafka

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

Check Docker Compose File:

Ensure that your docker-compose.yaml file is correctly configured with the necessary details for the "control-center" service. Check the service name, image name, ports, volumes, environment variables, and any other configurations required for the container to start.

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
```

```
Verify Python Version:
```

```
Make sure you are using a compatible version of Python with the Avro library. Check the Python version and compatibility requirements specified by the Avro library documentation.
```

... 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`

What is the use of Redpanda ?

Redpanda: Redpanda is built on top of the Raft consensus algorithm and is designed as a high-performance, low-latency alternative to Kafka. It uses a log-centric architecture similar to Kafka but with different underlying principles.

Negsignal:SIGKILL while converting data files to parquet format

Got this error because the docker container memory was exhausted. The data file was up to 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.

resources/rides.csv is missing

Copy the file found in the Java example:
data-engineering-zoomcamp/week_6_stream_processing/java/kafka_examples/src/main/resources/rides.csv

Kafka - python videos have low audio and hard to follow up

tip:As the videos have low audio so I downloaded them and used VLC media player with putting the audio to the max 200% of original audio and the audio became quite good or try to use auto caption generated on Youtube directly.

Kafka Python Videos - Rides.csv

There is no clear explanation of the rides.csv data that the producer.py python programs use. You can find that here https://raw.githubusercontent.com/DataTalksClub/data-engineering-zoomcamp/2bd33e89906181e424f7b12a299b70b19b7cfcd5/week_6_stream_processing/python/resources/rides.csv.

kafka.errors.NoBrokersAvailable: NoBrokersAvailable

If you have this error, it most likely that your kafka broker docker container is not working.

Use `docker ps` to confirm

Then in the docker compose yaml file folder, run `docker compose up -d` to start all the instances.

Kafka homework Q3, there are options that support scaling concept more than the others:

Ankush said we can focus on horizontal scaling option.

“think of scaling in terms of scaling from consumer end. Or consuming message via horizontal scaling”

How to fix docker compose error: Error response from daemon: pull access denied for spark-3.3.1, repository does not exist or may require 'docker login': denied: requested access to the resource is denied

If you get this error, know that you have not built your sparks and juypter images. This images aren’t readily available on dockerHub.

In the spark folder, run `./build.sh` from a bash cli to to build all images before running docker compose

Python Kafka: ./build.sh: Permission denied Error

Run this command in terminal in the same directory (/docker/spark):

chmod +x build.sh

Python Kafka: ‘KafkaTimeoutError: Failed to update metadata after 60.0 secs.’ when running stream-example/producer.py

Restarting all services worked for me:

```
docker-compose down

docker-compose up
```

Python Kafka: ./spark-submit.sh streaming.py - ERROR StandaloneSchedulerBackend: Application has been killed. Reason: All masters are unresponsive! Giving up.

While following tutorial 13.2 , when running ./spark-submit.sh streaming.py, encountered the following error:

```
…

24/03/11 09:48:36 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://localhost:7077...

24/03/11 09:48:36 INFO TransportClientFactory: Successfully created connection to localhost/127.0.0.1:7077 after 10 ms (0 ms spent in bootstraps)

24/03/11 09:48:54 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users should configure it(them) to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetrics.oldGenerationGarbageCollectors

24/03/11 09:48:56 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://localhost:7077…

24/03/11 09:49:16 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://localhost:7077...

24/03/11 09:49:36 WARN StandaloneSchedulerBackend: Application ID is not initialized yet.

24/03/11 09:49:36 ERROR StandaloneSchedulerBackend: Application has been killed. Reason: All masters are unresponsive! Giving up.

…
```

py4j.protocol.Py4JJavaError: An error occurred while calling None.org.apache.spark.sql.SparkSession.

: java.lang.IllegalStateException: Cannot call methods on a stopped SparkContext.

…

Solution:

Downgrade your local PySpark to **3.3.1** (same as Dockerfile)

The reason for the failed connection in my case was the mismatch of PySpark versions. You can see that from the logs of spark-master in the docker container.

Solution 2:

- Check what Spark version your local machine has
  - `pyspark –version`
  - `spark-submit –version`
- Add your version to `SPARK_VERSION` in **build.sh**

Python Kafka: ./spark-submit.sh streaming.py - How to check why Spark master connection fails

Start a new terminal

Run: **docker ps**

Copy the CONTAINER ID of the spark-master container

Run: **docker exec -it <spark_master_container_id> bash**

Run: **cat logs/spark-master.out**

Check for the log when the error happened

Google the error message from there

Python Kafka: ./spark-submit.sh streaming.py Error: py4j.protocol.Py4JJavaError: An error occurred while calling None.org.apache.spark.api.java.JavaSparkContext.

Make sure your java version is 11 or 8.

Check your version by:

java --version

Check all your versions by:

/usr/libexec/java_home -V

If you already have got java 11 but just not selected as default, select the specific version by:

export JAVA_HOME=$(/usr/libexec/java_home -v 11.0.22)

(or other version of 11)

Java Kafka: <project_name>-1.0-SNAPSHOT.jar errors: package xxx does not exist even after gradle build

In my set up, all of the dependencies listed in gradle.build were not installed in <project_name>-1.0-SNAPSHOT.jar.

Solution:

In build.gradle file, I added the following at the end:

```
shadowJar {

    archiveBaseName = "java-kafka-rides"

    archiveClassifier = ''

}
```

And then in the command line ran ‘gradle shadowjar’, and run the script from java-kafka-rides-1.0-SNAPSHOT.jar created by the shadowjar

Python Kafka: Installing dependencies for python3 06-streaming/python/avro_example/producer.py

confluent-kafka: `pip install confluent-kafka` or `conda install conda-forge::python-confluent-kafka`

fastavro: pip install fastavro

Abhirup Ghosh

Can install Faust Library for Module 6 Python Version due to dependency conflicts?

The Faust repository and library is no longer maintained - https://github.com/robinhood/faust

If you do not know Java, you now have the option to follow the Python Videos 6.13 & 6.14 here https://www.youtube.com/watch?v=BgAlVknDFlQ&list=PL3MmuxUbc_hJed7dXYoJw8DoCuVHhGEQb&index=80 and follow the RedPanda Python version here https://github.com/DataTalksClub/data-engineering-zoomcamp/tree/main/06-streaming/python/redpanda_example - NOTE: I highly recommend watching the Java videos to understand the concept of streaming but you can skip the coding parts - all will become clear when you get to the Python videos and RedPanda files.

Java Kafka: How to run producer/consumer/kstreams/etc in terminal

In the project directory, run:

```
java -cp build/libs/<jar_name>-1.0-SNAPSHOT.jar:out
src/main/java/org/example/JsonProducer.java
```

Java Kafka: When running the producer/consumer/etc java scripts, no results retrieved or no message sent

For example, when running JsonConsumer.java, got:

```
Consuming form kafka started

RESULTS:::0

RESULTS:::0

RESULTS:::0
```

Or when running JsonProducer.java, got:

Exception in thread "main" java.util.concurrent.ExecutionException: org.apache.kafka.common.errors.SaslAuthenticationException: Authentication failed

Solution:

Make sure in the scripts in src/main/java/org/example/ that you are running (e.g. JsonConsumer.java, JsonProducer.java), the StreamsConfig.BOOTSTRAP_SERVERS_CONFIG is the correct server url (e.g. europe-west3 from example vs europe-west2)

Make sure cluster key and secrets are updated in src/main/java/org/example/Secrets.java (KAFKA_CLUSTER_KEY and KAFKA_CLUSTER_SECRET)

Java Kafka: Tests are not picked up in VSCode

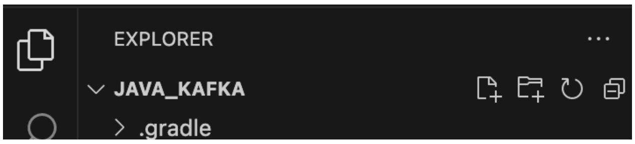
Situation: in VS Code, usually there will be a triangle icon next to each test. I couldn't see it at first and had to do some fixes.

Solution:

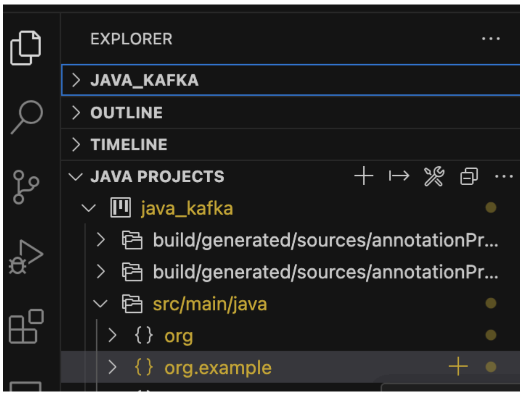
(Source)

VS Code

→ Explorer (first icon on the left navigation bar)



→ JAVA PROJECTS (bottom collapsable)



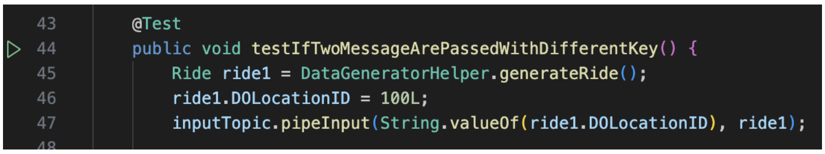
→ ··· icon next in the rightmost position to JAVA PROJECTS

→ clean Workspace

→ Confirm by clicking Reload and Delete

Now you will be able to see the triangle icon next to each test like what you normally see in python tests.

E.g.:



```
@Test
public void testIfTwoMessageArePassedWithDifferentKey() {
    Ride ride1 = DataGeneratorHelper.generateRide();
    ride1.DOLocationID = 100L;
    inputTopic.pipeInput(String.valueOf(ride1.DOLocationID), ride1);
```

You can also add classes and packages in this window instead of creating files in the project directory

Confluent Kafka: Where can I find schema registry URL?

In Confluent Cloud:

Environment → default (or whatever you named your environment as) → The right navigation bar → "Stream Governance API" → The URL under "Endpoint"

And create credentials from Credentials section below it

How do I check compatibility of local and container Spark versions?

You can check the version of your local spark using `spark-submit --version`. In the build.sh file of the Python folder, make sure that `SPARK_VERSION` matches your local version. Similarly, make sure the pyspark you pip installed also matches this version.

How to fix the error "ModuleNotFoundError: No module named 'kafka.vendor.six.moves'"?

According to https://github.com/dpkp/kafka-python/

"DUE TO ISSUES WITH RELEASES, IT IS SUGGESTED TO USE https://github.com/wbarnha/kafka-python-ng FOR THE TIME BEING"

Use pip install kafka-python-ng instead

How to fix “connection failed: connection to server at "127.0.0.1", port 5432 failed” error when setting up Postgres connection in pgAdmin?

Instead of using “localhost” as the host name/address, try “postgres”, or “host.docker.internal” instead

Alternative Solution: For those having installed postgres locally and disabling persist data on postgres-container in docker i.e. *volume:* removed, remember to use postgres port other than 5432 (e.g. 5433 is usable). And for **pgadmin host name/address**, if *localhost, postgres, and host.docker.internal* is not working, you can use your own *IPv4 Address* which can be found in Windows OS via: Command Prompt > ipconfig > Under Wireless LAN adapter WiFi 2. E.g.:

IPv4 Address. . . . . : 192.168.0.148

Why is my table not being created in PostgreSQL when I submit a job?

There could be a few reasons for this issue:

- Race Conditions: If you're running multiple processes in parallel.
- Database Connection Issues: The job might not be connecting to the correct PostgreSQL database, or there could be authentication or permission issues preventing table creation.
- Missing Table Creation Logic: The code responsible for creating the table might not be properly included or executed in the job submission process.

As a best practice, it's generally recommended to pre-create tables in PostgreSQL to avoid runtime errors. This ensures the database schema is properly set up before any jobs are executed.

Extra: Use CREATE TABLE IF NOT EXISTS in your code. This will prevent errors if the table already exists and ensure smooth job execution.

Project

How is my capstone project going to be evaluated?

- Each submitted project will be evaluated by 3 (three) randomly assigned students that have also submitted the project.
- You will also be responsible for grading the projects from 3 fellow students yourself. Please be aware that: not complying to this rule also implies you failing to achieve the Certificate at the end of the course.
- The final grade you get will be the median score of the grades you get from the peer reviewers.
- And of course, the peer review criteria for evaluating or being evaluated must follow the guidelines defined **here**.

Can I collaborate with others on the capstone project?

Collaboration is not allowed for the capstone submission. However, you can discuss ideas and get feedback from peers in the forums or Slack channels.

Project 1 & Project 2

There is only ONE project for this Zoomcamp. You do not need to submit or create two projects.

There are simply TWO chances to pass the course. You can use the Second Attempt if you a) fail the first attempt b) do not have the time due to other engagements such as holiday or sickness etc. to enter your project into the first attempt. Project evaluation - Reproducibility

The question is that sometimes even if you take plenty of effort to document every single step, and we can't even sure if the person doing the peer review will be able to follow-up, so how this criteria will be evaluated?

Alex clarifies: “Ideally yes, you should try to re-run everything. But I understand that not everyone has time to do it, so if you check the code by looking at it and try to spot errors, places with missing instructions and so on - then it's already great”

Certificates: how do I get it?

A: See the certificate.mdx file

Does anyone know nice and relatively large datasets?

See a list of datasets here:
https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/projects/datasets.md

How to run python as start up script?

You need to redefine the python environment variable to that of your user account

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
ex call. Works well when we are reading only from one topic.
```

Data Transformation from Databricks to Azure SQL DB

Transformed data can be moved in to azure blob storage and then it can be moved in to azure SQL DB, instead of moving directly from databricks to Azure SQL DB.

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.

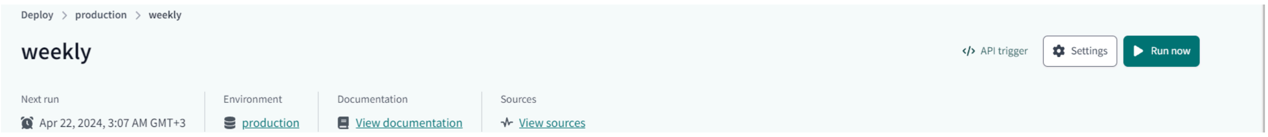
Orchestrating dbt cloud with Mage

You can trigger your dbt job in Mage pipeline. For this get your dbt cloud api key under settings/Api tokens/personal tokens. Add it safely to your .env

For example

```
dbt_api_trigger=dbt_**
```

Navigate to job page and find api trigger link



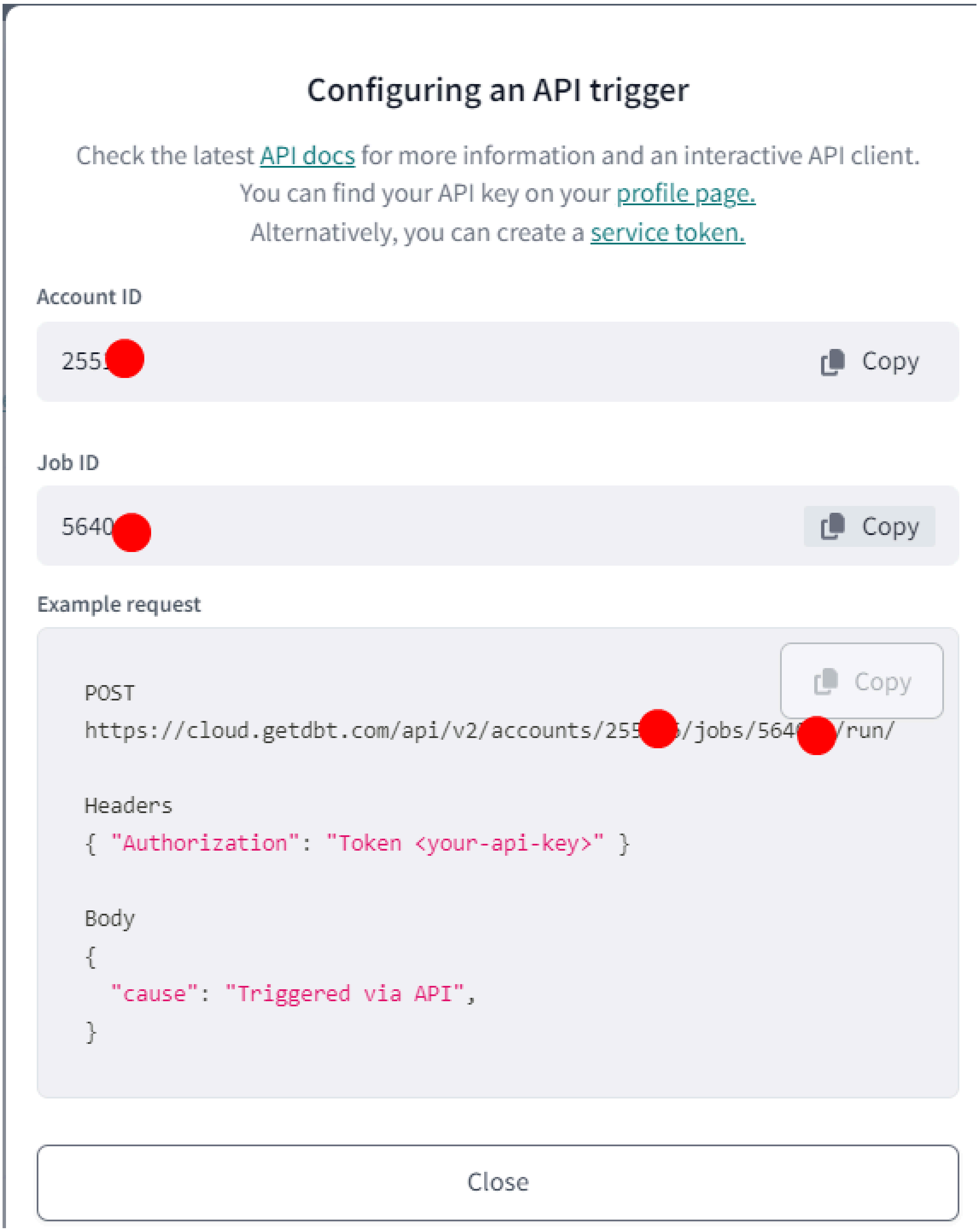
weekly

Next run
Apr 22, 2024, 3:07 AM GMT+3

Environment
production

Documentation
View documentation

Sources
View sources



Configuring an API trigger

Check the latest API docs for more information and an interactive API client.
You can find your API key on your profile page.
Alternatively, you can create a service token.

Account ID

255[redacted]

Job ID

5640[redacted]

Example request

```
POST
https://cloud.getdbt.com/api/v2/accounts/255[redacted]/jobs/564[redacted]/run/

Headers
{ "Authorization": "Token <your-api-key>" }

Body
{
  "cause": "Triggered via API",
}
```

Then create a custom mage Python block with a simple http request like here

```
from dotenv import load_dotenv
from pathlib import Path
dotenv_path = Path('/home/src/.env')
```

```
load_dotenv(dotenv_path=dotenv_path)
dbt_api_trigger= os.getenv(dbt_api_trigger)

url =
f"https://cloud.getdbt.com/api/v2/accounts/{dbt_account_id}/jobs/<job_id>/run/"

    headers = {
        "Authorization": f"Token {dbt_api_trigger}",
        "Content-Type": "application/json" }

    body = {
        "cause": "Triggered via API"
    }
    response = requests.post(url, headers=headers, json=body)
```

voila! You triggered dbt job form your mage pipeline.

Key Vault in Azure cloud stack

The key valut in Azure cloud is used to store credentials or passwords or secrets of different tech stack used in Azure. For example if u do not want to expose the password in SQL database, then we can save the password under a given name and use them in other Azure stack.

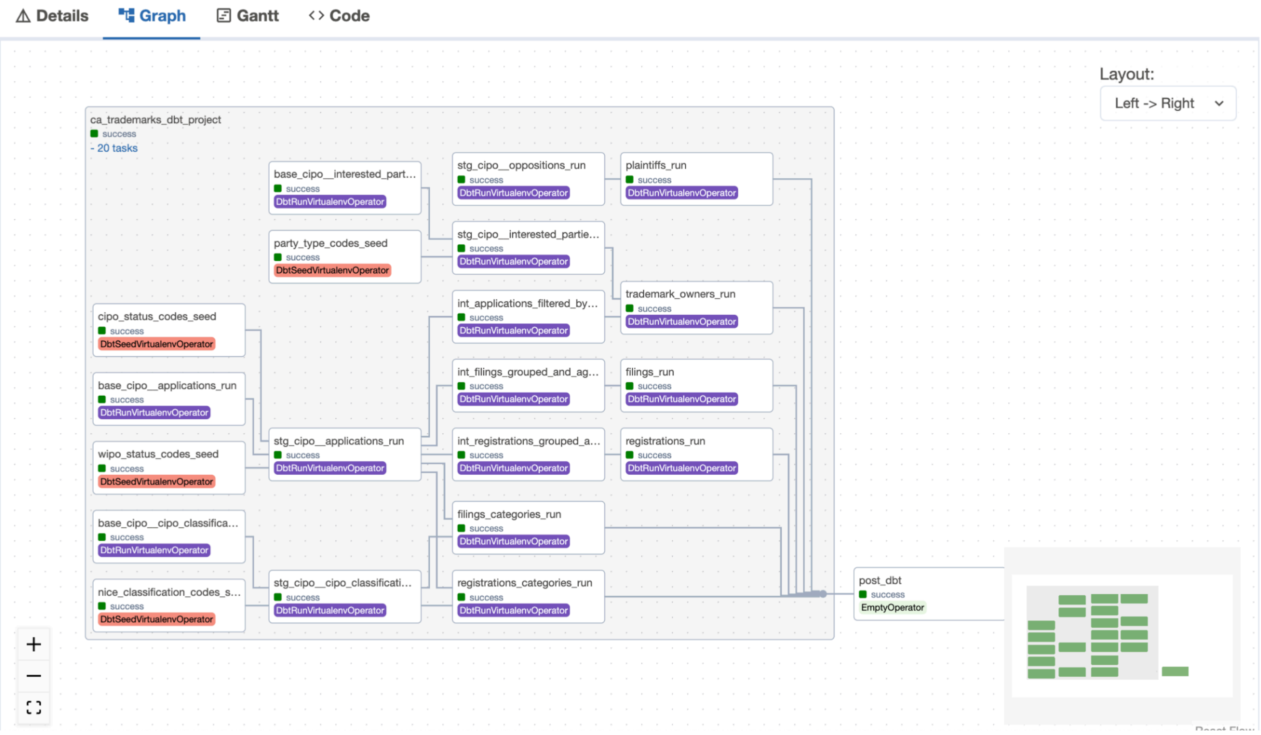
How to connect Pyspark with BigQuery?

The following line should be included in pyspark configuration

```
# Example initialization of SparkSession variable
spark = (SparkSession.builder
         .master(...)
         .appName(...)
         # Add the following configuration
         .config("spark.jars.packages",
"com.google.cloud.spark:spark-3.5-bigquery:0.37.0")
)
```

How to run a dbt-core project as an Airflow Task Group on Google Cloud Composer using a service account JSON key

1. Install the ***astronomer-cosmos*** package as a dependency. (see Terraform example).
2. Make a new folder, **dbt/**, inside the **dags/** folder of your Composer GCP bucket and copy paste your dbt-core project there. (see example)
3. Ensure your *profiles.yml* is configured to authenticate with a service account key. (see BigQuery example)
4. Create a new DAG using the **DbtTaskGroup** class and a **ProfileConfig** specifying a *profiles_yml_filepath* that points to the location of your JSON key file. (see example)
5. Your dbt lineage graph should now appear as tasks inside a task group like this:



How can I run UV in Kestra without installing it on every flow execution?

To avoid reinstalling uv on each flow run, you can create a custom Docker image based on the official Kestra image with uv pre-installed. Here's how:

- Create a Dockerfile (e.g., Dockerfile) with the following content:

```
FROM kestra/kestra:latest
```

```
USER root
RUN pip install uv

CMD ["server", "standalone"]
```

- Update your docker-compose.yml to build this custom image instead of pulling the default one:

```
# image: kestra/kestra:latest

build:
     context: .
     dockerfile: Dockerfile
```

This approach ensures that uv is available in the container at runtime without requiring installation during each flow execution.

Is it possible to create external tables in BigQuery using URLs, such as those from the NY Taxi data website?

Answer: Not really, only Bigtable, Cloud Storage, and Google Drive are supported data stores.

Is it ok to use NY_Taxi data for the project?

No

How to use dbt-core with Athena?

If you don’t have access to dbt Cloud which is already natively being supported by AWS, refer here: 1, 2, 3, & 4, you can use the community built dbt-Athena Adapter for dbt-Core.

Key Features

- Enables dbt to work with AWS Athena using dbt Core

- Allows data transformation using CREATE TABLE AS or CREATE VIEW SQL queries
- Not yet supported features:
    1. Python models
    2. Persisting documentation for views

This adapter can be a valuable resource for those who need to work with Athena using dbt Core, and I hope this entry can help others discover it.

Solving dbt-Athena library conflicts

When working on a dbt-Athena project, do not install dbt-athena-adapter. Instead, always use the dbt-athena-community package, ensuring it matches the version of dbt-core to avoid compatibility conflicts.

Best Practice

- Always pin the versions of dbt-core and dbt-athena-community to the same version.

- Example:

```
dbt-core~=1.9.3

dbt-athena-community~=1.9.3
```

Why?

- dbt-athena-adapter is outdated and no longer maintained.
- dbt-athena-community is the actively maintained package and is compatible with the latest versions of dbt-core.

Steps to Avoid Conflicts

1. Always check the compatibility matrix in the dbt-athena-community GitHub repository.
2. Update requirements.txt to use the latest compatible versions of dbt-core and dbt-athena-community.
3. Avoid mixing dbt-athena-adapter with dbt-athena-community in the same environment.

By following this practice, you can avoid the conflicts we faced previously and ensure a smooth development experience.

Workshop 1 - dlthub

Which set-up should I use for my dlt homework?

Technically you can use any code editor or Jupyter Notebook, as long as you can run dbt and answer the homework questions. A lot of code is provided by the instructor, on the homework page to give you a headstart in the right direction: https://github.com/DataTalksClub/data-engineering-zoomcamp/blob/main/cohorts/2025/workshops/dlt/dlt_homework.md

The most practical way is to use the provided Colabs Jupyter notebook called 'dlt - Homework.ipynb' which you can find here below, since all of the provided code is applicable in the Colabs set-up: https://colab.research.google.com/drive/1plqdl33K_HkVx0E0nGJrrkEUssStQsW7#scrollTo=BtsSxtFfXQs3

How do I install the necessary dependencies to run the code?

Answer: To run the provided code, ensure that the 'dlt[duckdb]' package is installed. You can do this by executing the provided installation command in a jupyter notebook: !pip install dlt[duckdb]. If you're doing it locally, be sure to also have duckdb pip installed (even before the duckdb package is loaded).

in zsh try:

```
pip install “dlt[duckdb]”
```

Other packages needed but not listed

If you are running Jupyter Notebook on a fresh new Codespace or in local machine with a new Virtual Environment, you will need this package to run the starter Jupyter Notebook offered by the teacher. Execute this:

Install all the necessary dependencies

pip install duckdb pandas numpy pyarrow

Or save it into a `requirements.txt` file:

dlt[duckdb]

duckdb

pandas

numpy

pyarrow # Optional, needed for Parquet support

Then run pip install -r requirements.txt

How can I use DuckDB In-Memory database with dlt ?

```
import dlt
import duckdb

conn = duckdb.connect()


def my_generator_fn():
   # implement your generator function
   pass

pipeline = dlt.pipeline(
   pipeline_name='my_pipeline',
   destination=dlt.destinations.duckdb(conn),
   dataset_name='dlt',
)


pipeline.run(
   my_generator_fn,
   table_name='my_table',
   write_disposition='replace',
)

print(conn.sql("SELECT * FROM dlt.my_table"))
```

```
conn.close()
```

Alternatively, you can switch to in-file storage with:

```
# In-Memory database storage
conn = duckdb.connect()

# File database storage
conn = duckdb.connect("/path/to/your/db.duckdb")
```

Homework - dlt Exercise 3 - Merge a generator concerns

After loading, you should have a total of 8 records, and ID 3 should have age 33

Question: ***Calculate the sum of ages of all the people loaded as described above***

The sum of all eight records' respective ages is too big to be in the choices. You need to first filter out the people whose occupation is equal to *None* in order to get an answer that is close to or present in the given choices. 😃

✅ **FIXED = use a raw string and keep the file:/// at the start of your file path** 🙂

‼️ I'm having an issue with the dlt workshop notebook. The 'Load to Parquet file' section specifically. No matter what I change the file path to, it's still saving the dlt files directly to my C drive.

```
# Set the bucket_url. We can also use a local folder
os.environ['DESTINATION__FILESYSTEM__BUCKET_URL'] =
r'file:///content/.dlt/my_folder'
url =
"https://storage.googleapis.com/dtc_zoomcamp_api/yellow_tripdata_2009-
06.jsonl"
# Define your pipeline
pipeline = dlt.pipeline(
    pipeline_name='my_pipeline',
    destination='filesystem',
    dataset_name='mydata'
)
# Run the pipeline with the generator we created earlier.
load_info = pipeline.run(stream_download_jsonl(url),
table_name="users", loader_file_format="parquet")

print(load_info)

# Get a list of all Parquet files in the specified folder
```

```
parquet_files =
glob.glob('/content/.dlt/my_folder/mydata/users/*.parquet')

# show parquet files
for file in parquet_files:
  print(file)
```

Problem with importing the dlt or dlt.sources module

Make sure you don’t have a dlt.py file saved in the same directory as your working file.

How to set credentials in Google Colab notebook to connect to BigQuery

In the secrets sidebar, create a secret ‘BIGQUERY_CRENTIALS’ with value being your Google Cloud service account key. Then load it with:

```
import os

from google.colab import userdata



os.environ["DESTINATION__BIGQUERY__CREDENTIALS"] =
userdata.get('BIGQUERY_CREDENTIALS')
```

How do I set up credentials to run dlt in my environment (not Google Colab)?

You can set up credentials for `dlt` in several ways. Here are the two most common methods:

1. Environment Variables (Easiest)

- Set credentials via environment variables. For example, to configure Google Cloud credentials:

```
export
GOOGLE_SECRETS__CREDENTIALS="/path/to/your/service_account_key.json"
```

- This method avoids hardcoding secrets in your code and works seamlessly with most environments.

2. Configuration Files (Recommended for Local Use)

- Use `.dlt/secrets.toml` for sensitive credentials and `.dlt/config.toml` for non-sensitive configurations.
- Example for Google Cloud in `secrets.toml`:

```
[google_secrets.credentials]
project_id = "<your-project-id>"
private_key = "-----BEGIN PRIVATE KEY-----\n...\n-----END PRIVATE KEY-----\n"
client_email = "<your-service-account>@<project-id>.iam.gserviceaccount.com"
```

- Place these files in the .dlt folder of your project.

Additional Notes:

- Never commit secrets.toml to version control (add it to .gitignore).
- Credentials can also be loaded via vaults, AWS Parameter Store, or custom setups.

For additional methods and detailed information, refer to the official dlt documentation

Make DLT comply with the XDG Base Dir Specification

You can set the environment variable in your shell init script (for Bash or ZSH):

export DLT_DATA_DIR=$XDG_DATA_HOME/dlt

Or for Fish (in config.fish):

set -x DLT_DATA_DIR “$XDG_DATA_HOME/dlt”

Embedding dlt into Apache Airflow

```
from airflow import DAG
from airflow.operators.python import PythonOperator
from datetime import datetime, timedelta
import dlt
from my_dlt_pipeline import load_data  # Import your dlt pipeline function

default_args = {
    "owner": "airflow",
    "depends_on_past": False,
    "start_date": datetime(2024, 2, 16),
    "retries": 1,
    "retry_delay": timedelta(minutes=5),
}

def run_dlt_pipeline():
    pipeline = dlt.pipeline(
        pipeline_name="my_pipeline",
        destination="duckdb",  # Change this based on your database
        dataset_name="my_dataset"
    )
    info = pipeline.run(load_data())
    print(info)  # Logs for debugging

with DAG(
    "dlt_airflow_pipeline",
    default_args=default_args,
    schedule_interval="@daily",
    catchup=False,
) as dag:
    run_dlt_task = PythonOperator(
        task_id="run_dlt_pipeline",
        python_callable=run_dlt_pipeline,
    )

    run_dlt_task
```

Embedding dlt into Kestra

```
id: dlt_ingestion
namespace: my.dlt
description: "Run dlt pipeline with Kestra"
tasks:
  - id: run_dlt
    type: io.kestra.plugin.scripts.python.Commands
    commands:
      - |
        import dlt
        from my_dlt_pipeline import load_data  # Import your dlt function

        pipeline = dlt.pipeline(
            pipeline_name="kestra_pipeline",
            destination="duckdb",
            dataset_name="kestra_dataset"

        )

        info = pipeline.run(load_data())

        print(info)
```

Loading Dlt Exports from GCS Filesystems

When using the filesystem destination, you may have issues reading the files exported because dlt will by default compress the files.

If you are using `loader_file_format="parquet"` then BigQuery should cope with this compression OK. If you want to use jsonl or csv format however, then you may need to disable file compression to avoid issues with reading the files directly in BigQuery. To do this set the following config:

```
[normalize.data_writer]

disable_compression = true
```

There is further information at https://dlthub.com/docs/dlt-ecosystem/destinations/filesystem#file-compression

[WARNING]: Test 'test.taxi_rides_ny.relationships_stg_yellow_tripdata_dropoff_locationid__locationid__ref_taxi_zone_lookup_csv_.085c4830e7' (models/staging/schema.yml) depends on a node named 'taxi_zone_lookup.csv' in package '' which was not found
solve: This warning indicates that dbt is trying to reference a model or source named taxi_zone_lookup.csv, but it cannot find it. We might have a typo in our ref() function.

```
tests:

  - name: relationships_stg_yellow_tripdata_dropoff_locationid

    description: "Ensure dropoff_location_id exists in taxi_zone_lookup.csv"

    relationships:

      to: ref('taxi_zone_lookup.csv')  # ❌ Wrong reference

      field: locationid
```

to:

```
 to: ref('taxi_zone_lookup')  # ✅ Correct reference
```

When I ran `df_spark = spark.createDataFrame(df_pandas)`, I encountered an error indicating a version mismatch between Pandas and Spark. To resolve this, I had two options: either downgrade Pandas to a version below 2 or upgrade Spark to version 3.5.5. I chose to upgrade Spark to 3.5.5, and it worked.

Avoiding Backpressure in Flink

- **What's Backpressure?**
    - It happens when **Flink processes data slower** than Kafka produces it.
    - This leads to **increased memory usage** and can **slow down or crash the job**.
- **How to Fix It?**
    - Adjust Kafka's **consumer parallelism** to **match the producer rate**.
    - **Increase partitions** in Kafka to allow more parallel processing.
    - Monitor **Flink metrics** to detect backpressure.
- `env.set_parallelism(4)  # Adjust parallelism to avoid bottlenecks`