Poliygon API ETL

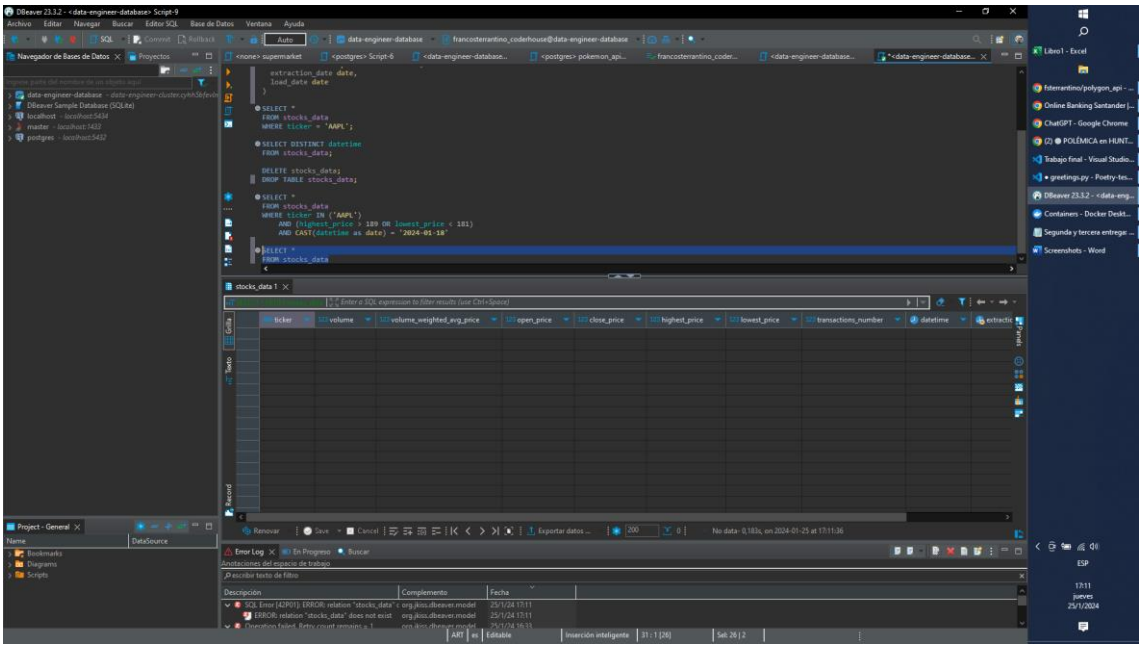
Starting configs

- Config.ini:

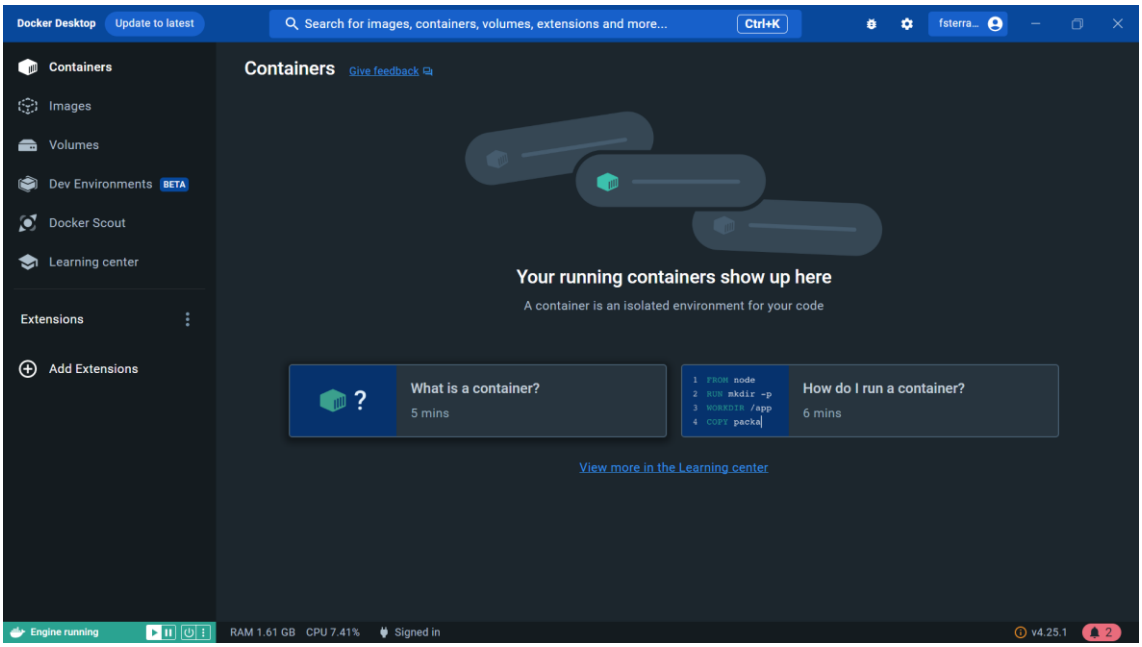
```
[API_PARAMETERS]
tickers_to_query_list = AAPL, AMZN, TSLA, BABA, MSFT, ACN, T, BAC, KO
time_frame = hour
frame_multiplier = 1
```

This config will query for that tickers info in an hourly bar.

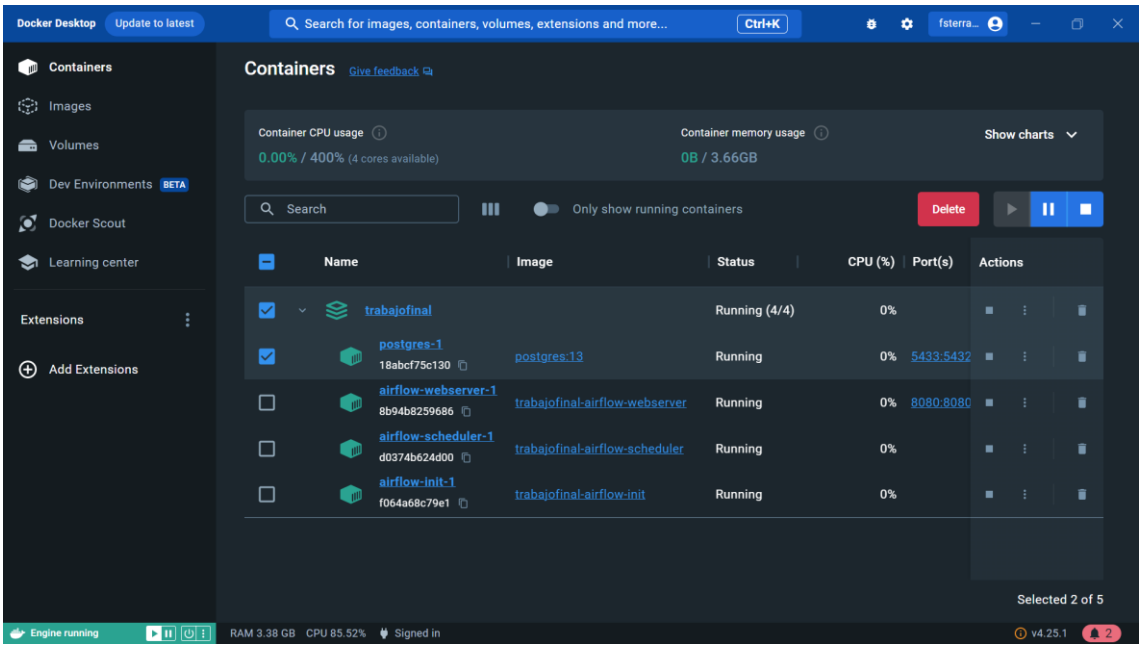
- Checking empty DB table:



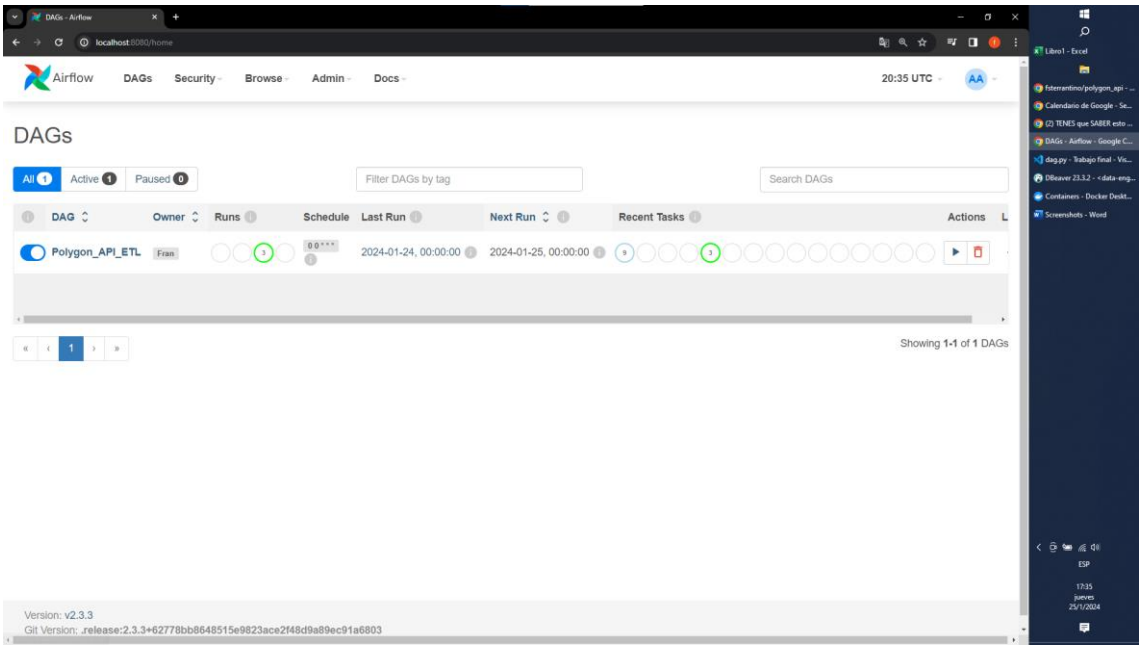
- Checking no containers in Docker:



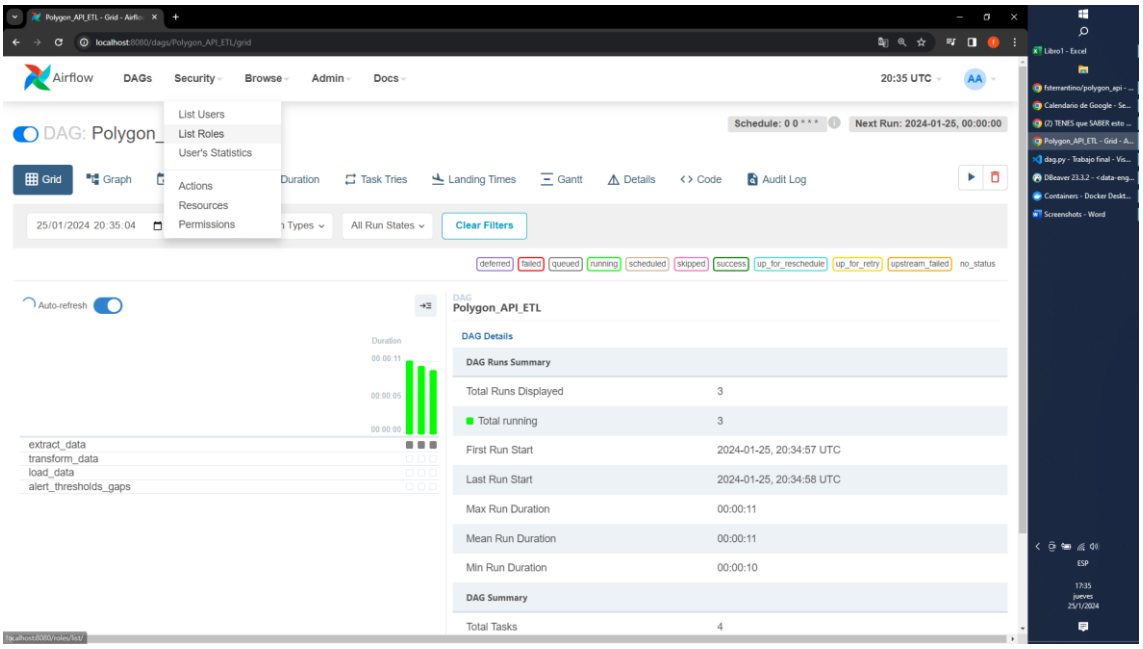
- Container started



Airflow

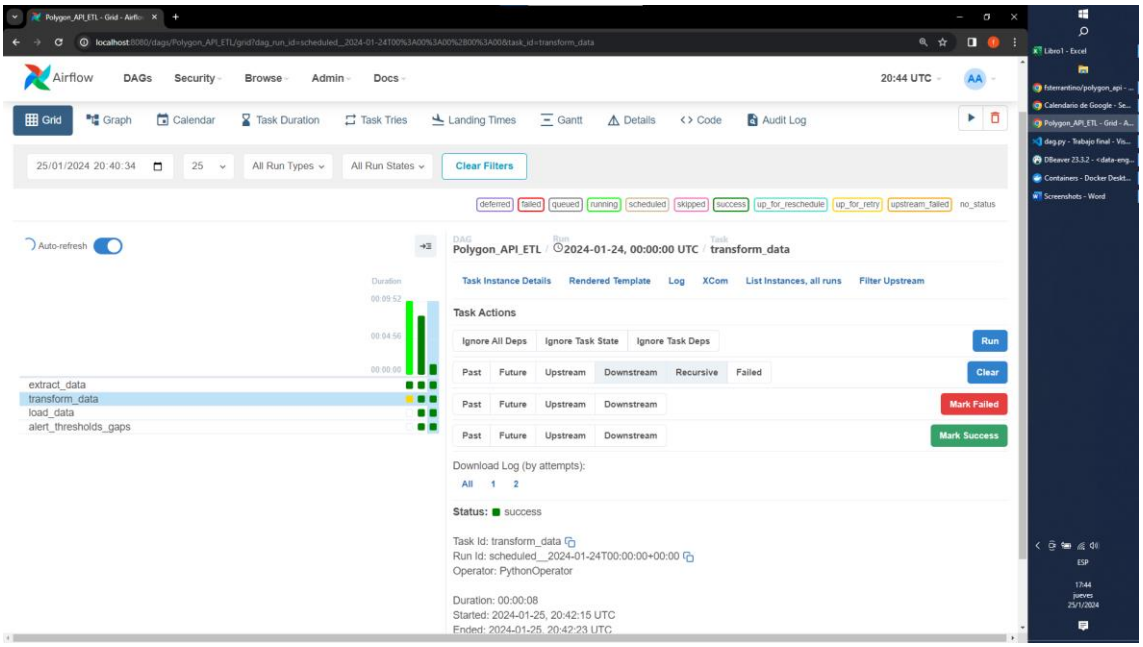


- Starting DAG execution (from startdate: 22/01/2024 to 24/01/2024)



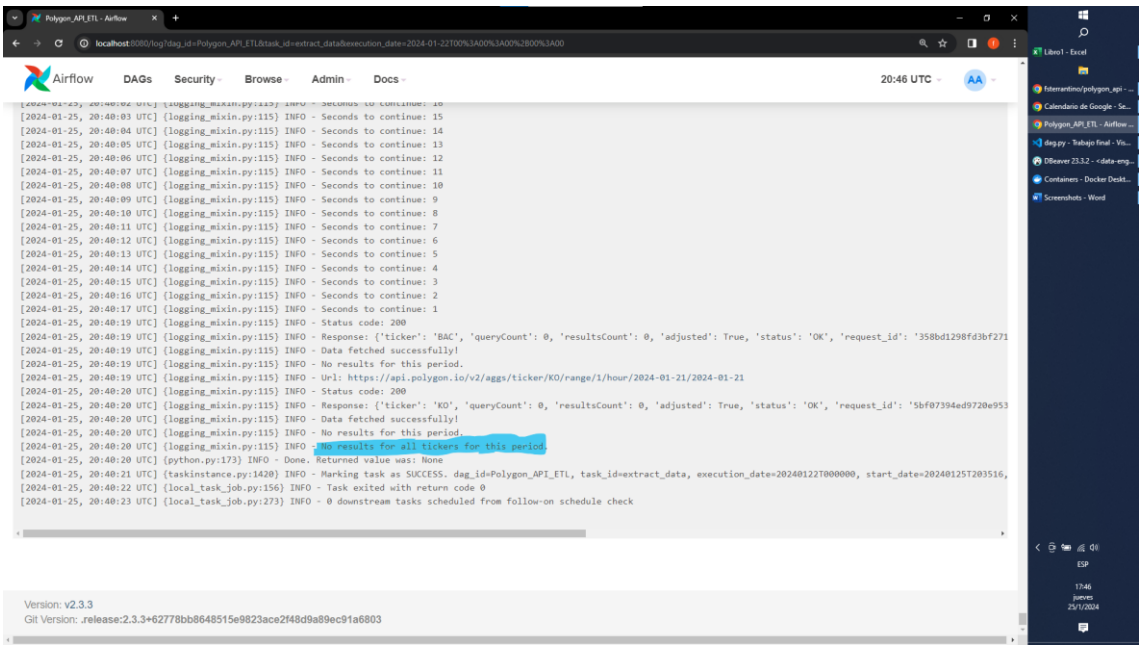
The expected result is the first transform to fail because it's retrieving data from Sunday, when stocks markets are closed.

The extraction delays some time because the API has a rate-limit of 5 requests per minute.

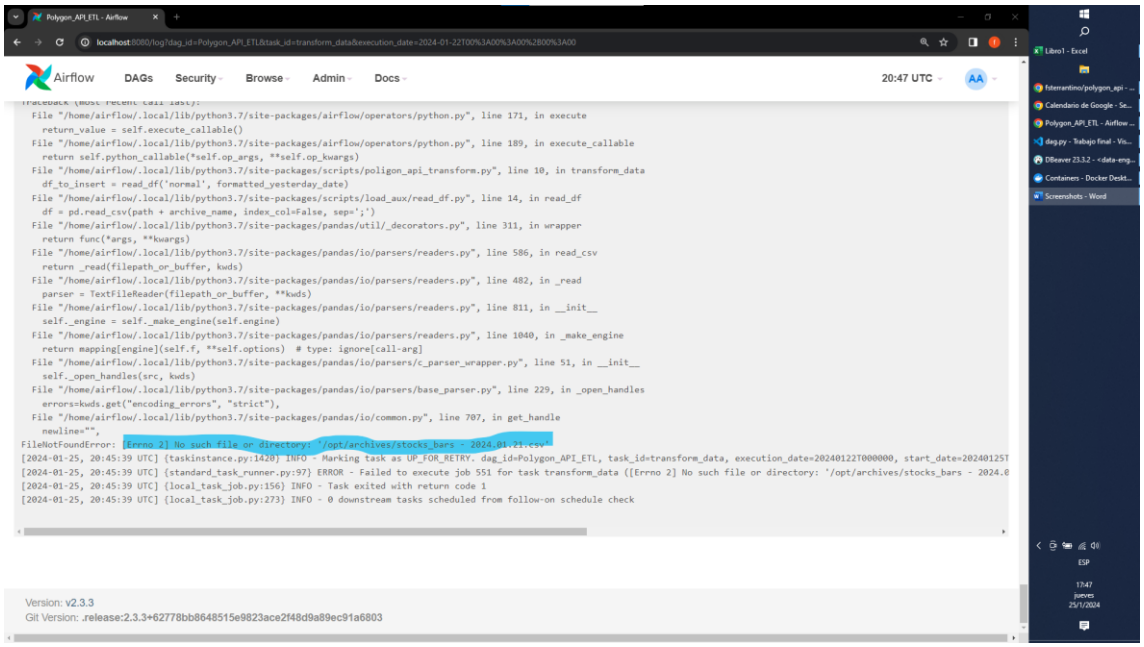


22/01/2024: Failed in transformation task as expected. The extraction did not generate the file as there is no activity on Sunday.

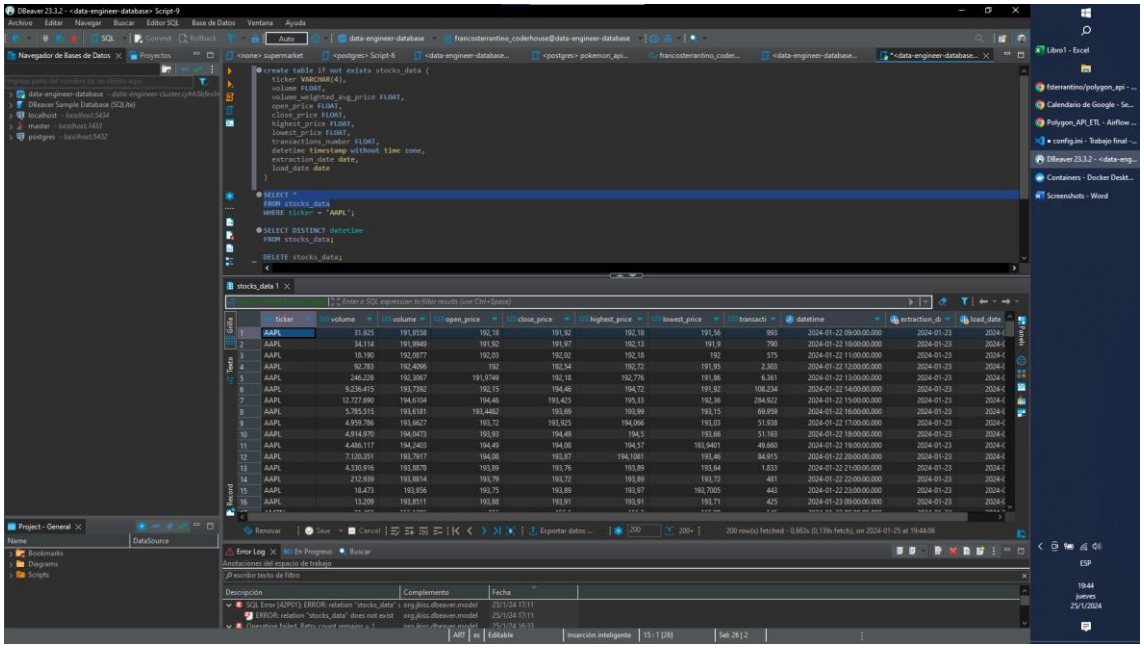
Extraction logs:



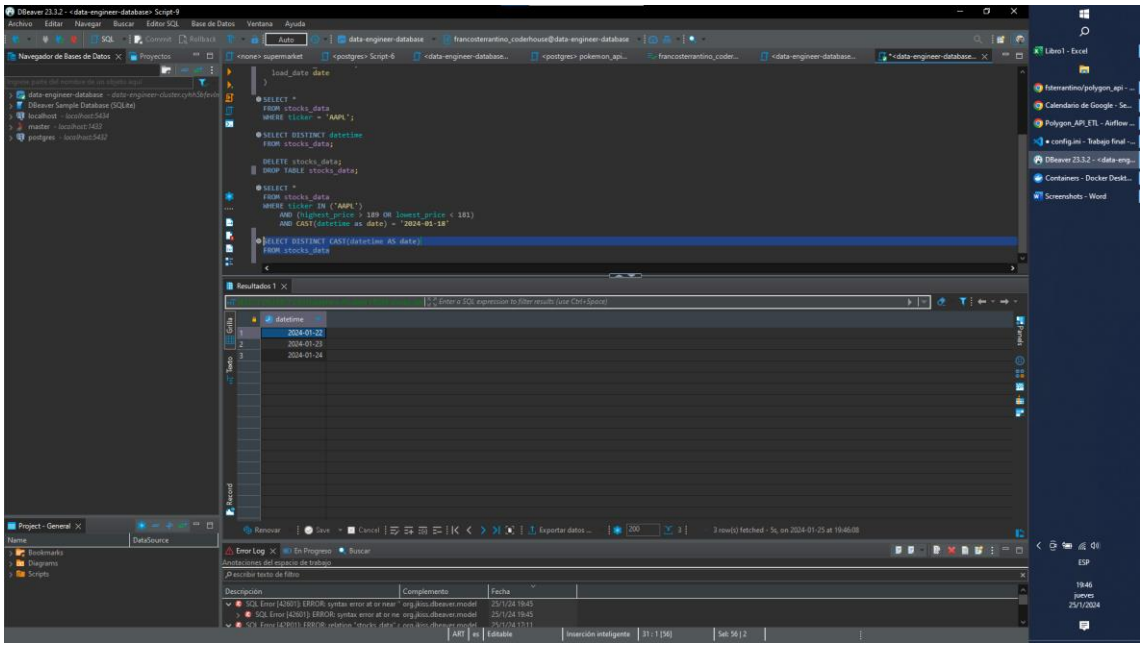
Transform logs:



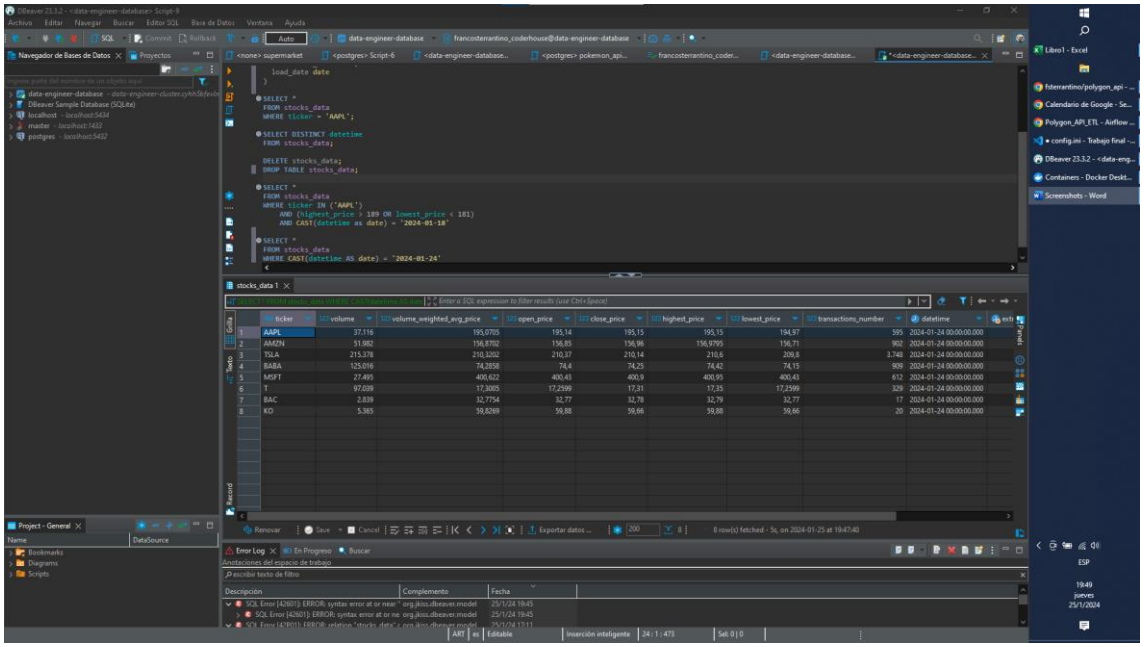
Load impact in DDBB:



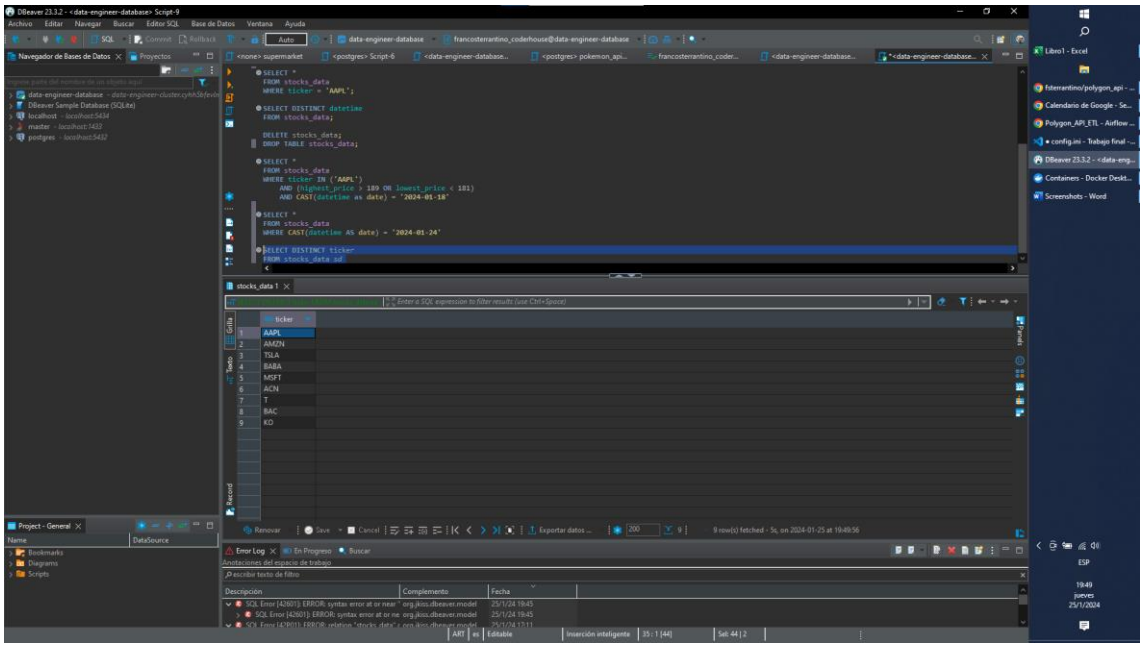
Distinct datetimes are: 22, 23, 24. As expected. We have some rows for January 24th because in the extraction of 23 it catches the Price for 24/01/2024 00:00:00.000.



January 24th screen:

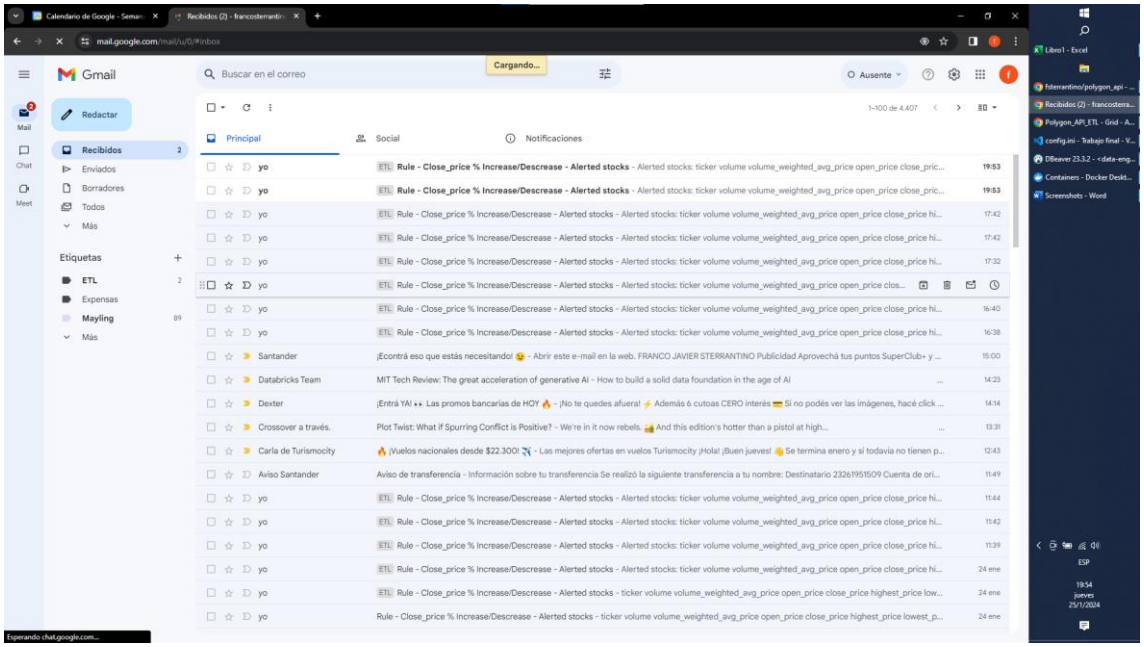


Checking distinct tickers: 9 tickers as expected.

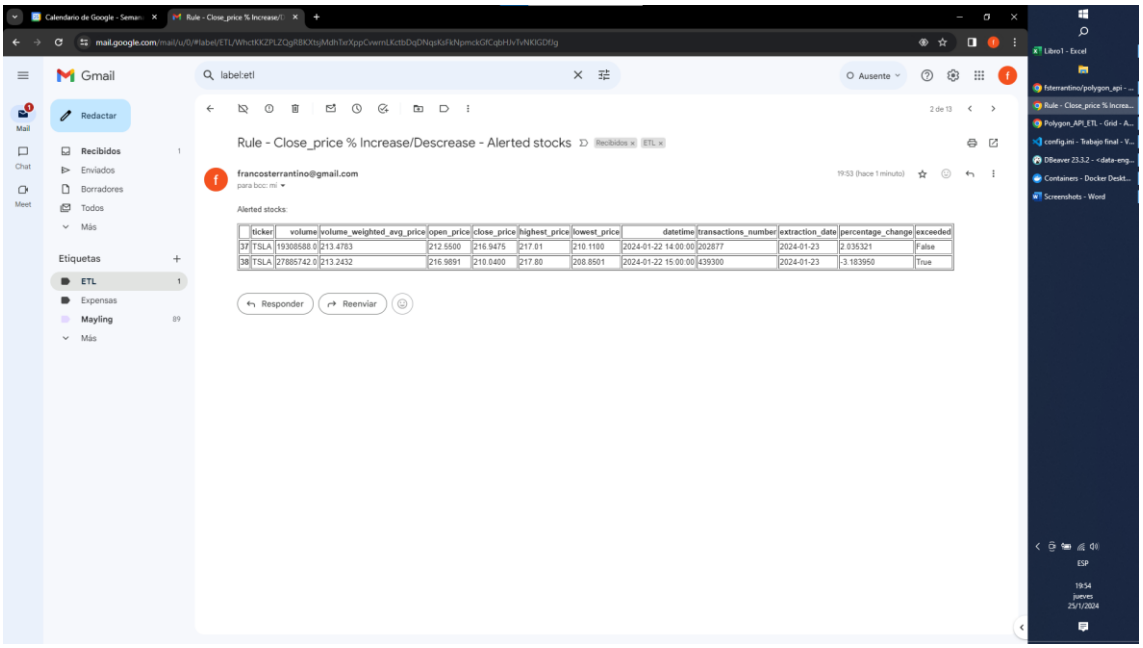


Alerted stock: I had to “clear” the tasks because one config.ini parameter was wrong. Decrease_threshold must be lower than 0.

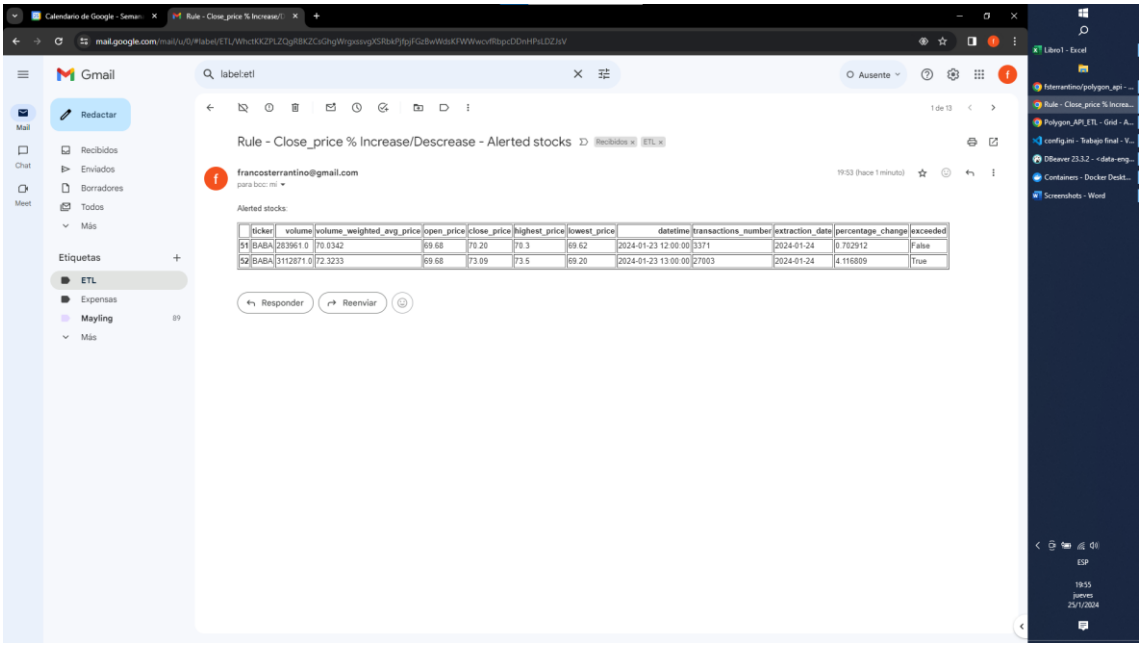
```
[ALERTS_PARAMS]
percentage_increase_threshold = 3
percentage_decrease_threshold = -2
```



TLSA alerted on January 23rd. The stock Price moved from 216.9475 to 210.04 (-3.183950%). More decrease than the threshold (-2%).



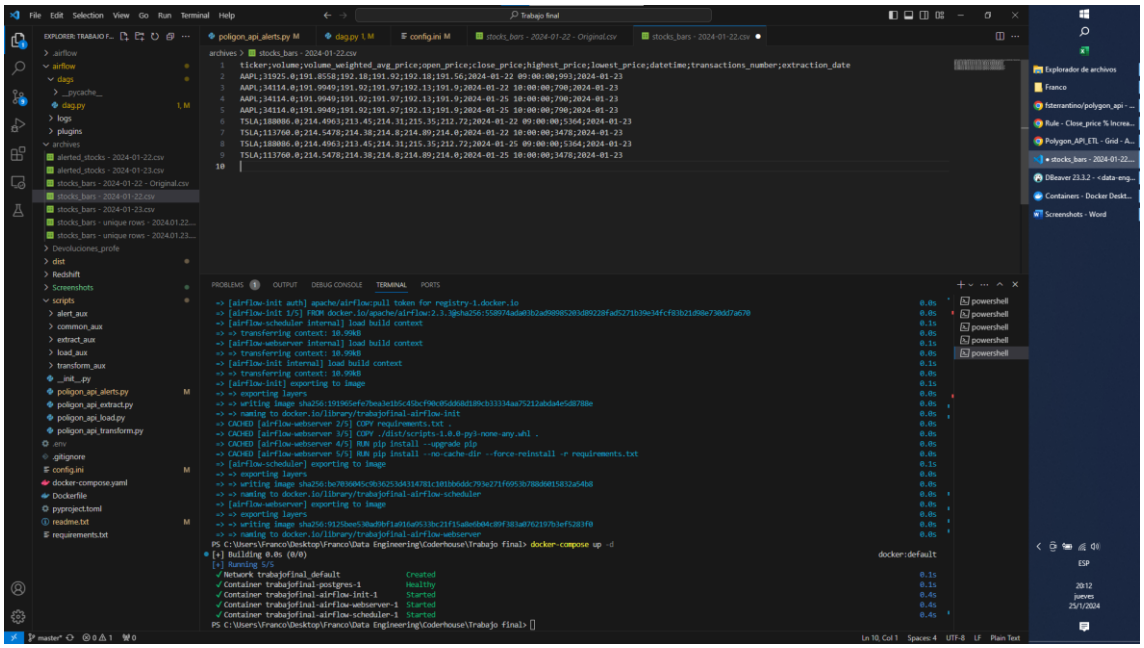
BABA alerted on January 24rd. The stock Price moved from 70.20 to 73.09 (4.116809%). More increase than the threshold (3%).



I downloaded a csv with the data loaded in the DB to check this control. The resulto f the script is correct:

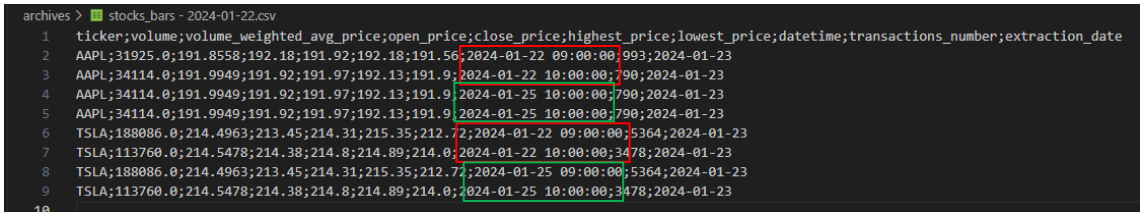
stocks_data_202401252000.csv

Now, I would try to test the integrity control. With that objective, I modificated the file that it's loaded on January 22nd.

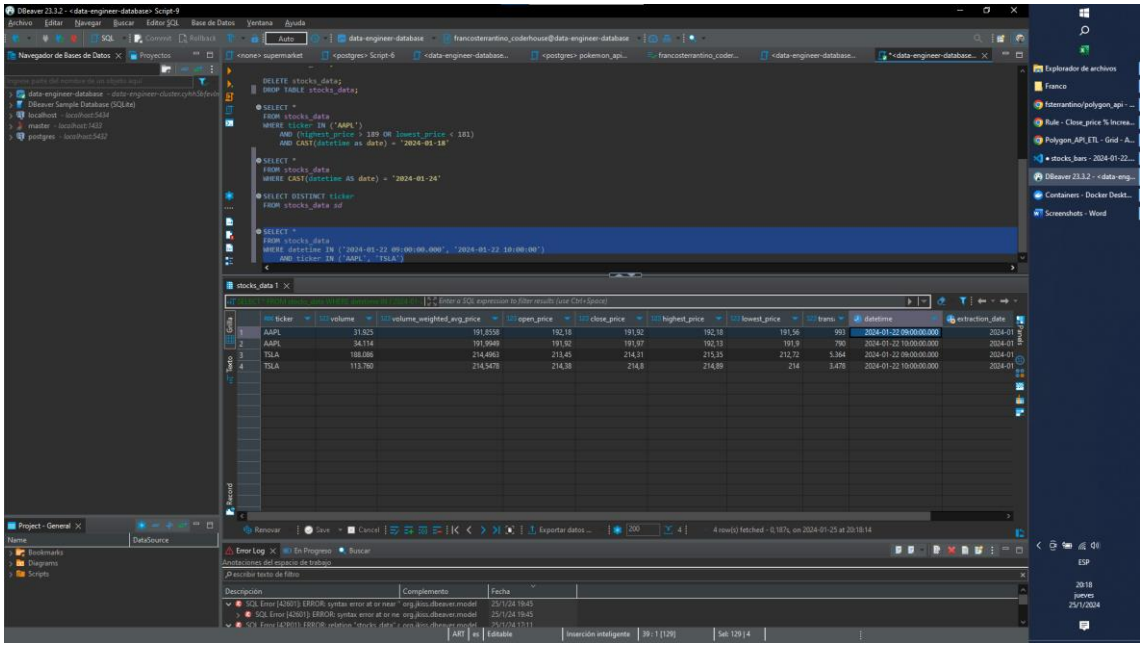


The first two rows of AAPL and TSLA, are the original ones, and they must be denied.

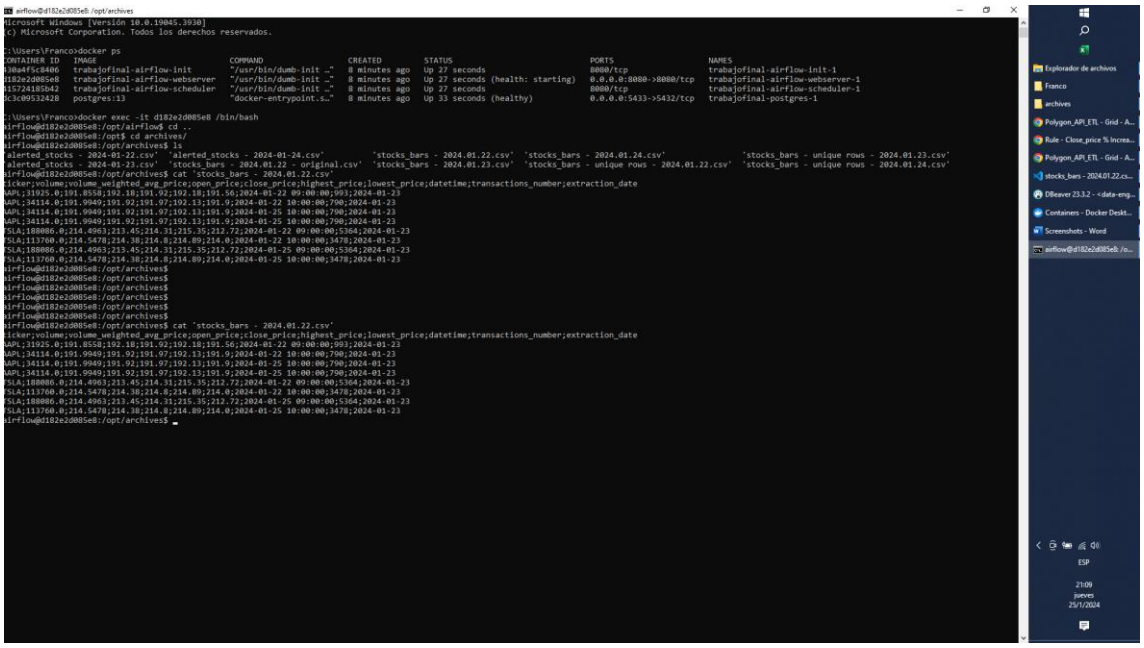
The other rows, are a copy with the datetime modified to 2024-01-25, in order to pass the control and be inserted.

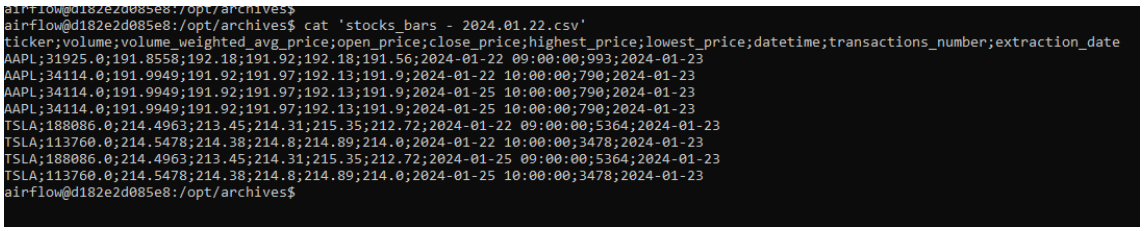


We can check that red rows are effectively loaded right now:

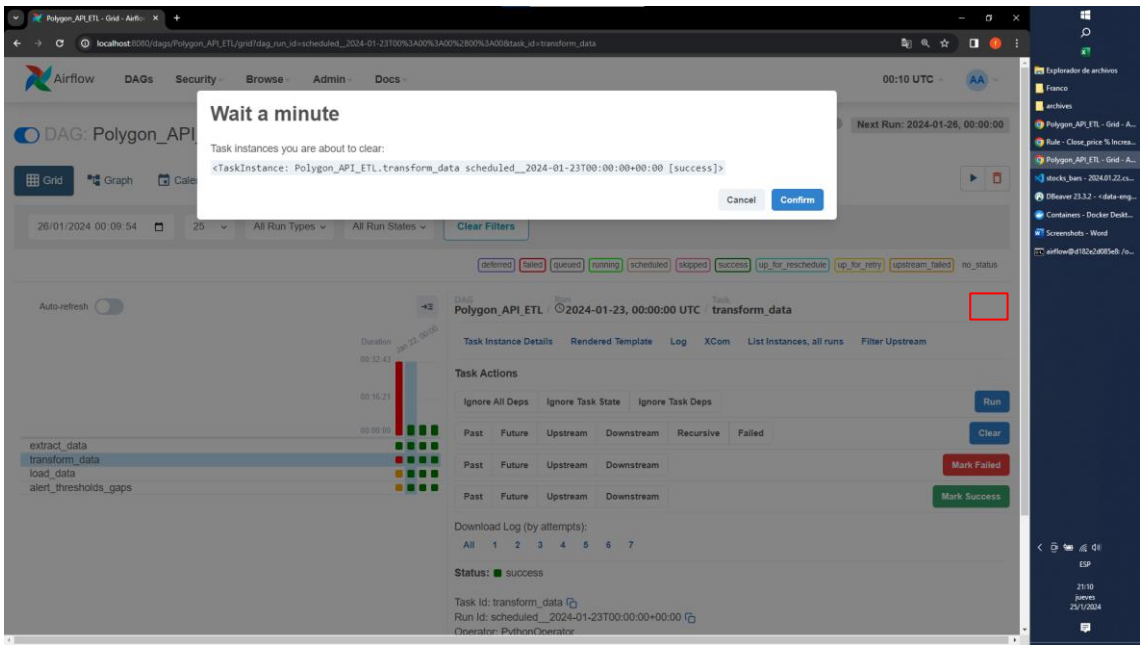


I replaced the file in the Docker container:

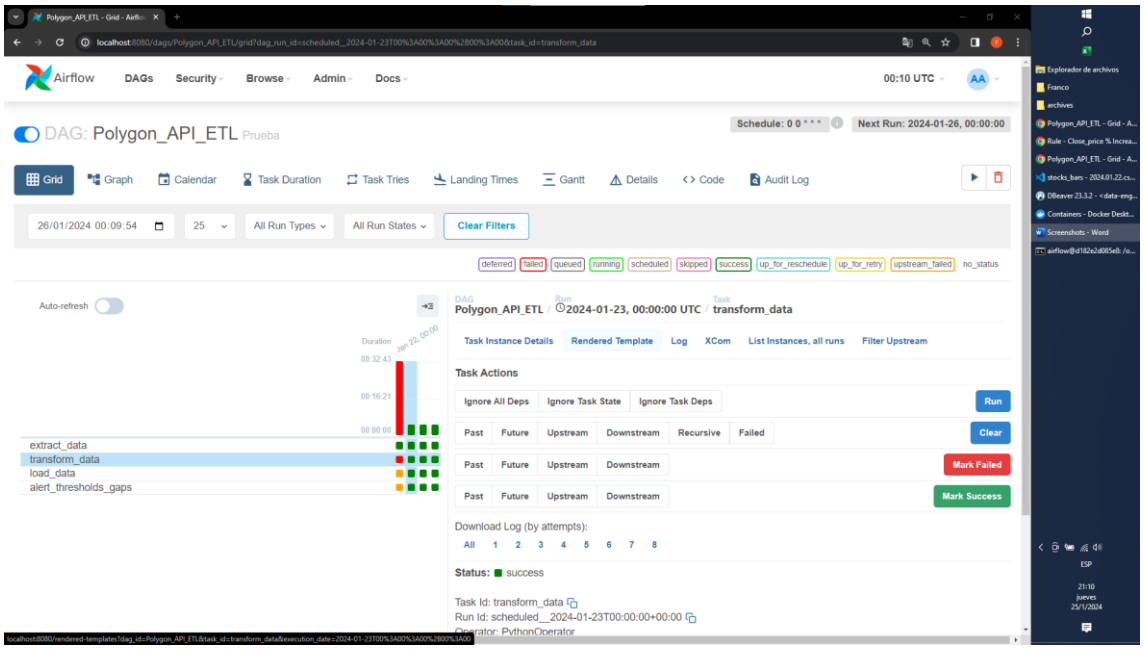




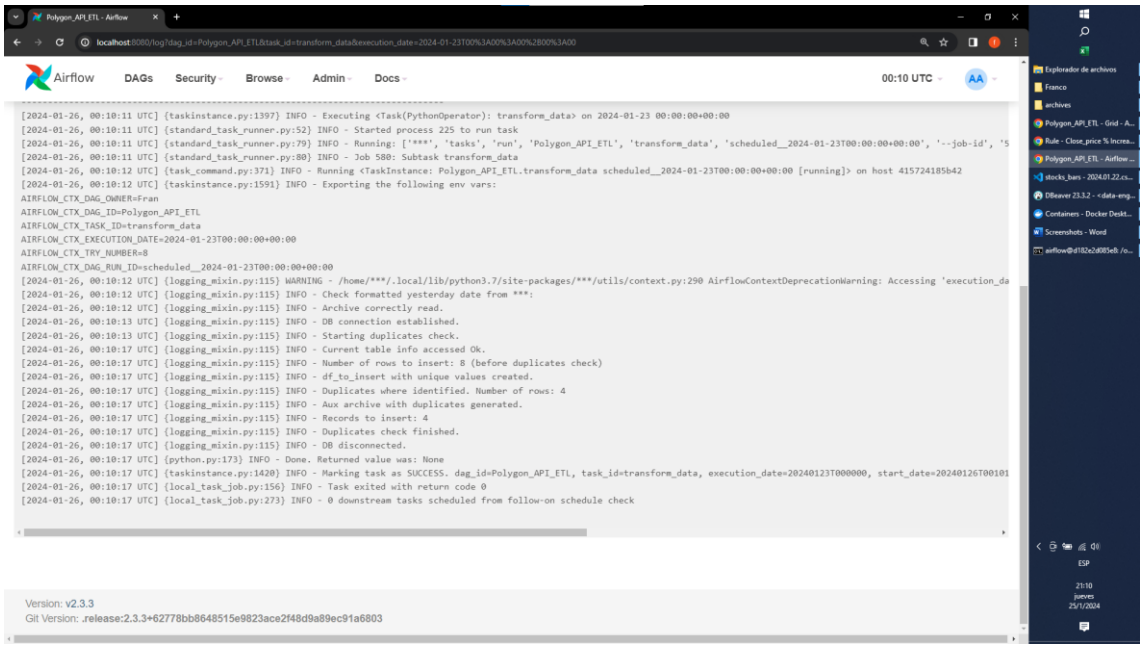
Now I'm going to clear the transform task from January 23rd (with info of 22nd) in order to re-execute it.



It finished with successful state:

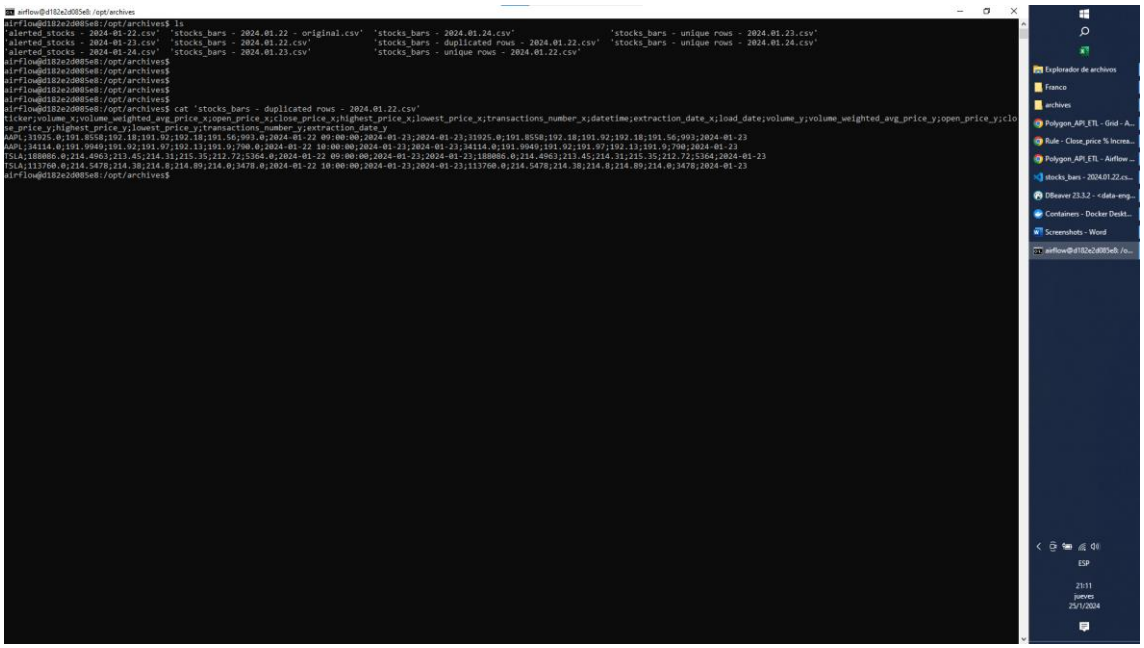


Let's see the logs:

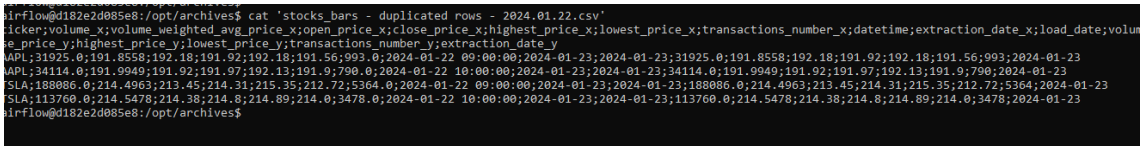


```
[2024-01-26, 00:10:12 UTC] {logging_mixin.py:115} WARNING - /home/***/.local/lib/python3.7/site-packages/***/utils/conte
[2024-01-26, 00:10:12 UTC] {logging_mixin.py:115} INFO - Check formatted yesterday date from ***:
[2024-01-26, 00:10:12 UTC] {logging_mixin.py:115} INFO - Archive correctly read.
[2024-01-26, 00:10:13 UTC] {logging_mixin.py:115} INFO - DB connection established.
[2024-01-26, 00:10:13 UTC] {logging_mixin.py:115} INFO - Starting duplicates check.
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - Current table info accessed Ok.
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - Number of rows to insert: 8 (before duplicates check)
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - df_to_insert with unique values created.
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - Duplicates where identified. Number of rows: 4
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - Aux archive with duplicates generated.
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - Records to insert: 4
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - Duplicates check finished.
[2024-01-26, 00:10:17 UTC] {logging_mixin.py:115} INFO - DB disconnected.
[2024-01-26, 00:10:17 UTC] {python.py:173} INFO - Done. Returned value was: None
[2024-01-26, 00:10:17 UTC] {taskinstance.py:1420} INFO - Marking task as SUCCESS. dag_id=Polygon_API_ETL, task_id=transf
[2024-01-26, 00:10:17 UTC] {local_task_job.py:156} INFO - Task exited with return code 0
[2024-01-26, 00:10:17 UTC] {local_task_job.py:273} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

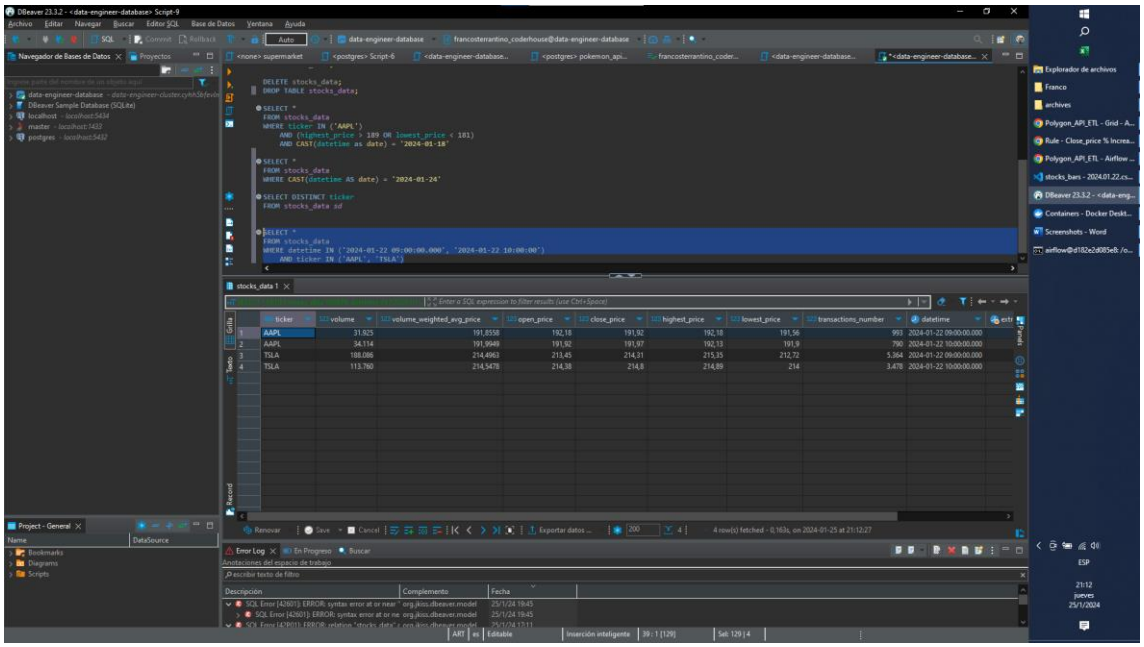
It detected successfully the 4 rows duplicated. Let’s see them in the file generated:



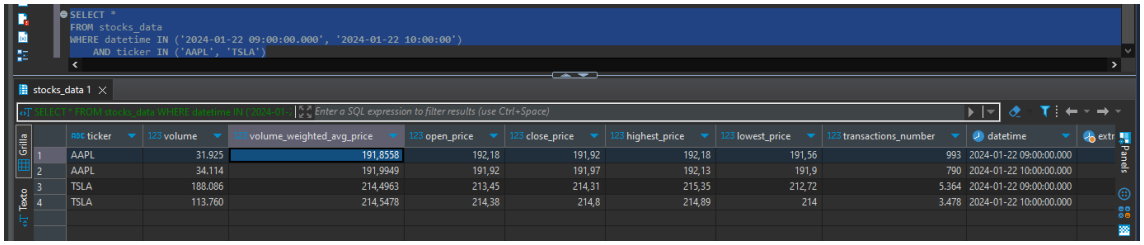
We can see the 4 duplicated rows:



And finally we can check the DB:

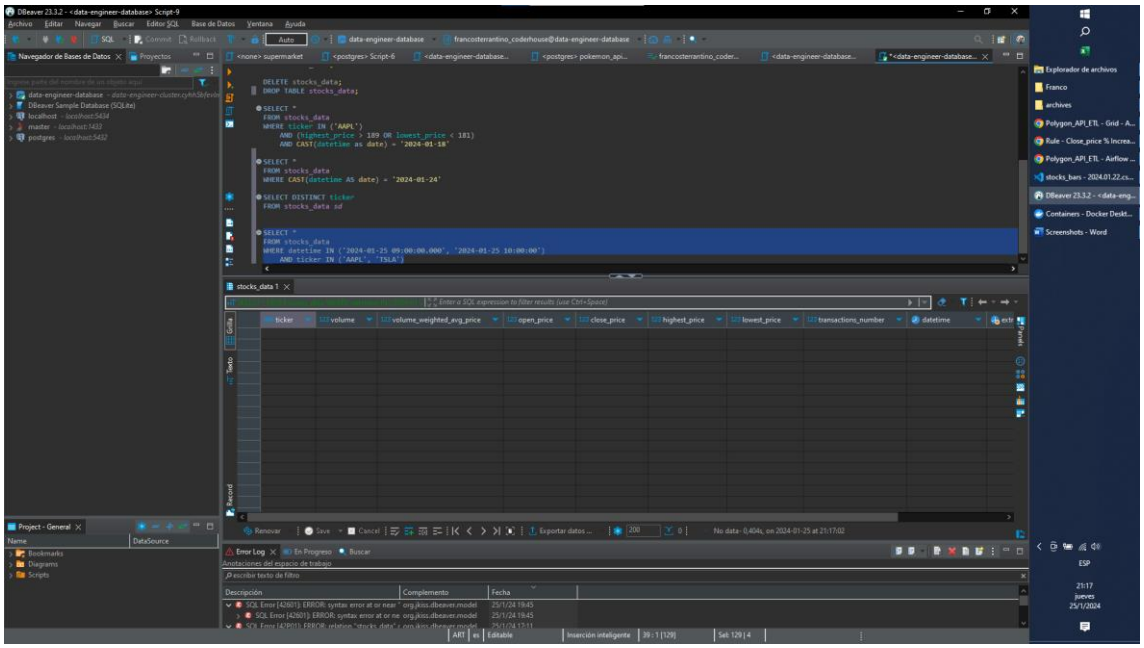


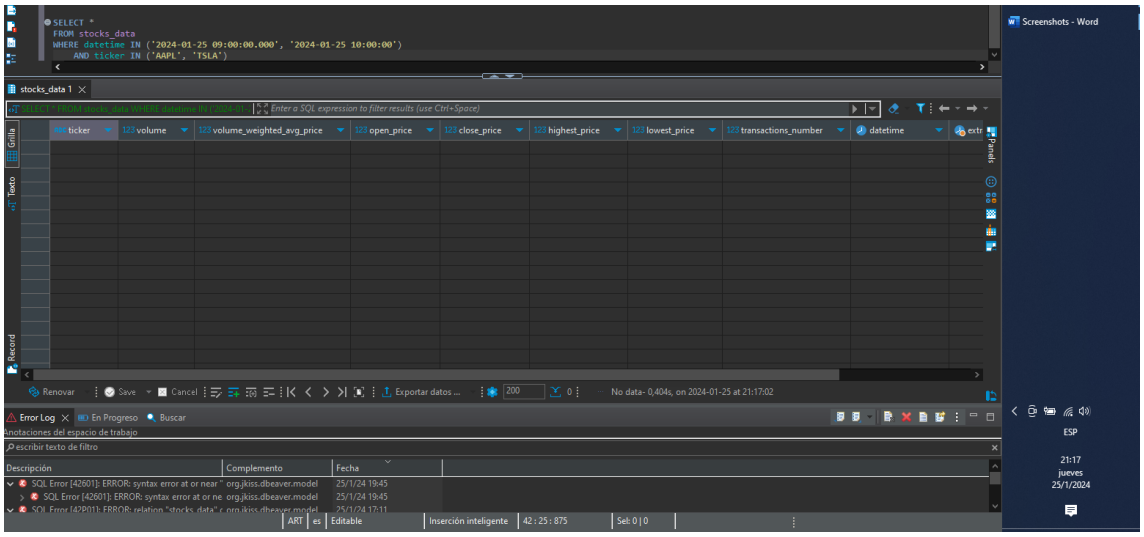
The query used before continues giving us the 4 original rows:



	ticker	volume	volume_weighted_avg_price	open_price	close_price	highest_price	lowest_price	transactions_number	datetime
1	AAPL	31.925	191,8558	192,18	191,92	192,18	191,56	993	2024-01-22 09:00:00.000
2	AAPL	34.114	191,9949	191,92	191,97	192,13	191,9	790	2024-01-22 10:00:00.000
3	TSLA	188.086	214,4963	213,45	214,31	215,35	212,72	5.364	2024-01-22 09:00:00.000
4	TSLA	113.760	214,5478	214,38	214,8	214,89	214	3.478	2024-01-22 10:00:00.000

The four additional rows are not inserted because the load task was not executed:




SELECT *
FROM stocks_data
WHERE datetime IN ('2024-01-25 09:00:00.000', '2024-01-25 10:00:00')
AND ticker IN ('AAPL', 'TSLA')
stocks_data 1
Enter a SQL expression to filter results (use Ctrl+Space)
ticker
volume
volume_weighted_avg_price
open_price
close_price
highest_price
lowest_price
transactions_number
datetime
Grilla
Texto
Record
Panels
Renovar
Save
Cancel
Exportar datos ...
200
No data- 0,404s, on 2024-01-25 at 21:17:02
Error Log
En Progreso
Buscar
Anotaciones del espacio de trabajo
escribir texto de filtro
Descripción
Complemento
Fecha
SQL Error [42601]: ERROR: syntax error at or near " org.jkiss.dbeaver.model 25/1/24 19:45
SQL Error [42601]: ERROR: syntax error at or ne org.jkiss.dbeaver.model 25/1/24 19:45
ART
es
Editable
Inserción inteligente
42 : 25 : 875
Sel: 0 | 0
Screenshots - Word
ESP
21:17
jueves
25/1/2024