

## Checklist tugas

<b>Bagian 1: Menyiapkan Lingkungan Airflow</b> <ul style="list-style-type: none"><li>• Siapkan lingkungan Airflow lokal (menggunakan Docker atau instalasi langsung)</li><li>• Konfigurasi koneksi yang diperlukan untuk sumber data dan DuckDB Anda</li></ul>	<ul style="list-style-type: none"><li>• Menggunakan Docker untuk menjalankan Airflow</li><li>• Airflow Web UI berjalan di <code>localhost:8080</code></li><li>• Database file DuckDB disimpan di volume Docker</li></ul>
<b>2.1 DAG Pipeline ETL</b> <ul style="list-style-type: none"><li>• Konversi pipeline ETL Anda dari tugas sebelumnya menjadi DAG Airflow</li><li>• Implementasikan dependensi tugas yang tepat menggunakan salah satu metode berikut:<ul style="list-style-type: none"><li>◦ Operator tradisional dengan operator <code>&gt;&gt;</code> dan <code>&lt;&lt;</code></li><li>◦ TaskFlow API dengan <i>decorators</i>(opsional)</li></ul></li><li>• Pastikan DAG Anda mencakup komponen berikut:<ul style="list-style-type: none"><li>◦ Tugas ekstraksi untuk setiap sumber data</li><li>◦ Tugas transformasi seperti yang didefinisikan dalam ETL sebelumnya</li><li>◦ Tugas <i>loading</i> data untuk mengisi data warehouse DuckDB Anda</li></ul></li></ul>	<ul style="list-style-type: none"><li>• Membaca sumber data dari CSV</li><li>• DuckDB digunakan sebagai target data warehouse</li><li>• DuckDB terhubung langsung dalam kode Python (tanpa Airflow hook)</li><li>• Gunakan <code>&gt;&gt;</code> atau TaskFlow API – <code>load(transform(extract()))</code></li><li>• Lebih lengkap ada pipeline</li></ul>

<ul style="list-style-type: none"> <li>○ Setidaknya satu <i>sensor</i> untuk memeriksa ketersediaan data</li> </ul>	
<p><b>2.2 Fitur Airflow Lanjutan</b></p> <ul style="list-style-type: none"> <li>● Implementasikan setidaknya dua dari fitur Airflow berikut: <ul style="list-style-type: none"> <li>○ Branching logic based on conditions</li> <li>○ Dynamic task generation</li> <li>○ Custom operators for specific business logic</li> <li>○ Error handling and retry mechanisms</li> <li>○ Email or Slack notifications for success/failure</li> <li>○ SLAs and monitoring setup</li> <li>○ Implementation of backfilling capabilities</li> </ul> </li> </ul>	<p>Error handling and retry mechanisms Implementation of backfilling capabilities</p>
<p><b>2.3 Strategi Penjadwalan dan Partisi</b></p> <ul style="list-style-type: none"> <li>● Rancang strategi <i>scheduling</i> yang tepat untuk DAG yang kalian buat</li> <li>● Dokumentasikan keputusan penjadwalan dan alasan kalian</li> </ul>	<ul style="list-style-type: none"> <li>● <code>schedule_interval set – 0 2 * * *</code> → daily at 2 AM</li> <li>● Uses <code>catchup=False</code></li> <li>● Backfilling supported run past dates via CLI/UI</li> </ul>
<p><b>3.1 Pengujian DAG</b></p> <ul style="list-style-type: none"> <li>● Uji DAG Anda menggunakan kemampuan pengujian bawaan Airflow</li> <li>● Demonstrasikan eksekusi yang berhasil dari DAG Anda dengan log yang tepat</li> </ul>	<p>Ini ada di folder dag, dan log-nya terlampir</p>

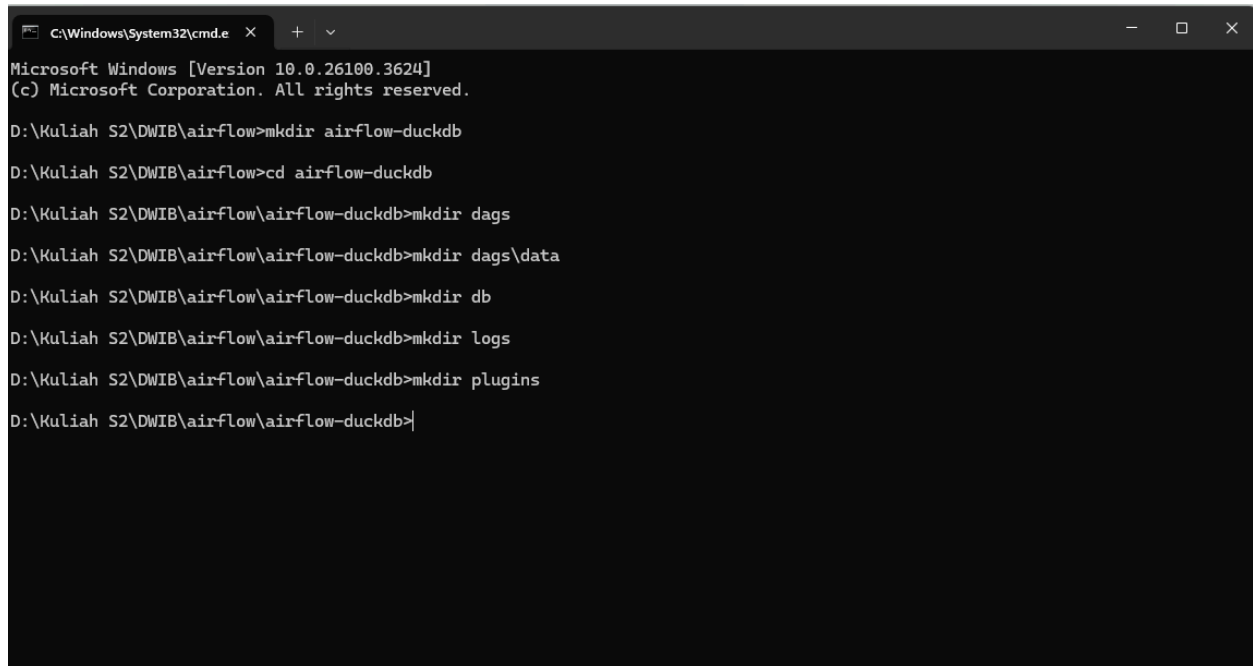
<p><b>3.2 Dokumentasi</b></p> <ul style="list-style-type: none"> <li>• Buat dokumentasi komprehensif untuk implementasi Airflow Anda: <ul style="list-style-type: none"> <li>◦ Diagram arsitektur yang menunjukkan dependensi tugas</li> <li>◦ Deskripsi tujuan setiap tugas</li> <li>◦ Informasi penjadwalan dan dependensi</li> <li>◦ Pengaturan pemantauan dan peringatan</li> <li>◦ Prosedur pemulihan kegagalan</li> </ul> </li> </ul>	
<p><b>Bagian 4: Kualitas Data</b></p> <ul style="list-style-type: none"> <li>• Implementasikan setidaknya dua pemeriksaan kualitas data (bisa menggunakan Airflow, Great Expectation, atau Python Native).</li> <li>• Buat DAG terpisah untuk memantau kualitas data</li> <li>• Dokumentasikan metrik kualitas data dan ambang batas (threshold)</li> </ul>	<ul style="list-style-type: none"> <li>• dag/data_quality_check.py</li> <li>• docs/data_quality_metrics</li> </ul>

## Create structure folder

```
mkdir airflow-duckdb
cd airflow-duckdb
```

```
mkdir dags
```

```
mkdir dags\data
mkdir db
mkdir logs
mkdir plugins
```



```
C:\Windows\System32\cmd.e
Microsoft Windows [Version 10.0.26100.3624]
(c) Microsoft Corporation. All rights reserved.

D:\Kuliah S2\DWIB\airflow>mkdir airflow-duckdb

D:\Kuliah S2\DWIB\airflow>cd airflow-duckdb

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>mkdir dags

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>mkdir dags\data

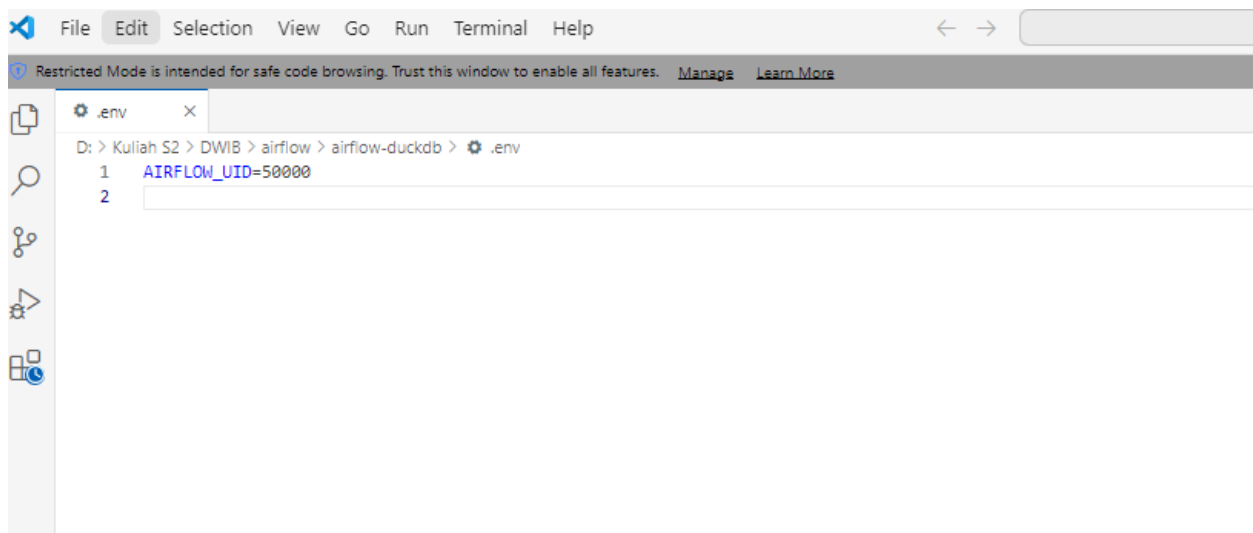
D:\Kuliah S2\DWIB\airflow\airflow-duckdb>mkdir db

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>mkdir logs

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>mkdir plugins

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>|
```

## Create .env file



```
File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

.env
D: > Kuliah S2 > DWIB > airflow > airflow-duckdb > .env
1 AIRFLOW_UID=50000
2
```

# Create Docker Compose yaml

## Initialize airflow database

docker compose up airflow-webserver airflow-scheduler --build

```
C:\Windows\System32\cmd.exe X + v
Microsoft Windows [Version 10.0.26100.3624]
(c) Microsoft Corporation. All rights reserved.

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>echo AIRFLOW_UID=50000 > .env

D:\Kuliah S2\DWIB\airflow\airflow-duckdb>docker compose up airflow-webserver airflow-scheduler --build
time="2025-04-05T08:38:22+07:00" level=warning msg="D:\\Kuliah S2\\DWIB\\airflow\\airflow-duckdb\\docker-compose.yaml: the attribute 'version' is obsolete, it will be ignored, please remove it to avoid potential confusion"
[+] Running 39/39
  ✓ airflow-scheduler Pulled                                197.1s
  ✓ airflow-webserver Pulled                                197.1s
  ✓ postgres Pulled                                         127.2s
[+] Running 5/5
  ✓ Network airflow-duckdb_default                          0.1s
  ✓ Volume "airflow-duckdb-postgres-db-volume"              0.0s
  ✓ Container airflow-duckdb-postgres-1                    0.4s
  ✓ Container airflow-duckdb-airflow-webserver-1            0.1s
  ✓ Container airflow-duckdb-airflow-scheduler-1            0.1s
Attaching to airflow-scheduler-1, airflow-webserver-1
airflow-webserver-1 |
airflow-scheduler-1 |
airflow-webserver-1 | /home/airflow/.local/lib/python3.8/site-packages/airflow/configuration.py:812 DeprecationWarning: The
airflow-webserver-1 | sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been us
airflow-webserver-1 | ed, but please update your config.
airflow-webserver-1 | /home/airflow/.local/lib/python3.8/site-packages/airflow/configuration.py:738 DeprecationWarning: The
airflow-webserver-1 | sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been us
airflow-webserver-1 | ed, but please update your config.
airflow-webserver-1 | /home/airflow/.local/lib/python3.8/site-packages/airflow/settings.py:194 DeprecationWarning: The sql_
airflow-webserver-1 | alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, b
```

In case error:

docker compose run --rm airflow-webserver airflow db init

Then restart:

docker compose up

## Create admin user

docker compose run airflow-webserver airflow users create

--username admin --password admin --firstname Admin

--lastname User --role Admin --email admin@example.com

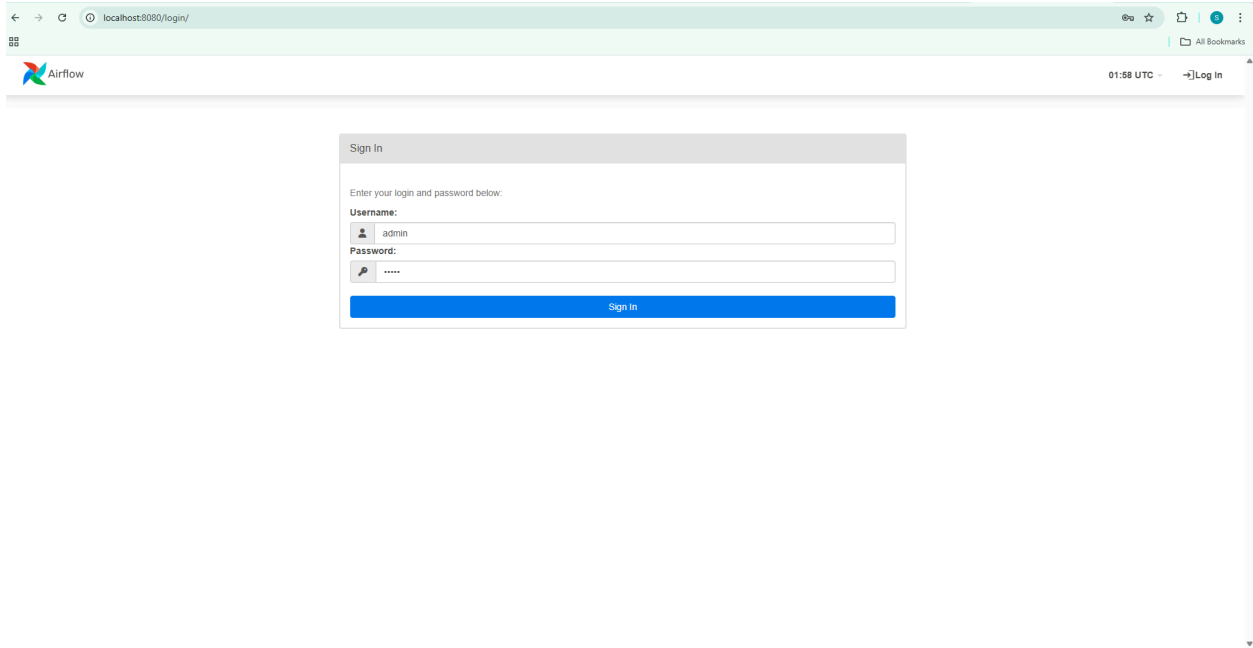
```
C:\Windows\System32\cmd.exe X Windows PowerShell X + v
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\silva> cd "D:\Kuliah S2\DWIB\airflow\airflow-duckdb"
PS D:\Kuliah S2\DWIB\airflow\airflow-duckdb> docker compose run airflow-webserver airflow users create --username admin --password a
dmin --firstname Admin --lastname User --role Admin --email admin@example.com
time="2025-04-05T08:47:54+07:00" level=warning msg="D:\Kuliah S2\DWIB\airflow\airflow-duckdb\docker-compose.yaml: the attribute
'version' is obsolete, it will be ignored, please remove it to avoid potential confusion"
time="2025-04-05T08:47:54+07:00" level=warning msg="Found orphan containers ([airflow-duckdb-airflow-webserver-run-58f3c9bf6959]) fo
r this project. If you removed or renamed this service in your compose file, you can run this command with the --remove-orphans flag
to clean it up."
[+] Creating 1/1
✔ Container airflow-duckdb-postgres-1 Running 0.0s

/home/airflow/.local/lib/python3.8/site-packages/airflow/configuration.py:812 DeprecationWarning: The sql_alchemy_conn option in [co
re] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.8/site-packages/airflow/configuration.py:738 DeprecationWarning: The sql_alchemy_conn option in [co
re] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.8/site-packages/airflow/settings.py:194 DeprecationWarning: The sql_alchemy_conn option in [core] h
as been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.8/site-packages/airflow/models/base.py:71 DeprecationWarning: The sql_alchemy_conn option in [core]
has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.

Please confirm database initialize (or wait 4 seconds to skip it). Are you sure? [y/N]
[2025-04-05T01:48:04.650+0000] {db.py:897} INFO - Log template table does not exist (added in 2.3.0); skipping log template sync.
y/home/airflow/.local/lib/python3.8/site-packages/flask_limiter/extension.py:336 UserWarning: Using the in-memory storage for tracki
ng rate limits as no storage was explicitly specified. This is not recommended for production use. See: https://flask-limiter.readth
edocs.io#configuring-a-storage-backend for documentation about configuring the storage backend.
[2025-04-05T01:48:04.935+0000] {override.py:855} INFO - Security DB not found Creating all Models from Base
[2025-04-05T01:48:05.537+0000] {override.py:857} INFO - Security DB Created
[2025-04-05T01:48:05.549+0000] {override.py:1369} INFO - Inserted Role: Admin
[2025-04-05T01:48:05.555+0000] {override.py:1369} INFO - Inserted Role: Public
[2025-04-05T01:48:05.558+0000] {override.py:868} WARNING - No user yet created, use flask fab command to do it.
```



In case error on init database and cannot go to airflow webpage do:

Stop container:

Ctrl + C

Cleanup:

`docker compose down --remove-orphans`

Initialize metadata DB:

`docker compose run airflow-webserver airflow db init`

Start service again:

`docker compose up`

# Implement the ETL DAG

Put dataset in — airflow-duckdb/dags/data/Online\_Retail.csv

Create new file - airflow-duckdb/dags/etl\_pipeline\_full.py

In case error duckdb module not found

Create file name Dockerfile (No extension) in airflow-duckdb folder and paste:  
FROM apache/airflow:2.8.1

USER airflow

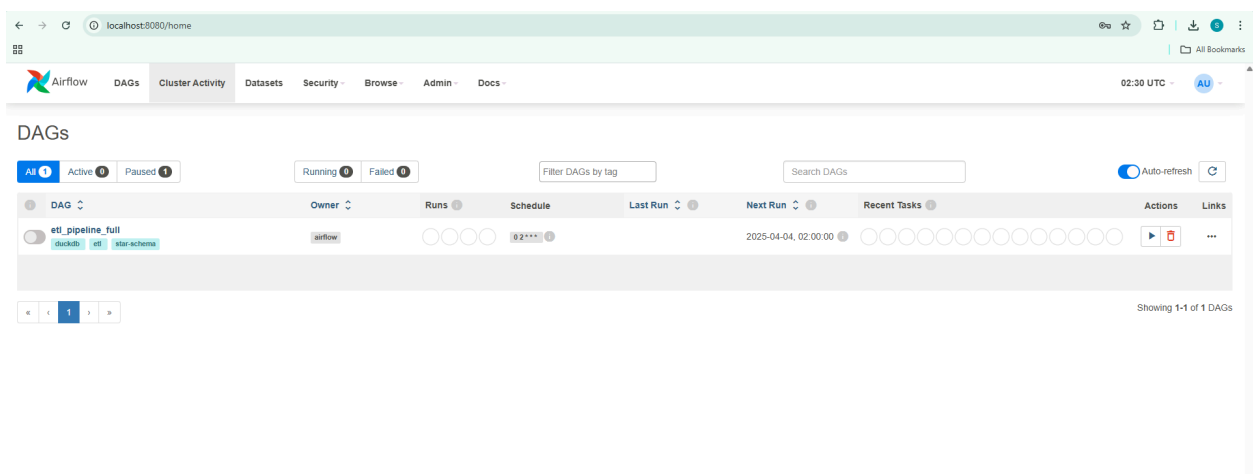
RUN pip install duckdb pandas

And the run this command

docker compose down --volumes --remove-orphans

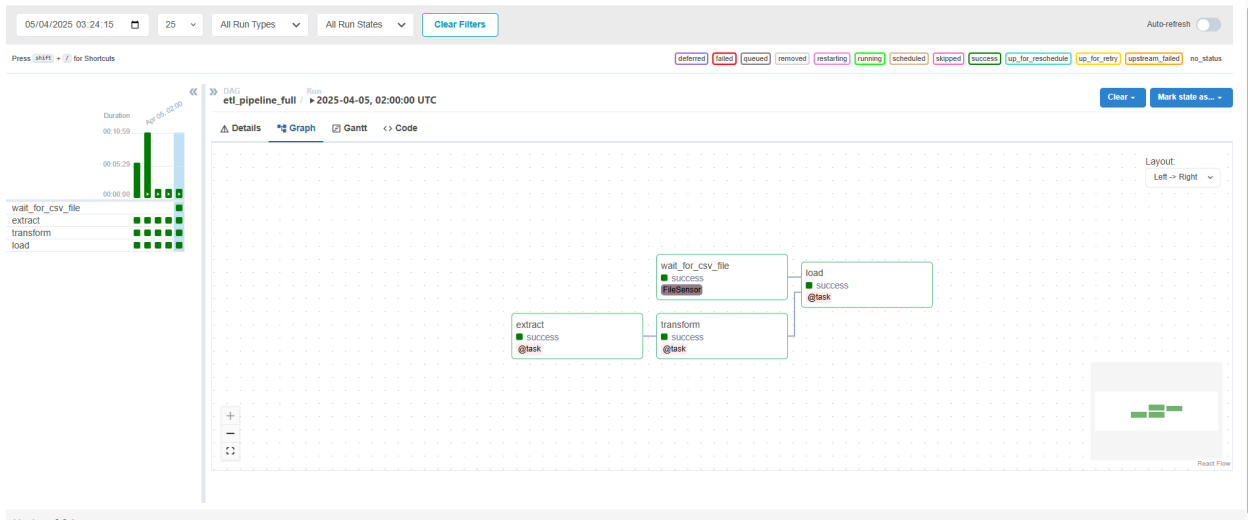
docker compose build

docker compose up



Running ETL pipeline is success





## Log sensor

decd9476825e

\*\*\* Found local files:

\*\*\* \*

/opt/airflow/logs/dag\_id=etl\_pipeline\_full/run\_id>manual\_\_2025-04-05T03:24:10.659781+00:00/task\_id=wait\_for\_csv\_file/attempt=1.log

[2025-04-05, 03:24:12 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep\_context=non-requeueable deps ti=<TaskInstance: etl\_pipeline\_full.wait\_for\_csv\_file manual\_\_2025-04-05T03:24:10.659781+00:00 [queued]>

[2025-04-05, 03:24:12 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep\_context=requeueable deps ti=<TaskInstance: etl\_pipeline\_full.wait\_for\_csv\_file manual\_\_2025-04-05T03:24:10.659781+00:00 [queued]>

[2025-04-05, 03:24:12 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 3

[2025-04-05, 03:24:12 UTC] {taskinstance.py:2191} INFO - Executing <Task(FileSensor):

wait\_for\_csv\_file> on 2025-04-05 03:24:10.659781+00:00

[2025-04-05, 03:24:12 UTC] {standard\_task\_runner.py:60} INFO - Started process 672 to run task

[2025-04-05, 03:24:12 UTC] {standard\_task\_runner.py:87} INFO - Running: ['\*\*\*', 'tasks', 'run', 'etl\_pipeline\_full', 'wait\_for\_csv\_file', 'manual\_\_2025-04-05T03:24:10.659781+00:00', '--job-id', '26', '--raw', '--subdir', 'DAGS\_FOLDER/etl\_pipeline\_full.py', '--cfg-path', '/tmp/tmpum6ldovg']

[2025-04-05, 03:24:12 UTC] {standard\_task\_runner.py:88} INFO - Job 26: Subtask

wait\_for\_csv\_file

[2025-04-05, 03:24:12 UTC] {logging\_mixin.py:188} WARNING -

/home/\*\*\*/.local/lib/python3.8/site-packages/\*\*\*/settings.py:194 DeprecationWarning: The sql\_alchemy\_conn option in [core] has been moved to the sql\_alchemy\_conn option in [database] - the old setting has been used, but please update your config.

[2025-04-05, 03:24:12 UTC] {task\_command.py:423} INFO - Running <TaskInstance: etl\_pipeline\_full.wait\_for\_csv\_file manual\_\_2025-04-05T03:24:10.659781+00:00 [running]> on host decd9476825e

[2025-04-05, 03:24:12 UTC] {taskinstance.py:2480} INFO - Exporting env vars:

AIRFLOW\_CTX\_DAG\_OWNER='\*\*\*' AIRFLOW\_CTX\_DAG\_ID='etl\_pipeline\_full'

AIRFLOW\_CTX\_TASK\_ID='wait\_for\_csv\_file'

AIRFLOW\_CTX\_EXECUTION\_DATE='2025-04-05T03:24:10.659781+00:00' AIRFLOW\_CTX\_TRY\_NUMBER='1'

AIRFLOW\_CTX\_DAG\_RUN\_ID='manual\_\_2025-04-05T03:24:10.659781+00:00'

[2025-04-05, 03:24:12 UTC] {base.py:83} INFO - Using connection ID 'fs\_default' for task execution.

[2025-04-05, 03:24:12 UTC] {filesystem.py:66} INFO - Poking for file

/opt/\*\*\*/dags/data/Online\_Retail.csv

```
[2025-04-05, 03:24:12 UTC] {filesystem.py:71} INFO - Found File
/opt/***/dags/data/Online_Retail.csv last modified: 20250405020601
[2025-04-05, 03:24:12 UTC] {base.py:295} INFO - Success criteria met. Exiting.
[2025-04-05, 03:24:12 UTC] {taskinstance.py:1138} INFO - Marking task as SUCCESS.
dag_id=etl_pipeline_full, task_id=wait_for_csv_file, execution_date=20250405T032410,
start_date=20250405T032412, end_date=20250405T032412
[2025-04-05, 03:24:12 UTC] {local_task_job_runner.py:234} INFO - Task exited with return code
0
[2025-04-05, 03:24:12 UTC] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from
follow-on schedule check
```

## Log Extract

All Levels ▼ All File Sources ▼ Wrap Download See More

▲ Large log file. Some lines have been truncated. Download logs in order to see everything.

```
dec9476825e
*** Found local files:
*** * /opt/airflow/logs/dag_id=etl_pipeline_full/run_id=manual_2025-04-05T02:36:24.327366+00:00/task_id=extract/attempt=1.log
[2025-04-05, 02:36:25 UTC] [taskinstance.py:1956] INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: etl_pipeline_full.extract manual_2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:25 UTC] [taskinstance.py:1956] INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: etl_pipeline_full.extract manual_2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:25 UTC] [taskinstance.py:2170] INFO - Starting attempt 1 of 3
[2025-04-05, 02:36:25 UTC] [taskinstance.py:2191] INFO - Executing <task(PythonDecoratedOperator): extract> on 2025-04-05 02:36:24.327366+00:00
[2025-04-05, 02:36:25 UTC] [standard_task_runner.py:60] INFO - Started process 175 to run task
[2025-04-05, 02:36:25 UTC] [standard_task_runner.py:87] INFO - Running: ['***', 'tasks', 'run', 'etl_pipeline_full', 'extract', 'manual_2025-04-05T02:36:24.327366+00:00', '--job-id', '7', '--raw', '--subdir', 'DAGS_FOLDER/etl_pipeline_full.py', '--cfg-path', '/t
[2025-04-05, 02:36:25 UTC] [standard_task_runner.py:88] INFO - Job 7: Subtask extract
[2025-04-05, 02:36:25 UTC] [logging_mixin.py:180] WARNING - /home/***/.local/lib/python3.8/site-packages/***/settings.py:194 DeprecationWarning: The sqlalchemy_conn option in [core] has been moved to the sqlalchemy_conn option in [database] - the old setting ha
[2025-04-05, 02:36:25 UTC] [task_command.py:423] INFO - Running <taskinstance: etl_pipeline_full.extract manual_2025-04-05T02:36:24.327366+00:00 [running]> on host dec9476825e
[2025-04-05, 02:36:26 UTC] [taskinstance.py:2480] INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER=*** AIRFLOW_CTX_DAG_ID='etl_pipeline_full' AIRFLOW_CTX_TASK_ID='extract' AIRFLOW_CTX_EXECUTION_DATE='2025-04-05T02:36:24.327366+00:00' AIRFLOW_CTX_TRY_NUMBER='1'
[2025-04-05, 02:36:27 UTC] [logging_mixin.py:180] WARNING - /opt/***/dags/etl_pipeline_full.py:29 UserWarning: Could not infer format, so each element will be parsed individually, falling back to 'dateutil'. To ensure parsing is consistent and as-expected, please
[2025-04-05, 02:36:31 UTC] [python.py:201] INFO - Done. Returned value was: {"columns":["InvoiceId","StockCode","Description","Quantity","InvoiceDate","UnitPrice","CustomerId","Country","Revenue"],"index":["0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,2
[2025-04-05, 02:36:35 UTC] [taskinstance.py:1138] INFO - Marking task as SUCCESS. dag_id=etl_pipeline_full, task_id=extract, execution_date=20250405T023624, start_date=20250405T023625, end_date=20250405T023635
[2025-04-05, 02:36:35 UTC] [local_task_job_runner.py:234] INFO - Task exited with return code 0
[2025-04-05, 02:36:36 UTC] [taskinstance.py:3280] INFO - 1 downstream tasks scheduled from follow-on schedule check
```

## Log transform

▲ Large log file. Some lines have been truncated. Download logs in order to see everything.

```
dec9476825e
*** Found local files:
*** * /opt/airflow/logs/dag_id=etl_pipeline_full/run_id=manual_2025-04-05T02:36:24.327366+00:00/task_id=transform/attempt=1.log
[2025-04-05, 02:36:37 UTC] [taskinstance.py:1956] INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: etl_pipeline_full.transform manual_2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:37 UTC] [taskinstance.py:1956] INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: etl_pipeline_full.transform manual_2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:37 UTC] [taskinstance.py:2170] INFO - Starting attempt 1 of 3
[2025-04-05, 02:36:37 UTC] [taskinstance.py:2191] INFO - Executing <task(PythonDecoratedOperator): transform> on 2025-04-05 02:36:24.327366+00:00
[2025-04-05, 02:36:37 UTC] [standard_task_runner.py:60] INFO - Started process 184 to run task
[2025-04-05, 02:36:37 UTC] [standard_task_runner.py:87] INFO - Running: ['***', 'tasks', 'run', 'etl_pipeline_full', 'transform', 'manual_2025-04-05T02:36:24.327366+00:00', '--job-id', '8', '--raw', '--subdir', 'DAGS_FOLDER/etl_pipeline_full.py', '--cfg-path', '
[2025-04-05, 02:36:37 UTC] [standard_task_runner.py:88] INFO - Job 8: Subtask transform
[2025-04-05, 02:36:37 UTC] [logging_mixin.py:180] WARNING - /home/***/.local/lib/python3.8/site-packages/***/settings.py:194 DeprecationWarning: The sqlalchemy_conn option in [core] has been moved to the sqlalchemy_conn option in [database] - the old setting ha
[2025-04-05, 02:36:37 UTC] [task_command.py:423] INFO - Running <taskinstance: etl_pipeline_full.transform manual_2025-04-05T02:36:24.327366+00:00 [running]> on host dec9476825e
[2025-04-05, 02:36:41 UTC] [taskinstance.py:2480] INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER=*** AIRFLOW_CTX_DAG_ID='etl_pipeline_full' AIRFLOW_CTX_TASK_ID='transform' AIRFLOW_CTX_EXECUTION_DATE='2025-04-05T02:36:24.327366+00:00' AIRFLOW_CTX_TRY_NUMBER='1'
[2025-04-05, 02:36:46 UTC] [python.py:201] INFO - Done. Returned value was: {'date': '{"columns":["InvoiceDate","DateKey","FullDate","Year","Month","Day","Quarter","DayOfWeek"],"index":["0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,
[2025-04-05, 02:36:48 UTC] [taskinstance.py:1138] INFO - Marking task as SUCCESS. dag_id=etl_pipeline_full, task_id=transform, execution_date=20250405T023624, start_date=20250405T023637, end_date=20250405T023648
[2025-04-05, 02:36:48 UTC] [local_task_job_runner.py:234] INFO - Task exited with return code 0
[2025-04-05, 02:36:48 UTC] [taskinstance.py:3280] INFO - 1 downstream tasks scheduled from follow-on schedule check
```

## Log Load

dec9476825e

```
*** Found local files:
*** * /opt/airflow/logs/dag_id=etl_pipeline_full/run_id=manual_2025-04-05T02:36:24.327366+00:00/task_id=load/attempt=1.log
[2025-04-05, 02:36:49 UTC] [taskinstance.py:1956] INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: etl_pipeline_full.load manual_2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:49 UTC] [taskinstance.py:1956] INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: etl_pipeline_full.load manual_2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:49 UTC] [taskinstance.py:2170] INFO - Starting attempt 1 of 3
[2025-04-05, 02:36:49 UTC] [taskinstance.py:2191] INFO - Executing <task(PythonDecoratedOperator): load> on 2025-04-05 02:36:24.327366+00:00
[2025-04-05, 02:36:49 UTC] [standard_task_runner.py:60] INFO - Started process 187 to run task
[2025-04-05, 02:36:49 UTC] [standard_task_runner.py:87] INFO - Running: ['***', 'tasks', 'run', 'etl_pipeline_full', 'load', 'manual_2025-04-05T02:36:24.327366+00:00', '--job-id', '9', '--raw', '--subdir', 'DAGS_FOLDER/etl_pipeline_full.py', '--cfg-path', '/tmp/
[2025-04-05, 02:36:49 UTC] [standard_task_runner.py:88] INFO - Job 9: Subtask load
[2025-04-05, 02:36:49 UTC] [logging_mixin.py:180] WARNING - /home/***/.local/lib/python3.8/site-packages/***/settings.py:194 DeprecationWarning: The sqlalchemy_conn option in [core] has been moved to the sqlalchemy_conn option in [database] - the old setting ha
[2025-04-05, 02:36:51 UTC] [taskinstance.py:2480] INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER=*** AIRFLOW_CTX_DAG_ID='etl_pipeline_full' AIRFLOW_CTX_TASK_ID='load' AIRFLOW_CTX_EXECUTION_DATE='2025-04-05T02:36:24.327366+00:00' AIRFLOW_CTX_TRY_NUMBER='1' AIR
[2025-04-05, 02:36:55 UTC] [python.py:201] INFO - Done. Returned value was: None
[2025-04-05, 02:36:55 UTC] [taskinstance.py:1138] INFO - Marking task as SUCCESS. dag_id=etl_pipeline_full, task_id=load, execution_date=20250405T023624, start_date=20250405T023649, end_date=20250405T023655
[2025-04-05, 02:36:55 UTC] [local_task_job_runner.py:234] INFO - Task exited with return code 0
[2025-04-05, 02:36:55 UTC] [taskinstance.py:3280] INFO - 0 downstream tasks scheduled from follow-on schedule check
```

dec9476825e

\*\*\* Found local files:

```

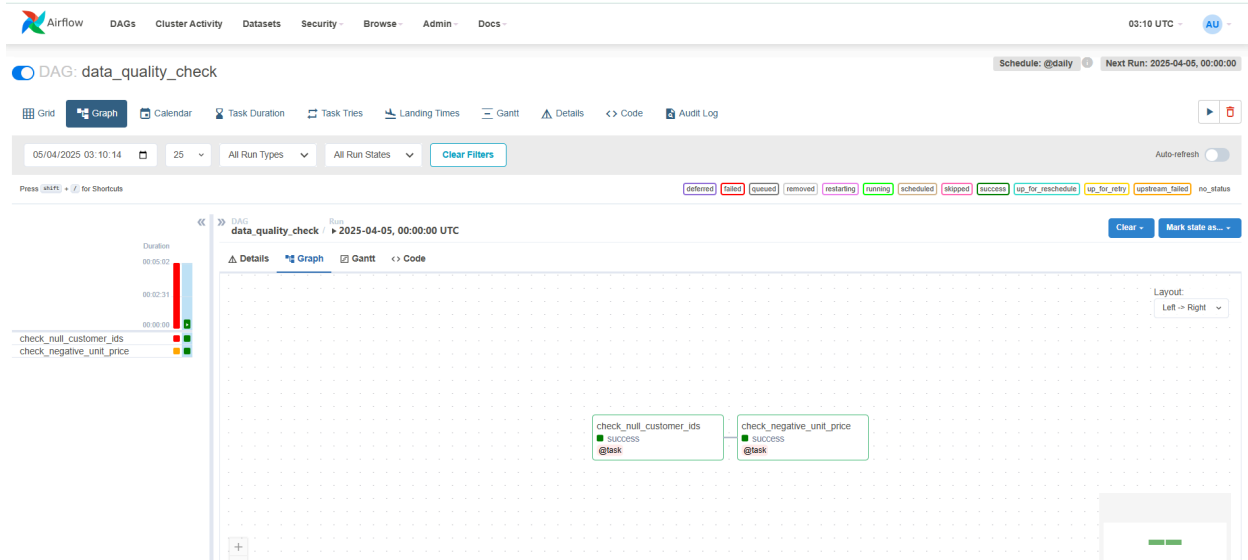
*** *
/opt/airflow/logs/dag_id=etl_pipeline_full/run_id>manual__2025-04-05T02:36:24.327366+00:00/task_id=load/attempt=1.log
[2025-04-05, 02:36:49 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: etl_pipeline_full.load manual__2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:49 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: etl_pipeline_full.load manual__2025-04-05T02:36:24.327366+00:00 [queued]>
[2025-04-05, 02:36:49 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 3
[2025-04-05, 02:36:49 UTC] {taskinstance.py:2191} INFO - Executing <Task(PythonDecoratedOperator): load> on 2025-04-05 02:36:24.327366+00:00
[2025-04-05, 02:36:49 UTC] {standard_task_runner.py:60} INFO - Started process 187 to run task
[2025-04-05, 02:36:49 UTC] {standard_task_runner.py:87} INFO - Running: ['***', 'tasks', 'run', 'etl_pipeline_full', 'load', 'manual__2025-04-05T02:36:24.327366+00:00', '--job-id', '9', '--raw', '--subdir', 'DAGS_FOLDER/etl_pipeline_full.py', '--cfg-path', '/tmp/tmp6hxoleiy']
[2025-04-05, 02:36:49 UTC] {standard_task_runner.py:88} INFO - Job 9: Subtask load
[2025-04-05, 02:36:49 UTC] {logging_mixin.py:188} WARNING - /home/***/.local/lib/python3.8/site-packages/***/settings.py:194 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.
[2025-04-05, 02:36:49 UTC] {task_command.py:423} INFO - Running <TaskInstance: etl_pipeline_full.load manual__2025-04-05T02:36:24.327366+00:00 [running]> on host decd9476825e
[2025-04-05, 02:36:51 UTC] {taskinstance.py:2480} INFO - Exporting env vars:
AIRFLOW_CTX_DAG_OWNER='***' AIRFLOW_CTX_DAG_ID='etl_pipeline_full' AIRFLOW_CTX_TASK_ID='load'
AIRFLOW_CTX_EXECUTION_DATE='2025-04-05T02:36:24.327366+00:00' AIRFLOW_CTX_TRY_NUMBER='1'
AIRFLOW_CTX_DAG_RUN_ID='manual__2025-04-05T02:36:24.327366+00:00'
[2025-04-05, 02:36:55 UTC] {python.py:201} INFO - Done. Returned value was: None
[2025-04-05, 02:36:55 UTC] {taskinstance.py:1138} INFO - Marking task as SUCCESS. dag_id=etl_pipeline_full, task_id=load, execution_date=20250405T023624, start_date=20250405T023649, end_date=20250405T023655
[2025-04-05, 02:36:55 UTC] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-04-05, 02:36:55 UTC] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from follow-on schedule check

```

## Data quality

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
data_quality_check <small>data_quality</small>	airflow	0	@daily	2025-04-04, 00:00:00	2025-04-04, 00:00:00	0	[Run] [Cancel]	...
etl_pipeline_full <small>data_quality etl star-schema</small>	airflow	1	@2*	2025-04-05, 02:36:24	2025-04-05, 02:00:00	1	[Run] [Cancel]	...

Showing 1-2 of 2 DAGs



## Log Check null customer

decd9476825e

\*\*\* Found local files:

\*\*\* \*

/opt/airflow/logs/dag\_id=data\_quality\_check/run\_id>manual\_\_2025-04-05T03:10:10.558855+00:00/task\_id=check\_null\_customer\_ids/attempt=1.log

[2025-04-05, 03:10:11 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep\_context=non-requeueable deps ti=<TaskInstance: data\_quality\_check.check\_null\_customer\_ids manual\_\_2025-04-05T03:10:10.558855+00:00 [queued]>

[2025-04-05, 03:10:11 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep\_context=requeueable deps ti=<TaskInstance: data\_quality\_check.check\_null\_customer\_ids manual\_\_2025-04-05T03:10:10.558855+00:00 [queued]>

[2025-04-05, 03:10:11 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 2

[2025-04-05, 03:10:11 UTC] {taskinstance.py:2191} INFO - Executing

<Task( PythonDecoratedOperator): check\_null\_customer\_ids> on 2025-04-05 03:10:10.558855+00:00

[2025-04-05, 03:10:11 UTC] {standard\_task\_runner.py:60} INFO - Started process 342 to run task

[2025-04-05, 03:10:11 UTC] {standard\_task\_runner.py:87} INFO - Running: ['\*\*\*', 'tasks',

'run', 'data\_quality\_check', 'check\_null\_customer\_ids',

'manual\_\_2025-04-05T03:10:10.558855+00:00', '--job-id', '23', '--raw', '--subdir',

'DAGS\_FOLDER/data\_quality\_check.py', '--cfg-path', '/tmp/tmpvjfzt012']

[2025-04-05, 03:10:11 UTC] {standard\_task\_runner.py:88} INFO - Job 23: Subtask

check\_null\_customer\_ids

[2025-04-05, 03:10:11 UTC] {logging\_mixin.py:188} WARNING -

/home/\*\*\*/.local/lib/python3.8/site-packages/\*\*\*/settings.py:194 DeprecationWarning: The

sql\_alchemy\_conn option in [core] has been moved to the sql\_alchemy\_conn option in [database]

- the old setting has been used, but please update your config.

[2025-04-05, 03:10:11 UTC] {task\_command.py:423} INFO - Running <TaskInstance: data\_quality\_check.check\_null\_customer\_ids manual\_\_2025-04-05T03:10:10.558855+00:00 [running]> on host decd9476825e

[2025-04-05, 03:10:11 UTC] {taskinstance.py:2480} INFO - Exporting env vars:

AIRFLOW\_CTX\_DAG\_OWNER='\*\*\*' AIRFLOW\_CTX\_DAG\_ID='data\_quality\_check'

AIRFLOW\_CTX\_TASK\_ID='check\_null\_customer\_ids'

AIRFLOW\_CTX\_EXECUTION\_DATE='2025-04-05T03:10:10.558855+00:00' AIRFLOW\_CTX\_TRY\_NUMBER='1'

AIRFLOW\_CTX\_DAG\_RUN\_ID='manual\_\_2025-04-05T03:10:10.558855+00:00'

[2025-04-05, 03:10:11 UTC] {logging\_mixin.py:188} INFO - CustomerKey NULL check passed.

[2025-04-05, 03:10:11 UTC] {python.py:201} INFO - Done. Returned value was: None

[2025-04-05, 03:10:11 UTC] {taskinstance.py:1138} INFO - Marking task as SUCCESS.

dag\_id=data\_quality\_check, task\_id=check\_null\_customer\_ids, execution\_date=20250405T031010,

start\_date=20250405T031011, end\_date=20250405T031011

```
[2025-04-05, 03:10:11 UTC] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-04-05, 03:10:11 UTC] {taskinstance.py:3280} INFO - 1 downstream tasks scheduled from follow-on schedule check
```

## Log Check Negative Unit

```
decd9476825e
*** Found local files:
*** *
/opt/airflow/logs/dag_id=data_quality_check/run_id>manual__2025-04-05T03:10:10.558855+00:00/task_id=check_negative_unit_price/attempt=1.log
[2025-04-05, 03:10:12 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: data_quality_check.check_negative_unit_price manual__2025-04-05T03:10:10.558855+00:00 [queued]>
[2025-04-05, 03:10:12 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: data_quality_check.check_negative_unit_price manual__2025-04-05T03:10:10.558855+00:00 [queued]>
[2025-04-05, 03:10:12 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 2
[2025-04-05, 03:10:12 UTC] {taskinstance.py:2191} INFO - Executing <Task(PythonDecoratedOperator): check_negative_unit_price> on 2025-04-05 03:10:10.558855+00:00
[2025-04-05, 03:10:12 UTC] {standard_task_runner.py:60} INFO - Started process 351 to run task
[2025-04-05, 03:10:12 UTC] {standard_task_runner.py:87} INFO - Running: ['***', 'tasks', 'run', 'data_quality_check', 'check_negative_unit_price', 'manual__2025-04-05T03:10:10.558855+00:00', '--job-id', '24', '--raw', '--subdir', 'DAGS_FOLDER/data_quality_check.py', '--cfg-path', '/tmp/tmpakq4t32']
[2025-04-05, 03:10:12 UTC] {standard_task_runner.py:88} INFO - Job 24: Subtask check_negative_unit_price
[2025-04-05, 03:10:12 UTC] {logging_mixin.py:188} WARNING - /home/***/.local/lib/python3.8/site-packages/***/settings.py:194 DeprecationWarning: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been used, but please update your config.
[2025-04-05, 03:10:12 UTC] {task_command.py:423} INFO - Running <TaskInstance: data_quality_check.check_negative_unit_price manual__2025-04-05T03:10:10.558855+00:00 [running]> on host decd9476825e
[2025-04-05, 03:10:12 UTC] {taskinstance.py:2480} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='***' AIRFLOW_CTX_DAG_ID='data_quality_check' AIRFLOW_CTX_TASK_ID='check_negative_unit_price' AIRFLOW_CTX_EXECUTION_DATE='2025-04-05T03:10:10.558855+00:00' AIRFLOW_CTX_TRY_NUMBER='1' AIRFLOW_CTX_DAG_RUN_ID='manual__2025-04-05T03:10:10.558855+00:00'
[2025-04-05, 03:10:12 UTC] {logging_mixin.py:188} INFO - ✅ UnitPrice negative check passed.
[2025-04-05, 03:10:12 UTC] {python.py:201} INFO - Done. Returned value was: None
[2025-04-05, 03:10:12 UTC] {taskinstance.py:1138} INFO - Marking task as SUCCESS. dag_id=data_quality_check, task_id=check_negative_unit_price, execution_date=20250405T031010, start_date=20250405T031012, end_date=20250405T031012
[2025-04-05, 03:10:12 UTC] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-04-05, 03:10:12 UTC] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

## Visualization

Alat	Alasan Rekomendasi
Metabase	Gratis, gampang diinstal, tinggal klik-klik buat dashboard
Power BI	Banyak fitur canggih buat analisis, bisa impor Excel langsung, cocok buat presentasi serius
Tableau	Visualnya keren dan interaktif, cocok buat nunjukin data ke dosen atau di seminar
DuckDB + Streamlit	Buat yang suka ngoding Python, bisa bikin dashboard interaktif sendiri dari nol

Contoh query yang akan di gunakan

Total Revenue per Month (using Date Dimension and Fact Sales)

```
SELECT d.Year, d.Month, SUM(f.Revenue) AS TotalRevenue
FROM retail.FactSales f
JOIN retail.DateDimension d ON f.DateKey = d.DateKey
GROUP BY d.Year, d.Month
ORDER BY d.Year, d.Month
```

Customer Lifetime Value (CLV): Histogram of total revenue per customer

```
SELECT c.CustomerID, SUM(f.Revenue) AS TotalRevenue
FROM retail.FactSales f
JOIN retail.CustomerDimension c ON f.CustomerKey = c.CustomerKey
GROUP BY c.CustomerID
```

Geographical Analysis: Revenue by Country (using InvoiceDimension)

```
SELECT Country, SUM(Revenue) AS TotalRevenue
FROM retail.FactSales f
JOIN retail.InvoiceDimension i ON f.InvoiceKey = i.InvoiceKey
GROUP BY Country
ORDER BY TotalRevenue DESC
```