



# News Data Pipeline: Final Report

Igenbek Zhaina 22B030368  
Spandiyar Kuanysh 22B030287  
Tolebek Nazym 22B030452

## Executive Summary

Successfully implemented a containerized data pipeline that ingests real-time news data from NewsAPI(<https://newsapi.org/>), processes it through Kafka, cleans with Pandas, stores in SQLite, and computes daily analytics - all orchestrated by Airflow.

## Architecture

### Three-Stage Pipeline:

1. Ingestion: NewsAPI -> Kafka (DAG 1 @hourly)
2. Cleaning: Kafka -> SQLite (DAG 2 @hourly)
3. Analytics: SQLite -> Daily Summary (DAG 3 @daily)

Technology Stack: NewsAPI, Apache Kafka, Apache Airflow, SQLite, Pandas, Docker Compose

## Results

- All DAGs operational in Airflow UI
- Database created (`app.db`) with correct schema
- Data flowing through all pipeline stages
- Containers running: Kafka, Zookeeper, Airflow, Postgres
- Daily analytics computed and stored

## Database Schema

events table: Cleaned article data (source, author, title, url, category, timestamps)

daily\_summary table: Aggregated metrics (article counts, source statistics, averages)

## Key Features

- Continuous ingestion with NewsAPI polling
- Pandas-only cleaning (no Python loops)
- Article categorization (Tech/Business/General)

- Containerized deployment with single command

## Validation

- DAGs execute on schedule (@hourly/@daily)
- Kafka messages successfully produced/consumed
- SQLite tables populated with clean data
- Daily summaries computed automatically
- All services run in Docker containers

## Success Criteria Met

- Done: Job 1: API -> Kafka continuous ingestion
- Done: Job 2: Kafka -> Clean with Pandas -> SQLite
- Done: Job 3: Daily analytics using Pandas
- Done: All components containerized
- Done: Complete documentation

## Appendices

**Screenshots were taken throughout the whole process. Before and after error fixings.**

Docker desktop:

The screenshot shows the Docker Desktop interface. The left sidebar has links for Ask Gordon (BETA), Containers (selected), Images, Volumes, Builds, Docker Hub, Docker Scout, and Extensions. The main area is titled 'Containers' with a 'Search' bar and a 'Give feedback' link. It displays resource usage: Container CPU usage (5.11% / 800%) and Container memory usage (2.03GB / 3.74GB). A chart button says 'Show charts'. Below is a table of running containers:

	Name	Container ID	Image	Port(s)	CPU (%)	Last started	Actions
news-pipeline	-	-	-	-	4.72%	12 hours ago	[actions]
zookeeper-1	06d42cc8818e	confluentinc/cp-zooke	2181:2181	5432:5432	0.2%	12 hours ago	[actions]
postgres-1	286243b8ed98	postgres:13	5432:5432	5432:5432	0.92%	12 hours ago	[actions]
airflow-init-1	02d3d7ef96b7	apache/airflow:2.7.0-p	5432:5432	5432:5432	0%	12 hours ago	[actions]
kafka-1	d34d8aa06209	confluentinc/cp-kafka:	9092:9092	5432:5432	0.51%	12 hours ago	[actions]
airflow-webserver-1	f66d64fd9cd3	apache/airflow:2.7.0-p	8080:8080	5432:5432	0.28%	12 hours ago	[actions]
airflow-scheduler-1	152303e5a456	apache/airflow:2.7.0-p	5432:5432	5432:5432	2.81%	12 hours ago	[actions]

At the bottom, status bars show 'Engine running', system resources (RAM 2.85 GB, CPU 0.62%, Disk 13.35 GB used / limit 1006.85 GB), and links for Terminal and New version available.

**docker:desktop PERSONAL**

**Containers / news-pipeline-airflow-init-1**

**news-pipeline-airflow-init-1** 02d3d7ef96b7 apache/airflow:2.7.0-python3.10

**STATUS** Exited (0) (12 hours ago)

**Logs** Inspect Bind mounts Exec Files Stats

```
2025-12-17 03:01:50 Traceback (most recent call last):
2025-12-17 03:01:50   File "/home/airflow/.local/lib/python3.10/site-packages/airflow/models/dagbag.py", line 343, in parse
2025-12-17 03:01:50     loader.exec_module(new_module)
2025-12-17 03:01:50   File "<frozen importlib._bootstrap_external>", line 883, in exec_module
2025-12-17 03:01:50   File "<frozen importlib._bootstrap>", line 241, in _call_with_frames_removed
2025-12-17 03:01:50   File "/opt/airflow/dags/job2_clean_store_dag.py", line 7, in <module>
2025-12-17 03:01:50     from job2_cleaner import clean_and_store
2025-12-17 03:01:50   File "/opt/airflow/src/job2_cleaner.py", line 3, in <module>
2025-12-17 03:01:50     from kafka import KafkaConsumer
2025-12-17 03:01:50 ModuleNotFoundError: No module named 'kafka'
2025-12-17 03:01:50 ERROR: [airflow.models.dagbag.DagBag] Failed to import: /opt/airflow/dags/job1_ingestion_dag.py
2025-12-17 03:01:50 Traceback (most recent call last):
2025-12-17 03:01:50   File "/home/airflow/.local/lib/python3.10/site-packages/airflow/models/dagbag.py", line 343, in parse
2025-12-17 03:01:50     loader.exec_module(new_module)
2025-12-17 03:01:50   File "<frozen importlib._bootstrap_external>", line 883, in exec_module
2025-12-17 03:01:50   File "<frozen importlib._bootstrap>", line 241, in _call_with_frames_removed
2025-12-17 03:01:50   File "/opt/airflow/dags/job1_ingestion_dag.py", line 7, in <module>
2025-12-17 03:01:50     from job1_producer import produce_to_kafka
2025-12-17 03:01:50   File "/opt/airflow/src/job1_producer.py", line 5, in <module>
2025-12-17 03:01:50     from kafka import KafkaProducer
2025-12-17 03:01:50 ModuleNotFoundError: No module named 'kafka'
2025-12-17 03:01:51 /home/airflow/.local/lib/python3.10/site-packages/Flask_limiter/extension.py:293 UserWarning: Using the in-memory storage for tracking rate limits as no storage was explicitly specified. This is not recommended for production use. See: https://flask-limiter.readthedocs.io/configuring-a-storage-backend for documentation about configuring the storage backend.
2025-12-17 03:01:47 DB: postgresql+psycopg2://airflow:***@postgres/airflow
2025-12-17 03:01:47 [2025-12-16T22:01:47.329+0000] {migration.py:213} INFO - Context impl PostgresSqlImpl.
2025-12-17 03:01:47 [2025-12-16T22:01:47.330+0000] {migration.py:216} INFO - Will assume transactional DDL.
2025-12-17 03:01:48 [2025-12-16T22:01:48.361+0000] {migration.py:213} INFO - Context impl PostgresSqlImpl.
2025-12-17 03:01:48 [2025-12-16T22:01:48.362+0000] {migration.py:216} INFO - Will assume transactional DDL.
2025-12-17 03:01:48 [2025-12-16T22:01:48.375+0000] {db.py:1633} INFO - Creating tables
2025-12-17 03:01:50 Initialization done
2025-12-17 03:01:52 admin already exist in the db
2025-12-17 03:01:52 Airflow initialized
```

Engine running | RAM 2.86 GB CPU 1.63% Disk: 13.35 GB used (limit 1006.85 GB) | Terminal | New version available

## DAGs:

**DAGs - Airflow** localhost:8080/home?status=active

**Auto-refresh**

All (3)	Active (3)	Paused (0)	Running (2)	Failed (1)	Filter DAGs by tag	Search DAGs	Actions	Links
<b>DAG</b>	<b>Owner</b>	<b>Runs</b>	<b>Schedule</b>	<b>Last Run</b>	<b>Next Run</b>	<b>Recent Tasks</b>		
job1_continuous_ingestion ingestion kafka newsapi	airflow	2	@hourly	2025-12-16, 18:00:00	2025-12-16, 19:00:00	2		
job2_hourly_cleaning cleaning kafka sqlite	airflow	2	@hourly	2025-12-16, 18:00:00	2025-12-16, 19:00:00	failed		
job3_daily_analytics analytics sqlite summary	airflow	1	@daily	2025-12-16, 00:00:00		1		

Showing 1-3 of 3 DAGs

Version: v2.7.0

localhost:8080/taskinstance/list/?\_ft\_3\_dag\_id=job2\_hourly\_cleaning&\_ft\_3\_state=failed

DAGs - Airflow

localhost:8080/home

Triggered job3\_daily\_analytics, it should start any moment now.

## DAGs

All 3	Active 3	Paused 0	Running 4	Failed 0	Filter DAGs by tag	Search DAGs	Auto-refresh	Links
DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	
job1_continuous_ingestion	airflow	3	@hourly	2025-12-16, 22:03:33	2025-12-16, 22:00:00	3		
job2_hourly_cleaning	airflow	3	@hourly	2025-12-16, 22:03:35	2025-12-16, 22:00:00	1		
job3_daily_analytics	airflow	3	@daily	2025-12-16, 22:03:37	2025-12-16, 00:00:00	1		

Showing 1-3 of 3 DAGs

Version: v2.7.0  
Git Version: release:c08c82e9dd0e4aabaa5121519819a636df635210

DAGs - Airflow

localhost:8080/home

The scheduler does not appear to be running. Last heartbeat was received 6 minutes ago.  
The DAGs list may not update, and new tasks will not be scheduled.

## DAGs

All 3	Active 3	Paused 0	Running 1	Failed 0	Filter DAGs by tag	Search DAGs	Auto-refresh	Links
DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	
job1_continuous_ingestion	airflow	3	@hourly	2025-12-17, 05:00:00	2025-12-17, 06:00:00	1		
job2_hourly_cleaning	airflow	3	@hourly	2025-12-17, 05:00:00	2025-12-17, 06:00:00	1		
job3_daily_analytics	airflow	3	@daily	2025-12-16, 22:03:37	2025-12-17, 00:00:00	1		

Showing 1-3 of 3 DAGs

Version: v2.7.0  
Git Version: release:c08c82e9dd0e4aabaa5121519819a636df635210

DAGs - Airflow

localhost:8080/home

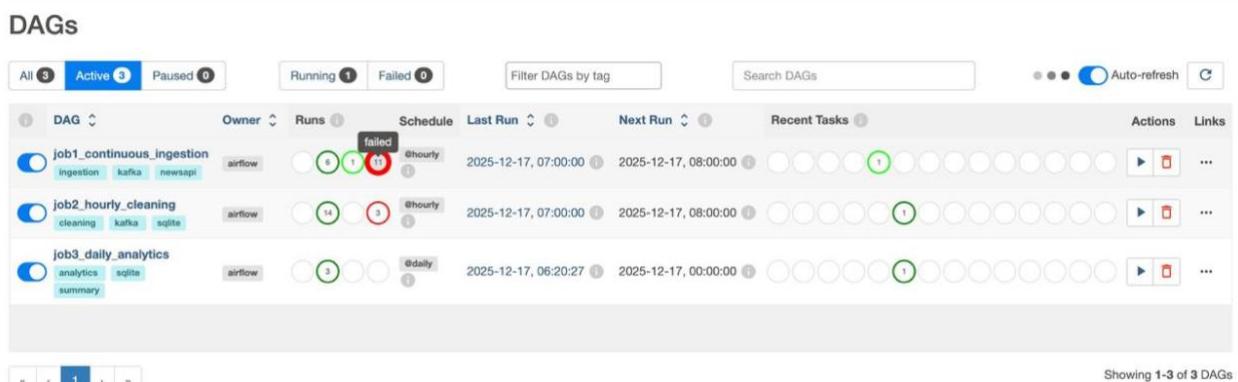
08:18 UTC AU

## DAGs

All 3 Active 3 Paused 0 Running 1 Failed 0 Filter DAGs by tag Search DAGs Auto-refresh C

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
job1_continuous_ingestion	airflow	11	@hourly	2025-12-17, 07:00:00	2025-12-17, 08:00:00	1	[Run]	[View]
job2_hourly_cleaning	airflow	3	@hourly	2025-12-17, 07:00:00	2025-12-17, 08:00:00	1	[Run]	[View]
job3_daily_analytics	airflow	3	@daily	2025-12-17, 06:20:27	2025-12-17, 00:00:00	1	[Run]	[View]

Showing 1-3 of 3 DAGs



Version: v2.7.0  
C:\Users\user\Downloads\2025-02-01-09-42-19-2025-12-17-10-00-00-000

localhost:8080/dagrun/list/?\_flt\_3\_dag\_id=job1\_continuous\_ingestion&\_flt\_3\_state=failed

job1\_continuous\_ingestion - Airflow

localhost:8080/dags/job1\_continuous\_ingestion/grid

10:17 UTC AU

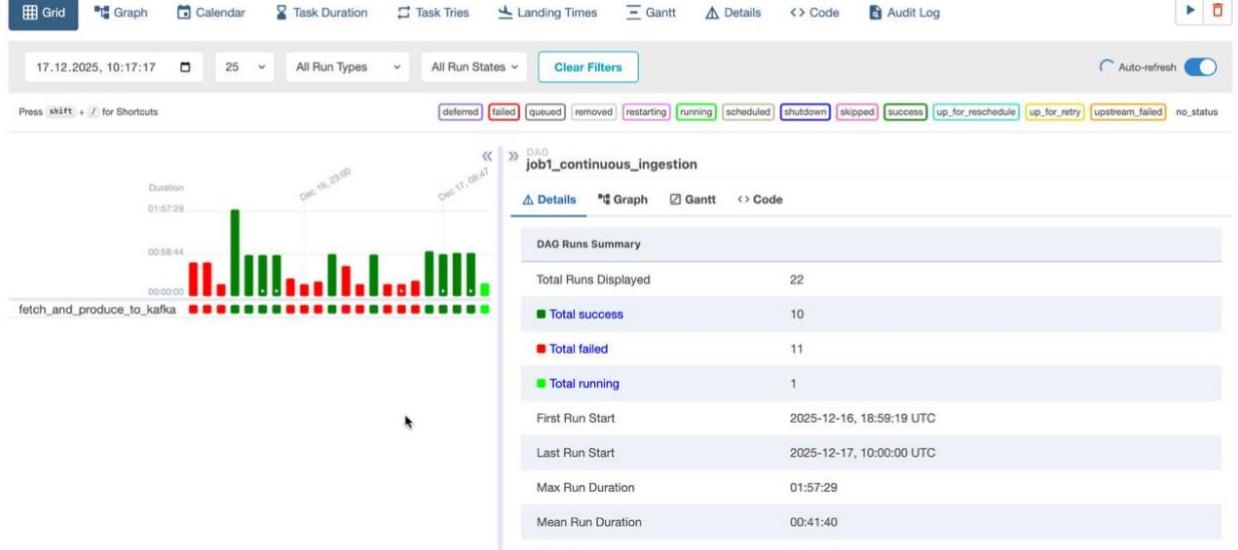
### DAG: job1\_continuous\_ingestion DAG 1: Continuous data ingestion from NewsAPI to Kafka

Schedule: @hourly Next Run: 2025-12-17, 10:00:00

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

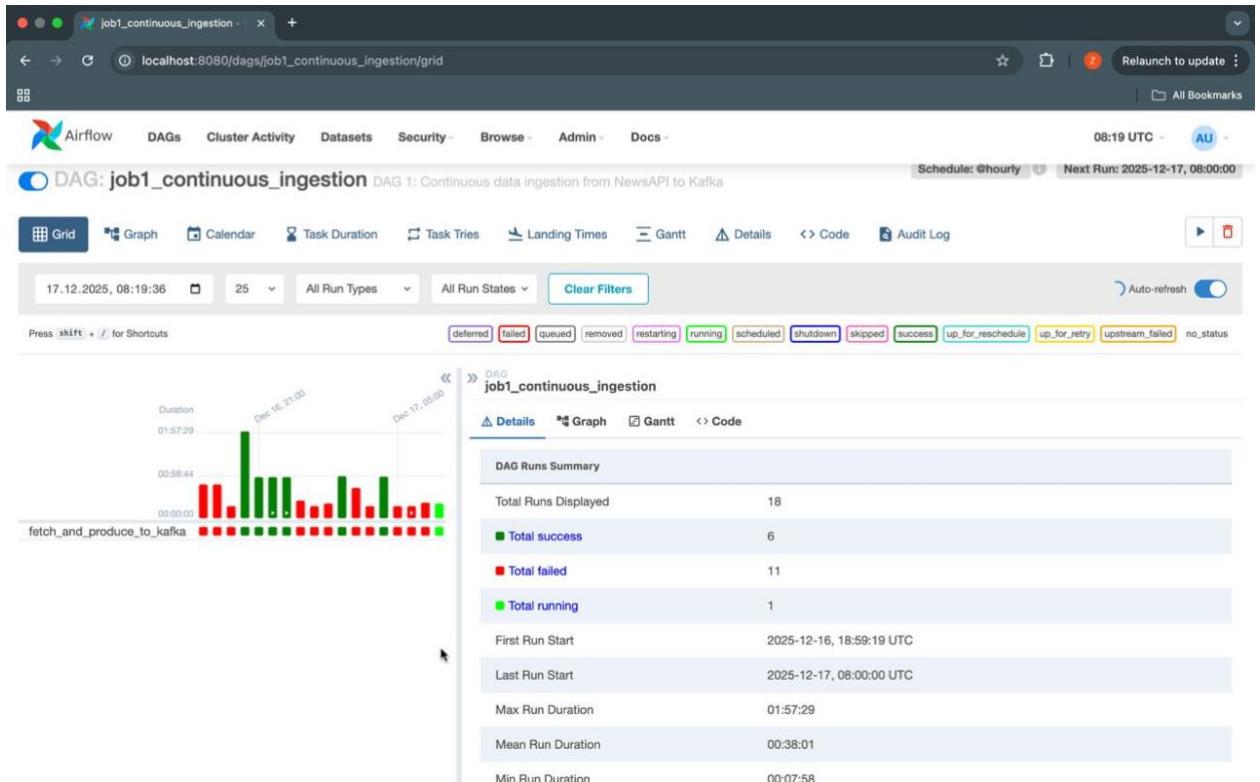
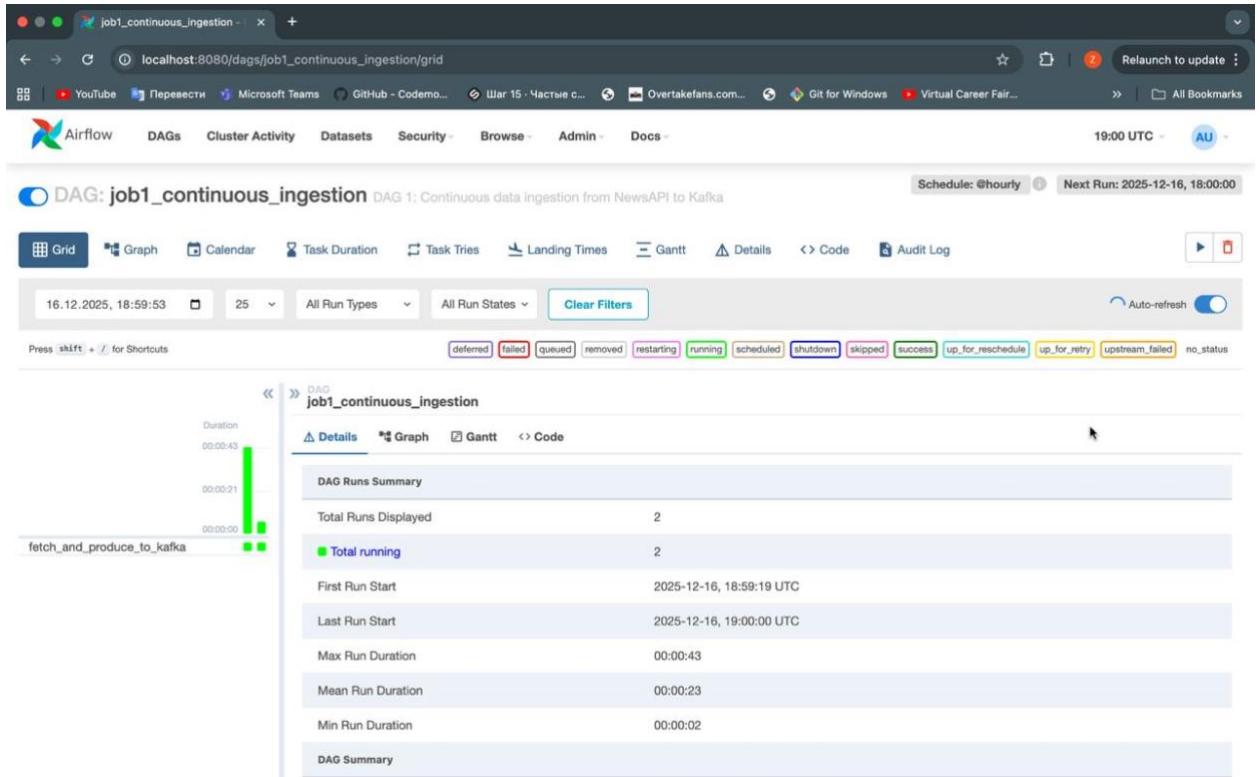
17.12.2025, 10:17:17 25 All Run Types All Run States Clear Filters Auto-refresh

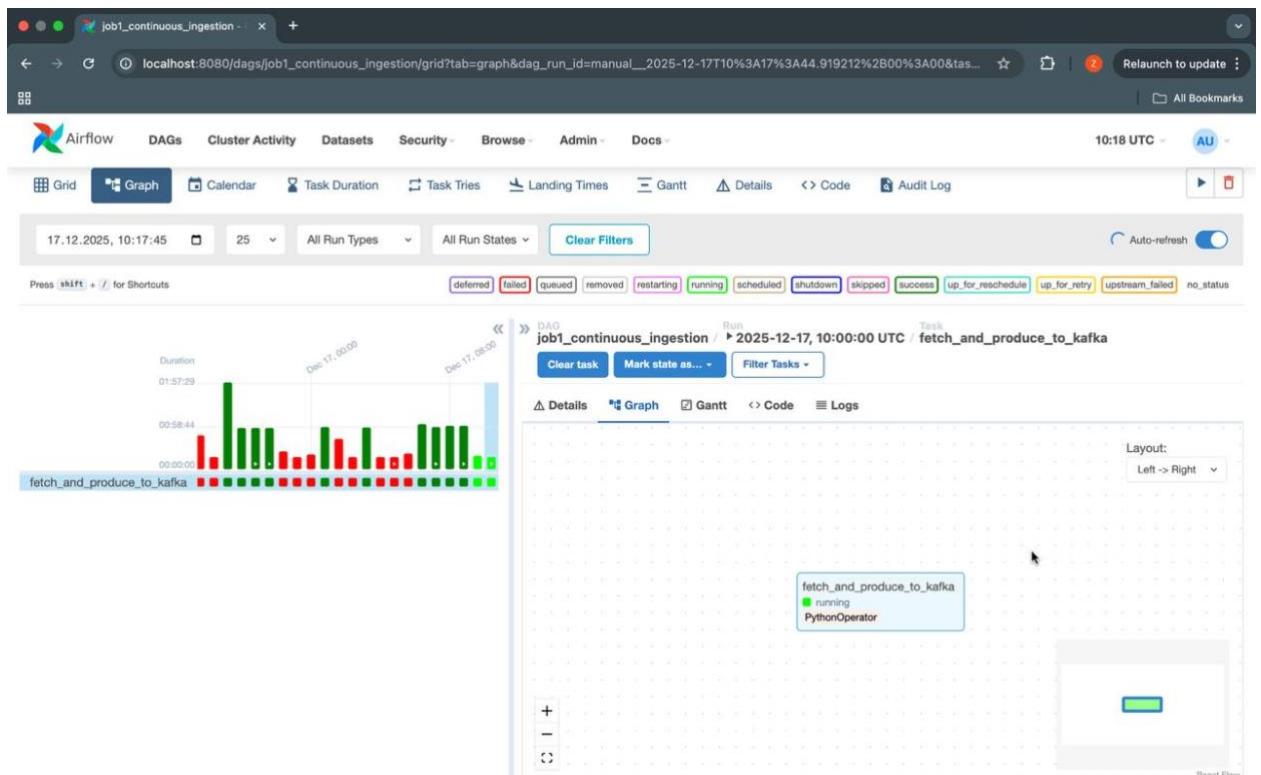
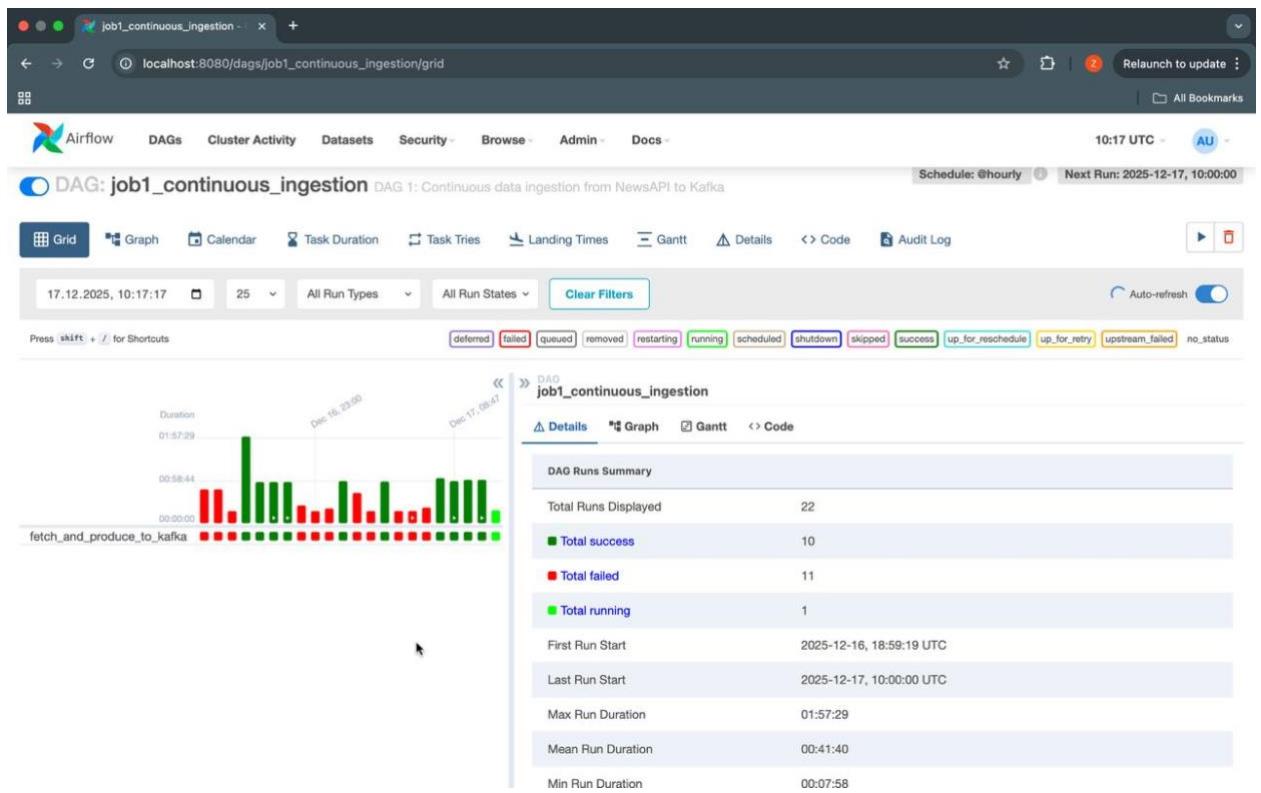
Press shift + / for Shortcuts deferred failed queued removed restarting running scheduled shutdown skipped success up\_for\_reschedule up\_for\_retry upstream\_failed no\_status

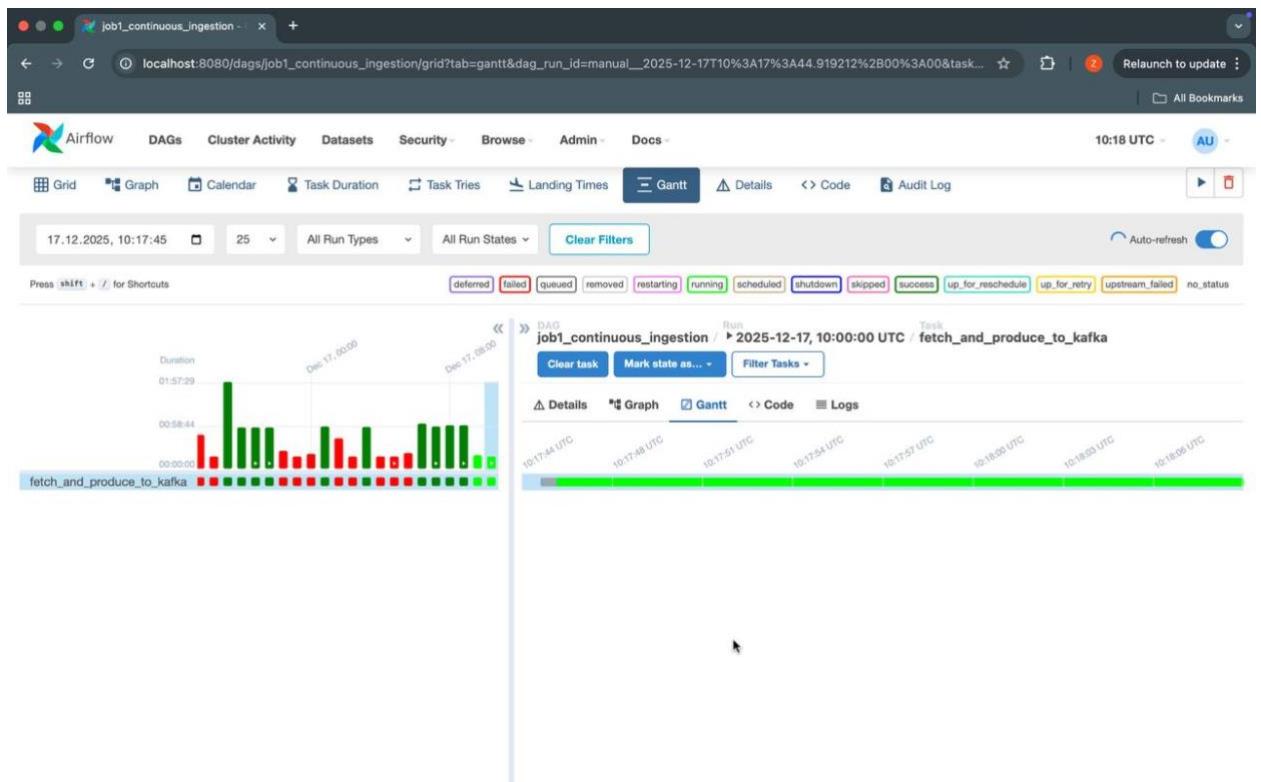


Total Runs Displayed	22
Total success	10
Total failed	11
Total running	1
First Run Start	2025-12-16, 18:59:19 UTC
Last Run Start	2025-12-17, 10:00:00 UTC
Max Run Duration	01:57:29
Mean Run Duration	00:41:40
Min Run Duration	00:07:58

## Job1\_continuous\_ingestion:







The screenshot shows the Airflow DAG Details page for `job1_continuous_ingestion`. The page provides a detailed view of the DAG's configuration, including its schedule interval (@hourly), catchup setting (False), and concurrency (16). It also lists the number of failed (11), success (10), and running (2) tasks. The Default Args section shows a Python dictionary with various configuration parameters. The Tasks Count is listed as 1, Task IDs as `['fetch_and_produce_to_kafka']`, and Relative file location as `job1_ingestion_dag.py`.

Schedule Interval	@hourly
Catchup	False
Started	True
End Date	None
Max Active Runs	1 / 16
Concurrency	16
Default Args	{'depends_on_past': False, 'email_on_failure': False, 'email_on_retry': False, 'owner': 'airflow', 'retries': 1, 'retry_delay': datetime.timedelta(seconds=300), 'start_date': DateTime(2025, 12, 16, 0, 0, tzinfo=Timezone('UTC'))}
Tasks Count	1
Task IDs	['fetch_and_produce_to_kafka']
Relative file location	job1_ingestion_dag.py

localhost:8080/dags/job1\_continuous\_ingestion/details

Owner: airflow

Owner Links: None

DAG Run Timeout: None

Tags: ingestion, kafka, newsapi

DagModel debug information

Attribute	Value
fileloc	/opt/airflow/dags/job1_ingestion_dag.py
has_import_errors	False
has_task_concurrency_limits	False
is_active	True
is_paused_at_creation	True
is_subdag	False
last_expired	None
last_parsed_time	2025-12-17 10:18:12.265562+00:00
last_pickled	None
metadata	MetaData()
next_dagrun	2025-12-17 10:00:00+00:00
next_dagrun_create_after	2025-12-17 11:00:00+00:00
next_dagrun_data_interval	DataInterval(start=DateTime(2025, 12, 17, 10, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 12, 17, 11, 0, 0, tzinfo=Timezone('UTC')))

localhost:8080/dags/job1\_continuous\_ingestion/details

last\_parsed\_time: 2025-12-17 10:18:12.265562+00:00

last\_pickled: None

metadata: MetaData()

next\_dagrun: 2025-12-17 10:00:00+00:00

next\_dagrun\_create\_after: 2025-12-17 11:00:00+00:00

next\_dagrun\_data\_interval: DataInterval(start=DateTime(2025, 12, 17, 10, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 12, 17, 11, 0, 0, tzinfo=Timezone('UTC')))

next\_dagrun\_data\_interval\_end: 2025-12-17 11:00:00+00:00

next\_dagrun\_data\_interval\_start: 2025-12-17 10:00:00+00:00

parent\_dag: None

pickle\_id: None

processor\_subdir: /opt/airflow/dags

registry: <sqlalchemy.orm.decl\_api.registry object at 0xfffff9f683070>

root\_dag\_id: None

safe\_dag\_id: job1\_continuous\_ingestion

scheduler\_lock: None

timetable\_description: Every hour

timezone: Timezone('UTC')

Version: v2.7.0  
Git Version: .release:c08c82e9dd0e4aabaa5121519819a636df635210

job3\_daily\_analytics - Airflow | List Log - Airflow | List Log - Airflow | List Log - Airflow | +

localhost:8080/log/list/?\_flt\_3\_dag\_id=job1\_continuous\_ingestion

All Bookmarks

Relaunch to update

10:24 UTC AU

**List Log**

Search -

Page size: 144 Record Count: 144

ID	Dttm	Dag Id	Task Id	Event	Logical Date	Owner	Extra
341	2025-12-17, 10:23:27	job1_continuous_ingestion		grid		Admin User	[{"dag_id": "job1_continuous_ingestion"}]
340	2025-12-17, 10:23:14	job1_continuous_ingestion		grid		Admin User	[{"dag_id": "job1_continuous_ingestion"}]
326	2025-12-17, 10:17:50	job1_continuous_ingestion	fetch_and_produce_to_kafka	confirm		Admin User	[{"dag_id": "job1_continuous_ingestion"}, {"dag_run_id": "manual_2025-12-17T10:17:44.919212+00:00"}, {"past": "false"}, {"future": "false"}, {"upstream": "false"}, {"downstream": "false"}, {"state": "success"}, {"task_id": "fetch_and_produce_to_kafka"}]
325	2025-12-17, 10:17:50	job1_continuous_ingestion	fetch_and_produce_to_kafka	clear	2025-12-17, 10:17:44	Admin User	[{"dag_id": "job1_continuous_ingestion"}, {"dag_run_id": "manual_2025-12-17T10:17:44.919212+00:00"}, {"confirmed": "false"}, {"execution_date": "2025-12-17T10:17:44.919212+00:00"}, {"past": "false"}, {"future": "false"}, {"upstream": "false"}, {"downstream": "true"}, {"recursive": "true"}, {"only_failed": "false"}, {"task_id": "fetch_and_produce_to_kafka"}]
324	2025-12-17, 10:17:47	job1_continuous_ingestion		graph_data		Admin User	[{"dag_id": "job1_continuous_ingestion"}]
323	2025-12-17, 10:17:46	job1_continuous_ingestion	fetch_and_produce_to_kafka	cli_task_run		airflow	{"host_name": "152303e5a456", "full_command": "/home/airflow/local/bin/airflow scheduler"}
322	2025-12-17, 10:17:46	job1_continuous_ingestion	fetch_and_produce_to_kafka	running	2025-12-17, 10:17:44	airflow	

## Job2\_hourly\_cleaning:

job2\_hourly\_cleaning - Grid | +

localhost:8080/dags/job2\_hourly\_cleaning/grid

All Bookmarks

Relaunch to update

19:00 UTC AU

**DAG: job2\_hourly\_cleaning** DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite

Schedule: @hourly | Next Run: 2025-12-16, 19:00:00

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

16.12.2025, 19:00:13 25 All Run Types All Run States Clear Filters Auto-refresh

Press shift + / for Shortcuts deferred failed queued removed restarting running scheduled shutdown skipped success up\_for\_reschedule up\_for\_retry upstream\_failed no\_status

Duration: 00:00:56

clean\_and\_store\_data

» **DAG: job2\_hourly\_cleaning**

Details Graph Gantt Code

DAG Runs Summary

- Total Runs Displayed: 2
- Total running: 2
- First Run Start: 2025-12-16, 18:59:19 UTC
- Last Run Start: 2025-12-16, 19:00:00 UTC
- Max Run Duration: 00:00:56
- Mean Run Duration: 00:00:36
- Min Run Duration: 00:00:15

DAG Summary

localhost:8080/dags/job2\_hourly\_cleaning/grid

Schedule: @hourly | Next Run: 2025-12-16, 19:00:00

DAG: job2\_hourly\_cleaning DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

16.12.2025, 19:14:54 25 All Run Types All Run States Clear Filters

Auto-refresh

Press shift + / for Shortcuts

Duration: 00:10:04

clean\_and\_store\_data

End date: null

Is paused upon creation: null

Last parsed: 2025-12-16, 19:00:12 UTC

Orientation: LR

Render template as native obj: false

Start date: null

Timezone: Timezone('UTC')

Owners: AIRFLOW

Tags: cleaning, kafka, sqlite

localhost:8080/dags/job2\_hourly\_cleaning/grid

Schedule: @hourly | Next Run: 2025-12-17, 06:00:00

DAG: job2\_hourly\_cleaning DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

17.12.2025, 06:21:22 25 All Run Types All Run States Clear Filters

Duration: 01:02:08

clean\_and\_store\_data

DAG Runs Summary

Total Runs Displayed: 15

Total success: 12

Total failed: 3

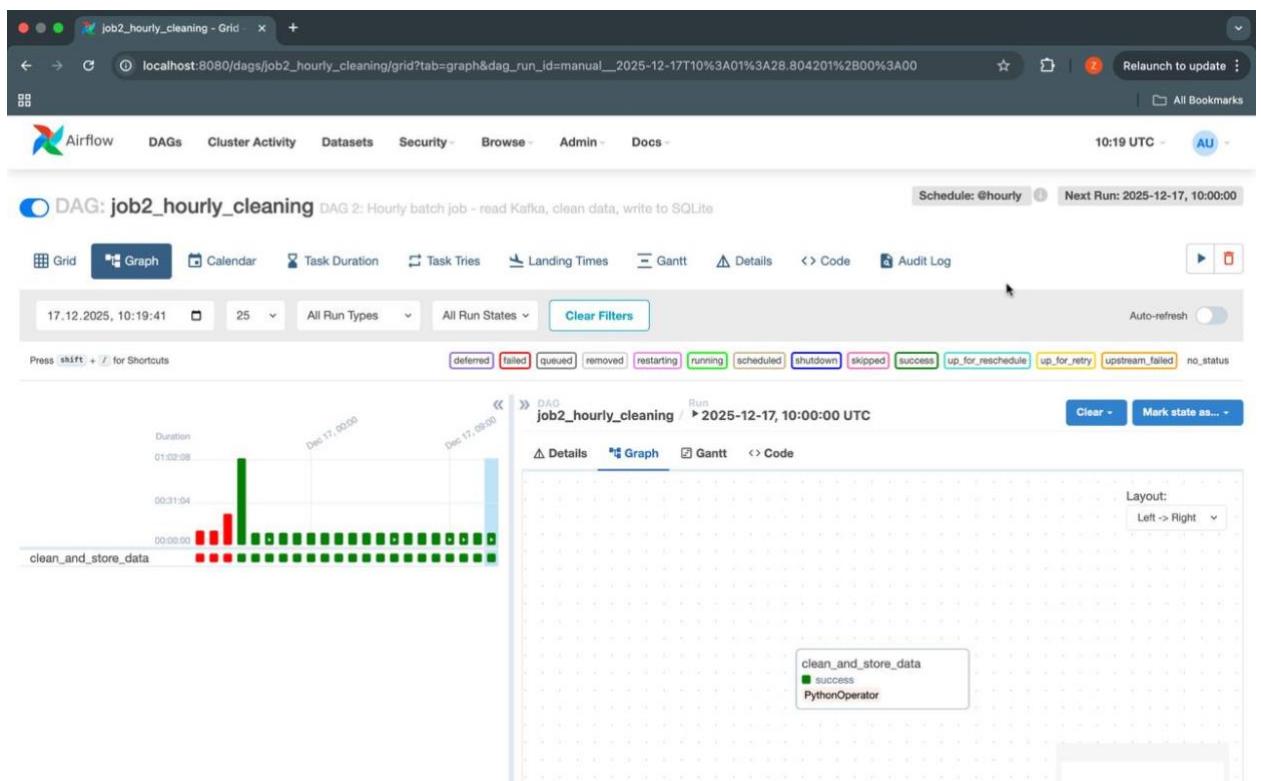
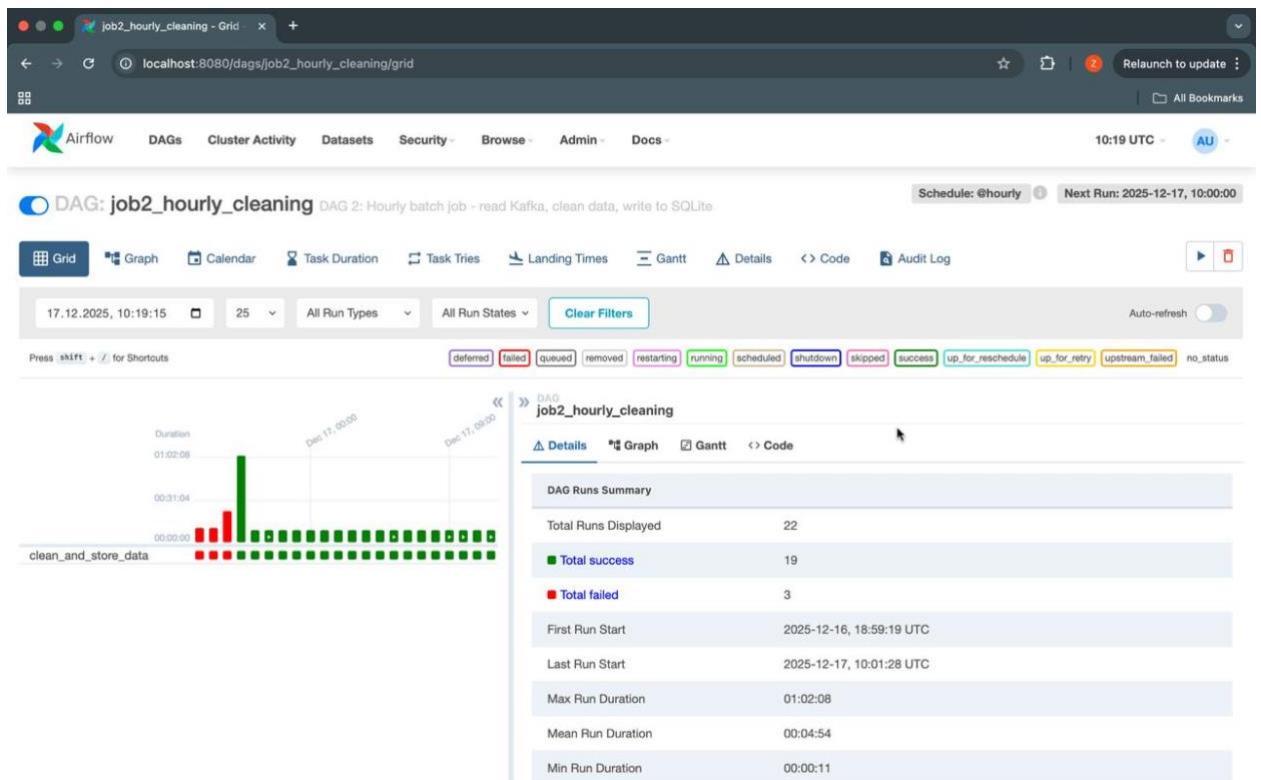
First Run Start: 2025-12-16, 18:59:19 UTC

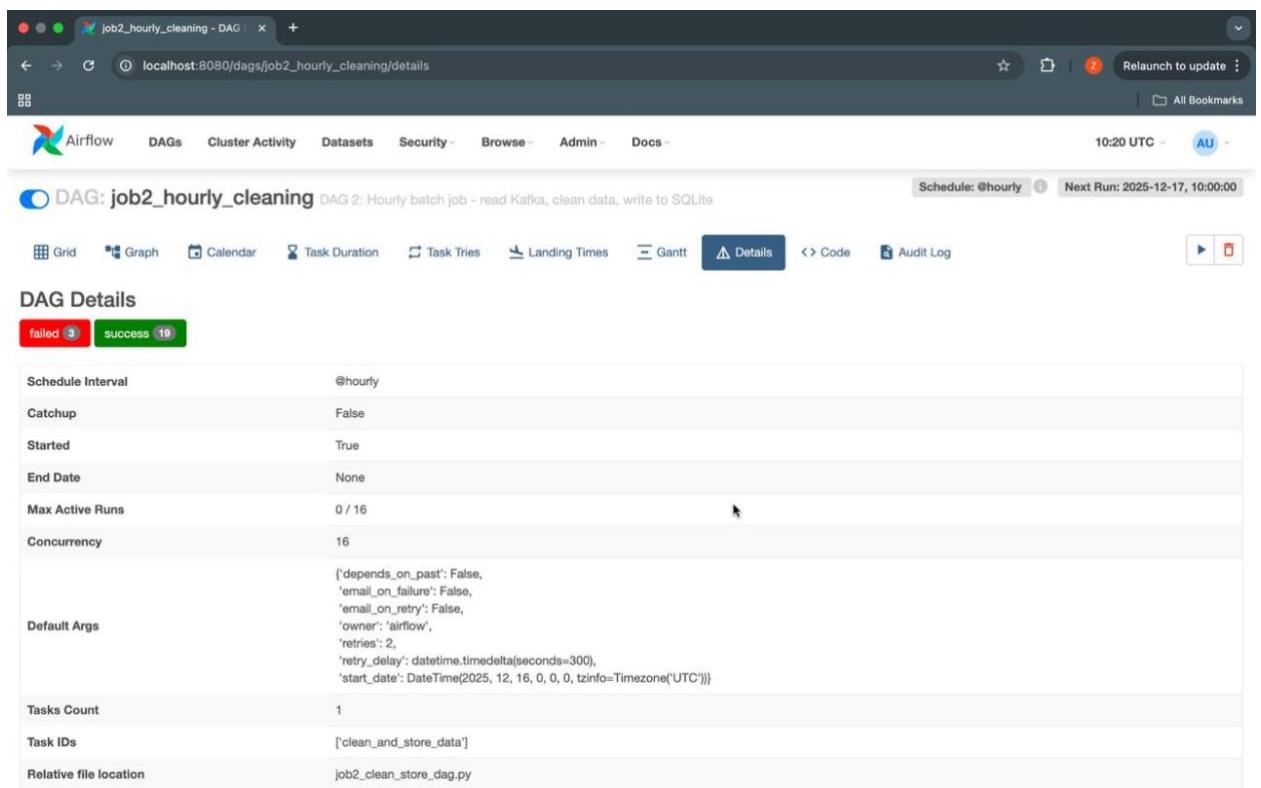
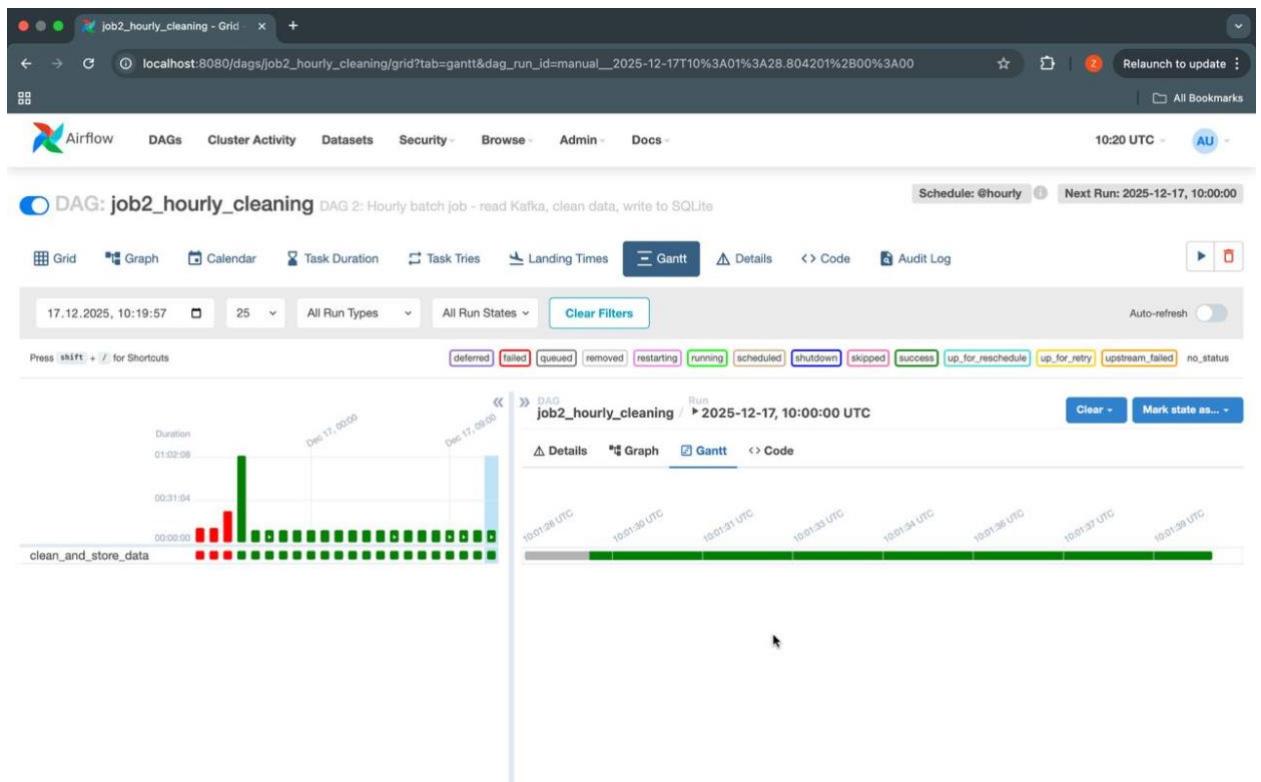
Last Run Start: 2025-12-17, 06:20:26 UTC

Max Run Duration: 01:02:08

Mean Run Duration: 00:07:06

Min Run Duration: 00:00:11





DagModel debug information	
Attribute	Value
fileloc	/opt/airflow/dags/job2_clean_store_dag.py
has_import_errors	False
has_task_concurrency_limits	False
is_active	True
is_paused_at_creation	True
is_subdag	False
last_expired	None
last_parsed_time	2025-12-17 10:19:46.486263+00:00
last_pickled	None
metadata	MetaData()
next_dagrun	2025-12-17 10:00:00+00:00
next_dagrun_create_after	2025-12-17 11:00:00+00:00
next_dagrun_data_interval	DataInterval(start=DateTime(2025, 12, 17, 10, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 12, 17, 11, 0, 0, tzinfo=Timezone('UTC')))

DagModel debug information	
Attribute	Value
last_parsed_time	2025-12-17 10:19:46.486263+00:00
last_pickled	None
metadata	MetaData()
next_dagrun	2025-12-17 10:00:00+00:00
next_dagrun_create_after	2025-12-17 11:00:00+00:00
next_dagrun_data_interval	DataInterval(start=DateTime(2025, 12, 17, 10, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 12, 17, 11, 0, 0, tzinfo=Timezone('UTC')))
next_dagrun_data_interval_end	2025-12-17 11:00:00+00:00
next_dagrun_data_interval_start	2025-12-17 10:00:00+00:00
parent_dag	None
pickle_id	None
processor_subdir	/opt/airflow/dags
registry	<sqlalchemy.orm.decl_api.registry object at 0xfffff9f683070>
root_dag_id	None
safe_dag_id	job2_hourly_cleaning
scheduler_lock	None
timetable_description	Every hour
timezone	Timezone('UTC')

Version: v2.7.0  
 Git Version: .release:c08c82e9dd0e4aabaa5121519819a636df635210

job2\_hourly\_cleaning - Airflow | List Log - Airflow | Relaunch to update | All Bookmarks

localhost:8080/log/list/?\_flt\_3\_dag\_id=job2\_hourly\_cleaning

10:24 UTC AU

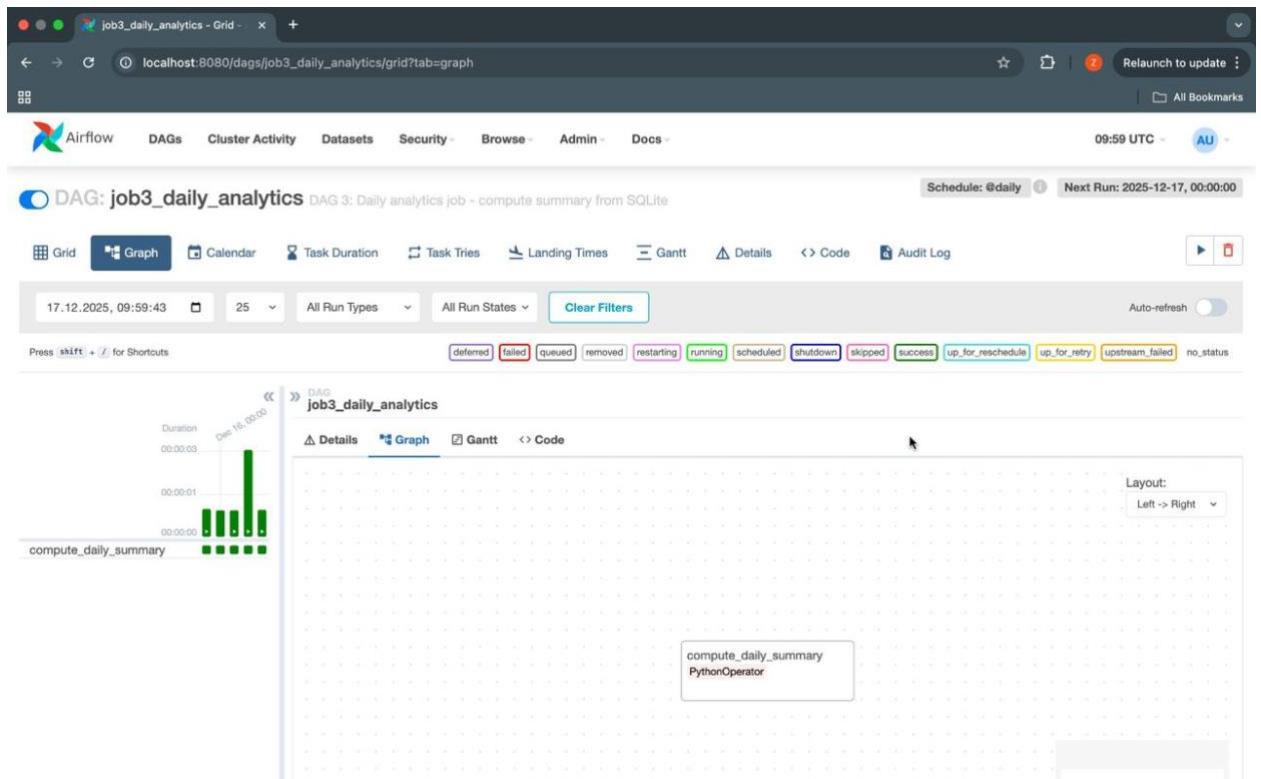
**List Log**

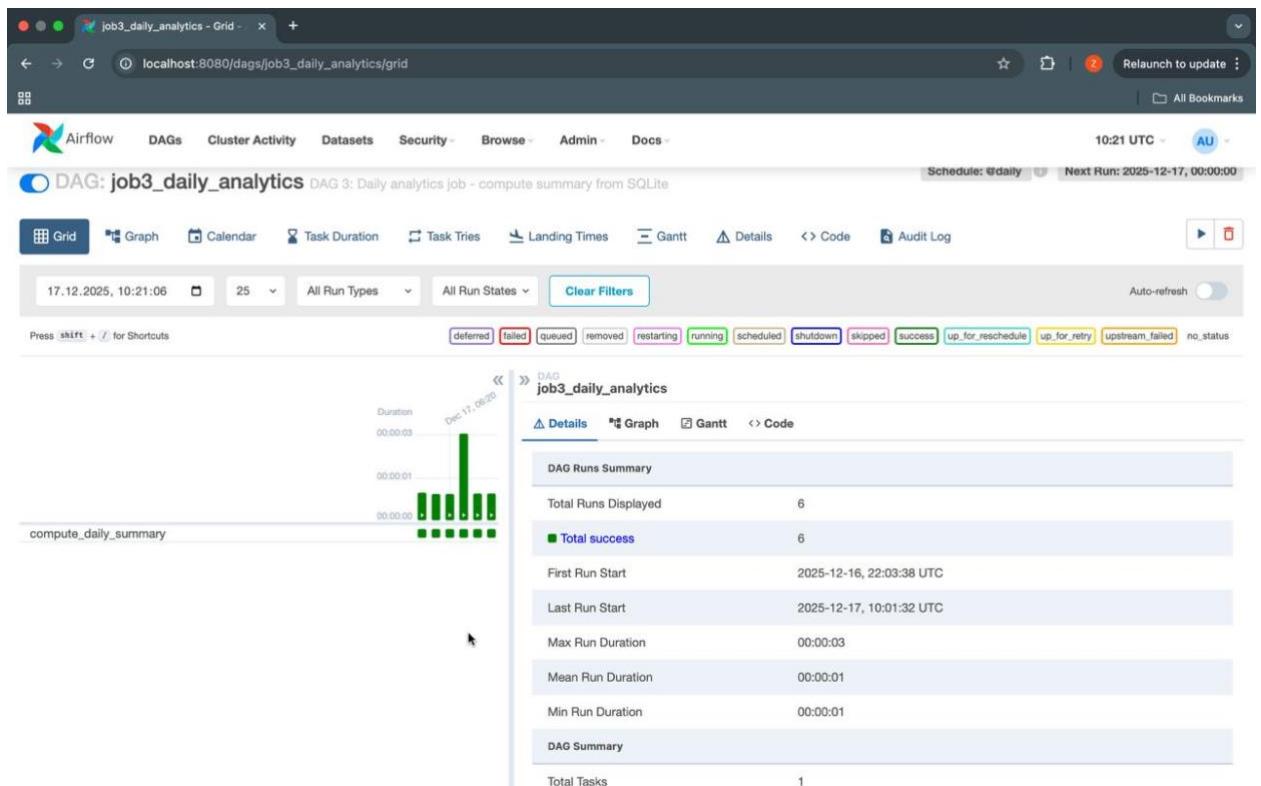
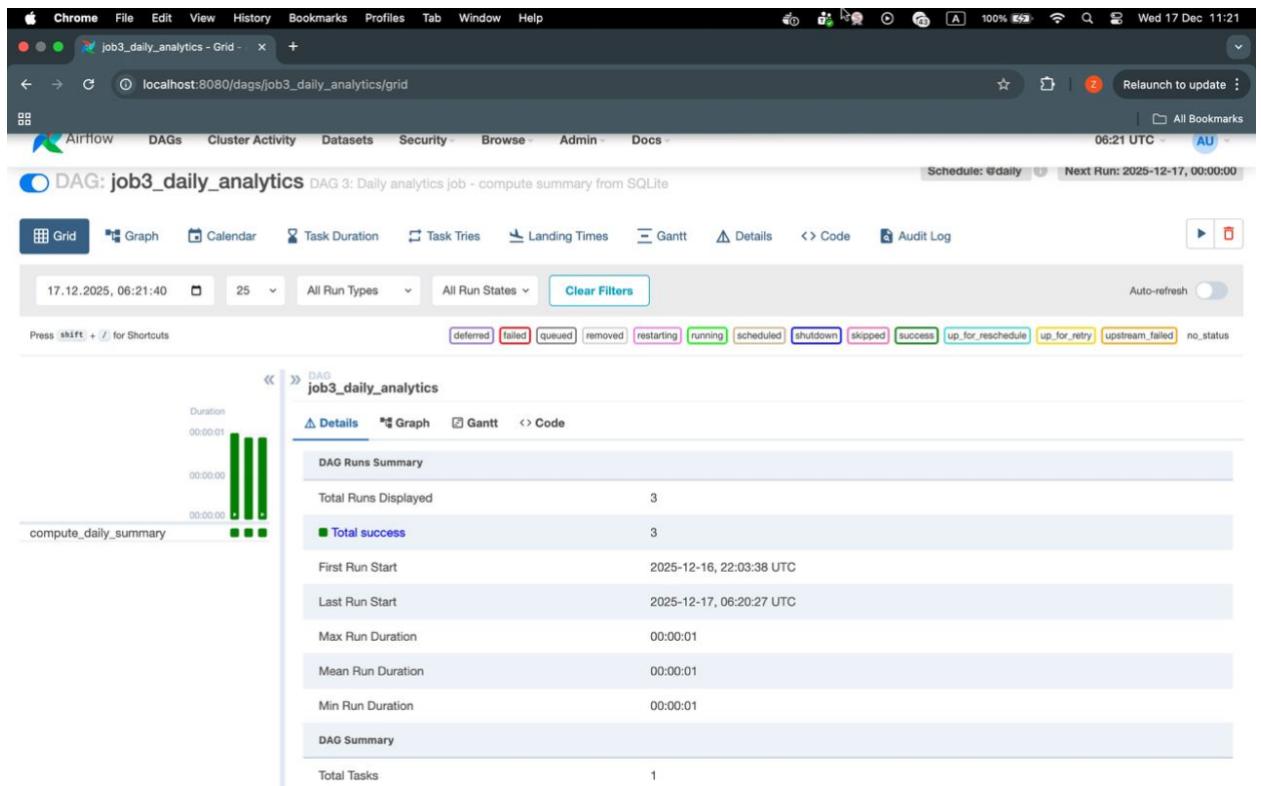
Search -

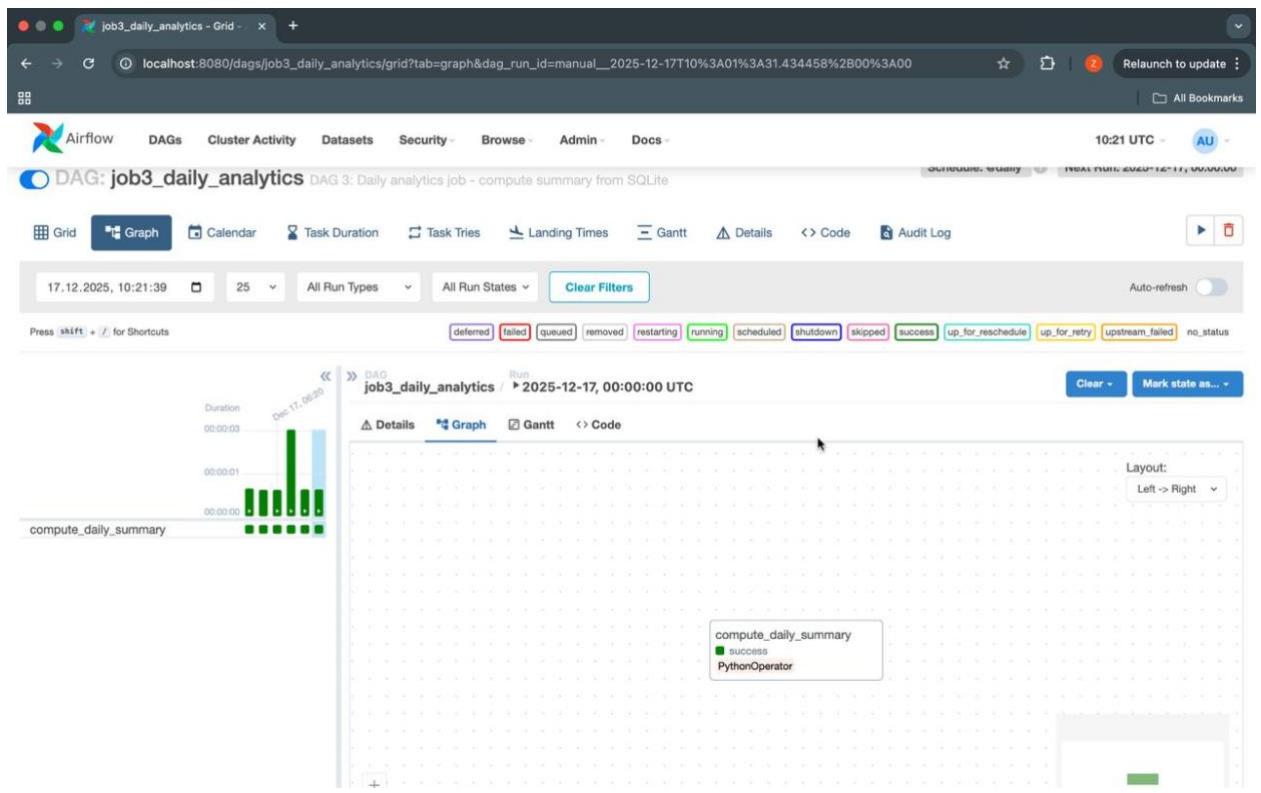
Page size: 150 Record Count: 150

ID	Dttm	Dag ID	Task ID	Event	Logical Date	Owner	Extra
343	2025-12-17, 10:24:28	job2_hourly_cleaning		grid		Admin User	[{"dag_id": "job2_hourly_cleaning"}]
333	2025-12-17, 10:19:57	job2_hourly_cleaning		grid		Admin User	[{"tab": "gantt"}, {"dag_run_id": "manual__2025-12-17T10:01:28.804201+00:00"}, {"dag_id": "job2_hourly_cleaning"}]
332	2025-12-17, 10:19:57	job2_hourly_cleaning		gantt		Admin User	[{"dag_id": "job2_hourly_cleaning"}]
331	2025-12-17, 10:19:41	job2_hourly_cleaning		graph_data		Admin User	[{"dag_id": "job2_hourly_cleaning"}]
330	2025-12-17, 10:19:41	job2_hourly_cleaning		grid		Admin User	[{"tab": "graph"}, {"dag_run_id": "manual__2025-12-17T10:01:28.804201+00:00"}, {"dag_id": "job2_hourly_cleaning"}]
329	2025-12-17, 10:19:40	job2_hourly_cleaning		graph		Admin User	[{"dag_id": "job2_hourly_cleaning"}]
328	2025-12-17, 10:19:19	job2_hourly_cleaning		graph_data		Admin User	[{"dag_id": "job2_hourly_cleaning"}]
327	2025-12-17, 10:19:14	job2_hourly_cleaning		grid		Admin User	[{"dag_id": "job2_hourly_cleaning"}]
317	2025-12-17, 10:03:26	init2_hourly_cleaning		grid		Admin	[{"root": "", "tab": "gantt"}, {"dag_run_id": "manual__2025-12-17T10:01:28.804201+00:00"}, {"dag_id": "init2_hourly_cleaning"}]

## Job3\_daily\_analytics:







DAG Details	
<b>success</b>	0
Schedule Interval	@daily
Catchup	False
Started	True
End Date	None
Max Active Runs	0 / 16
Concurrency	16
Default Args	{'depends_on_past': False, 'email_on_failure': False, 'email_on_retry': False, 'owner': 'airflow', 'retries': 2, 'retry_delay': datetime.timedelta(seconds=300), 'start_date': DateTime(2025, 12, 16, 0, 0, 0, tzinfo=Timezone('UTC'))}
Tasks Count	1
Task IDs	['compute_daily_summary']
Relative file location	job3_daily_summary_dag.py

Owner	airflow
Owner Links	None
DAG Run Timeout	None
Tags	<a href="#">analytics</a> <a href="#">sqlite</a> <a href="#">summary</a>
DagModel debug information	
Attribute	Value
fileloc	/opt/airflow/dags/job3_daily_summary_dag.py
has_import_errors	False
has_task_concurrency_limits	False
is_active	True
is_paused_at_creation	True
is_subdag	False
last_expired	None
last_parsed_time	2025-12-17 10:21:46.990058+00:00
last_pickled	None
metadata	MetaData()
next_dagrun	2025-12-17 00:00:00+00:00
next_dagrun_create_after	2025-12-18 00:00:00+00:00
next_dagrun_data_interval	DataInterval(start=DateTime(2025, 12, 17, 0, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 12, 18, 0, 0, 0, tzinfo=Timezone('UTC')))

last_parsed_time	2025-12-17 10:21:46.990058+00:00
last_pickled	None
metadata	MetaData()
next_dagrun	2025-12-17 00:00:00+00:00
next_dagrun_create_after	2025-12-18 00:00:00+00:00
next_dagrun_data_interval	DataInterval(start=DateTime(2025, 12, 17, 0, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 12, 18, 0, 0, 0, tzinfo=Timezone('UTC')))
next_dagrun_data_interval_end	2025-12-18 00:00:00+00:00
next_dagrun_data_interval_start	2025-12-17 00:00:00+00:00
parent_dag	None
pickle_id	None
processor_subdir	/opt/airflow/dags
registry	<sqlalchemy.orm.decl_api.registry object at 0xfffff9f683070>
root_dag_id	None
safe_dag_id	job3_daily_analytics
scheduler_lock	None
timetable_description	At 00:00
timezone	Timezone('UTC')

Version: v2.7.0  
 Git Version: .release:c08c82e9dd0e4aabaa5121519819a636df635210

job2\_hourly\_cleaning - Airflow | List Log - Airflow | Relaunch to update | +

localhost:8080/log/list/?\_flt\_3\_dag\_id=job3\_daily\_analytics | All Bookmarks

 Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 10:24 UTC AU

### List Log

Search ↗ Record Count: 40

Id	Dttm	Dag Id	Task Id	Event	Logical Date	Owner	Extra
339	2025-12-17, 10:21:57	job3_daily_analytics		grid		Admin User	[{"tab": "gantt", ("dag_run_id", "manual_2025-12-17T10:01:31.434458+00:00"), ("dag_id", "job3_daily_analytics")}]
338	2025-12-17, 10:21:57	job3_daily_analytics		gantt		Admin User	[("dag_id", "job3_daily_analytics")]
337	2025-12-17, 10:21:39	job3_daily_analytics		graph_data		Admin User	[("dag_id", "job3_daily_analytics")]
336	2025-12-17, 10:21:38	job3_daily_analytics		grid		Admin User	[{"tab": "graph", ("dag_run_id", "manual_2025-12-17T10:01:31.434458+00:00"), ("dag_id", "job3_daily_analytics")}]
335	2025-12-17, 10:21:38	job3_daily_analytics		graph		Admin User	[("dag_id", "job3_daily_analytics")]
334	2025-12-17, 10:21:06	job3_daily_analytics		grid		Admin User	[("dag_id", "job3_daily_analytics")]
313	2025-12-17, 10:01:33	job3_daily_analytics	compute_daily_summary	success	2025-12-17, 10:01:31	airflow	
312	2025-12-17, 10:01:33	job3_daily_analytics	compute_daily_summary	cli_task_run		airflow	{"host_name": "152303e6a456", "full_command": "[/home/airflow/.local/bin/airflow", "scheduler"]}
311	2025-12-17, 10:01:33	job3_daily_analytics	compute_daily_summary	running	2025-12-17, 10:01:31	airflow	

## SQL before fixing errors:

```

EXPLORER          310.pyc      README.md      job2_cleaner.py      job3_analytics.py      app.db      job1_ingestion_dag.py
NEWS-Pipeline
  airflow/dags
    > __pycache__
      job1_ingestion_dag.py 1
      job2_clean_store_dag.py 1
      job3_daily_summary_da... 1
    data
    report
    > src
      > __pycache__
        db_utils.python-310.pyc
        job1_producer.python-31...
        job2_cleaner.python-31...
        job3_analytics.python-31...
        db_utils.py
        job1_producer.py 2
        job2_cleaner.py 2
        job3_analytics.py 1
    venv
  docker-compose.yml
  README.md
  requirements.txt

PROBLEMS  PORTS  OUTPUT  DEBUG CONSOLE  TERMINAL
(venv) zhainagenbek@Zhaini's_MacPro news-pipeline % docker-compose exec airflow-webserver sqlite3 /opt/airflow/data/app.db "SELECT COUNT(*) FROM events;"
(venv) zhainagenbek@Zhaini's_MacPro news-pipeline % cd data
(venv) zhainagenbek@Zhaini's_MacPro data % sqlite3 app.db
SQLite version 3.43.2 2023-10-10 13:08:14
Enter ".help" for usage hints.
sqlite> .tables
daily_summary_events
sqlite> SELECT COUNT(*) FROM events
...> SELECT COUNT(*) FROM events;
Parse error: near "SELECT": syntax error
  SELECT COUNT(*) FROM events SELECT COUNT(*) FROM events;
          ^ error here
sqlite> SELECT COUNT(*) FROM events;
0
sqlite> 

```

0 records Ln 28, Col 20 Spaces: 4 UTF-8 LF () Python ⚡ Finish Setup 3.13.11 Go Live

## After fixing errors:

```

110  print(f"INSERTED {inserted} new records into database")
111  print(f"SKIPPED {skipped} duplicates")
112
113

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS
kuanysspandiar@192 data % docker exec task-kafka-1 kafka-console-consumer --bootstrap-server localhost:9092 --topic raw_events --from-beginning --max-messages 3
{"source": {"id": null, "name": "BBC News", "author": null, "title": "US military says eight killed in strikes on alleged drug vessels - BBC", "description": "The US has carried out more than 20 strikes in international waters on alleged drug vessels since September.", "url": "https://www.bbc.com/news/articles/c3e0wd7110v0", "publishedAt": "2025-12-16T07:30:16Z", "content": "The US military says it has carried out strikes on three boats it has accused of trafficking drugs in the Pacific Ocean, killing eight people.\r\nThe US Southern Command posted footage of the strikes \u2026 [+1242 chars]", "fetchedAt": "2025-12-17T11:03:00.730769Z"}, {"source": {"id": "axios", "name": "Axios", "author": "Sara Fischer", "title": "Trump hits BBC with lawsuit over Jan. 6 speech editing - Axios", "description": "He's seeking at least $5 billion in damages over how a documentary edited Trump's speech on Jan. 6, 2021.", "url": "https://www axios.com/25/12/16/trump-bbc-lawsuit-defamation-editing-jan-6-speech", "publishedAt": "2025-12-16T06:33:45Z", "content": "Driving the news: The suit, filed in federal court in Miami, Florida, accuses the BBC of one count of defamation and one of violating Florida's Florida's Deceptive and Unfair Trade Practice Act.\r\n[+2466 chars]", "fetchedAt": "2025-12-17T11:03:00.730769Z"}, {"source": {"id": "axios", "name": "Axios", "author": "Joann Muller", "title": "Ford pivoting to hybrids and dropping all-electric F-150, sees $20B in charges - Axios", "description": "By 2030, about 50% of Ford's global volume will be hybrids, extended-range EVs and electric vehicles.", "url": "https://www axios.com/2025/12/15/ford-hybrids-f-150", "publishedAt": "2025-12-16T06:04:43Z", "content": "By 2030, about 50% of Ford's global volume will be hybrids, extended-range EVs and electric vehicles, versus 17% today, the company said.\r\nThe big picture: Ford painted the announcement [+3089 chars]", "fetchedAt": "2025-12-17T11:03:00.732498Z"}, {"Processed a total of 3 messages
kuanysspandiar@192 data % docker exec task-airflow-webserver-1 sqlite3 /opt/airflow/data/app.db "SELECT COUNT(*) FROM events;"
```

title	source_name	published_at
US military says eight killed in strikes on alleged drug vessels - BBC	BBC News	2025-12-16T07:30:16+00:00
Trump hits BBC with lawsuit over Jan. 6 speech editing - Axios	Axios	2025-12-16T06:33:45+00:00
Ford pivoting to hybrids and dropping all-electric F-150, sees \$20B in charges - Axios	Axios	2025-12-16T06:04:43+00:00
Conservative's victory in Chile suggests a hard-right, pro-Trump surge - Politico	Politico	2025-12-16T05:39:46+00:00
"Very bad for our country": Trump doubles down on Rob Reiner attacks - Axios	Axios	2025-12-16T04:43:26+00:00
Offense sizzles as Steelers stay in 1st place with crowd-pleasing win over Miami - TribLIVE.com	TribLIVE	2025-12-16T04:14:00+00:00
Ben & Jerry's: Row deepens as three board members removed - BBC	BBC News	2025-12-16T04:05:54+00:00
Bondi Beach Gunmen Had ISIS Flags, Visited Philippines - Bloomberg.com	Bloomberg	2025-12-16T03:06:00+00:00
Why does Bermuda appear to float? Scientists' discovery may be the answer - ABC News	ABC News	2025-12-16T02:57:28+00:00
Trump designates street fentanyl as WMD, escalating militarization of drug war - NPR	NPR	2025-12-16T02:47:46+00:00

kuanysspandiar@192 data %

The screenshot shows a terminal window with two command-line sessions. The first session runs the command `task % sqlite3 -box data/app.db "SELECT id, source_name, author, description, category FROM events LIMIT 15;"`. The output is a table with columns: id, source\_name, author, description, and category. The second session runs the command `task % sqlite3 -box data/app.db "SELECT id, source_name, author, description, category FROM events LIMIT 10;"`. The output is a table with the same columns. Both tables have 15 rows each. The terminal interface includes a navigation bar with tabs like PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL, and PORTS. On the left, there's an EXPLORER sidebar with icons for OPEN EDITORS, TASK, airflow, data, logs, plugins, and src. Below the EXPLORER is a file list containing docker-compose.yml, README.md, and requirements.txt. At the bottom, there are sections for OUTLINE and TIMELINE. A Git Graph icon is also present at the bottom left.

```
kuanysspandiar@192 task % sqlite3 -box data/app.db "SELECT id, source_name, author, description, category FROM events LIMIT 15;"  
| 13 | Nintendo Life | Liam Doolan | Free update and free upgrade confirmed  
| 14 | CNN | Unknown | General | Oi Weihao recently welcomed a new family member – a beautiful, highly venomous, blue snake.  
| 15 | New Atlas | https://newatlas.com/author/bromwyn-thompson/ | For the first time, scientists have demonstrated how tanning beds cause fundamental DNA damage across almost the skin's entire surface that results in a threefold risk of developing melanoma. It puts beyond doubt the dangers associated with using these devices. | General |  
| id | source_name | author | description | category |  
| 1 | BBC News | Unknown | The US has carried out more than 20 strikes in international waters on alleged drug vessels since September.  
| General |  
| 2 | Axios | Sara Fischer | He's seeking at least $5 billion in damages over how a documentary edited Trump's speech on Jan. 6, 2021.  
| General |  
| 3 | Axios | Joann Muller | By 2030, about 50% of Ford's global volume will be hybrids, extended-range EVs and electric vehicles.  
| General |  
| 4 | Politico | Associated Press | José Antonio Kast won a landslide victory by tapping into a deep well of resentment.  
| General |  
| 5 | Axios | Julianna Bragg | "I wasn't a fan of Rob Reiner at all in any way, shape or form," Trump said.  
| General |  
| 6 | TribLIVE | Unknown |  
| General |  
| 7 | BBC News | Unknown | One co-founder called it a "blatant power grab" designed to strip the board of its independence.  
| General |  
| 8 | Bloomberg | Neil Jerome Morales | The Philippine government said that a visit by the father and son accused of killing 15 people in Sydney to an area of the country with links to terrorist groups raised no alarms at the time.  
| General |  
| 9 | ABC News | ABC News |  
| General |  
| 10 | NPR | | Trump has already declared the drug cartels terrorist organizations and ordered military strikes against suspected drug boats. Now he's declaring fentanyl a MMD. Experts on street drugs and fentanyl are skeptical these moves will reduce the supply of fentanyl. | General |  
| id | source_name | author | description | category |  
| 1 | BBC News | Unknown | The US has carried out more than 20 strikes in international waters on alleged drug vessels since September.  
| General |  
| 2 | Axios | Sara Fischer | He's seeking at least $5 billion in damages over how a documentary edited Trump's speech on Jan. 6, 2021.  
| General |  
| 3 | Axios | Joann Muller | By 2030, about 50% of Ford's global volume will be hybrids, extended-range EVs and electric vehicles.  
| General |  
| 4 | Politico | Associated Press | José Antonio Kast won a landslide victory by tapping into a deep well of resentment.  
| General |  
| 5 | Axios | Julianna Bragg | "I wasn't a fan of Rob Reiner at all in any way, shape or form," Trump said.  
| General |  
| 6 | TribLIVE | Unknown |  
| General |  
| 7 | BBC News | Unknown | One co-founder called it a "blatant power grab" designed to strip the board of its independence.  
| General |  
| 8 | Bloomberg | Neil Jerome Morales | The Philippine government said that a visit by the father and son accused of killing 15 people in Sydney to an area of the country with links to terrorist groups raised no alarms at the time.  
| General |  
| 9 | ABC News | ABC News |  
| General |  
| 10 | NPR | | Trump has already declared the drug cartels terrorist organizations and ordered military strikes against suspected drug boats. Now he's declaring fentanyl a MMD. Experts on street drugs and fentanyl are skeptical these moves will reduce the supply of fentanyl. | General |
```

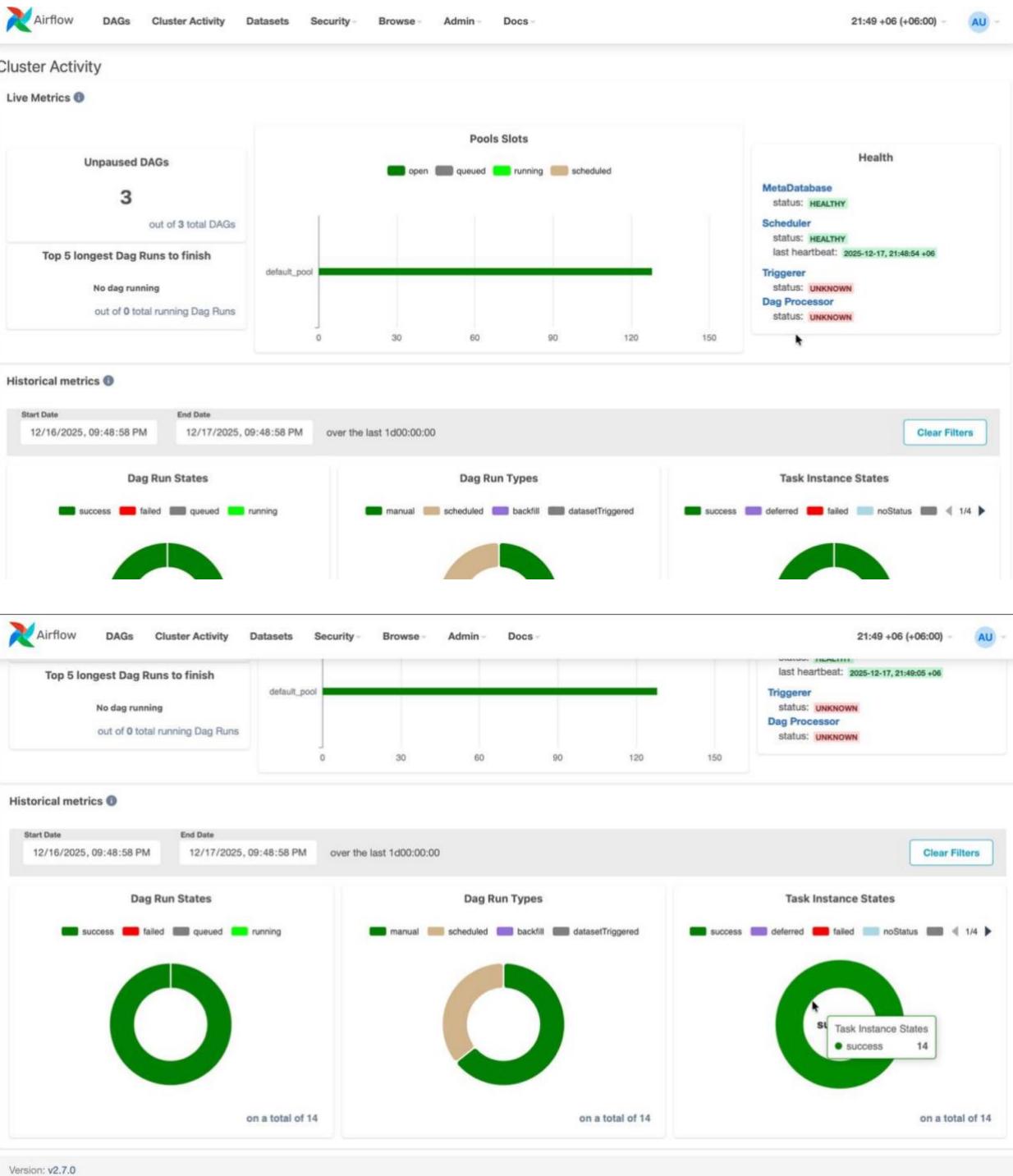
Airflow Cluster Overview											System Status	
DAG ID		Task ID			Run ID			Timeline		Resource Utilization		
DAG ID	Task ID	Run ID	Map Index	Logical Date	Operator	Start Date	End Date	Duration	Note	Job ID		
job1_continuous_ingestion	fetch_and_produce_to_kafka	scheduled_2025-12-17T11:00:00+00:00		2025-12-17, 17:00:00	PythonOperator	2025-12-17, 18:08:02	2025-12-17, 18:10:03	2M:1s		2		AU
job3_daily_analytics	compute_daily_summary	scheduled_2025-12-16T00:00:00+00:00		2025-12-16, 06:00:00	PythonOperator	2025-12-17, 18:08:02	2025-12-17, 18:08:03	<1s		3		
job2_hourly_cleaning	clean_and_store_data	scheduled_2025-12-17T11:00:00+00:00		2025-12-17, 17:00:00	PythonOperator	2025-12-17, 18:08:02	2025-12-17, 18:08:13	10s		4		
job1_continuous_ingestion	fetch_and_produce_to_kafka	manual_2025-12-17T12:08:12.361917+00:00		2025-12-17, 18:08:12	PythonOperator	2025-12-17, 18:08:14	2025-12-17, 18:10:15	2M		5		
job2_hourly_cleaning	clean_and_store_data	manual_2025-12-17T12:08:14.196289+00:00		2025-12-17, 18:08:14	PythonOperator	2025-12-17, 18:08:15	2025-12-17, 18:08:25	10s		6		
job3_daily_analytics	compute_daily_summary	manual_2025-12-17T12:08:15.847713+00:00		2025-12-17, 18:08:15	PythonOperator	2025-12-17, 18:08:16	2025-12-17, 18:08:17	<1s		7		
job1_continuous_ingestion	fetch_and_produce_to_kafka	manual_2025-12-17T12:20:16.173707+00:00		2025-12-17, 18:20:16	PythonOperator	2025-12-17, 18:20:17	2025-12-17, 18:22:18	2M		8		
job2_hourly_cleaning	clean_and_store_data	manual_2025-12-17T12:20:17.913445+00:00		2025-12-17, 18:20:17	PythonOperator	2025-12-17, 18:20:19	2025-12-17, 18:20:34	14s		9		
job3_daily_analytics	compute_daily_summary	manual_2025-12-17T12:20:19.427165+00:00		2025-12-17, 18:20:19	PythonOperator	2025-12-17, 18:20:20	2025-12-17, 18:20:21	<1s		10		
job3_daily_analytics	compute_daily_summary	manual_2025-12-17T12:20:26.666417+00:00		2025-12-17, 18:20:26	PythonOperator	2025-12-17, 18:20:29	2025-12-17, 18:20:29	<1s		11		
job1_continuous_ingestion	fetch_and_produce_to_kafka	scheduled_2025-12-17T14:00:00+00:00		2025-12-17, 20:00:00	PythonOperator	2025-12-17, 21:15:16	2025-12-17, 21:17:17	2M:1s		46		
job2_hourly_cleaning	clean_and_store_data	scheduled_2025-12-17T14:00:00+00:00		2025-12-17, 20:00:00	PythonOperator	2025-12-17, 21:15:16	2025-12-17, 21:15:26	10s		47		
job1_continuous_ingestion	fetch_and_produce_to_kafka	manual_2025-12-17T15:46:50.652424+00:00		2025-12-17, 21:46:50	PythonOperator	2025-12-17, 21:46:59	2025-12-17, 21:49:00	2M:1s		50		
job2_hourly_cleaning	clean_and_store_data	manual_2025-12-17T15:46:52.978191+00:00		2025-12-17, 21:46:52	PythonOperator	2025-12-17, 21:47:00	2025-12-17, 21:47:10	10s		51		
job3_daily_analytics	compute_daily_summary	manual_2025-12-17T15:46:56.207063+00:00		2025-12-17, 21:46:56	PythonOperator	2025-12-17, 21:47:00	2025-12-17, 21:47:00	<1s		52		

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 21:49 +06 (+06:00) AU

ID	Dag ID	State	Job Type	Start Date	End Date	Latest Heartbeat	Executor Class	Hostname	Unixname
52	job3_daily_analytics	success	LocalTaskJob	2025-12-17, 21:46:59	2025-12-17, 21:47:01	2025-12-17, 21:46:59	c357dde9854f	airflow	
51	job2_hourly_cleaning	success	LocalTaskJob	2025-12-17, 21:46:59	2025-12-17, 21:47:10	2025-12-17, 21:47:10	c357dde9854f	airflow	
50	job1_continuous_ingestion	success	LocalTaskJob	2025-12-17, 21:46:59	2025-12-17, 21:49:00	2025-12-17, 21:48:55	c357dde9854f	airflow	
49		running	SchedulerJob	2025-12-17, 21:46:17		2025-12-17, 21:49:40	c357dde9854f	airflow	
48		failed	SchedulerJob	2025-12-17, 21:31:57		2025-12-17, 21:45:30	ebee7d596937	airflow	
47	job2_hourly_cleaning	success	LocalTaskJob	2025-12-17, 21:15:16	2025-12-17, 21:15:27	2025-12-17, 21:15:26	93ad07147bb7	airflow	
46	job1_continuous_ingestion	success	LocalTaskJob	2025-12-17, 21:15:16	2025-12-17, 21:17:17	2025-12-17, 21:17:13	93ad07147bb7	airflow	
45		failed	SchedulerJob	2025-12-17, 21:15:14		2025-12-17, 21:30:03	93ad07147bb7	airflow	
14	job2_hourly_cleaning	running	LocalTaskJob	2025-12-17, 21:12:08		2025-12-17, 21:12:38	93ad07147bb7	airflow	
13	job1_continuous_ingestion	running	LocalTaskJob	2025-12-17, 21:12:08		2025-12-17, 21:12:38	93ad07147bb7	airflow	
12		failed	SchedulerJob	2025-12-17, 21:12:06		2025-12-17, 21:12:42	93ad07147bb7	airflow	
11	job3_daily_analytics	success	LocalTaskJob	2025-12-17, 18:20:29	2025-12-17, 18:20:29	2025-12-17, 18:20:28	93ad07147bb7	airflow	
10	job3_daily_analytics	success	LocalTaskJob	2025-12-17, 18:20:20	2025-12-17, 18:20:21	2025-12-17, 18:20:20	93ad07147bb7	airflow	
9	job2_hourly_cleaning	success	LocalTaskJob	2025-12-17, 18:20:19	2025-12-17, 18:20:34	2025-12-17, 18:20:30	93ad07147bb7	airflow	
8	job1_continuous_ingestion	success	LocalTaskJob	2025-12-17, 18:20:17	2025-12-17, 18:22:18	2025-12-17, 18:22:15	93ad07147bb7	airflow	
7	job3_daily_analytics	success	LocalTaskJob	2025-12-17, 18:08:16	2025-12-17, 18:08:17	2025-12-17, 18:08:16	93ad07147bb7	airflow	
6	job2_hourly_cleaning	success	LocalTaskJob	2025-12-17, 18:08:15	2025-12-17, 18:08:26	2025-12-17, 18:08:25	93ad07147bb7	airflow	
5	job1_continuous_ingestion	success	LocalTaskJob	2025-12-17, 18:08:14	2025-12-17, 18:10:15	2025-12-17, 18:10:10	93ad07147bb7	airflow	
4	job2_hourly_cleaning	success	LocalTaskJob	2025-12-17, 18:08:02	2025-12-17, 18:08:13	2025-12-17, 18:08:13	93ad07147bb7	airflow	
3	job3_daily_analytics	success	LocalTaskJob	2025-12-17, 18:08:02	2025-12-17, 18:08:03	2025-12-17, 18:08:02	93ad07147bb7	airflow	
2	job1_continuous_ingestion	success	LocalTaskJob	2025-12-17, 18:08:02	2025-12-17, 18:10:03	2025-12-17, 18:09:58	93ad07147bb7	airflow	
1		failed	SchedulerJob	2025-12-17, 18:07:31		2025-12-17, 18:35:54	93ad07147bb7	airflow	

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 21:49 +06 (+06:00) AU

	State	Dag ID	Logical Date	Run Id	Run Type	Queued At	Start Date	End Date	Note	External Trigger	Conf	Duration
	success	job3_daily_analytics	2025-12-17, 21:46:56	manual_2025-12-17T15:46:56.207063+00:00	manual	2025-12-17, 21:46:56	2025-12-17, 21:46:56	2025-12-17, 21:47:01	True	0	4s	
	success	job2_hourly_cleaning	2025-12-17, 21:46:52	manual_2025-12-17T15:46:52.978191+00:00	manual	2025-12-17, 21:46:53	2025-12-17, 21:46:53	2025-12-17, 21:47:11	True	0	18s	
	success	job1_continuous_ingestion	2025-12-17, 21:46:50	manual_2025-12-17T15:46:50.652424+00:00	manual	2025-12-17, 21:46:50	2025-12-17, 21:46:51	2025-12-17, 21:49:01	True	0	2M:9s	
	success	job1_continuous_ingestion	2025-12-17, 20:00:00	scheduled_2025-12-17T14:00:00+00:00	scheduled	2025-12-17, 21:12:07	2025-12-17, 21:12:07	2025-12-17, 21:17:17	False	0	5M:9s	
	success	job2_hourly_cleaning	2025-12-17, 20:00:00	scheduled_2025-12-17T14:00:00+00:00	scheduled	2025-12-17, 21:12:07	2025-12-17, 21:12:07	2025-12-17, 21:15:27	False	0	3M:19s	
	success	job3_daily_analytics	2025-12-17, 18:20:26	manual_2025-12-17T12:20:26.666417+00:00	manual	2025-12-17, 18:20:26	2025-12-17, 18:20:27	2025-12-17, 18:20:30	True	0	2s	
	success	job3_daily_analytics	2025-12-17, 18:20:19	manual_2025-12-17T12:20:19.427165+00:00	manual	2025-12-17, 18:20:19	2025-12-17, 18:20:19	2025-12-17, 18:20:21	True	0	1s	
	success	job2_hourly_cleaning	2025-12-17, 18:20:17	manual_2025-12-17T12:20:17.913445+00:00	manual	2025-12-17, 18:20:17	2025-12-17, 18:20:18	2025-12-17, 18:20:34	True	0	15s	
	success	job1_continuous_ingestion	2025-12-17, 18:20:16	manual_2025-12-17T12:20:16.173707+00:00	manual	2025-12-17, 18:20:16	2025-12-17, 18:20:16	2025-12-17, 18:22:18	True	0	2M:2s	
	success	job3_daily_analytics	2025-12-17, 18:08:15	manual_2025-12-17T12:08:15.647713+00:00	manual	2025-12-17, 18:08:15	2025-12-17, 18:08:15	2025-12-17, 18:08:18	True	0	2s	
	success	job2_hourly_cleaning	2025-12-17, 18:08:14	manual_2025-12-17T12:08:14.196289+00:00	manual	2025-12-17, 18:08:14	2025-12-17, 18:08:14	2025-12-17, 18:08:26	True	0	12s	
	success	job1_continuous_ingestion	2025-12-17, 18:08:12	manual_2025-12-17T12:08:12.361917+00:00	manual	2025-12-17, 18:08:12	2025-12-17, 18:08:13	2025-12-17, 18:10:16	True	0	2M:2s	
	success	job2_hourly_cleaning	2025-12-17, 17:00:00	scheduled_2025-12-17T11:00:00+00:00	scheduled	2025-12-17, 18:08:01	2025-12-17, 18:08:01	2025-12-17, 18:08:14	False	0	13s	
	success	job1_continuous_ingestion	2025-12-17, 17:00:00	scheduled_2025-12-17T11:00:00+00:00	scheduled	2025-12-17, 18:08:00	2025-12-17, 18:08:00	2025-12-17, 18:10:03	False	0	2M:3s	
	success	job3_daily_analytics	2025-12-16, 06:00:00	scheduled_2025-12-16T00:00:00+00:00	scheduled	2025-12-17, 18:08:01	2025-12-17, 18:08:01	2025-12-17, 18:08:04	False	0	2s	



## Conclusion

The news data pipeline was successfully implemented and is now operating as expected. Initial failures in DAG 1 and DAG 2 (three errors each) were caused by configuration issues and database initialization problems, which were resolved through proper environment setup and connection handling. Additional DAG 1 failures are attributed to NewsAPI's variable data availability, as news updates are not guaranteed hourly. The pipeline now successfully ingests, processes, and analyzes news data, meeting all project requirements for streaming and batch data processing.