



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 contains navigation options: Ask Gordon, Containers (selected), Images, Volumes, Builds, Docker Hub, Docker Scout, and Extensions. The main area is titled 'Containers' and shows a summary of container usage: CPU usage at 5.11% / 800% and memory usage at 2.03GB / 3.74GB. Below this is a table of running containers.

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

At the bottom of the interface, it shows system status: Engine running, RAM 2.85 GB, CPU 0.62%, Disk: 13.35 GB used (limit 1006.85 GB). There is also a 'Terminal' button and a 'New version available' notification.

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
2025-12-17 03:01:51 tracking rate limits as no storage was explicitly specified. This is not recommended for production use. See: https://flask-limiter.readthedocs.
2025-12-17 03:01:51 lo#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 PostgresqlImpl.
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 PostgresqlImpl.
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 1005.85 GB)

Terminal New version available

DAGs:

DAGs - Airflow

localhost:8080/home?status=active

Relaunch to update

19:23 UTC AU

DAGs

All 3 Active 3 Paused 0 Running 2 Failed 1 Filter DAGs by tag Search DAGs Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
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	1 failed		
job3_daily_analytics analytics sqlite summary	airflow	1	@daily	2025-12-16, 00:00:00				

Showing 1-3 of 3 DAGs

Version: v2.7.0

localhost:8080/taskinstance/list/?_flt_3_dag_id=job2_hourly_cleaning&_flt_3_state=failed

DAGs - Airflow

localhost:8080/home

Relaunch to update

All Bookmarks

Airflow

DAGsCluster ActivityDatasetsSecurityBrowseAdminDocs

22:04 UTCAU

Triggered job3_daily_analytics, it should start any moment now.

DAGs

All3Active3Paused0Running4Failed0

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
<div>job1_continuous_ingestion</div> <div>ingestionkafkanewsapi</div> <div>airflow</div>		<div>3</div> <div>3</div>	@hourly	2025-12-16, 22:03:33	2025-12-16, 22:00:00	<div>3</div>	<div>▶</div> <div>🗑</div>	...
<div>job2_hourly_cleaning</div> <div>cleaningkafkasqlite</div> <div>airflow</div>		<div>3</div> <div>3</div>	@hourly	2025-12-16, 22:03:35	2025-12-16, 22:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...
<div>job3_daily_analytics</div> <div>analyticssqlitesummary</div> <div>airflow</div>		<div>1</div> <div>1</div>	@daily	2025-12-16, 22:03:37	2025-12-16, 00:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...

Showing 1-3 of 3 DAGs

Version: v2.7.0

Git Version: .release:c08c82e9dd0e4aaba5121519819a636df635210

DAGs - Airflow

localhost:8080/home

Relaunch to update

All Bookmarks

Airflow

DAGsCluster ActivityDatasetsSecurityBrowseAdminDocs

06:19 UTCAU

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

All3Active3Paused0Running1Failed0

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
<div>job1_continuous_ingestion</div> <div>ingestionkafkanewsapi</div> <div>airflow</div>		<div>6</div> <div>9</div>	@hourly	2025-12-17, 05:00:00	2025-12-17, 06:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...
<div>job2_hourly_cleaning</div> <div>cleaningkafkasqlite</div> <div>airflow</div>		<div>11</div> <div>3</div>	@hourly	2025-12-17, 05:00:00	2025-12-17, 06:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...
<div>job3_daily_analytics</div> <div>analyticssqlitesummary</div> <div>airflow</div>		<div>2</div> <div>1</div>	@daily	2025-12-16, 22:03:37	2025-12-17, 00:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...

Showing 1-3 of 3 DAGs

Version: v2.7.0

Git Version: .release:c08c82e9dd0e4aaba5121519819a636df635210

DAGs - Airflow

localhost:8080/home

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

08:18 UTC

AU

DAGs

All 3

Active 3

Paused 0

Running 1

Failed 0

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
<div>job1_continuous_ingestion</div> <div>ingestion kafka newssapi</div> <div>airflow</div>		<div>9</div> <div>1</div> <div>11</div>	Hourly	2025-12-17, 07:00:00	2025-12-17, 08:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...
<div>job2_hourly_cleaning</div> <div>cleaning kafka sqlite</div> <div>airflow</div>		<div>14</div> <div>3</div>	Hourly	2025-12-17, 07:00:00	2025-12-17, 08:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...
<div>job3_daily_analytics</div> <div>analytics sqlite summary</div> <div>airflow</div>		<div>5</div>	@daily	2025-12-17, 06:20:27	2025-12-17, 00:00:00	<div>1</div>	<div>▶</div> <div>🗑</div>	...

Showing 1-3 of 3 DAGs

Version: v2.7.0
localhost:8080/dagrun/list/?_flt_3_dag_id=job1_continuous_ingestion&_flt_3_state=failed

job1_continuous_ingestion

localhost:8080/dags/job1_continuous_ingestion/grid

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

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

Duration

fetch_and_produce_to_kafka

01:57:29

00:58:44

00:00:00

Dec 16, 23:00

Dec 17, 08:47

DAG job1_continuous_ingestion

Details

Graph

Gantt

Code

DAG Runs Summary

Total Runs Displayed22

Total success10

Total failed11

Total running1

First Run Start2025-12-16, 18:59:19 UTC

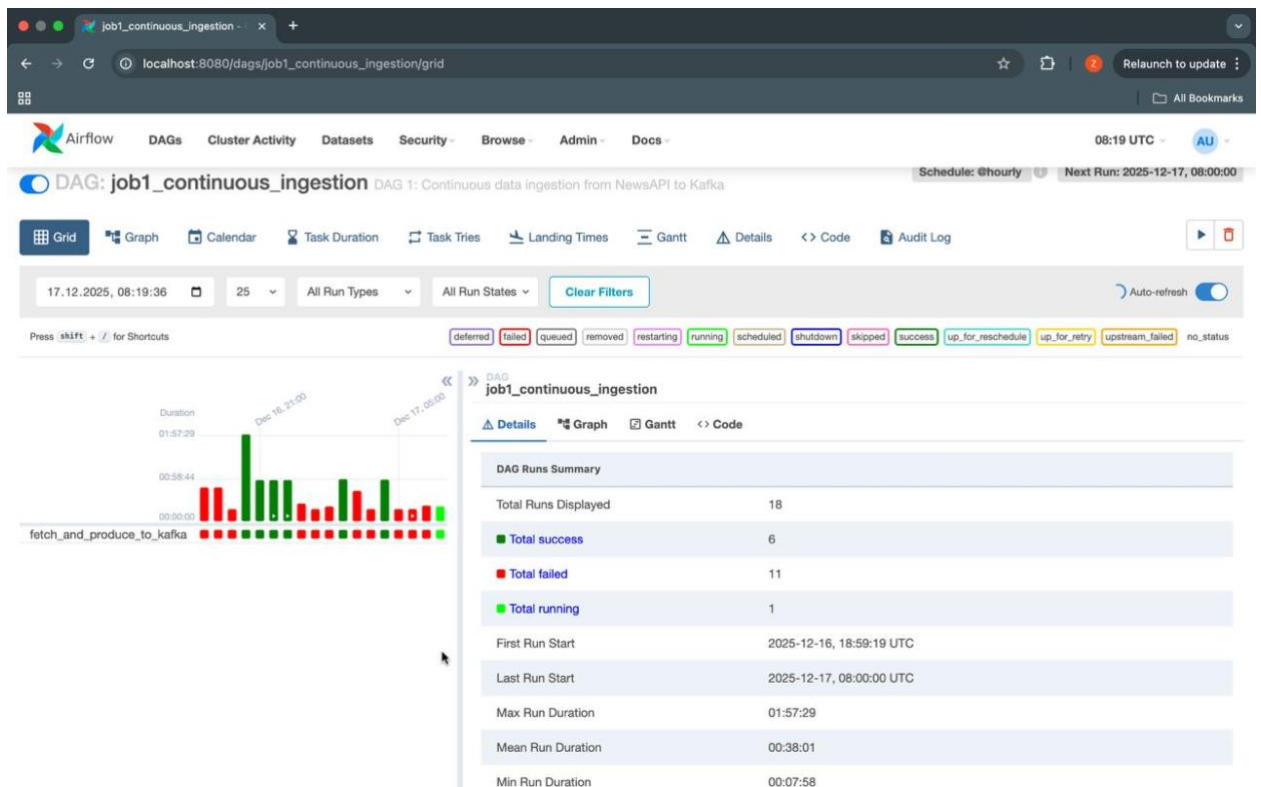
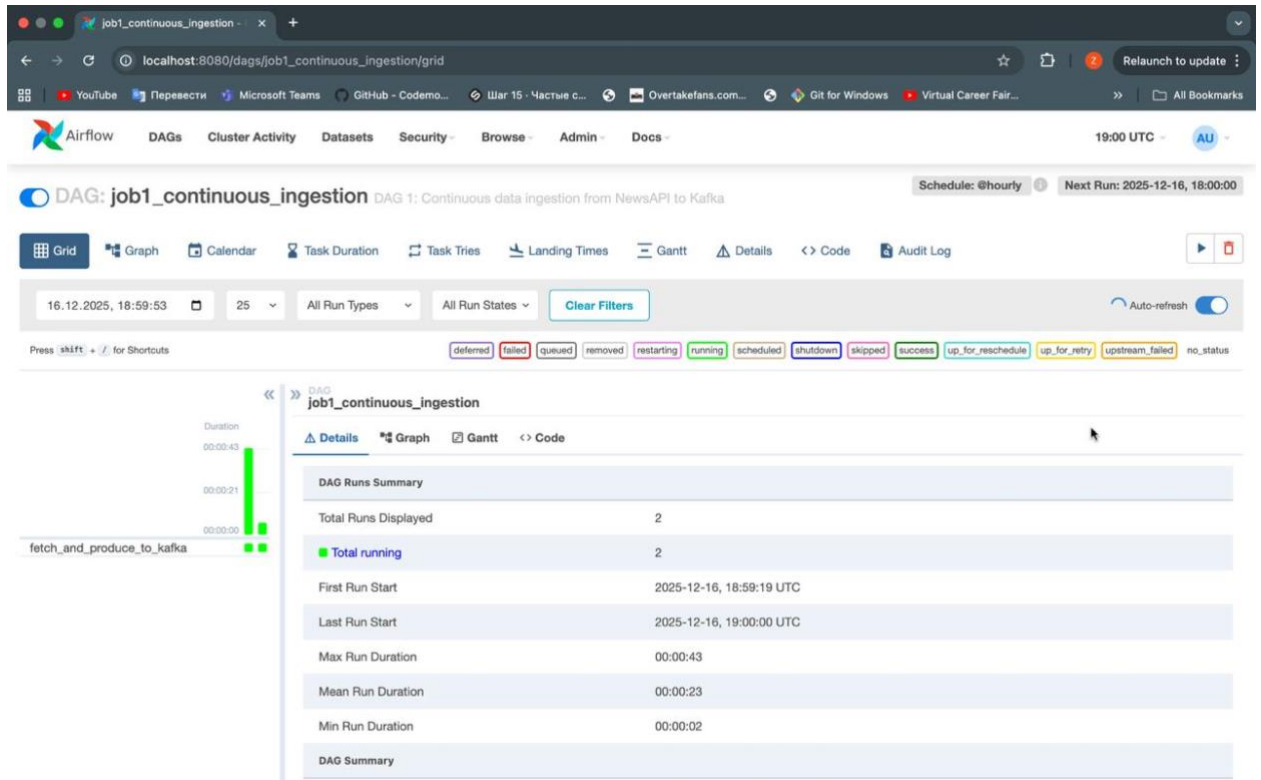
Last Run Start2025-12-17, 10:00:00 UTC

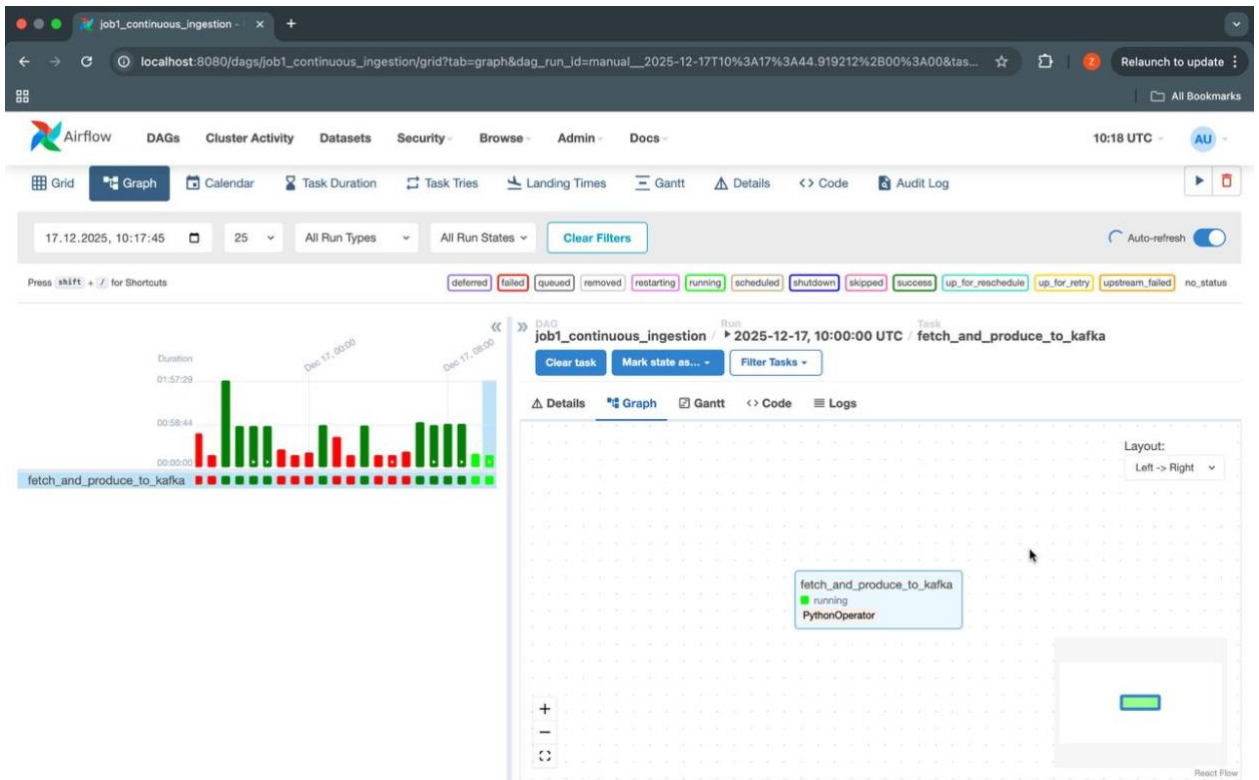
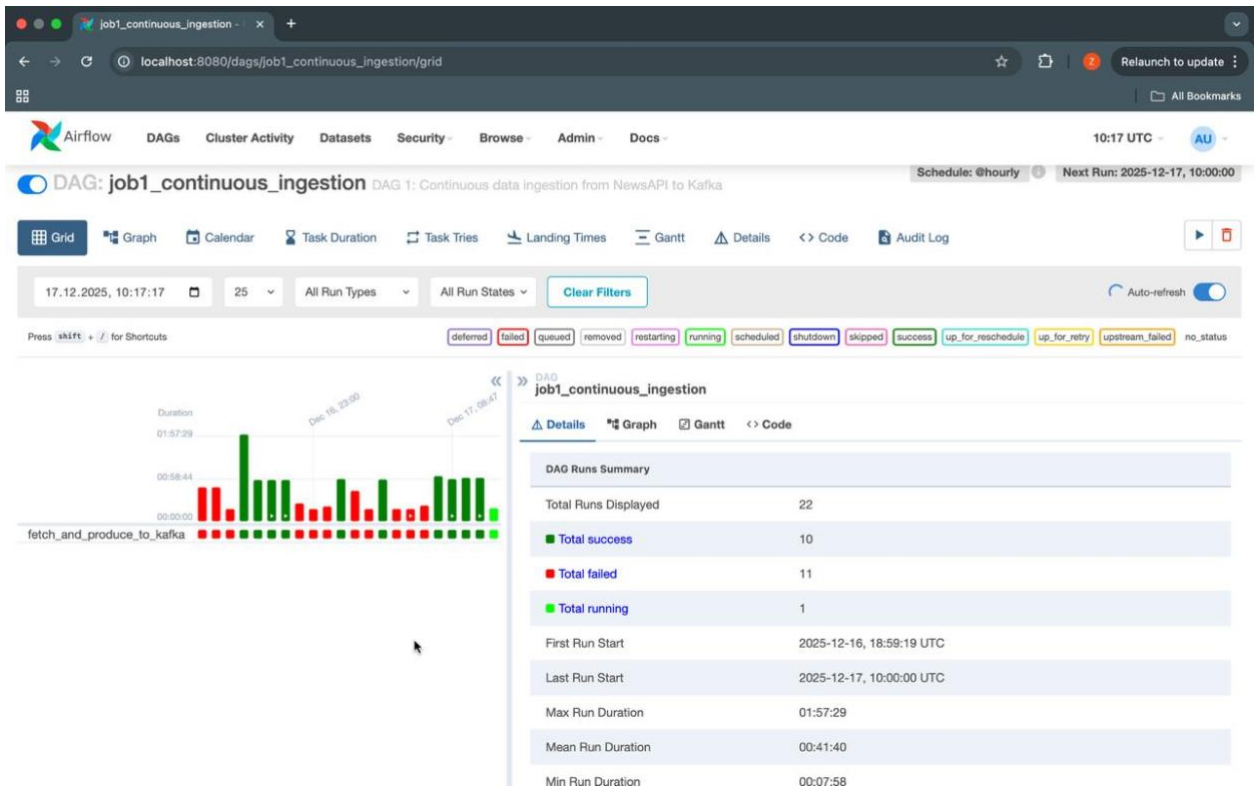
Max Run Duration01:57:29

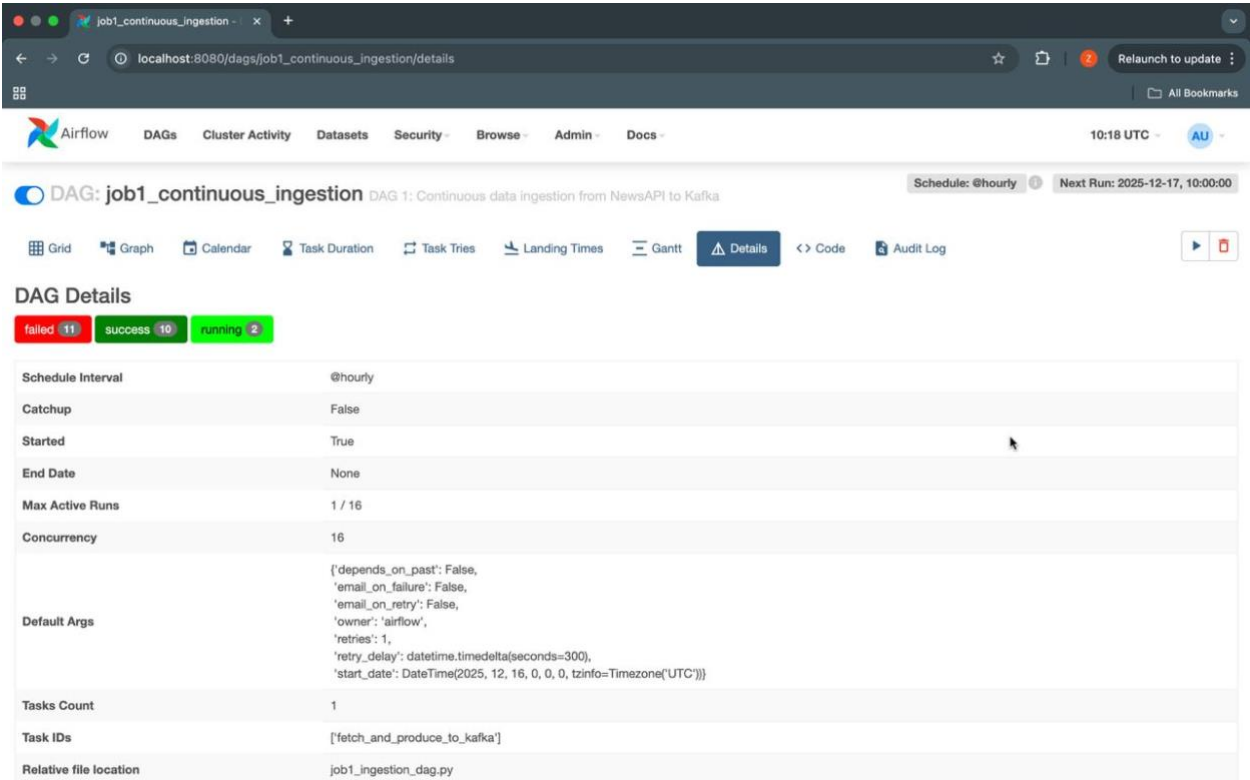
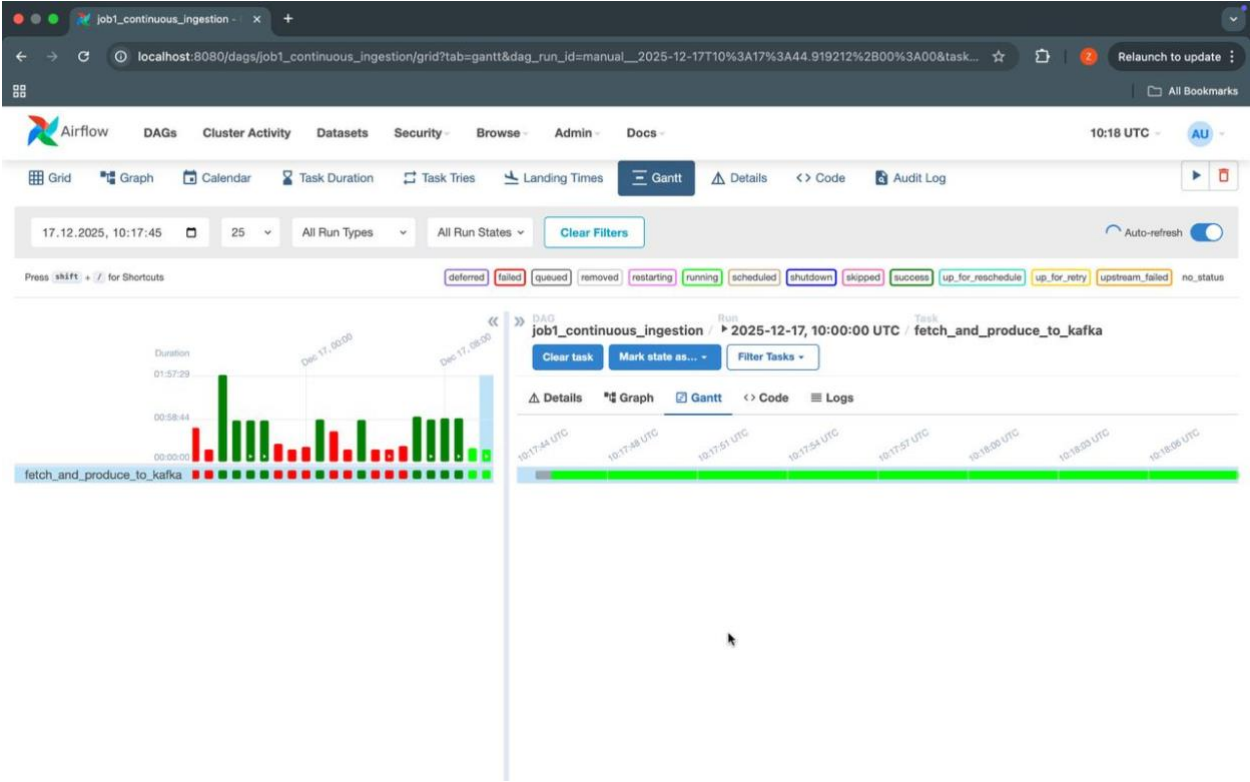
Mean Run Duration00:41:40

Min Run Duration00:07:58

Job1_continuous_ingestion:







job1_continuous_ingestion

localhost:8080/dags/job1_continuous_ingestion/details

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

10:18 UTC

AU

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')))

job1_continuous_ingestion

localhost:8080/dags/job1_continuous_ingestion/details

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

10:18 UTC

AU

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 0xffff9f683070>

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:c08c82e9dd0e4aaba5121519819a636df635210

job3_daily_analytics - Airflow x List Log - Airflow x List Log - Airflow x List Log - Airflow x +

localhost:8080/log/list/?_fit_3_dag_id=job1_continuous_ingestion

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

List Log

Search -

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 x +

localhost:8080/dags/job2_hourly_cleaning/grid

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 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

clean_and_store_data

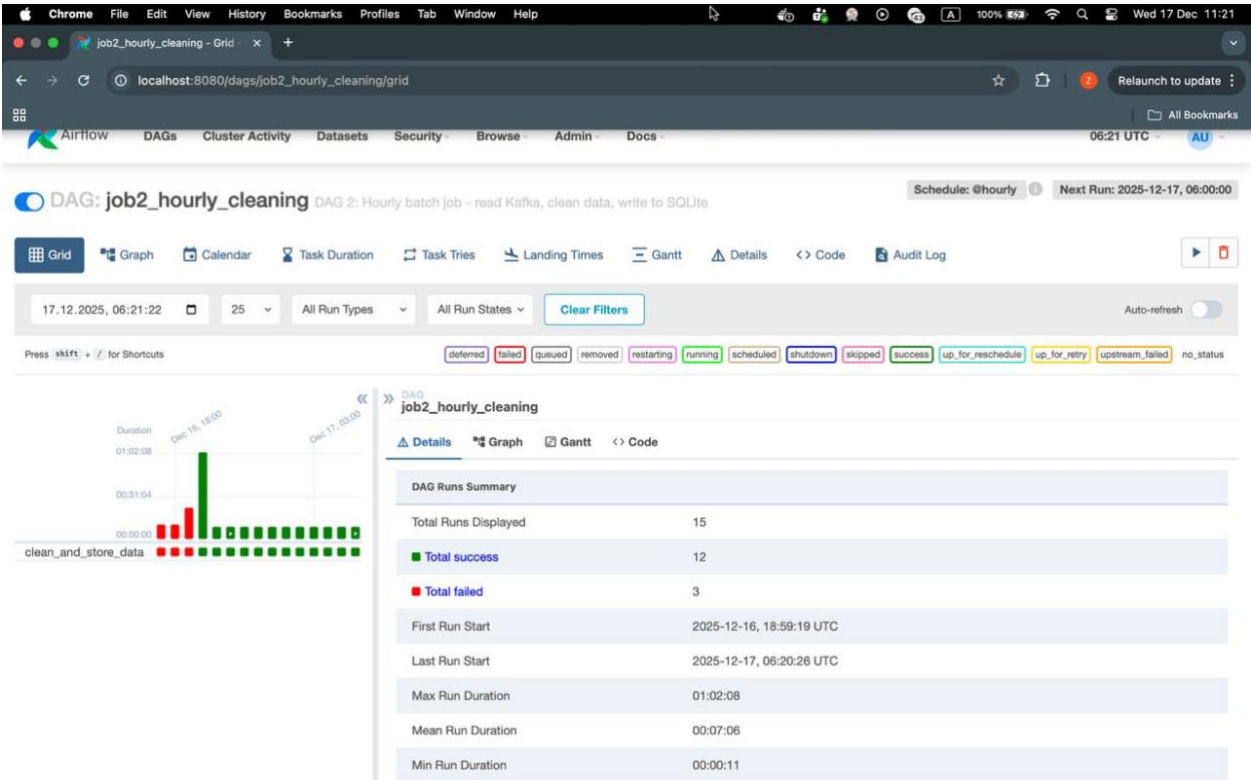
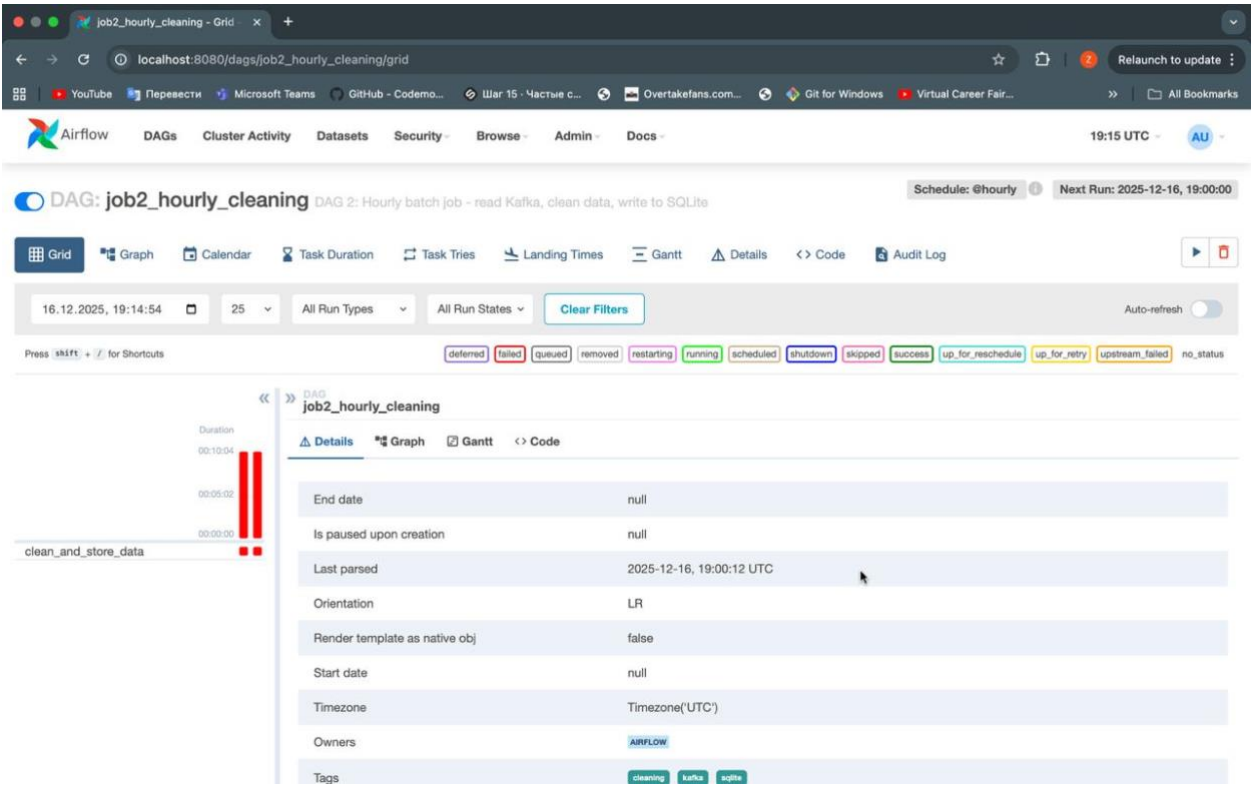
DAG job2_hourly_cleaning

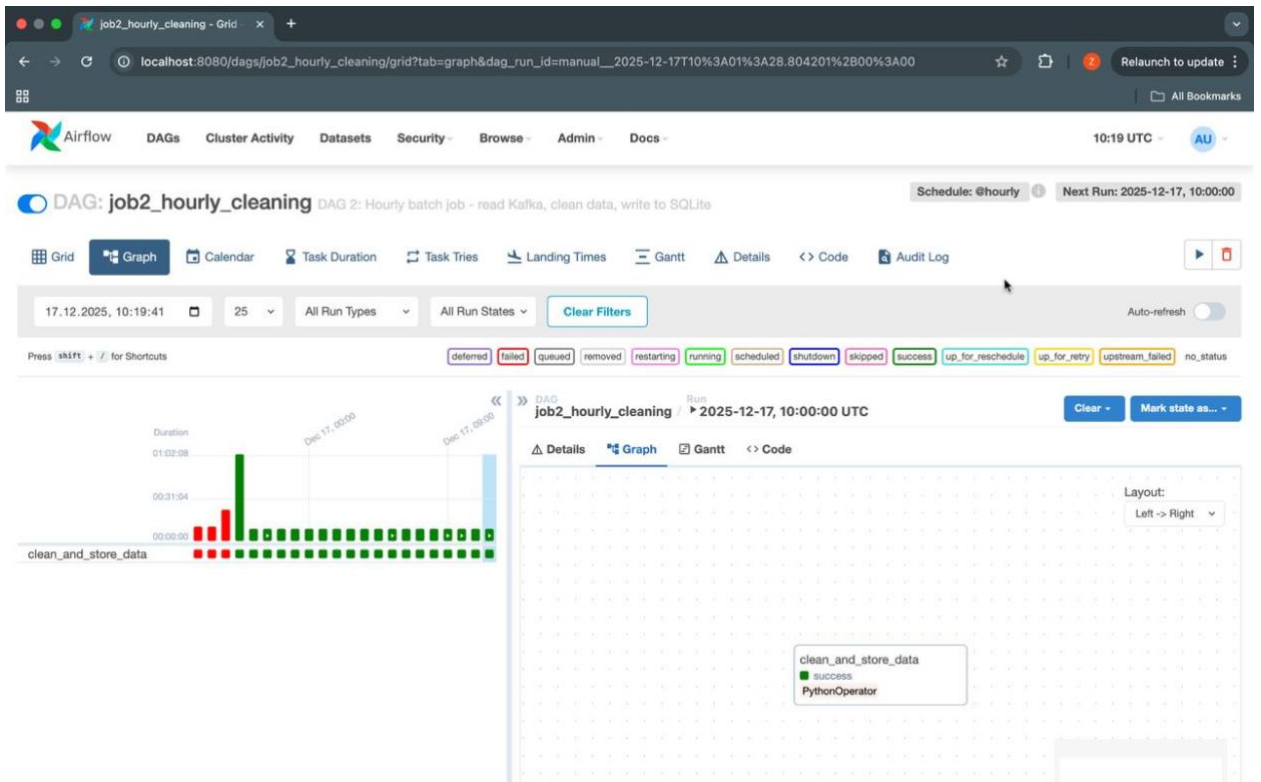
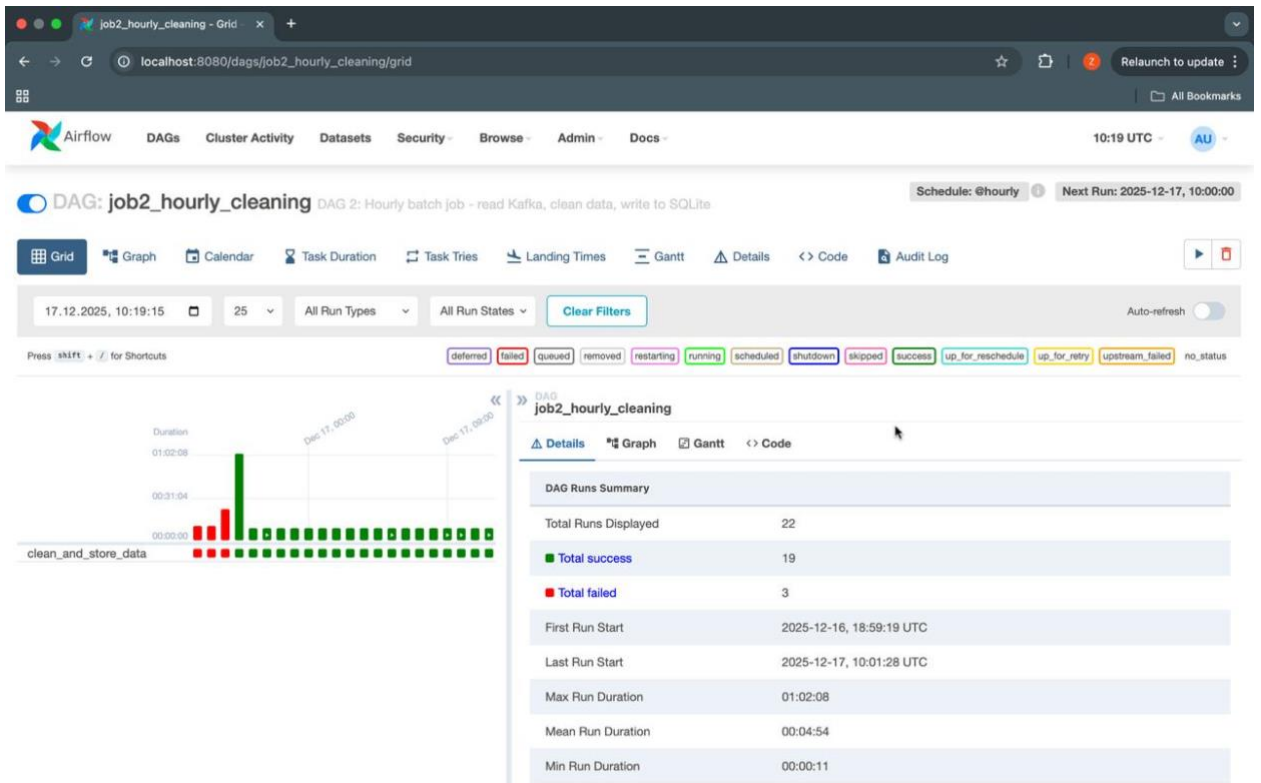
Details Graph Gantt Code

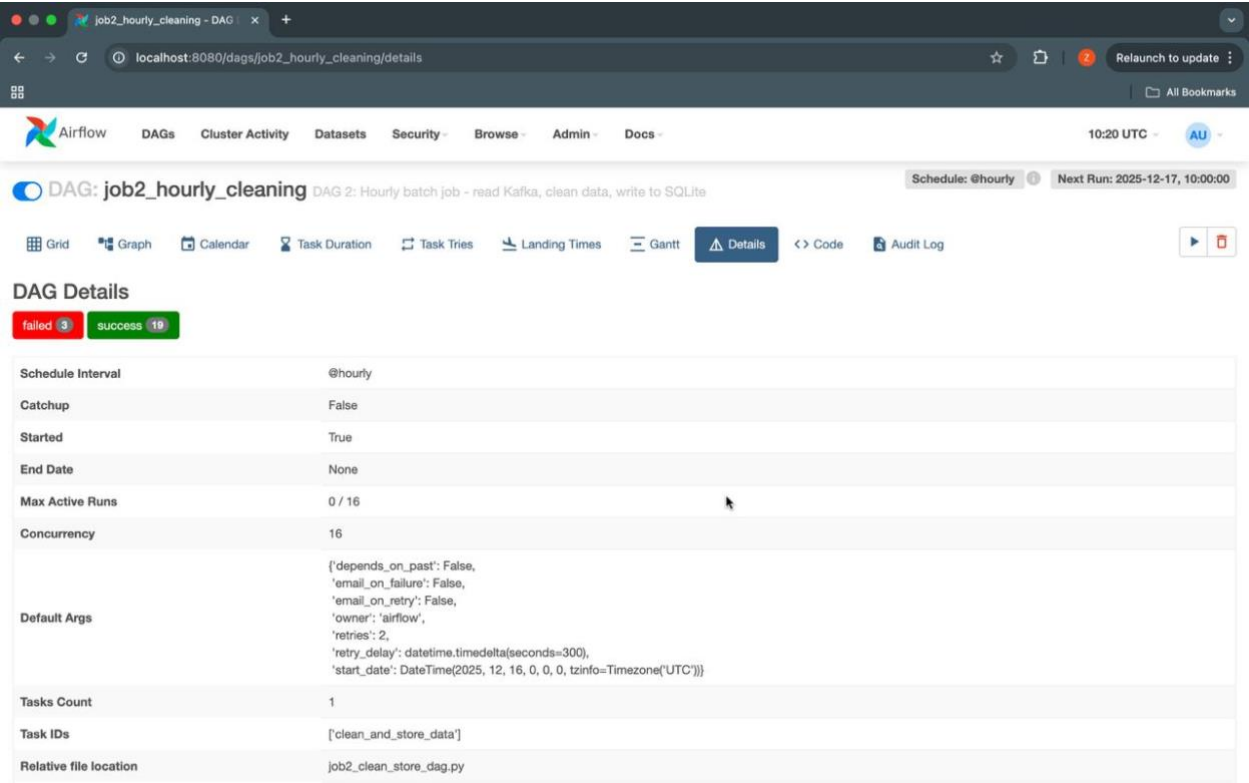
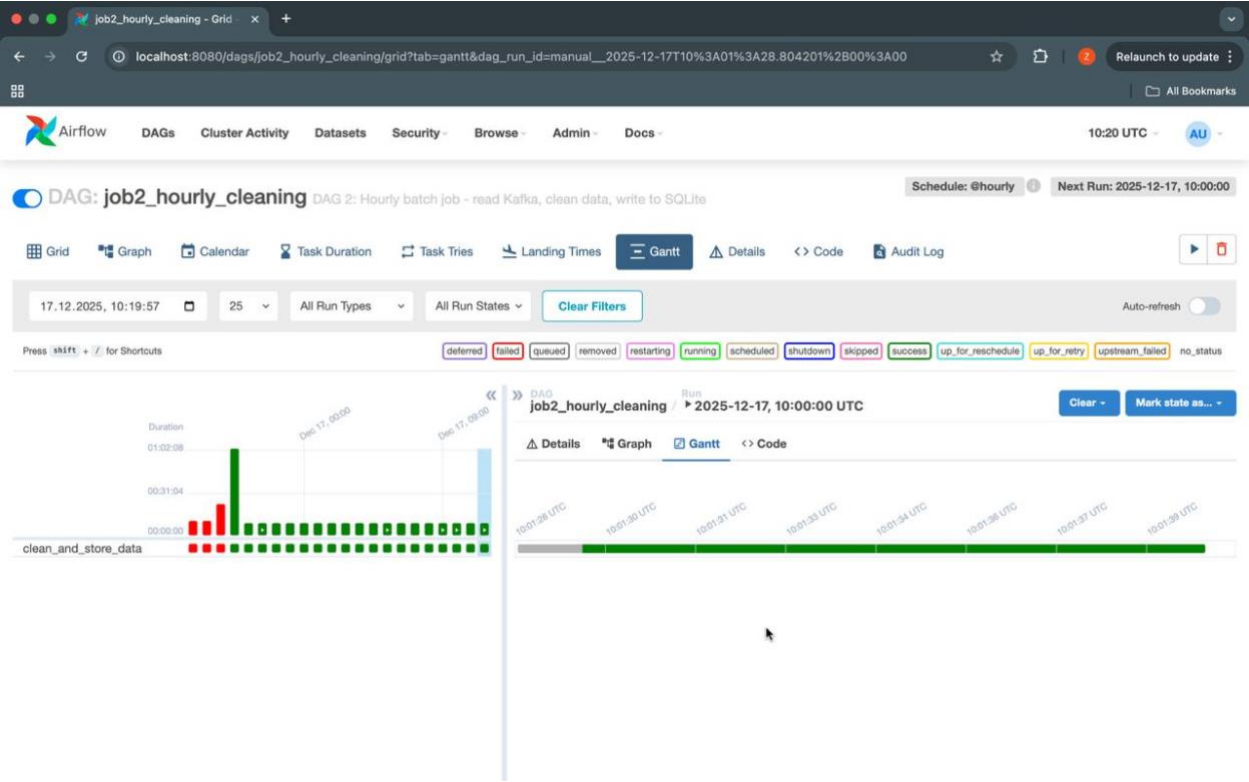
DAG Runs Summary

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







job2_hourly_cleaning - DAG

localhost:8080/dags/job2_hourly_cleaning/details

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

10:20 UTC

AU

Owner

airflow

Owner Links

None

DAG Run Timeout

None

Tags

cleaning

kafka

sqlite

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')))

job2_hourly_cleaning - DAG

localhost:8080/dags/job2_hourly_cleaning/details

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

10:20 UTC

AU

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 0xfffff683070>
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:c08c82e9dd0e4aaba5121519819a636df635210

job2_hourly_cleaning - Airflow

List Log - Airflow

List Log - Airflow

List Log - Airflow

List Log - Airflow

localhost:8080/log/list?_fit_3_dag_id=job2_hourly_cleaning

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

10:24 UTC

AU

List Log

Search

1

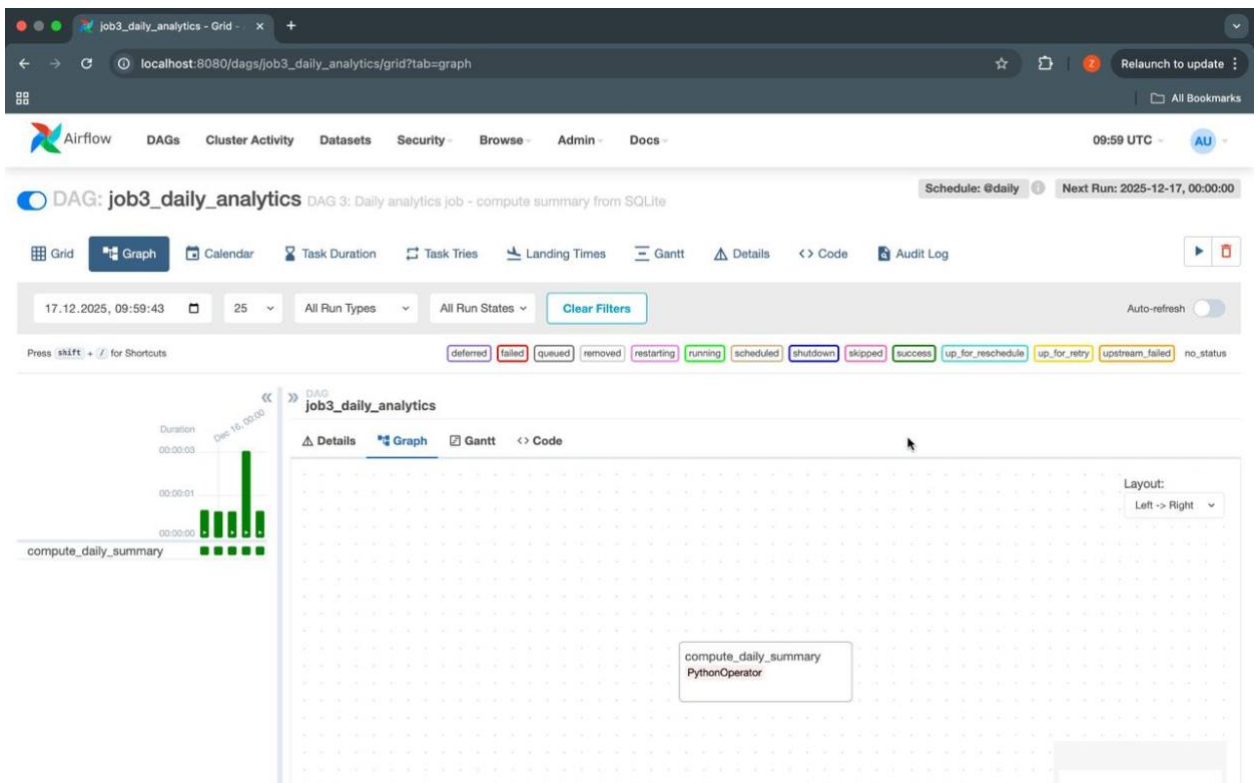
2

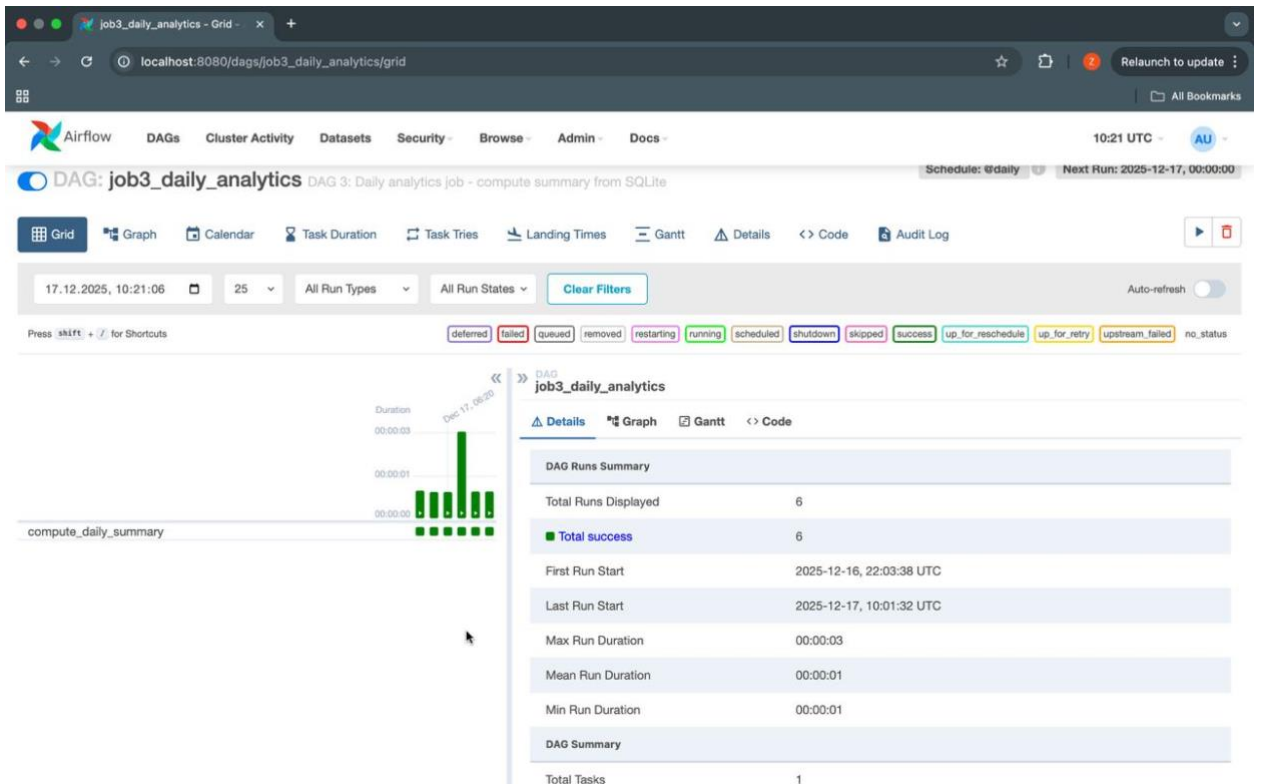
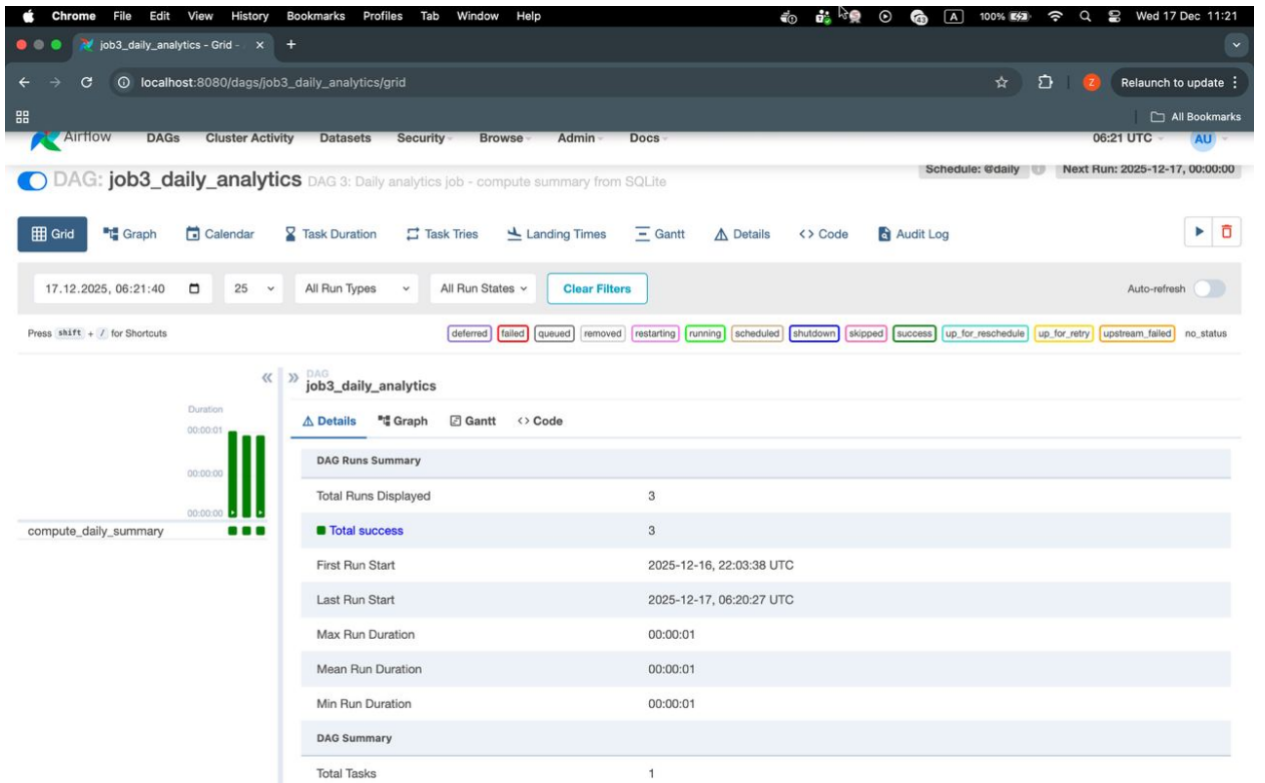
Page size

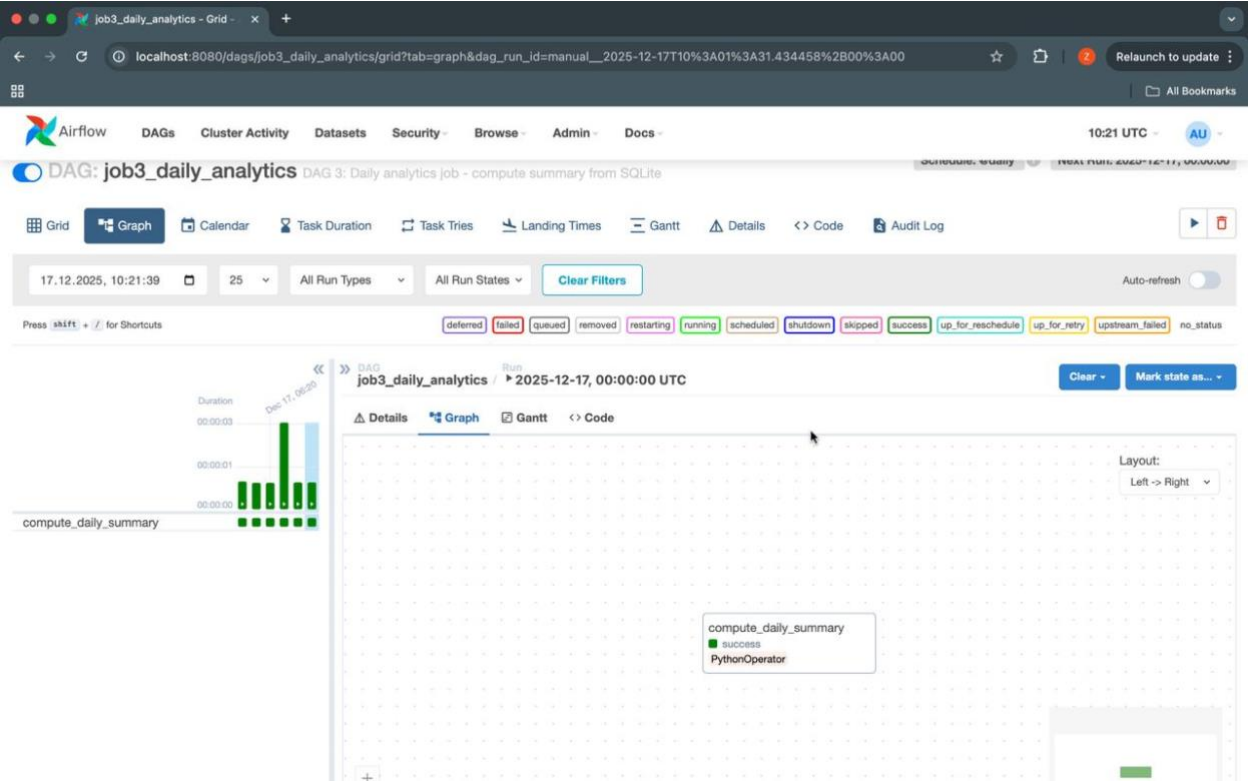
Record Count: 150

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	job2_hourly_cleaning		grid		Admin	[('root', ''), ('tab', 'gantt'), ('dag_run_id', 'manual__2025-12-17T10:01:28.804201+00:00'), ('dag_id', 'job2_hourly_cleaning')]

Job3_daily_analytics:







job3_daily_analytics - DAG

localhost:8080/dags/job3_daily_analytics/details

Relaunch to update

10:22 UTC AU

DAG: job3_daily_analytics DAG 3: Daily analytics job - compute summary from SQLite Schedule: @daily Next Run: 2025-12-17, 00:00:00

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

DAG Details

success 6


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

job3_daily_analytics - DAG

localhost:8080/dags/job3_daily_analytics/details

Relaunch to update

All Bookmarks

 DAGs Cluster Activity Datasets Security Browse Admin Docs

10:22 UTC AU

Ownerairflow

Owner LinksNone

DAG Run TimeoutNone

Tagsanalytics sqlite summary

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')))

job3_daily_analytics - DAG

localhost:8080/dags/job3_daily_analytics/details

Relaunch to update

All Bookmarks

 DAGs Cluster Activity Datasets Security Browse Admin Docs

10:22 UTC AU

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: .releasecc08c82e9dd0e4aaba5121519819a636df635210

job2_hourly_cleaning - Airflow

List Log - Airflow

List Log - Airflow

List Log - Airflow

List Log - Airflow

List Log - Airflow

localhost:8080/log/list?_flt_3_dag_id=job3_daily_analytics

Relaunch to update

All Bookmarks

Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

10:24 UTC

AU

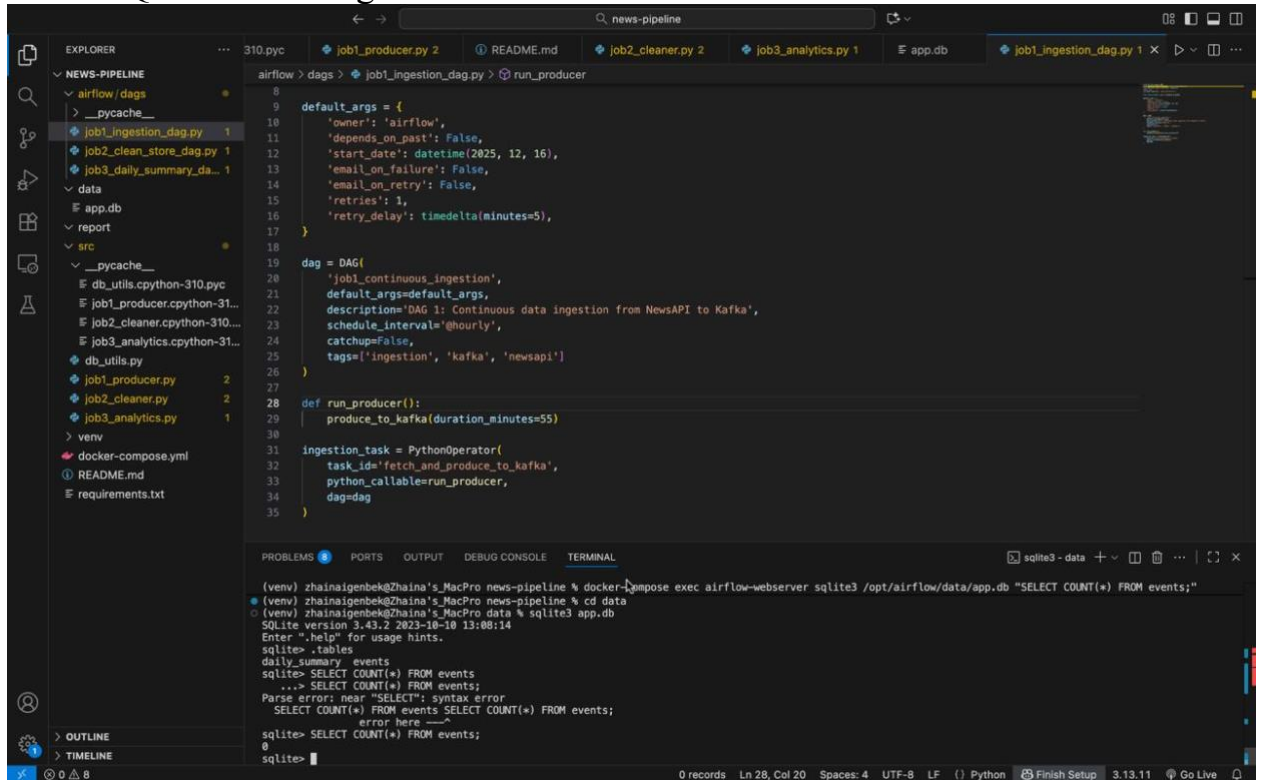
List Log

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": "152303e5a456", "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:

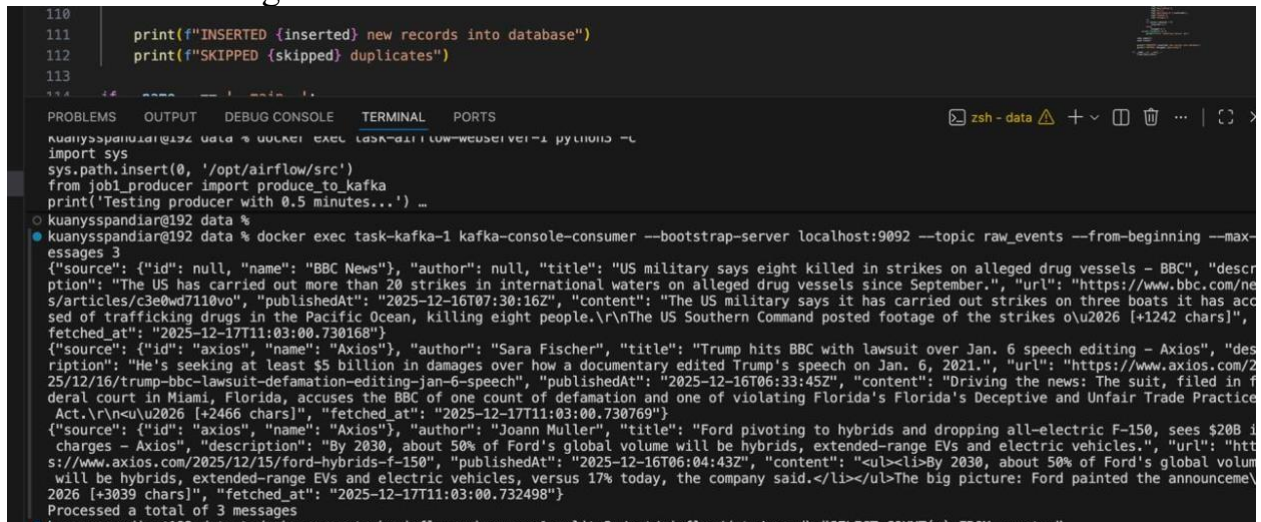


The screenshot shows a VS Code editor with a file explorer on the left displaying a project named 'NEWS-PIPELINE'. The main editor shows a file named 'job1_ingestion_dag.py' with the following code:

```
8
9
10 default_args = {
11     'owner': 'airflow',
12     'depends_on_past': False,
13     'start_date': datetime(2025, 12, 16),
14     'email_on_failure': False,
15     'email_on_retry': False,
16     'retries': 1,
17     'retry_delay': timedelta(minutes=5),
18 }
19
20 dag = DAG(
21     'job1_continuous_ingestion',
22     default_args=default_args,
23     description='DAG 1: Continuous data ingestion from NewsAPI to Kafka',
24     schedule_interval='@hourly',
25     catchup=False,
26     tags=['ingestion', 'kafka', 'newsapi'],
27 )
28
29 def run_producer():
30     produce_to_kafka(duration_minutes=55)
31
32 ingestion_task = PythonOperator(
33     task_id='fetch_and_produce_to_kafka',
34     python_callable=run_producer,
35     dag=dag
36 )
```

The terminal window at the bottom shows the command: `(venv) zhainagenbek@zhainas-MacPro news-pipeline % docker-compose exec airflow-webserver sqlite3 /opt/airflow/data/app.db "SELECT COUNT(*) FROM events;"`. The output shows the SQLite version and then a syntax error: `Parse error: near "SELECT": syntax error`.

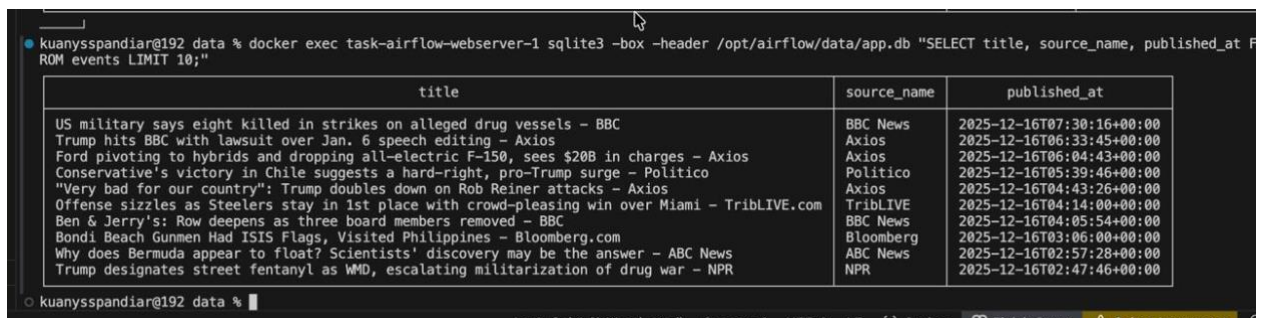
After fixing errors:



The screenshot shows a VS Code editor with a file named 'job1_producer.py' containing the following code:

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

The terminal window at the bottom shows the command: `kuanyspandiar@192 data % docker exec task-airflow-webserver-1 sqlite3 /opt/airflow/data/app.db "SELECT COUNT(*) FROM events;"`. The output shows the SQLite version and then the result of the query: `1`.



The screenshot shows a terminal window with the command: `kuanyspandiar@192 data % docker exec task-airflow-webserver-1 sqlite3 -box -header /opt/airflow/data/app.db "SELECT title, source_name, published_at FROM events LIMIT 10;"`. The output is a table with 3 columns: title, source_name, and published_at.

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

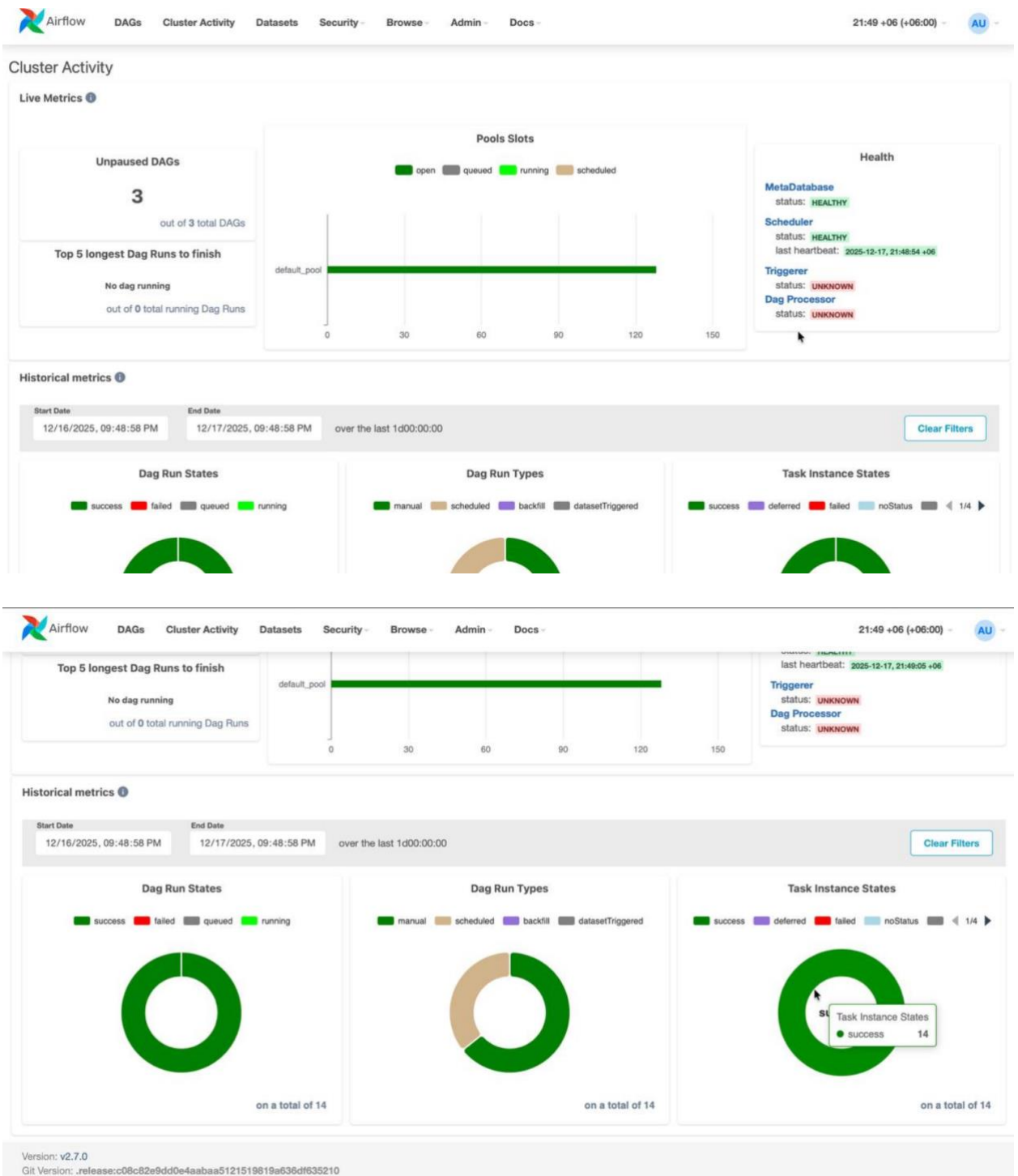
Terminal window showing a SQL query and its results in a table format. The query is: `sqlite3 -box data/app.db "SELECT id, source_name, author, description, category FROM events LIMIT 15;"`

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	Juliana Bragg	"I wasn't a fan of Rok 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 WMD. Experts on street drugs and fentanyl are skeptical these moves will reduce the supply of fentan...	General

Airflow											21:50 +06 (+06:00)	AU
DAGs Cluster Activity Datasets Security Browse Admin Docs												
	State	Dag Id	Task Id	Run Id	Map Index	Logical Date	Operator	Start Date	End Date	Duration	Note	Job Id
<input type="checkbox"/>	SUCCESS	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	2M1s		2
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	job3_daily_analytics	compute_daily_summary	manual_2025-12-17T12:08:15.647713+00:00		2025-12-17, 18:08:15	PythonOperator	2025-12-17, 18:08:16	2025-12-17, 18:08:17	<1s		7
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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	2M1s		46
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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	2M1s		50
<input type="checkbox"/>	SUCCESS	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
<input type="checkbox"/>	SUCCESS	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

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		success	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		ebee7d598937	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		934d07147bb7	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		934d07147bb7	airflow
45		failed	SchedulerJob	2025-12-17, 21:15:14		2025-12-17, 21:30:03		934d07147bb7	airflow
14	job2_hourly_cleaning	success	LocalTaskJob	2025-12-17, 21:12:08		2025-12-17, 21:12:38		934d07147bb7	airflow
13	job1_continuous_ingestion	success	LocalTaskJob	2025-12-17, 21:12:08		2025-12-17, 21:12:38		934d07147bb7	airflow
12		failed	SchedulerJob	2025-12-17, 21:12:06		2025-12-17, 21:12:42		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	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		934d07147bb7	airflow
1		failed	SchedulerJob	2025-12-17, 18:07:31		2025-12-17, 18:35:54		934d07147bb7	airflow

	State	Dag Id	Logical Date	Run Id	Run Type	Queued At	Start Date	End Date	Note	External Trigger	Conf	Duration
<input type="checkbox"/>	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		()	4s
<input type="checkbox"/>	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		()	18s
<input type="checkbox"/>	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		()	2M:9s
<input type="checkbox"/>	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		()	5M:9s
<input type="checkbox"/>	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		()	3M:19s
<input type="checkbox"/>	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		()	2s
<input type="checkbox"/>	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		()	1s
<input type="checkbox"/>	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		()	15s
<input type="checkbox"/>	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		()	2M:2s
<input type="checkbox"/>	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		()	2s
<input type="checkbox"/>	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		()	12s
<input type="checkbox"/>	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		()	2M:2s
<input type="checkbox"/>	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		()	13s
<input type="checkbox"/>	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		()	2M:3s
<input type="checkbox"/>	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		()	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.