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 every minute)
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)

Validation

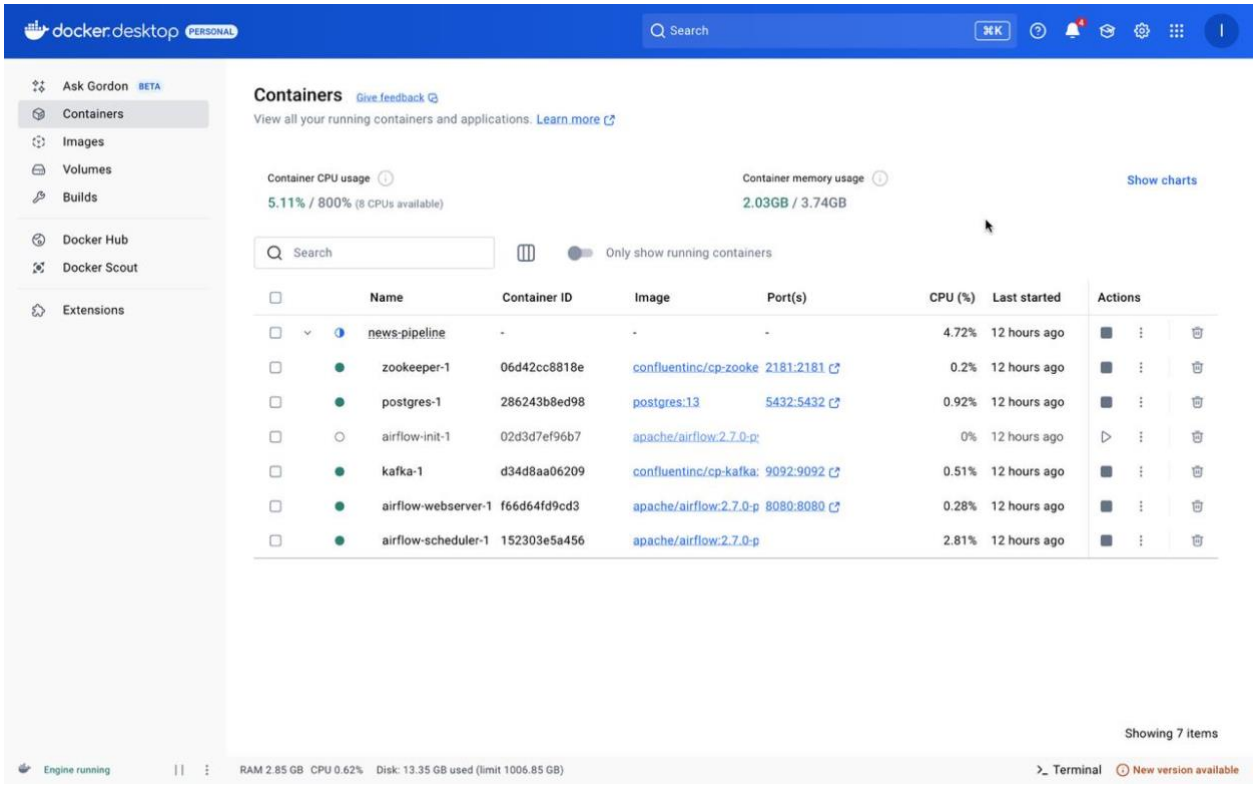
- DAGs execute on schedule (@hourly/@daily)
- Kafka messages successfully produced/consumed
- SQLite tables populated with clean data

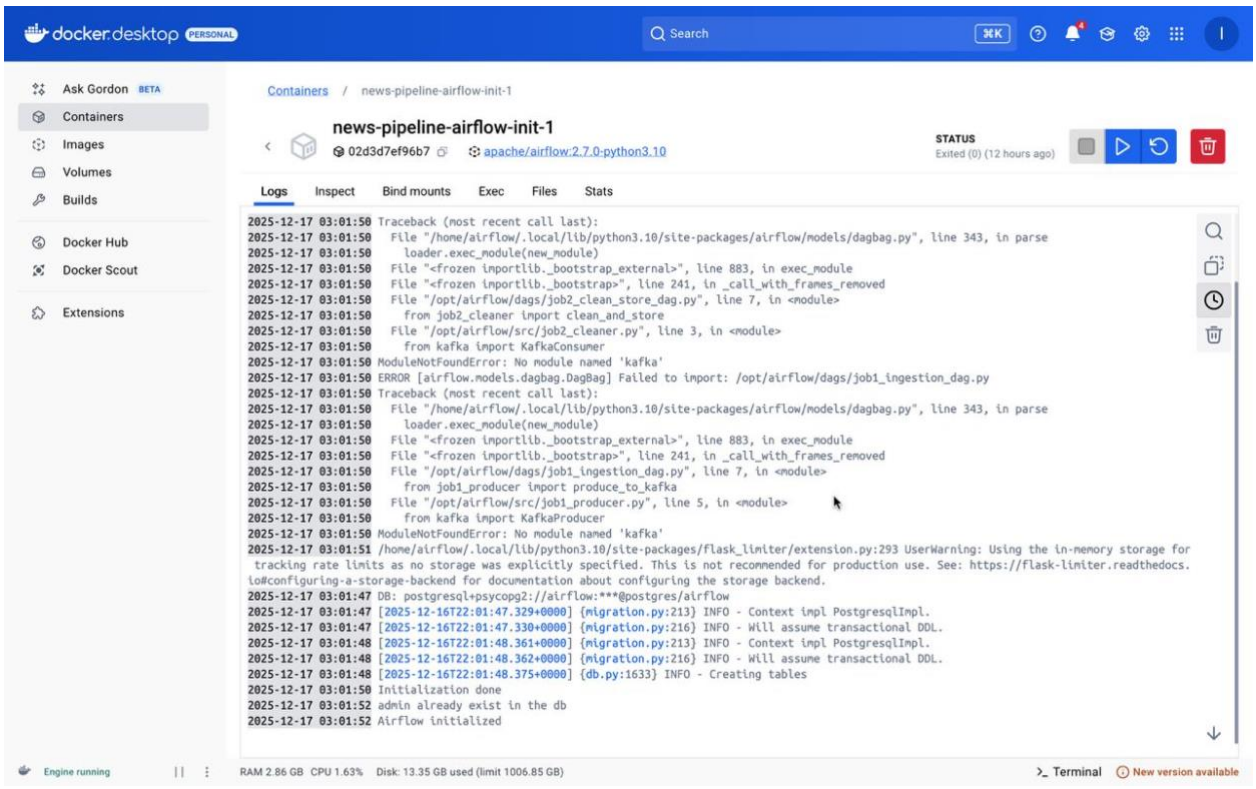
- Daily summaries computed automatically
- All services run in Docker containers

Appendices

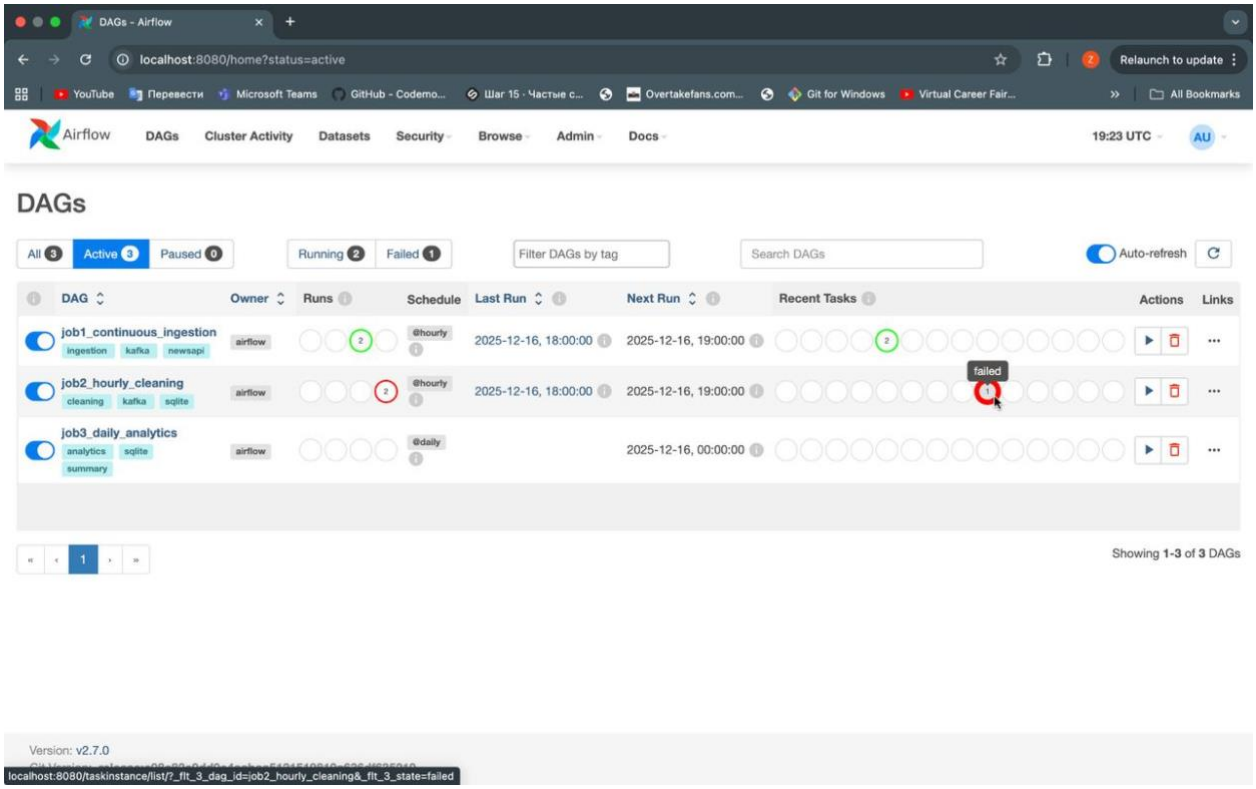
Screenshots were taken throughout the whole process. Before and after error fixings.

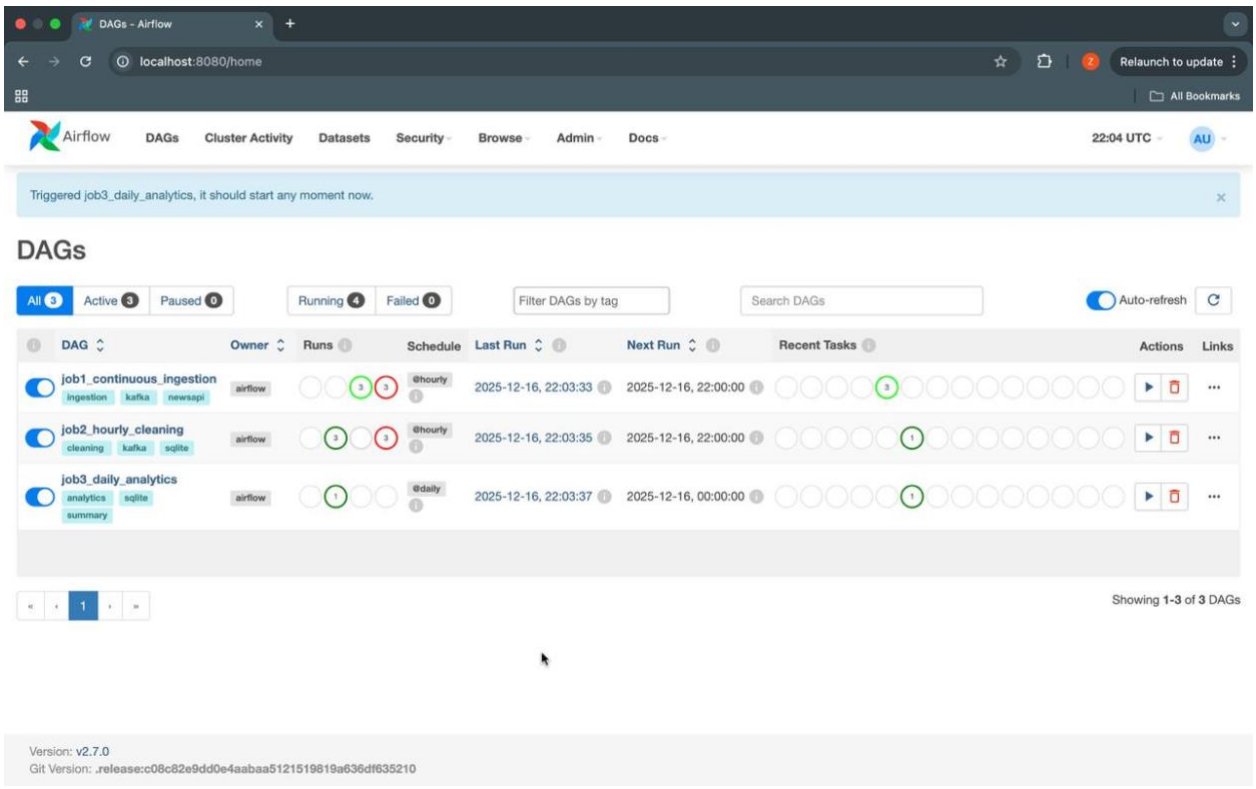
Docker desktop:

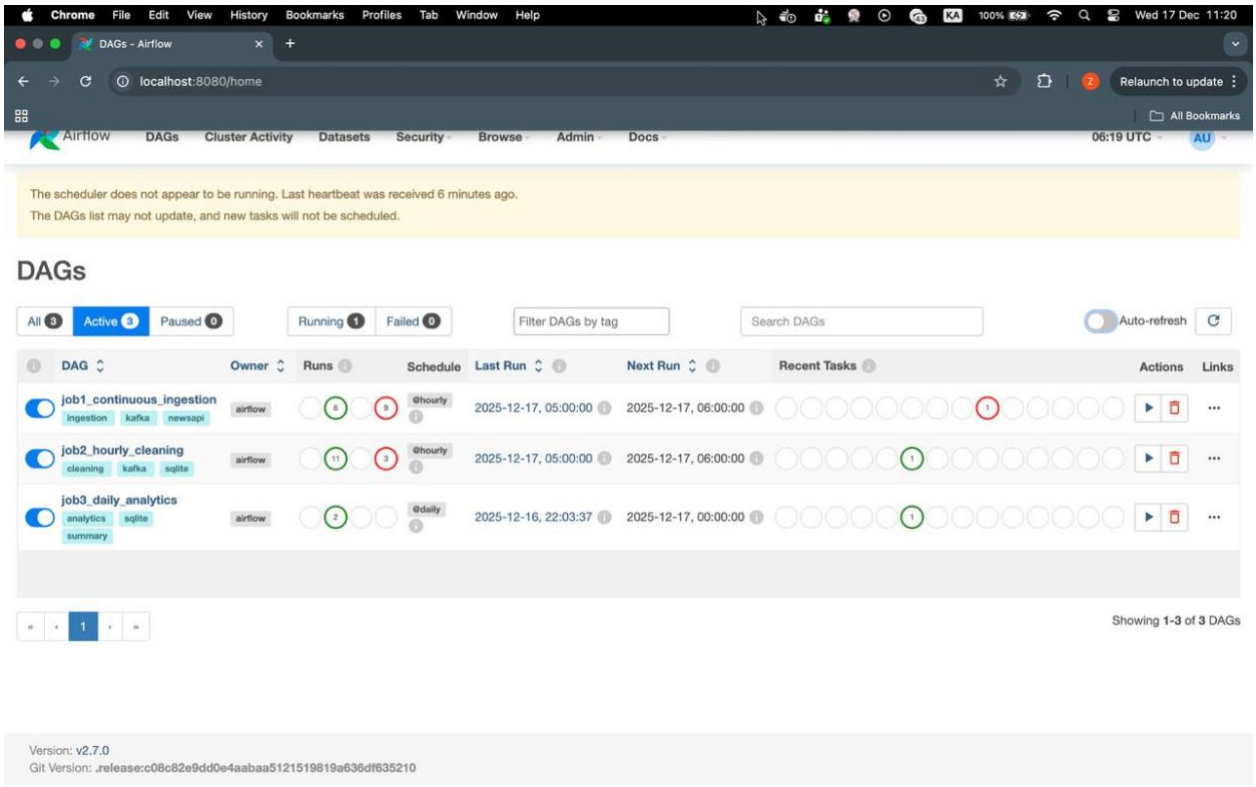


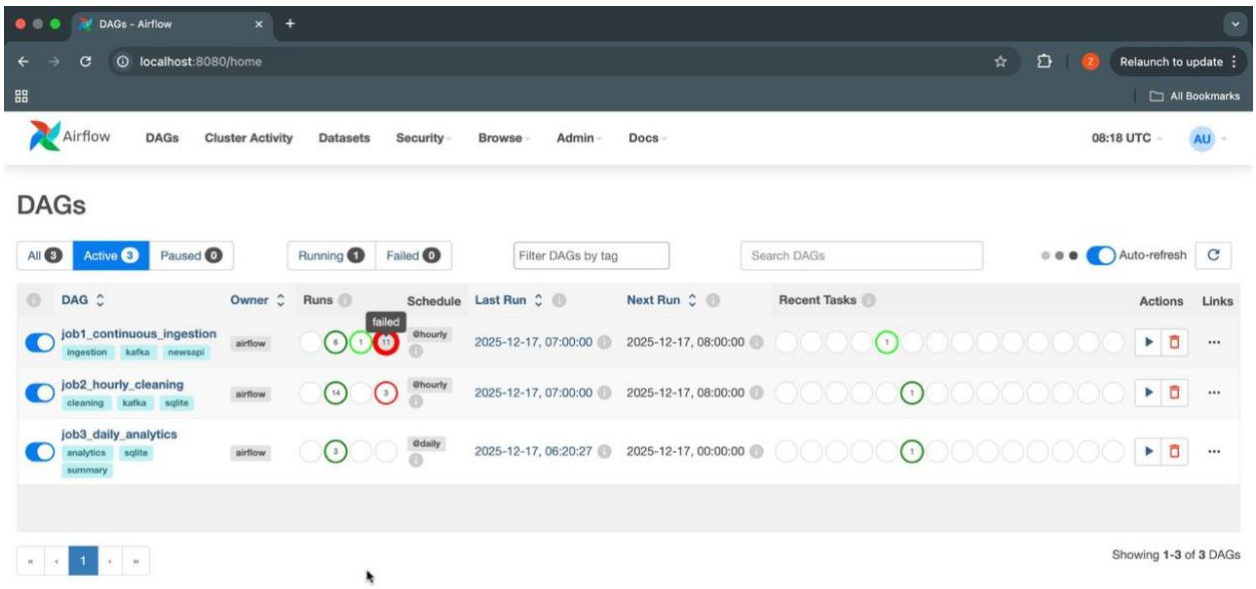

docker desktop PERSONAL
Search
Ask Gordon BETA
Containers
Images
Volumes
Builds
Docker Hub
Docker Scout
Extensions
Containers / news-pipeline-airflow-init-1
news-pipeline-airflow-init-1
02d3d7ef96b7
apache/airflow:2.7.0-python3.10
STATUS
Exited (0) (12 hours ago)
Logs
Inspect
Bind mounts
Exec
Files
Stats
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 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 1006.85 GB)
Terminal
New version available

DAGs:


DAGs - Airflow
localhost:8080/home?status=active
Relaunch to update
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
19:23 UTC
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
@hourly
2025-12-16, 18:00:00
2025-12-16, 19:00:00
job2_hourly_cleaning
cleaning
kafka
sqlite
airflow
@hourly
2025-12-16, 18:00:00
2025-12-16, 19:00:00
failed
job3_daily_analytics
analytics
sqlite
summary
airflow
@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

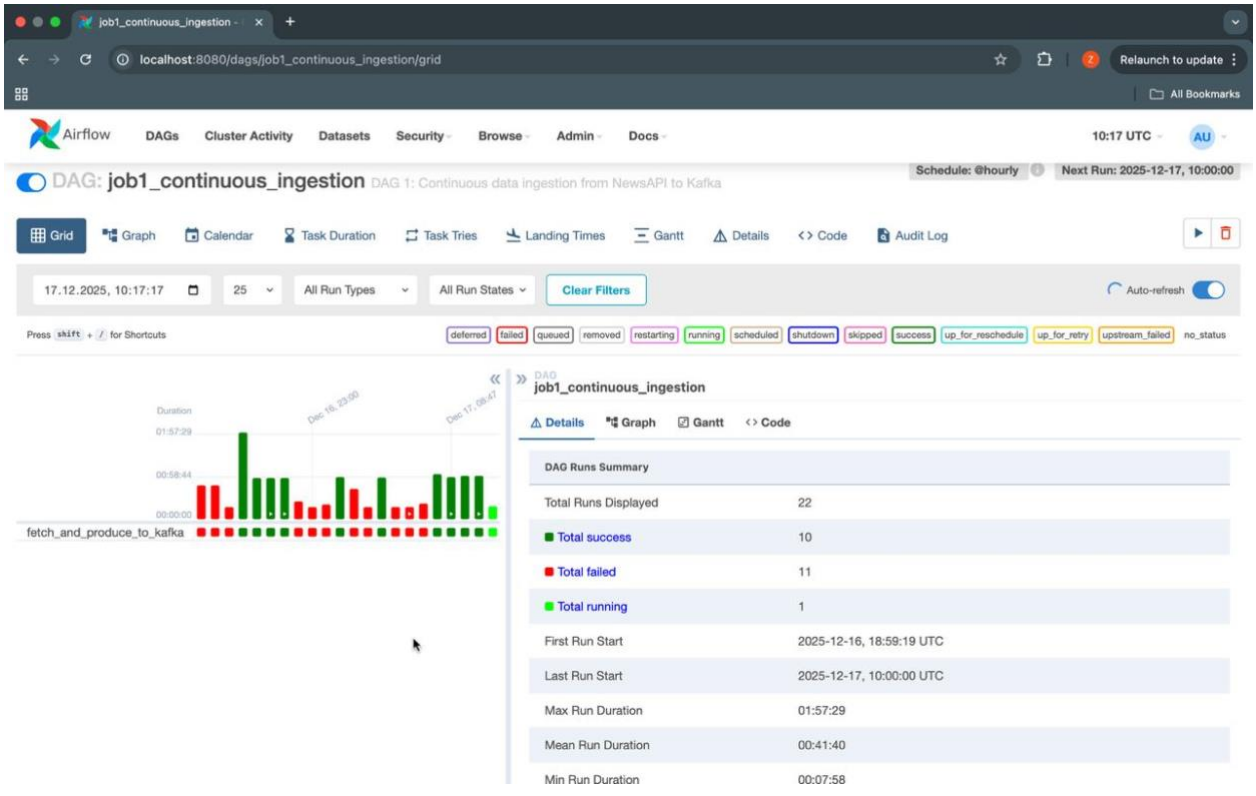

DAGs - Airflow
localhost:8080/home
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
22:04 UTC
AU
Triggered job3_daily_analytics, it should start any moment now.
DAGs
All 3
Active 3
Paused 0
Running 4
Failed 0
Filter DAGs by tag
Search DAGs
Auto-refresh
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
Actions
Links
job1_continuous_ingestion
ingestion
kafka
newsapi
airflow
@hourly
2025-12-16, 22:03:33
2025-12-16, 22:00:00
job2_hourly_cleaning
cleaning
kafka
sqlite
airflow
@hourly
2025-12-16, 22:03:35
2025-12-16, 22:00:00
job3_daily_analytics
analytics
sqlite
summary
airflow
@daily
2025-12-16, 22:03:37
2025-12-16, 00:00:00
Showing 1-3 of 3 DAGs
Version: v2.7.0
Git Version: .release:c08c82e9dd0e4aabaa5121519819a636df635210


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Wed 17 Dec 11:20
DAGs - Airflow
localhost:8080/home
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
06:19 UTC
AU
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
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
Actions
Links
job1_continuous_ingestion
ingestion
kafka
newsapi
airflow
@hourly
2025-12-17, 05:00:00
2025-12-17, 06:00:00
job2_hourly_cleaning
cleaning
kafka
sqlite
airflow
@hourly
2025-12-17, 05:00:00
2025-12-17, 06:00:00
job3_daily_analytics
analytics
sqlite
summary
airflow
@daily
2025-12-16, 22:03:37
2025-12-17, 00:00:00
Showing 1-3 of 3 DAGs
Version: v2.7.0
Git Version: .release:c08c82e9dd0e4aabaa5121519819a636df635210

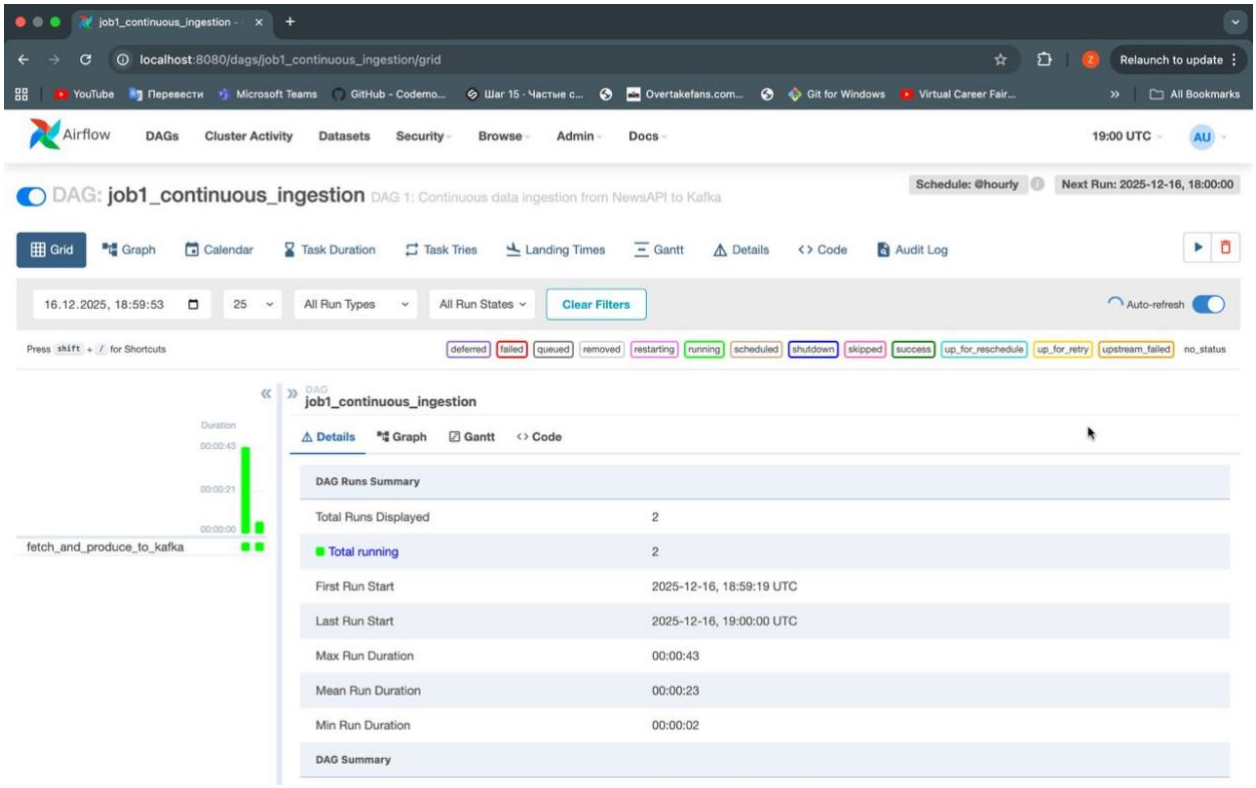


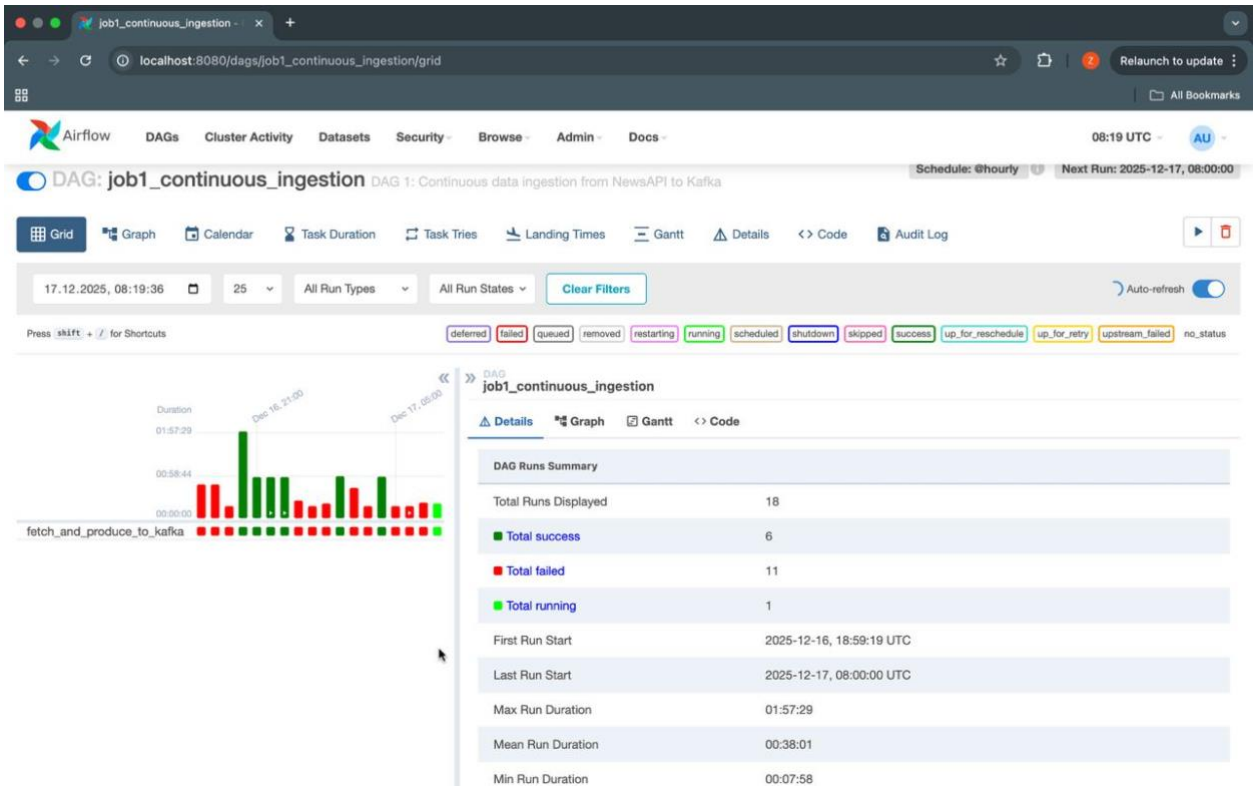
Version: v2.7.0

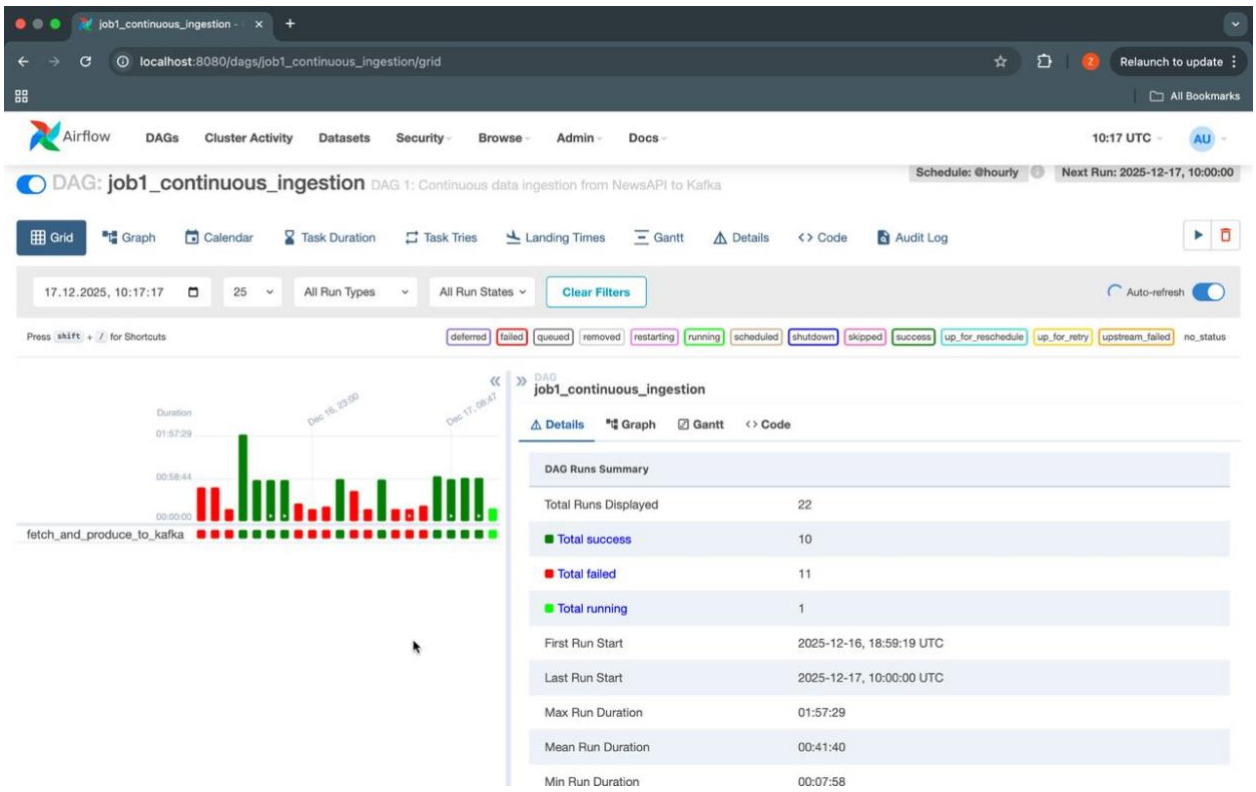
localhost:8080/dagrun/list/?_flt_3_dag_id=job1_continuous_ingestion&_flt_3_state=failed

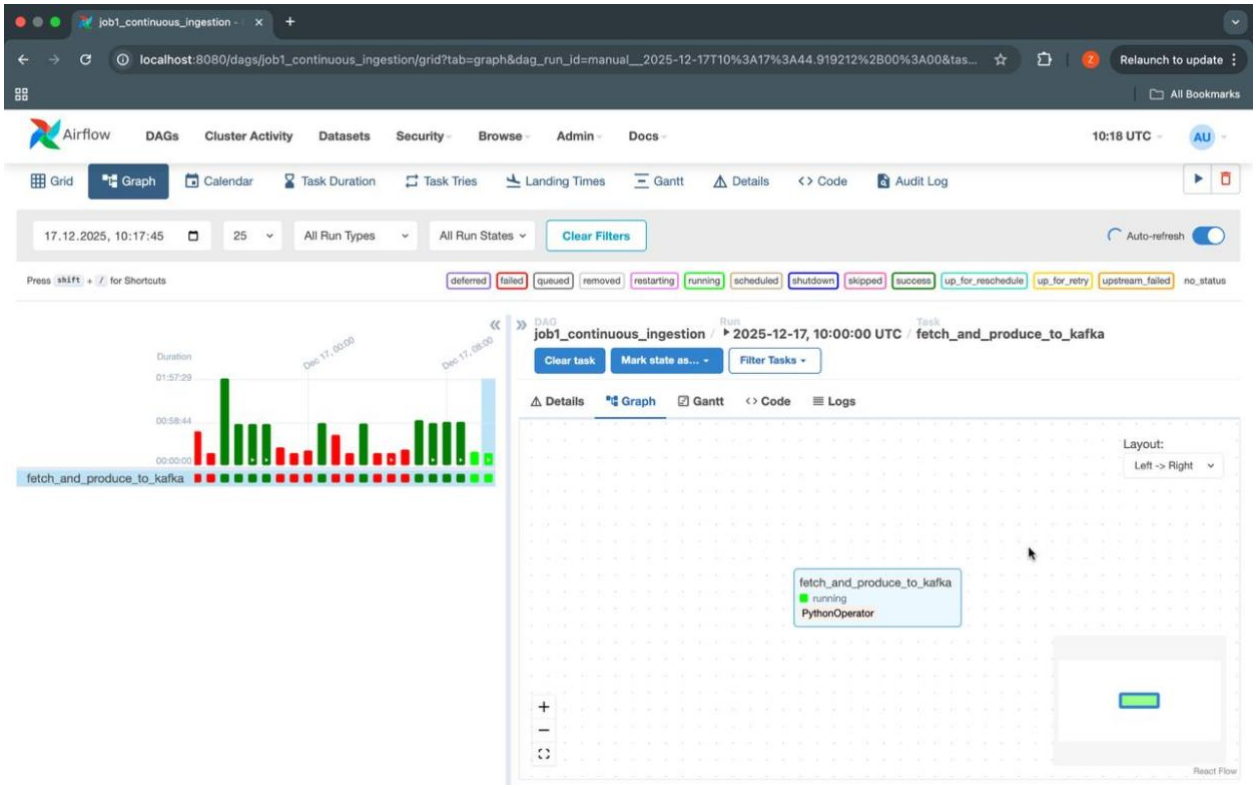


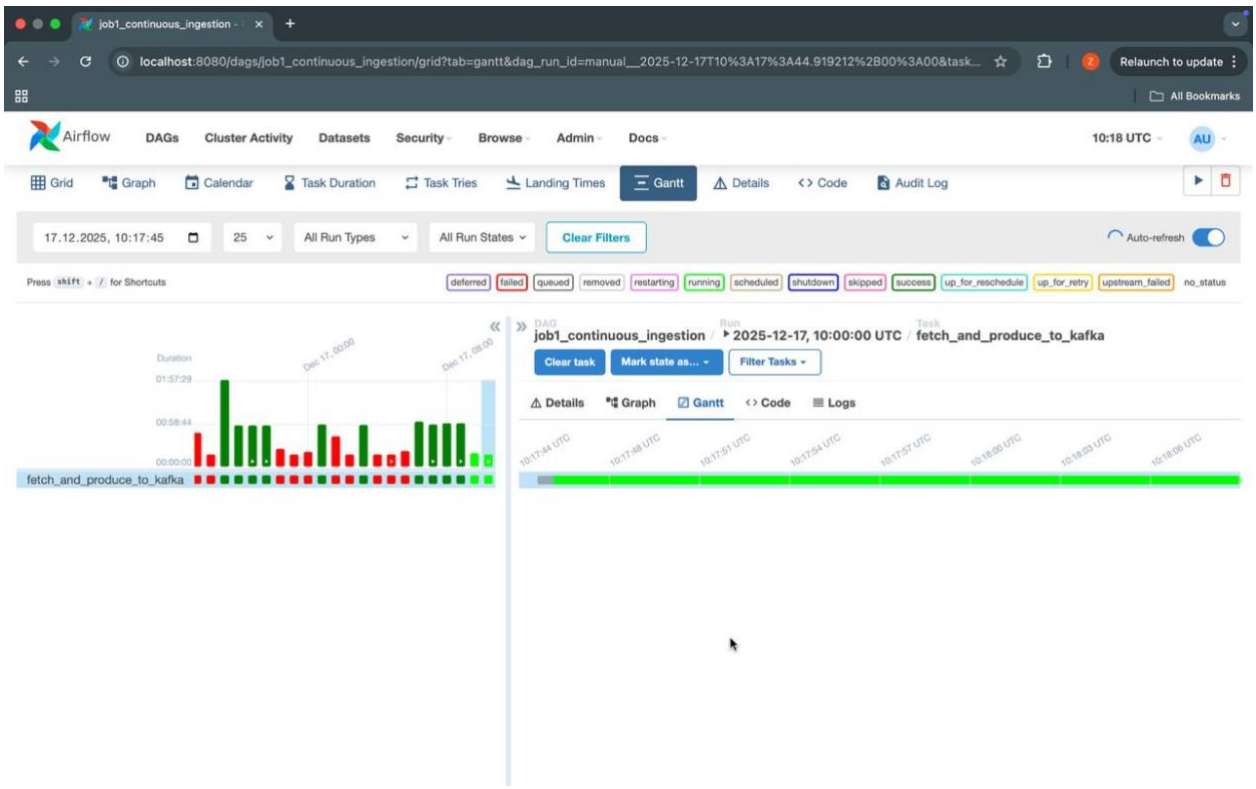
Job1_continuous_ingestion:

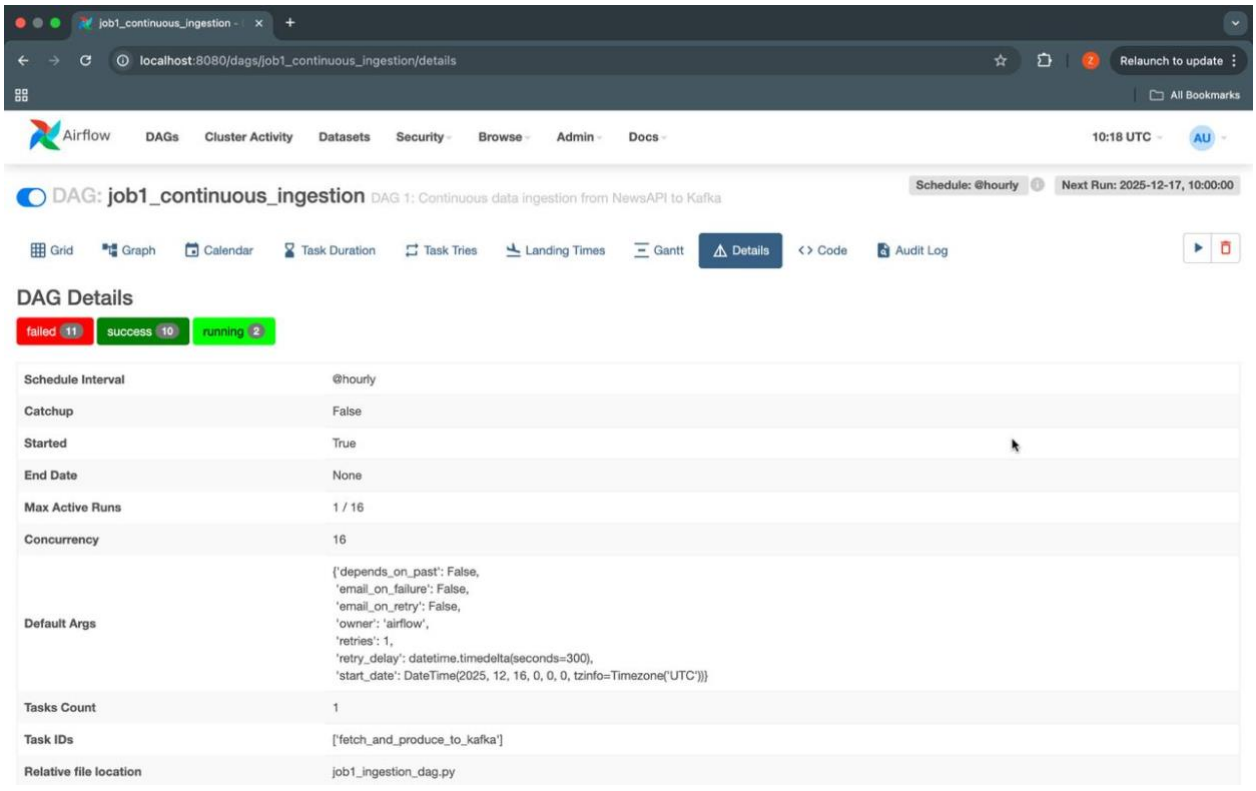


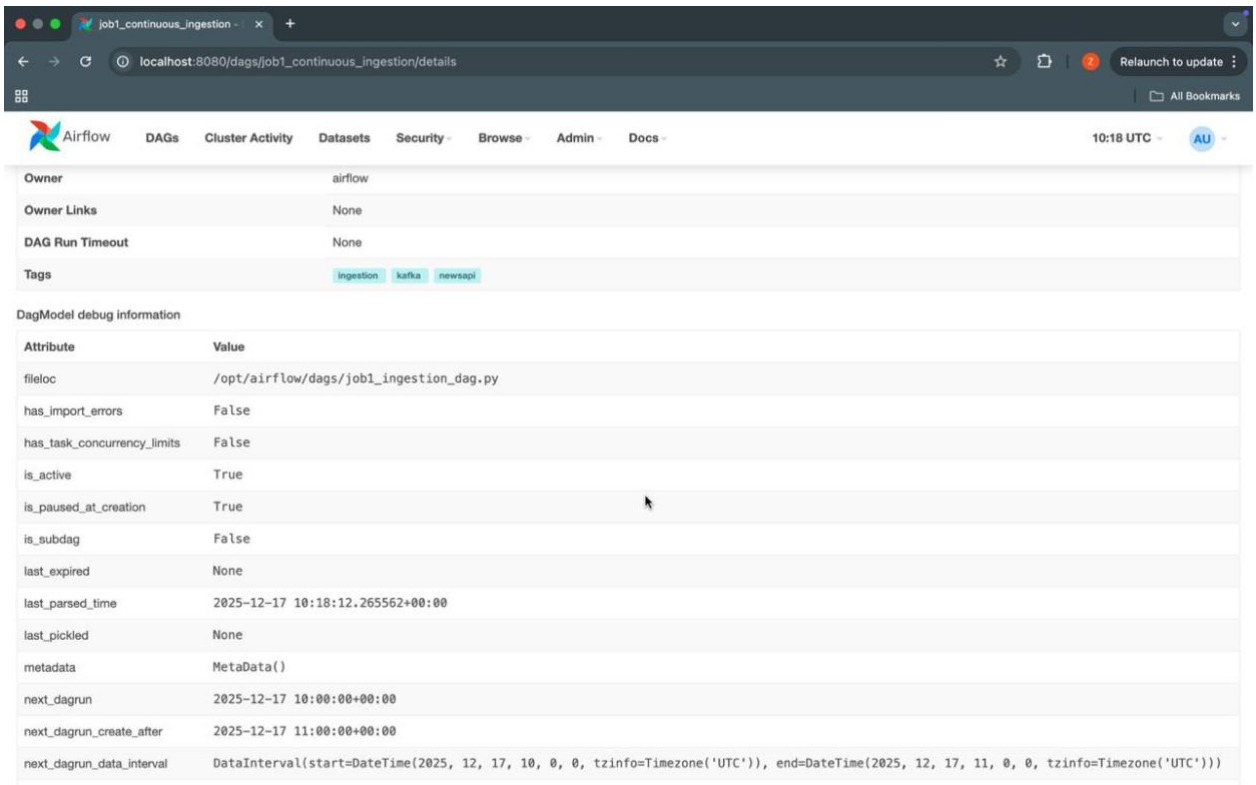


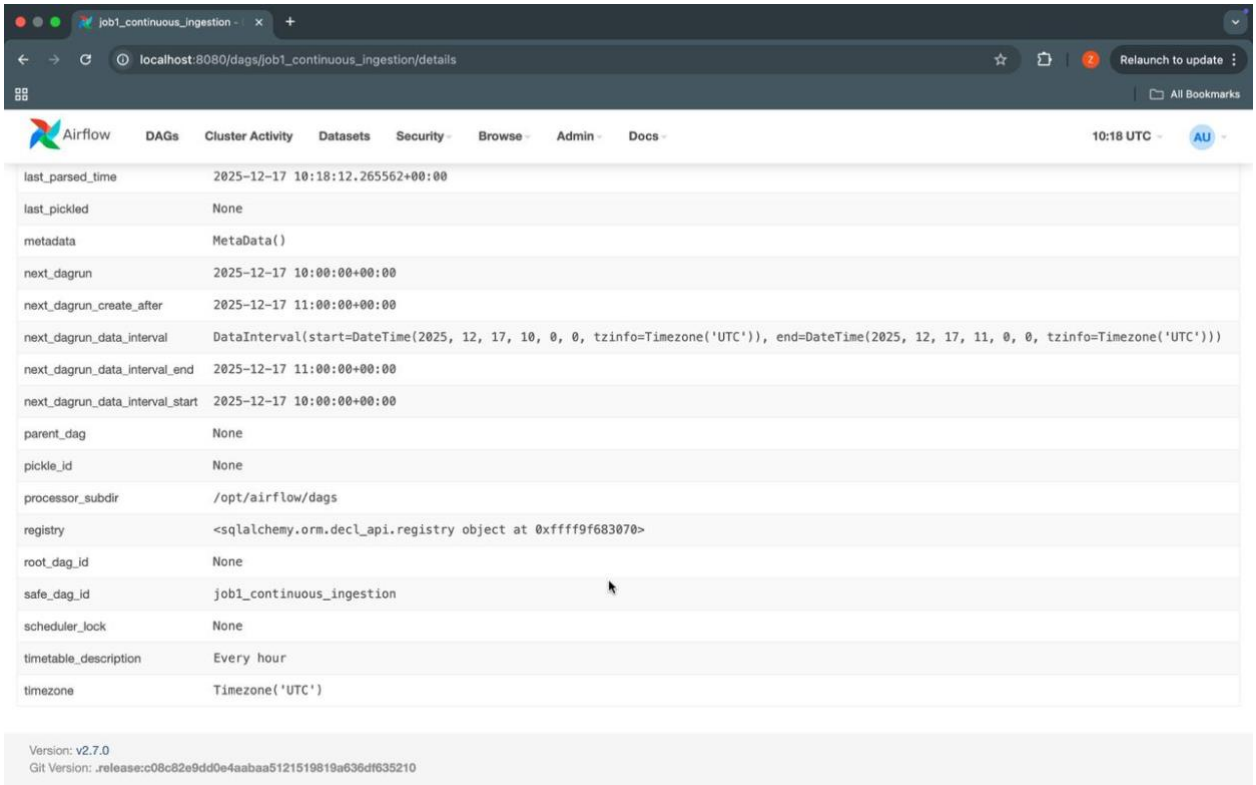

job1_continuous_ingestion
localhost:8080/dags/job1_continuous_ingestion/grid
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:17 UTC
AU
DAG: job1_continuous_ingestion
DAG 1: Continuous data ingestion from NewsAPI to Kafka
Schedule: @hourly
Next Run: 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
01:57:29
00:58:44
00:00:00
fetch_and_produce_to_kafka
DAG
job1_continuous_ingestion
Details
Graph
Gantt
Code
DAG Runs Summary
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

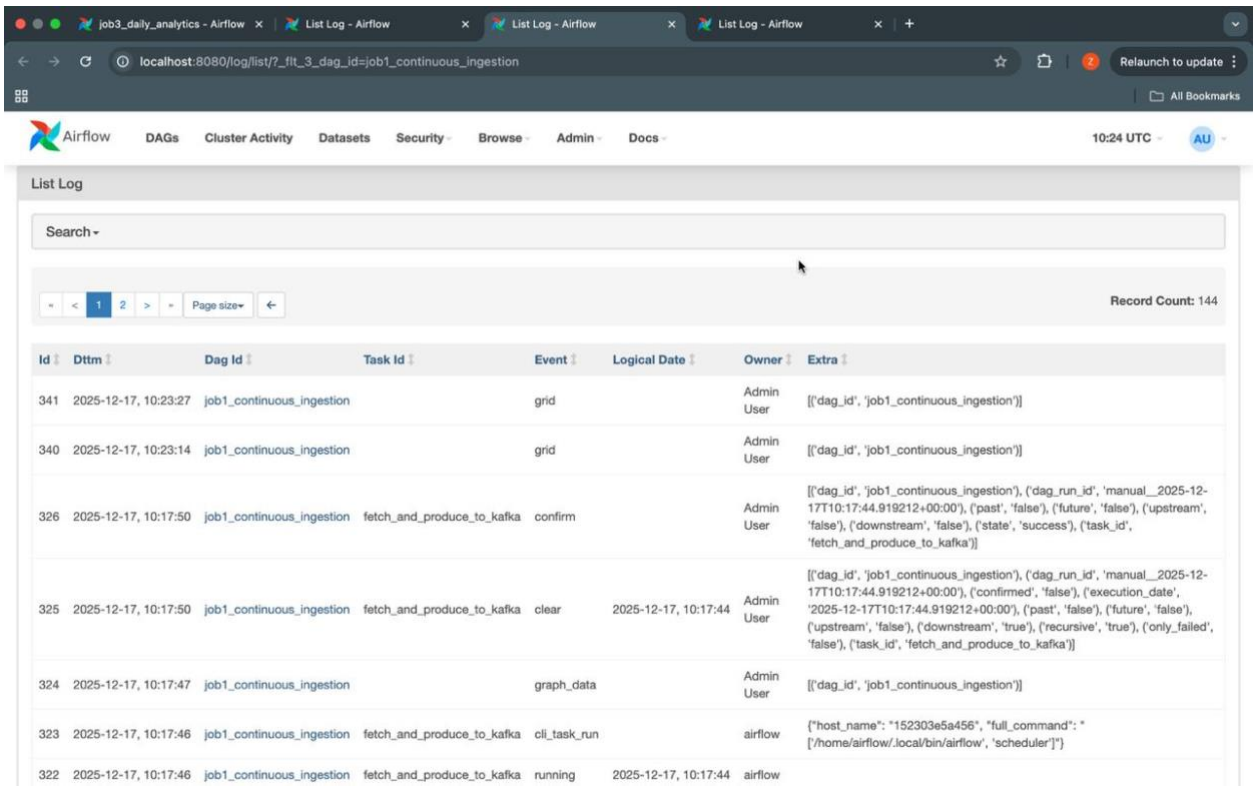

job1_continuous_ingestion
localhost:8080/dags/job1_continuous_ingestion/grid?tab=graph&dag_run_id=manual__2025-12-17T10%3A17%3A44.919212%2B00%3A00&tas...
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:18 UTC
AU
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
17.12.2025, 10:17:45
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
01:57:29
00:58:44
00:00:00
fetch_and_produce_to_kafka
DAG
job1_continuous_ingestion
Run
2025-12-17, 10:00:00 UTC
Task
fetch_and_produce_to_kafka
Clear task
Mark state as...
Filter Tasks
Details
Graph
Gantt
Code
Logs
Layout:
Left -> Right
fetch_and_produce_to_kafka
running
PythonOperator
React Flow


job1_continuous_ingestion
localhost:8080/dags/job1_continuous_ingestion/grid?tab=gantt&dag_run_id=manual__2025-12-17T10%3A17%3A44.919212%2B00%3A00&task...
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:18 UTC
AU
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
17.12.2025, 10:17:45
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
01:57:29
00:58:44
00:00:00
Dec 17, 00:00
Dec 17, 08:00
fetch_and_produce_to_kafka
DAG
job1_continuous_ingestion
Run
2025-12-17, 10:00:00 UTC
Task
fetch_and_produce_to_kafka
Clear task
Mark state as...
Filter Tasks
Details
Graph
Gantt
Code
Logs
10:17:44 UTC
10:17:48 UTC
10:17:51 UTC
10:17:54 UTC
10:17:57 UTC
10:18:00 UTC
10:18:03 UTC
10:18:06 UTC

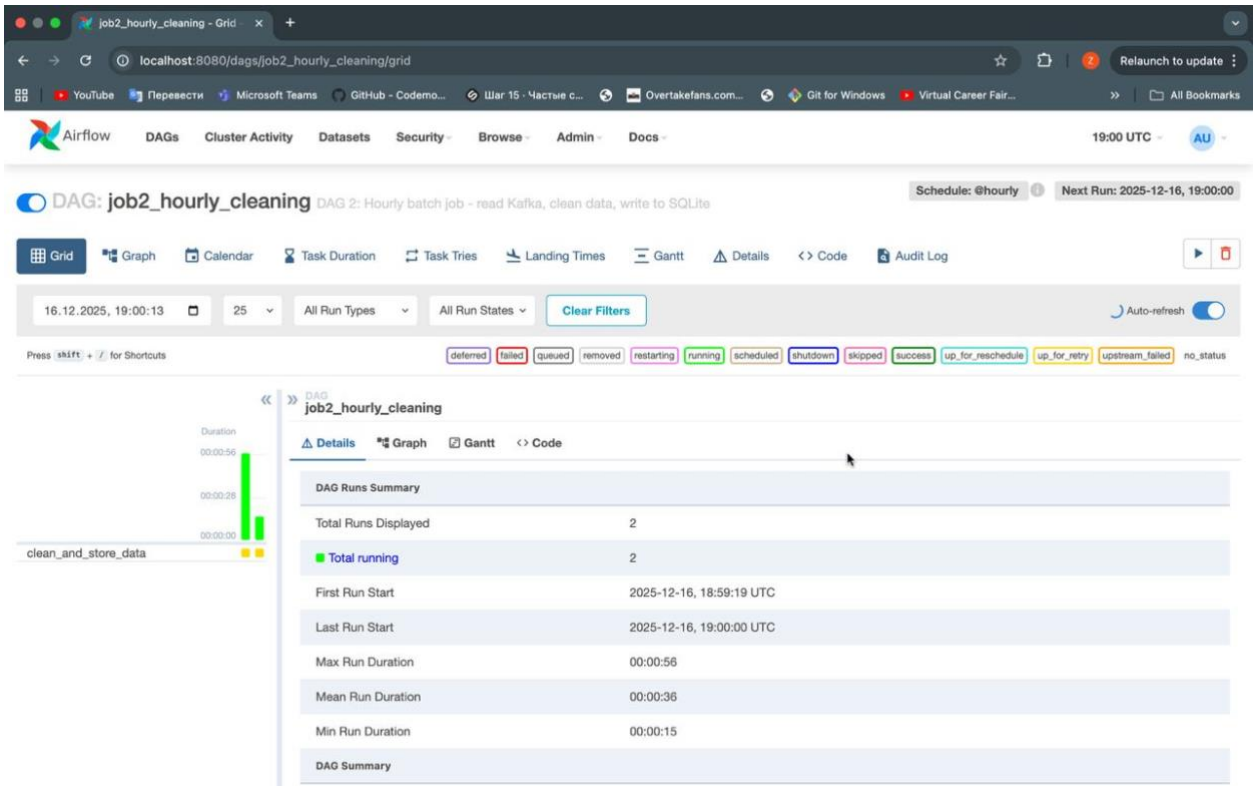

job1_continuous_ingestion
localhost:8080/dags/job1_continuous_ingestion/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:18 UTC
AU
DAG: job1_continuous_ingestion
DAG 1: Continuous data ingestion from NewsAPI to Kafka
Schedule: @hourly
Next Run: 2025-12-17, 10:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
DAG Details
failed 11
success 10
running 2
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, 0, tzinfo=Timezone('UTC'))}
Tasks Count
1
Task IDs
['fetch_and_produce_to_kafka']
Relative file location
job1_ingestion_dag.py

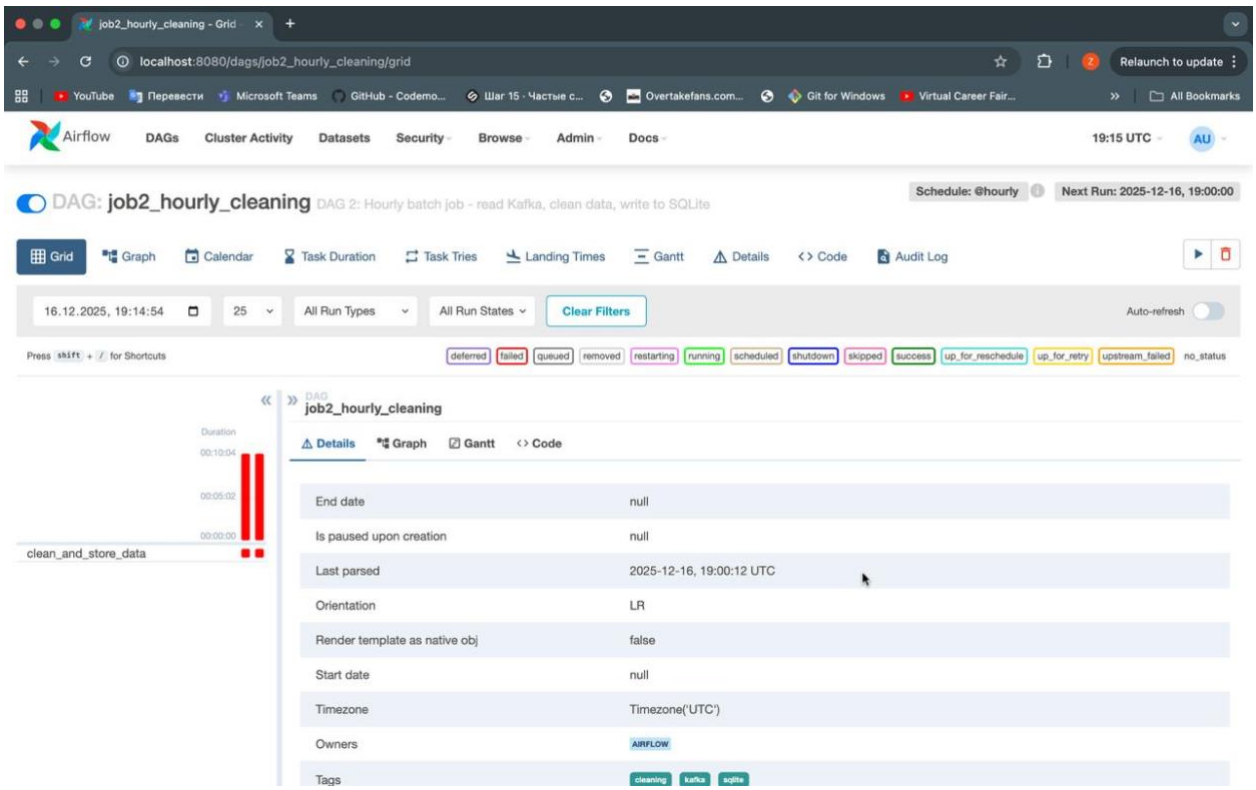

job1_continuous_ingestion
localhost:8080/dags/job1_continuous_ingestion/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:18 UTC
AU
Owner
airflow
Owner Links
None
DAG Run Timeout
None
Tags
ingestion
kafka
newsapi
DagModel debug information
Attribute
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')))

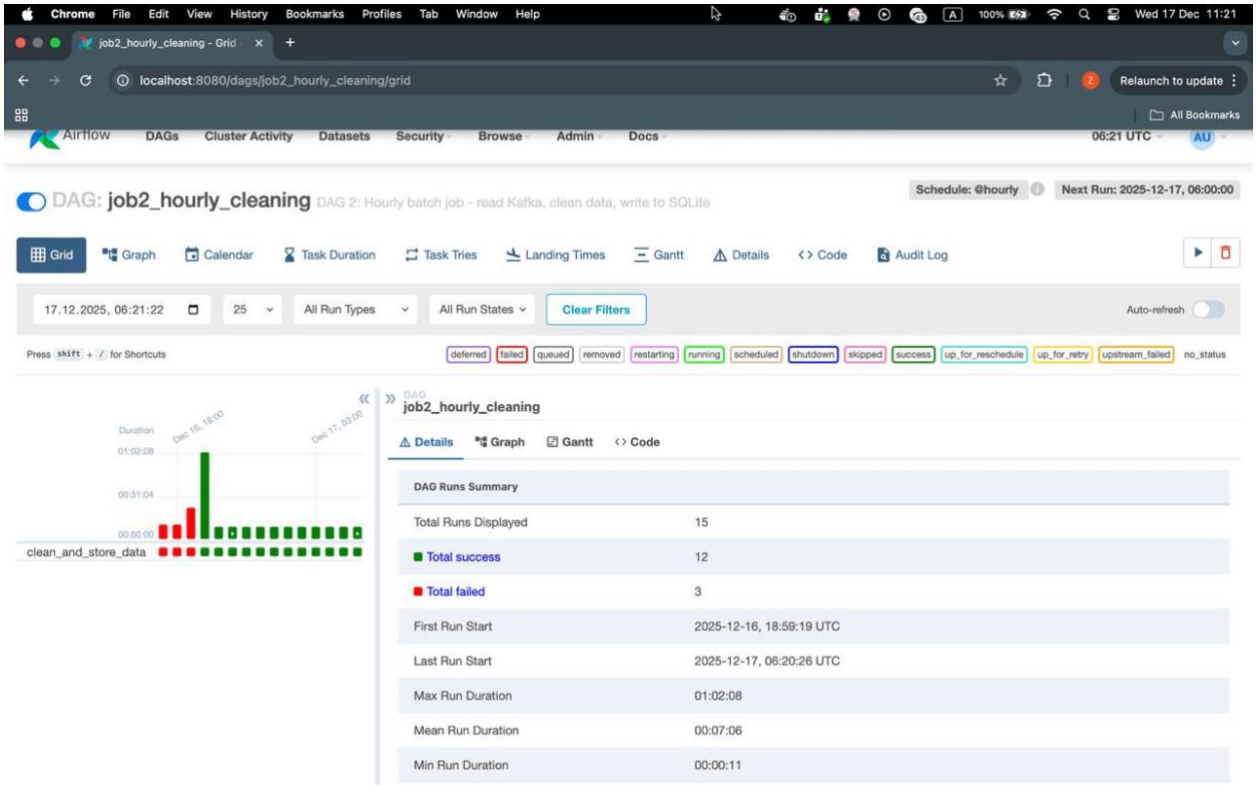

job1_continuous_ingestion
localhost:8080/dags/job1_continuous_ingestion/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:18 UTC
AU
last_parsed_time
2025-12-17 10:18:12.265562+00:00
last_pickled
None
metadata
MetaData()
next_dagrun
2025-12-17 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:c08c82e9dd0e4aabaa5121519819a636df635210

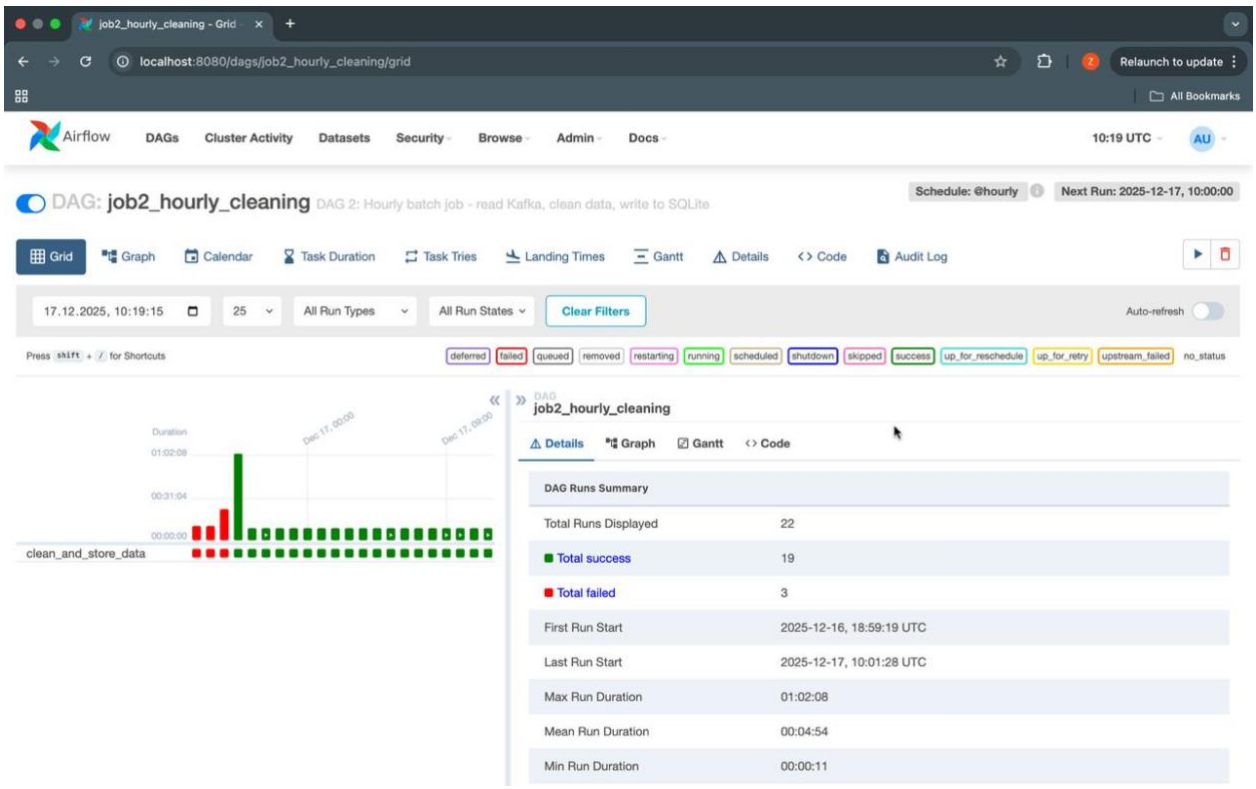


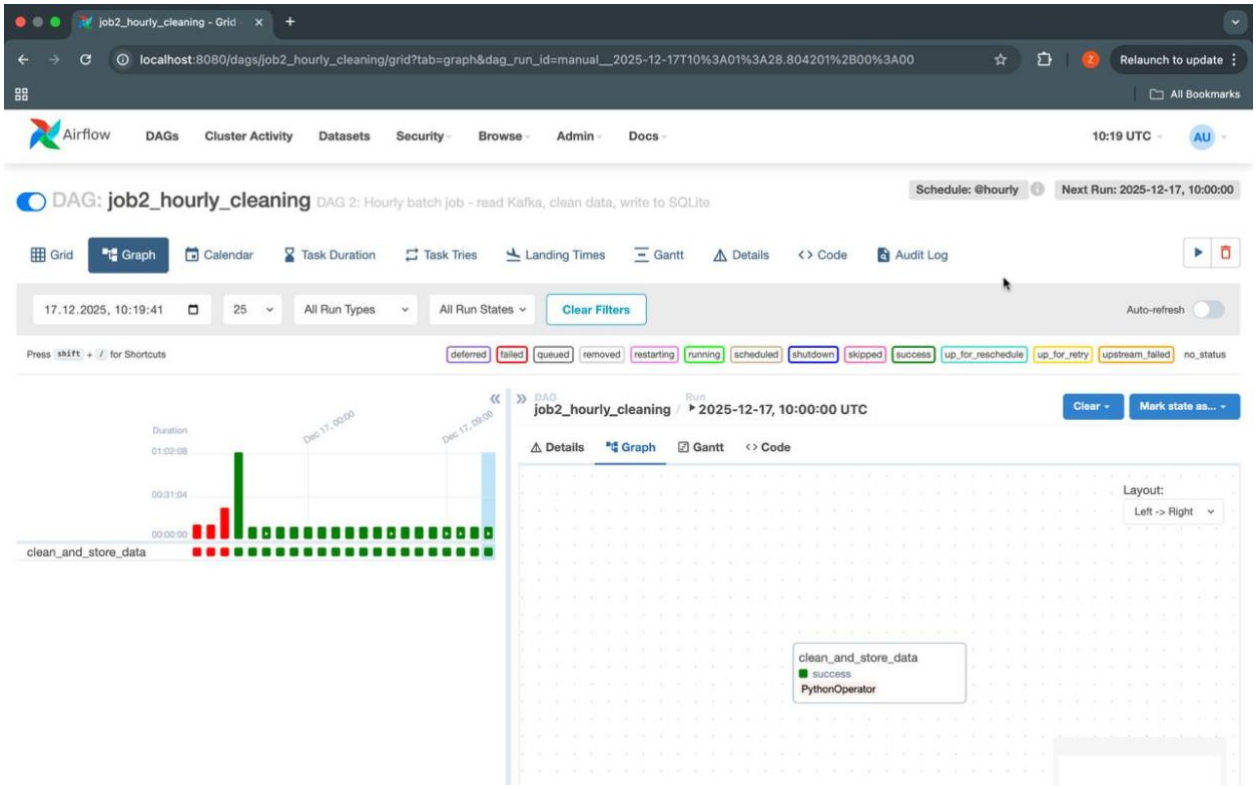
Job2_hourly_cleaning:

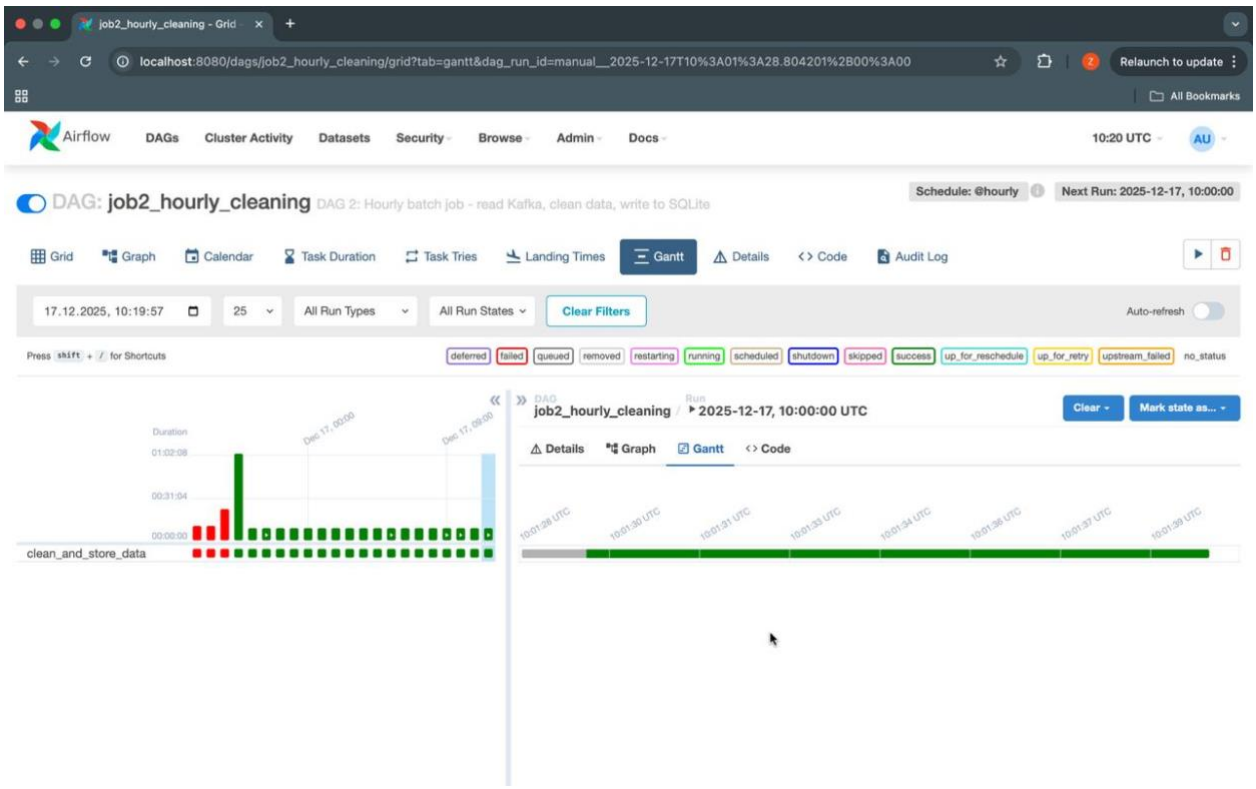


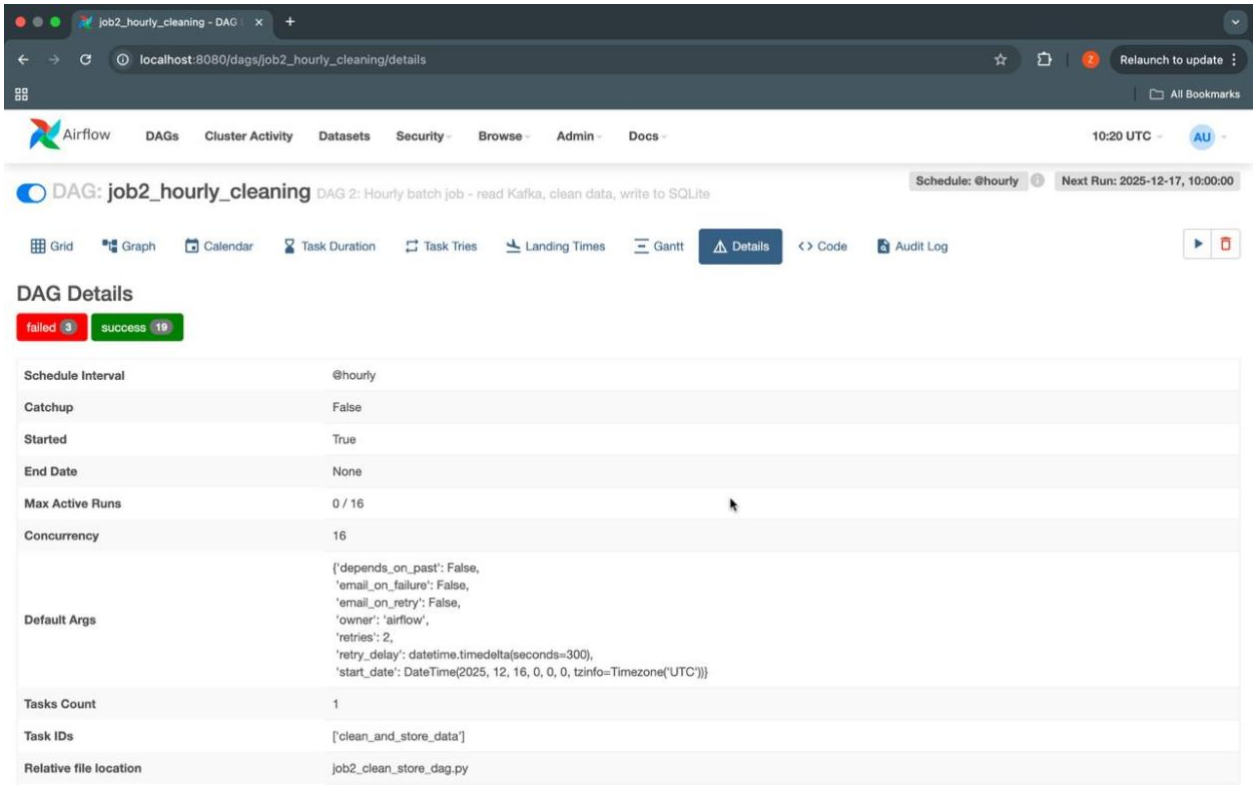

job2_hourly_cleaning - Grid
localhost:8080/dags/job2_hourly_cleaning/grid
Relaunch to update
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
19:15 UTC
DAG: job2_hourly_cleaning
DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite
Schedule: @hourly
Next Run: 2025-12-16, 19:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
16.12.2025, 19:14:54
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
00:10:04
00:05:02
00:00:00
clean_and_store_data
DAG
job2_hourly_cleaning
Details
Graph
Gantt
Code
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

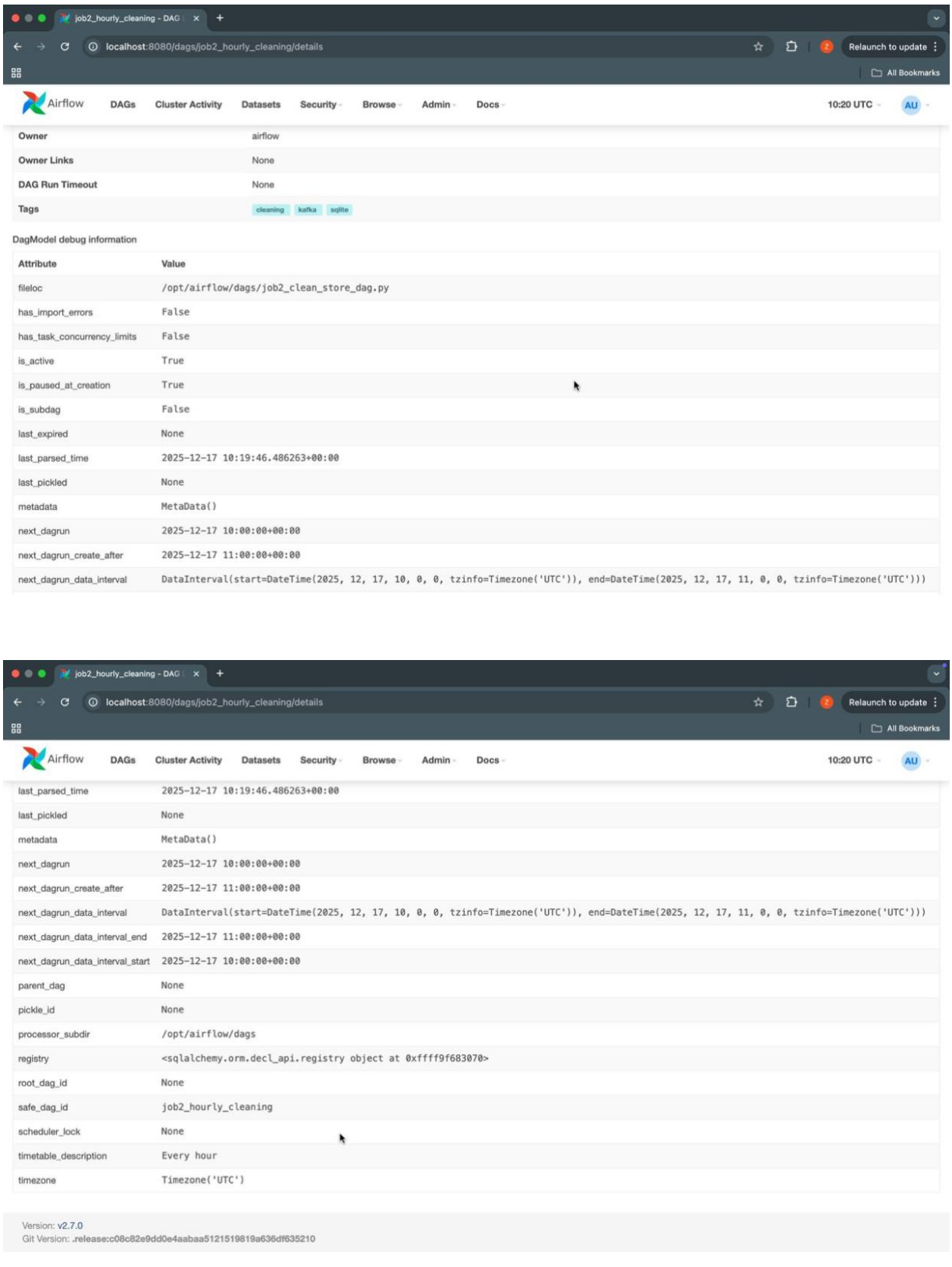

Chrome
File
Edit
View
History
Bookmarks
Profiles
Tab
Window
Help
Wed 17 Dec 11:21
job2_hourly_cleaning - Grid
localhost:8080/dags/job2_hourly_cleaning/grid
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
06:21 UTC
DAG: job2_hourly_cleaning
DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite
Schedule: @hourly
Next Run: 2025-12-17, 06:00:00
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
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
Dec 16, 18:00
Dec 17, 03:00
01:02:08
00:31:04
00:00:00
clean_and_store_data
DAG
job2_hourly_cleaning
Details
Graph
Gantt
Code
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


job2_hourly_cleaning - Grid
localhost:8080/dags/job2_hourly_cleaning/grid
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:19 UTC
AU
DAG: job2_hourly_cleaning
DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite
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, 10:19:15
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
01:02:08
00:31:04
00:00:00
clean_and_store_data
DAG
job2_hourly_cleaning
Details
Graph
Gantt
Code
DAG Runs Summary
Total Runs Displayed 22
Total success 19
Total failed 3
First Run Start 2025-12-16, 18:59:19 UTC
Last Run Start 2025-12-17, 10:01:28 UTC
Max Run Duration 01:02:08
Mean Run Duration 00:04:54
Min Run Duration 00:00:11

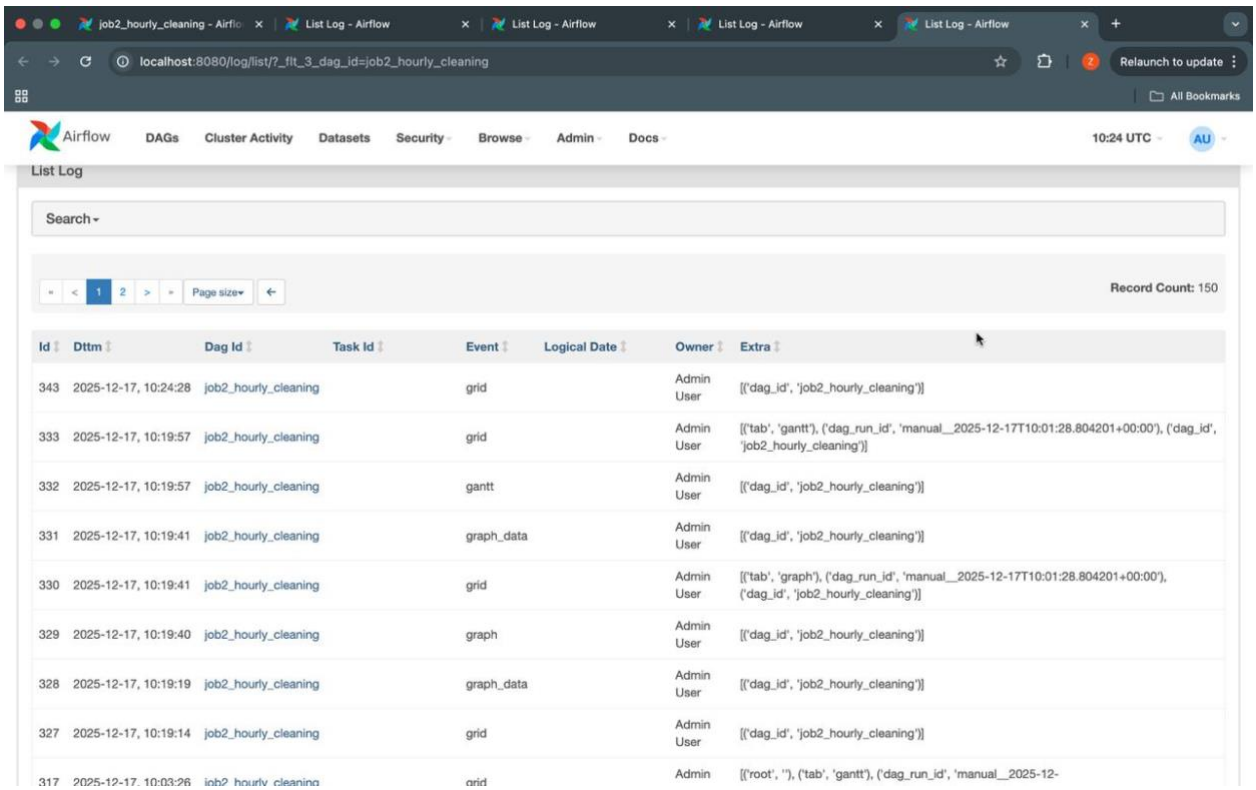

job2_hourly_cleaning - Grid
localhost:8080/dags/job2_hourly_cleaning/grid?tab=graph&dag_run_id=manual__2025-12-17T10%3A01%3A28.804201%2B00%3A00
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:19 UTC
AU
DAG: job2_hourly_cleaning
DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite
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, 10:19:41
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
01:02:08
00:31:04
00:00:00
clean_and_store_data
DAG
job2_hourly_cleaning
Run
2025-12-17, 10:00:00 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Layout:
Left -> Right
clean_and_store_data
success
PythonOperator


job2_hourly_cleaning - Grid
localhost:8080/dags/job2_hourly_cleaning/grid?tab=gantt&dag_run_id=manual__2025-12-17T10%3A01%3A28.804201%2B00%3A00
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:20 UTC
AU
DAG: job2_hourly_cleaning
DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite
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, 10:19:57
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
01:02:08
00:31:04
00:00:00
Dec 17, 09:00
clean_and_store_data
DAG
job2_hourly_cleaning
Run
2025-12-17, 10:00:00 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
10:01:28 UTC
10:01:30 UTC
10:01:31 UTC
10:01:33 UTC
10:01:34 UTC
10:01:36 UTC
10:01:37 UTC
10:01:39 UTC

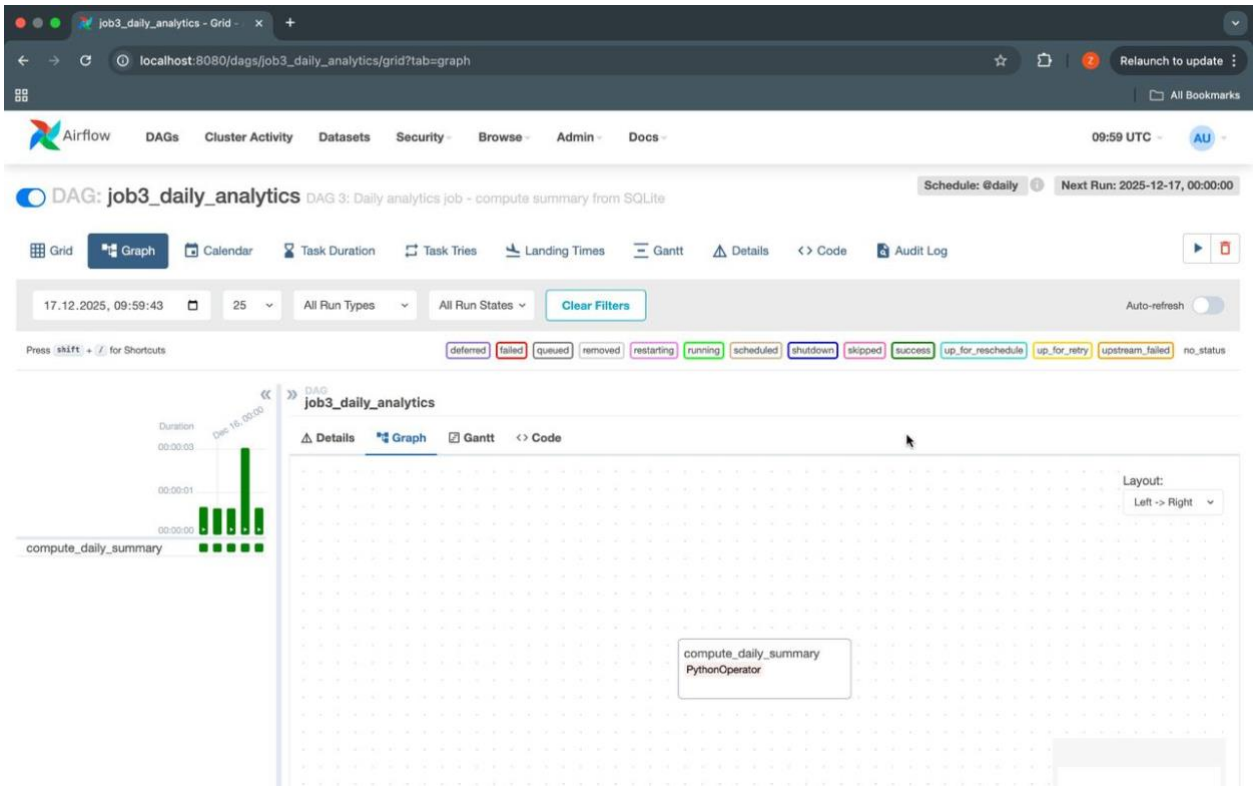

job2_hourly_cleaning - DAG
localhost:8080/dags/job2_hourly_cleaning/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:20 UTC
AU
DAG: job2_hourly_cleaning
DAG 2: Hourly batch job - read Kafka, clean data, write to SQLite
Schedule: @hourly
Next Run: 2025-12-17, 10:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
DAG Details
failed 3
success 19
Schedule Interval
@hourly
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
['clean_and_store_data']
Relative file location
job2_clean_store_dag.py

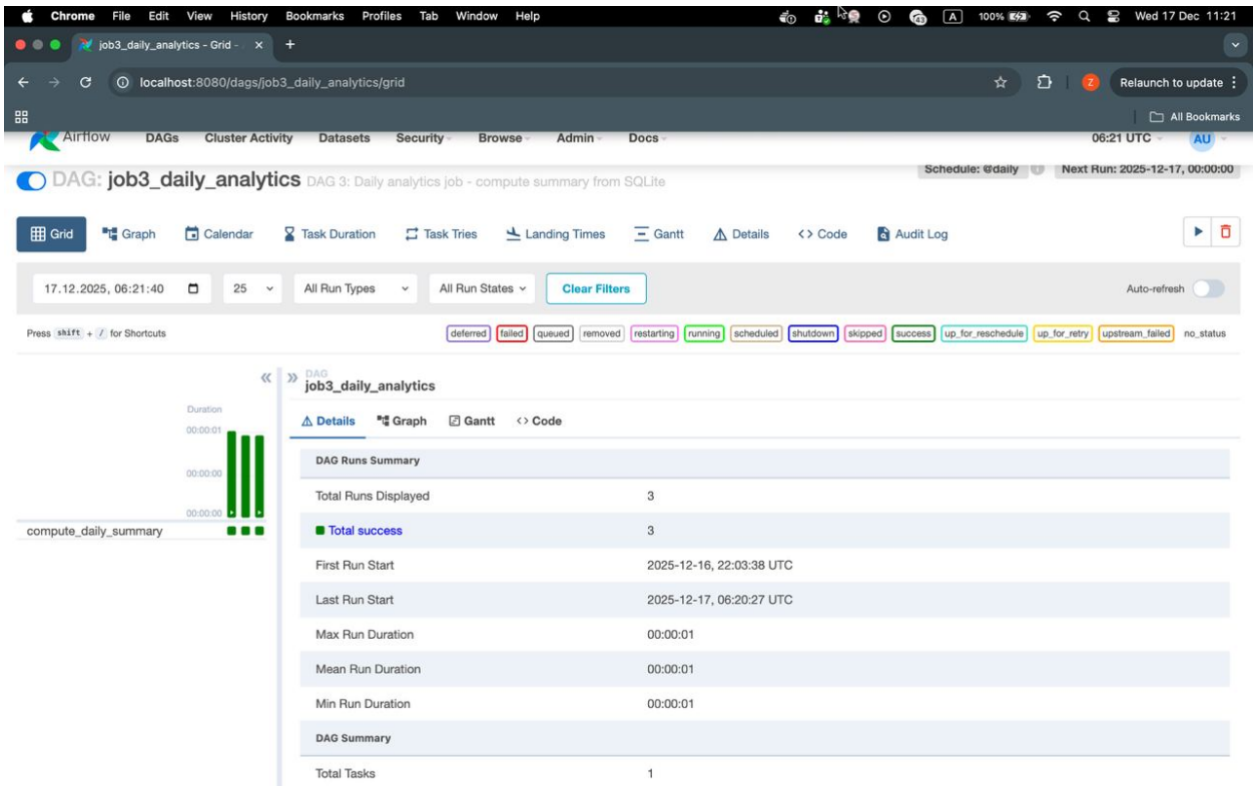

job2_hourly_cleaning - DAG
localhost:8080/dags/job2_hourly_cleaning/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:20 UTC
AU
Owner
airflow
Owner Links
None
DAG Run Timeout
None
Tags
cleaning
kafka
sqlite
DagModel debug information
Attribute
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')))

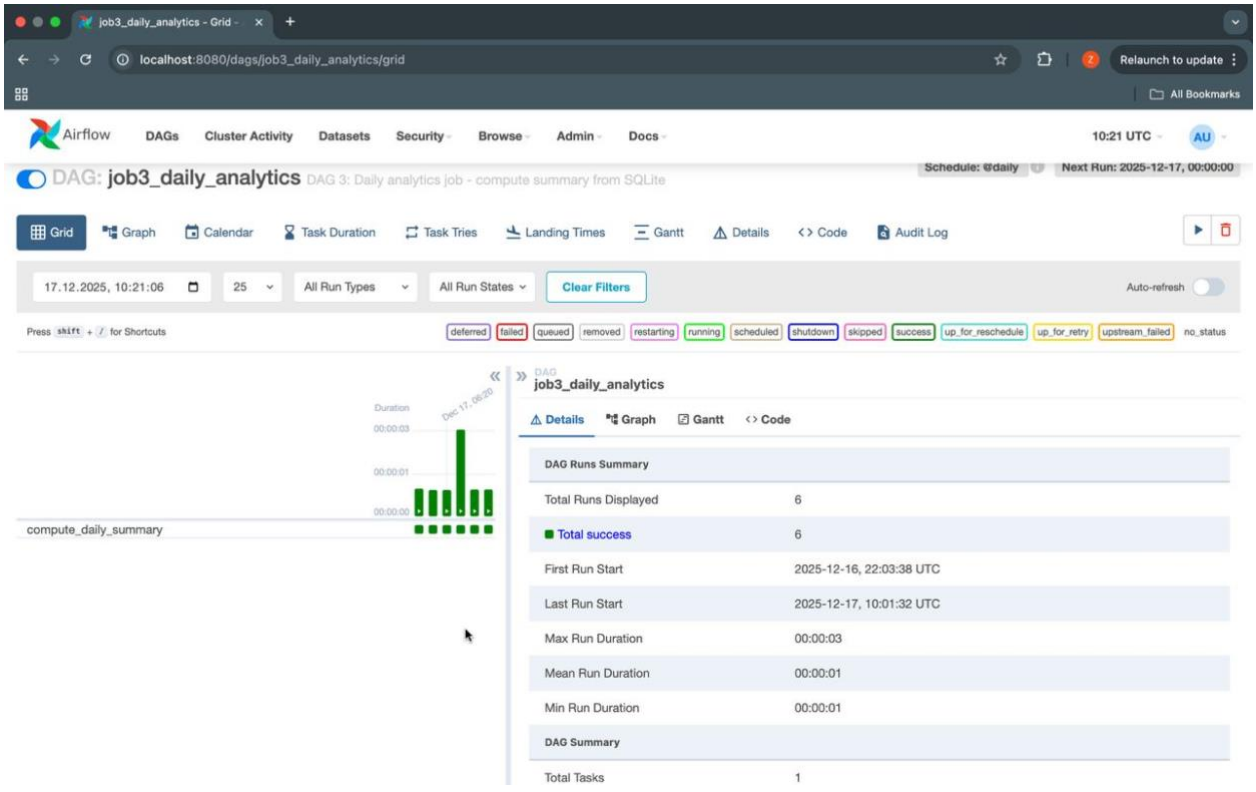
job2_hourly_cleaning - DAG
localhost:8080/dags/job2_hourly_cleaning/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:20 UTC
AU
last_parsed_time
2025-12-17 10:19:46.486263+00:00
last_pickled
None
metadata
MetaData()
next_dagrun
2025-12-17 10:00:00+00:00
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 0xffff9f683070>
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

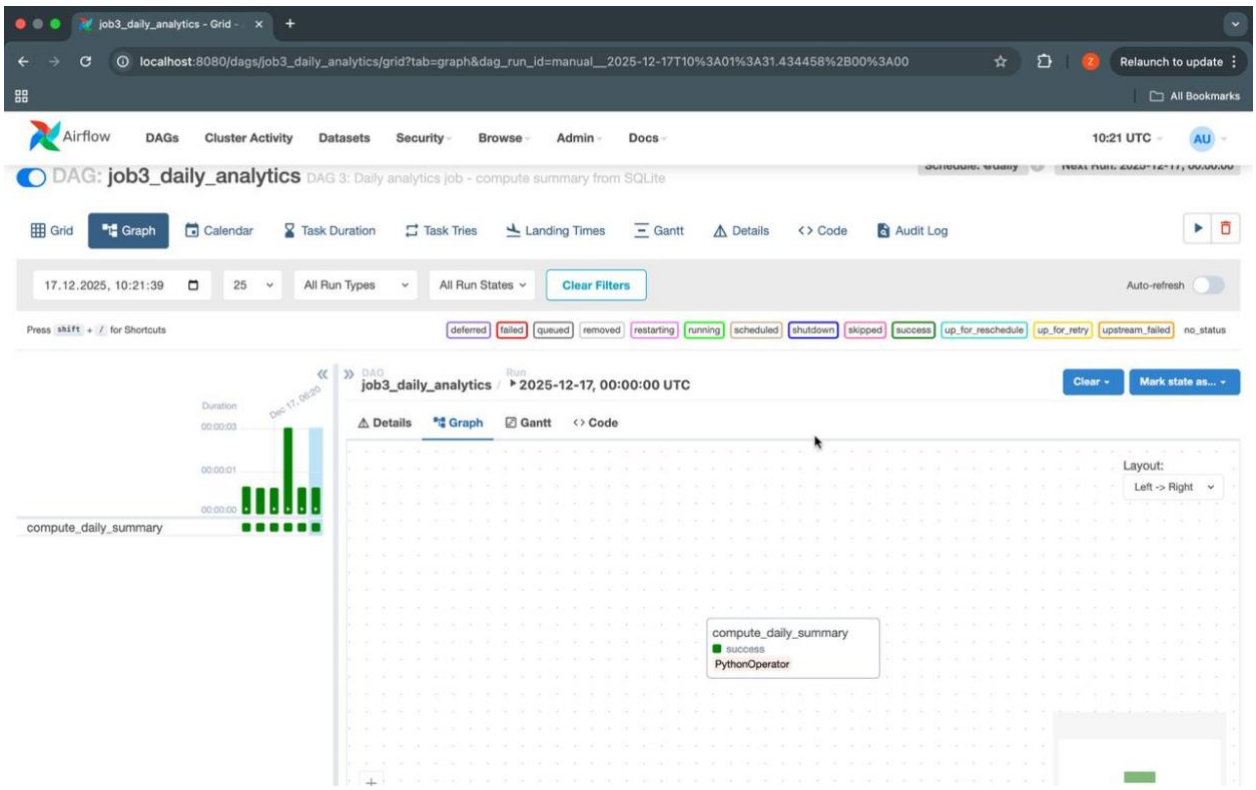


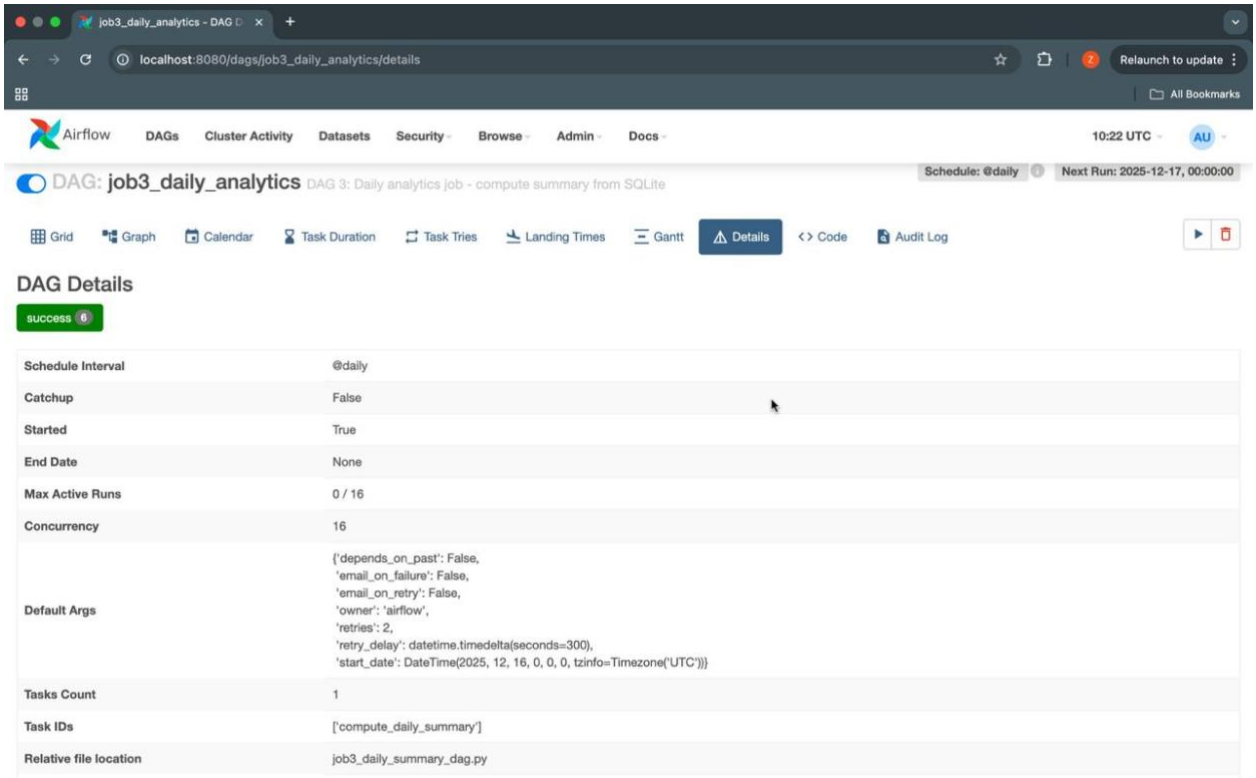
Job3_daily_analytics:

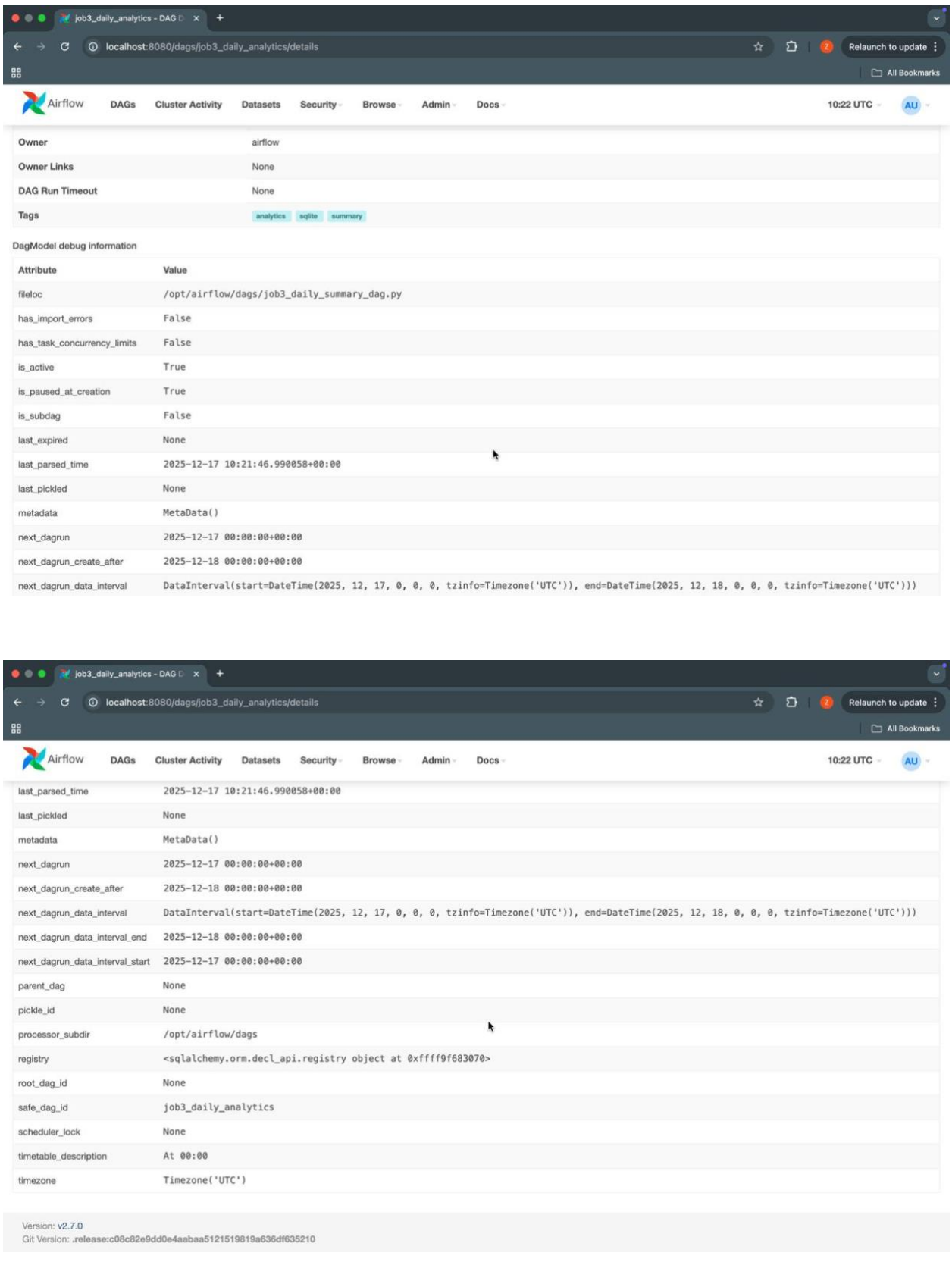


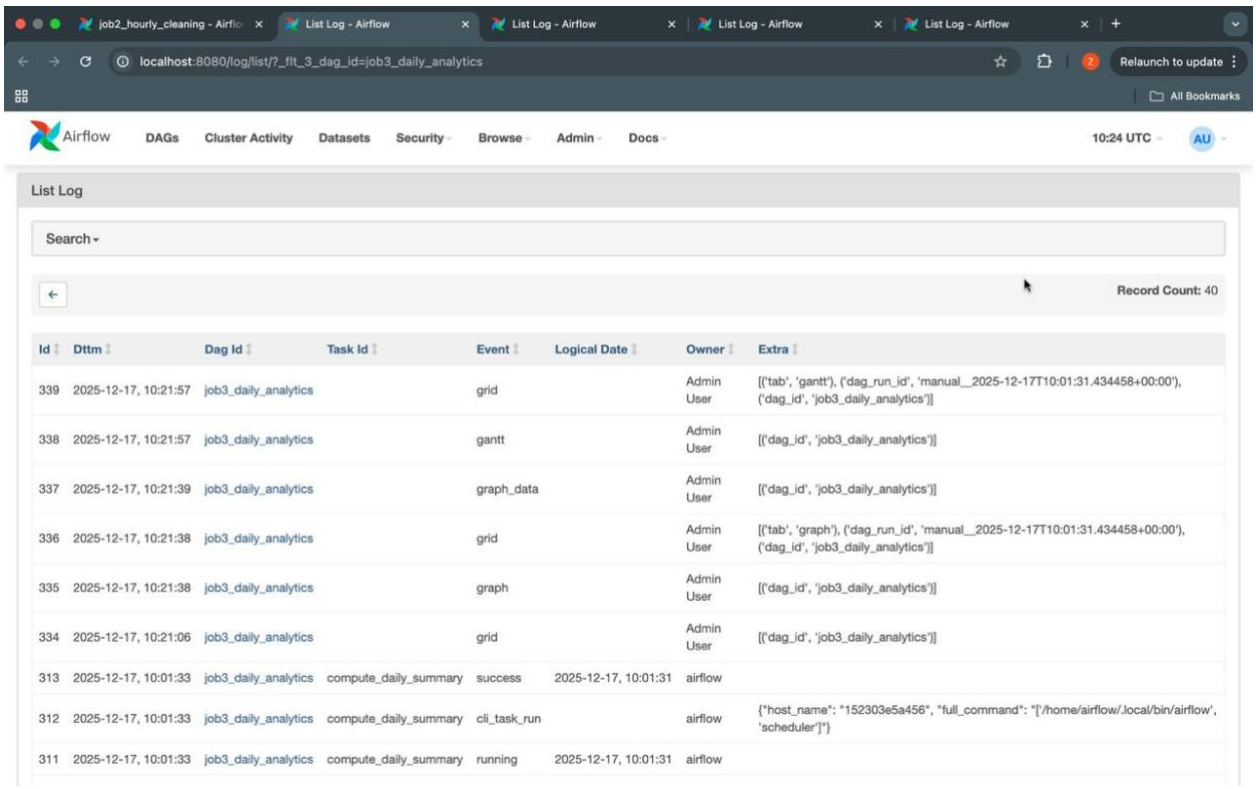

Chrome File Edit View History Bookmarks Profiles Tab Window Help
Wed 17 Dec 11:21
job3_daily_analytics - Grid -
localhost:8080/dags/job3_daily_analytics/grid
Relaunch to update
All Bookmarks
Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs
06:21 UTC
AU
DAG: job3_daily_analytics
DAG 3: 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
17.12.2025, 06:21:40
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
00:00:01
00:00:00
00:00:00
compute_daily_summary
DAG
job3_daily_analytics
Details Graph Gantt Code
DAG Runs Summary
Total Runs Displayed 3
Total success 3
First Run Start 2025-12-16, 22:03:38 UTC
Last Run Start 2025-12-17, 06:20:27 UTC
Max Run Duration 00:00:01
Mean Run Duration 00:00:01
Min Run Duration 00:00:01
DAG Summary
Total Tasks 1


job3_daily_analytics - Grid -
localhost:8080/dags/job3_daily_analytics/grid
Relaunch to update
All Bookmarks
Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs
10:21 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
17.12.2025, 10:21:06
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
Dec 17, 06:20
00:00:03
00:00:01
00:00:00
compute_daily_summary
DAG
job3_daily_analytics
Details Graph Gantt Code
DAG Runs Summary
Total Runs Displayed 6
Total success 6
First Run Start 2025-12-16, 22:03:38 UTC
Last Run Start 2025-12-17, 10:01:32 UTC
Max Run Duration 00:00:03
Mean Run Duration 00:00:01
Min Run Duration 00:00:01
DAG Summary
Total Tasks 1


job3_daily_analytics - Grid
localhost:8080/dags/job3_daily_analytics/grid?tab=graph&dag_run_id=manual__2025-12-17T10%3A01%3A31.434458%2B00%3A00
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:21 UTC
AU
DAG: job3_daily_analytics
DAG 3: Daily analytics job - compute summary from SQLite
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
17.12.2025, 10:21:39
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
DAG
job3_daily_analytics
Run
2025-12-17, 00:00:00 UTC
Clear
Mark state as...
Duration
Dec 17, 06:20
00:00:03
00:00:01
00:00:00
Details
Graph
Gantt
Code
Layout:
Left -> Right
compute_daily_summary
compute_daily_summary
success
PythonOperator

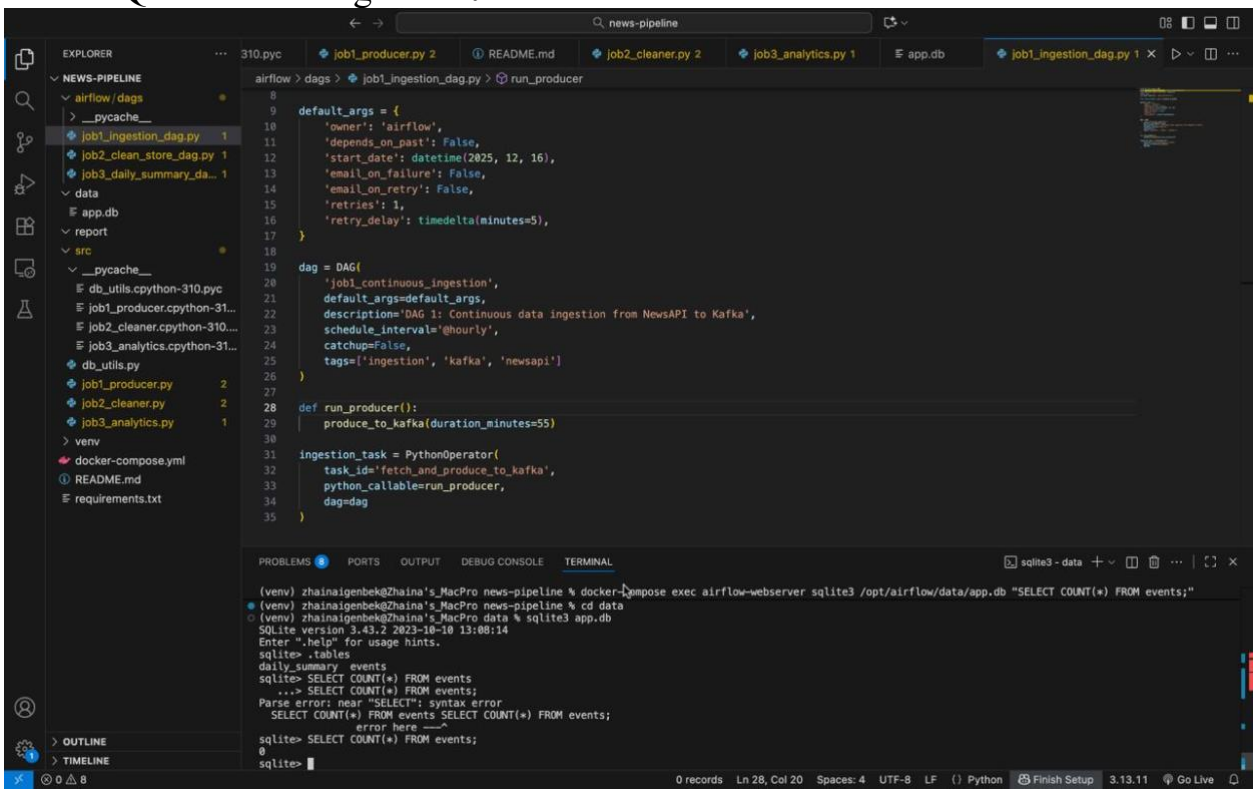

job3_daily_analytics - DAG D
localhost:8080/dags/job3_daily_analytics/details
Relaunch to update
All Bookmarks
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:22 UTC
AU
DAG: job3_daily_analytics
DAG 3: Daily analytics job - compute summary from SQLite
Schedule: @daily
Next Run: 2025-12-17, 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
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
10:22 UTC
AU
Owner
airflow
Owner Links
None
DAG Run Timeout
None
Tags
analytics
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
Airflow
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 0xffff9f683070>
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

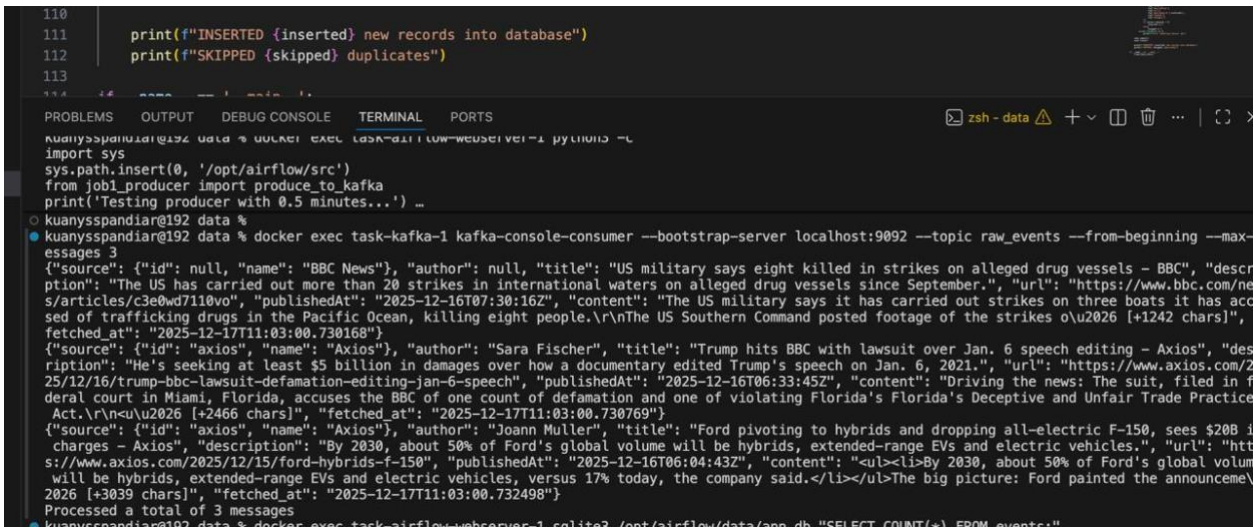


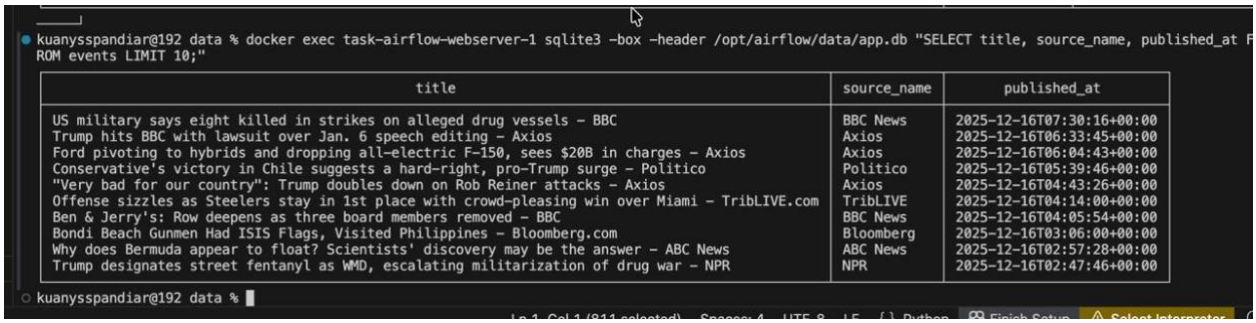
Id	Dttm	Dag Id	Task Id	Event	Logical Date	Owner	Extra
339	2025-12-17, 10:21:57	job3_daily_analytics		grid		Admin User	[('tab', 'gantt'), ('dag_run_id', 'manual__2025-12-17T10:01:31.434458+00:00'), ('dag_id', 'job3_daily_analytics')]
338	2025-12-17, 10:21:57	job3_daily_analytics		gantt		Admin User	[('dag_id', '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:

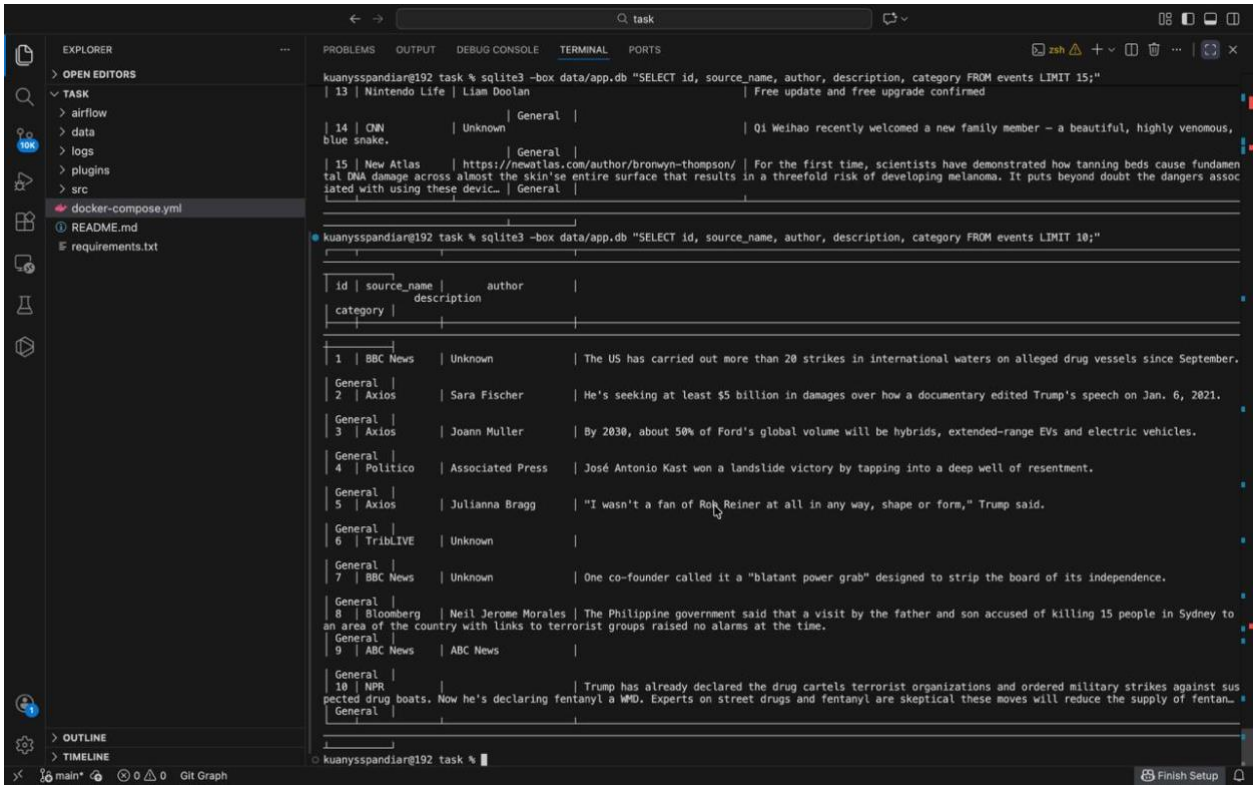


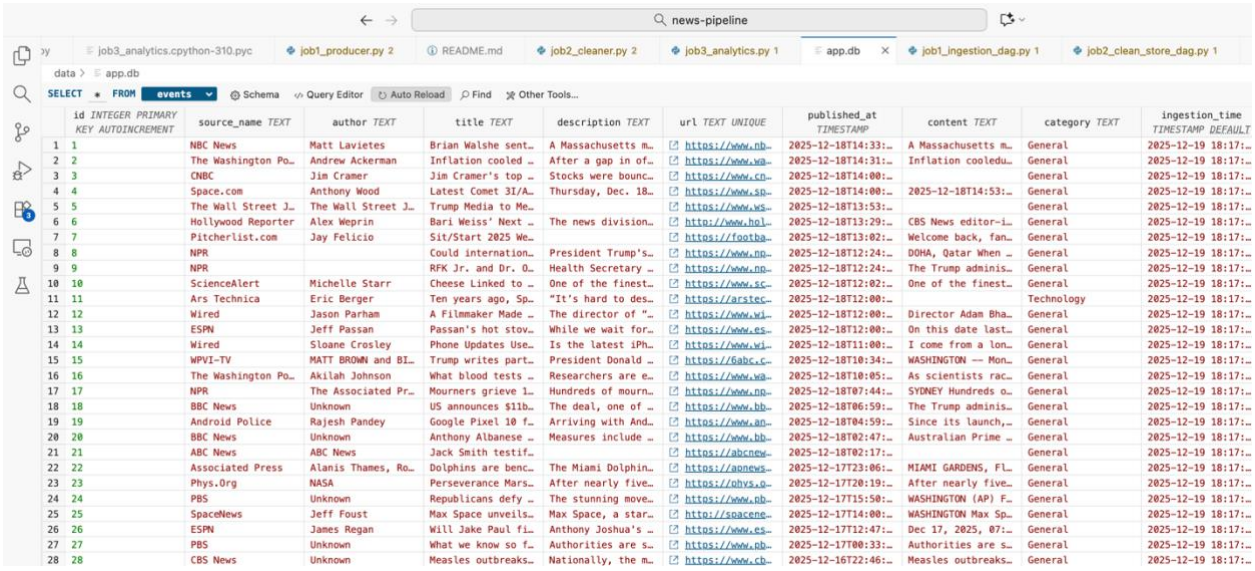
After fixing errors:

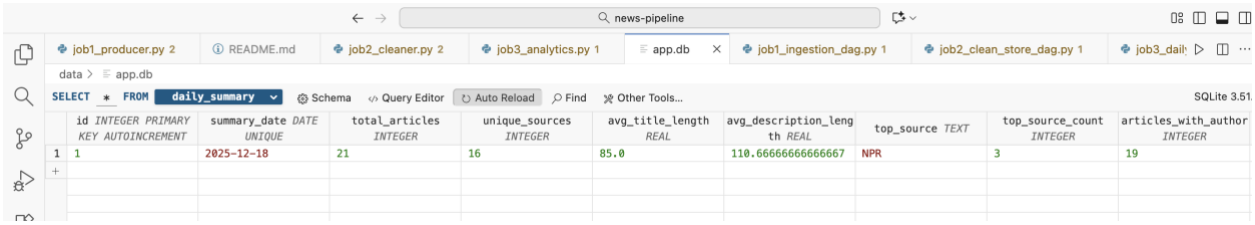


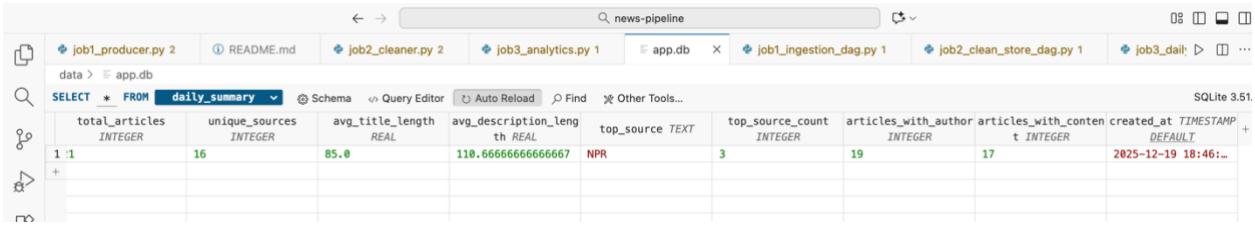


title	source_name	published_at
US military says eight killed in strikes on alleged drug vessels - BBC	BBC News	2025-12-16T07:30:16+00:00
Trump hits BBC with lawsuit over Jan. 6 speech editing - Axios	Axios	2025-12-16T06:33:45+00:00
Ford pivoting to hybrids and dropping all-electric F-150, sees $20B in 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

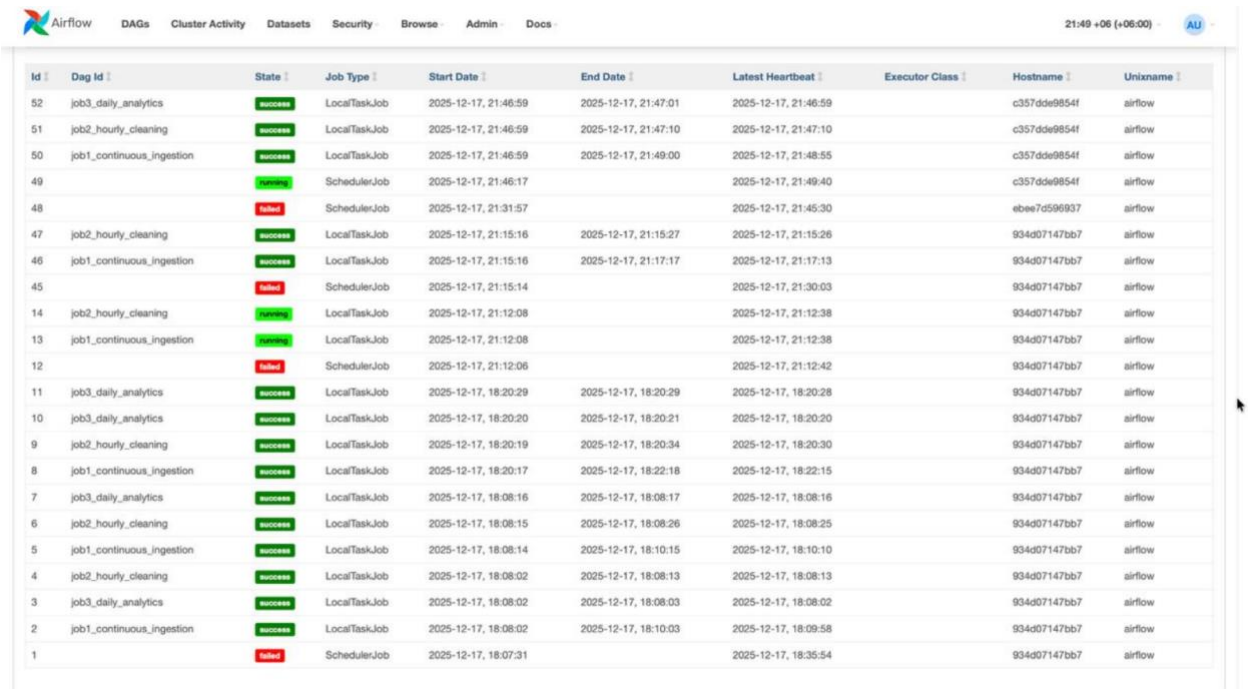



news-pipeline
job3_analytics.cpython-310.pyc
job1_producer.py 2
README.md
job2_cleaner.py 2
job3_analytics.py 1
app.db
job1_ingestion_dag.py 1
job2_clean_store_dag.py 1
data > app.db
SELECT * FROM events
Schema
Query Editor
Auto Reload
Find
Other Tools...
id INTEGER PRIMARY KEY AUTOINCREMENT
source_name TEXT
author TEXT
title TEXT
description TEXT
url TEXT UNIQUE
published_at TIMESTAMP
content TEXT
category TEXT
ingestion_time TIMESTAMP DEFAULT
1 1 NBC News Matt Lavietes Brian Walshe sent… A Massachusetts m… https://www.nb… 2025-12-18T14:33:… A Massachusetts m… General 2025-12-19 18:17:…
2 2 The Washington Po… Andrew Ackerman Inflation cooled … After a gap in of… https://www.wa… 2025-12-18T14:31:… Inflation cooledu… General 2025-12-19 18:17:…
3 3 CNBC Jim Cramer Jim Cramer's top … Stocks were bounc… https://www.cn… 2025-12-18T14:00:… General 2025-12-19 18:17:…
4 4 Space.com Anthony Wood Latest Comet 3I/A… Thursday, Dec. 18… https://www.sp… 2025-12-18T14:00:… 2025-12-18T14:53:… General 2025-12-19 18:17:…
5 5 The Wall Street J… The Wall Street J… Trump Media to Me… https://www.ws… 2025-12-18T13:53:… General 2025-12-19 18:17:…
6 6 Hollywood Reporter Alex Weprin Bari Weiss' Next … The news division… http://www.hol… 2025-12-18T13:29:… CBS News editor-i… General 2025-12-19 18:17:…
7 7 Pitcherlist.com Jay Felicio Sit/Start 2025 We… https://footba… 2025-12-18T13:02:… Welcome back, fan… General 2025-12-19 18:17:…
8 8 NPR Could internation… President Trump's… https://www.np… 2025-12-18T12:24:… DOHA, Qatar When … General 2025-12-19 18:17:…
9 9 NPR RFK Jr. and Dr. O… Health Secretary … https://www.np… 2025-12-18T12:24:… The Trump adminis… General 2025-12-19 18:17:…
10 10 ScienceAlert Michelle Starr Cheese Linked to … One of the finest… https://www.sc… 2025-12-18T12:02:… One of the finest… General 2025-12-19 18:17:…
11 11 Ars Technica Eric Berger Ten years ago, Sp… "It's hard to des… https://arstec… 2025-12-18T12:00:… Technology 2025-12-19 18:17:…
12 12 Wired Jason Parham A Filmmaker Made … The director of "… https://www.wi… 2025-12-18T12:00:… Director Adam Bha… General 2025-12-19 18:17:…
13 13 ESPN Jeff Passan Passan's hot stov… While we wait for… https://www.es… 2025-12-18T12:00:… On this date last… General 2025-12-19 18:17:…
14 14 Wired Sloane Crosley Phone Updates Use… Is the latest iPh… https://www.wi… 2025-12-18T11:00:… I come from a lon… General 2025-12-19 18:17:…
15 15 WPVI-TV MATT BROWN and BI… Trump writes part… President Donald … https://6abc.c… 2025-12-18T10:34:… WASHINGTON — Mon… General 2025-12-19 18:17:…
16 16 The Washington Po… Akilah Johnson What blood tests … Researchers are e… https://www.wa… 2025-12-18T10:05:… As scientists rac… General 2025-12-19 18:17:…
17 17 NPR The Associated Pr… Mourners grieve 1… Hundreds of mourn… https://www.np… 2025-12-18T07:44:… SYDNEY Hundreds o… General 2025-12-19 18:17:…
18 18 BBC News Unknown US announces $11b… The deal, one of … https://www.bb… 2025-12-18T06:59:… The Trump adminis… General 2025-12-19 18:17:…
19 19 Android Police Rajesh Pandey Google Pixel 10 f… Arriving with And… https://www.an… 2025-12-18T04:59:… Since its launch,… General 2025-12-19 18:17:…
20 20 BBC News Unknown Anthony Albanese … Measures include … https://www.bb… 2025-12-18T02:47:… Australian Prime … General 2025-12-19 18:17:…
21 21 ABC News ABC News Jack Smith testif… https://abcnew… 2025-12-18T02:17:… General 2025-12-19 18:17:…
22 22 Associated Press Alanis Thames, Ro… Dolphins are benc… The Miami Dolphin… https://apnews… 2025-12-17T23:06:… MIAMI GARDENS, Fl… General 2025-12-19 18:17:…
23 23 Phys.Org NASA Perseverance Mars… After nearly five… https://phys.o… 2025-12-17T20:19:… After nearly five… General 2025-12-19 18:17:…
24 24 PBS Unknown Republicans defy … The stunning move… https://www.pb… 2025-12-17T15:50:… WASHINGTON (AP) F… General 2025-12-19 18:17:…
25 25 SpaceNews Jeff Foust Max Space unveils… Max Space, a star… http://spacene… 2025-12-17T14:00:… WASHINGTON Max Sp… General 2025-12-19 18:17:…
26 26 ESPN James Regan Will Jake Paul fi… Anthony Joshua's … https://www.es… 2025-12-17T12:47:… Dec 17, 2025, 07:… General 2025-12-19 18:17:…
27 27 PBS Unknown What we know so f… Authorities are s… https://www.pb… 2025-12-17T00:33:… Authorities are s… General 2025-12-19 18:17:…
28 28 CBS News Unknown Measles outbreaks… Nationally, the m… https://www.cb… 2025-12-16T22:46:… Measles outbreaks… General 2025-12-19 18:17:…


news-pipeline
job1_producer.py 2
README.md
job2_cleaner.py 2
job3_analytics.py 1
app.db
job1_ingestion_dag.py 1
job2_clean_store_dag.py 1
data > app.db
SELECT * FROM daily_summary
Schema
Query Editor
Auto Reload
Find
Other Tools...
SQLite 3.51.
id INTEGER PRIMARY KEY AUTOINCREMENT
summary_date DATE UNIQUE
total_articles INTEGER
unique_sources INTEGER
avg_title_length REAL
avg_description_length REAL
top_source TEXT
top_source_count INTEGER
articles_with_author INTEGER
1 1 2025-12-18 21 16 85.0 110.66666666666667 NPR 3 19

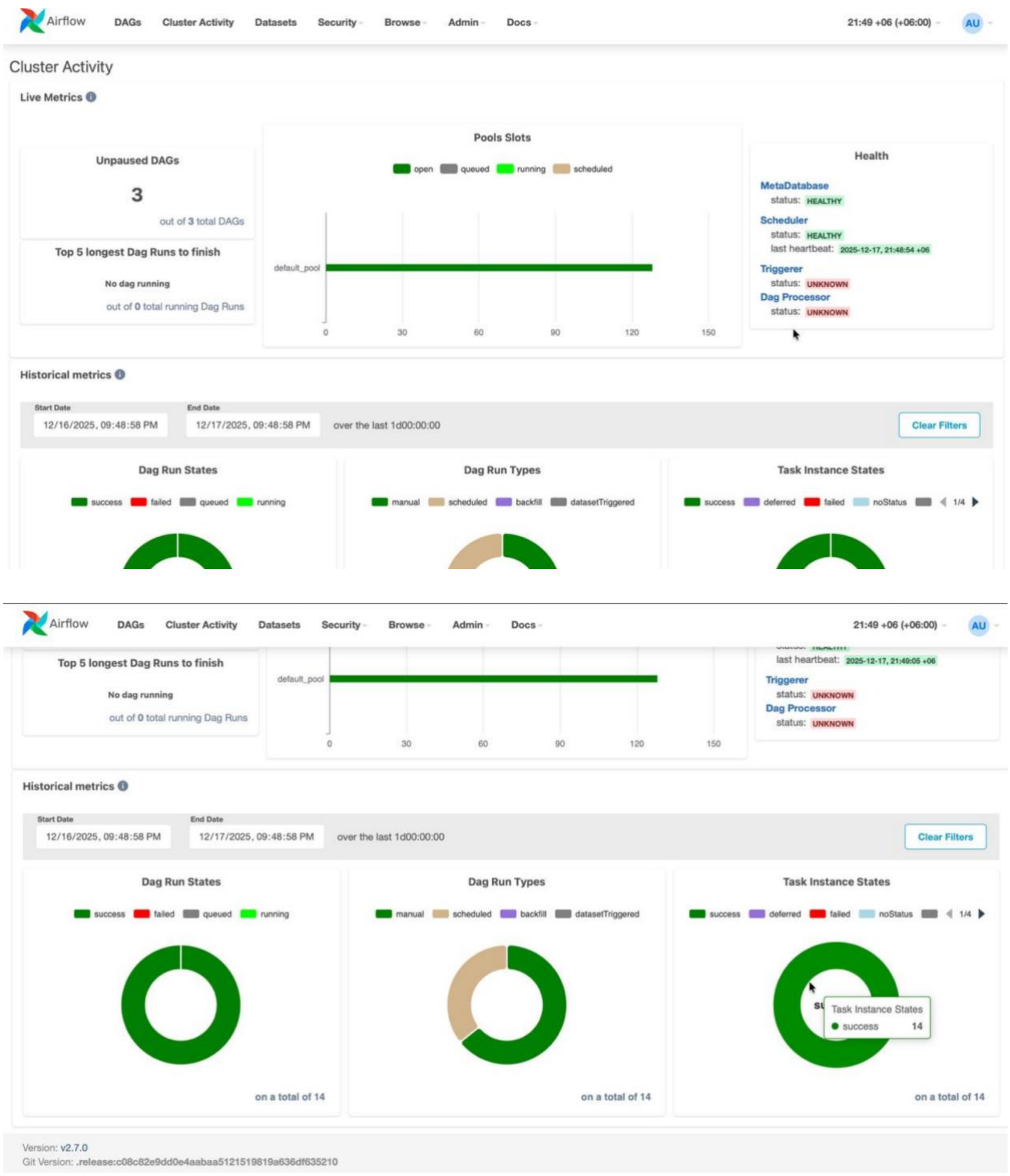

news-pipeline
job1_producer.py 2
README.md
job2_cleaner.py 2
job3_analytics.py 1
app.db
job1_ingestion_dag.py 1
job2_clean_store_dag.py 1
data > app.db
SELECT * FROM daily_summary
Schema
Query Editor
Auto Reload
Find
Other Tools...
SQLite 3.51.
total_articles INTEGER
unique_sources INTEGER
avg_title_length REAL
avg_description_length REAL
top_source TEXT
top_source_count INTEGER
articles_with_author INTEGER
articles_with_content INTEGER
created_at TIMESTAMP DEFAULT
1 1 16 85.0 110.66666666666667 NPR 3 19 17 2025-12-19 18:46:…


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
21:50 +06 (+06:00)
AU
State
Dag Id
Task Id
Run Id
Map Index
Logical Date
Operator
Start Date
End Date
Duration
Note
Job Id
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 2M:1s 2
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
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
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
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
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
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
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
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
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
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 2M:1s 46
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
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 2M:1s 50
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
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		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		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	running	LocalTaskJob	2025-12-17, 21:12:08		2025-12-17, 21:12:38		934d07147bb7	airflow
13	job1_continuous_ingestion	running	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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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.