Final Group Project Report

Team Members

Student 1: 23B151054 Khvan Anna
Student 2: 23B151041 Ilyassova Kamilla
Student 3: 23B031335 Kelis Karina

Course: Data Collection & Preparation
Submission Date: December 19, 2025

1. API Selection and Justification

Chosen API: GNews API (**https://gnews.io/**)

API Description

The GNews API is a real-time news aggregation service that collects articles from thousands of online news publishers worldwide. The API aggregates content from trusted media sources and exposes it through a clean, structured JSON interface. Because new articles are continuously published by news outlets, the data returned by GNews is frequently updated.

The GNews API satisfies all project requirements:

- It provides frequently updated news data, with new articles appearing every few minutes
- All responses are returned in structured JSON format
- The API is stable, well-documented, and widely used in production systems
- It delivers real, meaningful, non-random data suitable for analysis

Additionally, the ability to filter data by keyword, language, and time range makes the API highly suitable for continuous ingestion, downstream cleaning, and daily analytical aggregation tasks.

2. DAG 1 - Continuous Ingestion Job

This DAG implements pseudo-streaming ingestion by running a long-lived task that continuously polls the GNews API and streams raw news articles into Kafka. The job queries the API every 60 seconds and immediately publishes each fetched article as a JSON message to a Kafka topic.

Output: kafka topic (`raw_events`)
Message Format: JSON

Kafka Topic Schema

Each Kafka message represents one news article fetched from the API:

```
{
"article_id": "f3150c1a267ca89f938e06ca00cb631f",
"title": "Study: China Leads in 90 Percent of Critical Technologies",
"description": "A new Australian Strategic Policy Institute report flags...",
```

"content": "China is the world's leader in nearly 90% of critical technologies...",
"url": "https://www.newsmax.com/politics/china-leads-technologies/2025/12/14/id/1238262/",
"image": "https://www.newsmax.com/CMSPages/GetFile.aspx?...",
"published_at": "2025-12-14T21:08:34Z",
"source_name": "Newsmax",
"query": "technology",
"fetched_at_utc": "2025-12-15T09:09:12.370128+00:00"
}

3. DAG 2 - Hourly Cleaning and Storage Job

This batch job reads all new messages from Kafka, converts them into a Pandas DataFrame, applies cleaning and normalization rules, and stores the cleaned data in SQLite.

Cleaning Rules

- Drop records with missing critical fields: `article_id`, `published_at`, or `url`
- Convert `published_at` and `fetched_at_utc` fields to UTC-aware timestamps
- Remove records with invalid or unparsable publication timestamps
- Normalize `source_name` by trimming whitespace and converting to lowercase
- Trim whitespace from article titles
- Remove duplicate records based on the unique `article_id`

After cleaning, only the validated and normalized columns are selected and written to the database.

Output: SQLite table (`events`)

SQLite table schema:

Column Name	Type	Description
article_id	TEXT (Primary Key)	Unique article identifier
published_at	TEXT	Publication timestamp
fetched_at_utc	TEXT	Ingestion timestamp
source_name	TEXT	News source
title	TEXT	Article title
description	TEXT	Article description
url	TEXT	Article URL
query	TEXT	Search keyword

4. DAG 3 - Daily Analytics Job

This job performs batch analytical computations on the cleaned data stored in the SQLite database. The analytics are global over the entire dataset available at execution time and are recomputed once per day.

Analytical Metrics

The daily analytics job computes global aggregated statistics over all cleaned records available at execution time:

- Total number of collected articles (`total_articles`)
- Number of unique news sources (`unique_sources`)
- Average article title length (`avg_title_length`)

The job includes validation checks to ensure that the `events` table exists and contains data before performing analytics.

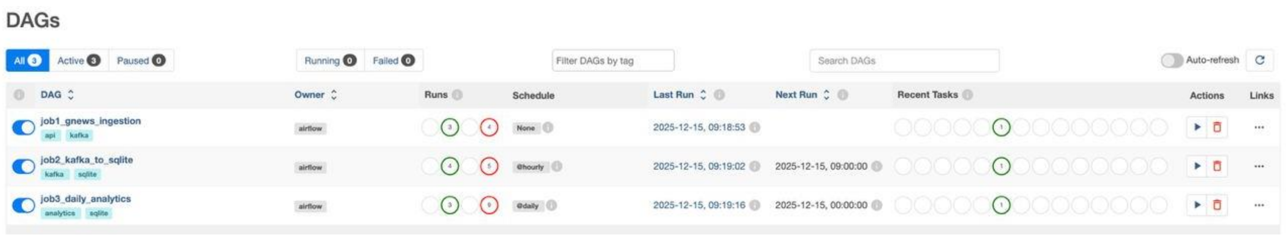
Output: SQLite table (`daily_summary`)

SQLite Database Schema

Column Name	Type	Description
total_articles	INTEGER	Total number of articles
unique_sources	INTEGER	Number of distinct news sources
avg_title_length	REAL	Average article title length

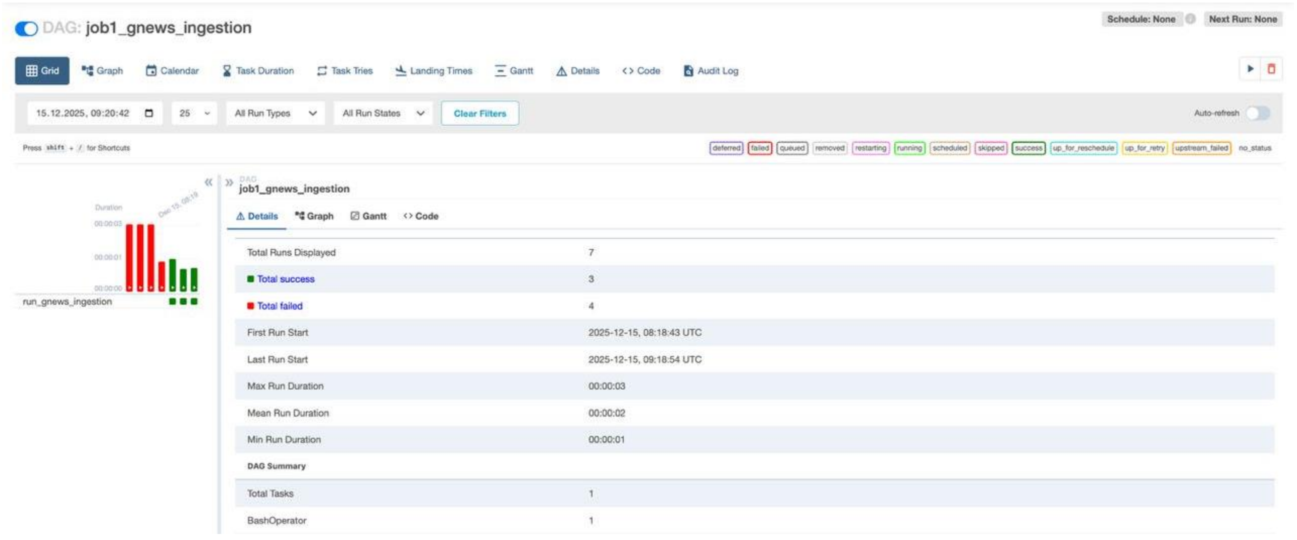
5. Airflow DAG Validation

All three DAGs were successfully executed:

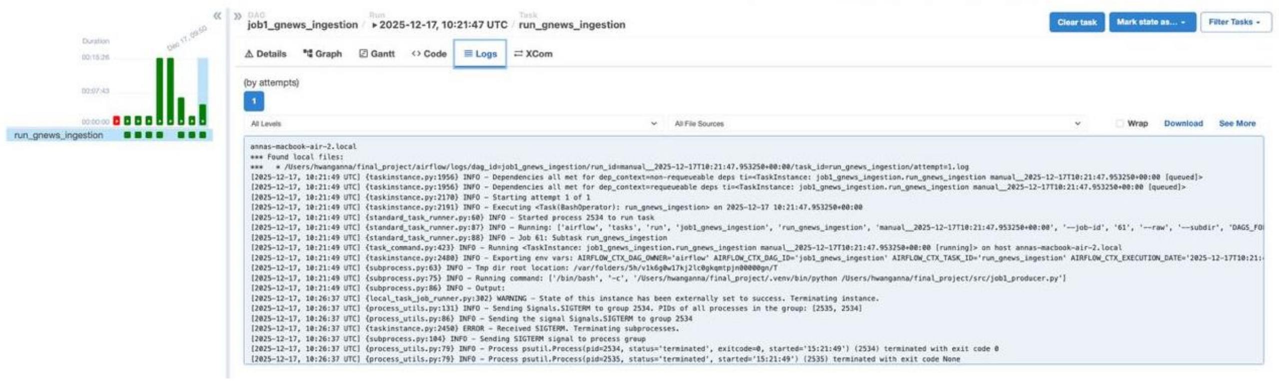


DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks
job1_gnews_ingestion (api, kafka)	airflow	3, 4	None	2025-12-15, 09:18:53		1
job2_kafka_to_sqlite (kafka, sqlite)	airflow	4, 5	@hourly	2025-12-15, 09:19:02	2025-12-15, 09:00:00	1
job3_daily_analytics (analytics, sqlite)	airflow	3, 9	@daily	2025-12-15, 09:19:16	2025-12-15, 00:00:00	1

1) DAG 1:

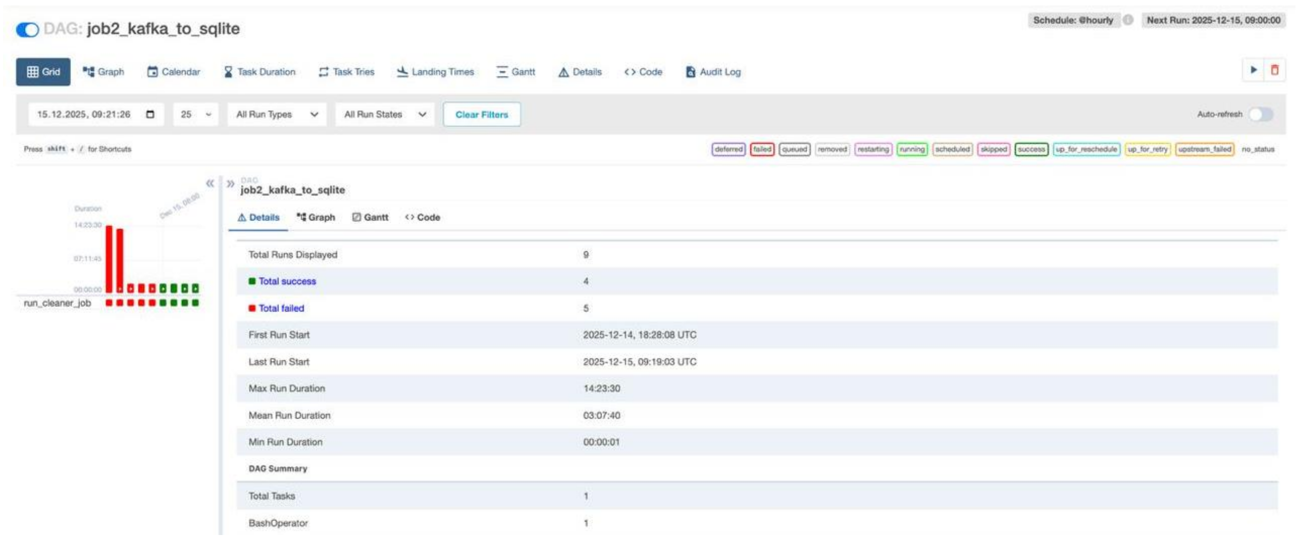


Total Runs Displayed	7
🟩 Total success	3
🟥 Total failed	4
First Run Start	2025-12-15, 08:18:43 UTC
Last Run Start	2025-12-15, 09:18:54 UTC
Max Run Duration	00:00:03
Mean Run Duration	00:00:02
Min Run Duration	00:00:01
DAG Summary	
Total Tasks	1
BashOperator	1

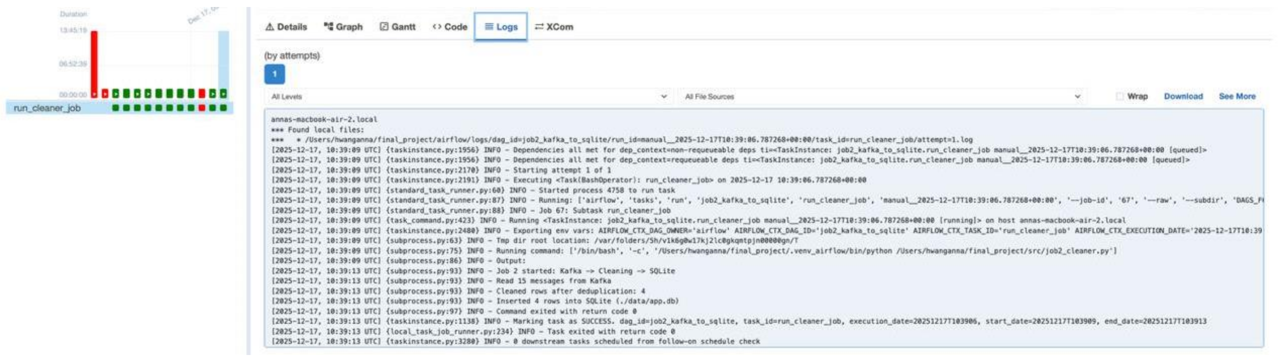


```
annas-macbook-air-2.local
*** Found local files:
***   * /Users/hwanganna/final_project/airflow/logs/dag_id=job1_gnews_ingestion/run_id=manual__2025-12-17T10:21:47.953250+00:00/task_id=run_gnews_ingestion/attempt=1.log
[2025-12-17, 10:21:49 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: job1_gnews_ingestion.run_gnews_ingestion manual__2025-12-17T10:21:47.953250+00:00 [queued]>
[2025-12-17, 10:21:49 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: job1_gnews_ingestion.run_gnews_ingestion manual__2025-12-17T10:21:47.953250+00:00 [queued]>
[2025-12-17, 10:21:49 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 1
[2025-12-17, 10:21:49 UTC] {taskinstance.py:2191} INFO - Executing <Task(BashOperator): run_gnews_ingestion> on 2025-12-17 10:21:47.953250+00:00
[2025-12-17, 10:21:49 UTC] {standard_task_runner.py:60} INFO - Started process 2534 to run task
[2025-12-17, 10:21:49 UTC] {standard_task_runner.py:87} INFO - Running: ['airflow', 'tasks', 'run', 'job1_gnews_ingestion', 'run_gnews_ingestion', 'manual__2025-12-17T10:21:47.953250+00:00', '--job-id', '61', '--raw', '--subdir', 'DAGS_FO
[2025-12-17, 10:21:49 UTC] {standard_task_runner.py:88} INFO - Job 61: Subtask run_gnews_ingestion
[2025-12-17, 10:21:49 UTC] {task_command.py:423} INFO - Running <TaskInstance: job1_gnews_ingestion.run_gnews_ingestion manual__2025-12-17T10:21:47.953250+00:00 [running]> on host annas-macbook-air-2.local
[2025-12-17, 10:21:49 UTC] {taskinstance.py:2480} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='airflow' AIRFLOW_CTX_DAG_ID='job1_gnews_ingestion' AIRFLOW_CTX_TASK_ID='run_gnews_ingestion' AIRFLOW_CTX_EXECUTION_DATE='2025-12-17T10:21:
[2025-12-17, 10:21:49 UTC] {subprocess.py:63} INFO - Tmp dir root location: /var/folders/5h/v1k6g0w17kj2lc0gkqmtpjn00000gn/T
[2025-12-17, 10:21:49 UTC] {subprocess.py:75} INFO - Running command: ['/bin/bash', '-c', '/Users/hwanganna/final_project/.venv/bin/python /Users/hwanganna/final_project/src/job1_producer.py']
[2025-12-17, 10:21:49 UTC] {subprocess.py:86} INFO - Output:
[2025-12-17, 10:26:37 UTC] {local_task_job_runner.py:302} WARNING - State of this instance has been externally set to success. Terminating instance.
[2025-12-17, 10:26:37 UTC] {process_utils.py:131} INFO - Sending Signals.SIGTERM to group 2534. PIDs of all processes in the group: [2535, 2534]
[2025-12-17, 10:26:37 UTC] {process_utils.py:86} INFO - Sending the signal Signals.SIGTERM to group 2534
[2025-12-17, 10:26:37 UTC] {taskinstance.py:2450} ERROR - Received SIGTERM. Terminating subprocesses.
[2025-12-17, 10:26:37 UTC] {subprocess.py:104} INFO - Sending SIGTERM signal to process group
[2025-12-17, 10:26:37 UTC] {process_utils.py:79} INFO - Process psutil.Process(pid=2534, status='terminated', exitcode=0, started='15:21:49') (2534) terminated with exit code 0
[2025-12-17, 10:26:37 UTC] {process_utils.py:79} INFO - Process psutil.Process(pid=2535, status='terminated', started='15:21:49') (2535) terminated with exit code None
```

2) DAG 2:

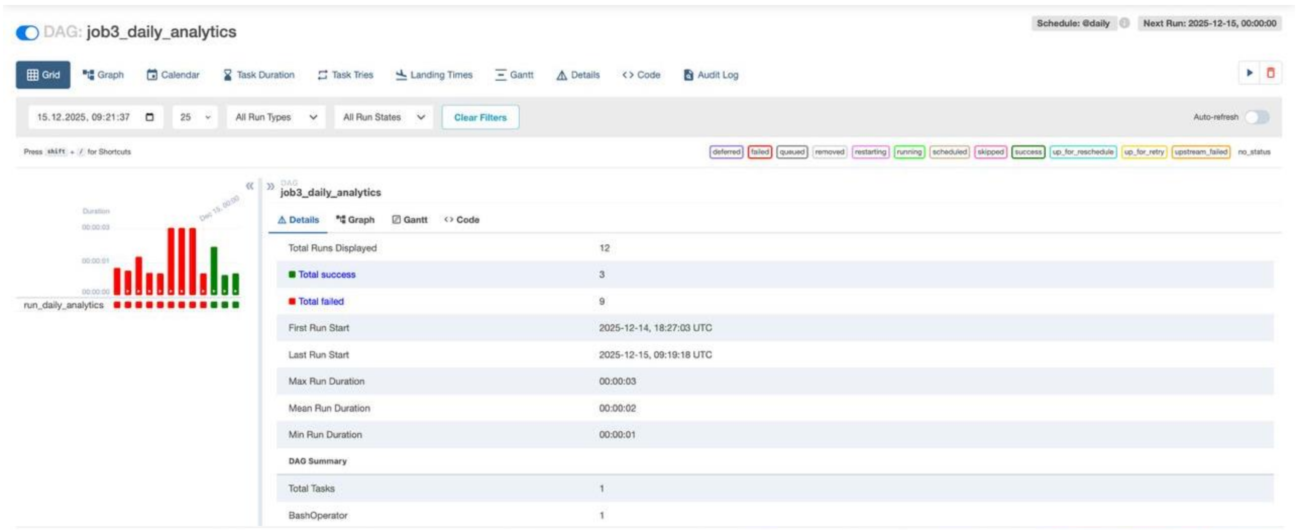


Total Runs Displayed	9
🟩 Total success	4
🟥 Total failed	5
First Run Start	2025-12-14, 18:28:08 UTC
Last Run Start	2025-12-15, 09:19:03 UTC
Max Run Duration	14:23:30
Mean Run Duration	03:07:40
Min Run Duration	00:00:01
DAG Summary	
Total Tasks	1
BashOperator	1

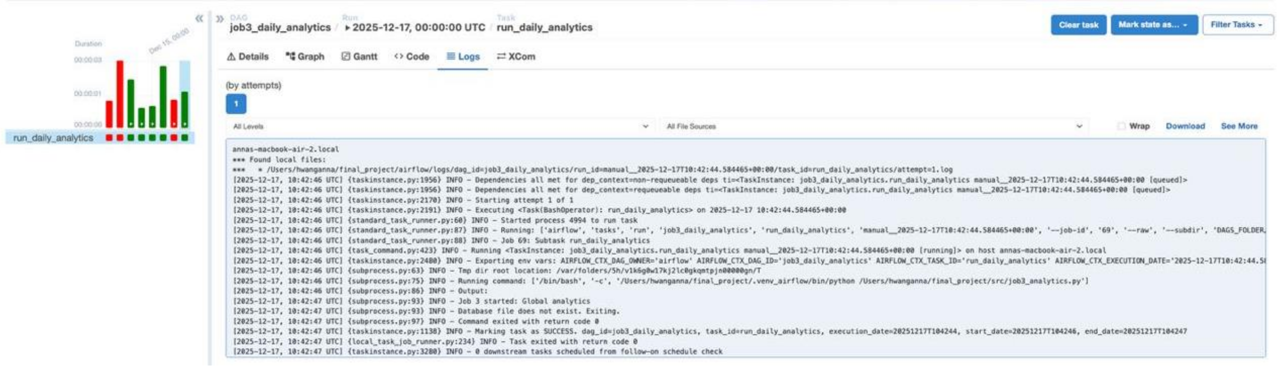


```
annas-macbook-air-2.local
*** Found local files:
***   * /Users/hwanganna/final_project/airflow/logs/dag_id=job2_kafka_to_sqlite/run_id=manual__2025-12-17T10:39:06.787268+00:00/task_id=run_cleaner_job/attempt=1.log
[2025-12-17, 10:39:09 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: job2_kafka_to_sqlite.run_cleaner_job manual__2025-12-17T10:39:06.787268+00:00 [queued]>
[2025-12-17, 10:39:09 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: job2_kafka_to_sqlite.run_cleaner_job manual__2025-12-17T10:39:06.787268+00:00 [queued]>
[2025-12-17, 10:39:09 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 1
[2025-12-17, 10:39:09 UTC] {taskinstance.py:2191} INFO - Executing <Task(BashOperator): run_cleaner_job> on 2025-12-17 10:39:06.787268+00:00
[2025-12-17, 10:39:09 UTC] {standard_task_runner.py:60} INFO - Started process 4758 to run task
[2025-12-17, 10:39:09 UTC] {standard_task_runner.py:87} INFO - Running: ['airflow', 'tasks', 'run', 'job2_kafka_to_sqlite', 'run_cleaner_job', 'manual__2025-12-17T10:39:06.787268+00:00', '--job-id', '67', '--raw', '--subdir', 'DAGS_F
[2025-12-17, 10:39:09 UTC] {standard_task_runner.py:88} INFO - Job 67: Subtask run_cleaner_job
[2025-12-17, 10:39:09 UTC] {task_command.py:423} INFO - Running <TaskInstance: job2_kafka_to_sqlite.run_cleaner_job manual__2025-12-17T10:39:06.787268+00:00 [running]> on host annas-macbook-air-2.local
[2025-12-17, 10:39:09 UTC] {taskinstance.py:2480} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='airflow' AIRFLOW_CTX_DAG_ID='job2_kafka_to_sqlite' AIRFLOW_CTX_TASK_ID='run_cleaner_job' AIRFLOW_CTX_EXECUTION_DATE='2025-12-17T10:39
[2025-12-17, 10:39:09 UTC] {subprocess.py:63} INFO - Tmp dir root location: /var/folders/5h/v1k6g0w17kj2lc0gkqmtpjn0000gn/T
[2025-12-17, 10:39:09 UTC] {subprocess.py:75} INFO - Running command: ['/bin/bash', '-c', '/Users/hwanganna/final_project/.venv_airflow/bin/python /Users/hwanganna/final_project/src/job2_cleaner.py']
[2025-12-17, 10:39:09 UTC] {subprocess.py:86} INFO - Output:
[2025-12-17, 10:39:13 UTC] {subprocess.py:93} INFO - Job 2 started: Kafka -> Cleaning -> SQLite
[2025-12-17, 10:39:13 UTC] {subprocess.py:93} INFO - Read 15 messages from Kafka
[2025-12-17, 10:39:13 UTC] {subprocess.py:93} INFO - Cleaned rows after deduplication: 4
[2025-12-17, 10:39:13 UTC] {subprocess.py:93} INFO - Inserted 4 rows into SQLite (./data/app.db)
[2025-12-17, 10:39:13 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2025-12-17, 10:39:13 UTC] {taskinstance.py:1138} INFO - Marking task as SUCCESS. dag_id=job2_kafka_to_sqlite, task_id=run_cleaner_job, execution_date=20251217T103906, start_date=20251217T103909, end_date=20251217T103913
[2025-12-17, 10:39:13 UTC] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-12-17, 10:39:13 UTC] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

3) DAG 3:



Total Runs Displayed	12
🟩 Total success	3
🟥 Total failed	9
First Run Start	2025-12-14, 18:27:03 UTC
Last Run Start	2025-12-15, 09:19:18 UTC
Max Run Duration	00:00:03
Mean Run Duration	00:00:02
Min Run Duration	00:00:01
DAG Summary	
Total Tasks	1
BashOperator	1



```
annas-macbook-air-2.local
*** Found local files:
***   * /Users/hwanganna/final_project/airflow/logs/dag_id=job3_daily_analytics/run_id=manual__2025-12-17T10:42:44.584465+00:00/task_id=run_daily_analytics/attempt=1.log
[2025-12-17, 10:42:46 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: job3_daily_analytics.run_daily_analytics manual__2025-12-17T10:42:44.584465+00:00 [queued]>
[2025-12-17, 10:42:46 UTC] {taskinstance.py:1956} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: job3_daily_analytics.run_daily_analytics manual__2025-12-17T10:42:44.584465+00:00 [queued]>
[2025-12-17, 10:42:46 UTC] {taskinstance.py:2170} INFO - Starting attempt 1 of 1
[2025-12-17, 10:42:46 UTC] {taskinstance.py:2191} INFO - Executing <Task(BashOperator): run_daily_analytics> on 2025-12-17 10:42:44.584465+00:00
[2025-12-17, 10:42:46 UTC] {standard_task_runner.py:60} INFO - Started process 4994 to run task
[2025-12-17, 10:42:46 UTC] {standard_task_runner.py:87} INFO - Running: ['airflow', 'tasks', 'run', 'job3_daily_analytics', 'run_daily_analytics', 'manual__2025-12-17T10:42:44.584465+00:00', '--job-id', '69', '--raw', '--subdir', 'DAGS_FOLDER
[2025-12-17, 10:42:46 UTC] {standard_task_runner.py:88} INFO - Job 69: Subtask run_daily_analytics
[2025-12-17, 10:42:46 UTC] {task_command.py:423} INFO - Running <TaskInstance: job3_daily_analytics.run_daily_analytics manual__2025-12-17T10:42:44.584465+00:00 [running]> on host annas-macbook-air-2.local
[2025-12-17, 10:42:46 UTC] {taskinstance.py:2480} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='airflow' AIRFLOW_CTX_DAG_ID='job3_daily_analytics' AIRFLOW_CTX_TASK_ID='run_daily_analytics' AIRFLOW_CTX_EXECUTION_DATE='2025-12-17T10:42:44.5
[2025-12-17, 10:42:46 UTC] {subprocess.py:63} INFO - Tmp dir root location: /var/folders/5h/v1k6g0w17kj2lc0gkqmtpjn00000gn/T
[2025-12-17, 10:42:46 UTC] {subprocess.py:75} INFO - Running command: ['/bin/bash', '-c', '/Users/hwanganna/final_project/.venv_airflow/bin/python /Users/hwanganna/final_project/src/job3_analytics.py']
[2025-12-17, 10:42:46 UTC] {subprocess.py:86} INFO - Output:
[2025-12-17, 10:42:47 UTC] {subprocess.py:93} INFO - Job 3 started: Global analytics
[2025-12-17, 10:42:47 UTC] {subprocess.py:93} INFO - Database file does not exist. Exiting.
[2025-12-17, 10:42:47 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2025-12-17, 10:42:47 UTC] {taskinstance.py:1138} INFO - Marking task as SUCCESS. dag_id=job3_daily_analytics, task_id=run_daily_analytics, execution_date=20251217T104244, start_date=20251217T104246, end_date=20251217T104247
[2025-12-17, 10:42:47 UTC] {local_task_job_runner.py:234} INFO - Task exited with return code 0
[2025-12-17, 10:42:47 UTC] {taskinstance.py:3280} INFO - 0 downstream tasks scheduled from follow-on schedule check
```