
Data Engineering Final Project Report

Project Title: Almaty Real-Time Weather Analytics Pipeline

Team Members:

- **Student 1:** Zhilikbay Arman - **22B030358**
- **Student 2:** Shakhizada Zgansulu - **22B030468**
- **Student 3:** Myrzakhankyzy Arailym - **22B030408**

---

1. Project Overview

This project involved developing a system for automatically collecting, cleaning, and analyzing real-time weather data for the city of Almaty. The system is built on a microservices architecture using Apache Airflow, Apache Kafka, and Docker containerization.

1. API Justification

For this project, we selected the **Tomorrow.io Weather API**.

- **Update Frequency:** The API provides real-time data. According to the limits (500 requests per day / 25 requests per hour), we configured data collection to be every 3 minutes (20 requests per hour), allowing us to maximize the free limit and obtain up-to-date data.
- **Data Quality:** It returns high-precision meteorological data (temperature, humidity, precipitation) for specific GPS coordinates (Almaty: 43.2220, 76.8512).
- **Format:** The API provides structured JSON responses, which are ideal for streaming into Kafka.

2. Kafka Topic Schema

We use a single Kafka topic named raw_weather_events. The producer sends messages in JSON format.

Example Message Payload:

JSON

```
{

"timestamp": "2023-12-18T10:00:00Z",

"location": "Almaty",

"temperature": 2.5,

"humidity": 78,

"wind_speed": 3.1,

"precipitation": 0.0,
"weather_code": 1000

}
```

3. Data Cleaning Rules (Pandas)

In **DAG 2**, we implemented the following cleaning logic using the **Pandas** library to ensure data integrity:

1. **Deduplication:** We remove any records with identical location and timestamp values to avoid double-counting.
2. **Type Conversion:** All numeric fields (temperature, humidity, etc.) are explicitly cast using pd.to_numeric with errors='coerce' to handle unexpected strings.
3. **Outlier Handling:** We filter out unrealistic values (e.g., temperatures outside the -50°C to +50°C range).
4. **Missing Values:** Any null values found in sensor data are filled with the **median** value of the current batch to maintain statistical consistency.
5. **Rounding:** Temperatures and humidity are rounded to 2 decimal places for storage efficiency.

4. Database Schema (SQLite)

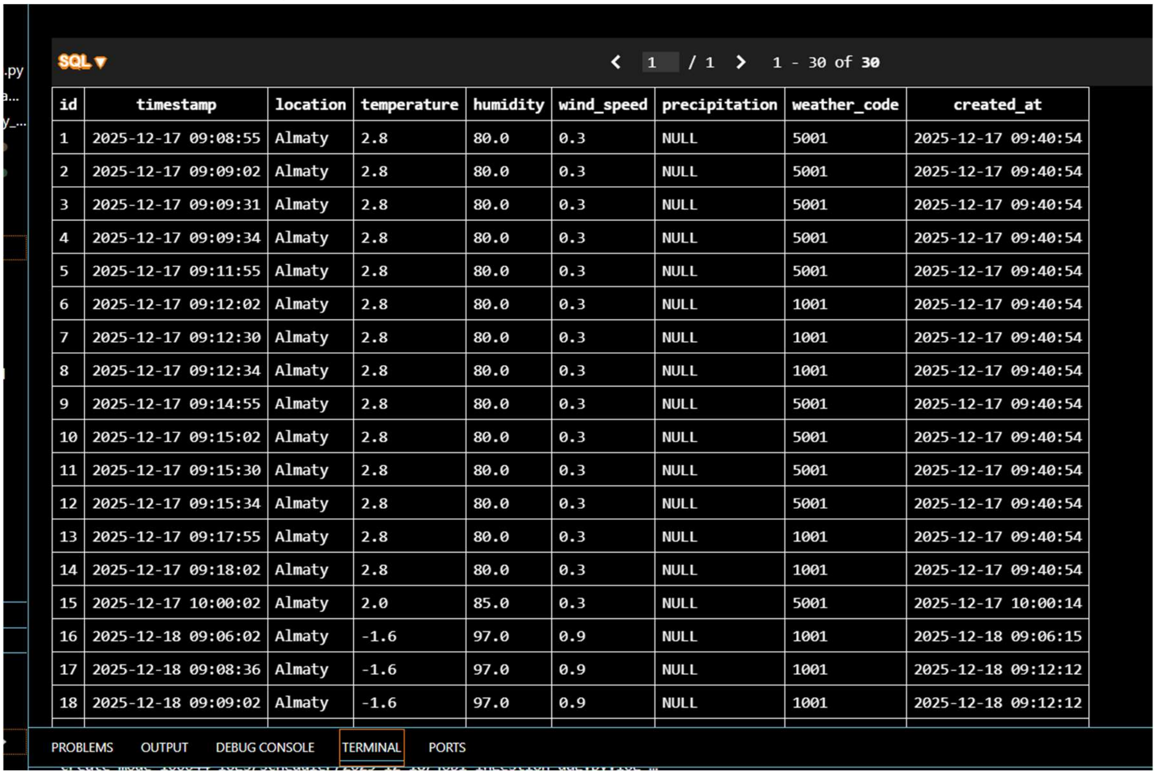
We use two tables in our app.db database:

Table: events (Cleaned Raw Data)

Column	Type	Description
id	INTEGER	Primary Key (Autoincrement)
timestamp	TEXT	ISO8601 string of the weather observation
temperature	REAL	Temperature in Celsius
humidity	REAL	Relative humidity percentage
precipitation	REAL	Precipitation intensity (mm/hr)

Table: daily_summary (Aggregated Data)

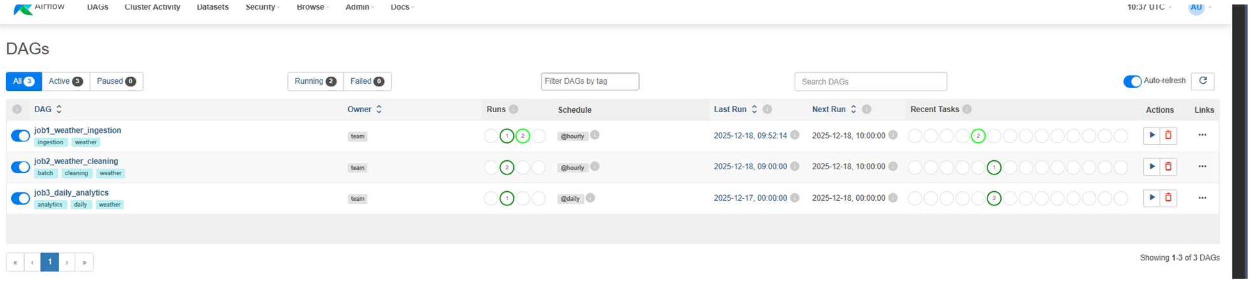
Column	Type	Description
date	TEXT	The date of the summary (Unique Key)
avg_temperature	REAL	Mean temperature for the day
min_temperature	REAL	Daily minimum
max_temperature	REAL	Daily maximum
record_count	INTEGER	Total number of samples collected that day



id	timestamp	location	temperature	humidity	wind_speed	precipitation	weather_code	created_at
1	2025-12-17 09:08:55	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
2	2025-12-17 09:09:02	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
3	2025-12-17 09:09:31	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
4	2025-12-17 09:09:34	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
5	2025-12-17 09:11:55	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
6	2025-12-17 09:12:02	Almaty	2.8	80.0	0.3	NULL	1001	2025-12-17 09:40:54
7	2025-12-17 09:12:30	Almaty	2.8	80.0	0.3	NULL	1001	2025-12-17 09:40:54
8	2025-12-17 09:12:34	Almaty	2.8	80.0	0.3	NULL	1001	2025-12-17 09:40:54
9	2025-12-17 09:14:55	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
10	2025-12-17 09:15:02	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
11	2025-12-17 09:15:30	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
12	2025-12-17 09:15:34	Almaty	2.8	80.0	0.3	NULL	5001	2025-12-17 09:40:54
13	2025-12-17 09:17:55	Almaty	2.8	80.0	0.3	NULL	1001	2025-12-17 09:40:54
14	2025-12-17 09:18:02	Almaty	2.8	80.0	0.3	NULL	1001	2025-12-17 09:40:54
15	2025-12-17 10:00:02	Almaty	2.0	85.0	0.3	NULL	5001	2025-12-17 10:00:14
16	2025-12-18 09:06:02	Almaty	-1.6	97.0	0.9	NULL	1001	2025-12-18 09:06:15
17	2025-12-18 09:08:36	Almaty	-1.6	97.0	0.9	NULL	1001	2025-12-18 09:12:12
18	2025-12-18 09:09:02	Almaty	-1.6	97.0	0.9	NULL	1001	2025-12-18 09:12:12

5. Implementation Screenshots

DAGs Inventory– Showing all three DAGs (job1, job2, job3) in the "Active" state.

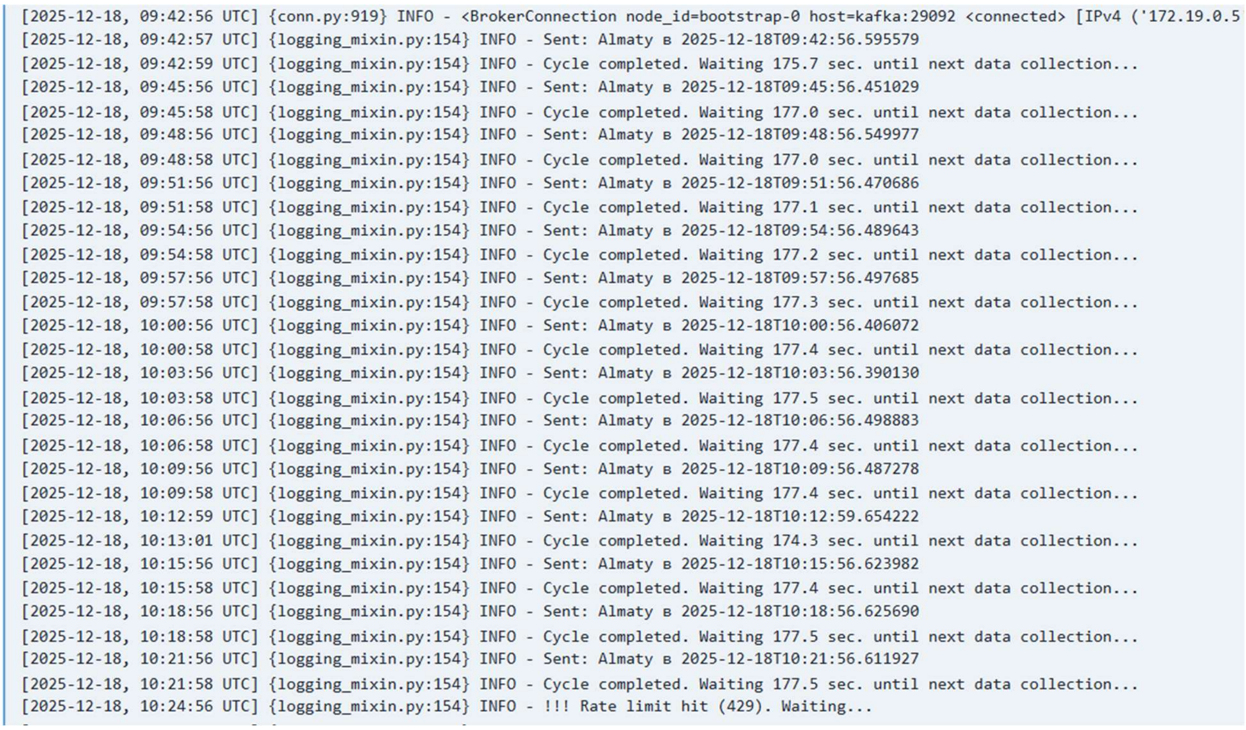


DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
job1_weather_ingestion ingestion weather	team	1 2	@hourly	2025-12-18, 09:52:14	2025-12-18, 10:00:00	2	▶ 🗑	···
job2_weather_cleaning batch cleaning weather	team	2	@hourly	2025-12-18, 09:00:00	2025-12-18, 10:00:00	1	▶ 🗑	···
job3_daily_analytics analytics daily weather	team	1	@daily	2025-12-17, 00:00:00	2025-12-18, 00:00:00	2	▶ 🗑	···

« ‹ 1 › »

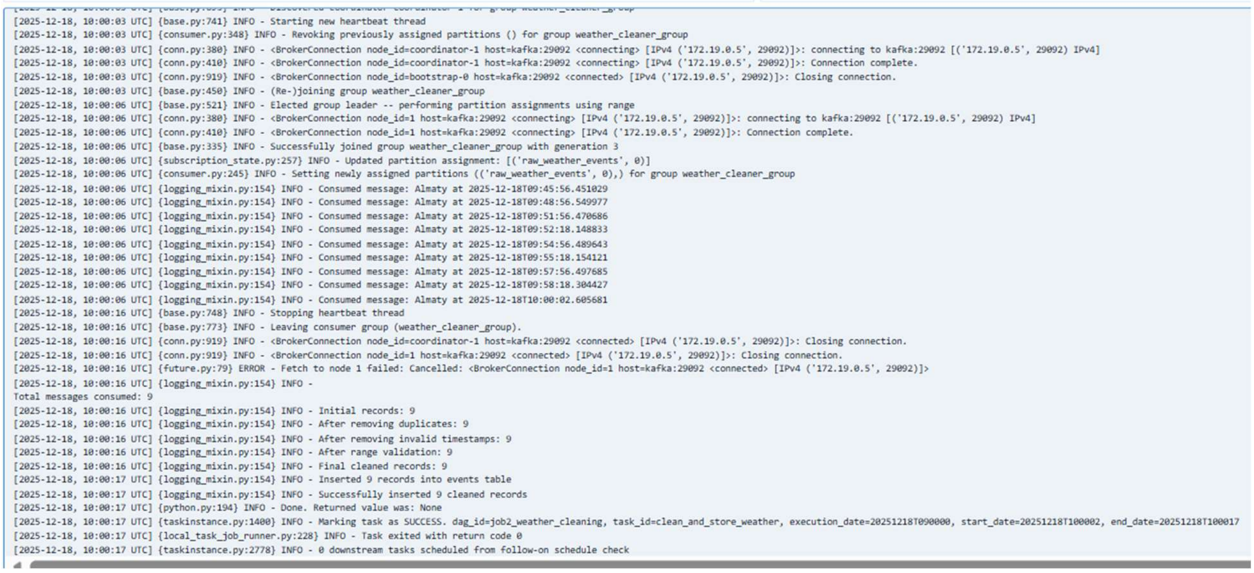
Showing 1-3 of 3 DAGs

DAG 1 Graph View – Showing the continuous ingestion task.



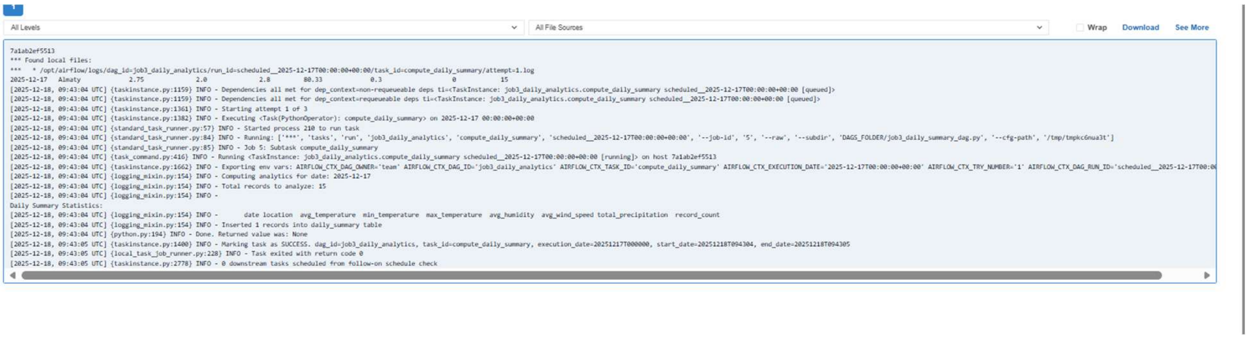
```
[2025-12-18, 09:42:56 UTC] {conn.py:919} INFO - <BrokerConnection node_id=bootstrap-0 host=kafka:29092 <connected> [IPv4 ('172.19.0.5
[2025-12-18, 09:42:57 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T09:42:56.595579
[2025-12-18, 09:42:59 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 175.7 sec. until next data collection...
[2025-12-18, 09:45:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T09:45:56.451029
[2025-12-18, 09:45:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.0 sec. until next data collection...
[2025-12-18, 09:48:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T09:48:56.549977
[2025-12-18, 09:48:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.0 sec. until next data collection...
[2025-12-18, 09:51:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T09:51:56.470686
[2025-12-18, 09:51:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.1 sec. until next data collection...
[2025-12-18, 09:54:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T09:54:56.489643
[2025-12-18, 09:54:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.2 sec. until next data collection...
[2025-12-18, 09:57:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T09:57:56.497685
[2025-12-18, 09:57:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.3 sec. until next data collection...
[2025-12-18, 10:00:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:00:56.406072
[2025-12-18, 10:00:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.4 sec. until next data collection...
[2025-12-18, 10:03:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:03:56.390130
[2025-12-18, 10:03:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.5 sec. until next data collection...
[2025-12-18, 10:06:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:06:56.498883
[2025-12-18, 10:06:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.4 sec. until next data collection...
[2025-12-18, 10:09:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:09:56.487278
[2025-12-18, 10:09:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.4 sec. until next data collection...
[2025-12-18, 10:12:59 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:12:59.654222
[2025-12-18, 10:13:01 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 174.3 sec. until next data collection...
[2025-12-18, 10:15:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:15:56.623982
[2025-12-18, 10:15:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.4 sec. until next data collection...
[2025-12-18, 10:18:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:18:56.625690
[2025-12-18, 10:18:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.5 sec. until next data collection...
[2025-12-18, 10:21:56 UTC] {logging_mixin.py:154} INFO - Sent: Almaty в 2025-12-18T10:21:56.611927
[2025-12-18, 10:21:58 UTC] {logging_mixin.py:154} INFO - Cycle completed. Waiting 177.5 sec. until next data collection...
[2025-12-18, 10:24:56 UTC] {logging_mixin.py:154} INFO - !!! Rate limit hit (429). Waiting...
```

DAG 2 Logs – A screenshot of the logs showing "Inserted X records into events table".



```
[2025-12-18, 10:00:03 UTC] {base.py:741} INFO - Starting new heartbeat thread
[2025-12-18, 10:00:03 UTC] {consumer.py:348} INFO - Revoking previously assigned partitions () for group weather_cleaner_group
[2025-12-18, 10:00:03 UTC] {conn.py:380} INFO - <BrokerConnection node_id=coordinator-1 host=kafka:29092 <connecting> [IPv4 ('172.19.0.5', 29092)]>: connecting to kafka:29092 [('172.19.0.5', 29092) IPv4]
[2025-12-18, 10:00:03 UTC] {conn.py:410} INFO - <BrokerConnection node_id=coordinator-1 host=kafka:29092 <connecting> [IPv4 ('172.19.0.5', 29092)]>: Connection complete.
[2025-12-18, 10:00:03 UTC] {conn.py:919} INFO - <BrokerConnection node_id=bootstrap-0 host=kafka:29092 <connected> [IPv4 ('172.19.0.5', 29092)]>: Closing connection.
[2025-12-18, 10:00:03 UTC] {base.py:450} INFO - (Re-)joining group weather_cleaner_group
[2025-12-18, 10:00:06 UTC] {base.py:521} INFO - Elected group leader -- performing partition assignments using range
[2025-12-18, 10:00:06 UTC] {conn.py:380} INFO - <BrokerConnection node_id=1 host=kafka:29092 <connecting> [IPv4 ('172.19.0.5', 29092)]>: connecting to kafka:29092 [('172.19.0.5', 29092) IPv4]
[2025-12-18, 10:00:06 UTC] {conn.py:410} INFO - <BrokerConnection node_id=1 host=kafka:29092 <connecting> [IPv4 ('172.19.0.5', 29092)]>: Connection complete.
[2025-12-18, 10:00:06 UTC] {base.py:335} INFO - Successfully joined group weather_cleaner_group with generation 3
[2025-12-18, 10:00:06 UTC] {subscription_state.py:257} INFO - Updated partition assignment: [('raw_weather_events', 0)]
[2025-12-18, 10:00:06 UTC] {consumer.py:245} INFO - Setting newly assigned partitions (('raw_weather_events', 0),) for group weather_cleaner_group
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:45:56.451029
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:48:56.549977
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:51:56.470686
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:52:18.148833
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:54:56.489643
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:55:18.154121
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:57:56.497685
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T09:58:18.304427
[2025-12-18, 10:00:06 UTC] {logging_mixin.py:154} INFO - Consumed message: Almaty at 2025-12-18T10:00:02.605681
[2025-12-18, 10:00:16 UTC] {base.py:748} INFO - Stopping heartbeat thread
[2025-12-18, 10:00:16 UTC] {base.py:773} INFO - Leaving consumer group (weather_cleaner_group).
[2025-12-18, 10:00:16 UTC] {conn.py:919} INFO - <BrokerConnection node_id=coordinator-1 host=kafka:29092 <connected> [IPv4 ('172.19.0.5', 29092)]>: Closing connection.
[2025-12-18, 10:00:16 UTC] {conn.py:919} INFO - <BrokerConnection node_id=1 host=kafka:29092 <connected> [IPv4 ('172.19.0.5', 29092)]>: Closing connection.
[2025-12-18, 10:00:16 UTC] {future.py:79} ERROR - Fetch to node 1 failed: Cancelled: <BrokerConnection node_id=1 host=kafka:29092 <connected> [IPv4 ('172.19.0.5', 29092)]>
[2025-12-18, 10:00:16 UTC] {logging_mixin.py:154} INFO -
Total messages consumed: 9
[2025-12-18, 10:00:16 UTC] {logging_mixin.py:154} INFO - Initial records: 9
[2025-12-18, 10:00:16 UTC] {logging_mixin.py:154} INFO - After removing duplicates: 9
[2025-12-18, 10:00:16 UTC] {logging_mixin.py:154} INFO - After removing invalid timestamps: 9
[2025-12-18, 10:00:16 UTC] {logging_mixin.py:154} INFO - After range validation: 9
[2025-12-18, 10:00:16 UTC] {logging_mixin.py:154} INFO - Final cleaned records: 9
[2025-12-18, 10:00:17 UTC] {logging_mixin.py:154} INFO - Inserted 9 records into events table
[2025-12-18, 10:00:17 UTC] {logging_mixin.py:154} INFO - Successfully inserted 9 cleaned records
[2025-12-18, 10:00:17 UTC] {python.py:194} INFO - Done. Returned value was: None
[2025-12-18, 10:00:17 UTC] {taskinstance.py:1400} INFO - Marking task as SUCCESS. dag_id=job2_weather_cleaning, task_id=clean_and_store_weather, execution_date=20251218T090000, start_date=20251218T100002, end_date=20251218T100017
[2025-12-18, 10:00:17 UTC] {local_task_job_runner.py:228} INFO - Task exited with return code 0
[2025-12-18, 10:00:17 UTC] {taskinstance.py:2778} INFO - 0 downstream tasks scheduled from follow-on schedule check
```

DAG 3 Analytics Table – A screenshot of the Airflow logs where the Pandas summary table is printed (the print(summary.to_string()) output).



```
7a1ab2ef5513
*** Found local files:
***   * /opt/airflow/logs/dag_id=job3_daily_analytics/run_id=scheduled__2025-12-17T00:00:00+00:00/task_id=compute_daily_summary/attempt=1.log
2025-12-17   Almaty          2.75            2.0             2.8        80.33           0.3                 0          15
[2025-12-18, 09:43:04 UTC] {taskinstance.py:1159} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: job3_daily_analytics.compute_daily_summary scheduled__2025-12-17T00:00:00+00:00 [queued]>
[2025-12-18, 09:43:04 UTC] {taskinstance.py:1159} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: job3_daily_analytics.compute_daily_summary scheduled__2025-12-17T00:00:00+00:00 [queued]>
[2025-12-18, 09:43:04 UTC] {taskinstance.py:1361} INFO - Starting attempt 1 of 3
[2025-12-18, 09:43:04 UTC] {taskinstance.py:1382} INFO - Executing <Task(PythonOperator): compute_daily_summary> on 2025-12-17 00:00:00+00:00
[2025-12-18, 09:43:04 UTC] {standard_task_runner.py:57} INFO - Started process 210 to run task
[2025-12-18, 09:43:04 UTC] {standard_task_runner.py:84} INFO - Running: ['***', 'tasks', 'run', 'job3_daily_analytics', 'compute_daily_summary', 'scheduled__2025-12-17T00:00:00+00:00', '--job-id', '5', '--raw', '--subdir', 'DAGS_FOLDER/job3_daily_summary_dag.py', '--cfg-path', '/tmp/tmpkc6nua3t']
[2025-12-18, 09:43:04 UTC] {standard_task_runner.py:85} INFO - Job 5: Subtask compute_daily_summary
[2025-12-18, 09:43:04 UTC] {task_command.py:416} INFO - Running <TaskInstance: job3_daily_analytics.compute_daily_summary scheduled__2025-12-17T00:00:00+00:00 [running]> on host 7a1ab2ef5513
[2025-12-18, 09:43:04 UTC] {taskinstance.py:1662} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='team' AIRFLOW_CTX_DAG_ID='job3_daily_analytics' AIRFLOW_CTX_TASK_ID='compute_daily_summary' AIRFLOW_CTX_EXECUTION_DATE='2025-12-17T00:00:00+00:00' AIRFLOW_CTX_TRY_NUMBER='1' AIRFLOW_CTX_DAG_RUN_ID='scheduled__2025-12-17T00:0
[2025-12-18, 09:43:04 UTC] {logging_mixin.py:154} INFO - Computing analytics for date: 2025-12-17
[2025-12-18, 09:43:04 UTC] {logging_mixin.py:154} INFO - Total records to analyze: 15
[2025-12-18, 09:43:04 UTC] {logging_mixin.py:154} INFO -
Daily Summary Statistics:
[2025-12-18, 09:43:04 UTC] {logging_mixin.py:154} INFO -       date location  avg_temperature  min_temperature  max_temperature  avg_humidity  avg_wind_speed total_precipitation  record_count
[2025-12-18, 09:43:04 UTC] {logging_mixin.py:154} INFO - Inserted 1 records into daily_summary table
[2025-12-18, 09:43:04 UTC] {python.py:194} INFO - Done. Returned value was: None
[2025-12-18, 09:43:05 UTC] {taskinstance.py:1400} INFO - Marking task as SUCCESS. dag_id=job3_daily_analytics, task_id=compute_daily_summary, execution_date=20251217T000000, start_date=20251218T094304, end_date=20251218T094305
[2025-12-18, 09:43:05 UTC] {local_task_job_runner.py:228} INFO - Task exited with return code 0
[2025-12-18, 09:43:05 UTC] {taskinstance.py:2778} INFO - 0 downstream tasks scheduled from follow-on schedule check
```