

Data Collection and Preparation - Final Project

Members:

1. Amreyeva Alina – 22B031240
2. Kemel Merey – 22B030615
3. Serik Dinmukhammed – 22B030439

1. Project Goal

The objective of this project was to build a robust, scalable data pipeline capable of handling real-world, frequently updating data. We implemented a streaming ingestion with processing to analyze the Carbon Intensity of the UK National Grid. The system automates the collection, cleaning, storage, and analysis of environmental data using Apache Airflow, Kafka, and SQLite.

2. API Justification

Chosen API: Carbon Intensity UK (National Grid ESO) - <https://api.carbonintensity.org.uk>

We selected the Carbon Intensity API (UK National Grid).

Validity: The API updates every 30 minutes, satisfying the requirement for "frequently updating" data (at least every hour).

Data Quality: It provides meaningful, real-world environmental data (forecast vs. actual carbon intensity), not random numbers.

Technical Fit: The API is stable, well-documented, and returns structured JSON without requiring complex authentication.

3. Data Schemas

Kafka Topic Schema (raw_events)

The raw message sent to Kafka follows this JSON structure:

code JSON

```
{  
  "ingestion_id": 123,  
  "timestamp": "2025-12-17T10:00:00.123456",  
  "source": "carbonintensity.org.uk",  
  "payload": {  
    "from": "2025-12-17T10:00Z",  
    "to": "2025-12-17T10:30Z",  
    "intensity": {  
      "forecast": 260,  
      "actual": 255,  
      "index": "moderate"  
    }  
  }  
}
```

```
name: raw_events  
MacBook-Air: DCPFinalProject % docker-compose exec kafka kafka-console-consumer --bootstrap-server kafka:29892 --topic raw_events --from-beginning  
{"ingestion_id": 1, "timestamp": "2025-12-17T10:00:00.000000", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 134, "index": "moderate"}}, {"ingestion_id": 2, "timestamp": "2025-12-17T10:00:00.000000", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 134, "index": "moderate"}}, {"ingestion_id": 3, "timestamp": "2025-12-17T10:48:33.388937", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 134, "index": "moderate"}}, {"ingestion_id": 4, "timestamp": "2025-12-17T10:48:33.388937", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 134, "index": "moderate"}}, {"ingestion_id": 5, "timestamp": "2025-12-17T10:48:33.388937", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 134, "index": "moderate"}}, {"ingestion_id": 6, "timestamp": "2025-12-17T10:42:06.404777", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 134, "index": "moderate"}}, {"ingestion_id": 7, "timestamp": "2025-12-17T10:42:37.383849", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 8, "timestamp": "2025-12-17T10:42:37.383849", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 9, "timestamp": "2025-12-17T10:43:39.360719", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 10, "timestamp": "2025-12-17T10:44:19.286549", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 11, "timestamp": "2025-12-17T10:45:13.975129", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 12, "timestamp": "2025-12-17T10:45:13.975129", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 13, "timestamp": "2025-12-17T10:45:44.636829", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 14, "timestamp": "2025-12-17T10:45:44.636829", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 15, "timestamp": "2025-12-17T10:46:40.732379", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 16, "timestamp": "2025-12-17T10:47:17.698617", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 17, "timestamp": "2025-12-17T10:47:17.698617", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 18, "timestamp": "2025-12-17T10:48:31.769369", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 19, "timestamp": "2025-12-17T10:48:31.769369", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 20, "timestamp": "2025-12-17T10:49:23.655639", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 21, "timestamp": "2025-12-17T10:49:23.655639", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 22, "timestamp": "2025-12-17T10:50:57.763374", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 23, "timestamp": "2025-12-17T10:50:57.763374", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 24, "timestamp": "2025-12-17T10:51:29.762471", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 25, "timestamp": "2025-12-17T10:51:29.762471", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 26, "timestamp": "2025-12-17T10:52:46.624627", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 27, "timestamp": "2025-12-17T10:51:06.655829", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}, {"ingestion_id": 28, "timestamp": "2025-12-17T10:53:36.595068", "source": "carbonintensity.org.uk", "payload": {"from": "2025-12-17T10:00Z", "to": "2025-12-17T10:30Z", "intensity": {"forecast": 130, "actual": 135, "index": "moderate"}}}
```

4. Cleaning rules

All data cleaning is performed in DAG 2 using the Pandas library, as required.

Handling Missing Values: We use df.dropna(subset=['forecast_intensity']) to remove records where the primary data point is missing.

Removing Duplicates: We use df.drop_duplicates() to handle any message replays from Kafka (ensuring exactly-once processing in the DB).

Text Normalization: We apply `.str.lower().str.strip()` to the source and index_intensity columns to ensure consistency (e.g., "Moderate" becomes "moderate").

Type Conversion: We explicitly convert forecast_intensity and actual_intensity to numeric types using pd.to_numeric(..., errors='coerce') to prevent database type errors.

5. SQLite Database Schema

Table 1: events (Cleaned Data)

Column	Type	Description
---	---	---
id	INTEGER PK	Auto-incrementing ID
ingestion_timestamp	TEXT	Timestamp of ingestion
forecast_intensity	INTEGER	Predicted carbon intensity
actual_intensity	INTEGER	Realized carbon intensity
index_intensity	TEXT	Category (e.g., 'moderate')
source	TEXT	API Source
created_at	TIMESTAMP	Record creation time

cid	name	type	notnull	dft_value	pk
0	id	INTEGER	0	<null>	1
1	ingestion_timestamp	TEXT	0	<null>	0
2	forecast_intensity	INTEGER	0	<null>	0
3	actual_intensity	INTEGER	0	<null>	0
4	index_intensity	TEXT	0	<null>	0
5	source	TEXT	0	<null>	0
6	created_at	TIMESTAMP	0	CURRENT_TIMESTAMP	0

Table 2: daily_summary (Aggregated Analytics)

Column	Type	Description
summary_date	DATE	Date of the analysis
part_of_day	TEXT	Morning, Afternoon, Evening, Night
avg_forecast	REAL	Average predicted intensity
avg_actual	REAL	Average actual intensity
max_intensity	INTEGER	Max intensity recorded
min_intensity	INTEGER	Min intensity recorded

	cid	name	type	notnull	dflt_value	pk
1		summary_date	DATE	0	<null>	0
2		part_of_day	TEXT	0	<null>	0
3		avg_forecast	REAL	0	<null>	0
4		avg_actual	REAL	0	<null>	0
5		max_intensity	INTEGER	0	<null>	0
6		min_intensity	INTEGER	0	<null>	0

6. Verification

6.1 Airflow DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
job1_ingestion_dag final [project]	dcp_team	○ ○ ○ ○ ○	/15 * * * *	2025-12-17, 16:45:00	2025-12-17, 16:45:00	○ ○ ○ ○ ○	▶ ⚡	...
job2_clean_store_dag final [project]	dcp_team	○ ○ ○ ○ ○	@hourly	2025-12-17, 17:22:11	2025-12-17, 17:00:00	○ ○ ○ ○ ○	▶ ⚡	...
job3_daily_summary_dag final [project]	dcp_team	○ ○ ○ ○ ○	@daily	2025-12-17, 17:22:13	2025-12-17, 00:00:00	○ ○ ○ ○ ○	▶ ⚡	...

6.2 Job 1 Logs (Ingestion)

6.3 Job 2 Logs (Cleaning)

Press shift + / for Shortcuts

job2_clean_store_dag > 2025-12-17, 17:00:00 UTC clean_and_store

Details Graph Gantt Code Logs

(by attempts) 1

All Levels All File Sources Wrap Download See More

```
[2025-12-17, 17:12:21 UTC] [taskinstance.py:1088] INFO - Exporting env vars: AIRFLOW_CTX_DAG_ID='job2_clean_store_dag' AIRFLOW_CTX_TASK_ID='clean_and_store' AIRFLOW_CTX_EXECUTION_ID='20251217T171221'
[2025-12-17, 17:12:21 UTC] [job2_cleaner.py:143] INFO - Starting Hourly Cleaning Job
[2025-12-17, 17:12:21 UTC] [taskinstance.py:1088] INFO - Database connection established.
[2025-12-17, 17:12:21 UTC] [comms.py:469] INFO - <brokerConnection client_id=kafka-python-2.3.8, node_id=bootstrap-0 host=kafka29892> connecting to kafka29892 [('172.24.0.4', 29892)] Connection complete.
[2025-12-17, 17:12:21 UTC] [taskinstance.py:1088] INFO - <brokerConnection client_id=kafka-python-2.3.8, node_id=bootstrap-0 host=kafka29892> connected [IPv4 ('172.24.0.4', 29892)] Connection complete.
[2025-12-17, 17:12:21 UTC] [comms.py:461] INFO - <brokerConnection client_id=kafka-python-2.3.8, node_id=bootstrap-0 host=kafka29892> connected [IPv4 ('172.24.0.4', 29892)] Connection complete.
[2025-12-17, 17:12:21 UTC] [taskinstance.py:1088] INFO - <brokerConnection client_id=kafka-python-2.3.8, node_id=bootstrap-0 host=kafka29892> connecting [IPv4 ('172.24.0.4', 29892)]> connecting to kafka29892 [('172.24.0.4', 29892)] Connection complete.
[2025-12-17, 17:12:21 UTC] [comms.py:461] INFO - <brokerConnection client_id=kafka-python-2.3.8, node_id=bootstrap-0 host=kafka29892> connected [IPv4 ('172.24.0.4', 29892)]> Connection complete.
[2025-12-17, 17:12:21 UTC] [taskinstance.py:1088] INFO - Coordinator for group/cleaner_group is 'coordinator-1' [kafka29892, None]
[2025-12-17, 17:12:21 UTC] [cleaner.py:895] INFO - Starting new heartbeat thread
[2025-12-17, 17:12:21 UTC] [cleaner.py:895] INFO - Starting new heartbeat thread for group cleaner_group
[2025-12-17, 17:12:21 UTC] [cleaner.py:433] INFO - Failed to join group cleaner_group; NodeNotReadyError: coordinator-1
[2025-12-17, 17:12:21 UTC] [cleaner.py:433] INFO - Failed to join group cleaner_group; NodeNotReadyError: coordinator-1
[2025-12-17, 17:12:21 UTC] [cleaner.py:461] INFO - <brokerConnection client_id=kafka-python-2.3.8, node_id=coordinator-1 host=kafka29892> connecting to kafka29892 [('172.24.0.4', 29892)]> connecting to kafka29892 [('172.24.0.4', 29892)] Connection complete.
[2025-12-17, 17:12:21 UTC] [cleaner.py:461] INFO - Received member id kafka-python-3.8-0x434354791093-4073-0ab7>+>0&#43;979fb5 for group cleaner_group will retry join group
[2025-12-17, 17:12:21 UTC] [cleaner.py:461] INFO - Failed to join group cleaner_group; [Errno 79] MemberIdRequiredError
[2025-12-17, 17:12:21 UTC] [cleaner.py:433] INFO - Successfully joined group cleaner_group due to 791 [Errno 79] MemberIdRequiredError
[2025-12-17, 17:12:21 UTC] [cleaner.py:433] INFO - Updated partition assignment: [(row_events, 0)]
[2025-12-17, 17:12:21 UTC] [subprocess.Popen:254] INFO - Updated partition assignment: [(row_events, 0)]
[2025-12-17, 17:12:21 UTC] [cleaner.py:461] INFO - Stopping heartbeat thread for group cleaner_group
[2025-12-17, 17:12:21 UTC] [cleaner.py:961] INFO - Leverageup request for group cleaner_group returned successfully
[2025-12-17, 17:12:21 UTC] [cleaner.py:961] INFO - Leverageup request for group cleaner_group returned successfully
[2025-12-17, 17:12:21 UTC] [fetcher.py:789] INFO - Fetch to node 1 failed: Canceled: <brokerConnection client_id=kafka-python-2.3.8, node_id=0 host=kafka29892> connected [IPv4 ('172.24.0.4', 29892)]>
[2025-12-17, 17:12:21 UTC] [fetcher.py:958] INFO - Fetch to node 2 failed: Canceled: <brokerConnection client_id=kafka-python-2.3.8, node_id=1 host=kafka29892> connected [IPv4 ('172.24.0.4', 29892)]>
[2025-12-17, 17:12:21 UTC] [job2_cleaner.py:172] INFO - Cleaning Job Finished
[2025-12-17, 17:12:21 UTC] [taskinstance.py:1088] INFO - Marking task as SUCCESS, due_id=job2_clean_store_dag, task_id=clean_and_store, execution_date=20251217T172211, start_date=20251217T172211, end_date=20251217T172211
[2025-12-17, 17:12:21 UTC] [taskinstance.py:2776] INFO - 0 downstream tasks scheduled from follow-on schedule check
```

6.4 Job 3 Logs (Analytics)

Press shift + / for Shortcuts

job3_daily_summary_dag > 2025-12-17, 00:00:00 UTC / calculate_daily_metrics

Details Graph Gantt Code Logs

(by attempts) 1

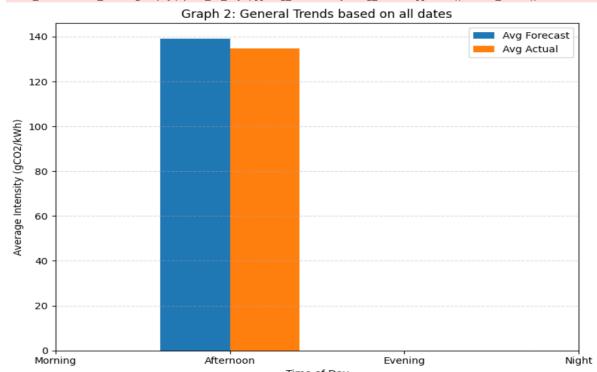
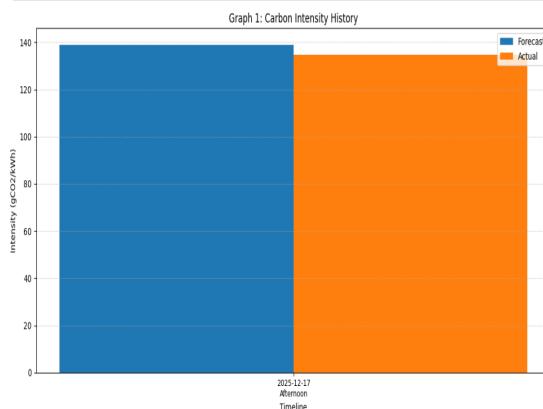
All Levels All File Sources Wrap Download See More

```
[79664619849] 
** Local files:
** + /opt/airflow/logs/dag_id=job3_daily_summary_dag/run_id=100/manual_2025-12-17T17:22:13.486640+00:00/task_calculate_daily_metrics/manual_2025-12-17T17:22:13.486640+00:00
[2025-12-17, 17:22:15 UTC] [taskinstance.py:1107] INFO - Dependencies all met for dep_context:metarequestable deps t1=tarikinstance: job3_daily_summary_dag.calculate_daily_metrics.manual_2025-12-17T17:22:13.486640+00:00
[2025-12-17, 17:22:15 UTC] [taskinstance.py:1359] INFO - Starting attempt of 1 for job3_daily_summary_dag.calculate_daily_metrics.manual_2025-12-17T17:22:13.486640+00:00
[2025-12-17, 17:22:15 UTC] [taskinstance.py:1359] INFO - Using operator PyxisOperator.
[2025-12-17, 17:22:15 UTC] [standard_task_runner.py:157] INFO - Started process 1558 to run task
[2025-12-17, 17:22:15 UTC] [standard_task_runner.py:157] INFO - Running "t1" for job3_daily_summary_dag.calculate_daily_metrics.manual_2025-12-17T17:22:13.486640+00:00
[2025-12-17, 17:22:15 UTC] [standard_task_runner.py:157] INFO - Subtask calculate_daily_metrics
[2025-12-17, 17:22:15 UTC] [logging_mixin.py:151] WARNING - The sol_alchemy_conn option in [core] has been moved to the sql_alchemy section of the configuration file. SolAlCHEMYConnWarning: The sol_alchemy_conn option in [core] has been moved to the sql_alchemy section of the configuration file.
[2025-12-17, 17:22:15 UTC] [logging_mixin.py:1688] INFO - Exporting env vars: AIRFLOW_CTX_DAG_ID='job3_daily_summary_dag' AIRFLOW_CTX_TASK_ID='calculate_daily_metrics' AIRFLOW_CTX_EXECUTION_ID='2025-12-17T17:22:13.486640+00:00' [running] on host 79664619849
[2025-12-17, 17:22:15 UTC] [job3_analytics.py:19] INFO - --- Starting Analytics Job ---
[2025-12-17, 17:22:15 UTC] [job3_analytics.py:52] INFO - --- SUCCESS: Written 1 rows to 'daily_summary'.
[2025-12-17, 17:22:15 UTC] [job3_analytics.py:52] INFO - summa_date part_of_day ... max_intensity min_intensity
@ 2025-12-17T17:22:15.334Z [1 rows x 8 columns]
[1 rows x 8 columns]
[2025-12-17, 17:22:15 UTC] [python.py:194] INFO - Done. Returned value was: None
[2025-12-17, 17:22:16 UTC] [taskinstance.py:1398] INFO - Marking task as SUCCESS, due_id=job3_daily_summary_dag, task_id=calculate_daily_metrics, execution_date=20251217T172213, start_date=20251217T172213, end_date=20251217T172213
[2025-12-17, 17:22:16 UTC] [local_task_job.py:226] INFO - Task ended with return code 0
[2025-12-17, 17:22:16 UTC] [taskinstance.py:2776] INFO - 0 downstream tasks scheduled from follow-on schedule check
```

6.5 Database Content & Visualization

```
In [4]: query_events = "SELECT * FROM events ORDER BY id DESC LIMIT 10"
df_events = pd.read_sql(query_events, conn)
print("Last 10 Recorded Events (Cleanned Data)")
display(df_events)

Last 10 Recorded Events (Cleanned Data)
   id ingestion_timestamp forecast_intensity actual_intensity index_intensity source created_at
0  80  2025-12-17T17:22:15.201899  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
1  79  2025-12-17T17:22:14.000700  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
2  78  2025-12-17T17:22:13.505457  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
3  77  2025-12-17T17:22:04.24334046  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
4  76  2025-12-17T17:20:13.199983  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
5  75  2025-12-17T17:19:40.334966  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
6  74  2025-12-17T17:19:09.158452  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
7  73  2025-12-17T17:18:38.321545  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
8  72  2025-12-17T17:18:07.356636  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
9  71  2025-12-17T17:17:36.220188  136      131  moderate carbonintensity.org.uk  2025-12-17T17:22:21
```



7. Conclusion

In this project, we built an end-to-end data pipeline that fully meets the project requirements. The system reliably collects, processes, and stores frequently updated data using Apache Airflow, Kafka, and SQLite. The pipeline is resilient to temporary failures and ensures stable data flow from ingestion to analytics. The final analytical results provide clear insights into carbon intensity patterns during different parts of the day, demonstrating a practical real-world data engineering solution.