Airflow ETL Pipeline for a Social or Environmental Dashboard

Phase 1 — Dataset Selection + Problem Justification

Task:

Students must choose **one public dataset** (medium-sized) OR generate synthetic data.
Examples:

- Air quality data
- Public transportation usage
- City noise levels
- Water consumption
- Wildlife sightings
- Renewable energy production
- World Bank datasets
- Government open data portals

Requirements:

Students must submit a **short justification (1 paragraph)** answering:

1. *Why this dataset matters socially, environmentally, or economically?*
2. *What problem could be improved by analyzing this data?*
3. *Who benefits from the insights (cities, communities, nature, health, etc.)?*

Phase 2 — Build an Airflow ETL Pipeline

Students must create a **fully working Airflow DAG** that includes:

Mandatory components:

▸ **Extract**

- Download the selected dataset OR
- Generate synthetic data using Python inside a task.

▸ **Transform**

- Cleaning (missing values, duplicates, type formatting).
- Aggregation or feature creation.
- Convert to a cleaned CSV or database table.

▸ **Load**

- Store the cleaned dataset into:
    - a local folder
    - a Postgres DB
    - or a cloud storage bucket (if allowed)

▸ **Scheduling**

- Add a schedule interval (daily, hourly, weekly, etc.).
- It can be real or simulated for class purposes.

▸ **Error Handling**

- Use at least one of:
    - `try/except` inside a PythonOperator
    - Airflow retry settings

    - task-level failure notifications (logging is enough)

▸ **Scaling Consideration**

Students must add one improvement like:

- Chunk processing
- Using efficient data formats (Parquet)
- Parallel tasks
- Filtering high-volume raw data before processing

Phase 3 — Dashboard Creation

Task:

Create a small dashboard using the **output of the Airflow ETL**.

Tools allowed:

- Power BI
- Looker Studio
- Tableau Public
- Python (Plotly/Matplotlib)
- Excel / Sheets (only if cleaned data is manageable)

Dashboard Requirements:

- At least **2 charts** and **1 KPI**
- Must use **only the cleaned/transformed data** from the ETL
- Must clearly support the justification stated in Phase 1

- Students must explain:
    - *Why these charts?*
    - *How does this dashboard help solve or understand the real-world problem?*

Final Output (Practical)

Students will present a **5–7 minute demo** showing:

1. Their Airflow DAG (running or visualized in the UI)
2. Their transformed dataset
3. Their dashboard
4. Their justification (verbally, not a document)