Weather Data Pipeline - GCP (Composer → BigQuery → Collab, Looker)

Objective

Fetch and visualize live weather data (temperature, humidity, conditions) for major Indian cities using the **Google Cloud Platform (GCP)** ecosystem.

This pipeline:

- Calls **OpenWeather API** hourly

- Orchestrates ingestion with **Cloud Composer (Airflow)**

- Stores curated records in **BigQuery** (partitioned & clustered for cost efficiency)

- Archives raw JSON to **Cloud Storage** (optional, for audit/backfills)

- Draw insights using **Google Collab**

- Visualizes insights in **Looker Studio**

Components

- 🌐 **OpenWeather API** → Source of live weather data (15 Indian cities)

- 🔑 **Secret Manager / Env Vars** → Secure storage of API keys

- ☁️ **Cloud Composer (Airflow)** → Orchestration of ingestion and checks

- 🗄️ **Cloud Storage (GCS)** → Raw JSON archive (`gs://weather-raw-archive/`)

- 📊 **BigQuery** → Analytical warehouse

  - Partitioned by `DATE(ts_utc)`

  - Clustered by `city`

- Unique key = (city, source_dt_utc)
- 📈 **Looker Studio** → Visualization dashboards

Architecture
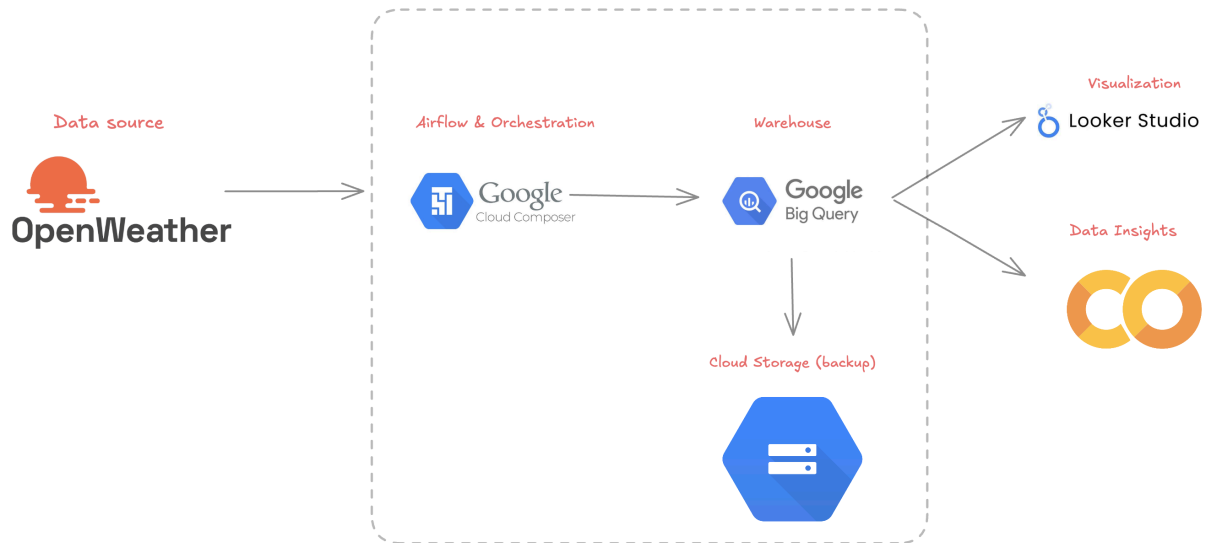
Setup Steps

1. OpenWeather API

- Sign up at https://openweathermap.org/api
- Generate and copy your **API key**

2. Cloud Composer

- Create a Composer environment in the **same region** as your BigQuery dataset
- Add environment variables:

```
OPENWEATHER_API_KEY = <your_api_key>
RAW_BUCKET = weather-raw-archive
```

3. BigQuery

- Create dataset:

  > airflow_bq_looker_project

- Table `openweather_15_cities_v2` will be created with partitioning + clustering by the DAG

4. Cloud Storage

- Create bucket:

  > weather-raw-archive

- Raw JSON payloads stored like:

  > gs://weather-raw-archive/openweather/YYYY/MM/DD/HH/payload.json

5. Deploy DAG

- Save DAG as `openweather_to_bq.py`
- Upload to Composer DAGs bucket:
- Trigger DAG in Airflow UI

6. Validate in BigQuery

Run test query:

```
SELECT city, temp, ts_utc
FROM `airflowbigqueryproject.airflow_bq_looker_project.openweather_15_cities_v2`
ORDER BY ts_utc DESC
LIMIT 20;
```

7. Looker Studio

- Connect Looker Studio to BigQuery

- Create dashboards for:
    - Temperature trends per city
    - Humidity & pressure comparisons
    - Latest snapshot across all cities

8. Google Collab

- Connect collab to BigQuery
- Generate insights

🔒 Reliability & Security

- **Secrets**: API key stored in Env Vars
- **Idempotency**: Stage + MERGE pattern prevents duplicates
- **Partitioning & Clustering**: Optimize query cost & performance
- **Monitoring**:
    - Row count check (≥ 80% of cities)
    - Temperature sanity check (−20°C ≤ temp ≤ 55°C)

🚀 Future Enhancements

- Add a dbt/python layer for transformations
- Enrich with other APIs (AQI, rainfall, alerts)