Carbon Emissions Batch Processing Pipeline - Final Portfolio

Name: Shivani Sinha
Student ID: 92124778

Abstract

The “Carbon Emissions Batch Processing Pipeline” is a serverless, fully-automated, batch data-processing system implemented using free-tier services on Google Cloud Platform (GCP). The goal is to simulate and analyze CO_2 emissions from different types of vehicles operating in various cities. The pipeline automates ingestion, processing, and visualization of synthetic ride data to track emissions trends and support sustainability insights.

Built entirely on cost-efficient GCP services—such as Cloud Storage, BigQuery, Pub/Sub, Workflows, and Looker Studio—the project showcases how to build a scalable and reproducible data solution suitable for portfolios and real-world applications.

Objectives

1. Build an end-to-end cloud-based batch data processing pipeline.
2. Use only GCP’s free-tier offerings.
3. Simulate realistic transportation data with timestamped ride-level emissions.
4. Automate ingestion and transformation of data without manual intervention.
5. Visualize trends and insights using the Looker Studio dashboard.

Methodology

1. Data is generated in CSV format simulating over 2,000 ride records.
2. Cloud Storage holds the raw data file (`rides.csv`).
3. Cloud Scheduler and Pub/Sub trigger the Cloud Workflows pipeline.
4. Cloud Workflows loads the file into BigQuery as a structured table.
5. BigQuery SQL View (`emissions_with_co2`) adds calculated `co2_grams` values.
6. Looker Studio visualizes insights such as emissions by city, type, and time.

Resources, Technologies, and Breakdown

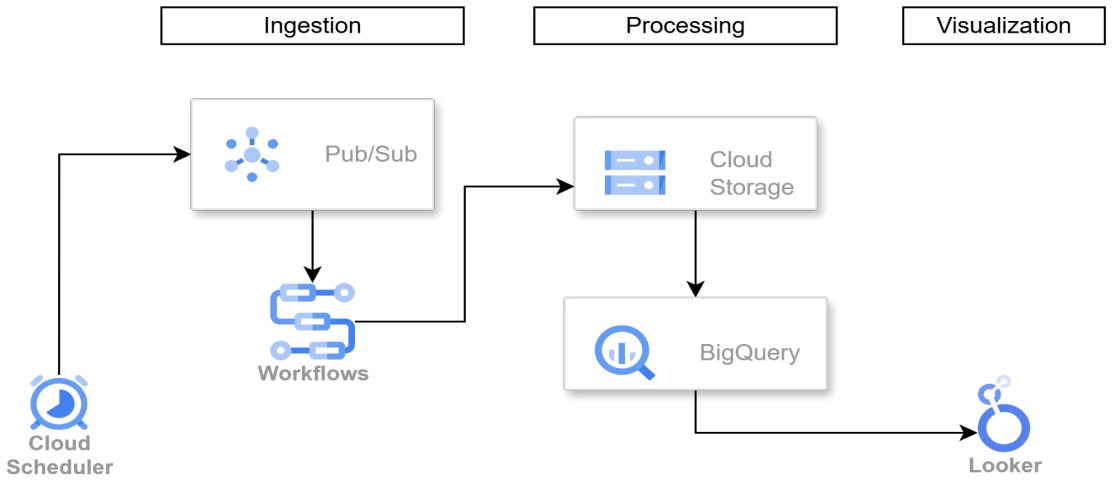
Data Storage : Google Cloud Storage
Data Warehouse : BigQuery
Automation : Cloud Workflows, Pub/Sub, Scheduler
Visualization : Looker Studio
File Format : CSV (rides.csv)
Version Control : GitHub
Language Used : SQL (for data transformation)

Implementation Summary

1. A synthetic dataset was designed with relevant fields: ride ID, vehicle type, timestamp, city, and distance.
2. A BigQuery table (`ride_emissions`) was created and populated via workflow triggers.
3. A separate SQL view (`emissions_with_co2`) performs CO_2 calculations based on `vehicle_type`.
4. A custom dashboard was created in Looker Studio with four main charts:

    a. Emissions by Vehicle Type
    b. Emissions by City
    c. Emissions trend by month

Design Architecture



Results Achieved

1. Successfully built and tested an automated batch data pipeline.
2. Verified data ingestion from GCS to BigQuery.
3. Designed a clean, informative dashboard that reflects live updates.
4. The project meets all objectives set out in the conception phase.

Reflection on Implementation

The project remained aligned with its original goals. While there were initial challenges—such as unsupported services in the free tier (e.g., Cloud Functions and PySpark)—they were effectively mitigated by redesigning the pipeline around Cloud Workflows.

Key Takeaways

1. GCP offers excellent free-tier resources for small-scale batch pipelines.
2. Designing workflows declaratively (YAML) enhances maintainability.
3. Using Looker Studio allows for quick dashboarding without coding.

4. Avoiding PySpark and Dataproc made the project more accessible and cost-efficient.

Conclusion

This project demonstrates how an effective, automated data pipeline can be developed using only free-tier cloud services. The simplicity and power of GCP's serverless ecosystem enabled seamless data processing, transformation, and visualization. The final product is reproducible, portable, and can be scaled up easily.

This batch-processing solution can serve as a practical template for similar sustainability or IoT data use cases in real-world settings.

Included Files for Submission

1. Abstract PDF (this document)
2. Phase 1 Concept Document
3. Phase 2 Reflection Document
4. Workflow YAML file
5. Sample CSV dataset (`rides.csv`)
6. BigQuery SQL View
7. Dashboard screenshots / link
8. GitHub link - GitHub Repo link
9. Walkthrough Video- Walkthrough Video Link