Data Engineering Assignment - Cloud Cost Intelligence

Submitted by: Bhuvanyu Geel

Date: November 2025

Overview

This repository contains the solution for the Cloud Cost Intelligence assignment. The solution covers data profiling, schema design, SQL transformations, pipeline architecture, and FinOps anomaly detection.

File Structure

- cloud_cost_assignment.ipynb: **(Main Submission)** A Jupyter Notebook containing the full analysis, Python code, SQL queries, and architectural diagrams.
- aws_line_items_12mo.csv: Provided dataset.
- gcp_billing_12mo.csv: Provided dataset.

How to Run

1. Ensure you have Python 3 and Jupyter installed.
2. Install dependencies:
   pip install pandas numpy matplotlib

3. Open cloud_cost_assignment.ipynb to view the analysis and run the code cells.

Quick Summary of Findings

Part A & B: Modeling

I utilized a **Star Schema** approach to unify the AWS and GCP datasets. This involves a central FACT_DAILY_CLOUD_SPEND table linked to dimensions for Service, Team, Environment, and Provider.

Part D: Pipeline

The proposed architecture is a **Daily Batch ELT** pipeline using **Airflow** for orchestration and **Snowflake + dbt** for transformation. This ensures handling of late-arriving billing data (up to 72 hours) and robust schema evolution.

Part E: FinOps Anomaly

- **Event:** A sudden cost spike in **Lambda (Data Team / Dev)** on Dec 28th.
- **Cause:** Likely a recursive lambda loop or an abandoned load test.
- **Action:** Implement **Automated Budget Actions** to throttle resources in the Dev environment if daily spend exceeds a threshold ($50).