Data Engineering Report

Project Name: Secure and Efficient Analysis of Global Health Data
Author: Sahil Panchal
E-portfolio: https://sahilpanchal83471.wixsite.com/insight
LinkedIn: https://www.linkedin.com/in/sahil-panchal-98735b191/
Date: 2024-11-05

Objective:

The goal of this project is to develop a robust and scalable data analytics solution to:

- Securely manage and filter a global health statistics dataset containing over 1 million records.
- Enable country-specific access to health ministers while ensuring data confidentiality.
- Provide tools for analyzing diseases without available treatments or vaccines.

Challenges Addressed:

- Large Dataset Size: Efficiently processing over 1 million records for secure sharing.
- Data Confidentiality: Restricting access to country-specific subsets of the dataset.
- Complex Analysis Requirements: Enabling meaningful insights into untreatable diseases.

Solution Overview

A data pipeline implemented using Apache Airflow and Google Cloud Platform (GCP) services ensures secure, scalable, and efficient data management:

- **Data Ingestion:** CSV file is uploaded to Google Cloud Storage (GCS) and ingested into Big Query.
- **Data Transformation:** Country-specific tables are created based on filtering criteria.
- **Reporting Views:** Aggregated views for each country focus on diseases without available treatments or vaccines.

Pipeline Components

1. **Data Ingestion:**
   - Source: Global health statistics file in GCS (e.g., global_health_data.csv).
   - Operator Used: GCSToBigQueryOperator for importing CSV data into a staging table in BigQuery.

2. **Data Transformation:**

   Country-Specific Tables:

   - Created in the transform_dataset.
   - SQL filters ensure only records relevant to each country are included.

   View Creation:

   - Views are stored in the reporting_dataset.
   - Focus on diseases without treatments or vaccines, with selected columns for analysis.

3. **Reporting and Security**

- Country-specific reporting views provide targeted insights for health ministers.
- Access is restricted to respective views, ensuring data confidentiality.

Workflow Details:

Tasks and Dependencies

1. **File Existence Check:**
   - Task: 'GCSObjectExistenceSensor' ensures the CSV file is available in GCS before proceeding.

2. **Data Load:**

- Task: GCSToBigQueryOperator ingests data into Big Query.
- Table: <your_project_id>.staging_dataset.global_data.

3. **Transformation:**

- Task: BigQueryInsertJobOperator creates country-specific tables.
- Example SQL Query:
  "CREATE OR REPLACE TABLE `<your_project_id>.transform_dataset.country_health_data` AS SELECT * FROM `<your_project_id>.staging_dataset.global_data` WHERE country = 'CountryName';"

4. **View Creation:**

- Task: BigQueryInsertJobOperator creates views with disease-specific filters.
- Example SQL Query:
  "CREATE OR REPLACE VIEW `<your_project_id>.reporting_dataset.country_view` AS SELECT Year, Disease_Name, Disease_Category, Prevalence_Rate, Incidence_Rate

FROM`<your_project_id>.transform_dataset.country_health_data` WHERE Availability_of_Vaccines_Treatment = FALSE;”

5. **Success Indicator:**

- Task: DummyOperator signals the successful completion of the pipeline.

Scalability and Optimization

- Dynamic Task Generation: Each country-specific table and view is handled independently, enabling parallelism.
- BigQuery Auto-scaling: BigQuery handles large dataset queries efficiently without additional infrastructure.
- Code Modularity: The DAG can be easily extended to include additional countries or data transformations.

Security Considerations

- Access Restrictions: BigQuery permissions restrict view access to authorized users.
- Service Account Permissions: Airflow’s service account requires appropriate roles for GCS and BigQuery.
- Data Confidentiality: Data segregation ensures that countries cannot access each other’s records.

Limitations and Future Enhancements

- Placeholder Values: Ensure all placeholders (<your_project_id>, <your_bucket_name>, etc.) are replaced with actual values.

- Error Handling: Add robust error-handling mechanisms for task failures.
- Monitoring and Alerts: Integrate logging and alerting for pipeline monitoring using Airflow's notification features.

Conclusion

- This data pipeline demonstrates a scalable and secure approach to processing and analyzing global health data. By leveraging GCP and Airflow, the solution ensures efficient data management and meets the stringent confidentiality requirements of global health research.