Data Engineering Test: Google Cloud Storage to PostgreSQL (4 hours)

Context

The assumption is a CSV file containing insurance claim data is stored in Google Cloud Storage. Your task is to extract this data, load it into PostgreSQL, perform some analyses, and store the results in PostgreSQL tables.

Input Data

The CSV file in Google Cloud Storage contains the following columns: IDpol, ClaimNb, Exposure, VehPower, VehAge, DrivAge, BonusMalus, VehBrand, VehGas, Area, Density, Region, ClaimAmount

Tasks

1. **Data Extraction and Loading**
    - Write a script (Python preferred, but any language is acceptable) to:
    - Extract the CSV data from Google Cloud Storage
    - Load the data into a PostgreSQL table named 'insurance_claims'
    - Ensure proper error handling and logging
2. **Query: Sum of Exposure by VehBrand and Area**
    - Write a SQL query in PostgreSQL to:
    - Calculate the sum of Exposure for each unique combination of VehBrand and Area
    - Store the results in a new table named 'exposure_summary'
3. **Query: Min and Max Density by Area**
    - Write a SQL query in PostgreSQL to:
    - Calculate the minimum and maximum Density for each Area
    - Store the results in a new table named 'density_summary'

Deliverables

1. The script for extracting data from GCS and loading it into PostgreSQL
2. SQL queries for tasks 2 and 3, including table creation statements
3. SQL commands for indexing and access control

4. A brief explanation of your approach, including error handling, performance optimization, and security considerations

Evaluation Criteria

1. **Data Extraction and Loading**
    - Correct extraction of data from GCS
    - Efficient and error-free loading into PostgreSQL
2. **SQL Query Efficiency and Accuracy**
    - Correct results for the required calculations
    - Efficient query design
3. **Common Security** in code
    - Understanding of performance optimization and security principles
4. **Code Quality and Documentation**
    - Clear, well-commented code (both script and SQL)
    - Comprehensive yet concise explanation of approach
5. **Error Handling and Data Integrity**
    - Robust error handling in the extraction and loading process
    - Consideration of data quality issues in SQL queries

(assuming script code use in production environment, provide the best practise based on your skill and experience)

If time permits, describe what should be the next step