DATA ENGINEER

Take-home Assignment

I. OVERVIEW

This take-home assignment is designed to evaluate your proficiency in data engineering tasks. It will test your skills in using Python, Apache Airflow, PostgreSQL, and ClickHouse to handle data ingestion, transformation, storage, transfer, and querying.

II. DETAILS

Objective:

Evaluate proficiency in data engineering tasks involving Python, Apache Airflow, PostgreSQL, and ClickHouse.

Time Required:

3 hours

Technology Stack:

- Python
- Apache Airflow
- PostgreSQL
- ClickHouse

Tasks:

Data Ingestion with Python

- Download a dataset from UCI Machine Learning Repository.
- Import it into a Python environment.

Data Transformation with Python

- Use Pandas to clean and transform the data.
- Generate summary statistics for three key variables.

Database Storage with PostgreSQL

- Create a PostgreSQL database and table.
- Store the cleaned dataset.

Data Transfer with Airflow

- Write an Apache Airflow DAG to transfer data from PostgreSQL to ClickHouse.
- Schedule the DAG to run every hour.

Data Querying with ClickHouse

- Write ClickHouse SQL queries to answer:
    1. How many unique values are in variable X?
    2. What is the average of variable Y grouped by variable Z?

Deliverables:

- Python script(s) for tasks 1 and 2.
- SQL schema and queries for task 3.
- Apache Airflow DAG for task 4.
- ClickHouse SQL queries for task 5.
- A README file detailing execution steps.

Evaluation Criteria:

- Code Quality: Readability, structure, modularity
- Data Transformation: Efficacy, efficiency
- Database Structure: Schema design, indexing
- Data Transfer: Airflow DAG functionality, scheduling
- Data Querying: Accuracy, efficiency in ClickHouse

III. SUBMISSION

- Create a Git repository for your project.
- Include a README with instructions on how to run the app and any additional information you'd like to provide.
- Share the Git repository link with us for evaluation.
- The deadline for submission is **48 hours**.
- You need to upload your assignment to Google Drive, and then **send a public link to the Google Drive file** to hr@gethomebase.com to submit your take-home assignment. Please make sure the **SHARING SETTING** is **PUBLIC**.