GSynergy Data Engineer Interview Challenge

Introduction

Thank you for your interest in advancing your career at GSynergy. We look forward to a productive and insightful interview process that informs both you and us if we are going to be a good fit for helping us deliver reliable and scalable data engineering pipelines to our clients, faster and cheaper than the competition.

We are a fast paced, high productivity team, where all our developers have high level of autonomy and ownership. We therefore require that all our teammates have strong development and communication skills.

The following challenge is meant to assess your abilities in

1. Inferring data models from loosely defined and presented data, and detecting and fixing data quality issues in raw data
2. Designing, developing and testing ELT pipelines using integration services from the major cloud services providers, i.e., Azure, AWS or GCP.
3. Communicating and presenting technical information succinctly and effectively.

If you have the appropriate level of proficiency, the challenge should take you between 3 and 5 hours of effort.

<u>General Rules and Guidance</u>

1. Do not seek assistance from anyone else, or plagiarize solutions from the web.
2. Structure and write your code as if you are writing it for production. Create appropriate branches and commit code as you would for real-world projects.
3. You may use Azure (preferred), AWS or GCP as your target cloud.
4. For your ELT development, you may use technologies / services available from these cloud providers. For example, you may use Azure Integration Services or Azure Fabric, but <u>not</u> Airflow, Pentaho, or Informatica.
5. For your target data warehouse technologies you may use Microsoft Fabric Data Warehouse, Databricks, Snowflake, or Redshift.
6. You may use dbt for writing your transformations.
7. Test your work thoroughly and ensure your submission is working before submitting it for review. You will not get a second chance.
8. Follow this guidance carefully, and submit as instructed. We may not reach out to you if you miss any steps.

9. Please reach out to us at careers@gsynergy.com if something requires clarification. We may take up to 24 hours to respond.

How to submit your solution

1. If writing code, create a public Git repository for all your code and design assets. Ensure that you have a well drafted readme.md file at the root level instructing us how to run and validate your code. We recommend Github, Azure Repos, and Azure Devops.
2. If using only the cloud console / UI for all your work, ensure you keep your work around for demonstration during in person (video call) interview.
3. Create a brief (2 – 5 mins) screen recording video demonstrating your solution and approach. Please use your own voice to describe your work. Upload this to a shared folder on Google Drive, Dropbox or a similar file sharing services, or to a video sharing service such as YouTube.
4. Send an email to careers@gsynergy.com with links to your Git repository and the video.

The Challenge: Data Warehouse and Data Pipeline

Challenge assets are available at:
https://drive.google.com/drive/folders/1TJeOLemuvVHLUf3O3wMkYcjeaVuwBwk4?usp=drive_link

The above includes a data folder that has several pipe delimited gzipped files of raw data. The names of the files start with either *hier* or *fact* to signify whether they have hierarchy (dimension) or fact data. The word following *hier* or *fact* indicates the table name for the raw data. Each file has a header row with column names.

The hier files have id and label columns for each level in the hierarchy. For the most part you can assume that the left most column is the primary key, but you should ensure that you draw out a proper structure by looking at the many-to-one relationships that the data manifests.

1. You must draw out an ER diagram showing raw table structure and any relationships between them that you can infer using column names. You may use schema inference tools, but you must document what you used and why. You must add the final ER diagram and any documentation explaining it to your submission's Github repository.
2. You must build a pipeline that
    a. Loads this raw data into the data warehouse from external storage such as Azure Blobs, AWS S3 or the like. You must write basic checks such as non-null, uniqueness of primary key, data types. Also check for foreign key constraints between fact and dimension tables. Do it for at least one hier (dimension), and one fact table.

b. Create a staging schema where the hierarchy table has been normalized into a table for each level and the staged fact table has foreign key relationships with those tables.

c. Create a refined table called *mview_weekly_sales* which totals *sales_units*, *sales_dollars*, and *discount_dollars* by *pos_site_id*, *sku_id*, *fsclwk_id*, *price_substate_id* and *type*.

d. BONUS: write transformation logic that will incrementally calculate all the totals in the above table for partially loaded data.

Evaluation Approach / Criteria

1. We will start by reviewing your screen video demonstrating your work. Please be sure to demonstrate how your solution meets all requirements. Please ensure that you speak clearly and audibly, and that your screen recording shows your entire screen. Showing the speaker in a small inset (pip) is preferred, but not required. There are many free screen recording tools available; you may use any as long as the video can be played in the browser using the link you provided. Please note that we will not download the video, or open any attachments to emails. You must upload the videos in ways that they will play in the browser itself without any need to downloads.
2. If we are satisfied with your presentation in the video, we will examine the code and assets in your Github repo and validate it’s working.