Glide Data Engineering Exercise

Overview

Welcome to the interactive exercise phase of the interview. We believe in evaluating candidates based on real problems they might encounter in this role along with providing you the opportunity to experience a scenario a Data Engineer would encounter at our company. The below interactive exercise is designed in that spirit.

Because none of us works in isolation, feel free to reach out to Nico (nicolas@glide.com) with any questions you may have. Because we know you are busy, we are aware that your questions may come at odd hours and have committed to being available to you. For timing, please let us know if you believe it will take longer than a few days to complete this exercise.

Scenario

As a data engineer you are tasked with ingesting data from our Human Resources system, transforming the data and storing it in our data warehouse. Data is ingested in batches and our batches come daily. Each batch is an export of all the employees from the Human Resources system as it existed on the previous day.

Description

You should have access to *“employee_data.zip”*. This contains 10 CSV files simulating the batch data from the Human Resources system from Jan 1, 2020 to Jan 10, 2020.

- Store the daily batches or snapshots in a Staging table. To keep it simple you can store this in a CSV file. Let’s call this the ***employee_snapshot*** table (you can name it to suit your needs).
- You will create a new table to model the slowly changing dimension, employee. To keep it simple, you can store this in a CSV file. Let’s call this the ***employee*** table (you can name it to suit your needs).

Deliverables

- The output of the exercise is to populate both the *employee_snapshot* and *employee* tables with the files in the *“employee_data.zip”*. These can be CSV files.
- All files (deliverables and scripts) should be uploaded to a private repo on Github. Please invite user **ngonik** when done.
- We prefer you use Python and PySpark to do the transformations. But you are welcome to use your language of choice.

Presentation

- You will present your deliverables and explain your choices of tool, algorithm and feedback on how to better do it, if you had more time.

As you work through the exercise:

- Brainstorm / use us as a sounding board.
- Get as much done as you can.
- If the process is not working for you, give us a shout out via email.

The format of the plan is whatever you decide, gdoc, slides, etc. At the end of the week, you will be sharing this presentation with us so we can take a look at it. The exercise usually takes a few days but we want to be respectful of your commitments, so let us know if you need more time.