Data Engineer Interview Test

Wisr is looking for a high quality data engineer which can deliver comprehensive solutions for our continuity and business growth.

You can be part of an amazing team which deals with data all the time using different processes, tools and technologies.

Following is a little challenge for those keen on joining this amazing company and team.

The Project

Build a small ETL process to ingest a few set of files into a data warehouse like project.

We are expecting an end-to-end ETL solution to deliver a simple star schema which an end user can easily slice and dice the data through a report or using basic ad-hoc queries.

Tools and Technologies

We are a Python and SQL workshop, we would like to see this project using just those tools.

However, we are open to other tools and technologies if we are able to easily replicate on our side.

For the database, use a simple and light engine like SQLite, MySQL or Postgres. If you have to use a licensed product, please choose MS SQL as we only use this at Wisr.

How to do it?

Instructions

* There are two sections
    1. ETL Exercise
    2. Data Reporting Exercise
* Complete as many steps as you can in each sections, with the best of your efforts
* Please include step by step instructions to run the code OR ideally build your solution using Jupyter Notebook
* Use the best practices in your Coding Style
* Be able to explain from the ground up the whole process in face to face interview

* Use only Python and SQL in your code

* Bonus points for writing Unit Tests

ETL Exercise

---

1. The data for this exercise can be found in the `data.zip` file. Can you describe the file format?

Super Bonus: The encoded file `bonus_etl_data_gen.txt` has instructions to generate your own data.

Follow those instructions. And, to get the bonus points, please encode the file with the same method that was used to generate the file.

2. Code your scripts to load the data into a database.

3. Design a star schema model which the data should flow.

4. Build your process to load the data into the star schema

Bonus point:

- add a field to classify the customer account balance in 3 groups

- add revenue per line item

- convert the dates to be distributed over the last 2 years

5. How to schedule this process to run multiple times per day?

Bonus: What to do if the data arrives in random order and times via streaming?

6. How to deploy this code?

Bonus: Can you make it to run on a container like process (Docker)?

Data Reporting

One of the most important aspects to build a DWH is to deliver insights to end-users.

Can you, using the designed star schema (or if you prefer the raw data), generate SQL statements to answer the following questions:

1. Which Market Segment has the largest customer base?
2. Which Nation has the lowest customer base?
3. What are the top 5 nations in terms of revenue?
4. From the top 5 nations, what is the most common shipping mode?
5. What are the top selling months?
6. Who are the top customer in terms of revenue and/or quantity?
7. Compare the sales revenue of a product in most recent current period against previous period?

ERD

--