Data Engineer Case Interview

Case Study Test

RESTRICTED: DO NOT CIRCULATE

Data Engineer Case Interview

Instructions for completing the case study

1. This case study is divided into **5** sessions, aiming to assess different areas of data engineering skills
2. The entire case is to be completed within 48 hours from the time you receive it.

Please submit the following within the allocated time:

1. A deck of slides for each section, preferably in *.pptx or *.pdf form.
2. The complete code used in your analysis, such that we should be able to run the code ourselves with appropriate languages / tools and reach conclusions comparable to what you include in your deck.

Instructions for interview

You will meet two or more current members of the data team for around an hour, where you will be invited to present your work **as though they were scenario-specific stakeholders**. In addition, you may be asked questions relating to your work for all sections.

You will be assessed on the quality of your delivery, as well as the validity of your answers.

Remarks

Please make assumptions if necessary when you encounter ambiguity while attempting the cases.

You are encouraged to implement data engineering best practices beyond the instructions.

You are advised to allocate more time to Section 1-4 than Section 5.

Section 1: Data Pipelines

The objective of this section is to design and implement a solution to process data files on a regular interval (e.g. daily).

Assume that there are 2 data files dataset1.csv and dataset2.csv, design a solution to process both files, along with the scheduling component.

The expected output of the processing task is a CSV file including a header containing the field names.

You can use common scheduling solutions such as cron or airflow to implement the scheduling component. You may assume that the data file will be available at 1am every day. Please provide documentation (a markdown file will help) to explain your solution.

Processing tasks:

- Split the name field into first_name, and last_name
- Remove any zeros prepended to the price field
- Delete any rows which do not have a name
- Create a new field named above_100, which is true if the price is strictly greater than 100

Note: please submit the processed dataset too.

Section 2: Databases

You are appointed by a car dealership to create their database infrastructure. There is only one store. In each business day, cars are being sold by a team of salespersons. Each transaction would contain information on the date and time of transaction, customer transacted with, and the car that was sold.

The following are known:

- Each car can only be sold by one salesperson.
- There are multiple manufacturers' cars sold.
- Each car has the following characteristics:
    - Manufacturer
    - Model name
    - Serial number
    - Weight
    - Price

Each sale transaction contains the following information:

- Customer Name
- Customer Phone
- Salesperson
- Characteristics of car sold

Set up a PostgreSQL database using the base docker image here given the above. Please provide a Dockerfile which will stand up your database with the DDL statements to create the necessary tables. Produce entity-relationship diagrams as necessary to illustrate your design.

Your team also needs you to query some information from the database that you have designed. You are tasked to write a SQL statement for each of the following task:

1) Output the list of customers and their spending.

2) Output the top 3 car manufacturers that customers bought by sales (quantity) and the sales number for it in the current month.

Section 3: System Design

You are designing data infrastructure on the cloud for a company whose main business is in processing images.

The company has a web application which collects images uploaded by customers. The company also has a separate web application which provides a stream of images using a Kafka stream. The company's software engineers have already written the code to process the images. The company would like to save processed images for a minimum of 7 days for archival purposes. Ideally, the company would also want to be able to have some Business Intelligence (BI) on key statistics including number and type of images processed, and by which customers.

Produce a system architecture diagram (e.g. Visio, PowerPoint) using any of the commercial cloud providers' ecosystem to explain your design. Please also indicate clearly if you have made any assumptions at any point.

Section 4: Charts and APIs

Your team decided to design a dashboard to display the statistic of COVID19 cases. You are tasked to display one of the components of the dashboard which is to display a visualisation representation of number of COVID19 cases in Singapore over time.

Your team decided to use the public data from https://covid19api.com/.

Display a graph to show the number cases in Singapore over time with reference to the APIs document from https://documenter.getpostman.com/view/10808728/SzS8rjbc#b07f97ba-24f4-4ebe-ad71-97fa35f3b683.

Section 5: Machine Learning

Using the dataset from https://archive.ics.uci.edu/ml/datasets/Car+Evaluation, create a machine learning model to predict the buying price given the following parameters:

- Maintenance = High
- Number of doors = 4
- Lug Boot Size = Big
- Safety = High
- Class Value = Good

(End of the Data Engineer Case Interview)