Data Scientist Technical Test

Task 1

MST Bank helps its coders in making virtual payments. Our bank records all transactions in the table Transaction, you must find out the current balance of all users and check whether they have breached their credit limit (If their current credit is less than 0).
Expected result table:

```
+------------+------------+------------+----------------------+
| user_id | user_name | credit | credit_limit_breached |
+------------+------------+------------+----------------------+
| 1 | Moustafa | -100 | Yes |
| 2 | Jonathan | 500 | No |
| 3 | Winston | 9900 | No |
| 4 | Luis | 800 | No |
+------------+------------+------------+----------------------+
```

Download dataset (db2.sql)

Notes:

- User with id (paid_by) transfer money to user with id (paid_to).
- Return the result table in any order.
- Column credit on expected result table is current balance after performing transactions.
- Column credit_limit_breached on expected result table is check credit_limit ("Yes" or "No")
- Attach your sql query file.
- Attach screen shot the result and query.

Task 2

You must deal with a complex matrix script. This script is essentially a grid of strings containing alphanumeric characters, spaces, and symbols. Your task is to decode this script by reading each column and selecting only the alphanumeric characters. The decoding process involves reading from the top to the bottom of each column, starting from the leftmost column.

If there are symbols or spaces between two alphanumeric characters in the decoded script, you must replace them with a single space to enhance readability. The rule **is not using** 'if' conditions for the decoding process.

The input consists of the dimensions of the matrix (rows and columns) followed by the elements of the matrix script. The output should be the decoded matrix script.

Give this matrix for the input:

```
7 3
My
Sa!
T-!
$j
#a-
jy_
aa-
```

Notes:

- Attach your python file
- Attach screen shot the result and the code

Task 3

Craigslist, a widely used platform for local classified ads, features nine main sections including jobs, housing, services, and more. Each section contains various categories, such as automotive or household services.

In this task, data from four sections for-sale, housing, community, and servicesis provided for sixteen different cities. Within these sections, sixteen specific categories have been selected.

Your goal is to predict the category of a Craigslist post based on the given city, section, and heading of the post. This involves text classification, where you analyze the text to determine its category.

You'll receive input in the form of JSON objects, with fields for city, section, and heading. Your task is to output the predicted category for each post.

To train your model, you have access to approximately 20,000 records, with each record including the category of the post. The training data is formatted similarly to the input data but includes the category field.

You can use the provided training file and sample tests to develop and evaluate your classification model. Build a simple UI in streamlit to display the input and the result.

Notes:

- Attach your folder of python code
- Attach screen shot the UI
- Attach link of your UI

Task 4

You are a Data Scientist at MST. One day, your company's SQL Server infrastructure reaches its limitations in handling large volumes of data and complex data structures. As a result, the management has decided to transition to a more scalable and flexible data storage solution.

As part of this transition, you are tasked with designing and implementing a Data Lake architecture to accommodate the company's growing data needs. Your role involves understanding the differences between Data Lake and Data Warehouse concepts and leveraging this knowledge to make informed decisions.

Explain how you would approach the implementation for MST in response to the limitations of SQL Server.

Good luck!