### FETCH REWARDS DATA ANALYST ASSESSMENT

### First: Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

### This data model has been designed on Toad Data Modeler and I am using SQL Server dialect for the SQL queries derived from this model.

### 

### Second: Write a query that directly answers a predetermined question from a business stakeholder

* What are the top 5 brands by receipts scanned for most recent month?

SELECT TOP 5 B.NAME AS BRAND\_NAME, COUNT(B.ID) AS COUNT\_BRAND

FROM BRANDS B

JOIN CPG C

ON B.ID = C.ID

JOIN ITEMLIST I

ON C.CPG\_ID = I.REWARDSPRODUCTPARTNERID

JOIN RECEIPTS\_NEW R

ON I.ID = R.ID

WHERE (CASE WHEN ISNUMERIC(R.DATESCANNED) = 1 THEN CAST(R.DATESCANNED AS FLOAT) ELSE NULL END) / 1000 >

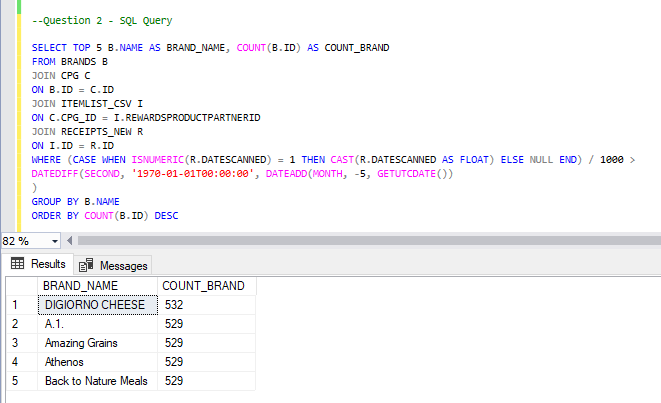
DATEDIFF(SECOND, '1970-01-01T00:00:00', DATEADD(MONTH, -5, GETUTCDATE())

)

GROUP BY B.NAME

ORDER BY COUNT(B.ID) DESC;

**Output:**



### Third: Evaluate Data Quality Issues in the Data Provided

Link to code:

<https://github.com/keertimurugan/FetchRewards/blob/main/Question3%20-%20Data%20Quality%20Issues.ipynb>

**Fourth: Communicate with Stakeholders**

Hello,

I am reaching out to share some of my findings with regards to the Fetch Rewards Data Analyst Assessment.

Here are the steps that I took while exploring the data and looking for any data quality issues:

1. **Missing Values**: I started with finding missing values and what that could mean. With the user data, the lastLogin column has missing values which could mean that there is no record of the user logging in to the app in the past.
2. **Duplicate records**: I looked for duplicate records in the dataset and all the datasets have duplicate records which can be dropped to avoid incorrect analysis in the future.
3. To avoid data quality issues, we can take further steps to make sure that the data coming in from the source system is clean to a certain extent.

I would also like to clarify some of my questions regarding the data provided.

1. What are the key metrics that we want to measure?
2. How large can the data get with time and how are we planning to store historical data?
3. What is the source of this dataset and is there any mechanism to clean the source data?

As our data continues to grow, these some of the key points to take into consideration:

1. The size of the data we are dealing with.
2. Understanding the most important KPIs and making sure to construct the data warehouse accordingly for better query performance.
3. Setting up data partitions for a scalable data warehouse.

Please let me know if you would like to discuss this over screen share.

Regards,

Keerti M