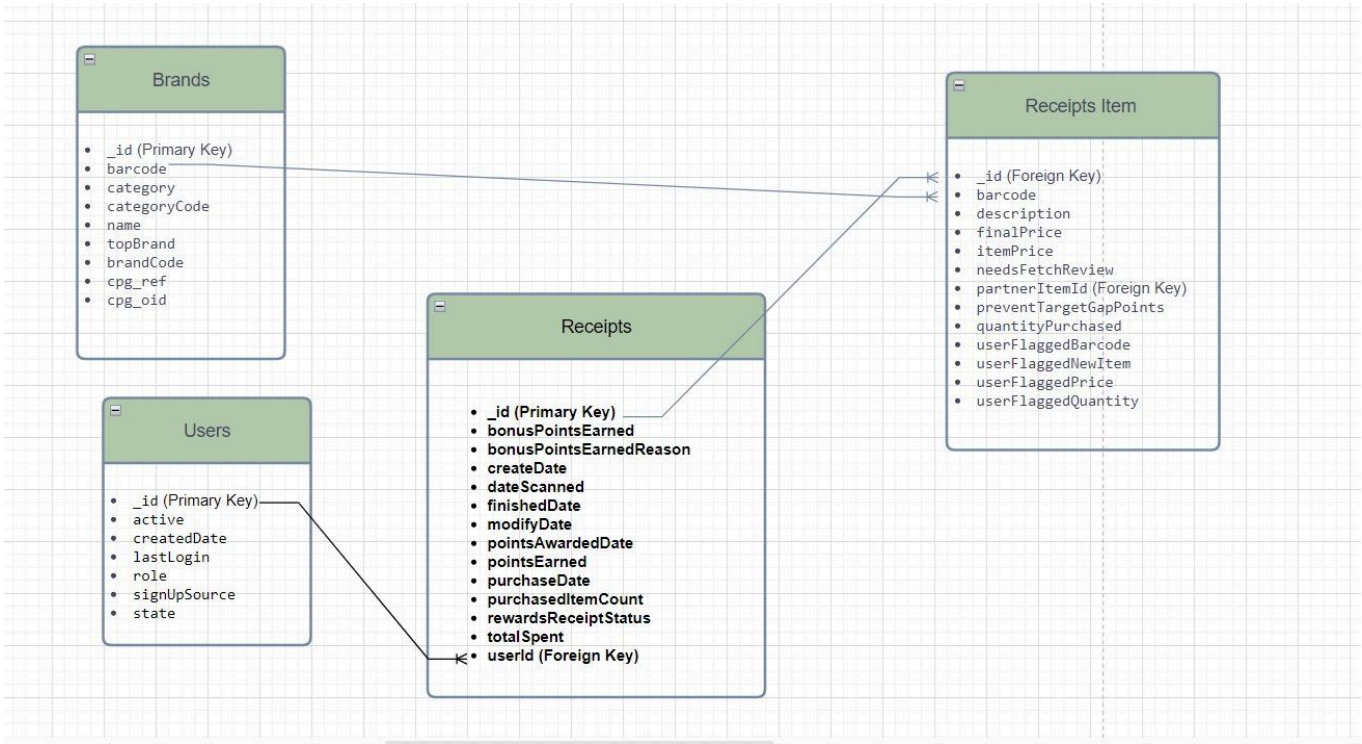
Fetch Assessment

Q1.Review Existing Unstructured Data and Diagram a New Structured Relational Data Model



Q2. Write queries that directly answer predetermined questions from a business stakeholder

I used PostgreSql.

- When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

```
86  SELECT
87      rewardsReceiptStatus,
88      AVG(totalSpent) AS avg_total_spent
89  FROM
90      receipts
91  WHERE
92      rewardsReceiptStatus IN ('FINISHED', 'REJECTED')
93  GROUP BY
94      rewardsReceiptStatus
95  ORDER BY
96      avg_total_spent DESC
97
```

Data Output | Messages | Notifications

	rewardsreceiptstatus character varying (255)	avg_total_spent double precision
1	FINISHED	80.85430501930502
2	REJECTED	23.326056338028184

- When considering *total number of items purchased* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

```
 99  SELECT
100      r.rewardsReceiptStatus,
101      SUM(ri.quantityPurchased) AS total_items_purchased
102  FROM
103      receipts r
104  JOIN
105      receipts_item ri
106  ON
107      r._id = ri._id
108  WHERE
109      r.rewardsReceiptStatus IN ('FINISHED', 'REJECTED')
110  GROUP BY
111      r.rewardsReceiptStatus
112  ORDER BY
113      total_items_purchased DESC;
114
```

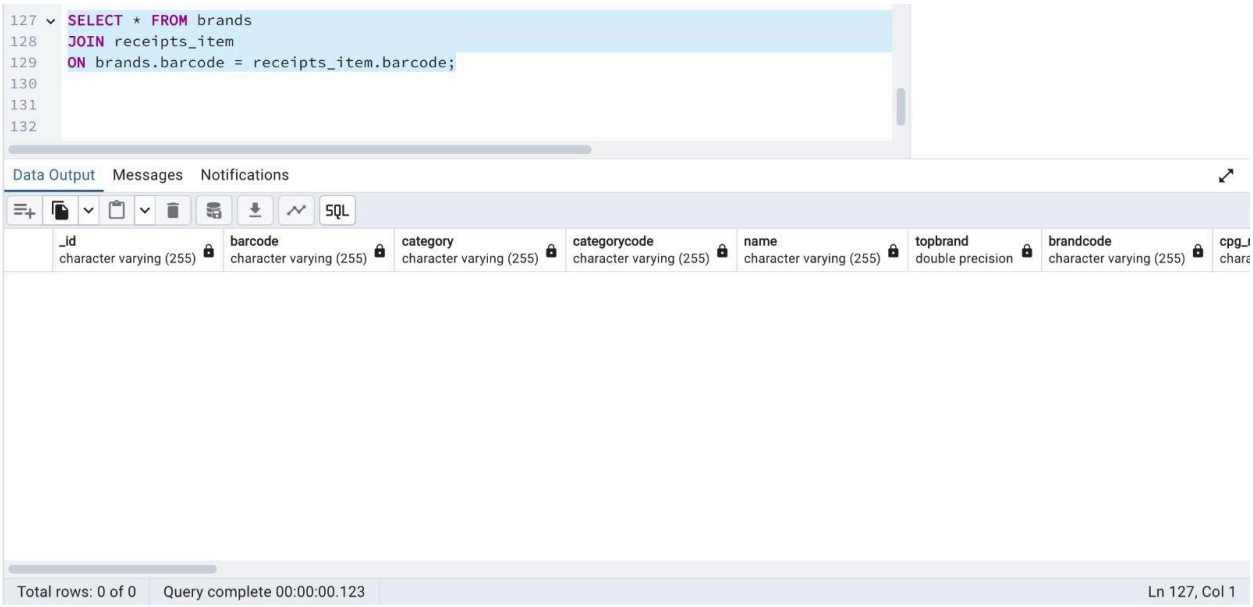
Data Output | Messages | Notifications

	rewardsreceiptstatus character varying (255)	total_items_purchased double precision
1	FINISHED	8176
2	REJECTED	141

Q3. Data Quality Issues

- The presence of duplicate values in the `_id` attribute of the `Users` table represents a significant data quality issue. This attribute is intended to serve as the primary key (PK) for the table, and primary keys must contain unique values to ensure the integrity and functionality of the database.
- There are missing values in the Category, CategoryCode, topBrand,BrandCode, and cpg_ref attributes of the `Brands` table.

- There are missing values in the lastLogin, signUpSource and state fields of the Users table
- There are missing values in bonusPointsEarned, finishedDate, and other fields of the Receipts table.
- In the Brands table, there are inconsistent values for the BrandCode attribute. There are duplicate entries, scientific notations, missing values, miscellaneous data etc.
- The barcode values in the receipts table do not match those in the brand table making it impossible to join these tables, pointing to a fundamental data quality issue related to data consistency and data integration.
- There are null values and “item not found” in the description attribute of the receipts table.



Q4.Communicate with Stakeholders

Hi,

I hope this message finds you well. As I perform my ongoing analysis of receipt, user, and brand datasets, several critical insights and challenges have surfaced that require your attention and expertise to address effectively.

1. Questions About the Data:

As part of my ongoing analysis of receipt, user, and brand data, I have identified a few areas where I need additional context and clarification, particularly regarding the ‘rewardsReceiptItemList’ field in the receipts dataset.The rewardsReceiptItemList field is critical for understanding the items purchased by users, as it contains detailed information about each item on a receipt. However, I have noticed that this field comprises numerous attributes that are not fully defined in our documentation. Additionally, there are inconsistencies and missing

values within these nested data fields.Could you provide a comprehensive definition and description for each attribute within the rewardsReceiptItemList? This will help me accurately interpret the data and ensure I am analyzing it correctly.

Additionally, are there any known issues with the data ingestion process that could affect data quality?

Could you please specify the priority data fields that should be focused on for deriving business insights? This includes fields from our receipts, users, and brands datasets that are most relevant for decision-making.

2.How did you discover the data quality issues?

While importing the receipts , user and brand JSON file, I encountered formatting errors that prevented a smooth load into the analysis environment. Upon closer inspection using Python and Pandas, and performing exploratory data analysis, I noticed inconsistencies and missing values within various fields in all the three tables.I also observed duplicate entries within the datasets (for example: 'Id' in the users table which is supposed to be a Primary key and should contain unique values has a lot of duplicates) ,which can lead to skewed analysis results and incorrect insights.These issues significantly affect our ability to perform reliable data analysis. Accurate data is essential for deriving meaningful insights.

3. Resolving Data Quality Issues

Resolving these issues is crucial for the success of the analysis and subsequent business strategies. To address these issues, we need access to any existing documentation that outlines the expected data format and validation rules.

Additionally, collaboration with the data engineering team to review and possibly adjust the data ingestion pipeline would be beneficial. Moreover, implementing a comprehensive data cleaning process to handle formatting errors, inconsistencies, and duplicates is also necessary.Your support in facilitating access to the necessary resources and expertise will significantly enhance the accuracy and effectiveness of the data analysis efforts. Thank you for your assistance.

4. What other information would you need to help you optimize the data assets you're trying to create?

As I focus on optimizing our data assets from the receipts, users, and brands datasets, I am requesting additional information crucial for maximizing the business insights and decision-making capabilities derived from these resources.

- Could you provide further details regarding the CPG collection referenced in the brands dataset (cpg field)? Understanding its structure and content will help me integrate this data seamlessly and leverage it effectively for market analysis and brand relationships.
- Are there specific metrics or benchmarks related to brand performance that I should consider? This could encompass market share trends, customer loyalty indicators, or effectiveness of promotional activities.
- Beyond last login frequency (lastLogin) and active status (active), are there other user engagement metrics or behavioral data points that are pivotal for our analyses? Clear insights into these factors will facilitate more refined user segmentation and targeted marketing strategies.
- Detailed data dictionaries for each dataset would be extremely helpful.

5. What performance and scaling concerns do you anticipate in production and how do you plan to address them?

Data Volume: With increasing data volume, we need to ensure our data processing pipelines are optimized for speed and efficiency. We plan to implement batch processing and leverage parallel computing where possible.
Real-Time Processing: If there is a requirement for real-time or near-real-time analytics, we will need to consider stream processing solutions and possibly scalable cloud services.
Data Storage: Ensuring that our data storage solutions are scalable and efficient will be critical. We may explore cloud storage options like AWS S3 or Google Cloud Storage to handle large datasets.

It would be great if we could discuss these considerations further to align on the importance of optimizing our data infrastructure. Your insights and guidance in this area will be beneficial as we prepare our data environment to support future business goals effectively.

Thank you for your attention to this matter. Looking forward to your feedback and support.

Best regards,
Charu Mangla