Diagram a New Structured Relational Data Model

Receipts	Users	Brands
_id	_id	_id
bonusPointsEarned	state	barcode
bonusPointsEarnedReason	createdDate	brandCode
createDate	lastLogin	category
dateScanned	role	categoryCode
finishedDate	active	cpg
modifyDate		topBrand
pointsAwardedDate		name
pointsEarned		
purchaseDate		
purchasedItemCount		
rewardsReceiptItemList		
rewardsReceiptStatus		
totalSpent		
userId		

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 b.name AS brand_name, COUNT(r._id) AS receipts_count
FROM brands b
JOIN rewards_receipt_item_list r ON b._id = r.brandId
WHERE EXTRACT(MONTH FROM r.dateScanned) = EXTRACT(MONTH FROM CURRENT_DATE)
GROUP BY b.name
ORDER BY receipts_count DESC
LIMIT 5;
```

The query establishes connections between the 'brands' and 'rewards_receipt_item_list' tables. It then filters the data to include only entries from the most recent month, groups the results based on brand names, calculates the count of receipts for each brand, and subsequently sorts the outcome in descending order. Finally, the query restricts the result set to the top 5 brands by receipts count.

Evaluate Data Quality Issues in the Data Provided

As part of our ongoing efforts to ensure data quality and integrity, I have conducted an analysis on the provided data. I utilized Python to identify and evaluate missing values within the dataset. Please find attached the Python Notebook named 'Data_Quality_Issue.ipynb' for a comprehensive view of the analysis.

Communicate with Stakeholders

I would like to share some preliminary insights from our recent data analysis endeavors. As we delve into the dataset, several observations and queries have emerged, which are crucial for ensuring the accuracy and dependability of our data.

Questions Regarding Data Quality:

1. Incomplete or Missing Data: Instances have been identified where certain data points are either missing or incomplete, raising concerns about the dataset's reliability and its potential impact on our analyses.

2. Outliers and Anomalies: Unusually high or low values have been observed, indicating potential outliers. Understanding the context behind these values is essential for accurate interpretations.

Discovery of Data Quality Issues:

1. Exploratory Data Analysis (EDA): Initial EDA has revealed discrepancies and irregularities in the data. By visualizing and summarizing key statistics, we aimed to comprehend the overall patterns and distributions.

2. Data Profiling: The application of data profiling techniques has helped identify missing values, outliers, and potential inconsistencies.

Information Needed for Resolution:

1. Data Source Details: Understanding the source of the data will aid in tracing and rectifying issues more effectively.

2. Data Collection Processes: Insights into how the data is collected will assist in addressing missing or inconsistent values.

Optimizing Data Assets:

1. Additional Attributes: Knowledge of specific business goals and KPIs will help determine if additional attributes or features are needed to optimize the dataset for analysis.

2. Data Usage Patterns: Understanding how stakeholders intend to use the data will guide the structuring process for maximum utility.

Performance and Scaling Concerns:

1. Volume of Data: As we progress towards production, it is crucial to consider the volume of data we will be handling. Implementing efficient storage and retrieval mechanisms may be necessary.

2. Scalability: Anticipating potential increases in data volume, exploring scalable solutions is vital to ensure the system's performance remains optimal.