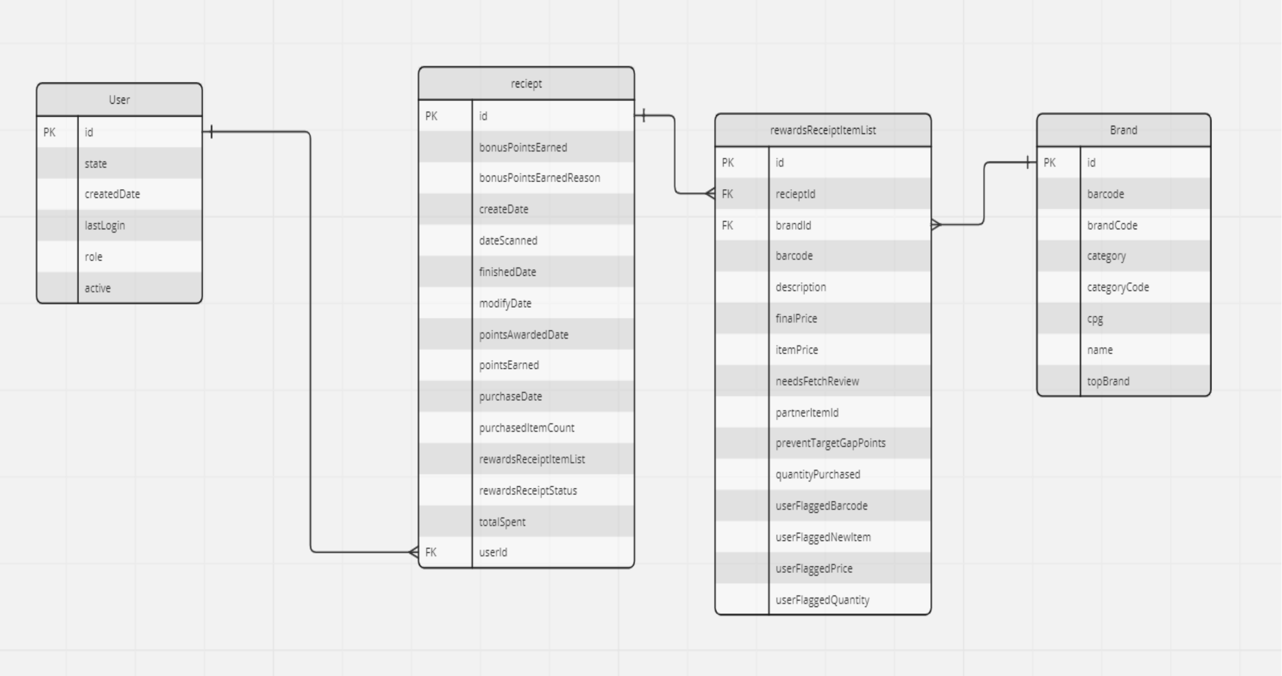
Data Analyst Assignment

Part 1: Review Existing Unstructured Data and Diagram a New Structured Relational Data Model.

Please find a structured data model with all the fields, joinable keys ensuring relations among data below, I have also included a clearer image file in the git link. I have ensured the data tables a normalized to an extend however, can be normalized further.



Part 2

Below are the queries to the questions asked in part 2 of the assignment. All the queries written are referenced to the structured data model in part 1 of the assignment.

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

```
SELECT TOP 5 brand.id as brandid
        ,brand.brand_name
        ,Count(brand.id) receipts_for_brand
FROM receipt r
INNER JOIN rewardsReceiptItemList ri ON r.id = ri.id
INNER JOIN brand b ON ri.brandid = b.id
```

```
WHERE Month(r.dateScanned) = Month(getdate)

GROUP BY brand.id

    ,brand_name

ORDER BY DESC receipt_for_brand
```

How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?

```
SELECT recent_month.id as brand_id

    ,recent_rank

    ,prevoius_rank

FROM (

    SELECT *

        ,DENSE_RANK() OVER (

            ORDER BY C1 DESC

            ) recent_rank

    FROM (

        SELECT TOP 5 brand.id

            ,brand_name

            ,Count(r.id) as C1

        FROM receipt r

        INNER JOIN rewardsReceiptItemList ri ON r.id = ri.id

        INNER JOIN brand b ON ri.brandid = b.id

        WHERE Month(r.dateScanned) = Month(getdate)

        GROUP BY brand.id

            ,brand_name

        ORDER BY C1 DESC

        ) t1

    ) recent_month
```

```
INNER JOIN (
    SELECT *
        ,DENSE_RANK() OVER (
            ORDER BY C2 DESC
            ) previous_rank
    FROM (
        SELECT brand.id
            ,brand_name
            ,Count(r.id) C2
        FROM receipt r
        INNER JOIN rewardsReceiptItemList ri ON r.id = ri.id
        INNER JOIN brand b ON ri.brandid = b.id
        WHERE Month(r.dateScanned) = DATEADD(MONTH, - 6, GETDATE())
        GROUP BY brand.id
            ,brand_name
        ) t2
    ) previous_month ON recent_month.id = previous_month.id
```

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

```
SELECT avg(CASE
            WHEN rewardsReceiptStatus = 'Accepted'
                THEN spend
            ELSE 0
            END) AS avg_accepted
    ,avg(CASE
            WHEN rewardsReceiptStatus = 'Rejected'
                THEN spend
            ELSE 0
            END) AS avg_rejected
FROM your_table_name
WHERE rewardsReceiptStatus IN (
        'Accepted'
        ,'Rejected'
        )
```

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

```
SELECT MAX(averageSpend) AS greaterAverageSpend
FROM (
    SELECT rewardsReceiptStatus
        ,AVG(totalSpent) AS averageSpend
    FROM Receipts
    WHERE rewardsReceiptStatus IN (
            'Accepted'
            ,'Rejected'
            )
    GROUP BY rewardsReceiptStatus
    ) AS subquery;
```

Which brand has the most spend among users who were created within the past 6 months?

```
SELECT brand.id
    ,bname
    ,sum(r.totalSpend)
FROM user u
INNER JOIN receipt r ON u.id = r.userid
INNER JOIN rewardsReceiptItemList ri ON r.id = ri.id
INNER JOIN brand b ON ri.brandid = b.id
WHERE u.createDate BETWEEN DATEADD(MONTH, - 6, GETDATE())
                    AND GETDATE()
GROUP BY brand.id
          ,brand_name
ORDER BY DESC 3
```

Which brand has the most transactions among users who were created within the past 6 months?

```
SELECT brand.id
          ,bname
          ,count(r.id)
FROM user u
INNER JOIN receipt r ON u.id = r.userid
INNER JOIN rewardsReceiptItemList ri ON r.id = ri.id
INNER JOIN brand b ON ri.id = b.id
WHERE u.createDate BETWEEN DATEADD(MONTH, - 6, GETDATE())
                    AND GETDATE()
GROUP BY bid
          ,brand_name
ORDER BY DESC 3
```

Part 3: Evaluate Data Quality Issues in the Data Provided

A data quality issue can be inconsistencies in the date format throughout the tables.

Below query is a sql qyery to identify if the dates are in specific order.

```
SELECT purchaseDate
FROM Receipts
WHERE purchaseDate IS NOT NULL
          AND purchaseDate NOT LIKE 'YYYY-MM-DD'
```

Part 4: Communicate with Stakeholders

I have answered this question considering some assumptions about the data.

Below is an email that address the part 4 of the assignment:

Hello,

I wanted to discuss some important aspects regarding our data and its quality. I have been exploring and analyzing the data assets we have and wanted to address a few points for your understanding.

Questions about the data:

As I delve deeper into our data, I have come across a few questions:

How was the data collected or sourced?

What are some specific data fields and their definitions- example(field-cpg)?

Are there any data transformations or aggregations applied before storing the data?

Discovery of Data Quality Issues:

During my exploration, I have identified some data quality issues:

Inconsistent date formats in the "purchaseDate" field of the receipts data.

Missing values in several fields across different datasets.

Duplicates in certain records of the users and brands data.

To address these data quality issues effectively, I would require the following information:

Clarification on the expected date format for the "purchaseDate" field.

Guidelines on how to handle missing values or any imputation strategies.

Criteria for identifying and handling duplicates in the datasets.

Optimizing Data Assets:

In order to optimize the data assets, we are working with, I would benefit from additional information, such as:

The specific business goals and objectives we aim to achieve using this data.

Any predefined data quality standards or benchmarks we should adhere to.

Any specific key performance indicators (KPIs) we should focus on to measure success.

Performance and Scaling Concerns:

While preparing for production, I anticipate the following performance and scaling concerns:

Increasing data volumes and the impact on storage requirements and processing times.

The need for efficient indexing and query optimization techniques as the data grows.

I believe addressing these questions and concerns will not only enhance our understanding of the data but also contribute to the overall success of our data-driven initiatives. I would appreciate your insights and guidance on these matters.

I look forward to discussing these aspects further in our next meeting.

Thanks,

Sanskruti Sasane.