## Question # 1: ER Diagram

A diagram of a database

Description automatically generated

## Question # 2

### 1. Top 5 brands by receipts scanned for the most recent month

WITH RecentMonthReceipts AS (

SELECT receipt\_id

FROM Receipts

WHERE purchase\_date >= DATE\_SUB(CURDATE(), INTERVAL 1 MONTH)

)

SELECT B.name AS brand\_name, COUNT(RI.receipt\_item\_id) AS total\_receipts

FROM RecentMonthReceipts RM

JOIN ReceiptItems RI ON RM.receipt\_id = RI.receipt\_id

JOIN Items I ON RI.item\_id = I.item\_id

JOIN Brands B ON I.brand\_id = B.brand\_id

GROUP BY B.name

ORDER BY total\_receipts DESC

LIMIT 5;

### 2) Average spend from receipts with 'Accepted' or 'Rejected'

SELECT RS.status\_name, AVG(R.total\_spent) AS average\_spend

FROM Receipts R

JOIN ReceiptStatus RS

ON R.receipt\_status\_id = RS.rstatus\_id

WHERE RS.status\_name IN ('Accepted', 'Rejected')

GROUP BY RS.status\_name;

### 3) Ranking Comparison: Top 5 brands by receipts scanned for the most recent month vs. the previous month

WITH RecentMonthReceipts AS (

SELECT receipt\_id

FROM Receipts

WHERE purchase\_date >= DATE\_SUB(CURDATE(), INTERVAL 1 MONTH)

),

PreviousMonthReceipts AS (

SELECT receipt\_id

FROM Receipts

WHERE purchase\_date BETWEEN DATE\_SUB(CURDATE(), INTERVAL 2 MONTH) AND DATE\_SUB(CURDATE(), INTERVAL 1 MONTH)

)

-- Recent Month Top 5 Brands

SELECT B.name AS brand\_name, COUNT(RI.receipt\_item\_id) AS total\_receipts, 'Recent Month' AS period

FROM RecentMonthReceipts RM

JOIN ReceiptItems RI ON RM.receipt\_id = RI.receipt\_id

JOIN Items I ON RI.item\_id = I.item\_id

JOIN Brands B ON I.brand\_id = B.brand\_id

GROUP BY B.name

ORDER BY total\_receipts DESC

LIMIT 5

UNION ALL

-- Previous Month Top 5 Brands

SELECT B.name AS brand\_name, COUNT(RI.receipt\_item\_id) AS total\_receipts, 'Previous Month' AS period

FROM PreviousMonthReceipts PM

JOIN ReceiptItems RI ON PM.receipt\_id = RI.receipt\_id

JOIN Items I ON RI.item\_id = I.item\_id

JOIN Brands B ON I.brand\_id = B.brand\_id

GROUP BY B.name

ORDER BY total\_receipts DESC

LIMIT 5;

### 4) Total number of items purchased for 'Accepted' vs. 'Rejected'

SELECT RS.status\_name, SUM(RI.quantity\_purchased) AS total\_items\_purchased FROM Receipts R JOIN ReceiptStatus RS

ON R.receipt\_status\_id = RS.rstatus\_id

JOIN ReceiptItems RI

ON R.receipt\_id = RI.receipt\_id

WHERE RS.status\_name IN ('Accepted', 'Rejected')

GROUP BY RS.status\_name;

### 5. Brand with the most spend among users created in the past 6 months

WITH RecentUsers

AS ( SELECT user\_id FROM Users WHERE created\_at >= DATE\_SUB(CURDATE(), INTERVAL 6 MONTH) )

SELECT B.name AS brand\_name, SUM(R.total\_spent) AS total\_spent FROM

RecentUsers RU JOIN Receipts R

ON RU.user\_id = R.user\_id

JOIN ReceiptItems RI

ON R.receipt\_id = RI.receipt\_id

JOIN Items I

ON RI.item\_id = I.item\_id

JOIN Brands B

ON I.brand\_id = B.brand\_id

GROUP BY B.name

ORDER BY total\_spent DESC

LIMIT 1;

## Question # 3

Below are the results of the exploratory analysis on the data:

### Missing Values

**Brands Data:**

* category: 155
* categoryCode: 650
* topBrand: 612
* brandCode: 234

**Receipts Data:**

* bonusPointsEarned: 575
* bonusPointsEarnedReason: 575
* finishedDate: 551
* pointsAwardedDate: 582
* pointsEarned: 510
* purchaseDate: 448
* purchasedItemCount: 484
* rewardsReceiptItemList: 440
* totalSpent: 435

**Users Data:**

* lastLogin: 62
* signUpSource: 48
* state: 56

### Duplicate Records Analysis

During the data quality evaluation, a check was performed for the presence of duplicate records, particularly focusing on the unique identifier (\_id) columns, which should uniquely identify each record. Below are the findings:

1. Brands Data

Duplicate \_id values: 0

The brands.json file does not contain any duplicate entries in the \_id field, ensuring that each brand record is uniquely identified.

2. Receipts Data

Duplicate \_id values: 0

The receipts.json file does not have any duplicate entries in the \_id field, ensuring that each receipt is correctly identified without redundancy.

3. Users Data

Duplicate \_id values: 283

The users.json file has 283 duplicate entries in the \_id field, which is problematic as this field should uniquely identify each user. This indicates a potential issue with data integrity, as duplicate user records may lead to errors in analysis and reporting.

### Investigating receipts with total spend = 0 and marked “Finished”

The investigation into receipts marked as "FINISHED" with a totalSpent of 0 revealed the following:

* There are **2 receipts** that have a "FINISHED" status but a totalSpent of 0.
* **None** of these receipts contain items flagged for review (needsFetchReview is absent in both cases).
* Both receipts have valid pointsEarned values, meaning they were rewarded points despite no recorded monetary expenditure.

Here’s a quick overview of these two receipts:

* **Receipt 1 (UserID: 6009e60450b3311194385009)**: Points Earned: 250, Purchase Date: 2021-01-21, Total Spent: 0.
* **Receipt 2 (UserID: 600afb2a7d983a124e9aded0)**: Points Earned: 500, Purchase Date: 2021-01-22, Total Spent: 0.

Both receipts seem to have earned points but show no details in the rewardsReceiptItemList. This suggests that:

1. These transactions might involve specific promotions, where points are awarded without any actual spending (e.g., receipt bonuses or special rewards).
2. There could be a data quality issue where the item list and spending information weren't recorded properly, but the receipt was still processed and points were awarded.

Investigating data types

The data types are consistent across all the tables.

Investigating Unique values in brands, states, user roles, and sign ups

While the values are unique for these fields, there are missing values that should be addressed

### Investigating outliers

During the analysis of the receipts data, a few key outliers and data quality concerns were identified. These issues may impact the accuracy of future analyses and should be taken into account when interpreting the results.

**1. Total Spent**

* The totalSpent field shows some significant outliers, with values ranging from $0.00 to $4,721.95.
* **Key Observation**: A number of receipts have a totalSpent value of $0.00, which could indicate issues such as refunds, incomplete transactions, or data entry errors.
* **Actionable Point**: These $0.00 entries should be investigated to determine their validity. Additionally, receipts with unusually high spending (e.g., $4,721.95) should be reviewed to confirm they are not outliers due to data entry mistakes.

**2. Points Earned**

* The pointsEarned field also exhibits notable outliers, with values ranging from 0 to 10,199.80 points.
* **Key Observation**: Several receipts have earned 0 points, even though a corresponding totalSpent amount is recorded. Conversely, there are receipts with extremely high points earned, which may represent promotions or data anomalies.
* **Actionable Point**: Further investigation is recommended for receipts with 0 points earned (but positive spending), as well as for those with unusually high points, to ensure the accuracy of the points system.

**3. Purchased Item Count**

* The purchasedItemCount field ranges from 0 to 689 items.
* **Key Observation**: Some receipts report a purchasedItemCount of 0 despite showing positive values for totalSpent and/or pointsEarned. Additionally, the maximum of 689 items may indicate unusual or bulk transactions.
* **Actionable Point**: Receipts with a 0 item count should be verified to determine if they are valid. Similarly, receipts with extremely high item counts should be reviewed to ensure they are not data entry errors.

## Question #4

**Subject:** Updates on Data Analysis Progress and Key Questions on Data Quality

Hi John,

I wanted to provide you with an update on our progress with the data analysis and the structured data model we're developing. We’ve made significant headway, and everything is moving in the right direction, but there are a few key areas where we could use your input.

**Key Updates:**

* We’ve built a structured relational model that captures the relationships between users, receipts, brands, and individual receipt items. This model allows us to answer critical business questions, such as identifying top brands by receipt scans and analyzing spending trends based on receipt statuses.
* While exploring the data, we uncovered some **data quality issues**, particularly around missing values and potential duplicates. For example:
  + Some receipts have missing purchaseDate fields, and in several instances, totalSpent values are zero despite having purchased items listed.
  + We also identified potential **duplicate user\_ids**, which could cause issues with linking receipts to the correct users.

**Questions About the Data:**

1. **Missing or Incomplete Data**: We found that a significant number of receipts are missing important fields such as purchaseDate and have totalSpent values of zero. Could you provide any insights into the processes that generate this data? Are there known issues or challenges in how receipts are captured or ingested into the system?
2. **Duplicate user\_ids**: We've identified some cases where the user\_id appears to be duplicated. Can you clarify if this is expected, or if this may indicate a data quality issue in how users are recorded?

**What We Need to Resolve the Issues:**

To resolve these issues, it would be helpful to have:

* **Clarification on how receipts and users are processed** and if there are any known limitations or sources of error that might explain the missing data and potential duplicates.
* **Access to any existing validation rules** or checks that could help us fill in these gaps, either through data cleansing or identifying records that need to be revisited.

**Additional Information to Optimize the Data Assets:**

Moving forward, it would also be helpful to have:

* **Details on data sources**: Are all datasets coming from a unified platform, or are there multiple sources contributing to the dataset? Understanding this will help us address any potential discrepancies early on.
* **Insight into future plans for data integration**: Are there additional datasets or new sources of information that we need to account for in the data model as we scale?

**Performance and Scaling Concerns:**

We’re confident that the current model works well for the current dataset. However, as the data continues to grow, particularly in terms of the volume of receipts and transaction details, we anticipate potential performance concerns due to the complexity of the joins involved in certain queries. To mitigate this:

* We’re focusing on **indexing key columns** (e.g., foreign keys like user\_id, brand\_id) to ensure query performance remains optimal.
* We're exploring the use of **views** to simplify some of the more complex queries, reducing the need for repeated joins.

If the data grows significantly or if we introduce new sources, we may consider **denormalization** in some areas to improve performance, but for now, we are focused on keeping the model clean, efficient, and easy to maintain.

Please let me know if we can schedule some time to discuss these points or if there’s someone else on the team who might have more context on these issues.

Looking forward to your thoughts!

Best regards,  
Raaiha Kabir