|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Receipts** | | \_id (PK) bonusPointsEarned bonusPointsEarnedReason createDate dateScanned finishedDate modifyDate pointsAwardedDate pointsEarned purchaseDate purchasedItemCount rewardsReceiptItemList rewardsReceiptStatus totalSpent userId (FK) |  1. E-R Business Model | |  | | --- | | **Users** | | \_id (PK) state createdDate lastLogin role active |   One to Many |

|  |
| --- |
| **Brand** |
| \_id (PK) barcode brandCode category categoryCode cpg topBrand name |

\*Can be created as an additional subset table from the brand schema if required.

|  |
| --- |
| **Brand Category** |
| \_id (PK) name category categoryCode |

**2.   
  
a. What are the top 5 brands by receipts scanned for most recent month?  
Assuming I can join receipt item list to brand code from a business use case.**

SELECT b.name AS Brand\_Name, COUNT(\*) AS Receipt\_Count

FROM Receipts r

JOIN Brands b ON r.rewardsReceiptItemList.brandCode =b.brandCode

WHERE FORMAT (r.dateScanned) = MONTH(CURRENT DATE)

GROUP BY b.name

ORDER BY Receipt\_Count DESC

LIMIT 5;

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

SELECT rewardsReceiptStatus, SUM(purchasedItemCount) AS Total\_Items

FROM Receipts

WHERE rewardsReceiptStatus IN ('Accepted', 'Rejected')

GROUP BY rewardsReceiptStatus;

**3. Evaluate Data Quality Issues in the Data Provided**

**Identify data quality issues using Python**

*Read Json Files (Receipts.json)*

def parse\_json\_file(file\_path):

with open('receipts.json','r') as file:

data = json.load(file)[0]

return data

-- Converting to a dataframe (Printing data to a Tabular Form)

data = parse\_json\_file('receipts.json','r')

import pandas as pd

df = pd.DataFrame(data)

print df

*Print Data Types*

//To validate if the fields have right data type such as all the date fields like Create Date, Scanned Date, Modify Date, Finished Date, Purchase Date has a date type of date; and Price associated fields are on currency or int; id as int and description as string//

data\_type = df.dtypes

print(data\_type)

// The above line will give all the data types, if a specific column is required then replace the createddate with any field. This allows us to determine and correct data types.//

print(df['Createddate'].dtypes)

*Print Distinct Values.*

print(df['id'].unique()

*Print Duplicates*

print(df.duplicated())

***MySQL on identifying data quality issues like duplicates, redundant data, relationships, and data integrity.***

*Duplicates:*

Select COUNT(\*) - COUNT(DISTINCT \*) AS Duplicates

From receipts;

*Verification of relationship between receipts and user’s table*

SELECT \*

FROM receipts as r

LEFTJOIN users as u on r.\_id = u.userId

WHERE u.uderId <> NULL;

**4. Communicate with Stakeholders**

To: x   
Subject: Data Quality Concerns in the data.  
  
Hope you are doing well. I want to bring your attention on some of the data quality issues I have encountered while analyzing our data. It’ is essential to ensure the data’s accuracy and reliability at this point before moving forward with any form of implementation or analysis.

Questions about the Data:  
1. What system or processes are responsible for capturing the data?   
2. How frequently the data gets refreshed or updated?  
3. What is the source of truth tables?   
4. What are the upstream/downstream sources flowing from these datasets.  
5. Is there any known issues or limitations that I should be aware of?

Data Quality Issues Identified:   
Inconsistent Data Types   
Missing Values   
Unstructured Data   
Redundant Data   
  
Solutions to Data Quality:  
1. Standardize and implement consistent data formatting.  
2. Implement daily/weekly validation checks on data accuracy and refreshes.   
3. Develop a process doc for data cleansing and loading.

Performance & Scaling Concerns:

I anticipate potential performance and scaling concerns due to the volume and complexity of the growing data in the future. I plan to do the following actions to address these concerns :  
  
1. Verify and implement efficient data storage and retrieval mechanisms.  
2. Optimize query performance and logics.  
3. Regularly monitor and validate system performance and conduct capacity planning to accommodate future growth.   
  
To optimize our datasets further, Appreciate any insights or feedback regarding the above-mentioned information. Please feel free to reach out if further clarification is required.