Subject: Data Modeling and Analysis for Fetch Rewards - Brands and Receipt Rewards

Hi Team,
Currently the data is being generated in a JSON structure having 3 main files, brands, receipts, and users. Owing to the unstructured nature of the json format, it is difficult to perform analysis on it directly. So, I propose to create a data warehouse to store this data in tables. This will make generating reports and performing analytics easier and faster. I've attached a diagram of the proposed data model for your review. Please email or slack me if you have any questions.

Before finalizing the database model, I would appreciate if you could clarify my questions regarding the following:

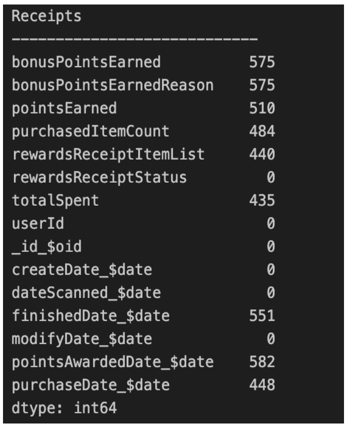
Relationship Questions:

1. I notice there's no direct relationship between brands and receipts in the JSON files except for the cpg_id (stored as rewards product partner ID in the receipts item list). Could we implement a clearer mapping between these tables to directly match brands with scanned receipt items?
2. The "brand code" in Receipts Items doesn't match the corresponding field in the Brands table. Are these meant to represent different concepts, or should we standardize them?

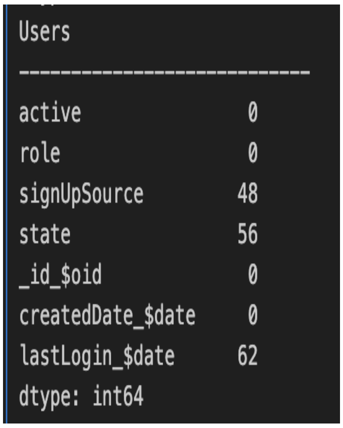
Data Quality Issues:

While processing the data in Python I found a few quality issues in the data:

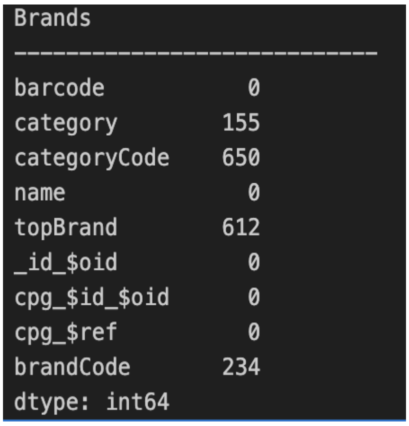
1. Missing Values: Many columns in the data do not have values present.



```
Receipts
---------------------------
bonusPointsEarned          575
bonusPointsEarnedReason    575
pointsEarned               510
purchasedItemCount         484
rewardsReceiptItemList     440
rewardsReceiptStatus         0
totalSpent                 435
userId                       0
_id_$oid                     0
createDate_$date             0
dateScanned_$date            0
finishedDate_$date         551
modifyDate_$date             0
pointsAwardedDate_$date    582
purchaseDate_$date         448
dtype: int64
```



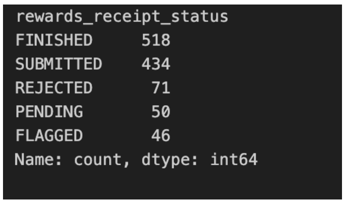
```
Users
----------------------------
active              0
role                0
signUpSource       48
state              56
_id_$oid            0
createdDate_$date   0
lastLogin_$date    62
dtype: int64
```



```
Brands
------------------------------
barcode           0
category        155
categoryCode    650
name              0
topBrand        612
_id_$oid          0
cpg_$id_$oid      0
cpg_$ref          0
brandCode       234
dtype: int64
```

As you can see, so many columns have missing/null values. We should evaluate which of these fields are critical for analysis and develop strategies to address these gaps.

2. Duplicates: Approximately half the records in the Users table appear to be duplicated. I recommend removing these redundancies before migration to the data warehouse.
3. Data Distribution: There are no receipts with "Accepted" status, and other receipt statuses show uneven distribution. This imbalance could impact analytics and future predictive modeling. We should collect additional data for more comprehensive analysis.



```
rewards_receipt_status
FINISHED     518
SUBMITTED    434
REJECTED      71
PENDING       50
FLAGGED       46
Name: count, dtype: int64
```

Processing Efficiency:

1. Currently, item lists are embedded within receipt data, requiring additional processing. Would it be possible to provide this information in a separate JSON file?
2. Our analytical queries currently require resource-intensive SQL operations. When finalizing the data model, we should optimize for our most frequent query patterns to improve performance.This will ensure efficient processing.

I believe addressing these issues will help us implement a robust data warehouse that can serve as a solid foundation for our analytics initiatives.

Please let me know your thoughts via Slack or email.

Thank you!