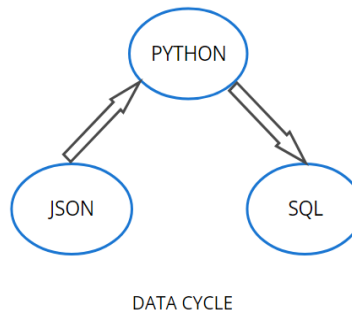


Subject: Data Modelling and Analysis for Fetch Rewards

I hope this message finds you well. I've conducted a detailed data modeling and analysis exercise as part of our ongoing efforts to enhance the Fetch Rewards system. This analysis focuses on the relationships between brand names and receipt rewards, specifically aiming to compare various metrics for receipts with different reward statuses.

Overview of the project:

Let's See the Data Transformation Cycle Diagram



From the diagram, we can analyze, the data was initially generated in json structure, then we transformed the data into a Python to perform the Exploratory Data analysis and performed the queries in my sql.

Currently data is structured in JSON format across three main files: brands, receipts, and users. Due to the inherent complexity of analyzing data directly from JSON, I propose creating a data warehouse to organize this data into tables. This approach will facilitate easier and faster generation of reports and analytics. Please refer to the attached data model diagram for your review. Feel free to message me with any questions.

Before finalizing the database model, I have some clarifications and concerns:

JSON Files:

Brand-Receipt Mapping Currently, there's no direct link between brands and receipts in the JSON files, except for the 'cpg_id' (rewards product partner ID) stored in the receipts file's item list. Can we establish a mapping between these two tables to directly associate the brand of an item scanned in receipts?

Brand Code Consistency: I noticed discrepancies where the 'brandCode' in the receipts item list does not match the 'brandCode' in the Brands table. Are these discrepancies intentional, or should we work towards aligning them for consistency?

Data Quality Issues:

1) **Missing Values** There are numerous columns in the data with missing or null values. These columns need further investigation to determine their relevance for current and future analyses. If critical, we should devise methods to populate these missing values.

Users		Receipts		Brands	
active	0	bonusPointsEarned	575	barcode	0
role	0	bonusPointsEarnedReason	575	category	155
signUpSource	48	pointsEarned	510	categoryCode	650
state	56	purchasedItemCount	484	name	0
id\$oid	0	rewardsReceiptItemList	440	topBrand	612
createdAt_\$date	0	rewardsReceiptStatus	0	_id_\$oid	0
lastLogin_\$date	62	totalSpent	435	cpg_\$id_\$oid	0
dtype: int64		userId	0	cpg_\$ref	0
		id\$oid	0	brandCode	234
		createDate_\$date	0	dtype: int64	
		dateScanned_\$date	0		
		finishedDate_\$date	551		
		modifyDate_\$date	0		
		pointsAwardedDate_\$date	582		
		purchaseDate_\$date	448		
		dtype: int64			

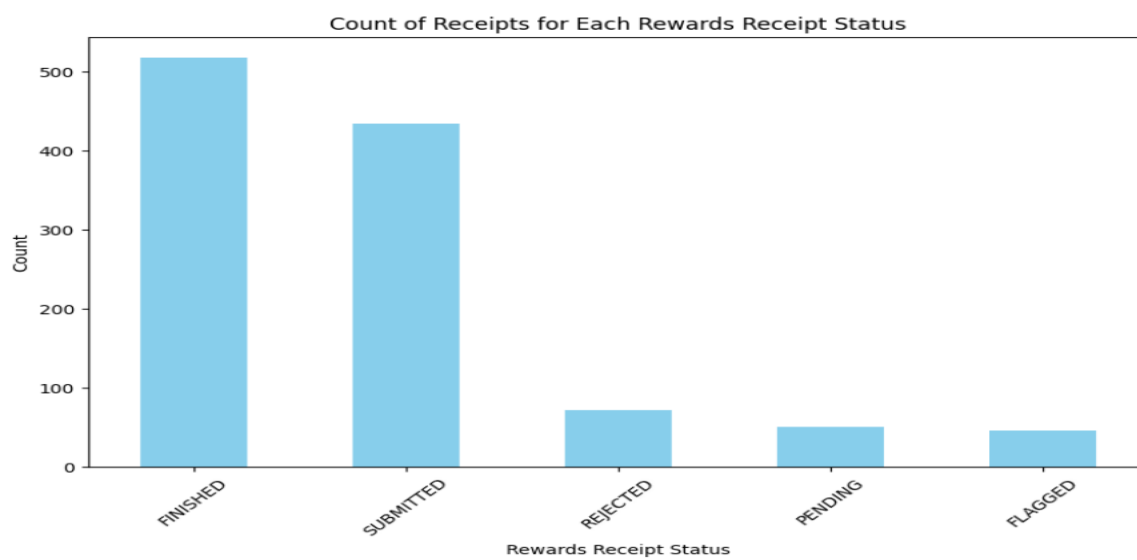
Duplicate Records: About half of the entries in the Users table appear to be duplicated. I recommend removing these duplicate rows before inserting data into the data warehouse to maintain data integrity.

```
print(users_df.duplicated().sum()) # duplicates in users table
print(brands_df.duplicated().sum())
print(receipts_df.duplicated(subset = ['_id_$oid']).sum())
```

✓ 0.0s

283
0
0

Receipt Status Distribution: There are no receipts with 'Accepted' status in the dataset, and other receipt statuses are unevenly distributed. This imbalance could pose challenges for future analytics and predictive modeling. Collecting more data is advisable to address these quality issues.



Other Considerations:

Receipt Item Lists:

Currently, each receipt stores its item list within its own data structure. To streamline processing, could we separate this information into a different JSON file?

Overall, resolving these issues will pave the way for implementing a robust data warehouse system that serves as the cornerstone for our analytics efforts.

Integration of Brand Details:

Incorporating detailed brand information will enhance analysis quality and accuracy, leading to better insights and informed decision-making.

Optimizing Query Performance:

Overall, resolving these issues will pave the way for implementing a robust data warehouse system that serves as the cornerstone for our analytics efforts.

Thank you!