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

Hi,

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 have attached the diagram for the data model for you to peruse. Please slack me with what questions you have.

Before moving ahead with finalizing the database model, I would appreciate if you could clarify my doubts regarding the following:

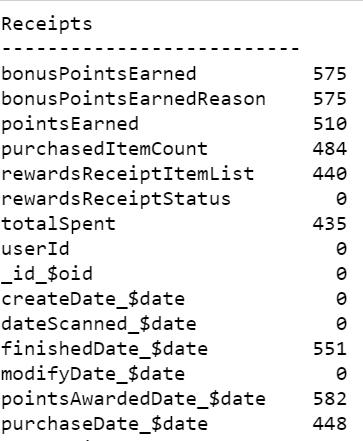
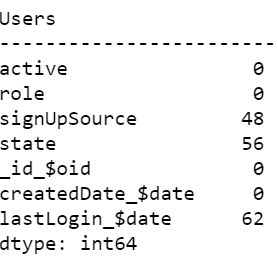
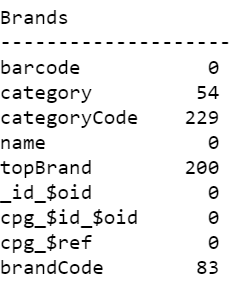
JSON files:

There seems to be no direct relation between brands and receipts in the JSON file apart from the cpg\_id which is stored in the receipts file as rewards product partner ID in the receipts item list. Can we have some sort of mapping between these 2 tables to directly match the brand of an item scanned in the receipts?

I noticed that Receipts Items list stores brand code however the values of this field does not match with the brand code in Brands table. Do they have different meaning? If not, can we please try and make them consistent?

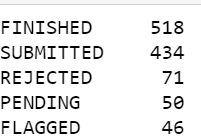
Data Quality Issues:

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.

As you can see, so many columns have missing/null values. We can investigate each one of these to understand their significance. If the column seems to be important for analysis, now or in future, we must find a way to fill it.

1. Duplicate values: Almost half the records in Users table, seem to be duplicated. I suggest we remove these rows before inserting them in the data warehouse.
2. There are no receipts of Accepted status in the dataset, moreover, the data for other receipt status is also not evenly distributed. This could be a problem when performing analytics and creating predictive analytical models in the future. We need to collect more data to perform quality analysis.



1. Other Issues:

The Item list for each receipt is currently stored within the receipt data itself. Additional time and processing are required to convert it into a different table and insert into the database. Can we get this information in a different JSON file to avoid additional processing?

To generate analytical reports and queries, we need to write processor heavy SQL statements which takes several minutes to run. When building a data model, we need to keep in mind the type of queries we are going to run most frequently and design the database accordingly. This will ensure efficient processing.

All in all, I think, if we get these blocks cleared, we can implement a very useful data warehouse systems that can be used as the basis of our Analytics.

Thank you!