Hello,

I am reaching out to you to share some insights regarding the unstructured data (.json files) that was shared as part of the Fetch Rewards Data Analyst Assessment.

I do have some clarifying questions regarding the data provided and your feedback would be appreciated.

1. What is the source of this unstructured data? And how frequently the data is fetched from these sources?
2. I observed a lot of missing values in the data, is it due to some data migration issue?
3. Are there any key metrics that we wish to measure?

As I did some initial data exploration, there were some data quality issues such as

1. There are some null and missing values in the data. Data had to be cleaned by removing these null and missing values which ever found irrelevant to the data and analysis.
2. User data file had duplicate User\_id.
3. The corresponding values of column name brandCode is present in brands.json and receipts.json files do not match each other. The column names should be changed if the both the columns have different meanings. Additionally, the brandCode column in the receipts table has 90% of missing values as compared to the brands table.
4. Brands dataset has no uniquely identifying attribute.

Some of the key pointers to be taken into consideration should be:

1. Construct the data warehouse in accordance to the most important KPI’s we wish to measure
2. Using data partitioning strategies for better efficiency and performance

Thank you! I look forward to hearing your thoughts.

Best Regards,

Labdhi Ghelani