Hi John(Product/Business Leader),

I explored the provided dataset and I have a few questions and concerns about the data.

1. Questions about the data :

Can you please tell how the relationship is maintained between items on the receipt and the brand? Is there a common field between Receipts and Brands JSON? Please confirm.

2. Data Quality Issues:

I found out the Data Quality issues by exploring the JSON files using Python. There are missing values and inconsistent data which need to be addressed. The same issues would be translated to relational model as well. If we could solve the problems in JSON, it can be fixed in relational model either.

3. Resolving Data Quality Issues:

Missing values issue can be fixed by building a regression model because mean imputation, median imputation techniques to fill the missing values do not apply in this case. If the business logic behind how the points is calculated using total spent, number of items on the receipt and brand information is known, then the regression model can be built to predict the missing values.

4. Information needed to optimize the data asset:

If the information on which brand the bonus points are earned, it would be easier to keep track of an information such as popular brand among the consumers rather querying 4 tables to answer such a question. This is one of the information requirements to optimize the data asset.

5. Performance and Scaling concerns:

For a few business questions, there is a need to query more than 2 tables which will increase the time to return the results. It can be solved by setting up a trigger and maintaining the user and brand statistics in separate tables which will decrease the time to query the data effectively.

As the business grows the data grows too which will bring the need of scaling the data warehouse. The scaling can be done either horizontally or vertically based on the budget, bandwidth and requirement. We can also look at the possibility of migrating to a cloud(AWS) which manages and scales the data warehouse automatically.

Please review the above concerns and questions about the data. We can setup a meeting later this week to address the above concerns and move forward accordingly.

Thanks,

P. Aravind

Data Analyst,

Fetch Rewards