Fourth: Communicate with Stakeholders

To: product/business leader

Sub: Questions about data

Hello,

I have developed a data model for querying business insights using the raw json files provided. The data model is normalized to 3NF i.e. I have incorporated additional tables (entities) for separating out columns that are not dependent on its default table. This is done to remove transitive/partial dependencies which could lead to inaccurate aggregations being performed on the data.

Although this model seems to be rigid for early-stage querying, it can be optimized further for better efficiency, scaling, and answering complex business questions in the future. Kindly answer the following questions about the data to improve the model:

1. What does the fields rewardsProductPartnerId, rewardsGroup, competitiorRewardsGroup signify? Are these field related to a third-party organization that rewards users based on the products in the transaction?
2. Does Metabrite change campaigns periodically? And if so, do the prices of products change depending on campaign?
3. Is CPG id unique across all products?

I have encountered the following data quality issues that I would like to bring to your attention:

1. Columns - barcode and brandCode which are used to uniquely identify products in a transaction (referring to collection rewardsReceiptItemList). Out of 1167 brand instances, 1160 are unique and 7 barcodes are duplicated. A remodeling is suggested on the brand table to have a non-null column that uniquely identifies each product. (Suggested column is brand uuid).
2. The users table has 485 instances of users records out of which 212 are unique and 283 are duplicated records. This could lead to potential double counting of users.
3. Data types of columns – pointsEarned and totalSpent are characters. Outlier records detected in columns – bonusPointsEarned and pruchasedItemCount.
4. Barcode “4011” is suggested to be added in the brand table to identify unknown item.
5. Collection rewardsReceiptItemList is an essential table for identifying products in a transaction. It has variable columns being reported by each rewardsPartner. Stabilizing this table could benefit during production in the long run.

Thank you for your time. Please let me know if you need further clarifications.

Best,

Hari