Subject: Clarifications and Data Quality Issues

Hi _____,

I have been working on the database design and trying to analyze the data of users, brands and receipts scanned. During the process, I had multiple questions about the business context of things that I think you would be the best person to answer.

Please verify that the following assumptions made during the data analysis are correct. If not, please flag the wrong ones to me in your response.

The data has been modeled with the following assumptions, please clarify if these are correctly assumed or if there are any misunderstandings within these:

1. A user may exist without ever having scanned even one receipt.
2. Each user may scan one or more receipts.
3. Each receipt must contain at least one or more items.
4. Each brand may have one or more items.
5. Each item with a unique barcode must belong to one particular brand only.
6. There could be items that may not have been purchased in any receipt ever scanned.
7. Each item belonging to a particular brand may have been bought more than once by one or more users.
8. When a receipt is scanned, it triggers one or more reward events based on the number of items in the purchase.
9. Each reward event is triggered for one particular item and specifies ‘quantityPurchased’ of that item.
10. A brand must sell products in only one category.
11. Many brands could sell in the same category.
12. ‘active’, ‘createdDate’, ‘role’ cannot be null (NOT NULL in the database model)
13. ‘brand_id’, ‘brand_barcode’, ‘brand_name’, ‘cpg_id’, ‘cpg_ref’ cannot be null (NOT NULL in the database model)
14. Each brand belongs to only one unique category (checked with current brands’ data
15. Each category is derived from a unique categoryCode (transitive dependency, hence split into a new table called CATEGORY)
16. Each category may have more than one brands
17. ‘Receipt_id’, ‘createDate’, ‘dateScanned’, ‘modifyDate’, ‘rewardsReceiptStatus’, ‘userId’ cannot be null (NOT NULL in the database model)

While conducting a detailed exploratory analysis of the given data sources, I came across a few findings that I would like to share and hope to get back a response to gain a better understanding of the following:

- I noticed that certain data elements related to user flags, user reviews are together with barcode, item price, and purchase count. I would appreciate some clarification on why these different aspects are combined in the same dataset, as it may affect the complexity and relevance of the analysis.

- I would also like to understand better the ‘barcode’ for brands as well as the barcodes of items and if they are different because through my analysis I found that 16 barcodes are coinciding among these: '511111001485', '511111001768', '511111003960', '511111004127', '511111101451', '511111104186', '511111104537', '511111204206', '511111502142', '511111518044', '511111602118', '511111704140', '511111802358', '511111901587', '511111902690', '511111904175'. Moreover, there are 1167 unique brands in the data but 1160 unique bar codes, why is there a mismatch of 7 barcodes? Is this an error?
- The same barcode has multiple descriptions for items that are scanned within receipts. Should we be devising a way to standardize the description? This will help in optimizing the item descriptions.
- Some receipts scanned do not have any items within the receipt. Can you explain better why there are such data points because a receipt should at least have one item (product purchased) in it.
- Should we maintain a separate database of all products that are scanned wherein products automatically get added to the database when a new receipt scan has a product that has never been scanned for?
- Can you explain the MetaBrite campaign’s business context and other fields in the data related to it?

- The current the users data has only two signup sources: ‘Google’ and ‘Email’. Can there be any others?
- Currently, the users data has users from only 8 states and the rest are NULL, is it an optional value while creating a new user? Is it a categorical choice among 50 states while creating a new user? (will model the state column as an ENUM in the database accordingly)
- Does CPG mean Consumer Packaged Goods in the brands' data? There are two possible values for CPG references in the data: ‘Cpgs’ and ‘Cogs’ - what do these mean?
  - The categoryCode seems to be just be an expansion of the category field and it might be a better idea to just have either as it is redundant information in the database.
  - When the needsFetchReview column is true and the reason for needing Fetch review is USER_FLAGGED, can you explain the user_flagged entries and what changes it should make to a users access to the app?

I would also like to know if the company has any plans for conducting A/B testing or bucketing users into VIP categories to effectively dash out bonuses. That way we can structure our data assets in a way ensuring time series analysis can be performed easily. Moreover, moving into the future, as the company and data grow, I was wondering if there are any plans of migrating the data to a data warehouse or a data lake house.

I’m looking forward to our call for a more detailed discussion and analysis of the findings mentioned above.

Thanks and Regards,
Debanjali Saha