Hello,

I have recently been tasked with building a relational data model for JSON data that we capture regarding our users purchases, receipts, and rewards activity. As part of this initiative I have documented some questions I have about the data as well as some potential data quality issues that may require attention.

Before getting into the questions and surfacing the potential data issues I would like to explain what has been done to this point. Thus far I have built a data model including 5 database tables using data samples from our Brands, Users, and Receipts JSON import files. Much of the data model takes the form of the JSON files with two important distinctions.

First, CPG values from Brands.json have been parsed into a separate table from the main Brands data. The CPG table is related to the Brands table Second, rewardsReceiptItemList values from the Receipts.json file have been parsed into a table separate from the base Receipts table. These model structure changes were done to make support indexing as these two parameters will likely collect a large number of data in the future.

The tables are named as follows:

- Brands – All data from the Brands.json sample except the CPG field, which has been parsed into a different table.
- CPG – Data in this table is parsed from "cpg" field in Brands.json. The CPG table is related to the Brands table via the CPG.ID field and the Brands.CpgID.
- Receipts – All data from Receipts.json except the rewardsReceiptItemList field; data from that field has been parsed into a separate table.
- RewardsReceiptItem – Data in this table is parsed from the "rewardsReceiptItemList" field in the Receipts.json sample.
- Users – This table contains all data from Users.json sample.

Questions

1. In the CPG data I noticed that all CPG records have a "ref" field value of "Cogs". Is this intentional or is this the result of an upstream data issue?
2. There are a several ID fields in the rewardsReceiptItemList data have ambiguous names which I am not able to easily connect to other parts of the data model. Those fields are partnerItemId, pointsPayerId, and rewardsProductPartnerId. It appears that these fields may be populated with CPG IDs. Can someone confirmation of that observation?
3. Is Brand Code the most reliable way to connect rewardsReceiptItemList data with Brands data? As pointed out below in the data issues I have uncovered, barcode data appears to be unreliable.
4. Should the **purchasedItemCount** in **Receipts** data match the number of items in the **rewardsReceiptItemList**? I considered a quality check for this condition but I would need clarity about this assumption.
5. How is the **topBrand** parameter in the **Brands** data used?

6. In the **rewardsReceiptItemList** parameter data there is a field named **rewardsReceiptStatus**. Two of the possible values for this field are “FLAGGED” and “REJECTED”. Do these values have the same functional meaning? If not, can you clarify how the cases should be treated for both values?
7. In the **Users** table there is a parameter named **active** that appears to be used for deactivating User accounts. How does a user account become deactivated? If a previously valid user account becomes deactivated how do we track the timing of the deactivation in order to perform time-based analytics (i.e. to identify fraudulent activity on the account after deactivation)?
8. Are user accounts only based in US territories? As discussed below in the data issues section, there are a number of NULL values in the State parameter which could be significant if all accounts are assumed to be US based.

Data Issues

As part of this initiative I also came across a number of potential data issues that I would like to 1) confirm with the team, and 2) offer a potential solution or resolution. I have detailed those issue in the table below:

Potential Issue	Potential solution
IDs in the **Users** table are not unique. This appears to be a result of duplicated **User** accounts in the JSON data.	Investigate whether there is a programmatic issue creating duplicate User accounts. Create script to remove or deactivate duplicated accounts using User ID.
48 NULL values in **SignUpSource** field of **Users** data.	Investigate whether issue is due to legacy data or potential bug in code. If bug in code resolve bug.
56 NULL values in **State** field of Users data.	Investigate whether issue is due to legacy data or potential bug in code. Prompt user account to update information on future log in.
Orphaned User IDs: There are 117 User IDs in the **Receipts** table that do not appear to be associated with user accounts according to the **Users** file.	Identify internal expert to assess whether User.json file contains all User records under scope. Backfill with more complete User.json file.
552 barcodes found in **rewardsReceiptItemList** that are not associated with a record in the Brands data.	Identify internal expert to assess whether Brands.json file contains complete data set under scope. Backfill with more complete Brands.json file.
186 **brandCodes** in **rewardsReceiptItemList** that are not associated with a record in the Brands data.	Identify internal expert to assess whether Brands.json file contains complete data set under scope. Backfill with more complete Brands.json file.
For 460 records in the Receipts table with a rewardsReceiptStatus of “Finished” the **pointsEarned** value in the Receipts does not reconcile with the total of the **pointsEarned** according to the **rewardsReceiptItemList**.	Confirm that this is in fact an issue with internal stakeholders. Look at upstream code to identify where pointsEarned calculations are having errors introduced.

Thank you!