Subject: Questions on Data Quality Issues

Hello Product/Business Leader,

I am writing this email to ask you about a few data issues we had during the data modeling and wanted more information on them. They are as follows:

1. We received users, brands, and receipts JSON files from the business. Though the data had null values in all of the files, there were duplicates found in the Users table. For this reason, we are unable to consider user_id as a primary key as it would be a problem in data analysis moving forward.

2. There is no direct link between brands and receipts in the JSON file apart from the cpg_id which is stored in the receipts file as rewards_product_partner_id in the receipts item list. Is it possible to have a mapping between these 2 tables to directly match the brand of an item scanned in the receipts?

3. There is a column name with ‘brandCode’ present in receipts and brands table, however, the value present in them does not match with each other. Do they have different meanings? If yes, can we try to change the names of the columns and make them more consistent?

4. We noticed that there are some receipts for which the points are awarded although there is no information on the receipt item list. This could be an issue with the OCR while reading barcodes on the items list and capturing records.

5. The item list for each receipt is currently stored in the receipt itself. This would require extra processing to convert into a table and insert into the database. Is it possible to get this information in a separate JSON to avoid additional processing?

As the volume of the uploaded data on the platform increases, we anticipate the performance will degrade. We plan to write efficient database queries and indexes for frequent queries. We may also want to separate the current data from the old data if we want to ensure it stays fresh.

We can schedule a call for 1 hr to discuss this further. Please let me know your availability so we can connect!

Thanks,
Vinay Kaushik