Subject: Data-related questions & Data quality issues

Hello Stakeholders,

I am Parth Parsana, currently working as a Data Analyst at Fetch Rewards. I was assigned a task to analyze the previous data (Receipts, Brands & users) and create a new relationship model to store this data in a data warehouse setting. After pre-processing the JSON file and performing feature engineering, I created an Items table (derived from receipts table). Upon analysis I have some questions, and I think you would be able to answer them for me:

* What is the Cpg\_id used for?
* What are the most important features present in the data provided?
* What KPIs does the department want to monitor?
* How is the user data collected? Since there are duplicates in user IDs
* In user's table, are all the provided user’s consumers or fetch staff?

Additionally, I performed an Exploratory analysis of the data and found many data quality issues like:

* Data Duplicity: The user table has half the user IDs as duplicates. Since we use the userId to link other tables, this will impact any analysis we perform because the primary key cannot have duplicate values.
* Data format issues: The date provided in the many fields of the receipt table is in a Unix numbered format, which is usually in date or date-time format. For any analysis to find top brands or top users in recent months or years, we need it in date format.
* Data inaccuracy: There were some inaccurate data in the receipts table where the total spend on a receipt did not match the number of items and price per item, which indicate the possibility of data loss. Additionally, between the brands and items table, the barcode has incomplete data (only four digits in a few).
* Missing Data: There was missing value present in all tables but most importantly there were nulls in pointsEarned, purchaseItemCount and totalSpend fields which means many users did not receive their rewards/bonus by using the app. I recommend we investigate how this data is processed/collected to find the underlying problem.

With the current data, there would be many concerns regarding the performance and scaling of the system, and how to handle them are as follows:

* Data not maintained correctly will lead to incorrect analysis – Make strict data format rules and apply valid constraints to allow only the correct data to be processed.
* Vertical scaling – There should be a strict duplicate check before entering any userId or receipt id to avoid inaccurate records or duplicates in primary key fields. Using a cloud platform will help manage the data systematically and accurately.
* No appropriate link between brands and items – We should use NLP techniques on the item description to match the brand it resembles if there are null.

I have attached my analysis with this email so you can review it if required. Please let me know if you need further clarification on any questions or the data quality issues I have mentioned. I will be happy to schedule a meeting to discuss it further.

Thanks,

Parth Parsana