Hello Product/Business Leader,

Hope you are doing well!

I was working on unstructured json data (contains users, receipts, brands json data) and after performing initial data analysis by exploring the data, I have few questions about the data and wanted to reach out to you regarding the same.

Firstly, I wanted to learn more about the source of data , where the data is stored and how the data is collected because data is not consistent and I have observed user_id's that are not present in Users data are present in receipts data. What should be the best approach to maintain data consistency ? Should those users in receipts be deleted ? The brand name is associated with two different barcodes for few brands and this might cause a problem going forward. Since barcode is unique and associated to single brand name.

In users data, the user_id is not unique and contains duplicates and many column names are not readable and contains special characters which can be resolved by creating ETL script to automate transformation from unstructured to structured data. The data has lot of null values and requires data cleaning but at this point it is difficult to proceed with data cleaning because information is not available for most of the attributes in rewardsreceiptsitemlist , how rewards and bonus points are awarded .Also, the rewardsreceiptsstatus field has values like pending, submitted, finished, rejected and flagged and what does FLAGGED status mean and are there any actions that needs to be performed when rewards status is REJECTED. To know the relevant columns we need more information about attributes. .

Since the unstructured data is in json format, the processing times would be large if we use postgresSQL. So we can leverage cloud and Bigdata for executing queries faster . As the data increases we can store the unstructured data by refactoring the data and storing the data in object store .
We can use distributed processing like Spark to perform analytics and data modeling.

Please let me know the best time to discuss in detail about the findings .

Thank You!
Regards,
Supritha