2. Queries that directly answer predetermined questions from a business stakeholder

- What are the top 5 brands by receipts scanned for most recent month?

Answer: Removing the "ITEM NOT FOUND" entries, the top 5 brands scanned for the month of Feb-2021 (the most recent month) are

1. FLIPBELT LEVEL TERRAIN WAIST POUCH, NEON YELLOW, LARGE/32-35

2. THINDUST SUMMER FACE MASK - SUN PROTECTION NECK GAITER FOR OUTDOORACTIVITIES

3. MUELLER AUSTRIA HYPERGRIND PRECISION ELECTRIC SPICE/COFFEE GRINDER MILLWITH LARGE GRINDING CAPACITY AND HD MOTOR ALSO FOR SPICES, HERBS, NUTS,GRAINS, WHITE

4. DELIMEX

5. GOOD SEASONS

- How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?

Answer: The top 5 brands scanned for the month of Jan-2021 are

1. *DIGIORNO CHEESE*

2. COOL WHIP

3. GODIVA INSTANT PUDDING MIX

4. DEVOUR

5. F. WHITLOCK & SONS BBQ SAUCE

The more popular brands scanned in Jan-2021 lie in the Food and Dining brands, whereas Feb-2021 brands are mixed with retail and dining

- When considering *average spend* from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

Answer: For rewardsReceiptStatus = 'FINISHED' (meaning ACCEPTED), the average spends (80.9) is higher than the average spends of rewardsReceiptStatus = 'REJECTED' (23.3)

- When considering *total number of items purchased* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

Answer: For rewardsReceiptStatus = 'FINISHED' (meaning ACCEPTED), the total number of items purchased (8184) is higher than total number of items purchased of rewardsReceiptStatus = 'REJECTED' (173)

- Which brand has the most *spend* among users who were created within the past 6 months?

Answer: Brand with most spends: PULL-UPS (4647)

- Which brand has the most *transactions* among users who were created within the past 6 months?

Answer: Brand with most transactions: DIGIORNO CHEESE (642)

3. Evaluation of Data Quality Issues in the Data Provided

- **Non-uniqueness of ids in users table**: of The number of unique users in the users table is 212 only. However the users table has 495 records. Ideally the table must be unique in terms of id column. Upon further inspection it was found out that theses are duplicated rows (other column values being the same). A unique key contraint would have helped to avoid that.
- **Null values in users table:** The fields of *lastLogin, signUpSource,* and *state* have null value percentage of 12.5%, 9.7% and 11.3% respectively.
- There are 117 users who have entries in the receipts table but are not present in the users table. Ideally, the userIds in receipts table should be a subset of all user ids present in the users table.
- The nested JSON structure for *rewardsReceiptItemList* field is not consistent. The keys change as per the value of the *rewardsReceiptStatus* field. For creating a relational database, we should have keys which are consistently present (even if the corresponding values are null values), and should be collectively exhaustive of all scenarios (example, when the status of the receipt is SUBMITTED, FLAGGED, or REJECTED).
- There should be a separate data source which contains the unnested keys of *rewardsReceiptItemList* field. This way, we can get analytical insights on a brand level in a better way. The volume of this table would be very high so we should think of creating indexes and partitions which would optimize analytical queries for brand level analyses.
- Just as *rewardsReceiptStatus* is present on a receipt level, the status should also be present on each item level so that we are able to identify the individual status of each item in the receipt.
- There are 3 cpg_ids which are present in the receipts table (rewardsProductPartnerId) but not present in the brands table. Ideally this should be a subset of the cpg_ids of brands table. This is a process flow that should be set up as part of brand onboarding

4. Communication with Stakeholders

Hi All,

I hope this message finds you well. We as an organization are striving to create an environment within all the teams to ensure that all key decisions are data-driven in nature, be it product or business or operations. For this, we have been working to enhance our data warehouse capabilities, to ensure that we take the help of data products and assets in decision-making. I am hereby sharing observations and insights for achieving these objectives, seeking your feedback on the same.

Data/Logs generation
We have so far discovered 3 data sources that will help us uncover various insights into product usage and business performance, namely user demographics, receipt scan history and brand information. I wanted to know if there are any other sources of data which we can utilize. On the top of my mind, I can think of more data sources such as

1. App usage performance: This data source will enhance product analytics and help us discover how customers see and use our product
2. Acquisition of user-level data: We can leverage more information about our users and their behavior to provide better experience of the product.
3. Acquisition of data of onboarded brands: This will help us analyze the portfolio of brand partnerships and bet on brands which are better-performing

Business Objectives and Key Results
To bridge the gaps for ensuring business continuity, it is necessary to understand how data is currently being used and any particular challenges people are facing. As a first step, there should be a cross-organization alignment in objectives and key results to be achieved. This will not only help in accomplishing common goals but also help in understanding the needs in terms of the following data initiatives:

1. data warehouse modeling (need for master data management, data cataloging and insights generation)
2. data platforms requirements (for heuristic and ML-driven decision-making)
3. database scaling and performance requirements (ensuring integrated schedules/flows for creating, updating and hosting the data in the most optimized way)

Data Quality Issue Identification and Resolution
In the existing data sources there were various issues identified using basic sanity checks and outlier detection. Stating some of them briefly

1. Duplicate records identified in data sources (namely in user information).

2. Inconsistencies regarding receipt history: There were users who are part of receipt scanned history who haven’t seemed to get onboarded. Same issue persists in brands table where brand identifiers haven’t seemed to get onboarded.
3. Inconsistent information of item-level information: The information about the items seem to be changing with, say, the status of the receipt.

These are issues which can hamper data accuracy and integrity. Thus, for resolving the issues, I would like to have a separate discussion on

- Business objectives and goals to be achieved
- Existing business process flows
- How process flows are tied to the logs generation process.

This should be a two-way discussion as to what optimizations need to be done to ensure better alignment in process and data.Hope that this will be a fruitful discussion and will set the path for the organization scaling greater heights

Best regards,
Ashish