Audience: Business Stakeholders

Subject: Exploratory data analysis (EDA) for the provided datasets (users, brand, receipts)

Summary:
The data team has been asked to analyze the datasets provided to understand the content, quality, usability, and integrity for each of the files.

The team uses the following process to analyze the data files:

- Data relationships
- Data modeling and objects
- Data pattern and trends
- Data assumption and validation
- Check for missing data for required fields
- Check that dates are in the correct format
- File layout
- File format

Once the data passes the initial check, the data is staged so that we can develop the initial models based on our EDA. During this process we may uncover additional issues which require the source system to resend corrected or missing files to get a working data model designed.

After review of the data, the team had the following questions about the data:

Some key findings:

- Transactions data is only through March 2021.
- Test data is included
- The data from receipts and brands is missing/incomplete key data that will allow this data to be analyzed for patterns/trends at a deeper level.
- Some the stakeholder questions could not be answered from the the provided queries as the data did not contain any records (ie. the request for analyzing records with rewardsReceiptStatus of ‘Accepted’ or ‘Rejected’)

Brand data schema

- Will the partner product file be integrated into the data warehouse?
- Will the brands data be augmented to include parent/child relationships so that each barcode can be grouped and filtered?
- How will the CPG collection be leveraged and will detailed analytics be required?

Receipts data schema

- Who will be the users of this data model? How the data gets modeled and developed into analytics will be based on the use cases, features, and metrics that the user is expecting to filter and report on.

- The initial prototype of the data model was developed with the idea to execute queries that are fast to answer the stakeholder questions - more aggregated reporting. As the business deep dives into the details there will be challenges with this original design to scale well.
- Will an items data set be provided that will allow the data ingested to separate out receipt items from the storing the items attributes in one table? The manner the data is ingesting will be problematic in getting to a high degree of data quality.
- Will the tracking of rewards be a specific purpose where you will need to analyze issues with rewarding and missing points, rejection, reasons, etc?

Users data schema

- Based on the definitions provided with the data this table should only contain consumers and not other types of roles. Will there be a need to track and audit internal users in the data?

How did we uncover the data issues?

During the EDA the team went through a series of tasks to identify the following types of issues:

- Missing values for key fields
- Missing values for fields that would be used to define the relationship between the objects
- Identify missing data that will allow for better data management and quality to reduce the records with many fields where the value is missing or blank
- Reviewed the questions from stakeholders and analyzed the data to ensure the necessary data elements was available.

What do you need to know to resolve the data quality issues?

- Meet with business stakeholders to understand the source systems, methods to extract the data (API, REST, S3, etc), business process for entering/updating data that will need to flow downstream into other systems like the data warehouse
- Review the business process for the transactions that will impact the datasets
- Review with the stakeholders the questions they will be asked to get answers for, KPI’s that will be monitoring the health and performance of the program, and who will be the different point of contact for SME, sign-off, and change management approvals
- Executive sponsorship of project, scheduled status updates with senior leadership to help clear roadblocks and make decisions
- Review design, orchestration, and data monitoring for the new pipelines that will be used to ingest the data
- Points of contact for each of the sources of data to help with clarifications and resolving issues when problems arise

What other information would you need to help you optimize the data assets you're trying to create?

- Metadata on the datasets to understand the data types, required fields, expected values

- Data dictionary and/or documentation for the data sources
- Use cases for the types of questions the business will want to understand better such as user purchase behavior, customer segmentation, rewards and LTV, product category and affinities.

What performance and scaling concerns do you anticipate in production and how do you plan to address them?

- The initial star schema relies on a highly de-normalized table (one big table) to optimize the performance for aggregate reporting.
    - Depending on the stakeholders and their use cases, the hybrid approach to some normalization will improve the data quality but impact reporting performance especially as the volume of data increases.
    - Some options to address the issue with volume growth include developing models tailored to specific use cases, sharding the data by time period as the need for analyzing all the data is not usually required for daily operations, and adding reporting views to precompile the aggregate reporting.
- Loading data in an incremental manner will be critical to ensure scalability and reliability of the data pipelines and downstream data transformations.
    - Required analysis to determine proper incremental window
    - Identify key fields to allow for data updates
    - Determine proper transactions activity for data whether it will be a full data replace or just update existing records