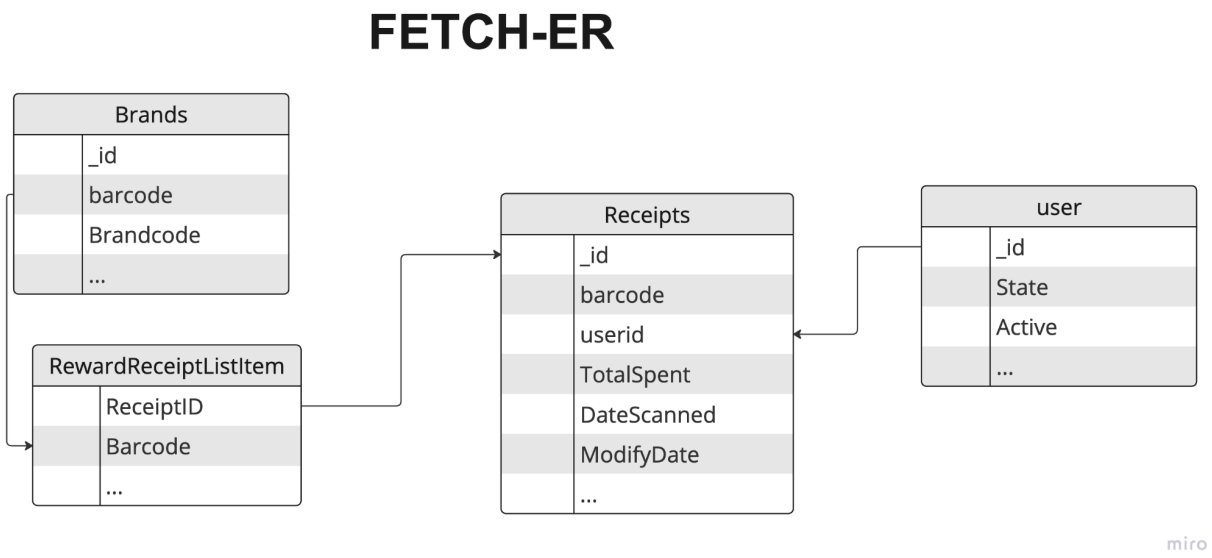
Fetch- Data Product Lead

Kajal Das
02/18/2024
Github: https://github.com/kajaldas22/

Questions to answer:

First: Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

ER Diagram (Image)



Second: Write a query that directly answers a predetermined question from a business stakeholder

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

   Comment: as data is for 2021, i have just selected for 2021-01 as the current month query returned no rows.

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

   The query should be union all between the current month vs previous month.

3. When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

Similar query as above, union all between accepted and rejected records to determine average spend

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

   Similar set of query as question#3

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

As data was not available for the last 6 months, in order to produce the result set, I used the 2021-01-01 date to bring the users.

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

   Similar query as previous one .

All the SQL i have uploaded into my github .

Third: Evaluate Data Quality Issues in the Data Provided

During the data analysis, I observed the presence of duplicates and sparse data within receipts and users tables. Following design I implemented in order to address the issues.

- Created a raw layer to ingest source data with basic cleansing
- In final layer, i have unpacked complex json data and removed the duplicates by using sql
- Designed a bridge table from one of the columns of the receipt table. This particular column contained packed JSON data containing receipts and associated scanned barcodes. This table served as an intermediary link to effectively manage and organize the data, ensuring that only unique and relevant information was retained in the final datasets.

Fourth: Communicate with Stakeholders

I wanted to reach out to discuss some key aspects of our data and how we can ensure it's in top shape to support our business goal.

- What questions do you have about the data?

  Are there any trends or patterns you would like to uncover from the data?
  Can you provide Insight into the specific business challenges or opportunities?

- How did you discover the data quality issues?

  While reviewing the data, observed high numbers of duplicates and sparse data in the transaction and bridge table(newly created) . These issues were identified through careful examination of data and validation. Are there any other data quality issues or inconsistencies we should be aware of?

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

  Resolving data quality issues requires a clear understanding of the expected standard or benchmarks you have in mind. Are there specific data quality metrics or criteria needed to focus?

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

As source or third party data is unstructured and with many data issue, we have to closely work with data providers to address the quality and accuracy of the data. This could involve implementing data cleansing and enriching before it is moved to the final table. Are there any specific optimization strategies you would like to consider?

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

As third party data would grow potentially specially transactions and bridge tables , additionally given source data would have duplicates. To address this concern and optimize our data pipeline and infrastructure , we are exploring parallel processing techniques , leveraging cloud based solutions, streaming etc. are there specific optimization or scalability requirements you would like us to consider or prioritize?