Dear Future Fetch Colleague,

I worked on the Fetch Rewards data provided and developed the data model for it. For moving ahead, it would be helpful if you could provide me with more details on the below topics.

A. Questions about the data
   1. Better working of Fetch Rewards ecosystem would be useful to understand the data cycle.
   2. I would like to understand the user journey and how that data we have captured represents that user journey.

B. Improvements in Data Quality

While creating the structured Entity Relationship Diagram for my data model, I discovered many data quality issues. I created Fact and Dimension tables using the data using the staging tables that I created with the raw data. While converting it into data marts, I found the following issues:

1. Users table had duplicate rows which were deleted using the SQL query.
2. There is a lack of relationship between the tables. The Item_receipts table has no key to uniquely identify each row. So, I created a key for it.
3. There are many nulls and missing values in the data making it harder to answer the business questions efficiently.
4. The dates mentioned in the receipts data were very redundant and very granular, requiring lot of space. Better understanding on which dates we're supposed to be used in the business questions would be beneficial.

C. To resolve the data quality issues, the following steps could have been taken:

1. Missing data could have been handled or a way to impute the missing values to treat it would have been better.
2. I would like to know how the receipts per item are generated and what links it to brands.

D. Other information to help optimize the data assets

1. Metadata and Documentation: The metadata or documentation on the data collection process, data definitions, and data lineage would help.
2. If you could also provide information on business rules and logic behind data generation and processing to ensure the analysis aligns with business expectations.
3. Access to historical data to analyze trends and patterns, which will improve the accuracy of our models.
4. Information on how frequently the data is updated and any related data refresh schedules.

E. Performance and scaling concerns:

1. As Fetch data grows, ensuring that our database can handle increased volume without degradation in query performance is critical. Our infrastructure must be able to scale horizontally to accommodate increased load. Utilizing cloud-based solutions with auto-scaling capabilities will be beneficial.
2. If real-time data processing is required, we need to ensure our systems can handle high-velocity data streams efficiently. Using technologies like Apache Kafka or similar for real-time data ingestion and processing can address this.
3. With large datasets, complex queries may lead to performance bottlenecks. Optimizing queries and using materialized views can help improve performance.

We can schedule a meeting to discuss these concerns if required. Kindly let me know if there are any questions from your side.

Thanks, and Best Regards,

Harsh Parikh

Analytics Engineer | Fetch Rewards