Gmail **Harika mangu <manguharika16@gmail.com>**

Fetch Rewards Project: Data Assessment Summary

Harika mangu <manguharika16@gmail.com> Thu, Feb 13, 2025 at 4:31 PM
Draft

Dear Product/Business Leader,

I hope this email finds you well. I've completed an initial analysis of our fetch rewards data and I have a few questions to ensure we are maximizing its value for our business objective.

1. Questions about Data?

- Are there any known gaps in our data collection process? I've noticed some inconsistencies in the data and want to ensure we're capturing all relevant touchpoints.
- What's our current data update frequency? Understanding this will help us provide the most timely insights and rewards to our users and come up with effective recommendations.
- What are the primary KPIs we're tracking? This will help us prioritize our data cleaning and analysis efforts
- Are there known seasonal trends in user behavior or reward redemption that we should be accounting for in our analysis
- I've observed some inconsistencies in how products are categorized within receipt item list.Is it because of inappropriate data entry or Is there a standard taxonomy we should be following to ensure accurate trend analysis?

2. How did you discover the data quality issues?

Data quality issues i have discovered during pipeline and analysis

- I found that date fields across different tables use varying formats.
- Upon inspection, I noticed that many essential fields such as user IDs,total spent, rewards partners sometimes contain null or empty values.This could impact our ability to calculate rewards and identify potential users
- Duplicate Entries: In all the tables, I identified multiple instances of duplicate records. It's unclear whether these are legitimate multiple submissions or data entry errors.Its confused me whether to drop or not while performing data quality evaluations.
- Outliers in Monetary Values: Some fields appear to have unusually high or low values, which may indicate data entry errors or could represent genuine outlier entries
- I noticed instances where foreign keys only referenced part of a composite primary key instead of the entire key. This situation highlighted the complexity of our data relationships and the importance of domain knowledge in making these decisions.

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

- Understanding whether problems stem from data entry errors, system integration issues, or process flaws is crucial for implementing lasting solutions
- Identifying issues such as duplicates, incomplete data, outdated information, or foreign key violations helps in targeting solutions
- Knowledge of data profiling tools, ETL processes, and data quality software capabilities is important for implementing fixes
- Particularly for issues like foreign key violations, understanding table relationships and data models is essential
- Understanding what different business units need from the data helps in setting appropriate quality thresholds

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

- Understanding how different teams and systems are using our data assets would help in prioritizing optimization efforts
- Establishing clear quality metrics for our data assets, such as accuracy, completeness, and consistency, would guide our optimization strategies
- Information on how each data asset contributes to specific business objectives or KPIs would help in prioritizing optimization efforts
- Understanding our current technology infrastructure and how data assets integrate with various systems would inform optimization strategies

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

- Projections of future data volume growth and new types of data the organization expects to handle would help in planning scalable optimization strategies by implementing data partitioning strategies and distributing data across multiple nodes to reduce load
- With increased data and user activity, query performance may degrade. We'll focus on query optimization and implement effective indexing strategies such as scaling horizontally on multiple clusters.
- Scaling our database architecture will incur additional costs. We'll explore serverless architectures and cloud solutions that can dynamically scale resources based on demand, optimizing both performance and cost
- To ensure continuous service, we'll set up database replication. This will improve fault tolerance.This helped me during pipelining

Understanding these aspects will greatly assist in refining our data strategy and ensuring that our insights align closely with business objectives. I'm happy to discuss these points further at your convenience

Thank you,

Best Regards,
M. Harika