Fourth: Communicate with Stakeholders

Hello Mike,

As an Analytics Engineer, my focus is on ensuring data accuracy, optimizing ETL processes, and improving performance for better reporting and decision-making. In reviewing the data, I have identified several key issues that impact usability and performance. I wanted to highlight these challenges and discuss next steps to enhance data quality and efficiency.

Key Data Quality Issues:

- Some fields contain multiple separate JSON objects, making it difficult to parse and process efficiently. We should explore whether we can structure this data differently to eliminate the need for complex parsing.
- Some fields contain mixed data types (e.g., numbers stored as strings), which can cause processing errors and impact reporting accuracy.
- Many fields, especially in Receipts and Brands tables, contain NULL or blank values.
- Points and rewards-related fields are frequently empty. Should these be replaced with 0, inferred from other data, or handled differently?
- The Users table has significant duplication, which could affect downstream analysis. We should clarify how users are being ingested and whether deduplication should occur at the source.
- Some columns contain identical information and may be unnecessary.
- Critical fields, especially around rewards and points, might be missing. We should define which fields are essential for reporting and which can be omitted to improve efficiency.
- Date-related fields require validation ensuring consistency in format and identifying which ones are meaningful for analysis.
- The ReceiptsItems table contains various columns currently residing in a single table, leading to data redundancy. If we can obtain relevant metadata, we could normalize this data, improving processing efficiency and optimizing our database.

Questions & Next Steps:

- Data Structure – Can we modify how this data is captured to eliminate JSON parsing issues?
- Business Logic on Rewards – What do different rewards-related fields represent, and how should missing values be handled?
- Essential vs Non-Essential Data – Can we define which fields are critical for reporting versus those we can remove to reduce redundancy and optimize queries?

Performance & Scalability Considerations:

> As our data volume grows, it’s important to capture it correctly to avoid issues down the line. Cleaning and restructuring the data will help improve query speed and processing efficiency. By taking the right steps now, we can keep our data fast, reliable, and scalable as it continues to grow.

Would love to discuss this further and align on the best approach. Let me know your thoughts.

Thanks,
Nitiraj