Subject: Insights and Next Steps on Data warehouse Project

Dear Business/Product Lead,

I hope this message finds you well.

I'm reaching out to share some updates on our data warehouse project, where we've been exploring our user engagement data with receipt and brands data to extract actionable insights.

1. Here are some questions that would help me understand the data better

    - How is the data collected across different platforms ?(e.g., mobile apps, websites) since I have found redundant records in users data.

    - How significant are the issues of missing data and outliers within our current data?

    - Are there specific business questions or goals that this data is intended to address?

2. I have a few questions regarding the Data Quality issues found during data validation using Python Scripts.(detailed analysis is done in the Github repo)

    - How are we planning to handle items which are not labelled in our database? For example, those items' barcodes are common i.e. '4011', which will make it difficult to identify different items.

    - Are there existing protocols for data validation and cleanup that we might improve or implement at earlier stages? Understanding this could help us pinpoint where inaccuracies are introduced, since there are a lot of missing values in all the data tables.

    - Data redundancy is also an issue in the current data, which needs to be solved by following Third Norm or probably using a non-relational database for specific purposes.

3. To optimize our data assets further and ensure robust analysis, additional information would be invaluable, such as:

    - Insights into user behavior and preferences not captured in our current datasets.

4. Looking ahead, as we scale our data analysis efforts, I anticipate potential challenges in handling larger datasets. To address these concerns, I plan to:

    - Implementing indexes on key columns used in search queries will be crucial for speeding up data retrieval times. For example, indexing user_id in receipts and brand_id in brands can significantly enhance performance.

    - We'll implement a data retention policy to archive old data and purge it from our active databases, which will help maintain performance and reduce storage costs.

I believe that addressing these points will significantly enhance the value we derive from our data and support informed decision-making across our projects.

Thank you for your time, and I look forward to hearing from you..

Best regards,
Aditya Sahu