Hi,

I am Apoorva Gupta, Data Analytics Engineer of ABC team. I recently conducted an analysis of our user activity, receipts, and purchase data to help the business to take data driven decisions. During this process, I identified several critical data integrity issues that could potentially impact our analytics and reporting. I wanted to share the key findings with you, along with their potential business impacts:

Key Issues & Business Impact

1. Missing User IDs from users database
    - Our users database doesn't have a complete list of user IDs.
    - Impact: This discrepancy can lead to data integrity issues, causing inaccurate reporting and inconsistencies in user-based analysis.
2. Duplicates data in user database
    - We have duplicate data in the users database.
    - Impact: This can cause performance issues, increased storage costs, and affect data integrity.
3. Redundant data and Inconsistent data in brands database
    - We are storing the same data in different fields, often with inconsistencies.
    - Impact: This increases storage costs and reduces database efficiency. Formatting differences can lead to inconsistent data retrieval and incorrect reporting.
4. Unstandardized naming convention of brand name
    - We have inconsistent naming conventions for brand names.
    - Impact: This can cause misalignment in reporting and tracking brand-level key metrics.
5. Missing brandcodes from brand database
    - Only Jan, 2021 brand codes are available in brands database
    - Impact: It can lead to inconsistency in your data, as some products may be linked to a brand, while others are not. This could result in incomplete records or mismatches during analysis, reporting, and decision-making.

The issues mentioned above, particularly redundant and inconsistent data, can slow down the system, increase storage costs, and cause performance issues as data grows. To address this, we need to:

- Clean and optimize the data.
- Ensure that only unique and standardized records are stored.
- Automate the data cleaning process to handle scaling challenges efficiently.

As our data volume increases, these issues will likely worsen, potentially slowing down queries and causing delays during high-traffic periods. To mitigate this, we need to enhance data storage, indexing, and optimize the system for handling large data loads.

There are some questions that I have regarding the data that might help us to resolve the data quality issues:

1. How are new users added, and is there a check to avoid duplicates?
2. How is brand data entered, and is there a system to standardize it?
3. How often do these errors happen? Can we set up alerts to catch them early?
4. Why are barcodes missing for recent products, and how can we update this regularly?
5. Do we have a process to clean and transform the data before using it?
6. How is this data being used by the business, and what are the key reporting or decision-making processes?

I have scheduled a call on Thursday to discuss the above in detail. Let me know in case of any time conflict. Please feel free to add any other team members who should be involved.

Thanks & Regards,
Apoorva Gupta
Data Analytics Engineer