Subject: Data Quality Assessment and Recommendations for Enhanced Data Assets

Dear [Stakeholder/ Leader],

I hope this message finds you all well. I am writing to provide an update on our recent data quality assessment and highlight the measures we recommend taking to improve the reliability and usability of our data assets. As we continue our data-driven initiatives, it is crucial to ensure the integrity and consistency of our data for informed decision-making.

First allow me to present a summary of my findings:

1. Data Completeness: I identified several instances of missing data across our three main data schemas: Receipts, Users, and Brand.
    - Receipts Table: Approximately half of the records are missing values for bonusPointsEarned, pointsEarned, finishedDate, pointsAwardedDate, and purchaseDate. These missing data points limit our ability to accurately track bonus points, transaction dates, and related metrics.
    - Brand Table: Over 50% of the records lack values for categoryCode and the topBrand indicator, which hampers our ability to categorize brands and identify top brands effectively.
    - Users Table: There are some missing values in the state and lastLogin fields, which may impact our user analytics and activity tracking.
2. Data Integrity and Consistency: During the assessment, I discovered 283 duplicate records in the Users Table. Also, some userIds in the Receipts table do not correspond to existing records in the Users table.
3. Data Reliability: During the analysis, I encountered instances where a single barcode was associated with multiple brands in the Brand table, which raises concerns about potential data entry errors. Additionally, some receipts with a 'FINISHED' status still have unknown bonus points. We need to address whether these missing values indicate zero bonus points or if there are updates pending.

To optimize the data assets we are creating, I kindly request the following information and collaboration:

1. To ensure data completeness, I kindly request the missing values in three main schemas. It would help if we could perform a comprehensive search for

missing brand information based on the barcodes recorded in the Receipts table. Additionally, I suggest enhancing our data collection processes to ensure the completeness of future data entries.

2. To ensure data integrity and optimize the unique identifier, I recommend removing duplicates when entering new data points and designating the '_id' field as the primary key of Users table.
3. To ensure consistency and accuracy across our data assets, I kindly request support in investigating the userIds that only exist in the Receipts table, and updating the userIds in Users table based on the existing records in the Receipts table.
4. To ensure reliability and timeliness of our data analytics, I need support in rectifying any data entry errors and checking the frequency of data updates.

Above all, It would be extremely helpful if you could provide us with specific business objectives and details regarding the key metrics. Understanding the will enables me and the team to focus on the most relevant data points and optimize the data schema accordingly.

As moving towards production, I anticipate potential challenges related to performance and scaling that could require proactive monitoring and suitable solutions. My first concern is data scalability. First of all, future data growth may require distributed computing or cloud-based solutions. Also, when querying large data sets, we need to implement appropriate indexing strategies, and consider partitioning to ensure efficient data retrieval. Data processing speed is another concern. During ETL, we need to continuously evaluate and optimize our data processing pipelines, encompassing data ingestion, pre-processing and analytics to ensure timely and responsive data insights. To further address these concerns, I'll schedule meetings and discuss plans with the data engineering team in detail based on our business goals.

Should you have any questions, require further clarification, or wish to discuss these matters in more detail, please do not hesitate to reach out. I look forward to your valuable input and collaboration.

Thank you for your time and consideration.

Best regards,

Wenxuan Zhou | Data Analyst
[Contact Information]