Email Subject: Addressing Data Quality Issues and Optimization Strategies for Data Assets

Dear Stakeholders,

I hope this message finds you well. I am reaching out to discuss some critical observations and questions regarding the data quality issues we've identified in our datasets. The goal is to ensure the integrity and accuracy of the data assets being created, which are vital for informed and data-driven decision-making, and optimal performance of the product.

During a recent data analysis, we discovered several inconsistencies and gaps across multiple datasets involved in the smooth performance of our product. I would like to outline these findings, pose some essential questions, and seek your guidance on the next steps to resolve these issues efficiently.

Questions Regarding the Data:

1. Receipts Data:
    a. Can you provide clarity on the conditions under which **`finishedDate`** and **`pointsAwardedDate`** might be missing?
    b. Are there specific business rules or processes that lead to the variability in **`bonusPointsEarned`** and **`pointsEarned`** values?
    c. How should we handle missing or inconsistent **`rewardsReceiptItemList`** entries to ensure accurate bonus point allocation?

2. Users Data:
    a. What are the possible reasons for gaps in the **`lastLogin`** field?
    b. Is there a defined period after which inactive users should be flagged, or is this process manual?

3. Brands Data:
    a. How critical is it to have complete **`categoryCode`**, **`brandCode`**, and category fields for our analysis?
    b. Are there any guidelines for imputing missing values in **`topBrand`** to maintain data consistency?

4. General Data Quality and Consistency:
    a. What protocols are in place for data validation and consistency checks before data is ingested into our systems?
    b. Are there historical data correction procedures we should be aware of to address these inconsistencies?

Discovery of Data Quality Issues:
The above issues were discovered through a series of systematic data analysis techniques, such as,

- Exploratory Data Analysis (EDA): Initial inspections of the datasets revealed anomalies in data types and missing values.
- Data Profiling: Detailed examination of each field highlighted inconsistencies.
- Null Value Assessment: Running null value checks indicated significant gaps in crucial fields.
- Data Type Verification: Discrepancies were found in timestamp fields.
- Duplication Check: Duplicate user lists were detected, which could potentially skew analysis and reporting.

Resolution and Optimization:
To effectively resolve the quality issues with our data assets, we need to address the following three key areas at a minimum.

1. Understand the business rules behind data generation and any acceptable ranges or values for the fields.
2. Determine if there are any automated or manual processes for flagging and correcting inconsistencies.
3. Clarify any knowledge from a domain expert that would help us accurately impute or appropriately handle missing values.

Performance and Scaling Concerns:
Given the volume of data and the need for real-time analysis, we anticipate challenges related to:

1. Efficiently processing large datasets while maintaining performance.
2. Ensuring scalability of our data validation and correction processes as the data grows.
3. Assuring accuracy of the results while updating the training data on periodic intervals to reflect current trends and consumer patterns.

Identifying these issues is a critical step towards ensuring the integrity and reliability of our data assets. I recommend a thorough review and correction process to address these inconsistencies. Please let me know how we can best collaborate to resolve these concerns and optimize our data for accurate analysis. I look forward to your thoughts on this matter.

Thank you for your time and assistance.

Best regards,
Niharika Chunduru