Subject: Rewards Data Quality Issues and Optimization Opportunities

Hi Team,

I hope you're doing well. As part of our current efforts to optimize the rewards data processing pipeline, I've analyzed our current data sources for receipts, users, and brands. Below are the key issues requiring attention and possible opportunities for improving the data structure to enhance usability:

1. **Data Quality Concerns:**
   An extensive Exploratory Data Analysis (EDA) was performed to gain insights into our available data sources and identify key quality concerns. The formal methodology used to identify these issues is detailed in the attached **"Data Quality Report.pdf."** Below is a summary of the findings:

   1. **Missing Values**
      - R**eceipts table** shows significant data gaps, with 60% of fields having more than 35% missing values.
      - **Brand table** has similar concerns with key columns like **categoryCode** and **topBrand** having over 50% null values.

   2. **Duplicate Records**
      A notable number of user records appear to be duplicated, with **identical _id, createdDate, lastLogin**, and other fields. This redundancy can lead to inaccurate user counts and skew our analysis.

   3. **Unmapped Users**
      Receipts table contains 117 **user_ids** that are missing in the user table, indicating gaps in the user-mapping process.

   4. **Incomplete Data**
      Some receipts contain placeholder entries, such as items marked as "ITEM NOT FOUND," which represent 5% of transactions by the total amount spent. For accurate analysis, these items need to be added.

5. **Anomalous Receipts**
   Certain transactions are marked as rejected (**status=REJECTED)**, but have valid points earned (**pointsEarned>0).** Additionally, there are valid receipt IDs with $0 spent values (**totalSpent** = 0). We need to understand the origin of these receipts to evaluate if these are expected cases or if further analysis is required.

2. Optimization Opportunities

1. **Brand Enrichment**
   Implement a dimensional table for brands containing comprehensive brand information. This approach ensures data consistency and simplifies brand associations across datasets.

2. **User Enrichment**
   Adding additional user profile details such as age, gender, preferences (product categories, brands), or loyalty program information, would aid in User Analytics / Segmentation.

3. **Data Standardization**
   Incorporating quality concerns from Step 1 such as missing user data, incomplete item information, and standardizing brand data would help us to strengthen our data model and build robust analytical pipelines.

3. Additional Information Needed for resolution/enhancement

To address the above data issues and optimize our available data assets, the following information would be invaluable:

1. **User Behavior Documentation**
   Addition of data sources that provides additional insights into user interaction with the mobile app would help with better segmentation and improve our analysis of available data feeds.

2. **Data Collection Process**
   An additional summary of the current data collection workflow (e.g., user input versus automated scanning) would help identify the root causes of

barcode inconsistencies, missing user data, and other anomalies. This would also allow us to suggest targeted improvements to the data pipeline.

3. **Additional Data Sources**
   Access to supplementary data on brands and user demographics would enable us to enrich existing datasets. This would be particularly beneficial for advanced use cases such as user segmentation and behavior analytics.

4. **Business Logic Input**
   A business definition document for bonus point calculations and user activity tracking would enhance our ability to validate receipt data and ensure consistency with the intended logic.

4. Performance and Scaling Considerations:

Once we gain clarity on the identified issues and receive the required information, we can implement data cleaning and schema enrichment procedures to enhance data model efficiency. To address potential performance bottlenecks, we can:

1. Explore data partitioning and indexing strategies based on key identifiers.
2. Optimize queries to improve processing efficiency.

To solve these data issues and incorporate the new optimization opportunities, I have prepared an **Agile-based** action plan, which I would like to discuss further with the team. Please let me know your availability so that I can schedule a quick internal session.

Thanks,
Gaurav Khatri