Subject: Action Required: Data Quality Issues and Optimization Opportunities

Hi Team,

I hope this message finds you well. As part of our ongoing efforts to optimize our data processing pipeline, I have analyzed our current data sources related to receipts, users, and brands, and have identified some issues that we need to address before we can fully utilize this data. Additionally, I've also identified some opportunities to optimize the data structure for improved usability, summarized below.

1. **Data Quality Issues:**
   I performed an extensive Exploratory Data Analysis suited to the nature of data, to understand available data sources, and noticed a few concerns. The formal methodology for identification of these quality issues is well described in the **"Data Quality Report.pdf"** document, attached to this email.

- **Missing Values:** Receipts and brand tables are highly prone to missing/null values. For the receipt table, **60% of the data fields** have more than **35%** missing values. Similarly, the brand table has significant missing columns, with **categorycode** and **topbrand** columns having more than **50%** null values.

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

- **Missing Users:** The receipts table has 117 **user_id's** that are missing from user table. This presents a clear gap in our user-mapping procedure.

- **Incomplete Data**: Certain receipts have missing or placeholder data, such as items marked as "**ITEM NOT FOUND**", which corresponds to **5%** of transactions by amount spent.

- **Anomalous Receipts**: Receipt data has transactions with **status=REJECTED**, yet with **pointsEarned** > 0, corresponding to . Similarly, receipt dataset consists of valid receipt id's with **totalSpent = 0.** We need to understand the origin of these receipts to evaluate how to handle these cases.

2. **Optimization Opportunities**

- **Brand Enrichment:** It would be beneficial to have a consistent **brand_id** field that links to a reference table containing additional brand information. This approach promotes data consistency and simplifies brand association across different datasets.
- **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.
- **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 solve the above data issues, and optimize the data assets, the following information would be highly useful.

- **User Behavior Documentation:** Understanding how users interact with the mobile app to flag new items or prices would help to interpret several data fields such as **userFlagged** data points.
- **Data Collection Process:** Information about the data collection process (e.g., user input vs. automated scanning) would aid in pinpointing the root cause of barcode inconsistencies and missing user data, and help in providing suggestions to improve the data pipeline.
- **Additional Data Sources:** Information about additional data sources regarding brands/user demographics, would aid in enriching these sources with relevant information for complex use cases such as user segmentation and behavior analytics.
- **Receipts Validity:** Further documentation on the origin of receipts and how each of the fields is populated is needed to validate receipts with 0 amount spent or the ones with invalid descriptions.
- **Business Logic Input:** Detailed business logic for bonus point calculation and user activity would enable us to validate the receipt data.

4. **Performance and Scaling Considerations:**

   Once these issues are understood and further information is provided, we can incorporate **data cleaning** and **schema enrichment** procedures as required to

improve the performance of the data model. To address performance bottlenecks, we can explore data **partitioning/indexing strategies** based on key identifiers and optimize queries to improve efficiency.

To approach the above concerns, I have prepared an **Agile-based** action plan, which I would like to discuss further with the team. Let me know what time works best for a quick meeting.

Thanks,
Gaurav Khatri