To: Product Manager
Cc: [Business Stakeholders]
From: Vishnusai Bhadramraju
Date: [Today's Date]

Subject: Data Quality Concerns and Optimization Strategy for Sales data

Hi Product Manager,

I hope this message finds you well. I wanted to update you on the recent data analysis project we have been working on, specifically regarding the receipts, brands, and users' data. Here are some key points I'd like to discuss:

Questions About the Data

1. **Data Completeness**: Are there specific fields in the receipts, brands, and users' data that are critical for your analysis?
2. **Brand Associations**: How do we ensure that each receipt entry correctly associates with the correct brand entry?
3. **User Insights**: What specific user metrics are you most interested in tracking over time?

Data Quality Issues

These issues were uncovered during our initial data exploration phase using descriptive statistics and visualization techniques. After a cross review confirmed the presence of these anomalies.

- **Missing Values**: There are numerous fields with missing data across the receipt, receipt items and users datasets.
- **Inconsistent Formats**: The formats of some fields (e.g., dates, prices) are inconsistent, leading to difficulties in standardizing the data.
- **Nested Structures**: The JSON data contains nested structures that are not directly compatible with our current analysis tools.
- **User ID Mismatch**: Some **user_id** values in the receipts table are not present in the users table.
- **Null Purchase Dates**: The **purchase date** column in the receipts table has null values, which cannot be dropped due to the uniqueness of receipt IDs.
- **No Relation Between Receipts and Brands**: There is no clear way to relate the brands table with the receipts table, making it difficult to link purchases to specific brands.
- **Require additional Info**: Require information about products that match products of each company might help to analyze more about different products and their performance

Key Concerns

1. **User ID Discrepancies**: What is the correct approach to handle **user_id** values that appear in receipts but are missing from the users table? Should these entries be excluded or is there a way to reconcile them?

2. **Brand and Barcode Relationship**: Can a single brand name have multiple barcodes? Understanding this will help in mapping receipts to the correct brand.
3. **Date Fields in Receipts**: We need clarification on the relationships between **create date**, **date scanned**, **date finished**, and **purchase date** in the receipts table. How do these dates interact, and which one should be prioritized for analysis?

Resolving Data Quality Issues

To address these issues, I need to understand:

Data Source Details: Understanding the origin and collection process of the data to identify any gaps or inconsistencies at the source.

Collaboration with Data Owners: Access to data owners or SMEs who can provide context and confirm the accuracy of critical fields.

And also mindful about these:

- **Field Importance**: Which fields are mandatory and which ones can tolerate missing values?
- **Validation Rules**: Are there any existing validation rules or data standards we should enforce during the data cleaning process?
- **Source Accuracy**: Can we get clarity on the most accurate source for resolving inconsistencies in the data?

Additional Information Needed

To optimize our data assets, it would be helpful to have:

- **Data Dictionary**: A detailed data dictionary explaining each field in the datasets.
- **Product info**: Details about products and their barcodes of a company's tables.
- **Business Rules**: Any business rules or logic that should be applied when transforming the data.
- **Historical Data**: Access to historical data to understand trends and outliers better.

Performance and Scaling Concerns

In production, we anticipate the following concerns:

- **Data Volume Growth**: As our data volume grows, ensuring that our processing pipelines can scale efficiently is critical. We plan to implement distributed processing frameworks (e.g., Apache Spark) to handle large datasets.
- **Query Performance:** Ensuring that our databases can handle complex queries without significant latency. We will optimize indexes and consider database partitioning strategies to maintain performance.
- **Data Storage**: Efficiently managing storage to balance cost and performance, potentially exploring cloud-based storage solutions with auto-scaling capabilities.
-

We plan to address these by leveraging cloud-based data solutions, optimizing our data queries, and using distributed computing frameworks.

Your input on the above points will be invaluable as we move forward. Let's schedule a meeting this week to discuss these aspects in detail.

Best regards,