Subject: Data Quality and Optimization Inquiry for Improved Analytics Performance

Hi [Business Leader's Name],

I hope this message finds you well. In the course of enhancing our data analytics capabilities, I have **conducted a preliminary data quality** assessment that **revealed a few areas requiring attention** to ensure our analyses drive the most accurate insights. I'd like to share these findings and seek further guidance to address these effectively.

1. **Data Quality Findings:**
    - A significant portion of our dataset in the `receipt_items` table, specifically in the `item_needsFetchReview` (88.29%) and `item_needsFetchReviewReason` (96.84%) columns, consists of missing values. This **high percentage of missing information may affect the reliability** of our analytics outcomes.
        - `receipt_items` is a table derived from originally `receipts` table for better database management
    - Additionally, our preliminary review of the `users` table indicates approximately 57% of the entries could be duplicates, which could **skew user engagement metrics and insights.**
2. **Discovery Method:**
    - The **missing values were identified through SQL queries** that calculated the percentage of NULL entries in specific columns.
    - The duplicate records in the `users` table were **detected using python's `pandas` library.**
3. **Information Required to Resolve Issues:**
    - For the missing data in `receipt_items`, **could you provide insight into whether these fields are critical for our analysis** and **if there are default or inferred values** that can be safely assumed?
    - Regarding the duplicates in the `users` table, the presence of such a high proportion of potential duplicates (approximately 57%) **raises concerns about our data collection and ingestion processes. Clarifying the criteria that define a unique user within our systems** (e.g., whether each `user_id` should be globally unique, or if email addresses are considered a secondary key) is crucial. This information would not only assist in refining our data cleaning protocols but also in identifying any underlying issues in how user data is collected or recorded. **Addressing this at the source could significantly improve the integrity of our user data** and, by extension, the accuracy of our user engagement analytics.
4. **Additional Information Needed:**
    - To optimize our data assets further, it would be **beneficial to know if there are specific analysis goals or questions** that the business aims to answer. This direction can guide our data preparation efforts to ensure we focus on the most relevant data points.
    - **Any historical context about common data collection** or entry issues could also help tailor our data cleaning strategies more effectively.
5. **Performance and Scaling Concerns:**
    - As we scale our data analytics efforts, maintaining data quality becomes increasingly challenging. **Implementing automated data validation checks** during ingestion could help mitigate these issues early on.

- For performance, **designing a more efficient data storage and querying structure**, such as **indexing critical columns**, can ensure our analytics queries remain **fast and reliable as the dataset grows**.

I believe addressing these points will significantly enhance the quality and reliability of our data-driven insights. Could we schedule a meeting to discuss these findings and the way forward in more detail? Your expertise and perspective would be invaluable in shaping our data strategy.

Thank you for considering these points. I look forward to your guidance on these matters.

Best regards,

Shivam