Subject: Data Quality Assessment and Optimization Plan

Dear [Business Leader],

I hope this email finds you well. I wanted to provide you with an update on our recent analysis of the data assets and highlight some key findings and considerations.

Questions About the Data:

In reviewing the data, several questions have arisen regarding the meaning and relevance of certain columns. For example, we are unclear about the purpose of many columns (partnerItemId,preventTargetGapPoints,pointsPayerId,rewardsProductPartnerId,originalMetaBriteBarcode,originalMetaBriteDescription,competitorRewardsGroup,originalMetaBriteQuantityPurchased,competitiveProduct,metabriteCampaignId) in the "rewardsReceiptItemList". Additionally, there is confusion surrounding the "createDate" and "modifyDate" columns in the "receipts" dataset. We need clarification on what events these dates correspond to.

Discovery of Data Quality Issues:

Our initial data exploration revealed numerous quality issues, including high percentages of null values in several columns of the "rewardsReceiptItemList". Specifically, we identified discrepancies such as some receipt IDs present in the 'rewardsReceiptItemList' table not being found in the 'receipts' table. Similarly, certain user IDs from the 'receipts' table were not matched with corresponding entries in the 'users' table". Furthermore, inconsistencies in date formats and improper representation of boolean values were noted. Duplicate records across multiple columns further complicated the analysis.

Resolution of Data Quality Issues:

To address these challenges, it would be beneficial to gain a deeper understanding of the underlying metadata and business model driving our data. Additionally, increasing the volume of data available for analysis would provide greater insights. Understanding the relationships between different tables and how they reflect real-world processes will be crucial for resolving data quality issues effectively.

Optimization of Data Assets:

To optimize our data assets, we need to consider additional information such as data lineage, data governance policies, and data retention requirements. By aligning our data assets with business objectives and ensuring data quality and integrity, we can create more valuable insights for decision-making.

Performance and Scaling Concerns:

As we move towards production, it's important to anticipate performance and scaling concerns. This includes ensuring that our infrastructure can handle increased data volumes and processing requirements. Implementing strategies such as data partitioning, indexing, and caching can help optimize performance and scalability.

I hope this overview provides clarity on our data analysis efforts and outlines the steps needed to address data quality issues and optimize our data assets. Please feel free to reach out if you have any further questions or if you require additional information.

Thank you for your attention to this matter.

Best regards,

Gops