Subject: Data Review and Next Steps

Respected Sir/Mam,

I've been reviewing our data to ensure it is accurate and reliable. I've identified some data quality issues that could impact our ability to effectively analyze user behavior and make informed decisions. I have few questions about the data like What system generated this data, and how is it collected? What constitutes a unique user in this dataset? Is the ‘\_id’ field in user, truly intended to be a unique identifier? What is the expected frequency of duplicate entries, if any? Are there existing data validation processes in place?

For getting some answers about the data quality, I run a script that identifies common issues in the data like duplicate values, outliers, inconsistencies, and unusual characters. That gave me some answers about the data inconsistency. For resolving some of these common issues, my suggestion would be to look for why the data is being duplicated, clarifying the intended unique identifier that will guide the de-duplication process, removing unnecessary characters that make the joins process problematic, and cleaning data so that null values, accurate consistent data can be available in the table. To further optimize our data assets, I would like to understand what the key business objectives we’re trying to achieve with this data are. What reports and dashboards are needed, and what metrics are most important? How is the data currently being used, and what other potential use cases exist? By understanding the growth rate of this data, I can ensure the solutions I build today work well into the future.

In a production environment, I anticipate the following performance and scaling concerns like large datasets can impact query performance and processing time, complex data transformations and aggregations can be resource-intensive, storing and managing growing data volumes can be challenging, ensuring the security and privacy of sensitive user data is critical. To address such concerns, we can take some measures like implement data indexing, optimize data processing pipelines, consider cloud-based data storage like AWS S3, implement data security measures, monitor performance and scale infrastructure, automate as much as possible.

Please let me know if you have any questions or would like to discuss this further.

Best regards,

Kirtan Pathak