Hi Product Leader,

Trust you are doing well and having great day.

In regards to the current project, I wanted to clarify certain key points regarding provided data. Below are the highlighted points:

* What questions do you have about the data?

1. Are there any metadata or documentation regarding this data?
2. What is the source and size of data?
3. What is the format of data from the source?
4. What is the data ingestion looks like real-time, near real-time or batch?
5. Are there any known limitations and biases that you know of when the data was collected?
6. In terms of feature selection is there certain feature which overweighs other considering business KPI?
7. In regards to null value handling, do you have any preferred method for pre-processing?
8. Are there any data security considerations while processing it?

* How did you discover the data quality issues?

1. I started with data profiling using python libraries to get statistical inferences like min, max, mean, etc.
2. Utilizing EDA techniques to visualizing the data.
3. Performed data completeness assessment to see the null values in features of the dataset.

* What do you need to know to resolve the data quality issues?

1. I would require domain expertise to understand the procedure for alteration of missing data.
2. Connect with data manager to understand the underlying data structure and reasoning for the same as the dates are stored in UTC format and several columns are stored as dictionary.
3. Understanding the current data types used for each column and reasoning behind it.

* What other information would you need to help you optimize the data assets you're trying to create?

1. Evaluation of need of data indexing with stakeholders for better speeds of data retrieval.
2. Assessing the possibility of data compression to lower the storage cost and improve data transfer efficiency.
3. Comprehending the possibility of aggregating data as it reduces data volume and complexity.

* What performance and scaling concerns do you anticipate in production and how do you plan to address them?

1. Increasing Data Volume: As data volume rises, query performance and storage needs may be impacted. Consider using data segmentation, indexing, and compression techniques to solve this problem. Larger data volumes can also be handled by scaling up physical resources or implementing distributed data processing platforms like Apache Spark.
2. Heavy data processing operations, such as aggregations or intricate transformations, can constitute bottlenecks in the data processing process. Use distributed computing frameworks to spread and parallelize processing activities, such as Apache Hadoop or Apache Spark. To reduce processing overhead, consider employing precomputed aggregates or materialized views for frequently requested data.
3. Monitoring and maintenance: As the project grows, it is critical to keep an eye on the system's functionality, stability, and dependability. Maintaining a scalable and reliable data science project requires the implementation of monitoring tools and procedures to find bottlenecks, errors, or anomalies.

These were all the questions and concerns regarding data files provided. Do let me know if we can connect to discuss the above findings.

Looking forward to hear from you.

Best Regards,

Eeshan Pancholiya