Hello ABC,

Hope you are doing well!

I am currently working on three datasets provided by your team. I have created a new structured relational data model where I encountered some issues and wanted to bring it to your notice and would like to request additional information from you so I can take it further.

Some of my questions are:

1. Can I know the source of the data? and how this data is being generated?

2. How is the data stored and where is the data located?

I have encountered some issues like duplicate values, null values, and more. I used Python to explore the data by using various methods, and techniques to understand the data better. Resolving the data quality issues by removing the null values, spatial data and sparse data is important, and so I performed some pre-processing to treat the data. Some of the issues which can be resolved are by understanding about the data generation process, attributes we store and their relation.

I also have some concerns with performance and scalability, when the data increases in huge amounts, we might move to a distributed environment like spark, where we can improve data processing. We can create a single pipeline to automate the whole process.

Please let me know if you'd like to have a quick call about this issue, please let me know your availability this week.

Kind Regards,

Tharun Kanamneni