**Subject: Follow-up on Data Quality and Optimization for AE\_Role Assignment**

Dear Stakeholder

I wanted to provide a brief update on the recent findings and next steps regarding the data quality and optimization of the dataset for AE\_Dataset

**Questions about the Data**  
Before proceeding with further optimizations, I have a few clarifications that would help streamline our approach:

* Could you confirm the data sources we are pulling from and whether any sources are subject to change?
* Are there any specific business rules or transformations that should be applied during the data preparation process that we may have missed?
* What is the refresh time period for the data ?
* Are there any active indicators or flags that I need to consider while performing the data analysis
* Are there any default values for dates like 1800-01-01 when there is no dates , are there any specific timezones datetime need to stored in ?
* What is the case type that needs to be followed for naming tables and columns
* What is the range of data you would need ? like past 12 months ?
* Is there any test data included in this ?

**Discovery of Data Quality Issues**  
During my analysis, I identified a few quality issues, including missing values, duplicates, incorrect formats, etc., These were discovered through a combination of automated validation checks and manual reviews of the datasets.

**Requirements to Resolve Data Quality Issues**  
To address these concerns, I need additional information regarding:

* The expected formats and value ranges for the impacted fields.
* Whether there are any upstream issues from the data sources that need attention.

**Performance and Scaling in Production**  
To ensure smooth performance in production, I'm exploring several options, such as:

* Implementing query optimizations and indexing to reduce query response times.
* Partitioning the dataset to enhance both storage and retrieval efficiency as it grows.
* Monitoring concurrency levels to ensure stable performance with multiple users querying the data.

I’ll continue to monitor the data quality and performance closely and provide further updates as we progress. Please let me know if you have any additional insights or requirements that can assist with this process.

Best regards,  
Karthik