What questions do you have about the data?

- What KPIs would you like to discover from the data?
- Could you please specify the columns per report?
- What will be the filter criteria per report?
- Are there any data formatting requirements?
- What are typical analytical requirements?

How did you discover the data quality issues?

I was analyzing the data using Pandas Dataframe and noticed several DQ issues related to Accuracy, Completeness, Consistency, and Uniqueness. I will need help with some Data Validation rules for capturing business metrics and its thresholds? I would like to use those in automation to detect data errors before reports are published.

Some data quality issues I discovered were:

1. Duplicate users.
2. Missing values for mandatory columns.
3. Missing links between reward items and referenced dim tables. For example, some of the barcode are missing for records in reward item table

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

1. Establish standard DQ check per table or the event payload for Data Accuracy, Consistency, Completeness, and Uniqueness
2. Review those checks with business users
3. Regularly get feedback from users
4. Applying a known json schema for the event payload?
5. Identify sequence of arrival of data in the system
6. Building knowledge on the schema and performing root cause analysis

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

- What are the sources of the data?
- The velocity at which data arrives in the system?
- What's the typical growth pattern of the data?
- What are refresh cycles of the data?
- Analyze logical and physical plan of the query and fine tune query ongoing basis
- What are the required SLA and performance requirement of each report
- KPIs and it's for each report

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

1. Increasing size of resultset of a query can slow down the performance
    a. Design efficient schema to remove unwanted data
    b. Optimize the query on going basis
2. Scalability issue due to concurrent access of data by multiple users
    a. Add more compute resources that can handle the concurrent query executions by multiple users
3. Query failure due to memory issue
    a. Conduct performance testing and identify what compute resources will be appropriate to handle the target load