Team 8: Michelle Xie, Elbek Gaybullaev​, Theophilus Ghartey​, Jingsi Hu​

Dr. Fogel

BAN5573

30 April 2023

**Interview Case**

**Communication with Stakeholder Questions**

1. What questions do you have about the data?

After working with the data, we have the following questions about the data:

* How was the data collected? What is the source of the data?
* What is the intended purpose of the data?
* Are their certain limitations that should be taken into consideration?
* Is there a reason that certain information is missing?

1. How did you discover the data quality issues?

**Data profiling**: involves analyzing the data to identify its structure, patterns, and quality characteristics. It could include reviewing data statistics such as min, max, and average values, checking data distributions, and analyzing missing or outliers.

**Data visualization:** Data visualization tools to create charts, graphs, and dashboards that make it easier to spot data quality issues. For example, to identify any correlation or trends between data points to identify data outliers.

**Data sampling:** Analyzing a subset of the data can be a valuable way to identify data quality issues quickly. By selecting a random sample of the data and analyzing it, data analysts can identify any inconsistencies or problems that may be present in the larger dataset.

**Data profiling tools:** There are various data profiling tools to automate identifying data quality issues. These tools can analyze data across multiple dimensions, such as completeness, consistency, accuracy, and highlight areas.

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

Understanding the context and characteristics of the analyzed data is essential. This may require additional information about the data sources, collection processes, transformations, and usage.

To identify the root causes of data quality issues, some areas of inquiry that may be helpful include:

1. Data collection: How was the data collected? Were the data sources reliable and consistent? Were there any gaps or errors in the data collection process?
2. Data transformation: Were any manipulations applied to the data before it was stored or analyzed? Were these transformations performed correctly and consistently? Were there any data losses or data type conversions that could have affected data quality?
3. Data validation and verification: Were quality control measures in place to ensure data accuracy and consistency? Were any data validation or verification checks performed on the data? Were there any data outliers or anomalies that could have affected data quality?
4. Data governance: Was a data governance framework in place to manage and monitor data quality? Were there any data governance policies or procedures that could have affected data quality?

By understanding these factors, we are taking steps to address the root causes of data quality issues and implement solutions to improve the accuracy and reliability of the data. Possible solutions include:

* Modifying the data collection process.
* Improving data validation and verification procedures.
* Implementing data quality checks and alerts.
* Improving data governance policies and procedures.

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

Context to what the columns labels represent. This includes knowing what OID represents, whether each transaction can consist of multiple items or if it is one to one in the data. This type of information would help better connect the 4 tables with one another and work on answering the business questions.

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

There are no indicators of what the ids or columns represent, and which are unique. Therefore, making it hard to create an index. To address this type of issue we would need to better understand how the data is collect and meaning of what is being connect. From there the data can be cleaned up and processed to provide a more accurate insight.

**Email Version Response:**

Dear Hiring Manager,

After conducting an analysis and working with of the data records provided, our team has created Entity Relationship Diagram (ERD), answered the given business problems, and identified some key data quality issues. This can be viewed in the attachment of this email.

While working with the data, we have the following questions about the data:

* How was the data collected? What is the source of the data?
* What is the intended purpose of the data?
* Are their certain limitations that should be taken into consideration?
* Is there a reason that certain information is missing?

When it came to identifying potential data quality issues, we did the following:

* **Data profiling**: involves analyzing the data to identify its structure, patterns, and quality characteristics. It could include reviewing data statistics such as min, max, and average values, checking data distributions, and analyzing missing or outliers.
* **Data visualization:** Data visualization tools to create charts, graphs, and dashboards that make it easier to spot data quality issues. For example, to identify any correlation or trends between data points to identify data outliers.
* **Data sampling:** Analyzing a subset of the data can be a valuable way to identify data quality issues quickly. By selecting a random sample of the data and analyzing it, data analysts can identify any inconsistencies or problems that may be present in the larger dataset.
* **Data profiling tools:** There are various data profiling tools to automate identifying data quality issues. These tools can analyze data across multiple dimensions, such as completeness, consistency, accuracy, and highlight areas.

In order to solve the data quality issues, understanding the context and characteristics of the analyzed data is essential. This may require additional information about the data sources, collection processes, transformations, and usage.

To identify the root causes of data quality issues, some areas of inquiry that may be helpful include:

1. Data collection: How was the data collected? Were the data sources reliable and consistent? Were there any gaps or errors in the data collection process?
2. Data transformation: Were any manipulations applied to the data before it was stored or analyzed? Were these transformations performed correctly and consistently? Were there any data losses or data type conversions that could have affected data quality?
3. Data validation and verification: Were quality control measures in place to ensure data accuracy and consistency? Were any data validation or verification checks performed on the data? Were there any data outliers or anomalies that could have affected data quality?
4. Data governance: Was a data governance framework in place to manage and monitor data quality? Were there any data governance policies or procedures that could have affected data quality?

By understanding these factors, we are taking steps to address the root causes of data quality issues and implement solutions to improve the accuracy and reliability of the data. Possible solutions include:

* Modifying the data collection process.
* Improving data validation and verification procedures.
* Implementing data quality checks and alerts.
* Improving data governance policies and procedures.

To optimize the data assets that we are trying to create, it would be use to have more context to what the columns labels represent. This includes knowing what OID represents, whether each transaction can consist of multiple items or if it is one to one in the data. This type of information would help better connect the 4 tables with one another and work on answering the business questions.

There are currently no clear indicators of what the ids or columns represent, and which are unique. Therefore, making it hard to create an index. To address this type of issue we would need to better understand how the data is collect and meaning of what is being connect. From there the data can be cleaned up and processed to provide a more accurate insight.

We would love to find some time to further discuss the data records and findings. Please let us know if you have any questions. We are looking forward to speaking with you soon!

Best,

Team 8