Google

Reading: Seven elements of quality testing

In this part of the course, you have been learning about the importance of quality testing in your ETL system. This is the process of checking data for defects in order to prevent system failures. Ideally, your pipeline should have checkpoints built-in that identify any defects before they arrive in the target database system. These checkpoints ensure that the data coming in is already clean and useful! In this reading, you will be given a checklist for what your ETL quality testing should be taking into account.

Checking for data quality involves ensuring the data is trustworthy before reaching its destination. When considering what checks you need to ensure the quality of your data as it moves through the pipeline, there are seven elements you should consider:

- **Completeness:** Does the data contain all of the desired components or measures?
- **Consistency:** Is the data compatible and in agreement across all systems?
- **Conformity:** Does the data fit the required destination format?
- **Accuracy:** Does the data conform to the actual entity being measured or described?
- **Redundancy:** Is only the necessary data being moved, transformed, and stored for use?
- **Timeliness:** Is the data current?
- **Integrity:** Is the data accurate, complete, consistent, and trustworthy? (Integrity is influenced by the previously mentioned qualities.) Quantitative validations, including checking for duplicates, the number of records, and the amounts listed, help ensure data's integrity.

Common issues

There are also some common issues you can protect against within your system to ensure the incoming data doesn't cause errors or other large-scale problems in your database system:

- **Check data mapping:** Does the data from the source match the data in the target database?
- **Check for inconsistencies:** Are there inconsistencies between the source system and the target system?
- **Check for inaccurate data:** Is the data correct and does it reflect the actual entity being measured?
- **Check for duplicate data:** Does this data already exist within the target system?

To address these issues and ensure your data meets all seven elements of quality testing, you can build intermediate steps into your pipeline that check the loaded data against known parameters. For example, to ensure the timeliness of the data, you can add a checkpoint that determines if that data matches the current date; if the incoming data fails this check, there's an issue upstream that needs to be flagged. Considering these checks in your design process will ensure your pipeline delivers quality data and needs less maintenance over time.

Key takeaways

One of the great things about BI is that it gives us the tools to automate certain processes that help save time and resources during data analysis– building quality checks into your ETL pipeline system is one of the ways you can do this! Making sure you are already considering the completeness, consistency, conformity, accuracy, redundancy, integrity, and timeliness of the data as it moves from one system to another means you and your team don't have to check the data manually later on.