Enhancing Data Quality Management: Introducing Capture and Cleanse Modes to the Data Quality Dashboard

Frank J DeFalco[1], Clair Blacketer[1]

[1] Janssen Research & Development

Background

Data quality is a complex challenge in various domains due to the intricate nature of information systems, diverse data sources, and inherent variability of data. While the existing Data Quality Dashboard[1] is a valuable tool for assessing data quality of standardized observational health data, it lacks the capability to capture and remove data records that fail data quality checks from a data source. The lack of this capability leaves records in the database that have failed data quality checks with no systematic way to collect them for further characterization or remove them to improve the research quality of the data.

Methods

New run modes named ‘capture’ and ‘cleanse’ were added to the Data Quality Dashboard package through a new parameter named ‘runMode’. These run modes provide alternatives to the existing mode called ‘execute’. Both capture and cleanse modes are optional and can be run in addition to the execution mode or not at all. The optional approach is critical so organizations can make determinations on if they want to remove records after data quality analysis. A subset of data quality checks require capture and cleanse options based on whether the data quality checks were specific to the contents of the data source or the structure of the data schema itself. All data quality checks are identified as eligible for capture and cleanse through the existing check description functionality.

Capture Mode: This mode provides the ability to identify data records that fail specific data quality checks and capture copies of the affected records to a user-specified schema. With capture mode, organizations can preserve and characterize the failing records separately, gaining valuable insights for further analysis and investigation of their data quality issues.

Cleanse Mode: This mode provides the ability to automatically remove failing records from a data source. When activated, this mode identifies data records that do not meet the defined data quality criteria and cleanses them from the common data model schema. By leveraging the cleanse mode, organizations can maintain a cleaner and more reliable dataset by eliminating records that fail data quality checks, ensuring data integrity and accuracy.

Results

The DataQualityDashboard package was updated with new features to support capture and cleanse. The implementation of the capture and cleanse feature has demonstrated promising results in improving data quality management.

Key results include:

- Enhanced Visibility: By capturing all failing data quality checks in capture mode, data owners can review the data quality failures in a dedicated schema. Deeper characterization of failed records that have been captured allows for further investigation of potential ETL or data issues.

- Proactive Data Cleansing: The cleanse mode allows data owners to automate the removal of failing records. Eliminating records that failed data quality checks allows organizations to improve the reliability of downstream analytics. A systematic approach to data cleansing provides a reproducible way to eliminate failing records as part of a data operations pipeline.

Future research will evaluate the impact of the data cleanse feature on subsequent studies that make use of a data source before and after data quality cleansing.

Conclusion

The capture and cleanse features represent a valuable addition to the Data Quality Dashboard, providing a proactive approach to data quality management. Capturing data records that fail data quality checks and enabling automated data cleansing allows organizations to improve data quality and prevent data quality issues from impacting the evidence generation processes that leverages their standardized data sources.

References

1. Data Quality Dashboard (https://github.com/ohdsi/dataqualitydashboard)