МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2024
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2024
Proceedings of the Fifty-Third Spring Conference
of the Union of Bulgarian Mathematicians
Borovets, April 1–5, 2024

CUSTOM DATA QUALITY MECHANISM IN DATA WAREHOUSE FACILITATED BY DATA INTEGRITY CHECKS*

Angel Georgiev[1], **Vladimir Valkanov**[2]

Faculty of Mathematics and Informatics,
Paisii Hilendarski University of Plovdiv, Plovdiv, Bulgaria
e-mails: [1]angel.georgiev@uni-plovdiv.bg, [2]vvalkanov@uni-plovdiv.net

In the era of data-driven decision-making, Data Warehousing (DWH) is crucial for organizations seeking to leverage extensive datasets. However, the success of DWH initiatives depends on the quality of the enclosed data. Insufficient quality data in Data Warehousing can impact the accuracy of analytical results, leading to misguided decisions and reduced business performance. This paper examines the significance of Data Quality Mechanisms in addressing challenges related to data quality. Data Quality Mechanisms play a crucial role in identifying, rectifying, and preventing data quality issues throughout the data lifecycle. This paper explores fundamental concepts, challenges, and impacts of data quality on business operations. It emphasizes the critical role of robust Data Quality Mechanisms in ensuring the accuracy, completeness, and reliability of data within the Data Warehousing ecosystem. As organizations increasingly recognize data as a strategic asset, it is imperative to implement effective data quality mechanisms to unlock the true potential of data warehouses and derive actionable insights.
Keywords: Data Warehouse, data quality, ground transportation, SQL, data, checks, data governance, data maintenance.

ПЕРСОНАЛИЗИРАН МЕХАНИЗЪМ ЗА КАЧЕСТВО НА ДАННИТЕ В ХРАНИЛИЩА НА ГОЛЕМИ ДАННИ, УЛЕСНЕН ОТ ПРОВЕРКИ ЗА ЦЕЛОСТТА НА ДАННИТЕ

Ангел Георгиев[1], **Владимир Вълканов**[2]

Факултет по математика и информатика,
Пловдивски университет „Паисий Хилендарски“, Пловдив, България
e-mails: [1]angel.georgiev@uni-plovdiv.bg, [2]vvalkanov@uni-plovdiv.net

*https://doi.org/10.55630/mem.2024.53.067-075
2020 Mathematics Subject Classification: 68T09, 62R07, 68P01, 68P15, 68P20.

В ерата на вземане на решения, базирани на данни, хранилищата за големи данни са от решаващо значение за организациите, които искат да използват обширни набори от данни. Успехът на технологията обаче зависи от качеството на приложените данни. Недостатъчно качествените данни могат да повлияят на точността на аналитичните резултати, което води до погрешни решения и намалена ефективност на бизнеса. Този документ разглежда значението на механизмите за качество на данните при справяне с предизвикателствата, свързани с качеството на данните. Механизмите за качество на данните играят решаваща роля в идентифицирането, коригирането и предотвратяването на проблеми с качеството на данните през целия жизнен цикъл на данните. Този документ изследва фундаментални концепции, предизвикателства и въздействия на качеството на данните върху бизнес операциите. Той подчертава критичната роля на стабилните механизми за качество на данните за осигуряване на точност, пълнота и надеждност на данните в екосистемата за съхранение на данни. Тъй като организациите все повече разпознават данните като стратегически актив, наложително е да се внедрят ефективни механизми за качество на данните, за да се отключи истинският потенциал на хранилищата за данни и да се извлекат полезни прозрения.
Ключови думи: Хранилища за големи данни, качество на данните, наземен транспорт, SQL, данни, проверки, следене на качеството да данните, поддръжка на развиетието на данните.

1. Introduction. In today's dynamic business intelligence environment, organizations are increasingly reliant on Data Warehousing (DWH) solutions to centralise, manage, and analyse large amounts of data for informed decision-making. However, the success of these efforts depends on the quality of the data they handle and store. Poor data quality can lead to inaccurate insights, misguided decision-making, and compromised business performance.

A reliable Data Quality Mechanism incorporates a range of procedures, tools, and best practices aimed at detecting, correcting, and preventing data quality issues throughout the data lifecycle. Data Quality Mechanisms have a vital role to the end client's decision-making data, which the Data Warehouse contains. The analytical results can be improved ultimately by implementing such techniques.

Data quality is one of the characteristics that cannot be omitted when designing a stable data platform. High-quality data can be distinguished by accuracy, completeness, consistency, timeliness, and relevance. The absence of these attributes may introduce errors, anomalies, and discrepancies, which can erode the reliability of analytics and decision support systems. [1]

This research examines the key points of data quality mechanisms in data warehousing. It explores the challenges associated with this field of information technology, the negative impact of poor data on business operations, and the techniques used to achieve and maintain high standards of data health in such ecosystems. As information becomes an increasingly important asset for organizations, its role is vital to the business development and growth of every data-oriented organization.

2. Current state. In today's world of data management and analytics, the role of Data Warehousing (DWH) has evolved significantly due to technological advancements and the growing importance of data-driven decision-making. As big tech companies continue to accumulate and leverage vast amounts of data, the importance of maintaining

high data quality within the DWH ecosystem has become increasingly critical. The state of Data Quality Mechanisms in Data Warehousing shows an increased focus on ensuring accurate, reliable, and complete data.

There are a lot of existing tools for controlling the state of information in a specific organization such as Talend Data Fabric, SAS Data Quality, SAP Data Intelligence, etc.

The first one is an integration platform offered by Talend, a software company specializing in data integration and management solutions. Talend Data Fabric is designed to address various aspects of the data management lifecycle, providing a comprehensive suite of tools for data integration, data quality, data governance, and more.

SAS Data Quality is a component of the SAS (Statistical Analysis System) software suite provided by the SAS Institute. SAS is a renowned software suite used for advanced analytics, business intelligence, data management, and statistical analysis. SAS Data Quality specifically focuses on ensuring the accuracy, completeness, and reliability of data within an organization's data ecosystem.

SAP Data Intelligence is a comprehensive data management solution provided by SAP, a leading enterprise software company. SAP Data Intelligence is designed to enable organizations to discover, connect, manage, and orchestrate disjointed and distributed data assets across the enterprise. It plays a crucial role in supporting data integration, processing, and governance in complex and dynamic data landscapes [10, 11].

All these software as a service provide ideal solutions for data quality issues, but nothing is a better fit for a company's data strategy than a custom one.

In summary, the state of Data Quality Mechanisms in Data Warehousing is constantly changing and evolving. Organizations are using advanced technologies and comprehensive governance approaches to tackle data quality issues and fully utilize their data for strategic decision-making [2, 3].

3. ETL Process Overview. The Extract, Transform, Load (ETL) process is a crucial component in data management, ensuring efficient and reliable data integration within organizations. ETL involves three stages: extracting data from source systems, transforming it, and loading it into target databases or data warehouses.

A. Extract

During the initial phase, data is extracted from various source systems, such as databases, applications, flat files, or external data sources. This process involves collecting raw data in its native format from different origins, capturing the information required for analysis and decision-making.

B. Transform

The data that has been extracted is then transformed. This involves cleaning, standardizing, and manipulating the data to meet the specific requirements of the target system or data warehouse. Transformations may include data cleansing to address errors or inconsistencies, data enrichment to add additional context, and data formatting to ensure uniformity.

C. Load

The last step is to load the transformed data into the destination system, which is usually a data warehouse or a database optimized for analytical queries. The data is organized and structured in a way that facilitates efficient querying and reporting. Loading can occur in batches or in real-time, depending on the organization's needs and the nature of the data.

The ETL process is closely linked to data quality, serving as a crucial gateway to ensure the accuracy, consistency, and reliability of data within an organization's information ecosystem. The connection to data quality comes in several aspects as: data cleansing in transformation, standardization for consistency, data enrichment for completeness, data profiling and validation, quality checks in loading, integration with data quality tools, etc. [4, 5, 6].

4. Custom Data Quality Mechanism. Data Quality Rules are a type of business rule that defines specific criteria for business requirements. They can be used to analyze and evaluate the quality of datasets, including how data behaves over time and its consistency across internal source systems. Data Quality Rules enable the measurement of various dimensions of data quality, such as:

- Accuracy – does the data match what is expected?
- Consistency – does the data match between alternative source systems?
- Completeness – has all values in the data been completed?

Data Quality Rules enable the identification of problem areas and trends. They facilitate continuous measurement, reveal opportunities for improvement, and provide the foundation for data cleansing. It is crucial to document Data Quality Rules clearly and make them easily accessible.

There are several milestone, which make the data quality rule mechanism: sign off process, new data quality rule requests and amendments, validation type, validation rule naming convention, validation rules for each semantic layer, key source tables, process for managing data quality alerts , validation rule severity server layer agreements definitions monitoring validation rules tools [7, 8].

The **sign-off process** for stakeholders must be completed before submitting a request for a new or updated Data Quality Rule.

A Validation Rule template cannot be submitted until the new rule or changes have been reviewed and approved by the domain owner. The requestor should document all details of the new rule or changes and send them to the domain owner for peer review and sign-off. If a domain owner makes a direct request, the details will still be subject to peer review. This review may be conducted by another domain owner or a business analyst.

New requests for Data Quality Rules: The sign-off email should be forwarded with the attached VR01 to the Data Steward. The innovation in this step of the validation process is the template structure, which provides flexible design and mobility in the rule creation. It consists of Excel Spreadsheet with the following attributes: **Date**, **Version**, **Validation type** (Data Integrity/ Quality Rule, to be completed by Data Steward), **Severity Status** (Top priority/High/Medium/Low this is the status when the validation check has failed, the threshold will define the severity status), **Golden Metric** (metric the validation rule is checking), **Semantic Layer** (Finance, CEXP, etc.), **Impacted Schema and tables, Validation Name** (QR/DI_Semantic layer_Metric) e.g. QR_Finance_GMV/DI_Finance_GMV, to be completed by Data Steward), **Event** (detail of what triggers this quality rule into action), **Description, Threshold (%, number), Frequency (daily, weekly, monthly), SQL Method, Statement, Peer reviewed by, Date**. For audit purposes, a copy of the VR01 and approval email will be saved in the Validation Rule requests folder as shown in Figure 1.

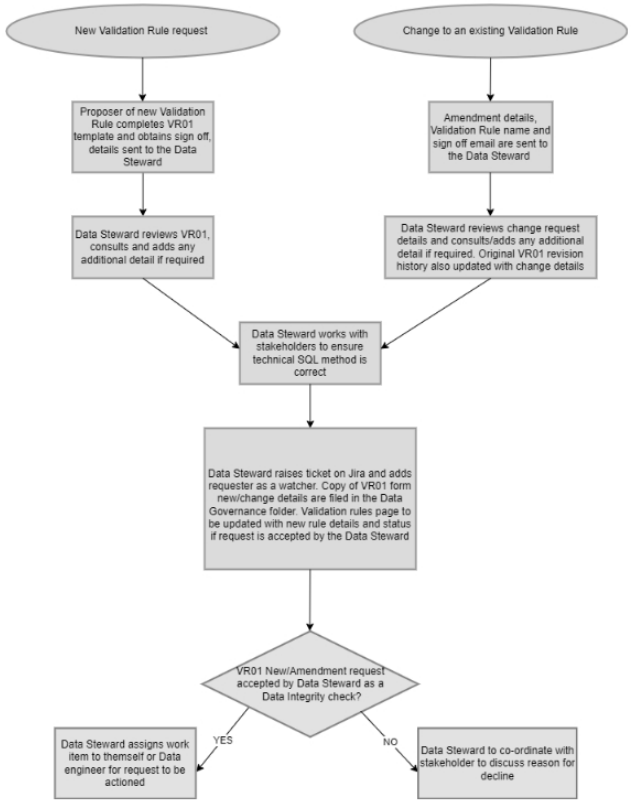

Fig. 1. Adding new validation rules

Amendments to existing Data Quality Rules: Please send details of changes, Validation Rule name, and sign-off email to the Data Steward. For audit purposes, the revision history of the original VR01 will be updated, along with the approval email.

The recipient of the VR01, known as the Data Steward, will review the details and collaborate with the proposer and other colleagues as necessary to ensure that all required information has been provided. It is important to note that while the template is designed to provide sufficient information for a request to be raised, in some cases the Data Steward may need to gather additional or supplementary information. The validation type for the Data Quality Rule will be classified as either Data Integrity or Quality Rule.

There are two validation types:

Data Integrity (DI) – refers to a technical validation, these rules are checking to see if the data can be loaded into the database, additionally if the data that is expected is available and received at the correct time.

Quality Rules (QR) – refers to a business validation, these rules are checking to see if the data results across different periods are consistent with expected results, and if the data is consistent across source systems.

The Data Steward will add the validation rule (DI/QR) to GitHub and the SQL logic will be peer reviewed by a member of the Data Team before it is pushed to production. A copy of the VR01 will be filed in the appropriate sub folder within the Data Governance (Validation Rule request) folder.

Requests for changes to an existing Validation Rule can also be submitted to the Data Steward, please ensure to reference the QR/DI validation name (see naming convention). The Data Steward will also capture the amendments to each rule within the revision history of the original VR01 [8, 9].

The Validation Rules in each semantic layer follow a common naming convention, which is captured in the Validation Name field. Quality Rules are formatted as 'QR_Semantic layer_Metric/description', for example, 'QR_Finance_Rides'. Rules that are considered a Data Integrity check are formatted as 'DI_Semantic layer_Metric'.

This format allows the rules to be displayed for each semantic layer within a single validation rules confluence page. Users benefit from having all validation rules for a specific semantic layer available on a single page.

Validation rules not related to a semantic layer and managed by the Data Team follow the format 'DI_SourceTable_description', such as DI_factoffersrex_nocancellations.

As described in Figure 2 alerts overview process consists of four steps:

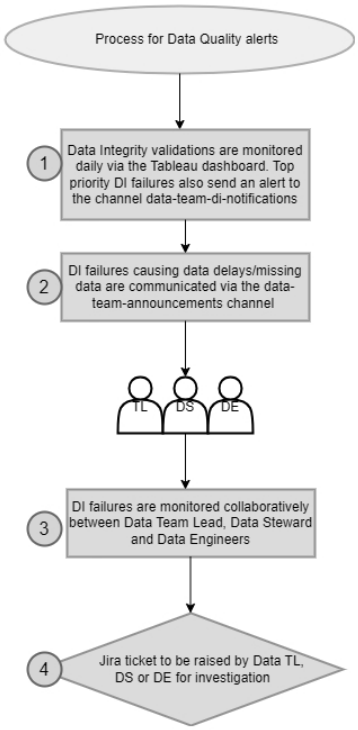

Fig. 2. Process for managing Data quality alerts

1. The Data Steward monitors alerts for Data Quality and Data Integrity through a Tableau dashboard. Top priority alerts for Data Integrity failures are also sent to dedicated Slack channel.

2. The Data Team Lead and Data Engineers will mostly monitor and manage top priority DI alerts that are not within the data stewardship ETL. The Data Steward will mostly manage top priority failures related to data stewardship. Any top priority failures that cause data delays or missing data will be communicated through the dedicated announcements channels.

3. The Data Steward will primarily monitor DI failures that are not top priority. However, during sick absences or vacations, the Data Team will monitor DI failures collaboratively.

4. A Jira ticket will be raised for any Data Integrity failures. In the case of Top priority DI failures outside of data stewardship ETL, the Jira ticket will be raised by the Data Team Lead or Data Engineer. For all other levels of severity and top priority data stewardship DI rules, the Data Steward will raise a Jira ticket.

The Service Level Agreement (SLA) specifies the timeframe for the Data Steward and Domain Owner to investigate a failed QR alert. Please note that the SLA does not cover the resolution or fix of the issue. The resolution or fix are not included within the definitions below:

- Top priority – 1 day when alert triggered.
- High – 2–3 days from when alert triggered.
- Medium – 5 days from when alert triggered.
- Low – 7 days from when alert triggered.

The unique contribution to the mechanism are the two database objects, containing the execution details of each rule. The table DI_ALERTS_TBL (***ID*** INTEGER, ***VALIDATION_NAME*** VARCHAR, ***VALIDATION_DESCRIPTION*** VARCHAR, ***QUERY_FOR_ACTUAL*** VARCHAR, ***QUERY_FOR_EXPECTED*** VARCHAR, ***ACTUAL*** DECIMAL, ***EXPECTED*** DECIMAL, ***THRESHOLD*** DECIMAL, ***IS_SUCCESS*** INT, ***AUTHOR*** VARCHAR, ***EXECUTION_DATE*** DATE), which stores the full information, plus historical data and the view DI_ALERTS_LAST_EXECUTION_V (***ID*** INTEGER, ***VALIDATION_NAME*** VARCHAR, ***VALIDATION_DESCRIPTION*** VARCHAR, ***QUERY_FOR_ACTUAL*** VARCHAR, ***QUERY_FOR_EXPECTED*** VARCHAR, ***ACTUAL*** DECIMAL, ***EXPECTED*** DECIMAL, ***THRESHOLD*** DECIMAL, ***IS_SUCCESS*** INT, ***AUTHOR*** VARCHAR), which holds data about the latest executed validations are the representation of the custom validation rule template in the DI mechanism itself. This provides the users with high flexibility in creating their own data integration checks and testing the business rules in an easy way.

Finally, we need to have in mind that all the DI validation failure or successes will be visualized in a specific Tableau dashboard, based on the table and view data, which we have previously transformed and processed.

5. Conclusion and future developments. The relationship between the Extract, Transform, Load (ETL) process and data quality is crucial in modern data management. ETL facilitates the movement of data and enhances data quality through meticulous extraction, thoughtful transformation, and precise loading. The processes are interconnected to enable organizations to obtain useful insights from reliable, consistent, and

accurate data, which promotes informed decision-making and strategic initiatives.

The ETL process acts as a guardian of data quality by incorporating cleansing, standardization, enrichment, and validation mechanisms to address discrepancies and fortify datasets against inaccuracies. The combination of ETL and data quality ensures the accuracy of information in data warehouses and supports strong analytics, business intelligence, and reporting. [9]

The current research approach provides custom mechanism from the perspective of enriching some existing techniques with uniquely designed attributes. The validation rule templates supply the users with the ability to set and create the rules in an flexible innovative way, which makes the requirements clear and depicts the desired outcome in an understandable way. The two tables, created specifically for the validation process hold the data, visualized later in the Tableau dashboard, presenting the needed information to the stakeholders.

The evolution of ETL processes and their integration with data quality initiatives are set to advance continually. Future ETL solutions may become more automated in identifying and addressing data quality issues in real-time, using artificial intelligence and specifically the newly occurring and quickly evolving LLMs and Generative Pre-trained Transformers. ETL processes are expected to become more dynamic and adaptive, adjusting transformations based on real-time insights into data quality. This will enable organizations to proactively manage and enhance data quality as conditions change.

Organizations are expected to see increased integration between ETL tools and specialized data quality platforms. This integration enables organizations to easily use advanced data quality features in their ETL workflows. Metadata management in ETL processes will become more sophisticated, providing comprehensive insights into the quality and lineage of data. Advanced metadata capabilities can help to make data quality management more transparent and traceable. ETL processes are likely to play a more important role in supporting data governance and compliance initiatives. Integrating with governance frameworks will ensure that data quality measures align with regulatory requirements and organizational policies. As organizations deal with increasing amounts of data, ETL solutions will need to adapt to handle diverse data sources while maintaining high data quality standards in dynamic environments.

In conclusion, ETL processes and data quality are continuously refined and innovated. As technology advances and the data landscape evolves, organizations can anticipate a future where ETL becomes an even more intelligent, adaptive, and integral part of their data management strategies, contributing to the ongoing pursuit of high-quality, actionable data.

REFERENCES

[1] W. H. Inmon. Building the Data Warehouse, 4th edn. Wiley India Pvt. Limited, 2005. ISBN 9788126506453.

[2] R. Kimball, M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 3rd edn. Indianapolis, John Wiley & Sons, Inc., 2013, ISBN 9780471153375.

[3] B. Moses, L. Gavish, M. Vorwerck. Data Quality Fundamentals. O'Reilly Media, Inc., 2022, ISBN 9781098112011.

[4] J. Reis, M. Housley. Fundamentals of Data Engineering: Plan and Build Robust Data Systems. O'Reilly Media, Inc., 2022. ISBN 9781098108304.

[5] Q. Liu, G. Feng, G. K. Tayi, J. Tian.. Managing Data Quality of the Data Warehouse: A Chance-Constrained Programming Approach. *Inf. Syst. Front.* **23** (2021), 375–389, https://doi.org/10.1007/s10796-019-09963-5.

[6] M. R. Sureddy, P. Yallamula. Data Quality Architecture for Data Warehouses. *International Journal of Research Culture Society*, **4**, no. 6 (2020), 95–100, DOIs:10.2017/IJRCS.2456.6683/202006017.

[7] A. Fattah, T. Ridwan, N. Sulistiyowati. Dimensional Data Design for Event Feedback Data Warehouse. *JISA (Jurnal Informatika dan Sains)*, **6**, no. 1 (2023) 69–73. https://doi.org/10.31326/jisa.v6i1.1648.

[8] H. Benkhaled, D. Berrabah, F. Boufarès. Data Warehouses and Big Data: How to Cope With Data Quality. *International Journal of Organizational and Collective Intelligence (IJOCI)*, 10, no. 3 (2020), 13 pp, https://doi.org/10.4018/IJOCI.2020070101.

[9] N. Gupta. Optimising data quality of a data warehouse using data purgation process. International Journal of Data Mining, Modelling and Management. 15, no. 1 (2023), 102–131, https://doi.org/10.1504/IJDMMM.2023.10055198.

[10] M. Altendeitering, T. Guggenberger. Data Quality Tools: Towards a Software Reference Architecture. Hawaii International Conference on System Sciences (HICSS), 2024.

[11] https://www.techtarget.com/searchdatamanagement/tip/Top-data-quality-management-tools (Last visited 23.12.2023)