CSIT Conference 2023, Yerevan, Armenia, September 25 - 30

Alert System for Data Quality in Data Lakes

Eliza Gyulgyulyan
IIAP NAS RA
Yerevan, Armenia
e-mail: eliza_gyulgyulyan@iiap.sci.am

Hrachya Astsatryan
IIAP NAS RA
Yerevan, Armenia
e-mail: hrach@sci.am

***Abstract*— Data lakes are popular repositories for storing large volumes of heterogeneous and unstructured data. However, ensuring data quality in data lakes poses challenges, potentially leading to inaccurate analyses and decisions. This article proposes an approach for a quality alert system based on a quality model and alert process. The system identifies and notifies users of data quality issues during analysis, reducing costs and facilitating data integration. It includes a dashboard for visualizing data quality metrics and measuring potential biases.**

***Keywords*—Data lakes, data quality, quality metrics, alert system.**

I. Introduction

Data lakes are popular for storing and analyzing large volumes of heterogeneous data. However, data quality issues can hinder their effectiveness, leading to suboptimal decision-making and increased costs. Existing research lacks a real-time alert system for identifying and notifying data quality issues during analysis. The proposed quality alert system fills this gap by providing a proactive approach to data quality management in data lakes.

The system measures critical dimensions based on predefined quality metrics to identify issues during analysis. It integrates with the data lake architecture, automating the detection and notification of quality issues. The system also measures potential biases resulting from poor-quality data, enabling users to assess their impact on decisions and outcomes.

The key contributions of this study include seamless integration with the data lake architecture, streamlining data processing, comprehensive quality metrics, measurement of potential biases, and a real-world case study to evaluate the system's effectiveness. These contributions reduce user burden, improve data processing efficiency, and enhance data-driven decision-making.

II. Related Work

Existing literature has addressed various aspects of data lake architecture and data quality management [1]–[4], but a comprehensive approach that encompasses quality alert system integration, reorganized data processing, comprehensive quality metrics, measurement of biased decisions, and real-world case studies is lacking. Previous research has explored data quality monitoring and alert system in different contexts, such as business intelligence or big data analytics [5], [6]. While some studies have focused on real-time detection and notification of quality issues [1], they often require manual assessment from users, leading to time-consuming and error-prone processes. In contrast, the proposed quality alert system automates data quality assessment within the data lake architecture, relieving users of the burden and seamlessly integrating with the analysis workflow. This automated approach enhances efficiency by proactively identifying and notifying users of quality issues. Additionally, while previous studies have explored reorganizing data processing in data lakes [7]–[9], there is a research gap in integrating an alert system to streamline data processing by identifying and notifying users of data quality issues. This research aims to fill this gap and develop a comprehensive set of quality metrics tailored to the needs of data lake environments. By addressing these gaps, the proposed approach provides a holistic solution for data quality management in data lakes.

III. Methods and Approach

A systematic approach involves several key components to develop the quality alert system for data lakes. This section describes the methodology and techniques used to implement the proposed system.

Quality Model Development: The first step in this approach was to develop a comprehensive quality model that captures the essential dimensions of data quality in data lakes [5]. As a continuation, quality dimensions such as completeness, accuracy, consistency, timeliness, and relevance are considered. Each dimension was further broken down. For example, the time-related dimension is measured through timeliness, currency and volatility. The quality model serves as the foundation for the quality alert system and guides the measurement and assessment of data quality during analysis.

Integration with Data Lake Architecture: An important aspect of the system is its seamless integration with the existing data lake architecture. The alert system should operate within the data lake infrastructure, leveraging its capabilities and minimizing the need for users to engage with quality metrics at a professional level. Integrating the alert system into the data lake ensures that data quality issues are identified and addressed during analysis without disrupting the overall data processing workflow.

Real-Time Alert Process: An alert process continuously monitors the data being analyzed to enable real-time detection

and notification of data quality issues. The alert process uses the quality model and associated metrics to assess the data quality during analysis. When a quality issue surpasses a predefined threshold, the system triggers an alert, notifying the user about the specific issue encountered. The real-time nature of the alert process allows users to address quality issues promptly, reducing the potential impact on decision-making.

Bias Measurement and Evaluation: In addition to detecting data quality issues, the system measures potential biases that may arise from analyzing poor-quality data, quantifies the level of bias in the analysis results. By providing users with insights into the biases present in the data, they can make informed decisions, taking into account the impact of data quality on the outcomes. This feature enhances the reliability and integrity of decision-making processes and helps mitigate the risks associated with biased analysis.

Quality Model Development: The dimensions are derived from the literature, primarily [10]–[12], and include consistency, uniqueness, accuracy, completeness, interpretability, and time-related dimensions. Additionally, six dimensions specific to the analysis context are specified: User Engagement and Enjoyment, Task Success, Information Novelty [13]–[15], user characterization, and System characterization [14]–[16]. However, these dimensions require adaptation to the analysis context, and their interrelationships need consideration for generating appropriate alert notifications. Each dimension is evaluated using a set of metrics. The complete set of metrics for each dimension will be defined in future work.

IV. Quality Model

The base for the System is the Quality Model designed from the classes described in Figure 1. The model represents the assessment of quality in the context of Big Data. It is based on quality *dimensions* and *metrics*, which are established through quality *questions* according to the *analysis* performed by the *user*. The assessment of a metric on a data *source* and/or *attribute* is performed through a predefined *measurement method.* Below, the component classes of the Quality Model are described more deeply.

Quality Dimension: a characteristic of quality and a measurable notion that is measured with quality metrics. Nonetheless, these dimensions are very specific to the context of data exploration and need to be adapted to analysis.

Quality Metric: The quality metric class refines the quality question, and is a quantitative way of data computation addressing the quality dimension. In other words, each dimension is considered by a set of metrics, e.g. NullValues is a metric of completeness. This means that the system should check the level of null values of data to alert about completeness. The complete set of metrics for each dimension needs to be defined in later works.

Quality Question: an option for a user to choose the dimensions and metrics that the System should alert about, according to an analysis. In the context of quality, it is a closed question as there is a limited list of quality dimensions and metrics predefined. If the user does not choose a question, the System considers all the dimensions and metrics by default. This makes the System be compatible with the requirements of the user and its analysis queries.

Analysis: This class specifies the type of analysis being performed. The quality questions (and their related dimensions and metrics) depend on the type of analysis. For instance, a NullValues metric will impact the interpretation of the results in an OLAP context, whereas it may be not the case in a context of data-mining (using particular classification techniques, for example).

Measurement Method: This class defines a quality formula, which measures the problem of quality for a particular metric. The measurement method represents the physical implementation of the metric computation, e.g. to compute the level of completeness, the System needs to compute the number of NullValues with CheckNull function. After the formula [(1-Number of not null values)/total number of values] is computed to alert about completeness level.

Data Source and Attribute: These two classes describe the Big Data substance on which the metrics are defined. The System needs to consider the different types of data structures and attribute formats to undertake the metric measurement method and parse the result.

User: This class indicates who performs the analysis and, if able, chooses quality questions. According to the "role" property, two types of users are described: 1) who is not sophisticated in the domain of quality, e.g. decision-maker or project analyst. This user does not choose quality questions but the System chooses instead by default. 2) who is sophisticated enough to choose quality questions and thus query quality reasonably.

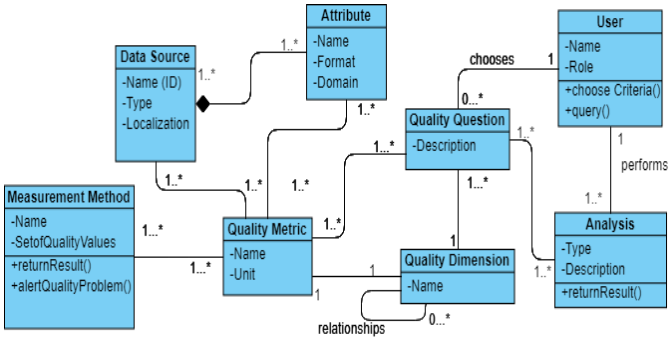

Figure 1. Quality model: The alert system engine

The instantiation of the quality model can be seen in Figure 2, when a business analyst - named Arsen, is sophisticated enough in the quality domain and establishes 3 quality questions for his analysis. For each question, there are three different dimensions, metrics and measurement methods. The object diagram of Figure 2 describes the use case of this paper about the analysis of export in Armenia.

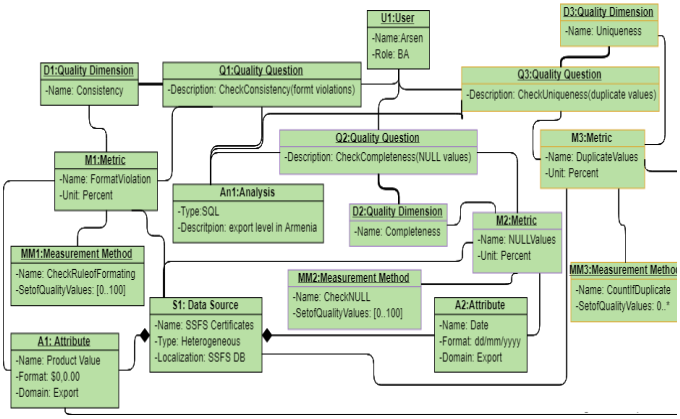

Figure 2. Object Diagram for Use Case

V. The Alert System

This subsection describes the Alert System roadmap (Figure 3). Depending on user types, there are two different roadmap processes. It is considered that the user a) is not sophisticated enough in the quality domain, and b) is sophisticated enough in the quality domain. Thus, when performing the analysis the user either a) directly queries the data, or b) chooses quality questions before querying.

1a. The user is not an expert in the quality domain, thus queries the data as a regular process of the analysis.

2a. Since no quality questions are chosen, the System considers all the possible dimensions and metrics.

1b. The user is sophisticated enough in the quality domain, thus choosing desirable Quality Questions from a predefined list with the [?user interface before querying data. On the interface, the user can also input values of desirable metric measures, e.g., "I need more than 90% of completeness".

2b. According to the quality questions, the system establishes quality dimensions and metrics, which need to be considered by the System to alert about.

3. After dimensions and metrics are established, the metric measurement methods are identified and the measurement process starts. The model operates on the data being analyzed.

4. After finishing the Metric Measurement process, the System alerts if the provided desired value is not satisfied. If there is no desired value, the System considers 100% as the desired value of quality and 0 as the desired value of quality problem.

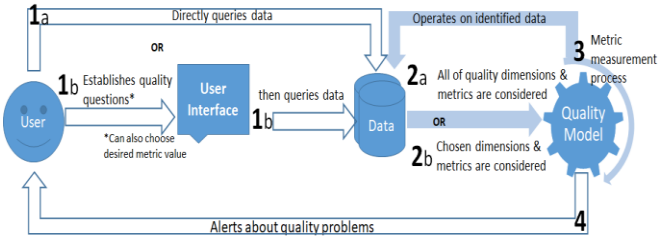

Figure 3. The Alert System roadmap

VI. Conclusion

The study proposes a quality alert system that addresses data quality issues during real-time data analysis. It enhances decision-making, reduces costs and risks associated with poor data quality, and improves the value and impact of data lakes. Future work includes further refinement of the quality metrics, describing the exact way of integration into data lakes. Additionally, advancing the framework to evaluate the biases inherent in data analysis comprehensively holds promise for enhancing the overall efficacy and reliability of the proposed quality alert system.

References

[1] M. Farid, A. Roatis, I. F. Ilyas, H.-F. Hoffmann, and X. Chu, "CLAMS: Bringing Quality to Data Lakes," in *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, San Francisco, California, USA: ACM Press, pp. 2089–2092, 2016.

[2] C. Campbell, "Top Five Differences between Data Lakes and Data Warehouses," Mar. 15, 2021. https://www.bluegranite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses (accessed Mar. 15, 2021).

[3] F. Ravat and Y. Zhao, "Data Lakes: Trends and Perspectives," in *Database and Expert Systems Applications*, S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A. M. Tjoa, and I. Khalil, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 304–313, 2019.

[4] "Data Lake Analytics | Microsoft Azure," Mar. 22, 2021. https://azure.microsoft.com/en-us/services/data-lake-analytics/ (accessed Mar. 23, 2021).

[5] E. Gyulgyulyan, J. Aligon, F. Ravat, and H. Astsatryan, "Data Quality Alerting Model for Big Data Analytics," in *New Trends in Databases and Information Systems*, T. Welzer, J. Eder, V. Podgorelec, R. Wrembel, M. Ivanović, J. Gamper, M. Morzy, T. Tzouramanis, J. Darmont, and A. Kamišalić Latifić, Eds., in Communications in Computer and Information Science. Springer International Publishing, pp. 489–500, 2019

[6] E. Gyulgyulyan, F. Ravat, H. Astsatryan, and J. Aligon, "Data Quality Impact in Business Inteligence" in 2018 Ivannikov Memorial Workshop (IVMEM), IEEE, pp. 47-51, 2018.

[7] I. Megdiche, F. Ravat, and Y. Zhao, "Metadata Management on Data Processing in Data Lakes," in SOFSEM 2021: Theory and Practice of Computer Science, T. Bureš, R. Dondi, J. Gamper, G. Guerrini, T. Jurdziński, C. Pahl, F. Sikora, and P. W. H. Wong, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 553–562, 2021.

[8] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena, "Data lake management: challenges and opportunities," *Proc. VLDB Endow.*, vol. 12, no. 12, Art. no. 12, Aug. 2019.

[9] "Data Lake Governance Best Practices - DZone Big Data," dzone.com, Mar. 14, 2021. https://dzone.com/articles/data-lake-governance-best-practices (accessed Mar. 14, 2021).

[10] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Business Horizons*, vol. 60, no. 3, Art. no. 3, May 2017.

[11] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, *Methodologies for Data Quality Assessment and Improvement*, vol. 41. 2009.

[12] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," Data Science Journal, vol. 14, no. 0, Art. no. 0, May 2015.

[13] M. Lopez, S. Nadal, M. Djedaini, P. Marcel, V. Peralta, and P. Furtado, "An Approach for Alert Raising in Real-Time Data Warehouses," in *Journées francophones sur les Entrepôts de Données et l'Analyse en ligne, E. Zimányi, Ed., in Revue de Nouvelles Technologies de l'Information*. Bruxelles, Belgium: Esteban Zimányi, Apr. 2015.

[14] S. Sarawagi, "User-Adaptive Exploration of Multidimensional Data," 2000.

[15] R. W. White and R. A. Roth, Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool, 2013. Accessed: Feb. 12, 2019.
[16] S. Idreos, O. Papaemmanouil, and S. Chaudhuri, “Overview of Data Exploration Techniques,” in *Proceedings of the 2015 ACM SIGMOD* International Conference on Management of Data, in SIGMOD ’15. New York, NY, USA: ACM, pp. 277–281, 2015.