Data Quality Management for Responsible AI in Data Lakes

Carolina Cortes
ccortes@fing.edu.uy
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay.

Camila Sanz
csanz@fing.edu.uy
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay.

Lorena Etcheverry
lorenae@fing.edu.uy
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay.

Adriana Marotta
amarotta@fing.edu.uy
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay.

ABSTRACT

Data lakes (DL) have become popular resources for organizations managing data science projects. As ethical debates around data-driven decision-making mechanisms grow, the concept of responsible AI has become more visible. Responsible AI frameworks heavily rely on high-quality data, which has brought into focus the assessment of the quality of the data used and new perspectives and dimensions of data quality. Moreover, data quality can only be assessed by considering the context in which it is evaluated. This paper presents our approach to context-aware data quality management, which offers a comprehensive solution to data quality issues in DL. Our approach is designed to adapt to different contexts, ensuring data quality management throughout the entire data lifecycle. We achieve this by defining different components of context and data quality management processing within the DL architecture, representing a novel contribution to existing work.

VLDB Workshop Reference Format:
Carolina Cortes, Camila Sanz, Lorena Etcheverry, and Adriana Marotta. Data Quality Management for Responsible AI in Data Lakes. VLDB 2024 Workshop: Tabular Data Analysis Workshop (TaDA).

1 INTRODUCTION

Data lakes (DL) are becoming increasingly popular as a data storage and management mechanism in organizations engaged in data science projects [9]. Unlike traditional data warehouses, DL store data in its raw format, providing flexibility for various data types and analytical processes. They can handle large volumes of data from diverse sources, making them ideal for data science projects that process massive datasets. DL also facilitates integrating diverse data sources, allowing data scientists to consolidate data from different departments, systems, and sources into a single repository without extensive data transformation upfront. This integration streamlines the data preparation process, enhancing data accessibility. By providing a unified platform for storing and accessing raw data, DL empower data scientists to perform complex analyses, build predictive models, and derive valuable insights from the data [6].

Proceedings of the VLDB Endowment. ISSN 2150-8097.

Data management and governance in DL present several challenges, such as the discovery of relevant datasets to answer a certain question [2, 5], the proper treatment of metadata [23], and definitely, the management of data quality throughout the lifecycle of data within the DL [9]. Data quality (DQ) is a multidimensional and highly context-dependent concept that the data quality and database communities have extensively studied, but few works deal with DQ management in DL [18].

On the other hand, the relationship between the quality of the data used to train and test models and the quality of the models obtained is undeniable. So much so that it is often stated that "an algorithm is only as good as the data it works with" [1]. In particular, poor-quality data can lead to biased or unreliable results, reducing the reliability and usefulness of AI applications. Conversely, high-quality data improves AI models' accuracy, fairness, and robustness, increasing their value and impact. Over time, various DQ dimensions were defined. Recently, some work has explored ethical or social dimensions of DQ, such as diversity and bias in data [7, 20, 27].

Complementary, the concepts of responsible AI [14] and responsible data management [28] have emerged in reaction to ethical discussions about the consequences of data-driven decision-making mechanisms, particularly those using machine learning techniques or AI. Responsible AI frameworks, such as fairness, transparency, accountability, and privacy, require high-quality data as a foundation for ethical AI development and deployment [12].

DQ requirements for data science projects are highly dependent on the problem to be solved. In particular, when dealing with human-related data, they may incorporate requirements on the representation and coverage of specific populations to mitigate biases in the models to be generated. Moreover, the notion of adequate representation may be different depending on the analysis to be performed. For example, in some instances, it may be necessary to ensure enough data from specific subpopulations (e.g., data from people with breasts in the case of a model used to diagnose thoracic diseases from X-ray images [15]).

Imagine the following example from the healthcare domain:

(Example 1.) *Consider a scenario where a health research institution wants to develop a model that can predict diseases based on a person's symptoms. To accomplish this, they gather datasets from various countries containing information about symptoms and diagnoses for different populations, including age, sex, ethnicity, etc. The user needs to ensure that each population subset is adequately represented.*

The reason behind this is that if there is a lack of sufficient representation, it can lead to representation bias, which can significantly impact the accuracy of the generated models [25]. Suppose the researchers need to analyze a disease that affects persons differently according to their biological sex. In this case, it is necessary to balance female and male persons' data. This need may be expressed through a DQ requirement that specifies that the DQ dimension "coverage" must satisfy the mentioned balance.

DL, and particularly data science applications, require flexible and effective mechanisms to manage DQ that capture diverse views and uses of data. These mechanisms should document DQ notions and enable traceability of decisions, promoting transparency. It's essential to consider multiple perspectives on DQ, requiring DQ management mechanisms that are flexible enough to capture and reconcile these distinct views.

This paper proposes a method for including context-aware DQ management in a DL by capturing DQ requirements and data characteristics as part of the context. DQ requirements consideration allows for preventing bias in the data that may lead to inaccurate analysis, while the DL zone as part of the data context allows dealing with heterogeneities, such as differences in underlying data types (text, images, videos, geospatial data, time-series data, and more), data models (relational data, document-oriented data, graph data), data volume, etc.

In Section 2, we begin with a brief introduction to related work. In Section 3, we illustrate our proposed approach through an example, and in Section 4, we present our conclusions.

2 BACKGROUND

Data quality management is crucial in big data platforms and has been extensively studied in the literature. Data quality has been described as "fitness for use" [29] for almost 30 years, which implies that it cannot be evaluated or enhanced without considering the context. There are several approaches to context-dependent data quality, such as assessing data quality based on user and location aspects [16, 17], defining contextual data quality dimensions [22, 29], and establishing context-aware data quality methodologies [24]. However, solutions still need to address big data quality management from a contextual perspective. Fadlallah et al. [3] reviewed the existing context-aware data quality assessment solutions, surveyed existing big data quality solutions, and then covered context-aware solutions. They found none of the current data quality assessment solutions could guarantee context awareness while handling big data. Some of their findings are mentioned below. The surveyed papers addressed only a partial view of the context, and some solutions focused only on data cleaning operations, which, unlike quality assessment solutions, do not allow determining whether the dataset can be used for its intended purpose. Some of the papers included in the survey use declarative operators based on the author's definition of quality dimensions, which may not match the quality dimensions required by the data context or the consumer's perspective. Finally, some of the papers reviewed only considered specific data quality dimensions.

Focusing on the weaknesses found in the literature review, in [4] the authors propose a big DQ assessment framework that considers context. In their framework data domain represents context, which is used to determine the quality dimensions that are being used. Our work propose a context definition that includes the data domain among other components. Furthermore, the quality dimensions are not determined by the data domain, instead are selected by an expert for each particular case. The main difference between the framework in [4] and our approach is that in we use the context to define the metrics for each data quality dimension, but not to select the dimensions that are being used.

3 CONTEXT-AWARE DATA QUALITY MANAGEMENT IN DATA LAKES

As mentioned above, DL give support to the co-existence of heterogeneous datasets, data storage for different levels of data processing, and various user profiles and requirements [9, 10]. Most authors agree that DL architectures should offer different zones for data, each meeting diverse storage and data consumption needs. It is also common for zone proposals to mention different processing and maturity levels of data. DL architectures should also consider how data is ingested into the DL and the metadata related to all the DL content and processes. Several proposals exist in the literature for DL architectures, which vary in many aspects. We suggest that readers interested in this topic refer to existent surveys [9, 11].

To ensure effective DQ management in the DL, it is important to identify specific locations and moments where DQ processes will be applied. Our approach to integrating DQ management in the DL considers the context in which the data is being processed. This means that we consider the data zone's characteristics, the data processing stage, and the user and task at hand. All of these factors significantly impact the data quality. This section will present a general DL architecture and our proposal for integrating context-aware DQ management into it.

3.1 Data Lake Architecture

We present a DL architecture designed to be as versatile as possible, considering the various needs that may arise in a big data analysis scenario. It draws inspiration from existing proposals, with the primary basis being the Zaloni Architecture [26]. However, we incorporate aspects from other DL architectures, such as those presented in [8, 13, 21, 30]. The following text describes our architecture's various zones and is illustrated in Figure 1. Single arrows indicate the flow of data, while double arrows indicate the flow of new data produced from analysis tasks. We provide the relevant reference if a component or aspect is based on an existing proposal.

(1) **Landing Zone:** This zone handles data ingestion and metadata extraction from the sources. If needed, it can run DQ, data compliance, and security (masking, anonymizing) checks. Data does not persist in this zone. [8, 26]
(2) **Raw Zone:** Data that has passed checks from the Landing Zone is stored in this zone in their raw form. [26]
(3) **Archival Zone:** Cold data (not frequently used data) from any of the zones is stored in this zone through offloading processes [11, 13]. In the future, if data is needed for analysis, it is sent to the same zone that it came from through offloading processes. Note that, to be able to send data to the same zone where it was previously stored, metadata about where data was persisted is needed.

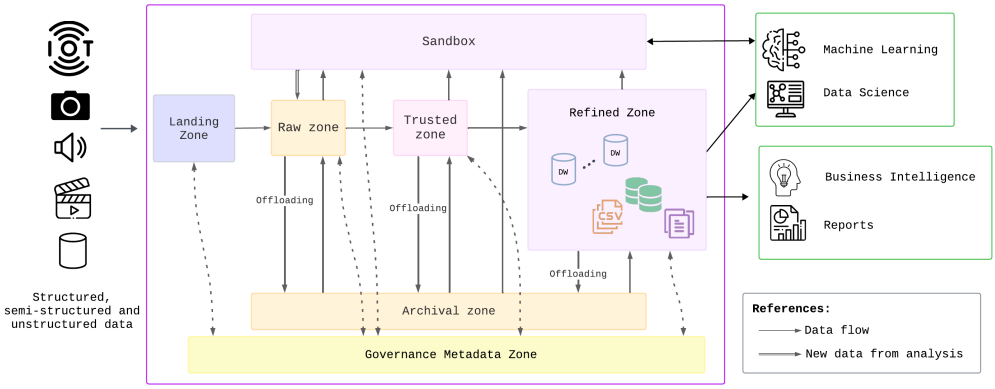

Figure 1: A general Zone-Based Data Lake architecture

(4) **Trusted Zone:** Data copied from the Raw Zone goes through transformations like data cleaning, integration, validation, etc. [8, 26] If new metadata is created, it should be stored in the metadata repository (Governance Zone)
(5) **Refined Zone:** Data copied from the Trusted Zone is modeled for business use cases and external tools [26]. This zone can store systems such as Data Warehouses to support Business Intelligence and Report tools.
(6) **Sandbox:** Data scientists access this zone to use and explore data from different zones in the DL [26]. Data obtained from analysis done in this zone can be sent to the Raw Zone.
(7) **Governance Metadata Zone:** This zone is in charge of managing general metadata (e.g.: dataset structure, processes, etc.) and DQ metadata (e.g.: DQ measures, DQ dimensions, etc.). It stores repositories for DQ and general metadata that any zone can access [19, 30].

3.2 Including Data Quality Management in the Data Lake

In this section, we emphasize the need for DQ management in the DL using an example to show the data lifecycle in the DL and the problems that arise. Then, we present our proposal to address them.

Consider the example presented in Section 1 and the DL architecture presented in Section 3.1. As mentioned above, an international health research institution wants to create a model to predict diseases based on a patient's symptoms. To achieve this, they receive heterogeneous data from multiple countries, which is then ingested into the DL. Each dataset contains information about the patient's age, gender, if the patient is a woman, whether she has given birth, chronic diseases, and existing illnesses in family members. This relevant information is extracted from the patient's clinical records.

We first present an example scenario without DQ management. Initially, data is sent to the *landing zone*, and metadata is extracted and stored in the *governance metadata zone*, while data is saved in the *raw zone*. Before any analysis is done, the data from different countries must be integrated to solve various heterogeneity issues. The *trusted zone* is used to achieve this because that is where the data is cleaned. The trusted zone is the zone where the data is supposed to be of reliable quality. As a result, the institution has a dataset that combines the symptoms of patients worldwide. The data is then sent to the sandbox to be used as an asset to train the model that predicts a patient's disease. However, the main problem in our example is that the training dataset contains biased data for some reason, with female population being under-represented. This biased data generates a model that predicts a patient's disease more accurately for males than females.

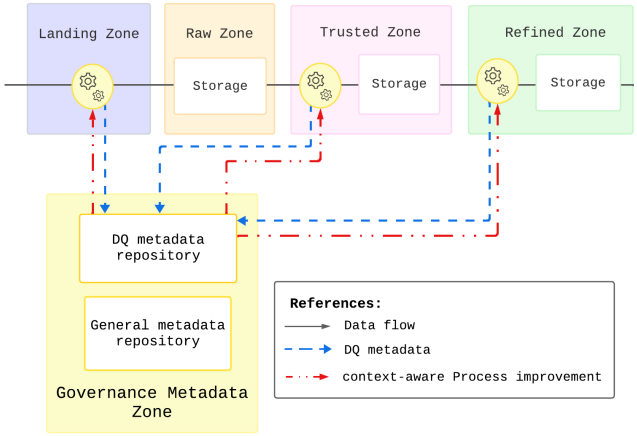

Figure 2: Data Quality Management in the Data Lake

We propose to include context-aware DQ in the DL, and Figure 2 shows how DQ processes are integrated into the DL. In this Figure, single black arrows indicate the flow of data, dotted blue arrows indicate the flow of DQ metadata and dotted red arrows indicate process improvements, derived from the DQ measurement assessment. DQ measurement and DQ metadata generation are carried out in the processing of each zone. The context relevant to this scenario includes DQ requirements and the DL zone from where the dataset will be extracted for usage. Accordingly, our DQ assessments depend on the DQ requirements posed by the researcher and the zone that stores the data being used.

Let C denote the context, such that $C =< R, Z, M >$, where R is a set of DQ Requirements, Z is the DL Zone, and M is the DQ metadata

repository, and let D denote the dataset. Consider a DQ assessment algorithm that, given a dataset D and a context C, returns the percentage of tuples of D that satisfy the DQ requirements R. The obtained DQ measure is stored in the repository M, which is located in the Governance Metadata Zone.

After the DQ assessment, DQ improvement may be applied. For this, the data processing applied in Z is modified, considering the DQ measures stored in M. When D is processed in the zone Z, it is cleaned so that it satisfies R, obtaining a DQ assessment of 100%.

Returning to our example, data arriving at the *trusted zone* is stored as an integrated dataset D. Considering that there is a DQ requirement R_1 that expresses the following: "This dataset must include the same quantity of female and male data registries", then $R = \{R_1\}$. We don't have any DQ metadata at this point so we can formalize our context as $C =< R, landing\ zone, \emptyset >$

For the DQ management process, the DQ dimension *coverage* is assessed considering the context. In this case, the context states a requirement, R_1, that the *coverage* dimension must include in its assessment. The value obtained in this assessment will reflect that the dataset D does not satisfy R_1 since it is not balanced between male and female data. This DQ metadata is stored in the *DQ metadata repository*, and the context is updated consequently (changing M). Later, DQ improvement is applied as a consequence of the assessment value stored in M, leading to modifications to the data processing of the *trusted zone*. These modifications cause, for example, data filtering so that the coverage requirements are satisfied and the resulting dataset becomes balanced. After that, the dataset is loaded into the *sandbox* and used for training the model.

4 CONCLUSION

This work introduced an adaptable approach for context-aware DQ management in a DL. Our mechanism accommodates various quality notions and is flexible enough to handle heterogeneous data types, models, and volumes. We also presented an example of a context-aware DQ metric that is geared towards enhancing fairness in the prediction model's result. In our approach, context includes user DQ requirements and the DL zone where data is utilized.

Future work is aimed at providing a more comprehensive proposal for DQ management in DL, with a specific focus on delineating the roles of the quality management components in each of the architecture zones and their interplay. Furthermore, we are working on implementing the proposal in a real-world scenario, with the potential to contribute to data quality management research.

ACKNOWLEDGMENTS

This work was partially funded by Comisión Sectorial de Investigación Científica (CSIC), Universidad de la República, Uruguay.

REFERENCES

[1] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal* (3 2016).

[2] Alex Bogatu, Alvaro A.A. Fernandes, Norman W. Paton, and Nikolaos Konstantinou. 2020. Dataset discovery in data lakes. *ICDE* 2020-April (4 2020), 709–720.

[3] Hadi Fadlallah, Rima Kilany, Houssein Dhayne, Rami El Haddad, Rafiqul Haque, Yehia Taher, and Ali Jaber. 2023. Context-aware Big Data Quality Assessment: A Scoping Review. *ACM Journal of Data and Information Quality* 15, 3 (8 2023), 33.

[4] Hadi Fadlallah, Rima Kilany, Houssein Dhayne, Rami El Haddad, Rafiqul Haque, Yehia Taher, and Ali Jaber. 2023. BIGQA: Declarative Big Data Quality Assessment. *ACM Journal of Data and Information Quality* 15 (8 2023). Issue 3. https://doi.org/10.1145/3603706

[5] Grace Fan, Jin Wang, Yuliang Li, and Renée J. Miller. 2023. Table Discovery in Data Lakes: State-of-the-art and Future Directions. (2023), 6975.

[6] Huang Fang. 2015. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. 820–824. https://doi.org/10.1109/CYBER.2015.7288049

[7] Donatella Firmani, Letizia Tanca, Politecnico Di Milano, and Riccardo Torlone. 2019. Ethical Dimensions for Data Quality. *J. Data and Information Quality* 12, 2 (2019), pages.

[8] Corinna Giebler, Christoph Gröger, Eva Hoos, Holger Schwarz, and Bernhard Mitschang. 2020. A zone reference model for enterprise-grade data lake management. In *2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC)*. IEEE, 57–66.

[9] Rihan Hai, Christos Koutras, Christoph Quix, and Matthias Jarke. 2023. Data Lakes: A Survey of Functions and Systems. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (12 2023), 12571–12590.

[10] Rihan Hai, Christoph Quix, and Matthias Jarke. 2021. Data lake concept and systems: a survey. *arXiv preprint arXiv:2106.09592* 10 (2021).

[11] Tomislav Hlupi, Draen Oreanin, Domagoj Ruak, and Mirta Baranovi. 2022. An Overview of Current Data Lake Architecture Models. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. 1082–1087. ISSN: 2623-8764.

[12] Oana Inel, Tim Draws, and Lora Aroyo. 2023. Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11, 1 (Nov. 2023), 51–64. https://doi.org/10.1609/hcomp.v11i1.27547

[13] Bill Inmon. 2016. *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications.

[14] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. *ACM International Conference Proceeding Series* (6 2022), 310–323.

[15] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* 117, 23 (2020), 12592–12594.

[16] Yang W. Lee. 2003. Crafting Rules: Context-Reflective Data Quality Problem Solving. *Journal of Management Information Systems* 20, 3 (Dec. 2003), 93–119.

[17] Anna L McNab and D Alan Ladd. 2014. Information quality: the importance of context and trade-offs. In *2014 47th Hawaii International Conference on System Sciences*. IEEE, 3525–3532.

[18] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data lake management: challenges and opportunities. *Proc. VLDB Endow.* 12, 12 (aug 2019), 19861989. https://doi.org/10.14778/3352063.3352116

[19] Lamya Oukhouya, Anass El Haddadi, Brahim Er-raha, and Hiba Asri. 2021. A generic metadata management model for heterogeneous sources in a data warehouse. *E3S Web of Conferences* 297 (2021), 01069.

[20] Evaggelia Pitoura. 2020. Social-minded Measures of Data Quality: Fairness,Diversity, and Lack of Bias. *Journal of Data and Information Quality* 12, 3 (7 2020).

[21] Christoph Quix and Rihan Hai. 2019. Data Lake. In *Encyclopedia of Big Data Technologies*, Sherif Sakr and Albert Y. Zomaya (Eds.). Springer International Publishing, Cham, 552–559.

[22] Galina Rogova, Melita Hadzagic, Marie-Odette St-Hilaire, Mihai C Florea, and Pierre Valin. 2013. Context-based information quality for sequential decision making. In *2013 IEEE (CogSIMA)*. IEEE, 16–21.

[23] Pegdwendé Sawadogo and Jérôme Darmont. 2021. On data lake architectures and metadata management. *Journal of Intelligent Information Systems* 56 (2 2021), 97–120. Issue 1.

[24] Flavia Serra, Veronika Peralta, Adriana Marotta, and Patrick Marcel. 2023. Context-Aware Data Quality Management Methodology. *Communications in Computer and Information Science* 1850 CCIS (2023), 245–255.

[25] Nima Shahbazi, Yin Lin, H V Jagadish, and Abolfazl Asudeh. 2021. Representation Bias in Data: A Survey on Identification and Resolution Techniques. (2021), 47.

[26] Ben Sharma. 2018. *Architecting Data Lakes* (2nd ed.). O'Reilly Media, Inc.

[27] Divesh Srivastava, Monica Scannapieco, and Thomas C. Redman. 2019. Ensuring High-Quality Private Data for Responsible Data Science. *Journal of Data and Information Quality (JDIQ)* 11, 1 (1 2019), 1–9.

[28] Julia Stoyanovich, Bill Howe, and HV Jagadish. 2020. Responsible Data Management. 13, 12 (2020), 3474–3489.

[29] Diane M. Strong, Yang W. Lee, and Richard Y. Wang. 1997. Data quality in context. *Commun. ACM* 40 (5 1997), 103–110. Issue 5.

[30] Yan Zhao, Imen Megdiche, Franck Ravat, and Vincent-nam Dang. 2021. A Zone-Based Data Lake Architecture for IoT, Small and Big Data. In *Proceedings of the 25th International Database Engineering & Applications Symposium* (Montreal, QC, Canada) *(IDEAS '21)*. Association for Computing Machinery, New York, NY, USA, 94102.