An Automated Tool for Multi-Dimensional Data Quality Assessment

Pedro Guimarães[2,1,†], Filipe Santos[2,1,†], António C. Vieira[1,2,†] and Maribel Y. Santos[1,2,†]

[1]ALGORITMI Research Centre, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
[2]CCG/ZGDV Institute, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal

Abstract

Ensuring high-quality data is critical for effective analytics and data-driven decision-making. There have been substantial advances in defining data quality dimensions and frameworks that provide conceptual guidance, as well as commercial tools that enable validation, profiling, and cleansing processes. However, there is still a lack of a tool that provides an automatic assessment of literature-compliant data quality indicators. This paper introduces a data quality tool capable of automatically measuring multiple data quality indicators across established dimensions and of generating a comprehensive, user-friendly classification. This way, the tool offers insights of data quality indicators compliant with established dimensions and frameworks. The tool is validated using three datasets: two from real industrial cases and another one from a public dataset, to ensure the replicability of the performed evaluation. The results revealed good indications regarding the capability of the tool in assessing data quality problems, and in evaluating data quality dimensions and the overall data quality in an automatic, interpretable, and user-defined manner.

Keywords

Data quality, Data quality dimensions, Data quality problems, Data Governance, Automatic Evaluation

1. Introduction

Data has become a critical asset for organizations across all sectors, driving decision-making, operational efficiency, and the development of AI-driven systems [1, 2]. In fact, the reliability of insights, predictions, and automated processes depends heavily on the quality of the underlying data, while at the same time, modern data environments are increasingly complex, characterized by heterogeneous sources, large-scale datasets, and real-time streams [3, 4]. In this context, ensuring that data is accurate, complete, consistent, timely, and fit for its intended purpose is a central concern in data engineering.

Despite its recognized importance, maintaining high-quality data remains a significant challenge. Common issues, such as inconsistencies, duplicates, missing values, and integrity violations frequently occur across both structured and unstructured datasets, impacting multiple data quality dimensions [5, 6]. While international standards and conceptual frameworks provide guidance for defining and measuring data quality, their practical implementation in dynamic, large-scale, or automated environments is limited. The formalized procedures and metrics outlined by standards, such as ISO/IEC 25012 and ISO/IEC 25024, may be difficult to operationalize in real-time pipelines, highlighting a critical gap between theoretical guidance and practical automated enforcement of data quality.

Conversely, a range of frameworks and software tools have been proposed to address these challenges. Methodological frameworks, including Total Data Quality Management (TDQM), Big Data Quality Management (BDQM) [7], and Luzzu [8], offer structured approaches for monitoring and improving data quality across different domains and data lifecycles. Complementing these frameworks, commercial

PoEM2025: Companion Proceedings of the 18th IFIP Working Conference on the Practice of Enterprise Modeling: PoEM Forum, Doctoral Consortium, Business Case and Tool Forum, Workshops, December 3-5, 2025, Geneva, Switzerland

[†]These authors contributed equally.

pedro.guimaraes@ccg.pt (P. Guimarães); filipe.santos@ccg.pt (F. Santos); avieira@dsi.uminho.pt (A. C. Vieira); maribel@dsi.uminho.pt (M. Y. Santos)
0000-0003-3390-8528 (P. Guimarães); 0000-0002-1059-8902 (A. C. Vieira); 0000-0002-3249-6229 (M. Y. Santos)

and open-source tools, such as Informatica[1], Talend[2], Great Expectations[3], and Apache Beam[4] provide operational capabilities for automated validation, profiling, and cleansing of datasets. While these tools enable automation, they leave a gap for a solution capable of automatically quantifying multiple data quality indicators and frame such indicators in data quality dimensions, hence delivering a holistic and interpretable assessment for end-users.

Grounded on the above, this paper proposes a methodology and a supporting tool for automated data quality assessment that integrates established quality dimensions into a unified and interpretable evaluation process. The tool automatically quantifies indicators across multiple dimensions and synthesizes them into a holistic classification, offering users an automatic and accessible view of their data quality. Furthermore, its modular architecture, combining a graphical configuration interface, a backend orchestrated with Apache NiFi[5], and a web-based visualization layer, ensures seamless interaction between configuration, processing, and presentation.

This paper is structured as follows. Section 2 discusses literature related to the topics of this work, namely the identification of relevant data quality issues, metrics and dimensions for data quality, as well as standards, frameworks and existing tools. Section 3 proposes the methodology that guided the assessment of the data quality indicators incorporated in the developed tool. Section 4 discusses the architecture of such tool and its development process. Section 5 presents the application of the tool in selected use cases. Finally, the Section 6 discusses the main conclusions.

2. Related Work

Data quality refers to the degree to which data meets the requirements for its intended use, encompassing aspects such as accuracy, completeness, consistency, and timeliness [9, 10]. This topic has been explored across diverse domains, such as management, computer science, statistics and medicine, where the integrity and reliability of information are crucial for data-driven decision-making [11, 12].

Recent research has highlighted increasing efforts to standardize data quality evaluation, notably through standards, such as ISO/IEC 25012 and ISO/IEC 25024 [13, 14], which define data quality dimensions and measurement criteria. In parallel, frameworks have been developed to guide the management of data quality across different organizational domains [15, 7]. Simultaneously, automated solutions have emerged that can integrate these evaluations into data pipelines and analytical systems, enabling continuous monitoring and validation of data quality [16, 17].

Dimensions are fundamental criteria used to assess whether data is fit for the analytical purpose. Across the literature, among the most widely accepted dimensions are: *Accuracy*, *Completeness*, *Consistency*, and *Timeliness*. The former evaluated the degree to which values reflect reality. *Completeness* measures the extent to which all expected values are present in the data. *Consistency* measures the extent to which data values are coherent and conform across different sources and formats. Finally, *Timeliness* measures whether data is available when needed for its intended use [18, 17, 19, 13]. Building on these core dimensions, some frameworks additionally consider *Accessibility* and *Integrity* as relevant dimensions, particularly in Big Data contexts [20, 21], with the former measuring the ease with which data can be retrieved and used by authorized users when needed, and the latter assessing the protection of data from unauthorized modification or corruption.

Data quality in real-world systems is often compromised by a variety of specific issues, including inconsistencies, duplications, missing values, and coding errors. These problems can arise at both the instance and schema levels, and may originate from single or multiple data sources [22, 23, 24]. Common problem types affect multiple dimensions of data quality, such as completeness, accuracy, consistency, timeliness, accessibility, and uniqueness, as illustrated in Table 1. Missing or null values primarily impact completeness and accessibility, while incorrect data types or formatting compromise accuracy

[1]https://www.informatica.com/products/data-quality/cloud-data-quality-radar.html
[2]https://www.talend.com/products/data-quality/
[3]https://greatexpectations.io/
[4]https://beam.apache.org/
[5]https://nifi.apache.org

and consistency. Duplications affect both uniqueness and consistency, outdated temporal values reduce timeliness, and ambiguous or poorly named columns hinder accessibility and can introduce errors in accuracy. Violations of uniqueness or referential integrity further threaten consistency and uniqueness.

These challenges have significant consequences for the trustworthiness of data and the reliability of analyses, ultimately influencing both operational and strategic decision-making processes [25, 26]. Understanding and addressing these issues is therefore critical in the design of data engineering pipelines and governance frameworks that aim to maintain high-quality, reliable datasets.

Regarding standards, international frameworks provide formal structures for defining, measuring, and assessing data quality, offering organizations a common language and methodology. ISO/IEC 25012, for instance, establishes a conceptual model comprising 15 data quality characteristics, categorized into *inherent* and *system-dependent* quality, providing a comprehensive perspective on both the intrinsic properties of data and its behavior within systems [13]. Building on this, ISO/IEC 25024 defines standardized metrics for each dimension, including formal definitions, calculation formulas, and recommended thresholds, thereby enabling systematic and repeatable evaluation of data quality across diverse contexts [14]. ISO 8000 complements these approaches by emphasizing data interoperability and consistency in business environments, particularly in supply chain management, ensuring that data exchanged across heterogeneous systems remains accurate, complete, and usable [27].

These standards have been successfully applied in domains, such as finance and healthcare, where robust data governance and regulatory compliance demand high-quality data. Nevertheless, applying these standards in dynamic or real-time systems can be challenging, particularly when integrating formalized metrics into automated data pipelines, which suggests opportunities for further research and engineering solutions. Therefore, addressing this gap requires strategies to translate standardized quality frameworks into scalable, monitored solutions within modern data engineering ecosystems.

In response to these challenges, several frameworks have been proposed to operationalize data quality assessment providing more practical structures that can be adapted to specific organizational and technical contexts. The *Data Quality Framework* focuses on continuous monitoring throughout the data lifecycle, providing mechanisms to evaluate and improve quality as data flows through systems [28]. The *Total Data Quality Management* approach [29, 30] emphasizes continuous improvement processes, aligning data quality initiatives with organizational management practices. In turn, for large-scale environments, Taleb et al. [1] proposed the *Big Data Quality Management Framework*, which addresses the specific challenges of volume, variety, and velocity in analytical settings. In parallel, the *Luzzu* framework has been developed for Linked Data scenarios, offering flexibility through the definition of custom metrics and extensible quality assessments.

The choice of framework depends on the type of data, the scale of operations, and the organizational context in which it is applied. However, these frameworks highlight the need for adaptable approaches that can be tailored to the requirements of specific domains and data engineering ecosystems. To operationalize these frameworks in practice, organizations often rely on data quality tools such as **Informatica Data Quality**, **Talend**, and **Great Expectations**, which provide concrete implementations for monitoring, validating, and cleansing data with automated pipelines, as listed in Table 1.

The literature demonstrates substantial progress in defining data quality dimensions, metrics, and assessment technologies. However, there is a need for tools that can automatically quantify data quality indicators across multiple dimensions and provide a holistic classification of datasets or data streams. Such tools would enable continuous monitoring, support informed decision-making, and reduce the manual effort required to interpret quality assessments. By integrating automated evaluation with intuitive summarization, these solutions would allow organizations to maintain high-quality, trustworthy data in complex, dynamic systems, bridging the gap between conceptual frameworks and practical, operational data engineering.

Table 1
List of technologies for automated data quality assessment

Technology	Key Features	Application Context	Limitations
Chug et al. (2021)	Rule-based scoring system	Academic and business environments	Lacks flexibility and domain adaptation
Informatica Data Quality	ML-based automatic evaluation	Regulated sectors (e.g., finance, health)	High cost; limited integration
Talend Data Quality	Validation rule definition and pattern recognition	Business organisations	Dependent on the Talend ecosystem
Apache Kafka Streams	Stream-based real-time analysis	Big Data, IoT	Requires robust infrastructure
Apache Beam	Distributed processing with built-in evaluation	Scalable analytical systems	Steep learning curve

3. Data Quality Assessment Methodology

This section introduces the data quality assessment methodology, which comprises three main steps: (i) selecting the data quality dimensions to be adopted, (ii) mapping specific data quality problems to these dimensions, and (iii) aggregating the resulting quantitative measures into an overall data quality score. This score is then associated with a qualitative rating that characterizes the dataset (Figure 1). In the final step, each dimension's score is calculated as the complement of a weighted, normalized error rate for the identified problems. The overall dataset score is obtained as a weighted average of all dimension scores. Finally, the quantitative result is mapped to a qualitative scale to provide an intuitive interpretation of the dataset's quality.

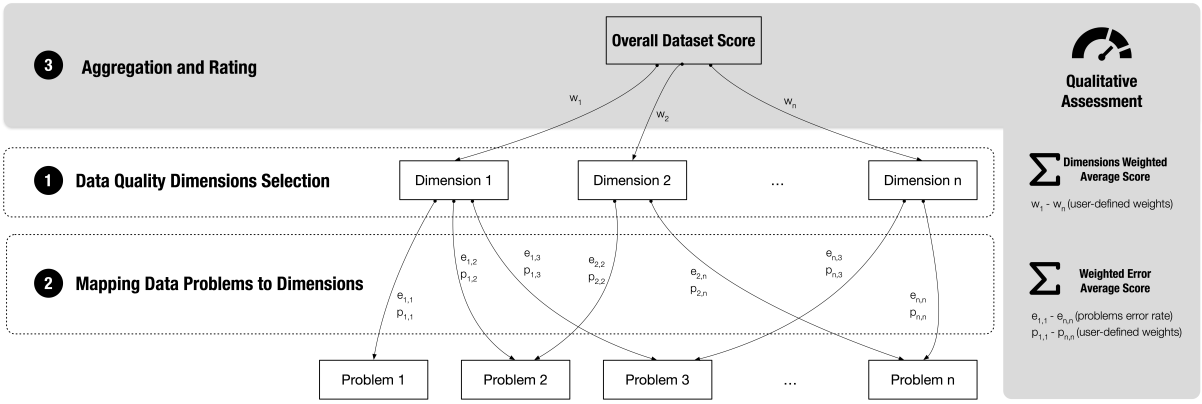

Figure 1: Proposed data quality assessment methodology.

3.1. Data Quality Dimensions Selection

As discussed in Section 2, the literature shows strong convergence around four fundamental data quality dimensions, namely *Accuracy*, *Completeness*, *Consistency*, and *Timeliness*, which had been originally established by Wang and Strong [18] and consolidated in the ISO/IEC 25012 standard. However, there are studies that mention other main data quality dimensions. For instance, Cichy and Rass [28] consider *Accuracy*, *Completeness* and *Timeliness* as the main ones, followed by *Accessibility* and *Uniqueness*. For Big Data contexts, Taleb et al. [7] also considered *Accuracy*, *Completeness*, *Consistency*, and *Timeliness*, which they unfied into a single one labelled as *Intrinsic*, along with *Representational*, *Contextual*, and *Accessibility*. Likewise, Debattista et al. [8] also suggest a view of data quality dimensions in the context of linked data. As can be seen, different authors propose different classifications of data quality dimensions, including different levels of detail, different requirements, and different criteria.

Given the above, the selection of dimensions in the context of the proposed tool was mainly guided by objectivity in the quality assessment. Other dimensions found in the literature, such as *Credibility*, *Understandability*, or *Reputation*, were excluded because they rely on subjective or contextual interpretation and cannot be evaluated through objective indicators on a dataset. Thus, this work considers six main dimensions that can be quantified using metrics derived directly from the dataset, enabling a transparent and replicable quality assessment. Furthermore, the selection of the dimensions was also guided by the intention of incorporating a comprehensive list of data quality specific problems or indicators, that can be associated with such dimensions. However, it should be noted that the purpose of this paper is to present the developed tool for the pre-selected set of dimensions, rather than arguing for the choices that guided the selection of the data quality dimensions, as these can be easily adapted in our tool. The following is the list of established data quality dimensions and their aim in terms of data quality assessment:

- **Accessibility** evaluates the structural usability of the dataset. Errors such as duplicate rows, empty rows, empty columns, or ambiguous column names are detected automatically as indicators of reduced usability and interpretability. In the tool, this dimension reflects how well a dataset is structured for both human interpretation and system-level processing.
- **Accuracy** evaluates the reliability of individual values in the dataset. It includes errors that indicate deviations from the expected or true values, such as incorrect data types, missing or empty entries, outliers, invalid or outdated dates, negative values, or ambiguous column names. Within the tool, these indicators are computed by scanning the dataset for value anomalies, incorrect typing, or structural inconsistencies, producing a quantitative measure of trustworthiness for analytical use.
- **Completeness** measures the extent to which all expected information is available. It focuses on missing data and empty strings, automatically calculated as the ratio of present versus expected values. In the tool, this dimension reflects the overall data availability, allowing users to determine whether their dataset provides sufficient information for downstream analysis.
- **Consistency** captures the internal coherence of the dataset by detecting patterns that violate formatting or logical uniformity. The tool checks for mismatched data types, irregular date formats, white spaces, or unexpected special characters, as well as outliers that break numeric coherence. The dimension therefore measures whether the dataset behaves predictably and can be integrated across sources or analytical processes without contradiction.
- **Timeliness** assesses temporal validity and data freshness. The tool automatically identifies invalid or outdated timestamps, comparing the detected dates against current system time or defined thresholds. This dimension provides an interpretable indication of how up-to-date the dataset is for its intended analytical or operational context.
- **Uniqueness** focuses on redundancy detection by identifying duplicate rows in the dataset. Within the tool, this dimension ensures that each record represents a unique entity, preventing double counting and data duplication. Although duplicate rows also affect *Accessibility*, they are treated here as integrity violations specifically related to identity.

3.2. Mapping Data Quality Problems to Dimensions

Although the literature identifies a large variety of potential data quality issues, many are described at an abstract level (e.g., semantic or referential inconsistencies) that cannot be automatically detected without domain-specific context. Therefore, the proposed approach focuses exclusively on data quality problems that are: i) directly observable in structured tabular data (e.g., CSV files); ii) quantifiable through objective, numerical metrics; and, iii) automatically detectable via Python-based validation scripts integrated into the processing pipeline. This methodological constraint ensures that every measured indicator corresponds to an error type that the tool can autonomously identify and quantify. The mapping between selected problems and their associated dimensions is summarised in Table 2.

Table 2
Data quality problems considered per data quality dimension.

Dimension	Identified Data Quality Problems
Accessibility	Duplicate rows; empty rows; empty columns; ambiguous column names.
Accuracy	Missing data; incorrect data types; spelling errors; invalid or outdated dates; empty strings; negative values; outliers; ambiguous column names.
Completeness	Missing data; empty strings.
Consistency	Incorrect data types; special characters; negative values; date formatting errors; spelling errors; white spaces; outliers.
Timeliness	Invalid dates; outdated temporal values.
Uniqueness	Duplicate rows.

As seen in Table 2, certain data quality problems, such as duplicate rows or ambiguous column names, appear in more than one dimension. This reflects the multidimensional nature of data quality, since, for instance, duplicate rows impact *Uniqueness* by violating the one-entity-one-record principle, while also affecting *Accessibility* by impairing structural usability. Similarly, ambiguous column names reduce both *Accuracy* (due to possible misinterpretation of values) and *Accessibility* (by hindering readability). These overlaps are handled through a weighting mechanism that allows users to control the relative influence of each problem on each dimension, ensuring flexibility and contextual relevance.

3.3. Quantification of the Data Quality Score: Aggregation and Rating

Standards such as ISO/IEC 25012 provide a conceptual model and illustrative ratio-based calculations but do not prescribe mandatory formulas. In line with this, the proposed approach considers each dimension score as the complement of the weighted, normalised error rates observed for that dimension. Formally, the score S for dimension d is computed as:

$$S_d \;=\; 1 - \sum_{j=1}^{m} \left(p_j \cdot e_j\right)$$

where m is the number of problems mapped to d, p_j is the user-defined weight of problem j, and e_j is the normalised error rate. As can be seen, we allow the inclusion of weights as a way for users to specify the most important data quality problems for each data quality dimension, according to their data contexts. Once the score of each dimension is obtained, the overall score for the dataset quality is calculated as a weighted average of all active dimensions:

$$S_{overall} = \frac{\sum_{i=1}^{n}(S_i \cdot w_i)}{\sum_{i=1}^{n} w_i}$$

where S_i is the score of dimension i, w_i is its user-defined weight, and n is the number of active dimensions. Thus, users can define weights for the quality problems per dimension, as well as for each dimension, in order to obtain a final quality score for the dataset.

Finally, this overall score is mapped into a qualitative scale inspired by the European Credit Transfer and Accumulation System (ECTS) (Table 3), as this classification enables intuitive communication of results, ensuring consistency when comparing different datasets.

Table 3
Qualitative scale for dataset quality assessment.

Overall Score	≥ 0.90	≥ 0.80	≥ 0.70	≥ 0.60	≥ 0.50	< 0.50
Qualitative Rating	A	B	C	D	E	F

4. Tool Architecture and Development

This section presents the technological architecture for the data quality tool that instantiates the presented methodology. The solution was designed to enable the automated assessment of data quality in CSV files, providing end users with an interactive interface for configuration and results visualization. The tool is built on a modular architecture, aiming to adapt to different analytical contexts. The architecture is structured into four main layers (see Figure 2):

- **Interaction Layer:** Enables users to upload CSV files, specify data types for each column, choose relevant data quality dimensions, and assign custom weights to tailor the assessment process, and was developed using *Gradio*.
- **Integration Layer:** Manages incoming data and parameters, stores files in HDFS (Hadoop Distributed File System), assigns a unique identifier (UUID) to files, and interfaces with the orchestration system (*Apache Nifi*) that acts as a middleware. It was implemented using *FastAPI*.
- **Storage and Processing Layer:** Detects data errors and computes quality scores. Results are stored in a MySQL database for further analysis and was developed using *Apache NiFi* and Python.
- **Visualisation Layer:** Enables users to view processed outcomes, including individual dimension scores, the overall quality score, and a qualitative evaluation. It was built with the support of *Angular*.

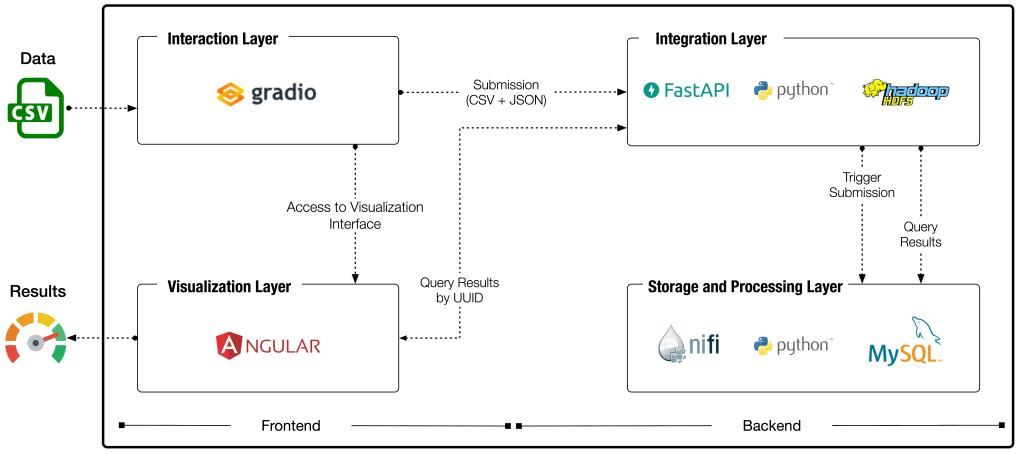

Figure 2: Technological architecture of the data quality assessment tool.

The developed solution includes a frontend based on Gradio and Angular, with a web interface for results visualisation. The backend is orchestrated with Apache NiFi. HDFS is used to physically store datasets, enabling large files to be split into chunks and processed efficiently in batches. This modular architecture enables automated, customisable, and accessible assessment of the quality of structured data, ensuring a smooth transition between configuration, processing, and results presentation.

The interface provides a graphical solution that abstracts the complexity of the technical process. The user can configure the evaluation parameters based on the uploaded CSV file, i.e., expected data types per column, additional constraints (e.g., no negative values or possible outliers), and assignment of weights to the selected quality dimensions and error types.

The tool performs a local validation to ensure that the sum of the weights is correct. If no inconsistencies are found, it generates two JSON objects: one containing the basic parameters and another detailing the selected dimensions and errors. These are sent, along with the original file, to the Integration Layer via HTTP, triggering the evaluation process. Additionally, upon successful submission, the interface dynamically generates a redirect button to the result visualisation frontend, using the UUID assigned to the submission. In the proposed solution, data processing is orchestrated using Apache NiFi, which enables controlled and modular task management. To handle dependencies between steps (such as

ensuring that error detection occurs before score calculation), the pipeline was divided into five steps, and one step for updating processing status, as illustrated in Figure 3.

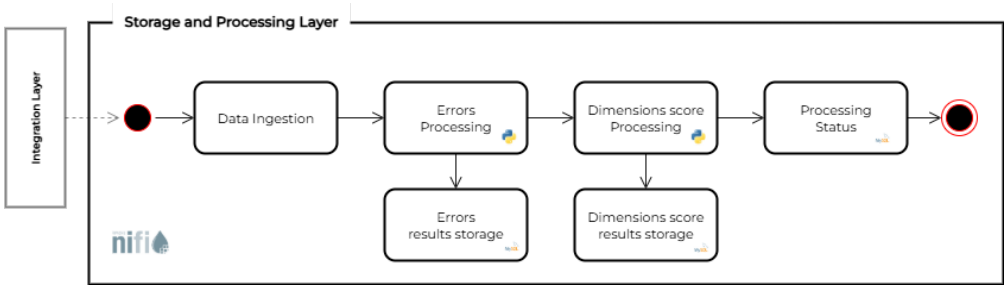

Figure 3: Modular orchestration of the pipeline in Apache NiFi.

- **Data Ingestion:** receives data from the Integration Layer, including the CSV file (stored in HDFS), the UUID, and the parameters defined by the user.
- **Error Processing:** runs a Python script that reads the file, applies the quality rules, and returns a JSON object containing the detected error metrics.
- **Error results storage:** stores the detected errors in a MySQL database, ensuring persistence and reusability for the next step.
- **Dimensions score Processing:** invokes a second Python script that reads the stored errors, applies the defined weights, and calculates the scores per dimension, the overall score, and the qualitative rating (A–F).
- **Dimensions score storage:** saves the final results in the MySQL database, making them available for later consultation.

This architecture ensures robustness and flexibility, allowing multiple submissions to be processed in parallel and avoiding dependency issues between steps. Regarding the presentation of results to the users, the application communicates with the Integration Layer components to retrieve the scores for each dimension, the overall score, and the corresponding qualitative classification. The results are displayed using interactive semicircular gauges (via *ngx-gauge*), featuring visual thresholds and descriptive explanations. Each metric is accompanied by a short textual description that helps the user understand its relevance, even without technical expertise.

5. Results: Data Quality Tool Evaluation

In order to validate the practical applicability of the developed tool and assess the reliability of the metrics produced, three distinct use cases were conducted, the first two using real-world datasets from industrial environments, and the last one using a public dataset to ensure replicability of this evaluation. The evaluation aimed to demonstrate the tool's ability to identify and quantify data quality issues, generate interpretable metrics per dimension, and ensure consistency with established normative references, namely the ISO/IEC 25012 standard.

5.1. Use Case A

The first use case is based on a dataset from a manufacturing company, containing records related to production orders, materials, operations, planned and actual times, and quality classifications. The dataset includes 23 columns and approximately 500 rows, combining integer, text, date, and decimal data types. During the initial analysis, several anomalies were identified, such as ambiguous column names, missing values in critical fields, and inconsistent date formatting. The tool was applied with all quality dimensions enabled and with the default pre-defined weights for all dimensions and their corresponding errors.

The evaluation produced a **global score of 87.8%**, corresponding to a qualitative classification of **Level B** (Figure 4). The dimensions *Completeness* (95.5%) and *Accessibility* showed the highest results, while *Consistency* recorded the lowest score (66%) due to formatting inconsistencies, spelling errors, and structural heterogeneity in the data.

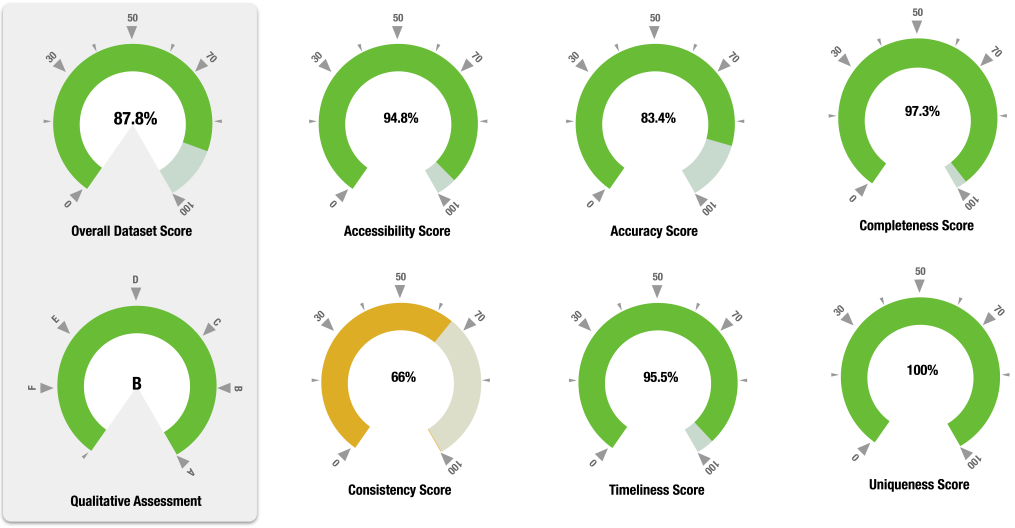

Figure 4: Data quality results for use case A.

5.2. Use Case B

The second validation scenario involves a dataset collected from a stone-cutting machine, containing technical variables such as speed, rotation, electric current, vibration, and temperature. The dataset comprises 12 columns and 499 records, each representing a timestamped measurement. This dataset proved to be structurally cleaner and more consistent, with few errors detected. As in the previous case, the tool was configured with all standard parameters and all six quality dimensions enabled.

The evaluation produced a **global score of 96.7%** and a qualitative classification of **Level A** (Figure 5). The dimensions *Completeness*, *Uniqueness*, and *Timeliness* achieved perfect scores of 100%, while the remaining dimensions also scored highly: *Accuracy* (94.3%), *Consistency* (93.5%), and *Accessibility* (95.7%).

Figure 5: Data quality results for use case B.

5.3. Use Case C

This validation scenario uses the IMDb (Internet Movie Database) Top-1000 Movies and TV Shows dataset, publicly available on Kaggle[6]. The dataset comprises information on 1,000 of the most highly rated films and television series according to IMDb rankings. Each record represents a unique title and includes a range of descriptive, numerical, and categorical attributes capturing various aspects of its metadata. Key attributes include the title, release year, age certification, runtime, genre, IMDb rating, Metascore, director, and up to four main cast members. Additional fields, such as number of votes, box office gross, poster link, and a short overview, provide both quantitative and qualitative dimensions for analysis.

The dataset exhibits several typical data quality challenges. These include missing values and inconsistent data types. Such imperfections make it particularly suitable for evaluating the performance of data quality assessment tools across multiple dimensions.

The dataset produced a **global score of 88%** and a qualitative classification of **Grade B**, indicating generally high data quality with some inconsistencies (Figure 6). *Accuracy* (90.7%) was high but affected by mismatch type errors (23% of cells), moderate missing values (2.7%), and minor textual noise. *Completeness* (98.4%) was excellent, with nearly all fields populated and no empty strings. In contrast, *Consistency* (57.3%) was the weakest dimension, due to frequent type inconsistencies, excessive spacing, and special characters. Both *Accessibility* and *Uniqueness* scored 100%, confirming a structurally clean and well-organized dataset. Overall, the results highlight that while the dataset is complete and easily accessible, improvements in formatting and data standardization are needed to enhance internal consistency and semantic accuracy.

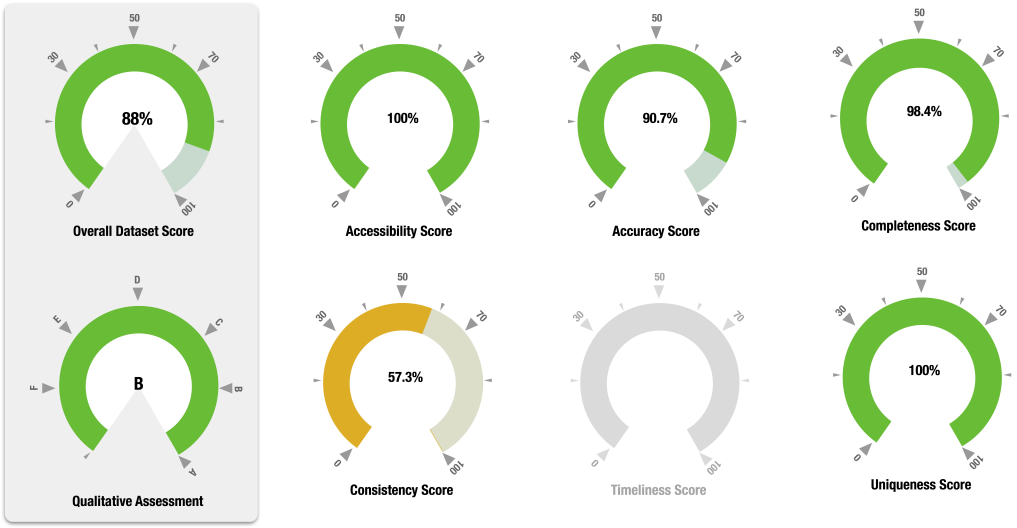

Figure 6: Data quality results for use case C.

5.4. Summary and Final Remarks

The three analysed use cases validated the tool's behaviour in distinct and realistic scenarios. In all contexts, the solution demonstrated: The ability to autonomously detect and quantify multiple types of errors; The generation of normalised scores per dimension and an overall score adjusted by weights; Conceptual compatibility with the ISO/IEC 25012 standard; Inclusion of practical dimensions such as *Uniqueness* and *Timeliness*, which are not covered by ISO/IEC 25012; The ability to allow users to specify different weights according to their domain knowledge and analytical requirements.

The ability to assign custom weights to each data quality problem and each dimension, as defined by the user in the configuration interface, grants the solution a high level of adaptability. This degree of

[6]https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows/data

personalization allows the evaluation to be tailored to the goals of each organisation or domain while preserving the objectivity of the results. The tool delivers a detailed evaluation, supported by both visual and analytical outputs, making it suitable for both technical and non-technical user profiles.

6. Conclusion

The critical role of data quality in decision-making underscores the need for automated, interpretable, and adaptable solutions. This paper addressed this challenge by proposing a supporting tool that operationalizes data quality assessment across multiple recognized data quality dimensions. The approach not only enables the detection and quantification of diverse data issues but also synthesizes the results into a comprehensive and user-friendly classification, offering a holistic view of data quality.

The tool was validated using two datasets from real industrial use cases and an additional public dataset. The latter ensures replicability of this evaluation, while the former datasets allow for testing its robustness in more complex cases. The obtained results of the proposed solution indicated that the tool is effective in assessing data quality across different contexts, promoting a more complete, adaptable, and interpretable model with strong potential for real-world decision-support applications.

A promising direction for future work involves applying the proposed tool across a variety of data quality frameworks and organizational contexts to further demonstrate its modularity and adaptability. By integrating the tool into diverse data governance ecosystems, it would be possible to evaluate how well its modular architecture supports different workflows, data life-cycles, and quality frameworks.

Acknowledgments

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Unit Project Scope UID/00319/Centro ALGORITMI (ALGORITMI/UM). This paper uses icons made available by www.flaticon.com.

Declaration on Generative AI

During the preparation of this work, the authors used CoPilot and Grammarly for sentence polishing, reword and rephrasing. All generated content was reviewed and edited by the authors, who take full responsibility for the final text.

References

[1] I. Taleb, M. A. Serhani, R. Dssouli, Big data quality: A comparative study of data quality frameworks, Big Data Research (2021).

[2] A. A. Vieira, L. M. Dias, M. Y. Santos, G. A. Pereira, J. A. Oliveira, Supply chain data integration: A literature review, Journal of Industrial Information Integration 19 (2020) 100161.

[3] Y. Cao, F. Q. A. Alyousuf, A new framework to assess the impact of new it-based technologies on the success of quality management system, Journal of Big Data 12 (2025) 8.

[4] H.-J. Zhang, C.-C. Chen, P. Ran, K. Yang, Q.-C. Liu, Z.-Y. Sun, J. Chen, J.-K. Chen, A multi-dimensional hierarchical evaluation system for data quality in trustworthy ai, Journal of Big Data 11 (2024) 136.

[5] V. Lindström, F. Persson, A. P. C. Viswanathan, M. Rajendran, Data quality issues in production planning and control–linkages to smart ppc, Computers in Industry 147 (2023) 103871.

[6] A. Polimeno, C. Braghin, M. Anisetti, C. A. Ardagna, Maximizing data quality while ensuring data protection in service-based data pipelines, Journal of Big Data 12 (2025) 62.

[7] I. Taleb, M. A. Serhani, C. Bouhaddioui, R. Dssouli, Big data quality framework: a holistic approach to continuous quality management, Journal of Big Data 8 (2021) 76.

[8] J. Debattista, S. Auer, C. Lange, Luzzu—a methodology and framework for linked data quality assessment, Journal of Data and Information Quality (JDIQ) 8 (2016) 1–32.
[9] A. Alizamini, F. Shams, A. Emrouznejad, Understanding data quality: A review of definitions and dimensions, International Journal of Information Management (2010).
[10] F. Sidi, P. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, A. Mustapha, Data quality: A survey of data quality dimensions, Journal of Information and Software Technology (2012).
[11] C. Batini, M. Scannapieco, Methodologies for data quality assessment and improvement, ACM Computing Surveys (2009).
[12] M. Hassany, S. S. Salim, H. Ibrahim, A. Mustapha, Review of data quality research: A survey approach, Journal of Theoretical and Applied Information Technology (2013).
[13] ISO/IEC, Iso/iec 25012: Software product quality requirements and evaluation (square) — data quality model, International Organization for Standardization (2008).
[14] I. Caballero, A. Caro, M. Piattini, Assessing data quality metrics with iso/iec 25024, Information Systems (2022).
[15] C. Cichy, S. Rass, A framework for data quality evaluation in big data environments, Journal of Big Data (2019).
[16] A. Chug, R. Sharma, R. Sehgal, A scoring model for domain-independent data quality evaluation, Journal of Big Data (2021).
[17] L. Ehrlinger, W. Wöß, Data quality: Research challenges and future directions, Journal of Data and Information Quality (2022).
[18] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems 12 (1996) 5–33.
[19] N. Laranjeiro, J. Bernardino, M. Vieira, A survey on data quality: Concepts, dimensions and challenges, Computer Science Review (2015).
[20] X. Zhang, X. Yang, H. Li, A survey on data quality: Classifications, assessment methods, and tools, ACM Computing Surveys (2019).
[21] Y. Ji, H. Wang, X. Zheng, X. Wang, Quality assurance of big data: A review, IEEE Access (2020).
[22] N. Laranjeiro, J. Bernardino, M. Vieira, A data quality assessment methodology for sql and nosql databases, Information Systems 63 (2016) 1–20.
[23] A. Hassenstein, P. Vanella, Challenges of poor data quality in organizations: An overview, Journal of Business Analytics (2022).
[24] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data Science Journal 14 (2015) 2.
[25] A. Haug, F. Zachariassen, D. Liempd, The costs of poor data quality, Journal of Industrial Engineering and Management 4 (2011) 168–193.
[26] B. T. Hazen, C. A. Boone, J. D. Ezell, L. A. Jones-Farmer, Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications, International Journal of Production Economics 154 (2014) 72–80.
[27] A. G. Carretero, F. Gualo, I. Caballero, M. Piattini, Mamd 2.0: Environment for data quality processes implantation based on iso 8000-6x and iso/iec 33000, Computer Standards & Interfaces 54 (2017) 139–151.
[28] C. Cichy, S. Rass, An overview of data quality frameworks, Ieee Access 7 (2019) 24634–24648.
[29] R. Y. Wang, A product perspective on total data quality management, Communications of the ACM 41 (1998) 58–65.
[30] G. Shankaranarayanan, Towards implementing total data quality management in a data warehouse, Journal of Information Technology Management 16 (2005) 21–30.