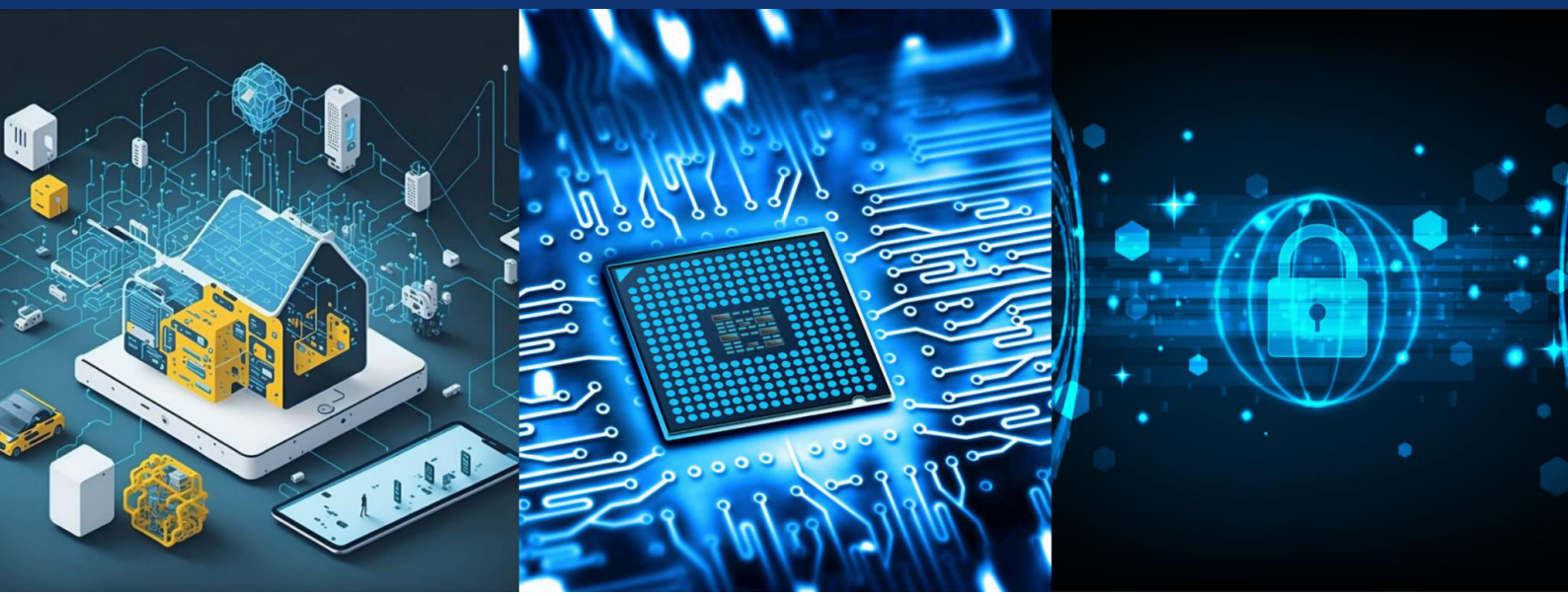
IJIRCCE

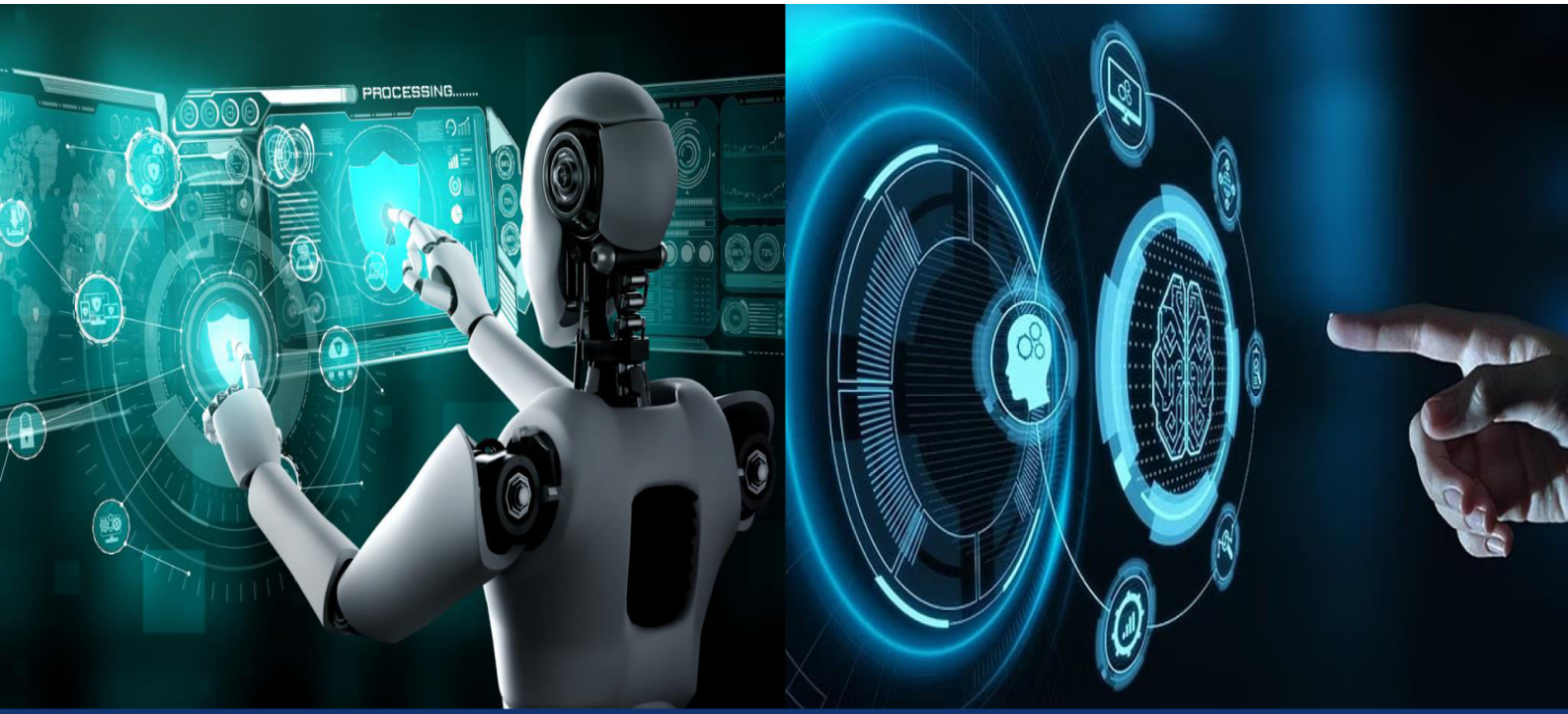
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 8, August 2026

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408013

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Autonomous Data Quality Assessment Frameworks: A Review of Methods, Tools, and Open Research Challenges

Nikhil R[1], Yogeshwaran K[2], Prof. Chandan Hegde[3]

Student, Department of Master of Computer Applications, Surana College (Autonomous), Bengaluru, India[1]

Student, Department of Master of Computer Applications, Surana College (Autonomous), Bengaluru, India[2]

Assistant Professor, Department of Master of Computer Applications, Surana College (Autonomous), Bengaluru, India[1]

ABSTRACT: With the increasing reliance on data for decision making, analytics and machine learning to generate actionable insights from data, the quality of the underlying data has become a matter of great importance. Evaluation of data quality has so far been a subjective and ad-hoc process, with few objective and systematic approaches to assessing it. In this article, we review and evaluate the cutting-edge methods and tools used for autonomous and score-based data quality assessment. Specifically, we discuss approaches that allow to combine objective assessments of individual data quality dimensions into a composite score and focus on rule-based data profiling and validation, modern data observability and anomaly detection methods. Following an overview of the theoretical foundations associated with each individual data quality dimension, we analyze and evaluate the five leading commercial products, covering both rule-based data validation/cleaning and data observability/anomaly detection methods in terms of their abilities in data profiling, data validation, data cleaning, data scoring, data workflow automation and transparency. We find that while all the five products reviewed have strengths in specific areas of the data pipeline, none of them provide truly end-to-end, autonomous and transparent quality assessment of data that would produce a composite score. We also find that while all commercial products reviewed have rule sets enabling them to perform data validation and data cleaning, these rules typically require labor-intensive configuration by data quality experts. We conclude that commercial data quality products identified lack several key features that would render them fit for purpose in terms of enabling high data quality. The most notable shortcomings are the absence of standardized and explainable approaches to composite data quality scoring and limited support for high-velocity and real-time data and domain-specific weighting of data quality dimensions. A unifying framework allowing to combine different data quality dimensions into a composite, automated and explainable score would greatly benefit the practice of data quality management.

KEYWORDS: Data Quality Assessment; Data Quality Dimensions; Automated Data Profiling; Data Validation and Cleansing; Score-Based Quality Framework; Data Observability; Data Governance; Machine Learning Data Readiness

I. INTRODUCTION

The value that analytics, business intelligence and machine learning can bring to the information-laden world of today is limited by the data that it uses. In a world where businesses are generating unprecedented volumes of various data from different sources, and where businesses can process this data in volumes previously unimaginable, the importance of the quality of this data has grown exponentially. A multitude of different types of information deficiencies, including imprecision, incompleteness, inconsistency, invalidity, redundancy and untimeliness severely limit the analytical value of this data to the business, resulting in analytical models that are built upon erroneous relationships and subsequently making businesses make bad decisions that they would not have made otherwise.

For large and complex data sets, which are the input for analytics, business intelligence and machine learning processes, manual data quality assessment is infeasible. At a larger scale the manual assessment of data quality is an expansion of the same tasks that a domain expert would perform when looking at smaller data sets, trying to find obvious inconsistencies, which would be too time-consuming and unproductive for large data sets, and therefore should be automated.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A score-based data quality framework allows for automating the assessment of all of the quality aspects mentioned above, by providing a score for every single one of them, from 1 to 100. By compiling individual scores into a general quality score for the data set, such a framework allows for automated checks that the data conforms to quality standards, allows prioritization of quality improvement tasks on the data set, and, most importantly, allows for comparison of multiple data sets to determine which one is of higher quality, and thus more suitable for analysis, business intelligence or machine learning processes.

This review will begin by detailing the different aspects of data quality, also known as data quality dimensions. A plethora of different data quality aspects have been described in the literature, most of which aim to provide a framework for manual data assessment, based on human understanding of what constitutes high quality data. The review will detail the different manual and automatic data quality assessment methods that have been proposed in the literature. Besides different frameworks that have been proposed, a variety of data quality tools have been developed, which are used to ensure data quality in businesses or for data validation and data cleaning purposes. An overview of different data quality tools and rule-based data validation and data cleaning methods is provided. Recent advances in artificial intelligence and machine learning fields have enabled the development of new data observability methods, which are used in data observability start-ups. In addition to providing a basis for reviewing existing data quality aspects and data quality assessment, the review will also outline the components of a pipeline for automatic data quality assessment, which includes data ingesting, assessing and monitoring, and feedback to the analyst.

The Review consists of five sections. The Review Methodology section describes the research process that formed the basis of this Review. The Literature Review section provides an overview of the data quality dimensions and data quality assessment methods that have been described in the literature. The Comparative Analysis of data quality tools section evaluates different data quality tools that have been developed to assess and monitor data quality within the score-based data quality framework. This section describes the different components of the automated data quality assessment pipelines. These components usually start with ingesting data and assessing its quality, followed by monitoring of the quality and providing feedback to the analyst. The section Research Gaps and Challenges describes the open questions in the field of data quality assessment and monitoring. Finally, the Conclusion section summarizes the findings of this study and suggests the directions for future research.

II. LITERATURE REVIEW

Foundational Data Quality Dimensions: An early work on data quality that laid the foundation for much of the data quality framework thinking was the definition of five dimensions, namely, Accuracy, Completeness, Consistency, Timeliness, and Relevance, for characterizing data quality. Furthermore, the work also highlighted the concept of data quality is fitness for use', which implies that there is no objective criteria, but rather data quality has to be judged based on its ability to meet the needs of the data users. As a consequence, one cannot determine data quality for a particular use by simply relying on objective measurements such as scores derived from rule-based assessments/statistical tests, but rather there is a need for frameworks that evaluate data quality according to different applications/tasks .

Manual Assessment versus Automated Assessment . As already outlined above, data quality manual assessment relies upon the so-called data quality dimensions. Nevertheless, for a considerable time also methods for the operational measurement of the data quality dimensions have been developed. The aforementioned illustration of inspection by humans for assessing data quality is in many occasions not sufficient since for very large databases such an approach is not scalable, for dynamic databases such an approach is not appropriate and, last but not least, such an approach is affected by possible human judgment bias. Therefore, in the recent years there has been a strong emphasis on so-called automated data quality assessment. The term automated denotes either fully- or semi-automated quantitative assessment of data sets by algorithms, which allows for truly continuous and objective control over data quality. As a result of the shift from the very conceptual definition stage in the early years of data quality management to the more practical-focused data quality management in the recent years, there is now a number of frameworks and methods for automated data quality assessment that have been developed, often incorporated in so-called automated assessment frameworks.[1]

Rule-Based Validation and Cleaning. Most of the current practices in data quality assessment and improvement rely on rule-based validation and rule-based data cleaning which include identification and correction of (duplicated) missing values and inconsistencies in data values according to prescribed rules. Such practices are very operationized and

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

automatized in numerous commercial data quality assessment tools such as, for example, IBM InfoSphere QualityStage and Informatica Data Quality. Rule-based approaches for data quality assessment, however, have a number of shortcomings. First, such approaches are often not general enough and have to be tailored to a particular database which is subject to data quality assessment. Here, by database I mean either a single database or databases that are used in conjunction for a particular analysis or decision in support of an enterprise. In other words, when talking about a database for data quality assessment we can also mean the whole analysis/scoping database or decision scope. Second, the rules for validation and cleaning have to be tailored to particular data quality problems that have to be addressed. Third, the maintenance of manually-written rules with corresponding thresholds across a very large number of data quality rules can be challenging.[2]

AI/ML-Driven Observability and Anomaly Detection. Recently, data quality management frameworks have been increasingly implemented as machine learning that either observes the data and/or detects anomalies in the data. Such data quality management tools are very effective in detecting of sporadic incidents in the data such as spikes in the data volume, significant shifts in data distribution or presence of very low quality records. Most of such tools require limited to no rule configuration since the rules are learned automatically by the system. Typically, such data quality management frameworks monitor all pipelines for incidents on a per-pipeline basis while providing a very rich set of information on the detected incident. Such tools, however, do not provide any comprehensive data profiling statistics at a data set level.[3]

Score-Based and AI-Assisted Platforms. The aforementioned tools mostly focus on enabling fully automatized end-to-end data quality workflow for previously unmanaged data in very large Enterprise settings. On the other hand there are also platforms that provide automatic calculation of a score for the single dimensions of the so far discussed data quality assessment framework based on Profiling, Validation and data cleaning in large as well as small data sets. Tools that provide such automated calculation can be found, for example, in Ataccama ONE among others. Again, the same limitation applies here since most of the tools for fully automatized data quality scoring do not address common data quality issues in small to medium size businesses but rather are in their infancy and are mostly focused on Enterprise customers. Furthermore, such tools are not fully transparent since there is no indication how the individual sub-scores for a particular data set have been combined into the final score in order to retrace the score in case of necessity.[4]

Standardization for Data Quality. While most of the practice-oriented work on data quality management focuses on the implementation of individual data quality tools, there is also research on standardization for total data quality management (TDQM) as well as data quality assessment. An example for the former can be seen in the IMF Data Quality Assessment Framework developed by the International Monetary Fund. Similar initiatives have taken place with regard to the definition of data quality standards, namely, the International Organization for Standardization (ISO) has issued a number of international standards for data quality management under the ISO 8000 portfolio.[5]

These standards, however, define the generic dimensions of data quality management as well as the processes for data quality management rather than specific methods for data quality assessment. A reference architecture for an automated data quality management follows the source data through a range of processing components that include ingestion, validation (validation engine), data profiling (profiling engine), data cleaning and transformation (cleaning and transformation module), quality monitoring, storage of different views of the data in versioned storage, and use of the data for training of artificial intelligence models with subsequent performance monitoring and feedback. This architecture reflects the fact that data quality assessment of large data sets has become a continuous process that is part of the life-cycle of machine learning model development.

To conclude, there is a number of different research lines on Data Quality Assessment that started with the very foundational definition of the dimensions of Data Quality. Most of the subsequent work built upon these early definitions, resulting in either rule-based or semi-automated manual validation practices. Nevertheless, the recent research on Data Quality Assessment builds upon the findings from the earlier studies but rather focuses on AI-assisted Data Quality Assessment practices. In other words, there are now a number of automated practices for scoring in multiple Data Quality Assessment frameworks. Still, there is a need for a fully transparent, standardized and generally applicable scoring framework(s) for Data Quality Assessment, which this research aims to contribute to.

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408013

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. REVIEW METHODLOGY

The current literature review is guided by the narrative-scoping method that allows for providing an in-depth analysis of the literature and industry data on autonomous data quality assessment. Databases, including Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect, and Scopus, were searched using the combination of keywords "data quality dimensions," "data quality assessment framework," "automated data profiling," "data quality scoring," "data observability," and "data validation and cleansing" with the use of Boolean operators. The sources of information were narrowed down to the documents published in the past decade to ensure the relevance of the findings to the modern approaches and tools used in autonomous and AI-enabled data quality assessment with the exception of conceptual papers on data quality dimensions, which were included regardless of the date of publishing due to their significance to the research problem.

The inclusion criteria were based on papers and industry resources that featured any of the measurable data quality dimensions, described an automated or semi-automated process of data quality assessment, or provided an insight into the features and capabilities of existing data quality tools. Initially, the retrieved sources were screened based on titles and abstracts or summaries in the case of industry resources, with the aim of eliminating the articles that had nothing to do with the research problem. Subsequently, the second screening was conducted, during which the full text of the short-listed papers and full versions of the documents provided by the industry sources were examined to extract the information needed for the comparative analysis and subsequent discussion.

IV. COMPARATIVE ANALYSIS

Table 1 compares five widely referenced data quality tools — Ataccama ONE, Monte Carlo, Anomalo, IBM InfoSphere QualityStage, and Informatica Data Quality (IDMC) — across five capability parameters central to an autonomous data quality assessment framework: data profiling, data validation, data cleansing, data quality scoring, and workflow automation. The comparison done highlights a clear split in the abilities of the tool especially in the context of AI-assisted, observability-focused tools (Monte Carlo, Anomalo), which emphasize pipeline monitoring and anomaly detection but offer little or no profiling, cleansing, or scoring functionality, and more traditional enterprise data quality platforms (IBM InfoSphere QualityStage, Informatica IDMC), which provide fuller profiling, validation, and cleansing capabilities but only partial or limited quality-scoring output. Ataccama ONE stands out as the only tool in this comparison that combines AI-driven profiling, validation, automated cleansing, and comprehensive multi-dimensional quality scoring within a single automated workflow, making it the closest existing approximation of a complete autonomous, score-based assessment framework among the tools reviewed.

Table 1. Comparison of Data Quality Tools

Parameter	Ataccama ONE	Monte Carlo	Anomalo	IBM InfoSphere QualityStage	Informatica Data Quality (IDMC)
Data Profiling	Provides AI-driven profiling to analyze data structure, quality, and statistics automatically.	Focuses on monitoring data pipelines rather than profiling datasets.	Primarily detects anomalies instead of profiling data.	Profiles enterprise datasets to identify quality issues before processing.	Profiles datasets to evaluate completeness, consistency, and accuracy.
Data Validation	Validates datasets using AI-assisted business rules and quality constraints.	Monitors pipeline health instead of validating individual records.	Detects validation issues through anomaly detection techniques.	Validates datasets against predefined business rules and standards.	Performs rule-based validation to ensure data quality and compliance.

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408013

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Parameter	Ataccama ONE	Monte Carlo	Anomalo	IBM InfoSphere QualityStage	Informatica Data Quality (IDMC)
Data Cleansing	Automatically identifies and corrects errors, inconsistencies, and missing values.	Does not perform data cleansing operations.	Identifies quality issues but does not clean datasets.	Cleanses, standardizes, and removes inconsistencies from enterprise data.	Cleanses and enriches data using predefined quality rules.
Data Quality Scoring	Generates comprehensive quality scores across multiple quality dimensions automatically.	Does not calculate data quality scores.	Does not provide quality scoring functionality.	Provides limited quality assessment metrics.	Calculates data quality scores and generates quality reports.
Workflow Automation	Automates profiling, validation, cleansing, scoring, and monitoring in a unified workflow.	Automates data pipeline monitoring and incident management.	Automates anomaly detection and alert generation.	Supports limited automation for quality processes.	Automates enterprise data quality workflows and governance processes.

V. RESEARCH GAPS AND CHALLENGES

Despite the identified strengths in data quality technology, the reviewed literature and comparative analysis demonstrate the absence of critical features that would enable the establishment of an objective data quality scoring system.

The following factors contribute to the limitations in implementing data quality assessment:
1. The lack of standardized scoring algorithms for combining different dimensions into a composite score.
2. The limited transparency and explainability of the calculation process even in commercial tools claiming to calculate data quality scores.
3. The absence of a tool which would provide both observability and profile plus cleansing while scoring data quality.
4. The absence of standardized, generalizable methods for automated rule creation for data quality validation and cleansing.
5. The limited focus on real-time or streaming data quality assessment.
6. The lack of domain-specific and task-specific considerations when calculating data quality scores.
7. The limited availability of benchmark data sets for training and testing data quality scoring algorithms.
8. The limited accessibility of data quality assessment tools for researchers and open-source machine learning communities due to their high cost.
9. The absence of a closed-loop system linking data quality scores to model training and monitoring.

The reviewed literature and comparative analysis indicate that while individual elements for an autonomous data quality assessment framework have been developed, a standardized, open, and accessible scoring system remains to be created. This system would holistically score data quality and enable automated rule creation for validating and cleansing data sets. The critical areas for further research include establishing standardized dimensions for calculating data quality scores, developing automated methods for rule creation that would generalize across application domains, and linking data quality scores to model performance in closed-loop systems.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. CONCLUSION

In this paper, we conduct a survey on current approaches to autonomous score-based data quality assessment. After a foundation discussion on the dimensionality of data quality, we review in detail three categories of approaches, namely, rule-based, AI-assisted, and most recently, observability-based approaches. Our assessment of various existing data quality tools reveals that there are many high-quality standalone tools supporting various aspects of data quality assessment such as anomaly detection and monitoring, rule-based data validation and data cleansing, and enterprise data quality, etc. However, there are very few tools that can support fully-automated score-based data quality assessment in a holistic manner, covering data profiling, data validation, data cleansing, and scoring.

This review has identified significant gaps in current tools: a) lack of transparent and standardized scoring; b) assessment of streaming data; c) quality dimension values weighing, which is highly domain-specific; d) simple, transparent, and reproducible frameworks which can be used for data science research purposes as well as in small-scale environments. Following work in this area should therefore aim at developing a score-based framework which autonomously assesses the data at hand on the basis of a normalized score. Such a framework should be fully integrated into the machine learning lifecycle, so that the quality score can be used not only for data cleaning, but also for selecting the best datasets for a given problem as well as for training models.

The assessment of data quality of datasets by using autonomous tools for score-based quality assessment is a matter which is becoming mature. The current state of tools for data quality assessment allows for an autonomous assessment of data, in particular for the huge amount of incoming data. There is still room for improvement, especially for a more standardised and transparent scoring method. This review confirms that, currently, most aspects of data quality can be assessed autonomously on a score, even outside of the data-science team. The higher the number of aspects that can be assessed completely and autonomously the higher the value for the data-operations as well as for the machine learning process. Thus, also for future projects autonomous data quality assessment with explainable scores will become even more important, as the data operations are scaling up and more and more models are being set up for decision making. Hence, autonomous data quality assessment is also a very valuable research area as well as application area.

REFERENCES

[1]. Al-Toq, Razan, and Abdulaziz Almaslukh. "DQMAF—Data Quality Modeling and Assessment Framework." *Information* 16.10 (2025): 911.
[2]. Miller, Russell, et al. "An Overview of Current and New Data Quality Dimensions under a Common Framework." (2024).
[3]. Gong, Youdi, et al. "A survey on dataset quality in machine learning." *Information and Software Technology* 162 (2023): 107268.
[4]. Alexakis, Theodoros, et al. "Evaluating Data Quality: Comparative Insights on Standards, Methodologies, and Modern Software Tools." *Electronics* 14.15 (2025): 3038.
[5]. Ramisetty, Gopinath. "Real-Time Data Integrity Nexus: Autonomous Quality Assurance Framework." *Journal of Computer Science and Technology Studies* 7.10 (2025): 665-671.
[6]. Podobnikar, Tomaž. "Bridging Perceived and Actual Data Quality: Automating the Framework for Governance Reliability." *Geosciences* 15.4 (2025): 117.
[7]. Bernardi, Filipe Andrade, et al. "Data quality in health research: integrative literature review." Journal of medical Internet research 25 (2023): e41446.
[8]. Liu, Mingyu, et al. "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook." IEEE Transactions on Intelligent Vehicles (2024).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details