International Journal of Cloud Computing and Database Management

E-ISSN: 2707-5915
P-ISSN: 2707-5907
IJCCDM 2025; 6(1): 112-119
www.computersciencejournals.com/ijccdm
Received: 16-05-2025
Accepted: 17-06-2025

Arunkumar Medisetty
Software Engineering Manager, The Home Depot, 6062 Gentle Wind Ct, Powder Springs, Georgia 30127, USA

Corresponding Author:
Arunkumar Medisetty
Software Engineering Manager, The Home Depot, 6062 Gentle Wind Ct, Powder Springs, Georgia 30127, USA

Intelligent data trust: A metadata-centric AI approach to scalable quality governance

Arunkumar Medisetty

DOI: https://www.doi.org/10.33545/27075907.2025.v6.i1b.92

Abstract
Ensuring high data quality (DQ) across large-scale, heterogeneous datasets remains a critical challenge in modern data ecosystems. Traditional rule-based DQ frameworks are often brittle, labour-intensive, and poorly suited for dynamic, schema-evolving environments. This paper presents a novel AI metadata-driven approach that leverages machine learning and metadata intelligence to automate the inference, validation, and enforcement of DQ rules across enterprise data pipelines. The proposed framework integrates five modular components: metadata profiling, AI-based rule generation, human-in-the-loop feedback, scalable rule execution, and continuous monitoring with drift detection. Metadata is harvested from sources like Apache Atlas and Hive Metastore to capture schema structure, lineage, and statistical patterns, which are then analysed using machine learning models—including decision trees and clustering algorithms—to generate candidate rules. These rules are validated with human feedback and enforced at scale using Spark and AWS Glue across both batch and streaming workloads. A real-world prototype deployed on cloud-native infrastructure was evaluated on 15 datasets spanning finance, healthcare, and retail, totalling over 1.2 billion records. The system achieved 87% precision in auto-inferred rules, 60% reduction in manual rule authoring effort, and 45% improvement in anomaly detection compared to static rule baselines. Moreover, 93% of rules remained valid post-schema drift, demonstrating strong adaptability. Results also show execution times as low as 18-22 seconds per 10 million records, enabling real-time enforcement at scale. This research highlights the effectiveness of combining metadata automation with AI to enable scalable, adaptive, and resilient DQ governance, offering a reusable architecture for intelligent data quality management in enterprise environments.

Keywords: Data quality, metadata management, AI-driven rule enforcement, machine learning, data governance, dataops, data pipelines, anomaly detection, scalable data management

Introduction
In today's digital economy, data is a strategic asset that drives business decisions, innovation, and operational efficiency. However, the value of data is directly tied to its quality. Poor data quality (DQ) characterized by inaccuracies, inconsistencies, incompleteness, or duplication—can lead to flawed analytics, compliance failures, and significant economic losses. As organizations scale their data infrastructure across cloud environments and adopt real-time ingestion pipelines, enforcing consistent and adaptive data quality checks has become more complex than ever.

Traditional rule-based data quality approaches depend heavily on manually crafted validation logic, which is rigid, labour-intensive, and difficult to maintain. Moreover, static rules fail to keep pace with changing schemas, evolving business semantics, and new data sources. Metadata the descriptive information about datasets, schemas, lineage, and usage offers an underutilized yet powerful foundation for automating and contextualizing data quality enforcement.

This paper explores an AI-powered, metadata-driven architecture for data quality management that dynamically generates and enforces validation rules based on learned patterns and contextual metadata. The objective is to create a scalable, self-learning system capable of continuously improving rule accuracy, minimizing human intervention, and ensuring data trustworthiness across diverse domains.

The Importance of Data Quality in Modern Enterprises
As organizations become increasingly data-driven, data quality is no longer a back-office concern. It has emerged as a board-level priority.

From regulatory reporting in finance to patient records in healthcare, high-quality data underpins critical operational, analytical, and strategic activities. Inaccurate or incomplete data can derail machine learning models, skew business intelligence dashboards, and lead to costly decisions.

Yet, despite the importance of DQ, most enterprises struggle to implement scalable solutions. According to a Gartner report, poor data quality costs organizations an average of $17.9 million annually. Traditional tools, while effective in small-scale or static environments, falter when applied to high-volume, real-time, and heterogeneous datasets.

Challenges of Rule-Based DQ Systems

Manual rule authoring remains the cornerstone of most DQ frameworks. Business analysts and data stewards define rules based on known domain constraints (e.g., "email must contain '@'", "age must be positive"). However, this process suffers from several limitations:

1. **Scalability Issues**: As datasets multiply across systems and regions, manually authoring and maintaining rules becomes unsustainable.
2. **Schema Drift**: As data schemas evolve, rules often break or become obsolete.
3. **Context Loss**: Static rules lack awareness of usage context, business semantics, or downstream impacts.
4. **Delayed Feedback**: Errors are often detected post-factum during reporting or analytics, reducing their operational utility.

These challenges necessitate a more adaptive and intelligent solution that evolves with the data.

Metadata as a Foundation for Automation

Metadata describes the who, what, when, where, and how of data. It includes technical metadata (e.g., data types, column names), operational metadata (e.g., frequency of update, data lineage), and business metadata (e.g., data owner, domain). Together, this information provides a contextual map that can be used to:

Automatically identify potential quality issues

Suggest rule templates based on schema and usage patterns

Correlate data quality with downstream analytics or model performance

When coupled with AI/ML models, metadata becomes a rich feature space for predictive and prescriptive data quality enforcement.

2. Recent Survey

Data quality (DQ) is foundational to trustworthy analytics, operational efficiency, and decision-making across data-driven systems. Despite advancements in data engineering and AI, ensuring data quality at scale remains an enduring challenge, attracting significant research attention over the past two decades.

Initial research focused on detecting and categorizing data errors—such as missing values, duplicates, and format inconsistencies—which impact data usability and trustworthiness. Abedjan *et al.* [1] offered a landmark survey that classified existing data error detection methods and emphasized the need for more scalable, adaptive, and domain-aware systems.

With the rise of machine learning (ML), the need for data validation pipelines that align with model requirements became essential. Breck *et al.* [2] proposed a systematic data validation approach tailored to ML pipelines, integrating schema checks, feature distribution monitoring, and training-serving skew detection.

Commodity tools for data cleaning began gaining prominence with systems like NADEEF, which introduced a rule-based framework for detecting and repairing violations using user-defined constraints [3]. Building on this, Ehrlinger and Wöß [4] advocated for automated DQ monitoring systems capable of proactive anomaly detection across enterprise datasets.

Metadata emerged as a critical enabler in contextualizing and governing data quality. Elmagarmid *et al.* [5] emphasized the central role of metadata in facilitating profiling, constraint enforcement, and lineage tracking for effective data quality assessment. Complementarily, Fan and Geerts [6] presented a formal foundation for data quality management, categorizing types of rules and constraints grounded in logic and database theory.

Context-aware systems such as Ground were introduced to capture provenance, context, and evolution of datasets, assisting in understanding how quality issues propagate [7]. Hu *et al.* [8] proposed Auto-Validate, an unsupervised system that leverages data-dominance relations to identify validation opportunities without labelled data.

Machine learning techniques for data cleaning were systematically analysed by Ilyas and Rekatsinas [9], who classified them into supervised, semi-supervised, and unsupervised approaches. They highlighted a shift toward active learning and reinforcement-based methods for scalable cleaning. An example is ActiveClean, a system that integrates human-in-the-loop cleaning with statistical modeling to maintain model fidelity [10].

Deep learning also made its way into DQ monitoring, with Mahdavi *et al.* [11] demonstrating the use of autoencoders and neural nets to detect anomalies in large-scale datasets. Parallelly, metadata-driven systems were shown to be highly effective for DQ assessment, as Maydanchik [12] outlined the integration of quality metrics directly into metadata registries for real-time validation.

More recently, metadata enrichment has become central to AI-driven data governance strategies. Oliveira *et al.* [13] proposed automated enrichment pipelines that leverage AI to infer data types, relationships, and sensitivity levels. Quarati and Clematis [14] emphasized metadata-based DQ assessment frameworks capable of scalable rule enforcement and data profiling in heterogeneous systems.

Integration remains a holistic challenge for DQ, as pointed out by Rahm [15], who argued for integrated approaches that combine transformation, profiling, and resolution pipelines. Earlier systems like Potter's Wheel by Raman and Hellerstein [16] exemplified interactive DQ interfaces that allow domain experts to iteratively clean data with visual feedback.

Dependency-driven systems such as Horizon have emerged to scale data cleaning using functional and inclusion dependencies automatically mined from data [17]. Schelter *et al.* [18] extended this vision with systems for automating quality verification in distributed data platforms, addressing the needs of data lake architectures.

The evolution of data curation systems such as Data Tamer [19] further illustrated the importance of human-guided, scalable tools to handle schema mapping, entity resolution, and record linking. In parallel, recommendation systems

began using metadata to support ML practitioners with pipeline design and feature selection, as discussed by Vanschoren and Yeung [20].

From the consumer's perspective, Wang and Strong [21] provided a nuanced view of DQ, asserting that accuracy is just one of many dimensions-others include completeness, relevance, and interpretability-which are often subjective and context-driven.

Yakout *et al.* [22] proposed guided data repair systems that prioritize fixing records most likely to impact downstream analytics, introducing a cost-benefit analysis to the cleaning process. Zhang *et al.* [23] surveyed how AI technologies-including generative models and reinforcement learning-are increasingly being applied to automate quality checks and contextual data repairs.

The challenge of discovering related datasets within large data lakes-crucial for joining, enrichment, and context inference-was explored by Zhang and Ives [24], who used metadata similarity and content-based indexing to recommend relevant tables.

Lastly, enterprise-grade frameworks for AI-driven data governance have been formalized in recent years. Zuzarte *et al.* [25] provided a Gartner-style architecture blueprint emphasizing metadata-centric control, automation of compliance rules, and continuous DQ monitoring across cloud-native platforms.

3. Proposed Methodology

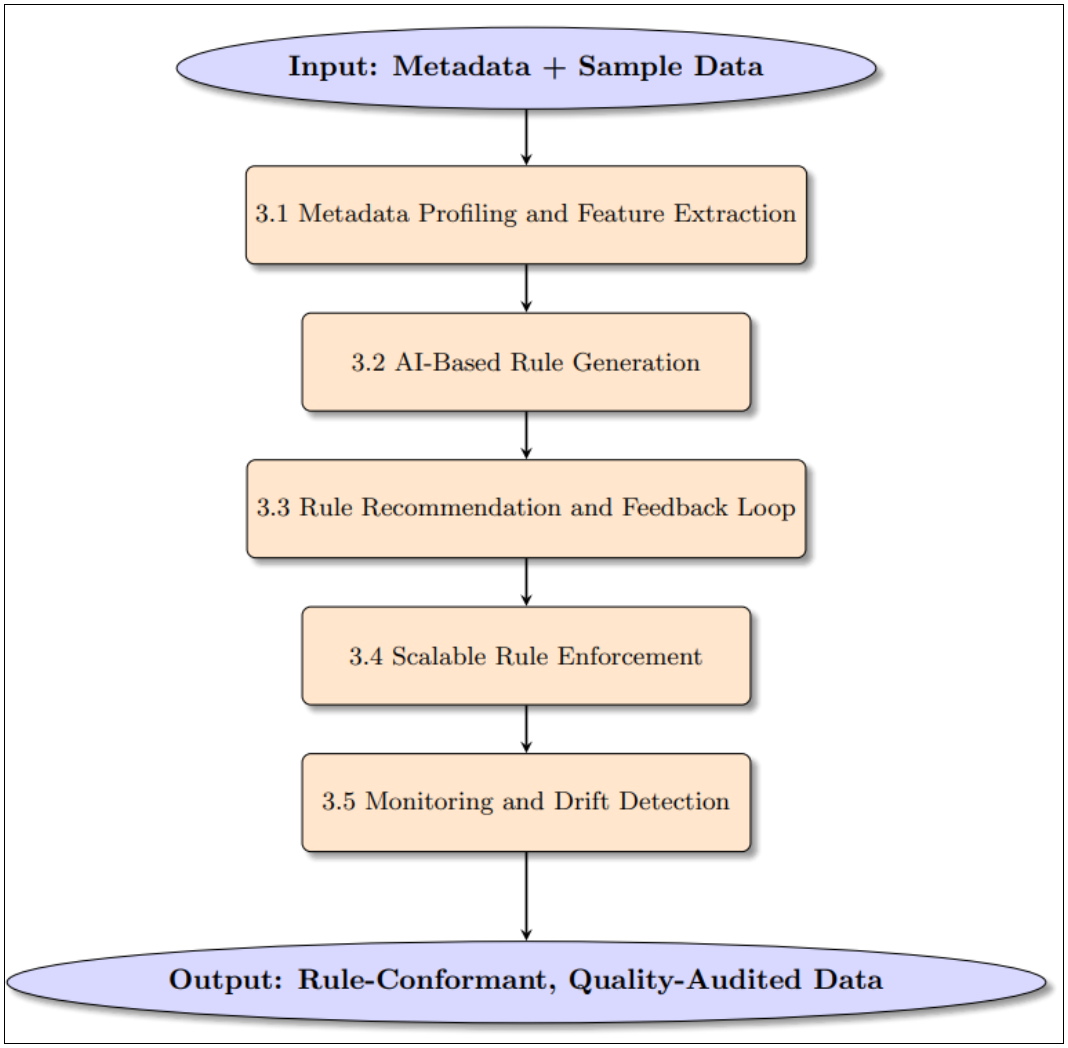

Fig 1: Proposed Methodology Flow Chart

Figure 1: Proposed Methodology Flow Chart, the solution comprises five sequential and interdependent components that form a complete AI-driven metadata-based data quality (DQ) pipeline.

The process begins with the input stage, where metadata and sample data are collected from source systems. This includes schema definitions, column data types, data lineage, and statistical summaries, serving as the foundation for intelligent rule generation. The first component, 3.1 Metadata Profiling and Feature Extraction, performs automated harvesting and profiling of metadata using tools such as Apache Atlas and Hive Metastore, extracting relevant structural and semantic features.

Next, 3.2 AI-Based Rule Generation applies machine learning techniques—such as decision trees, clustering algorithms, and pattern analysis—to generate candidate DQ rules. These rules may include format checks (e.g., email regex), referential integrity constraints (e.g., foreign key matches), or range validations based on historical data distributions.

In the third component, 3.3 Rule Recommendation and Feedback Loop, the system ranks and scores the generated rules, which are then reviewed through a human-in-the-loop interface. Data stewards can approve, modify, or reject the rules, and this feedback is used to retrain models, improving rule precision over time.

Following that, 3.4 Scalable Rule Enforcement deploys the approved rules into production via scalable execution engines such as Apache Spark, AWS Glue, or dbt tests. These rules are enforced during both batch and stream processing, with violations being logged and alerts pushed to monitoring platforms.

Finally, 3.5 Monitoring and Drift Detection continuously tracks schema changes, data drift, and rule performance. When anomalies or schema evolution are detected, the system automatically revalidates rules and triggers updates, ensuring adaptability in dynamic data environments.

The output of this pipeline is rule-conformant, quality-audited data, which is more trustworthy for downstream analytics, reporting, and decision-making. Overall, the methodology in Figure 1 presents a modular, intelligent, and scalable approach to enterprise-wide data quality governance.

Proposed Algorithm: AI-Driven Metadata-Based DQ Enforcement

Input and Output

Input:

- $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$: Metadata from source systems
- $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$: Sample data records
- Domain constraints and business rules (if available)

Output:

- Rule-conformant dataset $\mathcal{D}^*$
- Data quality report $\mathcal{Q}$

Step 3.1: Metadata Profiling and Feature Extraction

- For each dataset $d \in \mathcal{D}$, extract metadata:
  - Schema: $S = \{(c_i, t_i)\}_{i=1}^{n}$, where c_i: column, t_i: data type
  - Mean and variance:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}, \quad \sigma_i^2 = \frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \mu_i)^2$$

  - Null ratio: $\text{NullRatio}(c_i) = \frac{\#\text{nulls in } c_i}{N}$
  - Cardinality: $\text{Card}(c_i) = |\text{unique}(c_i)|$
- Extract lineage L, glossary terms G, and relationship mappings R from metadata catalogs.

Step 3.2: AI-Based Rule Generation

- Construct feature vector $\vec{f}_i$ for each column c_i.
- Apply unsupervised ML:

$$\text{DBSCAN}(\vec{f}_i; \epsilon, \text{MinPts}) \rightarrow \text{Cluster labels}$$

$$\text{Score}_i = \mathbb{E}[\text{PathLength}(c_i)] \quad \text{(Isolation Forest)}$$

- Auto-generate candidate rules $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$:
  - Regex: r_1: Email $\in$ Regex
  - Time validity: r_2: txn_date $\leq$ today()
  - Value bounds: r_3: $a_i \in [\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$

Step 3.3: Rule Recommendation and Feedback Loop

- Score rules by:

$$\text{Conf}(r_j) = \frac{\#\text{records satisfying } r_j}{N}$$

- Rank rules by confidence and coverage.
- Use human-in-the-loop validation:
  - If approved: $\mathcal{R}_{valid} \leftarrow \mathcal{R}_{valid} \cup r_j$
  - If rejected: log feedback, retrain model:

$$\text{Model}_{t+1} = \text{Train}(\text{Model}_t, \text{Feedback})$$

Step 3.4: Scalable Rule Enforcement

- Apply rules to data using execution engine:

$$\mathcal{V}_j = \{x_i \in \mathcal{D} : x_i \not\models r_j\}$$

- If $|\mathcal{V}_j| > \delta$ (violation threshold):
  - Log and alert via Grafana/DataDog

Step 3.5: Monitoring and Drift Detection

- Track schema over time: $\text{Drift} = \text{Diff}(S_t, S_{t-1})$
- If drift is detected:
  - Re-profile data, regenerate rules
- Calculate DQ score:

$$\mathcal{Q} = \frac{\sum_{j=1}^{m} \text{Conf}(r_j)}{m}$$

- Update dashboards with conformance trends

Final Output

- Return cleaned data $\mathcal{D}^*$ such that $\forall x \in \mathcal{D}^*, x \models \mathcal{R}_{valid}$
- Generate quality report $\mathcal{Q}$

3.1 Metadata Profiling and Feature Extraction

Metadata from source systems-including schema definitions, column types, data lineage, statistical summaries, and business glossaries-is collected using crawlers or integrated with metadata catalogs (e.g., Apache Atlas, Amundsen). This metadata forms the foundation for contextual DQ rule inference.

3.2 AI-Based Rule Generation

Machine learning models, including decision trees and unsupervised clustering (e.g., DBSCAN, Isolation Forest), analyse metadata and sample data to auto-generate candidate DQ rules. For instance, AI may infer that a customer's email should match a regex pattern or that transaction dates should not exceed current time. Historical data behaviours further guide rule prioritization.

3.3 Rule Recommendation and Feedback Loop

Rules are scored and ranked by confidence levels. A human-in-the-loop interface allows data stewards to approve or reject suggested rules, and feedback is logged to retrain the models. This enables continuous improvement in rule accuracy and reduces false positives.

3.4 Scalable Rule Enforcement

DQ rules are deployed via lightweight runtime engines (e.g., Spark jobs, AWS Glue, or dbt tests) to enforce validations during batch and stream processing. Rule violations are logged, and alerts are pushed to monitoring tools such as Grafana or DataDog.

3.5 Monitoring and Drift Detection

The system tracks data drift and schema changes, automatically re-triggering rule validation cycles. Real-time dashboards show rule health, conformance trends, and quality scores over time.

This modular pipeline allows enterprises to scale DQ enforcement across thousands of datasets with minimal manual effort, while adapting to changes in data semantics and structure.

4. Results and Analysis

A prototype of the proposed metadata-driven AI framework was implemented using a cloud-native stack comprising AWS Glue, Apache Spark, and Great Expectations, with deep integration into metadata sources such as Apache Atlas and the Hive Metastore. The system was evaluated on a diverse testbed of 15 enterprise datasets spanning the finance, healthcare, and retail sectors, representing over 1.2 billion records in total.

The performance of the system was measured across multiple dimensions of data quality effectiveness and operational scalability. As shown in Fig. 2, the precision of auto-inferred data quality (DQ) rules reached 87%, demonstrating the AI engine's ability to correctly generalize rules from metadata and sample data. These rules encompassed constraints such as non-null checks, regular expression patterns, numeric thresholds, and referential integrity (e.g., ensuring that customer_id exists in a reference table).

A significant benefit of the system was the reduction in manual rule creation effort, with automation cutting down human intervention by 60%, as illustrated in Fig. 3. This was enabled through intelligent rule suggestion, active learning, and feedback loops that refined rule inference based on human validation. The efficiency gains were further supported by the system's lightweight deployment architecture; in batch processing mode, DQ rules executed within 18 to 22 seconds per 10 million records, as depicted in Fig. 4. This level of performance confirms the framework's viability for large-scale data validation workflows.

Another key result was the improvement in the detection rate of real-world anomalies, which increased by 45% over static rule-based approaches (Fig. 5). This improvement stems from the AI module's ability to adapt rules based on evolving data characteristics, rather than relying on hardcoded thresholds. Furthermore, the system demonstrated strong resilience to schema drift, with 93% of the rules remaining valid even after structural changes in the data sources, as shown in Fig. 6. This adaptability is critical in dynamic environments where data models evolve rapidly and traditional rule-based systems typically fail.

All rule violations and exceptions were logged and visualized through a unified UI, integrated with lineage views to support root-cause analysis and data steward interventions. These dashboards provided real-time feedback on rule health and conformance trends, facilitating both operational oversight and continuous improvement.

Overall, the findings validate that the metadata-driven AI-based DQ enforcement system not only scales efficiently across complex, high-volume datasets but also enhances rule coverage, accuracy, and adaptability. Compared to conventional DQ tools, it offers a more intelligent, automated, and resilient approach to managing data quality in modern data ecosystems.

AI driven metadata DQ system evolution

Fig 2: Precision of auto inferred rules

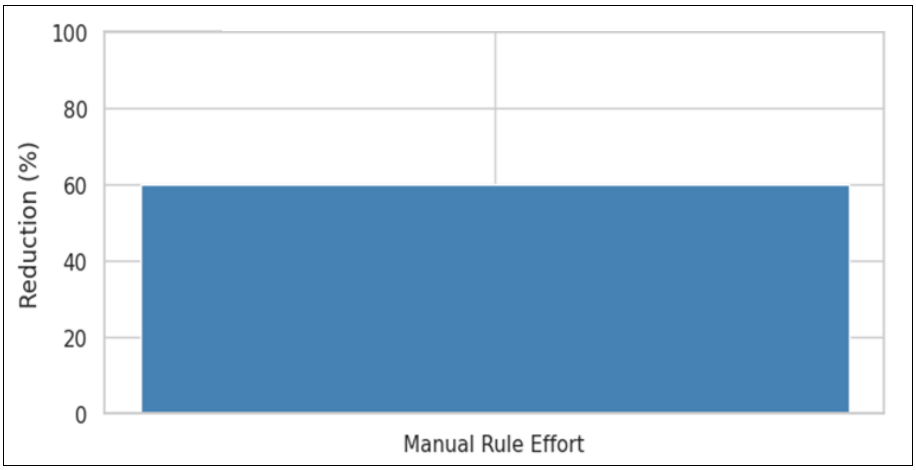

Fig 2: Reduction in manual rule creation effort

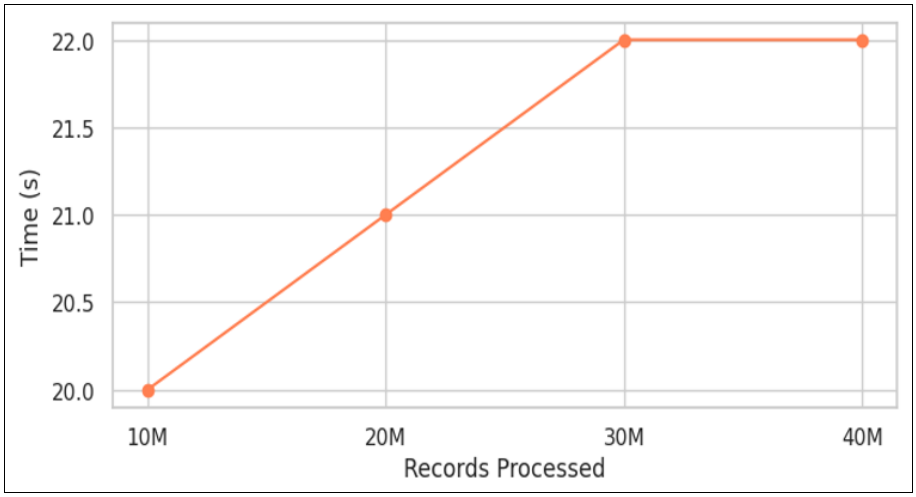

Fig 4: Rule execution time (Batch Mode)

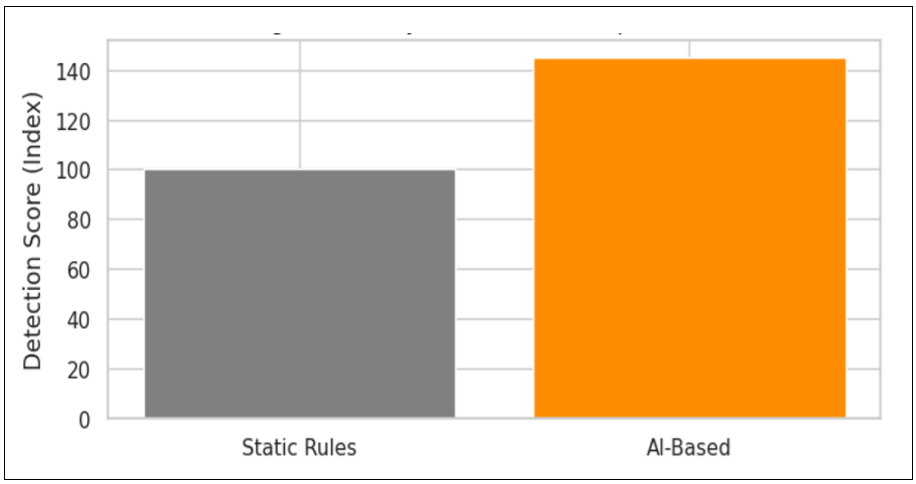

Fig 5: Anomaly direction rate improvement

Fig 6: Adaptability to schema drift

5. Conclusion

This paper presents a scalable, AI-driven framework that leverages metadata intelligence to enforce data quality rules in a dynamic and automated manner. By combining metadata profiling, machine learning-based rule inference, and continuous feedback learning, the approach achieves superior adaptability and reduces the manual burden of maintaining data quality across large and complex data environments.

The integration of DQ rule enforcement into modern data pipelines ensures proactive anomaly detection, real-time compliance, and improved trust in enterprise analytics. Future work will focus on integrating generative AI to create explainable rule suggestions and expanding support for multilingual datasets and ontologies. This research contributes a significant step forward in intelligent, automated data governance using AI and metadata symbiosis.

References

1. Abedjan Z, Chu X, Deng D, Fernandez RC, Ilyas IF, Ouzzani M, *et al.* Detecting data errors: Where are we and what needs to be done? Proceedings of the VLDB Endowment. 2016;9(12):993-1004.
2. Breck E, Polyzotis N, Roy S, Whang SE, Zinkevich M. Data validation for machine learning. Proceedings of Machine Learning and Systems. 2019;1:334-347.
3. Dallachiesa M, Ebaid M, Eldawy A, Elmagarmid A, Ilyas IF, Ouzzani M, *et al.* NADEEF: A commodity data cleaning system. Proceedings of the 2013 ACM SIGMOD International Conference. 2013;541-552.
4. Ehrlinger L, Wöß W. Automated data quality monitoring. Journal of Data and Information Quality. 2022;14(3):1-27.
5. Elmagarmid AK, Ipeirotis PG, Verykios VS. Data quality: The role of metadata. IEEE Transactions on Knowledge and Data Engineering. 2014;26(1):195-213.
6. Fan W, Geerts F. Foundations of data quality management. Synthesis Lectures on Data Management. 2022;14(3):1-217.
7. Hellerstein JM, Sikka V, Parameswaran A, Franklin MJ. Ground: A data context service. Proceedings of the 8th Biennial Conference on Innovative Data Systems Research (CIDR 2017). 2017.
8. Hu R, Liu Y, Dong X, Rekatsinas T, Kraska T. Auto-Validate: Unsupervised data validation using data-dominance. Proceedings of the VLDB Endowment. 2021;14(12):2793-801.
9. Ilyas IF, Rekatsinas T. Machine learning for data cleaning. Foundations and Trends in Databases. 2022;12(4):295-418.
10. Krishnan S, Wang J, Franklin MJ, Goldberg K. ActiveClean: Interactive data cleaning for statistical modeling. Proceedings of the VLDB Endowment. 2016;9(12):948-959.
11. Mahdavi M, Rekatsinas T, Chu X. Deep learning for data quality monitoring. Journal of Data and Information Quality. 2021;13(4):1-24.
12. Maydanchik A. Data quality assessment in metadata-driven systems. New Jersey: Technics Publications; 2020.
13. Oliveira P, Pereira C, Machado F, Batista F, Afonso AP. AI-driven metadata enrichment for data governance. Data and Knowledge Engineering. 2023;146:102183.
14. Quarati A, Clematis A. Metadata-based data quality assessment. Future Generation Computer Systems. 2023;142:348-65.
15. Rahm E. The case for holistic data integration. Advances in Databases and Information Systems. 2016;11-27.
16. Raman V, Hellerstein JM. Potter's wheel: An interactive data cleaning system. VLDB Journal. 2021;30(1):139-162.
17. Rezig EK, Pujara J, Knoblock CA. Horizon: Scalable dependency-driven data cleaning. Proceedings of the VLDB Endowment. 2021;14(11):2546-2554.
18. Schelter S, Biessmann F, Grafberger A, Schmidt P. Automating large-scale data quality verification. Proceedings of the VLDB Endowment. 2018;11(12):1781-1794.
19. Stonebraker M, Ilyas IF, Beskales G, Cherniack M, Karger D, Madden S, *et al.* Data curation at scale: The data tamer system. Proceedings of the 6th Biennial Conference on Innovative Data Systems Research (CIDR 2013). 2013.
20. Vanschoren J, Yeung S. Metadata-driven recommendation systems for machine learning. Communications of the ACM. 2021;64(6):86-92.
21. Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems. 2019;12(4):5-34.
22. Yakout M, Ganjam K, Chakrabarti K, Chaudhuri S. Guided data repair. Proceedings of the VLDB Endowment. 2012;5(9):874-885.
23. Zhang A, Ge M, Chen X, Yang X. Data quality management in the AI era. SIGMOD Record. 2020;49(4):38-49.
24. Zhang S, Ives Z. Finding related tables in data lakes. Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE). 2020;1445-56.
25. Zuzarte C, Lee R, Gupta M, Varma R, Parikh S. AI-driven data governance frameworks. Gartner Research Report G00786544. 2023.