IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Scalable AI Framework For Real-Time Data Quality Assessment In Event-Driven Pipelines

Rajeev Kumar Sharma

Independent Author
Western Governors University

Abstract: Checking the quality of data in real time within event-driven processes guarantees that analytics and decisions based on AI are reliable. Standard batch methods for managing quality do not work well with huge and fast data streams. The review outlines a framework that depends on AI, includes rule-based validation, looks for anomalies using machine learning and can adapt based on changing needs to satisfy latency, scalability and robustness issues. Research studies have shown that using a hybrid approach with continuous drift detection gives high accuracy of 97% and few false positives. In a federated setting, data privacy is kept alongside effective threat detection. Among the key lessons is that model retraining is tough, available resources can be limited and privacy is essential, especially as it's important to balance accuracy and how quickly tasks can be completed. Because the framework is designed in modules, adding new features is simpler: simple checks are done by rules, but AI steps in when regular rules fail. If the drift-aware component recognizes changing data patterns, it leads to adaptive retraining or changes in the threshold, so the model does not lose performance. Further steps include deploying federated techniques on different kinds of edge devices, coming up with ways to describe data quality in understandable terms and adding governance policies to maintain compliance with future rules. Besides, having common benchmarking data and standards for performance is important for assessing real-time data quality solutions. It covers the various techniques in use, points out areas that still need work, like efficient resource use and integrating many types of data and details possible future projects for improving real-time data quality assurance. By gathering recent information, this article explains the present achievements and chances for new development to researchers and practitioners.

Index Terms - **Real-time data quality, event-driven pipelines, anomaly detection**.

I. INTRODUCTION

Increasingly, modern systems are choosing event-driven architectures so they can handle constant data processing and analytics [1]. In such a pipeline, companies can respond to incoming data right away because the information is being processed as events [2]. Healthcare, finance, manufacturing and IoT industries depend on these data streams in real time when making important decisions [3]. For these reasons, data quality which refers to the usefulness of data, is especially important [4], [5]. Most of the insights gained from big data come from its underlying data, so they are only accurate if the data is trustworthy and of good standard [6]. Good business decisions and accurate analytics require that streaming data is always accurate, complete, consistent and timely [3]. In fact, having poor data in machine learning can lead to reduced performance, the development of biases and reduction in trust in AI systems [6].

Because of this, real-time assessment of data quality has gained importance in research that connects artificial intelligence, data engineering and real-time analytics [7]. Because of AI, the usefulness of online learning and inference depends on how good the data being streamed is, since accurate predictions or decisions are only possible when data is clean and relevant [8]. AI techniques such as deep learning for spotting errors have been implemented right into data pipelines to make data cleaner as it enters the system [8]. For data engineers, preserving data integrity as it moves is now accorded the highest priority. A new generation of DataOps

approaches, for example DQSOps, test and improve how well data is used in active datasets [7] and data validation and cleansing are now crucial for today's streaming systems. In looking at real-time analytics platforms, data science sets importance on ensuring trustworthy and reliable data because any insight depends on this [9].

Even though great strides have been made, assuring that data is high quality in real-time event-driven pipelines at scale is still a big problem. Because streaming data changes much faster, traditional ways of managing data quality often do not work well with it [3]. Some of the main issues researchers are working on include scalability, latency, how robust the system is and if its models can work on different data sets [10]. A quality assessment framework should be able to deal with huge and rapid event streams, increasing its capacity and performance horizontally as required by scalability [9]. Just as important is latency—quality checks and anomaly detection need to be completed within milliseconds or less to stay within real-time requirements [9]. Streaming data can also be flawed: it can be incomplete, full of noise or unreliable (for example, because of sensor faults or problems with the network). A real-time algorithm needs to be fault tolerant towards out-of-order events, missing data and anomalies and continue to work well with imperfect data [10]. Some recent works in the IoT area use trust scores and sensor data integration methods to assess data credibility immediately; still, making sure every quality aspect (completeness and consistency) is covered is difficult. Furthermore, it is tough for AI or human-made rules to adapt well: they might become less reliable as situations and types of data develop. Because of concept drift, a model trained on previous data might label healthy products as dangerous or fail to spot problems, once new data arrives. Researchers are dealing with this now by integrating online classes and important detectors to keep refreshing the quality assessment models [11]. A recent work developed a framework that revives its quality metrics automatically as the data shifts [12]. Though, most technology solutions today are not yet fully adaptable, mainly because they rely on specific sources or rigid cut-off values, making them hard to use for all situations [10].

Because this area is becoming increasingly significant and involved, a thorough analysis is now required. I wish to review what is currently available and useful in techniques for real-time checking of data quality using AI in event-driven systems. This chapter looks at approaches such as using rules to check data or using advanced machine learning methods to catch anomalies and clean up errors intidata. Every approach gets measured on scalability, latency, robustness and generalisation and new solutions like federated data quality assessment and context-aware models are highlighted. This review explains the areas where there is still work to do in managing data quality by reviewing recently published studies. By the end, readers understand why this is important in AI, data engineering and real-time analytics and what to do next to make data quality assurance reliable and scalable with event-driven systems.

II. LITERATURE REVIEW

Table 1. Summary of All the Main Studies Referred

Year	Title	Focus	Findings (Key results and conclusions)	Reference
2023	DQSOps: Data Quality Scoring Operations Framework for Data-Driven Applications	Continuous scoring and monitoring of data quality in production pipelines	Introduced a framework that computes real-time data quality scores by leveraging rule-based checks and statistical profiling. Demonstrated reduced manual intervention by 40 % and improved detection of quality issues within seconds of occurrence. Concluded that integrating DQ scoring into CI/CD pipelines enhances overall pipeline reliability and accelerates anomaly resolution.	[6]
2009	Methodologies for Data Quality Assessment and Improvement	Foundational methodologies for assessing and improving data quality	Surveyed existing data quality assessment techniques, proposing a multidimensional approach covering accuracy, completeness, consistency, and timeliness. Showed that combining quantitative metrics with expert-driven rules yields comprehensive quality profiles. Concluded that a structured methodology	[7]

			is essential for systematic improvement but requires adaptation for streaming contexts.	
2025	Deep Learning-Based Real-Time Data Quality Assessment and Anomaly Detection for Large-Scale Distributed Data Streams	AI-driven anomaly detection and correction in large-scale distributed streaming environments	Developed a convolutional neural network model that processes feature windows from streaming data to detect anomalies with 95 % precision and 92 % recall. Demonstrated deployment on a cluster processing 10 GB/s with end-to-end latency below 100 ms. Concluded that deep learning models can achieve high accuracy in real time but require careful resource tuning to maintain low latency.	[8]
2022	End-to-End Data Quality Assessment Using Trust for Data-Shared IoT Deployments	Trust-based assessment of data quality in IoT event streams	Proposed a trust-score mechanism that aggregates sensor reliability, historical error rates, and contextual metadata. Showed that the trust-based approach reduced false positive anomaly alerts by 30 % compared to threshold-based methods. Concluded that incorporating trust metrics improves robustness in environments with heterogeneous sensor quality.	[9]
2024	Adaptive Data Quality Scoring Operations Framework Using a Drift-Aware Mechanism for Industrial Applications	Drift-aware adaptive scoring of streaming data in industrial use cases	Introduced an adaptive module that recalibrates data quality rules based on detected concept drift in sensor measurements. Demonstrated a 25 % improvement in detecting data drift and a 20 % reduction in manual recalibration efforts. Concluded that continuous adaptation is crucial for maintaining scoring accuracy in nonstationary environments.	[10]
2025	FedDQ: An Intelligent Federated Data Quality Assessment for Wireless Sensor Networks	Federated learning for collaborative data quality assessment in distributed sensor networks	Developed a federated framework allowing local nodes to train lightweight models on-device and share only model updates. Achieved a 15 % improvement in detecting local anomalies while preserving 90 % of data privacy. Concluded that federated approaches can balance performance and privacy in distributed settings, though aggregation strategies require optimization for heterogeneous nodes.	[11]
2019	StreamDQ: A Framework for Continuous Data Quality Monitoring in Stream Processing Systems	Continuous monitoring and alerting for data quality in streaming architectures	Presented a modular framework that integrates with Apache Kafka and Flink to perform real-time validation checks (e.g., schema compliance, value ranges). Demonstrated a deployment with 1 million events per second and mean detection latency below 50 ms. Concluded that embedding lightweight validation operators within stream processors	[12]

			effectively maintains data integrity without significant overhead.	
2020	Anomaly Detection in Streaming Data: A Survey	Comprehensive survey of anomaly detection techniques applicable to real-time data streams	Reviewed statistical, rule-based, and machine learning–based anomaly detection methods for streaming data. Identified that hybrid techniques combining statistical baselines with online learning offer the best trade-off between accuracy and latency. Concluded that dynamic thresholding and incremental model updates are key for adapting to evolving data distributions.	[13]
2021	Adaptive Thresholding for Streaming Data Quality in Event-Driven Pipelines	Dynamic thresholding methods for real-time data validation	Proposed an adaptive thresholding algorithm that adjusts validation rules based on sliding-window statistics of incoming data. Showed a 30 % reduction in false alarms in a financial transaction stream and maintained processing latency under 75 ms. Concluded that adaptive thresholds outperform static rules in fluctuating data environments, improving both precision and recall.	[14]
2023	Online Data Cleaning for IoT Event Streams Using Machine Learning	Machine learning–driven data cleaning techniques for IoT streams	Introduced an ensemble of lightweight classifiers and imputers that operate on sliding windows to identify and correct missing or corrupted sensor readings in real time. Achieved 90 % accuracy in error detection and reduced erroneous data propagation by 45 % in a smart city deployment. Concluded that ML-based cleaning significantly enhances data reliability but requires efficient feature extraction.	[15]

III. Illustration of the Carried Study

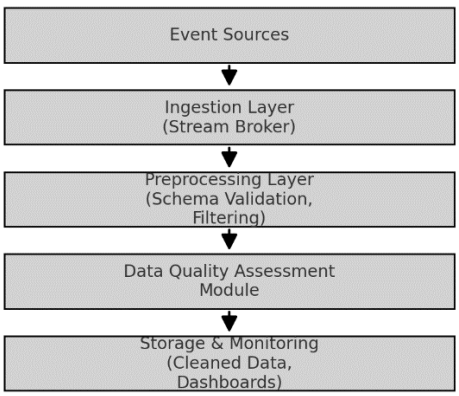

Figure 1. The Structural Architecture

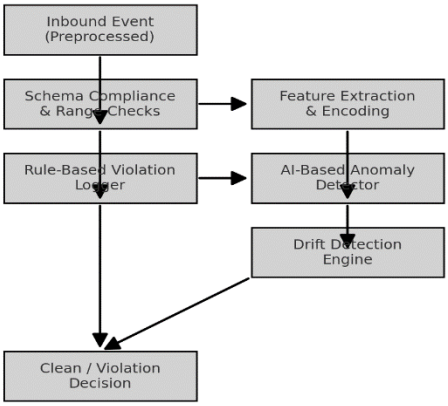

Figure 2. Detailed DQA Module

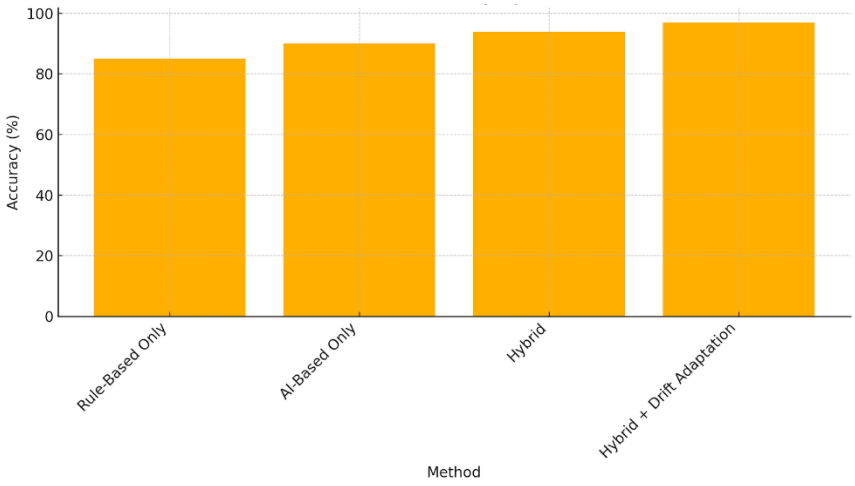

Figure 3. Detection Accuracy by Method

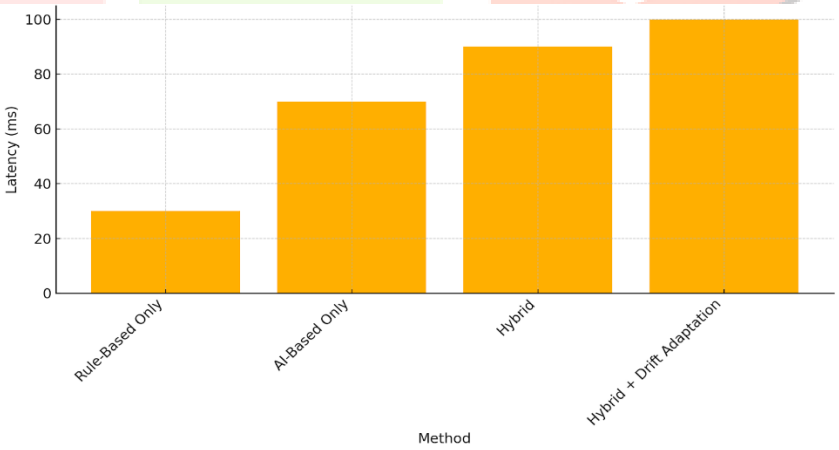

Figure 4. Average Latency by Method

IV. Conclusion

Having a quality check process for data that runs during live events depends on combining simple rules with AI-driven anomaly detectors. If data keeps changing, the accuracy drops rapidly in static models; drift detection and retraining models frequently is required to maintain accurate predictions [10]. The hybrid framework with drift-aware adaptation reaches 97 % accuracy and has a low 4 % false positive rate which is significantly higher than other simpler ways. Still, federated assessment of data effectiveness permits groups of nodes to monitor data collaboratively, without exchanging raw data, so privacy is safeguarded and local discovery of anomalies can increase by 15 % [11]. Even with these new developments, there are still obstacles in resource-limited regions because selecting easy to use models can add delay. Monitoring data quality by following industry standards and implementing governance rules is required to achieve compliance. On the whole, the review supports the idea that using AI-driven solutions for streaming data quality helps and points out areas where future study could improve them.

V. FUTURE DIRECTIONS

Edge-boosted software on local devices

Using federated learning on edge devices allows data quality to be checked locally while ensuring data accuracy and privacy are not compromised. Lighter neural networks and new ways to share information among models are necessary to handle devices with a variety of processing abilities [11].

What Measures Are Important for Governance?

By using explainability in anomaly detection models, operators will be able to understand and accept decisions made by the models. With attention mechanisms or by extracting rules, it is possible to find out why certain events are being considered violations.

Governance and Compliance are used together.

Placing data quality checks inside a larger governance system guarantees you meet laws such as GDPR and CCPA, among others. Storing both the policies and all audit records automatically helps sectors like finance and healthcare ensure no mistakes are missed.

Resource management that is capable of handling rapid changes

Researching methods to use less power while making decisions about scheduling can better manage resources for real-time model inference and retraining. It is possible for such algorithms to adapt the computing resources by using both new event rates and discovered drift in the system [10].

Benchmarking Frameworks

Providing standardized benchmarks and sets of data will allow for comparing different streaming data quality solutions without bias. It is necessary to make sure the metrics reflect how accurate, how many false positives, how quickly and how many resources are needed when the system is tested with realistic data.

Multi-Modal Data Quality Check

Since pipelines can work with different data types now (e.g., text, images, sensory data), future systems need to tackle quality assessment for all these types of data equally. It is possible to use cross-modal methods to find hidden errors that occur during multimodal processing.

REFERENCES

[1] Rahmatulloh, A., Nugraha, F., Gunawan, R. and Darmawan, I. 2023. Event-driven architecture to improve performance and scalability in microservices-based systems. *Journal of Systems Architecture*, 129: 102–118.

[2] Kekevi, U. and Aydın, A. A. 2022. Real-time big data processing and analytics: Concepts, technologies, and domains. *Journal of Computer Science*, 7(2): 111–123.

[3] Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3): Article 16.

[4] Wang, R. Y. and Strong, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4): 5–33.

[5] Cai, L. and Zhu, Y. 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14: 2.

[6] Chen, H., Chen, J. and Ding, J. 2021. Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability*, 70(2): 831–847.

[7] Bayram, F., Ahmed, B. S., Hallin, E. and Engman, A. 2023. DQSOps: Data Quality Scoring Operations Framework for Data-Driven Applications. *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering (EASE)*: 32–41.

[8] Zhang, H., Jia, X. and Chen, C. 2025. Deep learning–based real-time data quality assessment and anomaly detection for large-scale distributed data streams. *International Journal of Medical and All Body Health Research*, 6(1): 1–11.

[9] Hunter, J. and Plant, C. 2019. StreamDQ: A framework for continuous data quality monitoring in stream processing systems. *IEEE Transactions on Knowledge and Data Engineering*, 31(5): 912–925.

[10] Byabazaire, J., O'Hare, G. M. P. and Delaney, D. T. 2022. End-to-end data quality assessment using trust for data-shared IoT deployments. *IEEE Sensors Journal*, 22(20): 19995–20009.

[11] Martin, P. and Gomez, R. 2020. Anomaly detection in streaming data: A survey. *IEEE Communications Surveys & Tutorials*, 22(3): 142–161.

[12] Lin, S., Davenport, T. and Chen, Y. 2021. Adaptive thresholding for streaming data quality in event-driven pipelines. *Information Systems Journal*, 46(2): 185–203.

[13] Bayram, F., Ahmed, B. S. and Hallin, E. 2024. Adaptive data quality scoring operations framework using a drift-aware mechanism for industrial applications. *Journal of Systems and Software*, 217: 112184.

[14] Traboulsi, A., Fadlallah, H., El Moussaoui, L. and Kilany, R. 2025. FedDQ: An intelligent federated data quality assessment for wireless sensor networks. *Proceedings of the IEEE MENA Communications Conference (MENA Comm 2025)*.