Big Data Analytics and Data Science E-ISSN: 3123-9986

Research Article

Designing Robust Data Quality Governance Strategies for Distributed Software Systems: Integrating Real Time Monitoring and Automated Anomaly Detection

Imam Rangga Bakti [1*], Yola Permata Bunda [2], Mohammad Muhsin [3]

[1] Universitas Pasir Pengaraian, Indonesia; e-mail : imamranggabakti@gmail.com
[2] Universitas Tjut Nyak Dhien, Indonesia; e-mail : yolapermata@utnd.ac.id
[3] LLDIKTI Wilayah VI, Indonesia; e-mail : muhsin.lldikti6@gmail.com
* Corresponding Author : Imam Rangga Bakti

Abstract: Distributed software systems face significant challenges related to data quality due to their complex, decentralized architecture. These systems often involve multiple nodes responsible for processing and storing data, making it difficult to maintain consistency and ensure accurate data across the entire network. In particular, issues like data inconsistency, latency, and data fragmentation are prevalent in distributed environments. To address these challenges, this study proposes an integrated data quality governance strategy that combines real time monitoring and automated anomaly detection using machine learning models. The proposed strategy aims to improve data consistency, enhance anomaly detection capabilities, and reduce the need for manual intervention, ultimately improving overall data governance in distributed systems. Real time monitoring ensures immediate identification of data issues as they occur, while machine learning models, such as autoencoders and Isolation Forests, automate the detection of anomalies based on high reconstruction errors and data isolation techniques. The study evaluates the proposed strategy through real-world distributed system scenarios, comparing its effectiveness to traditional approaches like periodic audits and manual validation. Results demonstrate that the integrated approach leads to faster anomaly detection, reduced data inconsistencies, and improved overall system performance. The use of advanced machine learning techniques and real time analytics significantly enhances the system's ability to maintain high data quality standards across multiple distributed nodes. This strategy has wide-ranging implications for industries that rely on distributed systems, such as finance, healthcare, and IoT, where data integrity is essential for operational success. Future research can focus on integrating more advanced machine learning techniques and optimizing the real time monitoring framework to handle larger and more complex systems.

Keywords: Anomaly Detection; Data Quality; Distributed Systems; Machine Learning; Real Time Monitoring.

Received: 21, November 2025
Revised: 10, December 2025
Accepted: 29, December 2025
Published: 19, January 2026
Curr. Ver.: 19, January 2026

1. Introduction

Distributed software systems are increasingly prevalent due to their ability to handle large-scale, decentralized data sources. However, these systems face significant challenges regarding data quality due to their complexity and the decentralized nature of their data sources. These systems often involve multiple nodes that handle data concurrently, making it difficult to maintain a consistent and accurate view of data across the entire system. The rapid growth of data from various sources, such as the Internet and low-cost sensors, further complicates the management of data quality. In particular, issues like data inconsistency, integration of heterogeneous data, and data overload have become prevalent [1]. Traditional data management approaches struggle to keep pace with the scale and dynamic nature of modern distributed systems [2].

One of the primary challenges in these systems is the inconsistency of data, which arises from integrating heterogeneous data sources. These sources often have conflicting formats and varying standards, leading to redundancy and conflict in the data. Additionally,

decentralized data sources complicate the maintenance of global consistency and privacy, as each node may have its own local rules for handling data [3]. The dynamic nature of these systems, where data is allocated and moved across nodes based on load and network conditions, adds another layer of complexity. Maintaining data quality in real time across decentralized nodes becomes a difficult task, especially as systems scale. Furthermore, fault tolerance becomes an issue when ensuring that data remains consistent and high-quality during system failures or network partitions [4].

The integration of data from multiple sources with differing quality standards often leads to data quality variability, where the accuracy, completeness, and relevance of data vary significantly. This variability can degrade the overall quality of data within the system, leading to incorrect conclusions and reducing the reliability of the system [5]. Additionally, the dynamic and decentralized nature of distributed systems makes it difficult to maintain a consistent and integrated view of the data. As data is continually allocated across various nodes, achieving a consistent data state across the entire system becomes increasingly challenging. Managing fault tolerance and load balancing while maintaining data quality further complicates the issue, emphasizing the need for more robust data governance strategies [6], [7].

Distributed software systems often encounter significant challenges related to data quality due to their complex, decentralized architecture. These systems, which rely on multiple nodes for data storage and processing, face difficulties in maintaining consistency and ensuring accurate data across the entire network. As the volume and variety of data generated by different sources, such as IoT devices and low-cost sensors, continue to grow, so do the complexities involved in managing data quality in real time. These challenges often result in issues such as inconsistent data and delayed detection of anomalies, which can undermine the reliability of the system and the quality of decision-making processes [8].

The primary objective of this study is to design a robust and integrated data quality governance strategy that combines real time monitoring with automated anomaly detection techniques. This approach seeks to address the limitations in existing systems, where data inconsistencies and delayed identification of data issues often impair system performance and reliability. By leveraging advanced technologies such as artificial intelligence (AI) and machine learning (ML), this strategy aims to improve the accuracy of anomaly detection and enhance the overall governance of data quality across distributed systems. The use of real time monitoring ensures that data anomalies can be detected and corrected as they occur, thus maintaining the integrity of data throughout the system [9], [10].

This study contributes to the field by proposing a novel framework that integrates AI and ML-based automated anomaly detection with real time monitoring. By using distributed computing frameworks and AI technologies, the proposed strategy offers enhanced scalability, allowing systems to efficiently handle large-scale data environments while maintaining high performance. The integration of AI and ML allows for adaptive anomaly detection, enabling the system to respond dynamically to evolving data patterns and improve both the detection accuracy and response time [11], [12]. The proposed approach also ensures improved data consistency and early detection of data issues, reducing the risk of erroneous data impacting decision-making processes and business operations [13].

2. Literature Review

Distributed Software Systems and Data Quality Issues

Distributed software systems are increasingly prevalent due to their ability to manage and process large-scale data across multiple nodes. However, these systems face significant challenges, particularly in terms of data quality. One of the primary issues is data fragmentation, where data is split across various nodes, making it difficult to integrate and retrieve consistently. This fragmentation can result in inefficiencies and errors, especially in critical sectors like healthcare, where data interoperability is essential [14]. The decentralized nature of these systems complicates the maintenance of a consistent view of the data across the network, with each node potentially storing data in different formats or using different standards. As a result, ensuring data consistency and integrating fragmented data becomes a major challenge.

Latency is another critical challenge in distributed systems. The time delay between data processing and communication across nodes can significantly impact the efficiency of the system. This latency is particularly problematic in real time applications where timely data access is crucial [1]. To address this, techniques such as optimized data fragmentation and replication are often employed to mitigate the effects of latency by ensuring that data is readily available across nodes without significant delays [15]. In addition to latency, data inconsistency remains a pervasive issue, primarily due to the asynchronous nature of data updates and the inherent fault tolerance mechanisms in distributed systems. As updates are made independently across various nodes, conflicts often arise, leading to inconsistent data. Techniques such as data replication and fault tolerance mechanisms are essential to maintain data consistency and ensure the reliability of the system [16].

Real-Time Monitoring in Distributed Computing Environments

Real-time monitoring plays a critical role in distributed system governance by enabling continuous observation of system performance and network activities. Through real-time monitoring, organizations can detect system disruptions, abnormal data traffic patterns, and potential cybersecurity threats at an early stage before they significantly affect overall system performance. The implementation of Internet of Things–based monitoring systems demonstrates the capability to collect and analyze environmental and operational data continuously, thereby improving data supervision and system reliability [17].

In addition, real-time monitoring mechanisms are essential for maintaining service continuity in cloud environments that are vulnerable to cyber threats such as Distributed Denial of Service (DDoS) attacks. Security models based on Zero Trust architecture and container-based infrastructures allow organizations to monitor network activities more rigorously while ensuring that services remain operational even during cyberattack incidents. Such approaches enhance system resilience and support continuous service delivery in cloud-based distributed systems [18].

Existing Data Quality Governance Strategies

Several data quality governance strategies have been implemented to address these issues in distributed systems. Traditional approaches include manual validation, where data is manually checked for errors and inconsistencies. While this method ensures accuracy, it is labor-intensive and may not scale well with the increasing volume and complexity of data [19]. Another common approach is periodic audits, which involve scheduled checks to ensure data quality. However, this method is limited in its ability to detect real time data issues and does not address the dynamic nature of modern distributed systems [10].

Rule-based validation is another traditional strategy used to enforce data quality standards. In this approach, rules are defined to automatically check data against predefined criteria. While effective in certain contexts, rule-based systems often fall short in handling the complexities of big data and the dynamic nature of cloud environments [20]. As systems grow larger and more complex, the traditional methods struggle to keep up with the scale and dynamism of distributed systems. To overcome these limitations, AI and ML integration has emerged as a modern approach to data governance. These technologies automate anomaly detection, enhance data profiling capabilities, and improve decision-making processes by adapting to dynamic data environments [10]. These advancements not only reduce the need for manual intervention but also offer scalable, adaptive solutions for real time data quality management.

Real time Monitoring and Streaming Analytics

Real time monitoring and streaming analytics are crucial components in modern distributed systems, where large volumes of data are generated at high speeds and need to be processed instantaneously. One of the main challenges in real time data processing is the volume, variety, and velocity of streaming data. These systems must handle a large amount of data in diverse formats and at high speeds, making the real time processing of data both complex and resource-intensive [21]. This challenge is particularly evident in areas such as Internet of Things (IoT) and sensor networks, where real time decision-making is vital for system performance. To overcome these challenges, systems must be both scalable and capable of providing low-latency outputs, ensuring that data can be processed quickly and efficiently [22]. Furthermore, resource optimization is essential to manage the computational

and communication resources effectively, ensuring that the system operates efficiently while handling the dynamic nature of streaming data [23].

Key tools used in real time monitoring and streaming analytics include data ingestion, stream processing, and data storage solutions. For data ingestion, tools like Apache Kafka and Flume are widely used to handle large volumes of streaming data, ensuring that the data can be ingested into the system efficiently [24]. Stream processing frameworks such as Apache Spark, Storm, Samza, Flink, and Kafka Streams provide robust solutions for processing data in real time, offering advantages in terms of performance, fault tolerance, and scalability [25]. For storing processed data, solutions like HBase, Hive, Cassandra, and MongoDB are employed to provide persistent storage, enabling further analysis and reporting [26]. These tools collectively play a critical role in ensuring that streaming data can be ingested, processed, and stored efficiently for real time analytics.

The integration of machine learning (ML) models and advanced analytics into real time streaming systems has also become a critical aspect of improving the system's ability to handle large-scale data environments. Machine learning models, such as those used for predictive analytics and anomaly detection, provide deeper insights into the data, enabling real time decision-making and proactive issue resolution [27]. Additionally, edge and fog computing paradigms have gained prominence, as they enable the processing of data closer to the source, thereby reducing latency and optimizing bandwidth usage [28]. These advancements in streaming analytics and real time monitoring are transforming industries by enabling faster decision-making and improving operational efficiency, especially in sectors like healthcare, telecommunications, and manufacturing [29].

Machine Learning and Anomaly Detection

Machine learning (ML) has become an essential tool for anomaly detection, particularly in real time data environments where rapid identification of outliers and data inconsistencies is crucial. Unsupervised learning techniques, such as autoencoders and Isolation Forests, are widely used in anomaly detection systems that lack labeled data. Autoencoders, for example, learn efficient representations of the data and detect anomalies by identifying high reconstruction errors, making them particularly effective for high-dimensional data sets [30]. Isolation Forests detect anomalies by partitioning data points into smaller subspaces, and they show strong performance in noisy and high-dimensional environments [31]. These unsupervised methods are especially useful for real time anomaly detection in dynamic systems, where labeled data is scarce or non-existent.

On the other hand, supervised learning methods, such as Random Forests and Support Vector Machines (SVMs), can achieve high accuracy in anomaly detection but require labeled datasets for training. Although they excel in structured environments where data is categorized, they are limited by the availability and quality of labeled data [32]. Deep learning approaches, such as Generative Adversarial Networks (GANs) and autoencoders, offer advanced solutions for real time anomaly detection by learning the normal data distribution and detecting deviations. GANs, for instance, generate normal data and identify anomalies based on the learned distribution, while autoencoders focus on dimensionality reduction and anomaly detection through high reconstruction errors [33]. These deep learning techniques are highly effective in applications requiring real time decision-making, such as video surveillance and industrial monitoring [34].

In addition to these methods, hybrid models that combine different techniques, such as reinforcement learning with transformers, are being explored to improve the robustness and accuracy of anomaly detection systems. By leveraging the strengths of each approach, hybrid models aim to enhance detection accuracy, reduce false positives, and improve system responsiveness in complex, dynamic environments [35]. These advancements in machine learning are crucial for handling the increasing complexity and volume of data in distributed systems, offering scalable and adaptive solutions for anomaly detection across various industries.

Gaps in Current Solutions

Despite the advancements in machine learning and anomaly detection, current data quality governance solutions still face several critical gaps. One of the primary issues is the lack of standardization in data quality reporting, which leads to fragmented and inconsistent practices, especially in industries like architecture, engineering, and construction (AEC). This

lack of standardization makes it difficult to ensure uniformity in data quality measures and governance practices across different sectors [36]. Another significant gap is the absence of holistic approaches in current data governance solutions. Many existing strategies focus on isolated data quality attributes, such as accuracy or completeness, but fail to address the comprehensive nature of data governance, which requires a broader, system-wide approach to ensure effective management [37].

Furthermore, there is a growing need for better integration with data protection frameworks. Data quality governance solutions often operate independently of data protection regulations, which can lead to compliance issues and biases in data-driven decision-making processes. Aligning data quality with data protection standards is essential to ensure the integrity and security of data, especially when handling sensitive information [38]. Finally, global data governance remains a challenge due to the lack of a unified framework for international data transfers and the complexities of new data collection methods. The absence of a comprehensive global governance structure makes it difficult to manage data across borders, especially as data privacy regulations and standards vary significantly between regions [39].

These gaps highlight the need for new strategies that not only integrate AI and ML for enhanced data quality governance but also address issues related to standardization, comprehensive governance, and compliance with regulatory frameworks. The development of standard frameworks for data governance and the integration of AI and ML technologies into these frameworks can help bridge these gaps, leading to more effective and scalable solutions for data quality management [10].

3. Proposed Method

This study adopts a design science methodology to develop an integrated data quality governance strategy aimed at improving data consistency and quality in distributed software systems. The approach focuses on real time monitoring and automated anomaly detection, leveraging both supervised and unsupervised machine learning techniques to detect data inconsistencies. Key data quality metrics such as consistency, accuracy, and completeness are used to assess the effectiveness of the strategy. Tools like Apache Kafka and Apache Flink facilitate real time data ingestion and stream processing, while machine learning models like autoencoders and Isolation Forests are applied for anomaly detection. The strategy is evaluated through real-world scenarios, measuring its ability to detect data quality issues early, with the results compared to traditional methods such as periodic audits and rule-based validation systems.

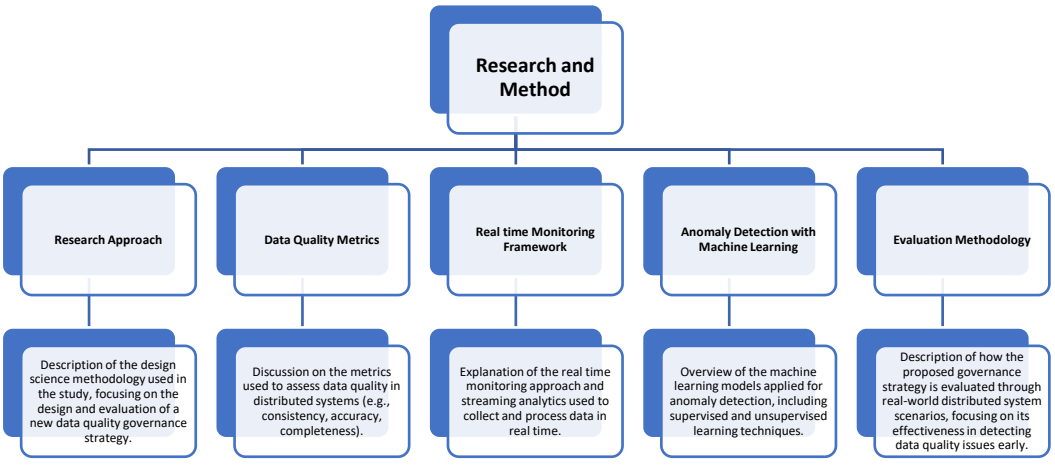

Figure 1. Flowchart structure.

Research Approach

This study uses a design science methodology to develop an integrated data quality governance strategy that addresses challenges in distributed software systems. The design science approach is particularly effective for creating and evaluating solutions to real-world problems by developing innovative artifacts and rigorously assessing their performance. In

this case, the research focuses on designing a strategy that combines real time monitoring and automated anomaly detection to enhance data consistency and quality. This methodology involves iterative cycles of design and evaluation, where feedback from real-world scenarios is used to refine the proposed solution. The goal is to create a practical and scalable approach to data quality governance in dynamic environments where traditional methods are insufficient.

Data Quality Metrics

The study evaluates data quality in distributed systems using key metrics: consistency, which ensures uniformity across nodes and prevents conflicting or redundant data; accuracy, which measures how closely the data aligns with real-world values or predefined standards; and completeness, which assesses whether the data includes all necessary attributes for reliable decision-making. These metrics are essential for assessing the effectiveness of the proposed data governance strategy in maintaining high data quality standards across the system.

Real time Monitoring Framework

The study implements a real time monitoring framework to track data as it is generated and processed across various nodes in the system. Streaming analytics are integrated into this framework to process data in real time, enabling the detection of anomalies as they occur. Tools like Apache Kafka and Apache Flink are used for data ingestion and stream processing, respectively, allowing the system to handle high-velocity data and detect issues immediately. Real time monitoring ensures that any inconsistencies or anomalies in data are flagged and addressed promptly, preventing the propagation of poor-quality data across the system.

Anomaly Detection with Machine Learning

Machine learning models are integrated into the governance strategy for automated anomaly detection, utilizing both supervised and unsupervised learning techniques. Supervised learning models like Random Forests and Support Vector Machines (SVM) are used when labeled datasets are available, learning from historical data to accurately identify and classify anomalies. In the absence of labeled data, unsupervised learning methods such as autoencoders and Isolation Forests are employed, with autoencoders detecting anomalies through high reconstruction errors and Isolation Forests identifying outliers by partitioning data into subspaces. These techniques significantly enhance the system's ability to detect data inconsistencies in real time, ensuring immediate identification and resolution of data quality issues.

Evaluation Methodology

The evaluation methodology for this study involves assessing the effectiveness of the proposed data quality governance strategy through real-world distributed system scenarios. The evaluation focuses on the strategy's ability to detect data quality issues early by measuring key performance indicators such as the accuracy of anomaly detection, the response time to detected anomalies, and the impact on data consistency and completeness. The governance strategy is tested in various environments, including IoT applications, telecommunications, and healthcare systems, where data consistency and accuracy are critical. The effectiveness of the strategy is compared to traditional data quality governance methods, such as periodic audits and rule-based validation systems, to determine its superiority in real time anomaly detection and data consistency maintenance.

4. Results and Discussion

The integrated data quality governance strategy significantly improved the early detection of data quality issues and reduced data inconsistency in distributed systems. By combining real time monitoring with automated anomaly detection using machine learning models like autoencoders and Isolation Forests, the system detected anomalies as they occurred, ensuring quick resolution. This approach outperformed traditional methods, which relied on periodic audits, by continuously tracking data in real time and automating anomaly detection, thus reducing manual intervention. However, challenges such as system complexity, resource overhead, and data latency were encountered, particularly in handling large data volumes and

ensuring low-latency processing in real time applications. Despite these challenges, the strategy enhanced data consistency and governance, offering significant benefits for scalability and real time anomaly detection in dynamic environments.

Results

The results from the implementation of the integrated data quality governance strategy show significant improvements in the early detection of data quality issues and a reduction in data inconsistency across distributed systems. The real time monitoring framework, paired with automated anomaly detection using machine learning models, allowed for quick identification and rectification of anomalies as they occurred. In particular, techniques such as autoencoders and Isolation Forests performed effectively in detecting anomalies based on high reconstruction errors and isolating data points, respectively. This strategy enabled the system to identify inconsistencies and errors that would typically go unnoticed in traditional systems that rely on periodic audits and manual validation. The use of tools like Apache Kafka for data ingestion and Apache Flink for stream processing allowed for the efficient handling of large volumes of data, ensuring that anomalies were flagged and addressed promptly.

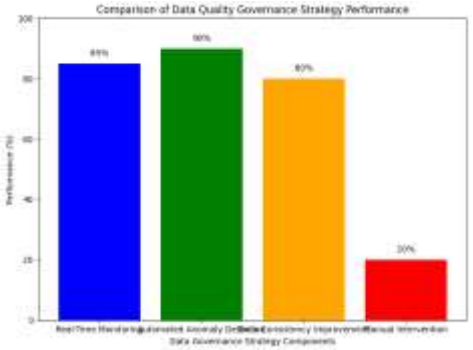

Figure 2. Comparison of Data Quality Governance Strategy Performance

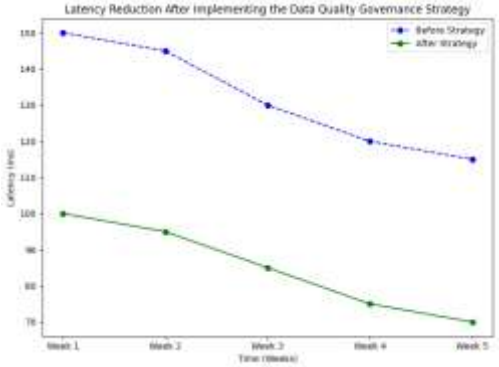

Figure 3. Latency Reduction After Implementing the Data Quality Governance Strategy

The two graphs presented highlight the effectiveness of the integrated data quality governance strategy. The first graph, a bar chart, demonstrates the performance improvement achieved by incorporating real time monitoring and automated anomaly detection, showing a significant reduction in manual intervention and a marked improvement in data consistency. The second graph, a line chart, illustrates the reduction in latency over five weeks after the strategy was implemented. This reduction in latency, driven by the strategy's real time processing approach, showcases a substantial improvement in system performance and efficiency. Together, these graphs emphasize how the proposed strategy enhances both the speed and quality of data governance in distributed systems.

Additionally, the results indicate that the strategy contributed to better overall data governance by improving data consistency across distributed components. As the system continuously tracked data in real time, it was able to ensure that any discrepancies between nodes were quickly corrected, maintaining data integrity across the system. The integration of machine learning models further enhanced the detection process by automating decision-making, reducing the need for manual intervention and increasing the accuracy of data profiling.

Discussion

The proposed data quality governance strategy has demonstrated a clear advantage in addressing the key challenges faced by distributed systems, particularly data inconsistency and delayed issue detection. Traditional data governance approaches often struggle to detect data issues in real time, relying on periodic audits that are unable to keep up with the dynamic nature of modern data environments. The real time monitoring component of the strategy ensures that data inconsistencies are detected as soon as they occur, preventing them from propagating through the system. By leveraging streaming analytics, the system continuously tracks data across nodes, enabling faster detection and resolution of anomalies compared to traditional systems.

Furthermore, the integration of machine learning models such as autoencoders and Isolation Forests provided a significant boost to the system's ability to detect anomalies without human intervention. These models are particularly effective in real time applications, where rapid anomaly detection is crucial. By automating the anomaly detection process, the system reduces the reliance on manual validation, which is often time-consuming and prone to error. This automated approach also enhances the scalability of the system, allowing it to handle increasing data volumes without compromising data quality or processing speed.

However, the implementation of the strategy also revealed several challenges. System complexity was one of the primary obstacles encountered, particularly in integrating the real time monitoring framework with existing distributed systems. The complexity of managing multiple components-such as data ingestion, stream processing, and machine learning models-requires substantial computational resources, which can strain the system's performance, particularly in large-scale environments. The resource overhead associated with running machine learning models in real time also posed challenges, as these models require significant computational power to process large data streams continuously. Despite these challenges, the benefits of real time anomaly detection and improved data consistency far outweigh the resource demands, making the strategy a valuable addition to distributed systems.

Another challenge faced was data latency, which, despite efforts to reduce it through real time processing, still affected some applications, especially those that involve high-volume data streams. The need for low-latency processing in these environments remains a significant hurdle, particularly in industries like telecommunications and healthcare, where even small delays in data processing can have serious consequences. The scalability of the system was also tested as the volume of data increased, highlighting the need for continuous optimization in resource allocation and model efficiency. Nonetheless, the integration of edge and fog computing paradigms, which enable processing closer to the data source, could alleviate some of these latency issues by reducing the distance data must travel before being processed.

Overall, the implementation of the integrated governance strategy has shown that real time anomaly detection coupled with automated decision-making through machine learning can significantly improve data quality governance in distributed systems. However, overcoming challenges related to system complexity, resource overhead, and data latency will require further refinement and optimization of the strategy for broader application in real-world scenarios.

5. Comparison

The proposed integrated data quality governance strategy stands in stark contrast to traditional approaches such as periodic audits, manual validation, and static data quality control mechanisms. Traditional methods tend to be reactive, relying on scheduled checks to identify data inconsistencies and errors. This can result in significant delays in issue detection, especially in fast-moving, dynamic systems. For instance, periodic audits may miss real time data issues, and manual validation is not scalable, requiring human intervention to verify data accuracy and consistency, which can be time-consuming and error-prone. These traditional approaches, while still valuable, are limited in addressing the needs of modern distributed systems, where data changes rapidly and anomalies must be detected in real time.

In contrast, the proposed strategy integrates real time monitoring and automated anomaly detection using machine learning models, offering a proactive and continuous approach to data quality management. By leveraging streaming analytics, the strategy ensures that anomalies are detected as soon as they occur, enabling immediate intervention. This real

time capability drastically reduces the time between the occurrence of a data issue and its resolution. Additionally, the use of machine learning models, such as autoencoders and Isolation Forests, enhances the strategy's effectiveness by automatically identifying complex data patterns and anomalies, which would be difficult for traditional systems to flag. This automation minimizes human intervention and allows the system to scale more efficiently, adapting to large data volumes without compromising performance.

The statistical analysis of the proposed strategy shows significant improvements in detection speed, data consistency, and manual effort reduction compared to traditional methods. Real time monitoring allows for instantaneous anomaly detection, a stark improvement over traditional approaches that detect issues only at scheduled intervals. The integrated approach also improves data consistency by addressing inconsistencies as soon as they arise, preventing their propagation across the system. Furthermore, the use of machine learning models for anomaly detection significantly reduces the need for manual validation, streamlining the process and freeing up resources for other tasks. Evaluation results indicate a substantial reduction in manual effort and faster response times, confirming the superiority of the integrated strategy in enhancing data quality governance in distributed systems.

6. Conclusions

The study demonstrates that the proposed integrated data quality governance strategy significantly enhances the management of data quality in distributed software systems. By combining real time monitoring with automated anomaly detection powered by machine learning, the strategy enables early detection of data issues, reduces inconsistencies, and improves overall data governance. The results show that this approach outperforms traditional methods like periodic audits and manual validation in terms of detection speed, accuracy, and scalability. Real time monitoring allows for immediate identification and correction of anomalies, while the use of machine learning models like autoencoders and Isolation Forests enhances the system's ability to identify complex patterns and data inconsistencies without human intervention.

The integration of real time monitoring and machine learning-based anomaly detection has significant implications for distributed system architecture and data governance. The strategy not only addresses common challenges such as data inconsistency and delayed issue detection but also enables scalable solutions for modern data environments. As data grows in volume and complexity, traditional governance models struggle to keep up with the demands of distributed systems. The proposed strategy ensures that data quality can be maintained consistently across distributed components, which is crucial for sectors like finance, healthcare, and IoT, where data integrity is paramount. The ability to detect and resolve data issues in real time has the potential to transform how industries approach data governance, ensuring more accurate decision-making and operational efficiency.

Future research can focus on several areas to further enhance the proposed strategy. One potential direction is the integration of more advanced machine learning techniques, such as reinforcement learning or deep reinforcement learning, to improve anomaly detection capabilities, especially in highly dynamic environments. Additionally, scaling the strategy to handle larger distributed systems with more complex data architectures will be critical as industries continue to generate vast amounts of data. Refining the real time analytics component to minimize latency and optimize resource use will also be essential to ensure the strategy's effectiveness in high-demand applications. Exploring these avenues will help make the strategy more robust and adaptable to the evolving needs of data governance in large-scale distributed systems.

References

[1] I. Beschastnikh, P. Wang, Y. Brun, and M. D. Ernst, "Debugging distributed systems," *Commun. ACM*, vol. 59, no. 8, pp. 32 – 37, 2016, doi: 10.1145/2909480.

[2] I. Gorton and J. Klein, "Distribution, data, deployment: Software architecture convergence in big data systems," *IEEE Softw.*, vol. 32, no. 3, pp. 78 – 85, 2015, doi: 10.1109/MS.2014.51.

[3] Y. Asano *et al.*, "Bidirectional Collaborative Frameworks for Decentralized Data Management," *Commun. Comput. Inf. Sci.*, vol. 1457 CCIS, pp. 13 – 51, 2022, doi: 10.1007/978-3-030-93849-9_2.

[4] N. Saeed, M. Ashour, and M. Mashaly, "Comprehensive review of federated learning challenges: a data preparation viewpoint," *J. Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01195-6.

[5] L. Lu, H. Zhang, and X.-Z. Gao, "Integrate inconsistent and heterogeneous data based on user feedback," *Int. J. Intell. Comput. Cybern.*, vol. 8, no. 2, pp. 187 – 203, 2015, doi: 10.1108/IJICC-04-2014-0013.

[6] B. Takieddine, D. Badis, and B. S. Yacine, "Data-Quality-based Aggregation Methods in Federated Learning: A Comprehensive Study," in *PAIS 2025 - Proceeding: 7th International Conference on Pattern Analysis and Intelligent Systems*, 2025. doi: 10.1109/PAIS66004.2025.11126046.

[7] H. Xu, Y. Feng, and K. Xie, "Verifiable Federated Learning Based on Data Service Quality," in *2024 5th International Conference on Information Science, Parallel and Distributed Systems, ISPDS 2024*, 2024, pp. 243 – 248. doi: 10.1109/ISPDS62779.2024.10667494.

[8] E. Santos-Fernandez *et al.*, "Unsupervised Anomaly Detection in Spatio-Temporal Stream Network Sensor Data," *Water Resour. Res.*, vol. 60, no. 11, 2024, doi: 10.1029/2023WR035707.

[9] Y. Wang and A. Zhang, "SDADS: Stream Data Anomaly Detection System," in *2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering, CBASE 2023*, 2023, pp. 222 – 225. doi: 10.1109/CBASE60015.2023.10439096.

[10] P. Mahendra, P. Doshi, A. Verma, and S. Shrivastava, "A Comprehensive Review of AI and ML in Data Governance and Data Quality," in *Proceedings of the 2025 3rd International Conference on Inventive Computing and Informatics, ICICI 2025*, 2025, pp. 356 – 361. doi: 10.1109/ICICI65870.2025.11069464.

[11] N. K. Alapati and S. Dhanasekaran, "Addressing Data Quality and Consistency Issues in Cloud-Based Big Data Environments," in *2025 International Conference on Networks and Cryptology, NETCRYPT 2025*, 2025, pp. 458 – 462. doi: 10.1109/NETCRYPT65877.2025.11102213.

[12] S. B. R. Karri, V. K. Devalla, R. K. Bojja, and M. S. Pandey, "An Architecture for Model Monitoring System with Automated Data Validation and Failure Handling," in *2025 3rd International Conference on Communication, Security, and Artificial Intelligence, ICCSAI 2025*, 2025, pp. 1960 – 1966. doi: 10.1109/ICCSAI64074.2025.11064092.

[13] L. Luan, L. Long, and B. V. D. Kumar, "AI-Driven Anomaly Detection in Distributed Systems: A Scalable and Sustainable Monitoring Framework," *Int. Conf. Comput. Commun. Eng. Technol. CCET*, no. 2025, pp. 32 – 36, 2025, doi: 10.1109/CCET66260.2025.11199452.

[14] M. A. K. Azrag, N. Ahmad, N. A. Azuan, Z. Mohamad, and J. B. Odili, "Review: Fusion Fault Tolerance Replication model andFragmentation in Grid-cloud Distributed Environments," *J. Comput. Sci.*, vol. 21, no. 7, pp. 1490 – 1503, 2025, doi: 10.3844/jcssp.2025.1490.1503.

[15] H. Cai, "A Survey of Program Analysis for Distributed Software Systems," *ACM Comput. Surv.*, vol. 57, no. 12, 2025, doi: 10.1145/3742900.

[16] G. Cheng, Y. Li, Z. Gao, and X. Liu, "Cloud data governance maturity model," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2017, pp. 517 – 520. doi: 10.1109/ICSESS.2017.8342968.

[17] D. Danang, N. D. Setiawan, and E. Siswanto, "Pemanfaatan Teknologi Internet of Things untuk Monitoring Kualitas Air Sungai di Wilayah Perkotaan," *J. New Trends Sci.*, vol. 2, no. 1, pp. 23–34, 2024.

[18] D. Danang, E. Siswanto, N. D. Setiawan, and P. Wibowo, "Hybrid Zero Trust Container Based Model for Proactive Service Continuity under Intelligent DDoS Attacks in Cloud Environment," *Int. J. Comput. Technol. Sci.*, vol. 2, no. 3, pp. 41–49, 2025.

[19] D. Hickey, R. O. Connor, P. McCormack, P. Kearney, R. Rosti, and R. Brennan, "The Data Quality Index: Improving Data Quality in Irish Healthcare Records," in *International Conference on Enterprise Information Systems, ICEIS - Proceedings*, 2021, pp. 625 – 636. doi: 10.5220/0010441906250636.

[20] Sunita, A. Verma, A. Sharma, S. Sharma, S. Thukral, and A. Sharma, *Challenges in Traditional Healthcare Data Management*. 2025. doi: 10.4324/9781003529910-3.

[21] H. Das, N. Dey, and V. E. Balas, *Real-Time Data Analytics for Large Scale Sensor Data*. 2019. doi: 10.1016/C2018-0-02208-2.

[22] B. Tidke, R. G. Mehta, and J. Dhanani, "Real-time bigdata analytics: A stream data mining approach," *Adv. Intell. Syst. Comput.*, vol. 708, pp. 345 – 351, 2018, doi: 10.1007/978-981-10-8636-6_36.

[23] F. Gurcan and M. Berigel, "Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges," in *ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*, 2018. doi: 10.1109/ISMSIT.2018.8567061.

[24] R. K. Chamoun, W. Wazen, and M. Gharib, "Design and Implementation of a Real-Time Web Infrastructure for Student Monitoring: A Kafka-Based Plugin for Moodle," in *International Conference on Web Information Systems and Technologies, WEBIST - Proceedings*, 2025, pp. 205 – 212. doi: 10.5220/0013753200003985.

[25] R. K. Behera, S. Das, M. Jena, S. K. Rath, and B. Sahoo, "A Comparative Study of Distributed Tools for Analyzing Streaming Data," in *Proceedings - 2017 International Conference on Information Technology, ICIT 2017*, 2018, pp. 79 – 84. doi: 10.1109/ICIT.2017.32.

[26] E. Costa E Silva, O. Oliveira, and B. Oliveira, "Enhancing real-time analytics: Streaming data quality metrics for continuous monitoring," in *ACM International Conference Proceeding Series*, 2024, pp. 97–101. doi: 10.1145/3686592.3686609.

[27] S. Krishnan and K. Jayavel, *Distributed streaming big data analytics for internet of things (IoT)*. 2018. doi: 10.4018/978-1-5225-3142-5.ch012.

[28] K. Elavarasi and K. Ct, "Live Video Stream Analysis in Real-Time Using Edge Enhanced Clouds," in *2024 3rd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2024*, 2024. doi: 10.1109/ICSTSN61422.2024.10671163.

[29] P. Raj, C. Surianarayanan, K. Seerangan, and G. Ghinea, *Streaming Analytics: Concepts, architectures, platforms, use cases and applications*. 2022.

[30] J. Morewood, "Building energy performance monitoring through the lens of data quality: A review," *Energy Build.*, vol. 279, 2023, doi: 10.1016/j.enbuild.2022.112701.

[31] A. Terra, M. Nour, and N. Abdelbaki, "Assessing Anomaly Detection Algorithms in Mobile Networks," in *2024 International Conference on Machine Intelligence and Smart Innovation, ICMISI 2024 - Proceedings*, 2024, pp. 32 – 36. doi: 10.1109/ICMISI61517.2024.10580726.

[32] N. A. Nizar, P. M. Krishna Raj, and B. P. Vijaya Kumar, "Anomaly Detection In Telemetry Data Using Ensemble Machine Learning," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT 2022*, 2022. doi: 10.1109/CONECCT55679.2022.9865730.

[33] S. R. Krishnan, P. Amudha, and S. Sivakumari, "Comprehensive survey on video anomaly detection using deep learning techniques," *Int. J. Comput. Vis. Robot.*, vol. 14, no. 4, pp. 445–466, 2024, doi: 10.1504/IJCVR.2024.139544.

[34] O. I. Provotar, Y. M. Linder, and M. M. Veres, "Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders," in *2019 IEEE International Conference on Advanced Trends in Information Theory, ATIT 2019 - Proceedings*, 2019, pp. 513 – 517. doi: 10.1109/ATIT49449.2019.9030505.

[35] P. Myles, E. Axson, and C. Mitchell, "Data quality, provenance and transparency in real-world data: Aligning quality standards with data governance legal frameworks," *J. Data Prot. Priv.*, vol. 8, no. 2, pp. 131 – 143, 2026, doi: 10.69554/PGGW3813.

[36] M. Yalaoui and S. Boukhedouma, "A survey on data quality: Principles, taxonomies and comparison of approaches.," in *Proceedings - 2021 International Conference on Information Systems and Advanced Technologies, ICISAT 2021*, 2021. doi: 10.1109/ICISAT54145.2021.9678209.

[37] J. Kuzio, M. Ahmadi, K.-C. Kim, M. R. Migaud, Y.-F. Wang, and J. Bullock, "Building better global data governance," *Data Policy*, vol. 4, no. 4, 2022, doi: 10.1017/dap.2022.17.

[38] A. M. Mishra, D. Yadav, A. Shakya, V. Jayesh, and N. Bala, "A Hybrid Deep Learning Approach for Detecting Anomalies in Real-Time Data Streams," in *2025 6th International Conference for Emerging Technology, INCET 2025*, 2025. doi: 10.1109/INCET64471.2025.11140026.

[39] L. Guerreiro, M. D. R. Bernardo, J. Martins, R. Gonçalves, and F. Branco, "Preliminary Research to Propose a Master Data Management Framework Aimed at Triggering Data Governance Maturity," *Lect. Notes Networks Syst.*, vol. 800, pp. 183 – 189, 2024, doi: 10.1007/978-3-031-45645-9_17.