JSSEIS

**Journal of Sustainable Software Engineering and Information Systems**

**Volume 1, Issue 1, 30-44.**
e_ISSN: xxxx-xxxx
https://e-journal.gomit.id/jsseis

# Understanding the Capabilities of Data Quality Measurement and Monitoring

**Normi Sham Awang Abu Bakar ***
International Islamic University Malaysia,
MALAYSIA

**Alea Syaffa Mohd Sabri**
International Islamic University Malaysia,
MALAYSIA

**Aisya Sufiah Hazlan**
International Islamic University Malaysia,
MALAYSIA

*** Corresponding author:**
Normi Sham Awang Abu Bakar, International Islamic University Malaysia, MALAYSIA. ✉email: nsham@iium.edu.my

**Article Info**

***Article history:***

Received: July 29, 2025
Revised: October 13, 2025
Accepted: October 19, 2025

***Keywords:***

Anomaly Detection
Data Governance
Data Quality
Data Quality Measurement
Data Quality Monitoring

**Abstract**

**Background of Study:** Ensuring high-quality data is essential for organizations that depend on analytics, automation, and regulatory compliance.
**Aims and Scope of Paper**: This paper explores the foundational concepts and evolving practices of two interrelated capabilities: data quality measurement and data quality monitoring. While measurement focuses on quantifying attributes such as accuracy, completeness, consistency, and timeliness, monitoring emphasizes the continuous detection and alerting of anomalies during data operations.
**Methods**: This paper examines the application of frameworks like Total Data Quality Management (TDQM), ISO 8000, and Data Management Association Data Management Body of Knowledge (DAMA DMBOK), alongside emerging tools such as rule-based engines, metadata-driven platforms, and AI-driven anomaly detection systems.
**Results**: Findings reveal a persistent gap in systems that integrate both measurement and monitoring effectively, hindering long-term data governance. This paper discusses a case study of the Data Quality framework implementation in the Healthcare sector. It was found that the healthcare organization implemented the Total Data Quality Management (TDQM) framework and Apache Griffin to ensure the accuracy, completeness, consistency, timeliness, and validity of clinical and IoT data through continuous monitoring, automated validation, and anomaly detection. Governance mechanisms aligned with ISO 8000 and HIPAA standards ensured full compliance, traceability, and accountability across all data quality and auditing processes. This study contributes to a deeper understanding of how integrated data quality practices can support digital transformation and operational resilience across industries.
**Conclusion:** The paper concludes by recommending the adoption of continuous quality measurement practices aligned with governance policies and supported by both human expertise and automation, arguing that data quality must be embedded as a dynamic and strategic function within the digital enterprise. While measurement emphasizes the quantification of data attributes such as accuracy, completeness, consistency, and timeliness, monitoring focuses on the continuous detection and alerting of data anomalies during data operations.

**To cite this article:** Awang Abu Bakar, N.S, et al (2025). Understanding the Capabilities of Data Quality Measurement and Monitoring. *Journal of Sustainable Software Engineering and Information Systems, 1*(1), 30-44.


## INTRODUCTION

The evolution of data as a strategic asset has significantly transformed how organizations design, operate, and manage their information systems. With increasing dependence on data to drive business intelligence, operational decisions, customer engagement, and regulatory compliance, the pressure to maintain high-quality data has never been more intense. Poor data quality leads to financial losses, misinformed decisions, and a degradation of trust in digital systems (Ehrlinger &

Wöß, 2022). Consequently, the disciplines of data quality measurement and data quality monitoring have emerged as critical pillars of data governance. In today's IT landscape, data quality is no longer the sole concern of database administrators. It now spans across departments, industries, and geographies, demanding organization-wide accountability.

Data quality measurement refers to the systematic process of assessing the state of data based on predefined dimensions such as accuracy, completeness, consistency, timeliness, and uniqueness (Ehrlinger & Wöß, 2022). This process involves deriving quantitative metrics that reflect how well a dataset conforms to business rules or expectations, forming the foundation for identifying quality issues, establishing baselines, and evaluating improvements (Bertossi & Geerts, 2020). As data grows in volume and complexity, IT systems from legacy databases and ERP platforms to real-time streaming analytics and AI pipelines require structured and scalable mechanisms to assess and ensure data quality.

In contrast, data quality monitoring is an ongoing, often automated process that continuously observes data as it flows through systems, raising alerts when anomalies or rule violations occur (Bertossi & Geerts, 2020). While measurement offers a snapshot of data quality at a specific point in time, monitoring ensures that data integrity is preserved during real-time operations. Together, these capabilities are essential not only for maintaining data trustworthiness but also for ensuring system reliability and compliance readiness.

Despite their complementary nature, measurement and monitoring often operate in silos. This fragmentation limits the organization's ability to achieve end-to-end data assurance and introduces challenges in governance and accountability. Many organizations still struggle to implement consistent quality checks or lack the tools and governance structures required for continuous quality assurance. Furthermore, while numerous frameworks and standards have been proposed over the years including Total Data Quality Management (TDQM), Data Management Association (DAMA) Data Management Body of Knowledge (DMBOK), and ISO 25012, translating these models into actionable, automated practices remains difficult in many operational contexts. Although many frameworks exist, few studies systematically integrate measurement and monitoring capabilities into an integrated governance framework.

Recent advancements in automation and machine learning are helping to close this gap. For instance, (Poon et al., 2021) introduced a semi-automated data quality control pipeline that uses unsupervised models like DBSCAN and KNN to detect anomalies in high-volume healthcare data, reducing manual oversight and increasing scalability. Similarly, modern tools such as dataquieR, pointblank, and assertr in R have emerged to support rule-based validation and profiling, yet many still lack essential features such as severity scoring, metadata integration, and intuitive interfaces for integrated workflows (Bertossi & Geerts, 2020).

This paper seeks to explore the interplay between data quality measurement and monitoring capabilities, especially within IT environments. It provides a critical overview of foundational frameworks, defines key quality dimensions and metrics, and identifies persistent challenges in implementation. It also highlights how emerging technologies, such as machine learning and real-time analytics, are reshaping the possibilities for integrated, scalable data quality assurance. Finally, the paper proposes a conceptual model for aligning measurement and monitoring functions within a unified data governance framework, thereby supporting more robust, efficient, and trustworthy information systems.

**Literatur Review**

One of the important aspects of data quality is measurements and metrics.

**Conceptual Frameworks and Theoretical Foundations**

At the heart of data quality measurement is the understanding that quality is not a singular attribute but a multidimensional construct. Early definitions, such as that proposed by Wang and Strong (1996) emphasized the notion of "fitness for use," suggesting that data must be evaluated in relation to the purpose for which it is being used. This contextual understanding has since become a defining characteristic of all major data quality frameworks. Figure 1 includes all the main activities in the

overall data quality frameworks and each individual activity can be further broken down into sub-activities to ensure the data correctness and integrity are preserved.


**Figure 1.** Data Quality Framework

Among the most widely cited frameworks is the Total Data Quality Management (TDQM) model, which adopts a cyclical approach to quality, defining, measuring, analysing, and improving. The TDQM framework emphasizes that quality must be integrated at every stage of the data lifecycle, from collection to consumption. This perspective aligns well with contemporary data pipeline architectures where data flows continuously through ingestion, transformation, storage, and visualization layers (Miller et al., 2025).

Another key framework is the DAMA DMBOK, which positions data quality as one of ten core functions in enterprise data management. DMBOK offers a practical structure for defining quality rules, assigning stewardship responsibilities, and establishing accountability. It also emphasizes the importance of aligning quality initiatives with business strategy, recognizing that data has value only when it supports real-world objectives (PLC, 2022).

In addition to these general-purpose frameworks, there are international standards such as ISO 8000 and ISO 25012, which formalize the definitions and categories of quality dimensions. ISO 8000 focuses on the characteristics of master data, promoting standardization in data exchange and system integration. ISO 25012 introduces a distinction between inherent and system-dependent quality, acknowledging that some aspects of quality are intrinsic to the data itself, while others depend on how it is stored, accessed, or maintained (Bertossi & Geerts, 2020).

Specialized domains often require additional adaptations. For example, the METRIC framework developed for medical AI datasets identifies critical quality dimensions such as bias, provenance, and representativeness (Mohammed et al., 2025). Similarly, the BCBS 239 principles in the banking sector focus on the accuracy, completeness, and timeliness of data used in risk reporting. These frameworks illustrate that while core quality principles are universal, their practical implementation must be tailored to domain-specific requirements.

Measuring data quality requires the conversion of abstract attributes into concrete, quantifiable indicators that can guide operations and strategic decision-making. Among the most widely assessed dimensions are accuracy, completeness, consistency, timeliness, uniqueness, and validity. Each dimension reflects a specific aspect of what makes data useful, trustworthy, and fit for its intended purpose.

Accuracy involves assessing the extent to which data values correctly describe the real-world entities they are intended to represent. For example, a customer's birthdate or billing amount should reflect actual, verifiable information (Mohammed et al., 2025). Completeness captures the presence or absence of required data fields. A record with all mandatory values filled in is deemed more complete than one with missing entries, and this has a direct impact on the reliability of analysis and reporting.

Consistency refers to the degree to which data remains uniform across systems. A product name that appears differently in inventory and sales systems introduces confusion and undermines confidence. Timeliness reflects the currency of data. Data that is accurate but outdated can result in flawed decisions, particularly in fast-paced environments such as stock trading or emergency response.

Uniqueness ensures that records do not duplicate entities already represented. Duplicate entries can lead to distorted insights, particularly in metrics such as customer counts or revenue forecasts. Validity, often enforced via schema rules or business constraints, checks that data conforms to expected formats. A phone number must match a defined digit pattern, and a date must fall within an acceptable range.

To operationalize these dimensions, organizations use a variety of metrics. For example, completeness might be assessed as the percentage of non-null fields across a dataset. Consistency may be evaluated through reconciliation between systems. Timeliness can be tracked by comparing update timestamps against required freshness thresholds. Many organizations visualize these metrics through dashboards and scorecards, allowing stakeholders to monitor changes over time and prioritize improvements accordingly (Miller et al., 2025).

The challenge, however, is that not all dimensions can be measured with equal precision or automation. Credibility and interpretability, for instance, often rely on human judgment. Even dimensions that are theoretically measurable may lack clear thresholds. What is considered "complete" in a customer service context may not meet regulatory expectations in a healthcare application. As such, organizations must establish context-sensitive benchmarks that align with business priorities and regulatory requirements (Miller et al., 2025; Schwabe et al., 2024).

**Challenges in Data Quality Management**

Despite the availability of frameworks and tools, the journey toward effective data quality measurement is fraught with challenges. One of the most significant is the inherent subjectivity in defining what constitutes "good enough" data. Different departments may have conflicting definitions of completeness or accuracy based on their needs. This lack of harmonization makes it difficult to implement organization-wide measurement standards (Schwabe et al., 2024).

The absence of ground truth complicates accuracy assessment. In many scenarios, such as sentiment analysis or predictive modelling, the "correct" value is unknown or constantly shifting. Organizations may rely on proxy data, expert reviews, or sampling techniques, but these introduce bias and reduce reliability. Moreover, without access to trusted reference data, even the most advanced validation tools cannot guarantee accuracy.

Data drift adds further complexity. In dynamic environments, data distributions may shift due to seasonal changes, user behaviour, or policy updates. A quality rule that was valid yesterday may no longer apply today. This necessitates continuous recalibration of rules and retraining of anomaly detection models. Static rules quickly become obsolete, and relying on them may create a false sense of security (Mohammed et al., 2025).

Scalability also poses significant hurdles. As data volumes grow into the petabyte range, measuring every record becomes computationally prohibitive. Organizations resort to sampling, which introduces the risk of missing critical edge cases. In streaming or event-based systems, the time available to validate each record is limited to milliseconds (Bertossi & Geerts, 2020).

Cultural and organizational factors exacerbate technical challenges. In many firms, data quality is seen as a technical issue to be addressed by IT, rather than a shared responsibility across business units. Without executive sponsorship, dedicated roles, and performance incentives, quality initiatives often lose momentum. Data stewards may be appointed but lack authority or resources to effect meaningful change. This misalignment between accountability and capability is a recurring theme in data quality literature (Precisely & University, 2024).

**The Role of Automation and Tooling**

In modern IT environments, the use of automation has become indispensable in the execution of data quality strategies. As data pipelines become more complex, the need for tools that can operate in real-time, detect issues as they occur, and remediate them without human intervention becomes critical. Automation enhances the speed, reliability, and consistency of data quality checks while reducing manual workload and the risk of human error (Miller et al., 2025).

A growing number of tools support automation in data profiling, validation, cleansing, and monitoring. Platforms like Talend, Informatica, Great Expectations, and Deequ (developed by AWS) provide users with the ability to define validation rules that are automatically applied to incoming data. These tools support features such as regular expression validation, duplication detection, range enforcement, null checks, and cross-field logic validation. In enterprise environments, they are often integrated into data lakehouses, ETL pipelines, and business intelligence platforms to ensure that poor-quality data does not contaminate downstream systems (Poon et al., 2021).

One notable capability in automated platforms is anomaly detection. Rather than relying solely on static rules, some tools use statistical models or machine learning to detect outliers or deviations from expected data patterns. This is especially useful when data variability is high, or business rules are difficult to define explicitly. For example, an automated system may learn that a certain customer typically makes 10 transactions per week and flag a spike to 500 as a potential anomaly. Such techniques can surface problems that would otherwise go undetected in manual reviews.

Automation also plays a key role in large-scale systems where data velocity is too high for manual validation. In industries like telecommunications, finance, or e-commerce, thousands of transactions may occur per second. Quality assurance mechanisms must keep up with this pace, which means rules must be optimized for performance and scalability. Stream processing frameworks such as Apache Kafka and Apache Flink are now being combined with quality engines that validate data on the fly before it is stored or consumed (Bertossi & Geerts, 2020).

Despite these advancements, automation is not a silver bullet. Automated systems are only as good as the rules, models, and logic they are based on. Poorly defined validation rules can generate false positives, eroding user trust in the tool. Overly strict rules can block legitimate data, while overly lax ones may allow issues to slip through. Therefore, even with automation, organizations must regularly audit their validation frameworks, tune their models, and update their rules based on feedback from business users and data analysts.

In addition, automated tools often struggle with semi-structured or unstructured data, such as free text, images, or audio. These data types are increasingly prevalent in fields such as healthcare, social media analytics, and customer feedback systems. Ensuring quality in these cases requires advanced techniques such as natural language processing, computer vision, or human-in-the-loop review systems. As the scope of data expands, automation must also evolve to address new formats and modalities.

The growing reliance on automation also brings governance implications. When automated systems act based on quality assessments, such as rejecting a data batch, suppressing an alert, or triggering an escalation. It becomes important to document the logic and maintain an audit trail. This transparency is essential not only for internal accountability but also for compliance with regulatory standards. Systems must not only act but explain their actions in a way that is understandable to stakeholders and auditors alike (Schwabe et al., 2024).

**Integration with Data Governance and Strategic Alignment**

Effective data quality measurement cannot succeed in a vacuum. It must be embedded within a broader data governance framework that defines roles, processes, and expectations for managing data across its lifecycle. Governance ensures that quality efforts are aligned with business objectives,

regulatory requirements, and accountability structures. Without governance, data quality becomes fragmented and reactive, rather than holistic and proactive (Precisely & University, 2024).

Organizations that treat data as an enterprise asset recognize the need for clear ownership. This involves assigning responsibilities to data stewards, custodians, and owners, who are empowered to maintain quality standards across domains. Measurement systems provide the evidence these roles need to monitor performance, enforce policies, and trigger remediation when necessary.

Modern governance platforms now integrate data quality measurement directly into their workflows. Solutions such as Collibra, Informatica, and Alation link business glossaries and metadata repositories to dashboards that visualize completeness, accuracy, and other dimensions. These tools allow stakeholders to trace quality issues to specific processes or systems. For example, a drop in data completeness may be linked to a malfunctioning upstream data feed or a new user interface that omits mandatory fields (Miller et al., 2025).

Regulatory compliance further strengthens the need for integrated measurement. Data privacy laws such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States require organizations to maintain accurate and timely personal data. Financial regulations, including BCBS 239, demand high levels of precision and auditability in risk reporting. In each case, data quality measurement serves as both a control and a compliance mechanism. It enables firms to document their efforts, demonstrate due diligence, and respond to audits or investigations with evidence-based reports (Fu et al., 2024).

In practice, the most successful organizations go beyond compliance. They embed quality metrics into their key performance indicators and executive dashboards. Business units track their own data quality scores, aligning performance with outcomes such as customer satisfaction, time-to-insight, or cost-to-serve. This creates a feedback loop where measurement leads to behavior change, and behavior change improves measurement outcomes.

Culture also plays a critical role. Organizations with a strong data culture foster shared ownership of data and recognize quality as a collective responsibility. They provide training on data literacy, create incentives for data stewardship, and reward proactive issue identification. Measurement in such environments becomes less of a burden and more of an enabler. It informs decisions, validates assumptions, and builds confidence across teams (Fu et al., 2024).

Nonetheless, many organizations still struggle with cultural resistance. Data governance is often seen as bureaucratic, and quality initiatives may be deprioritized in favour of speed or convenience. Overcoming this resistance requires leadership. Senior executives must advocate for quality, allocate resources, and model data-driven behaviour. When governance is supported from the top, and quality is measured transparently, organizations are more likely to achieve sustained improvement.

**Emerging Trends and Future Directions**

The landscape of data quality measurement is evolving rapidly, influenced by new technologies, regulatory developments, and shifting organizational needs. One of the most notable trends is the transition from periodic data audits to real-time monitoring. In the past, organizations would conduct quality assessments at fixed intervals, often quarterly or monthly, relying on static reports to detect issues. Today, the demand for real-time insights and operational agility means that data must be monitored continuously. This shift is being enabled by advances in cloud computing, event-driven architecture, and automated validation pipelines (Bertossi & Geerts, 2020; Mohammed et al., 2025).

Real-time monitoring allows for the early detection of anomalies and the immediate triggering of corrective actions. It supports use cases such as fraud detection, demand forecasting, and predictive maintenance, where the timeliness of data is essential to operational success. Platforms built on streaming technologies, such as Apache Kafka or Spark Streaming, now incorporate embedded

quality checks that validate data before it is stored, processed, or forwarded to downstream applications. This proactive approach minimizes the risk of flawed data entering decision-making systems and enhances confidence in real-time analytics.

Another key trend is the integration of data quality practices into DataOps, a methodology that applies DevOps principles to data workflows. In a DataOps environment, quality checks are treated as part of the software development lifecycle. Validation rules are version-controlled, tested automatically, and deployed alongside data transformation scripts. This enables continuous delivery of trusted data, reduces rework, and promotes collaboration between developers, analysts, and business users (Cai & Zhu, 2015).

As machine learning becomes more central to enterprise operations, organizations are also recognizing the importance of training data quality. Poor data not only compromises the accuracy of models but also introduces bias and undermines fairness. This has led to the rise of data-centric AI, a philosophy that prioritizes the quality of data over the complexity of algorithms. Data quality measurement in this context involves not only traditional metrics, such as completeness and accuracy, but also more nuanced considerations like representativeness, balance across classes, and labelling consistency (Buelvas et al., 2023).

Transparency in data quality processes is becoming increasingly important. Stakeholders want to understand not only whether data is reliable, but why it was deemed acceptable or rejected. This has given rise to explainable data quality systems that provide traceability, justifications for validation results, and user-friendly audit trails. These systems help build trust and support regulatory compliance, particularly in industries like healthcare and finance where decisions must be defensible (Mohammed et al., 2025).

Standardization is another area of active development. Despite the availability of frameworks such as ISO 8000, organizations often adopt different terminologies, rules, and metrics, leading to inconsistencies and duplication of effort. Industry groups are now working to harmonize best practices, define interoperable metrics, and promote shared vocabularies. These efforts aim to simplify tool integration, enhance benchmarking, and reduce the friction associated with multi-vendor environments (Miller et al., 2025).

Looking forward, the future of data quality measurement will be shaped by the convergence of automation, intelligence, and governance. Systems will become increasingly autonomous, capable of learning from feedback, adapting to new conditions, and orchestrating remediation without human intervention. Artificial intelligence will be used not only to detect anomalies but to predict where quality issues are likely to emerge based on historical trends, system changes, or user behavior. These predictive capabilities will allow organizations to shift from reactive correction to proactive prevention (Mohammed et al., 2025).

At the same time, ethical and regulatory considerations will demand stronger oversight. As data is used to make more consequential decisions, from credit approvals to medical diagnoses, the stakes of data quality will rise. Measurement systems will need to be both technically robust and ethically sound, balancing automation with accountability. Organizations will be expected to provide evidence that their data is accurate, unbiased, and responsibly managed.

In this evolving context, investment in data quality measurement is no longer optional. It is a prerequisite for innovation, trust, and resilience. Organizations that treat data quality as a living, strategic function will be better equipped to navigate complexity, seize opportunities, and protect their stakeholders.

## METHODOLOGY

Data quality monitoring capabilities refer to the technical and procedural mechanisms that enable organizations to observe, assess, and ensure the quality of data over time. These capabilities are essential for detecting anomalies, maintaining consistency, and ensuring that data remains fit for its intended use.

According to Miller et al., (2025), data quality monitoring involves not only measuring predefined quality metrics but also supporting continuous tracking of those metrics through automation, visualization, and alerts. This goes beyond one-time assessments by enabling scheduled evaluations, real-time validations, and trend analysis using tools such as Apache Griffin. Ehrlinger et al emphasize that effective monitoring requires aligning data quality metrics with business processes, including mechanisms for risk prioritization and loss event classification (Ehrlinger & Wöß, 2022).

In the context of big data, Gudivada et al. (2017) argues that traditional relational monitoring mechanisms are insufficient; instead, organizations must implement scalable systems that handle heterogeneous, voluminous, and dynamic data sources. Similarly, Poon et al. (2021) stresses that monitoring capabilities must be adaptive and dynamic, especially in high-velocity environments where data rapidly evolves.

Collectively, these studies underscore that data quality monitoring is a multi-faceted capability that integrates metric measurement, anomaly detection, historical tracking, and decision support to safeguard the long-term reliability and value of organizational data.

**Key Features of Data Quality Monitoring**

Data monitoring is a critical process in ensuring the quality, consistency, and usability of data over time. The following key features are essential in data quality monitoring capabilities, ensuring the effectiveness of the monitoring process. One such feature is task scheduling, which enables automated and regular execution of data quality checks, thus ensuring continuous assessment without manual intervention (Nguyen et al., 2025).

Another important capability is the storage of results, which allows organizations to retain historical data quality metrics for audit, compliance, and longitudinal analysis purposes (Miller et al., 2025). The retrieval of results further supports decision-making by enabling users to access and review past evaluations and diagnose recurring issues (Ehrlinger & Wöß, 2022).

Visualization over time is essential to detect trends, improvements, or degradation in data quality, allowing for proactive mitigation strategies (Miller et al., 2025). Finally, comparison of results across different time periods or rule versions is crucial for evaluating the impact of data quality initiatives and identifying areas for further improvement (Nguyen et al., 2025). These features collectively form the foundation for a robust data quality monitoring system that supports ongoing data governance and informed decision-making.

**Frameworks on Data Quality Monitoring Capabilities**

As organizations increasingly rely on large-scale, dynamic data ecosystems, the need for structured and scalable data quality monitoring (DQM) frameworks has become critical. These frameworks provide the architectural foundation and operational tools necessary to ensure continuous oversight of data integrity, supporting real-time validation, anomaly detection, and compliance with quality standards. This section reviews key DQM frameworks used in contemporary IT environments, focusing on their distinct capabilities and applications.

One of the most widely adopted frameworks is Apache Griffin, an open-source solution designed for both batch and streaming data quality monitoring. Integrated with big data platforms such as Apache Hadoop and Apache Spark, Griffin allows users to define data quality rules, perform real-time validation, and visualize quality metrics through a customizable dashboard. It emphasizes a model-driven approach that includes data profiling, rule-based validation, and temporal tracking of data quality metrics. These features make Apache Griffin particularly suitable for organizations operating in high-volume environments that require automated, end-to-end monitoring (Miller et al., 2025).

In the context of the Internet of Things (IoT), frameworks built around sensor fusion techniques are increasingly used to address the unique challenges of real-time sensor data monitoring. These frameworks combine data from multiple sensors to improve overall reliability and reduce discrepancies due to sensor drift or noise. Key components include sensor data aggregation, calibration monitoring, and data fusion algorithms such as Kalman filters or Bayesian inference. This

multi-layered approach is essential in industrial applications like smart factories, where decisions depend on precise, synchronized data streams (Segreto & Teti, 2023).

Another significant development in the field is the emergence of machine learning-based monitoring pipelines, which use unsupervised learning algorithms to identify anomalies without requiring predefined rules or labeled data. For example, Poon et al. (2021) proposed a semi-automated pipeline that integrates models like DBSCAN and K-Nearest Neighbors (KNN) to detect outliers in healthcare data. These pipelines are well-suited for complex, high-dimensional datasets where traditional rule-based validation may be insufficient. Additionally, they offer scalability and adaptability by learning patterns from the data itself, significantly reducing manual oversight and enabling more intelligent monitoring workflows.

Furthermore, big data monitoring frameworks are essential in distributed environments characterized by heterogeneous data sources, high velocity, and limited availability of gold-standard references. As noted by Poon et al. (2021), such frameworks must support real-time error detection, data integration across diverse systems, and methods for handling missing or inconsistent data. These are often implemented using distributed computing models and integrated with technologies like Apache Kafka or Apache NiFi to ensure monitoring is aligned with streaming and ETL processes.

Collectively, these frameworks demonstrate the evolving landscape of data quality monitoring. They provide organizations with the means to not only assess quality but to do so continuously and proactively, addressing the growing demands of data-driven operations.

**Benefits of Data Quality Monitoring Capabilities**

Implementing robust data quality monitoring capabilities provides significant benefits to organizations seeking to ensure reliable, timely, and actionable data. One of the foremost advantages is the early detection of anomalies and data integrity issues, allowing teams to proactively correct errors before they propagate into downstream systems or decision-making processes (Miller et al., 2025). Continuous monitoring ensures that data remains fit for purpose, supporting operational excellence, compliance, and customer satisfaction.

As highlighted by (Schwabe et al., 2024), monitoring is especially vital in big data and machine learning contexts, where poor data quality can lead to inaccurate model predictions, compliance failures, and financial losses. Furthermore, automated scheduling and visualization features allow stakeholders to track quality trends over time, aiding in the identification of recurring issues and the assessment of the effectiveness of data governance strategies.

Fast-changing big data environments, adaptive monitoring frameworks support real-time decision-making and ensure that data-driven processes are based on accurate and current information (Poon et al., 2021). Ultimately, data quality monitoring not only improves the trustworthiness and usability of data but also enhances the organization's ability to react to change, reduce risk, and drive innovation through analytics.

**Challenges in Comprehensive Data Quality Monitoring (CDQM)**

While Comprehensive Data Quality Monitoring (CDQM) plays a vital role in maintaining trustworthy and actionable data, its implementation is often fraught with technical, operational, and contextual challenges. These obstacles limit the scalability, usability, and effectiveness of monitoring systems, particularly in complex and dynamic data environments. The following points highlight the key challenges identified across academic and practical studies on CDQM:

1. Lack of Standardized and General-Purpose Metrics
   Many data quality tools fail to implement a wide range of standardized metrics for various quality dimensions (e.g., accuracy, completeness, timeliness). Most rely on domain-specific metrics, which hinders comparability and scalability (Miller et al., 2025).
2. Technical Complexity and Integration Overhead

Tools like Apache Griffin require complex setups involving multiple systems (e.g., Hadoop, Spark, Elasticsearch), which increases the barrier to adoption and limits their use to technically mature environments (Miller et al., 2025).

3. Inadequate Support for Unstructured or Heterogeneous Data
   CDQM frameworks often struggle with processing and validating unstructured or multi-source data, which is increasingly common in big data environments (Poon et al., 2021; Schwabe et al., 2024).
4. Scalability and Real-Time Monitoring Limitations
   While many frameworks support batch monitoring, real-time and streaming data quality checks are less developed, limiting their usefulness in dynamic and high-velocity environments (Schwabe et al., 2024; Buelvas et al., 2023).
5. High Human Dependency in Model Tuning and Validation
   Despite the use of automation and unsupervised techniques, human experts are still required to interpret results, validate anomalies, and fine-tune detection models (Buelvas et al., 2023).
6. Lack of Visualization and User-Friendly Interfaces
   Several tools either lack intuitive dashboards or only offer limited visual analytics, which makes it difficult for stakeholders to interpret trends and monitor quality metrics effectively (Poon et al., 2021).
7. Difficulty in Defining "Fitness for Use" Across Contexts
   The subjective and context-dependent nature of data quality complicates the development of universally applicable monitoring strategies.
8. Limited Automation in Data Profiling and Dependency Discovery
   Many tools provide basic data profiling but do not offer deep profiling capabilities (e.g., multi-column analysis, rule discovery), which are crucial for comprehensive monitoring.
9. Challenges in Monitoring Dynamic or Evolving Schemas
   NoSQL and schema-less systems in big data introduce additional challenges, as data structures change frequently, making predefined quality rules ineffective or obsolete.

**Future Trends in Data Quality Monitoring**

As data becomes increasingly central to business intelligence, regulatory compliance, and operational decision-making, the role of data quality monitoring (DQM) is evolving rapidly. No longer confined to static checks or manual validations, DQM is transitioning into a dynamic and intelligent discipline that leverages modern technologies to support continuous oversight, adaptability, and scalability. Several key trends are emerging that are expected to shape the future of DQM capabilities, particularly in the context of big data, artificial intelligence, and cloud-native infrastructures.

One of the most transformative developments in data quality monitoring is the integration of artificial intelligence (AI) and machine learning into anomaly detection and predictive monitoring. Traditional DQM systems depend heavily on predefined rules and thresholds, which may fail to detect novel or context-specific errors. In contrast, AI-driven systems use unsupervised learning algorithms such as DBSCAN, Isolation Forests, and autoencoders to detect anomalies in complex and high-volume datasets. For instance, Poon et al. (2021) introduced a semi-automated pipeline capable of identifying outliers in healthcare data without the need for labeled input. These predictive models are expected to become more common, enabling organizations to shift from reactive to proactive data quality assurance.

Martin et al presented a solution for assessing several quality dimensions of IoT data streams as they are generated. Additionally, the solution described in the paper actually improves the quality of data streams by curating them through the application of Artificial Intelligence techniques (Cortes et al., 2024).

Another significant trend is the growing demand for real-time monitoring in response to the rise of streaming data and event-driven architectures. Frameworks like Apache Griffin already support real-

time data validation as information flows through ingestion pipelines (Buelvas et al., 2023). Future DQM systems are expected to build on this foundation by offering sub-second latency, integration with edge devices in Internet of Things (IoT) environments, and dynamic dashboards that display quality metrics as they are generated. This will be particularly important in sectors such as finance, healthcare, and manufacturing, where immediate action is required when anomalies occur.

In addition to automation, the demand for explainability and interpretability in DQM systems is increasing. As machine learning becomes more integrated into monitoring frameworks, organizations must be able to understand and justify how certain data points are flagged as errors. Explainable AI will become a standard feature in future DQM solutions, offering transparent reasoning for anomaly detection and clear audit trails for decision-making processes. This need is particularly acute in regulated industries where transparency and traceability are essential (Cortes et al., 2024).

The future of DQM also lies in the use of metadata-driven and context-aware systems. Many current tools lack comprehensive integration with metadata repositories, limiting their ability to interpret data in context. Emerging frameworks will increasingly leverage data catalogs, semantic models, and schema registries to apply validation rules that are context-sensitive and adaptable to structural changes. According to (Poon et al., 2021), this capability is essential for improving the precision and responsiveness of monitoring systems, especially in dynamic data environments.

As cloud adoption accelerates, DQM tools will also evolve to support cloud-native and highly scalable architectures. These future tools will be optimized for containerized environments, serverless processing, and multi-cloud compatibility, allowing organizations to implement quality monitoring across distributed systems with minimal operational overhead (Miller et al., 2025). In addition, federated monitoring frameworks will become increasingly important in privacy-sensitive domains such as healthcare, where centralized data access is restricted. These frameworks allow for local data validation while maintaining centralized oversight, preserving data privacy without compromising quality control (Cortes et al., 2024).

Another critical trend is the integration of DQM with DataOps practices and continuous integration and deployment (CI/CD) pipelines. This integration will ensure that data quality checks are applied consistently across development, testing, and production environments. By embedding monitoring within the broader data lifecycle, organizations can adopt a "quality by design" approach that reduces error rates and promotes a culture of accountability and continuous improvement (Ehrlinger & Wöß, 2022).

In conclusion, the future of data quality monitoring is moving toward greater automation, intelligence, and integration. The convergence of machine learning, real-time analytics, metadata awareness, and cloud scalability is transforming DQM from a reactive checkpoint into a proactive, adaptive, and strategic function. As data continues to grow in complexity and importance, these trends will play a pivotal role in ensuring that organizations can maintain trustworthy, high-quality information that supports reliable and ethical decision-making (Batini et al., 2024).

## RESULT AND DISCUSSION

### Result:

This section highlights the implementation of the DQ measurement and monitoring in a Healthcare Analytics Platform.

### Framework Implementations

The healthcare organization adopted the Total Data Quality Management (TDQM) framework to evaluate core quality dimensions, including accuracy, completeness, consistency, timeliness, and validity (Martín et al., 2023). Measurement activities were focused on defining and operationalizing key metrics relevant to clinical and IoT-generated data. For instance, completeness was quantified as the percentage of patient records containing all mandatory fields, such as allergy history and demographic details. Accuracy was determined by matching patient identifiers against a centralized master reference index, thereby minimizing duplication and ensuring accurate patient linkage.

Timeliness was assessed as the elapsed time between IoT data capture and its ingestion into the enterprise analytics platform. These measurements were continuously monitored through interactive dashboards, which facilitated historical trend analysis and supported compliance auditing (Martín et al., 2023).

To complement measurement activities, the organization implemented Apache Griffin, an open-source data quality framework, enabling real-time validation within data ingestion pipelines. Beyond rule-based validation, the monitoring system integrated unsupervised anomaly detection models, including Isolation Forest and DBSCAN, to detect irregularities in patient vital signs streamed from IoT sensors (Kahn et al., 2012). These models allowed the detection of deviations that could signal device malfunction or abnormal physiological events. Additionally, automated alerting mechanisms were incorporated into clinical workflows to notify healthcare professionals when anomalies exceeded predefined thresholds, thereby promoting timely interventions and minimizing undetected data quality issues (Zhu et al., 2017).

Governance and compliance requirements were embedded throughout the framework, ensuring alignment with internationally recognized standards such as ISO 8000 for master data quality and HIPAA for data privacy and security (Miller et al., 2025). Comprehensive audit logs were maintained for all validation processes, anomaly detection outcomes, and corrective actions undertaken. These records facilitated regulatory audits and ensured traceability across the data quality lifecycle, reinforcing accountability and compliance with legal obligations (Nguyen et al., 2025).

**Outcomes and Benefits**

The implementation of this integrated data quality framework yielded significant improvements in data integrity and operational performance. One key achievement was the 90% reduction in duplicate patient records, which was facilitated by enhanced accuracy in patient identifier reconciliation across multiple systems. This improvement mitigated risks associated with fragmented medical histories and clinical errors. Moreover, the deployment of machine learning-based anomaly detection contributed to a 75% reduction in false alerts generated by IoT devices, enhancing the reliability of real-time patient monitoring and reducing the incidence of clinician alert fatigue (Nguyen et al., 2025). Regulatory compliance was substantially strengthened through the adoption of ISO 8000 standards and HIPAA-aligned governance mechanisms. The systematic maintenance of audit logs provided demonstrable evidence of compliance, enabling the organization to meet the stringent requirements of healthcare regulators (Martín et al., 2023). Additionally, the use of interactive dashboards to visualize key quality metrics facilitated greater transparency and accountability, enabling stakeholders to proactively monitor performance trends. Collectively, these benefits not only ensured compliance and operational efficiency but also enhanced clinician confidence in analytics-driven decision-making, thereby supporting improved patient care outcomes.

**Discussion:**

**Lessons Learned**

This case study illustrates the strategic value of integrating data quality measurement and monitoring capabilities within a single governance-oriented framework. While measurement provided the foundational benchmarks for quality evaluation and compliance, continuous monitoring enabled dynamic responsiveness to emerging anomalies in high-velocity data streams. The incorporation of machine learning models was particularly effective in addressing the complexity of IoT sensor data, where static rule-based systems would have been insufficient.

Furthermore, the implementation demonstrated that while automation reduces manual intervention, governance mechanisms remain critical to ensuring transparency, accountability, and interpretability of quality-related decisions. Audit trails, compliance reporting, and contextual interpretation of anomalies provided the necessary safeguards for both operational and regulatory assurance. Finally, aligning technological capabilities with established healthcare regulatory frameworks and clinical workflows proved essential for sustainable adoption. By embedding data

quality controls within routine operations, the organization achieved a robust and scalable approach to data governance, enhancing both compliance and patient-centric care.

Implications:
The findings of this study provide practical implications for organizations aiming to strengthen their data governance and quality assurance mechanisms. By integrating data quality measurement and monitoring within a unified framework, enterprises can achieve proactive control over data integrity, compliance, and operational reliability. For the healthcare sector, this integration enables real-time validation of critical patient data, reducing errors and supporting better clinical decision-making. From a managerial perspective, embedding data quality processes into governance structures ensures sustainable accountability and continuous improvement across departments.

Research Contribution:
This research contributes to the literature by presenting an empirical case study that demonstrates how Total Data Quality Management (TDQM) and Apache Griffin can be jointly implemented to achieve continuous data quality improvement in a real-world healthcare setting. It bridges the gap between theory and practice by showcasing how rule-based and AI-driven anomaly detection techniques can coexist under governance-oriented frameworks. The study also extends the understanding of data quality capabilities by highlighting their operational, technical, and organizational interdependencies.

Limitations:
Despite its contributions, this study has several limitations. First, the findings are based on a single case study within the healthcare industry, which may limit generalizability to other domains. Second, the research primarily focuses on technical and governance aspects, with limited attention to cultural or behavioral dimensions of data quality management. Additionally, the evaluation relied on organizational performance indicators rather than long-term outcome measures, suggesting a need for longitudinal analysis in future research.

Suggestions:
Future research should explore comparative analyses across multiple industries to identify sector-specific best practices in data quality measurement and monitoring. Expanding the study to include user perception, data literacy, and organizational culture could offer a more holistic understanding of implementation challenges. Moreover, future studies could examine how emerging technologies such as federated learning, blockchain, and explainable AI can enhance the transparency, traceability, and trustworthiness of automated data quality frameworks.

## CONCLUSION

Data quality measurement has become one of the defining capabilities of contemporary information systems. As data increasingly serves as the foundation for strategic decision-making, operational processes, and regulatory compliance, organizations can no longer afford to treat quality as a secondary concern. Instead, they must view it as a continuous, organization-wide responsibility that is integral to maintaining trust, achieving efficiency, and ensuring ethical and legal accountability.

Measuring and monitoring data quality is not just about identifying what is wrong with the data. It is about creating conditions in which data can be trusted, used confidently, and continuously improved. Measurement provides a benchmark, while monitoring ensures ongoing vigilance and responsiveness. Together, they are the mechanisms by which assumptions are tested, gaps are discovered, and governance becomes actionable. They are both technical processes and cultural signals that demonstrate an organization values data as a strategic resource and is willing to invest in its integrity.

In conclusion, data quality measurement and monitoring are not one-time exercises. They are evolving disciplines that must adapt to new technologies, regulatory environments, and business goals. They demand attention not only to metrics and tools but to people, processes, and purpose.

When treated as strategic and integrated functions, they become force multipliers, enhancing every downstream activity that depends on reliable information. In the years ahead, their relevance will only increase, making them among the most vital components of any successful IT and data strategy.

## ACKNOWLEDGMENT

This research was supported by the International Islamic University Malaysia (IIUM) under the Kulliyyah of Information and Communication Technology (KICT). The authors gratefully acknowledge this institutional support and the valuable suggestions provided by the reviewers and editors of the *Journal of Sustainable Software Engineering and Information Systems (JSSEIS)* during the review process

## AUTHOR CONTRIBUTION STATEMENT

NS edited the draft and wrote the final copy, ASM and ASH did the research and wrote the draft.

## REFERENCES

Batini, C., Scannapieco, M., & Rula, A. (2024). From data quality to data value: A new paradigm for data-driven organizations. *Information Systems Frontiers*, *26*(2), 445–462. https://doi.org/10.1007/s10796-023-10426-5

Bertossi, L., & Geerts, F. (2020). Data Quality and Explainable AI. *J. Data and Information Quality*, *12*(2), 1–9. https://doi.org/10.1145/3386687

Buelvas, J. H., Múnera, D., & Gaviria, N. (2023). DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems. *Internet of Things*, *22*(100769), 1–18. https://doi.org/10.1016/j.iot.2023.100769

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*, 2. https://doi.org/10.5334/dsj-2015-002

Cortes, C., Sanz, C., Etcheverry, L., & Marotta, A. (2024). Data Quality Management for Responsible AI in Data Lakes. *VLDB Workshops*. https://api.semanticscholar.org/CorpusID:273878300

Ehrlinger, L., & Wöß, W. (2022). A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, *5*, 1–30. https://doi.org/10.3389/fdata.2022.850611

Fu, Q., Nicholson, G. L., & Easton, J. M. (2024). Understanding data quality in a data-driven industry context: Insights from the fundamentals. *Journal of Industrial Information Integration*, *42*, 100729. https://doi.org/10.1016/j.jii.2024.100729

Gudivada, V. N., Apon, A., & Ding, J. (2017). Data Quality Considerations for Big Data and Machine Learning : Going Beyond Data Quality Considerations for Big Data and Machine Learning : Going Beyond Data Cleaning and Transformations. *International Journal on Advances in Software*, *10*(1), 1–20. https://www.researchgate.net/publication/318432363_Data_Quality_Considerations_for_Big_Data_and_Machine_Learning_Going_Beyond_Data_Cleaning_and_Transformations

Kahn, M. G., Raebel, M. A., Glanz, J. M., Riedlinger, K., & Steiner, J. F. (2012). A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Medical Care*, *50*(1), 28–39. https://doi.org/10.1097/MLR.0b013e318257dd67

Martín, L., Sánchez, L., Lanza, J., & Sotres, P. (2023). Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation. *Internet of Things*, *22*(April), 100779. https://doi.org/10.1016/j.iot.2023.100779

Miller, R., Hin, S., Chan, M., & Whelan, H. (2025). A Comparison of Data Quality Frameworks : A Review. *Big Data Cogn. Comput.*, *9*(93), 4–13. https://doi.org/10.3390/bdcc9040093

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, *132*, 102549. https://doi.org/10.1016/j.is.2025.102549

Nguyen, T., Nguyen, H.-T., & Nguyen-Hoang, T.-A. (2025). Data quality management in big data: Strategies, tools, and educational implications. *Journal of Parallel and Distributed Computing*, *200*, 105067. https://doi.org/10.1016/j.jpdc.2025.105067

PLC, E. (2022). *Global data management research: Data quality and business performance*. Experian

PLC.
Poon, L., Farshidi, S., Li, N., & Zhao, Z. (2021). Unsupervised Anomaly Detection in Data Quality Control. *In Proceedings of the 2021 IEEE International Conference on Big Data*, 2326–2336. https://doi.org/10.1109/BigData52589.2021.9671672
Precisely, & University, D. (2024). *Data quality and governance survey: The data quality confidence gap*. Precisely.
Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ Digital Medicine*, *7*(1), 203. https://doi.org/10.1038/s41746-024-01196-4
Segreto, T., & Teti, R. (2023). Data quality evaluation for smart multi-sensor process monitoring using data fusion and machine learning algorithms. *Production Engineering*, *17*(2), 197–210. https://doi.org/10.1007/s11740-022-01155-6
Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://web.mit.edu/tdqm/www/tdqmpub/beyondaccuracy_files/beyondaccuracy.html
Zhu, X., Xu, W., & Wang, Q. (2017). *Apache Griffin: An open-source framework for data quality solutions BT - Proceedings of the IEEE International Conference on Big Data*. 655–664. https://doi.org/10.1109/BigData.2017.8258015