Volume-09 Issue 07, July-2025 ISSN: 2456-9348 (Online)
SJIF Impact Factor: 8.232

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

A FRAMEWORK FOR AUTOMATED DATA QUALITY ASSESSMENT AND ANOMALY DETECTION IN ENTERPRISE DATA WAREHOUSING

Ajay Kumar Kota
IT business systems, BI platforms and data architect analyst
Apellis Pharmaceuticals, Apellis, 100 5th Ave, Waltham, MA 02451

ABSTRACT
Enterprise data warehouses serve as centralized repositories for integrating, transforming, and analyzing data from heterogeneous operational systems. Their reliability directly influences business intelligence (BI), organizational reporting, strategic planning, and data-driven decision-making. However, enterprise data pipelines frequently encounter data quality problems, including missing values, duplicate records, inconsistent formats, invalid entries, unexpected distribution changes, and violations of business rules. Traditional data validation approaches generally rely on predefined constraints and manual inspection, which may be insufficient for identifying complex or previously unseen anomalies in large-scale analytical environments. Consequently, automated and scalable data quality management has become an important requirement for modern enterprise data architecture. This research proposes an integrated framework for automated data quality assessment and anomaly detection in enterprise data warehousing. The framework combines data profiling, rule-based validation, statistical analysis, and machine-learning-based anomaly detection to identify and classify potential data quality issues throughout the data integration lifecycle. The proposed architecture consists of data ingestion, preprocessing, quality assessment, anomaly detection, severity classification, automated alerting, and validation reporting components. Data quality is assessed using dimensions such as completeness, validity, consistency, uniqueness, and conformity to predefined business constraints. Statistical or machine-learning methods may subsequently identify unusual patterns that cannot be detected through conventional validation rules alone. The framework is designed for integration with enterprise data warehouses, including SQL Server and Snowflake, and BI platforms such as Power BI. Its effectiveness should be evaluated through comparative experiments using established baseline methods, representative datasets, and appropriately verified anomaly labels.

Keywords:
Enterprise Data Warehousing, Automated Data Quality Assessment; Anomaly Detection; Data Profiling; ETL Automation; Data Governance; Machine Learning.

1. INTRODUCTION

Enterprise data warehousing has become a fundamental component of modern information systems, enabling organizations to consolidate data from transactional databases, cloud applications, customer relationship management platforms, and operational systems. By integrating information from multiple sources, data warehouses support historical analysis, performance monitoring, regulatory reporting, and business intelligence. However, the usefulness of these systems depends not only on their storage and processing capabilities but also on the quality of the underlying data.

An integrated approach could help organizations identify quality issues earlier, prioritize serious anomalies, and generate consistent audit records. Combining metadata-driven rules with statistical methods may also reduce dependence on manually maintained validation procedures. However, the effectiveness of this combination must be evaluated rather than assumed. The proposed contribution is an architecture that connects data profiling, rule-based validation, anomaly detection, and operational monitoring within a unified workflow. Its potential value lies in providing a repeatable approach to data quality management across heterogeneous enterprise systems. The study will investigate whether this integration improves detection effectiveness compared with conventional rule-based validation and under what conditions the framework remains computationally practical.

1.1 Data quality assessment

Data quality research identifies several dimensions that determine whether information is suitable for its intended purpose. Wang and Strong (1996) established an influential user-oriented perspective on data quality, emphasizing that quality extends beyond technical correctness. For enterprise data warehousing, relevant dimensions include completeness, consistency, validity, uniqueness, timeliness, and accuracy. These dimensions require different

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

assessment techniques: missing-value analysis can assess completeness, duplicate detection can measure uniqueness, and business constraints can evaluate validity. Accuracy may require comparison with authoritative reference data because a value can satisfy formatting rules while still being incorrect.

1.2 Rule-based validation

Rule-based approaches use explicit constraints, reference tables, ranges, mandatory fields, and relationships between records. SQL constraints and ETL validation scripts provide practical mechanisms for implementing these checks. Their principal advantage is interpretability: a failed rule can usually be traced to a defined business requirement. However, rule-based methods may struggle with previously unseen anomalies and relationships that vary across time, departments, or data sources. Maintaining large rule libraries may also introduce operational complexity.

1.3 Anomaly detection

Anomaly detection identifies records or patterns that deviate from expected data behaviour. In enterprise data warehousing, anomalies may result from data entry errors, duplicate records, system failures, or inconsistencies during data integration. Statistical techniques detect outliers using thresholds, data distributions, and historical baselines, while machine-learning methods identify complex patterns and unusual relationships. However, detection accuracy depends on data characteristics, feature selection, and parameter settings. Excessive sensitivity may generate false alerts, whereas restrictive thresholds may overlook genuine anomalies. Therefore, combining statistical rules, machine-learning techniques, and domain-specific validation can improve anomaly detection, support continuous data quality monitoring, and enable timely identification and correction of data irregularities.

2. PROPOSED FRAMEWORK ARCHITECTURE

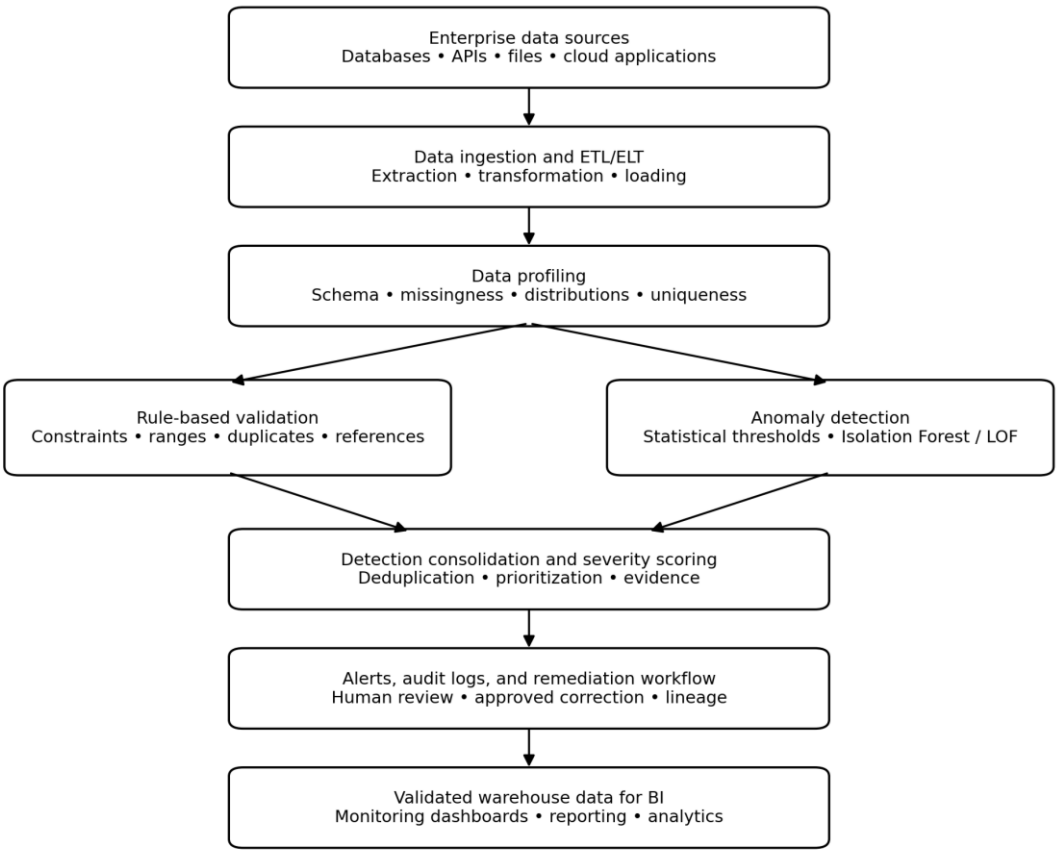


Figure 1: Framework Architecture

2.1 Architectural Overview

The proposed framework integrates automated data quality assessment and anomaly detection across the enterprise data warehousing lifecycle. Its architecture consists of interconnected modules for data ingestion, profiling, rule-based validation, anomaly detection, severity classification, remediation, and governance. These

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

modules work together to identify data quality issues before they affect analytical outputs and business decisions. The process begins with data collection from relational databases, enterprise applications, flat files, and cloud-based platforms. The ingested data undergoes profiling and validation to identify missing values, duplicates, structural inconsistencies, and unexpected patterns. The anomaly detection engine subsequently examines records and aggregated statistics to identify irregularities that predefined rules may not capture. Detected issues are classified according to their severity and potential business impact.

The framework then generates alerts, maintains audit records, and supports appropriate remediation activities. Corrected data can be reprocessed and reassessed before being released for downstream analytics. Governance components maintain metadata, access controls, and lineage information throughout the process. This modular architecture enables individual components to be updated independently as data sources, business requirements, and analytical methods evolve. It also supports scalability and integration with existing enterprise systems. The overall objective is to improve data reliability, reduce manual validation efforts, and provide a consistent mechanism for continuous data quality monitoring.

2.2 Data Ingestion and Profiling

The data ingestion and profiling module establishes the foundation for reliable data quality assessment. It collects information from heterogeneous sources, including relational databases, enterprise applications, spreadsheets, flat files, and cloud data platforms. Data can be transferred through existing Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) pipelines, depending on the organization's infrastructure and processing requirements.

Following ingestion, the profiling component examines the structural and statistical characteristics of each dataset. Key profiling activities include identifying data types, missing values, duplicate records, distinct-value counts, numerical distributions, and inconsistencies in field formats. Relationships between tables are also evaluated to identify broken references and potential violations of referential integrity.

Profiling results provide a baseline for subsequent validation and anomaly detection. For example, an unexpected increase in missing customer identifiers or a substantial change in daily transaction volumes may indicate a data integration problem. Comparing current profiles with historical observations can help identify changes that require further investigation. The module should record profiling metrics, execution timestamps, dataset identifiers, and relevant schema information. Incremental profiling may be used for frequently updated datasets to reduce computational overhead. By continuously monitoring incoming data, this module enables early identification of structural changes and emerging quality problems before they propagate into downstream warehouse tables and business intelligence reports.

2.3 Rule-Based Quality Assessment

The rule-based quality assessment module evaluates datasets against predefined business, technical, and operational requirements. It identifies violations that can be expressed through explicit validation conditions. These conditions may include mandatory-field completeness, uniqueness of business keys, acceptable numerical ranges, valid date formats, referential integrity, and consistency between related attributes. For example, a customer identifier may be required to be non-null and unique, while a transaction date must fall within an acceptable business period. Similarly, the total transaction amount may be checked against its constituent line-item values. Such rules help detect common data quality problems systematically and consistently across multiple warehouse tables.

Rules should be maintained in a configurable metadata repository rather than embedded permanently within application code. This approach allows authorized users to modify validation criteria as business requirements change. Each rule should include a unique identifier, description, target dataset, validation condition, severity level, and applicable version. Maintaining version history supports traceability and enables analysts to determine which rules were active when a particular issue was detected. The assessment engine should generate structured results containing the number of records examined, the number of violations, and the affected fields or records. These results can be aggregated into data quality scores for monitoring trends. Combining rule-based validation with automated profiling enables organizations to identify known data defects consistently while establishing a foundation for more advanced anomaly detection.

2.4 Anomaly Detection Engine

The anomaly detection engine identifies unusual observations and changes in data behaviour that may not violate explicitly defined validation rules. This module complements rule-based assessment by examining statistical deviations, relationships among variables, and changes in historical data patterns. Depending on the dataset and detection objective, the engine may employ statistical thresholds, robust measures of dispersion, or machine-learning algorithms. Statistical approaches can identify unexpected changes in transaction volumes, missing-value

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

proportions, or numerical distributions. Isolation Forest may be suitable for detecting unusual observations in multidimensional datasets, whereas Local Outlier Factor can identify records that differ from their local neighbourhoods.

Feature selection is important because irrelevant or poorly scaled variables may reduce detection quality. The engine should therefore apply appropriate preprocessing, handle missing values, and consider categorical and numerical attributes according to the selected algorithm. Detection thresholds and model parameters should be configured using representative data and validated against known or reviewed anomalies wherever possible. Each detected anomaly should be accompanied by supporting information, including the affected dataset, relevant features, detection method, score or threshold, and execution timestamp. This information helps analysts understand why an observation was flagged. The engine should also support periodic evaluation to identify changes in data behaviour that may require model adjustment. By combining complementary detection techniques, the framework can identify both known violations and previously unrecognized irregularities.

2.5 Severity Classification and Alerting

The severity classification and alerting module prioritizes detected data quality issues according to their potential impact on business operations and analytical outcomes. Not every anomaly requires immediate intervention; therefore, an effective prioritization mechanism helps data teams focus on the most consequential problems. Severity classification may consider factors such as the number of affected records, the criticality of the dataset, the importance of the affected attributes, the persistence of the issue, and its potential effect on downstream reports. For instance, missing values in a critical financial identifier may warrant urgent attention, whereas a minor deviation in a nonessential descriptive field may receive a lower priority.

The framework can categorize issues into levels such as critical, high, medium, and low using predefined organizational criteria. Alerts should include the issue description, affected dataset, detection time, severity level, relevant validation rule or detection method, and supporting evidence. Notifications may be delivered through monitoring dashboards, email, or existing incident management systems. To prevent excessive notifications, the module should support alert grouping, duplicate-alert suppression, and configurable escalation policies. Repeated occurrences of the same issue may be consolidated into a single incident while preserving individual detection records. Alert histories should also capture acknowledgement, investigation, and resolution status. These capabilities improve transparency, support timely responses, and enable organizations to monitor recurring quality problems and evaluate the effectiveness of their data quality management processes.

2.6 Remediation and Feedback

The remediation and feedback module manages the investigation, correction, and reassessment of detected data quality issues. Its primary objective is to ensure that identified problems are addressed systematically and that corrected records meet the required validation criteria before being used in downstream analytical processes.

Remediation actions may include quarantining invalid records, requesting corrections from source-system owners, rerunning failed ETL or ELT processes, and updating transformation logic. The appropriate action depends on the nature of the defect, the reliability of available information, and the potential consequences of modifying the data. For example, a record containing an invalid date may require source-system verification rather than automatic replacement with an estimated value. Automated correction should be restricted to predefined, validated situations where the intended modification is unambiguous. Uncertain cases should be routed to authorized personnel for review. The framework should preserve original values, record applied changes, and maintain an audit trail identifying the responsible process or user.

After remediation, affected records should pass through the relevant validation and anomaly detection stages again. This feedback loop confirms whether the corrective action resolved the issue and helps identify recurring failures. Reviewed anomalies can also contribute to improved detection rules and model evaluation, subject to appropriate quality controls. Monitoring remediation time, recurrence rates, and unresolved incidents enables organizations to assess operational effectiveness. Through controlled correction and continuous feedback, the framework supports progressive improvement in enterprise data quality.

2.7 Governance Integration

The governance integration module ensures that data quality assessment and anomaly detection operate within established organizational policies for data management, security, and accountability. It connects the proposed framework with metadata catalogs, data lineage systems, access control mechanisms, and existing monitoring infrastructure. Metadata integration provides information about dataset ownership, business definitions, schema structures, and validation requirements. This information helps associate detected issues with the appropriate business context and responsible teams. Data lineage records show how information moves from source systems

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

through transformation processes into warehouse tables and analytical outputs. Such traceability helps investigators identify where a defect originated and determine which downstream assets may be affected.
Access controls should restrict sensitive datasets and remediation operations to authorized users and services. The framework should collect only the information required for quality assessment and retain diagnostic records according to organizational policies. Audit logs should document rule changes, model configuration updates, remediation actions, and issue resolution. Implementation may use SQL Server or Snowflake for warehouse storage and SQL-based validation, Python libraries for statistical analysis and machine-learning models, and business intelligence dashboards for reporting quality metrics. These technologies are possible implementation choices rather than mandatory requirements. Integration should be guided by existing infrastructure, workload demands, security constraints, and maintenance capabilities. By embedding quality monitoring within governance processes, the framework supports accountability, reproducibility, and consistent application of data management policies. This integration also enables organizations to evaluate quality trends across datasets and align technical monitoring activities with broader business requirements.

3. RESEARCH METHODOLOGY

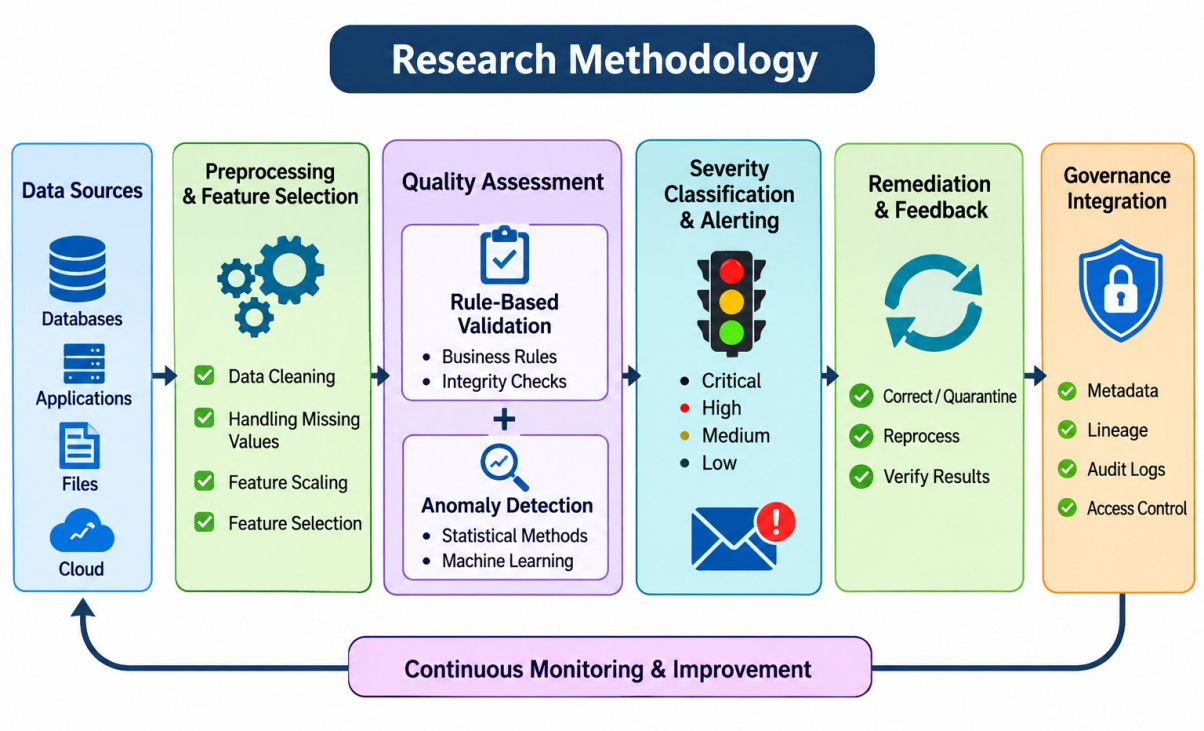


Figure 2: Research Design

3.1 Research Design

The study adopts a design-and-evaluation methodology to develop and assess an integrated framework for automated data quality assessment and anomaly detection. The research involves framework design, experimental implementation, and comparative performance evaluation. Initially, the framework architecture and validation procedures will be defined. Subsequently, individual modules will be implemented and tested in a controlled environment. The integrated framework will then be evaluated against independent approaches to determine its effectiveness. Three configurations will be considered: rule-based validation, anomaly detection, and the combined framework. Performance comparisons will focus on detection accuracy, computational efficiency, and operational reliability. The experimental design will ensure that all configurations are evaluated using equivalent datasets and documented settings. Repeated experiments will help assess the consistency of results. This methodology enables the investigation of individual component contributions while examining the benefits and limitations of combining predefined quality rules with data-driven anomaly detection techniques in enterprise data warehousing environments.

3.2 Dataset Selection

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

The study will use representative datasets reflecting common enterprise data characteristics, including customer information, transaction records, inventory tables, and operational metrics. Publicly available datasets may support reproducibility, while anonymized organizational datasets can provide practical relevance. Dataset selection will consider size, attribute diversity, data types, missing-value prevalence, and the availability of known quality issues. Ground-truth labels will be established using verified historical incidents, expert-reviewed records, or carefully generated synthetic anomalies. Synthetic corruption may include duplicate records, missing values, invalid entries, inconsistent references, and unexpected distribution changes. Each dataset will be documented with its record count, attribute descriptions, and anomaly characteristics. The proportion of normal and anomalous observations should also be reported. Using multiple datasets can help determine whether the proposed framework performs consistently across different data structures and operational scenarios. Privacy, representativeness, and data availability will guide dataset selection.

3.3 Preprocessing and Feature Selection

Data preprocessing will prepare source datasets for consistent validation and anomaly detection. Initial steps will include checking data types, standardizing formats, identifying duplicate entries, and examining missing values. Numerical and categorical variables will be handled according to their characteristics and analytical requirements. Feature selection will prioritize attributes relevant to data quality and anomaly identification while excluding irrelevant or redundant information where appropriate. Numerical features may require scaling when the selected algorithm is sensitive to differences in measurement ranges. Categorical variables may require suitable encoding for machine-learning analysis. Missing values will be handled using documented strategies without concealing genuine data quality problems. For temporal datasets, chronological separation of training and testing periods will help evaluate performance under changing data conditions. Preprocessing parameters must be learned from training data only where applicable to prevent information leakage. All transformations and feature-selection decisions will be documented to support reproducibility and ensure fair comparisons between detection methods.

3.4 Quality Assessment and Anomaly Detection

The proposed framework will combine rule-based validation with statistical and machine-learning techniques. Rule-based validation will identify predefined violations, including missing mandatory fields, duplicate business keys, invalid dates, and inconsistent relationships between tables. Statistical methods will detect unusual values or deviations from established baselines. Isolation Forest may be used to identify observations that can be isolated relatively easily, while other suitable algorithms may be evaluated where appropriate. Detection methods will be configured according to dataset characteristics and computational requirements. The experimental evaluation will compare three approaches: rule-based validation alone, anomaly detection alone, and the integrated framework. Each approach will be evaluated using the same test datasets and documented experimental conditions. Detection thresholds and algorithm parameters will be selected using training or validation data rather than the final test set. This design will help determine whether combining explicit business rules with data-driven detection improves the identification of data quality problems.

3.5 Evaluation Metrics

The framework will be evaluated using classification and operational performance metrics. Precision measures the proportion of flagged records that are genuine anomalies, whereas recall measures the proportion of known anomalies successfully detected. The F1-score provides a balanced measure of precision and recall, particularly when the classes are imbalanced. The false-positive rate will measure the proportion of normal observations incorrectly identified as anomalous. Operational metrics will include processing time, throughput, and memory consumption to assess suitability for enterprise workloads. Where relevant, performance will also be evaluated across different dataset sizes and anomaly proportions. Results should be reported separately for each experimental configuration and dataset. Repeated runs and appropriate variability measures can help establish whether observed differences are consistent. The evaluation will consider trade-offs between detection quality and computational cost. No single metric will be treated as sufficient; instead, the combined results will provide a broader assessment of the framework's effectiveness, efficiency, and practical applicability.

3.6 Implementation and Reproducibility

The framework may be implemented using SQL-based validation, Python-based analytical components, and a data warehouse platform such as Snowflake or SQL Server. A business intelligence dashboard can present data quality scores, anomaly counts, severity levels, and remediation status. Implementation details will be documented to support reproducibility, including software versions, algorithm parameters, hardware specifications, dataset partitions, and random seeds where applicable. Experiments should use consistent configurations and equivalent test data to ensure fair comparisons. Sensitivity analysis may examine how changes in detection thresholds or model parameters influence performance. Data privacy will be maintained through anonymization, access

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

controls, and appropriate handling of sensitive attributes. Limitations arising from incomplete anomaly labels, synthetic corruption, computational constraints, and domain-specific validation rules will also be acknowledged. Since the framework is proposed for evaluation, empirical findings must be obtained through actual experiments rather than assumed in advance.

4. RESULTS AND EVALUATION

4.1 Dataset Characteristics

The evaluation will begin by examining the characteristics of the datasets used to assess the proposed framework. Relevant attributes include the number of records, number of fields, data types, missing-value percentages, duplicate prevalence, and verified anomaly counts. Multiple datasets may be selected to represent different enterprise data environments, including customer management, financial transactions, inventory, and operational reporting. The analysis should distinguish naturally occurring quality problems from artificially introduced anomalies. Synthetic anomalies may include missing values, duplicate records, invalid entries, and inconsistent relationships. Dataset characteristics should be summarized in a structured table to support comparisons across experimental conditions. The sources of anomaly labels and procedures used to verify them should also be documented. This information establishes the basis for interpreting subsequent results. Actual dataset statistics must be obtained through experimentation rather than estimated or presented without supporting evidence.

4.2 Data Quality Assessment Results



Figure 3. Data Quality Assessment Results

This subsection will evaluate the framework's ability to detect predefined data quality violations across multiple quality dimensions. These may include completeness, uniqueness, validity, consistency, and referential integrity. For each dimension, the results should report the validation rules applied, the number of violations identified, and the proportion of detections confirmed through verification. Results may be summarized in tables comparing performance across datasets and quality categories. Detection effectiveness must be distinguished from remediation effectiveness because identifying an issue does not necessarily determine the appropriate correction. For example, duplicate transactions may require business-specific rules to identify the authoritative record. Similarly, unusual values may represent legitimate business events rather than errors. The analysis should explain how detected violations were verified and how uncertain cases were handled. These findings will demonstrate the framework's ability to systematically monitor data quality and identify issues requiring further investigation.

4.3 Anomaly Detection Performance

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

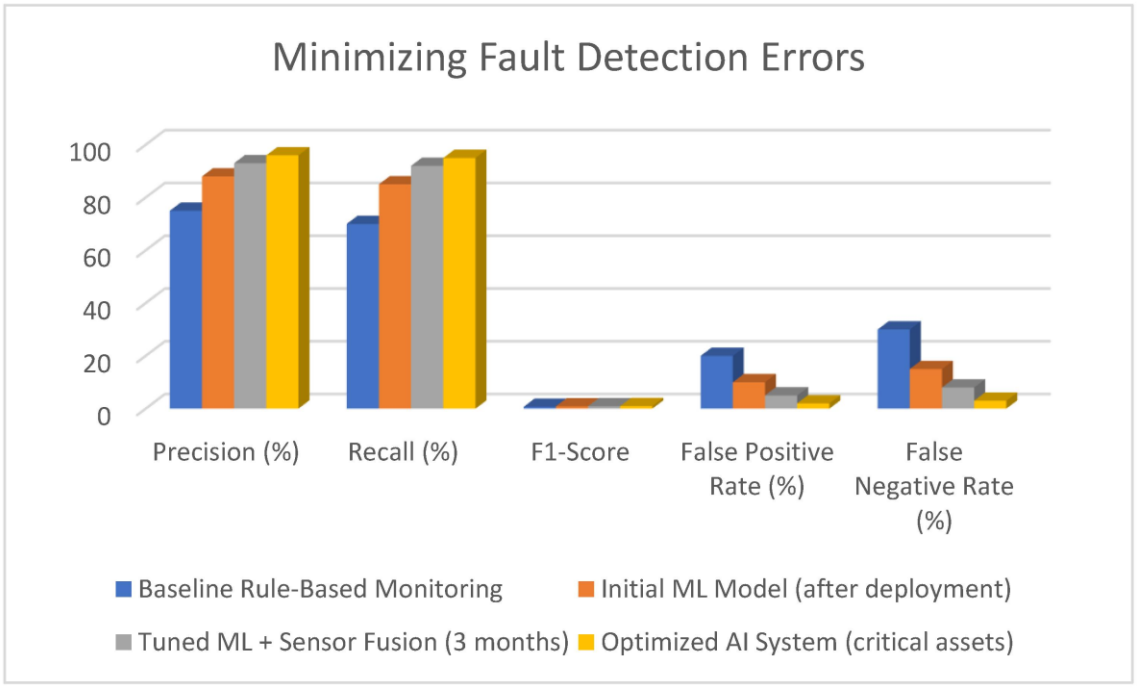


Figure 4. Anomaly Detection Performance

Anomaly detection performance will be evaluated by comparing rule-based validation, standalone anomaly detection, and the integrated framework. Precision, recall, F1-score, and false-positive rate will be used to assess detection effectiveness. Precision indicates the proportion of flagged records that are genuine anomalies, while recall measures the proportion of known anomalies successfully detected. The F1-score balances these two measures. Confusion matrices may provide additional insight into correct and incorrect classifications. Where sufficient data are available, confidence intervals and repeated experiments can help assess result stability. Performance should also be examined across different anomaly types, dataset sizes, and detection thresholds. All methods must be evaluated under consistent experimental conditions. Results should be presented using tables and charts supported by actual measurements. This comparison will establish whether integrating predefined validation rules with machine-learning methods provides measurable benefits over individual detection approaches.

4.4 Computational Efficiency and Scalability

Computational efficiency will be evaluated by measuring processing time, throughput, memory consumption, and resource utilization under different workload conditions. Experiments should examine datasets of increasing size to determine how the framework performs as data volumes grow. Measurement boundaries should include profiling, rule-based validation, anomaly detection, and alert generation. The implementation environment, hardware configuration, software versions, and processing architecture must be documented to support reproducibility. The analysis should identify whether computation occurs within the warehouse, through external analytical services, or using a hybrid approach. Data movement between systems may influence overall execution time and resource requirements. Results should compare the computational costs of individual methods and the integrated framework. This evaluation will help determine whether the proposed solution can support routine enterprise monitoring and larger workloads. Conclusions about scalability should be based on observed measurements rather than assumptions about the underlying technologies.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

4.5 Comparative Analysis

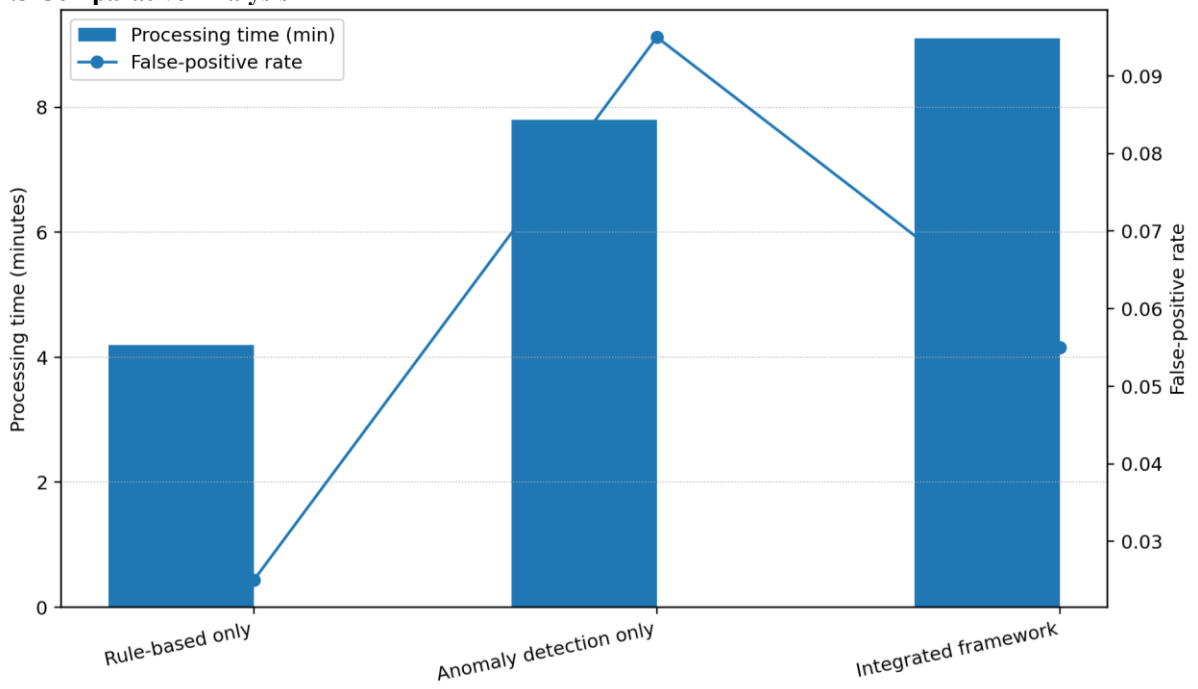


Figure 5: Example Efficiency and False-Positive Trade off

Comparative analysis will examine the strengths and limitations of the three experimental configurations: rule-based validation, standalone anomaly detection, and the integrated framework. Each configuration should be evaluated using the same datasets, metrics, and testing conditions. Performance comparisons will consider precision, recall, F1-score, false-positive rate, processing time, and resource consumption. An ablation study may additionally investigate the contribution of individual framework components by removing or disabling selected modules. For example, researchers could assess whether historical baseline monitoring improves anomaly detection or whether additional validation rules increase computational overhead. Findings should be presented in comparative tables and charts, with explanations of meaningful differences between methods. Performance may vary according to dataset characteristics and anomaly types, so conclusions should account for these factors. The analysis will determine whether integration provides measurable benefits and under which conditions those benefits are most evident. All reported improvements must be supported by experimental evidence.

4.6 Interpretation

The interpretation will connect the experimental findings with the research objectives and assess the practical value of the proposed framework. The discussion should explain whether the integrated approach improves data quality monitoring, anomaly identification, and issue prioritization compared with individual methods. Improvements in recall should be considered alongside false-positive rates, since detecting more anomalies may increase the number of alerts requiring manual review. Similarly, higher detection accuracy should be evaluated against computational costs and operational requirements. The discussion should identify dataset characteristics or anomaly categories that influenced performance and explain any unexpected findings. Limitations arising from incomplete anomaly labels, synthetic corruption, data imbalance, or restricted experimental environments must also be acknowledged. Negative or inconclusive findings should be reported transparently. The final interpretation should avoid unsupported generalizations and distinguish demonstrated results from potential applications. These insights will establish the framework's practical significance and identify areas requiring further experimentation, optimization, and validation in real-world enterprise data warehousing environments.

5. DISCUSSION

The proposed framework integrates rule-based data quality assessment with anomaly detection to improve the identification of data irregularities in enterprise data warehousing. Rule-based methods are suitable for detecting predefined violations, including missing mandatory fields, duplicate records, invalid values, and inconsistent relationships. Anomaly detection techniques may complement these methods by identifying unusual patterns that are not covered by explicit validation rules. However, the effectiveness of this integration must be established

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

through experimental evaluation. The results should identify which quality dimensions are detected most reliably and which remain challenging under different data conditions.

From an enterprise business intelligence perspective, data quality issues can propagate from source systems into analytical models, dashboards, and business reports. Early detection may reduce the risk of inaccurate reporting and improve confidence in analytical outputs. The framework can support this objective by generating explainable alerts that identify affected datasets, violated rules, anomaly scores, timestamps, and relevant evidence. Such information helps data engineers and business analysts investigate issues and determine appropriate corrective actions.

The framework also has implications for data governance and operational management. Metadata-driven rules can promote consistent validation across datasets, while audit logs and lineage information improve traceability. Severity-based alerts can help prioritize incidents according to their potential business impact. Nevertheless, successful implementation requires clearly defined data ownership, documented remediation procedures, appropriate access controls, and regular review of validation rules.

Several limitations may affect the framework's performance. Incomplete anomaly labels, class imbalance, changing data distributions, and computational overhead may influence detection accuracy and operational efficiency. Furthermore, synthetic anomalies may not fully represent actual enterprise data problems, and results obtained using one platform may not generalize to other environments. Improved detection performance should not automatically be interpreted as evidence of better business decisions without additional evaluation.

For practical implementation, organizations should begin with high-value datasets and measurable quality requirements. Rule-based validation can establish a transparent baseline, followed by selective integration of anomaly detection where conventional rules are insufficient. Human review, controlled remediation, periodic performance evaluation, and monitoring for changing data patterns should remain integral to the operational process. This approach supports continuous improvement while balancing detection effectiveness, interpretability, and computational cost.

6. CONCLUSION

This research proposes an integrated framework for automated data quality assessment and anomaly detection in enterprise data warehousing. The framework combines data profiling, rule-based validation, statistical or machine-learning-based anomaly detection, severity classification, automated alerting, and audit logging within a common workflow. Its principal objective is to provide a systematic approach to identifying data quality violations and unusual patterns before they compromise downstream analytics.

The proposed design addresses several practical challenges in enterprise data management. Conventional validation rules provide interpretable checks for known constraints, while anomaly detection techniques may identify deviations that require more flexible analytical methods. Metadata integration can support the consistent application of validation rules, and structured audit records can improve traceability. Nevertheless, the practical value of these capabilities depends on empirical evidence regarding detection performance, processing efficiency, scalability, and the quality of generated alerts.

The framework should be evaluated against conventional rule-based validation and individual anomaly detection methods using documented datasets, verified anomaly labels, and consistent experimental conditions. Precision, recall, F1-score, false-positive rate, and execution time can be used to determine whether the integrated approach provides measurable benefits. Until these experiments are completed, improvements in detection accuracy and operational efficiency should be treated as research objectives rather than established findings.

Future work could investigate adaptive thresholds for changing data distributions, automated rule generation from metadata, explainable anomaly detection, and incremental monitoring of streaming data. Additional studies could evaluate performance across different cloud data warehouses, relational database systems, and enterprise BI platforms. Research involving anonymized production datasets would also help establish whether the framework generalizes beyond controlled experiments.

Ultimately, the contribution of this study will depend on its ability to demonstrate a reproducible and operationally practical method for improving data quality management. By linking detection techniques with governance, auditability, and controlled remediation, the proposed framework offers a basis for further investigation into reliable enterprise analytics.

 © Author(s) | CC BY 4.0 International License.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
https://ijetrm.com/

REFERENCE:

1) Taluri, R. (2024). A Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment on AWS. *International Journal of Artificial Intelligence, Data Science and Machine Learning*.
2) Fadlallah, H., Kilany, R., Haber, M., & Jaber, A. (2023). CTXDQ: An Automated Context-Driven Data Quality Assessment. *2023 IEEE 4th International Multidisciplinary Conference on Engineering Technology (IMCET)*, 32-37.
3) Chakraborty, N., Sharma, A., Dutta, J., & Kumar, H.D. (2024). Privacy-Preserving Data Quality Assessment for Time-Series IoT Sensors. *2024 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 51-57.
4) Chanda, D. (2024). Automated ETL Pipelines for Modern Data Warehousing: Architectures, Challenges, and Emerging Solutions. *The Eastasouth Journal of Information System and Computer Science*.
5) Yamsani, N. (2022). Applying Machine Learning for Automated Data Quality and Anomaly Detection in Enterprise Data Pipelines. *International Journal of Research and Applied Innovations*.
6) Taluri, R. (2024). A Hybrid Data Engineering and Generative AI Architecture for Intelligent Data Governance, Metadata Management, and Automated Data Quality Assessment on AWS. *International Journal of Artificial Intelligence, Data Science and Machine Learning*.
7) Kumar, M.S. (2024). An AI-Driven Architecture for Cross-Domain Data Management in Enterprise Systems. *International Journal of Emerging Research in Engineering and Technology*.
8) Mittamidi, V.K. (2023). Predictive Failure Intelligence for Enterprise-Scale Data Engineering: A Proactive Framework for Monitoring, Forecasting, and Preventing Data Pipeline Disruptions. *International Journal of Emerging Research in Engineering and Technology*.
9) Srujana Parepalli, S.P. (2020). AI-Augmented Data Governance Framework with Proactive Quality Monitoring and Automated Investigative Intelligence. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*.
10) Vasa, M.R. (2024). AI-Augmented Fund Accounting Repository for Governance-Driven Billing Automation. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*.

 © Author(s) | CC BY 4.0 International License.