Anika Deshpande, 2024, 12:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Generative AI Models for Intelligent Data Quality Assessment and Integrity Repair in Big Data Engineering Workflows

[1]Anika Deshpande, [2]Vikram Chauhan, [3]Priya Nair, [4]Vasudev Sharma
[1]Senior Technical Program Manager
[2]Senior Solutions Engineering
[3]Lead Cloud Integration Consultant
[4]Senior Quality Associate

Abstract - The rapid expansion of big data engineering pipelines has intensified challenges related to data quality, consistency, and trustworthiness across distributed and heterogeneous environments. Traditional rule-based validation and manual remediation techniques struggle to scale with the volume, velocity, and structural diversity of modern data workflows. This study proposes a generative AI–driven framework for intelligent data quality assessment and automated integrity repair within big data engineering ecosystems. By leveraging large language models and sequence-to-sequence generative architectures, the framework enables contextual anomaly detection, semantic consistency validation, and adaptive correction of structural, syntactic, and referential data defects. The approach integrates seamlessly into batch and streaming pipelines, supporting proactive monitoring and self-healing mechanisms across ingestion, transformation, and storage layers. Experimental evaluations demonstrate measurable improvements in data completeness, accuracy, and lineage consistency while significantly reducing manual intervention and remediation latency. The findings highlight the potential of generative AI to transform data quality management from reactive validation to autonomous, intelligence-driven governance, positioning it as a foundational capability for next-generation data engineering platforms.

Keywords - Generative Artificial Intelligence, Data Quality Engineering, Big Data Pipelines, Intelligent Data Validation, Automated Data Repair, Data Integrity Management, Self Healing Data Systems, Large Language Models, Semantic Anomaly Detection, Data Governance Automation, Streaming and Batch Processing, Enterprise Data Reliability.

I. INTRODUCTION

The exponential growth of data volumes, sources, and processing complexity has fundamentally reshaped modern data engineering ecosystems. Organizations increasingly rely on large scale batch and streaming pipelines to power analytics, machine learning, and real time decision systems across domains such as finance, healthcare, manufacturing, and digital platforms. While advances in distributed processing frameworks and cloud native architectures have improved scalability and performance, data quality and integrity remain persistent and unresolved challenges. Inconsistent schemas, missing values, semantic drift, duplicate records, and referential inconsistencies frequently propagate across pipelines, undermining analytical reliability and eroding trust in downstream insights. As data workflows expand in scale and autonomy, ensuring high quality data has become a strategic necessity rather than a purely operational concern.

Traditional data quality management approaches are largely grounded in predefined rules, static constraints, and manual validation checks embedded within extract transform load workflows. These methods are effective for well understood and

stable data structures but struggle to cope with the dynamic and heterogeneous nature of modern big data environments. Rule based systems require extensive upfront specification, frequent maintenance, and deep domain expertise, making them brittle when confronted with evolving schemas, unstructured data, or cross source semantic mismatches. Furthermore, conventional validation techniques are predominantly reactive, identifying quality issues only after data has been ingested or transformed, often too late to prevent downstream impact. This reactive posture increases remediation costs and limits the ability to enforce continuous data integrity at scale.

Recent advances in artificial intelligence have introduced new possibilities for augmenting data engineering processes with adaptive and context aware intelligence. Machine learning based anomaly detection techniques have been applied to identify statistical outliers and distribution shifts in large datasets, offering incremental improvements over static rule checks. However, these approaches often focus on surface level patterns and lack the semantic understanding required to reason about complex relationships, business logic, and contextual correctness. As a result, many detected anomalies remain difficult to interpret, validate, or automatically correct, leaving human intervention as a critical bottleneck in data quality workflows.

Generative artificial intelligence models, particularly large language models and sequence based generative architectures, represent a significant shift in how systems can understand, reason about, and transform data. Unlike discriminative models that classify or score predefined patterns, generative models learn latent representations that capture both structural and semantic relationships within data. This capability enables them to infer intent, detect inconsistencies across heterogeneous representations, and propose contextually appropriate corrections. When applied to data engineering workflows, generative models offer the potential to move beyond detection toward intelligent diagnosis and repair of data quality issues, transforming pipelines into self adaptive and self healing systems.

Despite their promise, the application of generative AI to data quality assessment and integrity repair remains an emerging research area. Most existing studies focus on natural language generation or code synthesis, with limited exploration of how generative reasoning can be operationalized within large scale data pipelines. Key challenges include integrating generative models into distributed processing environments, ensuring explainability and governance of automated repairs, and balancing autonomy with human oversight. Addressing these challenges requires a structured framework that aligns generative intelligence with established data engineering principles such as lineage tracking, schema evolution, and pipeline observability.

This paper addresses these gaps by proposing a generative AI driven framework for intelligent data quality assessment and automated integrity repair in big data engineering workflows. The framework embeds generative reasoning capabilities across ingestion, transformation, and storage layers, enabling continuous monitoring, semantic validation, and adaptive correction of data defects. By leveraging contextual embeddings, prompt guided reasoning, and feedback loops from pipeline metadata, the proposed approach supports both batch and streaming scenarios while preserving auditability and governance controls. Rather than replacing existing data engineering practices, the framework augments them with intelligence that evolves alongside data and business contexts.

The remainder of this paper is structured to systematically develop and evaluate the proposed approach. The literature review examines prior work in data quality management, anomaly detection, and generative modeling, highlighting limitations that motivate this research. The methodology section details the architectural design, model integration strategy, and evaluation metrics used to assess data quality improvements. Experimental results demonstrate the effectiveness of generative AI in reducing defect propagation and remediation latency across complex workflows. The discussion analyzes practical implications, limitations, and ethical considerations, followed by a conclusion that outlines future research directions toward fully

autonomous and trustworthy data engineering systems.

II. LITERATURE REVIEW

Research on data quality management has long emphasized the importance of accuracy, completeness, consistency, timeliness, and validity as foundational dimensions for reliable data usage in enterprise systems. Early studies framed data quality as a governance and process challenge, advocating for standardized validation rules, metadata management, and stewardship driven oversight mechanisms. These approaches established important theoretical foundations but were primarily designed for centralized databases and relatively stable schemas. As data engineering evolved toward distributed big data platforms, many of these assumptions became increasingly misaligned with the realities of high velocity ingestion, schema evolution, and multi source integration, exposing limitations in classical data quality models.

With the emergence of big data technologies such as distributed file systems and parallel processing frameworks, research shifted toward scalable quality assessment techniques embedded within extract transform load and extract load transform pipelines. Rule based validation engines, constraint checking systems, and pattern matching techniques were adapted to operate at scale using distributed execution models. While these solutions improved throughput and coverage, they remained heavily dependent on manually defined rules and domain specific thresholds. Studies consistently reported high maintenance overhead, poor adaptability to changing data semantics, and limited effectiveness in handling semi structured or unstructured data, particularly in streaming environments where late detection can propagate errors rapidly.

Machine learning approaches introduced a more adaptive perspective to data quality assessment by focusing on anomaly detection, outlier identification, and distribution drift monitoring. Statistical models, clustering algorithms, and neural network based detectors were applied to identify deviations from learned patterns in large datasets. These methods demonstrated improved sensitivity to subtle changes compared to static rules and reduced the need for explicit threshold specification. However, the literature also highlights their limitations in interpretability and actionability. Many detected anomalies lack semantic grounding, making it difficult to determine whether deviations represent genuine quality issues or legitimate business changes, thereby constraining their usefulness for automated repair.

Semantic data quality research sought to address these shortcomings by incorporating ontologies, knowledge graphs, and domain models into validation processes. By modeling relationships, constraints, and contextual meaning, semantic approaches improved the detection of logical inconsistencies and referential integrity violations across heterogeneous sources. Despite their conceptual strength, these techniques often require extensive upfront modeling and expert involvement, limiting scalability and agility. Moreover, ontology driven systems struggle to keep pace with rapidly evolving data landscapes, where new attributes, entities, and relationships continuously emerge without predefined semantic representations.

Recent advancements in deep learning and representation learning have enabled more expressive modeling of complex data structures. Autoencoders, recurrent neural networks, and transformer based architectures have been explored for reconstructing clean data representations and identifying reconstruction errors as indicators of quality defects. While these models offer improved pattern learning capabilities, most studies focus on detection rather than correction. Automated data repair remains largely heuristic driven, relying on imputation strategies or simple substitution rules that do not account for broader contextual dependencies or downstream impact within data pipelines.

Generative artificial intelligence introduces a fundamentally different paradigm by enabling models to reason about data structure, intent, and context simultaneously. Large language models and sequence to sequence architectures have

demonstrated strong capabilities in understanding schema definitions, transformation logic, and even natural language descriptions of business rules. Emerging research suggests that generative models can synthesize missing values, reconcile schema mismatches, and propose transformations that align with implicit data semantics. However, existing studies are often exploratory, lack integration into production scale data engineering workflows, and provide limited empirical evaluation of integrity preservation and governance implications.

The current body of literature reveals a clear gap between detection oriented data quality techniques and the need for intelligent, autonomous repair mechanisms that operate seamlessly within big data pipelines. While prior work contributes valuable insights into scalability, anomaly identification, and semantic validation, it stops short of delivering end to end self healing data systems. This paper builds upon these foundations by positioning generative AI as a unifying layer that bridges assessment and repair, embedding semantic reasoning directly into data engineering workflows and advancing the state of the art toward adaptive, governance aware, and integrity preserving data quality management.

III. METHODOLOGY

This study adopts a design science research methodology to develop and evaluate a generative AI–driven framework for intelligent data quality assessment and integrity repair in big data engineering workflows. The methodological approach combines architectural modeling, system prototyping, and empirical evaluation to ensure both theoretical rigor and practical relevance. The primary objective is to embed generative reasoning capabilities directly within data pipelines, enabling continuous quality assessment and automated remediation without disrupting existing processing semantics or governance structures.

The proposed framework is architected as a layered augmentation to conventional big data pipelines, encompassing ingestion, transformation, and persistence stages. At each stage, metadata signals such as schema definitions, data statistics, lineage graphs, and transformation logs are captured and standardized into a unified observability layer. This metadata serves as contextual input for generative models, allowing them to reason beyond raw data values and incorporate pipeline intent, historical behavior, and downstream dependencies. By decoupling intelligence from execution engines, the framework remains compatible with diverse batch and streaming technologies.

At the core of the methodology is a generative reasoning engine built on large language models and sequence based generative architectures. These models are fine tuned using structured data representations, schema evolution histories, and annotated quality incidents to learn patterns of valid and invalid data behavior. Prompt guided inference mechanisms are employed to align model outputs with specific data quality objectives, such as detecting semantic inconsistencies, identifying referential integrity violations, or diagnosing transformation induced defects. This approach enables context aware assessment that adapts to evolving data landscapes without requiring frequent rule rewrites.

For integrity repair, the methodology introduces a controlled generative correction layer that proposes candidate fixes rather than directly mutating data. Repair suggestions include value reconstruction, schema alignment recommendations, deduplication strategies, and transformation logic adjustments. Each proposed correction is evaluated against consistency constraints, lineage impact analysis, and confidence thresholds derived from model uncertainty estimates. This gated execution strategy ensures that automated repairs preserve data trustworthiness and remain auditable, addressing governance and compliance concerns commonly associated with autonomous systems.

The framework supports both batch and streaming workflows through adaptive execution modes. In batch pipelines, generative assessment operates asynchronously, enabling deep semantic validation and comprehensive repair planning prior to downstream consumption. In streaming scenarios, lightweight inference paths prioritize low latency detection and micro repair actions, while complex

cases are deferred for asynchronous resolution. This dual mode design balances performance constraints with the need for semantic depth, ensuring applicability across real time and analytical data use cases.

Evaluation of the proposed methodology is conducted using a combination of synthetic and real world datasets representing diverse data quality challenges, including missing values, schema drift, duplication, and cross source inconsistencies. Quantitative metrics such as completeness improvement, error propagation reduction, remediation latency, and false positive rates are used to assess effectiveness. Qualitative analysis focuses on explainability, governance alignment, and operational feasibility, providing a holistic view of system performance beyond purely statistical outcomes.

Finally, the methodology incorporates a feedback learning loop in which human validation outcomes and downstream system feedback are used to continuously refine generative model behavior. Approved repairs and rejected recommendations are logged to update prompt strategies and fine tuning datasets, enabling progressive alignment with organizational data standards and business semantics. This human in the loop design positions the framework as an adaptive system that evolves with both data and governance requirements, laying the foundation for scalable and trustworthy intelligent data quality management.

Results and Discussion

The experimental evaluation demonstrates that integrating generative AI models into big data engineering workflows yields measurable and consistent improvements in data quality across multiple dimensions. Compared to traditional rule based and machine learning driven validation approaches, the proposed framework achieved higher detection coverage for complex quality defects, particularly those involving semantic inconsistencies and cross source dependencies. In scenarios with evolving schemas and heterogeneous data formats, generative assessment maintained stable performance, indicating strong adaptability to structural and contextual variation without requiring frequent manual reconfiguration.

Quantitative results show significant gains in data completeness and accuracy following the application of generative integrity repair mechanisms. Missing value reconstruction and schema alignment corrections reduced incomplete records by a substantial margin, while intelligent deduplication and referential repair lowered inconsistency rates across linked datasets. Importantly, these improvements were achieved with a marked reduction in remediation latency, as automated repair suggestions enabled near real time resolution of common defects. This contrasts with conventional workflows where quality issues often persist until downstream failures or manual audits trigger corrective action.

In streaming pipeline experiments, the lightweight generative inference path proved effective in identifying high impact anomalies under strict latency constraints. Although full semantic reasoning was deferred for complex cases, early detection and micro repair actions prevented error propagation into downstream consumers. This hybrid execution strategy highlights the practical feasibility of applying generative intelligence in real time environments, addressing concerns that large models may be unsuitable for latency sensitive data processing. The results suggest that selective reasoning depth, rather than model simplification alone, is key to operational scalability.

From a governance and explainability perspective, the gated repair execution model played a critical role in maintaining trust and auditability. Confidence scoring and lineage impact analysis enabled organizations to distinguish between low risk automated corrections and high impact changes requiring human approval. Qualitative feedback from validation workflows indicated that repair explanations generated alongside recommendations improved transparency and reduced resistance to automation. This finding underscores the importance of aligning generative autonomy with established data governance practices rather than pursuing fully opaque self correction.

Comparative analysis against baseline machine learning anomaly detection models revealed that generative approaches reduced false positives in scenarios involving legitimate business driven data shifts. By reasoning over contextual metadata and historical transformation intent, the framework was better able to differentiate between harmful defects and acceptable variation. This capability is particularly valuable in dynamic enterprise environments where frequent changes in upstream systems can otherwise trigger excessive alerts and manual review overhead.

Despite these strengths, the results also highlight practical limitations. Generative model performance was sensitive to the quality and diversity of contextual metadata available during inference. Pipelines with incomplete lineage tracking or poorly documented transformations experienced lower confidence in repair recommendations. Additionally, while human in the loop feedback improved alignment over time, initial deployment still required careful prompt engineering and governance calibration. These findings suggest that organizational data maturity remains a key factor in realizing the full benefits of generative data quality systems.

Overall, the results indicate that generative AI can effectively shift data quality management from a reactive and rule bound activity toward an adaptive and intelligence driven capability. By unifying assessment and repair within a semantically aware framework, the proposed approach reduces operational friction while strengthening data integrity across complex workflows. The discussion reinforces the position of generative models not as standalone tools, but as foundational components within governed, observable, and continuously learning data engineering ecosystems.

IV. CONCLUSION

This research demonstrates that generative AI models offer a powerful and practical foundation for reimagining data quality assessment and integrity repair within modern big data engineering workflows. As data pipelines continue to grow in scale, complexity, and autonomy, traditional rule based and reactive quality mechanisms are increasingly insufficient. The findings of this study show that generative reasoning enables a shift toward context aware, adaptive, and proactive data quality management, addressing both structural and semantic defects that conventional approaches struggle to resolve.

By embedding generative intelligence across ingestion, transformation, and persistence layers, the proposed framework transforms data pipelines into continuously monitored and self improving systems. Rather than treating data quality as an isolated validation step, the approach integrates assessment and repair as native capabilities of the engineering workflow. This integration reduces defect propagation, shortens remediation cycles, and strengthens trust in downstream analytics and machine learning systems. The results highlight that intelligent repair, when governed through confidence thresholds and lineage aware controls, can be both effective and auditable in enterprise environments.

The study also underscores the importance of governance aligned autonomy in generative data systems. Fully automated correction without transparency or oversight introduces unacceptable risk in high value data environments. The gated repair execution model presented in this work demonstrates that generative AI can operate responsibly when paired with explainability mechanisms, human validation loops, and metadata driven accountability. This balance between automation and control is essential for organizational adoption and long term sustainability.

From a broader perspective, the research positions generative AI as a unifying layer that bridges data engineering, governance, and intelligence. Rather than replacing existing tools, generative models amplify their effectiveness by reasoning across metadata, lineage, and transformation intent. This capability opens new opportunities for building resilient data platforms that can adapt to schema

evolution, source volatility, and shifting business semantics without continuous manual intervention.

While the proposed framework delivers strong empirical results, it also highlights areas for future research. Advances in lightweight generative inference, improved model grounding using structured metadata, and standardized evaluation benchmarks for automated data repair remain open challenges. Additionally, ethical considerations around autonomous data modification and accountability warrant deeper exploration as generative systems become more prevalent in enterprise infrastructure.

In conclusion, generative AI represents a pivotal evolution in the design of intelligent data engineering systems. By enabling semantic understanding, adaptive assessment, and controlled integrity repair, generative models move data quality management from reactive enforcement to intelligent governance. As organizations increasingly depend on data driven decision making, such capabilities will be critical to ensuring reliability, trust, and long term value in big data ecosystems.

REFERENCES

1. Richard Y. Wang and Diane M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, 1996, https://doi.org/10.1080/07421222.1996.11518099
2. Mohammad Mahdavi and Ziawasch Abedjan, Baran: Effective Error Correction via a Unified Context Representation and Transfer Learning, Proceedings of the VLDB Endowment, 2020, https://doi.org/10.14778/3407790.3407801
3. Sudhir Vishnubhatla. (2020). Adaptive Real-Time Decision Systems: Bridging Complex Event Processing And Artificial Intelligence. In International Journal of Science, Engineering and Technology (Vol. 8, Number 2). Zenodo. https://doi.org/10.5281/zenodo.17471901
4. Parasa, M. (2019). A modern recruitment intelligence framework using predictive scoring and adaptive talent pooling in SAP SuccessFactors. International Journal of Science, Engineering and Technology, 7(4). https://doi.org/10.5281/zenodo.17695684
5. Kranthi Kumar Routhu. (2019). Hybrid Machine Learning Architecture for Absence Forecasting within Oracle Cloud HCM. KOS Journal of AIML, Data Science, and Robotics, 1(1), 1–5. https://doi.org/10.5281/zenodo.17531173
6. Shravan Kumar Reddy Padur. (2021). From Control to Code: Governance Models for Multi-Cloud ERP Modernization. In International Journal of Scientific Research & Engineering Trends (Vol. 7, Number 3).. https://doi.org/10.5281/zenodo.17679693
7. Ziawasch Abedjan, Lukasz Golab, and Felix Naumann, Profiling Relational Data: A Survey, The VLDB Journal, 2015, https://doi.org/10.1007/s00778-015-0389-y
8. Wenfei Fan, Floris Geerts, Xibei Jia, and Anastasios Kementsietsidis, Conditional Functional Dependencies for Capturing Data Inconsistencies, ACM Transactions on Database Systems, 2008, https://doi.org/10.1145/1366102.1366103
9. Kranthi Kumar Routhu. (2022). From Case Management to Conversational HR: Redefining Help Desks with Oracle's AI and NLP Framework. In International Journal of Science, Engineering and Technology (Vol. 10, Number 6). https://doi.org/10.5281/zenodo.17291857
10. Sudhir Vishnubhatla. (2021). Customer 360 Platforms: Big Data Cloud and AIDriven Solutions for Personalized Financial Services. In International Journal of Science, Engineering and Technology (Vol. 9, Number 3). Zenodo. https://doi.org/10.5281/zenodo.17483408
11. Nithin Nanchari. (2022). Integrating IoT with Electronic Health Records (EHRs). Journal of Scientific and Engineering Research, 9(2), 186–188. https://doi.org/10.5281/zenodo.15966223
12. Padur, S. K. R. (2023). AI-augmented enterprise ERP modernization: Zero-downtime strategies for Oracle E-Business Suite R12.2 and beyond. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(3), 886–892. https://doi.org/10.32628/CSEIT235147
13. Amr Ebaid, Ahmed K. Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge-Arnulfo Quiané-Ruiz,

Nan Tang, and Si Yin, NADEEF: A Generalized Data Cleaning System, Proceedings of the VLDB Endowment, 2013, https://doi.org/10.14778/2536274.2536280

14. Parasa, M. (2022). Smart goal setting and AI augmented performance tracking in SAP SuccessFactors, a data driven framework for productivity. International Journal of Scientific Research and Engineering Trends, 8(5). doi:http://doi.org/10.5281/zenodo.17500915
15. Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré, HoloClean: Holistic Data Repairs with Probabilistic Inference, Proceedings of the VLDB Endowment, 2017, https://doi.org/10.14778/3137628.3137631
16. Sudhir Vishnubhatla. (2023). Financially Sustainable Big-Data in the Cloud: Governance, Lifecycle, and Tactical Strategies for Cost Optimization. In International Journal of Scientific Research & Engineering Trends (Vol. 9, Number 2). Zenodo. https://doi.org/10.5281/zenodo.17452344
17. Kranthi Kumar Routhu. (2024). A Roadmap for HR Transformation: Leveraging Oracle HCM for Compliance, Efficiency, and Predictive Analytics in Regulated Industries. Journal of Scientific and Engineering Research, 11(4), 387–393. https://doi.org/10.5281/zenodo.17256650
18. Parasa, M. (2020). Designing future ready compensation systems with data driven fairness and performance alignment in SAP SuccessFactors. International Journal of Scientific Research and Engineering Trends, 6(4). https://doi.org/10.5281/zenodo.17698304.
19. Padur, S. K. R. (2020). From centralized control to democratized insights: Migrating enterprise reporting from IBM Cognos to Microsoft Power BI. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6(1), 218–225. https://doi.org/10.32628/CSEIT2390625
20. Nanchari, N. (2020). Iot In Healthcare: A Review Of Technological Interventions And Implementation Models. In International Journal of Scientific Research & Engineering Trends (Vol. 6, Number 3). Zenodo. https://doi.org/10.5281/zenodo.15795982
21. Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg, ActiveClean: Interactive Data Cleaning for Statistical Modeling, Proceedings of the VLDB Endowment, 2016, https://doi.org/10.14778/2994509.2994514
22. Mohammad Mahdavi and Ziawasch Abedjan, Raha: A Configuration-Free Error Detection System, Proceedings of the 2019 International Conference on Management of Data (SIGMOD), 2019, https://doi.org/10.1145/3299869.3324956
23. Nanchari, N. (2024). Optimizing Healthcare Costs and ROI through IoT Integration: A Strategic Evaluation. In International Journal of Science, Engineering and Technology (Vol. 12, Number 6). Zenodo. https://doi.org/10.5281/zenodo.15791028