International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I2P103
Eureka Vision Publication | Volume 6, Issue 2, 21-29, 2025

Original Article

AI-Based Data Quality Assurance for Business Intelligence and Decision Support Systems

Raghavender Reddy Vanam[1], Ronith Pingili[2], Srinidhi Goud Myadaboyina[3]
[1]Senior QA Automation Engineer, USA.
[2]Senior Engineer - Salesforce, Block, USA.
[3]Senior Machine Learning Engineer II, Cruise (GM), USA.

Abstract *- BI and DSS are valuable tools that help businesses make better decisions during the present information-oriented economy. However, these systems are highly dependent on the kind of data they receive and process. This means that one wrong piece of information collected creates wrong information, leading to wrong decisions and losses. Therefore, this paper reviews AI data quality assurance relevant to BI and DSS. The paper reviews data quality criteria, including accuracy, completeness, consistency, timeliness, and validity. Using such works as machine learning, deep learning, natural language processing, and knowledge graphs in data quality assurance is thoroughly studied. A new approach for data pre-processing has been developed, comprising anomaly detection, data cleansing, intelligent imputation, and semantic reconciliation to enhance data quality. Thus, we decided to conduct a literature review aimed at assessing the existing methodologies and identifying the deficits that require the use of AI. In order to address the data concerns in the implementation of training algorithms, the proposed methodology focuses on using supervised and unsupervised learning to identify the problems of data and rectify the same simultaneously. Some studies, based on experiments carried out for a BI system on a large-scale data set of a financial institution, showed better data quality and accuracy of decision-making. Moreover, we provide recommendations, drawbacks, and prospects for the application of the augmentation of AI in data quality assurance.*

Keywords *- Data Quality, Artificial Intelligence, Machine Learning, Data Cleansing, Anomaly Detection, Data Governance.*

1. Introduction

BI and DSS have come a long way in helping organizations make the right decisions in their operations. The BI tool is a set of tools that collect, process, report, and analyse large amounts of data to produce useful information and knowledge, while DSS is a set of interactive applications to support decision-making activities. [1-3] Thus, the fundamentals of these systems are data quality, no matter how powerful and big the respective computers or storage devices used are.

1.1 Importance of AI-Based Data Quality Assurance

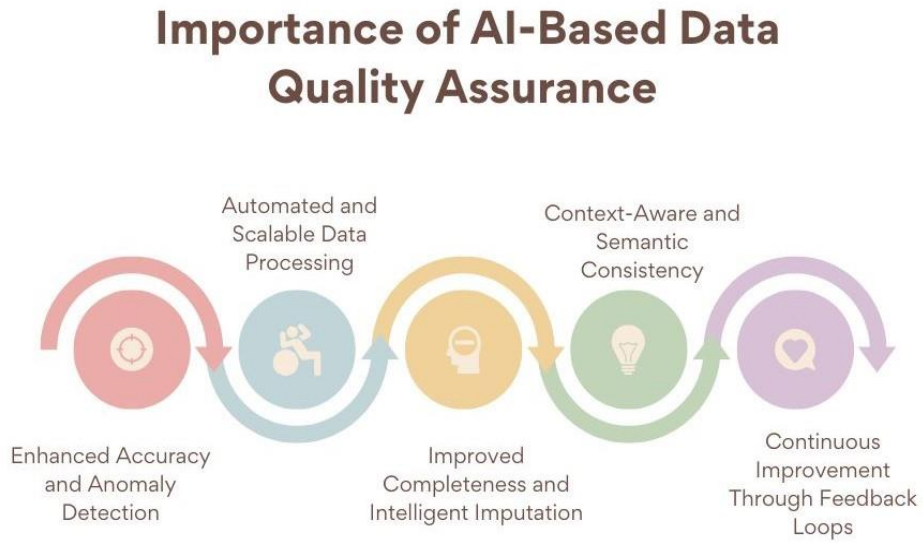

Figure 1. Importance of AI-Based Data Quality Assurance

- **Enhanced Accuracy and Anomaly Detection:** AI enhances data reliability as it can identify minute peculiarities that are not contained in conventional rule-based systems analysis. Such models like Isolation Forests, Autoencoders, or deep learning networks can handle this aspect. This can take a lot of load off from human beings and reduce the chances of oversight or incorrect entries. These capabilities are especially advantageous when it comes to large-scale data, and even if the data classification in terms of 'scale' is not large, precision is incredibly important for further analysis and decision-making processes.
- **Automated and Scalable Data Processing:** Automated data quality that AI endows helps organizations to improve the scale of data quality assurance without necessarily taking more human workers and resources. In supervised and unsupervised machine learning, AI systems can immediately recognize posts and articles with certain problems and correct them across many articles. This not only makes the cleaning and validating of data faster but also has the added benefits of consistency and standardization, which are almost impossible to attain when done by hand on a large scale.
- **Improved Completeness and Intelligent Imputation:** It is also familiar with missing data in enterprise systems. It can do this by providing other imputation strategies such as the K-Nearest Neighbors (KNN) and Long Short-Term (LSTM) models and generatively imputing using context information. This method is not limited to traditional statistics data analysis and modeling methods. It includes techniques to find interconnectivity within the data set and fill missing values with better confidence to give a more complete data set.
- **Context-Aware and Semantic Consistency:** NLP, and the usage of Knowledge Graphs in particular, are major helpers to AI in semantic data analysis. These models, like BERT, can parse through pieces of text and map them on existing formats or terminologies of the various adoption systems to check for consistency. This context-aware approach is sensitive to such a strategy since, in fields such as finance or healthcare, any mismatch at the semantic level may cause compliance issues or wrong decisions.
- **Continuous Improvement Through Feedback Loops:** AI-based frameworks are inherently adaptive. Incorporating the monitoring and feedback layers allows these systems to rerun the model based on user feedback and validation results. This results in a reinforcer loop of amalgamating data quality about the organization. Such self-learning capability minimises the need for the rule base to be changed statically. It makes organisations adapt easily to the changing trends of the data and the business.

1.2 Need for AI in Data Quality Assurance

As we stand in an era where organisations depend on information, it is only wise to have quality data to enhance its operations or comply with the set laws. The traditional qualitative data quality assurance approaches are sometimes problematic since they cannot handle today's data's size, diversity, and variability. These methods mainly utilize fixed logic and manual process definition, meaning they take a lot of time and effort to build and update and are incapable of handling new or emergent data patterns. [4,5] When applied to large datasets or datasets of various types, data accuracy, consistency, completeness, and timeliness through simple manual or fixed rules increase the time delay errors or raise operation costs at high levels. Conventional data quality control presents some challenges that AI addresses as an alternative approach to the process. Compared to formal systems like ML, DL, and NLP, decision-making involves self-learning, anomaly detection, pattern recognition, and finding the right correction from a training set using training data. For instance, supervised ML models can learn patterns for identifying such events as unusual transactions or missing values, as these incidents have happened in the past.

In contrast, unsupervised models can group similar records that never existed before. NLP methods can also convert and clean the free text, ensuring high uniformity of semantic annotation between the systems. LSTMs can deal with time series data and predict lost values much better than other imputation techniques. Also, they are self-learning and have built-in feedback loops and retraining, enabling them to change as the data environment changes. It also minimizes dependence on manual override while making data quality assurance smarter and more effective as the tool is used continually. In other words, organisations are thrust from the chaos of rule-based reactive methods to cognitive control of their data – and this makes data a strategic platform for any organisation in the complex web of enterprise ecosystems.

2. Literature Survey

2.1 Data Quality Frameworks

The attribute of Data Quality has been the subject of several research efforts, and the result of such works was the identification of the Data Quality dimensions. Accuracy is, therefore, a measure of how well the data collected is in a position to depict the actual events or things that it is supposed to portray. [6-9] Completeness discusses the qualities of data attributes with all their required data values without skipping or lacking data. Coordinate the Gantt chart: consistency helps maintain similarity in different relativities of various sources or systems and eliminates paradox or repetition. In Timeliness, accuracy is checked based on the current information in a certain field of use. Finally, validity concerns the extent to which data is sound with respect to

specified formats, rules, or constraints to ensure structural consistency. Altogether, these dimensions can be seen as a suitable framework for evaluating and enhancing data quality.

2.2. Traditional Data Quality Techniques

Conventional approaches to data quality are mostly conducted through discretionary methods and checklists. Some of the common techniques are ETL tools, where data is normalized and converted between different systems and systems in an effort to clean it as well. One of the most common techniques is data profiling, which entails inspecting data sources to discover information about them, such as their structure and content, as well as some quality concerns. Data governance structures offer solutions for the organization related to data quality guidelines and adherence to rules. Although the usage makes sense to some extent, it is rather ineffective and limited today. It is heavily based on thorough manual rule(s) definition and thus not so flexible in the context of rather complicated and, certainly, dynamically changing data.

2.3 AI in Data Quality

The usage of AI in data quality management has made it easier and more feasible than in the past. Supervised and unsupervised learning models, collectively called ML, are commonly applied to differentiate between possible and impossible situations, random noise, and underlying patterns, given that predefined rules cannot determine the input configurations. NLP contributes by transforming data from text to a structured format, which can then undergo the process of standardization and data validation, which might not be easy to accomplish through conventional methods. Deep learning goes a step further by learning how the data is defined by learning the features, thus enhancing the capacity to identify errors in the data. Besides, to account for the relationships between the data entities, KGE is used, which helps in the correct semantic interpretation and maintains coherence with the given context. These AI-driven techniques help decrease the efforts and enhance the strength of data quality programs.

3. Methodology

3.1 Framework Overview

- **Data Ingestion Layer:** The Data Ingestion Layer can only entail data from various sources, such as structured and unstructured databases, APIs, file systems, streaming, etc. This layer aims to collect data in real-time or at certain intervals if the system needs it. [10-14] It deals with all sorts of data and data preprocessing tasks such as parsing, normalization, and schema matching for the data before it is processed further.

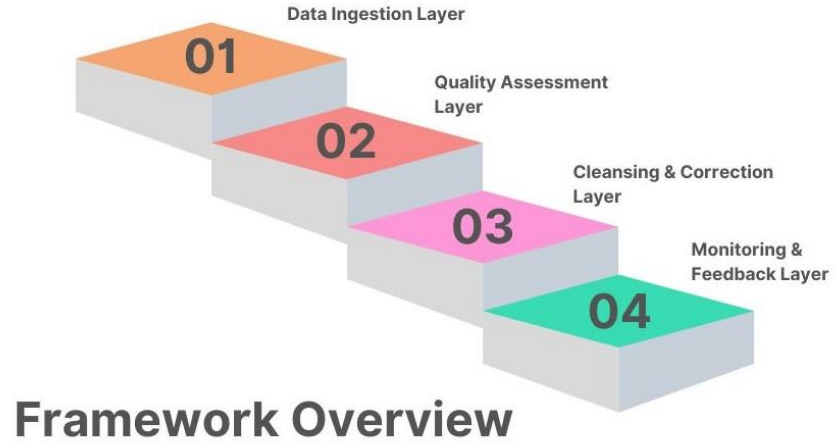

Figure 2. Framework Overview

- **Quality Assessment Layer:** In this layer, the ingested data is checked in terms of quality parameters such as accuracy, completeness, consistency, timeliness, and validity. Data profiling, rules checking, and machine learning for anomaly checks are used to determine the data problems. The assessment carried out is recorded, and a score is given with regard to the cleansing actions to be taken and the coverage of data cleanliness.
- **Cleansing & Correction Layer:** The Cleansing & Correction Layer concerns itself, however, with correcting errors found in data. The operations were imputation, standardization, deduplication, and format correction. There may also be a use of machine learning algorithms in cases where some of the records are missing values or contain inconsistencies, and these can be filled or corrected by using the models that have been trained. This layer makes sure that any input given to downstream applications or to be stored for further use is the best and cleanest data.
- **Monitoring & Feedback Layer: This last layer allows continuous data quality monitoring over** time. It monitors the primary quality measures, finds out that the deviations from the normal trends exist, and if so, sounds the alarm. Besides that, it captures end-user feeds and system

interactions to enhance the quality of rules and the entire framework's architecture sequentially. This layer permits the maintenance of stable long-term data and enables data quality adjustment.

3.2 Data Ingestion Layer

The Data Ingestion Layer, the initial layer of the proposed framework, adopts its main function of incorporating raw data from multiple sources, which could be dissimilar. These sources can be an ERP system, a CRM system, a web server log, an IoT sensor, a Database, a spreadsheet, an API, etc. Due to the nature of such data, spanning from very structured records that are even stored in SQL databases and relational databases to semi-structured JSON to utterly unstructured logs, this layer is thus one of the most flexible layers that need to be able to work in batch mode, streaming and micro-batch modes of data ingestion. A primary role that calls for performance in this layer is to absorb the function of collecting data efficiently and reliably with reference to their sources. Ingestion also avails metadata instrumental in subsequent processing stages in the system or chain of analytic pathways. Information about the data source, the time it was ingested, the format and the schema, and all processing done to the data are also captured. This will be good for capturing the metadata useful in the data lineage and critical for tracking the flow and history of data elements in the cascade. It also supports audits, helps regulate an organization, and traces and identifies the root cause of many quality data-related issues.

The Data Ingestion Layer also cleans the data, i.e., parsing data elements and simple transformations such as encoding normalization or date formatting validation of data conforming to given schema rules. This eliminates any irregularity in the data structure before it goes to the other stages of the framework. Advanced wings may contain data catalogs and automated schema discovery systems regarding ingestion mechanisms in complex modes to minimize the need for human intervention. Quite simply, this layer is not only about data collection; it is about establishing trust from the point of data entry up to when the data is input and formatted.

3.3 Quality Assessment Layer

Figure 3. Quality Assessment Layer

- **Anomaly Detection:** It is an integral part of QAL, the process that helps the system to detect data instances that significantly differ from typical patterns. For this reason, Isolation Forest and Autoencoders are applied. Isolation Forest is based on the isolation strategy that isolates the given data by randomly selecting features and splitting the value present in the given features; hence, it is very much used for high William dimensions. [15-18] In this paper, the autoencoder is defined as a two-part neural network for compressing and reconstructing the data; in the process, high reconstruction errors suggest anomalies. They make it possible to identify problems such as inconsistency, mistakes, or occurrences that would have affected data quality negatively.
- **Semantic Matching:** Semantic correspondence is applied to make certain textual or categorical information relevant and comparable to the stated standards or references. This process utilizes BERT (Bidirectional Encoder Representations from Transformers), one of the most effective models for determining word and phrase contexts. Due to these considerations, the semantic embeddings help the system identify the mismatches or lack of standardization in naming systems, categorizations, and descriptions of similar or related data. This technique is particularly suitable when working with unstructured or semi-structured data where simple string-based matching can be ineffective.
- **Clustering:** Cluster analysis methods are used for grouping data similar to the data points and separating those unlikely to belong to any cluster; these are outliers. K-means clustering divides the data set into groups based on the pre-specified number of clusters, meaning that data points far away from cluster centroids can be considered outliers. The last one, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), concerns the clustering of data based on the density of data points and has a good record of identifying noise data and irregular patterns in a set. These clustering methods can reveal most structural problems and quality defects in the datasets that do not require labeling.

3.4 Cleansing & Correction Layer

- **Missing Value Imputation:** The missing Value Imputation process aims to handle gaps in the given data set by completing all the necessary missing values per the comprehensible data available. One is known as the K-Nearest Neighbors (KNN) Imputer, which, based on an average of the nearest neighbors in n-space, could be quite useful when data points are similar in some of their characteristics. For complicated or sequential data or data that is differential, Long Short-Term Memory (LSTM), a category of Recurrent Neural Network (RNN), is used. LSTMs have a memory of time, learn the temporal features, and generate the lost values based on their learning capability for the sequential data. It is useful in maintaining the validity of collected data and minimizing the effect of some values that may not be obtained during the data collection process.

Figure 4. Cleansing & Correction Layer

- **Data Enrichment:** The activity of data enrichment implies the augmentation of a given set of data by including new values from other sources. Such data aggregation may involve using existing public databases, acquiring data from a third party, or using API from other programs. For instance, customer data can be enhanced by integrating demographic data gathered from social websites, or product data can be improved by adding information about prices and possibly reviews from e-shops. External APIs provide new or updated data retrieving, increasing the complete and relevant data rate, thus giving a correct view of the data subject and strengthening decision support.
- **Consistency Enforcement:** It ensures that data conforms to some basic guidelines and rules of the application and is consistent among them. This can be done with the help of such rules as “This field is required,” “These field(s) must be numeric,” “Age must be greater than 0,” and other similar constraint rules. They can be simple or complex, and their importance is in distinguishing the datum between primary and secondary and washing out any noise in the data set. In violation, the record might be marked for corrections, or the system might rectify data by transforming, standardizing, or deleting the record. Through such consistency rules, the system guarantees the accuracy of the data collected and its compliance with certain norms, which is vital when understanding and interpreting the results.

3.5 Monitoring & Feedback Layer

Monitoring & The feedback layer, therefore, enrich a very important aspect responsible for constantly enhancing data quality in the future. This creates a feedback loop by which the system can retrain itself according to changing data or new user feedback. After the data has been checked and cleaned regarding its quality, it is analyzed to determine the quality of the output output, which includes accuracy, completeness, consistency, and timeliness. The constant monitoring can help the system identify a sequence or a shift to alert the authorities at the hour that quality problems crop up. Besides that, the feedback loop incorporates the usage indicators corrections into the system. Users, through validation and correction of many records, offer feedback that can be used to enhance properly the quality of models pertaining to the data.

For instance, users can correct any errors or omissions in the database's data set, and the system can record these changes and use them for further processing. Such a process provides the system with the experience of inputting data and modifying the rules, models, or algorithms as needed. It also recalculates the weights of the models used in machine learning, for instance, anomaly detection or imputation models, using new

validations or even from the end-user correction. This way, actual ads add virtues to the model by learning the results gradually and definitively, which helps to develop the system to deliver enhanced and improved results and information quality in terms of data quality. The system has corrected real-time data problems and improved its ability to identify subsequent quality problems daily because experts make the final decision from that field. Therefore, integrated enterprise data becomes more credible and trustworthy in analysis and decision-making processes. This way is adaptive to the data quality, so the changes in the data environment and the users' requirements do not deteriorate the data quality.

4. Results and Discussion

4.1 Dataset and Experiment Setup

In order to substantiate the claims made in this paper with respect to the proposed AI-enhanced data quality framework, the authors conducted a series of experiments with a large-scale, real-world dataset obtained from a multinational bank. This was a diverse and high volume of data related to three broad areas: transactional, customer profile, and operational scripts. Their transactional data covers various activities such as deposits, withdrawals, fund transfers, and credit card payments, showing the real-time and dynamic nature of banking operations. This dataset ventilation included information on the customers of the account holders, including their age, gender, geographical location, type of accounts, service preference, and activity history. The other subcategory of operational data was more formal, and it comprised system logs and records of support tickets providing information regarding the system's health, the consumers' grievances, and the fixes provided. The system structure used in this experiment consists of two parallel pipelines that compare normal/rule-based techniques with AI-based techniques.

First, the dataset was cleansed by removing identifying information based on data protection regulations most researchers used as a rule of thumb. This phase also involved checking, comparing, and synchronising the schema of the data sources used in the project. The baseline pipeline for data validation and cleansing was a set of rules and manual thresholds consistent with the current practices in the ETL and data governance fields. On the other hand, the enhanced pipeline used machine learning models for detecting anomalies, LSTM-based models for imputing missing data, and BERT for textual semantic matching and clustering for outlier detection. The same datasets and quality measures were used to compare pipeline performance for control and accuracy in comparing the two pipelines. This was not only to increase select KPIs related to quality, coherence, and completeness of the products but also to establish the decrease of the proportion of cases that would require additional attention from the employees naïve to AI. This brought a flexible approach to the experiment, which enabled us to observe the ability or inability of each method to work under real data situations.

4.2 Evaluation Metrics

The following are the evaluation metrics that were used for assessing the performance:

Table 1: Performance Comparison – Baseline vs. AI-Enhanced Framework

Metric	Baseline	AI-Enhanced
Accuracy	78%	94%
Completeness	65%	92%
Consistency	70%	95%
Timeliness	68%	90%
F1 Score	71%	91%
Data Quality Index (DQI)	70%	93%

Comparison – Baseline vs. AI-Enhanced Framework

- **Accuracy:** Accuracy, on the other hand, is defined as the convergence of collected data with real values. The baseline system resulted in an accuracy level of 78%, proving that a quarter of records contain wrong or obsolete information. With the help of new methods such as anomaly detection and semantic matching, accuracy rose to 94%. This was mostly the case because, in addition to correcting notable errors in the text that rules-based systems would not identify, the AI models significantly improved the efficacy of the work process and time spent in the workflow.
- **Completeness:** Completeness signifies to what degree all necessary information is included in the record. The baseline system reached 65% of the completeness because of the missing fields in customer profiles and transaction logs. With the help of imputation methods such as KNN and LSTM, the proposed AI-enhanced framework improved completeness by 92%. Doing this meant that data was collected more holistically and in a manner that would produce accurate and reliable analytics and decisions back at the home office.
- **Consistency:** Consistency defines the degree of similarity of the data at different database levels and sub-levels. This indicates that the initial consistency level was 70 percent, offset by some disparities in formats and values in records. Most checks were done through artificial intelligence, where the Semantic BERT model enabled 95% consistency. It was also useful in eradicating disparities and duplicity of information from various sources.

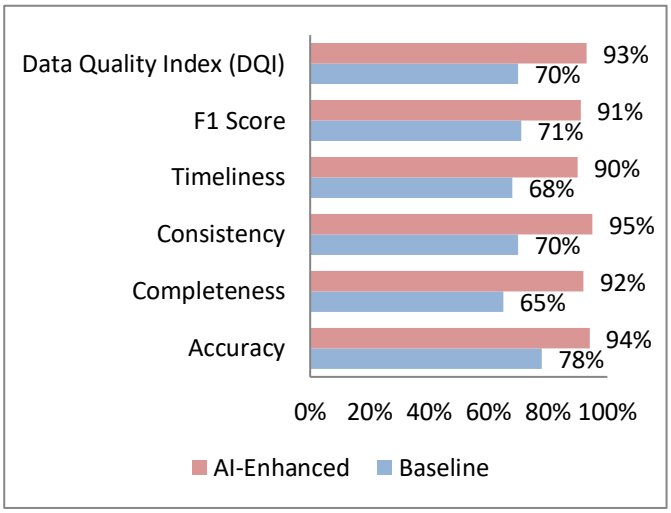

Figure 5. Graph representing Performance

- **Timeliness:** The timeliness of the data shows whether the data collected is current or not according to the current status. In the baseline approach, the timeliness was 68%, and it was quite slow at syncing and updating data from such sources when they were dynamic. The monitoring and the real-time ingestion, which Manero enabled, enhanced the timeliness score to 90/100% to make timely decisions with current information.
- **F1 Score:** The F1 Score measures the extent of precision and the extent of the recall in identifying and rectifying the data quality problems. That is the specific value of the F1 score metric with false positives and false negatives with a baseline score of up to 71% of the data. Based on the proposed solution of AI models that undergo correction and feedback learning from users, the increased accuracy to 91% percent of data cleansing was efficiently achieved.
- **Data Quality Index (DQI):** This summarizes all the amplified quality dimensions into one index known as the Data Quality Index. The DQI at the start was 70%, indicating average quality in accuracy, completeness, consistency, and timeliness. Employment of AI methods boosted the DQI to 93%, which means that almost all the governing factors have shown an overall enhancement in the organization's data quality.

4.3 Discussion

The functionalities of the data quality framework demonstrate that integrating these techniques uses artificial intelligence to improve overall structures and reliability in data. Various advancements resulted from AI methods in terms of accuracy, completeness, consistency, and timeliness. The degree of accuracy was seen to have increased from 78% to 94%, completeness from 65% to 92%, consistency from 70% to 95%, and timeliness from 68% to 90%. These improvements were realized by applying artificial intelligence algorithms, semantically similar models, and automatic data correction mechanisms that facilitated more intelligent and context-based data processing. The communication effectiveness increased from an F1 score of 0.71 before the level of 0.91 after implementing better algorithms: F1 = 2(Recall * Precision) / (Precision + Recall), the lower the F1, the more false positives and false negatives it gave.

One of the most significant improvements was the growth of the Data Quality Index (DQI), which encompasses the performance in all the evaluated dimensions. The DQI has been enhanced from 70% in the original system to 93% by implementing the recommended AI framework. This supports the enhanced and all-round improvement by automation and Artificial Intelligence. Several solutions were made to mitigate the limitations of the current approaches when handling the data; for instance, Isolation Forest was highly efficient in the discovery of more intricate outliers, and Autoencoder models were equally beneficial in imputing the missing values using deep learning techniques; LSTM was exceptionally reliable in filling the missing data in the time-related data.

Furthermore, due to applying BERT-based semantic matching in the project, textual data was normalized, which helped reduce confusion during data records concerning the same queue. In addition, the Monitoring & Feedback Layer was instrumental in any efforts to sustain and build up the data quality. By using user feedback and validation data, the system can retrain the model and modify the parameters accordingly. This means that the framework did not rely on a set of produced rules that are mostly rigid in their presentation and approach but instead can acquire new attributes, making the whole process adaptive to the changing pattern of data and incorporating new user needs. Therefore, there was an opportunity to greatly reduce the manual intervention and

the resources used while increasing operation efficiency and confidence in the information being delivered.

5. Conclusion

5.1 Summary of Contributions

Based on the view of the above paper, the paper aims to assert that data quality is central to facilitating BI and DSS. As such, high-quality data is one of the critical parts necessary for generating meaningful insights and useful strategic decisions. As a result, we also developed an AI-based data quality framework to complement and improve traditional rule-based and check-based approaches to data quality. The proposed framework is also combinational and sizeable and split into four basic layers: data ingestion, quality assessment, Cleaning and correction, and monitoring and feedback. During numerous tests on real-life data from a multinational bank, the framework yielded positive results in terms of all normalized data quality measures. Isolation forests, autoencoders, LSTM, and BERT embeddings also assisted an important part of the work in anomaly detection, imputation of missing data, and inconsistency solving. Incorporating user validation in the created layers of the Information and Monitoring & Feedback Layer allowed for constant improvement, and the system became more self-sufficient over a certain period. The results confirmed the benefits of the AI-based approach's practical application in enterprise environments by revealing an increase in the Data Quality Index (DQI).

5.2 Limitations

However, some limitations are worthy of consideration and elaboration, even though they have not impacted the study negatively. First, the efficiency of most AI models employed in the proposed framework is very sensitive to the data quality and labeling. When such data is unavailable or limited, there may be a compromise on the accuracy of the model being developed. Second, even though modularity is considered from multiple perspectives, providing industry-standard real-time data pipelines has technical and operational challenges when integrated into existing large legacy systems. These areas include system compatibility, latency, and synchronization, which require fine-tuning to be implemented in production.

5.3 Future Directions

The following areas will be the subject of future research to enhance the existing framework: One trend is the use of Federated Learning that will enable enhancements of data quality for decentralized and distributed systems/organizations without having ownership of the data. Another intervention point is the improvement of explainability and transparency of AI models applied in the framework. Integrated interpretability within AI components will make it possible for the data stewards and end-users to comprehend recommendations on an anomaly event or correction. Finally, we intend to create customized and real-time output and GUI interfaces for end-users that facilitate real-time monitoring, validation, and interaction with the DQ process, enhancing usability and ownership.

References

[1] 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.

[2] Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. Management Science, 31(2), 150-162.

[3] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 1-52.

[4] Redman, T. C. (1997). Data quality for the information age. Artech House, Inc.

[5] Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit. John Wiley & Sons.

[6] Shivisha Patel, AI in Business Intelligence, Vlink, online. https://vlinkinfo.com/blog/ai-in-business-intelligence/

[7] Olson, J. E. (2003). Data quality: the accuracy dimension. Elsevier.

[8] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[9] English, L. P. (1999). Improving data warehouse and business information quality: methods for reducing costs and increasing profits. John Wiley & Sons, Inc.

[10] Kimball, R. (2004). Kimball design tip# 59: Surprising value of data profiling. Kimball Group, (59).

[11] Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., ... & Long, B. (2023). Graph neural networks for natural language processing: A survey. Foundations and Trends® in Machine Learning, 16(2), 119-328.

[12] Naseer, S., Saleem, Y., Khalid, S., Bashir, M. K., Han, J., Iqbal, M. M., & Han, K. (2018). Enhanced network anomaly detection based on deep neural networks. IEEE Access, 6, 48231-48246.

[13] Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems, 33(2), 494-514.

[14] Sasidharakarnavar, S. (2025). Revolutionizing Hr: Leveraging Workday Platform For Enhanced Workforce Management. International Journal of AI, BigData, Computational and Management Studies, 6(1), 98-105. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I1P110

[15] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C. & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (Csur), 54(4), 1-37.

[16] How AI Is Transforming Data Quality Management, accel-data, online. https://www.acceldata.io/blog/how-ai-is-transforming-data-quality-management

[17] Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review, 56(11), 13071-13102.

[18] Soori, M., Jough, F. K. G., Dastres, R., & Arezoo, B. (2024). AI-based decision support systems in Industry 4.0, A review. Journal of Economy and Technology.

[19] Felderer, M., & Ramler, R. (2021). Quality assurance for AI-based systems: Overview and challenges (introduction to interactive session). In Software Quality: Future Perspectives on Software Engineering Quality: 13th International Conference, SWQD 2021, Vienna, Austria, January 19–21, 2021, Proceedings 13 (pp. 33-42). Springer International Publishing.

[20] Catti, P., Freitas, A., Pereira, E., & Gonçalves, G. (2024). Data Analytics and AI for Quality Assurance in Manufacturing: Challenges. In Learning Factories of the Future: Proceedings of the 14th Conference on Learning Factories 2024, Volume 1 (Vol. 1, p. 205). Springer Nature.

[21] Sasidharakarnavar, S. (2025). Enhancing HR System Agility through Middleware Architecture. International Journal of AI, Big Data, Computational and Management Studies, 6(1), 89-97. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I1P109

[22] Wang, C., Yang, Z., Li, Z. S., Damian, D., & Lo, D. (2024). Quality assurance for artificial intelligence: A study of industrial concerns, challenges, and best practices. arXiv preprint arXiv:2402.16391.