International Scientific Journal of Engineering and Management ISSN: 2583-6129
Volume: 03 Issue: 07 | July – 2024 DOI: 10.55041/ISJEM02160
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AI Enhanced Data Quality in Data Warehouses and Data Lakes for Efficient Data Driven Intelligence

Kiran Veernapu, Kiran_veernapu@yahoo.com

Abstract

Data quality is paramount in data-driven decision-making processes, especially when dealing with large volumes of data in environments like data warehouses and data lakes. These systems store vast amounts of raw and processed data from multiple sources, making data management and quality assurance complex yet critical. With the growing adoption of Artificial Intelligence (AI), new techniques and tools have emerged that can significantly enhance data quality. This paper discusses how AI can improve the quality of data within both data warehouses and data lakes by automating data cleansing, validation, anomaly detection, and ensuring consistency. It explores the benefits, challenges, and methodologies for integrating AI tools into these systems.

Keywords: Data quality, AI in data quality, data warehouse, data lakes, big data, data processing, data cleansing, data profiling.

1. Introduction

A data warehouse is a structured, centralized repository used to store historical data from multiple sources for reporting and analytics [1]. A data lake is a centralized repository that stores vast amounts of raw data in its native format (structured, semi-structured, or unstructured), enabling flexible analytics. Data Lakes were designed to host vast volumes of various data in their raw, unchanged from and almost process data in real-time [2]. While data lakes offer greater flexibility than data warehouses, data quality still plays a key role. Data warehouses and data lakes are foundational components of modern data architectures. These systems provide a centralized repository for analytics, reporting, and machine learning, yet they are often plagued with issues related to poor data quality. As the data is collected from multiple source systems, the data needs to be cleaned and updated before it is loaded to data warehouse. The process of identifying errors in data, correcting the errors, and replacing the missing data with an acceptable information is data cleansing [3]. Data is extracted, transformed and loaded to the data warehouse using a process called Extract Transform and Load (ETL). Studies shows that 40% of the data collected needs cleansing as it can be erroneous or not defined in the other system as needed for a common model of intelligence architecture [4]. This process of data cleansing can be accomplished either manual or by technological method. These data issues include inconsistencies, missing values, duplicates, and data entry errors, which can lead to inaccurate insights if they are loaded to the data warehouse or data lakes without cleaning the data and hinder effective decision-making.

Artificial Intelligence (AI) presents a powerful opportunity to address these challenges. By leveraging machine learning, natural language processing, and other AI techniques, organizations can automate the detection and correction of data issues, ensure higher quality data, and improve the performance of data storage systems.

2. The Importance of Data Quality in Data Warehouses and Data Lakes

Data quality refers to the condition of data based on factors like accuracy, completeness, consistency, reliability, and timeliness [5]. High-quality data is essential for making informed decisions, ensuring business operations run smoothly, and driving strategies that lead to success. Poor data quality can lead to errors, inefficiencies, and misguided decisions. High-quality data is essential for business intelligence (BI) systems to produce reliable insights for the accuracy of decision making. Maintaining high data quality ensures that organizations comply with regulatory requirements and protect sensitive data, ensuring data governance and compliance easy. Poor data quality can lead to inefficiencies in processing, analyzing, and decision-making which leads to operational Efficiency [6]. It is essential to understand the data quality issues and implement the needed data cleansing practices.

Data warehouse systems or data lakes collect the data from several source systems. As a first step data quality needs to be initiated in the source systems first. Data is transferred to data warehouse through ETL process. As a second step data quality metrics needs to be incorporated in the steps of data transfer. According to Zellal, N., & Zaouia, A. data quality is a multidimensional concept [7],

The main dimensions of data quality that are often considered include:

- **Accuracy**: In transactional systems for example in healthcare the Electronic Medical Records systems, data accuracy refers to the data which represents the must reflect the real-world entities it represents and as defined in the data model. It can also be defined as the measure of correctness of data which presents the authoritative source [7]. In the context pf data warehouse, the accuracy is defined as the subjectivity of data with respective to its definitions in the source system and also as defined in the data quality dimensions.
- **Completeness**: The source systems need to define the completeness with attribute values to be mandatory, optional, or inapplicable. Based on the definitions the data must be present for those attributes that are defined as mandatory to make the data set complete. In the data warehouse context, the necessary data needs to be present to satisfy the user requirements [6].
- **Consistency**: The consistency of data in the source system is defined as the data always present in the same format and compatible with previous format. Another quality is data to be of same type and similar value when they are stored by different applications or systems which makes the data equivalent. In the context of data warehouse there must be a single presentation of same data across the various data sets as defined by the business rules [7].
- **Timeliness**: Timeliness is defined as the time expectation for accessing the data to make it available at a defined timeline, in other words in the source system data should be updated as per the real-world events that impacts that data. For data warehouse perspective the data updated or populated in the data warehouse by a time period like day, week or a month can not be changed. This helps the decision maker to understand the data as on that time of the data creation [7].
- **Relevance**: The source system data capture needs to be relevant to the real-world definition of the data even in the system. The data model must be defined, and the data to be captured as relevant to the business needs. In the data warehouse systems data captured needs to be relevant to the business requirements also to be relevant as on the time frame defined when it is combined with several source systems when it is made as single entity.
- **Accessibility**: In source systems data is protected by different roles and privileges. Defining the roles and privileges to make data easily accessible by the right business owner with right data that they may need. Datawarehouse systems need to reflect the definitions of source system data to make the data available, or easily and quickly retrievable by several role players.

3. Best Practices to Ensure data quality

Ensuring data quality is essential for accurate reporting, effective decision-making, and reliable business insights.

3.1 Implement Data Profiling

Before using datasets for any purpose, it is important to know the dataset and its metadata. The process of finding metadata is called data profiling [8]. Data profiling involves analyzing and understanding the data's structure, content, and relationships. By profiling the data, you can identify anomalies, missing values, data patterns, and discrepancies before they become bigger issues. Regular data profiling helps spot errors early in source systems or during the ETL process. According to Jang et al, Profiling means figuring out possible mistakes in data by looking at data numbers and patterns. There are mainly three types of data profiling methods: column profiling, single-table profiling, and cross-table profiling. This research focuses mostly on column and single-table profiling methods [9].

3.2 Data Cleansing (ETL Process)

Data cleansing should be an integral part of the ETL process to detect and correct errors in the data before it reaches the data warehouse. Identify and eliminate duplicate records, Fix inconsistent data formats like Standardize date formats, currency, and other units of measurement, use automated tools or rules-based systems to flag and correct common errors, such as typos or invalid data entries [10]. Data Cleaning is a gradual process that goes through steps such as Data Analysis, Definition, Data Mapping Rules, Transformation, Verification, and Backflow Data Analysis [11].

3.3 Data Validation

Create validation rules to check the accuracy and consistency of data as it enters the system and throughout its lifecycle. Ensure data coming from source systems meets quality requirements (e.g., checking if a phone number matches the expected format or if an email address is valid). During ETL, verify that transformations (e.g., aggregations, calculations) are applied correctly. Ensure the data complies [12,13] with specific business rules (e.g., a "sales" record should have a valid customer ID or product code).

3.4 Automate Data Quality Monitoring

Set up automated data quality monitoring systems that continuously check data quality in real-time. Use tools that can monitor for inconsistencies, missing data, or unusual patterns. Implement triggers or alerts that notify data stewards of potential issues before they affect the data warehouse or downstream systems [14].

3.5 Data Governance

Establish a robust data governance framework that defines roles, responsibilities, and processes to maintain data quality [15]. Assign data owners or stewards responsible for overseeing the quality of specific data domains. Designate data stewards to monitor, manage, and correct data quality issues. Maintain metadata to ensure transparency about data sources, transformations, and lineage. Conduct regular data quality audits to identify and address ongoing data issues. Perform checks periodically to validate whether data still adheres to quality standards [16]. Use the audit results to improve the data governance framework and the overall data management process.

3.6 Data Lineage Tracking

Tracking data lineage ensures that data is traceable from its original source through all transformations to the data warehouse [17]. Knowing the origin of data and the steps it has gone through helps identify where issues might arise and makes it easier to correct errors. Implement lineage tracking for both data flow and transformation logic.

4. AI Tools and Techniques for Improving Data Quality

AI tools can significantly enhance data quality by automating tasks, improving accuracy, and identifying patterns that may be missed by traditional methods. AI-driven solutions bring a range of capabilities that make data management more efficient, scalable, and intelligent. Here's how AI tools can help improve data quality: AI offers a wide range of tools and techniques that can automate and improve the process of managing data quality. Below are key AI techniques used in enhancing data quality:

4.1. Data Cleansing and Preprocessing

The primary tasks of data cleansing is detecting the error, where various violations of data rules are identified. Another step in the process is to repair the data [18]. One of the primary ways AI can improve data quality is through automatic data cleansing and preprocessing. AI-based tools can:

- Detect and Correct Errors: Machine learning models can be trained to identify common data entry mistakes, such as typos or incorrect formatting. By learning from historical data, these models can automatically correct inaccuracies in datasets.

- Handle Missing Data: AI models can predict missing values based on patterns in the existing data. Imputation techniques, such as regression-based methods or k-nearest neighbor (KNN), are commonly used.
- Remove Duplicates: AI-powered algorithms can automatically detect duplicate records across data sources and merge them into a single, consistent record.

4.2. Anomaly Detection

Anomaly detection, which is called outlier detection, finds uncommon events or patterns that are different from the normal behavior expected in a system or dataset. Data quality anomalies related to six quality dimensions: Accuracy, Consistency, Completeness, Conformity, Uniqueness, and Readability [19]. Conventional data cleaning tools are used to fix problems with data quality. Yet, these tools might not be enough to find hidden issues that require more advanced and smart methods for identifying them. AI can play a critical role in identifying anomalies or outliers within data, which can often be indicative of poor data quality. Unsupervised learning techniques, such as clustering, can detect data points that deviate significantly from normal behavior. For example, an AI model could identify data entry errors such as Outliers might indicate erroneous inputs or system glitches. Fraudulent activities such as unusual patterns may point to potential fraud or security breaches.

Ilyas, I. F., & Chu, X. proposed a model called Isolation Forest is an algorithm for learning without supervision that finds anomalies in regular data by making isolation trees. It is a decision tree-based method that creates several binary trees. These trees split data by randomly choosing a feature and a value from that feature's range. The splitting goes on until all data points are individually separated from others. This results in shorter paths in the trees for anomalies, as these outliers need fewer splits to be isolated compared to normal data points, making them easier to spot. By using this splitting method, Isolation Forest can quickly find outliers without needing to create a normality model, unlike many other outlier detection methods.

By flagging anomalies in real-time, AI can help prevent the use of faulty data in business decisions.

4.3. Data Validation and Consistency Checking

AI techniques can be applied to automatically validate data across different sources, ensuring consistency. Data validation involves ensuring that data entries adhere to predefined rules, such as:

- Format Consistency: Ensuring that fields, such as phone numbers or dates, follow a consistent format.
- Range Validation: Checking that numerical data lies within acceptable [20] ranges (e.g., age should not be negative).
- Cross-Source Consistency: AI can compare data across different systems to verify that data values are consistent and accurate.

4.4. Natural Language Processing (NLP) for Unstructured Data

Data lakes often contain large amounts of unstructured data, such as text documents, images, and social media feeds. NLP techniques can help process this unstructured data and improve [20] its quality by:

- Text Mining: Identifying and extracting meaningful information from large volumes of text data, filtering out irrelevant or noisy information.
- Sentiment Analysis: Determining the sentiment or opinion expressed in text, useful for business analytics involving customer feedback.
- Named Entity Recognition (NER): Identifying key entities, such as names, locations, and organizations, in unstructured text to ensure accurate data categorization [20].

4.5. Data Profiling and Metadata Management

AI can assist in data profiling, which involves assessing the quality of data in terms of its completeness, uniqueness, and conformity to expected standards. Machine learning models can analyze data and generate comprehensive reports on the current state of the data, identifying potential quality issues [21]. Moreover, metadata management tools powered by AI can help organize and manage data more effectively, enhancing the overall data quality framework.

4.6 AI enabled Data Quality Tools

AI-enabled data quality tools leverage machine learning and other AI technologies to automate and enhance data cleaning, validation, profiling, and enrichment. These tools help organizations ensure the accuracy, completeness, consistency, and reliability of their data. Here are some of the top data quality tools:

- **Trifacta**: Trifacta uses machine learning to analyze data and recommend transformations for cleaning and enriching the data [22]. It offers a user-friendly interface for data wrangling and supports automation of data prep tasks. The tool helps in data preparation, profiling, and automation for ETL- [23] (extract, transform, load).
- **Talend Data Quality**: Key Features: Talend uses AI to automate data profiling, monitoring, and cleansing. It offers real-time data integration and quality assessments across cloud and on-premises environments [23]. Examples like, data governance, data integration, and data cleansing.
- **Informatica Data Quality**: Key Features: Informatica's platform integrates AI and machine learning to help with data profiling, anomaly detection, and automated data cleansing. The platform includes real-time monitoring and workflow management for ongoing data quality checks [24]. Examples like, data governance, enterprise data management, and compliance.
- **IBM InfoSphere**: Information Server: Key Features: IBM's AI-enabled tool helps automate data profiling, cleansing, and validation. It also offers data lineage, monitoring, and integration with machine learning for pattern detection and anomaly identification [25]. Some of the use cases are, data governance, integration, and data stewardship.
- **Microsoft Azure Data Quality Services (DQS)**: Key Features: Microsoft Azure DQS leverages AI and machine learning algorithms for data profiling, data cleansing, and standardization. It helps ensure data quality within Azure-based environments [26]. Some of the use cases are, data quality management within cloud applications and databases.
- **SAS Data Quality**: Key Features: SAS's solution integrates machine learning algorithms for predictive data quality, anomaly detection, and data profiling. It provides comprehensive data governance and data quality management capabilities [25]. Some examples are, advanced analytics, data governance, and predictive data management.

5. Benefits of Using AI for Data Quality Improvement

Using AI to make data better can have many good effects, especially now when data is very important for good decision-making. AI helps all parts of the data process, from gathering to looking at it, by making tasks easier, finding patterns, and ensuring consistency.

- Automation and Efficiency: AI can automate many of the time-consuming tasks related to data cleaning, validation, and monitoring, significantly reducing the need for manual intervention.
- Scalability: AI tools are capable of handling massive datasets at scale, making them particularly suitable for data lakes, which often contain petabytes of raw, unstructured data.
- Accuracy: By learning from patterns in the data, AI can provide more accurate insights and detect issues that traditional methods might miss.
- Real-time Monitoring: AI-powered systems can continuously monitor data streams, providing real-time feedback on the quality of data being ingested into data warehouses and lakes.

6. Challenges and Considerations

While AI tools offer significant advantages in improving data quality, there are several challenges to consider. AI tools may need to access sensitive data, which requires robust data security measures to prevent unauthorized access.

Many AI models, particularly deep learning, are considered "black boxes" and may lack transparency. This can be an issue in contexts where decision-making requires understanding the rationale behind predictions. Implementing AI tools in existing data infrastructures can be complex, requiring customization and integration with legacy systems. AI models are prone to biases in the data they are trained on, which could perpetuate existing data quality issues.

7. Conclusion

AI tools are revolutionizing data quality management in both data warehouses and data lakes. By automating processes such as data cleansing, anomaly detection, and data validation, AI can significantly enhance the accuracy, completeness, and consistency of data. However, challenges like integration complexity, data privacy, and model interpretability need to be carefully managed to maximize the benefits of AI. As data volumes continue to grow and organizations depend more heavily on data-driven decisions, the role of AI in improving data quality will only become more essential. In the future, AI will continue to evolve, offering even more advanced techniques for improving data quality and making data storage systems like data lakes and data warehouses more efficient, secure, and intelligent.

References

[1] Efendi, T. F., & Krisanty, M. (2020). Warehouse Data System Analysis PT. Kanaan Global Indonesia. International Journal of Computer and Information System (IJCIS), 1(3), 70-73.

[2] Hlupić, T., Oreščanin, D., Ružak, D., & Baranović, M. (2022, May). An overview of current data lake architecture models. In 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 1082-1087). IEEE.

[3] Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 international conference on management of data (pp. 2201-2206).

[4] T. Z. Ali, T. M. Abdelaziz, A. M. Maatuk and S. M. Elakeili, "A Framework for Improving Data Quality in Data Warehouse: A Case Study," 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, 2020, pp. 1-8, doi: 10.1109/ACIT50332.2020.9300119.

[5] C. Cichy and S. Rass, "An Overview of Data Quality Frameworks," in IEEE Access, vol. 7, pp. 24634-24648, 2019, doi: 10.1109/ACCESS.2019.2899751.

[6] Benkhaled, H. N., & Berrabah, D. (2019). Data Quality Management For Data Warehouse Systems: State Of The Art. JERI.

[7] Zellal, N., & Zaouia, A. (2016, October). A measurement model for factors influencing data quality in data warehouse. In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 46-51). IEEE.

[8] Abedjan, Z., Golab, L., & Naumann, F. (2016, May). Data profiling. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE) (pp. 1432-1435). IEEE.

[9] Jang, W.-J., Lee, S.-T., Kim, J.-B., & Gim, G.-Y. (2019). A Study on Data Profiling: Focusing on Attribute Value Quality Index. Applied Sciences, 9(23), 5054. https://doi.org/10.3390/app9235054

[10] Swapna, S., Niranjan, P., Srinivas, B., & Swapna, R. (2016, March). Data cleaning for data quality. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 344-348). IEEE.

[11] Fatima, A., Nazir, N., & Khan, M. G. (2017). Data cleaning in data warehouse: A survey of data pre-processing techniques and tools. Int. J. Inf. Technol. Comput. Sci, 9(3), 50-61.

[12] Gao, J., Xie, C., & Tao, C. (2016, March). Big data validation and quality assurance--issuses, challenges, and needs. In 2016 IEEE symposium on service-oriented system engineering (SOSE) (pp. 433-441). IEEE.

[13] Di Zio, M., Fursova, N., Gelsema, T., Gießing, S., Guarnera, U., Petrauskienė, J., ... & Walsdorfer, K. (2016). Methodology for data validation 1.0. Essnet Validat Foundation.

[14] Ehrlinger, L., & Wöß, W. (2017, October). Automated Data Quality Monitoring. In ICIQ.

[15] Koltay, T. (2016). Data governance, data literacy and the management of data quality. IFLA journal, 42(4), 303-312.

[16] Barker, J. M. (2016). Data Governance: the missing approach to improving data quality. University of Phoenix.

[17] Tang, M., Shao, S., Yang, W., Liang, Y., Yu, Y., Saha, B., & Hyun, D. (2019, April). Sac: A system for big data lineage tracking. In 2019 IEEE 35th International Conference on Data Engineering (ICDE) (pp. 1964-1967). IEEE.

[18] Ilyas, I. F., & Chu, X. (2019). Data cleaning. Morgan & Claypool.

[19] Widad, E., Saida, E., & Gahi, Y. (2023). Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis. IEEE Access.

[20] Kiefer, C. (2016, September). Assessing the Quality of Unstructured Data: An Initial Overview. In LWDA (pp. 62-73).

[21] Aikoh, K., Isoda, Y., & Sugimoto, K. (2020, October). Data profiling method for metadata management. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 779-780). IEEE.

[22] Mba, C. C. (2021). Evaluation of data quality tools (Doctoral dissertation, Politecnico di Torino).

[23] Petrova-Antonova, D., & Tancheva, R. (2020). Data cleaning: A case study with OpenRefine and Trifacta Wrangler. In Quality of Information and Communications Technology: 13th International Conference, QUATIC 2020, Faro, Portugal, September 9–11, 2020, Proceedings 13 (pp. 32-40). Springer International Publishing.

[24] Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. B. (2019). Data quality in ETL process: A preliminary study. Procedia Computer Science, 159, 676-687.

[25] Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. Frontiers in big data, 5, 850611.

[26] Eswararaj, D. (2023). Developing a Data Quality Framework on Azure Cloud: Ensuring Accuracy, Completeness, and Consistency. International Journal of Computer Trends and Technology, 71(5), 62-72.