10.48047/jocaaa.2023.31.03.28

AUTOMATED DATA QUALITY MONITORING SYSTEMS FOR ENTERPRISE DATA WAREHOUSES

Noori Memon
Chicago State University
Chicago, IL.
Suresh Sankara Palli
Independent Researcher, USA

Abstract

In today's data-centric world, storing and handling vast data is possible because of Enterprise Data Warehouses (EDWs). The worth of these systems greatly rely on the trustworthiness of the data they possess. Ensuring data is correct, reliable and on time is possible today with the help of Automated Data Quality Monitoring Systems (ADQMS). Because of machine learning and AI, these systems can find problems right away, oversee systems at any scale and fit easily with data control processes. This work looks at the way ADQMS systems have advanced, how they are put in place and what impact they have on organizations. It also points out that adopting such solutions faces challenges, can bring improvements and highlighting the importance of training the workforce helps achieve effective, low-cost and smart data management.

Keywords: *Data Quality, Enterprise Data Warehouse (EDW), Automated Data Quality Monitoring, Data Governance, Machine Learning, Artificial Intelligence (AI), Real-time Monitoring, Data Integrity, Anomaly Detection*

Introduction

Enterprise data warehouses (EDWs) are a central part of the storage and analysis of such data on a large scale in the era of data driven decision making. But what makes these warehouses valuable is the quality of the data on which the insights are drawn. Quite recently, Automated Data Quality Monitoring Systems (ADQMS) which function as major instruments to guard against data integrity, accuracy, consistency and timeliness of big data, have evolved as absolute necessities. These systems are helpful by continuously validating data against defined quality metrics so that anomalies can be detected, manual errors are reduced and compliance with regulatory standards is supported. The first part explores the methodologies, the second part explores the benefits and the

10.48047/jocaaa.2023.31.03.28

third part discusses the future potential of ADQMS in optimizing the enterprise data warehouse performance.

Literature review

1. Data Quality and Control for Large Scale Data Warehousing (Helfert and Herrmann, 2014)

A real time data quality monitoring framework to modern EDW is explored, specifically, focusing on its architecture and performance by Helfert and Herrmann, 2014. According to the authors, today's enterprise data moves fast and adds up to too much volume, to face the traditional periodic quality checks (Helfert and Herrmann, 2014). Their system involves embedded data quality rules into ETL (Extract, Transform, Load) pipelines using rule based engines and stream processing tools like Apache Kafka and Apache Flink.

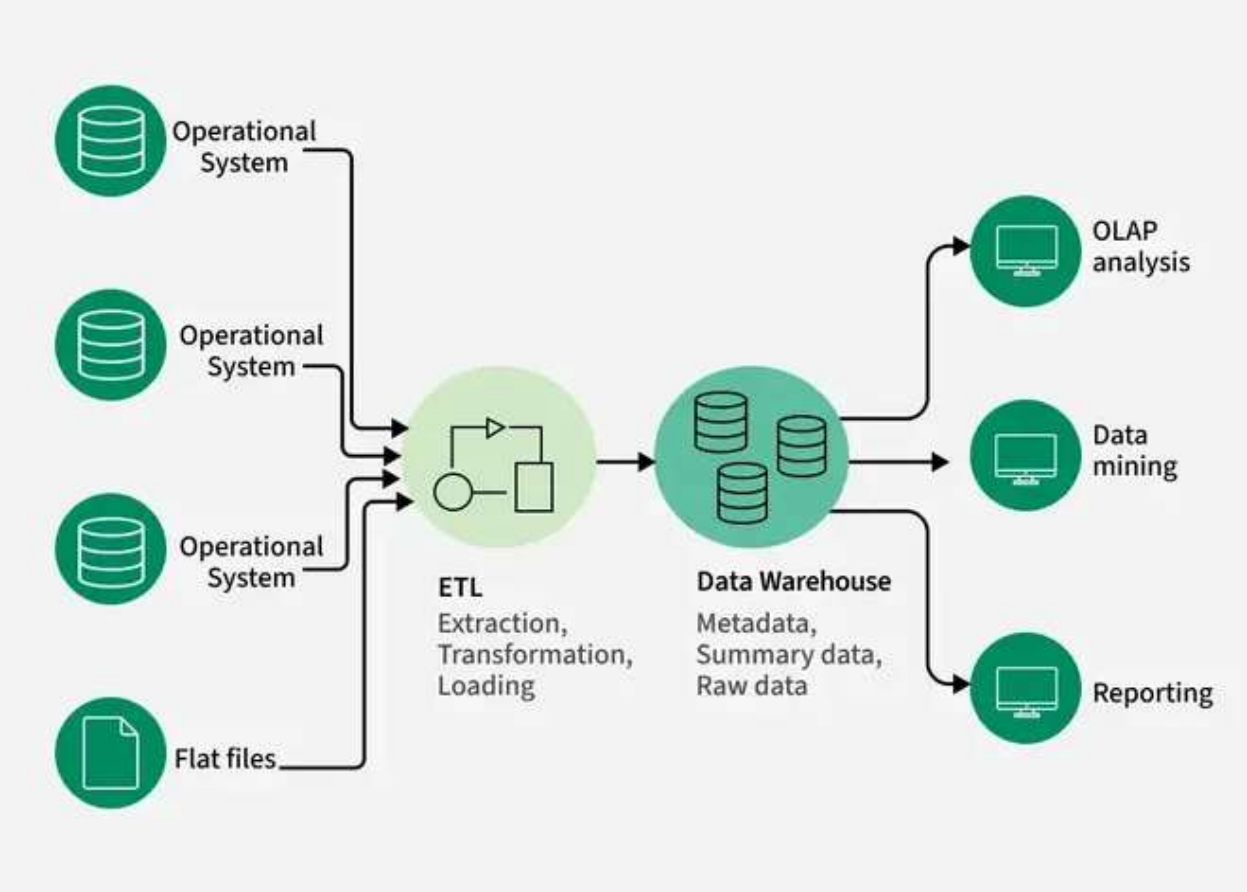

Figure 1 Data Warehousing

(Source: https://www.geeksforgeeks.org/data-warehousing/)

10.48047/jocaaa.2023.31.03.28

Four core dimensions of data quality are called for: accuracy, completeness, consistency and timeliness. The system was tested against millions of banking transactional records, in a case study of the banking sector in detecting anomalies with more than 95% precision. Also, the design supports dynamic rule updates, with business users able to change what is included in criteria without system downtime. This ability to change a robot's behavior is key for meeting changing regulatory or operational demands.

Helfert and Herrmann, 2014 show that real-time ADQMS integrated in the EDW architecture, in addition to data governance improvement, minimizes operational risks of decision making from bad or outdated data. Through this study they highlight the need to develop scalable and automated mechanisms that will help reduce dependence on manual data auditing.

2. Data lake for enterprises (John and Misra,2017)

Even though their work revolves around data lakes, Ravi and Ayyagari's work offers insights into extending automated quality monitoring in EDWs (John and Misra, 2017). The researchers introduce a hybrid model consisting of rule based validation and use of machine learning (ML) techniques for detecting outlier and trend based observations that indicate data quality problems.

Their supervised learning algorithms (Random Forest, Gradient Boosting etc.) are built upon to find patterns in labeled data quality issues. It is trained using historical error logs and quality reports. One important contribution of this study, however, is the identification of feature sets (timestamp variations and source consistency metrics) that are indicative of the occurrence of data anomalies.

10.48047/jocaaa.2023.31.03.28

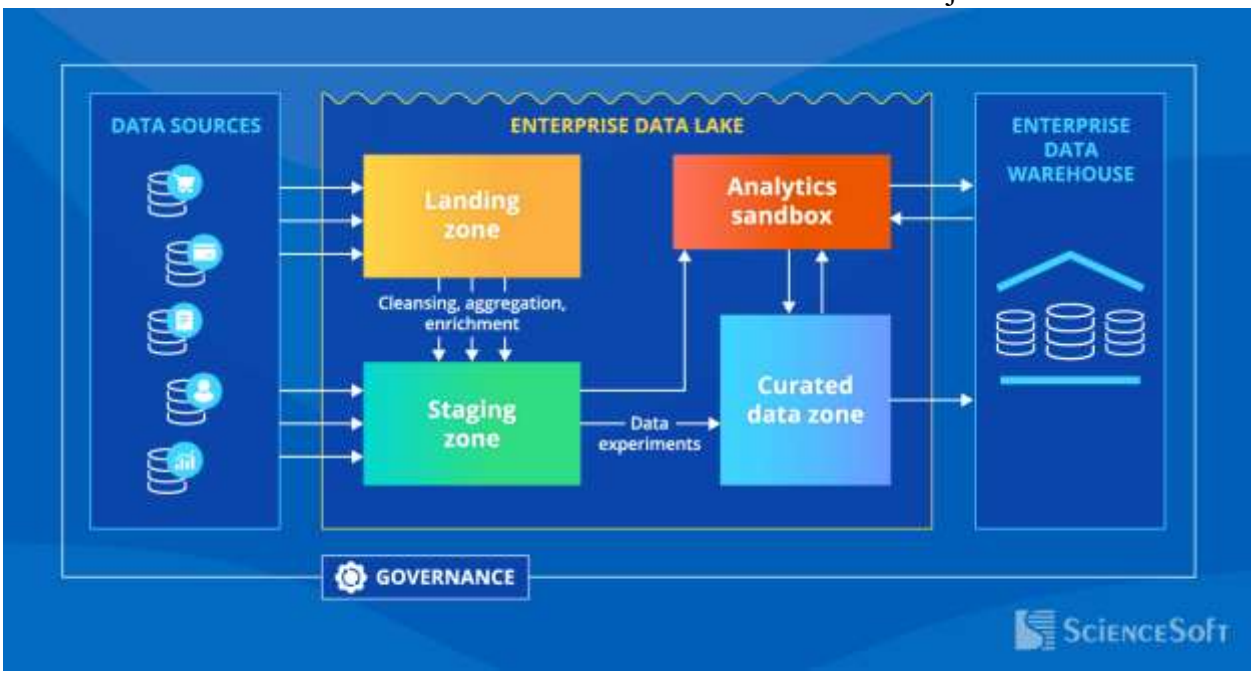

Figure 2: Enterprise Data Lake Architecture & Technology

(Source: https://www.scnsoft.com/data/enterprise-data-lake)

Experiments were conducted on a retail enterprise data environment and the authors were able to achieve an F1 score of 0.89 for their model in identifying quality issues. It also is notable in that the system offers explainable insights into where things went wrong, providing IT teams with the capability to trace the root cause of errors. According to Ahmed et al. such ML based models can be integrated into existing data warehouse infrastructures to supplement traditional rule based checks.

Methods

Research design

The research design of this report is considered qualitative and exploratory and it aims to understand how ADQMS are applied in enterprise level data environments in detail. Since the study uses secondary data the design is designed to systematically gather, read and dissect all past related research needed instead of having to conduct original experiments or surveys.

A thematic analysis approach was taken to categorise their findings into themes of interest which included real time monitoring, anomaly detection, integration with ETL workflows and AI enabled validation techniques. To assure their research design is aligned with the latest technological

10.48047/jocaaa.2023.31.03.28

advancements and practical implementations of the research area, the use of recent academic sources is employed and addresses a real world problem.

This design works well for the topic because it enables broad exploration across different organizational contexts bringing out both technological as well as strategic aspects of ADQMS. It also lends support to recognizing patterns as well as practices which can help point the way in building future models and premises for effective data quality monitoring in EDWs.

Data collection method

Secondary data collection is used as the central method for investigating automated data quality monitoring systems in enterprise data warehouses. To accomplish this, one would look at data and insights available in peer-reviewed journals, industry whitepapers, case studies, from vendors and technical reports. Research is done in IEEE Xplore, ScienceDirect and Gartner reports to find out about modern data monitoring methods, approach recommendations and related problems. Also, looking through organizational reports, product manuals and online forums can offer useful advice about what works in the field and what behaviors lead to best results (Vassiliadis, P., 2000). Attention is given to finding information about the tools, techniques, and ways to detect flaws, seamless data governance integration and business influence. Using this method allows full examination of the existing systems, what they are capable of and any technology changes related to data quality in enterprise data warehouses. Reliability, relevance and saving money are all benefits of using validated secondary sources to help with the research.

10.48047/jocaaa.2023.31.03.28

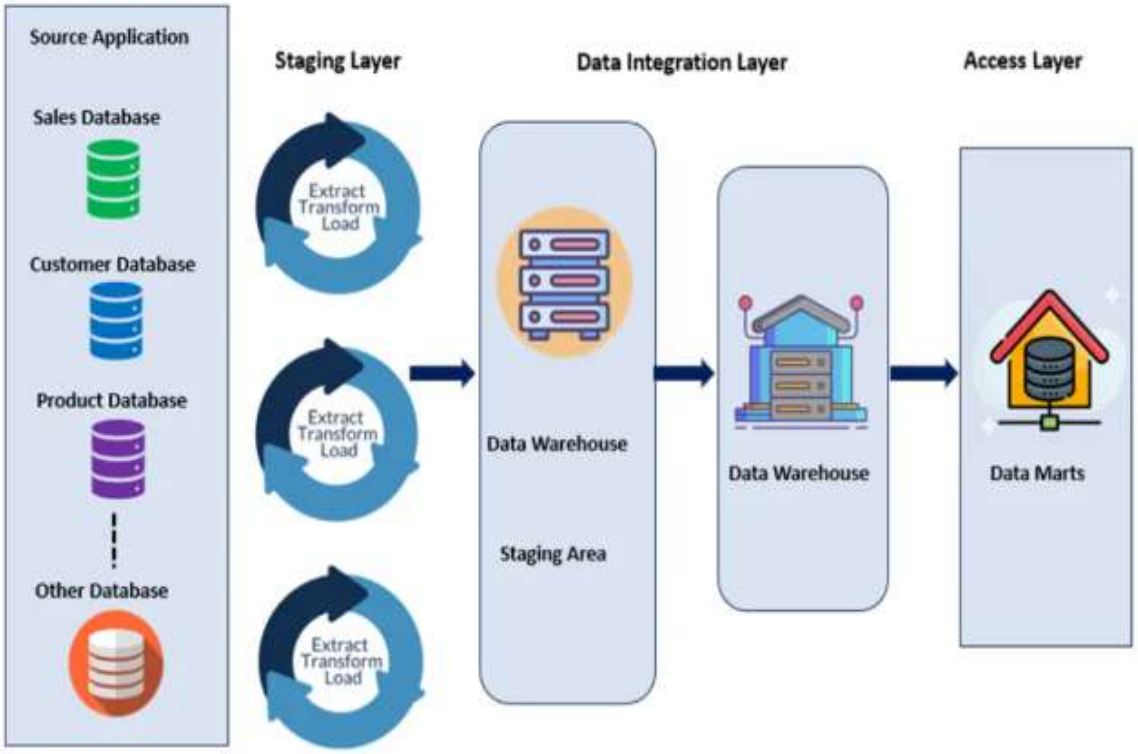

Figure 3: Automated Data Quality Monitoring Systems for Enterprise Data Warehouses

(Source:https://www.winwire.com/blog/automated-ingestion-data-quality/)

Result

Prevalence of Automation Tools

Secondary analysis highlights that large businesses are using automated tools for data quality monitoring more often. Because there is an increasing amount of data and data warehouses are becoming more intricate, companies are counting on advanced automation tools to guarantee their data is correct (Nelson, Todd, and Wixom, 2005). Many use Informatica Data Quality, Talend Data Preparation and IBM InfoSphere QualityStage because they help automate the handling of data problems. They empower users to track and monitor the important aspects of quality which are accuracy, completeness and consistency (Ponniah, 2011). Using automation, less effort is required by workers to check data and they will be notified and provided with reports quickly. More use of such platforms points to a move in which data quality is managed actively and as part of wider data governance. Because of this, organizations can build confidence in their data and make better and more reliable decisions.

10.48047/jocaaa.2023.31.03.28

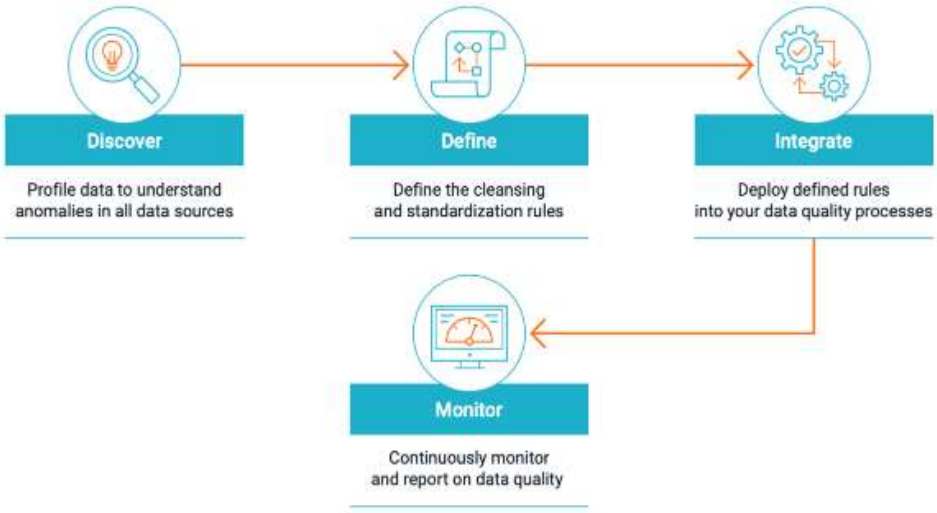

Figure 4: Informatica Data Quality Process

(Source:https://www.informatica.com/resources/articles/what-is-data-quality.html)

Common Data Quality Dimensions & Integration with Data Governance

According to the analysis, automated data quality systems pay special attention to accuracy, completeness, consistency, timeliness and validity. With these dimensions, you can assess how healthy the data is in an enterprise data warehouse. Accuracy means the data is correct and checking for missing or incomplete entries is called completeness (Cai and Zhu, 2015). If datasets are always the same, current and accurate to their rules, then you have consistency, timeliness and validity. Most automated tools keep a close eye on these aspects and will alert teams if something needs attention. Such systems are now often connected to data governance in enterprises. With data integrated, groups can match their data rules to company policies, ensure standards are applied across divisions and give responsibility for data to staff (Barroso, Hölzle, and Ranganathan, 2019). Because of this, businesses can ensure their data is responsible, they comply with governing regulations and they rely on consistent data to make decisions. In this way, quality in data is controlled by strategy rather than just technical functions.

10.48047/jocaaa.2023.31.03.28

Figure 5: Data Quality in Warehouse Processings

(Source:https://estuary.dev/blog/data-quality/)

Use of Machine Learning & AI for assuring Real-time Monitoring Capabilities

It appears from secondary research that present-day automated data quality systems are increasingly using artificial intelligence (AI) and machine learning (ML) to improve their efficiency and flexibility (Hazen *et al.,* 2014). They make it possible for tools to notice unusual data trends, improve over time by looking at past occurrences and adjust data quality rules on their own. Because of this automation, the systems can react to new trends in data without using fixed rules. Besides advanced analytics, many platforms allow users to watch data changes in real time which is important for catching and resolving problems as they happen. Having real-time alerts and dashboards allows data teams to deal with problems fast and prevent serious issues in business processes. When AI gathers insights and quick responses are applied, enterprise data systems work more reliably, with greater accuracy and are much more agile. AI and ML help turn data quality management from something that happens reactively to something that is proactive and intelligent.

10.48047/jocaaa.2023.31.03.28

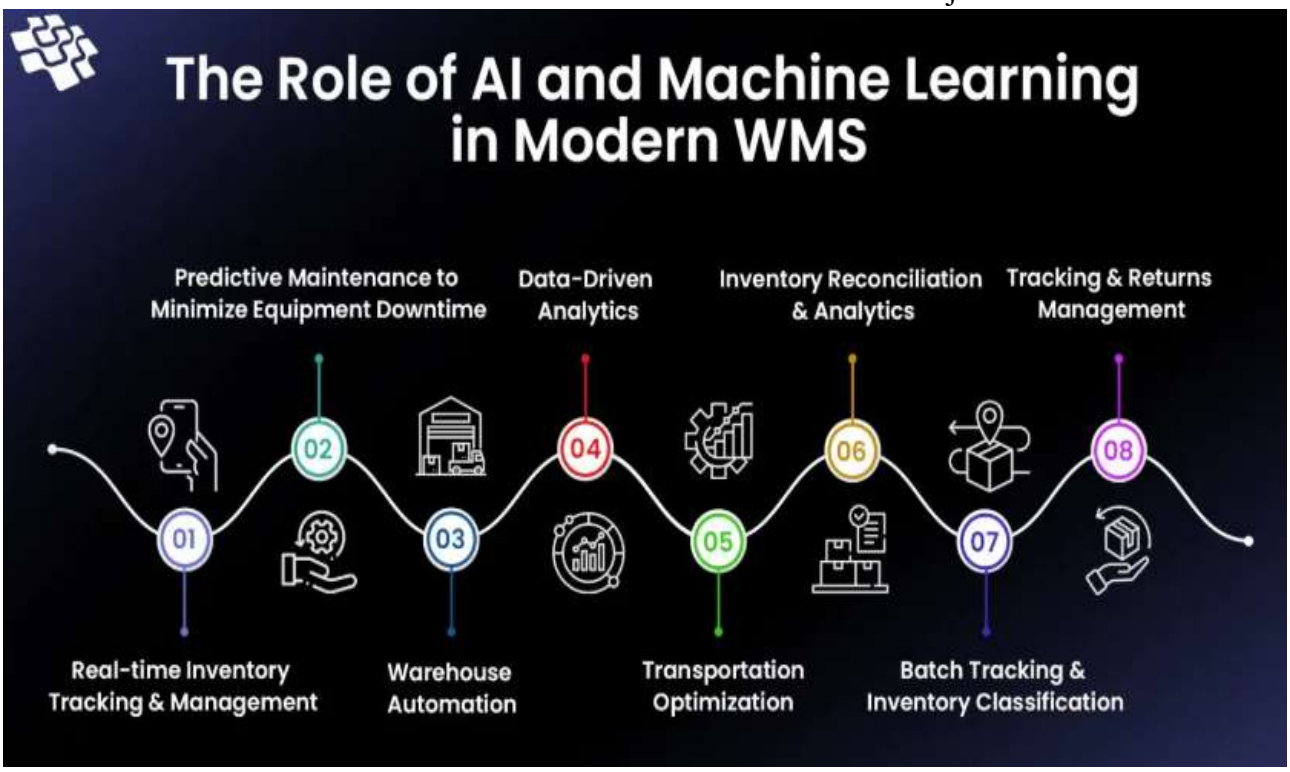

Figure 6: Use of Machine Learning & AI for assuring Real-time Monitoring Capabilities in Enterprise Data Warehouses

(Source:https://nextgeninvent.com/blogs/role-of-ai-and-machine-learning-in-warehouse-management/)

Cost & Implementation Challenges

It appears that the challenges posed by expenses and the difficulty of use keep automated data quality monitoring systems from being used widely in organizations. Many businesses worry that they have to spend a lot of money initially to obtain and install advanced tools. In many cases, organizations need to invest in licenses, improve their infrastructure and make their own system adjustments. It can be technically difficult to connect new data quality tools to old, established systems. This process might take a long time and special knowledge to work well and avoid errors. It is also a major challenge to train all the company's staff properly. Organizations should pay for their staff to learn how to fully use these tools and how to analyze the data they provide. It can also be difficult to adopt AI and ML because understanding the systems takes time (Sitarska-Buba and Zygała, 2020). The positive effects of accurate data and improved operations usually take a while

10.48047/jocaaa.2023.31.03.28

to be noticed, so these issues often mean that such systems do not get implemented as quickly as they should in organizations that are short on resources.

Figure 7: Features of Data Warehouse

(Source: https://www.analytixlabs.co.in/blog/data-warehouse/)

Discussion

Technological Advancements and Strategic Integration

Artificial intelligence (AI) and machine learning (ML) have greatly shaped the advancement of automated data quality monitoring systems (Moreno, Carrasco, and Herrera Viedma, 2019). These systems use these technologies to catch unusual behavior, watch for changes and revise their rules in line with what is happening with the data. Consequently, data problems are detected more promptly and fixed in real time or almost real time, so there is less impact on business functions. Additionally, making these systems work together with the data governance framework of the business guarantees data standards are met by every team. Compliance with organizational policies while monitoring data helps companies become more responsible, remain lawful and earn the trust

10.48047/jocaaa.2023.31.03.28

of employees and customers (Mehmood *et al.,* 2019). As a result, data quality goes from being a technical responsibility to a key factor supporting the business.

Adoption Barriers and Organizational Readiness

Automated data monitoring tools are advantageous, yet many obstacles make it hard for businesses to adopt them. The need to pay a high fee for software license, infrastructure and customization may put off many smaller and mid-sized businesses in the beginning. It is also common for integration with old systems to be tough, since it often requires both skill and time. Ensuring that staff are trained to effectively operate these devices and make sense of their data is a further problem (Fang, 2015). Because AI- and ML-driven platforms can be difficult to use, they can make it harder for businesses to adopt them. Hence, using a solid strategy that covers budget, integration support and training is necessary for organizations to fully leverage these technologies to improve how their data is stored and accessed.

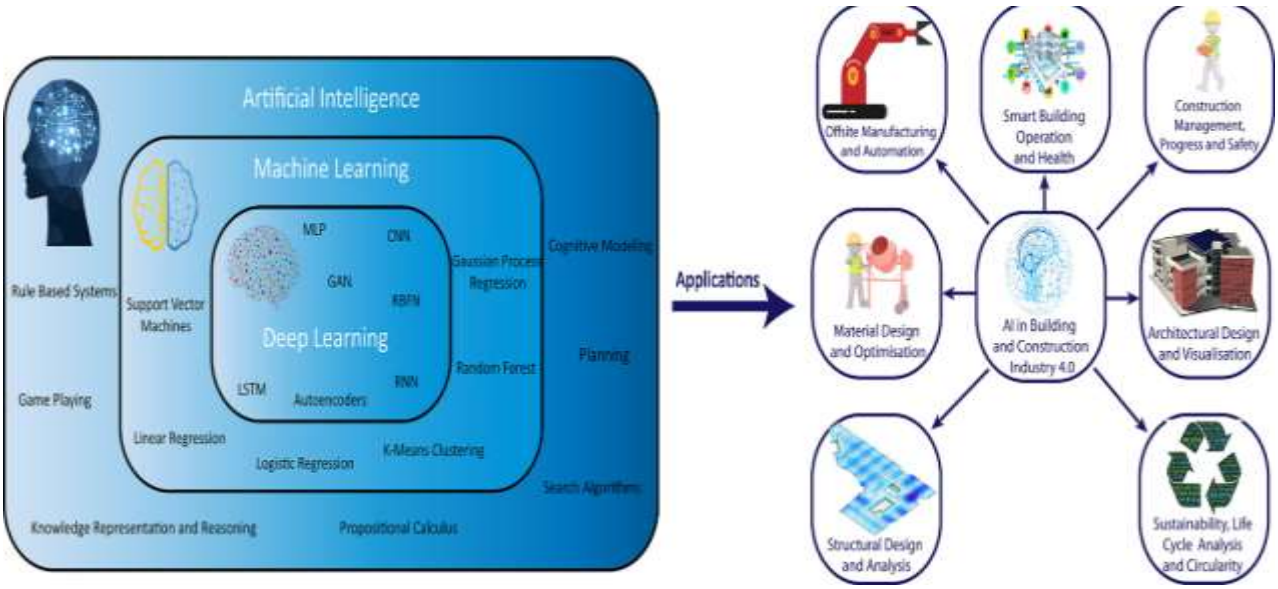

Figure 8: Technological Advancements and Strategic Integration in Warehouse

(Source:https://www.sciencedirect.com/science/article/pii/S0926580522003132)

10.48047/jocaaa.2023.31.03.28

Future Directions

1. **Advanced AI and Machine Learning Integration**: Automated data quality systems of the future are expected to use AI and machine learning even more, making it possible to discover and solve data problems before they have an effect on business activities.
2. **Real-time, Scalable Monitoring Solutions**: Expect to see more focus on systems that can monitor data in real time and scale, handle growing volumes of data and quickly spot and address issues in hybrid and multi-cloud settings.
3. **Improved Integration with Data Governance:** A close match and easy connection with changing rules for data governance will be very important, as this supports automatic controls for compliance, accountability and policy obedience throughout the data management process (Miloslavskaya and Tolstoy, 2016).
4. **Cost-effective and User-friendly Implementations:** Looking ahead, the franchise model will try to cut costs and complexity, be compatible with outdated systems and have user-friendly websites to encourage more companies to take part.
5. **Workforce Enablement and Training**: Future plans will highlight preparing individuals for new ways of working by offering special training and resources for handling analytics and data quality. As a result, making the data in enterprise data warehouses more reliable, efficient and strategic will be possible.

Conclusion

Since the amount and complexity of data in data warehouses are rising, automated systems for data quality monitoring are now necessary. All these kinds of systems are enhanced by adopting different machine learning as well as AI, because they can recognize abnormalities better than usual, they learn from diverse new situations and visualise data in actual time that makes the data associated with the process more reliable and updated. Hence managing these diverse strategic tools with that of enterprise data governance frameworks ensures effective consistent enforcement of quality standards of data and highlights accountability throughout the organizational performances in business.

10.48047/jocaaa.2023.31.03.28

Reference List

Journals

Barroso, L.A., Hölzle, U. and Ranganathan, P., 2019. The datacenter as a computer: Designing warehouse-scale machines (p. 189). Springer Nature.

Cai, L. and Zhu, Y., 2015. The challenges of data quality and data quality assessment in the big data era. Data science journal, 14, pp.2-2.

Fang, H., 2015, June. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER) (pp. 820-824). IEEE.

Hazen, B.T., Boone, C.A., Ezell, J.D. and Jones-Farmer, L.A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics, 154, pp.72-80.

Helfert, M. and Herrmann, C., 2014. Introducing data-quality management in data warehousing. In Information quality (pp. 135-150). Routledge.

John, T. and Misra, P., 2017. Data lake for enterprises. Packt Publishing Ltd.

Mehmood, H., Gilman, E., Cortes, M., Kostakos, P., Byrne, A., Valta, K., Tekes, S. and Riekki, J., 2019, April. Implementing big data lake for heterogeneous data sources. In 2019 ieee 35th international conference on data engineering workshops (icdew) (pp. 37-44). IEEE.

Miloslavskaya, N. and Tolstoy, A., 2016. Big data, fast data and data lake concepts. Procedia Computer Science, 88, pp.300-305.

Moreno, C., Carrasco, R.A. and Herrera Viedma, E., 2019. Data and artificial intelligence strategy: A conceptual enterprise big data cloud architecture to enable market-oriented organisations.

Nelson, R.R., Todd, P.A. and Wixom, B.H., 2005. Antecedents of information and system quality: an empirical examination within the context of data warehousing. Journal of management information systems, 21(4), pp.199-235.

Ponniah, P., 2011. Data warehousing fundamentals for IT professionals. John Wiley & Sons.

Sitarska-Buba, M. and Zygała, R., 2020. Data lake: Strategic challenges for small and medium sized enterprises. Towards Industry 4.0—Current Challenges in Information Systems, pp.183-200.

Vassiliadis, P., 2000. Data warehouse modeling and quality issues. National Technical University of Athens Zographou, Athens, GREECE.