North American Journal of Engineering and Research

Etd. 2020

Data quality frameworks for real-time data pipelines

Girish Ganachari
Email: girish.gie@gmail.com

Abstract

This study analyzes trends in developing data quality frameworks for real-time data pipelines, which are crucial for various industries today. The paper acknowledges the significant benefits of real-time data processing but also highlights potential issues with data quality, integrity, completeness, and timeliness. Existing architectures designed for micro-batch processing and homogenous data sources are still used, despite the challenges they pose in these varied environments. The paper proposes methodologies like machine learning and AI for automating data quality control. Intersystem synergism is therefore central to developing robust solutions for real-time data quality control.

Keywords: Data quality framework, real-time data, data pipelines, real-time.

Introduction

In today's world, organizations have a lot of dependence on the real-time information processing which necessitates handling data at the source. These pipelines are important to make good and suitable decisions in almost every field from finance to healthcare, manufacturing and telecommunication. Data quality in these pipelines is highly critical because low quality data will result to wrong information and thus, lead to operational problems and financial losses. Although data quality is a valuable concept, sustaining it in operational systems for Real-Time data is an even greater challenge since it has to handle with the velocity, and volume in variety.

Research Background

Data quality refers to the characteristics of data, including accuracy, completeness, consistency, timeliness, and validity. In conventional batch processing systems, data quality control is performed at specific intervals during data validation. However, real-time data pipelines process data continuously, demanding more dynamic and reliable approaches. This is crucial because the data itself can sometimes be of low quality. These pipelines are essential for modern applications like stream processing, real-time fraud detection, predictive maintenance, and real-time customer engagement.

Several challenges follow the need to constantly update the quality of data in real-time applications: Real-time processing implies nearly instantaneous data processing, data integration arises due to the incorporation of data of different sources and nature, and volatility refers to the common characteristic of data that fluctuates rapidly in real-time applications. Some of the conventions for data quality, prescriptively available in current DQ frameworks, are developed for batch data processing methods. This has created the need for the formation and implementation of paradigms of early frameworks and methods that incorporate real-time information processing only.

Aims and objectives

Aim

The core aim of the study is on to analyze the data quality framework for the real-time data pipelines and accessing the data pipeline and the quality across.

Objectives

- To study the existing data quality framework.

- To understand the key data quality dimensions for real-time pipelines.
- To access the challenges prevailing in maintaining the data quality in real-time.
- To understand data quality can drive the real-time data pipelines.

Literature Review

To study the existing data quality framework

According to Lindig, et al., (2021), it is stated that there is the emergence of various frameworks for data quality as a result of the increase in quantity and quality of data in organizations. Originally, these frameworks were designed for environments involved in batch processing, where the quality control could be conducted remotely (Lindig, et al., 2021). Popular frameworks like the Data Management Association's (DAMA) model from the USA offer detailed procedures to follow within the data governance, quality, and improvement plans (Mehmood and Anees, 2020). The model that encompasses all these stages is identified as Total Data Quality Management (TDQM) and is continuous in nature. Future developments have however made it possible to incorporate real time data into these frameworks. For example, the Data Quality Management Framework developed by IBM and the Open Data Quality (ODQ) framework use real-time measurements and technique automation for dynamically applied data quality management. These frameworks deal with the quality of data in terms of correctness, integrity, uniformity and timelessness of the data however these frameworks sometimes fail in the live environment as data from different sources come at a large speed and with different formats.

According to Paganelli, et al., (2022), it is stated that although various reference frameworks are useful for a solid start, it is becoming increasingly apparent that more and more specific solutions are required for more specific needs such as those of real-time data pipelines (Paganelli, et al., 2022).

To understand the key data quality dimensions for real-time pipelines

Khan, et al., (2021) mentioned that since quality data is a fact, data quality dimensions are required to map the framework and criterion for measuring it. As mentioned before, in real-time data pipelines, some dimensions are more critical than others because of the procession's timeliness and data volume (Khan, et al., 2021). Accuracy enables the collected information to resemble real-life conditions thus it is appropriate for real-time decision-making.

Raj, et al., (2020), mentioned that completeness as a dimension is the one that deals with absences of data with an aim of avoiding any form of a gap that might lead to carrying out wrong analysis. Consistency makes sure that the different datasets and the sources are similar and standardized that is very important for integrated systems that deal with different data streams (Raj, et al., 2020). Timeliness which had arguably emerged as the most important dimension in a real-time environment establishes how up to date the data is for purposes of action.

To access the challenges prevailing in maintaining the data quality in real-time

According to Mehmood and Anees, (2020), it is given that the concerns regarding data quality in real-time data pipelines are quite different from those of batch processing. Some of them include; Firstly, the amount and rate of data that is being generated is massive and this is overwhelming to conventional data quality management techniques (Mehmood and Anees, 2020). To maintain this high-speed data flow real time checks and corrections are needed which are frequently intricate and call for a lot of computing power. The phenomenon of data heterogeneity due to the combining of various data sources like sensors, social media feeds, and operational systems in the main challenge associated with the maintenance of consistent and accurate data.

Moreover, Taleb, et al., (2021), also gave that real-time environments are becoming more sensitive with the latency and performance issues; with processing time delays, the data can be less up-to-date or even contain missing some parts of information. The trouble shooting needs to be real-time to prevent propagation through the pipeline, implying the need for complicated algorithms and infrastructure (Taleb, et al., 2021). The other challenge is the ability to merge structured and unstructured data because it is very difficult to design frameworks that would monitor different types of data.

To understand data quality can drive the real-time data pipelines

Real-time data consumers strongly depend on high-quality data as one of the key success factors. Credible information improves the dependability of the near real-time analysis that is critical in decision making. In certain industries including, finance, healthcare, and telecommunications, real-time insights make a huge difference hence the quality of the data will always have a positive influence on the performance of the firm and consequently a competitive advantage. Accuracy also has a part to play in improving the standards and avoiding mistakes that lower the standard of data used in decision making. Also, it optimizes the data flow in the pipelines through reducing the frequency of reprocessing as well as corrections on them hence the latency in the system will be reduced making it more efficient. Modern advanced data quality frameworks utilize machine learning and artificial intelligence to perform quality check and adapt to the new format of the data and thus enhance real-life data pipelines. With the growing focus of organizations on real-time data the supremacy of data quality as a core enabler of performance at the operand and strategic level assumes more importance.

Literature Gap

A major gap in the current literature on data quality in real-time data pipelines is the lack of standard frameworks within which the issues of data quality management in high velocity, heterogeneous data systems might be properly contained. There is a lot that can be discovered using existing frameworks, yet they are predominantly focused on the batch processing concepts or vice versa offer insufficient specifics about validating real-time data, controlling latencies, and incorporating different forms of data. More studies can be aimed at constructing integrated, actionable conceptual models that utilize the state-of-the-art approaches, such as machine learning for the ongoing process of data quality monitoring to fill the gap between existing conceptual models and implementation in changing big data environments.

Methodology

The research has used the secondary data collection to get the real, authentic and validated data for the research. All the secondary sources relevant to the data quality frameworks, real-time data and the relevant papers, journals, and articles to deliver validated outcomes. The thematic qualitative analysis of the data is done to understand the data quality frameworks, in the real-time data pipelines to gain valuable outcomes of the research. The research throughout is done ethically without affecting the sentiments of any person or group and within the research code of conduct. The research is within the ethical purview, beginning from the data collection to the final outcomes.

Results and Discussions

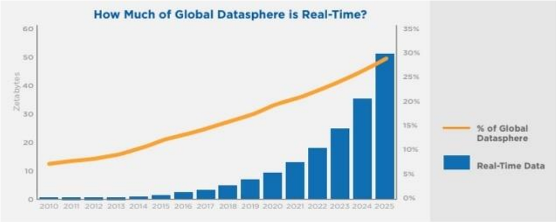

Figure 1: Growth in real-time data

Source: (Condon, 2018)

According to the research, the demand from customers will drive the expansion of real-time data in part: As the digital world of customers becomes more integrated with their physical realities, they anticipate being able to access products and services regardless of where they are, regardless of the connection they have, and on any device, they choose (Condon, 2018). Personalized data that is available in real time and on the go is what they desire.

When it comes to the needs of a company, automated equipment on a production floor is dependent on real-time data for the purpose of process management and improvement, as stated in the research (Lindig, et al., 2021).

All things considered, according to IDC, the majority of the 150 billion gadgets that will be connected throughout the world in 2025 will be producing data in real time (Condon, 2018). It is anticipated that the worldwide datasphere would increase from 23 Zettabytes (ZB) in 2017 to 175 ZB by the calendar year 2025. There are one trillion gigabytes that are comparable to one zettabyte (Condon, 2018).

The data infrastructures of businesses who want to provide an exceptional customer experience and increase their market share need to be able to accommodate the growing volume of real-time data.

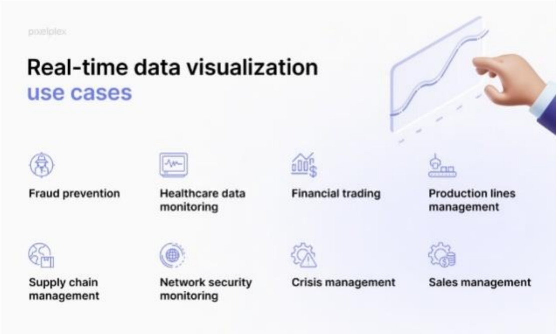
Real-time data visualization
use cases
Fraud prevention
Healthcare data monitoring
Financial trading
Production lines management
Supply chain management
Network security monitoring
Crisis management
Sales management

Figure 2: Realtime data visualization
Source: (Yetchenko, 2024)

From the above image, it is evident that the real-data visualization is multifaceted in every aspect, indicating that the data quality framework ingrained within the data pipelines are driving the ways towards effective fraud prevention, making ease in monitoring the healthcare data, financial tracking and trading, streamlining the production lines, effective supply chains, embrace security and monitoring, efficiency in handling the crisis management, and robust sales management (Yetchenko, 2024). The data pipelines are driven with the data quality framework in managing the data across.

Conclusion

Thus, in conclusion, the study established the significance of data quality frameworks in real-time data pipelines regardless of the industry type. Real-time data processing has potential in decision-making and in saving time yet it has its downside when it comes to data quality. Modern frameworks, that were built to perform batch processing, gradually adapt to the real-time one, but the problem of handling high volume, high variety data still persists.

In terms of the results, it is pointed out that accuracy, completeness, consistency, and punctuality are the significant dimensions for controlling data quality in real-time contexts. The nature of challenges that data quality faces include those arising from the volume, velocity, and heterogeneity of data require robust and agile approaches to managing data quality. Real-time data is instrumental in modern applications that rely on machine learning and AI techniques; thus, real-time data pipelines greatly benefit from using AI-driven quality control and the ability to work with different data formats.

The further studies should be undertaken to create the notions of the IM model that would take into consideration the specific requirements to the management of real-time data quality. These models should make use of high-end technology to solve the problem of latency and improve the quality of data during the ever-evolving data environment.

In summary, as organizations depend on the real-time data for its strategic management and operational intelligence, the organizations should invest in adequate data quality framework for sustaining its competitiveness and innovation.

Recommendations

- Investment in Advanced Technologies: Thus, organizations need to focus on funding the implementation of the machine learning and artificial intelligence technology to improve the monitoring and management of real-time data quality.
- Enhanced Framework Development: From the analyzed researches it can be concluded that more investigation into the frameworks suitable specifically for real time data environments is needed, with the emphasis put on adaptability and extendibility.
- Collaboration and Knowledge Sharing: Promoting cooperation between schools, business organizations, and companies with help of relevant IT suppliers can facilitate development of progressive approaches and techniques concerning real-time data quality.

Future Work

The further research should be conducted in order to create the flexible data quality frameworks that are aligned with the implementation of the contemporary technologies such as blockchain and IoT. Future work could look at the use of research within the context of specific sectors such as healthcare, finance, and manufacturing industries to tackle the issues of real-time data quality management. It is crucial in the long-term examination of such frameworks' efficacy and adaptability to emerging contexts as they are applied on a larger scale. Academic institutions, businesses, and technology suppliers should persist with cooperation in developing more efficacious analytical models, which must be based on easily implemented, real-time quality assurance in data pipelines.

References

[1] Condon, S. (2018). By 2025, nearly 30 percent of data generated will be real-time, IDC says. https://www.zdnet.com/article/by-2025-nearly-30-percent-of-data-generated-will-be-real-time-idc-says/

[2] Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior research methods*, 1-20. https://link.springer.com/content/pdf/10.3758/s13428-021-01694-3.pdf

[3] Khan, F., Yarveisy, R., &Abbassi, R. (2021). Risk-based pipeline integrity management: A road map for the resilient pipelines. *Journal of Pipeline Science and Engineering*, *1*(1), 74-87. https://www.sciencedirect.com/science/article/pii/S2667143321000123

[4] Lindig, S., Moser, D., Curran, A. J., Rath, K., Khalilnejad, A., French, R. H., ... & Luo, W. (2021). International collaboration framework for the

calculation of performance loss rates: Data quality, benchmarks, and trends (towards a uniform methodology). *Progress in Photovoltaics: Research and Applications*, *29*(6), 573-602. https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.3397

[5] Mehmood, E., & Anees, T. (2020). Challenges and solutions for processing real-time big data stream: a systematic literature review. *IEEE Access*, *8*, 119123-119143. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9126812

[6] Paganelli, A. I., Mondéjar, A. G., da Silva, A. C., Silva-Calpa, G., Teixeira, M. F., Carvalho, F., ... &Endler, M. (2022). Real-time data analysis in health monitoring systems: A comprehensive systematic literature review. *Journal of Biomedical Informatics*, *127*, 104009. https://www.sciencedirect.com/science/article/pii/S1532046422000259

[7] Raj, A., Bosch, J., Olsson, H. H., & Wang, T. J. (2020, August). Modelling data pipelines. In *2020 46th Euromicro conference on software engineering and advanced applications (SEAA)* (pp. 13-20). IEEE. https://research.chalmers.se/publication/521248/file/521248_Fulltext.pdf

[8] Taleb, I., Serhani, M. A., Bouhaddioui, C., &Dssouli, R. (2021). Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data*, *8*(1), 76. https://link.springer.com/content/pdf/10.1186/s40537-021-00468-0.pdf

[9] Yetchenko, D. (2024). Real-time data visualization: use, cases, tools, and best practices. Pixelplex. https://pixelplex.io/blog/real-time-data-visualization/