Journal of Information Systems Engineering and Management
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Optimizing Real-Time Data Pipelines for AI-Driven Decision-Making: Architectures, Challenges, and Future Trends

Deepak Chanda
Sr Data Analyst SERCO, INC VA, USA
journalpublications.dc@gmail.com

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Real-time data pipes underlie AI-driven systems for decision-making, enabling instant insights and response in areas such as self-driving vehicles, financial trade, smart cities, and health. The following is an overview of optimizing data pipes in real-time using models such as Lambda, Kappa, and Event-Driven Microservices, and leading technologies such as Apache Kafka, Spark Streaming, and data warehousing in the cloud. The document evaluates primary challenges such as latency, data drift, scalability, and integration complexities and prescribes strategic interventions such as in-memory computation, asynchronous computation, adaptive learning models, and automatic scaling of cloud resources. The document also presents a comparative analysis of technologies, metrics, and cost. Future trends such as Edge AI, federated learning, AI-optimized systems, quantum computing, and adoption of blockchain technologies are analyzed for future data pipes in real time. Optimizing data pipes in real-time, thus, makes AI systems accurate, scalable, and secure, enabling smart digital ecosystems and industry disruption.

Keywords: Real-time data pipelines, AI optimization, low-latency processing, data drift, scalability, Lambda architecture, Kappa architecture, microservices, edge computing, in-memory computing, adaptive learning models, autonomous vehicles

1. INTRODUCTION

In the current world, where technology has advanced and become more dynamic in its delivery, data processing is now a core asset. Real-time AI means that a system is capable of making decisions within an instant, which is very useful in many critical applications (Tien, 2017). For example, self-driving cars require AI in real time to analyze the data coming from the car's sensors to safely steer the vehicle, or the trading operations that use AI to perform transactions in mere microseconds to capture gains in the fluctuating financial markets. Likewise, smart cities use AI in real time to regulate the use of resources such as traffic and energy. The effectiveness of these applications relies on the optimization of a data pipeline that can process large amounts of data with small delays.

Data pipelines are the foundation of real-time AI, ensuring that data streams from sources to processing elements and then to decision-making. The architectures of these pipelines should be designed for scalability to handle large amounts of data and for low latency for quick processing time and also for reliability when faced with the different conditions. As more data is being generated, the proper handling of this data becomes a necessity, particularly in defining the most efficient data pipeline. Latency, data drift, and other issues can slow down these pipelines, and therefore, optimizing them is not just desirable but necessary for effective AI operations.

2. ANATOMY OF REAL-TIME DATA PIPELINES

A real-time data pipeline can be explained as a methodology of data processing that implies the immediate provision of information for analysis and decision-making based on the data that is being constantly produced. This is different from other batch process systems that process data in blocks as this takes time, which is inherent in the system. Real-time processing pipelines are aimed at providing low latency and high throughput, which facilitates the reaction to the event as soon as it occurs (Tantalaki, 2020).

The core components of a real-time data pipeline include data ingestion, data processing, data storage, and analytics and integration (Figure 1):

3. DATA INGESTION

This involves the collection of raw data from various sources, such as sensors, application programming interfaces (APIs), and user interactions. The ingestion layer must support high throughput and low-latency to prevent bottlenecks at the entry point of the pipeline.

4. DATA PROCESSING

Once ingested, data undergoes transformation and analysis. Real-time processing can be achieved through stream processing, where data is processed as it arrives, or micro-batching, which processes data in small, frequent batches. Stream processing is preferred for ultra-low-latency requirements, while micro-batching offers a balance between latency and resource utilization.

5. DATA STORAGE

Efficient storage solutions are necessary for persisting processed data and supporting rapid retrieval. In-memory databases and NoSQL databases are commonly employed to meet the demands of real-time access speeds (Badgujar, 2023).

6. ANALYTICS AND AI INTEGRATION

The final component involves applying analytical models and AI algorithms to the processed data to generate actionable insights. This integration must be seamless to ensure that the AI models receive timely and relevant data inputs.

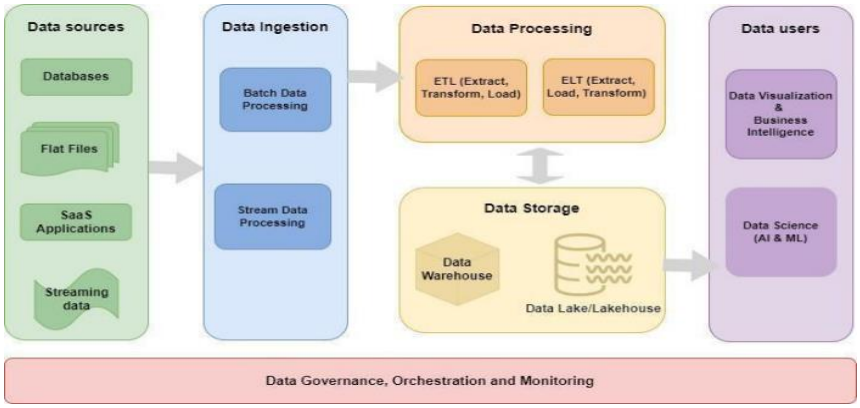

Figure 1: Fig 1 (Devineni, 2022)

7. THE CHALLENGES: BREAKING DOWN THE BARRIERS

Latency: Latency issues are still present because when there is a delay in the processing of data, it may produce irrelevant information. Sources of latency include the network latency, the processing overhead, and the latency that is incurred during the serialization of data. For instance, in financial trading systems, even a microsecond delay could lead to high financial losses.

Data Drift: Data drift, also known as concept drift or context drift, is the slow and gradual change in the distribution of the data over time, which can have a negative impact on the performance of the AI models. There is also the problem of reducing model effectiveness as the data changes over time because the models are built based on the existing data.

Scalability and Cost: The major challenge is to achieve scalability while at the same time maintaining cost efficiency. Proportioning for high traffic volume means that during off-peak periods, resources are unused, hence, wastage of resources. However, a lack of resources during such times is likely to lead to reduced performance.

8. CRAFTING THE OPTIMAL PIPELINE: STRATEGIES AND BEST PRACTICES

In order to deal with the challenges of a real-time data pipeline, there are several measures that should be taken, and the following are some of the best practices that should be adopted. Architectural optimization is basic in this respect.

Choosing the right architecture that has to be either Lambda, Kappa, or Microservices should be made depending on certain parameters such as latency, the degree of complexity, and scalability (Karabey Aksakalli et al., 2021). Lambda architecture can be described as a batch and real-time data processing architecture, while Kappa architecture is just a real-time processing architecture. However, microservices are more scalable and flexible since they operate in a modular fashion. Also, introducing edge computing can help to minimize the latencies due to data transfer to the centralized processing nodes since the data processing will be performed closer to the sources.

Reducing latency is equally important in ensuring that real-time systems are responsive, as is the case in real-time applications. Memory computing enhances the way in which the data is accessed and worked upon because it eliminates the time spent in accessing the disk storage and retrieval. The asynchronous processing also improves performance by addressing the issue of data processing by separating the data ingestion from the processing tasks so that the system can process new data as it comes in while the other tasks are being processed. Also, when the data is exchanged between systems, it is in the most useful format, thanks to the efficient data serialization formats like Avro or Google Protocol Buffers, therefore reducing the time taken to encode and decode the data, thus reducing the overall latency.

Another important feature of AI maintenance is data drift and model decay prevention and monitoring. Real-time observation of data streams and model performance indicators allows identifying data drift and avoiding a decline in model performance (Demšar & Bosnić, 2018). The proposed learning models are adaptive learning models that are capable of updating themselves with fresh data, hence providing updated results. Furthermore, this approach makes feature engineering processes scalable and adaptive to new trends as the models do not require data scientists to manually update the feature selection. This is useful to prevent issues that may lead to model degradation and shorten its useful life cycle.

9. REAL-WORLD APPLICATIONS AND USE CASES

The enhancement of real-time data pipelines has general implications in numerous industries. In self-driving cars, real-time data processing is important for the sensor fusion where the data gathered from LIDAR, radar, and cameras are combined in order to have a full picture of the car. Low latency is important when identifying obstacles to avoid and to ensure the safety of the passengers on board (Tedeschi & Sciancalepore, 2019). For instance, Tesla's self-driving car's autopilot feature and Waymo's self-driving car system involve the use of real-time data to make split-second decisions on how to navigate. In financial trading systems, high-frequency trading platforms are heavily dependent on real-time data pipelines to analyze the changes in the markets and make trades in the millisecond range. Artificial intelligence models enable traders to forecast the market and make appropriate decisions that will curtail their losses and enhance profitability. This is due to the fact that only a 1-3ms delay in a pipeline can lead to huge losses.

10. COMPARATIVE ANALYSIS: TECHNOLOGIES AND ARCHITECTURES

The comparison allows to identify advantages and disadvantages of various architectures and technologies in question. Comparing lambda architecture to other architectures it is evident that though it provides robustness due to batch and real time processing thereby complicating it by having two layers. nonetheless, Kappa Architecture is more straightforward since it is designed for handling only real-time data and could be less efficient in analysis of historical data. Event-Driven Microservices enhance scalability and flexibility of the pipeline by breaking it into segments that can be developed and released individually. Tool and platform evaluation also provides the distinction as follows: Apache Kafka is one of the most popular messaging platforms because of its high throughput and fault tolerance and Pulsar is also a strong competitor to it. Kafka is famous for its stability and fault-tolerance mechanism while Pulsar performs better in different tenancy and geographical replication (Sharma & Atyab, 2022). For stream processing, Apache Flink outperforms Spark Streaming in latency-sensitive applications due to its native event-driven architecture.

11. FUTURE TRENDS: THE NEXT FRONTIER OF REAL-TIME AI PIPELINES

The future of real-time AI pipes is shaped by several emerging trends. Edge AI and Federated Learning are becoming increasingly crucial, enabling responsiveness through edge-processed data, keeping time-sensitive application latencies to a minimum. Federated learning is also keeping data private by allowing AI models to learn on federated data sources without sharing raw data. This is crucial for IoT devices and wearable health monitors. AI-optimized scaling is also a game-changer. Applying AI to prediction and resource scaling, real-time pipes optimize for optimal

performance and cost. Smart workload load-balancing assures dynamic computation resource allocation in accordance with current demands, preventing bottlenecks and optimizing for low latencies.

12. CONCLUSION: DRIVING THE FUTURE WITH OPTIMIZED DATA PIPELINES

Therefore, the optimization of real-time data pipelines is essential for the enhancement of AI in decision-aiding processes in various businesses and sectors. Thus, the appropriate selection of the architectural models, the usage of advanced technologies, and the implementation of the proper strategies for latency, scalability as well as security will help to ensure that the data pipelines are highly efficient and dependable. Real-time AI systems supported by optimal pipelines are revolutionizing self-driving cars, stock markets, smart cities and cities' health systems, and the overall quality of life. Some of the strategic considerations are the use of certain architectures that correspond with the application needs, in-memory computing and the asynchronous processing for a decrease in latency, and using autoscaling for cost-effective scale up. Moreover, it is required to track the changes in the data environment and apply the corresponding changes to the model for continued accuracy.

REFERENCES

[1] Badgujar, P. (2023). Optimizing Data Storage and Retrieval in NoSQL Databases Strategies for Scalability. Journal of Technological Innovations, 4(2). https://doi.org/10.93153/w824cx89

[2] Demšar, J., & Bosnić, Z. (2018). Detecting concept drift in data streams using model explanation. Expert Systems with Applications, 92, 546–559. https://doi.org/10.1016/j.eswa.2017.10.003

[3] Devineni, S. (2022, May 21). An Overview of Data Pipeline Architecture. DZone. https://dzone.com/articles/an-overview-of-key-components-of-a-data-pipeline

[4] Karabey Aksakalli, I., Çelik, T., Can, A. B., & Teki̇nerdoğan, B. (2021). Deployment and communication patterns in microservice architectures: A systematic literature review. Journal of Systems and Software, 180, 111014. https://doi.org/10.1016/j.jss.2021.111014

[5] Sharma, R., & Atyab, M. (2022). Cloud-Native Microservices with Apache Pulsar. Springer. https://link.springer.com/book/10.1007/978-1-4842-7839-0

[6] Tantalaki, N. (2020). A review on big data real-time stream processing and its scheduling techniques. International Journal of Parallel, Emergent and Distributed Systems.

[7] Tedeschi, P., & Sciancalepore, S. (2019). Edge and fog computing in critical infrastructures: Analysis, security threats, and research challenges. 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), 1–10. https://doi.org/10.1109/eurospw.2019.00007

[8] Tien, J. M. (2017). Internet of Things, Real-Time Decision Making, and Artificial Intelligence. Annals of Data Science, 4(2), 149–178. https://doi.org/10.1007/s40745-017-0112-5