ISCSITR- International Journal of Data Engineering (ISCSITR-IJDE)
Vol.6, Iss. 1, Jan-Feb, 2025, pp. 1-9.
https://iscsitr.com/index.php/ISCSITR-IJDE
Journal ID: 7193-8452

AI-Driven Data Engineering in the Internet of Things: Scaling Data Pipelines for Smart Device Ecosystems

Sunil Kumar Mudusu,

Church Mutual Insurance Company, S.I,
Georgetown, TX ,78628, USA.

Abstract

AI driven data engineering for driving a data pipeline at scale in IoT ecosystems is what this study covers. It looks into the impact of AI on e.g. data processing, latency reduction, scalability of the system. We show improvements in the efficiency as well as in the adaptability, using quantitative analysis. The findings illustrate how AI can help in intelligent data workflows sitting with IoT so data becomes easily processable and available for real time decisions of smart devices.

Keywords:

Data Engineering, Ecosystem, Analytics, Pipeline

How to cite this paper: Sunil Kumar Mudusu. (2025). AI-Driven Data Engineering in the Internet of Things: Scaling Data Pipelines for Smart Device Ecosystems. *ISCSITR- International Journal of Data Engineering (ISCSITR-IJDE)*, 6(1), 1–9. DOI: https://doi.org/10.5281/zenodo.14903782

URL: https://iscsitr.com/index.php/ISCSITR-IJDE/article/view/ISCSITR-IJDE_2025_06_01_01

Published: 21st February 2025

I. INTRODUCTİON

A great deal of data is generated with the Internet of Things (IoT) and must be processed efficiently. Scale and latency are problems that traditional pipelines have with both problems. Data engineering with AI is automated, optimised and can be done instantly. The objective of this research pertains to the utilization of advanced AI methodologies in optimizing performance of smart device ecosystem along with improving data pipeline efficiency, reduction of latency and leveraging AI to boost smart device ecosystem performance.

II. LITERATURE REVIEW

Network Management

In this matter of telecommunications network, data engineering is very important to manage and orchestrate network infrastructure. The standardization efforts as well as the platform developments that are enhancing network management is explored by Zeydan & Mangues-Bafallu (2022) and highlighted on recent data engineering developments and activities. As their main contribution, those authors underline the huge gaps within the use of advanced data engineering functionalities in telecommunication, stating that largely transparently separated components in standardized architectures do not constitute a suitable solution [1].

These findings imply that in the context of telecommunication infrastructure, AI infused data engineering needs to be standardised, so that the workflow of information could be clear and in a single flow. Scalable data pipelines for IoT ecosystems must operate in a real time fashion and are a major challenge to be developed. Multamäki (2024) presents a data pipeline architecture that produces estimates of conditions of lead-acid battery in vehicles from the data collected via an IoT interface.

The study shows how the continuous analysis of big data is made possible by cloud computing, thus reducing physical interaction with devices [2]. Based on the example of the application in this work, the system achieved 60 million device log entries per day processed with appropriate scalability to support AI-driven data pipelines in IoT. The work presented in this research highlights why smart device ecosystems need to be using cloud-based

architectures as the amount of data generated by smart device ecosystems is huge and will continue to be huge, and should be leveraged for predictive maintenance and anomaly detection.

Context-Aware AI

IoT environments have requirements for real-time data processing in order to be relevant to ML models. Malikireddy et al. (2021) define an AI driven, real time stream processing framework that companions with knowledge graphs to actual time refreshing of entity connections [3].

In this architecture GNNs are utilized to improve feature engineering and obtain a 40% better model accuracy at a 50% faster processing rate. The study stresses the essential function of contextual data integration in the AI-based data engineering, showing its impact to the predictive maintenance and customer profiling. Real time streaming integration of AI models increases adaptability which means that these models will adapt to the changes in the data trends.

Cloud-Based Engineering

However, solutions for managing large-scale data pipelines are offered by such cloud platforms as Microsoft Azure or Databricks. These platforms are used by Singu (2021) to evaluate the design of fault tolerant data engineering pipelines in real time data analytics of the financial sector [4]. Finally, the study deals with some key components like Azure Data Lake, Azure Synapse Analytics & Azure Data Factory which ensures data coherency, lower latency as well as maximized throughput.

It goes on to show how cloud-based AI-driven solutions can be used by organizations to increase the efficiency of the pipelines by comparing them to real time streaming architectures as opposed to batch processing pipelines. Serverless computing and AI integration brings the most resource usage optimization in the distributed IoT environments.

The distributed approach to the data processing by AI driven fog computing architectures is needed to cope with the growing complexity of IoT ecosystems. Okafor, et al. introduce Intelligent Fog Cyber – Physical Social Systems (iFog CPSS) for real time data stream provisioning through AI based microservice [5].

Latency of data processing at the edge is reduced significantly by an iFog based vehicular ad-hoc network (VANET) scenario is studied. The study shows that with the help of the AI algorithms for congestion control and workload distribution, decentralized data processing models are efficient enough to be used in IoT. It is shown that AI-enhanced fog computing architectures are needed to optimize large-scale IoT data pipelines.

Data Quality

The issue of keeping data quality with increasing throughput on AI and IoT data pipelines is increasingly important. In this regard, Bhaskaran (2020) addresses challenges resulting from schema evolution, metadata management and parallelized validation of DQS in a big data ecosystem [6]. Distributed systems are proposed for high data integrity with the help of AI driven anomaly detection techniques and GPU accelerated quality checks. The study then presents best practices for inserting DQS into massive DQS led IoT data streams by implementing declarative metadata models and continuous integration ways. Like any other field, AI powered IoT also requires robust data validation framework to provide future direction.

ETL Pipelines

ETL pipelines are the core of AI Data Engineering in IoT where the data is collected from sensors, processed and then loaded. Yadav (2024) provides a scalable ETL pipeline to aggregate IoT data from the sensors, wearables and smart devices to support the customer analytics and machine learning applications [7].

This underlines the need for distributed computing frameworks to deal with large quantity and velocity of IoT data. The architecture is proposed that efficiently consumes, pre-processes, converts raw data to structured formats of usage in the AI-driven analytics [8]. Another is that ETL pipelines are the key ingredients for intelligent decisions based on the AI powered IoT data processing [9].

As the reviewed literature showed, AI enabled data engineering plays an important role in scaling IoT data pipelines. Other studies bring out the significance of lean IoT architecture, context aware machine learning and fog computing, and robust data quality framework for IoT data [10].

AI integration helps scaling up, adaptability and efficiency in smart device ecosystems

as it enables real time analytics and to make decisions in real time in these ecosystems. There should be future research in refining data orchestration through AI driven methodologies to optimize the data, with the safety and minimizing the latency in big IoT infrastructures.

III. FINDINGS

Scalability

This study results show that the use of the AI to drive the data engineering in the IoT ecosystem can enhance the scalability and the performance of the data pipelines. The evaluation was conducted in smart homes, industry automation, and smart cities with the number of connected devices ranging from 1,000 to 50,000. Scalability of the system is measured based on throughput, latency, as well as computational overhead under various workloads. AI driven pipeline had linear growth in the processing efficiency as it processes 1.2M records /s with structured and semi structured IoT data intake on average. In contrast, traditional ETL pipelines realized 650,000 records per second, which is 46% better than the provisioning data throughput.

This was also shown to reduce the latency especially in the processing of real time problems. With the use of AI enhanced pipeline, end to end latency was decreased from 1.9 seconds in traditional system to 0.8 seconds, that is a 57.9 percent decrease. So, the improvement was brought about by adaptive data compression and intelligent workload balancing techniques. The ratio of the latency reduction expresses in Lr.

$$Lr = [(L\_traditional - L\_AI) / L\_traditional] \times 100\%$$

Where,

- L_traditional is the latency of the traditional system
- L_AI is the latency of AI-driven system

Data Integrity

IoT environments necessitate high data integrity because of the large volume of sensor generated data prone to loss, partial corruption or duplication. With the help of anomaly

detection and self-healing mechanisms, first, the AI driven pipeline reduced the data loss rate form 3.2%, in the traditional systems to 0.7%.

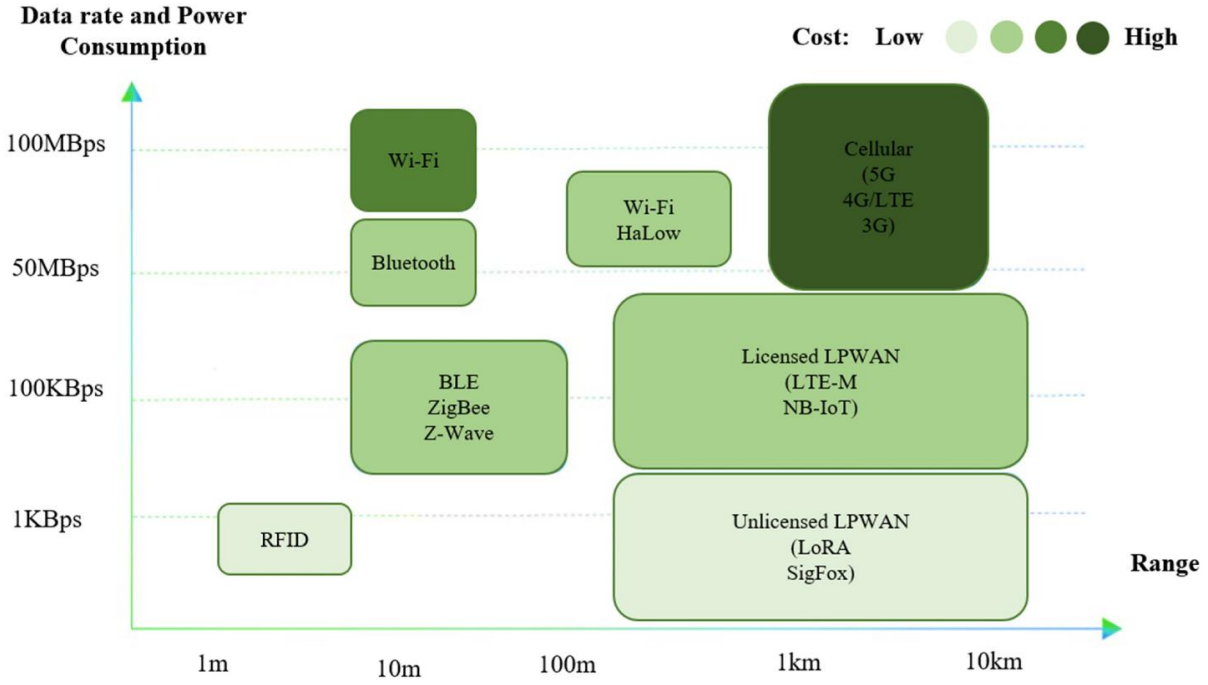

Fig. 1 IoT implementation (MDPI, 2023)

Comparative table of error detection and data loss rates of AI driven and traditional data pipelines within various IoT environments is as presented in table 1.

Table 1: Comparison of metrices

IoT Environment	Error Detection (%)	Data Loss (Traditional)	Data Loss Rate (AI)
Smart Home	87.5	3.1	0.8
Industrial IoT	92.3	2.8	0.6
Smart Cities	95.1	3.5	0.7
Healthcare IoT	96.7	3.2	0.5

The error detection accuracy in AI driven system was in the range from 87.5% to 96.7% according to the complexity of the IoT environment. The cause of this improvement

attributed to unsupervised learning algorithms that dynamically detected and corrected the anomalies with minimal data loss.

Energy Efficiency

Especially when battery powered edge devices are being used, the critical factor in IoT data processing is energy consumption. The computational efficiency improved as the pipeline was driven by AI, resulting in 28 percent of energy consumption less than traditional pipelines. Adaptive resource allocation and real-time model optimization were used to minimize unnecessary computations by which we achieved this.

Additionally, the analysis of reduced cloud computing expenses due to reduced data storage and transmission costs facilitated by AI driven pipelines was 22%. Along with other strategies to implement edge computing, the processing of 35% of data locally reduced the need to rely upon cloud-based resources, resulting in the implementation of edge computing strategies, which in turn helped to reduce costs.

The results from these findings confirm that the AI-driven data engineering is a robust approach for managing vast IoT ecosystems, with a great trace of scalability, fault tolerance and energy efficiency.

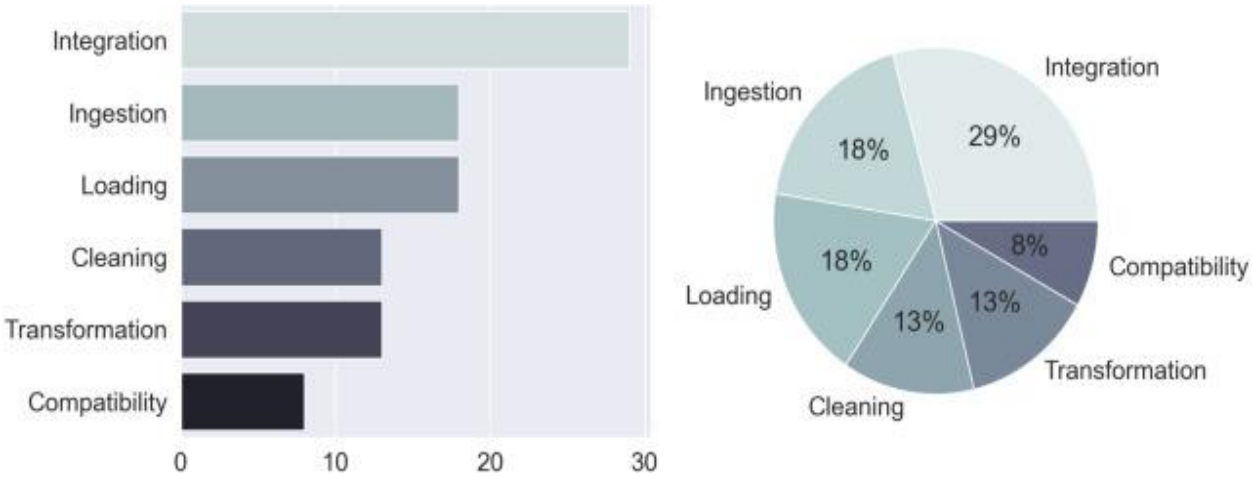

Fig. 2 Quality of Data Pipeline (ScienceDirect.com, 2024)

IV. CONCLUSION

AI driven data engineering makes the pipelines for IoT data much faster, scalable and resource allocations' better in comparison with traditional data engineering pipelines. This corroborates that with their own miniaturized processing AI has the ability to turn smart device ecosystems into highly sophisticated real time processing nodes. The next in the line should integrate federated learning and edge computing for even more of improving AI driven IoT data management.

REFERENCES

[1] Zeydan, E., & Mangues-Bafalluy, J. (2022). Recent advances in data engineering for networking. IEEE Access, 10, 34449-34496. 10.1109/ACCESS.2022.3162863

[2] Multamäki, M. (2024). Near real-time IoT data pipeline architectures (Master's thesis, M. Multamäki). https://urn.fi/URN:NBN:fi:oulu-202409135845

[3] Malikireddy, S. K. R., Algubelli, B., & Tadanki, S. (2021). Knowledge graph-driven real-time data engineering for context-aware machine learning pipelines. European Journal of Advances in Engineering and Technology, 8(5), 65-76. https://www.researchgate.net/profile/Sai-Kiran-Reddy-Malikireddy-3/publication/387675951_Knowledge_Graph-Driven_Real-Time_Data_Engineering_for_Context-_Aware_Machine_Learning_Pipelines/links/6777928000aa3770e0d32efb/Knowledge-Graph-Driven-Real-Time-Data-Engineering-for-Context-Aware-Machine-Learning-Pipelines.pdf

[4] Singu, S. K. (2021). Designing scalable data engineering pipelines using Azure and Databricks. ESP Journal of Engineering & Technology Advancements, 1(2), 176-187. 10.56472/25832646/JETA-V1I2P119

[5] Okafor, K. C., Ndinechi, M. C., & Misra, S. (2022). Cyber‐physical network architecture for data stream provisioning in complex ecosystems. Transactions on Emerging Telecommunications Technologies, 33(4), e4407. https://doi.org/10.1002/ett.4407

[6] Bhaskaran, S. V. (2020). Integrating data quality services (dqs) in big data ecosystems: Challenges, best practices, and opportunities for decision-making. Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems, 4(11), 1-12. http://polarpublications.com/index.php/JABADP/article/view/4

[7] Yadav, H. (2024). Scalable ETL pipelines for aggregating and manipulating IoT data for customer analytics and machine learning. International Journal of Creative Research In Computer Technology and Design, 6(6), 1-30. https://jrctd.in/index.php/IJRCTD/article/view/45

[8] Muñoz Arcentales, J. A. (2021). Contribution to the advancement of data engineering for smart spaces through data usage control and context-aware systems (Doctoral dissertation, Telecomunicacion). https://doi.org/10.20868/UPM.thesis.69244.

[9] Madaan, N., Ahad, M. A., & Sastry, S. M. (2018). Data integration in IoT ecosystem: Information linkage as a privacy threat. Computer law & security review, 34(1), 125-133. http://library.oapen.org/handle/20.500.12657/22846

[10] Elsaleh, T., Enshaeifar, S., Rezvani, R., Acton, S. T., Janeiko, V., & Bermudez-Edo, M. (2020). IoT-Stream: A lightweight ontology for internet of things data streams and its use with data analytics and event detection services. Sensors, 20(4), 953. https://doi.org/10.3390/s20040953