# BIG DATA INFRASTRUCTURE: INTEGRATING LEGACY SYSTEMS WITH AI-DRIVEN PLATFORMS

Aeshna Kapoor

Lead Data Scientist, BNY Mellon, New York, USA

***ABSTRACT***

*The rapid evolution of data-driven technologies has led to the proliferation of big data systems capable of managing and analyzing vast amounts of data. However, many organizations continue to rely on legacy systems that are deeply entrenched in their operations. The challenge lies in integrating these legacy systems with new, AI-driven platforms to create a cohesive, hybrid infrastructure that leverages the strengths of both. This paper presents a comprehensive approach to designing and implementing a hybrid big data infrastructure that combines legacy systems with advanced AI technologies. We explore the challenges, architectural considerations, and the potential benefits of such an integration, aiming to provide a roadmap for organizations seeking to modernize their data infrastructure without completely abandoning their existing investments..*

***KEYWORDS***

*Big Data, Hybrid Infrastructure, Legacy Systems, AI Integration, Data Platforms*

## 1. INTRODUCTION

### 1.1. Background

The advent of big data has redefined how organizations collect, manage, and analyze information. AI and machine learning-powered data platforms offer unprecedented capabilities in real-time analytics, predictive modeling, and data-driven decision-making [3]. However, many organizations remain dependent on legacy systems, which were not designed to manage the scale, variety, or velocity of contemporary data streams [4]. While legacy systems continue to play critical roles in day-to-day business operations, replacing them is often impractical due to their deep integration and reliability [4].

### 1.2. Problem Statement

Organizations face a significant challenge in integrating these legacy systems with AI-powered platforms [5]. While legacy systems operate with monolithic architectures, batch processing, and structured data formats, modern platforms rely on microservices, real-time data handling, and unstructured or semi-structured data [2]. The objective of this paper is to propose a hybrid infrastructure that allows for the seamless integration of legacy systems and AI-driven platforms, maximizing their strengths while minimizing their limitations [1].

David C. Wyld et al. (Eds): NWCOM, ASOFT, ITCCMA, CSITY, BDIoT, MLNLP, SIGPRO, AIFZ– 2024
pp. 145-152, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141913

### 1.3. Objectives

The core objective of this research is to propose a hybrid architecture that supports the gradual modernization of legacy systems through AI integration. The paper also aims to examine compatibility concerns between legacy infrastructures and new technologies while addressing key challenges related to scalability, data interoperability, and cost-effective implementation.

## 2. LITERATURE REVIEW

### 2.1. Legacy Systems in Big Data

Legacy systems are often associated with outdated technologies and rigid architectures, which lack the flexibility and scalability needed for modern data operations [4]. Despite these limitations, they continue to be used due to their stability and deep integration within business processes [5]. Various studies have highlighted the difficulties organizations face in integrating legacy systems with AI-based platforms, particularly in terms of data compatibility, interoperability, and migration costs [5]

### 2.2. AI-Driven Platforms

Modern AI platforms leverage distributed computing, machine learning algorithms, and real-time data analytics to generate actionable insights [1]. These platforms are well-equipped to manage big data, making them suitable for tasks such as predictive maintenance, customer behavior analysis, and personalized marketing strategies [2]. However, integrating these advanced platforms with legacy systems presents technical and organizational challenges [6].

### 2.3. Hybrid Infrastructure Approaches

Various approaches have been suggested in the literature for hybrid infrastructure design. These range from middleware solutions that bridge legacy systems and modern platforms to more comprehensive frameworks that re-engineer legacy applications for modern data architectures [6]. The solutions proposed vary in complexity, cost, and effectiveness, and there is no one-size-fits-all approach [7].

## 3. PROPOSED HYBRID INFRASTRUCTURE

### 3.1. Architectural Overview

The proposed hybrid infrastructure consists of key architectural components that ensure a smooth integration of legacy systems and AI-driven platforms:

- **Data Integration Layer**: This layer serves as a bridge between legacy systems and modern platforms. A block diagram representation of this integration process is shown below. The diagram demonstrates how data from legacy systems flows through various stages before being processed by AI. It employs ETL (Extract, Transform, Load) processes, API gateways, and data lakes to facilitate data exchange [9].
- **Microservices Architecture**: Legacy systems are gradually re-engineered into microservices, allowing them to interact more effectively with modern platforms [7]. This approach enables the incremental modernization of legacy applications without disrupting ongoing operations [10].

- **AI and Machine Learning Layer**: The integration of AI in legacy systems is one of the key contributions of this paper. The specific machine learning algorithms used in our AI integration include XGBoost for predictive modeling, Random Forest for classification tasks, and Neural Networks for deeper pattern recognition. These models were selected due to their versatility in handling structured and unstructured data alike, ensuring seamless integration with legacy systems [3].
- **Orchestration and Management:** A centralized orchestration layer governs data processing workflows and resource allocation, ensuring optimal performance and real-time monitoring across both legacy and modern platforms [8].

### 3.2. Data Flow and Processing

In the proposed hybrid infrastructure, data from legacy systems is first ingested into the data integration layer, where it is transformed into a format compatible with modern platforms [4]. This data is then processed by the AI and machine learning layer, where advanced analytics are performed [5]. The results of these analyses can be fed back into legacy systems or used directly by modern applications [6].

### 3.3. Security and Compliance

Security and compliance are critical considerations in the hybrid infrastructure [7]. The integration of legacy systems, which may have weaker security controls, with modern platforms necessitates a comprehensive security framework [8]. This framework should include data encryption, access controls, audit trails, and compliance with relevant regulations such as GDPR and HIPAA [9].

## 4. CHALLENGES AND SOLUTIONS

### 4.1. Data Interoperability

One of the primary challenges in integrating legacy systems with AI platforms is ensuring data interoperability [1]. Legacy systems often use proprietary data formats incompatible with modern AI platforms, making it necessary to implement robust data transformation solutions [9].

### 4.2. Scalability

Scaling a hybrid infrastructure to handle large volumes of data is another critical challenge [4]. This can be achieved by designing the architecture to support horizontal scaling, leveraging cloud-based platforms and distributed computing solutions [6].

### 4.3. Performance Optimization

Performance optimization is crucial to ensure that the hybrid infrastructure operates efficiently [6]. This requires careful management of data processing workflows, load balancing, and resource allocation [7]. The orchestration layer plays a key role in monitoring and optimizing system performance [8].

### 4.4. Cost Management

Implementing a hybrid infrastructure can be costly, particularly if extensive re-engineering of legacy systems is required [9]. Organizations should conduct a thorough cost-benefit analysis to

determine the most cost-effective approach to integration [10]. This may involve prioritizing certain systems for modernization while leaving others in their legacy state [1].

# 5. CASE STUDIES

## 5.1. Financial Services

In the financial services industry, many organizations rely on legacy systems for core banking operations [2]. By integrating these systems with AI-driven platforms, banks can enhance their fraud detection capabilities, improve customer service through personalized recommendations, and optimize their trading strategies [3].

## 5.2. Healthcare

Healthcare organizations often use legacy systems for patient records and billing [4]. A hybrid infrastructure allows for the integration of these systems with AI-driven platforms that can analyze patient data in real-time, enabling predictive diagnostics, personalized treatment plans, and improved patient outcomes [5].

## 5.3. Manufacturing

In the manufacturing sector, legacy systems are commonly used for supply chain management and production control [6]. Integrating these systems with AI-driven platforms can lead to more efficient production processes, predictive maintenance of equipment, and better demand forecasting [7].

# 6. CONTRIBUTION AND NOVELTY

The proposed hybrid infrastructure offers a unique solution that enables legacy systems to coexist with modern AI-driven platforms. It provides a cost-effective method for modernization, allowing organizations to leverage the power of AI without needing to replace their legacy systems entirely [3]. This research focuses on sectors such as financial services and healthcare, where legacy systems are deeply entrenched, and the cost-benefit analysis of modernization is crucial.

# 7. EMERGING TRENDS AND FUTURE CONSIDERATIONS

## 7.1. Edge AI and IoT: Bringing AI Closer to the Data Source

With the growing demand for real-time data analysis and decision-making, Edge AI is emerging as a transformative force across industries like manufacturing, logistics, utilities, and telecommunications. While cloud-based AI remains widespread, Edge AI enables organizations to deploy AI models closer to the data source, significantly reducing latency and improving response times [6].

Challenges for Legacy Systems: Many legacy systems lack the real-time processing power required for edge computing. Integration with modern Edge AI frameworks often requires re-engineering and updating existing infrastructure [8]. Despite these challenges, the benefits of predictive maintenance in manufacturing, real-time optimization in logistics, and smart grid management in utilities make Edge AI integration critical [9].

### 7.2. Quantum Computing: The Future of Data Processing

Quantum computing holds the potential to revolutionize data processing, enabling faster, more complex computations than classical systems. As quantum technology becomes more accessible, it could significantly impact AI models and big data analytics, providing solutions to previously unsolvable problems [10].

Challenges for Legacy Systems: Legacy systems are built on classical computing architectures and may not be compatible with quantum-enhanced AI models. As quantum computing becomes more prevalent, organizations will need to upgrade or redesign their data systems to handle quantum-safe encryption and large-scale quantum computations [9]. Despite these hurdles, industries such as healthcare, logistics, and finance are expected to benefit from the capabilities of quantum AI in areas like drug discovery, supply chain optimization, and financial risk analysis [10].

## 8. CONCLUSION

The integration of legacy systems with AI-driven platforms is a pivotal step for organizations aiming to remain competitive in today's fast-paced, data-driven world. Legacy systems, while often reliable and integral to operations, were not designed to handle the massive volumes and complexity of modern data streams. However, replacing them entirely can be a costly, disruptive process. A hybrid infrastructure, as proposed in this paper, offers a seamless approach that allows organizations to integrate cutting-edge AI capabilities without abandoning their legacy systems. This approach ensures that companies can take advantage of the latest advancements in AI while maintaining the stability and continuity of their existing operations.

This hybrid solution provides several key benefits. First, it offers the ability to modernize data processing and analytics capabilities, allowing organizations to glean real-time insights from their data that were previously unattainable with legacy systems alone. The implementation of AI models—such as machine learning algorithms like XGBoost, Random Forest, and neural networks—enables businesses to forecast trends, optimize operations, and improve decision-making. At the same time, the reliance on legacy systems is minimized but not eliminated, enabling companies to protect their existing investments in hardware and software while still moving toward a more data-centric future.

Furthermore, the integration of AI technologies such as Edge AI and Quantum Computing provides a significant leap forward for organizations that need to process data at scale and in real-time. Edge AI, which brings computation closer to the data source, reduces latency and increases the speed of decision-making—critical for industries like manufacturing, logistics, and utilities. Quantum computing, with its unparalleled processing power, has the potential to revolutionize areas such as optimization, encryption, and large-scale data analysis. As these technologies continue to mature, organizations will need to ensure that their infrastructure is adaptable and future-proofed to accommodate these innovations, making the hybrid approach an essential strategy.

The hybrid infrastructure also mitigates the challenges that typically accompany data integration, including issues related to interoperability, scalability, and performance. The use of ETL (Extract, Transform, Load) processes within the data integration layer allows for the smooth flow of data between disparate systems, ensuring that legacy systems and AI-driven platforms can work together harmoniously. By adopting microservices architecture, organizations can break down monolithic systems into smaller, more manageable components that can be upgraded

incrementally. This gradual modernization process ensures minimal disruption to business operations while maintaining flexibility for future technological advancements.

In terms of performance, the orchestration and management layer plays a crucial role in balancing workloads, optimizing resource allocation, and monitoring system performance in real-time. This ensures that the hybrid infrastructure remains efficient, scalable, and resilient even as the volume of data and complexity of processes increase. By implementing AI-driven orchestration, organizations can automate many of the routine tasks associated with managing data infrastructure, reducing the burden on IT teams and freeing up resources for innovation.

From a financial perspective, a hybrid infrastructure offers a cost-effective solution by allowing organizations to phase their investments in new technology. Rather than undertaking a complete overhaul of their systems—which can be prohibitively expensive and risky—companies can adopt a gradual approach, upgrading specific components as needed. This incremental strategy not only preserves existing investments but also allows for continuous innovation, ensuring that organizations remain agile in the face of rapidly evolving technological landscapes.

## 8.1. Summary

The integration of legacy systems with AI-driven platforms presents both challenges and opportunities for organizations. A hybrid infrastructure, as proposed in this paper, provides a practical and scalable solution that enables organizations to leverage the best of both worlds—maintaining the stability of legacy systems while benefiting from the advanced capabilities of AI technologies. By carefully addressing issues such as data interoperability, scalability, performance optimization, and cost management, organizations can build a robust data infrastructure that supports their long-term strategic goals.

The hybrid approach also allows for future-proofing, as it enables the gradual integration of emerging technologies like Edge AI and Quantum Computing, ensuring that organizations are prepared for the next wave of innovation. By implementing a flexible, adaptive infrastructure, companies can improve operational efficiency, enhance decision-making, and drive innovation, ultimately gaining a competitive edge in the digital age.

## 8.2. Future Work

Looking ahead, future research should focus on developing more advanced tools and frameworks to further improve hybrid infrastructure designs. Key areas of interest include the automation of data transformation processes, the implementation of AI-driven orchestration systems, and the development of more cost-effective and scalable solutions for managing large datasets. Additionally, further case studies across a diverse range of industries are essential to validate the effectiveness of these hybrid infrastructures in various contexts. For example, industries such as healthcare, finance, manufacturing, and logistics could provide valuable insights into how different sectors adapt to the challenges and opportunities presented by hybrid data infrastructures.

Moreover, future work could explore the potential of Edge AI and Quantum Computing in more detail, examining how these technologies could be integrated into hybrid infrastructures to unlock even greater capabilities. Understanding the security, ethical, and operational implications of these technologies will be critical as they become more widely adopted.

## 8.3. Final Thoughts

In the context of today's rapidly evolving technological landscape, the need for organizations to adopt flexible and adaptive infrastructures cannot be overstated. The complexities of big data, combined with the exponential growth in data volumes and the increasing demand for real-time insights, require a new approach to data infrastructure. By embracing a hybrid model, organizations can modernize their systems without the risks and costs associated with complete system replacements, preserving the value of their legacy systems while simultaneously positioning themselves at the forefront of technological innovation.

The hybrid infrastructure not only addresses current challenges but also sets the stage for future growth, ensuring that organizations are well-equipped to navigate the demands of the digital age. As companies continue to evolve, those that successfully integrate AI-driven platforms with their legacy systems will be better positioned to lead their industries, drive innovation, and create sustainable competitive advantages. By modernizing their data infrastructure, organizations can unlock new opportunities for growth, efficiency, and success in an increasingly data-driven world.

## REFERENCES


[1] Marz, N., & Warren, J. (2015) Big Data: Principles and best practices of scalable real-time data systems, Manning Publications.

[2] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015) "The rise of 'big data' on cloud computing: Review and open research issues", Information Systems, Vol. 47, pp. 98-115. https://doi.org/10.1016/j.is.2014.07.006

[3] Chen, M., Mao, S., & Liu, Y. (2014) "Big data: A survey", Mobile Networks and Applications, Vol. 19, No. 2, pp. 171-209. https://doi.org/10.1007/s11036-013-0489-0

[4] Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014) "Trends in big data analytics", Journal of Parallel and Distributed Computing, Vol. 74, No. 7, pp. 2561-2573. https://doi.org/10.1016/j.jpdc.2014.01.003

[5] Khondoker, R., Patwary, M., & Islam, R. (2016) "A comprehensive study on big data issues and challenges", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, pp. 427-433. https://doi.org/10.14569/IJACSA.2016.070255

[6] Ruparelia, N. B. (2010) "Software development lifecycle models", ACM SIGSOFT Software Engineering Notes, Vol. 35, No. 3, pp. 8-13. https://doi.org/10.1145/1809218.1809221

[7] Sallam, R. L., Richardson, J., Schlegel, K., & Sood, B. (2020) "Magic Quadrant for Analytics and Business Intelligence Platforms", Gartner Research. https://www.gartner.com/doc/reprints?id=1-24VT0K9Z&ct=200327&st=sb

[8] Chen, J., & Zhang, Y. (2014) "A study of machine learning in wireless sensor networks", International Journal of Distributed Sensor Networks, Vol. 10, No. 8, pp. 597098. https://doi.org/10.1155/2014/597098

[9] Zhang, D., & Zhou, L. (2020) "Data lakes and analytics platforms in the age of big data: A systematic review", Journal of Big Data, Vol. 7, No. 1, pp. 1-15. https://doi.org/10.1186/s40537-020-00353-9

[10] Thota, C., Prasad, M., & Srivastava, A. (2017) "Big data analytics and machine learning for better healthcare", In EAI/Springer Innovations in Communication and Computing, Springer, pp. 103-117. https://doi.org/10.1007/978-3-319-52491-7_7


## AUTHOR


**Aeshna Kapoor** is a Lead Data Scientist at BNY Mellon USA With c ience, she focuses on integrating legacy systems with modern AI platforms to d financial industry. Aeshna is a frequent speaker at industry conferences.