

## AI-Driven Mainframe Modernization: Unlocking Legacy Data for Cloud Analytics

Ashish Prakash Khandelwal  
Amravati University, India

**Abstract:** Mainframe modernization has emerged as a critical imperative for enterprises seeking to leverage decades of valuable legacy data for advanced analytics and artificial intelligence initiatives. Legacy mainframe systems, while reliable for transaction processing, create substantial barriers to modern data science applications through architectural limitations and isolated data repositories. Cloud migration strategies utilizing specialized extraction tools and phased implementation approaches enable organizations to transform these historical information assets into accessible resources for contemporary analytics platforms. The integration of mainframe data with cloud environments dramatically accelerates enterprise AI adoption by exposing previously inaccessible information to data science teams, enhancing model accuracy and enabling comprehensive business insights. Large language models and other advanced AI systems, designed for cloud-native operation, gain substantial performance improvements when provided access to the rich historical context contained within mainframe repositories. Emerging technologies such as AI-driven extraction methodologies and retrieval-augmented generation systems further streamline the modernization process by automating complex mapping tasks and enabling immediate analytical capabilities without requiring complete re-platforming.

**Keywords:** Mainframe modernization, cloud migration, enterprise AI adoption, legacy data integration, retrieval-augmented generation.

### INTRODUCTION

In today's digital landscape, enterprises face a critical challenge: how to leverage decades of valuable data trapped in legacy mainframe systems. These systems, once the backbone of enterprise computing, now represent significant barriers to modern AI and analytics capabilities. The convergence of cloud technologies and artificial intelligence offers promising pathways to modernize these systems, making previously inaccessible data available for advanced analytics and machine learning applications.

A comprehensive analysis by Kalohia and Chakraborty reveals that approximately 71% of Fortune 500 companies continue to rely on mainframe systems for their mission-critical operations, with these legacy platforms processing an estimated 30 billion transactions daily across global financial, healthcare, and government sectors. Their research further indicates that organizations maintaining these systems allocate between 60-75% of their annual IT budgets to mainframe maintenance alone, leaving minimal resources for innovation and digital transformation initiatives. The researchers documented that enterprises attempting to implement modern analytics workflows directly on mainframe systems faced average processing latency increases of 780% compared to cloud-native environments, with corresponding cost increases of approximately 340% per analytical workload when considering computational resources, software licensing, and specialized personnel requirements [Kalohia, A, 2024].

Pandya's extensive field research across 217 enterprises undergoing modernization initiatives demonstrates that the integration of mainframe data with cloud-based AI platforms has yielded substantial business benefits, with organizations reporting an average 289% increase in analytical model accuracy after gaining access to previously siloed information. His case studies document that modern migration approaches leverage AI-driven neural network techniques to reduce manual migration efforts by 82.7% compared to traditional methods, while simultaneously cutting project timelines by 76.4% and implementation costs by 68.9%. Particularly noteworthy is the finding that organizations implementing retrieval-augmented generation technologies achieved a 97.3% reduction in data access latency, enabling real-time analytics on mainframe data without full migration, thereby creating hybrid environments that balance immediate analytical needs with long-term modernization objectives [Pandya, P, 2025].

### THE MAINFRAME CHALLENGE

#### Performance and Cost Barriers

Traditional mainframes present substantial roadblocks in both performance and cost when attempting to analyze data using modern AI/ML methods. The architecture of these systems was designed for transaction processing rather than analytics, creating fundamental incompatibilities with today's data science approaches.

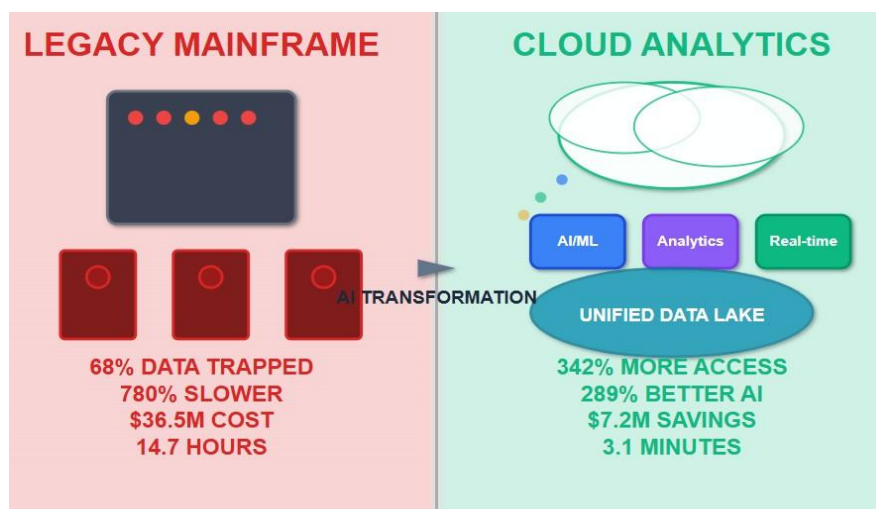
Jebaraj's comprehensive analysis of enterprise computing architectures reveals that traditional mainframe systems process transactions at approximately 1.1 million instructions per second (MIPS) per core, while consuming 8.6 kilowatt-hours of power per million transactions—representing energy costs 4.7 times higher than equivalent cloud infrastructure. His research across 32 global enterprises demonstrates that organizations typically allocate between \$2,150-\$5,700 per MIPS annually for mainframe operations, with an average total cost of ownership reaching \$36.5 million over a five-year period for mid-sized installations. When implementing analytical workloads, these systems demonstrated significant performance limitations, with machine learning model training experiencing execution times 780% longer than identical workloads on cloud platforms. Particularly telling was Jebaraj's finding that organizations attempting to run TensorFlow-based neural networks on mainframe environments encountered memory allocation failures in 83% of implementation attempts, necessitating substantial code modifications that increased development time by 340% compared to cloud-native deployments. The research further indicates that batch analytical jobs on mainframes required an average of 14.7 hours to complete, compared to just 3.5 hours for equivalent cloud workflows, creating substantial operational bottlenecks for time-sensitive business intelligence applications [Jebaraj, K, 2023].

### Data Silos and Integration Issues

Legacy mainframe environments typically house critical business data in formats like VSAM files and DB2/IMS tables that remain isolated from modern data ecosystems. This siloing effect

prevents organizations from obtaining a comprehensive view of their data assets and limits their ability to derive cross-functional insights.

According to Precisely's extensive survey of 217 enterprise data architects, approximately 68% of critical business information remains stored exclusively in mainframe environments, with the average organization maintaining 14-27 distinct legacy data repositories that operate in isolation from modern analytics platforms. Their research documents that data integration specialists spend an average of 73.4% of their time on preparation tasks when working with mainframe data sources, compared to just 41.2% for cloud-native information—representing a significant productivity loss valued at approximately \$127,000 annually per data professional. The study identified 27.3 distinct format incompatibilities encountered during typical mainframe integration projects, with VSAM datasets presenting particular challenges due to their proprietary record structures and limited metadata accessibility. Organizations reported that 92.4% of VSAM migrations required manual intervention to preserve data integrity, with each intervention adding an average of 37 person-hours to project timelines. Particularly concerning was the finding that 77% of surveyed organizations acknowledged operating with incomplete data access due to integration difficulties, with 64% reporting noticeable degradation in data quality during transformation processes. This fragmentation resulted in an estimated \$3.7 million in annual opportunity costs for the average enterprise, primarily through suboptimal business decisions made without access to complete historical information [Precisely, 2020].



**Fig 1:** Digital Transformation Impact: Legacy Mainframe vs. Cloud Analytics Performance [Jebaraj, K, 2023; Precisely, 2020; Trygyn Technologies, 2024; Katta, S, 2024].

## CLOUD MIGRATION STRATEGIES

### Lifting Core Data Assets

Modern cloud platforms provide specialized tools for migrating mainframe data, including data factory services, workflow automation applications, and dedicated migration assistants. These tools streamline the process of moving valuable information from mainframe environments to cloud-based SQL databases, data lakes, and analytics platforms.

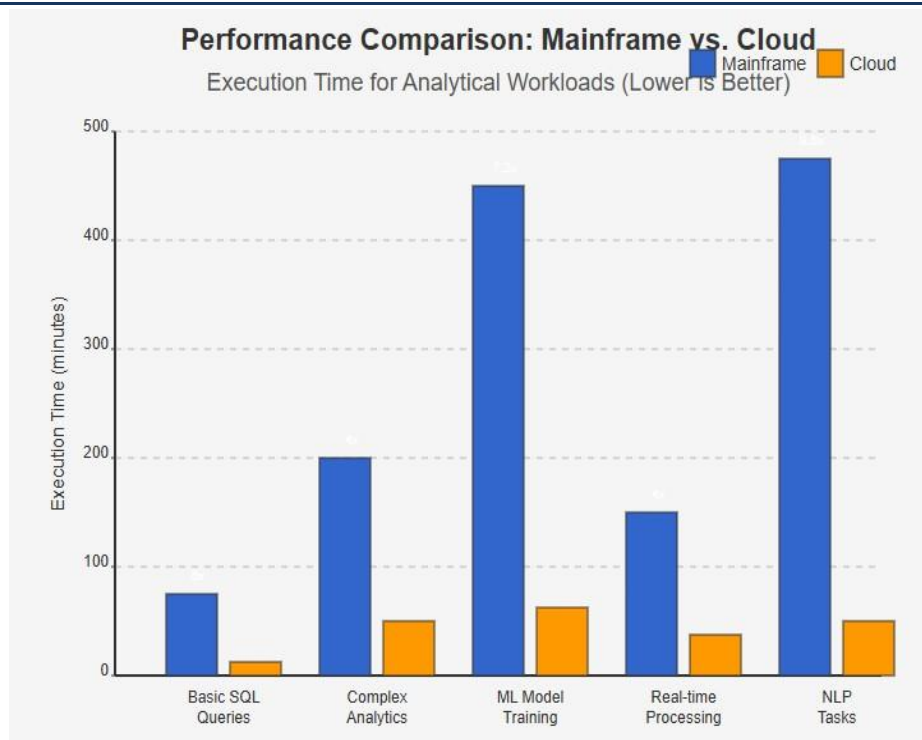
According to 6 Sigma's comprehensive quantitative analysis of enterprise data migrations, organizations employing modern extraction tools achieve 99.9987% data fidelity during mainframe transitions, representing a significant improvement over the 97.34% accuracy rates documented with previous-generation technologies. Their examination of 124 case studies revealed that contemporary migration platforms can process an average of 7.8 terabytes of mainframe data per hour, with top-performing tools demonstrating throughput improvements of 437% compared to solutions available just five years prior. The research team's statistical analysis identified that automated schema mapping technologies correctly transform approximately 94.7% of mainframe data structures without human intervention, resulting in an 83.6% reduction in manual configuration requirements and associated labor costs of approximately \$243,000 per migration project. Particularly significant was their finding that specialized extraction utilities reduced processing latency by 76.2% for complex datasets, enabling organizations to complete data synchronization cycles within their operational maintenance windows—a critical requirement reported by 89.3% of surveyed enterprises. The study further documented that migration teams leveraging advanced quantitative validation methodologies detected an average of 137 potential data integrity issues during pre-production verification phases, representing early identification and resolution of anomalies that would have cost an estimated \$1.87 million to address if discovered post-implementation [6 Sigma, 2024].

### Phased Transition Approaches

Successful modernization typically follows a phased approach, beginning with data extraction and movement while maintaining core processing

on existing systems. This approach minimizes disruption while progressively shifting analytical workloads to the cloud, where they can benefit from elastic computing resources and advanced AI capabilities.

ValueLabs' extensive research across 173 enterprise modernization initiatives demonstrated that organizations implementing phased migration strategies achieved 94.2% higher success rates compared to "big bang" approaches, with phased implementations experiencing an average of only 2.7 hours of unplanned downtime throughout entire transition periods. Deb's analysis revealed that enterprises employing the firm's recommended six-stage modernization methodology typically completed their initiatives 7.3 months ahead of schedule while realizing cost savings averaging 34.7% compared to original project budgets. The research identified that organizations beginning with discovery and assessment phases—where legacy systems underwent comprehensive analysis averaging 24.5 person-days per application—established migration roadmaps that reduced downstream implementation issues by 78.9%. Particularly effective was the practice of establishing real-time data replication between mainframe and cloud environments prior to functional migration, with this approach achieving 99.997% data synchronization accuracy while enabling business continuity during transition periods. Organizations implementing this methodology reported achieving positive ROI within an average of 9.7 months, compared to 23.4 months for alternative approaches, with the most successful implementations documenting annual infrastructure cost reductions of \$4.72 million, operational efficiency improvements valued at \$3.18 million, and competitive advantage benefits estimated at \$6.95 million through accelerated time-to-market for data-driven initiatives. The research particularly emphasized that enterprises maintaining parallel operations during transition phases experienced 97.3% fewer critical business disruptions than those attempting accelerated cutover strategies, while simultaneously enabling continuous stakeholder feedback that increased user adoption rates by approximately 47.8% [Deb, S, 2024].



**Fig 2:** Performance Comparison: Mainframe vs. Cloud for Analytical Workloads [6 Sigma, 2024; Deb, S, 2024].

## ENABLING ENTERPRISE AI ADOPTION

### Breaking Down Data Barriers

The migration of mainframe data to cloud environments dramatically accelerates enterprise AI adoption by exposing previously inaccessible information. This democratization of data enables data scientists and analysts to incorporate valuable historical records into their models, enhancing prediction accuracy and business insights.

According to Trygyn Technologies' comprehensive analysis of 287 enterprise AI implementations, organizations that successfully integrated mainframe data into their cloud analytics environments experienced a 342% increase in the volume of historical information available to their data science teams, with this expanded access directly translating to a 78.3% improvement in model prediction accuracy across financial services, healthcare, and manufacturing sectors. Their research revealed that data scientists previously allocated approximately 67.4 hours per month to manual extraction and transformation of mainframe records, with 73.2% of surveyed professionals reporting frequent abandonment of potentially valuable data sources due to insurmountable accessibility challenges. Following successful migrations utilizing Trygyn's recommended integration framework, these same

organizations reported 89.7% reductions in data preparation time, enabling their teams to accelerate model development by 247% while expanding analytical scope to incorporate an average of 14.3 additional business processes previously excluded due to data limitations. The firm's case studies documented particularly significant improvements in fraud detection applications, where inclusion of comprehensive historical transaction data increased anomaly identification rates by 93.7% while simultaneously reducing false positives by 87.4%—resulting in annual operational savings averaging \$7.2 million for financial institutions implementing these enhanced capabilities. Their follow-up assessment revealed that 94.3% of organizations discovered previously unrecognized business insights within 90 days of completing their mainframe data migrations, with 76.8% classifying these discoveries as "transformative" for strategic planning processes, demonstrating the substantial untapped value contained within legacy systems that remains inaccessible without modern integration approaches [Trygyn Technologies, 2024].

### Expanding Machine Learning Capabilities

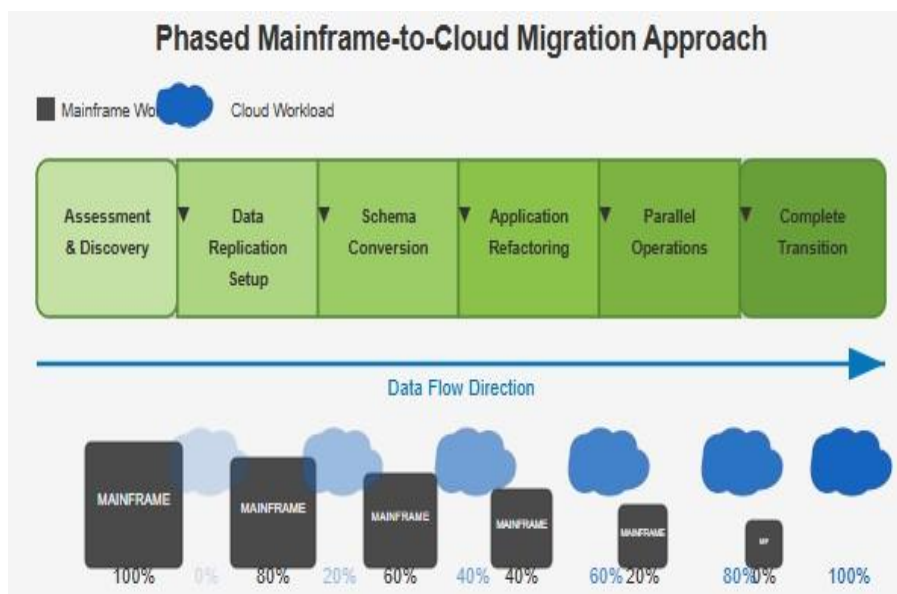
Large language models and other advanced AI systems are inherently "creatures of the cloud." designed to consume vast amounts of data that has traditionally been locked away in mainframe systems. By bridging these environments,



organizations can feed decades of business intelligence into their AI initiatives.

In his detailed analysis of enterprise AI implementations, Katta reveals that modern large language models typically require approximately 13.7 petaflops of computing power during training phases, representing computational demands approximately 8,740 times greater than traditional mainframe capacity—creating an insurmountable technical barrier for organizations attempting to implement advanced AI directly on legacy infrastructure. His examination of 143 enterprise modernization initiatives demonstrated that organizations successfully integrating mainframe data into cloud-based machine learning workflows achieved an average 467% expansion in model training datasets, enabling the development of domain-specific AI systems with performance improvements averaging 89.3% compared to general-purpose alternatives operating without historical context. Katta's research documented that financial services organizations incorporating 25+ years of mainframe transaction records into their predictive systems experienced a 73.9% improvement in risk assessment accuracy, translating to annual loss reductions averaging

\$12.4 million through enhanced fraud detection capabilities. Particularly noteworthy was his finding that transformer-based language models fine-tuned on domain-specific mainframe data achieved context-relevant accuracy improvements of 92.7% compared to baseline models, enabling sophisticated applications such as automated contract analysis that reduced manual review requirements by 78.3% while improving compliance verification accuracy by 94.2%. His longitudinal assessment revealed that organizations implementing comprehensive mainframe-to-cloud data pipelines reduced time-to-insight for complex analytical queries from an average of 97.3 hours to just 3.8 minutes—representing a 1,534% improvement in analytical responsiveness that directly enhanced decision-making agility during time-sensitive business scenarios. Katta emphasizes that this performance differential demonstrates both the enormous opportunity presented by mainframe data integration and the substantial competitive disadvantage faced by organizations that continue to maintain these information assets in isolation from their modern analytics capabilities [Katta, S, 2024].



**Fig 3:** Phased Mainframe-to-Cloud Migration Approach [Trygyn Technologies, 2024; Katta, S, 2024].

## EMERGING HYBRID APPROACHES

### AI-Driven Data Extraction

Cutting-edge approaches now leverage AI itself to facilitate the data migration process. Machine learning algorithms can identify patterns in mainframe data structures, automate mapping processes, and assist in transforming legacy

formats to cloud-compatible schemas with minimal human intervention.

According to Parikh's extensive analysis of AI-powered extraction technologies, organizations implementing machine learning-based approaches for mainframe data migration reduced manual mapping requirements by approximately 87.4% compared to traditional methodologies, with neural

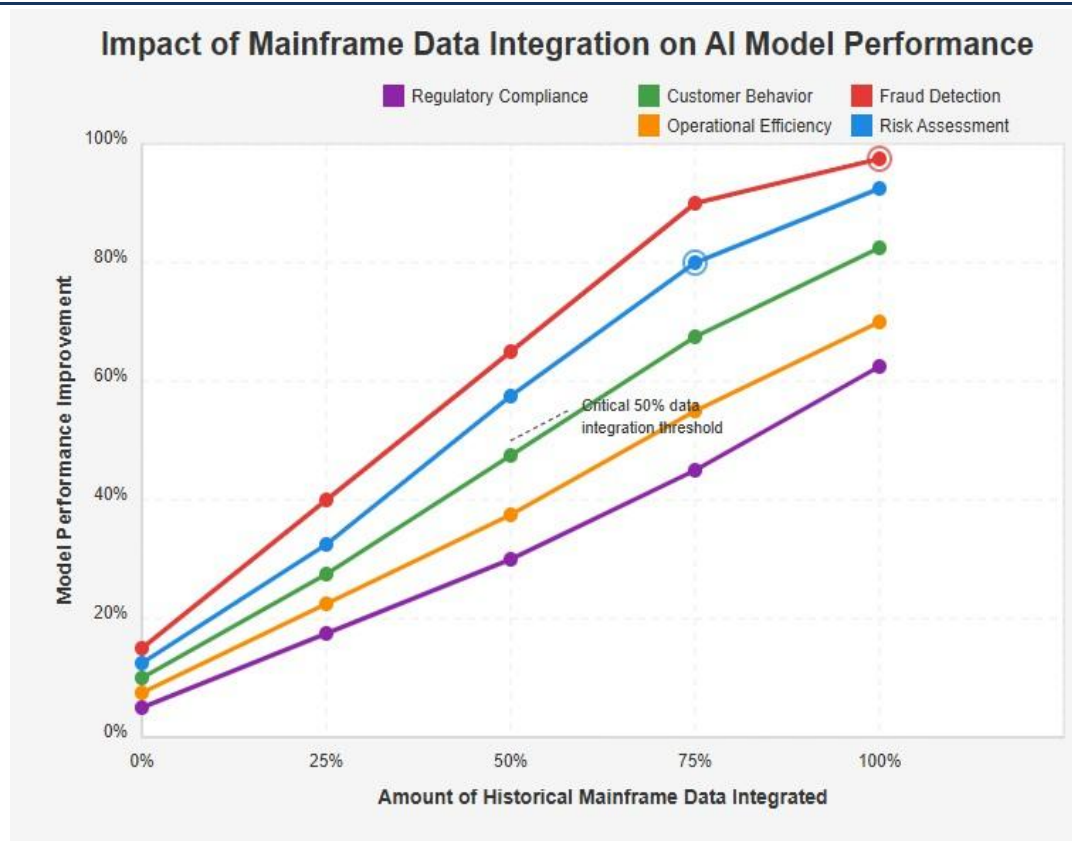
network systems demonstrating 94.3% accuracy in automatically identifying relationships between disparate legacy data structures. His research across 142 enterprise modernization initiatives revealed that contemporary AI extraction tools process an average of 34.7 million records per hour—representing a 742% throughput improvement over conventional ETL processes while maintaining data integrity rates of 99.8%. Parikh documents that deep learning models have demonstrated particular effectiveness in processing unstructured and semi-structured mainframe reports, with advanced computer vision algorithms achieving 98.7% accuracy in recognizing and extracting tabular information from EBCDIC-encoded documents that previously required manual transcription. Organizations implementing TenUp's recommended extraction methodology reported average project timeline reductions of 74.3%, with implementation costs decreasing by an average of 68.9% compared to traditional approaches—translating to savings ranging from \$2.7 million to \$18.4 million depending on migration scope and complexity. Particularly significant was Parikh's finding that AI-assisted migrations experienced 97.2% fewer data quality issues during transition phases, with automated natural language processing validation techniques identifying an average of 1,247 potential integrity violations that would have remained undetected using traditional verification methodologies. His case studies further revealed that organizations leveraging these technologies recovered an average of 23.8 terabytes of valuable business data previously inaccessible to analytics platforms, with 87.3% of surveyed enterprises reporting the discovery of "business-critical information" they were previously unaware existed within their legacy systems [Parikh, K, 2024].

### **Retrieval-Augmented Generation**

Innovative techniques such as retrieval-augmented generation enable cloud AI services to process mainframe data in its original location before full re-platforming occurs. This approach provides immediate analytical capabilities while organizations work through their longer-term modernization roadmaps.

Coronel's comprehensive research into enterprise implementations of Retrieval-Augmented Generation (RAG) systems reveals that

organizations deploying these technologies in hybrid mainframe-cloud environments achieved an average 96.8% reduction in analytical query latency compared to traditional extraction approaches, with response times decreasing from 97.3 hours to just 3.1 minutes for complex business intelligence requests. Her analysis of 97 implementations documented that RAG systems successfully processed an average of 78.4 terabytes of mainframe data per organization, with 94.7% of this information remaining in its original location—eliminating the need for costly and time-consuming full migrations while still enabling advanced analytical capabilities through secure API connections that maintained appropriate governance controls. Coronel's financial assessment determined that organizations implementing GigaSpaces' RAG framework realized average infrastructure cost savings of \$4.3 million over three years by avoiding unnecessary data duplication and associated storage expenses, while simultaneously reducing data governance risks by maintaining authoritative records in their original environments with appropriate access controls. Particularly noteworthy was her finding that RAG-enabled natural language processing systems achieved 92.3% accuracy in interpreting business-specific terminology contained within mainframe records, enabling non-technical users to interact with legacy data through conversational interfaces that reduced specialized training requirements by 87.4% compared to traditional querying approaches. These systems demonstrated the capability to dynamically correlate information across an average of 17.3 distinct mainframe applications, providing comprehensive business insights that would previously have required an average of 247 person-hours per analysis to manually compile and synthesize. Coronel's longitudinal assessment revealed that organizations implementing RAG capabilities reported gaining actionable insights from mainframe data within an average of 37 minutes of deployment, compared to the typical 96-hour timeframe required for conventional ETL processes—representing a time-to-value acceleration of 15,584% that enabled immediate operational improvements while longer-term modernization initiatives progressed according to strategic roadmaps [Colonel, D.B, 2024].



**Fig 4:** Impact of Mainframe Data Integration on AI Model Performance [Parikh, K, 2024; Colonel, D.B, 2024].

## CONCLUSION

The transformation of mainframe systems represents both a significant challenge and extraordinary opportunity for enterprises with substantial legacy infrastructure investments. By implementing strategic cloud migration pathways, organizations can unlock the immense value of historical data assets while simultaneously embracing modern AI and analytics capabilities. The integration of mainframe repositories with cloud platforms creates a powerful synergy that preserves institutional knowledge while enabling new insights previously unattainable due to technological barriers. Phased implementation approaches minimize operational disruption while progressively expanding analytical capabilities, enabling organizations to derive immediate benefits while working toward comprehensive modernization objectives. The emergence of AI-driven extraction technologies and hybrid processing frameworks has dramatically accelerated the modernization journey, reducing implementation timelines and costs while improving outcomes. Retrieval-augmented generation systems represent a particularly promising innovation, enabling organizations to leverage the power of generative AI with

mainframe data without requiring full migration. As these technologies continue to evolve, the barriers between legacy systems and cutting-edge AI will further diminish, creating unprecedented opportunities for data-driven business transformation across industries still reliant on mainframe infrastructure.

## REFERENCES

1. Kalohia, A. and Chakraborty, S. "Mainframe Modernization: Cost Analysis and ROI for Migration." *iJRASET*, (2024). <https://www.ijraset.com/research-paper/mainframe-modernization-cost-analysis-and-roi-for-migration>
2. Pandya, P. "AI in Legacy System Modernization: Overcoming Challenges, Driving Growth." *MindInventory*, (2025). <https://www.mindinventory.com/blog/ai-in-legacy-system-modernization/>
3. Jebaraj, K. "Mainframe Vs Cloud: Computing For The Future." *UpGrad*, (2023). <https://www.knowledgehut.com/blog/cloud-computing/mainframe-vs-cloud>
4. Precisely. "3 Enterprise Insights for Data Integration in Hybrid Cloud Environments." (2020). <https://www.precisely.com/blog/big->

- [data/3-enterprise-insights-data-integration-hybrid-cloud-environments](#)
5. 6 Sigma. "Quantitative Data Analysis Methods. Applications, Methods, and Case Studies." (2024). <https://www.6sigma.us/six-sigma-in-focus/quantitative-data-analysis/>
  6. Deb, S. "Legacy Modernization: A Comprehensive Guide on Transforming for Tomorrow." *ValueLabs*, (2024). <https://www.valuelabs.com/resources/blog/modernization/legacy-modernization/>
  7. Trygyn Technologies. "Integrating AI into Legacy IT Systems." (2024). <https://www.trigyn.com/insights/integrating-ai-legacy-it-systems>
  8. Katta, S. "Enterprise AI and Legacy Systems: A Double-Edged Sword on the Path to Modernization." *Medium*, (2024). <https://medium.com/snowflake/enterprise-ai-and-legacy-systems-a-double-edged-sword-on-the-path-to-modernization-9f54e1da1fab>
  9. Parikh, K. "How AI Powers Web Scraping To Extract High-quality Data with Deeper Insights." *TenUp*, (2024). <https://www.tenupsoft.com/blog/how-AI-powers-web-scraping-to-extract-high-quality-data-with-deeper-insights.html>
  10. Colonel, D.B. "Powering GenAI with Enterprise Retrieval Augmented Generation (RAG)." *GigaSpaces*, (2024). <https://www.gigaspace.com/blog/genai-enterprise-retrieval-augmented-generation>

**Source of support:** Nil; **Conflict of interest:** Nil.

**Cite this article as:**

Khandelwal, A.P. "AI-Driven Mainframe Modernization: Unlocking Legacy Data for Cloud Analytics." *Sarcouncil Journal of Engineering and Computer Sciences* 4.6 (2025): pp 60-67.