OPEN
MAINFRAME
PROJECT

Modernization Working Group

From Legacy to Leaders: Mainframe Data meets Artificial Intelligence

v2 Dec 2024

Executive Summary

In today's era of data-driven decision-making powered by Artificial Intelligence (AI) and Machine Learning (ML), the availability of high-quality data is paramount. Mainframe Computers hold vast amounts of data going back sometimes to decades, a goldmine for AI/ML and analytics. However, leveraging this data poses challenges due to the differing computer architecture on mainframes compared to modern commodity servers. This paper aims to summarize options available to allow AI workloads to use mainframe data. It also discusses tools that enable AI workloads to run on mainframes, as well as those designed to transfer data for utilizing AI tools not available in mainframe environments.

Introduction

Mainframes have dominated enterprise computing for decades. They are being used extensively in industries such as banking, finance, healthcare and utilities for processing vast amounts of data in real-time and for running mission-critical software because of their high reliability, availability and serviceability (RAS) features. According to a 2021 report by IBM, “45 of the top 50 banks, 4 of the top 5 airlines, 7 of the top 10 global retailers, and 67 of the Fortune 100 companies leverage the mainframe as their core platform”.[1] They follow the High Throughput Computing (HTC) paradigm which prioritizes executing numerous simple, independent tasks concurrently. Furthermore, features such as high security, high scalability and retro compatibility with legacy applications keep mainframes as critical systems in large-scale business computing.[2] While IBM is often synonymous with mainframes, other respected companies like Broadcom, Hitachi, Dell, and Rocket Software (including the Micro Focus acquisition) also play significant roles in this space. [3]

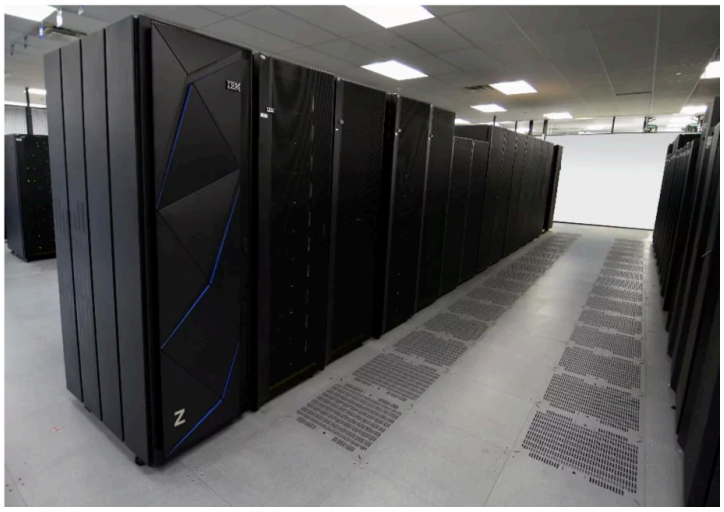

Fig 1. Source: IBM. Modern Mainframes. .

Over the decades, there have always been predictions about the imminent phase-out of mainframes which have been proven wrong. IBM's report shows that mainframes still handle almost 70% of the world's production IT workloads, living up to their nickname "Big Iron".[1] The report's survey also states that 90% percent of IT and business executives view their mainframe as a growth platform, with more than half reporting an increase in transaction volumes over the year.

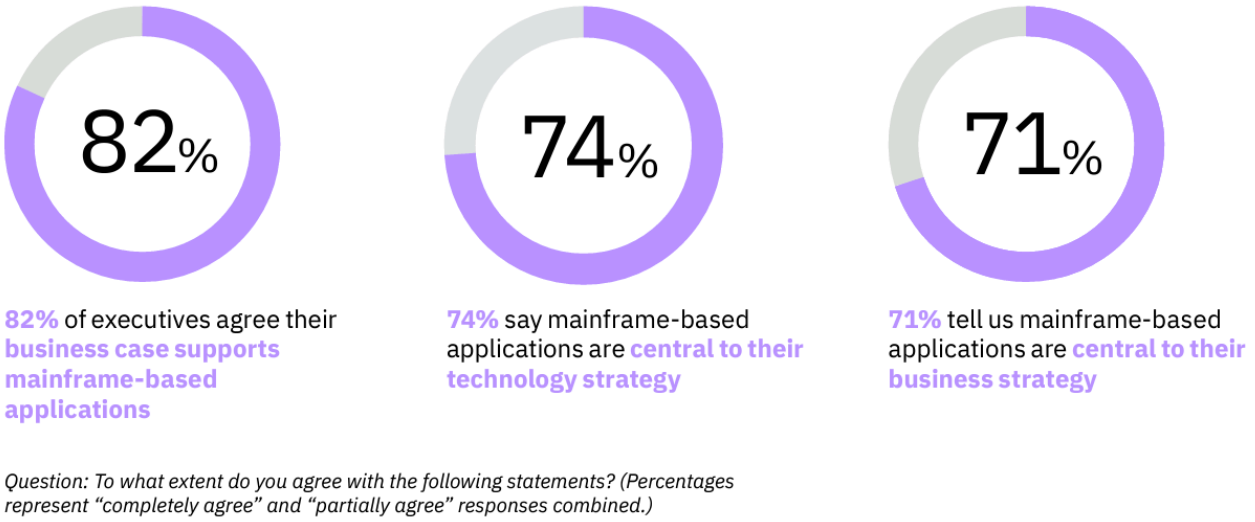

Fig 2. Source: IBM. Most organizations agree that mainframes are central to their business and technology strategy.

Since the 2010s, there has been a push toward digital transformation driven by the Cloud. It is interesting to note that the Cloud's features like the client-server model, usage of thin clients and redundancy were practices which first appeared in mainframes over half a century ago.[4] Currently, the adoption of a hybrid cloud approach in combination with mainframe modernization is gaining traction. [5]

More recently, rapid advancements in artificial intelligence, particularly in generative AI and machine learning (ML) have renewed organizations' focus on accessing and leveraging the wealth of legacy data stored in mainframes.

Thus, the question arises: How can organizations utilize terabytes of invaluable data that was collected by them and take advantage of AI and ML tools to drive data-driven decision-making, generate new revenue streams and gain new insights powered by generative AI?

Current State

Artificial Intelligence (AI) is transforming industries by enhancing operational efficiency, driving innovation and providing competitive advantages to companies that effectively leverage it.

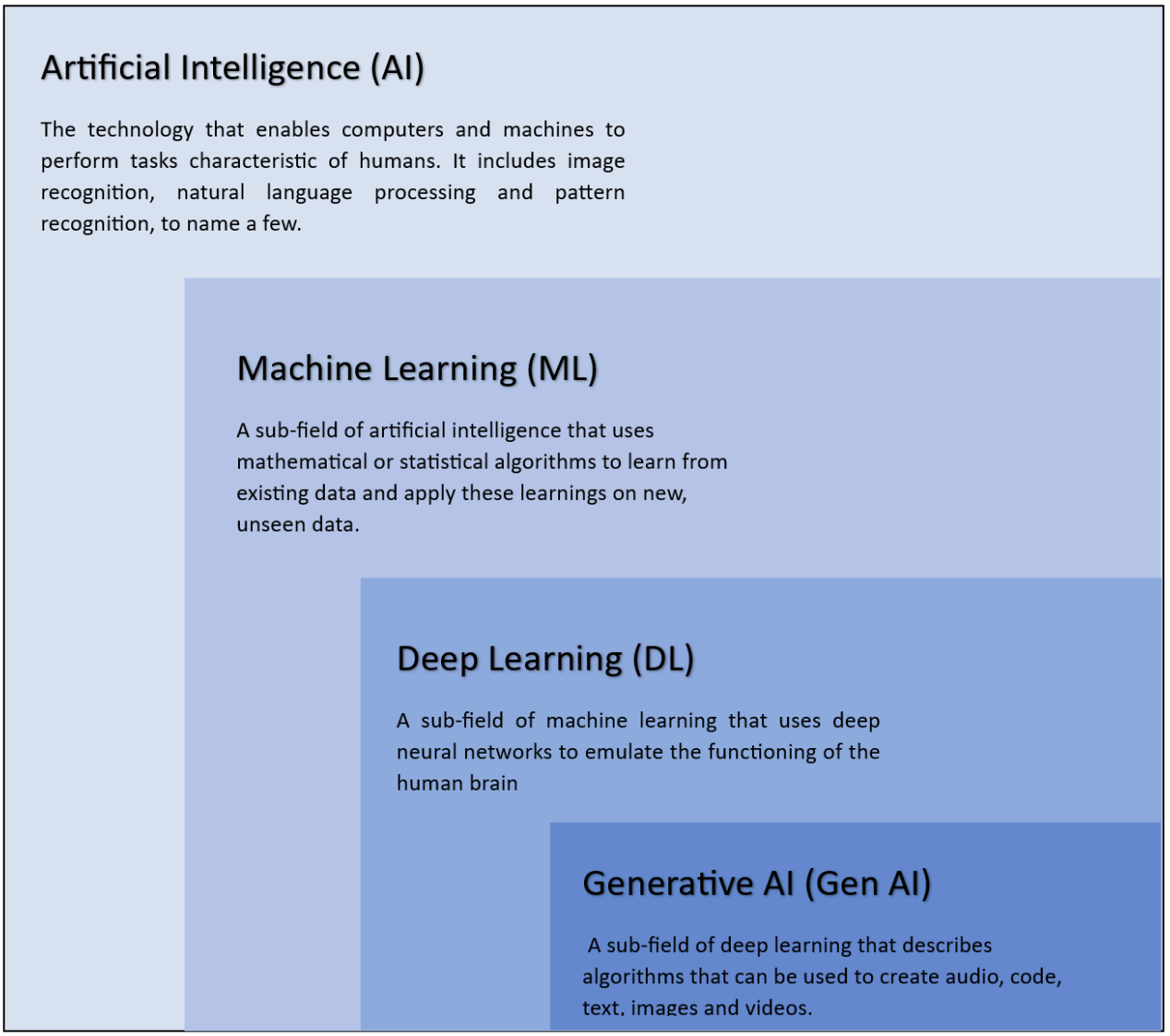

Fig 3. Artificial Intelligence, Machine Learning, Deep Learning and Generative AI at a glance

Despite some concerns, AI is recognized for boosting productivity by automating tasks. Applications like chatbots, recommendation systems, and fault diagnostics have proven AI's versatility across fields such as retail[6], customer service, fault[7] and fraud detection. [8] [9] More recently, Generative AI, powered by Large Language Models (LLMs) has gained prominence for its ability to create new content with minimal human input.

Machine Learning (ML)is a subset of AI rooted in theoretical computer science and applied mathematics, and uses data and algorithmic models to replicate human learning. A typical ML pipeline consists of six stages - Data Collection, Data Cleaning, Feature Engineering , Model Training, Model Testing, Model Verification and Model Deployment.

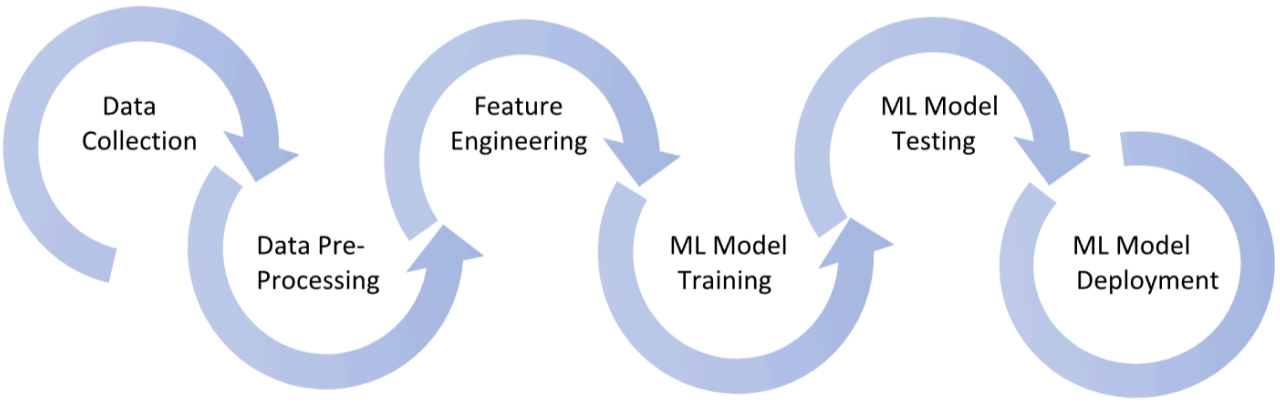

Fig 4. A typical Machine Learning Pipeline. The process is iterative as the model gets trained on new data at regular intervals.

Although ML algorithms have been around since the 1950s, they became industry-relevant in the 2000s due to advances in computational resources and data availability. [10] As British mathematician Clive Humby noted, “Data is the new oil” highlighting the critical role of data in driving industries. Mainframes, across various industries, have for long hosted fairly good quality business data collected over decades which is a gold mine for ML tasks ranging from traditional algorithms to cutting-edge generative AI applications.[11] [12]

When looking at mainframes with its large amount of data, broadly three main approaches exist to allow such data to be consumed by AI workloads, be it generative AI, ML or general analytics and decision-making:

- Running AI on mainframes.
- Running AI outside mainframes without migrating mainframe data.
- Running AI workloads on mainframe data outside mainframes.

Running AI on Mainframes

There are several benefits of running AI workloads on the mainframe. First and foremost, no data is moved out which inherently makes this approach easy to implement, more secure and reliable and at the same time reduces costs involved in setting up separate servers and cloud environments for ML. Additionally, it reduces latency as the data processing happens where the data resides, leading to faster decision-making.

ML models, particularly during the training phase, demand significant computational resources. While modern computers leverage dedicated Graphics Processing Units (GPUs) to parallelize and accelerate this process, older mainframes traditionally lack this capability. To address this, Newer IBM mainframes have the "IBM Telum" chip which allows the CPU to handle code

execution while GPUs manage data inference [13]. Two tools that allow running ML models on mainframes are:

- Machine Learning for z/OS
- SQL Data Insights on DB2

Machine Learning for z/OS:
This is a transactional AI solution that runs natively on z/OS, IBM's traditional operating system for mainframes. It allows the implementation of Python and R's ML and Data Science libraries compatible with the s390x mainframe architecture directly on mainframes and its acceleration is aided by use of the IBM Telum chip. The implementation can leverage pre-built packages of libraries like PyTorch and distributions such as Anaconda[14] [15] or use them within a s390x container.

SQL Data Insights on DB2:
The popular mainframe Database, DB2 offers a feature called "SQL Data Insights" [16] which enables users to run SQL queries to gain insights into their data, such as identifying similarities, clustering, and finding analogies. This helps organizations to discover new patterns in their data, group similar data points and hence leverage the data stored on their mainframe to its full potential.

Running AI outside Mainframes without migrating Mainframe data

This approach is well-suited for three specific scenarios. First, it's a great fit for organizations with older mainframe systems that lack AI-enabled chips like IBM's Telum, where upgrading the mainframe hardware would not justify the cost compared to running AI/ML workloads on traditional GPU-powered servers. Second, it works well in cases where the mainframe's spare capacity isn't sufficient to handle the compute resources needed for AI/ML workloads. Finally, it's suitable when multiple data sources outside the mainframe are involved, especially when challenges like large external data volumes or cross-border data transfers make it difficult to bring that data into the mainframe. Some of the popular tools available which can be used for this approach are:

- Data Virtualization Tools
- z/OS Connect

Data Virtualization - These tools allow for the creation of "virtual views" of data from the mainframe and other enterprise data sources. A virtual view in the context of databases is a representation of data in any customized way without actually creating a new, separate copy of the data with the added advantage of real-time access to data without the need for complex Extract, Transform and Load(ETL) procedures. It also serves as a proxy to the original data, allowing for additional features like data manipulation and content scrubbing, such as masking customer PII data. This is crucial for maintaining data privacy and complying with regulations governing the use of consumer-related data. Rocket Software's mainframe Data Virtualization and Stonebond Technologies' Enterprise Enabler are examples of common data virtualization tools. [17] [18]

z/OS Connect - This tool allows for the creation of APIs that make it easy to access and use mainframe data and applications. The APIs adhere to industry standards(OpenAPI v3) for compatibility and integration. z/OS Connect also supports cloud native development and offers a simple web-interface, which reduces the learning curve for professionals using this tool. Thus it allows applications running outside mainframes to seamlessly integrate with mainframes.

Running AI on Mainframe data outside Mainframes

Organizations might opt to move their data outside mainframes as there exists a well-developed suite of proprietary and open-source ML libraries and tools. In that case, they make use of tools to move the data outside the mainframes in batches or near-real-time. These data movement tools are classified as batch and near-real-time tools.

Batch Tools - Data can be moved in and out of the mainframe through batch jobs which are defined in Job Control Language(JCL). Batch processing on mainframes enables the efficient handling of vast data volumes, often in terabytes, without user interaction. After processing millions of records, these batch jobs generate outputs. Once the data is extracted, it can be seamlessly transferred to the cloud using tools like GCP's BigQuery Connector or Microsoft's Host Integration Server. This integration allows organizations to leverage mainframe data with cloud-native solutions, enhancing analytics and decision-making capabilities. [19] [20]

Near Real Time Tools - These tools capture near-real-time data changes and deliver them to various targets outside the mainframe such as databases, message queues, or ETL solutions. These tools capture changes on the source database and transfer the change data to the target. IBM's InfoSphere Change Data Capture, Rocket Software's Data Replicate and Sync, IBM's Watsonx.data and Precisely's Iron Stream are commonly used near real-time data movement tools.[21] [22] [23] [24]

Once the data is moved out of mainframes, several ML tools can be used on the data for analysis.

Cloud-based Machine Learning Solutions - Cloud-based machine learning (ML) services provide an end to end, fully managed environment for building, training, and deploying ML models. They simplify the ML workflow by abstracting server management thanks to which data scientists and developers can focus on the development and deployment of models. They support collaborative development and can handle big data in distributed environments. Users can choose from a variety of pre-built algorithms or build custom ones from scratch, deploying models securely and at scale with minimal effort.

In addition to these, generative AI services offer pre-trained models, such as foundation models that can be further customized with organization specific data. This enables the creation of specialized AI applications without starting from scratch.

Several major cloud providers, such as Google, Amazon and Microsoft offer similar services, ensuring that regardless of the specific platform, users have access to robust tools for managing the entire ML lifecycle. While opting for these solutions, it's important for organizations to remember that these services are typically tied to a specific vendor, leading to a potential lock-in with the provider's ecosystem.

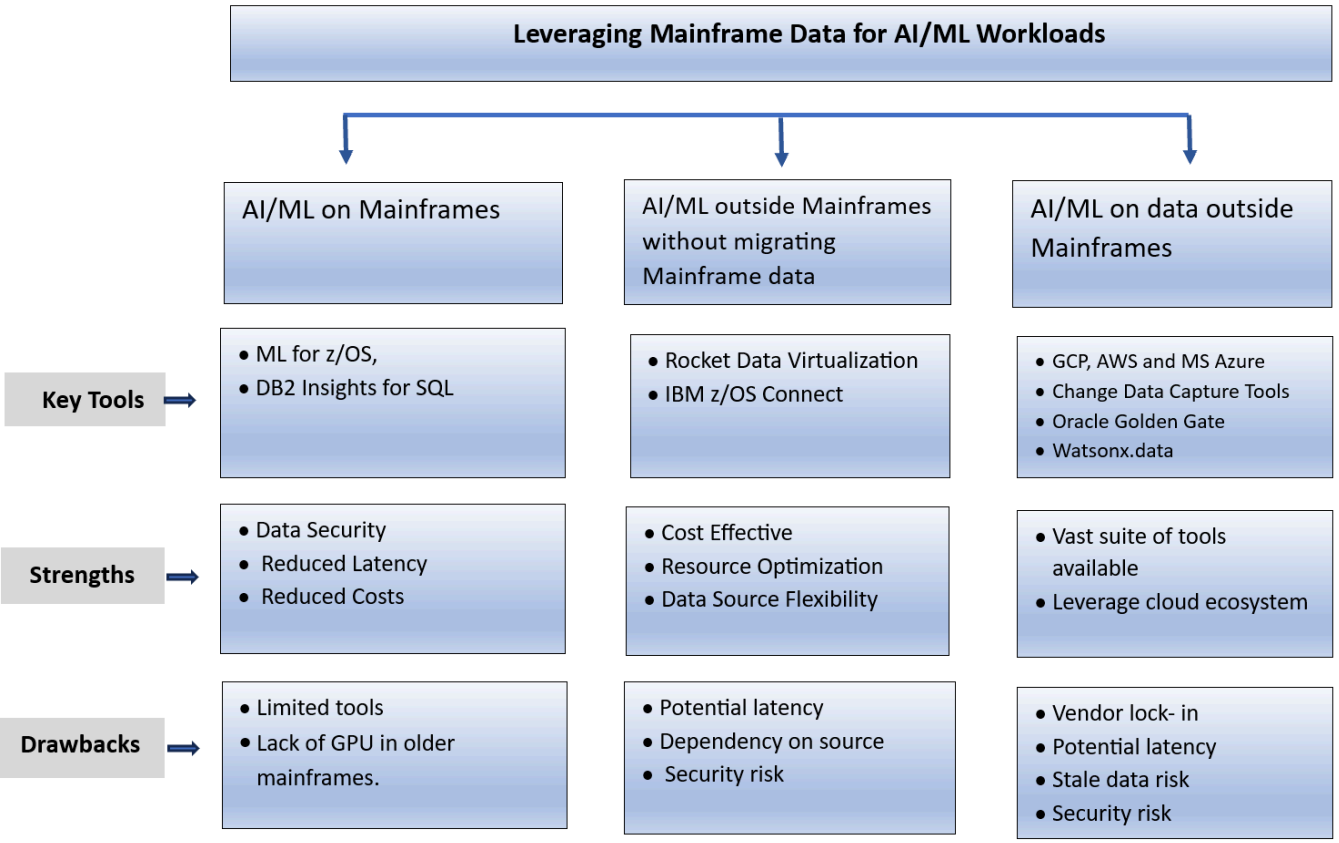

Fig 3. Methods to leverage mainframe data for AI/ML workloads at a glance.

Discussion

In principle, running AI on mainframes seems to be the best option for most enterprises since they are able to leverage the RAS (Reliability, Availability and Serviceability) properties of a mainframe. However, in practice this faces some challenges as older mainframes are not equipped with a GPU (Graphics Processing Unit) by default, which hinders the process of effective inferencing of data. Besides this, the number of tools available to run ML models on mainframes are few. This area has a lot of scope for improvement since popular ML and Data Science libraries such as PyTorch and Anaconda have distributions for the s390x architecture used by mainframes.

The AI ecosystem is highly advanced outside mainframes. Therefore, transferring the data out of mainframes or maintaining a copy of the data on mainframes while executing ML models externally is a common approach for running analytics and deriving insights. This enables organizations to continue utilizing the high throughput processing capabilities of mainframes and running their existing applications on them, while also leveraging modern technologies for AI/ML and analysis of their existing data to aid in making business decisions. However, moving sensitive data outside the mainframe poses security risks and introduces a degree of latency which may pose a problem in large datasets.

Challenges with AI/ML

Despite the numerous benefits of leveraging mainframe data for ML , there are several challenges that can hinder the smooth progression and effective implementation of ML models at the Enterprise level. Some of these challenges include:

1. Data Quality - Since mainframes store data that are decades old, it is important to understand the schema changes of the data. It is also important to effectively pre-process the data to handle missing values and incorrect data. [25] It would be prudent to implement data governance practices on mainframes that ensures that high quality data is available for all downstream tasks.

2. Bias in Data - ML models require large quantities of data for the training stage, it is important to make sure that the data is free from bias, so that ML models don't learn from biased data and end up making biased predictions.

3. Model Interpretability - Interpretability in ML is the extent to which a cause and effect can be observed within a system. All ML models are fundamentally based on mathematics and statistics. While some models, like regression, are straightforward to understand in terms of both their functioning and results, more complex models such as Neural

Networks operate as "black boxes," making it challenging to understand and explain the resulting insights to stakeholders and regulatory bodies. It is important to include mechanisms to explain the reasoning and rationale behind an output. LIME(Local Inference Model-Agnostic Explanation) is an example of one such technique that breaks down the output function of a black box model into local, interpretable models to help understand it better.

4. Privacy - On proprietary ML tools on and outside the mainframe such as AWS SageMaker and WatsonX, metadata and error information are collected to improve services and enhance customer experience and in case of Generative AI solutions, user prompts are stored to train and improve the performance of the model all of which raises concerns about compromising data privacy. Hence it is imperative to understand how ML platforms use customer data and opt out of them where necessary. In the case of LLMs, it's important to carefully review the API's terms of use and fully comprehend their implications for data privacy. [26] It is also important to ensure compliance with relevant state and federal regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA).

5. Less focus on Deployment and MLOps - ML is a rapidly evolving field with new state of the art models and discoveries being made every week. It is important to focus not only on research and development of cutting-edge ML models, but also focus on their scaling, deployment and supportability to ensure that they can be effectively used for real-world applications. [25]

6. Carbon Footprint - Training and running ML models requires tremendous amounts of computational capabilities which consumes abundant resources like electricity and the carbon emissions from such processes are high and detrimental to the climate. With large organizations aiming to reach carbon neutrality within the coming decades, it is a challenge to maintain a balance between the need of staying at the forefront of technological innovation and caring for the environment. A silver lining of running ML workloads on mainframes is that more modern mainframes tend to have a lower environmental impact compared to traditional commodity servers having the same compute capacity. [27]

Conclusion

This paper examines approaches for enabling the use of mainframe data in downstream tasks such as machine learning and data analytics.

The first approach involves running machine learning models directly on mainframes using tools like IBM's ML on z/OS and DB2 SQL Insights. This option allows organizations to keep their data on mainframes and leverage the RAS properties of mainframes.This option is also more environmentally friendly, as mainframes have a lower environmental impact compared to commodity servers. At present, the range of available tools for mainframes are limited and older mainframes typically lack a GPU to efficiently handle modern AI/ML workloads. However, with evolving technologies, like modern mainframes having GPUs and ML Libraries such as Pytorch providing mainframe- specific builds, this approach holds immense potential.

The second approach involves migrating data from mainframes to the cloud. Data movement tools enable data transfer to the cloud. Once in the cloud, the data can be pre-processed and transformed, with the option to maintain in-sync copies on the mainframe using tools like IBM's DB2 data gate. This approach offers access to a suite of good quality proprietary and open-source libraries and tools but introduces potential risks related to data privacy and security due to the usage of third party applications. This approach can end up being overly complex due to data being moved across various systems and also introduces latency.

Overall, all the approaches discussed have their pros and cons. It is important for organizations to do a thorough analysis in terms of the requirements, feasibility and cost before opting for any of these approaches. In conclusion, as AI and ML continue to advance rapidly with cutting-edge breakthroughs, it is crucial to capitalize on the wealth of historical data stored in mainframes to stay competitive.

References

[1] IBM(2021).*Application Modernization on the Mainframe*. https://www.ibm.com/downloads/cas/7BJPNGND?_ga=2.72317459.1696084635.1710142763-2067957453.1707311480

[2] *z/OS Basic Skills*. (n.d.). https://www.ibm.com/docs/en/zos-basic-skills?topic=vmt-who-uses-mainframes-why-do-they-do-it

[3] Zolman, S. (n.d.). *Top 20 Mainframe software suppliers*. https://www.netnetweb.com/content/blog/top-20-mainframe-software-suppliers#:~:text=Major%20hardware%2C%20software%2C%20and%20services,IT%20Services%20Companies%20for%202022

[4] Das, T. (2022, December 20). *Mainframe vs cloud computing: Know the similarities and differences*. Planet Mainframe. https://planetmainframe.com/2022/12/mainframe-vs-cloud-computing/

[5] Kyndryl(2023). *Kyndryl's 2023 State of Mainframe Modernization Survey Report*. https://www.kyndryl.com/content/dam/kyndrylprogram/cs_ar_as/state-mainframe-modernization.pdf

[6] *Decking the aisles with data: How Walmart's AI-powered inventory system brightens the holidays*. (n.d.). Decking the Aisles With Data: How Walmart's AI-powered Inventory System Brightens the Holidays. https://tech.walmart.com/content/walmart-global-tech/en_us/blog/post/walmarts-ai-powered-inventory-system-brightens-the-holidays.html

[7] S. Gupta, A. Venugopal and M. Jidhu Mohan, "Fault Detection and Diagnosis using AutoEncoders and Interpretable AI - Case Study on an Industrial Chiller," 2022 IEEE International Symposium on Advanced Control of Industrial Processes (AdCONIP), Vancouver, BC, Canada, 2022, pp. 198-203, doi: 10.1109/AdCONIP55568.2022.9894262

[8] Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. In *McKinsey & Company*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

[9] Nielsen, J. (2024, January 30). *AI improves employee productivity by 66%*. Nielsen Norman Group. https://www.nngroup.com/articles/ai-tools-productivity-gains/

[10] *Machine learning, explained | MIT Sloan*. (2021, April 21). MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[11] *Mainframe data integration for digital innovation & cloud analytics*. (n.d.). Software AG. https://www.softwareag.com/en_corporate/resources/mainframe-modernization/wp/mainframe-data-integration.html

[12] Forrester(n.d.) *Mainframes are a part of modern IT strategies.* https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Alliances/forrester-mainframes-critical-part-modern-it-strategies.pdf

[13] Jacobi, C.,(2021, August 31). *IBM Telum Processor: the next-gen microprocessor for IBM Z and IBM LinuxONE*. IBM Blog. https://www.ibm.com/blog/ibm-telum-processor-the-next-gen-microprocessor-for-ibm-z-and-ibm-linuxone

[14]Ibm. (n.d.). *GitHub - IBM/pytorch-on-z: Prebuilt PyTorch Packages for IBM Z and LinuxONE*. GitHub. https://github.com/IBM/pytorch-on-z

[15] Anaconda. (2023, March 8). *Anaconda | Anaconda brings data science to Linux on IBM Z and LinuxONE*. https://www.anaconda.com/blog/anaconda-on-ibm-z-and-linuxone

[16] *IBM Developer*. (n.d.). https://developer.ibm.com/articles/use-ibm-db2-sql-data-insights-to-uncover-hidden-relationships-in-your-data/

[17] *Rocket® Data Virtualization*. (n.d.). Rocket Software. https://www.rocketsoftware.com/products/rocket-data-virtualization

[18] Admin. (2018, June 27). *Mainframe Data Possibilities with Enterprise Enabler® |Stone Bond Technologies*. Stone Bond Technologies. https://stonebond.com/mainframe-data/

[19] *z/OS Basic Skills*. (n.d.-b). https://www.ibm.com/docs/en/zos-basic-skills?topic=processing-mainframes-working-after-hours-batch

[20]*Mainframe Connector documentation | Google Cloud*. (n.d.). Google Cloud. https://cloud.google.com/mainframe-connector/docs

[21] *InfoSphere Data Replication 11.4.0*. (n.d.). https://www.ibm.com/docs/en/idr/11.4.0?topic=replication-data-vsam-zos-remote-source

[22]*Rocket® Data Replicate and sync*. (n.d.-b). Rocket Software. https://www.rocketsoftware.com/products/rocket-data-replicate-and-sync

[23]Precisely, Inc. (2024, August 21). *Ironstream: Integrate mainframe & IBM i systems into IT analytics platforms*. Precisely. https://www.precisely.com/product/precisely-ironstream/ironstream

[24]*IBM Data Gate*. (n.d.). https://www.ibm.com/products/data-gate

[25]Bankhwal, M., Bisht, A., Chui, M., Roberts, R., & Van Heteren, A. (2024). AI for social good: Improving lives and protecting the planet. In *McKinsey & Company*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/ai-for-social-good

[26] OpenAI API Terms of Use(2023)
https://openai.com/policies/terms-of-use/

[27] Bandy, M. (2023, September 28). Mainframe & enterprise sustainability: Become an energy efficient juggernaut. *blog.share.org*.
https://blog.share.org/Article/mainframe-enterprise-sustainability-become-an-energy-efficient-juggernaut

Glossary of Abbreviations

This list highlights some common abbreviations used throughout the white paper.

- AI- Artificial Intelligence
- API- Application Programming Interface
- AWS- Amazon Web Services
- CPU- Central Processing Unit
- ETL- Extract, Transform and Load
- GCP- Google Cloud Platform
- GPU- Graphics Processing Unit
- Gen AI- Generative Artificial Intelligence
- LLMs- Large Language Models
- ML- Machine Learning
- RAS- Reliability, Availability and Serviceability
- SQL- Structured Query Language

About this paper

Authors

This document was developed under the Open Mainframe Project's Summer 2024 Mentorship program and sponsored by the Mainframe Modernization Working Group.

Swathi Rao is a student currently pursuing her Bachelor of Technology in Computer Science and Engineering at PES University, Bengaluru, India. Her areas of interest are software development, enterprise business systems, fintech, game development and technical writing. She is keen to work on innovative solutions in emerging technologies.

Dr. Vinu Russell N. Viswasadhas is the Associate Director, Data Consulting at Kyndryl. He holds a Doctorate in Computer Science with a concentration in big data analytics from Colorado Technical University. He is also an Adjunct Professor teaching machine learning at South College, Tennessee. He specializes in data modernization, designing custom data strategies, custom data movement patterns to migrate and democratize data and enterprise data governance.

Acknowledgements

We are incredibly grateful for the useful suggestions and feedback from members of the Mainframe Modernization Working Group, part of the Open Mainframe Project. We would like to thank Bruno Azenha, Misty Decker and Aditi Rai for their helpful feedback that greatly helped shape the content in this white paper. We would also like to thank Mr. Ramesh Vishveshwar, Client Architect, IBM for his valuable insights on the AI ecoystem on mainframes. We hope this white paper will serve as a useful reference document for developers, project managers and decision makers who are looking to gain insights in the AI for mainframe data space.

About the Modernization Working Group

The Modernization Working Group, part of the Open Mainframe Project, intends to be a focal point for thought leadership around what it means to modernize mainframe applications.

We are an open group passionate about knowledge sharing, leading constructive discussions, challenging status quo, exploring creative solutions, and generating contents that will help the community in their journey.

Learn more about the working group at:

https://openmainframeproject.org/our-projects/working-groups/modernization-working-group/