Bringing AI Workloads to the Mainframe

Why high-fidelity, low-latency ML data and AI inferencing can thrive on mainframe platforms like IBM Z

Jason English
Principal Analyst, Intellyx

October 2021

What will it take to move AI from pilot projects to real enterprise operation?

We've been pitched a fantastic vision over the past decades about the expected capabilities of artificial intelligence (AI).

Maybe AI appeared in our collective consciousness via enthusiastic TED talks, or perhaps the thread started in our favorite sci-fi shows and novels. According to some storytellers, we should be immersed in a general form of AI soon and have 'robots with personalities' at our service.

While that's fun to think about, there's much more productive work to be done by AI today. AI is being used to change the game in some of the most sophisticated arenas in business and science.

The most relevant applications of AI today involve near real-time data interpretation and decision-making—or *inferencing*—to meet critical business needs. We must lean on AI to make inferences about what exactly should be done to resolve complex problems, help customers, or drive new opportunities when moments matter.

- *To an AI engineer,* the speed and performance of how AI is trained to respond to huge volumes of incoming data at each decision point are critical.
- *For a business leader,* AI's value for improving quality of service, bringing new competitive differentiation to market, and avoiding risk are paramount.
- *For all enterprise constituents,* the causes and effects of these critical decision points are already manifested on the mainframe—the digital backbone of the business.

What is bringing sophisticated AI work and machine learning data together on the IBM Z platform?

Challenges to AI productivity

There's no such thing as a useful general-purpose AI—nor is there a simplistic way to design or deliver AI functionality that would be of use to any given company. The bar is set high for achieving productivity and meaningful value from AI investments.

Common challenges for productive enterprise AI include:

- **Picking the right time and place.** Companies are already expected to deliver software and digital capabilities to market faster than ever to meet customer demand, so the success of AI is highly dependent on when and where it is utilized – it's not magical. Whether the new AI-enabled service is a new claims processing phone app or a retail product lookup, if the results fall behind competitor services in value, customers will still jump over to a competitor in an instant.

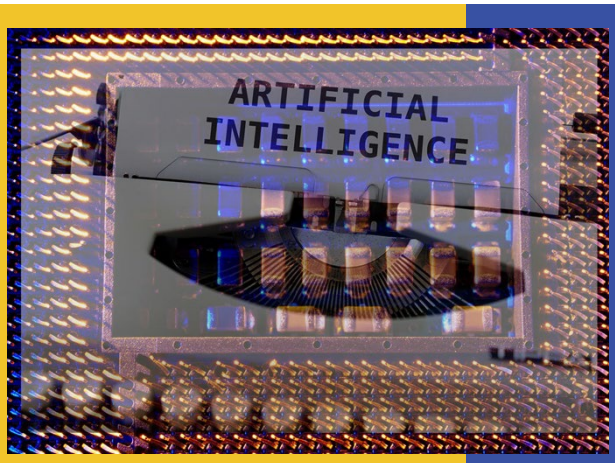
- **Real-time requirements.** There is no tolerance for timeouts and errors in front of customers, and mistakes are costly to undo. Companies are losing money on their digital investments due to slow decisions and faulty data, and AI must deliver real-time improvements.

- **Training and deep specialization.** Industry consortiums and individual businesses will need to train AI to perform very specific functions and recognize a specialized ontology. Therefore, much of the initial labor and cost of implementing an AI program is on the machine learning (ML) side of training data preparation, while that is only the entry point for migration of the resulting data to deep learning models for inference workloads.

- **Trust and regulation.** Regulated industries can't afford shortcuts in deploying AI workloads. Any unforeseen AI miscalculation could lead to a customer data leak, or an accusation of bias, for instance. Such errors can lead to service level agreement violations and regulatory penalties in some cases, so it is important to monitor how AI vendors and service providers prioritize trusted AI as a continuous improvement goal.

- **Latency and performance.** Contrary to conventional wisdom, utilizing distributed x86 infrastructures and cloud resources is not optimal for all use cases. Timeouts can

occur when executing inference logic, especially in mission-critical, time sensitive and privacy-intensive cases. Maintaining high security and low latency at the same time is extremely difficult when data and workloads must move between architectures.



While mainframes weren't initially designed with AI inference in mind, they are already trusted to handle the most mission-critical, secure, data-intensive and high-performance business and government workloads.

All of the above challenges to AI productivity have answers in the mainframe.

Solutions: Realizing the benefits of mainframe for AI work

Modern mainframe platforms like IBM Z are improving in power and capacity all the time, and they can be a central part of an AI strategy, bringing private cloud elasticity and cost flexibility to bear alongside new innovations and the contributions of an open source community.

Companies who implemented AI programs realized increased employee productivity during the recent pandemic, outperforming peers on revenue growth by a 6 percent margin. Rather than leaning on financial definitions of ROI, which don't reflect the economics of adding unique AI capabilities, we can put the benefits of AI on the mainframe into three main categories: *Standardization, Co-location,* and *Optimization*.

Benefit #1: Standardization

The mainframe offers both the easiest place to start making initial investments in AI, as well as a bulwark to future-proofing the value of those investments.

Best-in-class provisioning specifications come standard on the modern mainframe. Within the Z environment, applications, along with attendant storage and compute workloads sit side by side with ultra-fast virtual networking, hypersecurity, and built-in resiliency and redundant backup for disaster recovery and ransomware response scenarios.

Elastic scalability is another feature of mainframes that you might have thought only belonged to cloud infrastructure. In essence, the mainframe usually has enough capacity to scale up like a private cloud, with additional low-cost scaling enabled by containerization techniques or turning on many Z boxes in a sysplex for another dimension of availability.

Interoperability and open standards play into selecting environments for AI work, as data scientists look for ways to compile months or years of previous data analysis and machine learning into new models for solving business problems. Current AI solutions on Z leverage can use TensorFlow PyTorch libraries and other open-source solutions to build

and train on the platform of their choice, so investments are rooted in the data science community.

Compliance tests are par for the course in most critical mainframe application environments, which allows teams to run AI functionality that enhances the organization's ability to prevent account takeover and identity theft, as well as providing an audit trail for reporting to regulatory boards.

Use case: Healthcare review

A healthcare provider uses AI to review medication requests and recommend treatment options for clinicians. The insights generated from this data can improve the provider's patient services, while simultaneously processing reports for HIPAA patient information audits. FFIEC data privacy reviews can also backstop the billing process to ensure secure information stays on the mainframe and is never exposed by AI processes.

Benefit #2: Co-location

Massive AI workloads such as image processing and DNA sequencing are well-suited for public cloud IaaS and highly distributed architectures such as hyperscalers or virtual server farms. Still, the performance, security, and cost benefits of co-locating AI inference workloads alongside other critical enterprise work are often underestimated.

Microservices native. Did you know that much of what we know as microservices architecture didn't start in the cloud? Microservices-based deployments to multi-tenant systems with ephemeral workloads and compute jobs are native to the mainframe.

Today's mainframe delivers excellent performance for containerized deployments. IBM hosts an IBM Z Container Image Registry of trusted Docker images, and with the open zCX container protocol, teams can deploy existing containers and converted Linux boxes right next to CICS systems of record with super-fast, low-latency connections.

Data scientists doing TensorFlow modeling in a Docker container atop Apache Spark data may iteratively model behaviors on a container in their local platform, leverage the zCX container specification and platform-agnostic ONNX framework standards, then convert the model through the IBM Deep Learning Compiler (DLC), and deliver the model into production on a target IBM Z LinuxONE system.

Data gravity. Data gravity should draw inference work to the data, and not the other way around. Even the fastest form of data transfer across distributed and cloud systems

must respect the speed-of-light barrier. Minor pauses for connections and the movement of the AI inference workload itself can quickly add up to an unacceptable condition.

Most core systems still have a large portion of their data originating or passing through Z, and so that data can be harvested in place without the security risks, costs and latency of off-site replication.

The need to rapidly process high volumes of incoming data in the context of a complex inference model—which may incorporate millions of decision criteria and data points of its own—means data gravity exerts a strong pull toward running AI workloads next to data on the mainframe.

Solutions like IBM Cloud Pak for Data further improve this data gravity versus latency equation by allowing queries and interpretation of data while it is at rest in systems of record.

Secure yet portable. Proprietary AI training data, along with the sensitive underlying business and customer data that informs inferences, should remain secure and private—while the inference workload itself should be portable and able to run quickly where best suited. The mainframe features encryption and hyper protect services to protect data at rest, in flight and in use.

AI assets of any known model framework should be easily portable to the mainframe for efficient deployment and reuse. Furthermore, the ability to offload some less time-sensitive workloads to lower-consumption CPU regions like zIIPs can help enterprises realize cost benefits.

Use case: Fraud detection

Credit card fraud is increasing at a rate of more than 35% a year—with companies writing down more than $27 billion in annual losses. Decisions on whether to accept or flag a transaction for fraud interception must be made within seconds. Neither losing customers to interminable wait times nor losing transaction money are viable options.

To counteract this constant fraud problem while keeping service levels intact, a leading bank put their inference engine right next to transactional systems on their Z platform. Their AI was able to examine almost every transaction in real time, at an average rate of 2ms or less—saving approximately $2M a month previously lost to fraud.

Benefit #3: Optimization

When people think of strategic AI, they tend to think of game theory and deep learning systems—such as when Deep Blue was able to defeat the world's leading human chess masters. But the power of AI for optimization goes deeper than that.

Why stop with making tactical improvements, when AI has the power to help companies realize strategic differentiators?

Insight at scale. The fundamental strategic value of AI on the mainframe is the ability to process, filter, correlate and categorize massive amounts of data to assist humans in executing policy-based decisions and making scientific discoveries at a scale and speed we could never achieve via conventional technology platforms.

In some ways, conducting the most strategic AI activities such as collaborative forecasting and core research atop the beating heart of the enterprise—the mainframe—is as much a sign of a next-generation business architecture as the advent of cloud.

Modernizing applications and logic. Thought-leading architects and IT executives are seeking to modernize applications, data and the entire enterprise computing estate with AI as a critical enabler.

Any serious business depends upon a heterogeneous portfolio of applications –from custom-developed applications and vendor-supported ones in multiple clouds – to legacy apps and new microservices apps, SaaS productivity tools and secured networks.

Several SLAs (service level agreements) are usually in place with customers and partners if any part of this complex ecosystem of systems fails. Why not bring the data and modeling of all of these interdependencies onto the mainframe as a fulcrum for insights and improvement?

Conducting AIOps on the IBM Z mainframe (see my previous paper here) offers enterprises a great way to control the incoming storm of security alerts and system monitoring data from this application estate, in order to remediate whatever issues can be automated away and inform administrators when an issue needs human intervention.

Combining AI ensembles. Like a well-practiced band, the most impressive AI applications combine several inference models for even better decision making and handling of very complex processes. Some AI might be delivered in a federated model,

for instance if clients want certain business insights shared with partners, while the data is kept private to keep it away from competitors.

Take for instance a securities transaction. Once a broker or app places an order to buy or sell, a complex multi-party dance between their brokerage, trading exchanges, custodians, banks and regulatory concerns must take place before a final settlement happens. Such trades may be batched into groups with other trades, but each may also have specific timing SLAs or regulatory requirements.

Use case: Mortgage scoring

If we're talking about complex AI optimization, consider the work required for a legal and responsible mortgage and loan scoring process. The 'yes or no' decision on a loan application is anything but binary for an industry that generates thousands of pages of paper documents per transaction.

Fast processing of a home loan to meet transaction deadlines requires an AI that can balance the lender's business rules for acceptance with regulatory requirements on the privacy of credit ratings, while still respecting fair and unbiased housing practices.

The Intellyx Take

The primary obstacle to AI adoption success in the enterprise is risk—even though in some competitive and highly regulated industries, business failure is inevitable without the help of AI.

Executive leaders worry about the risk of wasted time and investment in AI models that aren't feasible or sustainable in today's hybrid IT environments. This uncertainty can get in the way of progress.

For the known future, critical enterprise applications and data will reside on the mainframe—while innovation in machine learning and AI modeling will continue to happen collaboratively, all over the world.

Doing the hard AI inference work on the mainframe strikes the right balance of power, performance, and security.

About the Author

Jason "JE" English (@bluefug) is Principal Analyst and CMO at Intellyx, a boutique analyst firm covering digital transformation. His writing is focused on how agile collaboration between customers, partners and employees can accelerate innovation.

In addition to several leadership roles in supply chain, interactive and cloud computing companies, Jason led marketing efforts for the development, testing and virtualization software company ITKO, from its bootstrap startup days, through a successful acquisition by CA in 2011. JE co-authored the book Service Virtualization: Reality is Overrated to capture the then-novel practice of test environment simulation for Agile development, and more than 60 thousand copies are in circulation today.

Resources: Learn More

Visit the **IBM Z Real-Time Analytics** webpage for more information and assessment tools: https://www.ibm.com/it-infrastructure/z/capabilities/real-time-analytics