

# AI Expedition

## Steering Business Innovation on IBM Z and LinuxONE

Tabari Alexander

Colton Cox

Joy Deng

Purvi Patel

Andrew Sica

Shin (Kelly) Yang



Artificial Intelligence

Data and AI





## Executive overview

Enterprises today are confronted with a spectrum of challenges in the AI domain ranging from complex considerations to fundamental decisions on where to start implementing AI. Navigating this landscape requires not only strategic thinking but also adept decision making as more enterprises strive to unlock the full potential of AI for their business.

This IBM Redbooks® Redguide publication introduces validated strategies that ensure effective and successful outcomes for AI implementation in enterprise workloads. These approaches include cost-effective entry points with robust data management capabilities across IBM Z® and LinuxONE all tailored and seamlessly integrated into enterprise AI workloads.

# Introduction to artificial intelligence in enterprises

Artificial intelligence (AI) is a transformational force that is reshaping the way enterprises operate, innovate, and compete with each other. Since ChatGPT's release over a year ago, AI has captivated the minds of individuals and enterprises to show the vast potential and power that this technology can unleash. The IBM Global AI Adoption Index 2023 shows that, "42% of IT professionals at large organizations report that they have actively deployed AI while an additional 40% are actively exploring using the technology. Additionally, 59% of IT professionals deploying or exploring AI indicate that their company has accelerated their investments in or rollout of AI in the past 24 months."<sup>1</sup> As organizations navigate through the complexity between technology and business, it becomes apparent that AI adoption is more than just a trend. It is a strategic move necessary to stay competitive in any industry.

A key driver behind the sudden importance of AI in the world today is that enterprises need to process vast amounts of data with speed and precision to extract insights that drive business results. In some specific use cases, organizations need to glean real-time business insights from their mission critical, enterprise workloads, which reside on IBM Z and LinuxONE. Embrace optimized open source technology that brings many popular AI frameworks and tools natively and seamlessly into your enterprise. Train your model anywhere and deploy the model on IBM Z and LinuxONE for inferencing to use the industry's first integrated on-chip AI accelerator designed for high-speed, latency-optimized inferencing. AI inferencing can be run at scale with low latency, embedding AI into every transaction with no impact on service-level agreements (SLAs), which empowers the enterprise to make informed decisions in real-time.

This IBM Redguide publication walks you through how to navigate your AI adventure on IBM Z and LinuxONE. First, we cover infusing AI into critical business applications. Second, we discuss cost-effective entry points for using AI on IBM Z. Finally, we show how to manage data across platforms for use on IBM Z with Feature store. Whether it is fraud detection or anti-money laundering, find out ways to easily consume AI capabilities with industry common skills to integrate AI into your enterprise today.

## Infusing AI into critical business applications

Mission-critical applications are types of software applications that are vitally important for business operations of an enterprise, where failure would have a disastrous effect on the revenue, operations, and trust in the organization. Examples of mission-critical applications vary by industry, but include online banking systems, health care systems, and stock exchanges.

Many mission-critical and core business applications have been running on the mainframe for decades. IBM Institute for Business Value (IBV) reveals that nearly 7 in 10 IT executives affirm mainframe-based applications are central to their business and technology strategies. 68% of respondents also assert that mainframe systems are central to their hybrid cloud strategy.<sup>2</sup>

With all the data and applications that run on IBM Z and LinuxONE, enterprises can use AI to solve complex business problems and drive customer delight. AI thrives on data, and organizations have accumulated and saved heaps of data from their transactions over the years. Hence, businesses can generate more value from their data by infusing intelligence into applications.

---

<sup>1</sup> <https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Wide-spread-Deployment-by-Early-Adopters>

<sup>2</sup> <https://www.ibm.com/blog/new-study-reveals-why-mainframe-application-modernization-is-key-to-accelerating-digital-transformation/>

In the banking industry, the use of digital transactions rose during the pandemic as people moved to online banking and shopping. However, this rise in digital transactions also raises the chance of fraud in online transactions. According to a Celent study, fraud generated an estimated US \$385 billion in losses globally to the banking, cards, and payment sectors in 2021.<sup>3</sup> To reduce the risk of revenue loss, organizations began to proactively detect and prevent fraudulent transactions in real time. To achieve the best outcome, it is essential to infuse AI in every transaction without impacting transactional performance. This raises a requirement for a resilient, low-latency, and high-performance platform that is capable of real-time inferencing on mission-critical data.

## **The business solution**

The requirements for a resilient, low-latency, and high-performance platform point to AI infusion into mission-critical applications colocated on IBM Z and LinuxONE.

The recently available IBM z16™ processor has an industry-first on-chip AI accelerator that is designed for high-speed and latency-optimized inferencing. It accommodates 300 billion inference requests per day with a 1-millisecond response time. It delivers consistent response times with optimized inferencing that scales with workloads and scores every transaction while still meeting the most stringent application SLAs. Instead of getting insights after the transaction occurs, IBM z16 helps organizations create value with accelerated AI insights that are applied to each transaction, in real-time, before the transaction completes.

In addition to the optimized inferencing, colocating AI applications with the data on IBM Z systems provides data gravity benefits. IBM Z servers process approximately 30 billion transactions each day, up to 19 billion of which are encrypted. So, it colocates analytics, securely, with the data. Tight integration of AI with data and core business applications that reside on IBM Z allows organizations to leverage the quality of service you expect from IBM Z platform: AI resiliency, 99.99999% availability, and IBM Z flagship security for enterprise data.

Let's revisit the use case of fraud detection in online transactions. Due to SLA requirements, most banks run fraud detection algorithms only on a fraction of transactions in real-time. Many of these transactions trigger fraud detection analysis on a post-transaction basis, drastically limiting the ability to detect fraud and avoid losses. However, with IBM z16 on-chip AI accelerator, banks can score 100% of transactions in real-time, and get response times within application SLAs, to detect and prevent fraudulent transactions. This results in significant cost savings and improved customer satisfaction. Celent estimates that scoring every transaction on the z16 processor can potentially reduce banking, card, and payments fraud losses by US \$161 billion globally.<sup>3</sup>

## **Key products and technologies**

Are you ready to derive insights from the enterprise data in real-time? Then you can put AI to work using the following products offered by IBM®.

IBM Machine Learning for z/OS (MLz, formally known as Watson Machine Learning for z/OS (WMLz)) is an enterprise machine learning solution that runs on the IBM Z platform. This end-to-end machine learning platform enables organizations to infuse machine learning (ML) and deep learning (DL) models to score transactional workloads running on IBM Z to derive real-time business insights at scale. Easy to use web user interface allows you to build and train your ML/DL models on any platform using your framework of choice, including MLz, and easily import them to MLz with a single click. From there the models can be deployed on IBM Z into your most demanding transactional workloads to drive business value in every transaction, without impacting application SLAs or user experience. Tight integration of data and AI provides transactional affinity, a key in achieving low latency.

---

<sup>3</sup> <https://www.ibm.com/downloads/cas/D0XY3Q94>

The Administration dashboard in MLz can manage and monitor models for inaccuracy, such as bias and drift, which helps simplify the efforts that are required to maintain a production level model accuracy.

In addition to REST API calls to the scoring service, MLz also provides:

- ▶ A scoring service that is integrated with native IBM CICS® runtime through the ANLScore program that can be called through a CICS LINK command. This can be used to easily modernize CICS COBOL applications to infuse AI. This solution yields optimized performance as the scoring server is colocated with the transaction within the CICS runtime itself.
- ▶ A scoring service through the IBM WebSphere® Optimized Local Adapters (WOLA) interface for IMS COBOL programs. This enables optimized performance for in-transaction inferencing in a high-volume IMS transactions environment.

Trust in AI is a critical imperative for enterprises as it ensures that AI decisions are transparent, reliable, and unbiased. This is essential for building trust with customers and maintaining brand reputation. By prioritizing trustworthy AI, enterprises can avoid the potential risks that are associated with AI and comply with regulations and leverage the technology to drive growth and innovation. Release 3.1 of Machine Learning for z/OS supports deploying AI models confidently with AI model explainability capability natively available. MLz also integrates with Openscale deployed on Linux on IBM Z for model monitoring capabilities. It also supports scheduling periodic reevaluations of model performance with new data to ensure model accuracy over time and sends alerts when performance deteriorates, allowing you to maintain model accuracy through automatic model refresh.

IBM provides the full-featured machine learning platform, training anywhere or on IBM Z and readily deploying those models on IBM z/OS® applications, colocated with enterprise transaction data and business logic for in-transaction scoring in near real-time without impact to applications' SLAs. It takes advantage of the security, performance, and resiliency of the IBM Z platform to deliver business insights when and where they are needed, at the point of impact, in real-time.

## Cost-effective entry points for leveraging AI on IBM Z

While the potential benefits and business opportunities of enterprise AI are well-documented, it can be difficult to identify the best starting point. And like any major technology, investing in AI requires careful consideration of the costs and returns. In this section, we lay out strategies for identifying the most cost-effective approach for a first-time enterprise AI use case, as well as offer an example of a well-considered starting point.

First, it is worth recognizing that there is no single way to measure cost-effectiveness with a technology investment; it can relate to various explicit and implicit costs, including the areas of licensing costs, program management, and return-on-investment (ROI).

- ▶ How much are you willing to invest in new enterprise software?

While software licensing costs might be one of the largest explicit expenditures undertaken by an enterprise IT team, it also provides assurances that may be critical for success, for example certainty that the technology has the necessary capacity and also access to lines of support and documentation. In the case of enterprise AI, this may also enable faster time-to-market for a first-time use case, with the necessary infrastructure ready for immediate use to support model deployment. It is possible that your team would be seeking a more flexible solution that allows for a more situation-specific approach to your AI use case.

- ▶ How familiar is your team with developing on open source software?

There is no denying that the open source community provides the greatest variety of AI solutions, frameworks, and programming languages available for anybody to pick up and use, individuals and enterprises alike. The fast-paced and collaborative nature of open source development means that the resulting projects often live closest to the “bleeding edge” in terms of all available AI software. It also means that such software can be freely accessed, used, and modified to fit a particular use case. However, to use such software, your team may need to be versed in modern programming languages, as well as skilled enough to assemble your own solutions. All of this assumes that your organization and industry are tolerant of the use of open source software.

- ▶ Are you prepared to develop and maintain your own AI pipeline? If choosing to go the open source route, the investment of time and resources that are needed to set up AI operations for your organization comes into play. Depending on the solution mix, this could include:
  - Data gathering/wrangling solutions
  - Data transformation solutions
  - Model development languages and software (including model training)
  - Model deployment pipeline
  - Model evaluation software

Each of these pieces may require some degree of integration and automation. While utilizing open source can allow you to custom-build each component to your specific needs, you also need to expend time and capital toward planning the right solution mix, development of the pipeline, and maintenance of the result.

The creation of a bespoke AI solution might be exactly what you are looking for, rather than something prepared by an enterprise software vendor.

- ▶ Do you have the available hardware needed to scale an enterprise-grade AI use case?

Importantly, scaling an artificial intelligence pipeline to meet the demands of an enterprise-grade workload can be computationally intensive. In fact, an O’Reilly Media survey, “AI Adoption in the Enterprise 2022”, found that “32% of respondents with AI in production reported that their companies spent over 21% of their IT budget on AI.”<sup>4</sup>

That being said, investing in the right hardware might be one of the easiest ways to realize dividends in the form of enterprise AI at scale. With the IBM Integrated Accelerator for AI in IBM z16 and IBM LinuxONE 4, you can gain more value from incorporating AI processing into runtime applications on IBM Z, colocating inferencing with your core transactional systems. In comparison to the use of distributed platforms for inferencing, you can achieve significantly faster response times through such collocation, allowing for various use cases to operate within SLA windows. Given this, you may more quickly achieve ROI for the use case you are considering when factoring in the right hardware. Later, we provide an example of a fraud detection use case that derived from an IBM Z customer that meets this criteria.

Consider how your organization would respond to each of the previously mentioned questions before determining the architectural approach to your use case.

## **Leveraging open source frameworks for AI on IBM Z and LinuxONE**

Open source might be the most attractive option for many of the reasons previously indicated: flexibility of the solution, total investment cost for software, and operational familiarity for a wider range of developers. If the downsides do not represent a serious concern to your team, then there are a couple IBM offerings worth considering that leverage highly popular open source frameworks.

---

<sup>4</sup> Source: <https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2022/>

### ***Python AI Toolkit for IBM z/OS***

Python AI Toolkit for IBM z/OS is a product delivering industry-leading AI Python packages for availability on z/OS that also leverages IBM supply chain security. This ensures that the Python packages have been scanned and vetted for vulnerabilities that might compromise the safety of the operational environment, which may help to mitigate the concerns of warier organizations who hesitate to tread into open source. With a familiar, flexible, and agile delivery experience, you can access a range of 180+ packages that are built for data science and AI use cases, including:

- ▶ **matplotlib:** A comprehensive library for creating static, animated, and interactive visualizations in Python.
- ▶ **Pandas:** Providing fast, flexible, and expressive data structures that are designed to make working with 'relational' or 'labeled' data both easy and intuitive.
- ▶ **XGBoost:** An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable.

### ***AI Toolkit for IBM Z and LinuxONE***

AI Toolkit for IBM Z and LinuxONE is a family of popular open source AI frameworks adapted for IBM Z and LinuxONE hardware. The AI Toolkit consists of IBM Elite Support and IBM Secure Engineering that vet and scan open source AI serving frameworks and IBM certified containers for security vulnerabilities and validate compliance with industry regulations.

Consisting of serving frameworks, machine learning and deep learning frameworks, and deployment options, all sitting atop the power of the IBM Z Integrated Accelerator for AI, the AI Toolkit enables you utilize and deploy AI frameworks with confidence.

### ***AI applications for IBM Z and LinuxONE***

Perhaps your organization is better suited for a ready-built option, allowing for a use case to be more quickly assembled without the skills needed for a bespoke AI pipeline. With this option, you can also depend on a maintenance stream and development roadmap to continually improve the software. Also, you can depend on the solution to be suited for enterprise-grade workloads, meeting the demands of your mission-critical applications.

### ***Machine learning for IBM z/OS***

Machine learning for IBM z/OS (MLz) allows you to infuse machine learning and deep learning models with z/OS applications, facilitating real-time business insights at scale. When leveraged with the IBM z16 and the IBM Telum® AIU, users can colocate key applications with inferencing requests, allowing for rapid response times that deliver intelligence while meeting service-level agreements. With MLz Enterprise Edition 3.1, users can also access trustworthy AI features, allowing for the generation of visualized explanations for MLz scoring results.

### ***Utilizing zIPs and IFLs with AI workloads***

One challenge for enterprises introducing AI into their mission-critical operations is identifying available, optimal hardware to process the workloads. Acquiring new resources can be a significant expense and a mature IT operation might see full or near-full utilization of hardware resources.

Relief can come in the form of optimized processors available for IBM Z and LinuxONE machines, dedicated to run certain specialized workloads alongside normal operations, including workloads for many common artificial intelligence use cases.



The IBM z Integrated Information Processor (zIIP) is a dedicated processor that is designed to operate asynchronously with the general processors in a mainframe to process new workloads without affecting the million service unit (MSU) rating or the machine model designation that often influences software license charges. zIIPs are designed for select database, cloud, and transaction (Java) processing workloads

Examples of AI software that can utilize zIIPs to achieve optimized hardware costs include:

- ▶ ONNX: Library calls and compiling AI models on Open Neural Network Exchange (ONNX) where the ONNX operators are defined to run directly on z/OS are eligible for zIIP workloads.
- ▶ Python: Up to 70% of Python AI and ML workloads are considered eligible for zIIP.
- ▶ IBM Z AI Data Embedding library of z/OS: When invoked by using the Java native application programming interface, the IBM Z AI Data Embedding library of z/OS is considered eligible for zIIP.
- ▶ Machine learning for IBM z/OS: The availability of zIIP engines allows all training and inferencing conducted with MLz to be eligible for zIIP. You can also deploy Apache Spark on IBM zIIPs to run analytics and machine learning on large, complex data sets.

The IBM Integrated Facility for Linux (IFL) is a processor that is dedicated to Linux workloads on IBM Z and LinuxONE. It allows for the reduction of operational, software, facility, and energy expenses with a high Linux server density. Similar to zIIPs, it does not increase charges for IBM Z software running on “standard” processors, nor does it affect the MSU rating or the IBM Z systems model designation.

## **Prioritizing return on investment**

While you can easily become ensnared in the debate over the optimal software and hardware investment, it may ultimately be easier to consider cost-efficiency for AI by weighing the return on investment of different use cases. From there, you can determine the appropriate technology mix to meet the use case and recoup your investment, thereby opening the window for further use cases. Importantly, many popular first-time use cases for AI across a range of industries have been documented as case studies, providing patterns to try in your own organization.

An example from the banking industry concerns the issue of fraud detection. One US bank found themselves unable to score all their transactions in real time with their existing off-platform scoring engine due to latency. Without the ability to scale fraud detection, 80% of all transactions were unscored for fraud, resulting in millions of dollars exposed to fraud annually.

To mitigate this issue, the bank architects a fraud detection and prevention solution running on IBM Z that utilizes a Python-based machine learning model and LightGBM, a framework that increases model efficiency and reduces memory usage. With this solution, the client achieved 100% real-time scoring, ultimately saving >\$20 M annually in exposure risk.

Such a use case can be optimized with the Integrated Accelerator for AI on the IBM z16 and LinuxONE Emperor 4, colocating the model with transactional systems.

In a study commissioned by IBM, Celent determined that “running advanced AI models directly in the mainframe environment is a powerful innovation in an industry where an estimated 70% of global transaction value runs on IBM mainframes.”

For US banks, where fraud losses average 9.3¢ per \$100 transacted, an advanced inferencing model on the IBM z16 could reduce losses to 3.7¢, a 60% improvement.<sup>5</sup>

<sup>5</sup> <https://www.ibm.com/downloads/cas/D0XY3Q94>

### **Example architecture: Fraud detection with AI Toolkit for IBM Z**

While there are multiple ways to achieve real-time fraud detection with z/OS applications, one approach you may consider utilizes the AI Toolkit for IBM Z, allowing access to a suite of lightweight and free to download tools and runtime packages. Coupled with zIIP eligibility through IBM z/OS Container Extensions (zCX), and making use of the IBM Z Integrated Accelerator for AI, this architecture may offer a rapid path toward a return on investment.

The example is detailed in Figure 1, and can be broken down as follows:

- ▶ A credit card authorization application running on IMS can be REST-connected to a zCX instance colocated on an IBM z16.
- ▶ Within the zCX instance, the AI Toolkit for IBM Z allows for use of the IBM Z Accelerated for NVIDIA Triton Inference Server.
- ▶ A Neural Network Fraud Detection model can be deployed and loaded with the IBM Z Deep Learning Compiler through the Triton Python Backend.
- ▶ Once loaded, the model can invoke the IBM Z Integrated Accelerator for AI with every model inference, achieving real-time scoring.

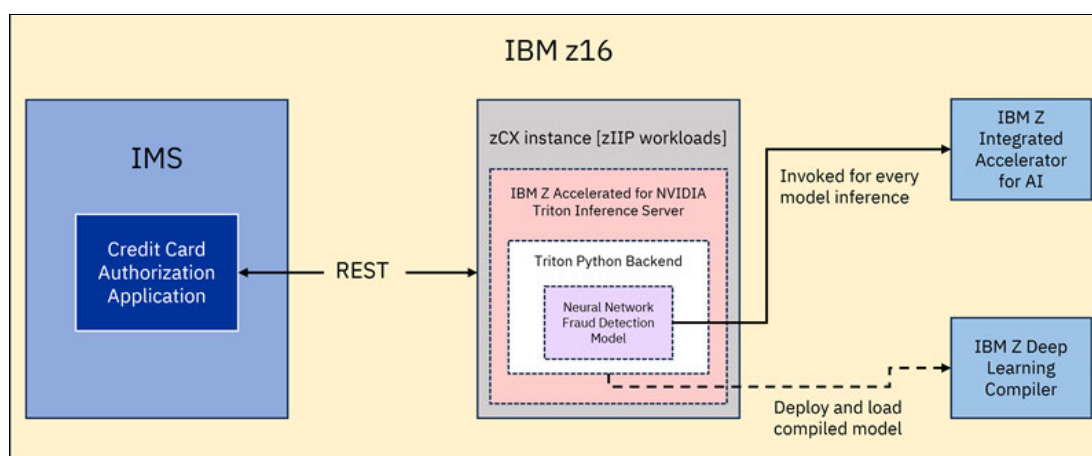


Figure 1 Example architecture: Fraud detection with AI Toolkit for IBM Z

## **Feature store: Managing data across platforms for use on IBM Z**

Data is at the core of any AI use case. Not only does data provide the historical content and context that a given model is trained on, but it also provides the input and basis for new inferences that are made with a trained model. It should come as no surprise then that data and data access may also create a great number of challenges as you explore an AI use case. For example, the quality and quantity of the data sources that are used for training an AI model can greatly impact the insights that are delivered when the model is put to productive use - 'garbage in, garbage out' as the saying goes.

The challenges with data are not only limited to data quality and quantity. Data privacy and security must also be considered end to end for any enterprise use case, and the ability to govern data used both for training and inference is a key benefit of products such as IBM Cloud Pak for Data.

Another common production challenge that is overlooked is the availability of additional input data (features) that does not reside on the platform. This is critical when dealing with a real-time use case that has strict SLA requirements. Some common architectures can be used to address the issue, including the use of a feature store that enables the use of pre-transformed features in real-time.

## Understanding the scenario

In the early stage of an AI use case, a data scientist generally explores and analyzes data pertinent to the specific problem they are addressing. As they analyze, they consider different modeling techniques with the goal of using AI to infer insights (such as making a prediction on whether a transaction is likely to be fraudulent). Their goal through this exercise is to identify an approach (data pipeline and AI model) that produces the most accurate prediction.

In several cases, this exploration may result in the creation of new, additional features that augments the set of existing data points. These new features serve to augment available real-time data to enable the AI model to produce more accurate insights. Common examples include:

- ▶ Recent, historical events (transactions) to produce a more accurate prediction for use cases involving a sequence of events. This enables models like Long-short term memory (LSTM) to build up a state (or context) based on the recent events.
- ▶ Aggregated or enriching features based on historical data.

Consider a real-time credit card fraud detection scenario. In practice, as credit card transactions move through authorization processing, there is a set of “new” data that is associated with that specific transaction. This data includes a set of features that are associated with the cardholder, the merchant processing the request, as well as the transaction details (date, time, amount, and so on).

While some of these features may be critical input for the AI model detecting fraud, a data scientist may find they are able to improve accuracy by using additional data based on historical transactions. For example, short term (30 day) averages of the card holder transaction value, or a recent history of the geographies where their transactions originated, can enable the AI model to take into account recent behaviors. This additional context can help produce a more accurate prediction.

These additional features can be preprocessed and transformed before model execution. Many clients store these preprocessed features in an existing AI feature store, a repository that is commonly used to make preprocessed data available for reuse.

As these use cases progress toward production, some challenges may arise. One challenge comes from such implementations where there are also strict application SLA requirements. In these cases, a key benefit of deploying AI alongside a business application on IBM Z is to keep latency low and improve scalability; thus, calling to an off-platform feature store to gather additional historical data introduces intolerable latency and scalability challenges.

The good news is there are strategies to address these concerns and bring outside data back to the platform.

## Leveraging feature store on IBM Z for low latency inference use cases

Bringing data back to IBM Z can enable a more accurate prediction, the benefits of which can be astronomical.

Colocating a high performing feature store on IBM Z alongside the AI inference service and business application provides for the most cohesive solution.

In this context, the feature store serves as a localized, high performing data store containing the pre-transformed set of features. The data store technology that is used should be capable of serving real-time requests with low single digit millisecond response times while scaling to meet demands.

In a typical end-to-end solution flow for a real-time use case, the inference service would expose an API that is invoked by the business application, as shown in Figure 2. The business application would provide the required raw transaction data on the API request. When invoked, the API would pass this data to a scoring handler which, based on key data from the transaction, queries the feature store to extract relevant pre-transformed features as shown in step 1. The scoring service then builds the AI model input tensor, which is now inclusive of raw data and the pre-transformed features. Finally, the inference request is issued in step 2. An overview of this architecture is shown in Figure 2.

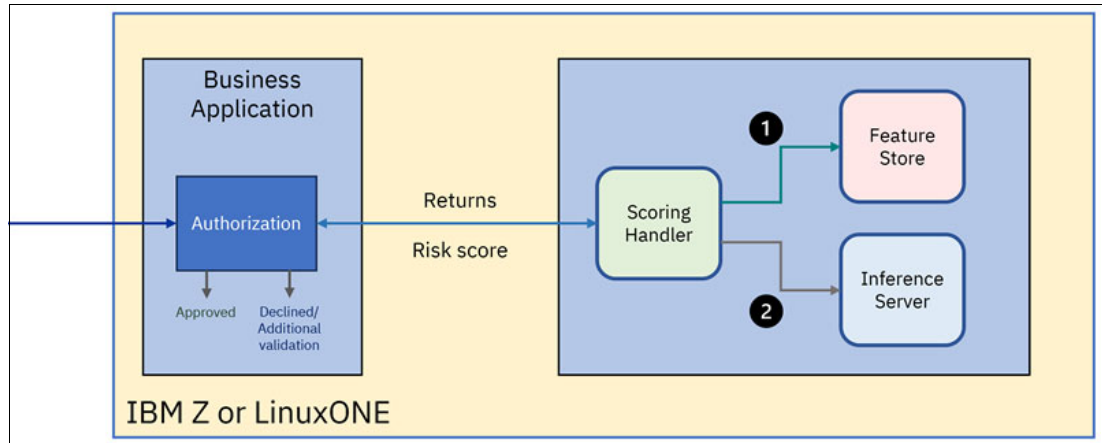


Figure 2 Leveraging feature store on IBM Z for low latency inferencing architecture overview

Various specific technologies can be leveraged to create a feature store, both on z/OS and for Linux on IBM Z environments. Many clients have chosen to use in-memory databases for a real-time use case feature store. There are a number of potential technologies that can be leveraged, including Hazelcast, Redis, and Red Hat DataGrid - all of which can scale to satisfy high rates of read requests in parallel with the qualities of service needed for online scoring solutions. In z/OS environments, these Linux based options may be deployed in z/OS environments by using zCX.

As mentioned prior in this chapter, in addition to storing these additional data points, it is advantageous to preprocess them in advance as well - that is, transform them to the format required by the AI model. This further decreases the processing overhead during an inference request.

### **Addressing consistency between IBM Z and off-platform feature store**

In addition to these real-time usage considerations, another common concern is consistency between the IBM Z and off-platform feature store. This concern arises when the enterprise has off-platform (x86 or cloud-based) feature stores that are in use and contain the additional needed features. In practice, this has been addressed by replicating the features that are needed for the IBM Z use case to a local, IBM Z feature store. This local copy can then be updated as needed. An example architecture is shown in Figure 3 on page 11.

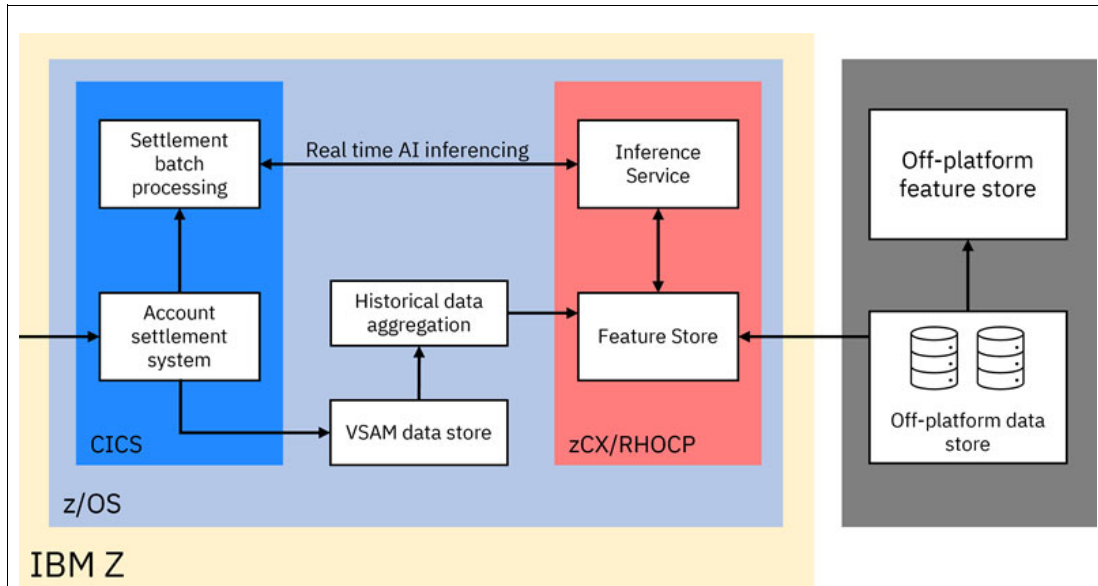


Figure 3 Architecture overview for addressing consistency concerns between the IBM Z and off-platform feature store

In this view, real-time inferencing is being invoked during a financial institution’s clearing and settlement processing for fraud detection. As each batch request is processed, a scoring service is being invoked (“Real-time AI inferencing”). These requests are processed by an inference service that is deployed to Red Hat OpenShift Container Platform hosted in zCX. As shown previously, the inference service gathers additional data from the feature store before processing the inference request.

In Figure 3, the feature store is populated from 2 sources:

- ▶ Short-term (recent) features are sourced and aggregated from local z/OS application data.
- ▶ Longer term historical features are processed in an off-platform data store. In the architecture shown, the off-platform data store is used to populate the off-platform feature store. The specific features that are needed for the IBM Z use case are replicated to the zCX for Red Hat OpenShift based feature store for quick access.

## Summary

Today artificial intelligence presents a huge opportunity to turn data into actionable insights and actions, amplify human capabilities, decrease risk, and increase return-on-assets by achieving breakthrough innovations. To remain competitive, the adoption of AI is not so much a choice for organizations as it is a necessity. By embracing AI, organizations can gain a competitive edge, improve operational efficiencies, enhance customer experiences, and deliver innovative solutions. This IBM Redguide discussed AI infusion into mission-critical applications, different cost-effective entry points for leveraging AI on IBM Z, and data management across platforms for use on IBM Z with Feature store.

Initiating the AI journey on IBM Z and LinuxONE may be challenging and full of uncertainties but there are many steps that can be taken. No matter where you are on your AI journey, you can participate in a Discovery Workshop tailored and hosted by the AI on IBM Z Solutions Team at no charge. In this workshop, get hands-on experience with the IBM team in exploring use cases through design thinking sessions and ideate on possible architecture.

Build out a proof-of-concept with foundational capabilities and develop a practical implementation plan for a clear view of how AI can be leveraged on IBM Z and LinuxONE. For more information, email [aionz@us.ibm.com](mailto:aionz@us.ibm.com).

For teams that are further along in their AI exploration with developed use cases and who are eager to dive into more specifics, there are many options to explore potential collaborations. Run highly optimized, no-charge libraries like TensorFlow, TensorFlow Serving, Snap ML, IBM Z Deep Learning Compiler and Triton Inference Server natively on IBM Z and LinuxONE. The AI Toolkit offers an optional IBM Elite Support while offering IBM Secure Engineering that vets and scans each package for security vulnerabilities that are compliant with industry regulations. Find out more here:

<https://www.ibm.com/products/ai-toolkit-for-z-and-linuxone>

Reach out for any questions. IBM is ready to help you and your organization embarking on this pathway toward AI on IBM Z and LinuxONE.

## Resources for more information

Discovering valuable resources is key to your journey. Stay informed with some reliable sources to find out more. The resources that are listed in this section are considered suitable for a more detailed discussion of the topics that are covered in this book.

Other books and publications:

- ▶ *Demystifying Data with AI on IBM Z*, REDP-5633
- ▶ *Finding an On-ramp to Your AI on IBM Z Journey*, REDP-5723
- ▶ *IBM Cloud Pak for Data on IBM Z*, REDP-5695
- ▶ *Optimized Inferencing and Integration with AI on IBM zSystems: Introduction, Methodology, and Use Cases*, REDP-5661
- ▶ *Turning Data into Insight with Machine Learning for IBM z/OS*, SG24-8552
- ▶ *What AI Can Do for You: Use Cases for AI on IBM Z*, REDP-5679

Explore what IBM industry leaders are blogging about:

- ▶ Accelerating TensorFlow Inference on IBM z16 by Elpida Tzortzatos  
<https://www.ibm.com/blog/accelerating-tensorflow-inference-on-ibm-z16/>
- ▶ IBM accelerates enterprise AI for clients with new capabilities on IBM Z by Elpida Tzortzatos and Meeta Vouk  
<https://www.ibm.com/blog/announcement/ibm-accelerates-enterprise-ai-for-clients-with-new-watsonx-capabilities-on-ibm-z/>
- ▶ IBM Telum Processor: the next-gen microprocessor for IBM Z and IBM LinuxONE by Christian Jacobi and Elpida Tzortzatos  
<https://www.poc.ibm.com/blog/ibm-telum-processor-the-next-gen-microprocessor-for-ibm-z-and-ibm-linuxone/>

Learn more:

- ▶ AI on IBM Z and LinuxONE Technology 101  
<https://ibm.github.io/ai-on-z-101/>
- ▶ AI Toolkit for IBM Z and LinuxONE  
<https://www.ibm.com/products/ai-toolkit-for-z-and-linuxone>
- ▶ Cambrian AI Research Analyst Paper - What If IBM Z Could Help Stop Fraud?  
<https://www.ibm.com/downloads/cas/PRKLA3GV>
- ▶ IBM Global AI Adoption Index 2022  
<https://www.ibm.com/downloads/cas/GVAGA3JP>
- ▶ IBM Z and LinuxONE Community - AI on IBM Z and IBM LinuxONE  
<https://community.ibm.com/community/user/ibmz-and-linuxone/groups/public?CommunityKey=038560b2-e962-4500-b0b5-e3745175a065>
- ▶ IBM Z and LinuxONE Container Registry  
<https://ibm.github.io/ibm-z-oss-hub/main/main.html>
- ▶ Journey to AI on IBM Z and LinuxONE Content Solution  
<https://ibm.biz/AIonIBMzCS>

## Authors

This guide was produced by a team of specialists from around the world working with the IBM Redbooks team.

**Tabari Alexander** is a Senior Technical Staff Member for IBM Z AI and Analytics, where he works on bridging AI acceleration capabilities with the IBM Z software ecosystem. He has a master's degree in Computer Science from Columbia University with a focus in machine learning, and nearly 20 years of development experience within IBM Z. Previous to his involvement in the AI space, Tabari was the Product Owner for PDSE, and brought forth PDSE encryption and zEDC compression of PDSE data sets.

**Colton Cox** is the manager of the AI on IBM Z design team and oversees efforts to drive human-centric practices across a portfolio of AI and data science products. In his 5 years at IBM, he has worked across multiple products, including the z/TPF operating system, the IBM Z Security and Compliance Center, and Machine learning for IBM z/OS. He holds a bachelor's degree in English from SUNY Oneonta and a master's degree in Information Design & Strategy from Northwestern University. His areas of expertise include design leadership, content design, and storytelling in technical domains.

**Joy Deng** is an enterprise product manager for AI on IBM Z based in Raleigh, North Carolina. She has 5 years of experience in tech product management, and before that had experience in market research, strategy, and operations finance across CPG and retail. She holds a bachelor's degree in Marketing and Psychology from Washington University in St. Louis and also a Masters of Business Administration from Duke University Fuqua School of Business with concentrations in Strategy and Tech Management. Her areas of expertise include customer-centered product design, and launching data and AI offerings.

**Purvi Patel** is a senior client engagement leader on the IBM AI on Z Solutions team, where she is leading the development team. She spearheads the initiative to infuse AI into clients' mission-critical workloads. She holds a master's degree in Computer Science from New York Institute of Technology and has over 25 years of experience in the core mainframe technologies. She has a passion for problem solving and possesses an amazing talent for thinking creatively in high-stressed situations and thrives on client interactions. Previously, she was the chief product owner for z/OS Diagnostics Aids, where she designed and implemented many solutions in SVC and Stand-alone dumps. Most recently, she took on the challenge of securing the sensitive personal information in the diagnostics dumps through IBM Data Privacy for Diagnostics product. Purvi holds many patents in the various areas of operating systems and was recognized with the Outstanding Technical Achievement Award by IBM.

**Andrew Sica** is an IBM Senior Technical Staff Member and Chief Architect of the AI on IBM Z and LinuxONE Solutions team. In his nearly 24 years at IBM, Andrew has led several innovative platform initiatives, with areas of expertise including AI infrastructure, operating system development, and platform economics. In Andrew's current role, he works extensively with IBM customers who are interested in leveraging AI to improve their business insights.

**Shin (Kelly) Yang** is a AI on IBM Z Product Manager based in Poughkeepsie, NY. She has 8 years of experience product management and is responsible for the strategy and development of products for the AI on IBM Z organization. She holds a bachelor's degree in Computer Science and an MBA from Clarkson University. Her areas of expertise span from AI technology to business with ACs in Management and Leadership, and Supply Chain Management.

Thanks to the following people for their contributions to this project:

Makenzie Manna and Elias Luna  
**IBM Redbooks, US**

## Now you can become a published author, too!

Here is an opportunity to spotlight your skills, grow your career, and become a published author, all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:  
[ibm.com/redbooks/residencies.html](https://ibm.com/redbooks/residencies.html)



## Stay connected to IBM Redbooks

- ▶ Find us on LinkedIn:  
<http://www.linkedin.com/groups/2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:  
<https://www.redbooks.ibm.com/subscribe>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:  
<http://www.redbooks.ibm.com/rss.html>



# Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US*

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.


## COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

# Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <https://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

CICS®	IBM z16™	z/OS®
IBM®	Redbooks®	z16™
IBM Telum®	Redbooks (logo)  ®	
IBM Z®	WebSphere®	

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Red Hat, OpenShift are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.





REDP-5713-00

ISBN 0738434507

Printed in U.S.A.

Get connected

