

# Finding an On-ramp to Your AI on IBM Z Journey

Stefanie Lauria

Evan Rivera

Andrew Sica



Artificial Intelligence

IBM Z





## The mainframe to power AI in core business workloads is in your data center

### Highlights

- ▶ “Action 1: Who to talk to and how to talk to them”
- ▶ “Action 2: Use case discovery and assessment”
- ▶ “Action 3: Identify differentiating features that can bring additional value”
- ▶ “Action 4: Key communication for an AI project”

With every generation, innovative technology breaks through in a way that fundamentally shifts business and even society at large, and we have already experienced this with the advent of smartphones and the internet. These breakthroughs stem from technological innovation as well as the problems (or use cases) they address.

With the release of ChatGPT in 2022 to consumers, AI has garnered widespread public attention in an unseen way. Even more critically, the direct and conscious use of generative AI capabilities by consumers has become commonplace. For example, ChatGPT reached 100 million active users within 2 months after launch.

However, while consumers are enthralled by generative AI technologies such as ChatGPT, it is essential that enterprises take a more cautious approach, especially considering the use cases in question. The large models powering these capabilities require immense compute resource to run, take a long time to produce a result, and can suffer from a variety of issues, such as hallucination (essentially producing a result that has no relevance to the input or input context). Additionally, explaining the results these models produce is challenging; this is problematic when dealing with highly regulated industries and sensitive use cases where the ability to explain what factors led to a business decision is necessary.

For many core business enterprise workloads, traditional AI provides the needed characteristics to get new insights while still meeting critical application performance goals or Service Level Agreements (SLAs). Traditional AI refers to AI using focused machine learning or deep learning models which have been trained for a specific problem (e.g., predicting fraud).

### Leveraging traditional AI on IBM Z

The AI capabilities on IBM® Z are optimized for use alongside the core business workloads that are so critical to the success of mainframe clients. In this environment, the ability to achieve business insights at scale is critical. This is especially true in the high volume, transactional workloads that IBM Z® clients run where every millisecond matters.

With these core business workloads in mind, IBM designed the Telum processor used by the z16™ and LinuxONE Emperor 4. The IBM z16™ delivered new AI capabilities, including an

on-chip AI accelerator designed to enable the infusion of AI into your most critical business applications while still meeting SLAs. Clients that have purchased an IBM z16 or LinuxONE 4 have these on-chip accelerators available; the accelerator is enabled on chips with active cores and come at no additional charge.

Key features and results include:

- ▶ An industry-first, integrated, on-chip, AI accelerator designed for high-speed, latency-optimized inferencing with sustainability as a core principle.

**DISCLAIMER:** Cited by a third-party industry analyst.

- ▶ With IBM LinuxONE Emperor 4, process up to 300 billion inference requests per day with 1 ms response time using a Credit Card Fraud Detection model.

**DISCLAIMER:** Performance result is extrapolated from IBM internal tests running local inference operations in an IBM LinuxONE Emperor 4 LPAR with 48 cores and 128 GB memory on Ubuntu 20.04 (SMT mode) using a synthetic credit card fraud detection model (<https://github.com/IBM/ai-on-z-fraud-detection>) exploiting the Integrated Accelerator for AI. The benchmark was running with 8 parallel threads each pinned to the first core of a different chip. The `lscpu` command was used to identify the core-chip topology. A batch size of 128 inference operations was used. Results may vary.

- ▶ IBM z16 is designed to score business transactions at scale delivering the capacity to process up to 300B deep learning inference requests per day with 1ms of latency.

**DISCLAIMER:** Performance result is extrapolated from IBM internal tests running local inference operations in an IBM z16 LPAR with 48 IFLs and 128 GB memory on Ubuntu 20.04 (SMT mode) using a synthetic credit card fraud detection model (<https://github.com/IBM/ai-on-z-fraud-detection>) exploiting the Integrated Accelerator for AI. The benchmark was running with 8 parallel threads each pinned to the first core of a different chip. The `lscpu` command was used to identify the core-chip topology. A batch size of 128 inference operations was used. Results were also reproduced using an IBM z/OS® V2R4 LPAR with 24 CPs and 256GB memory on IBM z16. The same credit card fraud detection model was used. The benchmark was executed with a single thread performing inference operations. A batch size of 128 inference operations was used. Results may vary.

- ▶ IBM z16 with z/OS delivers up to 20x lower response time and up to 19x higher throughput when co-locating applications and inferencing requests versus sending the same inferencing requests to a compared x86 cloud server with 60ms average network latency.

**DISCLAIMER:** Performance results based on IBM internal tests using an IBM CICS OLTP credit card workload with in-transaction fraud detection. A synthetic credit card fraud detection model was used: <https://github.com/IBM/ai-on-z-fraud-detection>. On IBM z16, inferencing was done with IBM MLz on IBM zCX. Tensorflow Serving was used on the compared x86 server. A Linux on IBM Z LPAR, located on the same IBM z16, was used to bridge the network connection between the measured IBM z/OS LPAR and the x86 server. Additional network latency was introduced with the Linux "tc-netem" command to simulate a remote cloud environment with 60ms average latency. Measured improvements are due to network latency. Results may vary.

*IBM z16 configuration:* Measurements were run using a z/OS (v2R4) LPAR with MLz (OSCE) and zCX with APAR– oa61559 and APAR - OA62310 applied, 8 CPs, 16 zIIPs, and 8GB of memory. x86 configuration: Tensorflow Serving 2.4 ran on Ubuntu 20.04.3 LTS on 8 Skylake Intel Xeon Gold CPUs @ 2.30 GHz with Hyperthreading turned on, 1.5 TB memory, RAID5 local SSD Storage.

A key additional focus by IBM has been to create capabilities that allow AI services to be easily consumed by business applications; for example, the Machine Learning for z/OS solution allows you to use local shared memory APIs in native COBOL to call model scoring services. Not only does this greatly simplify updates to application code, but it also is a significantly more efficient path that utilizing REST APIs. Additionally, IBM has optimized many popular open-source frameworks and libraries to transparently take advantage of hardware investment such as the on-chip accelerator and the on-core SIMD (vector) processing units.

Another key aspect of IBM's strategy is 'Build and Train anywhere, Deploy on IBM Z.' This approach ensures that your data scientists can build and train their models in their preferred environment, whether that be IBM Z or any other model development environment. When they are ready to deploy on IBM Z, they are able to bring their model and other AI assets to the platform and deploy them for use. Along with functional portability, our approach is to ensure that they will be able to seamlessly leverage the best acceleration targets without having to change their models.

This approach has allowed numerous IBM Z clients to implement use cases on IBM Z; examples include fraud detection in both financial and insurance sectors, clearing and settlement, credit card overlimit risk scoring, insurance claims processing, and many more. See Figure 0-1 for additional examples.

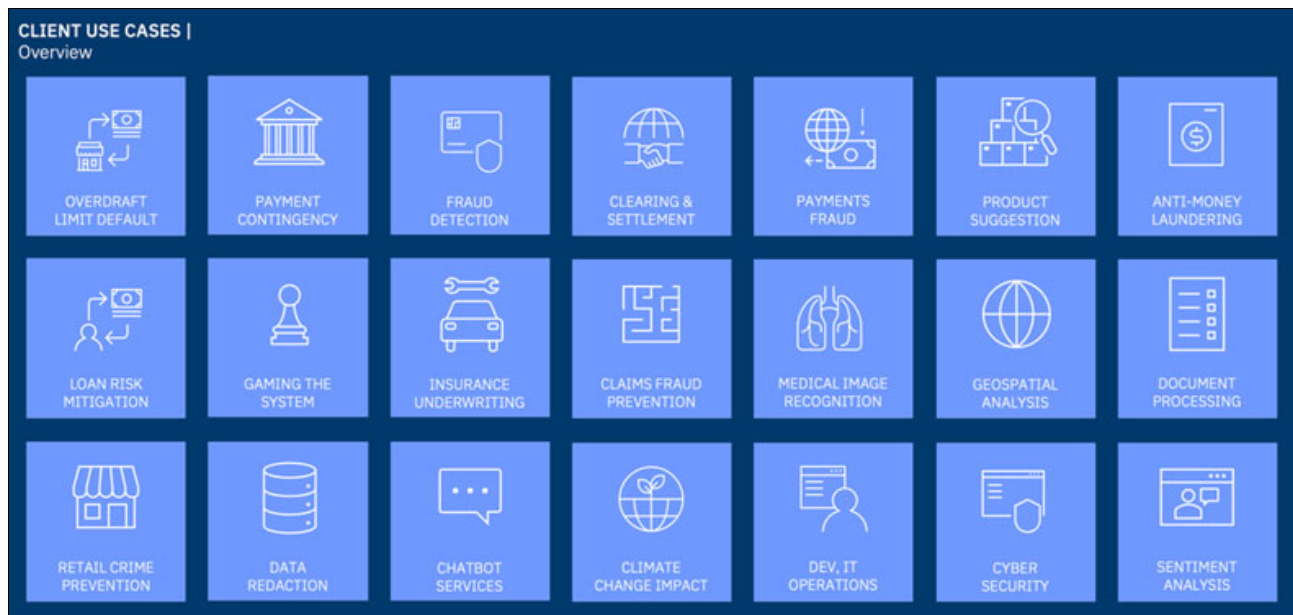


Figure 0-1 Overview of client use cases for AI on IBM Z

Achieving your goals on any project is a challenge - even more so when new or unfamiliar technologies are involved. AI brings its own requirements, software stack and ecosystem. Additionally, new AI specific personas such as data scientists are typically involved to analyze data, model, and create related assets.

Finding the right use case is another critical challenge. There are many considerations when identifying and analyzing use cases; these include:

- ▶ The availability and quality of data.
- ▶ The feasibility of using AI to solve the problem.
- ▶ The SLA requirements that must be achieved.
- ▶ The potential Return on Investment (ROI).
- ▶ The risk associated with the use case, i.e., What would a wrong decision cost? What regulations are in place to govern the use of AI?

AI on IBM Z is designed to help you leverage AI in your most critical workloads and with qualities of service unachievable anywhere else. This publication presents a framework to help you in getting jump started with your AI projects on IBM Z; we will touch on many of the challenges presented above and provide you with resources to get jump started.

This publication is written for those who play a strategic role within an organization; they hold senior positions and impact project decisions. This publication is also useful for consultants and IT architects.

## The framework

In the next several sections, we will detail how your enterprise can be successful with an AI workload. We will guide you by helping you identify the right stakeholders needed, and how to identify potential use cases that are best suited for AI. Also, outlined here are tools on how to leverage IBM Z architecture to run AI models effectively.

### Action 1: Who to talk to and how to talk to them

Successful deployment of an AI application in production starts with involving the right stakeholders. AI capabilities need to be asserted with the right use cases and address the needs of your users and business. One of these key stakeholders is someone that represents the line of business. This person oversees the business operation, such as an officer in the bank critically analyzing impact to the balance sheet. They will make sure that business priorities and the AI solution align, and SLA requirements are met. They also have a stake in making sure there is a return on investment when leveraging AI on Z.

With the AI on IBM Z use cases, the team will have to rally to build an AI model to solve a business need. A data scientist will be a critical stakeholder to not only help construct this model, but also to the entire lifecycle management of the project. A data scientist will help manage the data needed to create an AI model. They are trained to review the needs of the business and align the features of data to train and build an AI model that accomplishes the goals of the use case. The AI on IBM Z architecture is very flexible and will allow your data science team to work with well-known industry tools of their choice and still deploy the model closest to your critical workloads on Z.

The team will also need to include application architects and developers. These individuals are well versed with the applications that run on Z, including the middleware that process your enterprise's transactions. These transaction-based applications can, in most use cases, leverage AI to rate the transactions to meet a business need. So, it is important to involve them from proof of concept to production.

Certainly, you should also be involving the IBM Z infrastructure team, from system programmers to system administrators to database administrators to security administrators and architects, to name a few. The infrastructure team may need to set up the AI scoring/inference environment required, depending on the reference architecture selected and will directly interact with the mainframe and any workloads running on IBM Z and help the data science team deploy AI models on the platform. They will also work with the application development team to ensure that the AI model deployed is rating the business transactions and will work with the line of business to ensure that the scoring of the transactions meet SLAs.

Creating the right culture to ensure collaboration with this cross functional team is paramount. This requires IBM Z infrastructure teams to approach new colleagues and advocate for the benefits of running AI models next to IBM CICS® or IMS transactions. Some benefits include latency and throughput, and better consumption which helps enterprises meet stringent SLA requirements. There are many reasons why investing resources into a proof of concept would be beneficial for the business. For example, most of the data needed to create the AI model already resides within Z. This will allow the model to rate transactions in real time. To keep the engagement with the rest of the team ensure that you have benchmarks and metrics and that can be achieved throughout the lifecycle of the project. Notwithstanding the team must also have buy-in from the decision-makers and leaders in the enterprise such as CIOs, CDOs (Chief Data Officer), and CTOs.

## **Action 2: Use case discovery and assessment**

Once the team comes together, it is important to assess the applications and workloads running on Z, have a clear understanding of the business challenge, and internally determine how to leverage the z16 and Telum chip to improve latency, throughput, and ROI.

### **Discovery - What applications and workloads run on the mainframe?**

This involves taking inventory of the applications and workloads that currently run on the mainframe. For example, most clients have CICS or IMS applications processing numerous types of transactions. These applications are potential opportunities for AI to deliver new business value.

Also, be cognizant of where your data resides for these applications. Perhaps, the data is being stored in IBM Db2® or IMS, but in most client environments data comes from other distributed sources. Therefore, working with the database administrator is going to be essential so that you can gather all the data points. Additionally, consider the dependencies of these applications, such as identifying real time or batch processing requirements.

### **Discovery - What are the business challenges?**

Having a clear and agreed upon problem statement is critical to the success of any AI project. AI has powerful capabilities and implications, and it needs to be clear that AI is the appropriate and ethical solution to the business problem.

AI is well suited for certain types of problems such as giving recommendations based on past behavior, anticipate and preempt disruption, detect liability and mitigate risk, comb through complex topics to help in research and discovery, collecting large amounts of knowledge and distilling it at scale, aiding in the personalization of experiences through natural language.<sup>1</sup>

### **Recommendations**

Based on the historical data and patterns of user behavior, AI is highly effective in making targeted recommendations confidently. This can be used for marketing campaigns and product recommendations.<sup>1</sup>

### **Anticipating and preempting disruption**

AI can monitor your systems and analyze when potential issues are likely to arise. It can even mitigate issues for cost cutting measures. For example, AI can very well identify future peak consumptions in the system and either create new or free-up existing resources.<sup>1</sup>

### **Detect liability and mitigate risk**

This type of challenge is best fit for the mainframe and the workloads running on Z. AI can rate risk with incredible precision, giving the enterprise confidence that risk and losses can be reduced. As previously mentioned in this paper, AI can help with detect fraud and prevent money laundering schemes.<sup>1</sup>

### **Research and discovery**

Leverage AI to sift through copious amounts of data and information, get to insights faster, and hone into domain-specific areas of interest. This can be used to help system administrators analyze the system and get the analytics necessary to inspect it.<sup>1</sup>

### **Distilling knowledge at scale**

AI can easily collect enterprise-wide domain knowledge and combine it with industry wide expertise to create a sole source for insights, helping employees with various levels of education gain understanding about a subject in a self-service manner. This can be in a form of conversational AI that help junior staff attain answers to questions that only very senior staff may know.<sup>1</sup>

---

<sup>1</sup> <https://www.ibm.com/design/thinking/page/toolkit/activity/ai-essentials-intent>

### **Personalization of experiences**

Using AI, the enterprise can collect historical data to make more targeted recommendations to the end user that can tailor their experience with your product. The most common example of this is Netflix recommendations based on what movies and shows the user has watched in the past.<sup>2</sup>

### **Discovery - Which infrastructure can best support AI capabilities and workloads?**

Once the team identifies the key business problems that AI can solve, the next step is to establish which platform is best suited to solve these challenges. Leveraging AI on the mainframe is part of an overall infrastructure strategy that aligns with other distributed architectures. Data scientists can train the model anywhere they like using the tooling of their choice, and then deploy it on IBM Z to get the benefits of inferencing alongside existing IBM Z workloads. Running AI in this fashion will ensure that SLA requirements are met in addition to the inherent security benefits of being on-prem.

The IBM Z infrastructure is flexible and allows a data scientist to use industry standard tools such as TensorFlow, PyTorch, and Huggingface, to name a few, to create an AI model. The data scientists then convert the AI model to the ONNX format and deploy on the platform to get the best inferencing and throughput to meet the business' SLA requirements. If the enterprise uses z/OS as their environment, then deploying a machine learning or deep learning model on IBM Z can be done seamlessly with Machine Learning for IBM z/OS (MLz). For more information on Machine Learning for IBM z/OS, see:

<https://www.ibm.com/products/machine-learning-for-zos>

IBM Z is ideal for solving complex transactional problems and deploying AI models. Since many use cases in the discovery process rely on IBM Z workloads, such as CICS, IMS, and Db2, and much of the data resides on the IBM Z platform, transactional applications can score every transaction in real time.

### **Assessment - ROI**

Data science, IBM Z infrastructure, and application teams collaborating can achieve the goals of their use case and solve complex problems for their enterprise. For the lines of business there are real benefits in running AI on Z, as SLA requirements can be met with the power of the IBM Z infrastructure. For some of the use cases there are actual cost saving and revenue growth opportunities. Not to mention increase customer satisfaction with quicker, more precise, and objective decisions which increase brand loyalty.

### **Action 3: Identify differentiating features that can bring additional value**

Once the business challenge is understood and the use case is discovered, identifying differentiating features on IBM Z will maximize additional value and ROI.

#### **Differentiating feature - Transactional AI for real-time scoring**

Many AI use cases require the need for the model to respond in real-time to make a positive business impact. For example, the difference between fraud detection and fraud prevention is the ability to address the fraud in real-time. IBM z16 delivered the IBM Z Integrated Accelerator for AI which is an industry-first, integrated, on-chip, AI accelerator designed for high-speed, latency-optimized inferencing. This AI accelerator allows for transactional AI that positions clients to meet stringent service-level agreements (SLA).

#### **Differentiating feature - Optimized open-source AI frameworks**

Open-source software is foundational to the AI industry and AI on IBM Z is no exception. A key focus by IBM has been to create capabilities that allow AI services to be easily consumed by business applications. IBM has optimized many popular open-source frameworks and libraries to transparently take advantage of hardware investment such as the on-chip accelerator and the on-core SIMD (vector) processing units. A wide range of industry standard AI frameworks are available on IBM Z, such as TensorFlow, XGBoost, scikit-learn, and many more.

<sup>2</sup> <https://www.ibm.com/design/thinking/page/toolkit/activity/ai-essentials-intent>



### **Differentiating feature - Build and train anywhere and deploy on IBM Z**

The ability for clients to leverage their existing AI investments can be beneficial for the success of an AI project. Enabling the AI team to build and train their models anywhere ensures the preferred environment, including IBM Z, can be used. When they are ready to deploy on IBM Z, they can bring their model and other AI assets to the platform and deploy them for use. Along with functional portability, our approach is to ensure that they will be able to seamlessly leverage the best acceleration targets without having to change their models.

### **Differentiating feature - Core business applications and data reside on IBM Z**

AI projects require data to be successful. IBM Z is an excellent candidate for AI workloads because it is where the data resides. In addition to the data, IBM Z is where the core business applications run. Deploying AI workloads next to the data and alongside the core business applications are beneficial from a real-time perspective as well as security and privacy because the data remains within the same environment.

A key focus by IBM has been to create capabilities that allow AI services to be easily consumed by business applications; for example, the Machine Learning for z/OS solution allows you to leverage local shared memory APIs in native COBOL to call model scoring services. Not only does this simplify updates to application code, but it also is a significantly more efficient path than utilizing REST APIs.

### **AI solution templates on IBM Z and LinuxONE**

AI solution templates allow clients to accelerate their on-ramp to their AI journey on IBM Z and LinuxONE. A suite of AI solution templates that span various industries, use cases, and capabilities can be leveraged based on the discovered use case.

AI solution templates provide hands on experience on IBM Z across the full AI lifecycle. With sample open-source datasets, AI model training capabilities are provided to build the AI model. Once the AI model is built, guidance on deploying the model to available AI model deployment frameworks, such as MLz, can be used. With the model deployed, sample business application, such as CICS-COBOL or web-based applications, can be utilized to start putting the AI model to work. Lastly, sample applications for model analysis are available to have visibility into the decision the model is making and why.

Based on the specific industry, different AI Solution Templates provide an on-ramp to jump starting your AI journey on IBM Z. Some industries include finance, insurance, and health care. Within these industries, there is many different AI use cases such as credit card fraud detection, clearing and settlement, health insurance claims, and more. Based on the business problem that is being addressed, specific AI Solution Templates can assist in solving it.

### **Action 4: Key communication for an AI project**

Running a project involving AI shares many of the same considerations of any technical project. However, there are some additional factors that should be considered to ensure the project is successful.

In our work with clients on AI projects, we have found the first and most crucial factor to a successful project is communication across the stakeholders. While this may seem obvious, it is a quite common failure point. A key reason for this is that frequently the early (pre-deployment) AI project stages involve specialized personas like data scientists that tend to work in isolation from infrastructure and application architects; this is especially true of mainframe personas who often have not interacted with their enterprises data science teams in the past.

This isolation between key project personas at early project stages can often lead to major rework at deployment time. As data scientists create AI assets like data pipelines and models for potential production use, they are creating relationships and requirements that will need to be handled in the production environment. A couple of common examples of this to consider:

- ▶ The data scientist may be creating data preprocessing pipelines or models that are too compute intensive given the SLA requirements of the workload.
  - Data preprocessing is used to transform raw data into the format needed for model execution. These transformations are an often-overlooked complexity: they are typically implemented in python which can add overhead in real-time, and they are often difficult to optimize in python. There are techniques that can be used to create optimized or optimizable pipelines - however, this is best considered at an early stage.
  - AI model selection for real-time use cases can be a similarly fraught exercise. Ideally, the most accurate model for a problem would be selected and used; in practice, throughput and latency requirements play a key factor. If an inference request takes too long, it will be abandoned, and the transaction will complete without the benefit of AI. There are a variety of strategies that can be used in these circumstances; however, these are best considered early in the process before a substantial data science effort is spent on a model that cannot be deployed for a problem.
    - There is some good news in these circumstances for mainframe clients: the IBM z16 Integrated Accelerator for AI often enables the use of more complex models that provide better accuracy while still meeting application SLA requirements.
- ▶ Another common example is on introducing new production data requirements.
  - While in the initial stages of data analysis, engineering, and model creation, a data scientist is often working with a variety of historical data. They may develop models that require highly engineered characteristics, including aggregated data fields and historical data sources. They may also use data that originates or is stored in other sources outside of the core business application.
    - These data architectures can introduce additional latency and complexity... However, there are a variety of well-known techniques to implement them in production architectures. Again, the most critical point is that these architectures must be planned for - ideally in the early project stages, to avoid substantial delays later.

There are numerous other examples of potential roadblocks and stumbling points. Figure 0-2 on page 9 is an example AI project flow - at each stage of the project, make sure that the communication lines are open, and application, data and infrastructure personas are engaged in the planning discussion.

As we discuss in the next section, IBM Client Engineering for Systems has a detailed workshop where we can help you avoid these common pitfalls.

01	02	03	04	05
<b>Define</b> <i>your project goals</i>	<b>Prepare</b> <i>the data</i>	<b>Choose</b> <i>a tool</i>	<b>Train</b> <i>your model</i>	<b>Deploy</b> <i>your model</i>
What do you want to find out?  Do you have the data to analyze?	Refine the data  Add the data as a project asset or in a data repository	Pick the tool that matches your data and desired outcome  Choose between an automated process, a graphical editor or code your own model	Train the model with the data you supply  Let a model building tool choose estimators and optimize or choose your own	Score the model to generate predictions  Make your model available in production  Retrain as needed
<b>Key Persona(s):</b> - Line of Business - Data Scientist  <b>Stakeholder:</b> - IT Architect - Application Architect  <b>Influencer:</b> - Data Engineer	<b>Key Persona(s):</b> - Data Engineer  <b>Stakeholder:</b> - Data Owners - Data Scientist  <b>Influencer:</b> - Line of Business - Data Wrangler	<b>Key Persona(s):</b> - Data Scientist  <b>Stakeholder:</b> - IT Architect - Application Architect  <b>Influencer:</b> - Line of Business - Data Engineer	<b>Key Persona(s):</b> - Data scientist  <b>Stakeholder:</b> - IT Architect - Application Architect  <b>Influencer:</b> - Data science manager	<b>Key Persona(s):</b> - IT Architect - Application Architect  <b>Stakeholder:</b> - Data Scientist - ML Engineer  <b>Influencer:</b> - Line of Business

Figure 0-2 Example AI project flow

## What’s next: How IBM can help

We have covered a wide range of topics thus far at a very high level. As you look towards your first (or next!) AI project, the details and decisions can still be daunting. However, you do not have to worry, because IBM has resources available to not only help you better understand AI technologies, but also help jump start your project. We will touch on each of these.

### No-charge workshops run by IBM Client Engineering for Systems

An unbelievably valuable resource that IBM has made available are no-charge discovery workshops. These workshops are intended to both enable you on IBM Z and LinuxONE AI capabilities and help you in getting jump started on an AI project.

The workshop can help across a range of potential areas, including:

- ▶ Those who are interested in leveraging AI and analytic capabilities to gain new insights from their IBM Z based workloads.
- ▶ Those interested in understanding how IBM Z can enable them to score every transaction, so no opportunity is missed due to latency.
- ▶ Those interested in seeing how IBM Z enables scoring (or inference) using industry standard open-source frameworks directly on platform which allows for minimal impact to transactional workload SLAs while still supporting all the qualities of service of IBM Z.

During the workshop we will work together with your team to ideate on use cases that will bring the return on investment that meets your organizations expectations while meeting your SLAs. Coming out of the workshop we will achieve common goals, such as giving you a clear understanding of the technology requirements to run AI on IBM Z or LinuxONE and how the capability of the infrastructure is ready to support your use cases. You will come

out having named, defined, and prioritized the most advantageous use case for your business. In addition, we will help you scope an MVP and give you a reference architecture so that together we will deliver a proof of concept (POC) that successfully achieves the agreed upon use case.

To deliver on the promise of the POC we must have the right personas engaged from the onset, as we have outlined earlier in this document. As such, for the workshop IBM typically suggests the following roles or personas if available, although not require having the workshop:

- ▶ Line of business
- ▶ Data scientist
- ▶ Application architect
- ▶ Data architects
- ▶ Infrastructure Architect

## Summary

Having the right people come together to align on the best use case for your organization will deliver the best ROI and ensure that SLA requirements are met. Given all the resources provided here, you and your organization can get started on creating AI models and deploying them on IBM Z to achieve scoring in real time. Should you need additional assistance or if your organization would like a free discovery workshop, contact [aionz@us.ibm.com](mailto:aionz@us.ibm.com) or [ce4s@ibm.com](mailto:ce4s@ibm.com).

## Resources for more information

For more information that can help you get started with AI on IBM Z in a self-service manner, see the following community resources developed by our team (include links if applicable):

- ▶ AI on IBM Z 101 GitHub page - Contains a plethora of resources that will help you learn more about AI and the about the capabilities of AI on Z:  
<https://ibm.github.io/ai-on-z-101/>
- ▶ Journey to AI on IBM Z and LinuxONE Content Solution page - Here we offer added guidance on how to get started on AI on IBM Z solutions and provide you with the technical resources needed to get your AI models deployed in production:  
<https://www.ibm.com/support/z-content-solutions/journey-to-ai-on-z/>

# 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 Z®	z/OS®
Db2®	IBM z16™	z16™
IBM®	Redbooks (logo)  ®	

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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.

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





REDP-5723-00

ISBN DocISBN

Printed in U.S.A.

Get connected

