Infuse AI into IBM Z Applications

Yichong Yu, STSM, IBM Cloud for Financial Services
Surya V Duggirala, IBM Distinguished Engineer and CTO, Enterprise Cloud Architecture

Businesses across the globe are increasingly relying on cloud computing to modernize their current operations. Majority of these businesses across the industries like Banking, Insurance, Healthcare, Retail, Government etc., run their core transactional and batch workloads on IBM Z as it provides unmatched qualities of service like security, reliability, and availability. To combine the core strengths of IBM Z platform and IBM Cloud and to keep the applications and data secure and compliant, IBM recommends a hybrid cloud approach to mainframe application modernization.

Artificial Intelligence (AI) is being infused in every aspect of the business processes to make them intelligent and efficient. AI is pervasive not only across the business processes, but it is infused across all layers of hybrid cloud, right from the processor chips, cloud fabric, cloud services and even in the cloud operations. According to PwC research, global GDP could be up to 14% higher in 2030 as a result of AI – the equivalent of an additional $15.7 trillion – making it the biggest commercial opportunity in today's fast changing economy.

AI in hybrid cloud

Mainframe computers play a central role in the daily operations of most of the world's largest corporations, including top worldwide banks, many of world's largest insurers, airlines & retailers. Mainframes process billions of banking transactions per day including many billions of ATM transactions per day.

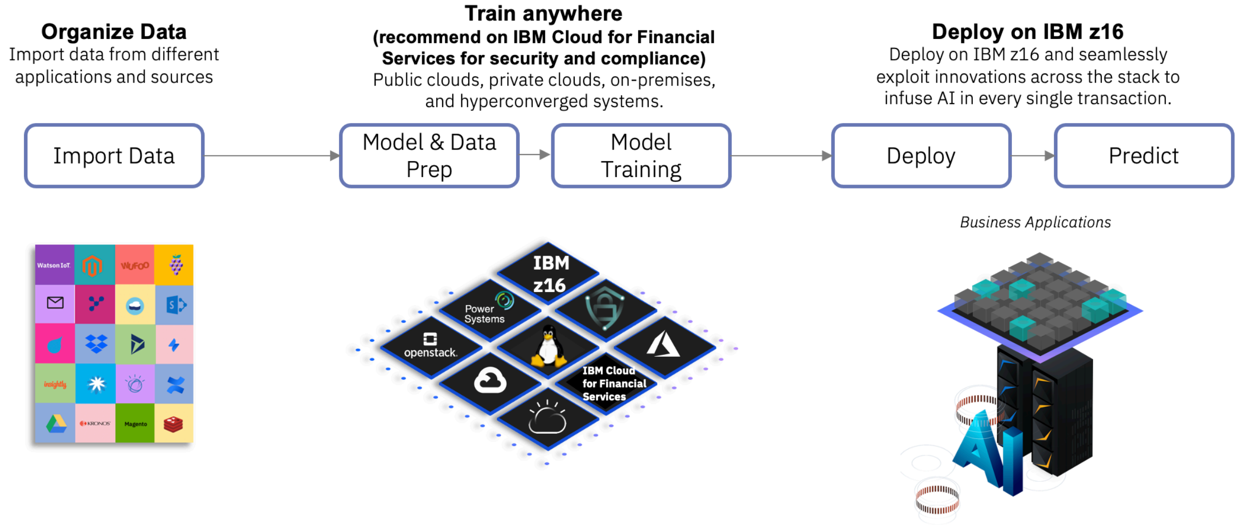
Companies need to have intelligence built into the transaction processing logic to handle instant payments, increasingly stricter regulations that require real time Anti-Money Laundering (AML) and increases in fraud. Implementing AI at the transaction level without impacting SLAs is now more critical than ever. Even for clearing and settlement use cases that are processed in batch workloads the volume of transactions is increasing and the time windows are shrinking as the industry is moving to same day settlements. Throughput and latency have become critical for both real-time and batch processing.

Companies have been struggling with the performance of invoking AI model inferencing over the network. For example, one of our clients, a large US based Bank, was attempting to integrate real-time fraud detection into their credit card processing application running on the IBM Z platform.

Originally, they had built a machine learning model and deployed it on separate platform from the transaction processing. However, this was not scalable, as invoking the scoring engine over network resulted in inconsistent response times, ranging from 50-100 milliseconds(ms). This resulted in only less than 10% of the credit card transactions being able to acquire a risk score from the model due to the time sensitive nature of the process. To solve this gap, client took their machine learning model and deployed it directly, co-locating with their z/OS based credit processing application on IBM Z. This has reduced their latency from over 50 ms per call to under 2 ms and enabled them to score 100% of all transaction and save over $20M per year in reduced risk exposure.

Innovations like the IBM Telum processor chip with the IBM Z Integrated Accelerator for AI are positioning IBM as the premier platform provider for inference workloads. Client can leverage the incredible speed to meet mission critical SLAs with IBM Z. Meanwhile, cloud provides flexibility and scalability, enables effective collaboration, and supports flexible pay-as-you-go pricing model. Data scientists can access the various tools and data sources available in cloud to develop, train, and test the AI models with the framework they are familiar with on cloud platform.

Financial data are highly sensitive. Data scientists need secure cloud infrastructure to protect the data and training jobs, and meet security and compliance requirements, that is where the IBM Cloud for Financial Services can help. IBM Cloud Framework for Financial Services is designed to help address the needs of financial services institutions with regulatory compliance, security, and resiliency requirements. The Financial Services Cloud framework defines a comprehensive set of control requirements and provides automation and configuration of proven reference architectures. Cloud services or ecosystem partner services can evidence compliance to the controls and become IBM Cloud for Financial Services Validated. The Financial Services Validated designation signifies that you have successfully evidenced compliance to the control requirements of the IBM Cloud Framework for Financial Services. IBM Cloud Security and Compliance Center enables continuous compliance and protect customer and application data. Financial institutions can confidently host their mission-critical applications on IBM cloud and transact quickly and efficiently.



Organize Data
Import data from different applications and sources

Train anywhere
(recommend on IBM Cloud for Financial Services for security and compliance)
Public clouds, private clouds, on-premises, and hyperconverged systems.

Deploy on IBM z16
Deploy on IBM z16 and seamlessly exploit innovations across the stack to infuse AI in every single transaction.

Import Data → Model & Data Prep → Model Training → Deploy → Predict

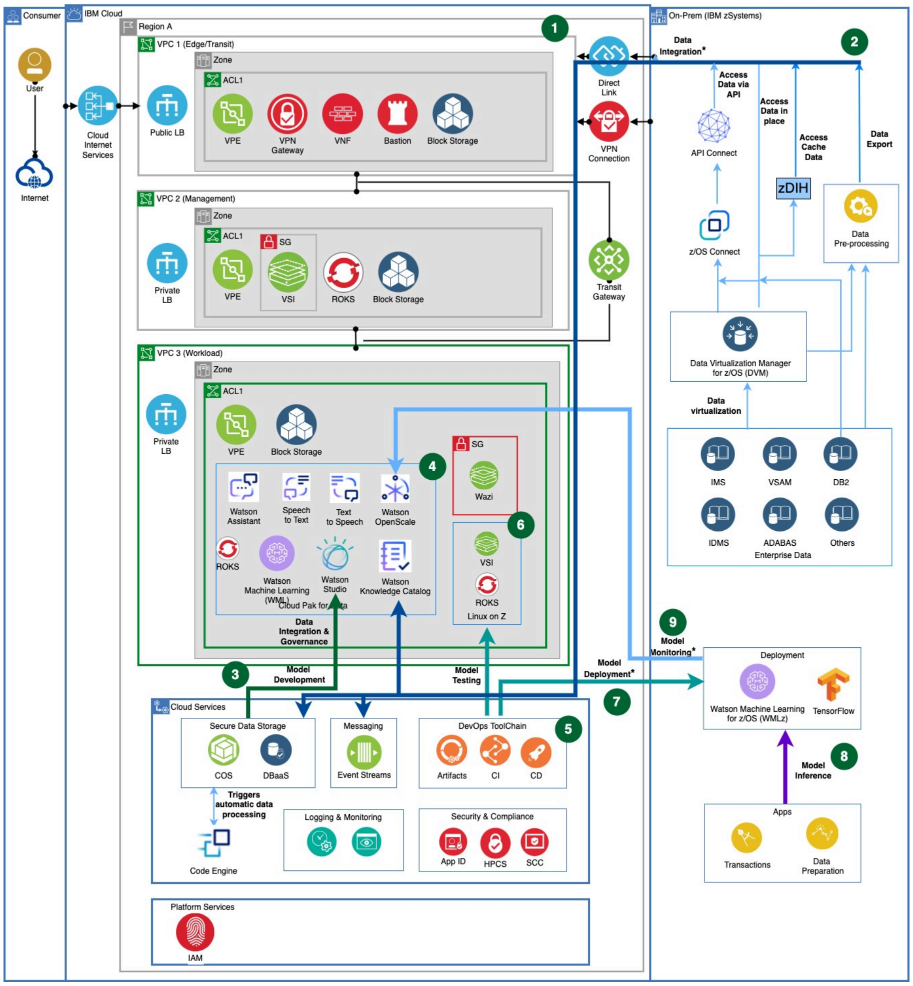
Business Applications

AI reference architecture in hybrid cloud

With advances in hybrid cloud that make modernization less risky, it’s an essential platform for any digital transformation. This accelerates decision velocity and provides the agility to move your business forward.

Financial data are highly sensitive and needs to be protected, and AI workloads also need to be deployed in cloud environment that meets regulatory compliance, security, and resiliency requirements. IBM Cloud for Financial Services is designed to help clients mitigate risk and accelerate cloud adoption for even their most sensitive workloads.

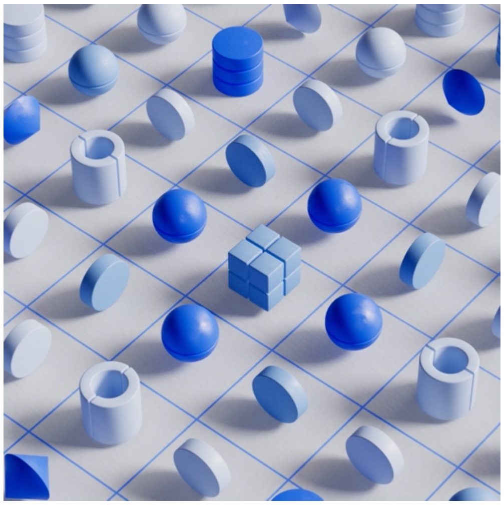
AI model building is an iterative process. It normally starts with understanding the business problem at hand, and iterate through understanding and preparation of data, training and testing the models, model deployment, model inference, and model monitoring.



The IBM Cloud for Financial Services provides different flavors of the reference architectures. Here is the AI pipeline in VPC reference architecture in hybrid cloud.

1. Set up the infrastructure by creating DevOps toolchain for "Deploy infrastructure as Code for the IBM Cloud for Financial Services", which provides a secure environment for model training jobs. For more information, refer to reference architecture for IBM Cloud for Financial Services.
2. Make the enterprise data on-prem accessible in cloud. There are different ways to expose the data on-prem. Please refer to data integration blog for details.
3. Understand, prepare, and manage the governance of the data with tools in IBM Cloud Pak for Data, for example, organizing and managing data with Watson Knowledge Catalog.
4. Training jobs can be run in the workload VPC. Various tools in IBM Cloud Pak for Data can be used. Watson Studio can be used as the model development environment. Different machine learning platforms and tools can be used to train the model (example, TensorFlow, pyTorch, SnapML, etc.). Model can be tested in Watson Machine Learning. For conversational AI, tools like Watson Assistant, Speech-to-Text, and Text-to-Speech can be used.
5. IBM DevOps Toolchains can be used to manage the model source codes and tested models. Refer to the DevOps blog for more details.
6. Linux on IBM Z or IBM Wazi as a Service can be used to test the AI models and AI applications before pushing them to on-premises IBM Z.
7. Tested AI models are deployed on IBM Watson Machine Learning for z/OS.
8. AI model inference is invoked to gain insights from the data.
9. AI models are continuously monitored via IBM Watson OpenScale, which is an enterprise-grade environment for AI applications that provides enterprise visibility into how your AI is built and used, and delivers return on investment. Its open platform enables businesses to operate and automate AI at scale with transparent, explainable outcomes that are free from harmful bias and drift.

AI model data preparation in hybrid cloud

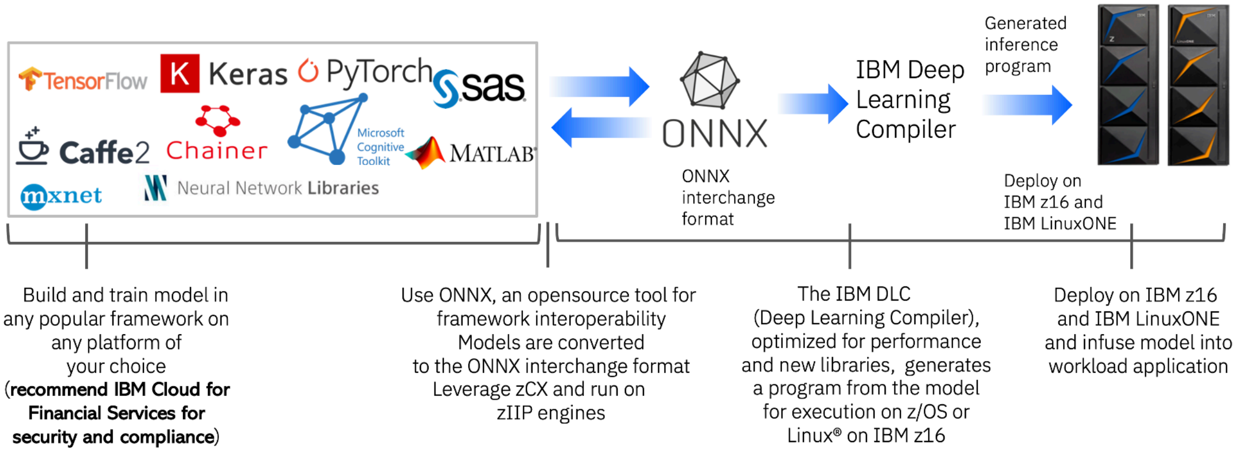


Most transaction data are on IBM Z. Companies could adopt different data integration strategies to make the data accessible to model training in cloud. For relatively small amount of data, data can be made available via APIs. To enable users and applications read/write access to IBM Z data in place, without having to move, replicate or transform the data, data virtualization strategy can be adopted. IBM Data Virtualization Manager (DVM) for z/OS provides virtual, integrated views of data residing on IBM Z. For one-time big volume historical data extraction, data pre-processing can be done on-prem to filter and mask sensitive data. Refer to the data integration blog for different data integration strategies.

Tools from IBM Cloud Pak for Data can be used to integrate and manage data from various sources in cloud. IBM Watson Studio provides different types of connectors to connect to various data sources. Watson Knowledge Catalog provides the data governance and privacy capabilities of the data fabric architecture. The data refinery tool, available with IBM Watson Studio and IBM Watson Knowledge Catalog, saves data preparation time by quickly transforming large amounts of raw data into consumable, high-quality information that's ready for analytics. A knowledge core can be developed by curating data assets and enriching them with governance artifacts that describe their properties and meaning. Business vocabulary can be used to further enrich the data assets. Data in the catalog can be consumed by data scientists for model building purpose.

AI model development in IBM Cloud

There are different machine learning (ML) frameworks to train AI models, and different model serving frameworks to host the AI models. To allow data scientists flexibility when choosing the ML framework and to allow model hosting framework to be independent of the model training framework, AI models can be converted to Open Neural Network Exchange (ONNX) format. ONNX defines a common set of operators - the building blocks of machine learning and deep learning models - and a common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers.


TensorFlow
Keras
PyTorch
sas
Caffe2
Chainer
Microsoft Cognitive Toolkit
MATLAB
mxnet
Neural Network Libraries
ONNX
ONNX interchange format
IBM Deep Learning Compiler
Generated inference program
Deploy on IBM z16 and IBM LinuxONE
Build and train model in any popular framework on any platform of your choice (recommend IBM Cloud for Financial Services for security and compliance)
Use ONNX, an opensource tool for framework interoperability Models are converted to the ONNX interchange format Leverage zCX and run on zIIP engines
The IBM DLC (Deep Learning Compiler), optimized for performance and new libraries, generates a program from the model for execution on z/OS or Linux® on IBM z16
Deploy on IBM z16 and IBM LinuxONE and infuse model into workload application

These are the typical steps to build and persist the model in IBM Cloud:

- **Project setup** – Various tools in IBM Cloud Pak for Data can be used for AI model training in IBM Cloud. One can start with creating a Watson Studio project, which is the logical organization of resources and can be used to manage access to the resources.
- **Data connection** – Watson Studio supports data connectors to various data sources. One can access data from various sources securely in IBM Cloud.
- **Model training** – Data scientists can use the framework of their choice and build AI models in cloud to get the benefits of cloud like collaboration, flexibility, and scalability. IBM Cloud for Financial Services provides the secure and compliant environment for sensitive workloads and allows for continuous monitoring.
- **Model persistence** – Trained model is persisted, so that it can be brought to the deployment environment for model inference. Hyper Protect Crypto Services provides the highest level of protection for data at rest in IBM Cloud.
- **Model conversion** – Trained and tested models can be converted to Open Neural Network Exchange (ONNX) format. By converting the model into ONNX format, the model can be hosted in a framework that is independent of the model training framework, so that data scientists have more flexibility to choose the framework they are most familiar with to train the model.

AI model testing in IBM Cloud

AI models can be tested in IBM Cloud before being deployed to on-premises IBM Z Platform.

Linux on IBM Z is the collective term for the Linux operating system compiled to run on IBM Z Systems, especially System Z machines. Linux on IBM Z supports containers and wide range of tools on Linux. AI models can be tested on Linux on Z virtual servers in IBM Cloud.

IBM Wazi as a Service provides cloud native development and testing environment for z/OS on IBM Cloud. It enables clients to have self-serve access to z/OS systems and shorten their development and testing cycles. It can be further integrated into a secure toolchain for continuous delivery. Clients can set up their own custom z/OS environment within minutes. Applications developed for z/OS can be tested on Wazi before being deployed to on-prem. Please check with Wazi documentation regarding the support on AI models.

AI model deployment to IBM Z

There are different ways to deploy AI models on IBM Z, for example:

- Deploying with IBM Watson Machine Learning for z/OS (WMLz)
- Deploying with WMLz Online Scoring Community Edition (OSCE)
- Deploying with TensorFlow Serving
- IBM zDNN plugin for TensorFlow

IBM Watson Machine Learning for z/OS (WMLz) is an enterprise-grade and production-ready platform that enables embedding ML and DL models into transactional applications for real-time insights. With WMLz, organizations can build, deploy, and run models on IBM Z and leverage several essential enterprise-grade ML features, such as model versioning, auditing, and monitoring. Models trained on other platforms can be converted to PMML or ONNX format and deployed on WMLz. Refer to IBM Watson Machine Learning for z/OS for details.

The WMLz Online Scoring Community Edition (OSCE) is a special no-charge version of WMLz that is intended for simple, non-production testing of the real-time scoring function of pretrained ONNX models. WMLz OSCE can be used for rapid use case evaluation of embedding DL models in transactional z/OS applications while leveraging the Integrated Accelerator for AI. WMLz OSCE is packaged as an s390x Docker container image that is easily deployed in a zCX.

TensorFlow Serving is an open-source, high-performance deployment option that is a good fit for enterprises that are heavily invested in the TensorFlow ecosystem or have complex model pipelines. TensorFlow Serving is available as a container image in the IBM Z Container Image Registry, and it can be used in a zCX or a Linux on IBM Z environment. Any z/OS application can access the TensorFlow model by using a REST API call.

IBM zDNN Plugin for TensorFlow can also be used to deploy TensorFlow model and utilize the IBM Integrated Accelerator on Z. IBM-zDNN-Plugin will detect the operations in your model that are supported by the Integrated Accelerator for AI and transparently target them to the device.

AI model inference on IBM Z

Just like model training, data needs to be preprocessed and transformed to the right format before invoking the model. The data transformations should be consistent with the ones that are done at model training time. Data transformation application can be collocated with the model inference on IBM Z to get better performance.

WMLz delivers predictive analytics capabilities to the platform and enables the generation of real-time insights at the source of a transaction. Models can be imported into WMLz with REST APIs or directly to a z/OS CICS region, and model inference can be invoked via REST APIs or CICS calls correspondingly. There is also a new capability available now to invoke WMLz scoring service **for real-time inferencing in IMS COBOL applications** through the WebSphere Optimized Local Adapter (WOLA) interface. This enhancement is also applicable for ***Batch COBOL*** and ***CICS COBOL*** applications.

AI model monitoring

Enterprises need to monitor the AI models for bias, fairness, and trust with added transparency on how the AI models make decisions.

IBM Watson OpenScale is an open, enterprise-grade platform that enables businesses to build, operate, and manage production AI. IBM Watson OpenScale detects and mitigates bias and drift, increases the quality and accuracy of your predictions, and explains transactions and perform what-if analysis.

IBM Watson OpenScale can be set up on Linux on Z or in cloud to monitor the AI models.

Conclusion

To protect clients' investments and help accelerate their digital transformation journey, IBM Cloud provides a set of solution patterns that will help remove the inhibitors and modernize IBM Z workloads. The AI solution patterns described here focuses on AI model development and deployment in hybrid cloud. The IBM toolchain deploys IBM Cloud for Financial Services infrastructure as code and can be used as the secure training environment in cloud. IBM Integrated Accelerator on Z16 enables real-time low-latency model inference at scale and makes in-transaction model inference feasible. Combining the power of IBM Z and IBM Cloud for Financial Services allows companies to benefit from both worlds and achieve their goals in hybrid cloud specially to infuse AI into IBM Z workloads.