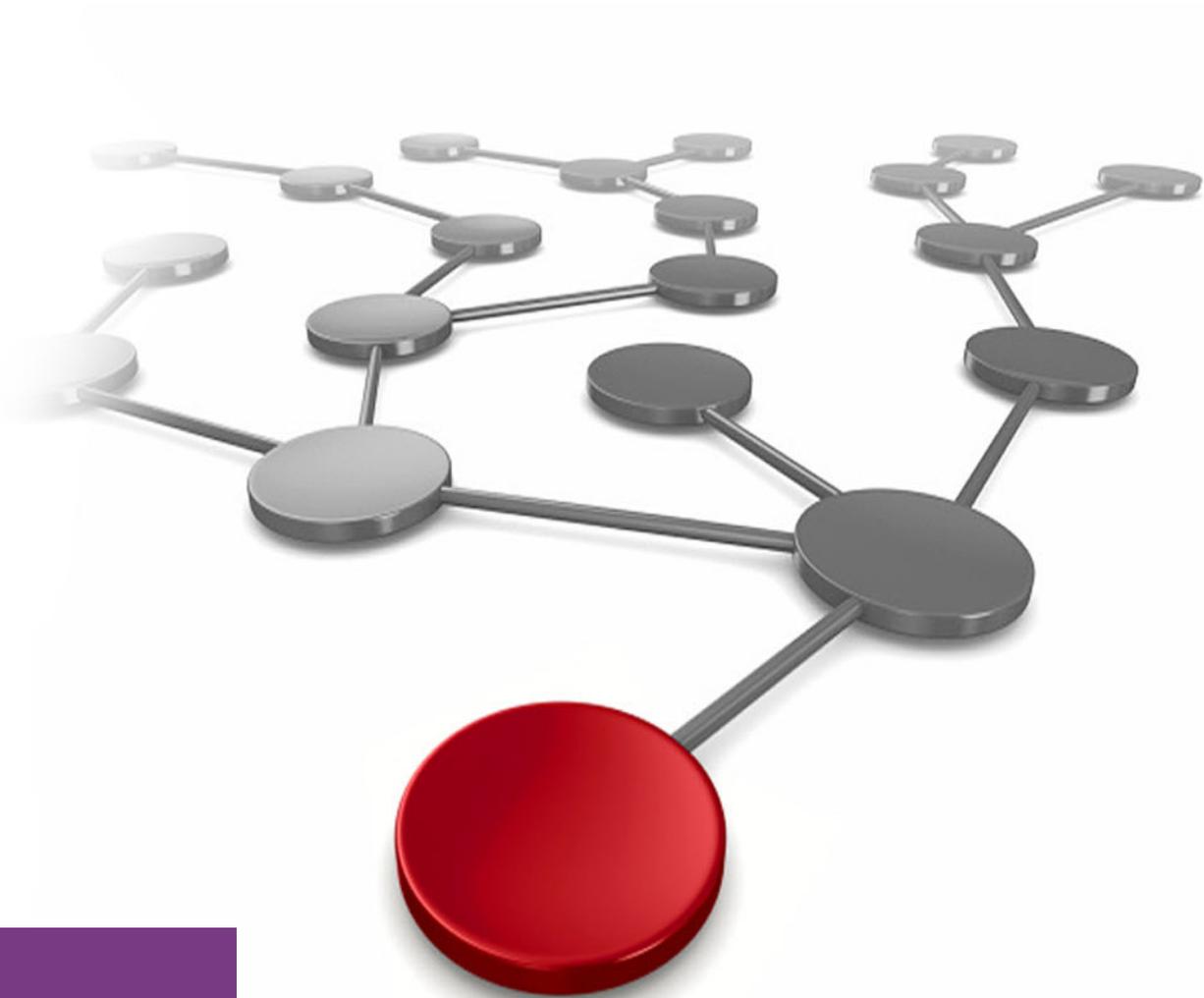


Demystifying Data with AI on IBM Z

Richard Ruppel
Roger E. Sanders
Khadija Souissi



IBM Z



The need to compete: Harnessing related data with low latency

Highlights

When it comes to AI, the industry recognizes the need for transparency, trust, security, speed, and gaining actionable, real-time responses without impacting the client experience. As your organization develops and proceeds on the journey to AI, consider the following benefits laid out in this paper:

- ▶ A key pillar of the AI on IBM Z strategy is to build a world class inference platform that supports a diverse set of business and IT use cases.
- ▶ The AI Ladder is the IBM guiding principle for organizations to embrace and transform their business by connecting data and AI.
- ▶ The IBM Watson portfolio includes a rich set of tools and runtimes that make it easy to use data from a variety of sources; build, train, and deploy AI models anywhere; trust model recommendations and predictions; and get more value from AI quicker.
- ▶ IBM infuses AI into its offerings to help clients manage environments and enhance security.

Artificial Intelligence (AI) is a profoundly transformative technology that is poised to revolutionize practically every industry because of its broad applicability to a wide set of use cases. It is already impacting our personal lives through so many real-world applications that we barely notice it. Moreover, innovative companies are working to infuse AI into their applications to automate repetitive tasks, improve customer experience, enable them to make smarter decisions faster, and help them gain competitive advantages. However, AI still faces many practical challenges.

AI is generally thought of as the all-encompassing computer science concept that is concerned with building smart machines that can perform tasks that normally require human intelligence. Machine Learning (ML), a subset of AI, refers to a broad set of techniques that give computers the ability to "learn" by themselves, using existing data and "trained" algorithms or models. ML is used primarily for things like email spam filtering, virtual assistants, recommender systems, customer care, and online fraud detection. Deep Learning (DL) is a technique for implementing ML that relies on deep artificial neural networks - modeled loosely after the neural networks found in the human brain - that are designed to recognize hidden patterns in data to perform complex tasks such as image recognition, object detection, and natural language processing.

AI is only as good as the data it has been trained on, and the more data models consume, the better they get. But obtaining large datasets that are comprehensive enough to be used for model training can be difficult. In some cases, even small datasets can be hard to come by due to privacy laws or security concerns. Another hurdle to overcome is latency. To make real-time decisions, you need the real-time insights that are generated when AI models are scored directly in transactional workloads every time a transaction is executed. Without the correct information architecture and sufficient compute resources available, difficulties with implementation can arise when AI systems are scaled from one use case to another, such as when going from predicting customer churn or analyzing sales data for a particular store, to a company-wide deployment. Additionally, the "black box" complexity of DL technology creates the challenge of being able to show which factors led to a decision or prediction, and how. This is particularly important in applications where trust and transparency are critical and the decisions made carry societal implications, such as in criminal justice situations and financial lending. IBM® recognizes these challenges and is leading efforts to overcome them in enterprises worldwide.

The IBM Z® platform is the business world's preferred system of record, responsible for more than 70% of the operational data generated by governments, financial institutions, retailers, and other large enterprises around the world. Yet, some organizations have evolved to a strategy of moving huge volumes of data off IBM Z for AI development and deployment, daily. Data movement brings its own unique challenges, costs, and security risks. Worse, the more data is moved away from its source of origin, the more outdated the insights that can be gleaned from that data become due to latency. Outdated insights can lead to poor decisions, resulting in the loss of revenue or potentially harmful actions. Today's business opportunities and questions demand answers in real-time, requiring trusted analytic and AI insight from the most current enterprise data available.

IBM Z is a top-grade AI infrastructure that offers the combination of low-latency, high performance, reduced complexity, and resiliency within a security-rich environment that enterprises demand. Therefore, our goal is to provide a comprehensive and consumable AI experience for operationalizing (deploying) AI on Z, as well as build a world class AI inferencing platform. Often, the landscape for AI model development and training is quite different from the one used for model deployment and inferencing. Therefore, the approach IBM is taking is to enable organizations to build and train models on their platform of choice, including Anaconda, RStudio, PyTorch, TensorFlow, MXNet, Caffe, and more, and be able to deploy those models to an environment that has data affinity to mission-critical applications running on IBM Z. To that end, IBM has made significant investments in Open Neural Network Exchange (ONNX) technology, which enables AI models to be easily transferred to an IBM Z system for deployment. Furthermore, upstream contributions to opensource by IBM has resulted in a growing ecosystem of community-supported builds for the IBM Z platform. By combining existing IBM Z resources with concepts like data virtualization, Hybrid Transaction/Analytical Processing (HTAP), and new technology provided by IBM Watson®, businesses can infuse AI models directly into their transactional processes to deliver trusted and transparent actionable insights in real-time, while ensuring their applications adhere to strict service level agreements (SLAs).

Climbing the AI Ladder

Data is the fuel that powers digital transformation and it is AI that unlocks the value of data and transforms the way businesses operate and deliver value. Yet, the adoption of AI remains formidable. To successfully scale AI throughout an organization, companies must overcome three major challenges: data complexity, talent scarcity, and a lack of trust in AI systems.

IBM recognizes this and using information gleaned from thousands of AI engagements, we have built a prescriptive approach to successful AI implementation that we call "The IBM AI Ladder," illustrated in Figure 0-1. The AI Ladder™ provides a framework that helps organizations overcome these and other challenges so they can turn their AI aspirations into real business outcomes. It enables companies to simplify and automate how they turn data into insights by unifying the collection, organization, and analysis of data, regardless of where that data lives. Additionally, it serves as a technology roadmap that unifies how the data and AI products and services offered by IBM work together to accelerate the journey to AI.

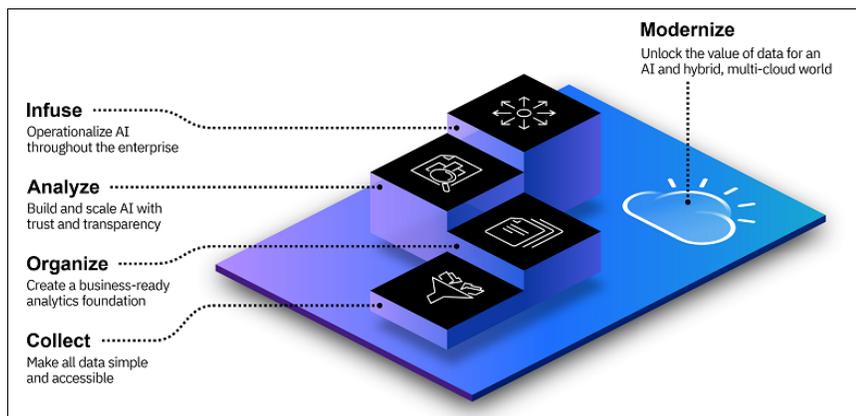


Figure 0-1 AI ladder

The AI Ladder consists of four steps (or "rungs"). The first step, Collect, is about making data simple and accessible. Data can come from anywhere (public sources, public clouds, private clouds, on-premises servers)- and can be stored in a variety of formats. However, not all data is accessible due to technical, regulatory, or other reasons. So, the first step of the AI ladder focuses on collecting all types of data from everywhere and making it as broadly accessible as possible, subject to internal policies and regulatory requirements. This step is also concerned with providing uniform access to data, regardless of its source or type. Users should not have to resolve integration problems to use the data after it has been made available.

The second step, Organize, is concerned with creating a trusted, business-ready analytics foundation with built-in data governance, data protection, and compliance. If you look at data, even casually, you will discover that most data has its share of problems: ambiguity, missing values, and values that are obviously incorrect, to name a few. Not only do these and similar problems need to be resolved, but redundancies and contradictions need to be eliminated. Steps must also be taken here to ensure that data is correctly cataloged and that data sources are properly documented. Moreover, sensitive data must always be kept private and secure, and appropriate data governance must be put in place to ensure that data is kept consistent and trustworthy.

After the data has been properly collected and organized, it is ready for use with AI. So, the third step on the AI Ladder, Analyze, focuses on helping organizations engage with AI to gain insights from their data that can help them transform their business. To this end, it is important that organizations follow the AI lifecycle, which consists of three stages:

- ▶ **Build:** This is where AI algorithms, or models, are created. It is important that the proper algorithms are chosen and evaluated here so the final model selected is the model that is best suited to solve the problem at hand.
- ▶ **Run:** After a model has been built and verified, it needs to be put into production, often as part of a more complex application workflow or business process. At this stage it is critical that the model built can make correct decisions and predictions when presented with "real-world" data.
- ▶ **Manage:** After put into production, a model must be continuously evaluated to ensure its results align with the key performance indicators and business metrics that drove its development. At this stage, to avoid bias and provide transparency, companies must be able to explain exactly how and why the model makes its decisions and recommendations.

The last step, Infuse, is where AI gets "operationalized." In other words, where AI models get deployed in production environments throughout the enterprise, under standards that guarantee quality is high, errors are low, SLAs are met, and feedback loops are in place to ensure model quality does not degrade over time. To take advantage of everything AI has to offer, organizations must embed it across as many of their workflows as possible; they need to push it into every department, every business process, every activity. This requires an AI foundation that can easily be scaled as business needs grow. IBM Z helps in this area by enabling AI applications to run close to where system of record data originates, eliminating the need for data movement. It may also require existing workflows to undergo significant revisions or modifications. Or it may force entirely new workflows to be developed. While this work effort may ultimately prove to be complex, it represents the last step towards an effective, AI-infused, digital transformation.

Spanning the four steps of the AI Ladder is the concept of Modernize, which refers to the building of a proper information architecture for AI, with the aim of gaining the flexibility needed to meet today's demands, and provide the ability to remain competitive in the future. This requires a shift towards hybrid cloud technologies. Many companies remain reluctant to embrace the public cloud, preferring to adopt private cloud solutions or keep their mission-critical data and applications on dedicated servers so they have greater control of their environment and can more easily comply with regulations. However, by opting for a hybrid multi-cloud platform like IBM Cloud® Pak for Data, these organizations can use their data and applications on-premises and across public or private clouds by exploiting the power of containers. Hybrid multi-cloud platforms provide greater operational agility, deliver access to more data that can be used to fuel smarter AI, and enable users to identify, analyze, and respond quickly to changes. Additionally, data management can be carried out effectively, even when extreme data proliferation spreads information across multiple clouds and data centers. This type of infrastructure is flexible enough to accommodate new types and sources of data as they become available.

The IBM Z advantage

Today, organizations have already begun exploring AI capabilities to get actionable business insights from data. When it comes to operationalizing AI, there is a need for organizations to address the requirements related to infusing AI models into applications efficiently without impacting transactional performance and causing complexity. AI production implementations require a resilient, low latency, high performance environment to apply AI model inferencing within existing operational processes. IBM is playing a key role in enabling clients to deliver enterprise-scale AI and infuse intelligence into applications, keeping data in-place and bringing AI to that data. This allows organizations to deliver insight when it is needed, where it is needed at the point of impact. All while reducing latency, minimizing cost and complexity, and improving data governance and security.

IBM Z runs up to 19 billion encrypted transactions a day.¹ This implies a significant level of potential benefits when applying AI where those transactions take place, making it possible to inference every transaction to enrich it with trusted, actionable insights and be able to make the correct decision at the right time and tune the transaction. These transactions come in many forms, from many different industries. Be it a banking transaction that requires fraud detection before it can be approved, an online retail transaction for which an AI model can propose personalized offerings at the point of sale, an insurance claim adjudication process where overpayment detection can help avoid loss, or an AI model to detect if the transaction is taking too much time. These examples help illustrate how applying AI at the point of transaction in any industry can bring great value compared to getting the insights after the transaction has occurred. For enterprise clients running on IBM Z, the platform supports applications and services that produce a wealth of valuable transactional data originating and residing on the platform. The Z platform is uniquely positioned to offer the combination of low latency, high performance, reduced complexity, and resiliency within a security rich environment.

When it comes to the question of where this data should be leveraged for AI, many organizations have started copying and moving data to other systems that are used to apply AI and analytics. This approach leads to challenges like data latency, data incoherence, security intrusion points, and increased complexity and costs. Data latency is the time between when data originates and an action is taken based on that data. To make real-time decisions, clients need real-time insight. When data is moved away from its original source, the insights can become outdated quickly. Every copy of data has its own cost, latency, and decision risk. Outdated insights can lead to poor quality decisions and potentially harmful actions. Furthermore, multiple copies of data can cause security and compliance concerns, especially if data is moved outside the firewall. When insights are needed fast, calling off-platform is sub-optimal. This brings additional latency risk and can cause only a subset of data to be eligible for inference. Thus, the insights cannot be counted on reliably and opportunities can be lost.

To reduce the data latency gap and avoid these challenges, the IBM AI strategy centers around enabling clients to mine the insights close to the data, where it originates, driven by data gravity. As organizations reimagine their approach to digital transformation and use AI and ML to extend their competitive advantage, it is the ideal time to consider the benefits of bringing AI to the transactional data instead of bringing data to AI. When deploying AI on the IBM Z platform, you can get the most out of your data. Generating insights where the data originates ensures that current insights are available in real-time, at the point of interaction. This also minimizes the movement of sensitive enterprise data, drives cost efficiencies, improves governance and security, and reduces decision latency risk. Reliable in-transaction inferencing requires consistent response times within SLAs. AI on IBM Z addresses this requirement by ensuring a low latency inference and a high throughput. IBM Z provides consistent response times with optimized inference that can scale with clients' workloads. This eliminates the overhead acquired when going off-platform for insight.

Furthermore, most organizations need to work with existing infrastructure investments. IBM is enabling these organizations to improve productivity by leveraging their existing investments in people, processes, and infrastructure. By deploying AI on Z, you can get the most out of your infrastructure investment, minimize IT complexity and costs, and benefit from the same high-level of resiliency for AI as you have with your applications. In fact, when applications are dependent upon AI insights, the AI workload requires the same resiliency as the applications it complements. Tight integration of AI with data and core business applications that reside on IBM Z allows you to leverage the qualities of service you expect from IBM Z, ensuring resiliency for both applications and

¹ <https://www.ibm.com/it-infrastructure/z/capabilities/transaction-processing>

Tight integration of AI with data and core business applications that reside on IBM Z allows you to leverage the qualities of service you expect from IBM Z, ensuring not only applications but also AI resiliency, 99.99999% availability, enhancing security by encrypting data everywhere, and leveraging technologies like Data Privacy for Diagnostics and Hyper Protect Data Controller. Encryption enables secure AI scoring on IBM Z for your most sensitive data be it on-premises or in-flight analysis.

New personas like data scientists and data engineers can continue to work with their favorite state of the art and opensource technologies which IBM embraces in its offerings. IBM Z supports the most popular ML algorithms, so that your data scientists can continue working with their tools of choice, in turn eliminating the time and education required to learn a new tool. They can also leverage the collaboration capabilities offered by the IBM Z AI offerings, and scale to build and deploy many AI models in the same environment.

Major technologies that enable AI on Z

IBM Z systems process approximately 30 billion transactions a day, up to 19 billion of which are encrypted. So, it is no wonder that IBM Z is where 70% or more of the world's corporate data originates. For large organizations that rely on IBM Z to process their mission critical workloads, moving data to another platform for AI is not a viable approach. Data movement impedes both data usage and time to insight; therefore, it makes more sense to co-locate analytics, securely, with the data. Moreover, AI must be able to process enormous amounts of data and is very compute intensive. The Single Instruction Multiple Data (SIMD) accelerator that is integrated in the newest Z system CPUs can act on multiple data items simultaneously (in response to a single instruction), enabling AI algorithms to extract real-time insight from financial and consumer transactions as they are processed.

The use of Linux on IBM Z over the course of 20+ years has opened the doors to a vast ecosystem of opensource software that has been compiled and validated by IBM for s390x architecture. More recent IBM Z features, such as RESTful APIs and z/OS® Container Extensions (z/CX), support emerging workloads and make it possible to deploy Linux on Z applications in Docker containers for workflows that require an affinity to z/OS. Additionally, the opensource project Zowe has created new opportunities to use modern interfaces to interact with IBM z/OS, via plug-ins and extensions created by third-party vendors and IBM clients. Optimized libraries and compilers ensure that frameworks like TensorFlow, Spark, scikit-learn, and so forth, can be deployed on IBM Z, enabling organizations to seamlessly leverage their hardware and software investments. IBM Z now includes a cloud-native ecosystem with Kubernetes and containers that requires no Z platform-specific development skills. Additionally, Red Hat OpenShift on IBM Z extends those capabilities with integrated tooling and a feature-rich ecosystem.

Often, the landscape for AI model development and training is quite different from the one used for model deployment and scoring. Therefore, the approach to AI IBM is taking is to enable organizations to build and train models on their platform of choice, yet be able to deploy those models to an environment that has data affinity to mission-critical applications - such as the transactional applications that run on IBM Z. As a result, data professionals who build and train AI models somewhere other than IBM Z can take advantage of predictive model markup language (PMML) or the ONNX to easily move those models to an IBM Z system for deployment. PMML is an open-source XML format for describing data mining and statistical models, including inputs to the models, transformations used to prepare data for data mining, and parameters that define the models themselves. ONNX is a new open-source industry ecosystem, established by Facebook and Microsoft (IBM Research® contributed the TensorFlow/Keras translator, which is used to convert TensorFlow DL models to ONNX format) that provides a common open format for representing DL models. ONNX enables data scientists to build and train models in one tool stack, such as PyTorch, TensorFlow, MXNet, and Caffe, and then deploy those models on a different platform for scoring without worrying about downstream inference implications. This is the heart of our "train anywhere, deploy on Z" strategy and it is why we are architecting solutions that facilitate model portability to IBM Z and enabling organizations to seamlessly leverage their existing hardware and software investments.

Databases used for transactional workloads are optimized for processing complete rows in a table, whereas databases used for analytics should be optimized for processing individual columns. Consequently, HTAP implementations utilize technology that reorganizes data stored in rows in a transactional system into columnar format for analytics so both workload types are fully optimized. Thus, there is virtually no latency because the

columnar data used for analytics is based on the latest committed transactional data. The IBM Db2® Analytics Accelerator (IDAA) is a cost-effective, high-speed, in-memory query engine that is designed to run business reporting and analytics workloads efficiently. It transforms IBM Db2 for z/OS, which is purpose built for transactional workloads, into a highly efficient HTAP environment, thereby aiding in the Organize step of the AI Ladder. It also drives out cost and complexity while enabling real-time analytics on transactional data as it is generated. As part of its unique design, IDAA utilizes breakthrough technologies that route queries typically found in transactional workloads to IBM Db2 for z/OS and queries typically found in analytics applications to IDAA. Thus, each query executes in its optimal environment for maximum speed and cost efficiency.

The surge of AI application development is driving an exponential growth in the need for data access. As a result, organizations often need to leverage data that originates on IBM Z for information-driven projects, which may run on non-Z platforms. To that end, many organizations have embarked on projects that attempt to copy data from Db2 for z/OS to traditional RDBMSs, data warehouses, Hadoop data lakes, Kafka streaming hubs, and so forth, in real-time. To assist with the Collect step of the AI Ladder, IBM Db2 for z/OS Data Gate (Db2 Data Gate) offers a new way for IBM Z clients to provide read-only access to data originating in Db2 for z/OS – without having to access Db2 for z/OS and without impacting source systems on IBM Z. Instead of accessing data in the IBM Z data source directly, an application accesses a synchronized copy of Db2 for z/OS data, hosted by a separate system running in IBM Cloud Pak® for Data, a fully integrated data and AI platform that modernizes how businesses collect, organize, and analyze data to infuse data-powered AI throughout their organizations. The target system can be established anywhere IBM Cloud Pak for Data is supported, enabling a wide range of target platforms that include on-premises, public cloud, and private cloud deployments.

IBM Analytics Engine is a cloud-based service that is designed to eliminate many of the key pain points associated with trying to grow Big Data analytics capabilities. It helps with the Analyze step of the AI Ladder by enabling data professionals and application developers to rapidly provision, manage, run, and retire deeply complex Hadoop and Spark clusters that might distract them from their core responsibilities and impede their AI efforts. It also allows them to develop and deploy advanced analytics applications in minutes. Rather than use Hadoop as both a computation engine and a persistence layer for long-term data storage, Analytics Engine handles compute resources and storage separately. This separation allows both to be scaled independently to meet business needs. As data volumes grow, more storage can be added, as the number and scope of data science projects increase, more nodes can be added to a cluster, or clusters with different node sizes can be quickly spun up. It also enables organizations to scale their Big Data landscape in line with business requirements while minimizing unnecessary infrastructure investments. The use of object storage helps to protect data and improve availability, and the dynamic spin-up/down of clusters, as needed, helps simplify maintenance and upgrade operations. Because Analytics Engine is a supplemental service of IBM Watson Studio, it integrates seamlessly with other Watson tools and runtimes, resulting in a simplified, streamlined, user experience with best-of-breed solutions that support every step of the AI Ladder.

To drive insights and outcomes at scale, data science teams need access to the most relevant data available. But all too often they encounter barriers that prevent them from finding and accessing the data they need. Even when the proper data is located, data privacy laws and regulatory compliance requirements can restrict how that data is used and shared. Consequently, a data professional can spend much of their time searching for the proper data and preparing it for use with AI. The solution to these challenges on IBM Z is IBM Watson Knowledge Catalog, running on Linux on Z. Assisting with the Collect and Organize steps of the AI Ladder, Watson Knowledge Catalog is a collaborative, AI-powered data catalog that is designed to discover and catalog enormous volumes of complex data and analytic assets, spread across multiple sources. It can connect to a diverse set of data sources like relational database management systems (RDBMSs), NoSQL databases, Hadoop Distributed File Systems, and more. It can also curate and catalog a wide variety of analytical and AI assets, including structured and unstructured data, notebooks, models, and dashboards. Thus, it works to eliminate barriers to data access while ensuring that every bit of information collected is securely indexed, classified, and accessible. Because understanding and trusting data is essential to delivering quality AI, Watson Knowledge Catalog utilizes Watson technology to enrich its metadata index of information with classification, profiling, and quality assessment content. It also delivers active policy enforcement capabilities that can both control access to information assets and provide automatic and dynamic masking of sensitive data elements, enabling data professionals to build and train models using sensitive data they normally do not have access to. This means that no matter the team or the situation, the

correct people have access to the correct data at the right time without running the risk of violating regulatory compliance requirements.

Data professionals use different tools to extract the full value from data for a number of reasons. Some prefer tools that offer a drag-and-drop approach to model building while others want a more powerful integrated development environment (IDE) like Spyder, PyCharm, RStudio, and Jupyter Notebook. So, it is not uncommon for an organization to end up juggling a large, complex collection of tools. However, the larger the collection, the more difficult it becomes to integrate them into workflows. This is where IBM Watson Studio helps clients with the Analyze step of the AI Ladder. Watson Studio delivers a simplified, all-in-one data science and ML environment that enables data professionals to prepare and analyze data, build and train AI models anywhere, and optimize decisions by uniting teams, automating AI lifecycles, and accelerating time to value. It accommodates those who prefer the drag-and-drop approach with IBM SPSS® Modeler and Visual Model Builder, while offering enterprise-hardened versions of popular opensource tools such as Jupyter Notebook and RStudio for those who work with popular programming languages like Python, R and Scala in an IDE. With Anaconda's expansion to IBM Z, data professionals can grow their opensource data experience while continuing to interface with other key tools and frameworks, such as conda, xGBoost, and scikit-learn. Coupled with project management features like programmatic access and version control, Watson Studio creates a workflow that is incredibly efficient. Data professionals can share assets, results, and models they created with different tools within the same project, in an integrated way, enabling users to collaborate with confidence. Additionally, with its newest feature, AutoAI, users can build models faster, scale experimentation and deployment, and increase trust and transparency. AutoAI automates data preparation, model development, feature engineering, and hyperparameter optimization. Furthermore, because Watson Studio is seamlessly integrated with IBM Watson Knowledge Catalog, users can easily transform data and AI models into trusted enterprise resources.

Moving AI beyond research papers and experimentation to real-world environments can be intimidating because of the complexity involved with integrating AI models with existing, mission-critical applications. There can also be concerns about the impact AI integration might have on the ability to meet SLAs, and challenges with maintaining the production-level accuracy expected from a model after deployment. IBM Watson Machine Learning (WML) solves these challenges and aids with the Infuse step of the AI Ladder by accelerating AI and ML model deployment, at scale. With WML, data professionals can deploy ML, DL, and decision optimization models on-premises or across any cloud, with a single click. Users can mix and match models from IBM Watson Studio, SPSS Modeler, and open-source notebooks. They can also dynamically keep deployed models accurate through continuous learning and retraining. Additionally, WML can manage and monitor models for inaccuracies such as drift, bias, and risk. All of which greatly simplifies the effort required to maintain production-level model accuracy. Finally, WML helps to scale AI production by automatically generating the application programming interfaces (APIs) that are frequently required to infuse AI models into new or existing applications.

On IBM Z, Watson Studio and WML have been combined to create a single product, IBM Watson Machine Learning for z/OS (WMLz). WMLz offers the same feature set, functionality, tooling, and collaborative work environment as Watson Studio and WML, thereby aiding with the Analyze and Infusion steps of the AI Ladder. However, it also enables users to work directly where 70% of enterprise data originates, on IBM Z. It delivers a high speed, low overhead model scoring engine on z/OS, making it possible for ML models to be scored directly in transactional workloads running on IBM Z. WMLz utilizes a z/OS Apache Spark cluster that continuously ingests data into the WMLz pipeline, in real-time. It also has an application cluster that can run on z/OS or Linux that provides a web user interface and an administration dashboard. Additionally, a Linux-based custom IDE is available, if desired. To make real-time decisions, you need real-time insights that are based on the most recent and relevant data available. With so much of an enterprise's critical information being accessible from z/OS, WMLz is profound in terms of enabling organizations to apply predictive analytics in-transaction.

IBM Watson OpenScale™ is an open platform that gives businesses a clear and accurate view of their AI systems, helping them monitor and fine-tune performance throughout a system's lifecycle. It helps eliminate trust and transparency concerns that prevent companies from fully adopting AI, assisting with the Infuse step of the AI Ladder. For example, if an insurance company utilizes AI to process claims and a claim is rejected, an explainability feature enables a claims processor to get detailed, understandable answers to questions clients or regulators might ask. Traceability lets anyone retrace the process that was used to make the decision, step-by-step, right back to the original documents and data the AI system drew from. Additionally, if harmful bias swayed the outcome, a bias

feature pinpoints how it occurred and automatically works to lessen its impact. Watson OpenScale puts businesses in complete control of their AI's full lifecycle. It makes creating, scaling, and evolving AI models much simpler, helping companies bridge the data science skills gap. Moreover, it gives organizations the confidence to make fast, accurate decisions based on AI they know is free from harmful bias, while being open and easily pluggable into existing workflow processes.

Many of the offerings for IBM Z discussed above and how they tie together to help with each step of the AI ladder can be seen outlined in Figure 0-2.

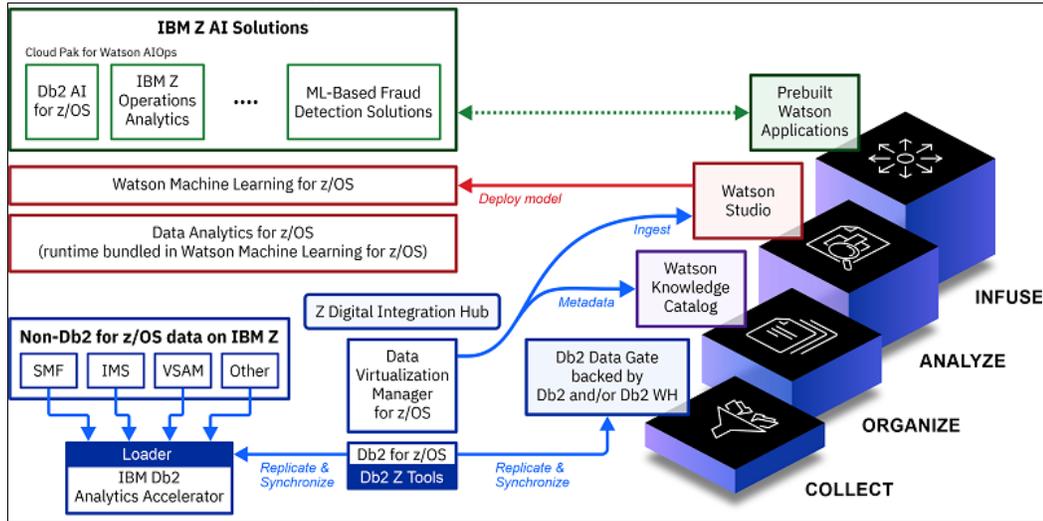


Figure 0-2 IBM products and solutions that enable clients to climb the AI Ladder

Understanding the flow of AI on IBM Z

Before diving into the architecture, let us first look at what it takes to create and use AI models. When it comes to building AI models and incorporating them into applications, there are two distinct phases: developing and training the models and deploying the models, as illustrated in Figure 0-3. Clients can potentially take different approaches for each phase of the process.

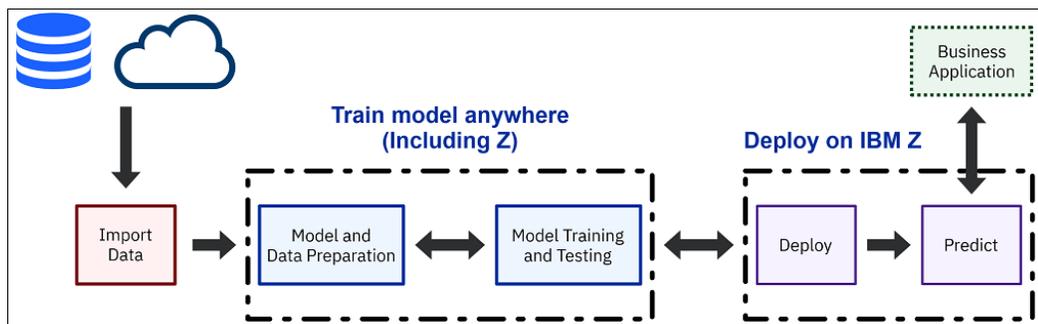


Figure 0-3 Flow of AI models

The development and training phase is where data scientists bring together data from various sources and work to detect correlations and patterns that ultimately will be used in business logic to make better decisions. Data scientists must first prepare the data, transforming structured and unstructured data into a usable format. Then a model (algorithm), based on the relationships and patterns, is created. Next, data scientists train and test the model to improve its accuracy. This can be a time consuming and resource intensive process as data scientists test new parameters and train with additional data to get the best possible outcome. It has been said that the

development and training is more of an art than a science and the results of improvements in model accuracy may have real impact on a company's bottom line.

Data scientists may use a variety of toolsets, frameworks, and platforms during the development and training of models. The choices are typically based on what the data scientists are familiar with or what their company has invested in. IBM Z has enabled popular open-source data science packages and have optimized libraries and compilers of AI frameworks and runtimes that leverage the IBM Z architecture, allowing data scientists to complete their work on the IBM Z platform. Advantages like data gravity, security, and data governance makes IBM Z a good option. However, IBM enables clients to build and train models where it makes sense for them, leveraging existing skills and resource investments.

The other phase is on the deployment of models. Here the models created by the data scientists are made available for applications to exploit. Applications can then be enhanced to use the inference result that is returned from the model. Take a simple loan approval application as an example. A company can spend resources coding criteria and trying to keep the code up to date that is used to determine if a loan should be approved. Alternatively, the application logic can simply rely on an AI model as shown in Example 0-1.

Example 0-1

```
approveLoan = 'Y' if model returns '1' or approveLoan = 'N' if model returns '0'
```

The model is deployed into production and the inference can occur with every transaction. Additionally, it is important to validate the accuracy of the model. In this example, if loan defaults increase, data scientists can retrain the model with new data and work to improve its accuracy. After changes to the model are made, it can be re-deployed, with potentially no changes to the application.

The key point on the deployment side of the equation has to do with where the model is deployed. The ecosystem used for development and training may not offer the capabilities needed to score every transaction, particularly in a high-volume application. Deployment of models on a different platform from where an application runs, for example, may introduce intolerable latency into the transaction. A key pillar of the AI on IBM Z strategy is to build a world class inference platform. The kind of platform where clients can infuse AI into their most demanding transactional workload with minimal resources and cost. Also, the kind of platform that gives clients the ability to leverage the insights they have gained from their data to drive business value in every transaction, without impacting end-user experience.

As stated, IBM has enabled popular toolsets and frameworks to run on the IBM Z platform. Building and training models on IBM Z has many advantages but the reality is that this work can be done anywhere. What cannot be found just anywhere is a world class inference platform. Deploying models on IBM Z provides low latency and scalability for the in-transaction inference that is needed for real-time or near real-time business cases, such as fraud detection, that can result in significant savings or losses. To that end, IBM has made investments to allow model training to be done anywhere and easy deployment on IBM Z. IBM is focusing on making models, and other assets created with these models, seamlessly portable to IBM Z. We do this by taking advantage of ONNX technology, allowing data scientists to build models in any popular framework and convert them into a common, portable format. ONNX models can then be deployed on IBM Z through the use of an ONNX model compiler developed by IBM research. This compiler builds on ONNX-MLIR (Multi-Level Intermediate Representation) project and serves to create a highly optimized inference program. Additionally, IBM is leveraging z/OS Container Extensions (zCX) as an integral part of the AI on IBM Z strategy. zCX provides the ability to deploy containerized AI frameworks and workloads under z/OS. TensorFlow, one of the most popular opensource frameworks today for both model creation and training and an inference platform, is available for both Linux on Z and z/OS clients via zCX. With zCX, TensorFlow is available to applications through TensorFlow Serving. TensorFlow on IBM Z has received platform-specific hardware and software optimizations to improve performance. The container-based strategy coupled with zCX will help bring new frameworks to the platform more quickly.

With an understanding of how models are created and a focus on the idea that deploying models on IBM Z is advantageous, let us look at the components of the architecture on IBM Z. Figure 0-4 illustrates how AI can be infused into a IBM CICS® and Db2 for z/OS application running in z/OS.

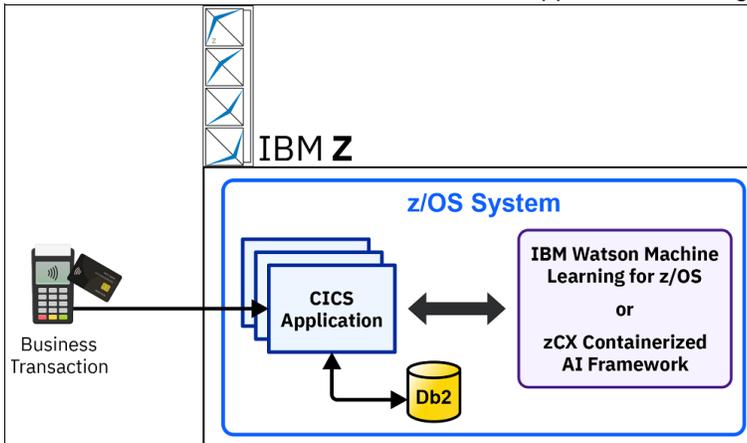


Figure 0-4 Architecture components

Before adding AI, a transaction is initiated and a CICS application processes the request. Data stored in Db2 for z/OS is retrieved, application logic is executed, and a result is provided back to the requestor. As illustrated in Figure 0-4, the application can be enhanced with a REST (representational state transfer) API call to the AI inference service running in z/OS. The service can be IBM Watson Machine Learning for z/OS, running in an IBM WebSphere® Liberty server, or the service can be a containerized AI framework, such as TensorFlow Serving, running in zCX. In either case, a result from the inference is returned to the calling program and application logic can act based on the value returned.

Transactional affinity, or keeping the inference service close to the transaction, is imperative in achieving low latency and processing large volumes of data. Running CICS and either zCX or the Liberty server on the same z/OS logical partition provides the lowest network latency possible. Accessing the inference service in the cloud may have a one second or more response time whereas accessing the service in a different area of memory in the same operating system may take microseconds. Of course, running the inference service in z/OS can provide the same level of resiliency as the application itself. Consider the lost opportunity if the inference service is unavailable for even a single transaction.

Scenerios: AI on IBM Z

AI and ML can be applied to a variety of use cases across problem types and industries to drive business insights and competitive advantages. While sales and marketing can leverage AI to offer personalized offerings to a customer or forecast customers' demands, an operations department can apply AI to optimize processes and enhance performance. AI can also have a focus on issue detection and problem prevention – for example, to detect and prevent fraud and customer churn. Whatever the use case pattern is, organizations can apply AI to get business advantages in the areas of revenue increase, cost reduction, and turnover speed.

Figure 0-5 demonstrates these key areas of which AI can be applied to create new opportunities and gain business advantages.

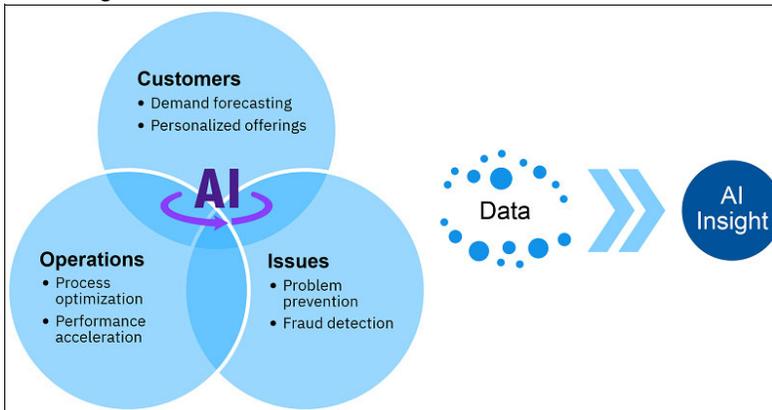


Figure 0-5 AI use cases

Specifically for IBM Z, there is a spectrum of use case patterns and AI applications which can be categorized in two types of usage clusters, namely business use cases and IT related use cases.

Business

One common use case within the financial industry consists of learning patterns in the data to identify fraud before a transaction closes, instead of after the transaction has taken place. In fact, performing AI inferencing off-platform can lead the organization to take on unwanted risk and approve transactions without additional fraud inspection due to system unavailability or network latency. Being able to scale and apply fraud detection for every transaction in real-time allows a bank, for example, to mitigate risk, help prevent money loss, and protect the organization's reputation and client relationships. Applying AI for anti-money laundering can also help banks minimize risk by enabling the ability to detect multiple account adversarial money wiring patterns. The same bank can mine data to determine whether to provide or extend a loan to a customer in real-time. As customers are getting more and more open to change, the bank can also use AI to assess whether a customer is likely to stop doing business with them.

AI can also be applied to infuse intelligence in the trade clearing and settlement process for risk mitigation, for example, using AI models to predict which trades or transactions have high risk exposures and propose solutions for a more efficient settlement process. The expedited remediation of questionable transactions can prevent costly consequences, regulatory violations, and negative business impact. With this, without impacting SLAs and batch process window, organizations can proactively work to prevent losses, and lower operational, regulatory, and compliance costs.

Insurance companies can infuse intelligence into their claim processing applications to detect exceptional claims or overpayment at the point of claim adjudication. Thus, an insurance provider can prevent inappropriate payments at the right time instead of going through the process and incurring the related overhead of collecting the money back, with a risk that the money is never returned. Additionally, insurance companies can use yet another AI model to predict if a customer is underinsured, thereby minimizing risk. Furthermore, applying AI against customer data allows the insurance company to perform customer classification and segmentation, a valuable feature when it comes to positioning new insurance services and predicting acceptance likelihood.

Government agencies can use AI algorithms to help address compliance requirements by detecting non-compliant behavior, such as tax payment fraud or tax evasion. To optimize their digital services, the same agency can use AI for pre-filing tax forms. A Social Services Department can use AI to identify fraud in benefits applications and assess the likelihood that individuals will need multiple agency support and proactively engage with other agencies to create the best outcome, and manage costs. In the automotive industry, AI models can be applied not only in sales and marketing for customer segmentation, inventory, and recall analysis, but also to optimize production processes, or perform product performance analysis. The support of DL model inference on IBM Z opens new opportunities across industries. For example, to apply AI for document processing or even to apply natural language processing.

Retail companies can leverage ML in the context of promotions and apply real-time scoring for personalized marketing to aid in up-sales and cross-sales. In retail, customer loyalty and satisfaction are critical, when customers order online, it is primarily for convenience and cost savings. Most are willing to forgo the immediate satisfaction from physically seeing and touching the item and walking out the door with the new product in-hand. When online shoppers receive merchandise that fails to meet their needs or expectations, their disappointment and the effort required for resolution may discolor their impression of the retailer, undermining both the convenience and the cost savings. First-time buyers who return an order are less likely to purchase from the same retailer again. Therefore, retailers can use AI and advanced analytics to identify both the most frequently returned products, and ineffective marketing tactics that elevate the number of returns. Using statistical and ML models, a company can take advantage of rapid alerts and automated information feeds on returned products to quickly identify problems with sales, anticipate, and quickly resolve evolving issues. Such a solution can allow retailers to develop a customized response system, enabling unique incentive offers from customer service specialists to dissatisfied buyers. This personalized, near real-time response can make a tremendous difference to customer perceptions of retailers, enabling the company to help improve customer loyalty.

As you might have noticed, every industry can profit from various benefits when deploying AI technologies on IBM Z. Most of the use case examples described require real-time insights based on data created and residing on the IBM Z platform. Based on the in-transaction inferencing and the tight integration with the core business systems running on IBM Z, AI insight can be delivered timely and in an optimal manner, during the transaction or at the point of interaction with the customer or user. When milliseconds matter, AI on Z technologies offer an optimized and performant approach to AI, on the most current transactional data, and integrated with transactional applications on an enterprise scale.

IT Operations and AI for Operations (AIOps)

AI is already impacting the role of IT and has become necessary in today's complicated IT environments. This has required organizations to learn how to use it, so that both IT professionals and businesses can obtain the benefits possible. The extensive amount of data created and residing on the Z platform, perceived as Big Data, represents an expansive potential of opportunities. However, this data needs to be unlocked and understood to detect insights that can be leveraged to help prevent outages, optimize processes and workloads running on Z, and even prevent business problems. Therefore, AIOps and IT Operational Analytics share many similarities with transactional applications, from a throughput and resiliency perspective.

It is well known that an outage can cost organizations money, prestige, and customers. Thus, incredible value can be delivered if organizations can proactively prevent an outage, or even predict one in advance by applying AI and ML to operational analytics. To achieve this, IT departments can apply AI mechanisms that detect anomalies at transaction, application, subsystem, and system levels. This can be applied to every kind of application based on the corresponding IT data. Historical IT data can be used for baseline determination to identify what is normal/abnormal for a certain transaction, application, or workload, resulting in a model that can be used to detect anomalies in real-time. Furthermore, AI allows you to perform correlation discovery, identifying abnormal behavior in databases, such as Db2 for z/OS, and transaction servers, such as CICS, for a transaction that is taking longer than expected. This can considerably help IT personnel quickly detect issues, and work to prevent further problems.

When it comes to batch workloads, IT departments are usually faced by the challenge of how to tune their batch job runs to optimally use resources, and avoid long running batch jobs that lead to additional costs. AI can help IT operators perform relevant batch workload analysis, identify patterns, and perform predictions for batch job elapsed times based on actions they can take, such as tune the processing by adding further resources. In addition, enterprises might face challenges related to missing IT operations domain skills. To increase the productivity of novice system administrators, AI capabilities on the Z platform can help interpret the content of IT data and derive insights efficiently without the need to acquire deep skills in the operating systems and workloads. In this context, IBM has already started the journey of augmenting IBM Z systems software with AI capabilities by offering turnkey solutions that can be deployed without any AI skills.

IBM and the infusion of AI in our products

IBM envisions organizations enhancing their applications by embedding AI to leverage the vast amount of data available, and generate new value and opportunities. The same vision exists for IBM offerings, utilizing AI to deliver innovation and additional value to our clients. So, it should come as no surprise that we rely on IBM WMLz to enhance our own offerings.

Companies that strive for resiliency of their systems must focus on operations, but this can be challenging as IT environments continue to become more complex. This is where AIOps comes into play. The idea is to use artificial intelligence for IT operations (AIOps) by leveraging operational data to more efficiently and effectively monitor and support systems. IBM Z provides large amounts of operational data through event and performance metrics, system management data and log files. This data can be utilized to identify issues and notify operations. IBM Z Operations Analytics (IZOA) leverages ML and operational data to gain greater visibility into systems, detect irregularities in workloads, and help perform root cause analysis more quickly. Identifying, alerting, and acting on an issue before it impacts users is a key component of providing a resilient system.

For IBM Z clients that rely on Db2 for z/OS, resiliency comes not only in the form of data being available, but also in the form of consistent performance when the data is accessed. A change resulting in slower response time of queries can be perceived by users as an outage. IBM Db2 AI for z/OS (Db2ZAI) utilizes ML and AI to help improve application performance and efficiency by collecting data and learning workload patterns. This results in SQL query optimizations through better data access path selection and sort enhancements leading to CPU savings. Db2ZAI can also detect SQL performance regression and potentially resolve situations so users are not impacted. System Health Assessment for the Db2 subsystem is also available to ensure the overall environment is running optimally. All of this is done without the need for data science skills – it is built into IBM WMLz. It greatly reduces the time of costly DBA resources to tune and monitor Db2 systems, allowing them to focus on more important tasks.

Like resiliency, security is at the top of every company's list of priorities. Protecting sensitive client data is necessary across the enterprise, especially in areas that might be easily overlooked. To provide better serviceability, IBM and third-party vendors may rely on diagnostic tools, such as dumps, for problem determination and resolution. However, the data collected from these diagnostic tools is frequently sent to vendors and it may contain sensitive data. IBM Data Privacy for Diagnostics (DPfD) utilizes ML to identify sensitive and non-sensitive data and provides tooling that clients can use to redact information in the data that should not be shared. This allows vendors to receive the data they need while enabling clients to comply with corporate data governance policies.

What's next: How IBM can help

IBM has created the roadmap for your AI journey with the AI Ladder. By collecting your valuable data and organizing it in a meaningful way, you will be able to analyze the data to gain insights that can be infused into your applications. IBM Z is the platform many organizations count on to get them where they want to go. In addition to the IBM solutions available, IBM has brought many of the popular opensource toolsets and AI frameworks to the platform. This allows companies to leverage investments in skills and technology they have already made. Furthermore, IBM offers pre-built solutions and products already embedded with AI, making it easy to get started without requiring data science skills. Where should you begin? The IBM Content Solution website lays out four steps for your "Journey to AI on IBM Z and LinuxONE":

1. Identify your use case and learn how AI on Z can help
2. Learn about available AI technologies
3. Choose a solution that fits your needs and explore how it works in your business
4. Integrate into your applications and utilize AI driven insights

Then what? Take advantage of our IBM Garage™ to build a small-scale, real world AI project to start you on your journey. IBM can help you every step of the way.

Resources for more information

For more information about the concepts highlighted in the paper, see the following resources:

- ▶ The AI Ladder by Rob Thomas:
<https://www.oreilly.com/library/view/the-ai-ladder/9781492073130/>
- ▶ IBM Data Privacy for Diagnostics product solution brief:
<https://www.ibm.com/downloads/cas/RDGELBZV>
- ▶ IBM Db2 AI for z/OS product page:
<https://www.ibm.com/products/db2-ai-for-zos>
- ▶ IBM Z Operations Analytics product page:
<https://www.ibm.com/products/z-operations-analytics>
- ▶ Journey to AI on IBM Z and LinuxONE content solution:
<https://www.ibm.com/support/z-content-solutions/journey-to-ai-on-z/>
- ▶ Data and AI Software on IBM Z and LinuxONE:
<https://www.ibm.com/analytics/data-and-ai-on-ibm-z>
- ▶ Real-time analytics on the IBM mainframe:
<https://www.ibm.com/it-infrastructure/z/capabilities/real-time-analytics>
- ▶ A Forrester Consulting Thought Leadership Paper Commissioned By IBM, "Leverage Data Where It Originates To Drive Substantial Business Benefits":
<https://www.ibm.com/downloads/cas/ZE0ENRB1>
- ▶ Deploy AI with trust and confidence with Explainable AI:
<https://www.ibm.com/watson/explainable-ai>
- ▶ Leveraging AI on IBM Z and LinuxONE for Accelerated Insights:
<https://www.ibm.com/blogs/systems/leveraging-ai-on-ibm-z-and-ibm-linuxone-for-accelerated-insights/>
- ▶ zCX Content Solution:
<http://ibm.biz/z0SContainerExtensions>
- ▶ TensorFlow Blog:
<https://ibm.biz/BdfPFH>

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](http://www.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 <http://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 Research®	The AI Ladder™
Db2®	IBM Watson®	Watson OpenScale™
IBM®	IBM Z®	WebSphere®
IBM Cloud®	OpenScale™	z/OS®
IBM Cloud Pak®	Redbooks (logo)  ®	
IBM Garage™	SPSS®	

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.

Zowe, are trademarks of the Linux Foundation.

Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

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

RStudio, and the RStudio logo are registered trademarks of RStudio, Inc.

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



REDP-5633-00

ISBN 0738459887

Printed in U.S.A.

Get connected

