Redbooks
ibm.com/redbooks

What AI Can Do for You

Use Cases for AI on IBM Z

Makenzie Manna

Diego Cardalliaguet

Mehmet Cuneyt Goksu

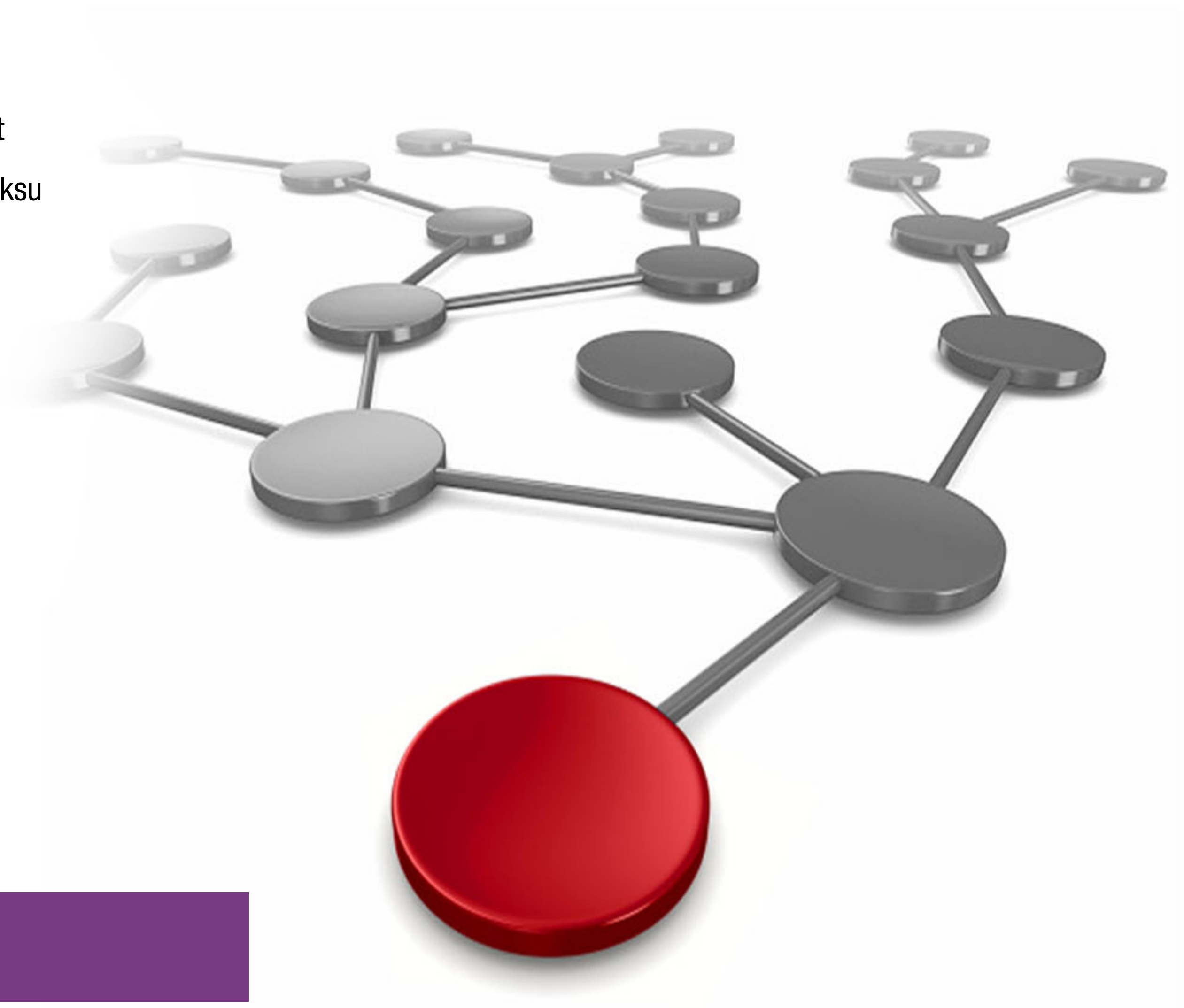
Alex Osadchyy

Lih M Wang

Sherry Yu

Poonam Zham

Erica Ross



IBM Z

IBM

Redpaper

IBM Redbooks

What AI Can Do for You: Use Cases for AI on IBM Z

July 2022

REDP-5679-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (July 2022)

This edition applies to IBM z15 and IBM z16 unless otherwise noted.

© Copyright International Business Machines Corporation 2022. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

© Copyright IBM Corp. 2022.

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.

© Copyright IBM Corp. 2022.

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 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®
Cognos®
Db2®
IBM®
IBM Cloud®
IBM Cloud Pak®
IBM Watson®
IBM Z®
IBM z16™
Insight®
OMEGAMON®
Parallel Sysplex®
RACF®
Redbooks®
Redbooks (logo) ®
WebSphere®
z/OS®
z16™

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.

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

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

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

UNIX is a registered trademark of The Open Group in the United States and other countries.

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

Preface

Artificial intelligence (AI) and machine learning (ML) are talked about as though they are in a distant future. That future is here. We live in a world where technology is fully integrated with how we live. People own smartphones, smart wearables, smart TVs, and so on. With the integration of technology into almost every aspect of our everyday lives, there is an ever-growing, massive amount of data coming from each digital interaction. This data is the critical fuel powering deep enterprise insights that can be used to expand AI capabilities and infuse those capabilities into mission-critical applications and processes to gain a competitive advantage.

With the relatively novice AI and ML technologies in demand by most industries today, there is still confusion around how to most efficiently use these capabilities to realize the most benefits. There is a diverse scope of AI and ML solutions ranging from models that aid in medical discoveries to models that detect fraudulent banking behavior to protect consumers and financial institutions from costly financial breaches. Adding to the complexity of how AI and ML can help enterprises solve their evolving business demands is the growing number of available tools and product offerings to bring these capabilities to life on the IBM Z® platform.

This IBM® Redpaper publication can help you better understand the uniquely advantageous roles AI and IBM Z can play in helping your organization realize your business goals. It introduces some of the most critical AI on IBM Z use cases currently being worked on across multiple industries, and describes component suggestions with high-level reference architectures for the implementation of each use case. We cover the following use cases (see each chapter for more specific implementations for each overarching topic):

- “Artificial intelligence for IT operations use cases” on page 19
- “Data and artificial intelligence operations use cases” on page 33
- “Use cases for artificial intelligence in chatbots” on page 47
- “Financial services sector use cases” on page 51
- “Public sector use cases” on page 61
- “Healthcare sector use cases” on page 65
- “Retail and insurance sector use cases” on page 71

This paper is intended for IT architects to understand the integration and components of the solution at a high level, and IT management and decision makers to better understand how these AI on IBM Z use cases can be applied to help solve your business problems.

Authors

This paper was produced by a team of specialists from around the world working at IBM Redbooks, Poughkeepsie Center.

Makenzie Manna is an IBM® Redbooks® Project Leader in the United States with a primary focus on IBM Z software and solutions. She holds a Master of Science degree in Computer Science Software Development from Marist College. Her areas of interest and expertise include AI solutions, IBM Z, educational course content development, video editing and production, video and cartoon animation, and technical content development.

© Copyright IBM Corp. 2022.

Diego Cardalliaguet is the Data Fabric EMEA squad leader for IBM Z. He has spent the last 22 years working with new technologies on IBM Z. Diego has worked with Java and application servers, Linux on IBM Z, Python on IBM z/OS®, and now AI. With wide technical skills on IBM Z, he coordinated complex projects and international highly skilled teams. Diego is a L3 Certified IT Specialist with a background in theoretical physics and mathematics.

Mehmet Cuneyt Goksu an Executive IT Specialist for IBM Z solutions in the context of data and AI at the IBM Germany development lab. He holds a bachelor degree and PhD in Computer Science, and an MBA in International Business. He works with IBM Db2® for z/OS and IBM Z, and has more than 30 years of experience in these areas. Cuneyt participates in IBM Db2 Analytics Accelerator and IBM Db2 for z/OS Data Gate (Db2 Data Gate) proof of concepts (PoCs), customer deployments, migrations, resolution of critical situations, and enjoys working with diverse projects and customers. Before joining IBM, Cuneyt was a member of the IBM Db2 Gold Consultant Team and named an IBM Champion. He was a member of the board of directors in the IDUG community, and worked as a regular instructor for IBM Education Programs. He is a L3 Certified IT Specialist, and actively attends certification board reviews. He is a Db2 for z/OS Liaison@IDUG EMEA.

Alex Osadchyy is a Solutions Architect and Leader at IBM Systems. He is an Honored member of National Society of Leadership and Success (NSLS). In his role at IBM, Alex works with key independent software vendors (ISVs) to create joint products for the unique value proposition for IBM Z. His focus areas include data AI and ML collocated solutions, containerized applications on Red Hat OpenShift, cloud workloads with IBM Hyper Protect technologies, IBM z/OS products modernization, and other solutions that span hybrid cloud and data solutions that are based on the enterprise platform for mission-critical applications, such as IBM Z and LinuxONE. He is a solutions architect, engineering manager, and software developer with over 19 years of experience in creating software and solutions ranging from low-level network drivers to multi-cloud platforms. With business insights, doctoral backgrounds, and research and development experience, he uses leadership and the scientific approach to create repeatable Software Development Life Cycle (SDLC) pathways for partner-building solutions with strategic AI and ML and hybrid cloud technologies on IBM Z.

Erica Ross is a Summit Technical Solution Specialist in Dallas. She holds a bachelor degree in Business Information Systems from Texas Christian University. This is her third IBM Redbooks publication contribution. Her areas of expertise include IBM Z software and technical content development.

Lih M Wang is currently an Executive Client Technical Specialist for the IBM Technology business unit. She is an advocate for AI innovations in IT Operation initiatives supporting cross-industry clients. She has worked with many clients in cross-system performance monitoring and availability management for z/OS, IBM Db2, IBM CICS®, IBM Information Management System (IMS), and IBM MQ. For AI use cases, Lih has implemented rapid (PoC) projects with real customer system data and demonstrated how AI machine learning technology can help accelerate problem prevention. Lih is an IBM certified L3 thought leader, co-authored *Using zEnterprise for Smart Analytics: Volume 2 Implementation*, SG24-8008, a frequent presenter at technical conferences, and sponsors zCouncil user groups.

Sherry Yu is a data scientist at IBM Australia. She works with clients to understand business problems and design solutions by using AI technologies. She is a researcher in the field of deep learning (DL) with a recent focus on its application for the detection of epileptic seizures by using wearables. She holds a PhD in Computer Science from Monash University. Before joining IBM 2 years ago, Shuang did research with the Defence Science and Technology Group Australia on deep reinforcement learning and natural language processing (NLP).

Poonam Zham is working as a Data Scientist - Client Engineering at IBM Australia. In her role, she helps customers unlock the potential of AI and ML by using IBM Cloud Pak® for Data and the latest AI technologies. She holds a PhD in Biomedical Engineering and received worldwide media coverage for her research on the early diagnosis of Parkinson's disease by using AI and ML. Before joining IBM Australia, she worked with Software Development Labs at IBM China and Dell EMC. She has over 12 years of experience focusing on data platforms, enterprise software architecture, and product development for cloud infrastructure management applications and disaster recovery (DR) solutions.

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

Robert Haimowitz and Lydia Parziale
IBM Redbooks, Poughkeepsie Center

Patrik Hysky
IBM Systems Technical Sales Services, Austin

Tom Ambrosio and Bill Lamastro
IBM CPO

Shuang Yu and Ke Wei Wei
IBM China

Joy Deng, Andrew Sica, Suhas Kashyap, Elpida Tzortzatos
IBM US

Iris Baron
IBM Canada

Now you can become a published author, too!

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

Obtain more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Where artificial intelligence fits into your industry

Artificial intelligence (AI) and machine learning (ML) were technologies leaving us in utter awe 10 years ago. Despite being studied since 1952, AI was relegated to the research departments of enterprises and universities. What happened since circa 2010 that brought AI and ML to the forefront of investment by enterprises from many industries?

The first thing that happened was that we always have computing power with us. Our smartphones are computers that are permanently connected to high-speed networks that can exchange data at rates that have never been seen before.

Second, the amount of data that is created with so many computers exchanging information grows every day. According to a study conducted by Forbes a few years ago, data creation rates were as high as 2.5 quintillion bytes of data, daily.[1] Data is the food feeding AI and ML algorithms, and we have enough data to rear several generations of algorithms.

To feed algorithms, we need computing power. General availability of powerful systems with much memory and CPU has driven the process of expansion of AI and ML to enterprises that some years ago only dreamed of this possibility. Based in observation, researchers differentiate three eras: the pre-deep learning (DL) era, the DL era, and the large-scale era.[2] The need for compute power has multiplied by several figures in the last 5 years.

The result is that we are now used to seeing recommendation systems when we shop online, stock forecasting when we invest, analyses of which are the most financially efficient companies, or chatting with robots that find the correct information for us.

IBM with IBM Z is fully aware of this reality, and being at the core of many companies around the world has renewed capacities to achieve the work of using AI and ML.

[1] https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=17869a8760ba

[2] Compute Trends Across Three Eras of Machine Learning by Jaime Sevilla et al. (March 2022), found at: https://arxiv.org/abs/2202.05924

© Copyright IBM Corp. 2022.

1.1 General aspects of using AI in your organization

As businesses continue to embed AI and ML technology into consumer products, government agencies, such as the US Consumer Product Safety Commission, recognize the need to develop a regulatory process to assess and analyze safety concerns and opportunities that are related to the usage of these technologies in consumer products.[3] The idea behind generalized use of AI and ML is solving problems and answering specific questions by using data that is generated by operations or data that is extracted from customer management.

In this section, we describe some of the key considerations when incorporating the usage of AI and ML in your enterprise. Many of these key considerations are not directly related to specific techniques but with the consequences of applying the techniques. Based on industry observations, here are the following eight core themes driving enterprises to leverage AI on the IBM Z platform:

- Democratizing AI
- Green AI
- Secure AI
- AI agility
- Decision velocity
- AI-infused software and AI-powered solutions (for IBM Z)
- Data governance and explainability
- Intelligent infrastructure

When we extract information from raw data, we must be careful about the process that we are following, which is why we are describing these eight themes. We do not go deep into the description of each theme, but it is important to understand fully the role of each them driving the adoption of AI and ML solutions within industries and businesses.

1.1.1 Democratizing AI

When we think about AI democratization, we find two main characteristics that rise above the rest: democratization as general availability of AI and ML for the public, and the ability for enterprises to create their own services and models without deep knowledge.

General availability of AI and ML

Companies are deploying several kinds of AI-infused applications for clients to use, such as recommendation systems, self-driving transports, customer care chats, and healthcare diagnostic systems. This paper described a few examples of these applications.

The way that you make AI and ML a product of general usage is important because it might generate issues such as unfairness in the data analysis or several types of bias or problems in explaining how a decision was assessed by an algorithm. We describe the most important issues in 1.2, “Desirable AI attributes” on page 5.

AI can also be used in a way that can generate social impact and compensate for the lack of equilibrium in human behavior, as the IBM Data Science and AI Elite Team demonstrated with its Incubator for Social Impact. Other examples apply to education, climate change mitigation, and public sector management.

Even if they are not intended for the public, functions such as auto AI, which is intended to lower the steep learning curve of creating models, are available from vendors for developers to innovate with data at a faster pace.

[3] US Consumer Product Safety Commission Report: AI and ML in Consumer Products

The key for democratization of AI to succeed is good data management in all of its stages. Data fabric is the best working framework that we have now because it establishes the path for governance and quality of data (essential to support good modeling with the needed attributes) from wherever it comes. With the amount of data that is created and stored on IBM Z, it is necessary to provide good, secured access.

1.1.2 Green AI

As the volume of data that is available continues to increase, it is logical to assume a resulting rise in data centers. Unfortunately, data centers are major contributors to an enterprise's carbon footprint due to the massive amounts of electricity that are required to power them. When this electricity is generated from non-renewable energy sources, like fossil fuels, they release greenhouse gases into the atmosphere and contribute to the global climate crisis. Through the combination of increased consumer demand for energy-efficient business policies and practices, more widespread adoption of clean energy and sustainability legislature and regulations, and the expectation of power requirements continuing to increase, business leader's are prioritizing efficiency and focusing on reducing their carbon footprint.

It might come as a surprise that there is a large carbon footprint that is associated with the resource-intensive tasks in AI, such as the training and running of models. The amount of computing power that is needed to leverage AI technology grows every day, with calculations stating that computing power doubles every month.[4] This situation adds even more pressure to optimize AI in a way that reduces your enterprise's carbon and electric footprint. *Green AI* is the training and deploying of AI models in a carbon and electricity footprint-optimized manner.

Leveraging AI applications and frameworks running on IBM Z can help your enterprise achieve energy-efficiency goals and green AI with the added benefit of enabling uncompromised model accuracy. This enhanced model accuracy is achieved through scheduled, periodic re-evaluations of newly available data to monitor the accuracy of the model over time, with the ability to receive alerts when performance degradation is observed. One critical benefit contributing to green AI that is enabled by the IBM Z platform is the ability to conduct resource-intensive model training on your platform of choice and still have the ability to deploy that model on IBM Z. By performing training on a platform other than IBM Z, for example, on a public cloud, private cloud, or on-premises, you can achieve reduced energy consumption while maintaining the ability to operate the model within transactional applications on the platform that is optimized for AI inferencing. Additionally, with the first-in-industry IBM z16™ integrated AI-accelerator, you can achieve low-latency, high-throughput inferencing of all transactions in real time that is supported by open-source tools and frameworks.

Green AI is not necessarily an industry-specific theme. Instead, it is a pervasive pattern that is observed across industries with use cases spanning from energy-efficient training to image and natural language processing (NLP).

[4] Green AI by Roy Schwartz et al. (2019), arXiv:1907.10597v3.

1.1.3 Secure AI

Secure AI consists of providing industry-leading protections for sensitive personal data that is leveraged in AI models. Secure AI includes the following considerations:

- Data privacy

  It has been said that data is the new oil. Certainly, AI algorithms need data to live. However, some data lives under regulated conditions, especially if it is personal data. In Europe, there is the General Data Protection Regulation (GDPR) that clearly states that personal data is owned by the person that is affected. Other regulations might apply. From this perspective, IBM Z is where much of this type of data is and where it must be (see 2.2.2, "Data gravity" on page 12 and 2.4.2, "Encryption and compression dynamically and at rest on IBM Z" on page 17). Face recognition is also a technology where privacy is a must.

- Anonymization

  Sometimes, it is necessary to use anonymization tools that preserve data types and formats and still are useful for training some AI algorithms.

In fact, both considerations are closely related: AI can be a powerful tool to re-identify and de-anonymize data because of the ease of finding patterns that AI provides. We must ensure that we use all means that are possible to provide a good level of privacy and respect for the data. Additional considerations include the following ones:

- Input validation

  When we feed an algorithm with data, we must ensure good and fair results because they depend on what the input data looks like. Knowing what the right data is and what format is needed is essential to achieving what we want.

- Encryption

  To enhance the results that are needed when an algorithm works with encrypted data, the use of homomorphic encryption is a good technique. IBM Z provides the IBM Fully Homomorphic Encryption Tool Kit that can be used within z/OS Container Extensions (zCX) containers on z/OS or in Linux for IBM Z. With this solution, we can provide accuracy in algorithm training and data privacy concurrently. IBM Z can provide a full range of solutions to protect data (see 2.4.2, "Encryption and compression dynamically and at rest on IBM Z" on page 17).

- Robustness

  If we are making AI and ML a key area to assess decisions and help understand the information behind data that is generated in companies, we must have models that work consistently and give the same result with the same input. Some of the dangers we face are data poisoning when training or feeding models, generation of adversarial examples to fool DL models, or extraction of data from features. Keeping models and data in a secured environment, such as IBM Z, is helpful to avoid many of these problems.

1.1.4 AI-infused software solutions for IBM Z

As AI and ML research advances, we see more use cases that are applied to every area of technology: autonomous vehicles, robotics, industry automation, or city management, among others. The IT industry is not oblivious to this trend and provides much benefit in some areas. Some of the tools that IBM provides include the following ones:

- AI operations (AIOps) and intelligent infrastructure

  AIOps provides a set of tools that are based on AI and ML that can help with problem identification, isolation, and resolution by analyzing data that is generated by IBM Z in logs, System Management Facility (SMF) records, and other structured and unstructured data. IBM Operational Log and Data Analytics (IZLDA) is intended for this case. It helps you to detect anomalies in the system as they occur, and tries to resolve them before running into trouble.

- In-line transaction scoring

  Core business applications running on transactional managers like CICS, IBM WebSphere® Liberty, or IBM Information Management System (IMS) now can send requests inside transactional time to IBM Watson® Machine Learning for z/OS engines. You can enrich your core applications with AI and ML models that can assess or score, in real time, and send back the results to the transactional managers.

- Systems optimization

  IBM Z Anomaly Analytics with Watson (ZAA) 5.1 is intended to work with the main subsystems running in IBM Z like IBM Db2, IBM CICS, and IBM MQ. It helps in AIOps for transactional managers and integrates with Watson AIOps on IBM Cloud Pak to correlate events from the rest of the enterprise. Also, you can find specific software to analyze performance and find anomalies for Db2 with IBM Db2 for z/OS Artificial Intelligence.

1.2 Desirable AI attributes

The movement to start using AI and ML in an enterprise is an effort that requires much work from different departments. We have seen that, depending on the area that you want to infuse with AI, using AI and ML might involve many different profiles, such as data scientists, data engineers, data architects, systems administrators, DBAs, and business lines. All these people working to deploy and use AI inside the company must be aware of the final goal that they want to achieve. Some models are developed to profile customers, and others assign probabilities of events that might happen. There are many possibilities. There are some common properties that are shared by all of them. We describe a few of them in this section.

These properties or goals must be concrete and measurable, and you must act if they are not.

- Trustworthiness

  AI influences human decision making or sometimes even substitutes for it. Humans must not distrust or overtrust the decisions that are taken by an algorithm. It is necessary to find the point where none of these ends happen.

- AI agility with AI and hybrid cloud

  Project lifecycles must be quick enough to avoid model degradation. Data creation is mostly dynamic by nature, and models should be modified or retrained to maintain their accuracy. Adaptation to the changes of the infrastructure with the fabric keeping data manageable is another reason to address how models behave.

- Decision velocity with real-time AI

  Specifically in an enterprise case, AI and ML models are embedded during process or transaction executions. Performance of the models is understood as accuracy in calculations and the ability to contribute to a transaction without delaying it, which is important when you intend to process thousands of records each second. The new Telum processor, which is embedded in the IBM Z processors and specific to the IBM z16, offers this important property to AI models that are deployed on IBM Z.

- Bias embedded in models

  This important research topic is one where IBM and other providers are putting in much effort. We name a few concepts here because this matter is beyond the scope of this paper.

  Policies to avoid bias, together with toolkits and analysis of model behavior are good tools to build fairer AI and ML models. Some examples of embedded bias in models can be race, gender, color, religion, nationality, education level, and age. For more information about policy priorities to strengthen the adoption of testing, assessment, and mitigation strategies to minimize instances of bias in AI systems, see Mitigating Bias in Artificial Intelligence. For an extensible open source toolkit to help you examine, report, and mitigate discrimination and bias in ML models throughout the AI application lifecycle, see AI Fairness 360 toolkit.

1.3 Popular industries for AI

AI and ML are becoming pervasive technologies that eventually will expand into all homes through intelligent assistants like Amazon Alexa and Microsoft Cortana. But before that happens, there are several industries at the forefront of adoption today, leveraging AI for practical efficiencies and optimizations to better their business.

Innovation and usage of all types of data that are generated in a company is the driver for most enterprises to adopt an AI strategy. In this section, we describe some of the most widespread AI implementations and trends of today, which are based on modern business problems in the following industries:

- IT and consumer technology
- Banking and finance
- Healthcare
- Retail and commerce
- Insurance
- Transportation

IT and consumer technology

In 1.1.4, “AI-infused software solutions for IBM Z” on page 5, we describe using AI and ML inside of IBM Z, but it is just one example of what the IT industry is doing for all kinds of systems and applications. We are using AI to get better, more efficient systems, design better chips, and simplify software solutions. For more information about use cases for IT and consumer technology, see Chapter 3, “Artificial intelligence for IT operations use cases” on page 19 and Chapter 4, “Data and artificial intelligence operations use cases” on page 33.

Technology companies offer AI-based solutions such as intelligent assistants that use neural networks, automatic translation services (again, neural networks), face recognition services, systems automation and management, correction of grammar as you type, and others.

Banking and finance

This industry was one of the first ones to start applying all types of AL and ML technologies, and it invests much money in finding use cases. The most well-known cases for this industry are fraud detection and using AI and ML for credit assessment. For more information, see Chapter 6, “Financial services sector use cases” on page 51.

Healthcare

Healthcare was one of the first industries where AI was applied and appeared in the media announcing new ways of diagnosis and aids to medical staff. Image recognition, design of medicines, and management of documentation are some of these use cases. For more information, see Chapter 8, “Healthcare sector use cases” on page 65.

Retail and commerce

The features that were developed for this industry are not apparent. The greatest star of AI and ML technologies for retail is recommendation systems. Under the sentence of “others also bought”, there is an algorithm that finds what related objects or services that you might need. Personalization of prices is also extended when buying online. For more information, see Chapter 9, “Retail and insurance sector use cases” on page 71.

Insurance

Chat bots, customer churn, personalized offerings, and risk assessment are some of the AI functions that this industry uses regularly. Claims processing is described in Chapter 9, “Retail and insurance sector use cases” on page 71, and chat bot implementations are described in Chapter 5, “Use cases for artificial intelligence in chatbots” on page 47.

Transportation

One of the most incredible things in futuristic films is autonomous driving. It is not strange to read about this AI miracle in the news. Some other uses of AI for transportation are maintenance prediction for complex systems such as planes or trains, route planning, or fleet organization.

2

Artificial intelligence, machine learning, and data on IBM Z

The role of IBM Z in the context of artificial intelligence (AI) and machine learning (ML) is essential to provide digital transformation, micro-services, and hybrid cloud solutions to an analytics- and AI-driven company. To become an analytics- and AI-driven company and use data fully, an information architecture (IA) is a mandatory but often neglected step. An IA for AI helps organizations to standardize and make clearly defined and prescriptive decisions, which are especially helpful in today's hybrid cloud world. This IA should reflect a modern and open design, enabling organizations to work in a flexible environment by using any tool running on any cloud or infrastructure to establish a solid data foundation for an analytics and AI world.

IBM Z technology has proven to be adaptable throughout its history and continues to introduce new capabilities to support critical engagement, analytics, and AI applications, including support for a hybrid cloud infrastructure. New capabilities demonstrate that IBM is continuing to invest in IBM Z, which is a vibrant, adaptable platform that can be depended on to differentiate an organization. IBM Z is the most resilient platform that is available for core transactional systems, and now it offers integration with IBM Cloud Pak for Data, which is the IBM strategic platform for data and AI.

The term *AI* refers to techniques that embody the purpose of mimicking human behavior. The purpose is to enable computers to perform tasks that are regarded as uniquely human, that is, processes that require intelligence. ML is a subfield of AI techniques that is sometimes referred to as classical or non-deep ML. It focuses on the usage of statistical methods for problem solving by recognizing patterns in the data and uses that information to make some predictions.

Traditionally, when you think of running applications and services on the IBM Z platform, you think of doing so within z/OS. Being an operating system, z/OS offers significant benefits, including a secure environment for high availability (HA) enterprise workloads. Within z/OS are z/OS Container Extensions (zCX) to run Linux-based containers, such as Docker, within a z/OS environment. This capability also allows the deployment of Linux on IBM Z applications as Linux based containers within a z/OS system to directly support workloads with an affinity to z/OS.

© Copyright IBM Corp. 2022.

Also, running these containerized applications on z/OS enables you to use existing z/OS skills, streamline your DevOps environment, and retain the benefits of z/OS qualities of service, such as security and HA. Compounding these benefits is the ability to leverage Red Hat OpenShift Container Platform on IBM Z, providing an enterprise Kubernetes-based platform to develop, deploy, manage, and run new containerized applications. When running Red Hat OpenShift Container Platform on IBM Z, you can leverage the data gravity of the IBM Z platform through the colocation of containerized applications with traditional workloads to optimize latency, response times, deployment, security, service, and costs. Additionally, you can achieve economic and operational efficiencies, business continuity with HA and disaster recovery (DR), and vertical solutions that used heavily in industry-specific use cases, like dealing with scalability and peak workloads in retail.[1]

2.1 From AI experimentation to business at scale

One of the many benefits of implementing AI solutions on IBM Z is the broad range of offerings that are available on the platform to support each stage of your data and AI solution. The IBM Z infrastructure also offers a robust, secure, and scalable infrastructure for enterprise applications and transactions.

Because most business transactions of large enterprises are processed on IBM Z, performing the AI analytics and predictions on the platform positions those processes close to the source of the transaction, which is the data. This phenomenon of moving AI applications and frameworks close to the business data is called *data gravity*. Data gravity enables real-time analytics, leaves sensitive data in a secure data environment, and reduces latency and network traffic because the need for data movement is removed. Developments, such as the IBM Telum processor chip with an embedded AI accelerator, are positioning IBM Z as the premier platform provider for inference workloads. This hardware innovation speaks to the commitment and focus IBM made to support and enable AI and improve inferencing at scale.

2.2 Real-time business insights

For mostly IT architectures and the transaction systems, insight and decision-making processes were designed decades ago. Traditional architectural approaches that depend solely on data movement can impede the ability to rapidly adapt to changing business cycles and adopt new development methods.

To deliver modern applications, organizations facilitate and simplify access to relational and non-relational IBM Z transactional data and combine this data with data originating off platform. They access and update the system of record in IBM Z data through traditional APIs, such as SQL, and modern RESTful APIs (when combined with IBM z/OS Connect EE). They reduce the cost and delay of moving data to non-IBM Z platforms.

Data, its usage, and the insights it provides have a shelf life. Some decisions require the most current data and real-time insight (at the point of interaction or transaction) to improve decision-making in areas like fraud detection and up-sell or cross-sell efforts, and supporting real-time opportunities such as digital moments.

[1] https://public.dhe.ibm.com/software/dw/linux390/docu/RHOCP-reference-architecture.pdf

Separating the systems of record from modern systems of engagement can result in customer defections and an increased risk and lost opportunity for improved operational productivity. Digital opportunities are momentary. They must be based on the exact real inventory status to make offers to customers, which is different than traditional approaches that use extensive data movement. Doing nothing or merely maintaining these existing approaches opens the door for the competition to disrupt your business and attract your customers. New business with increased data volumes leads to higher resource usage and exacerbates a challenging issue. Throwing hardware at the problem just pushes off the problem to another day. A change in approach is needed. Organizations can address the challenges of their current architecture to better leverage data where it originates, close the data latency gap, and embed data and insight into business processes where appropriate.

2.2.1 Time-to-value of data

As the world becomes increasingly connected, digitalization, that is, the usage of digital technologies to transform business and manufacturing operations, is a key differentiator that enables companies to remain competitive. Digitalization promises lower costs, improved production quality, flexibility and efficiency, shorter response time to customer requests and market demands, and opens new and innovative business opportunities. Turning data into value is a critical success factor.

Take condition monitoring as an example, which requires analytics for manufacturing equipment to ensure productivity and availability of production anytime. Smooth operation of machines in production environments is the key. Early detection of upcoming faults is essential for optimized condition-based maintenance plans.

A typical manufacturer on their digitalization journey must address several avenues before they get return on investment and intended benefits. Typical challenges include how to handle the volume of required machine data and the data that is collected through connecting to automation systems or adding special sensing equipment. Then, this data must be stored, analyzed to build monitoring models, or used through AI-driven anomaly detection approaches.

Data is specific to operation domains coming with high-frequency, big volume, and industry domain-specific formats. This specificity is one factor in preventing IT groups from gaining insights into this data over the last decade. Another challenge is the global distribution of manufacturing or operations sites and their remote locations. Typically, manufacturing sites are placed close to resource supplies (for example, mines and water plants), or as far away as rural or residential areas. Internet bandwidths were not a driver in the past for a factory site, so available Internet bandwidth is often limited and used for outbound business transaction workloads.

The time-to-value of data is the notion that the more time that elapses between the time of data generation and the time that data is leveraged, the business value of the data decreases. The simple formula states that if capturing, analyzing, and acting on that information can be made faster, the value to the business increases. A single piece of data might provide invaluable insight in the first few seconds of its life, indicating that it should be processed rapidly, in a streaming fashion. However, that same data, when stored and aggregated over time alongside millions of other data points, can also provide essential models and enable historical analysis. Even more subtly, in certain cases, the raw streaming data has little value without historical or reference context. So, the real-time data is worthless unless the older data is available to it.

There are also cases where the data value effectively drops to zero over a short period. For these *perishable insights*, if you do not act on them immediately, you lose the opportunity. The most dramatic examples are detecting faults in power plants or airplanes before they explode or crash. However, many modern use cases such as fraud prevention, real-time offers, real-time resource allocation, geo-tracking, and many others also depend on up-to-the-second data.

After you have pipelines in place feeding your "data lake" (a traditional repository for semi-structured log and device data) from enterprise sources in a streaming fashion, you can move to real-time analysis, and even predictive analytics, seamlessly for little marginal cost.[2]

2.2.2 Data gravity

Today's business opportunities and questions demand answers in real time, requiring trusted analytic and AI insight from the most current enterprise data. Organizations can gain an advantage when their leadership finds a way to provide ready access to this data as part of a data fabric architecture. This approach reduces the time between insight and decision.

According to IBM studies, roughly 70% of enterprise data originates behind the firewall.[3] Leveraging that data where it originates aligns with the theory of *data gravity*, which suggests that data can be thought of as having mass and therefore attracting other objects. The more data that exists, the greater the mass, and the more likely that the data attracts other objects, such as applications, services, and other data.

However, the more the data is moved away from where it originates, the more outdated the insights become. Therefore, any action that is based on replicated data can never be real-time and might be suboptimal. Every copy of data has its unique cost, latency, and risk, and such risk can include data governance, security, and decision latency. Data movement impedes both data usage and time-to-insight, and it can undermine some of the value of many modernization efforts.

A data fabric approach attempts to address these challenges in three ways:

- Simplify data access.
- Push queries to where the data resides.
- Take advantage of intelligent metadata to facilitate and simplify data usage.

Data gravity is a compelling concept for organizations that want to bring computation to the data, that is, analyzing data where it originates.

You can access data at its source so that critical data-driven decisions can be made before an interaction or transaction completes and before your customer abandons their interaction with your organization. Developers can readily combine IBM Z data with other enterprise data sources to gain real-time insight, accelerate deployment of new web and mobile applications, modernize the enterprise, and take advantage of today's API economy. Your architectural strategy can ensure access to data across platforms and across multiple clouds, whether the data is structured or unstructured. Data virtualization is emerging as an exciting, cost-effective substitute for and augmentation to traditional data collection (incremental copy, data movement, and extract, transform, and load (ETL)). With data virtualization, you can access data where it originates to reduce the time and resources that are used to combine data from multiple systems. Less time and fewer resources can translate into savings. Leveraging data virtualization technology can support greater flexibility and agility, which are core to digital transformation.[4]

[2] https://insidebigdata.com/2016/04/08/why-time-value-of-data-matters/

[3] https://www.ibm.com/downloads/cas/QO6AM6PV

[4] https://www.ibm.com/support/pages/introduction-ibm-data-virtualization-manager-zos

2.2.3 Data movement

Data movement is the ability to move data from one place in your organization to another through technologies that include ETL, extract, load, and transform (ELT), and data replication and change data capture (CDC), primarily for the purposes of data migration and data warehousing. Data movement, data synchronization, and data replication are complementary methods of data integration. Together, they enable the ability to deliver fresh data to keep databases, data warehouses, big data, and cloud systems current.

Replication offers a dependable, low-impact method of creating an accurate and up-to-date copy of your single- or multiple-source data, which can be deployed to any person who needs access to it from wherever and whenever they work. Users can stay in control of replicated databases with flexible configurations of archiving and retention rules, data relocation, and storage.

Synchronization keeps replicated data fresh, so users and applications are working from the best information. Some companies update replicated databases with either a batch-oriented (pull) or real-time (push) configuration. For many relational databases, you can synchronize new data instantly with a capability called CDC. Comprehensive data movement and transformation capabilities might also be required to modernize and extend the IT portfolio. For example, you might need data movement tools to move data from earlier systems and platforms to cloud databases. ELT is a data integration process that combines data from multiple sources into a single data store. This capability is valuable to meet hybrid integration requirements, such as connecting and transforming earlier data sources to a data warehouse environment or moving data from transactional databases to a big data or data lake environment.

2.2.4 IBM Z hardware for transaction and data volume

Since its inception, IBM Z has always been light years ahead of its time with systematic updates. Originally in high demand because of its unique ability to run any application without modification, IBM Z remain prized for its unprecedented utilization rates, long-term cost advantages, security, scalability, and resiliency. Today, IBM Z works with open-source languages, databases, and development tools, runs in the cloud, and can be accessed from Windows, web, mobile, Internet of Things (IoT), and web service interfaces.

The mainframe has adapted and reinvented itself, staking and defending its claim as the most powerful business computer in the world. IBM Z is renowned for processing transactions and serving enterprise applications. It runs up to 19 billion encrypted transactions a day, provides 99.99999%[5] availability with resiliency and instant recovery, protects data at rest and in flight with security and data privacy, and simplifies life for developers through enterprise DevOps and cloud native development.

Each of the new launches of IBM Z has delivered an accumulation of important new capabilities to the platform in order for the platform to participate in and drive businesses' digital transformation. Capabilities that common today on IBM Z are Java, mobile enablement, APIs, and web enablement. New generations of IBM Z also lower cost of operations, improve the capacity to scale and drive new workloads, and make the platform more cost-efficient while delivering increased security and resilience.

[5] https://www.ibm.com/it-infrastructure/z/capabilities/transaction-processing

2.3 Accelerated approach to AI and ML with IBM Z

At its simplest form, AI is a field that combines computer science and robust data sets to enable problem solving. It also encompasses the subfields of ML and deep learning (DL), which are frequently mentioned with AI. These disciplines are composed of AI algorithms that seek to create expert systems, which make predictions or classifications based on input data.

Here are the types of AI:[6]

Weak AI Also called narrow AI or artificial narrow intelligence (ANI), this AI is trained and focused to perform specific tasks. Weak AI drives most of the AI that surrounds us today. *Narrow* might be a more accurate descriptor for this type of AI because it is anything but weak. Narrow AI enables robust applications like Apple Siri, Amazon Alexa, IBM Watson, and autonomous vehicles.

Strong AI Made up of artificial general intelligence (AGI) and artificial super intelligence (ASI), AGI, or general AI, is a theoretical form of AI where a machine would have an intelligence equal to humans. This AI would have a self-aware consciousness that can solve problems, learn, and plan for the future. ASI, also known as super intelligence, would surpass the intelligence and ability of the human brain. Although strong AI is still entirely theoretical with no practical examples in usage today that does not mean that AI researchers are not exploring its development. In the meantime, the best examples of ASI might be from science fiction, such as HAL, the superhuman, rogue computer assistant in *2001: A Space Odyssey*.

IBM has been a leader in advancing AI-driven technologies for enterprises, and it has pioneered the future of ML systems for multiple industries. Based on decades of AI research, years of experience working with organizations of all sizes, and on the learning experiences of over 30,000 IBM Watson engagements, IBM has developed the AI ladder for successful AI deployments.

2.3.1 Why AI cannot exist without information architecture

AI is about mimicking and improving the human function, that is, bringing human features to technology. In the consumer world, these features are mimicking speech, vision, and daily interactions. In enterprises, these features are mimicking and improving enterprise functions, such as logistics, marketing, finance, operations, and HR.

Enterprise AI is about solving sophisticated business problems in highly dynamic environments, which requires an understanding of defined use cases and starting points.[7]

AI is a journey that begins with data. Thus, AI cannot exist without an IA. Gaining business value and insights from data is not always easy. A traditional infrastructure is inadequate for AI workloads, and data silos make it difficult to get a holistic view of all your information, which limits the value of AI. Organizations are moving toward hybrid cloud to respond to evolving business needs. As data is increasingly distributed, it becomes a struggle to provide adequate protection and management. An infrastructure that was not built for AI and hybrid cloud is not flexible enough to respond to modern workloads and demands without adding complexity.[8]

[6] https://www.ibm.com/cloud/learn/what-is-artificial-intelligence
[7] https://www.ibm.com/blogs/think/2018/02/ibm-ai-ladder/
[8] https://www.ibm.com/downloads/cas/XNXLLX6D

IBM Cloud Pak for Data is the only comprehensive offering in the industry that delivers an AI IA on any cloud as an integrated infrastructure, network, storage, and server platform (including specialized servers) that can be deployed in a set of predefined configurations on various cloud platforms. IBM Cloud Pak for Data, as an integrated collection of data and analytics microservices, is built on a cloud-native architecture that enables users to collect, organize, and analyze data with unprecedented simplicity and agility and leverage AI infused analytics applications, all within a governed environment.

2.3.2 Prescriptive approach to AI with the IBM AI Ladder

The best AI is built on a foundation of data that is collected, organized, and analyzed carefully, and then infused into the business. This foundation should be open, flexible, and allow access to data of every type regardless of where it is. Every successful AI project goes through a multi-step process that starts with having the correct data and progresses to using AI broadly. As companies modernize, they seek to provide an architecture that will propel them into the future. The journey to AI is about moving data from ingest to insights with an IA that can easily be integrated throughout the organization. Each part of the AI ladder must provide integration to the entire journey. Starting a project on one part of the journey is fine, but it is critical to ensure that the project considers an overall IA for AI to optimize resources and modernize your infrastructure for expanding AI workloads.[9]

Collecting

Data is the fuel that powers AI, but it can become trapped or stored in a way that makes it difficult or cost-prohibitive to maintain or expand. You must unleash that data so it can expand from edge to core to public cloud within a simple and cost-efficient infrastructure.

Organizing

AI is only as good as the data on which it relies. Businesses must fully understand what data they have so they can leverage it for AI and other organizational needs, including compliance, data optimization, data cataloging, and data governance.

Analyzing

Analysis is critical to the AI journey, and it must provide high performance for fast analysis and seamless connection to both the data lake and the storage catalog. Organizations must plan for issues beyond the deployment of AI, that is, you must build AI infrastructures that give you confidence in your data and enable you to access it wherever it is.

Infusing

Business challenges can become an opportunity to explore, understand, predict, and bring an AI infrastructure to every organization.

Modernizing

A solid IA is the foundation for AI and hybrid cloud. Modernizing your infrastructure means building a foundation that takes advantage of cloud-native technologies and drives AI throughout the organization. IBM Storage for data and AI delivers the flexibility that is needed to respond to AI workloads. It supports integration with Kubernetes and Red Hat OpenShift Container Platform, making it easier to deploy cloud-native applications.[10]

[9] https://z7solutions.com/partners/ibm/

[10] https://illuminatesolutions.net/ai/Beginner/Building-a-Strong-AI-Foundation.aspx

2.4 Intelligent platform and infrastructure

A hybrid cloud computing environment is created by an infrastructure that connects on-premises, private cloud, and public cloud services, enables the portability of applications, and provides orchestration and management capabilities between them. A hybrid cloud infrastructure provides the flexibility and cost-efficiency that is required to become a digital enterprise that, when combined with AI capabilities, can unlock competitive industry innovations to help achieve business objectives.

The IBM hybrid cloud approach is composed of Red Hat OpenShift and Red Hat Enterprise Linux, which brings together hybrid cloud and core open source technologies for flexible application development. Red Hat OpenShift provides the enterprise-ready Kubernetes container platform for managing deployments, and Red Hat Enterprise Linux provides the operating system that enables the deployment of applications to hybrid cloud environments. Together, these technologies enable a hybrid cloud platform that provides an environment to leverage integrated offerings and capabilities, such as developer, security, and operations tools for enterprise production use.

Building on the hybrid cloud platform is hybrid cloud software. Hybrid cloud software provides the technology and capabilities to build, modernize, and scale applications more quickly to improve business agility and accelerate the delivery of enterprise applications. IBM provides hybrid cloud software offerings that are designed to run anywhere and integrate everywhere. These IBM offerings are packaged as IBM Cloud Paks and are built on Red Hat OpenShift. They facilitate a single control point for managing your hybrid IT environment. IBM Cloud Paks leverage pre-integrated data, automation, and security capabilities to accelerate application modernization across any cloud or IT infrastructure. Additionally, there is Red Hat Marketplace, which serves an open cloud marketplace of certified hybrid cloud software from IBM and IBM Business Partners, which makes it easier to build solutions and deploy them in container-based environments in public clouds and on-premises. This approach provides consistent visibility, governance, and automation across the entire hybrid cloud landscape, which bridges traditional virtual machine applications with new cloud-native container applications.

Not all hybrid cloud approaches are alike. Many cloud infrastructures run versions of open-source technology that are optimized and customized for their specific platform. Although this approach allows you some level or application portability, not all functions can be moved from one cloud vendor to another one, which means you can potentially be locked into a cloud platform that appears to run open-source software. Through its abstraction layer, the Red Hat OpenShift Container Platform provides application portability across any cloud that support the platform, so there is no lock-in.

The IBM Z platform includes several technologies that work seamlessly with Db2 for z/OS and other data while minimizing the need to move around data. Organizations wanting to make the most of their Db2 for z/OS data while keeping it on IBM Z should consider a Hybrid Transaction Analytical Processing (HTAP) approach to support the in-place analytical usage of data.

HTAP, which is a term that was devised by Gartner, refers to a system that supports both online transaction processing (OLTP) and online analytical processing (OLAP) within a single environment. The purpose of a transactional system is to maintain transactional service-level agreements (SLAs), and the system is dedicated and optimized specifically for this purpose. Analytics processing on the same platform as a transactional system can consume precious resources that are required by the operational applications.

HTAP allows organizations to modernize application infrastructure while decreasing data latency, cost, complexity and security risk, so that they can generate analytic insight from real-time data, at the transactional source, without moving that data off platform. Instead of generating analytic insight in downstream systems, insight is generated at the point of data origination, which means faster time to insight for better decisions and results.

2.4.1 Security, data privacy, and IT operations with AI

IBM interviewed hundreds of CIOs and CTOs and found that the challenges of managing complexity and change while trying to keep up with innovation cuts across CIO experiences. A CIO managing the balance between innovation and stability might have been investing for decades in getting a handle on this problem. But even with dozens of tools for monitoring and automation, less than half report an end-to-end understanding of these critical environments. Managing and processing thousands of incidents monthly, including sifting through daily logs for minor incidents and reducing major business impacting outages, is time-consuming and expensive. Market research firm Aberdeen estimates an outage cost about $260,000 per hour.[11]

With the digital transformation and shift to hybrid cloud architectures, a potential 100x increase in machine and human data output from these systems can be realized. It can take days to sift through reams of data for the golden signals[12] of monitoring, which leads to taking days to diagnose and fix an issue when every minute costs money with an unknown impact to customer perception.

Organizations are spending their highly skilled and scarce talent on putting out often preventable fires, which contributes to mounting burnout, attrition, and lost opportunity to build for the future. Furthermore, with the looming retirements of many IBM Z administrators, along with an increasing growth in IBM Z workloads, organizations are trying to find tools to optimize both costs and skills. With IBM Watson AI operations (AIOps), time to resolution is significantly shortened by the application of AI to accelerate the problem resolution process.

The main differentiator of Watson AIOps is its core capabilities of discovering hidden insights and connecting the dots across sources, including the ones originating on IBM Z. Watson AIOps specifically aligns to those problems of complex system failures (dark debt[13]) with natural language processing (NLP).

2.4.2 Encryption and compression dynamically and at rest on IBM Z

The proper application of encryption can significantly reduce the losses from data theft. By the end of 2019, securing data in transmission over the Internet had greatly improved, with nearly 95% of website-focused traffic encrypted. For more information about the intersection of confidentiality, technology, and transmission, see *Maintaining Data Protection in a Hybrid, Multi-Cloud World*.

IBM Z pervasive encryption enables extensive encryption of data in-flight and at-rest to substantially simplify encryption and reduce costs that are associated with protecting data and achieving compliance mandates. The IBM Z platform is designed to provide pervasive encryption capabilities to help you protect data efficiently in the digital enterprise with up to 19 billion fully encrypted transactions per day.[14]

[11] https://www.machinemetrics.com/blog/the-real-cost-of-downtime-in-manufacturing#:~:text=According%20to%20Analyst%20firm%20Aberdeen,much%20as%20%24260%2C000%20an%20hour!

[12] https://www.ibm.com/garage/method/practices/manage/golden-signals/

[13] https://www.bmc.com/blogs/dark-debt/#:~:text=Derived%20from%20the%20physics%20term,planned%20for%20or%20defended%20against

[14] https://www.ibm.com/blogs/systems/no-worries-with-pervasive-encryption/

Data set encryption enables encryption of files in bulk through the access method, as opposed to encrypting a single field or row at a time. z/OS data set encryption is designed to offer high throughput, low-cost encryption. This type of encryption is intended to be more accessible to the organization than many other forms of encryption. For example, z/OS data set encryption is transparent to the application, which requires no changes to application code. z/OS data set encryption enables customers to encrypt data at course scale without performing data identification and classification first.[15]

The implementation of security across the enterprise in a hybrid multi-cloud environment can be leveraged through the comprehensive IBM zSecure Suite and applications and IBM Z foundational security capabilities, such as IBM Data Privacy Passport, IBM Z pervasive encryption, and IBM Resource Access Control Facility (IBM RACF®).

[15] https://www.ibm.com/support/pages/system/files/inline-files/DataSetEncryptionFAQzOSV2R2.pdf

3

Artificial intelligence for IT operations use cases

Today, digital transformation has drastically changed the way that we do business. The exponential growth and complexity of business applications has imposed unprecedented pressure on IT operations and places high demand on agile support. Industry leaders have their visions set, and they are actively looking for a competitive edge by using artificial intelligence (AI) technologies to accelerate IT operational excellence to support their business needs. Some of the more common challenges that are faced in IT operations include the following ones:

- Digital transformation driving exponential business growth

  Transactional workloads have grown unpredictably high from millions to *billions per day*, making it challenging to predict the right amount of capacity that is required while still maintaining the lowest operating cost.

- Complexity of business applications across hybrid environments

  This complexity imposes significant challenges on delivering new applications on time, across platforms, and across public and private multi-cloud environments. Adding to this challenge is the need for businesses to support and meet stringent service-level agreements (SLAs).

- High demands by unpredictable business events

  As a result of recent global events, such as the COVID-19 pandemic, businesses have been experiencing extreme surges in demand for online products and services. This increase in business demand requires much more dynamic resource provisioning and timely support responses of the business.

- Dynamic changes in complex business applications

  Such changes often require changes to IT infrastructure to support the new business logic and processes across the hybrid environment. These frequent changes impose high pressure upon CIOs and IT managers to sustain application and system resilience. Holistic insight and superior intelligence to manage systems and application health has become a critical success factor for managing IT systems to support 24x7 availability.

© Copyright IBM Corp. 2022.

- IBM Z knowledge and skill gaps

  Subject matter and technical experts for IBM Z are becoming increasingly harder to find. With experienced talent reaching retirement and an unbalanced ratio between higher transaction volume and a scarce next-generation talent pool, IBM Z skills are becoming less prevalent and more difficult to recruit. Industry visionary leaders are looking to leverage automation to reduce requirements for some of the manual, labor-intensive work on the platform. AI and machine learning (ML) have become must-have alternatives for achieving IT operational excellence to deliver business outcomes for the long term.

Like fraudulent cases in the financial services industry, IT systems performance and health posture might be confronted by resource intruders or bad actors lurking in the systems or inside of application programs. As a result of these bad actors, your systems likely experience sporadic slow responsive issues, or unknown risks, until they degrade or damage the core fundamentals and lock out the entire system. The following use cases are extracted from recent rapid proof of concept (PoC) projects with clients to illustrate the compelling opportunity AI provides to mitigate the challenges that are experienced in modern, complex IT operations.

3.1 Anomaly analytics and detection on IBM Z

Enterprise clients implemented alert monitoring and automated remediation processes for decades to support mission-critical business systems. However, with rapid digital transformation and the increasing complexity of hybrid application workloads, most companies face challenges in preventing unpredictable outages. For example, a global financial services company experienced a major outage that took more than 3 weeks to resolve. Two months after the first outage, there was yet another significant performance degradation persisting for more than 4 weeks, and it was not clear what exactly the root cause was and how to prevent another unpredictable disruption.

Working with IBM, client leaders look for innovation with AI, ML, and automation to support business growth in modern agile efficiency.

3.1.1 The solution: Auto-detecting anomalous system behavior by using ML

For the discussion about how the company can leverage an AI-infused solution to solve complex problems, the clients had the following questions:

- Can AI-ML give early warning before the outage occurs?
- Can AI-ML detect anomalies in advance?
- Can AI-ML predict potential impact or risk level?
- Can AI-ML provide prescriptive recommendations?

One approach to an AI-infused solution is to automate the process of anomaly detection for the complex and huge bulk of system management data to achieve system resiliency and maintain SLAs. You can leverage IBM Z Anomaly Analytics with Watson (ZAA) to apply IBM Watson Machine Learning for z/OS (WMLz) technology to build a historical model of normal operations for your environment. ZAA processes standard System Management Facility (SMF) data to train a model as a baseline, which reflects normal operations that are based on key performance indicator (KPI) values over varying times of the day and across multiple days of the week.

Furthermore, ZAA performs continuous scoring with current real-time SMF data to auto-detect anomalous behavior in your IBM Z environment. ZAA analyzes the data on IBM Z without the latency and impact that is related to streaming huge volumes of data to third-party analytics platforms, which might cause network congestion, which helps enterprises avoid costly incidents and service disruptions. When anomalous behavior is detected, alerts can be generated optionally to notify IT operations about the anomalies.

Figure 3-1 illustrates how ZAA leverages ML-enabled processes to break through some of the human obstacles and automate anomaly detection efficiently without depending on manual efforts. The features of ZAA include the following ones:

1. Load Historical Data for Model Training: Months of historical raw SMF data is processed by ZAA to train the historical normal baseline model by using Watson Machine Learning for z/OS.
2. Scoring: ZAA processes real-time SMF metrics, launches a continuous scoring process and compares live metrics against the historical model, scores anomalies on a scale 0 - 100 through an internal algorithm, and stores the resulting score cards in the ZAA Enterprise Data Warehouse (EDW).
3. Visualize Anomaly Scores: The anomaly scoring results are visualized on the ZAA Problem Insight GUI, and then saved in the Db2 EDW where historical scorecards can be retrieved and viewed online through ZAA user interface.

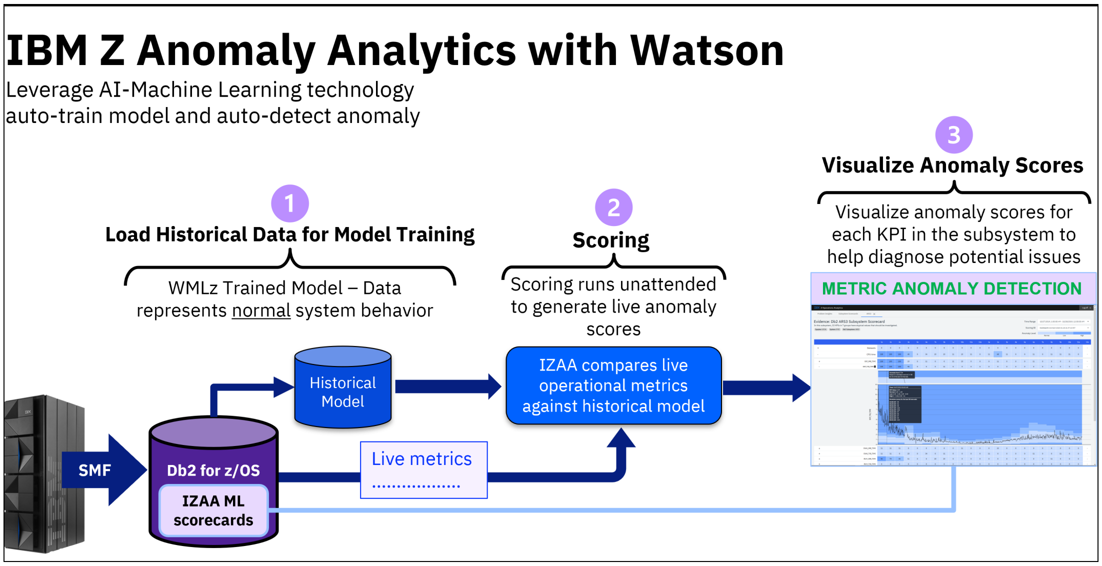

IBM Z Anomaly Analytics with Watson
Leverage AI-Machine Learning technology
auto-train model and auto-detect anomaly
1
Load Historical Data for Model Training
WMLz Trained Model – Data represents normal system behavior
2
Scoring
Scoring runs unattended to generate live anomaly scores
3
Visualize Anomaly Scores
Visualize anomaly scores for each KPI in the subsystem to help diagnose potential issues
METRIC ANOMALY DETECTION
Historical Model
IZAA compares live operational metrics against historical model
SMF
Db2 for z/OS
IZAA ML scorecards
Live metrics

Figure 3-1 Architecture for IBM Z Anomaly Analytics metric-based anomaly detection

> **Note:** To use continuous scoring, you must configure IBM Z Common Data Provider to stream SMF data sources from ZAA into the EDW. Additionally, continuous scoring jobs run at an interval that is specified by the configuration property `IZOA_SCORE_FREQ` in the `izoaml.config` file. The `IZOA_SCORE_FREQ` value is specified in seconds and must be greater than or equal to 60.
>
> For more information, see *IBM Z Anomaly Analytics with Watson 5.1: User Guide*.

ZAA can perform scoring in batch mode or continuous real-time mode. With batch scoring, it performs analysis by ingesting historical SMF data. In continuous mode, it launches a long-running scoring process that analyzes real-time SMF data at configurable time intervals. ZAA indicates observed anomalous behavior through calculated anomaly scores for the near real-time operational data from the systems that it is monitoring. It produces scoring results every minute the data is available, and rolls up the highest score to the KPI group to which the score belongs every 5 or 10 minutes, as specified.

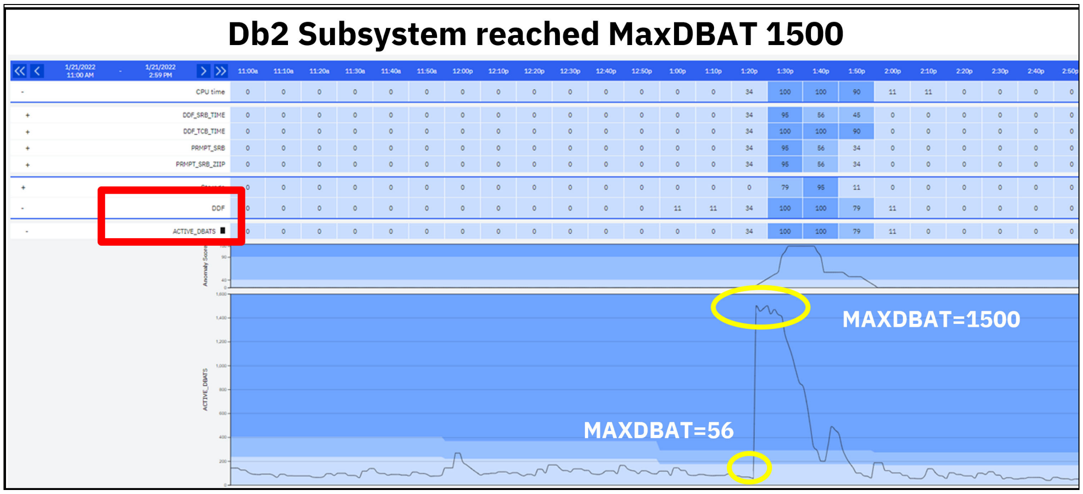
These scores are sent to subsystem-specific analysis routines for further analysis. In our example, anomaly scores are stored in an IBM Db2 database, where the Problem Insights server can interact with the database to retrieve and visualize near-real-time and historical analysis results. If the anomaly score exceeds a specified threshold, an anomaly event record is created, and analysis results are made visible in the Problem Insights interface through the **Problem Insights** tab, which is provided by the Problem Insights server. Optionally, you can have these anomaly event records forward to an event management system for further cross-platform event correlation to determine whether there are potential system-damaging or application-performance threats.

> **Note:** For metric anomaly detection, ZAA includes IBM Insight® Packs for IBM CICS Transaction Server for z/OS, IBM Db2 for z/OS, IBM Information Management System (IMS) Transaction Manager, and IBM MQ for z/OS. Depending on the insights that you want, you *must enable the appropriate Insight Pack* in the appropriate server:
>
> 1. The Problem Insights server so that it can provide visualizations of subsystem-specific insights.
> 2. The ML software system so that it can provide z/OS and subsystem-specific ML capabilities.
>
> For more information, see *IBM Z Anomaly Analytics with Watson 5.1: User Guide*.

Anomaly score ranges that are produced by ZAA are presented in the Problem Insights interface with the following associated anomaly levels:

- *Normal* anomaly level: Anomaly score range of *0 – 39*
- *Low* anomaly level: Anomaly score range of *40 – 89*
- *High* anomaly level: Anomaly score range of *90 – 100*

For this problem use case, ZAA produced scorecards representing the normal, low, or high anomalous posture. We quickly discovered that all four IBM Db2 data-sharing members reached the maximum threshold of 1500 active database access threads (ADBATs), as shown in Figure 3-2 on page 23. In the Db2 system configuration DSNZPARM, the administrator predefines maximum concurrent DBATs (MaxDBATs), and in this case the ceiling number is 1500. When the active DBATs reach MaxDBATs, any upcoming remote Db2 requests are immediately put in DBAT_Queued.

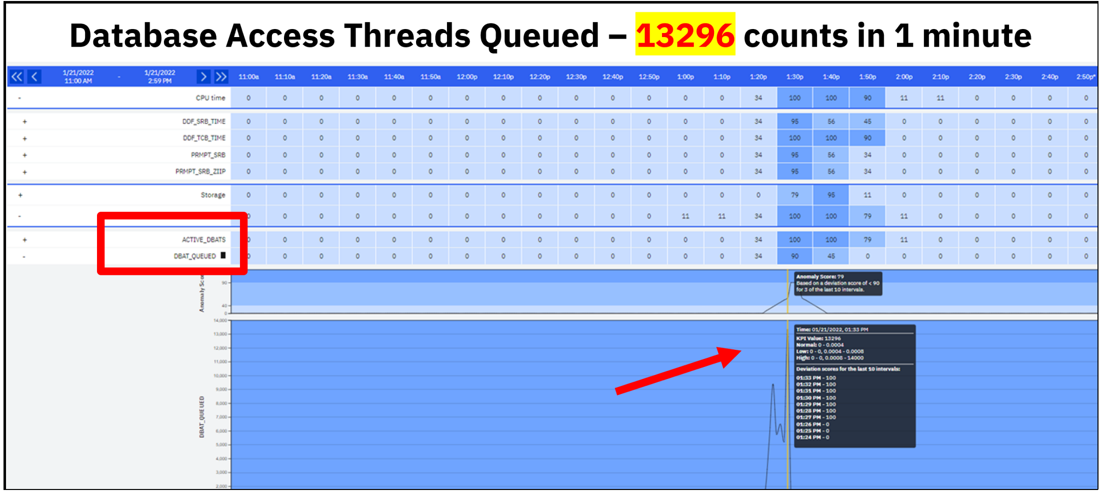


Db2 Subsystem reached MaxDBAT 1500

	1/21/2022 11:00 AM - 1/21/2022 2:59 PM	11:00a	11:10a	11:20a	11:30a	11:40a	11:50a	12:00p	12:10p	12:20p	12:30p	12:40p	12:50p	1:00p	1:10p	1:20p	1:30p	1:40p	1:50p	2:00p	2:10p	2:20p	2:30p	2:40p	2:50p
-	CPU time	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	100	100	90	11	11	0	0	0	0
+	DDF_SRB_TIME	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	95	56	45	0	0	0	0	0	0
+	DDF_TCB_TIME	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	100	100	90	0	0	0	0	0	0
+	PRMPT_SRB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	95	56	34	0	0	0	0	0	0
+	PRMPT_SRB_ZIIP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	95	56	34	0	0	0	0	0	0
+	Storage	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	95	11	0	0	0	0	0	0
-	DDF	0	0	0	0	0	0	0	0	0	0	0	0	11	11	34	100	100	79	11	0	0	0	0	0
-	ACTIVE_DBATS ■	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	100	100	79	11	0	0	0	0	0

Figure 3-2 IBM Z Anomaly Analytics detects total maximum Db2 access threads reached

During that 5-minute range, there also were scores indicating a high anomaly level for the DBAT_Queued indicator, as shown in Figure 3-3. In this example, more than 13,000 requests were queued up per minute in a selected Db2 subsystem to await the release of an available DBAT thread. This anomaly detection communicates a workload surge on the system, in this case, the total number of queued threads spiked from zero to 13,296 within a 1-minute interval.



Database Access Threads Queued – 13296 counts in 1 minute

	1/21/2022 11:00 AM - 1/21/2022 2:59 PM	11:00a	11:10a	11:20a	11:30a	11:40a	11:50a	12:00p	12:10p	12:20p	12:30p	12:40p	12:50p	1:00p	1:10p	1:20p	1:30p	1:40p	1:50p	2:00p	2:10p	2:20p	2:30p	2:40p	2:50p*
-	CPU time	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	100	100	90	11	11	0	0	0	0
+	DDF_SRB_TIME	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	95	56	45	0	0	0	0	0	0
+	DDF_TCB_TIME	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	100	100	90	0	0	0	0	0	0
+	PRMPT_SRB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	95	56	34	0	0	0	0	0	0
+	PRMPT_SRB_ZIIP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	95	56	34	0	0	0	0	0	0
+	Storage	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	95	11	0	0	0	0	0	0
-		0	0	0	0	0	0	0	0	0	0	0	0	11	11	34	100	100	79	11	0	0	0	0	0
+	ACTIVE_DBATS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	100	100	79	11	0	0	0	0	0
-	DBAT_QUEUED ■	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	90	45	0	0	0	0	0	0	0

Anomaly Score: 79
Based on a deviation score of < 90
for 3 of the last 10 intervals.

Time: 01/21/2022, 01:33 PM
KPI Value: 13296
Normal: 0 - 0.0004
Low: 0 - 0, 0.0004 - 0.0008
High: 0 - 0, 0.0008 - 14000
Deviation scores for the last 10 intervals:
01:33 PM - 100
01:32 PM - 100
01:31 PM - 100
01:30 PM - 100
01:29 PM - 100
01:28 PM - 100
01:27 PM - 100
01:26 PM - 0
01:25 PM - 0
01:24 PM - 0

Figure 3-3 IBM Z Anomaly Analytics auto-detects spiky DBAT queued-up requests

Another valuable observation that is obtained from the ML-produced scorecards is that all these Db2 subsystems encountered high anomalies in latch suspensions. Usually, this situation is an indicator that CPU resources might be in contention or there might be a shortage. On the contrary, in this problem case the CPU Group KPI indicated that most Db2 address spaces were not starving for CPU, with one exception: the Db2 Distributed Data Facility (DDF) CPU_TCB time.

The symptoms led to two questions:

1. What was causing those remote Db2 requests to flood in and queue up so quickly? We seldom reached the maximum number of active concurrent threads.
2. Why were other address spaces not suffering from a CPU shortage? Does this situation imply that this Db2 subsystem was not acknowledged by the z/OS dispatcher or IBM z/OS Workload Manager (WLM) for its CPU needs?

This anomalous pattern of DDF_TCB time prevailed among other Db2 data sharing members. The team decided to verify with the z/OS administrator how the WLM services were defined in the WLM policy definition. It is possible that with the new substantial workload changes, the z/OS administrator was not notified to refine the WLM policy to ensure that the WLM was realigned with more precise priority and importance settings for each service class.

In addition to ML auto-detecting key performance metric anomalies, you can use the *log analytics tool* and its relevant content search with efficient index functions. IBM Z Operational Log and Data Analytics (IZLDA) makes it possible to inspect millions of system log messages with a few clicks.

Figure 3-4 illustrates how SYSLOG message content analytics can be used to speed up trend analysis and find hot spots. For example, when quantifying an anomalous trend out of multiple subsystems who suffered from excessive lockouts and lock escalation symptoms, the task can be labor-intensive among myriads of system and application logs. Those log message contents often reveal the application program or transaction names, a database object name for quick victim identification waiting for a table space, or an index page.

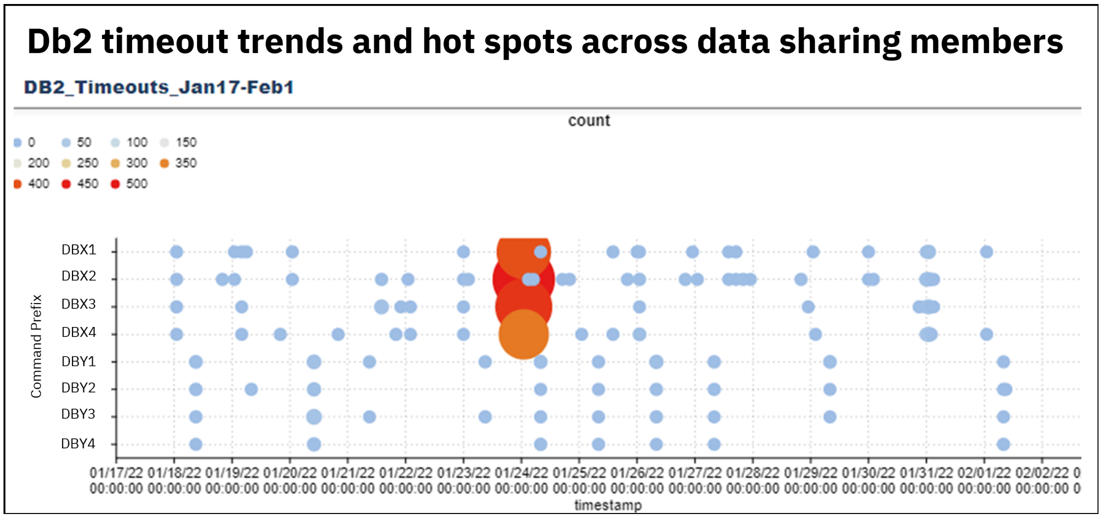


Figure 3-4 SYSLOG message content analytics highlights anomalous timeout trends

In our example, you might wonder why the maximum threshold of Db2 access threads were reached. When attempting to identify the root cause of an anomaly, you also look at whether the SYSLOG reveals more evidence. For example, in this problem scenario, the analytics identifies a repetitive warning message ID "`DSNX908I`" that a Db2 Stored Procedure was terminated frequently due to exceeding the CPU resource limitation. This information might be valuable for your root cause identification efforts.

Customers often ask if AI-ML can help detect unusual behavior in advance. In this use case, it was discovered in the days before the outage that IBM MQ applications were experiencing errors due to the maximum connection limit being reached. The IBM MQ bridge to CICS transactions was queued up while Db2 threads encountered numerous deadlocks and timeouts. An hour earlier, the network system indicated that the name server was unresponsive, the hardware Coupling Facility (CF) issued 80% storage full alerts, and IMS business transactions were experiencing problems with IMS Connect error messages. All those symptoms might provide relevant information and critical insights for determining the root cause that might have been otherwise missed.

With the usage of ML-enabled anomaly analytics, a data analyst can quickly identify trends with holistic views by using preprocessed scoring and auto-comparison results against the historical model to detect unusual performance behaviors. The resulting scorecards allow you to visually observe pervasive trends earlier than the traditional manual monitoring and evaluating process, which often rely on subject matter experts (SMEs) to diagnose and inspect the problem.

3.1.2 Business outcome

AI and ML technology can accelerate evidence gathering, sift out noises so that analysts can focus on real problems, and speed up root cause discovery. This technology can help your organization identify the source that is causing unusual behaviors that are experienced by underpinned infrastructure and applications that are impacting your bottom line.

Traditional monitoring tools were designed more for siloed subsystems, and those alert monitoring tools require SMEs to predefine monitoring policies. If the rules or thresholds are not activated or accurate, enterprises might miss the problem symptoms, which resulted in less-than-optimal timing of remedial efforts.

In general, by design the traditional monitoring or trace tools do not provide built-in ML capabilities that are equipped to auto-compare against the normal historical model and automatically observe your system indicators to assess a potential culprit. To complement the usage of classical monitoring and automation processes, AI can play a pivotal role to infuse unsupervised monitoring and measurement at the machine speed. As a result, the IT operational efficiency for improving mean-time-to-know (MTTK) and mean-time-to-resolve (MTTR) can be accelerated from days or weeks to hours or minutes. This approach can help greatly reduce the duration of downtime and minimize the magnitude of negative impact and disruption to the business services.

3.2 Hybrid incident management

During a commercially popularized shopping day in the US, a major fashion retailer with a strong online presence began experiencing significantly high volumes of online transactions. Because the retail site was experiencing more traffic than expected, they hit a database sizing limit that they had not seen before, which resulted in an outage. It took the company 40 minutes to realize they were experiencing an outage, and then another 36 minutes to notify the correct personnel to investigate and confirm the issue. It took another 17 minutes for change approval, remedial action, and verification that the resolution worked properly. The outage lasted 93 minutes, which is an hour and 33 minutes of downtime and countless lost profits because of it.

This major IT outage amid a wildly popular retail holiday left their customers experiencing a lagging website user experience. Eventually, the customers who were loyal enough to deal with lagging website speeds were entirely unable to access the website, rendering the site useless.

The retailer's IT director explained that this database encountered this unforeseen capacity limit that halted all website transactions. As a part of the digital transformation that businesses are experiencing, customers are no longer satisfied with mediocre online experiences. They do not want "quick enough" experiences, they want fast, almost instantaneous experiences. The window of opportunity to make a sale grows increasingly smaller in an era where consumers have easy access to a growing number of suppliers. Those 93 minutes of downtime meant that the retailer missed their targeted profit margin during a peak online shopping season.

3.2.1 The solution: Proactive incident management with AI

Most modern business applications run across open systems and IBM zSystems in hybrid environments. The complexity between middleware, databases, and networks requires many manual hours of human analysis and multiple domain experts to solve a problem. In this use case, it happened on a Saturday, and for the first 40 minutes, IT operations did not know about the problem until their customers called about website issues. With AI and ML anomaly and SYSLOG analytics, the company might have proactively captured the database error message much sooner and improved its MTTK efficiency.

Figure 3-5 illustrates the Log Analytics tool that might have visualized the error and its impact radius on the relevant subsystems. This relevant information could have helped accelerate problem discovery and provided remedial precision with a much faster response.

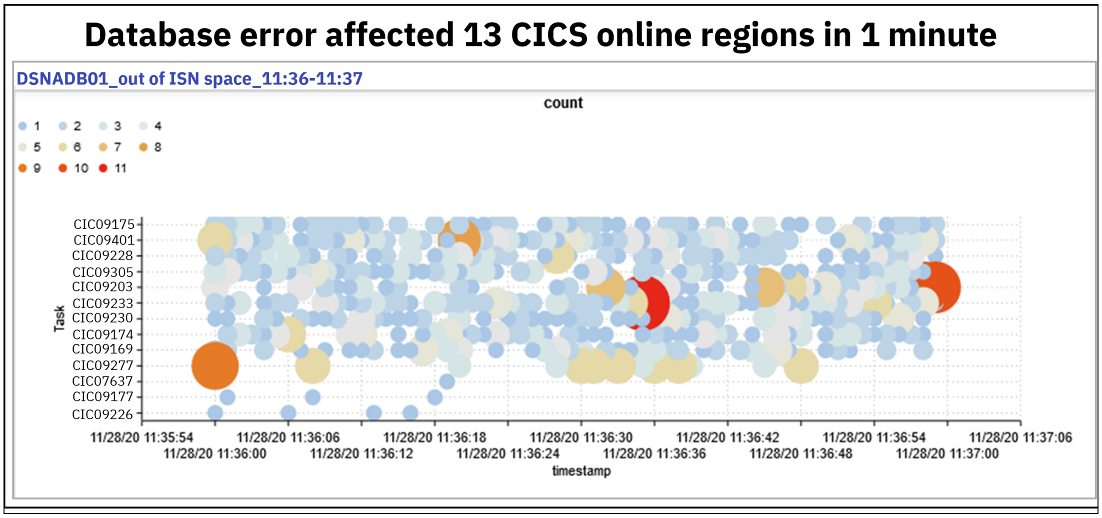


Figure 3-5 Log Analytics detects an anomaly and visualizes the affected CICS regions

Additionally, during a postmortem review, the IT operations saw that several warning signs surfaced a week earlier across the surrounding subsystems. Earlier event messages and the CPU capacity shortage from SMF data indicated the following significant warning signs:

- Numerous spikes of CPU resources exceeded capacity in the previous week. As shown in Figure 3-6, you can see that the consumed MSU exceeded the capped capacity during peak workloads.
- Clustered IBM MQ transactions were queued up, and there was high IBM Z Integrated Information Processor (zIIP) engine utilization during the batch job cycle in the third shift, which signified unusual higher workload volume and resource contention.

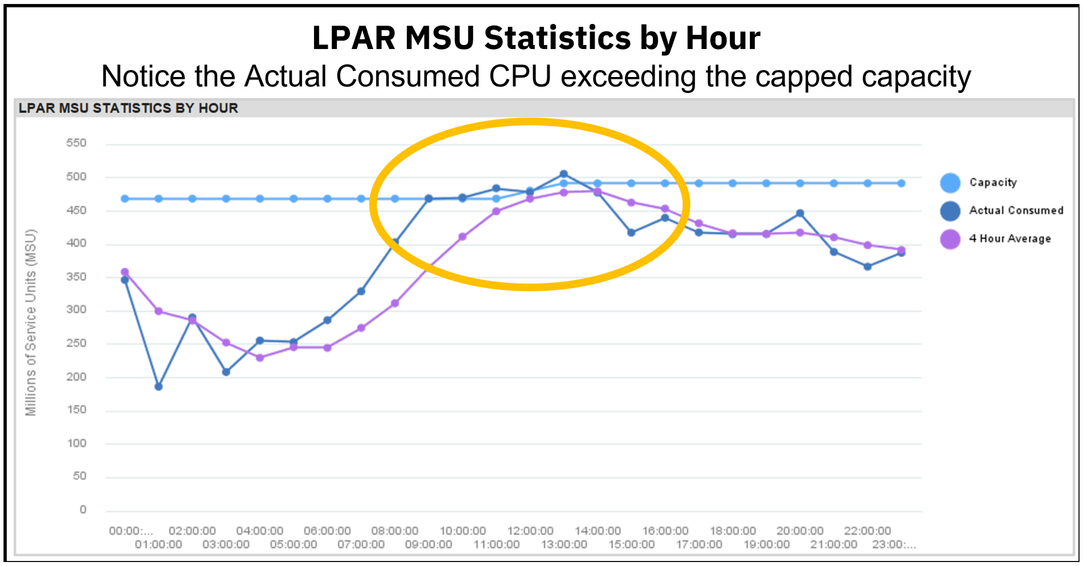


Figure 3-6 Log Analytics detects CPU consumption exceeding the capped capacity

For hybrid applications, the IBM Watson AI operations (AIOps) tool also showed sporadic slow response times.

For this hybrid application outage use case, the operations can configure IBM Problem Insight Server to forward anomalous event alerts to IBM Cloud Pak for Watson AIOps, where Watson AI Manager oversees the enterprise-wide events and correlates events based on the topology and relationships of infrastructural configuration items. This holistic view with deeper understanding of the application and infrastructure relationships further accelerates problem determination much more efficiently.

Figure 3-7 illustrates an IBM Cloud Pak for Watson AIOps integrated work flow with AI and ML and ZAA:

1. Events and alerts are triggered from ZAA or the IBM OMEGAMON® performance monitoring suite.
2. IBM Cloud Pak for Watson AIOps receives alerts from ZAA.
3. Watson AI Manager correlates events and metric anomalies, forwards the event story to the ChatOps component, and notifies IT operations personnel or the Site Reliability Engineer (SRE).
4. Watson AIOps integrates relevant IBM zSystems insights with the ServiceNow ticket history, and expedites incident handling with relevant history and a prescribed recommendation runbook.

The SRE can optionally drill down and launch in context back to the ZAA Problem Insight console for more diagnostic information and take corrective actions on the z/OS system console.

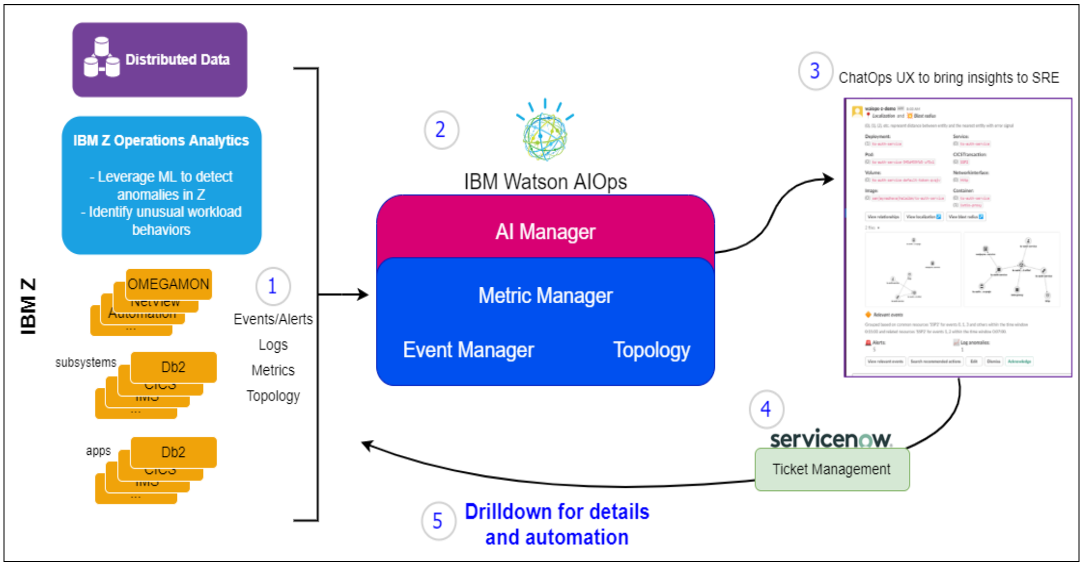


Figure 3-7 Integrated incident management flow from IBM Z with IBM Cloud Pak for Watson AIOps

3.2.2 Business outcome

With AI-infused automation, IBM Cloud Pak for Watson AIOps improved the MTTR by 43% for this use case. For other cases with similar causes, such as the database running out of storage, index page excessive splits, or an out of internal space map sequence record, AI-automated database KPI monitoring and comparison against the historical space growth rate can empower IT operations staff with much more insightful information, allowing them to become proactive in their resolution efforts without manually searching and sifting vast data sources. Business continuity objective can be achieved, and customer satisfaction can remain uncompromised.

3.3 Release and change management

When there is a system problem occurrence, your first reaction is “What changed”? Often, change is required to support new business needs. Whether it is a system bulk maintenance change or a major new application rollout, change might introduce risks to system stability and business continuity, especially with the agile pace of business demands today. The IT manager frequently asks how to be proactive to assess change impact before deciding to approve a release roll-out plan.

Another challenge is that comprehensive test cases are not easily built, and often a surprising condition disrupts the normal production operations.

Use case: Insurance company

A large insurance company began encountering periodic timeouts of their regular batch jobs after system administrators turned their bulk of z/OS system and Db2 maintenance to production over the weekend. However, operators did not recognize any abnormalities until users called in with complaints. Initially, the problem was not obvious, and it was difficult to determine which domain expert should be called. It was not until hundreds of failed jobs flooded in that the operator called the senior database administrator, who eventually called IBM Support.

After a few Severity 2 tickets were opened, IBM cross-domain teams from Db2 and TCP/IP support started collecting diagnostic and trace data. This problem was not easily isolated, and it was difficult to determine the root cause. The client rolled back part of the maintenance change, but problems continued. A final system dump diagnosis led to a test fix to close out the issue. This intermittent disruption impacted their operations over 3 days. The total revenue and outage cost was estimated over USD $100 million.

Use case: Bank

A similar case occurred at a large European bank, which rolled out a new version of Db2. Throughout the month after the rollout, the users noticed frequent problems in IBM MQ applications that were queued up unpredictably. In addition, Db2 appeared to have more unusual lockouts. These sporadic symptoms occurred for 3 weeks, and then one peak day, Db2 experienced excessive timeouts among 16-way data sharing members. This behavior drove z/OS global resource serialization (GRS) to its limits, and the entire sysplex was locked up. The bank had to run an initial program load (IPL) on all the affected z/OS logical partitions (LPARs) and subsystems within this production IBM Parallel Sysplex®.

The total duration of downtime was over 22 hours, with a substantial negative impact to their business operations.

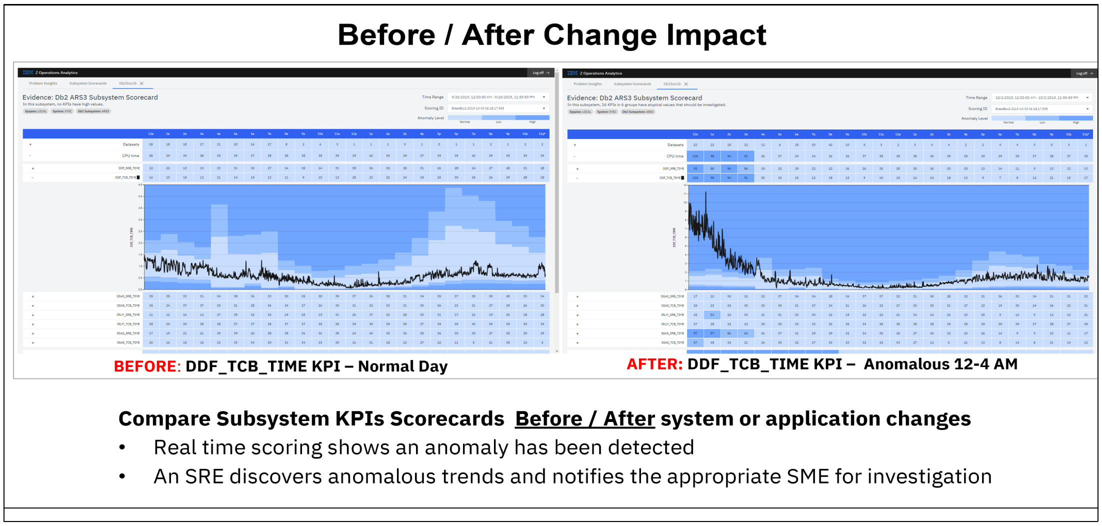
3.3.1 The solution: Assessing change impact with AI

The challenges were two-fold for this use case: Was the problem caused by the infrastructural system change or by application code changes? Most failed jobs were victims of the same problem. It took a substantial amount of time between organizations and domain specialists to pinpoint the root cause lurking inside of millions of lines of software component code. With such a massive bulk of maintenance, it was difficult to isolate whether the bad actor was in the z/OS base, network, or Db2 component.

With AI-ML infused technology, you can visualize and quantify unusual spikes and trends, and highlight impacted areas with the same failure return code and with auto-comparison against the baseline normal operations model. In this use case, both Db2 and IMS job failures showed substantial abnormal trends with TCP/IP bad reason codes. These abnormalities might point to a direction beyond only diving into the silo-subsystem component, which led to a faster discovery to find a fresh HIPER APAR in the support knowledge portal. When z/OS and the UNIX System Services KPI set are activated with ZAA auto-detecting anomalies, clients gain even more benefits with a holistic view of their broader IT system and subsystem's health indicators.

In “Use case: Bank” on page 29, the symptoms initially surfaced through IBM MQ indicators, and it was not clear whether the IBM MQ problems originated with a client or something else. Db2 problem symptoms became more obvious and prominent after higher transaction volume of workloads came online.

Figure 3-8 shows ZAA output where Db2 DDF_TCB_Time shows abnormally high suspension trends. This built-in comparison viewing feature can help simplify manual tasks that are dependent on an SME to process a vast number of raw SMF records and compare system performance differences before and after a change.



Before / After Change Impact

BEFORE: DDF_TCB_TIME KPI – Normal Day

AFTER: DDF_TCB_TIME KPI – Anomalous 12-4 AM

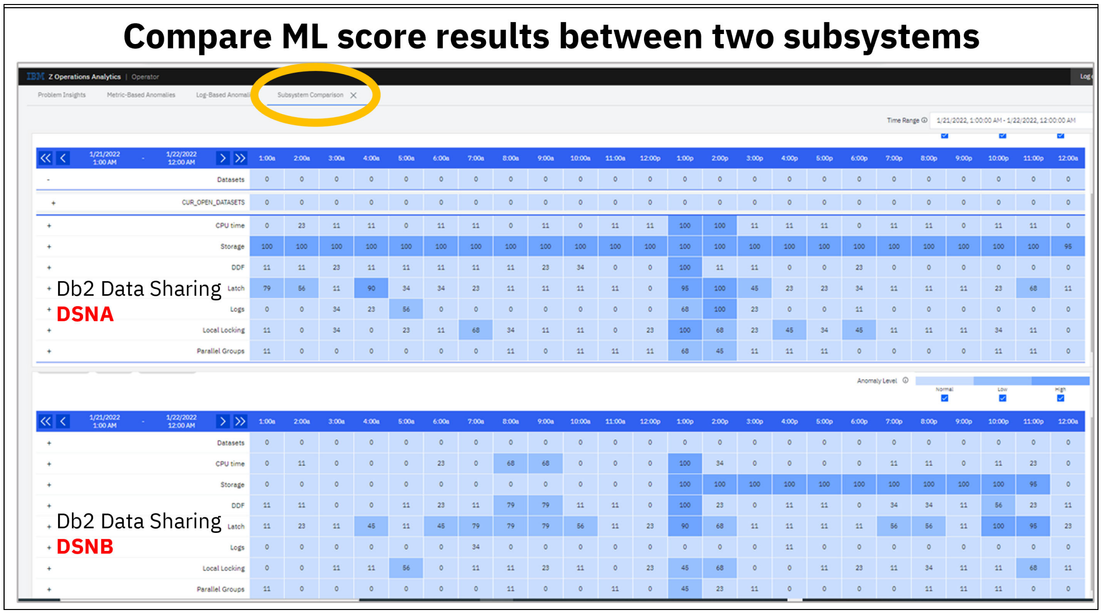
Compare Subsystem KPIs Scorecards Before / After system or application changes

- Real time scoring shows an anomaly has been detected
- An SRE discovers anomalous trends and notifies the appropriate SME for investigation

Figure 3-8 Comparing ML results before and after a system or application change

In addition to ML-based anomalous scoring, IBM Z log analytics can be used to demonstrate unusual spikes with excessive timeouts among Db2 data sharing members. Simultaneously, you can quickly determine that the IBM MQ queue depth was backed up because of high message arrival rates. By using the cognitive content analysis capability, you see that both Db2 and IBM MQ abnormal trends trigger alert notifications much sooner than in a reactive mode.

Another common use case is to compare an application or a system release change. For example, if tuning exercises were done to refine the z/OS workload manager policy, you want to validate whether the workload balance is achieved among data sharing members. In Figure 3-9 on page 31, the upper key performance scorecards compare Db2 data sharing member DSNA to data sharing member DSNB. The scores are the result of an ZAA continuous scoring feature evaluation with real-time SMF data. Using this ZAA built-in "Comparison" feature, you can view KPI scores between subsystems side-by-side at a glance, and determine how the anomaly results differ before and after a major change, either in the system configuration or in an application program.



Compare ML score results between two subsystems

Time Range ⓘ 1/21/2022, 1:00:00 AM - 1/22/2022, 12:00:00 AM

Db2 Data Sharing **DSNA**

1/21/2022 1:00 AM - 1/22/2022 12:00 AM	1:00a	2:00a	3:00a	4:00a	5:00a	6:00a	7:00a	8:00a	9:00a	10:00a	11:00a	12:00p	1:00p	2:00p	3:00p	4:00p	5:00p	6:00p	7:00p	8:00p	9:00p	10:00p	11:00p	12:00a
- Datasets	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
+ CUR_OPEN_DATASETS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
+ CPU time	0	23	11	11	0	11	11	0	11	0	11	11	100	100	11	11	11	0	11	11	0	11	11	0
+ Storage	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	95
+ DDF	11	11	23	11	11	11	11	11	23	34	0	0	100	11	11	0	0	23	0	0	0	0	0	0
+ Latch	79	56	11	90	34	34	23	11	11	11	11	0	95	100	45	23	23	34	11	11	11	23	68	11
+ Logs	0	0	34	23	56	0	0	0	0	0	0	0	68	100	23	0	0	11	0	0	0	0	0	0
+ Local Locking	11	0	34	0	23	11	68	34	11	11	0	23	100	68	23	45	34	45	11	11	11	34	11	0
+ Parallel Groups	11	0	0	0	0	0	0	11	0	11	11	11	68	45	11	11	11	0	0	0	0	11	11	0

Anomaly Level ⓘ
☑ Normal ☑ Low ☑ High

Db2 Data Sharing **DSNB**

1/21/2022 1:00 AM - 1/22/2022 12:00 AM	1:00a	2:00a	3:00a	4:00a	5:00a	6:00a	7:00a	8:00a	9:00a	10:00a	11:00a	12:00p	1:00p	2:00p	3:00p	4:00p	5:00p	6:00p	7:00p	8:00p	9:00p	10:00p	11:00p	12:00a
+ Datasets	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
+ CPU time	0	11	0	0	0	23	0	68	68	0	0	0	100	34	0	0	0	0	11	11	0	11	23	0
+ Storage	0	0	0	0	0	0	0	0	0	0	0	0	100	100	100	100	100	100	100	100	100	100	95	0
+ DDF	11	11	0	0	11	23	11	79	79	11	11	0	100	23	0	11	11	0	34	34	11	56	23	11
+ Latch	11	23	11	45	11	45	79	79	79	56	11	23	90	68	11	11	11	11	56	56	11	100	95	23
+ Logs	0	0	0	0	0	0	34	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0
+ Local Locking	0	0	11	11	56	0	11	11	23	11	0	23	45	68	0	0	11	23	11	34	11	11	68	11
+ Parallel Groups	11	0	0	0	0	0	0	11	0	0	11	0	45	23	11	0	0	0	0	11	11	11	0	0

Figure 3-9 Using the comparison viewing feature to assess DSNA and DSNB KPIs (displayed side by side)

3.3.2 Business outcome

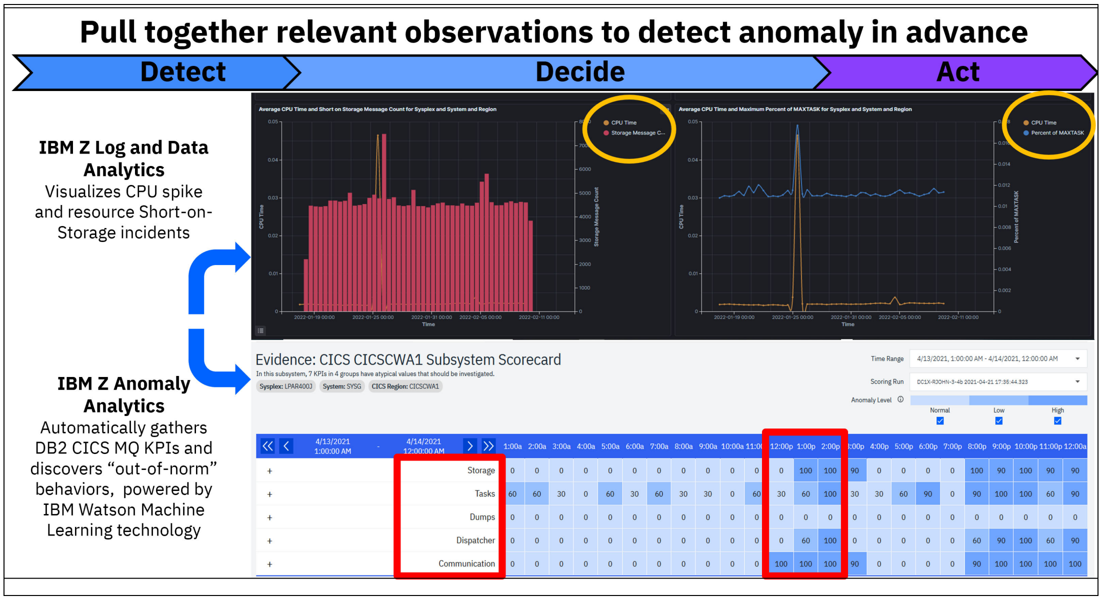
With automated, out-of-norm detection and evidence gathering by AI-ML capabilities, IT support teams can assess a problem's magnitude and prepare themselves with more data reference points. AI inferencing can speed up the problem determination process of your business with scientific evidence at machine speed. This process helps minimize the mean-time-to-identify (MTTI) and MTTR for faster recovery with a minimal impact. SREs and support teams can be empowered with AI that uses deeper insights to pro-actively pinpoint the root cause of problems.

For client change release management, this use case shows how AI can help provide relevant evidence across systems and components by leveraging auto-comparison against a historical baseline to assess before and after change impact for making informative business decisions. This scope expands from operations to DevOps application management.

3.4 Summary

The purpose of this chapter was to describe some of the real customer use cases that might be more efficiently resolved by using AI and ML technology. The SMF and log data for your systems are natural data sources for gaining valuable IBM Z insights without imposing extra resource impact because the data does not require expensive trace instrumentation. ML can take advantage of scientific rules to detect system-wide performance and resource anomalies, and SYSLOG content can add eventful, detailed information that is associated with impacted applications, transactions, infrastructure components, and database objects. Using AI, ML, and log content analytics technologies together, you can empower your operations with deeper insights about system health and out-of-norm conditions much earlier and have better chances of mitigating your risks and achieving high resiliency goals.

Figure 3-10 illustrates the concept of how an ensemble with relevant information (from log content and ML anomaly scores) can turn siloed data into actionable insights. The upper part of Figure 3-10, IBM Z Log Analytics produces a visualization of a CPU spike with a CICS Short-on-Storage occurrence. The lower part of the figure shows the detection of anomalies in CICS KPIs that are related to storage, maximum number of tasks, and TCP/IP socket connections.


Pull together relevant observations to detect anomaly in advance
Detect
Decide
Act
Average CPU Time and Short on Storage Message Count for Sysplex and System and Region
CPU Time
Storage Message C...
CPU Time
Storage Message Count
Time
2022-01-19 00:00
2022-01-25 00:00
2022-01-31 00:00
2022-02-05 00:00
2022-02-11 00:00
Average CPU Time and Maximum Percent of MAXTASK for Sysplex and System and Region
CPU Time
Percent of MAXTASK
Time

IBM Z Log and Data Analytics
Visualizes CPU spike and resource Short-on-Storage incidents

IBM Z Anomaly Analytics
Automatically gathers DB2 CICS MQ KPIs and discovers “out-of-norm” behaviors, powered by IBM Watson Machine Learning technology

Evidence: CICS CICSCWA1 Subsystem Scorecard

In this subsystem, 7 KPIs in 4 groups have atypical values that should be investigated.

Sysplex: LPAR400J System: SYSG CICS Region: CICSCWA1

Time Range 4/13/2021, 1:00:00 AM - 4/14/2021, 12:00:00 AM

Scoring Run DC1X-RJOHN-3-4b 2021-04-21 17:35:44.323

Anomaly Level Normal ☑ Low ☑ High ☑

« ‹ 4/13/2021 1:00:00 AM - 4/14/2021 12:00:00 AM › »		1:00a	2:00a	3:00a	4:00a	5:00a	6:00a	7:00a	8:00a	9:00a	10:00a	11:0	12:00p	1:00p	2:00p	:00p	4:00p	5:00p	6:00p	7:00p	8:00p	9:00p	10:00p	11:00p	12:00a
+	Storage	0	0	0	0	0	0	0	0	0	0	0	0	100	100	90	0	0	0	0	100	90	100	90	90
+	Tasks	60	60	30	0	60	30	60	30	30	0	60	30	60	100	30	30	60	90	0	90	100	100	60	90
+	Dumps	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
+	Dispatcher	0	0	0	0	0	0	0	0	0	0	0	0	60	100	0	0	0	0	0	60	90	100	60	90
+	Communication	0	0	0	0	0	0	0	0	0	0	0	100	100	100	90	0	0	0	0	90	100	100	100	100

Figure 3-10 Enriching observability insights with IBM Z Anomaly Detection and Log Analytics

In summary, AI and ML can empower your IT operations to discover unknown or hidden issues before they damage your system health. Traditional alert monitoring requires human knowledge to predefine thresholds for what to monitor and when to alert. With an AI-infused solution, ZAA auto-learns your system and detects out-of-norm behavior in near real-time mode, which provides extra monitoring to ensure your system’s optimal resiliency.

You might ask, what is the cost to implement AIOps? For today’s modern complex applications running across platform infrastructure, automating anomaly detection in real time requires much higher computing power to process high volumes of system performance records across the enterprise and produce deep insights without impacting workload performance. With the recent introduction of IBM z16 hardware, which is embedded with an industry first on-chip integrated AI accelerator, clients can leverage the platform’s hardware advantages to automate system anomaly detection and analytics among billions of unstructured log and metric data in real time at unprecedented speed and scale. This AI accelerator on-chip innovation enables AIOps to deal with petabytes of system management records and millions of log messages daily in real-time mode a reality.

4

Data and artificial intelligence operations use cases

Data artificial intelligence operations (AIOps) are a continuation of the efforts to improve the agility, efficiency, and decision-making data operations (DataOps) by harnessing artificial intelligence (AI) and machine learning (ML) solutions. Like development delivery transformation in DevOps approaches, DataOps looks into improving automation and quality of data intelligence systems by data analysts and data engineers. DataOps is a collection of practices and architectural templates that focus on the integration of data flows between data sources and producers and data consumers across an organization. Application of AIOps automation for data further improves the efficiency and quality of DataOps.

To address the key challenges of Data AIOps that are introduced with recent developments within hybrid cloud, AI, the Internet of Things (IoT), and edge computing, IBM aligned its solutions with a data fabric data management design concept. An exponential growth of big data and data complexity increasingly prompts enterprises to unify and govern their data environments. Although data growth is exponential, knowledge growth remains mostly linear. data fabric is meant to improve the situation. It is an architecture that facilitates the end-to-end integration of various data pipelines and cloud environments by using intelligent and automated systems. The data fabric ensures universal data access and helps to create efficiencies and enforceable policies around data to allow organizations to manage exponential data increase.

Combining the concept of data gravity of IBM Z along with data fabric architectural solutions can help your organization modernize the application infrastructure while decreasing latency, cost, complexity, and security risks. A data fabric can simplify data complexity by automating data integration, governance, and consumption. Both application and data can be colocated on IBM Z while also benefiting from best-in-class data privacy, security, availability, scalability, resiliency, and sustainability. IBM Z is also a key component of the IBM hybrid cloud strategy with Red Hat OpenShift Container Platform, which spans across hardware, on-premises data centers, and on-cloud deployments. Choosing IBM Z and Red Hat OpenShift as a target platform for data fabric implementation synergizes the values of all three components.

© Copyright IBM Corp. 2022.

4.1 Simplifying data complexity with a data fabric on IBM Z

Over time, organizations created several data silos by using the tools that were available at the time, whether an on-premises data lake, data warehouse, or any type of relational database, or a cloud data warehouse or other type of business intelligence warehouse. The challenge with multiple siloed data stores is how to make timely decisions that are also holistic and data-centric. In an attempt to enable decision-making with this data, organizations start creating replicas, offload the data for analysis, extract, transform, and load (ETL) the data, and continuously implement sources of various complexity and technological relevance. Increasingly, this process leads to issues of latency, security, complexity, resiliency, and cost, which further complicates the data landscape and fails to reach the objective of making timely, holistic, and data-centric decisions.

A data fabric helps to alleviate the boundaries between siloed data by connecting data at its source rather than collecting the data in one place. This task is achieved by using a data virtualization technique that consists of three layers with a distinct purpose (Figure 4-1). A connection layer on top of data structures that were designed independently provides various adapters to connect them. A data virtualization layer simplifies data complexity by using intelligent and automated integration. Finally, the consumer layer provides a data access interface to the analytics projects, dashboards, data catalogs, and other applications. Once implemented, lower layers remain intact with minimal maintenance while all new development can focus on the layers above.

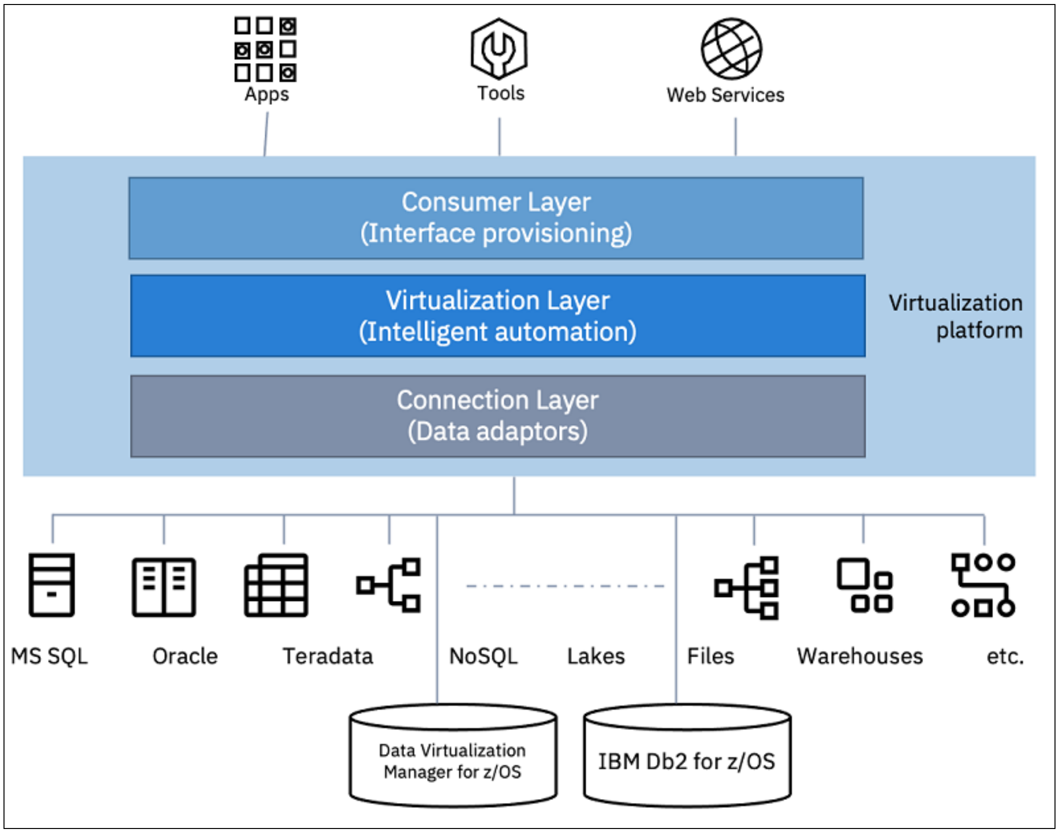


Figure 4-1 Data fabric layered approach to virtualizing independent data sources

4.1.1 Read-only cloud access to Db2 for z/OS data

Businesses with data in Db2 on IBM Z want a simplified way to access the data from new hybrid cloud applications for analytics purposes. These businesses experience the typical challenges of high latency and high impact, so they create complex and expensive in-house code, and have difficulty in getting transactional consistency. When data is separated from its source, it becomes outdated compared to the current state of the System of Record (SOR). More frequent polling of the data creates extra workload on the CPU, and it increases memory, disk, and network usage. To reduce the latency and validity of the data, organizations often try to come up with a complex, in-house solution that runs on IBM z/OS, which leads to more complexity and increases code maintenance. Such businesses are on a journey to the cloud with their Db2 for z/OS data, but are struggling to make it work effectively.

To improve the integration of data originating in Db2 for z/OS into modern hybrid cloud applications, IBM presents IBM Db2 for z/OS Data Gate (Db2 Data Gate) as a solution. It delivers data for hybrid cloud use cases, such as high-volume inquiry workloads. For online banking applications, whether in a web interface, mobile application, or even a partner API, the inquiries to account balances, delivery statuses, and claim statuses can be performed with superior performance characteristics to ensure better currency of data. For applications retaining control on Db2 for z/OS, Db2 Data Gate, which is shown in Figure 4-2, allows the running of off-platform analytics and warehousing. Organizations can access Db2 for z/OS data on the cloud or from Linux and IBM Cloud Pak for Data (version 4.0 at the time of writing), which are hosted on-premises. Pre-integrated data and AI services in IBM Cloud Pak for Data on Red Hat OpenShift Container Platform provides a complete range of capabilities within an open and extensive cloud-native platform that can be quickly leveraged to build AI and ML applications on top of Db2 for z/OS data.

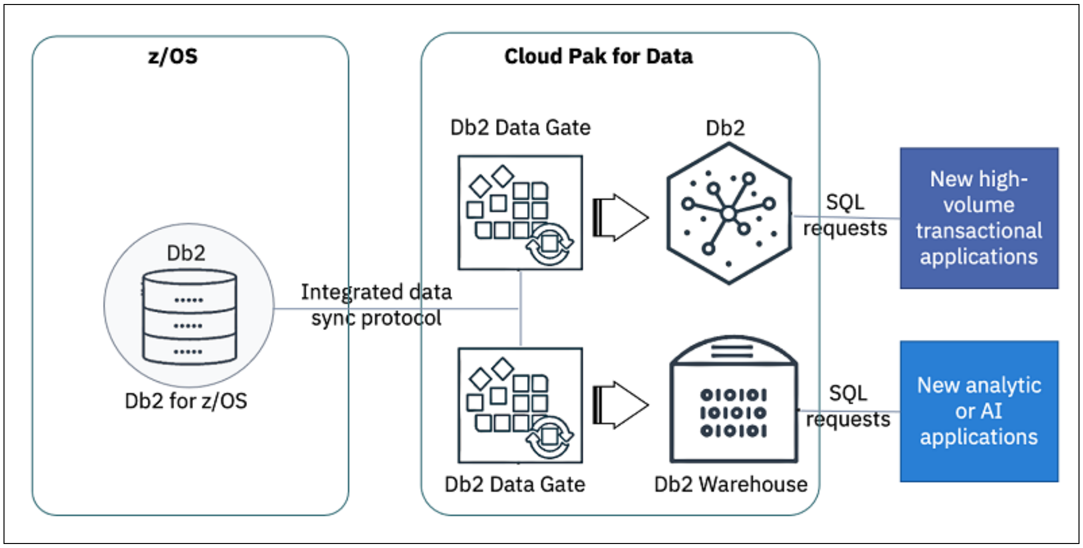


Figure 4-2 Db2 Data Gate for read-only data queries from cloud-native applications

Db2 Data Gate is an integrated synchronization solution for read-only data queries that is optimized in several ways compared to any general-purpose replication. It is built into the mainline of Db2 for z/OS v12 and uses a IBM Z Integrated Information Processor (zIIP) enabled synchronization protocol. The protocol is lightweight, high-throughput, and low-latency, with a minimal impact on the existing SOR in terms of central processors (CPs), memory, and I/O. The Db2 Data Gate service runs on the IBM Cloud Pak for Data side and receives the data from a Db2 for z/OS source. The target database in IBM Cloud Pak for Data can be Db2 or an IBM Db2 Warehouse (an analytics data warehouse). New high-intensity transactional workloads, such as mobile banking applications, justify the use of Db2, while for analytic or AI workloads, they are an instance of a Db2 Warehouse. As a result, the data in Db2 for z/OS can be accessed in near real time from new cloud-native applications without degrading the performance of the core transaction engine.

4.1.2 Leveraging data virtualization on IBM Z in a data fabric

Businesses have existing investments in infrastructure, which hold tremendous value and should not be wasted in this new hybrid cloud world. How can you transform into an insights-driven enterprise, create your cloud IT strategy, and leverage and expand on proven platforms such as IBM Z and other IBM technology? How can you make data accessible whether the source is a new or existing or relational or non-relational, including the critical business data on IBM Z? How can you deploy ML models within transactional applications as output from your data sources? These questions and many others arise when companies think about the next generation of data organization. Many companies choose to move data off the source platform and put it all in one place, which introduces latencies and creates an impact on the network and storage side, and also increases security risks. Any form of replication and synchronization leads to significant development and maintenance efforts as the amounts of data increase.

The answer lies in data virtualization techniques that make connections to data structures that were designed independently possible. Data virtualization is one of the key elements in a data fabric, which simplifies data complexity and consumption by using intelligent and automated integration. When you must integrate existing data assets on IBM Z, one popular solution is IBM Data Virtualization Manager (DVM) for z/OS. It provides virtual, integrated views of data on IBM Z, and enables users and applications to gain read/write access to IBM Z data in place without moving, replicating, or transforming the data. With a centralized view of data in the virtualization layer, including IBM Z data, DVM for z/OS can be used seamlessly in Watson Studio on IBM Z in the Consumer layer to readily build, test, and evaluate AI models on your platform of choice.

Traditional data movement approaches can negatively impact the opportunity to benefit from data where and when it is needed. By unlocking IBM Z data, you can access data in place from applications, and update and join IBM Z data in real time with other enterprise data. Interfacing with IBM DVM for z/OS occurs through a modern set of APIs that includes SQL, NoSQL, REST, and SOAP. Mainframe development skills are not required. Connecting to data rather than replicating it minimizes data movement. Having DVM optimized for IBM Z can lead to lower cost and reduced risks.

Figure 4-3 on page 37 illustrates how data virtualization enables consumption of various and otherwise difficult to access z/OS data sources.

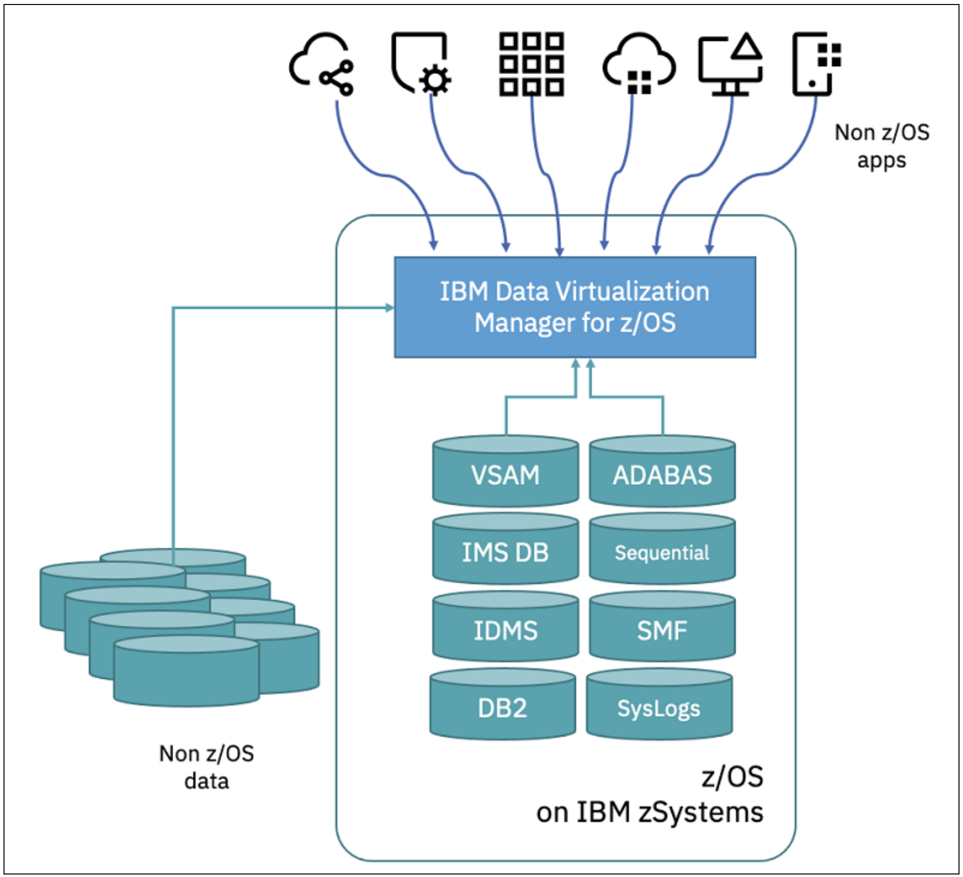


Figure 4-3 Data virtualization for various data sources

IBM DVM for z/OS can simplify the development of applications that access relational and non-relational data types, including IBM Db2 for z/OS, Virtual Storage Access Method (VSAM), IBM Information Management System (IMS), ADABAS, Integrated Database Management System (IDMS), and System Management Facility (SMF). Modern applications can interface with DVM by using an SQL data interface. Watson Machine Learning on z/OS can simplify analytics and AI and ML development by importing AI models in Predictive Model Markup Language (PMML) and Open Neural Network Exchange (ONNX), enabling the applications to address more complex information needs within business services and in colocation with the data. This approach minimizes the need for specific mainframe development skills and allows you to enrich modern applications with live transactional data. DVM is optimized for IBM Z, including a high degree of zIIP and memory usage for high performance.

4.1.3 Connecting the correct data to the correct people from anywhere with IBM Cloud Pak for Data

Companies are increasingly dealing with complexity when trying to understand and resolve the “Version of Truth” among a growing number of data sources. The requirement to bring data closer to the business insight applications is met by copying the data from the source. Collecting the data in one place for more holistic decision-making leads to the creation of a central data lake or data hub. The downsides of such solutions are latency, security, complexity, resiliency, and cost issues. The data becomes stale and is no longer trusted, and there are increased storage and integration costs as the data and number of sources increase. Regional data protection rules might prohibit central hub style data collection.

Rather than collecting the data in one place for processing, the approach to address the issues should be to connect the data at its source. A modern cloud-native implementation of a data fabric architecture that follows this approach is provided by IBM Cloud Pak for Data (version 4.0 at the time of writing). It connects the correct data at the correct time for faster, trusted AI outcomes. IBM Cloud Pak for Data uses a unified platform that spans hybrid and multi-cloud environments to ingest, explore, prepare, manage, govern, and serve petabyte-scale data for business-ready AI.

IBM Cloud Pak for Data is a fully integrated data and AI platform that modernizes how businesses collect, organize, and analyze data. It modernizes the way how they infuse AI throughout their organizations. IBM Cloud Pak for Data is built on the Red Hat OpenShift Container Platform and integrates with key IBM Z technology, which means that you can use your IBM Z enterprise data more securely in place for use in applications such as engagement, AI, and analytics. You can develop AI models within IBM Cloud Pak for Data and then readily deploy them into enterprise production applications for actionable, real-time insight at the point of a transaction.

When data originates on-premises, query results in real time can be achieved only if the data is accessed at its source on IBM Z. Available bandwidth and cloud capacity cannot compensate for the data latency that is associated with data movement. By incorporating IBM Z into a data fabric (see Figure 4-4), your organization can access transactional data at its source to deliver real-time insight from that data for real-time decisions.

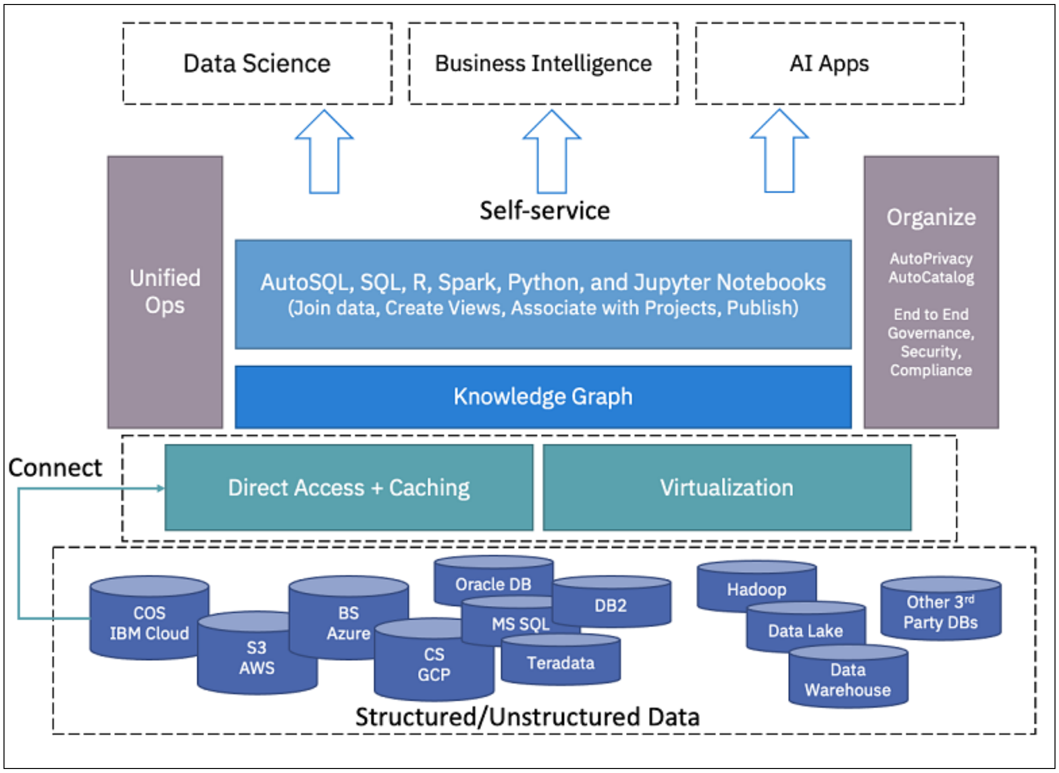


Figure 4-4 Data fabric implementation of data collection or connection in IBM Cloud Pak for Data

Data virtualization mechanisms in IBM Cloud Pak for Data integrate data sources across multiple types and locations and turns them into one logical data view. This virtual data lake makes the job of getting value out of your data easy. By creating connections to your data sources, you can quickly view across your organization's data. This virtual data platform enables real-time analytics without moving data or by using duplication, ETLs, and other storage requirements, so processing times are greatly accelerated. This approach brings real-time insightful results to decision-making applications or analysts more quickly and dependably than existing methods.

With data virtualization, you can manage users, connect to multiple data sources, create and govern virtual assets, and then consume the virtualized data. Connect, join, create views, and consume are the main actions that are needed for data virtualization. Data virtualization supports queries by using standard SQL through common interfaces by including R, Spark, Python, and Jupyter Notebooks, in addition to most common analytics application tools, including IBM Watson Studio (formerly IBM Data Science Experience) and IBM Cognos® Analytics.

IBM Cloud Pak for Data comes with the AutoSQL technology, which automates the access, integration, and management of data for AI. AutoSQL is a high-performance, universal query engine that simplifies the data landscape by enabling clients to use the same query across disparate data sources, including data warehouses, data lakes, and streaming data, which saves time and resources that typically go into moving data and maintaining multiple query engines. With the platform’s existing data virtualization capabilities, AutoSQL empowers users to easily query data across hybrid, multi-cloud, and multi-vendor environments. AutoSQL includes pre-integrated data governance capabilities, so data consumers are assured of the quality and validity of the data.

4.1.4 Consolidating approaches with IBM Cloud Pak Business Automation

In response to marketplace pressure to be agile and competitive, businesses must effectively respond to customers and provide great customer experience. Different parts of an organization focusing on their respective areas naturally lead to a piecemeal approach for automation. Such an approach often fails to capture all customer data. Also, the data that is captured is of varying levels of detail and quality, and it requires validation before it can be used for decision-making. Actions need insights, and inability to provide business users with easy access to key insights creates a lag in meeting evolving customer needs. Already, data and content must be embedded within the business application to monitor operations through key performance and workforce dashboards.

IBM Cloud Pak for Business Automation runs on Red Hat OpenShift, including Red Hat OpenShift on the IBM Linux on Z and LinuxONE platforms. By leveraging Red Hat OpenShift, you may easily move or colocate the application closer to the data by deploying on Linux on IBM Z. Both AI and ML and rules-based applications can run on the same platform, where the former can augment the latter to enrich business-critical initiatives, such as the personalization of customer interactions, processing credit authorizations, processing insurance claims, and detecting fraud in real time.

IBM Cloud Pak for Business Automation, while running on the IBM Z and LinuxONE platforms, can leverage data from various areas of business operation on and off the platform. Data from event or transaction logs from enterprise applications like SAP can be used to process mining services to understand the real work that is done by employees, locate hidden bottlenecks, and pinpoint where automation leads to the biggest process improvements.

Data that is generated by operating systems can be captured and presented in dashboards of operational intelligence services to data scientists for analyzing and gaining insights by using AI and ML (for example, in Business Automation Insights and Business Performance Center, as shown in Figure 4-5). A full lifecycle of enterprise content is another area that can be securely managed. Content can support unstructured or semi-structured data that is composed of documents, text, images, audio, and video. Other features of IBM Cloud Pak for Business Automation allow you to consolidate document processing and automate decisions and workflows on the same cloud-native platform.

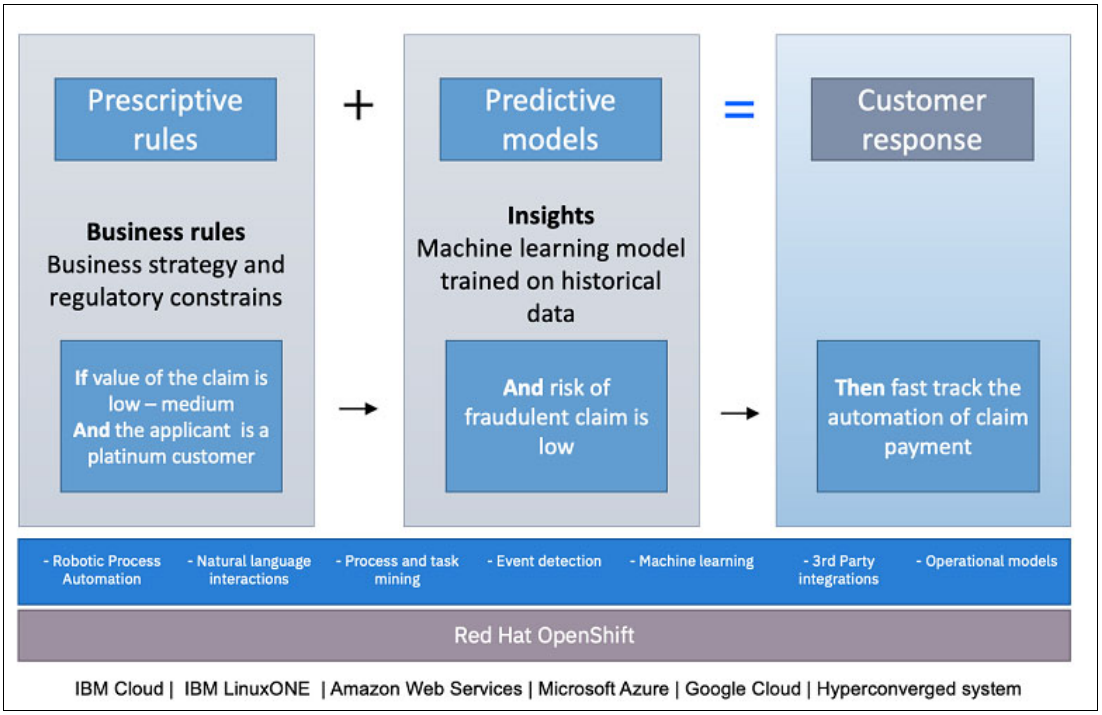

Prescriptive rules
+
Predictive models
=
Customer response
Business rules
Business strategy and regulatory constrains
Insights
Machine learning model trained on historical data
If value of the claim is low – medium
And the applicant is a platinum customer
And risk of fraudulent claim is low
Then fast track the automation of claim payment
- Robotic Process Automation
- Natural language interactions
- Process and task mining
- Event detection
- Machine learning
- 3rd Party integrations
- Operational models
Red Hat OpenShift
IBM Cloud | IBM LinuxONE | Amazon Web Services | Microsoft Azure | Google Cloud | Hyperconverged system

Figure 4-5 Combining business rules and machine learning in IBM Cloud Pak for Business Automation

To achieve effective customer response, services in IBM Cloud Pak for Business Automation combine business rules and ML. ML models (for example, in Watson Machine Learning or through third-party ML providers) are accessible to business analysts and augment rules-based decisions with extra insights into data. For example, an insurance company can create a claim automation solution with a prescriptive rule to check into the value of the claim in the incoming data stream. They might have an ML model that trained on previously processed claims. The ML in production looks into the claim details and predicts how low the risk of fraud might be. If both criteria are satisfied, the decision service “fast tracks” the automation of a claim payment. The data and all components of such a solution run on the IBM Z or LinuxONE platform, with an Red Hat OpenShift cloud-native platform, which can also span across other on-premises or on-cloud infrastructure options.

4.2 Efficient data governance with AIOps in an IBM Z data fabric

Complexity in data management is made worse when there is not a holistic data governance practice. If a data-driven customer experience is sought, then data itself must be treated as a legitimate business asset, that is, a governance practice is essential. How do you ensure trusted and secure insights with leading governance capabilities on the cloud for enterprise? How can you eliminate general bottlenecks and speed decision making? How do you achieve data governance, data quality, and data privacy?

A data fabric simplifies adherence to defined rules and processes, including data privacy restrictions for business-ready data. A data fabric helps you to know what data an organization has, where that data is, and how that data can be used while adhering to data privacy restrictions. IBM ML-assisted data cataloging and governance tools at the core of its data fabric offering provides capabilities that allow customers to ensure compliance with business terms, data privacy (for example, General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Health Insurance Portability and Accountability Act (HIPAA), and Fair Credit Reporting Act (FCRA)) and protection rules, and the lineage of the data. Cataloging maintains active metadata, and provides masking and redacting capabilities, role-based access control (RBAC), and lineage tracking (see Figure 4-6).

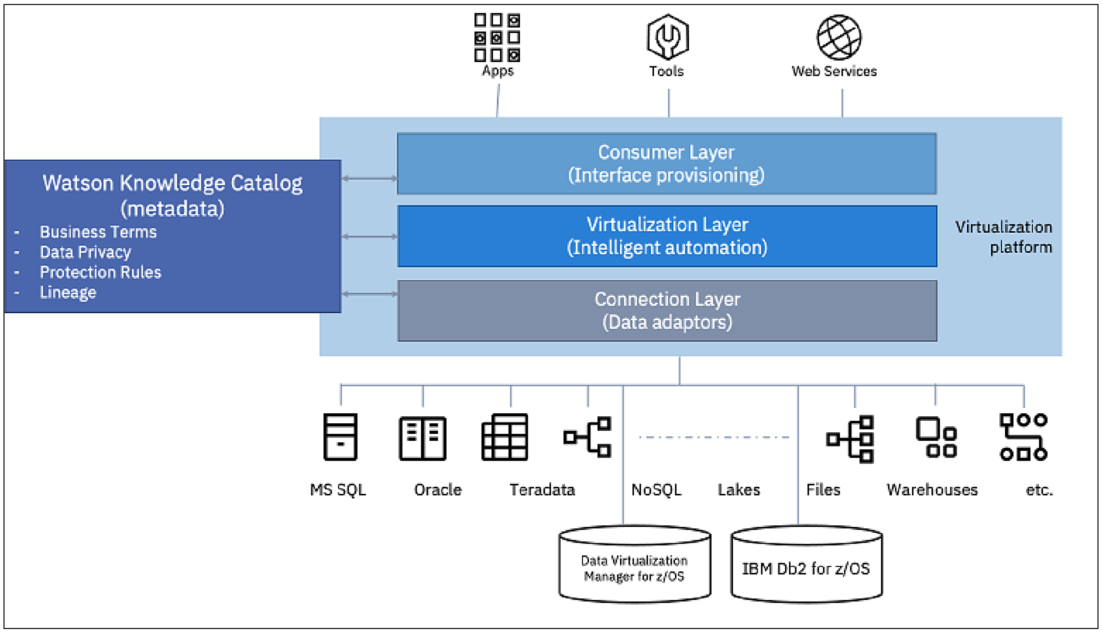

Apps
Tools
Web Services
Watson Knowledge Catalog (metadata)
- Business Terms
- Data Privacy
- Protection Rules
- Lineage
Consumer Layer (Interface provisioning)
Virtualization Layer (Intelligent automation)
Connection Layer (Data adaptors)
Virtualization platform
MS SQL
Oracle
Teradata
NoSQL
Lakes
Files
Warehouses
etc.
Data Virtualization Manager for z/OS
IBM Db2 for z/OS

Figure 4-6 Cataloging in data fabric for adherence to defined rules and processes

4.2.1 Enabling governance of z/OS data sources

To meet business requirements and client expectations, applications require information that is composed and aggregated across data from multiple systems of records. It is important to provide the access *and* apply industry-specific regulatory policies and rules across platforms and track how data is used. The data in Db2 for z/OS (and other IBM Z data) must be integrated in data governance scenarios.

IBM Watson Knowledge Catalog (WKC), which is a part of IBM Cloud Pak for Data, can be used as a central governance point for enterprise data. WKC provides capabilities for cataloging, discovering, browsing, and understanding the data. It establishes Db2 for z/OS and other IBM Z data as good candidates for data fabric initiatives. WKC requires DVM to expose a few z/OS data formats that it cannot handle on its own (for example, VSAM). By using WKC with DVM for z/OS, an organization can readily incorporate data that originates in VSAM.

WKC can automatically apply industry-specific regulatory policies and rules to your data. It secures data across the enterprise. WKC is an AI-augmented data catalog allowing business users to easily understand, collaborate, enrich, and access the correct data. Metadata and a governance layer for all data, analytics, and AI initiatives increases visibility and collaboration on any cloud. WKC allows masking data dynamically and consistently at a user-defined granular level. It can facilitate anonymized training data and the creation of test sets while maintaining integrity of the data.

WKC can import metadata directly from the Db2 for z/OS catalog. For non-Db2 z/OS data, it requires DVM for z/OS (see Figure 4-7). Virtual table definitions of VSAM data sets in DVM for z/OS can be imported into WKC. Virtualization of VSAM data sets is key to making them readily accessible to applications and governed within WKC. Although IBM Cloud Pak for Data can be deployed on either x86 or IBM Z, WKC at the time of writing is available only on x86.

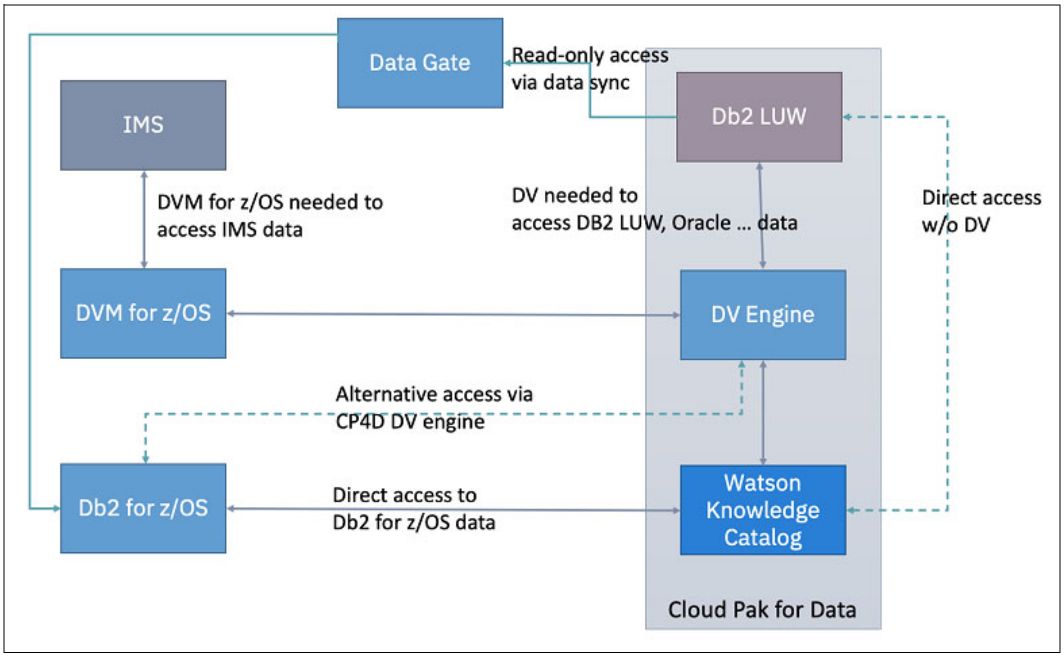

Data Gate
Read-only access via data sync
IMS
Db2 LUW
DVM for z/OS needed to access IMS data
DV needed to access DB2 LUW, Oracle ... data
Direct access w/o DV
DVM for z/OS
DV Engine
Alternative access via CP4D DV engine
Db2 for z/OS
Direct access to Db2 for z/OS data
Watson Knowledge Catalog
Cloud Pak for Data

Figure 4-7 Watson Knowledge Catalog accessing data on z/OS

4.2.2 Data governance with IBM Cloud Pak for Data

Regardless of whether a company is storing the data with one of the ecosystem data stores, or they connected to an established system with data virtualization, all this data must be governed from a central point of view. Data organization is synonymous with DataOps, which represents an automated, process-oriented methodology that improves the quality and creates a business-ready analytics foundation.

IBM Cloud Pak for Data provides automated cataloging and data privacy capabilities. Data in continuous data delivery processes such as change data capture (CDC), streaming data pipelines, ETL, and extract, load, and transform (ELT), can be “sliced and diced” in different ways from how it was originally presented. In IBM Cloud Pak for Data, WKC captures the details around both the operational DataOps work from the ETL processing, and the self-service DataOps work from the ELT processing. For both cases, the data being transformed is fully governed. With governed DataOps on the operational side and self-service on the business-user side, data flows more efficiently across an entire information value chain.

Figure 4-8 shows the organize phase from ETL processing through to self-service preparation for analyzing the data.

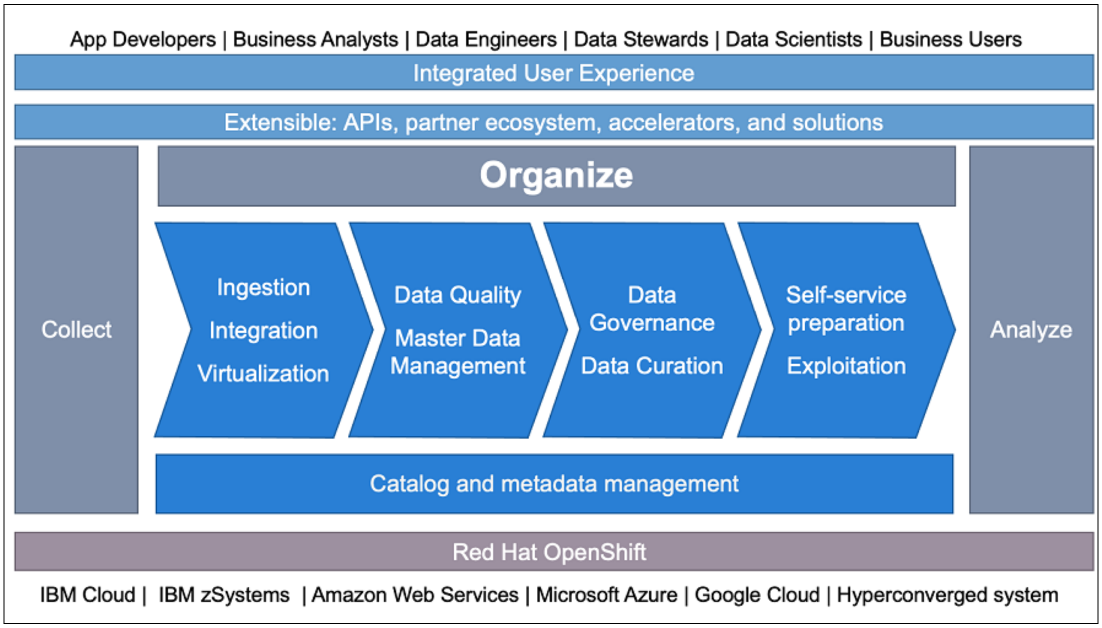

App Developers | Business Analysts | Data Engineers | Data Stewards | Data Scientists | Business Users
Integrated User Experience
Extensible: APIs, partner ecosystem, accelerators, and solutions
Organize
Collect
Ingestion
Integration
Virtualization
Data Quality
Master Data Management
Data Governance
Data Curation
Self-service preparation
Exploitation
Analyze
Catalog and metadata management
Red Hat OpenShift
IBM Cloud | IBM zSystems | Amazon Web Services | Microsoft Azure | Google Cloud | Hyperconverged system

Figure 4-8 Cataloging and data privacy of the organize phase in IBM Cloud Pak for Data

IBM Cloud Pak for Data comes with an AI-powered solution that is named AutoCatalog, which automates how data is discovered and classified to maintain a real-time catalog of data assets from across disparate data landscapes. It is designed so that teams across the business can easily find data. AutoCatalog is a critical capability of the intelligent data fabric within the platform. AutoCatalog helps overcome the challenges that are faced by managing a complex hybrid and multi-cloud enterprise data landscape, and it helps ensure that data consumers can easily find and access the correct data at the correct time regardless of location.

4.3 Achieving trusted outcomes with AIOps in an IBM Z data fabric

Another factor that slows the adaptation of AI and ML by businesses is issues with trust. Companies around the world value their reputation and must be in control of the risks after a new technology migrates from experimentation to production. How do your AI models make decisions? Is there any bias? Are the decisions fair? Can you explain transactions and perform what-if analysis? Is there a feedback loop to continuously improve the quality and accuracy? Without being able to answer such questions through processes and tools, AI and ML models that are created by data scientists and implemented by engineers would remain experiments with a degree of risk to them.

The IBM governed data and AI technology is structured on the application of our fundamental principles for ethical AI: transparency, explainability, fairness, robustness, and privacy. These five focus areas are how we define trustworthy AI. These principles, and the solutions that practice them, allow practicing AI efficiently and effectively, and doing it in a way that instills confidence in the outcome. You can analyze your AI with trust and transparency, and understand how your AI models make decisions; detect and mitigate bias and drift; increase the quality and accuracy of your predictions; and explain transactions and perform model analysis.

4.3.1 Trustworthy AI for ML models in IBM Cloud Pak for Data

Many organizations report that their data science teams have issues getting access to the correct governed data and putting their models into production. It is hard to establish trust in the decisions that AI models make and eliminate the security risks. How can such organizations confidently build an end-to-end AI workflow?

- Healthcare companies might be seeking to provide equitable care to their members while using AI and ML models.
- Large retailers with AI- and ML-empowered automated hiring might want to ensure that their practices are fair and in line with their social responsibility posture.
- Perhaps large banks are automating the otherwise manual audit process for regulatory compliance across the thousands of AI and ML models that they have.

All of these organizations are looking to use AI and ML that is trustworthy.

A data fabric and its implementation in IBM Watson Studio within IBM Cloud Pak for Data provides a solution to these challenges with three key pillars: trust in data, trust in models, and trust in process. Companies can unlock trustworthy AI by starting with governed data access for data scientists. Automated model operations (MLOps) are infused with trust throughout the entire AI lifecycle in IBM Cloud Pak for Data. Each stage of the AI lifecycle offers transparency and has monitoring to ensure proper AI governance.

The architecture of IBM Cloud Pak for Data is built with compliance and scalability that use first design principles. Model monitoring is provided for multiple aspects, including *explainability*, *draft*, *bias*, *fairness*, and `quality`. To provide end-to-end trust, data governance is integrated with data science in IBM Cloud Pak for Data.

IBM Watson OpenScale (Figure 4-9 on page 45) is an enterprise-grade environment for AI applications that provides your enterprise visibility into how your AI is built, 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.

IBM Cloud Pak for Data with Watson OpenScale can work with models that are deployed on IBM Z.

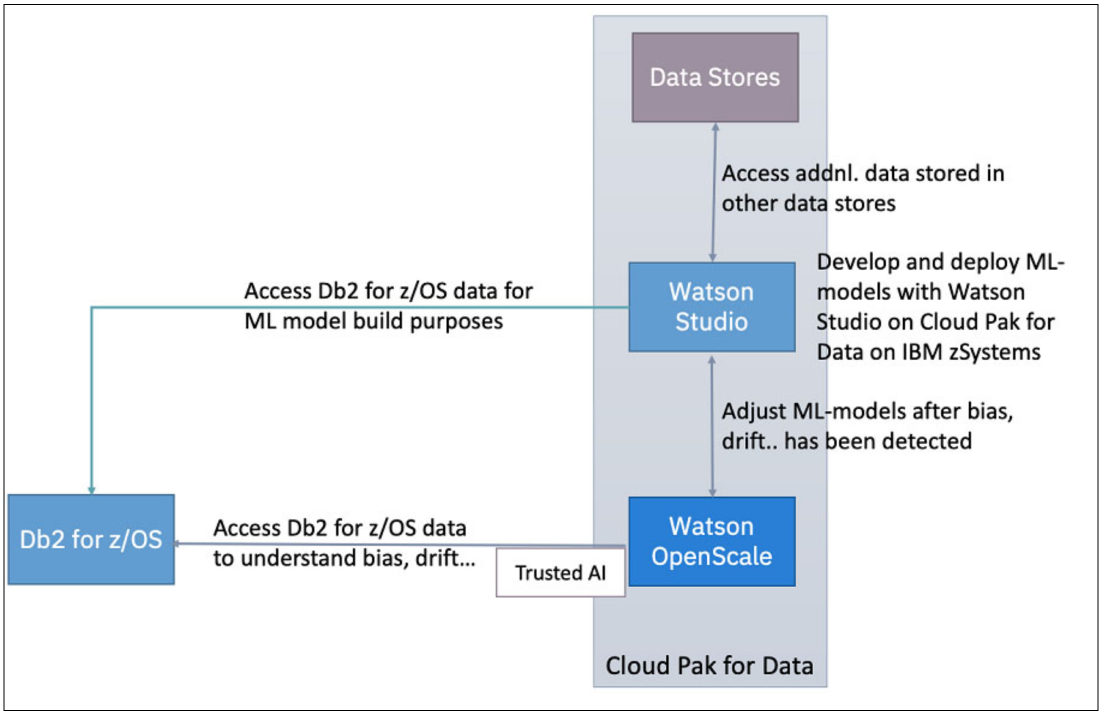

Data Stores
Access addnl. data stored in other data stores
Develop and deploy ML-models with Watson Studio on Cloud Pak for Data on IBM zSystems
Access Db2 for z/OS data for ML model build purposes
Watson Studio
Adjust ML-models after bias, drift.. has been detected
Db2 for z/OS
Access Db2 for z/OS data to understand bias, drift...
Trusted AI
Watson OpenScale
Cloud Pak for Data

Figure 4-9 IBM Watson OpenScale for Trusted AI on z/OS and in IBM Cloud Pak for Data

To deal with privacy concerns, IBM Cloud Pak for Data offers a universal data privacy framework that is named AutoPrivacy, which employs AI to intelligently automate the identification, monitoring, and enforcement of policies on sensitive data across the organization. Spanning the entire data and AI lifecycle, this framework allows business leaders to provide the self-service access data consumers need without sacrificing security or compliance. Build a better strategy for governance risk and compliance by eliminating compliance “blind spots” and minimizing risk.

5

Use cases for artificial intelligence in chatbots

Chatbots are virtual agents that communicate with customers and perform tasks by using natural language. A chatbot can be on a website, an online chat service, a mobile device, or voice-activated platforms.

With the help of artificial intelligence (AI) and natural language processing (NLP), chatbots can understand a wide range of natural language queries and contexts. Queries that are made in natural language can vary from person to person, and hardcoding the inputs and responses is not practical (for example, providing every linguistically valid way of asking ‘What is today’s special?’). AI helps to recognize user intents and extract the key entities that are mentioned in natural language input, where the language can vary substantially from the expected form. Recent advancements in NLP, such as Bidirectional Encoder Representations from Transformers (BERT) and transformer models, made identifying intent and entities in natural language more accurate.

Modern chatbot development tools do not require developers to build the language models from scratch. A developer needs to provide only a few training examples to fine-tune a built-in language model to suit their use case. For example, you can provide the chatbot an intent “pay a bill” with examples such as “I need to pay my bill”, “Pay my account balance”, and “make a payment”. The chatbot learns to generalize these examples from a sentence it has never seen, such as “I would like to pay a bill”, and will be able to identify the user’s intent from the new query. Chatbots can take over many repetitive tasks from human operators and are available 24x7.

The common tasks chatbots perform include the following ones:

- Question answering: Handle customer inquiries and provide immediate answers and information.
- Calendar manipulation: Schedule or cancel a meeting, or create an event.
- Collect information to complete a form: Can be used for onboarding, applying for a job, booking, or placing orders.
- Process customer requests: Making a payment or transaction or a refund.

© Copyright IBM Corp. 2022.

- Automate processes: Play media, or make a phone call.
- Collect feedback.
- Make recommendations.

For more information about creating intents by using IBM Watson Assistant, see the IBM Cloud® Documentation.

Some chatbots can be multipurpose or general-purpose to perform a combination of tasks. These more complex chatbots are usually organized with a decision tree, and they require the chatbot to recognize different intents from the user. The intents move the chatbot into corresponding conversation branches to handle dialogue that is related to the identified task. These chatbots use machine learning (ML) and NLP to automatically classify text into intent categories, extract the entities that contain key information, and make recommendations based on history or other user behaviors.

For most commercial chatbots, there are three important components: intents, entities, and dialogs.

AI plays an important part in discovering intents and entities. Intents are what the user asks for, for example, a query “Where can I see the menu?” can be classified into an intent “read menu”. ML models can be trained to automatically classify sentences into intents. Usually, the model must learn from at leave five real user examples of the same intent to classify an unknown sentence into the correct intent category.

Entities are collections of information that can be captured from the customer’s input. For example, an entity “City” can be a collection of city names, such as “New York”, “Melbourne”, and “Toronto”. A chatbot developer can define an entity as a dictionary, a regular expression pattern, or characterize it by using a natural language model. Entities help the chatbot capture key information and save them as context or knowledge for the chatbot to determine the best response.

Dialogs are usually in the form of decision trees or flow diagrams that represent how different scenarios of user conversations are handled. For example, a dialog can diverge into three branches when there are three intents and can further branch when more information is received from entities that are extracted from the user input. The dialogs are usually manually designed to fit a use case rather than automatically generated by AI. We highly recommend planning a conversational flow before implementing the chatbot because the flow greatly helps the chatbot to handle conversations seamlessly and to make sure that there are no loose ends in dialogs.

5.1 Customer service chatbots

Customer service can be an expensive and human labor-intensive operation, especially when there is a need to optimize service availability and customer wait time. Chatbots have the benefit of providing immediate and 24x7 support. With the help of AI, chatbots can handle many easy customer service tasks, such as answering frequently asked questions, payments and refunds, and filling out forms. Chatbots free human operators from simple and repetitive service tasks, so humans can focus on more complicated and higher value cases.

Chatbot services are also easier to scale than human customer support teams. Demand for customer service in the insurance industry, the entertainment industry, and for public sectors can be impacted by climate, festive seasons, and disasters. The ability to quickly scale up or scale down translates to cost-saving.

Recently, disasters such as the COVID-19 pandemic and brush fires have overwhelmed companies and governments with an unusual surge of questions. Waiting times were drastically increased on hotlines, but the customers still expected the same amount of waiting time before the disaster. As a result, customer service satisfaction was severely impacted. Scaling up the team of human agents to face this challenge would require a long time for recruiting and training and a large cost for paying more staff. However, chatbots are faster to build, deploy, and scale. They automatically can crawl trusted websites to stay up to date in dynamic environments, which abrogates the costs of extra staff training and logistics.

IBM Z allows customer service chatbots to scale easily, and provides chatbots a security-rich environment in which to operate.

Chatbots can be built with commercially available tools and open source tools. For example, Rasa2 is an open-source tool for building chatbots, which allows developers to use SpaCy and Natural Language Understanding pipelines to customize a chatbot. The development and testing of the chatbot can be done in any Linux environment, either inside or outside IBM Z. When the chatbot is ready to be deployed, it can be packaged into a Docker container and deployed in a z/OS Container Extensions (zCX) instance. zCX enables Linux applications to be deployed as containers on z/OS and zCX instance hosts containers that run products and applications.

Figure 5-1 shows a reference architecture for customer service chatbots on IBM Z.

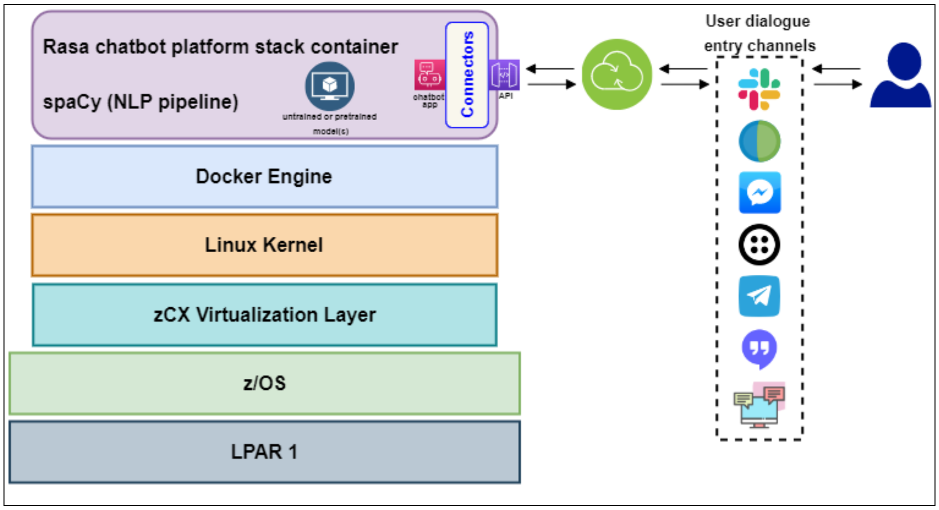


Figure 5-1 Reference architecture for customer service chatbots on IBM Z

Deploying a customer service chatbot on IBM Z allows it to enjoy the general benefits of the IBM Z environment, which brings security, scalability, high-availability and resiliency. For some organizations, chatbots need access to sensitive or privacy data, and moving them around can involve great security risks. IBM Z is one of the most secure systems in the industry.Deploying the chatbot on IBM Z means that data can stay in one place, which reduces security risks. High availability (HA) and resiliency are also important factors for customer service chatbots because their uptime is directly related to customer satisfaction. Another potential benefit is that IBM Z can reduce inferencing latency by using its on-chip AI accelerator if the chatbot uses neural network-based language models, which can give customers faster responses to chat queries.

6

Financial services sector use cases

In a highly dynamic environment such as banking and finance, where there are billions of daily transactions, it is key to have real-time knowledge of what is happening. The IBM Real Time Insights strategy is key to analyzing each of the transactions running in an organization.

As data becomes more dynamic and fast changing, there is more need to be aware of what changes are happening and react as needed. To ease these operations, the best strategy is to put artificial intelligence (AI) models and applications where the data resides. There are two main benefits for this strategy: You do not waste time and resources with copies of data, and you always act on the latest version of the data. This situation is known as *data gravity*.

Most of the fraud that is committed online are transactions running on-premises in core applications for many different industries, such as finance and banking, payments, e-commerce, or B2B, which need real-time surveillance.

IBM z16 has new features to help prevent fraud. The incorporation of Telum to on-chip execution provides computing power to run AI and machine learning (ML) models to score many transactions per second with significantly reduced latency. As fraud becomes more sophisticated, data scientists must develop more complex models, and shift from asynchronous scoring to real-time scoring as it happens.

Deployment of AI and ML models on IBM Z provides data security and a reduction of data management to maintain consistency and data quality and the best performance when using real-time evaluation while avoiding remote calls during transaction time.

© Copyright IBM Corp. 2022.

6.1 Methods to leverage AI and ML on IBM Z

There are two ideal ways to leverage AI and ML on IBM Z. Which method you use, or a setup that leverages both, depends on the use case that you need.

IBM Watson Machine Learning for z/OS (WMLz) is an end-to-end, enterprise ML platform. It helps you create, train, deploy, and monitor ML models to extract value from your mission-critical data on IBM Z while keeping the data where it resides.

A critical feature that is provided by WMLz is the inference engine, where AI and ML models are deployed to work in real time with CICS, IBM Information Management System (IMS), and any application that may request its services from anywhere in your company, and not from IBM Z. It is called the *online scoring service*, and it is depicted in the WMLz Base in Figure 6-1. IBM z/OS UNIX System Services and CICS region scoring engines are similar in that they run the same kind of AI and ML general models. At the time of writing, you are required to use IBM z/OS Container Extensions (zCX) to deploy imported Open Neural Network Exchange (ONNX)-based models on WMLz versions before version 2.3.

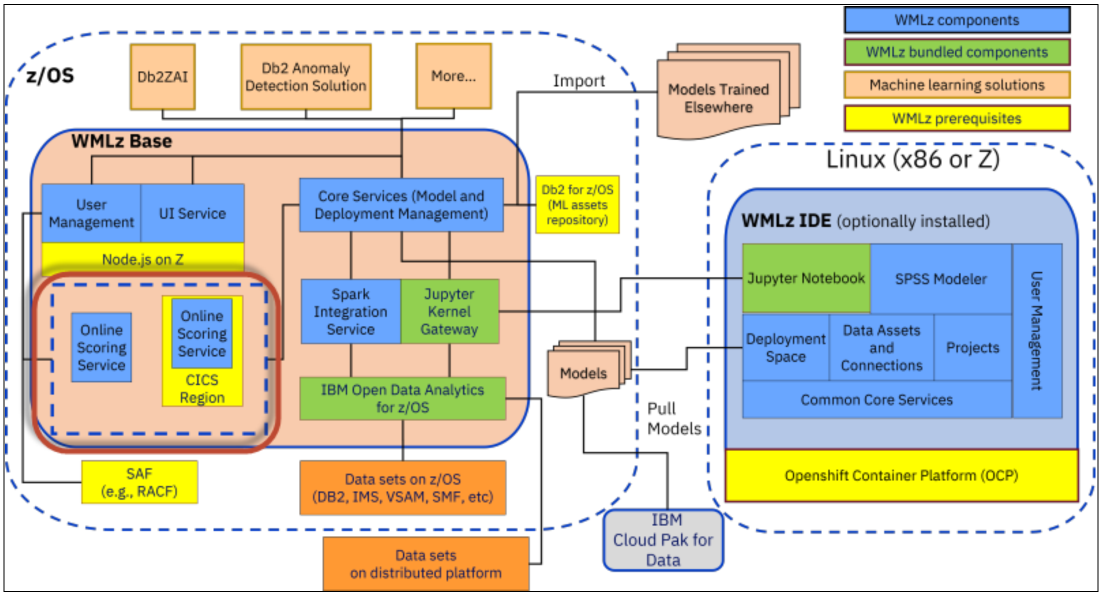


Figure 6-1 Overall architecture for WMLz 2.4

As shown in Figure 6-1, it is possible to cover the entire lifecycle of a model by using WMLz.

Online scoring services use IBM WebSphere Liberty servers that contain all the needed inbound and outbound connections to communicate with requesters and back ends. The Liberty server can be embedded in CICS regions (as shown in Figure 6-2 on page 53), which is an option that gives you the most value when transactions interacting with WMLz are running on CICS, or the server can be an independent running Liberty profile running directly on z/OS.

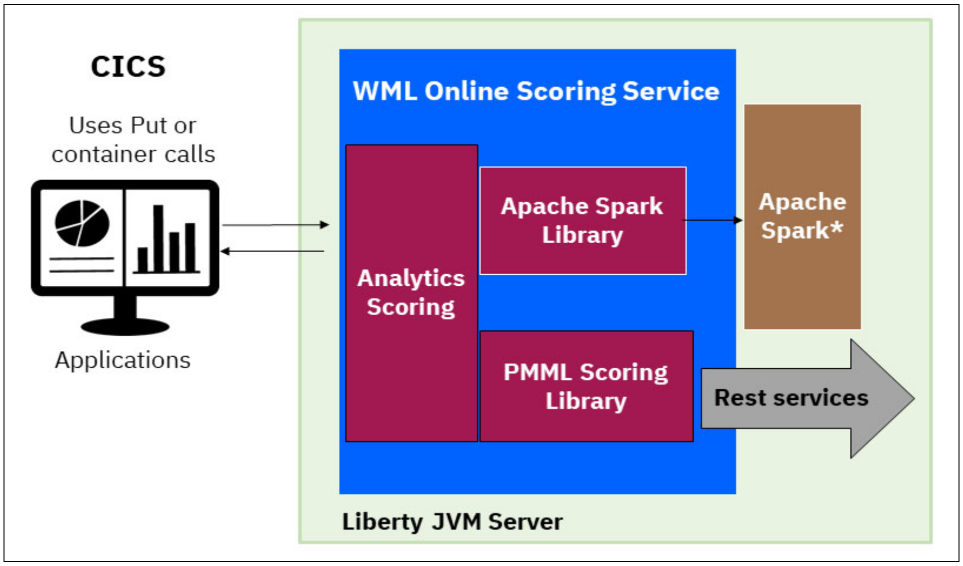


Figure 6-2 Online scoring services in WMLz

In both cases, Liberty can be run in high availability (HA) clusters, and the generation of APIs that invoke the models that are deployed in them make the APIs available for all the applications running on your enterprise, and not only IBM Z applications. API requests can be managed and routed by using z/OS Connect EE for inbound requests outside IBM Z.

Also, using zCX allows you to deploy open-source solutions that are available in a container format (for example, TensorFlow, as shown in Figure 6-3). This technology allows you to simplify deployments with memory-fast connections between the components. zCX can support HA, and the containers are automatically managed.

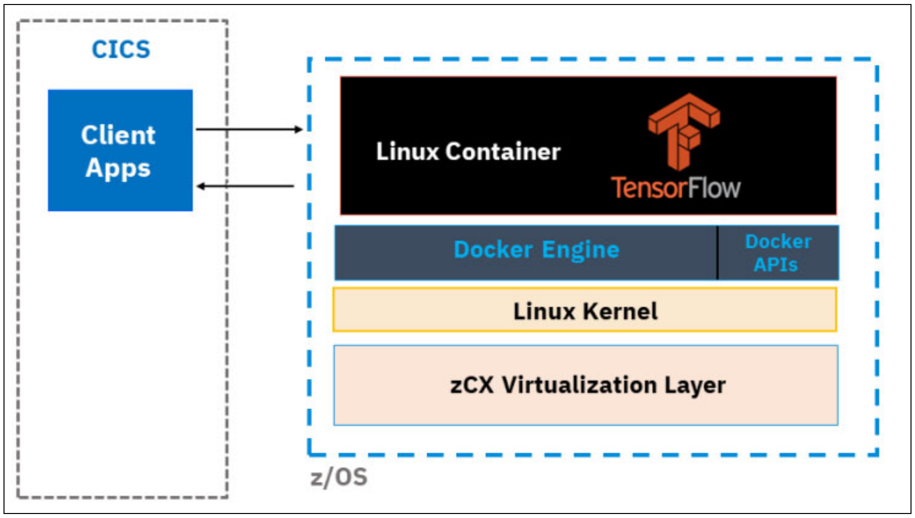


Figure 6-3 zCX with a TensorFlow container

AI and ML models can be used to enrich transaction execution and add value to the applications that already run within the core of financial institutions and banks. The overall architecture is the same because although transactions are business-oriented with some dedicated to payments and others to credits and managing accounts, the way they run is the same. They run inside a high-performing application server such as CICS, IMS, or WebSphere, and they must send requests to the AI and ML models and receive answers.

The main difference in selecting a solution is the technology that you use for the models, such as the language that is used for development and whether the modules are deployable or not when using some of the industry formats that are becoming the *de facto* standards.

6.2 Real-time fraud detection

Models are also data, and as such, business-dependent. Fraud data sets can be highly skewed. In this section, we describe an example that already is set up and deployed in production. It is an architecture that leverages patterns to prevent fraudulent transactions.

This customer moved from a static, rule-based engine to a dynamic scoring architecture that is based on self-developed models that are programmed in Go. The initial architecture for AI inferencing was built off IBM Z, but the client noticed that many transactions were running at risk because of the delays that were experienced by the off-platform AI engine. Few transactions were run with the fraud scoring done.

After moving the model to the IBM Z, the client can score every transaction within their target time of less than 5 milliseconds. Figure 6-4 shows the basic architecture setup for this transactional scoring. Banking applications start the send requests to CICS, which then initiates the transactions as usual (1). If the transaction requires AI scoring, a REST request is sent with the features vector to any of the instances of model execution (2), which then sends the data with the result back to the transaction. In this case, there is one instance of the model running on Docker containers on top of zCX, and a second instance of the same model running on a Linux on IBM Z logical partition (LPAR) inside of the same system. Both instances communicate with the z/OS LPAR by using in-memory communications, which allow the high-speed communication to preserve the reduced transactional time.

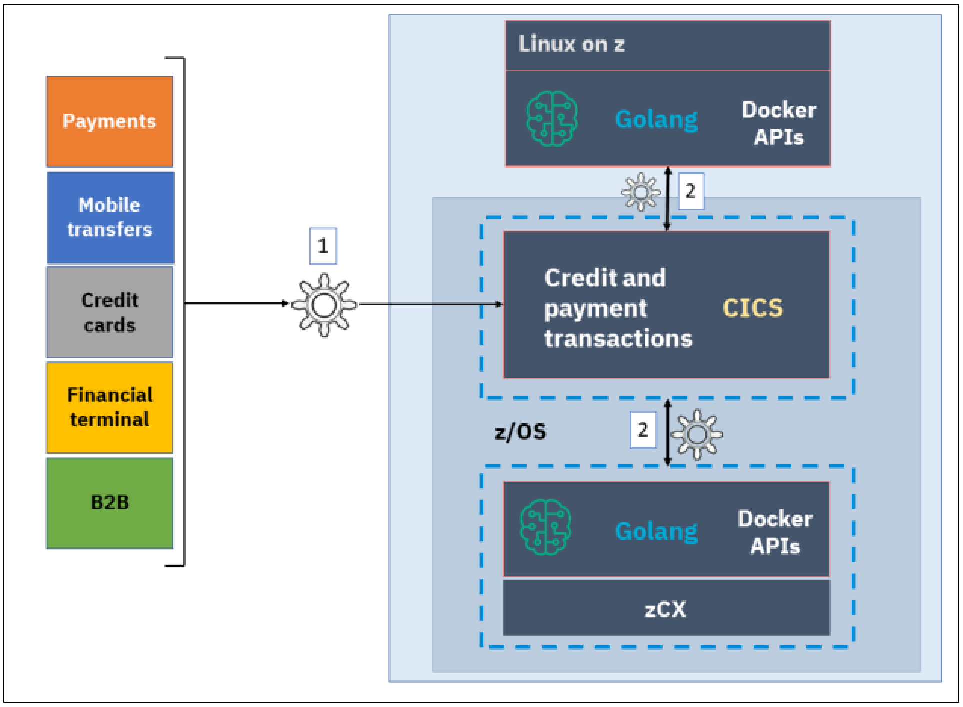


Figure 6-4 Architecture for an AI model on zCX and Linux on IBM Z

A second model of architecture that is commonly used for fraud cases relies on WMLz engines to deploy AI and ML models.

There are two options for where to deploy the scoring engine. You can see the first option, where you deploy the scoring engine in a zCX instance on IBM Z, in Figure 6-5.

The transactional system in our example is a CICS application server, but it also could be IMS or WebSphere Liberty application servers, or any other subsystem that can send and receive RESTful communications. For more information about the process that is shown in Figure 6-5, see *Optimized Inferencing and Integration with AI on IBM Z Introduction, Methodology, and Use Cases*, REDP-5661.

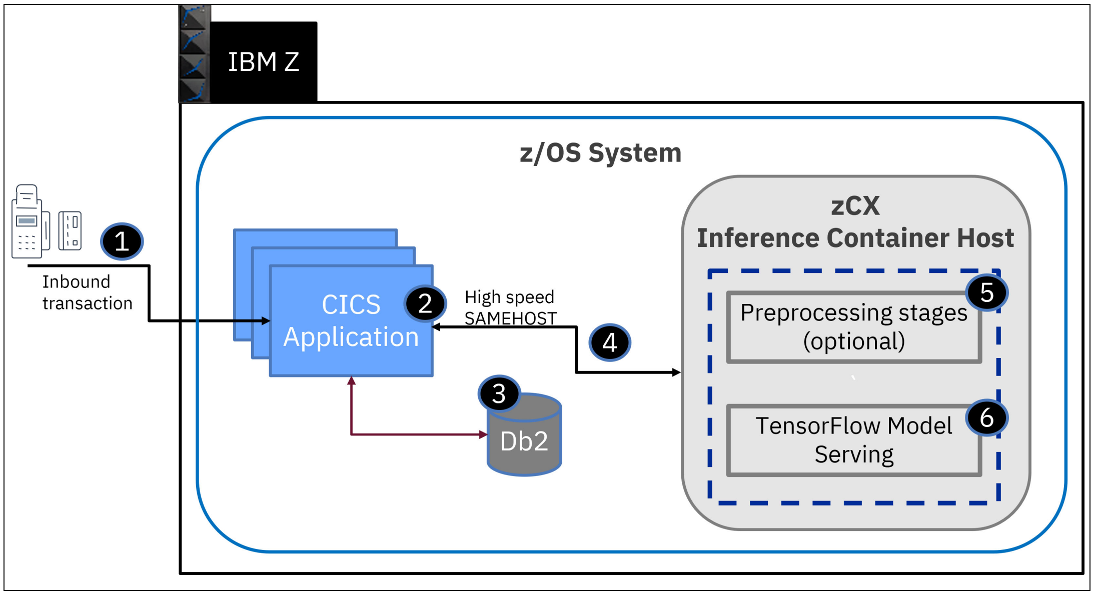


Figure 6-5 WMLz deployment option for fraud detection models

Models are frequently developed and tested in many platforms and languages, such as Python, Scala, R, and Go. Training a model can be done on any platform if you have enough computing power for complex models, but moving that model into production requires careful testing to ensure that transactions are not delayed, especially if you plan to run the model within a transaction. To ensure success, you may install IBM WMLz Online Scoring Community Edition[1] to work with the in-transaction scoring approach.

In summary, the approach to use in-transaction AI and ML assessments require the following items:

- In-platform execution of the models, which ensures that your transactions run as fast as always.
- WMLz scoring engines on CICS regions or WebSphere Liberty to deploy ONNX or Predictive Model Markup Language (PMML) models that re developed and trained off platform.
- Linux on IBM Z or zCX technologies with containers if you are using models that are not deployable with the previous formats. If the software is available for Linux on IBM Z, you can build your own containers for zCX if they already are not available.
- IBM z16, where you have on-chip AI acceleration with Telum that is embedded together with regular central processors (CPs), which provides an execution speed for your transactions that cannot be achieved by other means.

[1] https://www.ibm.com/products/machine-learning-for-zos

6.3 Credit assessments

Calculating the risk of granting a loan or credit is a complex task. It relied traditionally on scoring rules that are defined internally by each company and applied by trained people that had to decide based on rules and intuition. This process takes time and is prone to errors. Today, financial institutions are rushing to grant or deny credit on the same day that they are requested. The user experience was poor because customers had to wait days for a decision in the best cases.

The scoring process can be accelerated by helping risk assessment with AI and ML models that can analyze large amounts of data that is related to one customer. This analysis can reveal expense patterns, historical behavior, and even geopositioning for each payment. The process of granting or denying can be automated or sent to a human analyst that has more information to decide.

Figure 6-6 shows the process flow for a credit risk evaluation.

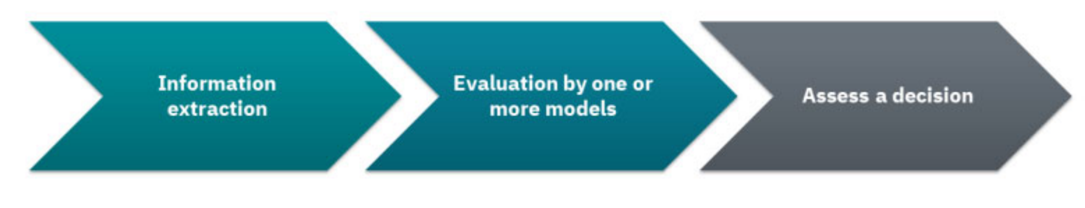


Figure 6-6 Process of a credit risk evaluation

Information extraction is aimed at studying client behavior by analyzing account and credit card records. There is much information in the quantities that are spent and in the text that is associated with each of the records of spending that you can add to the classical risk analysis variables like annual income or address. In fact, this case is a good one for mixing traditional rule-based profiling with AI and ML methods to increase accuracy. For more information about this approach, see *Machine Learning with Business Rules on IBM Z: Acting on Your Insights*, REDP-5502.

Again, there are two architectural models that you can apply to this case if you decide to use WMLz. The first one is the one that is shown in Figure 6-5 on page 55, where you can score in real time with deployed models on one of the three engines that are provided by WMLz.

A second model to evaluate credit is to use batch inference. With the less dynamic nature of this kind of assessment, you can use this feature to prepare profiles in advance and store the results in regular databases for them to be available for applications such as mobile banking or financial terminals.

The architecture model for batch inference can be seen in Figure 6-7 on page 57.

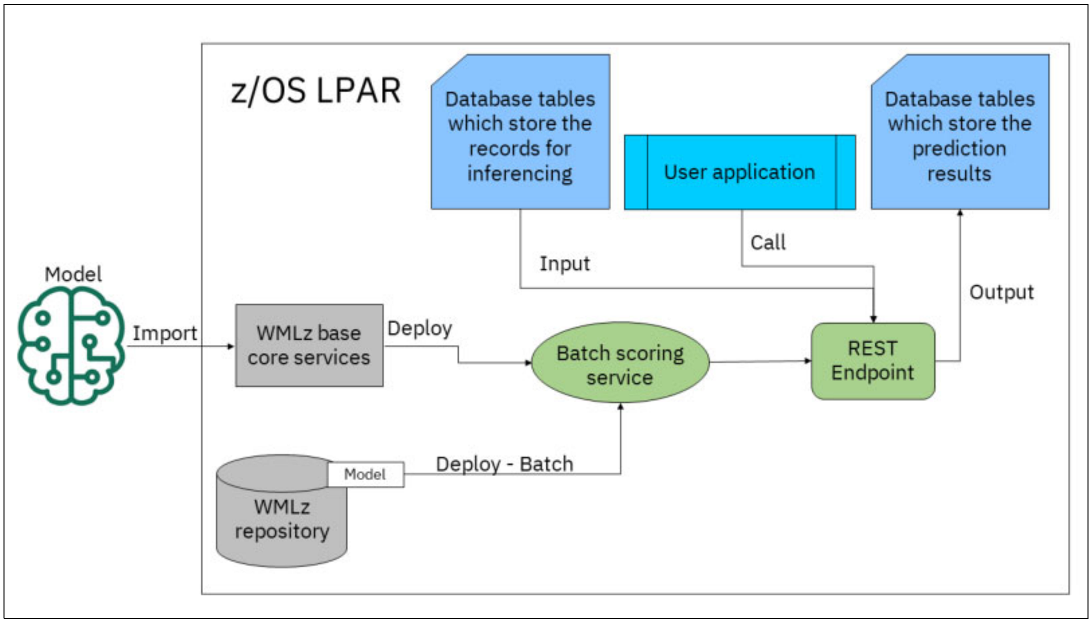


Figure 6-7 WMLz batch scoring architecture

6.4 Money laundering prevention

Money laundering is the process of processing money that is obtained through illegal activities into legitimate income. This process is an international crime and impacts both the social and economic growth of any country. The total money that is laundered annually worldwide is almost 2 - 5% of global GDP.

Anti-money laundering (AML) consists of policies that guide financial institutions in the prevention, detection, and reporting of money laundering activities. While money laundering is an international crime, many rules are local, and they can sometimes conflict with federal policies, making it difficult for financial institutions to remain compliant with rules and regulations.

In recent years, ML and AI technologies have been increasingly deployed to counter money laundering activities. Here are a few of the scenarios where AI on IBM Z can help:

- Behavioral changes: Determining and learning the transaction movement of customers with similar behaviors. ML algorithms can analyze large sets of transaction data and generate suspicious activity and anomaly reports. To perform this rigorous analysis, WMLz can be used with a similar approach to the one that is shown in Figure 6-5 on page 55. When the model flags fraudulent transactions or accounts, further SQL data insights can help you understand why the alert was raised.
- Customer insights: The Know Your Customer (KYC) process can be made more robust and faster by using a client risk profile that is based on a wide array of other external sources. Watson Knowledge Catalog (WKC) is one such mechanism.[2]
- Automating AML compliance: To perform AML compliance, a financial institute must examine large amounts of unstructured data from external sources. AI can help manage and analyze unstructured data with great speed. IBM Z Security and Compliance Center provides features for automated compliance checks, tracking dashboards, and historical compliance score analytics.

[2] https://www.ibm.com/au-en/cloud/watson-knowledge-catalog

AI helps monitor transactions, provide customer insights, and handle unstructured data. Due to the amount of data, the algorithm needs a massive data processing and storage infrastructure. Response times are critical because delays in the detection mechanism can lead to financial losses.

With AI integrated into the new Telum chip, large sets of data can be accessed in real time inside the chip, which cuts down on any time lag. So, running an AI application on IBM z16 can identify suspicious activities and prevent them from happening in real time. IBM z16 on-chip inferencing delivers up to 300 billion requests per day with a 1 ms response time and prevents fraud before it happens by scoring up to 100% of transactions in real time without impacting service-level agreements (SLAs).

6.5 Claims processing

Claims processing is one of the most important services for any kind of insurance company. Insurance companies set up a process that consists of various steps, as shown in Figure 6-8. This process needs time and effort to perform these steps. Insurance companies streamline behind the scenes by using AI to alleviate administrative and operational spending.

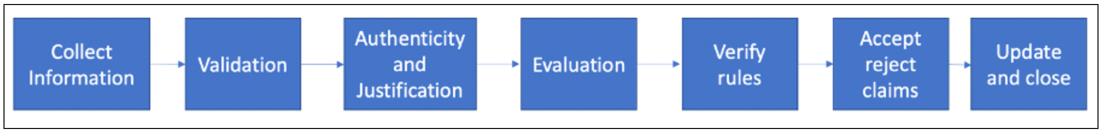


Figure 6-8 Steps for a claim process

6.5.1 Intelligent automation for insurance claims by using IBM Z

This section covers methods to automate claims by using accelerated AI on IBM Z with an architectural approach similar to what is shown in Figure 6-9 on page 59. The business goal that is enabled by intelligent automation is the reduction of the manually intensive claims handling effort by automating 75% of the claims processing steps with faster processing time, which improves the customer experience.

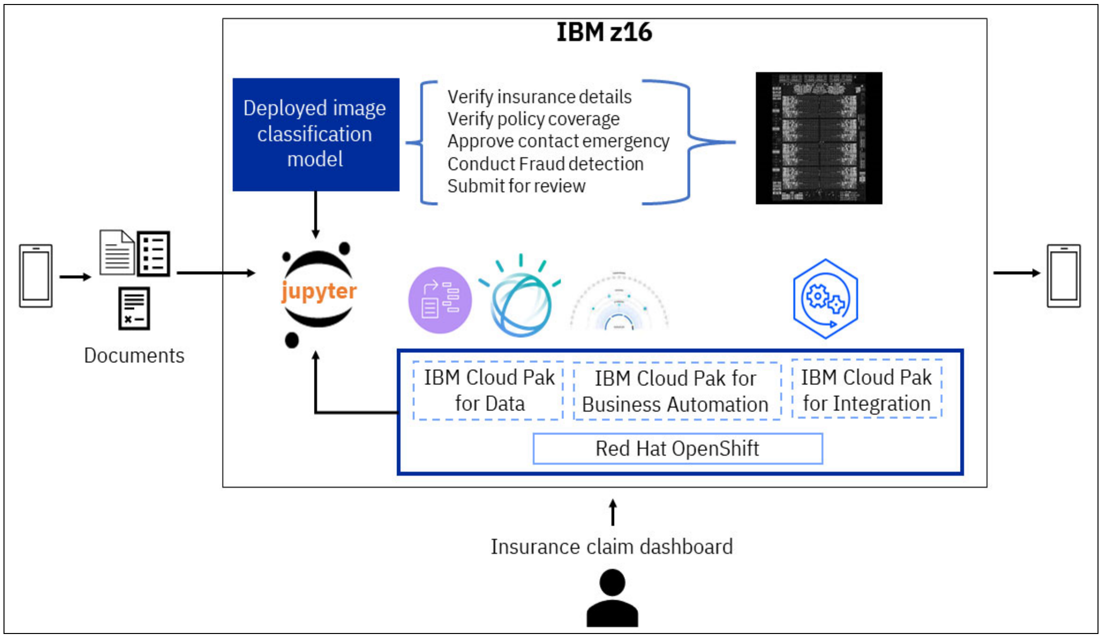

IBM z16
Deployed image classification model
Verify insurance details
Verify policy coverage
Approve contact emergency
Conduct Fraud detection
Submit for review
Documents
jupyter
IBM Cloud Pak for Data
IBM Cloud Pak for Business Automation
IBM Cloud Pak for Integration
Red Hat OpenShift
Insurance claim dashboard

Figure 6-9 Automating claims

IBM developed an intelligent workflow for processing claims by using different IBM Cloud Pak services that can be easily deployed on IBM Z by using Red Hat OpenShift.

The customer starts a claim process for an accident by using a mobile application:

- The customer provides details about insurance, location, proof of identification, and images of the accident from the accident site. This information is gathered by a chatbot.
- Completing the collection step automatically triggers a new claims process by using the IBM Cloud Paks for Business Automation capability.
- When the claims document is received and identified, text extraction is done by using optical character recognition (OCR). One of the methods to achieve OCR is by using open-source libraries like PIL and tesseract.
- Watson Natural Language Understanding, which is based on custom configuration, delivers automation by identifying entities and understanding the relevant information based on the type of documents.
- To evaluate the damage that is caused to the vehicle, a deep learning (DL) model is deployed. The model can be trained anywhere (including IBM Z), deployed and optimized in IBM Z by using ONNX and the IBM Z Deep Learning Compiler (DLC) compiler (for more information, see Chapter 9, “Retail and insurance sector use cases” on page 71).
- During the process, the application invokes a task that is called Fraud Detection, which can be performed with AI capabilities by tuning the fraud evaluation model.

IBM Cloud Pak for Integration is the “glue” that ties all the pieces of this process together by using App-Connect. API-Connect and KafkaFlow enable mobile applications and integrate with the claim dashboard, claim management application, and data center.

6.6 Clearing and settlement

Market exchanges are performed in a 3-step process that involves execution, clearing, and settlement. Millions of operations are performed every market session without any trouble for the buyer and the seller, but there are some that fail in the settlement step. This situation does not mean that those trades are not executed, but instead means that they must be analyzed to identify the problem and resolved. A team of people studying what happened and solving the issues with the operation is time-consuming, and this manual study is stressful because they need to find a solution as quickly as they can.

In this case, AI and ML are good tools that can be applied to finding the problem of why an operation was not settled by analyzing the history of failures and by using DL models. Analysts can get a ranking of reasons for the failure in a fraction of time that they would need. You can use models to analyze and predict what transactions would fail and give a list of causes and solutions for the settlement to be faster.

For this clearing and settlement scenario, with the kind of models that are needed, a good implementation would use TensorFlow. At the time of writing, TensorFlow can run on IBM Z as a Linux package, which means that the pattern to run AI and ML models must be on a Linux on IBM Z LPAR or inside zCX containers, as described in Figure 6-3 on page 53 and Figure 6-4 on page 54.

You can find an example of TensorFlow implementation with a container inside zCX on GItHub.

As with other use cases, you can combine the usage of a rule engine with AI and ML models to get better accuracy (see 6.3, “Credit assessments” on page 56).

7

Public sector use cases

The usage of artificial intelligence (AI) is growing in the public sector. Use cases range from image processing to fraud detection and decision making.

AI can help with such items as welfare payment, resource planning, and infrastructure planning. Using machine learning (ML) and automated processes to help governments with decisions allows the process to be completed much faster than manually assessing each case, which might mean that customers get much faster responses and actions for issues that are time-sensitive to them, for example, healthcare and unemployment payments. Automated decision-making reduces labor costs and allows the decision-making process to scale easily.

When using AI to make decisions in the public sector, fairness is an important factor. It is important that AI does not make decisions that are biased toward certain criteria, such as age, gender, and ethnicity. IBM Watson OpenScale provides an enterprise-grade environment to understand and analyze how AI models make decisions. Watson OpenScale provides trust and transparency to ML models by monitoring model bias and drift. Therefore, the AI-assisted decision-making process can be immediately adjusted and de-biased when bias toward any group is detected.

Fraud and anomaly detection is also a common use case for AI in the public sector. When processing applications and claims that relate to welfare, immigration, and tax, it is time-consuming to review each case manually. ML models can help human reviewers in reading the documents, forms, and processing images that attached to the applications. The ML models can use natural language processing (NLP) and image-processing techniques to detect abnormal cases based on training on past documents. Then, human reviewers can review these cases to determine whether they are fraudulent. ML is effective in automatically detecting anomalies and inconsistencies in tax returns, therefore helping detect tax evasion and fraud.

Similarly, medical and health records can be read by machines and have key information extracted from them. This information can be useful for detecting fraud in claims and planning resources and budgets.

© Copyright IBM Corp. 2022.

Monitoring and analyzing social media data by using AI also helps agencies understand the general publics' sentiment toward certain topics, such as vaccination and climate change. AI can help enforcing social media rules and policies. Because social media information is generally time-sensitive, automating the analysis with NLP allows the latest information to be processed in large quantities.

The public sector also uses AI to predict and mitigate risk. For example, AI can use climate data to predict weather events and extreme weather conditions. Governments can take measures in advance to prepare for these events and reduce a natural disaster's impact.

When deploying AI systems for the public sector, security is one of the main concerns. The data that is used for ML training and inference often can include sensitive and personal data, and bringing AI close to data in IBM Z helps minimize the movement of data and reduce security risks. To deal with sensitive and personal data, public sector users can also use IBM Data Privacy for Diagnostics (DPfD), which uses ML to identity sensitive and non-sensitive data. It tags the sensitive data and provides tools to redact it. IBM Z also provides a security-rich environment with 100% encryption of all data.

7.1 Claims fraud detection

It is labor-intensive for governments to discover fraudulent claims on unemployment payments, tax, and insurance. A state government in the US took up to 40 hours per case on determining fraudulent claims, and the process could not scale. This government realized that it needed automation to help their limited team of investigators target the correct set of transactions. By leveraging AI on IBM Z to automate the fraud detection process, the government reduced the processing time to under 1 minute per claim, which enabled the investigators to focus on higher value work. Additionally, investigators can now work on each case without having to move data or make copies, which open up security concerns.

Anomaly detection models are often used to detect fraudulent claims. There are classical ML models that can be used for anomaly detection, such as the logistic regression, decision tree, random forest, and ensemble methods. Neural network models can be effective, such as the auto-encoder and convolutional models.

Claims applications can include documents in natural language, and ML models can be trained to extract important features in free text that help identify fraud.

These ML models can be trained anywhere and deployed on IBM Z. Figure 7-1 on page 63 demonstrates the flow of deploying deep learning (DL) models such as fraud detection and image-processing models onto IBM Z.

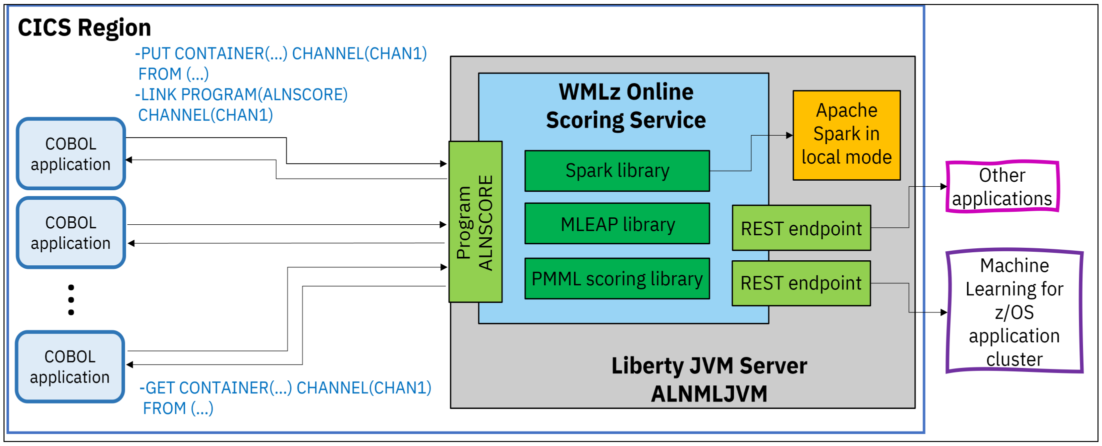

CICS Region
-PUT CONTAINER(...) CHANNEL(CHAN1) FROM (...)
-LINK PROGRAM(ALNSCORE) CHANNEL(CHAN1)
COBOL application
COBOL application
COBOL application
-GET CONTAINER(...) CHANNEL(CHAN1) FROM (...)
Program ALNSCORE
WMLz Online Scoring Service
Spark library
MLEAP library
PMML scoring library
Apache Spark in local mode
REST endpoint
REST endpoint
Liberty JVM Server ALNMLJVM
Other applications
Machine Learning for z/OS application cluster

Figure 7-1 Deploying deep learning models

After a model is trained, it must be converted to Open Neural Network Exchange (ONNX) format to be compiled by Watson Machine Learning for z/OS (WMLz). For more information about how to convert a model that built with common frameworks to an ONNX format, see the ONNX website. WMLz uses the IBM Z Deep Learning Compiler (DLC), which compiles a model in ONNX format into a WMLz model so that it can be deployed on IBM Z and be optimized for its on-chip AI accelerator. The user can import the anomaly detection model into WMLz and click **Deploy**, and a model endpoint is created automatically and made available for scoring.

AI on IBM Z supports tools and frameworks that include PyTorch, TensorFlow, Keras, Anaconda, Spark, and others.

The system architecture for the detection of claims fraud is similar to general fraud detection use cases, where a reference architecture can be found in Figure 6-4 on page 54. In this claims fraud use case, the input to the system is claims data.

7.2 Geospatial image analysis

Another use case that governments find beneficial on IBM Z is performing geospatial image analysis. A state agency in a European country wanted to automate the detection of building additions, modifications, and demolitions for land surveys and tax purposes. Deploying an object detection model to perform this task on IBM Z is a natural choice because IBM Z is a data secure platform and provides stringent access control, which is vital in ensuring proper care of sensitive aerial images. Another benefit of deploying the AI models on IBM Z is that b by using the AI accelerator, there is low to no extra cost for increasing computational power, which can help to optimize IT and operational costs.

Figure 7-2 shows an inference application on z/OS with WMLz that is used for this use case.

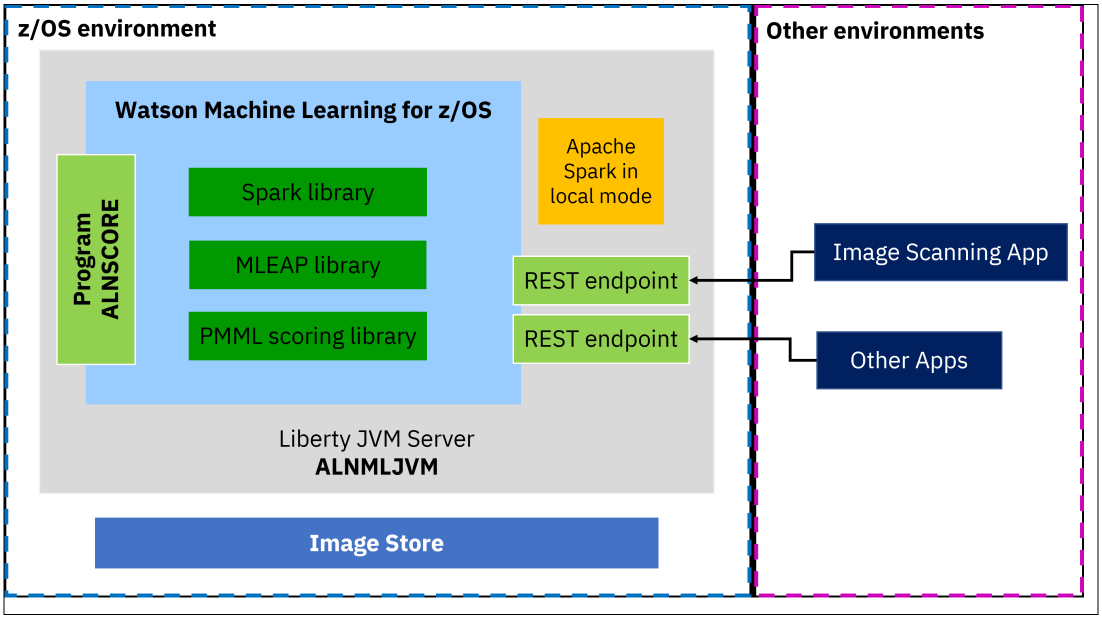


Figure 7-2 Inference application on z/OS with WMLz

The object detection models are trained with aerial geospatial images, where buildings are labeled with bounding boxes. The models can be trained both on and off IBM Z, therefore clients can choose the ideal ML package and training environment for their model and use case. Most object detection models for images are convolutional neural network models, and the inference latency for those models can be largely reduced by the on-chip AI accelerator. Like the fraudulent claims and detection use case, the models can be deployed with DLC on WMLz after being converted to the ONNX format.

8

Healthcare sector use cases

Artificial intelligence (AI) and machine learning (ML) have been used since the 1950s, but AI contributed to the healthcare industry in the early 1970s. The main challenge at that time was managing and processing the large size and volumes of data sets. This data abundance stemmed from digital advancements that led to improvements in the availability of healthcare data from various triggering sources, so the role of AI evolved. More recently, advancements in healthcare continue to evolve as enterprises invest in and incorporate ML and AI technologies into their data, applications, and processes. These AI- and ML-infused solutions continue to yield impressive business results. This chapter explores the AI trends in the healthcare sector, and the unique value proposition that IBM Z delivers to this industry.

© Copyright IBM Corp. 2022.

8.1 Changing the role of AI in healthcare

AI in healthcare now covers a wide range of modern applications, from predicting hospitalization and decision-making to classifying and helping to detect diseases and monitoring, as shown in Figure 8-1.

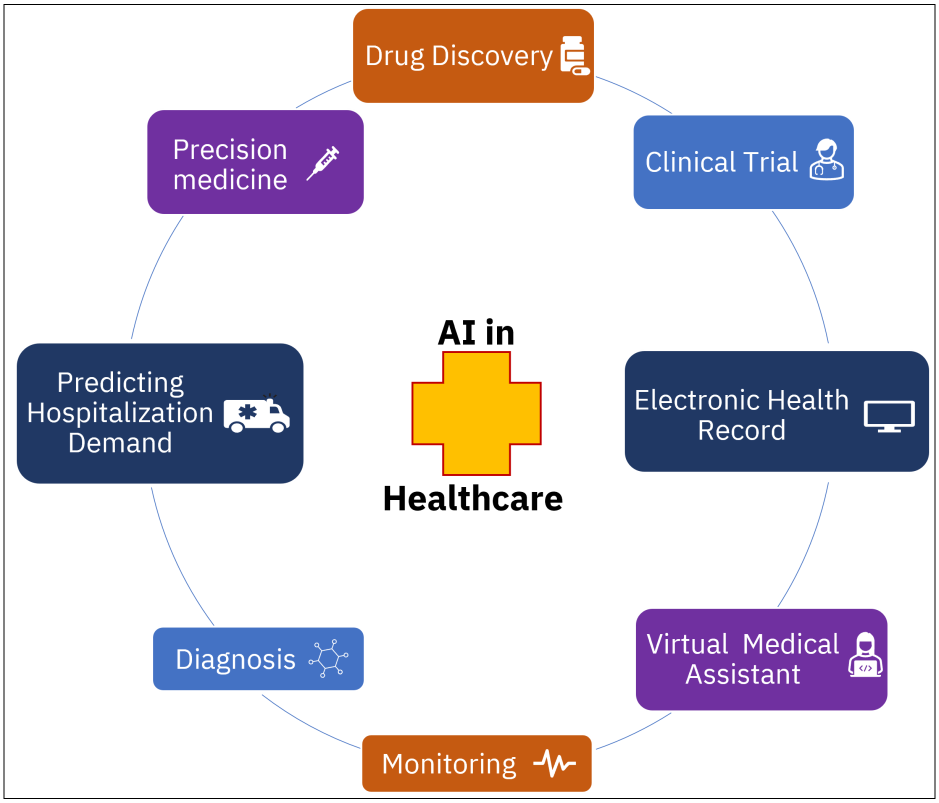


Figure 8-1 Healthcare applications in AI

AI is helping pharmaceutical companies to speed up the discovery of new drugs. The vast amount of data can be processed much faster, providing valuable insights to the researcher. Deep learning (DL) and neural networks are two of the most widely used analytical tools in drug discovery.

The way clinical trials were conducted changed significantly over the last couple of years with the adoption of AI. Medical acceptance of remote devices and Internet of Things (IoT) technology now enables drug trials to be completed virtually. Clinical trials can be made more efficient and effective by using ML algorithms like Logistic Regression.[1]

In the last decade, hospitals adopted Electronic Health Records (EHR) systems. Recently, EHR data has been used in a wide range of applications, such as hospital admission predictions, diagnosis of diseases, and risk stratification. Initially, ML was used, but now with an exponential increase in the amount of data, more projects began using DL algorithms.

Large amounts of valuable information can be extracted from clinical narratives by using natural language processing (NLP) techniques. Some reports claim that AI Assistant reviews save 18% of the time that was used to answer clinical questions compared to a standard patient record review.[2]

[1] https://www.sciencedirect.com/science/article/abs/pii/S1059131121002685

Virtual medical assistants are another evolving field that requires communication in real time and quick data insights to make better decisions.

8.1.1 Impact of COVID-19

COVID-19 has further shown us how critical it is to explore emerging technologies, develop and assimilate them into production environments, and meet real-life business requirements. AI tools and applications are expected to respond faster to this uncertain health crisis. AI was already helping the discovery of novel drugs and vaccines, but pharmaceutical companies are now under unprecedented pressure to speed up the discovery of vaccines and automate the process to better handle pandemic-like situations. Usage of telehealth care, AI-powered chatbots with clinical scenarios, and the mHealth app for clinical decision use has become more vital than ever. With limited medical resources, accurate predictions, quick diagnosis, and an enhanced monitoring system are a necessity for hospitals. Due to the devastating impact on health and economics, getting accurate projections of behavior changes and impact is important for governments to preempt and plan for contingencies.

8.2 Challenges in healthcare data processing

The current challenges in running AI in the healthcare sector include the following items:

- Computing complex data, such as EHR, which consists of medical history, radiology images, laboratory test results, treatment plans, medications, and speech data.
- Running DL algorithms on high-resolution images, such as Optical Coherence Tomography (OCT), Digital Imaging and Communications in Medicine (DICOM), and the Neuroimaging Informatics Technology Initiative (NifTi).
- Running DL models with a large amount of data needs more resources and consumes much energy and processing time. One of the solutions is using smaller data sets, but it might impact the accuracy of models and provide incorrect results, which make the whole experience futile, and can even have adverse effects on medical management.
- Real-time computation and insights for monitoring for Health Monitoring Systems, which can compute in real time and notify back to doctors.
- Compatibility of systems with an open-source framework.
- Data availability is important for mission-critical applications, and most healthcare delivery and support demands it.
- Scalability is important for mission-critical applications, especially to handle pandemic-like situations.
- Data security of healthcare data is kept confidential and sensitive and should be stored and shared carefully.

[2] Chi, E. A., et al, “Development and Validation of an Artificial Intelligence System to Optimize Clinician Review of Patient Records”, JAMA network open, 4(7), e2117391, 2021, found at: https://doi.org/10.1001/jamanetworkopen.2021.17391

8.3 Efficacy in data processing with IBM Z

DL is one of the most powerful tools that are used in neural networks. Processing of the model might take a long time if it is not optimized. The IBM Deep Learning Compiler (DLC) verifies that the code is compatible to run on the IBM Z platform and optimized for it, which helps to boost performance and get results faster.

Medical data contains various types of reports where most of the data is unstructured and complex. These kinds of data complexity and structure challenges can be solved by using IBM Cloud Pak for Data and its supported integrations. Specifically, when looking to integrate complex and disparate data sources on IBM Z, you can leverage IBM Data Virtualization Manager for z/OS in IBM Cloud Pak for Data.

To handle sensitive data, the IBM Z platform provides the IBM Data Privacy for Diagnostics (DPfD) tool, which uses ML internally and provides the capability to redact patient information.

More benefits of using IBM Z for your data processing needs include the following ones:

- Real-time insights on IBM Z that support healthcare monitoring can be gained by using in-memory computing power in a secure environment on the platform.
- Scaling up is one of the major challenges in digital healthcare AI models and applications.[3] To help overcome this challenge, IBM Z has a built-in fabric.
- The Open Neural Network Exchange (ONNX) is an open source for ML interoperability. ONNX provides a format that helps you reuse trained neural network models across multiple frameworks. IBM Z supports ONNX, which further helps to improve the performance of ML models.
- IBM Z is equipped with pervasive encryption technology to protect your sensitive data and manage privacy based on your policies.
- IBM Watson Machine Learning for z/OS is an ideal solution to run an end-to-end ML cycle, that is, create, train, and deploy ML models to extract value from your mission-critical data on IBM Z while keeping the data where it resides.

8.4 Running data processing on IBM Z

This section describes a hospital use case for diagnosing a disease by running DL model inference on MRI scan images, and then predicting when patients will visit a hospital to estimate the demand for hospital services.

8.4.1 Background and objective

A neurological hospital has large sets of MRI scans, images of the brain, and medical reports that provide details about patients. The goal is to train and develop a DL model for images and deploy it on IBM Z so that this model can run for every new patient getting an MRI scan, which helps doctors with diagnosis.

[3] Xu, J., et al., “Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives”, Hum Genet 138, 109–124 (2019), found at: https://doi.org/10.1007/s00439-019-01970-5

The hospital wants to perform analyses that are based on the EHR and medical reports of the patient. These reports consist of some structured and unstructured data. Because the data is sensitive, the hospital does not want to migrate data to perform data analytics activity, and they want to save data in encrypted form.

Architecture

The use case consists of the following two objectives:

1. Image processing by using DL and deploying it on IBM Z.
2. Predicting hospital demand by using ML to analyze EHR and medical reports, which are on IBM Z, and run ML algorithms to predict when patients will visit the hospital.

Figure 8-2 shows the architecture diagram to run this scenario on IBM Z.

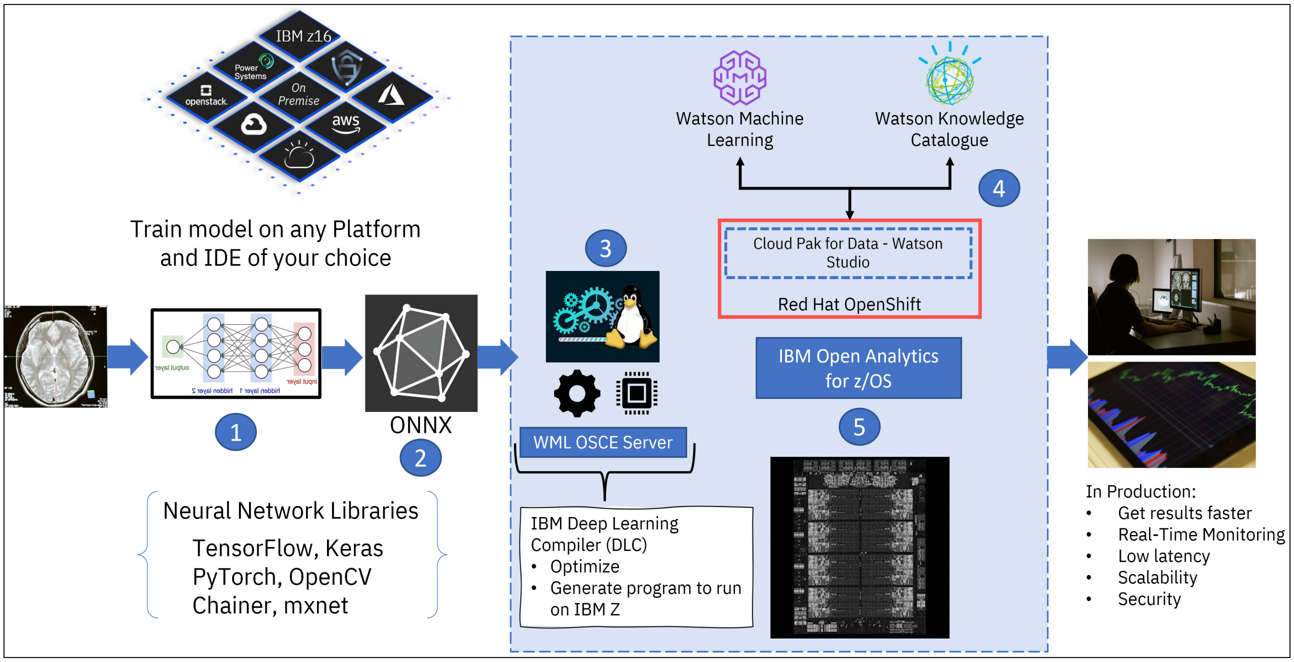


Figure 8-2 Data processing on IBM Z

Image processing with deep learning

While training and developing the model, the hospital had the flexibility of training the model anywhere or by using any integrated development environment (IDE) on Linux (either Linux on IBM Z or LinuxONE) on IBM Z.

When the model is trained, the next step is to convert the trained model to the ONNX format, for example, by using the ONNX-MLIR open-source compiler project.

The IBM Watson Machine Learning for z/OS (WMLz) Online Scoring Community Edition (OSCE) server can import the trained model, which uses DLC technology to compile and deploy the compiled model to the server. For each deployed model, a REST API is generated, which can be used for online inferencing of the model.

Because the data is sensitive, IBM Z provides a secure environment to train and run the model. Training large data sets requires a high-computing resource, and with in-memory computing power, IBM Z can generate faster results.

Predicting hospital demand by using ML

To extract meaningful data from EHR and the medical records of the patients visiting the hospitals and external information like weather conditions, use IBM Watson Knowledge Catalog (WKC). WKC helps to catalog and analyze assets, including ML models and structured and unstructured data.

IBM Open Data Analytics for z/OS and the Watson Machine Learning capability on IBM Z can be leveraged to get the analytics results within IBM Z itself at run time, which eliminates data migration of sensitive data. The unparalleled pervasive encryption that IBM Z provides with secure data practices in the enterprise help protect data from cybersecurity threats.

9

Retail and insurance sector use cases

Consumer diversity and market conditions are pushing the retail industry to a “world of extremes” that creates clear winners and losers. In particular, the consumer marketplace is being shaped by the following five deep-seated trends:

- Customers fragmenting into micro-segments as a result of pronounced shifts in demographics, attitudes, and patterns of behavior. Furthermore, they are “trading down” to low-cost commodities on one end and “trading up” to high-value, premium brands and companies at the other end. Retailers operate in a world where there are no established norms or serving the needs of the “average” customer.
- Overwhelmed and time-strapped customers are seeking greater control over their interactions with businesses. Empowered by new technology and regulation, they actively shield themselves from “me-too” marketing tactics. Only retailers offering differentiated, relevant value gain access to customers’ mind-share and personal information.
- Customers are increasingly empowered by unparalleled access to information virtually wherever, whenever, and however they want it. Retailers must provide value propositions and shopping experiences that keep customers coming back even in a world of total information transparency.
- Mega-retailers break the boundaries. The world’s top retailers are rapidly expanding across the world, channel formats, and product and service categories, blurring market segments and devouring market share. Aspiring “mega-retailers” must discover how to maintain growth at ever-increasing levels of scale and complexity. Meanwhile, competitors must successfully differentiate themselves to survive.
- Partnering becomes pervasive. Competition is no longer a solo game. Leading retailers are morphing their enterprises into flexible “value networks” that are based on strong integration and collaboration with alliance partners. Industry competitors face increased pressure to match the responsiveness and agility of these connected and mutually dependent business models.

© Copyright IBM Corp. 2022.

The fundamental challenge for retailers is to become truly customer-centric in strategy and execution by acting on the following four key strategic imperatives:

- Craft an exceedingly focused, distinctive brand proposition. Build a clear position in the minds of customers
- Drive customer-valued innovation through deeper insight. Bring successful innovations to market before the competition by using new tools, techniques, and information sources to understand the true drivers of customer needs and preferences.
- Optimize core activities through systematic intelligence. Improve performance in core functions (for example, merchandising, pricing, and HR) by augmenting traditional practices with advanced analytics.
- Realign the organization to use customer centrality. Drive change into all the basic building blocks of the organization to achieve true customer focus. Shift from disconnected multi-channel operations to a unified orchestration of the customer experience.

9.1 Customer profiling based on shopping behavior

Customer profiling always has been central to effective brand positioning and market strategy. Today, customer profiling takes on a whole new role. Consumers no longer accept blanket messaging. They expect more from brands, that is, they want a personalized experience at every touch point that speaks to them alone. The demand for brands to tailor marketing in this way has spurred the need for in-depth data that goes far beyond basic demographics, developing consumer and customer profiles based on their perceptions, interests, attitudes, and behaviors.

Access to this data eliminated the need for guesswork and paved the way for more data-driven creativity and decision-making.

Most people are now blocking advertisements on their mobile and desktop devices. Online shopping is the mainstream activity for online consumers, and most people opt for multi-device approaches to purchasing. The average internet user now has up to eight social media accounts.

These facts tell us much. As digital consumer become more fragmented and challenging to reach, they are taking more control. With the power to choose what advertising and marketing that they are exposed to and when they are exposed to it, they are demanding more of what they want from brands, and less of what they do not want.

However, there is not less potential for marketers to reach them. In fact, the situation is quite the opposite.

With more consumers now taking a multi-channel, multi-platform approach and spending increasing amounts of time online, the many ways to connect with them is greater than ever. The many ways to connect with consumers exemplify the need for a more consumer-centric approach to marketing and customer profiling that puts authentic branding at the forefront and respects an individual consumer’s right to tailored content that they genuinely value.

Customer profiling is the only way to gather the insights that are needed to identify, segment, and define target audiences. Going far beyond basic demographics, this profiling means getting as close to your consumer as possible so that you can reach them in the correct way.

But the answer does not lie in extensive research: It lies in good quality insight. There is a difference.

Customer profiling is about creating value from data to understand everything that there is to know about your target consumers and the market that surrounds them. Leading brands are putting insight in the driver's seat to put consumers at the heart of messaging, guiding everything from campaign planning to brand positioning.

This process starts with focusing on current customers, followed by the target audience and target market so that you get the validation that you need that you are looking at the correct people. Then, it is about understanding and defining these consumers to get the messaging correct, and focusing on the ideal customer insights that can drive meaningful creativity.

9.1.1 Know Your Customer

Know Your Customer (KYC) information typically serves the following three purposes:

- Industry regulations and compliance requirements to detect and prevent financial crimes.
- Opportunity identification, such as selling more products or identifying and moving unprofitable customers.
- Understanding the customer to provide appropriate customer service across all products and relationships with the organization versus siloed data and relationships.

Figure 9-1 shows the relevant business data domains that source supporting data.

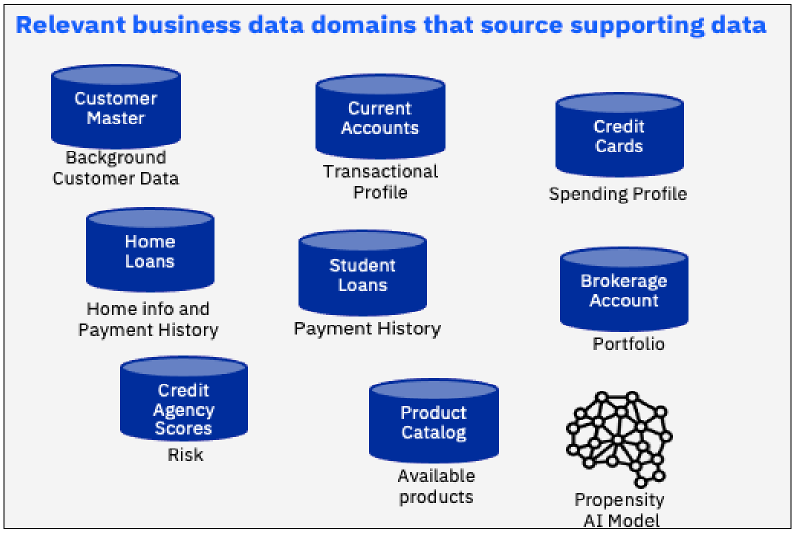


Figure 9-1 Relevant business data

Assuming that the data sources are Virtual Storage Access Method (VSAM), IBM Information Management System (IMS), Db2 for z/OS, IBM Db2 Distributed, Oracle, and MongoDB, we provide the following example:

Identify the client relationships' participation across the portfolio of products. Firms struggle to connect customer and product information because they are protected by lines of business (LOB) silos. To do this task, complete the following steps:

1. Browse the IBM Z data and virtualize or replicate only what is needed.
2. Create global connections to Db2 for z/OS, IBM Data Virtualization Manager (DVM), and Db2 for Linux, UNIX, and Windows (LUW).
3. Use IBM Db2 for z/OS Data Gate (Db2 Data Gate). The differentiator of Db2 Data Gate is that there is no impact to existing applications on IBM Z when building the use case with Db2 for z/OS data.

4. Catalog browsed connections to understand the data. Use the IBM Watson Knowledge Catalog (WKC) to provision data sources in Db2 for z/OS and other data sources in IBM Z through DVM.
5. Apply protection rules to protect sensitive data. Use WKC (Privacy) with Db2 for z/OS and DVM.
6. Virtualize data from VSAM, Db2 for z/OS, and Oracle. Use Data Virtualize through Db2 for z/OS, DVM Oracle, and MongoDB.
7. Discover quality and business terms and publish to the catalog. Use WKC (Quality) through the Data Virtualize tables.
8. Consolidate the information for the same client from disparate systems. Use IBM Match 360 directly from tables while accessing base Db2 for z/OS tables or virtual tables through DVM.

9.2 Individualizing pricing

The days of traditional insurance are fading as technology transforms insurance products and services and their delivery method. Some customers say that their insurance providers do not offer any customization. Many customers are willing to trade their behavioral data in exchange for lower premiums and quicker settlements. Even more customers are craving personalized offers.

When offering personalized insurance, the focus must be on core growth areas, including customer experience, product development, and improved online services.

9.2.1 Customer experience

With customer experience, it is important to learn what the customer’s needs are and then use that data to improve the customer experience through personalization and recommendations. Personalization includes things like finding out what a particular customer, or a segment of customers, want from your business and other businesses.

Here is a list of questions to consider when seeking to understand your customer’s needs and wants:

- Do they want more payment choices?
- Do they want a dedicated call center?
- Do they want advice about how they can keep their premiums down?
- Do they find your billing process to be long-winded?

9.2.2 Product development: Insurance package personalization

Because different industry subdomains are distinct from one another, their approach to personalization is diverse. Consider the following two cases of product personalization in the insurance sector: health insurance personalization and car insurance personalization.

Health insurance personalization

As with any personal targeting, people-first healthcare solutions need person-specific data, which means obtaining an individual’s email, phone number, and residential address, and it might include requiring them to use wearables and smart devices to collect real-time health data, such as their heart rate. Smart devices and other wearables that send data to insurers are key to providing a real-time insurance experience.

When a patient's vital signs indicate the emergence of a serious condition, it is possible to leverage the data that is collected by these devices and offer early intervention recommendations or even immediate treatment. Additionally, some companies create and send out personalized communications and recommendations to their members on how to stay healthy. The content can be in the form of videos on diabetic prevention, fitness membership discounts, medication consumption reminders or tracking, and even targeted prevention campaigns.

Car insurance personalization

Car insurance personalization aims to provide coverage that is based on a customer's own unique needs and associated risk profile. This personalization approach diverges from the traditional one-size-fits-all policy allocation that is often perceived as unfair. Insurers can create a custom policy that is based on an individual's driving record, annual mileage, expectations, car model, and age. Newer car models are equipped with telematics devices with artificial intelligence (AI) capabilities to deliver accurate data collection of factors, such as driving speed, distance that is covered, current location, and other items to further strengthen the integrity of the data that is collected by the insurer.

Improving online services

Through connected experiences, some companies allow their customers to file insurance claims or request some level of assistance through AI assistants. For example, for a car insurance claim, the customer might simply send pictures of their car damage and the assistant guides them through the filling process, verifying the legitimacy of the case, and initiating next steps, such as dispatching funds.

9.2.3 Implementing individualization on IBM Z

All these individualizations require rule engines to be in place. The decision logic is generally hidden in the code of one or more applications, and potentially across different platforms and development teams. Over time, as changes are added to the business policy, the code becomes more complex, making changes and auditability increasingly difficult. A good solution is to use decision automation to externalize the logic that implements the decision from the application (see Figure 9-2). This solution allows for changes to the business policies to be made in a much more focused and agile way that is separate from traditional application changes.

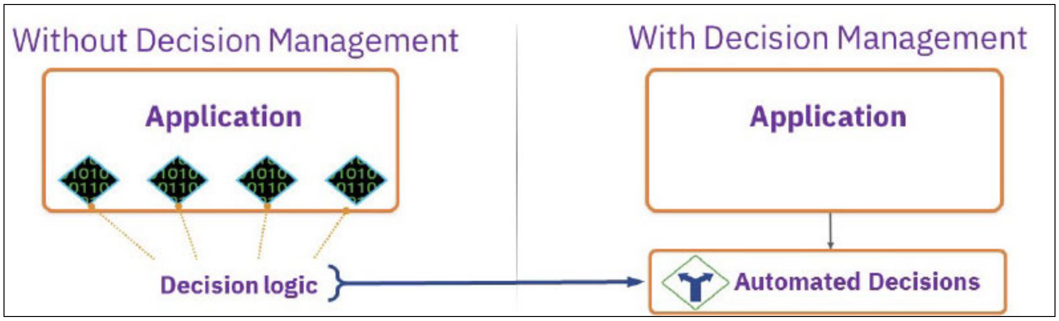


Figure 9-2 Decision externalization

IBM Operational Decision Manager for z/OS is a comprehensive decision automation platform that helps you capture, analyze, govern, and automate rules-based business decisions with optimized runtime options for integration with z/OS applications. Through the integration of decision management capabilities and machine learning (ML) technology, you can achieve smarter decisions.

Figure 9-3 depicts the simplification of architecture that is enabled by the Operational Decision Manager (ODM) to facilitate the orchestration, rules, and data for intelligent decision services. The client application interacts with the Decision Services Orchestration component in ODM, which then orchestrates the connection to the data preparation services, business rules execution services, and the predictive model execution services to analyze every incoming transaction.

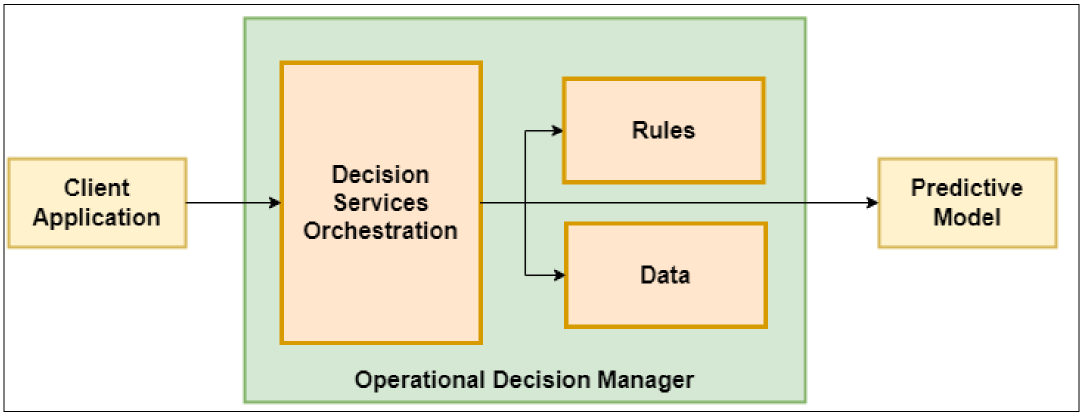


Figure 9-3 Operational Decision Manager

Here are sample insurance use cases and solutions that can help to realize a faster return on investment with ODM for z/OS:

- Automation of underwriting decisions in the insurance industry.
- Automation of claims processing in the insurance industry.

If you compare “ODM use cases” and “ML use cases”, there is a large amount of overlap.

The approach that is taken by business rules in making decisions is prescriptive. In the prescriptive approach, explicit rules are evaluated against the incoming data. Different algorithms might determine how exactly those rules are applied, but it is a deterministic approach because someone who knows the rules can know what response the rules return. This approach is both the strength and the weakness of business rules.

ML uses past data to take a predictive approach and make a calculated guess about a potential future occurrence. There is great power in this approach when you have the historical data to use as a unique data source for future decisions.

Prescriptive and predictive approaches can complement each other.

The solution architecture can be simplified when you allow ODM to handle the orchestration, rules, and data. The client application that runs in CICS or IMS on z/OS interacts with a component called “Decision Services Orchestration.” This component is responsible for orchestrating the calls to the following three services:

- Data Preparation
- Predictive (“Scoring”) Model Execution
- Business Rules Execution

Figure 9-4 on page 77 shows an optimized infrastructure solution that is based on the following components:

- IBM Machine Learning for the scoring engine
- ODM on z/OS for the business rules management system

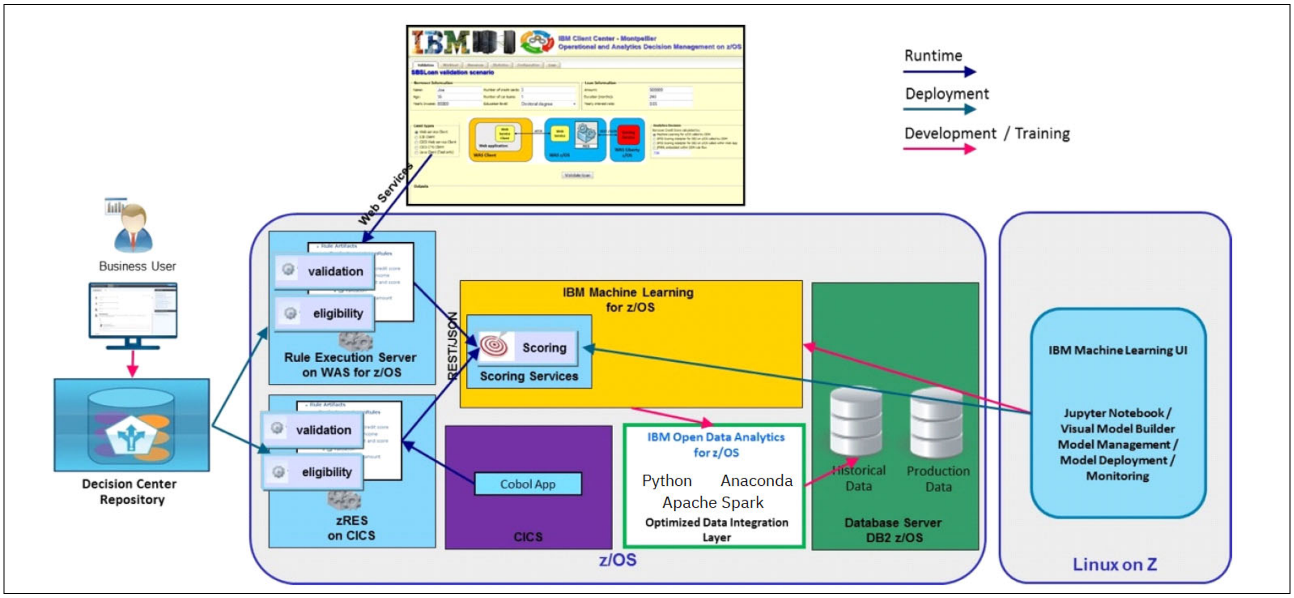

IBM Client Center - Montpellier
Operational and Analytics Decision Management on z/OS
SBSLoan validation scenario
Runtime
Deployment
Development / Training
Web Services
Business User
Decision Center Repository
validation
eligibility
Rule Execution Server on WAS for z/OS
REST/JSON
IBM Machine Learning for z/OS
Scoring
Scoring Services
validation
eligibility
zRES on CICS
Cobol App
CICS
IBM Open Data Analytics for z/OS
Python
Anaconda
Apache Spark
Optimized Data Integration Layer
Historical Data
Production Data
Database Server DB2 z/OS
z/OS
IBM Machine Learning UI
Jupyter Notebook / Visual Model Builder Model Management / Model Deployment / Monitoring
Linux on Z

Figure 9-4 Optimized infrastructure

In this example, Decision Service Orchestration is implemented in ODM, and ODM calls the ML scoring service running on z/OS.

The combination of increased requirements and the development of advanced new technologies has brought forth a new era: Scoring is based on ML and business rules systems.

Improved accuracy and speed with systems that can constantly analyze feeds of data and events reveal the risks and opportunities, and let you make real-time decisions.

By combining these two disciplines, you get the best of both a deterministic and a probabilistic approach. As a result, you can make the best decisions with the data at any specific point in time.

Real-time authorization and decisions that apply thousands of data points and sophisticated modeling techniques help increase the accuracy of approvals of genuine transactions.

For more information about the technical specifications about leveraging ML for business rules on IBM Z, see *Machine Learning with Business Rules on IBM Z: Acting on Your Insights*, REDP-5502.

Abbreviations and acronyms

AGI	artificial general intelligence
AI	artificial intelligence
AIOps	artificial intelligence operations
AML	anti-money laundering
ANI	artificial narrow intelligence
ASI	artificial super intelligence
BERT	Bidirectional Encoder Representations from Transformers
CCPA	California Consumer Privacy Act
CDC	change data capture
CF	Coupling Facility
CP	central processor
DataOps	data operations
DBAT	database access thread
DDF	Distributed Data Facility
DICOM	Digital Imaging and Communications in Medicine
DL	deep learning
DLC	IBM Z Deep Learning Compiler
DPfD	Data Privacy for Diagnostics
DR	disaster recovery
DVM	Data Virtualization Manager
EDW	Enterprise Data Warehouse
EHR	Electronic Health Records
ELT	extract, load, and transform
ETL	extract, transform, and load
FCRA	Fair Credit Reporting Act
GDPR	General Data Protection Regulation
GRS	global resource serialization
HA	high availability
HIPAA	Health Insurance Portability and Accountability Act
HTAP	Hybrid Transaction Analytical Processing
IA	information architecture
IBM	International Business Machines Corporation
IDE	integrated development environment
IDMS	Integrated Database Management System
IMS	Information Management System
IoT	Internet of Things
ISV	independent software vendor
IZLDA	IBM Z Operational Log and Data Analytics
KPI	key performance indicator
KYC	Know Your Customer
LOB	lines of business
LPAR	logical partition
LUW	Linux, UNIX, and Windows
ML	machine learning
MLOps	model operations
MTTI	mean-time-to-identify
MTTK	mean-time-to-know
MTTR	mean-time-to-resolve
NifTi	Neuroimaging Informatics Technology Initiative
NLP	natural language processing
OCR	optical character recognition
OCT	Optical Coherence Tomography
ODM	Operational Decision Manager
OLAP	online analytical processing
OLTP	online transaction processing
ONNX	Open Neural Network Exchange
OSCE	Online Scoring Community Edition
PMML	Predictive Model Markup Language
PoC	proof of concept
RBAC	role-based access control
SLA	service-level agreement
SME	subject matter expert
SMF	System Management Facility
SOR	System of Record
SRE	Site Reliability Engineer
VSAM	Virtual Storage Access Method
WKC	Watson Knowledge Catalog
WLM	Workload Manager
WMLz	Watson Machine Learning for z/OS
ZAA	IBM Z Anomaly Analytics
zCX	z/OS Container Extensions
zIIP	IBM Z Integrated Information Processor

© Copyright IBM Corp. 2022.

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics that are covered in this paper.

IBM Redbooks

The following IBM Redbooks publication provides more information about the topics in this document. The publication that is referenced might be available in softcopy only.

- *Machine Learning with Business Rules on IBM Z: Acting on Your Insights*, REDP-5502

You can search for, view, download, or order this document and other Redbooks, Redpapers, web docs, drafts, and additional materials at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- BBVA reduces CO2 emissions and energy consumption of its Data Center processors by 50% with IBM technology.

  https://www.bbva.com/en/sustainability/bbva-reduces-co-emissions-and-energy-consumption-of-its-data-center-processors-by-50-with-ibm-technology/
- Can machine learning improve randomized clinical trial analysis?

  https://www.sciencedirect.com/science/article/abs/pii/S1059131121002685
- Compute Trends Across Three Eras of Machine Learning.

  https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=17869a8760ba
- Green AI.

  https://arxiv.org/abs/1907.10597v3
- How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.

  https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=17869a8760ba
- Mitigating Bias in Artificial Intelligence.

  https://www.ibm.com/policy/wp-content/uploads/2021/05/AI_Bias_IBMPolicyLab.pdf
- Why Time Value of Data Matters.

  https://insidebigdata.com/2016/04/08/why-time-value-of-data-matters/

© Copyright IBM Corp. 2022.

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5679-00

ISBN 0738460699

Printed in U.S.A.

Get connected

Redbooks®

ibm.com/redbooks