Deep Dive into AI on IBM Z

February 24, 2026 (9:30AM)

Purvi Patel
Senior AI Engagement Leader
AI on IBM Z and LinuxONE
purvi@us.ibm.com

Steve Warren
STSM, Principal Solution Architect
Worldwide System Center
swarren@us.ibm.com

Agenda

- Overview
- Thinking about AI on IBM z17…
  - New IBM Telum II AI capabilities
  - What will I be able to do with IBM Spyre?
- Navigating the mainframe AI ecosystem
- Agentic AI
  - AI agents in z/OS
- How can I get started?
- Miscellaneous topics

Mainframe is crucial in the world today

70%

of all financial transactions go through an IBM mainframe[1]

8 of 10 top payment companies

$8.5T

payment volume by value being processed on mainframe[2]

AI on IBM Z: real time insights with superior performance and optimal latency for faster decision making

Mission critical business transactions run on IBM Z®

>70%

of the world's transactions run on the mainframe.[1]

Co-locate data and transactions with AI

84%

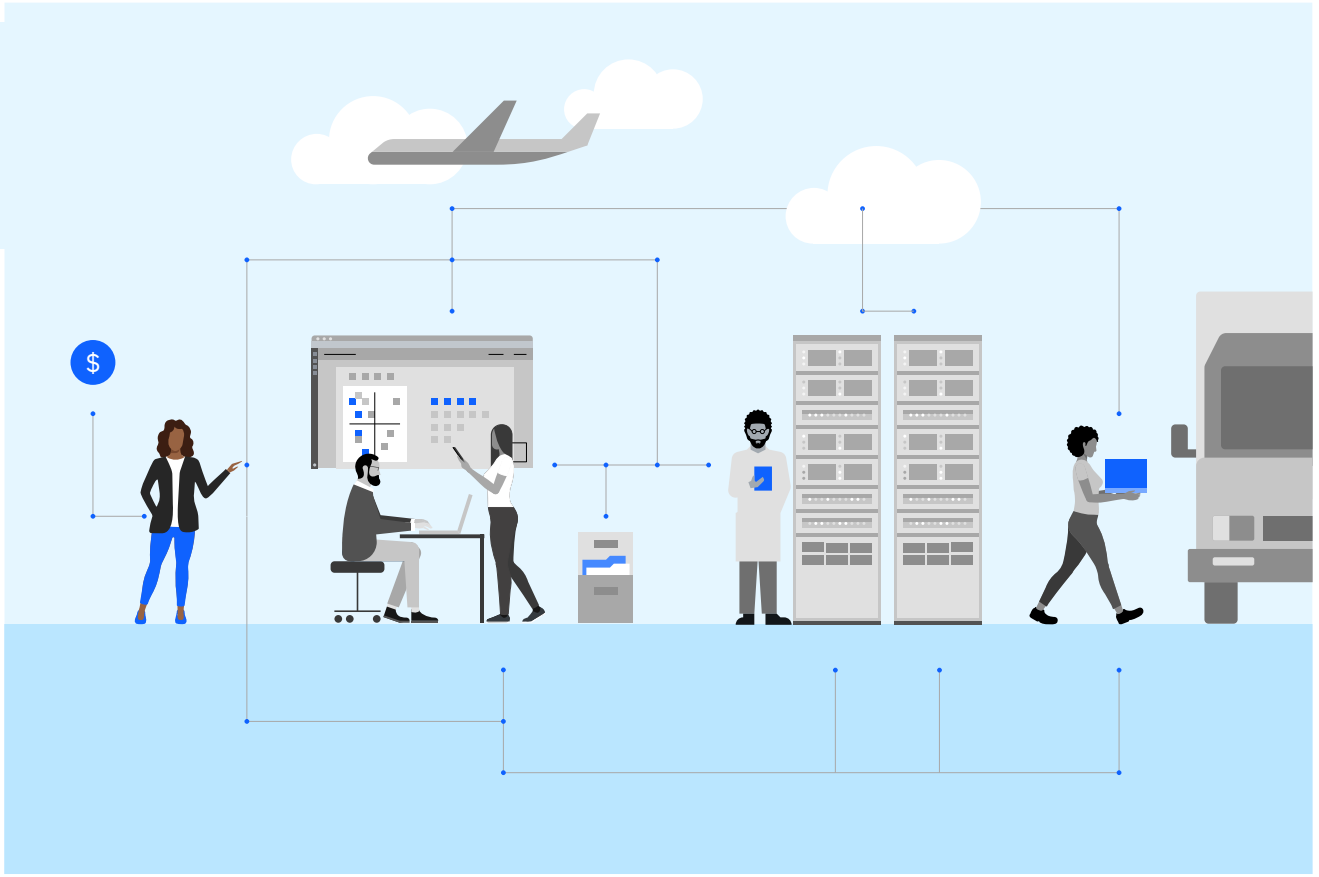
- of IT Executives say incorporating AI into mainframe transactions is important.[2]

Drive innovation without compromising security and resiliency

79%

- of IT executives agree that mainframes are essential, enabling AI-driven innovation and value creation.[3]

There is rich untapped AI potential in your transactions



Your enterprise relies on the transaction processing strength of IBM Z mainframe to execute millions of decisions daily

Opportunity for AI in transactions

- Decisions in payments, credit, claims, AML, etc.
- Improvements in precision
  x high-volume transactions
  = huge economic outcomes

Market Trends in Fraud in Financial Sector

SHARE
EDUCATE • NETWORK • INFLUENCE

72%

of alerts from traditional rules-based systems flagged are false positives[1]

40%

of false positives cut by banks using AI[1]

87%

of financial institutions deploying AI-powered fraud detection systems report fraud prevention efforts now save more money than they cost[2]

1. The Fintech Mag. "AI in Fraud Detection: How Banks Reduce False Positives by 40%" May 5, 2025 https://thefintechmag.com/ai-in-fraud-detection-how-banks-reduce-false-positives-by-40/
2. Alloy Fraud Report 2025: https://www.alloy.com/reports/fraud-report-2025

Market Trends in Insurance Industry

SHARE
EDUCATE • NETWORK • INFLUENCE

$308B

Insurance fraud stolen every year from American consumers[1]

21%

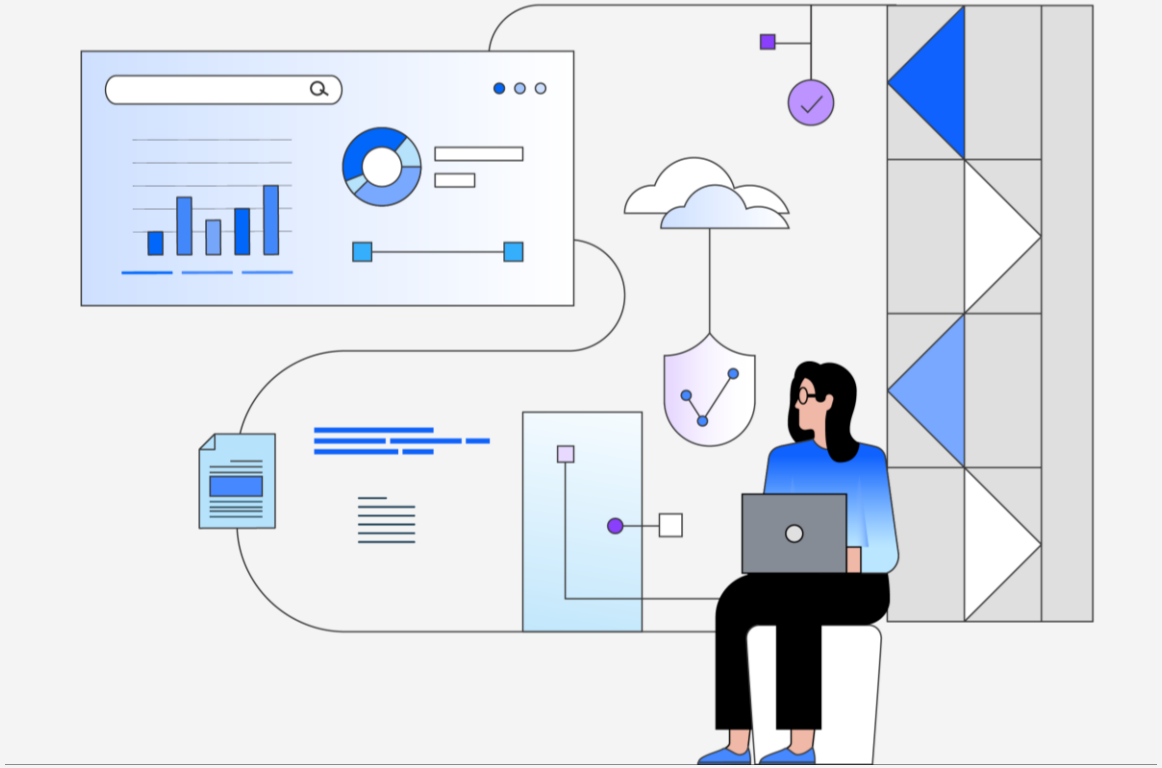
of insurance companies plan to invest in AI in the next two years[1]

3-5%

Accuracy improvement in insurance claims from domain-level rewiring with AI[2]

1. Coalition Against Insurance Fraud https://insurancefraud.org/fraud-stats/
2. McKinsey: The Future of AI in the Insurance Industry https://www.mckinsey.com/industries/financial-services/our-insights/the-future-of-ai-in-the-insurance-industry

AI inside transactions, where milliseconds matter

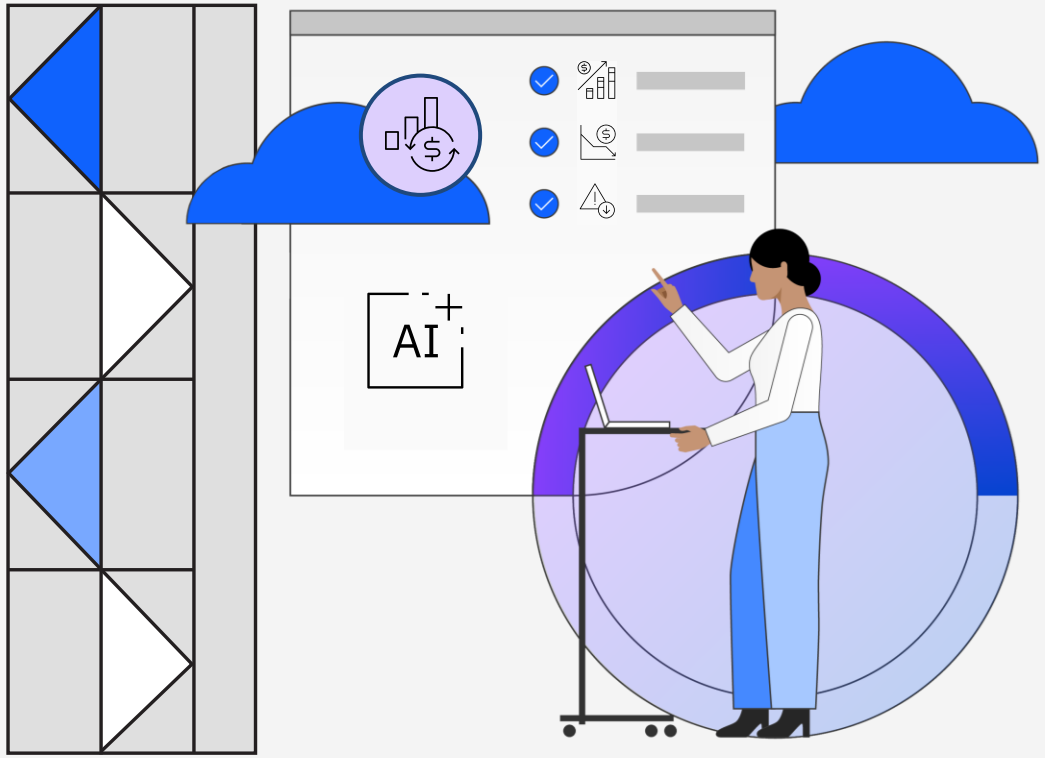


Transactional AI embeds machine learning models directly into your *transactional systems*

Instead of waiting for off-platform AI, insights are injected right into the flow:

- ✓ Fraud flagged before approval
- ✓ Payments rerouted before failure
- ✓ Claims adjusted before submission errors
- ✓ Customer offers are optimized while they're interacting

Transactional AI enhances the business drivers that matter most – revenue, cost, and risk



Revenue Growth

- Intelligent decisioning
- Real-time personalization
- Dynamic pricing

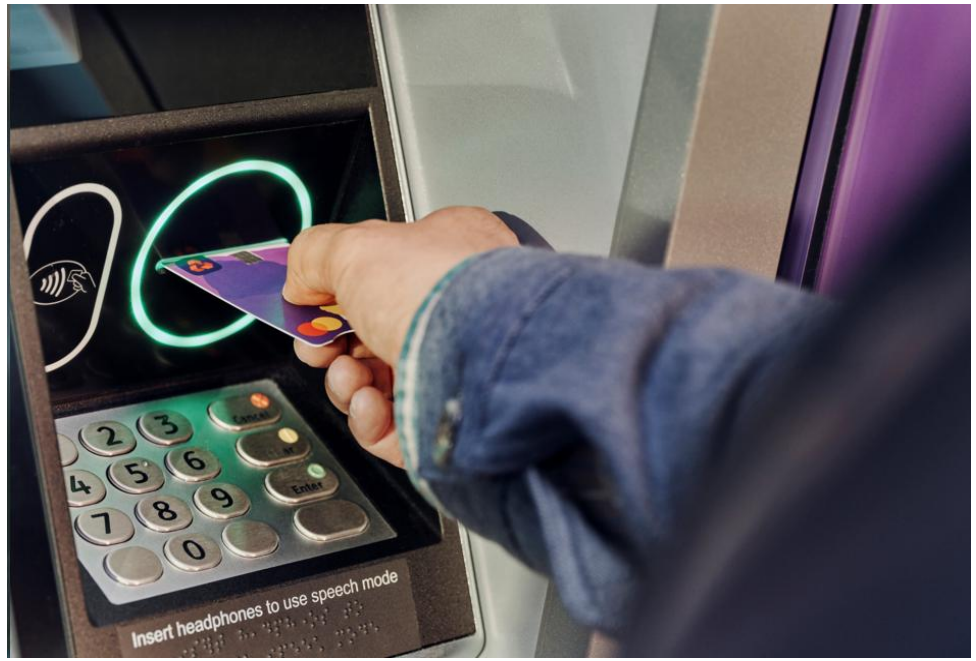
Cost Reduction

- Process automation
- Exception handling
- Resource optimization

Risk Management

- Fraud prevention
- Compliance monitoring
- Operational resilience

Transactional AI can help save costs and lower business risk



In high-volume environments, even small decisions uplifts compounds into multimillion-dollar value.

Business Case – Card Payments (5M Transactions/day at $10 1% fraud rate)

Approach	Off-platform AI	Transactional AI on Z	Value Creation moving AI to on-prem
Fraudulent Transactions	50K	50K	
Coverage	20% sampled	100% real-time	
Daily Fraud Prevented	10K*$50 = $100K	50K*$10 = $500K	$400K Daily
Annual Fraud Prevented	$36.5M	$182.5M	$146M Annually

IBM z17

Fully engineered stack for AI where it matters most

Transaction processing platform

Operating systems & firmware

IBM Z infrastructure

Built on a foundation of security, resiliency, and high availability.

450 billion

Inference operations per day with 1ms response time vs. 300B on IBM z16[9]

7.5x AI throughput

Utilizing 8 AI processing units vs. one on IBM z16[10]

83% less power

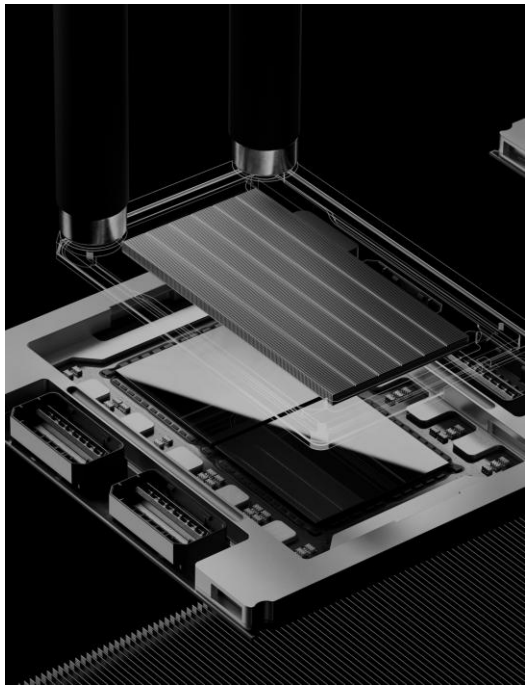
Moving AI-infused OLTP workloads from compared 2 year old x86 servers[11]

ADVANCING AI USE CASES WITH IBM Z17

Unlock potential with accelerating computing on IBM z17

IBM Telum® II Processor

Accelerate in-transaction AI with **encoder LLMs** and multiple AI model techniques



- Support for LLM compute primitives
- Improved quantization and matrix operations
- Improved AI processing over IBM z16

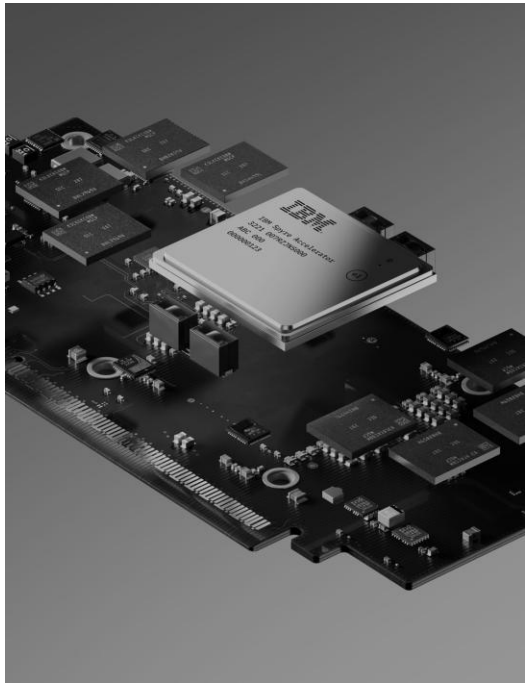
In-drawer intelligent routing

AI **workload balancing** during peak usage

- Remote AI processing
- Up to 8x AI processing available

IBM Spyre™ Accelerator

Optimize **generative AI** and **LLM** use cases



- 32 Gen AI-ready cores per adapter card
- Up to 48 adapter cards per system

New AI Concepts + Terminology for this session

Encoder

Non-generative LLMs focused on natural language understanding.

Useful for tasks requiring understanding like text classification, named entity recognition, sentiment analysis.

Example models:
BERT, RoBERTa, IBM Slate

Decoder

Generative LLMs able to generate new content based on a given prompt.

Useful for tasks like content creation, chat bots and virtual assistants, summarization.

Example models:
Llama, Mistral, IBM Granite

Quantization

Technique used to reduce precision of model values. (e.g. fp32 -> int8).

Goal is to reduce size and compute requirements while minimizing loss in accuracy.

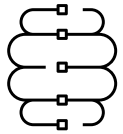
Techniques is generally used with larger models; thus, it is common with LLMs.

Encoder + Decoder models container both types of layer

Provides a foundation to:

Unlock greater AI productivity, and business agility

- Apply multiple AI models to transactions
- Increase practitioner productivity
- Predict and reduce IT incidents for greater resiliency
- On-board new Z skills in less time



Inferencing

Draw inferences from patterns in data at transaction scale

Multi-model (predictive AI + LLMs) inferencing

- Advanced fraud detection
- Anomaly detection

Outcomes:

- Improved accuracy
- Fewer false positives

Assistants

Helps complete tasks for you based on your requests

Expert assistance leveraging contextual information with Gen AI

- Business Assistants
- Code & Operations Assistants

Outcomes:

- Increased productivity
- Faster skills onboarding

Agents

Proactively works to achieve business and IT goals

Leverage AI-insight and tools to reason, decide and problem-solve

- Automated trading
- Application health

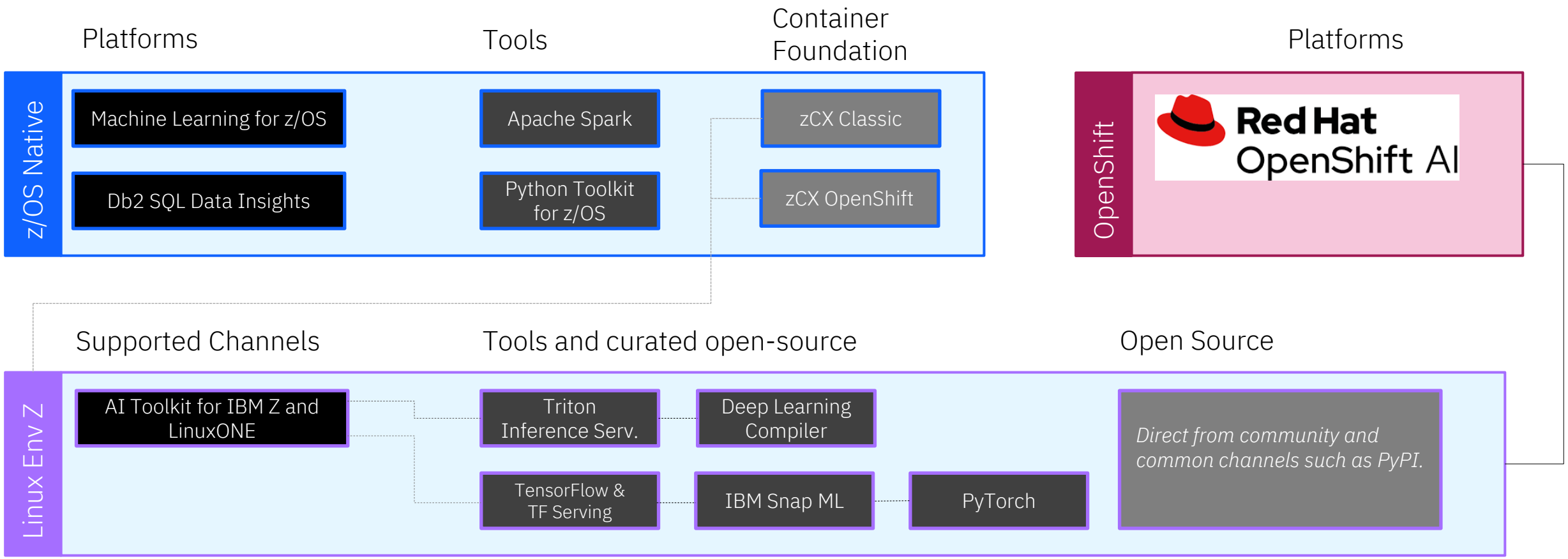
Outcomes:

- Optimized workflows
- Autonomous operations

Effective combination of AI inferencing, assistants, and agents to transform and simplify the mainframe experience

AI MODEL DEPLOYMENT OPTIONS ON IBM Z

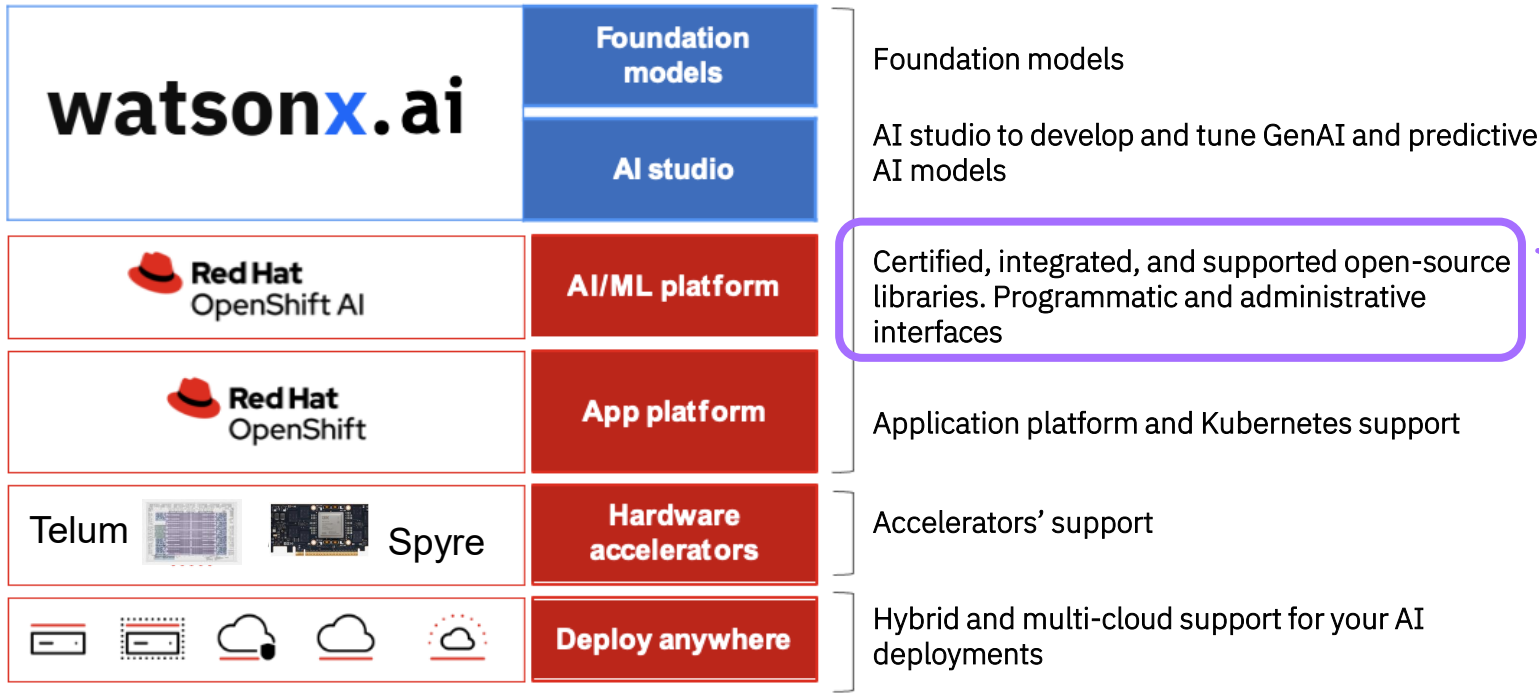
Transactional AI Stack- Overview



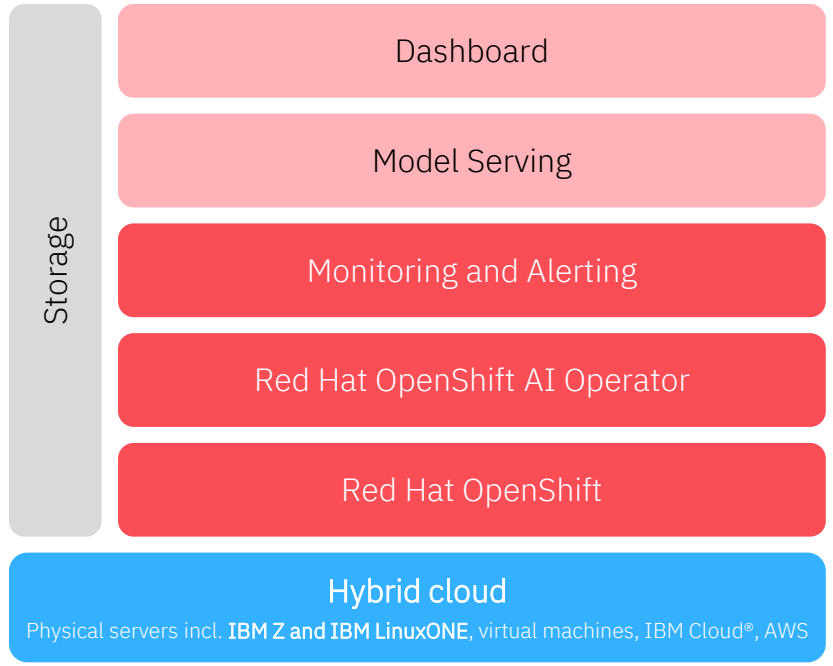
IBM Z & LinuxONE GenAI Stack - Red Hat OpenShift AI

Build, serve, and monitor your AI model, and manage data science pipelines

End-to-end Stack

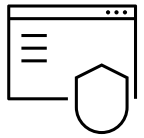


Features and capabilities



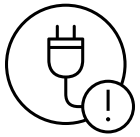
Red Hat OpenShift AI on IBM Z and LinuxONE

Designed to enable secure, scalable AI across hybrid cloud



Accelerate AI adoption with accessible, open-source tools

- Integrate widely-used AI/ML frameworks (KServe, Triton inference server, TrustyAI, vLLM, Kueue) to leverage existing skills and streamline collaboration.
- Simplify the AI lifecycle with built-in MLOps capabilities for model development, deployment, monitoring.
- Reduce integration complexity by co-locating AI workloads with mission-critical applications on IBM Z and LinuxONE.



Provide hybrid cloud consistency for AI model deployment

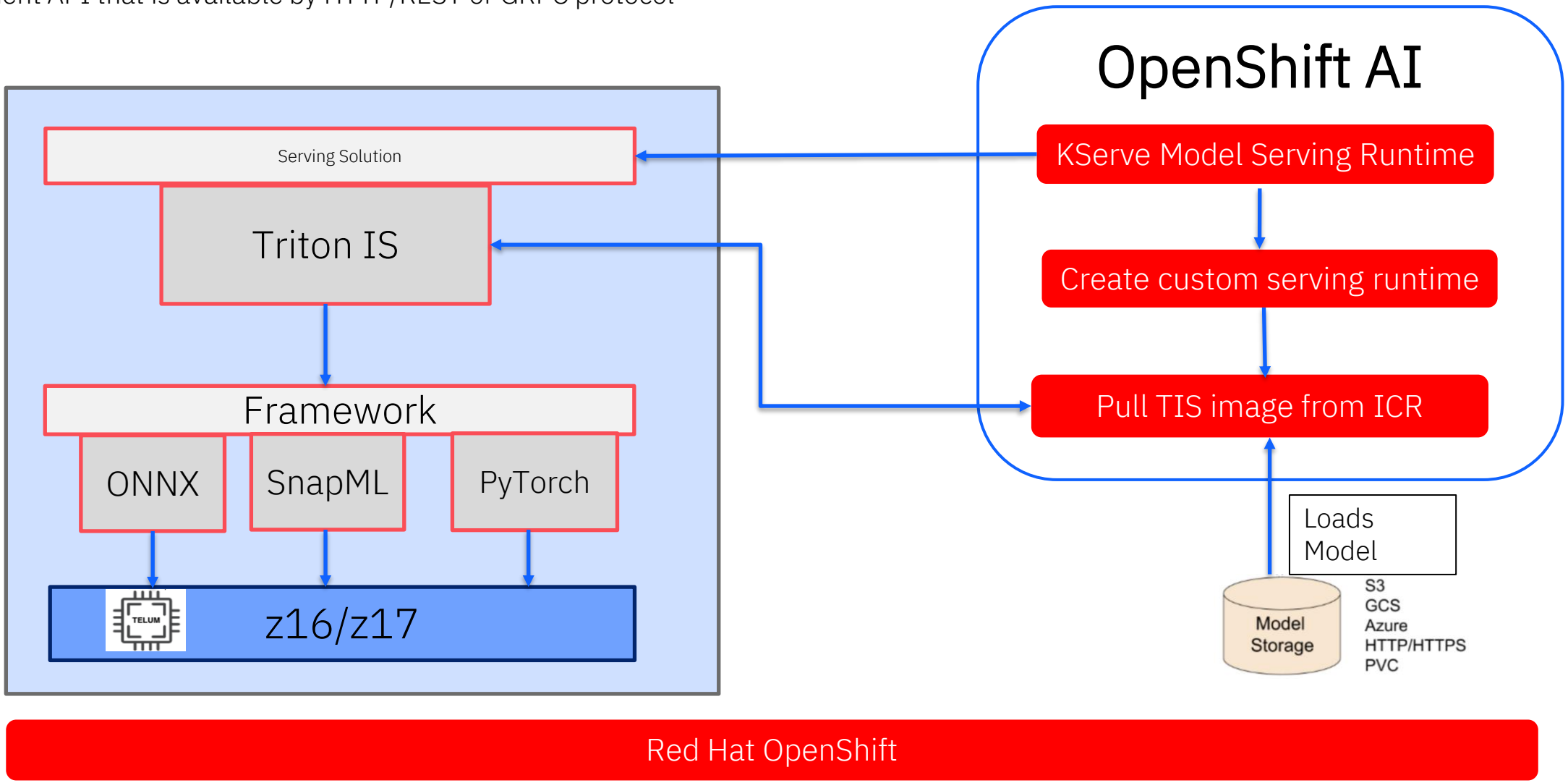
- Provide a unified Kubernetes-based platform across on-premises IBM Z and LinuxONE and hybrid cloud environments.
- Standardize tooling for workload portability and predictable performance.
- Speed production rollout with automated serving, monitoring for AI models.

Boost AI performance with enterprise-grade infrastructure

- Run containerized AI workloads alongside traditional applications and VMs on the same IBM Z or LinuxONE system.
- Leverage IBM Telum processors with on-chip AI acceleration and dedicated Spyre PCIe cards for high-volume predictive and generative AI inferencing.
- Scale AI performance for demanding use cases such as real-time fraud detection, agentic AI, and complex analytics.

RedHat OpenShift AI Integration with Triton Inference Server (TIS)

- Third party integration of IBMZ TIS with RHOAI is similar to NVIDIA TIS for Telum I and Telum II
- TIS will support PyTorch , SnapML , ONNX backend
- The models being served by Triton Inference Server can be queried and controlled by a dedicated model management API that is available by HTTP/REST or GRPC protocol



AI Infusion - Platform Entry Points on z/OS

Machine Learning for z/OS

Full-featured machine learning platform
Build and train anywhere, deploy on z/OS

- Telum on-chip AI Accelerator and zIIP exploitation for AI inferencing at scale
- Score transactions natively in CICS, IMS and Batch applications with near-zero latency
- Trust and Explainability in every decision

- AI-driven insights for accurate and real-time automation
- Infuse AI into products for a truly intelligent platform to propel your business forward
- Integrate with IBM ODM to augment rules-based systems with AI-driven insights.

Join us to learn more about MLz

Accelerated Business Decisions with Machine Learning for z/OS
Salon 22
Tues, Feb 24, 10:30am – 11:30am
Speakers: Steve Warren, Purvi Patel

Db2 SQL Data Insights

Industry first AI capability can power any Db2 13 for IBM z/OS application
Leverage AI enhanced SQL with no data science skill.

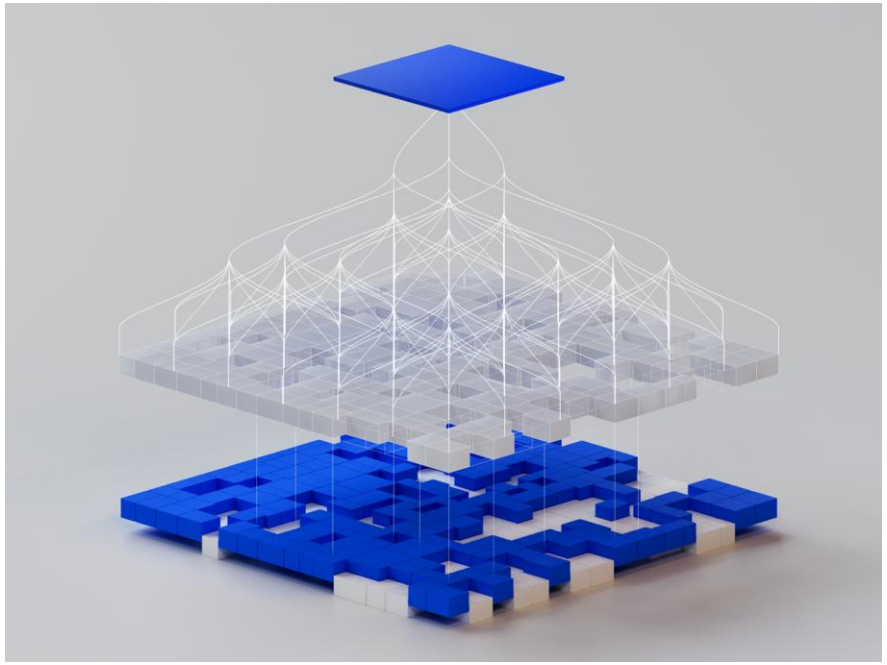
- Use built-in AI models to understand underlying semantics of the data
- New queries to identify similarities, dissimilarities and correlations.
- Interpretability provided out of the box

- Minimal AI deployment complexity; no mode lifecycle management.
- No data science skills needed - invoked via simple SQL queries.
- Seamlessly leverage available IBM Z hardware acceleration.

AI Toolkit for IBM Z and LinuxONE

IBM's Elite support offering around key open-source AI frameworks

- IBM Elite Support
- High performing serving frameworks
- Optimized to leverage IBM Z Integrated Accelerator for AI
- Certified software for security



IBM Z Accelerated TensorFlow

Popular ML and DL lifecycle management platform optimized to run on Z and leverage Telum on-chip accelerator.

IBM Z Accelerated Serving for TensorFlow

Flexible and high-performing serving platform for ML/DL models optimized to leverage Telum on-chip accelerator.

IBM Z Accelerated Snap ML

A library that optimizes the training/scoring of popular ML models optimized to leverage Telum on-chip accelerator.

IBM Z Accelerated Triton Inference Server

High-performance inference server that supports the deployment of ML or DL models at scale optimized to leverage Telum on-chip accelerator.

IBM Z Deep Learning Compiler

Generates a program from any ONNX DL models to execute on z/OS or Linux on Z optimized to leverage Telum on-chip accelerator.

IBM Z Accelerated PyTorch

A popular Machine Learning framework based on Torch library, used for applications such as language processing and computer vision optimized to leverage Telum on-chip accelerator.

AI on IBM Z Ecosystem Stack

Designed for Business Insights and Intelligent Infrastructure

BUSINESS INSIGHTS

Infuse AI in real time into every transaction

MACHINE LEARNING FOR IBM z/OS

Deliver AI solutions at an unprecedented speed

DB2 FOR Z/OS WITH SQL DATA INSIGHTS

Uncover hidden patterns from data locked in z/OS

Red Hat OpenShift AI

watsonx.ai

IBM's Integrated AI Platform

watsonx.governance

End-to-end AI Governance

AI TOOLKIT FOR IBM Z AND LINUXONE

Support popular open-source AI tools on IBM Z

INTELLIGENT INFRASTRUCTURE

Improve automation, security, privacy and ITOps with AI

AI Framework for IBM z/OS

Enable intelligent admin, ops and automation

IBM DB2 AI FOR Z/OS

Enhance database performance with ML

IBM Concert for Z

Streamline mainframe operations with AI-powered efficiency

watsonx Assistant for IBM Z

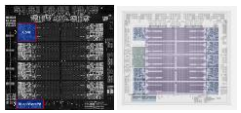
Enable conversational AI, automate tasks and build skills

IBM Threat Detection for z/OS

Next gen protection for your most crucial data

ONNX, scikit-learn, Snap ML, python, spaCy, RASA, TensorFlow, Kubeflow, OpenCV, vLLM, watsonx, Keras, PyTorch, Hugging Face, MAGE, NVIDIA Triton Inference Server, Red Hat OpenShift AI, Apache Spark, mlflow

Enable market leading AI / ML ecosystem on IBM Z



Telum I & Telum II

On-chip AI accelerator

Spyre Accelerator

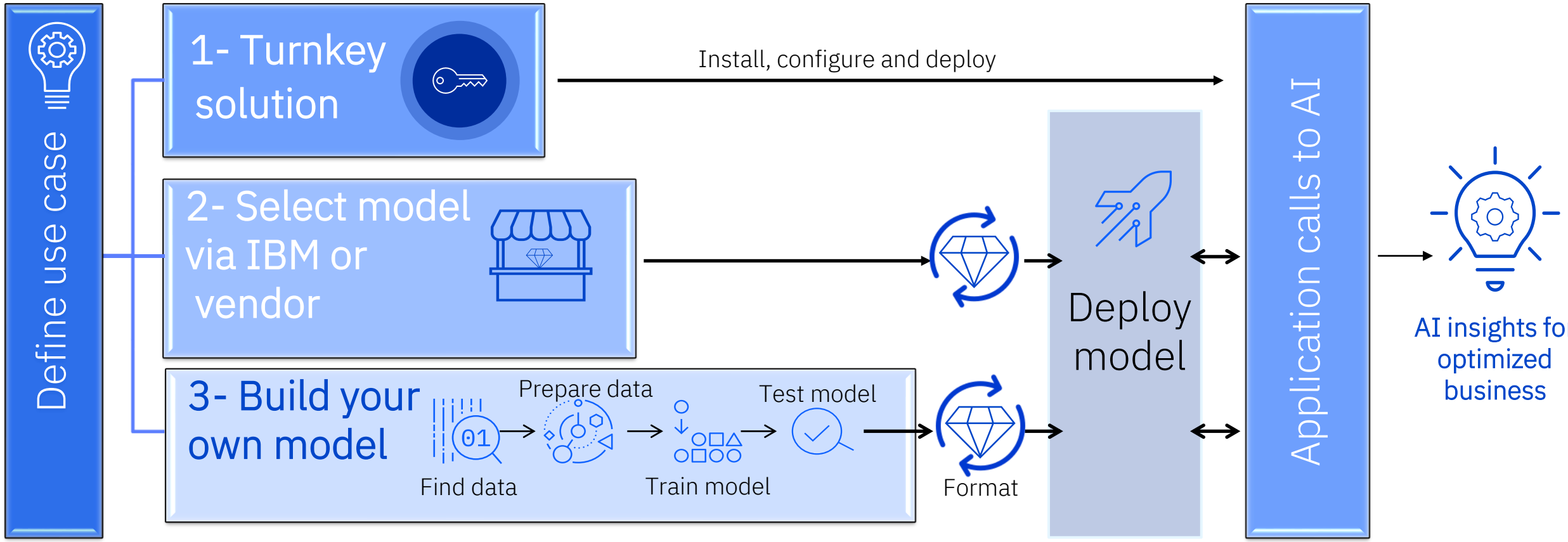
Host attached AI accelerator

Deliver billions of inference requests per day in real time

AI on Z & LinuxONE Entry points

Software High Level AI Entry Points

AI Client Engagement experiences from multiple starting options



Entry Point 1: Examples of ISV Turnkey AI Solutions

Optimized AI Ecosystem partner solutions

Defined use case

1- Turnkey solution

Install, configure and deploy

Application calls to AI

Core Banking & Payment	ISV Point Solutions	
FIS	Anti-Money Laundering & Payments Fraud	Credit /Lender Risk Management
Luxoft A DXC Technology Company	Clari5 CustomerXPs™	finbots.ai
fiserv.	Claims Processing (e.g. Healthcare)	Customer Experience Management
WORLDLINE	EXPONENTIAL AI	OvationCXM

Entry Point 2 - Select pre-built model from IBM or vendor

Bring Training Data → Use pre-built model from IBM or vendor → Model conversion & formatting → Deploy & serve models for inferencing acceleration on Z and IBM® LinuxONE → Infuse insights into Applications

Defined use case

IBM Synthetic Data Sets

IBM AI on Z

Github examples (e.g. Credit Card Fraud Detection)

2- Select model via IBM or vendor

Client real data

Other masked data

watsonx. ai

Models from other vendors / sources

Predictive Model Markup Language

PMML, ONNX, or JSON

ONNX

Open Neural Network Exchange

IBM Z Deep Learning Compiler* (for ONNX)

IBM Z Accelerated for TensorFlow Serving*

IBM Z Accelerated for PyTorch*

IBM Z Accelerated for Nvidia Triton Inference Server*

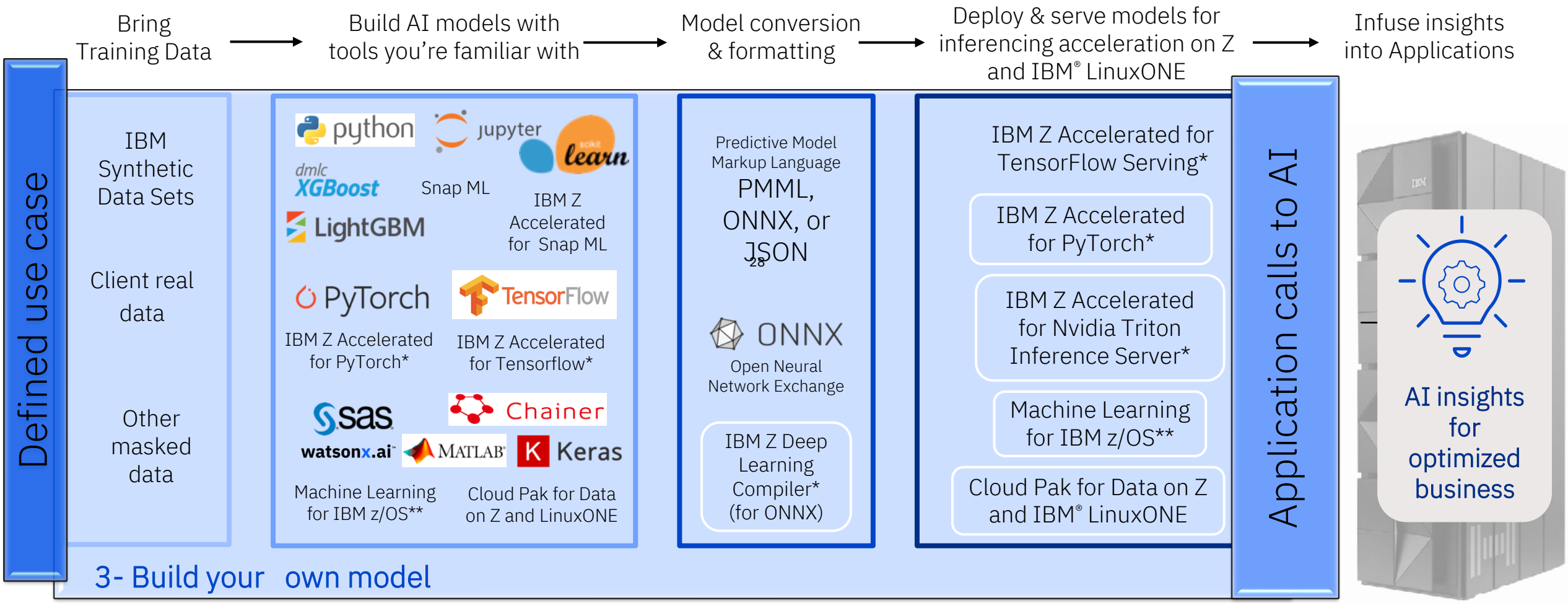
IBM Cloud Pak® for Data on Z and IBM® LinuxONE

Application calls to AI

https://ibm.github.io/ai-on-z-101/ For details on which tools are compatible/included with others

* Included with AI Toolkit for IBM Z and LinuxONE

Entry Point 3- Build your own models



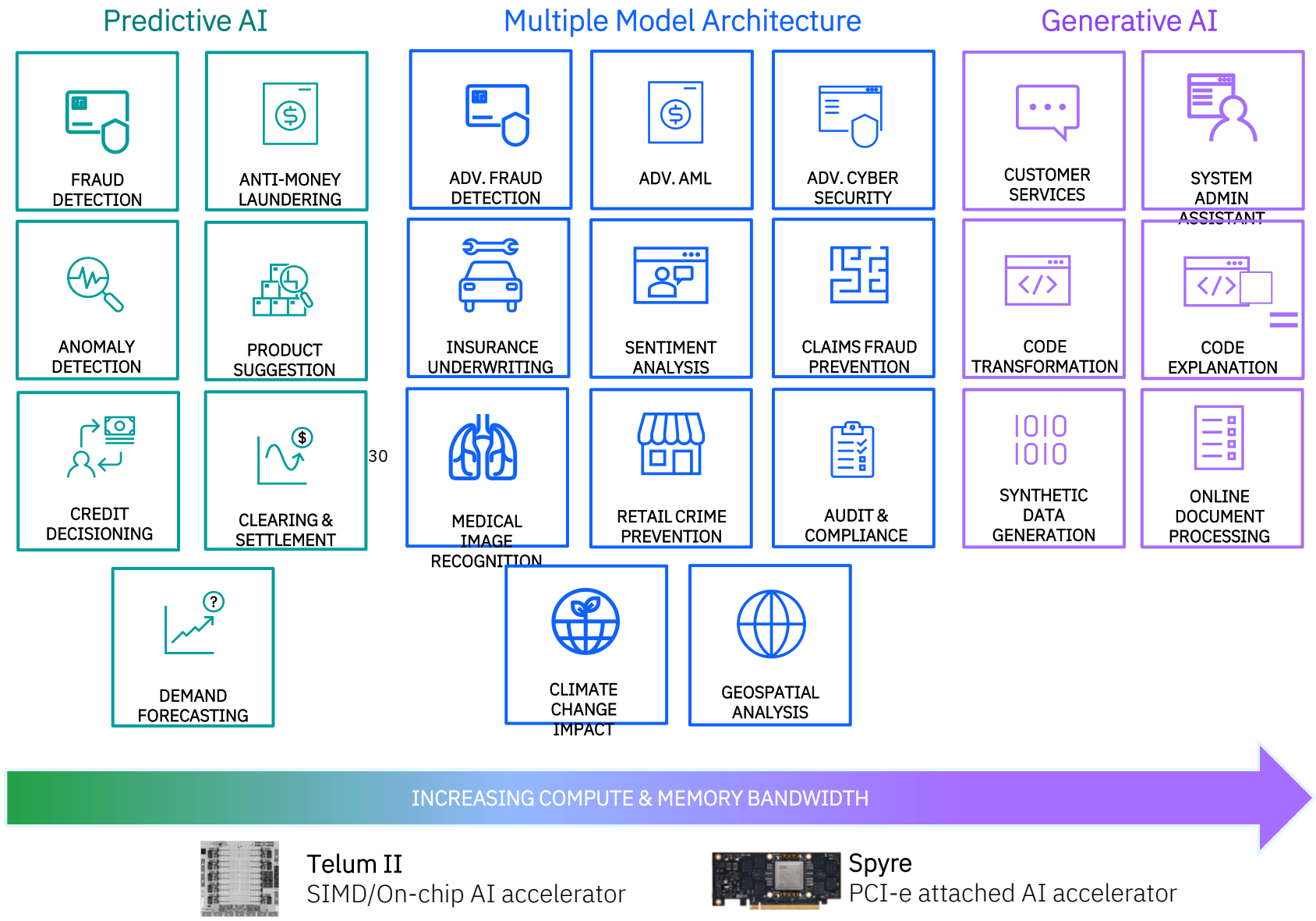
https://ibm.github.io/ai-on-z-101/ For details on which tools are compatible/included with others

* Included with AI Toolkit for IBM Z and LinuxONE

** IBM z/OS only

ENABLING NEW USE CASES
IBM TELUM II & IBM SPYRE

Telum™ II & Spyre Designed to accelerate enterprise AI use cases at scale

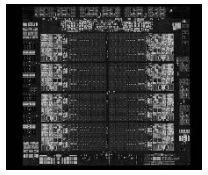


AI acceleration technologies to address business requirements

Predictive AI

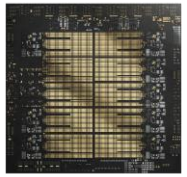
- Available on z16 w/ Telum and z17 with Telum II
- Inferencing in every transaction
- Speed and scale are critical
- Low latency and high throughput
- Precision and accuracy matter

IBM Telum



OR

IBM Telum II

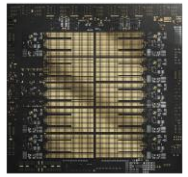


Predictive AI for real-time, high-volume transaction processing unlocking business growth

Multiple model AI

- Available with z17 and Telum II for most use cases, can add Spyre for more complex encoder LLMs
- Advanced AI with higher predictive accuracy and expanded insights
- Generates high value by combining the power of traditional AI models and LLMs in real-time on every transaction

IBM Telum II



Predictive AI and Encoder LLMs for advanced real-time, high-volume transaction processing unlocking business growth w/ optimized outcome

Generative AI (GenAI)

- Available with z17 + Spyre Accelerator
- Larger LLMs require much more compute
- Performance measured in tokens per second
- Security is a priority to ensure client data and models are protected
- Energy efficiency is essential to scale workloads responsibly

IBM Spyre Accelerator

Decoder LLMs to accelerate and scale Generative AI workloads

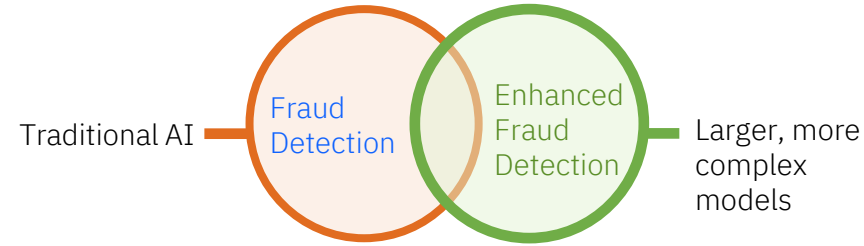
INCREASING COMPUTE & MEMORY BANDWIDTH

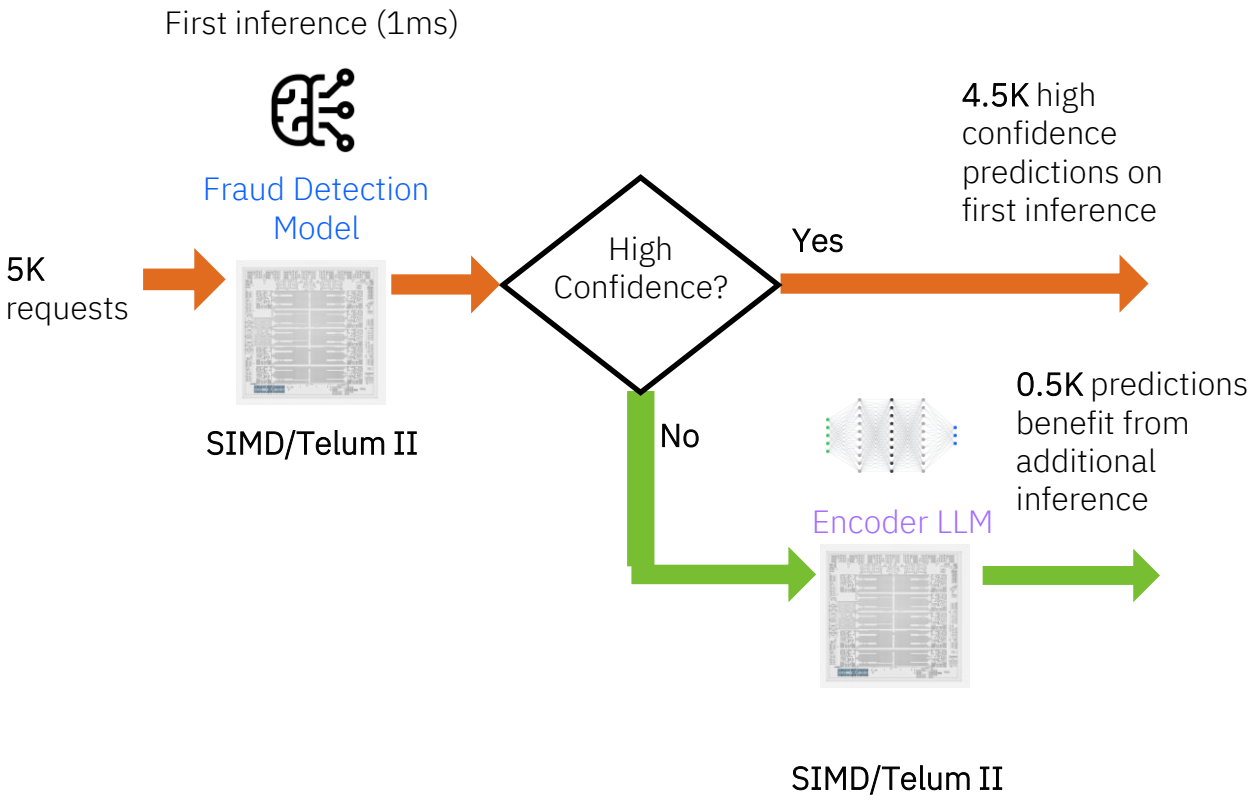
Claims Quality Use case – Traditional AI

Advanced AI for improved accuracy and deeper insights

Multiple Model AI architecture

Payment fraud detection example





Benefits

Improved accuracy

Reduced false positives and negatives

Greater ROI

Achieved by reducing loss or increasing gains with AI

Client loyalty

Improved and customized customer experiences

Improved efficiency

Reducing time spent on manual processing

Score 100% transactions in real-time
while transactions are happening

Advanced Audit & Compliance

Business Scenario:

State government agency wants to modernize and automate application that approves custom license plates and disabled placards. Current process is manually and requires administrators to review through thousands of applications and attached documents to screen for profanities or out of policy requests and identify and reject potentially fraudulent applications. For audit and compliance, the state agency must prove that applications that are being approved/rejected adhere to their state government policies

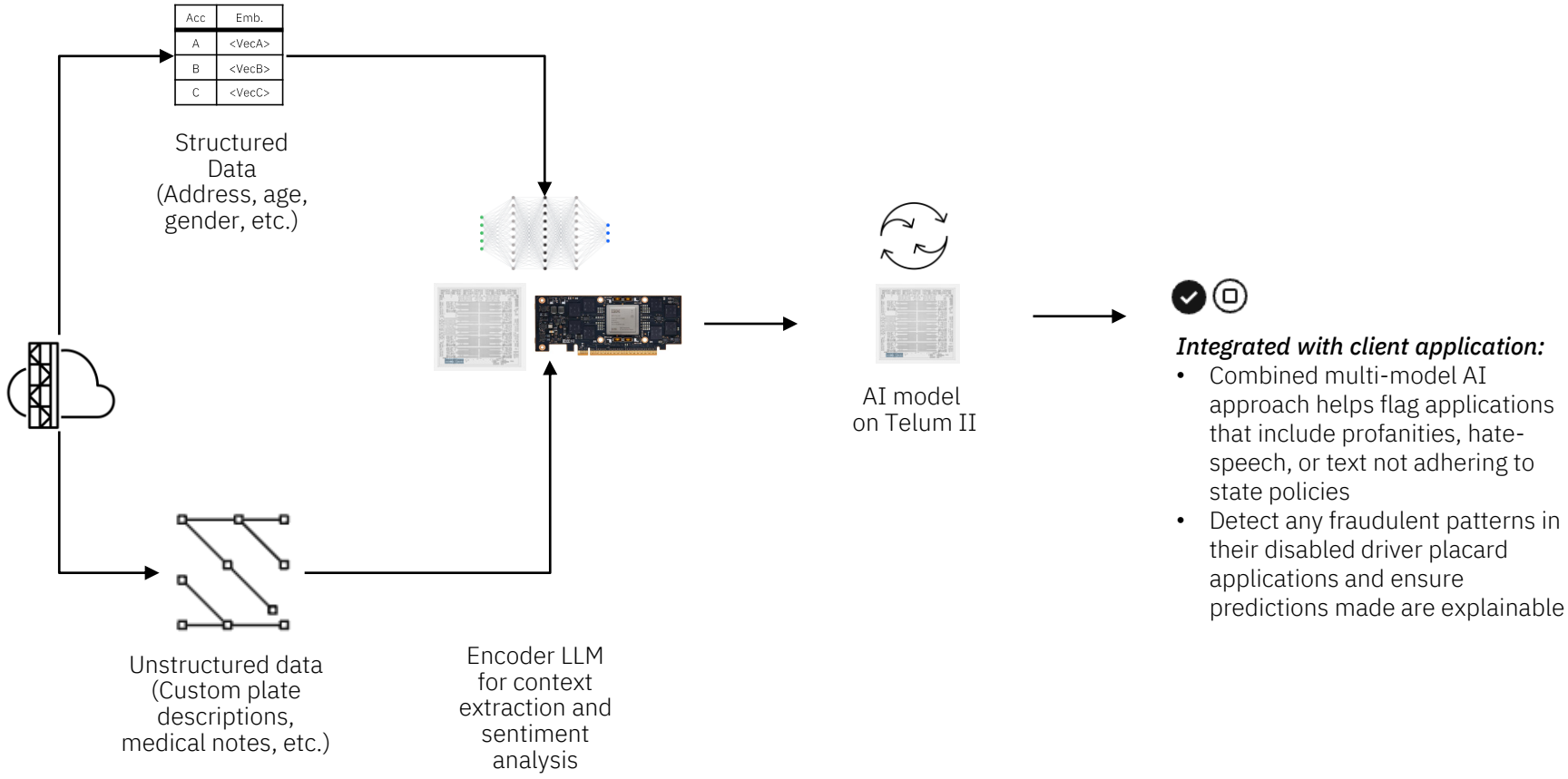
Business Impact:

- Reduced manual investigations
- Manage adherence to compliance regulations
- Faster processing
- Fraud and error reduction

Achieved by deploying an encoder LLM model extract insights and generate embeddings from the unstructured input data and a predictive AI model to predict potential fraudulent applications

Example Solution

IBM z17
with Telum II



Products	***Components***	***Models***
Machine Learning for z/OS, AI Toolkit for IBM Z & LinuxONE, RHOAI	watsonx.ai, watsonx.governance	BERT, Random Forest, Anomaly Detection

Home Insurance Claim Fraud on IBM z17 with Spyre Accelerator*

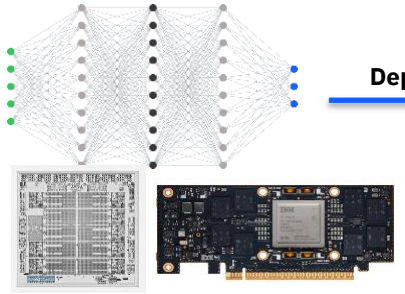
Claim Narrative

Two days ago, there was **malicious damage.** Here is my initial summary of items for you to cover: the dormers, the outside walls, and the decorative benches. Probably **not very urgent.** My losses were $3,920

Claims Image Data

Unstructured data

1



Encoder Large Language Model
Telum II or Spyre

Deployment

Machine Learning for z/OS | **AI Toolkit(LonZ)**

Inference

Cause of claim = **Vandalism**
Urgency = **Not Urgent**
Damage Severity = **Low**

IBM Synthetic Datasets Product

Structured data

Policy ID	Date	Total Paid	...	Claim Amount
6AA823A34	1/2/23	$2500	<value>	$1500

2

Cause of Claim	Urgency	Damage Severity
Vandalism	Not Urgent	Low

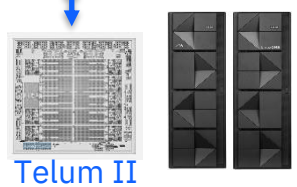
New features generated in real-time integrate with structured data

3

Machine Learning for z/OS | **AI Toolkit(LonZ)**

ML/DL Model

AI Explainability



Telum II

4

High Risk

Low Risk

High Risk Claim ID

IBM SQL DI (z/OS)

5

AI Similarity Query

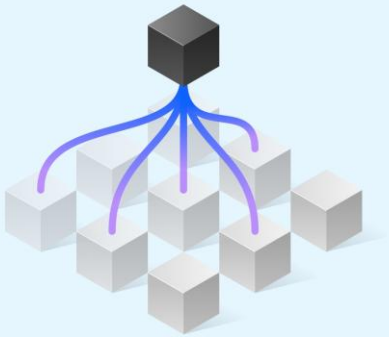
Similar Claims based on High Risk

Policy ID	...	Similarity Score
6AA354A43	<value>	0.93
6AB464B21	<value>	0.82

GENERATIVE AI ON IBM Z & LINUXONE

Insurance Claims Demo

AI Serving Platforms, Runtimes, Environments and Performance/ Compute Differences



Transactional AI vs. Generative AI vs. Agentic AI

Transactional AI

- ✓ Speed and scale are critical.
- ✓ Low latency and high throughput are absolute requirements.
- ✓ Today: smaller models less compute needed (ML/DL model and encoder LLMs).

Generative AI (GenAI):

- ✓ The focus is on acceptable performance, measured in tokens per second, rather than the best performance in the industry.
- ✓ Larger Models requiring a lot more compute (GenAI decoder LLMs).
- ✓ Security is a priority, ensuring the protection of client data and models.

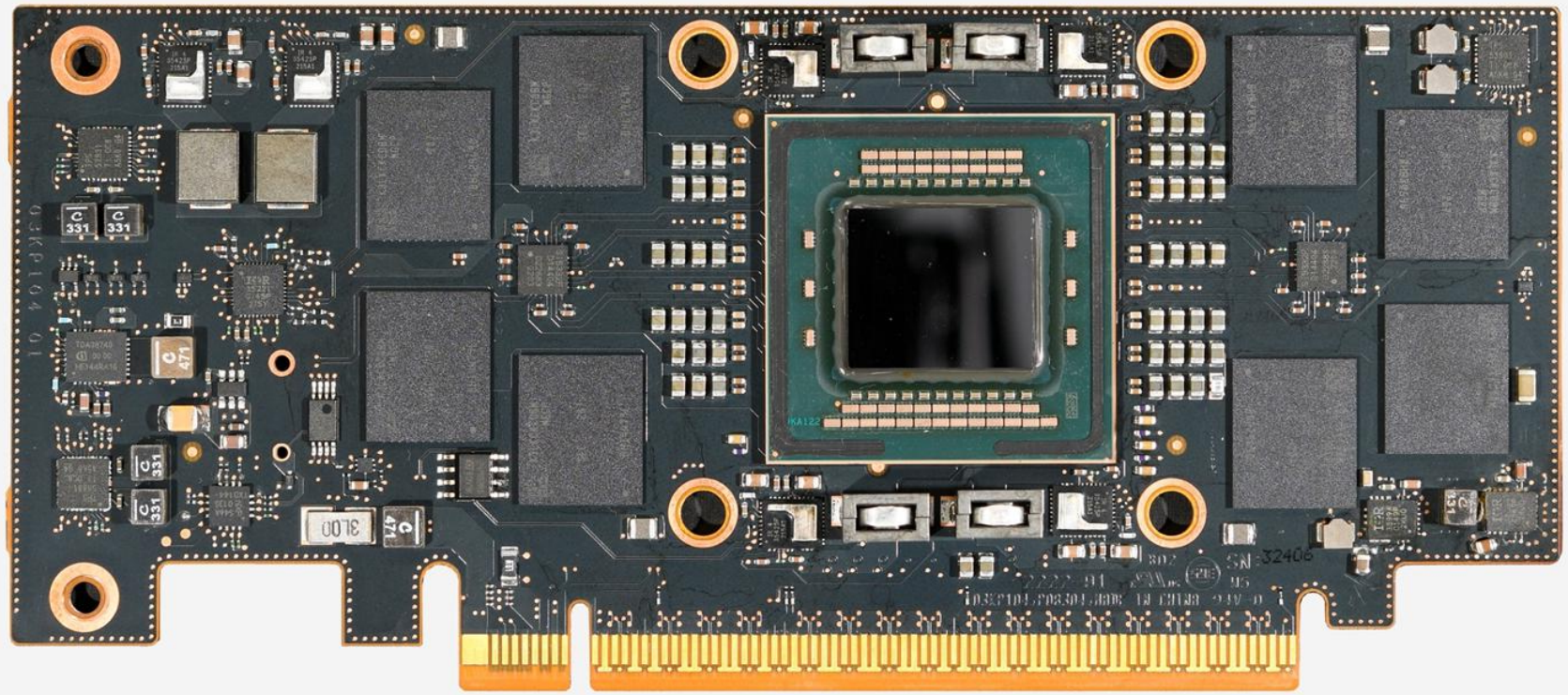
Agentic AI*

- ✓ Agentic AI is a framework for accomplishing goals with limited supervision that consists of AI agents.
- ✓ AI Agents use LLMs (LLM in a loop) to reason and can interface with tools, other models, and other IT systems to fulfill user goals.
- ✓ Capacity planning shifts from single-model sizing to workflow-level concurrency.
- ✓ Throughput becomes orchestration bound the bottleneck is parallelizing agent steps, not raw model throughput.

IBM Spyre™ Accelerator PCIe attached card

- 75W PCIe gen5 x16 adapter
- 128GB of LPDDR5 memory
- Up to 8 cards per I/O drawer
- Generative AI: 8 cards form a logical cluster
- 1TB of memory
- 1.6TB per second aggregate memory bandwidth

Designed to handle large language models and generative AI use cases



IBM watsonx Assistant for Z

Mainframe experience, reimagined.

Generative AI solution, transforming and simplifying the way IBM Z® users of all experience levels engage and interact with the mainframe to be more productive.

Conversational AI

Quick and accurate answers to questions that leverage IBM Z domain-specific and your own documentation.

Integrate Automation

Confidently perform both routine and complex tasks. Connect and drive execution of tasks in other tools; all initiated through AI conversation.

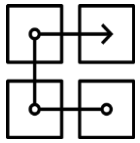
Configurable Assistant

Personalize based on business process and job roles. Seamlessly integrate your own documentation, processes and best practices to answer proprietary questions.

Benefits: | Reduce learning curve | Increase productivity | Increase efficiency | Reduce errors

What makes IBM watsonx Assistant for Z different?

Fast, current, Z context-aware, automated, and secure.

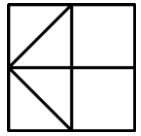


Agents with Built-In Domain Expertise

Foundational Agentic Framework providing a scalable architecture for building and managing AI agents across the mainframe ecosystem

IBM Z AI agents are specialized for mainframe workflows and embeds with decades of expertise

Out-of-the-box agents helps accelerate time-to-value and simplifying deployment



Purpose-Built for IBM Z

Uses RAG* to ground responses with both IBM Z and your enterprise knowledge - if we don't know, we don't guess!

Gen AI chat backed by ability to collaborate with agents across your ecosystem for deeper insights.

Customizable to Your Environment

Easily integrate your own documentation, automations, APIs, and tools.

Support and extend your agents with Model Context Protocol (MCP)

Enforce compliance guardrails and ensure only authorized user can execute automations

.

Simplified Agent Management and Flexible Builder Experience

Single place to manage, deploy, and run all AI agents—IBM-built, third-party, and custom

Democratizes expert knowledge. Create no-code agents in minutes or build pro-code agents for advanced, custom workflows.

*Retrieval-augmented generation (RAG) Link
Pre-ingested documents and Redbooks covering over 200+ IBM Z products

Generative AI use cases

Tax Document Processing

Leverage GenAI to summarize submitted tax documents and highlight key findings such as mismatched income reporting, unusually high deductions, etc.

Mistral Small: Used for high-fidelity image-to-text transcription and object identification within documents

Financial Documents Processing

Leverage GenAI to summarize financial documents and business reports to extract key data points such as financial metrics, and performance indicators; and identify essential information needed for compliance processes like financial audits

Qwen 2.5 VL 7B

Insurance Document Processing

Extract key points from content that has been stored as unstructured text. Search and identify essential information needed. For insurance, the primary challenge is digitizing and analyzing "messy" real-world data: handwritten claims forms, tables, and photos of physical damage.

Qwen 2.5-VL (Vision-Language) is considered one of the strongest models
mistral-small-3-1-24b-instruct-2503

Customs Screening for Suspicious Cargo

Utilize Generative AI to identify potentially suspicious cargo through sophisticated image processing techniques and analysis of textual descriptions associated with each shipment

Mistral Small: Used for high-fidelity image-to-text transcription and object identification within documents

Key Encoder and Decoder Models on the roadmap for optimization support on IBM z17 (List is regularly updated)

Transactional use cases

Use cases	Model
Advanced Fraud detection (Core Payments)	BERT (110m)
Advanced fraud detection (credit cards)	BERT (110m)
Advanced AML (online)	BERT (110m)
Advanced AML (batch)	BERT (110m)
Advanced Clearing and Settlement	BERT-Large (340m)
Advanced Claims processing	DistilBERT/ Distilroberta (82m)

Non-Transactional use cases

Use cases	Model
Sensitive data tagging	Slate (153m) XLM-RoBERTa (270m)
AI for coding (WCA4z)	Granite (20B)
Dev & IT advisor (ChatOps,WXA4Z)	Granite(8B)
Cybersecurity	CyBERT (130m)
Recommendation engines	DistilBERT (66m)
AIOps - incident detection, RCA, etc.	BERT (110m)
Sentiment analysis	Slate (153m)
Document processing	Llama-3 (8B), Mistral/Mixstral, Qwen, Gemma
Summarization	QWEN (Qwen2.5-Coder-32B-Instruct), Mistral/Misxstral (mistral-small-3-1-24b-instruct-2503), Granite,Llama 3-8B

Takeaway:

Transactional AI on IBM Z delivers real-time inferencing in 100% of transactions compared to off-prem AI

Co-location Benefits

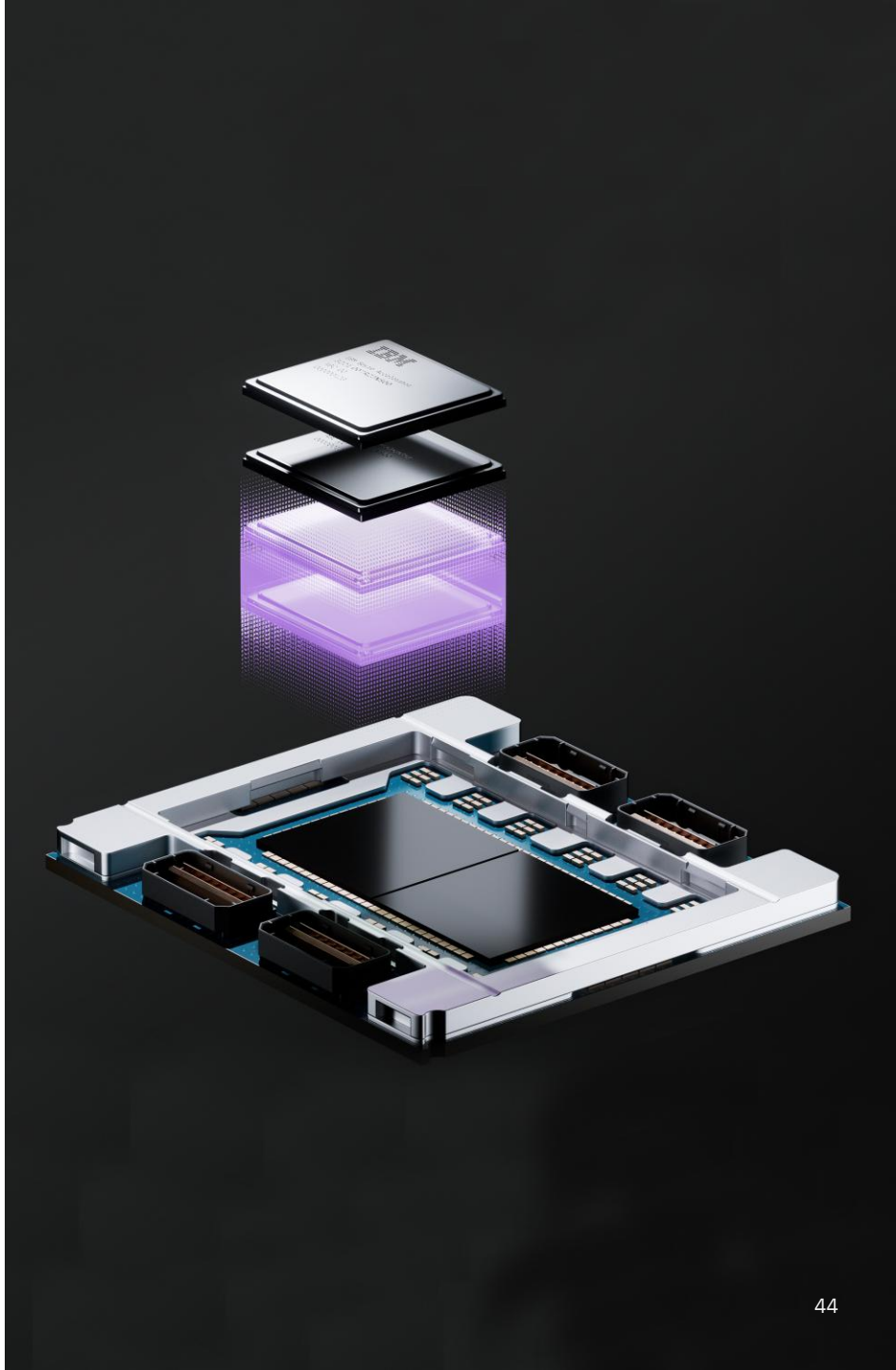
- Inference runs next to the data
- No latency
- No transfer risk

Hardware Acceleration

- Telum, Telum II and Spyre Accelerator deliver inferencing acceleration for various types of use cases

Qualities of Service

- Models operate on complete, trusted transactional data
- Availability
- Security
- Resiliency



Deploy AI at scale with efficiency

Adapt to ever growing AI workloads without sacrificing performance or efficiency

Deploy AI models at scale with built-in AI accelerators

Enable scalable deployments for advanced AI solutions with built-in accelerators to gain business insights at the speed of decision making

Optimize AI Ecosystems

Leverage common industry tools like OpenShift AI to simplify the AI implementation and deployment across the hybrid stack

Leverage optimized tools for built-in accelerators

Build and deploy AI models using cutting-edge technologies, AI frameworks and open-source tools in AI Toolkit for IBM Z® and IBM® LinuxONE that are optimized for built-in AI accelerators

Elevate AI model accuracy

Optimize outcome and increase model robustness by combining traditional AI with non-generative AI LLMs for better results, leveraging Telum™ II and Spyre acceleration

Drive innovation by deploying GenAI on-premises

Create intelligent applications and embrace generative AI solutions close to your data and leveraging Spyre acceleration

Minimize AI energy impact with power-efficient technologies

Optimize power consumption, reduce operational cost, contributing to overall savings by leveraging power-efficient built-in accelerators and consolidating AI workloads, with over 5x energy savings on a sample OLTP application vs x86

GETTING STARTED

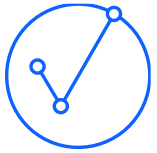
The path to get started with AI on IBM Z and LinuxONE today

IBM’s investment in partnering with clients

1. Client briefing

Discussion and demonstration of AI on IBM Z or LinuxONE strategy and capabilities. Understand where how AI can be leveraged for impact in your business.

1 - 2 hours
onsite or virtual

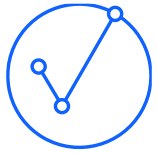


2. Discovery Workshop

Engage with an AI on IBM Z expert team to identify and ideate on candidate use cases, identify the value delivered by AI on IBM Z or LinuxONE, and scope a project architecture.

1/2 - 1 day
onsite or virtual

Contact: aionz@us.ibm.com



3. Pilot / Proof of Concept

Collaborate with a multi-disciplinary IBM team to jointly innovate and rapidly demonstrate AI on IBM Z or LinuxONE value.

4 - 6 weeks

AI Business Value Discovery Workshop on IBM Z & LinuxONE

Making transaction processing smarter and more valuable

By applying AI decisioning directly within your high-volume, mission critical workloads and applications already running your business, this free-of-charge, 1-2 day workshop will identify and demonstrate high-value that matters to your business: revenue growth, operating margin, customer experience, risk reduction, and more.

Working with IBM experts, selected opportunities are rapidly validated without disrupting operations through a focused Proof of Concept, demonstrating measurable improvements in decision velocity, consistency, and business outcomes on your core transactional systems, IBM Z ® and LinuxONE ™.

Contact us at *purvi@us.ibm.com, swarren@us.ibm.com*

Required Client Personas for Workshop

- Line of Business and/or Business Analyst
- Chief Data Officer
- AI Leader | Data Scientist
- Application Architect
- Infrastructure Architect

Identify and Assess Business Value

Pinpoint high-impact AI opportunities within your mission-critical transaction flows to drive growth, reduce risk, and unlock deeper operational and market insights from your core systems.

Observe Transactional AI in Action

Witness intelligence elevate at the transaction level as AI infusion transforms routine business decisions into valuable insights that unlock competitive differentiation and untapped business potential.

Design Your Proof of Concept

Co-create your Transactional AI proof of concept with IBM and your key leaders to move beyond ideas, proving value early and creating momentum toward confidently deploying intelligence into everyday business execution.

Experience more with IBM

Visit us at the IBM Booth #113

After a full day of technical sessions, take a break with us!

Connect with our experts, snap a photo with the z17 Plexi or the latest Telum II, and get an up-close look at our Spyre Accelerator.

Come back each day for fresh topics and demos at our expert stations.

Think 2026

Join 5000+ senior business and technology leaders who are seizing the AI revolution to unlock unprecedented growth and productivity at Think 2026.

Find out more information using the QR code below.

IBM Digital Asset Haven

IBM Digital Asset Haven is the operational backbone for financial institutions and regulated enterprises entering the digital asset economy.

Find out more information using the QR code below.

Additional AI on IBM Z AI Practitioner Resources

Key Resources

AI on IBM Z Seismic Page
One stop shop for all AI on IBM Z sales kits, enablement information and others…

AI on Z Github 101
Repository of information on AI tooling and frameworks available in the IBM Z and IBM® LinuxONE ecosystem

Journey to AI Content Solution
Hands on guides to get started with top use cases

Solution Templates
Hands on guides to get started with top use cases

IBM Z and IBM® LinuxONE Container Registry
Download open-source frameworks and tooling vetted for security from this registry

Redbook Papers

- What AI Can Do for You: Use Cases for AI on IBM Z
- Optimized Inferencing and Integration with AI on IBM zSystems: Introduction, Methodology, and Use Cases
- Enriching Linux on IBM Z Workloads with AI
- Finding an On-Ramp to Your AI on IBM Z Journey
- IBM Synthetic Data Sets
- Turning Data into Insight with Machine Learning for IBM z/OS
- Securely Leverage Open-Source Software with Python AI Toolkit for IBM z/OS
- IBM Cloud Pak for Data on IBM Z

AI on Z and LinuxONE Sessions at SHARE

Generative AI, Assistants and Agents

watsonx Assistant for Z - for Db2 Systems Programmers and DBAs
Salon 19
Tues, Feb 24, 10:30am - 11:30am
Speaker: Maryela Weihrauch

IBM Concert for Z, an AI-Powered Mainframe Resilience Platform: Solutions vs. Alerts
Salon 22
Wed, Feb 25, 9:15am - 10:15am
Speakers: Domenico D'Alterio, Fabricio Miatto

Using IBM AI Generative Tooling to Assist with Code Development
Salon 22
Thurs, Feb 26, 10:30am - 11:30am
Speakers: Ben Hicks, Mary Julian

MQ Meets Agentic AI: Intelligent Workflows on z/OS
Salon 19
Wed, Feb 25, 3:45pm - 4:45pm
Speaker: Toby Keegan

An Introduction to Leveraging AI to Increase Productivity in CICS
Salon 17
Wed, Feb 25, 9:15am - 10:15am
Speakers: Kye Maloy, James Davies

CICS 6.3 is AI Ready!
Salon 17
Mon, Feb 23, 3:45pm - 4:45pm
Speaker: Kye Maloy

Data Center Automation - Z System Automation and Agentic AI
Salon 21
Mon, Feb 23, 9:45am - 10:05am
Speakers: Johannes Hausch, Ash Mahay

AI on Z in General

AI on IBM Z and LinuxONE Strategy for 2026 and Beyond!
Salon 22
Mon, Feb 23 2:30pm - 3:30pm
Speaker: Elpida Tzortzatos

Unlocking Business Insights: Deep Dive Into AI on IBM Z and LinuxONE
Salon 22
Tues, Feb 24 9:15am - 10:15am
Speakers: Purvi Patel, Steve Warren

Securing AI on Z: Addressing Emerging Threats and Building Trustworthy Systems
Salon 20
Wed, Feb 25, 2:30pm - 3:30pm
Speakers: Gregg Arquero, Elijah Swift

AI-Infused z/OS

AI Infused z/OS: Overview and Updates
Salon 14
Thurs, Feb 26, 9:15am - 10:15am
Speaker: Anastasiia Didkovska

The 'Next-Generation z/OS Experience' Overview
Salon 13
Mon, Feb 23, 2:30pm - 3:30pm
Speaker: Roger Bales

ML/AI Business Insights

Accelerated Business Decisions with Machine Learning for z/OS
Salon 22
Tues, Feb 24, 10:30am – 11:30am
Speakers: Steve Warren, Purvi Patel

Hands-on Labs

BYOD Lab: Advanced AI Insights Leveraging Hugging Face Large Language Models as Part of a Multi-Model Approach on IBM Z and LinuxONE
STE Lab
Tues, Feb 24, 3:45pm - 4:45pm
Speakers: Steve Warren, Purvi Patel

BYOD Lab: WXA4Z Agentic Hands-on Workshop
Salon 1
Mon, Feb 23, 1:15pm - 2:15pm
Speaker: Dan Snyder

BYOD Lab: AI Enabled Proactive Monitoring to Get the Most From Your System With IBM Concert for Z
Salon 22
Wed, Feb 25, 1:15pm - 2:15pm
Speakers: Domenico D'Alterio, Fabricio Miatto

BYOD Lab: Using the Spyre Accelerator (and Telum II) to Operationalize AI at Scale on IBM z17
STE Lab
Tues, Feb 24, 10:30am - 11:30am
Speaker: Artem Minin

Your feedback is important!

SHARE
EDUCATE • NETWORK • INFLUENCE

Submit a session evaluation for each session you attend:

www.share.org/evaluation