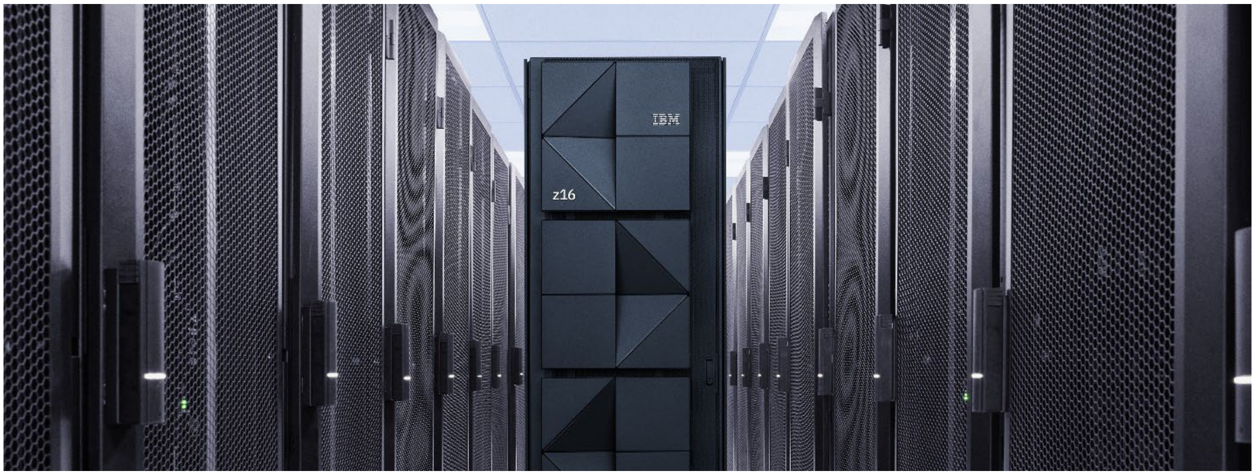
Cambrian AI Research
explore the explosion

What If IBM Z Could Help Stop Fraud?

Last year, IBM launched the z16 with an integrated AI accelerator on each CPU chip. Now, with the infusion of AI into IBM z/OS and a robust AI open-source toolkit, IBM Z customers can realize low-latency AI on a highly trustworthy and secure enterprise system: the modernized IBM Mainframe.

A year ago, we wondered aloud why every IBM Z customer wouldn't leverage the new on-platform AI capabilities the new IBM Telum processor offers, integrating AI analytics with operational data and core business transactions in a single system. Increasingly, institutions using IBM Z are deploying or investigating the opportunities IBM z16 provides for AI analytics. This paper explores the motivations and benefits of integrating AI inference processing on the IBM z16 with a new AI Toolkit, AIOps, and an AI refresh for IBM z/OS.

The advent of AI is revolutionizing many industries. Now, it has entered the domain of mainframe computing, where many of the world's largest banks and insurance companies run their mission-critical workloads. With the launch of the IBM Telum processor on the IBM mainframe, customers have a compelling reason to embrace AI technology on IBM Z, integrating AI into transaction processing and other vital workloads where the data resides on the mainframe and where security and reliability protect the backbone of the business.



AI on the Mainframe

Mainframe owners and operators are notoriously conservative. They must be able to survive as they manage the core business processing of banks, insurance companies, and other critical workloads across industries. Mistakes are not allowed, and they avoid risk at all costs. Consequently, to help these clients realize more business value from their mainframe investments, IBM had to check all the boxes. The company needed to deliver high value (like scoring every transaction instead of sampling) and near-zero risk.

They started with high performance, translating to low-latency AI processing, enabling in-transaction AI analytics for areas like credit card fraud detection. Second, IBM engineered the IBM z16 hardware and software to deliver massive scale, providing millions or even billions of AI inferences daily. Finally, AI had to become accessible. IBM turned to the vast open-source AI models, then battle-tested them and provided elite support.

Enterprises have been watching the growth of AI for non-mission-critical applications over the last few years. And while ChatGPT opens a new chapter for enterprises, promising exceptional value in areas like customer service chatbots, the core of business operations has remained off-limits due in part to the unacceptable latency of off-platform analyses. In such environments, AI can raise as many questions as it answers or even more.

An enterprise must infuse AI into its operations and transactions for better decision-making while keeping the data and processing on a secure and trusted platform. That means that instead of bringing the data to AI processing, you need to get AI processing to the data. And that means, for many companies, to IBM Z.

The Role of IBM Z for AI in the Enterprise

For many companies, IBM Z stores and processes the company's heartbeat. Every transaction, every fraud check, and every claim goes through IBM Z. Since AI inference processing can now be conducted efficiently in situ on the IBM z, companies can leverage AI into core business workloads and data repositories without incurring additional infrastructure expenses and avoiding potential security and privacy risks. Many organizations already execute their business decisions on the platform through the IBM Operational Decision Manager (ODM). Adding AI to the ODM decision processing helps refine decision processing even further. A widely held industry belief is that combining decisions and AI is a best-of-both-worlds approach.

Integrating AI into IBM z/OS

IBM z/OS 3.1 ushers in a new era of operating system intelligence. This latest version of IBM z/OS incorporates AI capabilities throughout the system, facilitating intelligent systems administration guidance and operations that continuously learn and enhance performance.

The introduction of AI System Services for IBM z/OS empowers the system to optimize IT processes, streamline system management, boost performance, and reduce the need for complex skill sets. The initial implementation of this advancement is in AI-powered WLM (Workload Manager), which intelligently anticipates upcoming workloads and dynamically adjusts system resources for optimized efficiency. These capabilities leverage Machine Learning for IBM z/OS and the IBM z16 processor improvements to enhance the capabilities of IBM z/OS.

The New AI Toolkit for IBM Z

The AI Toolkit for IBM Z and IBM LinuxONE offers clients IBM Elite Support for popular open-source and IBM non-warranted AI programs. This support gives clients confidence in deploying open-source AI in production, including programs like TensorFlow, TensorFlow Serving, NVIDIA's Triton Inference Server, IBM's Snap ML, and IBM Z Deep Learning Compiler.

The solution adopts a "free to download and use but only pay for support" approach to open-source AI, ensuring clients receive the correct value at different stages of their AI journey. It extends IBM Elite support to Linux on IBM Z clients. It offers assistance with high-performing open-source AI serving frameworks and other frameworks that leverage IBM Z Integrated Accelerator for AI on IBM z16.

The offering simplifies support licensing and provides clients with IBM Certified containers for popular open-source and IBM non-warranted AI programs. These containers are optimized to run seamlessly on IBM Z and IBM LinuxONE, having undergone thorough container vulnerability and security tests conducted by IBM, enabling clients to deploy open-source AI with confidence.

AIOps Improves IT Efficiency

Another critical attribute of AI on IBM Z is helping to drive efficiencies in IT operations. The IBM z16 leverages AI built into the system and the operating system for AI-powered infrastructure designed to enable automation and efficiencies for productivity in critical tasks. For example, AI can help an IBM z/OS practitioner to free up time from administrative tasks to focus on innovation.

AIOps includes a new Chatbot called ChatOps, which supports collaborative incident remediation. ChatOps integrates with various IBM z/OS subsystems, including WLM, NetView, OMEGAMON, IBM Z Log, Data Analytics, and IBM Z Anomaly Analytics.

IBM watsonx Now Provides Code Generation for Application Modernization

IBM has added a new 20 billion-parameter generative AI model to help refactor, transform, and validate COBOL code, speeding time-to-value and augmenting skills for critical application modernization on IBM Z. Potential benefits include

- Accelerating code development and increasing developer productivity throughout the application modernization lifecycle
- Managing total cost, complexity, and risk of application modernization initiatives, including translation and optimization of code in-place on IBM Z
- Expanding access to a broader pool of IT skills and accelerating developer onboarding
- Achieving high-quality, easy-to-maintain code through model customization and the application of best practices

For more information, see IBM Unveils watsonx Generative AI Capabilities to Accelerate Mainframe Application Modernization

Prime Uses Cases for AI on IBM Z

IBM clients are already adopting AI on their IBM Z systems. One of the most compelling and financially attractive is fraud detection. Typically, a bank can only sample transactions randomly to determine likely fraud. With AI integrated into IBM Z, companies can detect deception in nearly every transaction. Another benefit of this approach is to improve customer satisfaction, with fewer false positives and avoiding becoming fraud victims. It is also reasonable to expect significant banks to generate additional service revenue by offering customers fraud detection as a paid-for service. They can reduce losses due to fraud and recapture revenue associated with rejected transactions.

Another everyday use case is processing insurance claims and identifying potentially fraudulent claims in real time. This leverages on-chip AI inferencing to enable claims processing at speed and scale while stopping fraud before payout rather than chasing debt.

Conclusions

The IBM z16 and IBM Telum processors' integration of AI capabilities began a new and compelling value proposition for mainframe customers. With its excellent performance, efficiency, seamless integration, robust security, and future-proof design, deploying AI on the Telum processor brings transformative benefits to mainframe environments. IBM has now extended the analytic solutions available on IBM Z with an AI Toolkit, a new IBM z/OS infused with AI, Code Generation and AIOps to manage the AI workflow.

IBM Z customers stand at the forefront of integrating AI and transactional processing. They are embracing this innovative technology stack, unlocking the full potential of AI and achieving new levels of efficiency and competitiveness in the AI era.

IMPORTANT INFORMATION ABOUT THIS PAPER

Author and Publisher

Karl Freund, Founder and Principal Analyst, Cambrian-AI Research LLC

Inquiries

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Cambrian-AI Research". Non-press and non-analysts must receive prior written permission by Cambrian-AI Research for any citations.

Licensing

This document, including any supporting materials, is owned by Cambrian-AI Research. This publication may not be reproduced, distributed, or shared in any form without Cambrian-AI Research's prior written permission.

Disclosures

This document was developed with IBM Inc. funding and support. Although the document may utilize publicly available material from various vendors, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.