

**Celent.**

# **Mitigating Fraud in The AI Age**

Supporting Transaction Fraud Detection at Scale on IBM z17

Neil Katkov, PhD

April 2025

This report was commissioned by IBM, which asked Celent to design and execute a Celent study on its behalf. The analysis and conclusions are Celent's alone, and IBM had no editorial control over report contents.

A part of GlobalData

# Contents

<b>Executive Summary .....</b>	<b>3</b>
<b>The Expanding Fraud Threat.....</b>	<b>4</b>
<b>Advanced Tools for Anti-Fraud .....</b>	<b>6</b>
Machine learning .....	6
Transformer technology and generative AI.....	6
<b>Enabling Real-Time Fraud Detection in High-Volume Transaction Environments.....</b>	<b>7</b>
<b>Quantifying the Benefits of Running AI on the Mainframe .....</b>	<b>9</b>
Capturing banking fraud, transaction by transaction.....	9
Fraud prevention for insurers.....	11
<b>Path Forward: The Future of Fraud Prevention on the Mainframe .....</b>	<b>12</b>
<b>Leveraging Celent’s Expertise .....</b>	<b>13</b>
Support for Financial Institutions.....	13
Support for Vendors.....	13
<b>Related Celent Research .....</b>	<b>14</b>

# Executive Summary

---

To counter the increasingly sophisticated and technology-enabled fraud targeting the financial services and other industries, these institutions and their technology partners are developing ever-more advanced solutions to detect and prevent fraudulent activity, including techniques that leverage multiple predictive deep learning and generative AI models. However, advanced anti-fraud models typically run on peripheral hardware separated from the transactions and data, creating a “last mile” challenge in deploying AI models at scale for large institutions that run mission critical transactions on mainframe computers. In addition, the need for more processing power and technology resources continues to grow as the models become larger and more complex.

IBM went a long way in overcoming this barrier with the release of its IBM Telum<sup>®</sup> processor at the core of IBM z16<sup>™</sup> in 2022. The Telum processor incorporated an AI accelerator that supports high-throughput AI inferencing directly within the mainframe environment, at 3 million inferences per second, according to IBM.

This breakthrough enables even the largest banks and businesses to run complex models against 100% of transactions in real time—not just a sample of as little as 10 or 20%, as is often the case. It also solves the latency and throughput issues that are often caused by sending core transactions off the mainframe for inferencing, and without additional security risks. This enabled businesses to capture millions of dollars more of fraud before the payment or transaction was complete.

Three years later, IBM is releasing its second-generation AI accelerators, the Integrated Accelerator for AI in the IBM Telum<sup>®</sup> II processor and IBM Spyre<sup>™</sup> Accelerator (available in Q4 2025, according to IBM), as part of the new IBM z17<sup>™</sup> mainframe. This new chipset and architecture enable sharing of AI processing power, so IBM z17 transactions can access eight times the AI processing units compared to IBM z16. Moreover, the accelerators are tuned to handle large language models (LLMs), and with significantly increased AI computational throughput compared to z16.

A straightforward measure of the potential benefit of IBM’s AI accelerators is how much additional fraud institutions can expect to capture by running advanced AI inferencing on all transactions instead of a sample. Celent estimates that banking, cards, and payments transactions are worth US\$1,471 trillion annually. This is a US\$340 trillion—or 30%— increase over 2021. 70% of all transactions, or US\$1,025 trillion—roughly US\$1 quadrillion—run on IBM Z mainframes, roughly the same proportion as in 2021.

Celent further estimates that banking, card, and payments fraud totaled a whopping US\$533 billion in 2024, with US\$374 billion occurring at firms that run Z. Running all transactions through AI models using IBM z17’s advanced AI acceleration at scale—instead of sampling a small number— could improve fraud capture in banking, cards, and payments by US\$190 billion annually.

For a tier 1 bank using IBM z17, this is equivalent to US\$208 million dollars in stopped fraud—before it happens—per year, and US\$35 million in stopped fraud for a tier 2 bank.

Financial institutions, insurance companies, and indeed any business or organization that is running payments and credit cards or evaluating fraud within transactions that run on IBM Z will need to make their own assessment of the business case before proceeding. Nevertheless, the ability to perform the most advanced AI inferencing techniques directly in the mainframe environment is a promising tool in the fight against fraud. Moreover, there are numerous other use cases, such as anti-money laundering and anomaly detection, which require secure, compliant, and protected AI operations that could benefit from on-chip AI inferencing in IBM Z environments.

# The Expanding Fraud Threat






Fraud losses have grown steadily since being thrown into overdrive by the pandemic. Aided and abetted by AI and automation, fraud in the banking, cards, and payments industries reached US\$535 billion globally in 2024.

Over the past five years, fraud has become increasingly sophisticated as criminals use machine learning and automation technologies to boost both the effectiveness and the pervasiveness of their scams. While traditional typologies like check fraud are still a major threat, fraud in the 2020s is characterized by social engineering-based trickery, including authorized push payment fraud, invoice fraud, and business email compromise (BEC). Impersonation scams target businesses, charities, and government agencies alike.

Scams are assuming an increasingly human face as criminals use readily available generative AI tools to power text and voice-based deceptions, at scale. Appallingly, romance scams and other confidence plays are perpetrated by industrialized criminal organizations operating out of at-risk jurisdictions using human-trafficked labor.

Virtually every step in the financial services value chain, from account opening onwards, and every product—including retail and corporate payments, loans, and cards—is targeted by fraud. At the same time, the ongoing digitization of financial services offers new opportunities to fraudsters. A good example is account-to-account transfers. While providing individuals with an exceptionally convenient means of payment, account-to-account transfers have also generated billions of dollars in fraud.

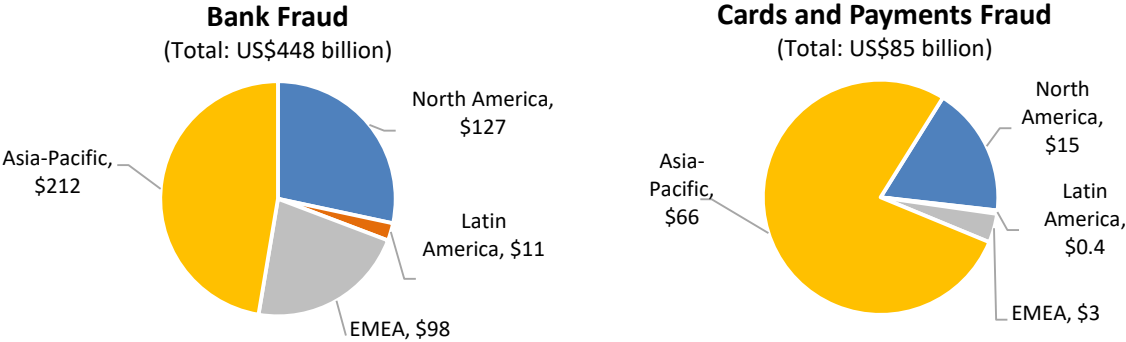
Figure 1: Trending Fraud Threats

Tactical Fraud		Social Engineering	
	APP fraud		Confidence scams (family, relatives)
	Business email compromise		Impersonation scams (business, gov't, CEO, etc.)
	Invoice fraud		Romance scams, pig butchering

Source: Celent

These and other trends have contributed to steady annual growth of 11.1% in banking fraud between 2021 and 2024 and 14.2% in cards and payments fraud (CAGR basis). Celent estimates that banking losses to fraud reached \$448 billion globally in 2024, a whopping \$120 billion more than in 2021. At the same time, cards and payments fraud soared to \$85 billion in 2024, \$28 billion more than in 2021.

**Figure 2: Banking, Cards, and Payments Fraud Losses in 2024**



Source: Celent estimates based on BIS transaction data and central bank fraud data.  
Note: Bank fraud includes credit transfers, direct debits, and checks. Cards and payments fraud includes credit and debit cards, e-payments, and other payments.

# Advanced Tools for Anti-Fraud

---

Machine learning models for combatting fraud continue to evolve, leading to higher fraud capture rates. In addition, transformer technologies such as generative AI and quantitative foundation models are now pushing fraud detection even farther.

## Machine learning

Financial institutions are making use of deep learning to analyze transaction data at scale to detect potentially fraudulent activity, including new, previously unseen typologies. New techniques being applied to the fraud problem include recurrent neural networks (RNN), which enhance the ability of the model to analyze time sequences and have been demonstrated to enhance fraud detection results considerably. Another development is advanced AI techniques that deploy multiple machine learning models simultaneously against transaction data in order to enhance results.

Fraud models are also making use of increasingly extensive data sets, in addition to core transaction data. Customer data, biometric data, and consortium data are all being leveraged to add features to models to increase their ability to detect fraud.

## Transformer technology and generative AI

Since transformer technology burst onto the scene in 2022 through the release of readily accessible generative AI tools, banks and payments processors have moved quickly to apply the new technology to fraud detection. Institutions are applying large language models (LLMs) to support model development, create synthetic data for training models, and derive insights from qualitative fraud data. Quantitative foundation models—also called large transaction models (LTMs)—are being used to analyze quantitative transaction data to predict fraud and apply these insights to machine learning models to enhance fraud detection.

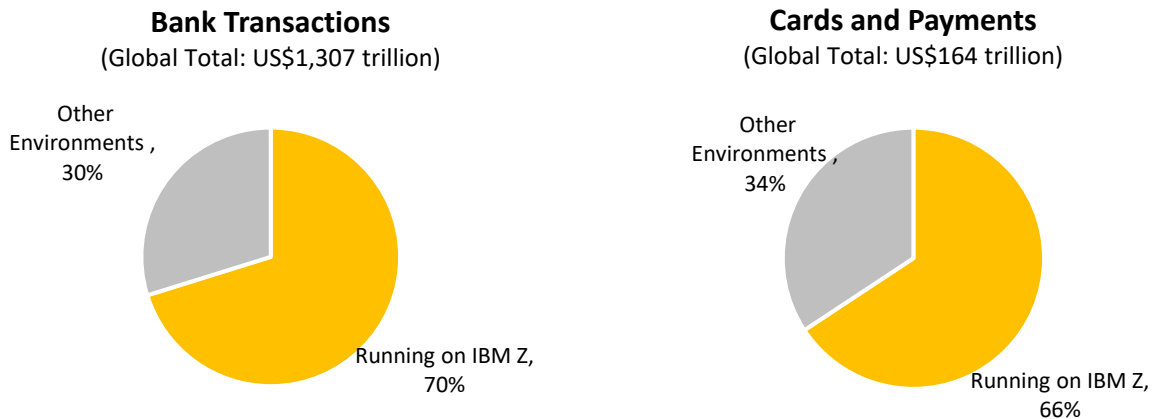
The application of transformer models directly in the detection environment is on the horizon. Fraud detection will only be able to fully benefit from advances in machine learning and transformer models if financial institutions can run these increasingly complex models in real time on all of their transactions, a “holy grail” use case that is likely to first be achieved with proprietary, tightly focused foundation models.

# Enabling Real-Time Fraud Detection in High-Volume Transaction Environments

Operational approaches to fraud detection for mainframe environments often rely on off-platform fraud detection systems that limit fraud scoring to a sample of—typically high-risk—transactions. The inability to apply advanced AI models to all transactions means that a significant amount of fraud—possibly a majority of fraud—slips out the door.

Celent estimates that banking, cards, and payments transactions are worth US\$1,471 trillion annually, a US\$340 trillion—or 30%—increase over 2021. Many large financial institutions rely on mainframe computing platforms for their core operations. At least two-thirds of the top 100 banks run on IBM Z mainframes. Most of the major cards and payments processors—including a number of new entrants—also use IBM Z mainframes. Celent estimates that, globally, 70% of transactions by value, or US\$1,025 trillion—roughly US\$1 quadrillion—run on IBM Z.

**Figure 3: Banks, Cards, and Payments Transaction Value on IBM Z**



Source: Celent

Large financial institutions running on mainframes typically send transactions to anti-fraud solutions running on peripheral systems. The round-trip introduces a degree of latency which, in high-volume, real-time environments, can impact SLAs or even cause transactions to time out. As a result, large financial institutions are often forced to score only a fraction of transactions in real time, which greatly limits their ability to detect fraud.

IBM went a long way in overcoming this barrier with the release of its Telum processor for the IBM z16 mainframe in 2022. The Telum processor incorporated an AI accelerator that supports high-throughput AI inferencing directly within the mainframe environment. This breakthrough enables even the largest banks to run complex models against all transactions in real time, and without the latency and throughput issues that are often caused by sending core transactions off the mainframe to peripheral fraud systems.

This ability to apply advanced models to all transactions—not just a sample as is often the case—enables more fraud capture compared to sampling. The logic is straightforward: More transactions scored will result in more fraud being prevented. It is even better when you can score more with minimal to no impact to performance or SLAs, as IBM asserts.

Anti-fraud models have become more complex in the years since, with data scientists leveraging internal and external data sets and deploying multiple models to detect fraud. Significantly, generative AI is now being used to enhance machine learning models, and Celent believes that before long transformer models will be utilized directly in real time inferencing. All this complexity puts added compute demands on the systems running the models.

IBM has not lost sight of these trends. As part of its IBM z17 system, they developed the Telum II processor, which incorporates an on-chip AI accelerator with eight times the AI processing units compared to IBM z16. In the fraud context, IBM claims that, using a deep learning model for credit card fraud detection, IBM z17 can process up to 5 million inference operations per second—or up to an impressive 450 billion inference operations per day—with 1 millisecond or less response time.<sup>1</sup>

IBM has also developed an off-chip accelerator delivered via PCIe card, the IBM Spyre Accelerator, which boasts 32 accelerator cores. Moreover, the new accelerators are tuned to handle LLMs for both generative and predictive use cases. The use of the predictive analytics capabilities of LLMs will be increasingly relevant as generative AI develops into an integral tool for fraud detection.

The architecture of IBM z17 enables sharing of AI processing power across this chip set, providing compute powerful enough to run multiple AI models in the mainframe environment. Banks, payments processors, and other organizations are increasingly deploying multiple model strategies to improve accuracy and reduce false positives in fraud detection, anti-money laundering, anomaly detection, and other use cases. According to IBM, the AI accelerators in IBM z17 enable organizations to run increasingly larger, more complex models—including LLMs—for real-time workloads to gain better outcomes. Moreover, running these models directly in the mainframe helps safeguard data and the intellectual property of their models.

---

<sup>1</sup> Performance result is extrapolated from IBM internal tests running on IBM Systems Hardware of machine type 9175. The benchmark was executed with 1 thread performing local inference operations using a LSTM based synthetic Credit Card Fraud Detection (CCFD) model (<https://github.com/IBM/ai-on-z-fraud-detection>) to exploit the IBM Integrated Accelerator for AI. A batch size of 160 was used. IBM Systems Hardware configuration: 1 LPAR running Red Hat® Enterprise Linux® 9.4 with 6 IFLs (SMT), 128 GB memory. 1 LPAR with 2 CPs, 4 zIIPs and 256 GB memory running IBM z/OS® 3.1 with IBM z/OS Container Extensions (zCX) feature. Results may vary.

# Quantifying the Benefits of Running AI on the Mainframe

The IBM z17 mainframe environment incorporates its second-generation AI accelerator, designed to run advanced AI inferencing directly on the mainframe at industrial scale. Celent estimates that, by supporting advanced fraud detection for all transactions running on IBM Z, the new AI accelerator could potentially reduce banking, cards, and payments fraud losses by US\$190 billion globally, on an annual basis.

We have seen that a majority of global financial transactions run on IBM Z. At the same time, AI-based fraud detection typically takes place on peripheral systems, constraining the ability to run fraud detection in real time without adversely impacting SLAs. IBM’s Telum II and Spyre Accelerator eliminate this barrier by supporting advanced AI inferencing directly in the mainframe environment. Financial institutions using the new accelerators can run the most advanced inferencing routines against 100% of transactions, not just a sample, and thereby significantly reduce their vulnerability to fraud.

Quantifiable benefits of advanced AI fraud detection on IBM z17 mainframes:	Reduce industry fraud losses by...		Reduce losses per bank by...		Reduce insurance fraud by...
	<u>US</u> 6.3¢ per \$100	<u>Globally</u> 2.5¢ per \$100	<u>Tier 1 US Bank</u> US\$208 million	<u>Tier 2 US Bank</u> US\$35 million	US\$83 billion (digital channels only)

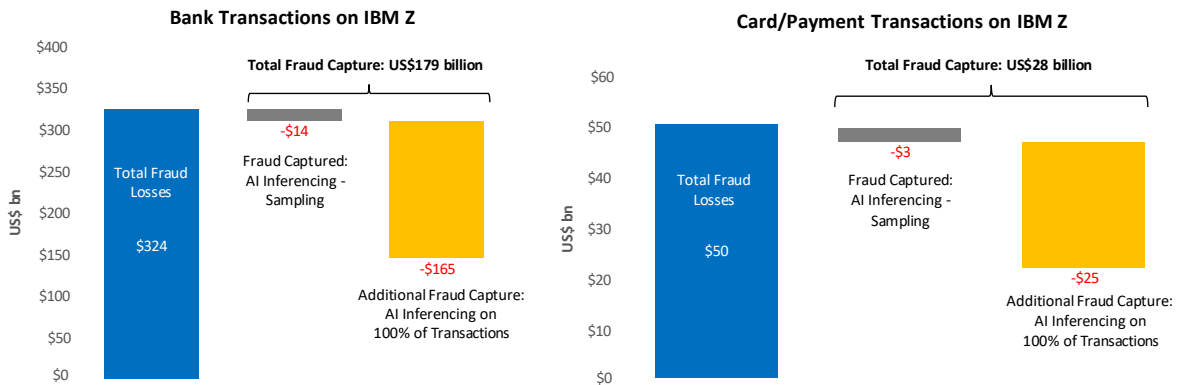
## Capturing banking fraud, transaction by transaction

Banks can reap the full potential of modern inferencing technology by running advanced models against all transactions. Looking at the benefits at the transaction level, Celent estimates that applying advanced AI models to all banking transactions would potentially reduce fraud losses by 2.5¢ cents for every \$100 of transactions globally (2.5 basis points).

In the US, where fraud rates are higher than the global average—9.3¢ for every \$100 compared to 4.0¢ globally—fraud losses could be reduced by even more: 6.3¢ for every \$100. This is equivalent to saving the bank US\$1.64 for an average transaction of US\$2,612.

If all banking, cards, and payments institutions currently running on IBM Z were to make use of this capability, what would the benefits be? Celent estimates that sending all transactions—not just a sample—currently running on IBM Z mainframes through advanced AI models could result in US\$190 billion in additional fraud captured globally, compared to sampling, for a total US\$207 billion in fraud reduction. This includes a US\$165 billion reduction in bank fraud losses and a \$25 billion reduction in cards and payments fraud. In the US alone, bank fraud losses could be reduced by \$51 billion and cards and payments fraud losses by \$7 billion.

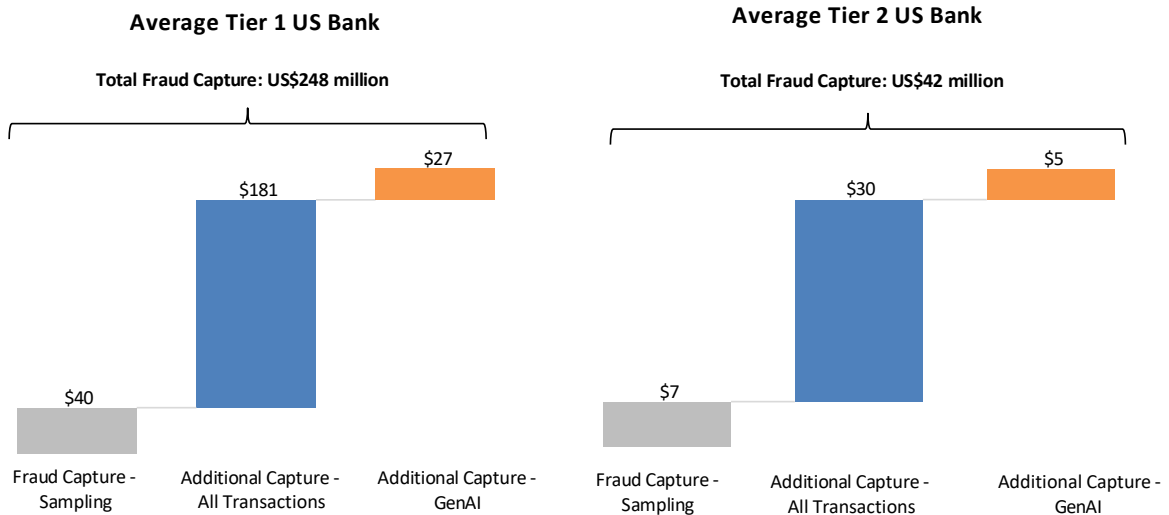
**Figure 4: Potential Fraud Loss Reduction with Advanced AI Inferencing on IBM Z**



Source: Celent

What do the benefits look like for an average financial institution? Celent estimates that for a Tier 1 US bank on IBM z17, running all transactions through advanced inferencing models—compared to applying AI models to only about a tenth of transactions—could reduce fraud losses by an additional US\$181 million compared to sampling. Factors contributing to this uplift include the ability to run multiple models and datasets on all transactions, thanks to the multiple AI accelerators running in parallel in IBM’s Spyre configuration. Additionally, for banks that apply generative AI to fraud detection, banks can expect a minimum 15% in incremental fraud capture, or \$27 million, for a total of US\$208 million in additional fraud capture compared to sampling.

**Figure 5: Potential Fraud Loss Reduction with Advanced AI Inferencing Per Bank**



Source: Celent

We also estimate that Tier 2 US banks could avoid an additional US\$30 million in fraud losses on average, with an incremental uplift of US\$5 million if leveraging generative AI models, for a total additional fraud capture of \$35 million over sampling.

Because of the very high processing capability of IBM’s AI accelerators, they should be able to support expected increases in computing requirements for fraud detection as machine learning and transformer models continue to improve and become more powerful.

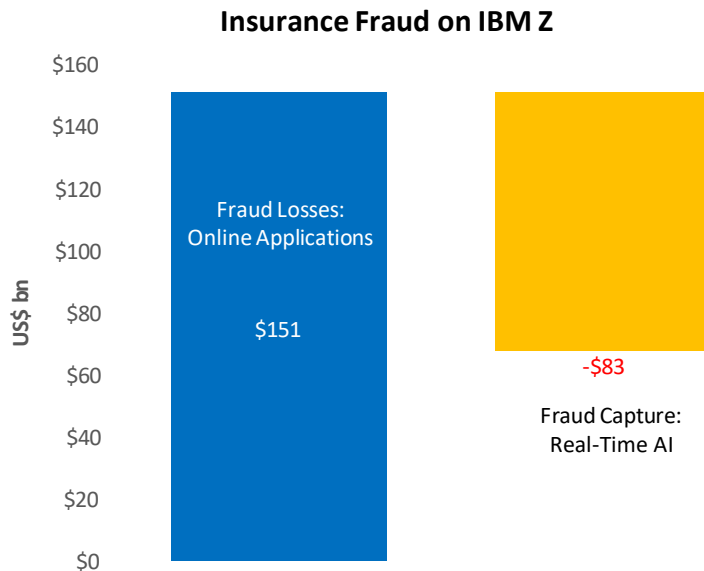
## Fraud prevention for insurers

The insurance industry is also heavily reliant on mainframes; in fact, many top insurers worldwide run on IBM Z. This provides insurers with an opportunity to maximize their fraud reduction efforts by taking advantage of IBM’s Telum II and Spyre Accelerator AI acceleration technology.

The insurance industry has a huge fraud problem. In the property and casualty sector, as many as 10% of claims may be fraudulent.<sup>1</sup> However, insurance claims are typically reviewed before they are paid out. Because of the comparatively lengthy cycle times, insurers do not face the kind of SLA pressure that is leading banks to put their claims fraud models on the mainframe.

To the degree that fraudulent claims are tied to fraudulent applications, though, the logic changes. In particular, insurance applications through digital channels are issued quickly and at scale and could benefit from the speed and accuracy of advanced fraud detection on the mainframe. Celent estimates that applying advanced fraud detection to online insurance applications alone could potentially prevent downstream fraud losses of US\$83 billion globally.

**Figure 6: Potential Insurance Fraud Loss Reduction with Advanced AI Inferencing on IBM Z**



Source: Celent

<sup>1</sup> Coalition Against Insurance Fraud

# Path Forward: The Future of Fraud Prevention on the Mainframe

---

Since the introduction of IBM z16 in 2022, IBM Z mainframes have been able to support AI inferencing directly in the mainframe, making it possible to score 100% of real-time transactions even in high-volume production environments. IBM's second generation of AI accelerators provide further computing power and extreme parallel processing that can run multi-model AI and other advanced, compute-hungry model and data strategies.

These enhanced capabilities will help support a number of evolving challenges that banking, card, and payments institutions are facing in their efforts to reduce fraud.

- **An increasingly scaled payments environment.** The digitization of financial services continues to drive a relentless expansion in the volume, speed, and variety of transactions. Firms will need to ensure their fraud systems can keep up with these scale challenges.
- **The faster “speed to market” of new fraud typologies.** Criminals are increasingly using AI and automation technology to perpetrate fraud at industrial scale. In addition to supporting the scale of real-time transactions, firms will need to focus on fast model development—such as leveraging autoAI techniques—in order to counter quickly emerging fraud typologies.
- **The need to enlist transformer technology in the fight against fraud.** The next opportunity for moving the needle in fraud prevention lies in transformer technologies such as LLMs, LTM, and multi-modal generative AI. Financial institutions will need to keep up with these quickly developing technologies in order to optimize fraud capture—particularly in a world where these same technologies are readily available to the criminals perpetrating fraud.

Fraud is not just a bottom-line issue but can also have profound negative effects on customer experience, customer perception, and reputational risk. Financial institutions running IBM Z should take a close look at what can be gained by moving fraud detection to the mainframe.

# Leveraging Celent's Expertise

---

If you found this report valuable, you might consider engaging with Celent for custom analysis and research. Our collective experience and the knowledge we gained while working on this report can help you streamline the creation, refinement, or execution of your strategies.

## Support for Financial Institutions

Typical projects we support include:

**Vendor short listing and selection.** We perform discovery specific to you and your business to better understand your unique needs. We then create and administer a custom RFI to selected vendors to assist you in making rapid and accurate vendor choices.

**Business practice evaluations.** We spend time evaluating your business processes and requirements. Based on our knowledge of the market, we identify potential process or technology constraints and provide clear insights that will help you implement industry best practices.

**IT and business strategy creation.** We collect perspectives from your executive team, your front line business and IT staff, and your customers. We then analyze your current position, institutional capabilities, and technology against your goals. If necessary, we help you reformulate your technology and business plans to address short-term and long-term needs.

## Support for Vendors

We provide services that help you refine your product and service offerings. Examples include:

**Product and service strategy evaluation.** We help you assess your market position in terms of functionality, technology, and services. Our strategy workshops will help you target the right customers and map your offerings to their needs.

**Market messaging and collateral review.** Based on our extensive experience with your potential clients, we assess your marketing and sales materials—including your website and any collateral.

# Related Celent Research

---

[Mitigating Fraud in The AI Age: Understanding the Challenge](#)  
March 2025

[Dimensions: Financial Crime IT Pressures and Priorities 2024](#)  
August 2024

[IT Spending on Risk Management in Banks: 2024 Edition](#)  
July 2024

[GenAI-oneers in Risk & Compliance: Cross-Sector Survey and Spotlights](#)  
June 2024

[Dimensions: Risk & Compliance IT Pressures & Priorities 2024 Edition](#)  
May 2024

[IT and Operational Spending on Fraud: 2024 Edition](#)  
April 2024

[Operationalizing Fraud Prevention on IBM z16: Reducing Losses in Banking, Cards, and Payments](#)  
April 2022

## Copyright Notice

Copyright 2025 Celent, a division of GlobalData Plc. All rights reserved. This report may not be reproduced, copied or redistributed, in whole or in part, in any form or by any means, without the written permission of Celent, a part of GlobalData ("Celent") and Celent accepts no liability whatsoever for the actions of third parties in this respect. Celent and any third party content providers whose content is included in this report are the sole copyright owners of the content in this report. Any third party content in this report has been included by Celent with the permission of the relevant content owner. Any use of this report by any third party is strictly prohibited without a license expressly granted by Celent. Any use of third party content included in this report is strictly prohibited without the express permission of the relevant content owner. This report is not intended for general circulation, nor is it to be used, reproduced, copied, quoted or distributed by third parties for any purpose other than those that may be set forth herein without the prior written permission of Celent. Neither all nor any part of the contents of this report, or any opinions expressed herein, shall be disseminated to the public through advertising media, public relations, news media, sales media, mail, direct transmittal, or any other public means of communications, without the prior written consent of Celent. Any violation of Celent's rights in this report will be enforced to the fullest extent of the law, including the pursuit of monetary damages and injunctive relief in the event of any breach of the foregoing restrictions.

This report is not a substitute for tailored professional advice on how a specific financial institution should execute its strategy. This report is not investment advice and should not be relied on for such advice or as a substitute for consultation with professional accountants, tax, legal or financial advisers. Celent has made every effort to use reliable, up-to-date and comprehensive information and analysis, but all information is provided without warranty of any kind, express or implied. Information furnished by others, upon which all or portions of this report are based, is believed to be reliable but has not been verified, and no warranty is given as to the accuracy of such information. Public information and industry and statistical data, are from sources we deem to be reliable; however, we make no representation as to the accuracy or completeness of such information and have accepted the information without further verification.

Celent disclaims any responsibility to update the information or conclusions in this report. Celent accepts no liability for any loss arising from any action taken or refrained from as a result of information contained in this report or any reports or sources of information referred to herein, or for any consequential, special or similar damages even if advised of the possibility of such damages.

There are no third party beneficiaries with respect to this report, and we accept no liability to any third party. The opinions expressed herein are valid only for the purpose stated herein and as of the date of this report.

No responsibility is taken for changes in market conditions or laws or regulations and no obligation is assumed to revise this report to reflect changes, events or conditions, which occur subsequent to the date hereof.