PURESTORAGE®

WHITE PAPER

FSI GenAI Pod: Empowering Financial Services with GenAI

From Game-changing Use Cases to Turnkey RAG Solutions

Contents

Foreword

The financial services industry is in the midst of a generational shift. Speed is no longer measured only in trades per second—it's measured in how quickly you can turn raw data into the right decision. KX has built its reputation on delivering real-time, time series intelligence for the world's most demanding markets. Now, KX, NVIDIA, Pure Storage®, and World Wide Technology are partnering to take the next step—empowering institutions with the ability to reason over their data, past and present, through generative AI (GenAI).

The FSI GenAI Pod, a turnkey RAG solution for financial services, brings together the best in AI models, graphics processing unit (GPU) acceleration, high-performance storage, and time-aware analytics in a platform that's ready for the realities of capital markets. It enables research to be conducted in minutes, not hours; client interactions to be informed by live portfolio insights; and trading strategies to be shaped by market signals the instant they emerge. This is not just about faster answers—it's about better answers, grounded in truth, delivered at the speed of the market.

Our goal is simple: to give financial institutions the competitive edge they need in the AI era, with technology they can deploy today, at scale, with confidence.

Introduction

Financial services are at a pivotal crossroads. Traditional banks and asset managers face make-or-break competition from AI-driven upstarts. FinTechs, quantitative hedge funds, and other early AI adopters are rapidly eroding market share with record-breaking returns. In this landscape, it's no longer enough to trade fast—speed of insight and action are the new edge. Firms must extract insights faster than ever before. GenAI offers a path for incumbents to thrive by unlocking smarter, quicker decisions and personalized services. This paper explores five high-impact GenAI use cases in finance and introduces the turnkey retrieval-augmented generation (RAG) solution for financial services—FSI GenAI Pod—that can dramatically accelerate AI adoption. We start with business cases and then dive into the technical architecture enabling them.

Five Game-changing GenAI Use Cases in Financial Services

Leading financial institutions are exploring GenAI to transform operations and customer experiences. Five state-of-the-art use cases stand out.

AI-powered Research Assistant

An AI-powered research assistant automates research, analysis, and report generation. Rather than manually sifting through filings or news, analysts can pose questions to an AI assistant that combs through millions of documents and data points and then delivers a response with supporting evidence. This dramatically speeds up research workflows; for example, answering complex queries in minutes instead of hours. By relying on a rich knowledge base, the assistant provides targeted insights and recommendations, freeing analysts to focus on higher-value tasks.

AI-powered Relationship Manager

An AI-powered relationship manager enhances client interactions by equipping relationship managers with real-time insights and personalized content to enhance interactions with their large institutional clients. An AI-based relationship assistant/copilot can summarize an institutional client's portfolio, market movements, or relevant research in real time, allowing relationship managers to proactively drive new order flows and deepen client engagement. This leads to more informed conversations and faster response to client needs, strengthening loyalty and sales.

Personalized Portfolio Construction

GenAI enables the delivery of AI-assisted, tailor-made investment strategies at scale. By analyzing an individual's objectives and the full spectrum of market data, a GenAI system can propose customized portfolio allocations or trade ideas. It blends real-time and historical data to identify strategies suited to each client, which is far beyond generic model portfolios. Portfolio managers can then refine these AI-suggested strategies, improving outcomes and efficiency in portfolio construction.

Real-time Alpha and Beta Extraction

Leverage GenAI, built on top of temporal AI from KX, to continuously scan market data, identify trading opportunities (alpha), and manage risk/exposure (beta) in real time. KX's high-performance, time series engine ensures that GenAI models are not only context-aware but also temporally precise, enabling the system to spot subtle patterns, microstructure anomalies, or arbitrage opportunities before the wider market reacts. For instance, the combined stack can accurately detect a regime shift or liquidity fragmentation that human traders or legacy quant models may miss. By catching these opportunities and risks first, firms can act preemptively—a significant competitive advantage in fast-moving markets. Early pilots have shown that GenAI-plus-KX temporal AI strategies can capture 150 to 250 basis points of excess annualized alpha compared to traditional quant approaches, particularly during volatile regimes where timing and millisecond-level context are critical.

Deep Learning for Time Series Forecasting

Leverage advanced neural networks to forecast market trends and asset price movements. By training deep learning models on expansive time series data sets (for example, tick data and economic indicators), the system can predict short-term and long-term trends with higher accuracy. This improves everything from algorithmic trading (anticipating volatility or volume spikes) to risk management and scenario planning. The solution continuously refines its models as new data arrives, blending historical context with real-time information for dynamic forecasting.

These use cases are currently delivering real-world return on investment in forward-thinking firms. The common thread is that GenAI is automating complex analysis and decision-making processes—from research to trading—that were previously time-consuming or impractical. The result is improved insights, faster decisions, and even entirely new products and services for customers.

A Differentiated RAG Architecture: KX, NVIDIA, and Pure Storage

Achieving these AI capabilities in production requires a robust and modern architecture. The **FSI GenAI Pod** is built on a differentiated RAG stack that combines the strengths of three industry leaders: **KX**, **NVIDIA**, and **Pure Storage®**. This architecture is designed to deliver accurate, real-time insights by marrying high-performance analytics with state-of-the-art AI models.

KX and KDB.AI (AI Vector Database and Analytics)

At the core is KX technology including the KDB.AI vector database, which can handle both structured and unstructured data and has a rich analytics library. KX is a **recognized leader in high-performance, high-velocity time series analytics**—winning 17 out of 18 STAC benchmark tests—and plays a mission-critical role in capital markets by blending real-time and historical data. This pedigree carries into the GenAI realm. The KDB.AI vector database serves as the RAG **knowledge base**, storing embeddings of both structured and unstructured data (for example, research reports, market data, and customer data). This enables the system to retrieve relevant facts with ultra-low latency whenever the AI model needs to ground its responses. The KX analytics engine also provides in-line data processing and "**time-aware**" analysis so the AI model can reason not just over static text but also dynamic time series patterns (crucial for financial use cases).

NVIDIA NeMo Framework and AI Models

NVIDIA contributes the AI "brain" of the solution. A **fine-tuned, 70-billion-parameter large language model (LLM)**—Nemotron 70B—serves as the generative core capable of understanding finance-specific queries and producing detailed natural language responses. This LLM is augmented by the NVIDIA NeMo Agent toolkit for RAG, which includes a retriever model to fetch relevant context from KDB.AI, a NeMo reranker to sort results by relevance, and an embedding model (NV Embed v2) to convert text into high-dimensional vectors for similarity search. These components work in concert—when a user asks a question, the system generates an embedding of the query and retrieves the top relevant documents from the vector database. The LLM then uses that context to produce an answer with cited sources.

In addition, any derived mathematical calculations required on the time series market data are performed by the KDB.AI database and results are passed to the LLM so it can embed the derived values within the final LLM response to the prompt. This **RAG loop** ensures responses are grounded in real data (reducing hallucinations) while still leveraging the LLM's natural language prowess. NVIDIA's GPU infrastructure underpins the whole pipeline, delivering the compute needed for both training and inference of these deep learning models.

Pure Storage High-performance Data Pipeline

Pure Storage provides the data foundation that keeps this AI engine fed with information at speed and scale. All stages of the AI workflow—from ingesting raw data to training models and serving live queries—depend on fast, reliable storage. Pure Storage Enterprise Data Cloud is a **leader in all-around performance**, offering massive read/write throughput and microscopic latency (tens of millions of file operations per second). This means the FSI GenAI Pod can load training data sets, write model checkpoints, and retrieve vector embeddings without I/O bottlenecks. The Pure Storage platform is also **extremely efficient**, delivering up to 5x better power efficiency (only 0.9kW per petabyte) and 5x better space efficiency (0.5PB per rack unit) than traditional SSD—figures that are doubling in 2025.

This efficiency translates to lower operational costs and a smaller data center footprint, which is crucial because financial AI workloads often involve petabytes of data. Equally important, Pure Storage brings enterprise-grade reliability (six nines availability and far fewer component failures than commodity storage). In a 24X7 trading environment, this resilience ensures AI services remain always-on. Finally, Pure Storage platform storage integrates seamlessly. Features like integrated networking and nondisruptive expansion make it simple to scale performance or capacity on demand, with no complex reconfiguration. In short, Pure Storage provides a **fast, efficient, and rock-solid data pipeline** for the FSI GenAI Pod.

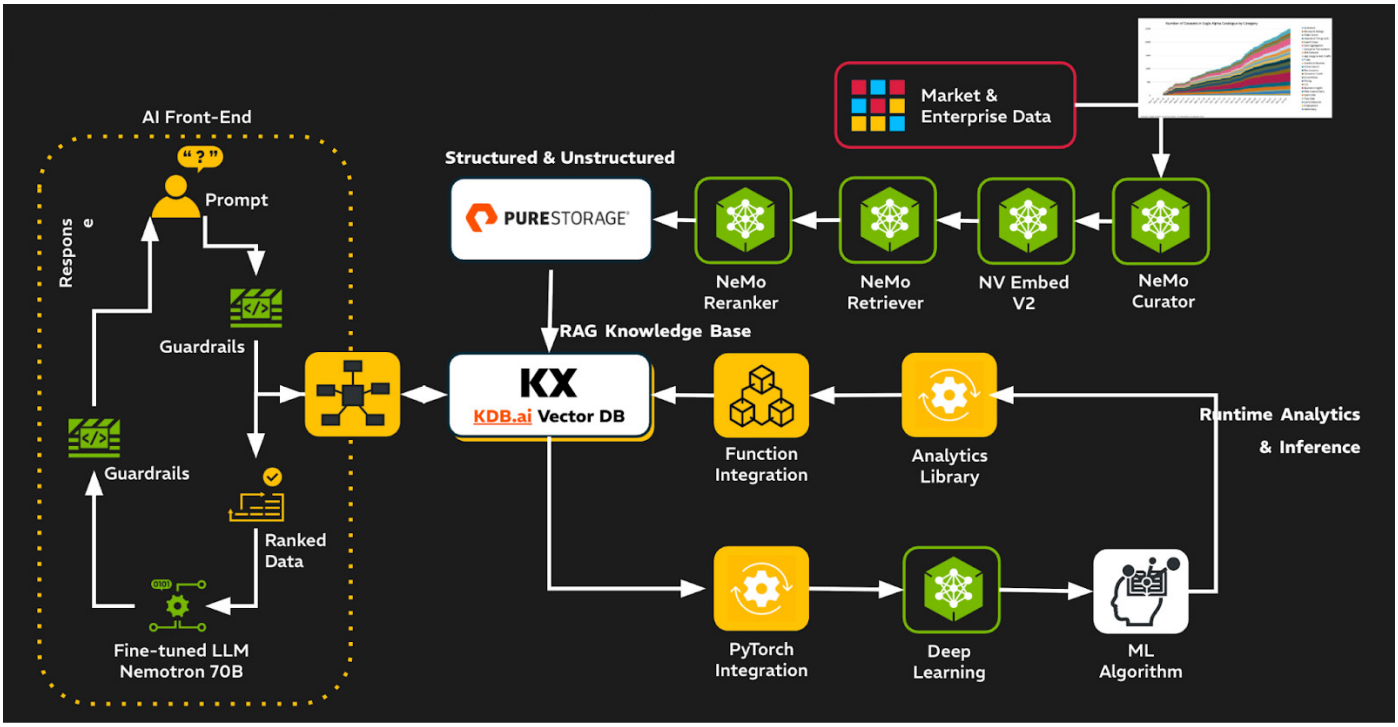

FIGURE 1 AI-powered research assistant architectural diagram

By tightly integrating KX real-time analytics, the Pure Storage data pipeline, and NVIDIA AI models, this architecture delivers **RAG tailored for financial services**. The system can ingest **all data**—from market feeds to internal research—and serve up insights in natural language with the speed and accuracy that modern markets demand. Notably, it produces accurate and efficient RAG responses with no hallucinations by grounding answers on a shared knowledge base. This differentiated stack turns the promise of GenAI into a practical reality for finance.

Covering the Complete AI Pipeline

One strength of the FSI GenAI Pod is that it spans the **entire AI pipeline**, not just model training or a point solution. It provides an end-to-end platform for data-driven insight generation, covering every stage from **data ingestion** to **analysis and persistence**. In practice, this looks like:

1. **Ingestion and processing**: The pipeline can ingest data from myriad sources (including market data feeds, analyst reports, SEC 10-K/10-Q filings, proprietary documents, customer relationship management systems, and news feeds) and then curate and preprocess it for AI consumption. The KX platform is adept at streaming real-time data as well as batch processing historical data. This ensures the **raw material** for the AI model—be it structured time series data or unstructured text—is continuously fed and organized.
2. **Training**: The environment supports training and fine-tuning AI models using the ingested data. Whether it's refining the LLM on proprietary research reports or training a new time series forecast model, the system's GPU-accelerated compute and high-throughput storage enable iterative training cycles. Checkpointing and versioning of models is backed by Pure Storage platform storage for reliability and speed.
3. **Inference**: When deployed, the LLM and associated models run inference to serve user queries or automated tasks. Thanks to the RAG setup, inference isn't just a raw LLM response—it involves retrieving relevant facts from the vector database (ingestion makes those facts available) and then generating a context-aware answer. Data from tables, graphs, and images in PDF documents is also extracted as required and forms part of the response. This **fusion of search and generation** provides intelligible, up-to-date answers to research analysts, relationship managers, portfolio managers, traders, and clients in real time.
4. **Analysis and decision support**: The outputs of the AI model (forecasts, recommendations, and extracted insights) feed into analysis tools. The KX analytics library can perform further calculations on AI outputs (for example, evaluating a strategy suggested by the AI model against historical data). Users can filter and customize results, visualize trends, or run "what-if" scenarios all within the same ecosystem, turning AI insights into actionable decisions.
5. **Persistence**: All data and results are persistently stored and managed. Every intermediate embedding, model artifact, and generated report can be kept in the system's datastores for auditability and reuse. The Pure Storage solution excels at this **persistent data layer**, ensuring that petabytes of vectors, model checkpoints, and historical records are stored with **consistent low-latency access and robust data protection** (including built-in backup and restore capabilities). This persistence also means the AI models can continuously learn from new data (since nothing is lost) and that outputs can be traced back for compliance and verification, a key concern in finance.

By covering *ingestion*, *processing*, *training*, *inference*, *analysis*, and *persistence*, the FSI GenAI Pod functions as a unified AI factory. It eliminates the typical silos and integration headaches of assembling an AI pipeline from scratch. Data flows seamlessly from raw source to refined insight within one platform.

Fast Deployment and Integration with Turnkey GenAI Pods

Knowing the value of GenAI is one thing—deploying it in a production environment is another challenge entirely. Financial institutions often struggle with the complexity and risk of integrating AI systems into legacy infrastructure. The FSI GenAI Pod is designed to remove these barriers with a **turnkey deployment model** and built-in best practices:

- **Rapid time to production**: This solution can **cut deployment time by 90% or more** compared to a do-it-yourself approach. All the components—data ingestion, KX software, Pure Storage, NVIDIA GPUs, NVIDIA AI Enterprise software, and NeMo models—come preintegrated and optimized for immediate use. Instead of spending months knitting together data pipelines, databases, and AI frameworks, financial firms can get a GenAI pilot up and running in weeks. This speed to deploy means faster time to value and less project risk.
- **Single pane of glass management**: The FSI GenAI Pod provides a unified user experience to monitor and manage the entire stack. From one dashboard, IT teams and data scientists can oversee the vector database (KDB.AI), the LLM and microservices, GPU utilization, and storage performance. This single pane of glass drives efficiency by giving end-to-end visibility "per token"—you can track how data flows to the model and back. It simplifies operations and troubleshooting, as there's no need to switch between separate tools for storage, database, and AI models.
- **Built-in RAG templates**: To further accelerate development, the platform includes **prebuilt workflow templates for common RAG use cases**. These are like blueprints for tasks such as research Q&A, document summarization, or trade idea generation, already wired with the retrieval and generation components. Teams can use these templates as starting points, customizing them to proprietary data or specific business logic. It's a "don't-reinvent-the-wheel" approach that enables institutions to leverage proven patterns and focus on the unique aspects of their use case.
- **Kubernetes-orchestrated microservices**: At its core, the FSI GenAI Pod runs on a cloud-native architecture with Kubernetes orchestration. All major functions (including the ingestors, vector database, LLM, and retriever/reranker services) are containerized as microservices. Kubernetes manages these, handling scaling, resilience (auto-restarting failed pods), and upgrades with minimal downtime. For the end user, this means the solution is **inherently scalable and portable**—it can run on premises, on bare metal, or in a private cloud, and it can flex from a small proof-of-concept cluster to a large multi-node deployment as demand grows. This containerized approach also supports **flexible form factors**. Financial firms can deploy the FSI GenAI Pod on existing hardware, as an integrated appliance, or even consume it via the Pure Storage **Evergreen//One**™ as-a-service model for AI infrastructure. In all cases, the heavy lifting of orchestration is handled for you.
- **Unified storage, compute, and network**: Having a performance-tuned, right-sized unified hardware stack of storage, compute, and network significantly boosts GPU utilization, increasing inference throughput in terms of tokens per second and reducing the latency and time to first token. This in turn improves user experiences and end customer productivity.

- **Seamless integration and expertise**: The solution is designed to slot into enterprise environments with minimal friction. It offers standard APIs and function integration points so that financial firms can connect their existing applications (for example, trading systems, customer relationship management systems, and portals) to the GenAI services. Moreover, Pure Storage and its partners, like World Wide Technology (WWT), provide **deployment and architecture expertise** out of the box. This means best practice guidance on networking, security, and data governance for the FSI GenAI Pod, as well as onsite assistance to ensure everything works in harmony with legacy systems. Such support greatly reduces integration risk and accelerates the journey from pilot to production.

Thus, the FSI GenAI Pod is enterprise-ready. It's not just a theoretical reference design, but a practical, **turnkey GenAI/RAG solution for financial services** that addresses real-world deployment challenges. With far less effort, IT leaders can stand up GenAI capabilities that would otherwise require extensive engineering—and they can do so knowing they're built on hardened, industry-leading components.

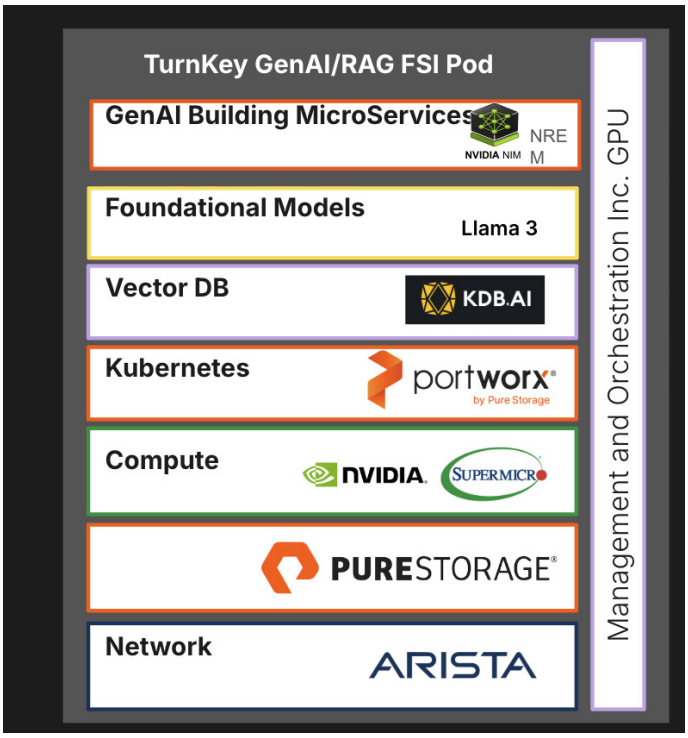

FIGURE 2 Turnkey GenAI/RAG solution for financial services

Backed by Industry Leaders: KX, NVIDIA, Arista Networks, SuperMicro, and Pure Storage

The FSI GenAI Pod is built on a foundation of trusted, enterprise-grade technologies from four industry leaders—KX, NVIDIA, Arista Networks, and Pure Storage—each bringing deep domain expertise and proven performance.

KX

Known for the kdb+ time series database, KX has long been the analytics engine behind trading floors and hedge funds. The company has **established itself as a leader in capital markets** and other data-intensive industries by managing mission-critical data at massive speed and scale. Unlike generic vector databases, KDB.AI is natively time-aware, enabling LLMs to reason over dynamic market data in context—a critical capability in financial services. KX systems enable customers to make decisions in a fraction of the time that competing technologies require. By extending this platform to support vector search and AI, KX brings its real-time pedigree to the world of GenAI. In other words, the same technology that excels at millisecond-level tick data analysis is now empowering your AI models with instantaneous data retrieval of structured and unstructured data in real time. This deep domain expertise in capital markets ensures the FSI GenAI Pod truly speaks the language of finance.

"In today's markets, every microsecond and every decision counts," says Ashok Reddy, CEO of KX. "By bringing together KX's real-time analytics, NVIDIA's AI leadership, and Pure Storage's data performance, the FSI GenAI Pod gives financial institutions the power to reason over their data—past and present—with unprecedented speed and accuracy. This isn't just about faster answers—it's about better answers, grounded in truth, delivered at the speed of the market."

NVIDIA

In the age of AI, financial services customers are adopting NVIDIA's accelerated platforms for extracting critical market intelligence and key insights from multi-modal data. NVIDIA's integrated hardware/software platform, powered by NVIDIA AI Enterprise software and CUDA-X, can operate on both time series market data and unstructured data, providing financial services customers the ability to develop a unique IP and generate ROI for their firm as well as their end customers.

"With this partnership with industry leaders such as KX, AI database leaders in time series AI, on the NVIDIA platform, we are excited to make those goals and objectives achievable for our end customers," says Prabhu Ramamoorthy, Global Partner Manager for NVIDIA Financial Services. "Unified compute, networking, and storage is key to enabling such quantitative AI platforms; so great to partner with Pure Storage to ensure accuracy, speed, throughput for mining key predictive intelligence, and data patterns that are essential for our end customers."

Arista Networks

Arista delivers the high-performance networking backbone required to move vast amounts of financial data at low latency. In GenAI environments, Arista's cloud-grade switches and extensible operating system (Arista EOS) ensure predictable, secure, and high-bandwidth connectivity between GPUs, storage, and compute layers. With proven deployments across global financial centers, Arista brings scalability and operational consistency to AI workloads. It plays a critical role in maximizing GPU utilization and minimizing data movement delays, which are essential in time-sensitive financial use cases.

SuperMicro

Supermicro (SMCI) has carved out a leading position in the GenAI server market by combining speed, flexibility, and efficiency. Its close partnerships with NVIDIA give it first-mover access to the latest GPUs, while its modular Building Block architecture enables highly customized systems tuned for everything from large-scale LLM training to edge AI inference.
A pioneer in direct liquid cooling, Supermicro delivers denser, more energy-efficient systems than many competitors, cutting costs while sustaining peak GPU performance. With massive production capacity, a $1B+ inventory buffer, and rapid time-to-market, SMCI consistently outpaces rivals, making it the go-to choice for enterprises racing to deploy AI at scale.

Pure Storage

Pure Storage is an **industry leader in data storage**, with the highest customer satisfaction (Net Promoter Score of 81 in 2024) of any IT infrastructure vendor. Its technology is trusted by nine of the top 10 global investment banks, seven of the top 10 asset managers, and five of the top 10 insurance firms—a testament to reliability and innovation. Pure Storage has also been a Gartner® Magic Quadrant™[1] leader in storage for 11 consecutive years, reflecting its completeness of vision and execution. For financial institutions, partnering with Pure Storage means their GenAI data pipeline is built on **battle-tested, enterprise-grade infrastructure** with world-class support. Moreover, the Pure Storage modern approach—Enterprise Data Cloud and the Evergreen® model of nondisruptive upgrades and as-a-service offerings—aligns well with the agility needs of AI initiatives.

Together, the collaboration of KX, NVIDIA, and Pure Storage delivers a solution greater than the sum of its parts. Institutions get the **real-time analytics power of KX**, the **data performance and efficiency of Pure Storage**, and the **AI prowess of NVIDIA**, all preintegrated with the specific demands of financial markets in mind. This combination of market leaders provides assurance to IT decision-makers and AI practitioners alike: the technology is *cutting-edge*, and the partners behind it are *experienced* and *dependable*.

Conclusion

Generative AI is ushering in a new era for financial services, from dramatically faster research and personalized client service to smarter trading and risk management. The five use cases outlined—and others on the horizon—illustrate the transformative potential. But realizing that potential requires an end-to-end solution that bridges the gap between high-level AI promise and day-to-day production reality. The FSI GenAI Pod is exactly that bridge. It offers a holistic, RAG-based AI platform that financial institutions can deploy with speed and confidence. By leveraging the strengths of KX, NVIDIA, and Pure Storage, this solution enables firms to harness the power of generative AI—with turnkey speed and ease of deployment—to gain the competitive insights and agility needed in the AI era of finance.

To experience the power of the FSI GenAI Pod through live demonstrations, book a hands-on session at the WWT ATC.

1 | GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally, and is used herein with permission. All rights reserved

CM2727-02-en-09/25