White Paper

Engineering the AI Trusted Stack for Financial Services

How Financial Institutions Can Move from Pilot Purgatory to Enterprise Production AI Deployment

NUTANIX

Table of Contents

Executive Summary

The financial services industry has moved beyond the "AI honeymoon". While 2023 and 2024 were defined by experimentation with Generative AI, 2025 established a call for measurable return on investment (ROI) and enterprise-scale execution. The strategic pivot has shifted from experimentation to industrialization, from chatbots to Agentic AI: autonomous reasoning engines capable of executing multi-step workflows from credit underwriting, compliance monitoring, to market analysis.

Yet while ambition has matured, most infrastructure has not.

Industry data suggests that up to 95% of AI pilots never reach production, a phenomenon increasingly termed "Pilot Purgatory." The root cause is not a shortage of innovative models or executive enthusiasm. Often the largest implementation challenges in financial services are around data governance, security, risk and lack of interoperability across the technology ecosystem and technical debt. As one major bank puts it: AI infrastructure is the hidden bottleneck and next battleground.

The Cloud Native Computing Foundation (CNCF) research reveals a profound gap between AI ambition and infrastructure reality and highlights that the real competitive advantage lies not in algorithms but rather in the unglamorous infrastructure capabilities.

The Banker notes that in short, banks need to get out of pilot purgatory and become AI-first organisations. The infrastructure challenge is that financial institutions are attempting to run next-generation AI workloads on infrastructure designed for an earlier era. Legacy technical debt, siloed data, security vulnerabilities, and emerging regulatory requirements can create compounding barriers that better models alone may not solve. AI infrastructure change, as Deloitte sees it, as a strategic differentiator.

To move forward, FinServ leaders are coalescing around a new industry standard: the "Trusted Stack". This foundational architecture incorporates security, governance, and transparency into every layer of AI infrastructure, from data ingestion to model output, to mitigate adoption from stalling against regulatory and internal resistance.

This report examines the infrastructure challenges financial services face when dealing with AI "Pilot Purgatory", the architectural aspects of a trusted stack, and outlines how the Nutanix Cloud Platform (NCP) and Nutanix Enterprise AI (NAI) solutions can deliver the capabilities needed for enterprise AI at scale. By providing a unified operating model across core data centers, edge environments, and public clouds, Nutanix provides institutions a way to address the fundamental gaps that hinder enterprise production deployment, supporting the transition from deliberation to deployment.

* Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris sit amet scelerisque velit. Donec quis arcu non tellus bibendum elementum et sit amet leo. Nam vulputate tempor hendrerit. Nam accumsan. sit amet leo.

The Situation: Financial Services AI at an Inflection Point

Market Adoption Reaches Critical Mass

The adoption of AI in financial services has reached a significant tipping point. By late 2025, market intelligence indicated that 85% of global banks have integrated some form of AI into their operations. Adoption is particularly concentrated among larger institutions; research suggests that 75% of institutions managing over $100 billion in assets have reported full integration of AI strategies as of late 2025. To support this momentum, industry AI spending has projections suggesting the financial sector AI infrastructure investment will reach $97 billion by 2027.

The Shift to Agentic AI

The nature of this adoption is pivoting from Generative AI (content creation) to a large focus on Agentic AI (autonomous action). While reports suggest that only approximately 10% of firms have scaled agents enterprise-wide, this represents the primary strategic focus for 2026.

These autonomous reasoning engines are designed to facilitate complex processes within banks, insurance and capital markets. Here are a few examples of the emerging agentic workflows:

- **Autonomous Relationship Manager Sales Agents:** McKinsey reports banks deploying agents that prioritize prospects, tailor outreach with context-rich messaging, and even negotiate within guardrails, all while continually learning from outcomes. These systems sense, decide, and act in real time, instead of waiting for humans to analyze dashboards or chase leads. The result: potential to return ten to 12 hours a week to each banker, improving the coverage ratio by about 40 percent.
- **Perpetual Know-Your-Customer (pKYC) Monitoring Agents:** Moody's reports that, unlike periodic manual reviews, pKYC agents continuously monitor customer risk profiles, triggering alerts when significant changes occur, such as a sudden spike in cross-border transactions or a change in beneficial ownership. The agent autonomously decides what constitutes a material change and initiates appropriate due diligence.
- **Agentic Security Operations Center (SOC) for Financial Institutions:** Morgan Stanley reports that companies are piloting agents that replicate specialized roles, such as threat hunters, detection engineers and credential managers. These agents simulate SOC workflows, streamline triage, investigation and remediation, operating at machine speed while maintaining human oversight for orchestration.

1 Thomson Reuters: Strategic Insights from Momentum AI Finance Official Post-Event Report | New York 2025

The shift to enterprise AI and autonomous action fundamentally alters infrastructure requirements. Each of these use cases shares common demands: continuous real-time operation, autonomous decision-making within defined guardrails, cross-system data access, and auditability for regulatory scrutiny. This level of operational intensity requires a foundation far more robust than infrastructure designed for batch processing or human-initiated queries.

The Bifurcation: Strategic Divergence in Financial Services

A distinct strategic divergence currently characterizes the financial services industry as institutions adopt different infrastructure models based on scale and resource allocation. Some analysts report that AI readiness will separate leaders from laggards.

For example, according to recent analyst reporting, three US Banks account for 75% of global banking AI patents. Adoption varies significantly across tiers, which may create a widening AI capability gap across the sector: Tier 1 banks (assets exceeding $100B) show roughly 75% to 80% AI integration; mid-tier institutions reach 50% to 60%; and regional banks currently sit at 30% to 40%.

Tier 1 Strategy: The "AI Factory" Approach

Leading global financial institutions have effectively positioned themselves as AI technology leaders and have publicly prioritized large-scale AI infrastructure investments. They are not merely deploying models; they are building proprietary "AI Factories" with infrastructure investments measured in billions.

- **One major Global Systemically Important Bank's (G-SIB)** large language model (LLM) suite reportedly processes meaningful work for over 200,000 employees daily, with a stated target of nearly $2 billion in annual value.
- **Another major US bank**, with nearly 1,000 AI-specific patents filed, aims to innovate in areas such as customer service, security protocols, and the provision of efficient and personalized banking services.
- **Another top 3 US bank** sees its chatbot handle approximately 2 million customer interactions daily, the equivalent of what 11,000 employees could do, on infrastructure the bank owns and operates.

The Tier 1 strategy is characterized by:

- **Internal Development:** Significant focus on proprietary platforms and models.
- **Platform Centralization:** Centralized AI environments designed to provide secure access to vetted models across the enterprise.
- **Intellectual Property Management:** Treating data and AI architecture as a potential competitive differentiator through patenting and trade secret protections.
- **Workforce Transformation:** Enterprise-wide reskilling programs and dedicated AI centers of excellence.

Tier 2 and Tier 3: Pragmatic Adoption and Integration

Mid-tier financial institutions (regional banks with assets roughly $10B–$100B+) are progressing beyond pilots toward practical AI deployments, though scaled enterprise adoption remains early and uneven. Yet they possess potential advantages: large enough to gain significant value from operational consolidation, while remaining more agile than Tier 1 peers in deploying modern AI infrastructure.

Without the deep pockets or specialized talent of global banks, mid-tier firms prioritize high-value use cases: Anti-Money Laundering/Know Your Customer (AML/KYC) automation, real-time risk scoring, document processing, compliance reporting, and increasingly bring development in-house or through partnerships rather than depending on external vendors.

This pragmatic approach requires careful architectural planning but presents a strategic opportunity. Mid-tier institutions can unify fragmentation left by years of M&A and accumulated technical debt.

Tier 3 institutions, (i.e. community banks, credit unions, and regional players typically under ~$10B in assets) have more slowly adopted AI, with industry estimates suggesting roughly 30–40% adoption rates. Partnership models dominate: for example, 67% of credit unions prefer fintech partnerships over internal development, the highest rate among all institution types. Implementations focus on back-office efficiency (document classification, loan processing) and targeted credit decisioning rather than enterprise-wide systems. Smaller and resource-constrained institutions may struggle to move beyond pilots, contending with data quality issues, governance gaps, and early deployment decisions that prioritized speed over architectural adaptability.

The strategic question for mid-tier and regional institutions is not whether to adopt AI, but how to build infrastructure that delivers Tier 1 capabilities without Tier 1 complexity and cost, precisely the challenge enterprise AI infrastructure is designed to address.

The Regulatory Vise Tightens

The Regulatory Landscape

As AI adoption accelerates, financial services institutions operate within an increasingly complex regulatory landscape that informs infrastructure requirements.

European Union: AI Governance Frameworks

The EU AI Act entered into force in August 2024, with various obligations scheduled to take effect through 2026 and 2027. For financial services, some key considerations include:

- **High-risk applications**: Organizations utilizing AI for credit decisioning or insurance underwriting may be subject to requirements for risk management systems, data governance, and technical documentation.
- **Digital Operational Resilience Act** (**DORA**): Effective January 2025, DORA introduced requirements for Information and Communications Technology (ICT) risk management and third-party oversight, that encompass AI systems and deals with concentration, driving interest in hybrid cloud strategies to address cloud concentration risk.
- **European Insurance and Occupational Pensions Authority** (**EIOPA**) **guidance**: Recent opinions (August 2025)provide additional guidance on supervisory expectations for AI governance under existing insurance-specific frameworks.

United States: Regulatory Guidance & State Legislation

A patchwork of state and agency requirements have emerged, as federal frameworks continue to evolve:

- **New York State Department of Financial Services** (**NYDFS) Industry Letter (October 2024)**: Outlines how existing cybersecurity regulations (23 NYCRR Part 500) apply to AI risks, emphasizing the role of risk assessments and access controls within institution-owned compliance obligations.
- **Model Risk Management (Supervisory Letter SR 11-7)**: Federal Reserve guidance regarding model validation is increasingly referenced in the context of LLMs and generative AI tools.
- **State AI Acts:** Legislation such as the Colorado AI Act (effective June 2026) introduces additional compliance complexities for institutions.

United Kingdom: Principles-Based Oversight

The United Kingdom (UK) maintains a pro-innovation, principles-based approach. The UK Financial Conduct Authority (FCA) AI Sprint (January 2025) and the Critical Third Parties regime (effective January 2025) establish expectations for AI governance and oversight of third parties. Similar to DORA, interest in hybrid architectures is seen as one path to manage provider dependency.

control.While Nutanix solutions provide the infrastructure to possibly support various compliance requirements, the customer remains solely responsible for ensuring their specific implementation meets all applicable legal and regulatory obligations.

Infrastructure Implications for Compliance Support

Evolving regulatory standards highlight the value of infrastructure capabilities that support flexible and resilient operations.

- **Data Sovereignty:** In light of frameworks like DORA and the General Data Protection Regulation (GDPR), organizations often seek localized, regionally-isolated or sovereign data storage solutions.
- **Auditability & Lineage:** Infrastructure should integrate governance into the system architecture so that transparency, traceability, and control are integral to the design.
- **Explainability Support:** Systems should be designed to help demonstrate how AI-driven decisions were reached, particularly in critical financial processes such as credit decisioning and underwriting.
- **Identity Controls:** The management of AI agents necessitates identity protocols to help mitigate unauthorized data access.
- **Resilience:** To support operational continuity, hybrid architectures, such as those leveraging hyperconverged infrastructure, can be used to mitigate single points of failure and support tested failover procedures.

These regulatory requirements do not exist in isolation, they can compound the infrastructure challenges financial institutions already face. Data mandates may present integration challenges for certain legacy and cloud environments when attempting to implement policy-driven data placement. Furthermore, rigorous auditability requirements can highlight the limitations of governance models that were not integrated by design. As organizations adopt autonomous agents, traditional identity controls may face hurdles. In this landscape, the regulatory environment emphasizes the importance of infrastructure architectures designed to support AI workloads at enterprise scale.

1 Thomson Reuters: Strategic Insights from Momentum AI Finance Official Post-Event Report | New York 2025

The Trusted Stack

As one Tier 1 bank CIO observed, institutions cannot simply layer AI onto inefficient processes and expect transformative results. Financial institutions are attempting to run next-generation AI workloads on systems designed for an earlier era.

Infrastructure challenges in financial services often stem from the interaction between legacy systems and the compute demands of modern AI. The recent Nutanix State of Enterprise AI report, indicates that 90% of enterprises view AI as a priority, and 91% of surveyed IT leaders suggest current infrastructure requires modernization.

Infrastructure Challenge	The Core Problem	Why It Impacts Pilots to Production
1. Legacy Technical Debt and System Incompatibility	Many banks still rely on 1980s code (COBOL/Perl) with slow, 18-month update cycles.	1/3 of legacy systems cannot support AI. Old hardware often crashes under high-compute AI demands and many environments will hit breaking points as AI workloads increase
2. The "Pilot Purgatory" Scaling Gap	Teams build fragmented "point solutions" that don't talk to each other	A staggering 98% of firms struggle to move GenAI from the lab to production due to disjointed architecture
3. Data Gravity and Silo Entrenchment	Essential data can often be siloed in systems and environments. As AI systems grow, they create "data gravity," making it increasingly difficult and expensive to move data between the edge, on-premise, and cloud.	Moving massive datasets can get expensive. Initial AI project costs are expected to increase by more than 30% solely due to the infrastructure lift required for data preparation (cleaning, organizing, and sanitizing).
4. Acute Identity and Security Vulnerabilities	The shift toward Agentic AI requires a new type of security; machine identities already outnumber human ones in the financial sector by a ratio of 80:1.	Approximately 68% of organizations lack the necessary identity controls for AI/LLMs slowing pilots from reaching safety clearance
5. Regulatory Sovereignty and the Cloud Impasse	Increasing "data sovereignty" rules (like DORA/EU AI Act) are impacting where data must reside	Without "air-gapped" or sovereign infrastructure, up to 30% of projects are abandoned after the pilot stage due to inadequate risk controls.

The Emergence of the Trusted Stack

The challenges outlined above share a common thread: they cannot be solved by better models or organization improvements alone. They require a foundational rethinking of how AI infrastructure is architected. Industry leaders at the 2025 Momentum AI Finance conference framed trust as the "foundational currency" of AI adoption; without it, initiatives stall against resistance from internal users, customers, and regulators alike. integral to the design.

The concept of a "trusted stack" for AI infrastructure emerged from industry dialogue at events like the 2025 Momentum AI Finance conference in New York, where practitioners, analysts, and technologists convened to define what enterprise-ready AI requires. A subsequent Reuters analysis distilled these conversations into a four-layer framework that serves as a conceptual framework. Trust, of course, is not new to financial services, it is the industry's foundational currency. But what constitutes trust in the context of autonomous AI systems is being defined in real time. The framework below reflects this emerging consensus.

Layer	Function	Requirement
Data	Ingestion, curation, and quality management	Provenance, lineage, and access controls
Models	Algorithms powering inference and generation	Auditability, bias monitoring, version control
Applications	End-user tools and autonomous agents	Identity management, permissioning, guardrails
Observability	Monitoring and auditing of AI decisions	Transparency, explainability, regulatory reporting

Each layer plays a role in the overall strength of the system. A model built on poorly governed data inherits its flaws. An application lacking identity controls cannot scale past pilot. Observability bolted on after deployment is unlikely to satisfy regulators demanding auditability by design.

For financial services CIOs, the strategic question has shifted. It is no longer "Should we adopt AI?" It is:

> *How do we control and scale infrastructure and operations to meet GenAI demand, securely, adjacent to our data, without being crushed by costs or regulatory penalties?*

The Infrastructure Imperative

The Reuters report's central message is blunt: AI success depends on prioritizing work that technology leaders often dismiss as unglamorous. Data governance, legacy modernization, and integrated observability are not strategic afterthoughts, they are prerequisites. Organizations treating infrastructure as a downstream concern are, in the report's framing, "building on sand." As McKinsey has noted, capturing AI value requires organizations to "rewire" their operations, a transformation that begins with infrastructure, not just models.

Three dynamics appear to be shaping how leading institutions are responding:

The real competitive moat isn't the model, it's the data beneath it. As large language models commoditize, proprietary datasets become the defensible advantage. But realizing that advantage requires what the report calls "AI-ready

content": curated, cleansed, and linked repositories accessible for model consumption. This is painstaking, unglamorous work, and it's where competitive differentiation lives.

Legacy systems remain the greatest barrier. Platforms designed for quarterly release cycles cannot support the rapid feedback loops AI demands. Worse, they trap critical data in siloed formats that reduce the enterprise-wide integration boards now expect. Modernization lacks glamour, but it's non-negotiable.

The architecture is shifting toward platforms plus customization. Hybrid infrastructure that brings AI compute to data rather than forcing data migration to edge or clouds. Rather than building from scratch, firms are leveraging foundational capabilities from major providers while focusing internal resources on domain-specific applications and agents. Technology leaders are also adopting multi-model strategies rather than committing to single solutions, preserving flexibility as the landscape evolves.

The through-line is consistent: governance, security, and observability must be designed in from the start, not bolted on later. The organizations that escape "Pilot Purgatory" will be those willing to do the foundational work first, translating the abstract layers of the 'Trusted Stack' into a concrete operational reality

The Path Forward: Engineering the Trusted Stack

The 'Trusted Stack' provides the conceptual framework required for regulated innovation. However, a framework alone does not move pilots to production. Financial institutions seeking an AI architecture designed to support trust at every layer while addressing the typical blockers, legacy incompatibility, data gravity, and security control, that can often stall initiatives.

To bridge this gap, leaders are increasingly adopting hybrid AI architectures that bring compute to the data rather than moving sensitive data to the cloud. This approach translates the Trusted Stack into four operational capabilities that can be facilitated through the Nutanix Cloud Platform solution.

The Trusted Stack Framework

Trusted Stack Layer	Operational Capability	Key Financial Services AI architecture Considerations
Data	Control & Observability	Full lineage and policy-based placement; visibility to mitigate leakage to public models.
Models	Governance & Validation	Immutable versioning; audit trails for every inference; "pre-flight" testing for accuracy.
Applications	Security & Resilience	Hardened infrastructure; "Dark Site" capabilities; operational continuity across edge-core-cloud.
Observability	Cost & Scale Predictability	Right-sized resources (CPU/GPU); transparent TCO; linear scaling without "cloud bill shock."

Realizing the Stack: The Nutanix Architecture

The architecture spans three operational domains: core datacenters for training and heavy inference workloads, edge locations (branches, trading floors, operations centers) for latency-sensitive and sovereignty-constrained workloads, and public cloud for burst capacity, experimentation, and globally distributed teams. Each domain requires consistent governance, security, and observability, a unified operating model rather than three separate environments.

The Nutanix stack delivers this consistency, helping institutions place workloads where they make sense without sacrificing control and while supporting customers' compliance efforts.

1. Data Control & Observability (Nutanix Unified Storage & Data Lens)
Nutanix provides a unified data service that allows AI workloads to access data in native formats (file, block, object) without costly transformations.

- Proactive Governance: Nutanix Data Lens provides the essential "look forward" capability, offering global visibility and auditing to help proprietary financial data from being accessed by unauthorized models.
- Sovereignty-First Replication: Data can be restricted to specific jurisdictions or on-premises GPU clusters through policy-based controls, supporting customers' compliance with local residency laws.
- Encryption Everywhere: Integrated encryption at rest and in transit maintains a high security posture even when compute occurs in a hybrid cloud environment.

2. Governance & Validation (Nutanix Enterprise AI)
Nutanix Enterprise AI (NAI) serves as the centralized control plane, designed to support customers' auditability requirements.

- Enterprise Model Repository: NAI maintains an immutable registry of validated models (from Hugging Face or NVIDIA NIM). This facilitates the use of only authorized, "clean" models used in production.
- Pre-flight Testing: Before a model is deployed to a branch or trading floor, NAI allows teams to test model accessibility and response quality via a natural language interface, reducing the risk of "hallucinations" in critical workflows.
- Standardization: By providing secure API endpoints (compatible with OpenAI standards), NAI allows institutions to swap models or update versions without rewriting application code.

3. Security & Resilience (Nutanix Cloud Infrastructure & NKP)
Nutanix Cloud Infrastructure (NCI) and Nutanix Kubernetes Platform (NKP) provide the hardened, self-healing foundation required for mission-critical AI applications like fraud detection.

- Untethered Edge Operations: NCI supports "Dark Site" or air-gapped deployments. A trading floor can continue running fraud-detection agents even if the connection to the central cloud or data center is severed.
- Zero-Trust Microsegmentation: Using Nutanix Flow, institutions can isolate AI workloads at the network level, facilitating that an AI agent can only communicate with the specific datasets and APIs it is authorized to see.
- Identity-Centric Access: Native RBAC integration provides capabilities so only authorized data scientists and developers can modify model parameters or access inference logs.

4. Cost & Performance Predictability (Nutanix GPT-in-a-Box)

To avoid the "Pilot Purgatory" caused by infrastructure complexity, Nutanix offers GPT-in-a-Box, a turnkey, validated full-stack solution.

- Right-Sized Compute Economics: On-premise inferencing for cost-efficient per-query operations at scale. Though acquiring GPUs and provisioning the power and cooling to support this may represent significant capital and operational expense. Nutanix addresses this through a tiered approach: Small Language Model(s) (SLMs) and lighter inference workloads can run on standard CPUs with acceleration, avoiding GPU dependency only; heavier training and peak-demand inference can burst to cloud, leveraging elastic capacity without permanent infrastructure commitment. The result is a hybrid model that optimizes TCO across the full AI lifecycle, not just at deployment, but as workloads evolve.
- GPU-Optional Deployment: While heavy training requires GPUs, Nutanix allows institutions to run "right-sized" inference models on standard CPUs with acceleration, promoting cost efficiency.
- Linear Scaling: The Nutanix web-scale architecture allows banks to start small and scale AI capacity predictably, node-by-node, which can reduce the upfront CapEx or unpredictable OpEx of public cloud-only strategies.

The Reference Architecture

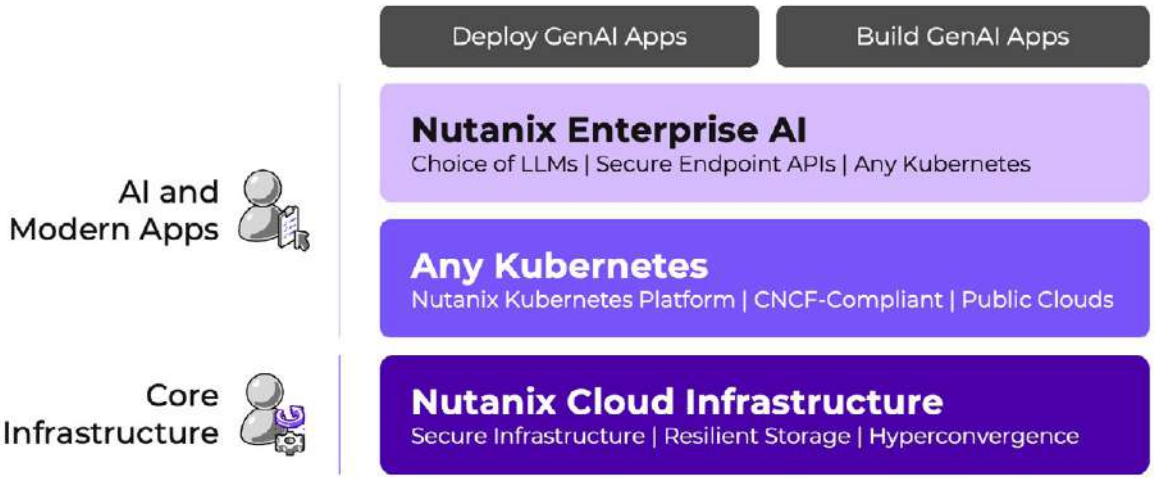
The Trusted Stack is realized through four integrated layers that move in lockstep:

1. Nutanix Cloud Infrastructure (NCI): The resilient, agile hardware foundation from edge to cloud.
2. Nutanix Unified Storage (NUS): Data storage services with integrated governance via Data Lens.
3. Nutanix Kubernetes Platform (NKP): Orchestration for containerized AI agents and microservices.
4. Nutanix Enterprise AI (NAI): The control plane for model management, security, and API delivery.

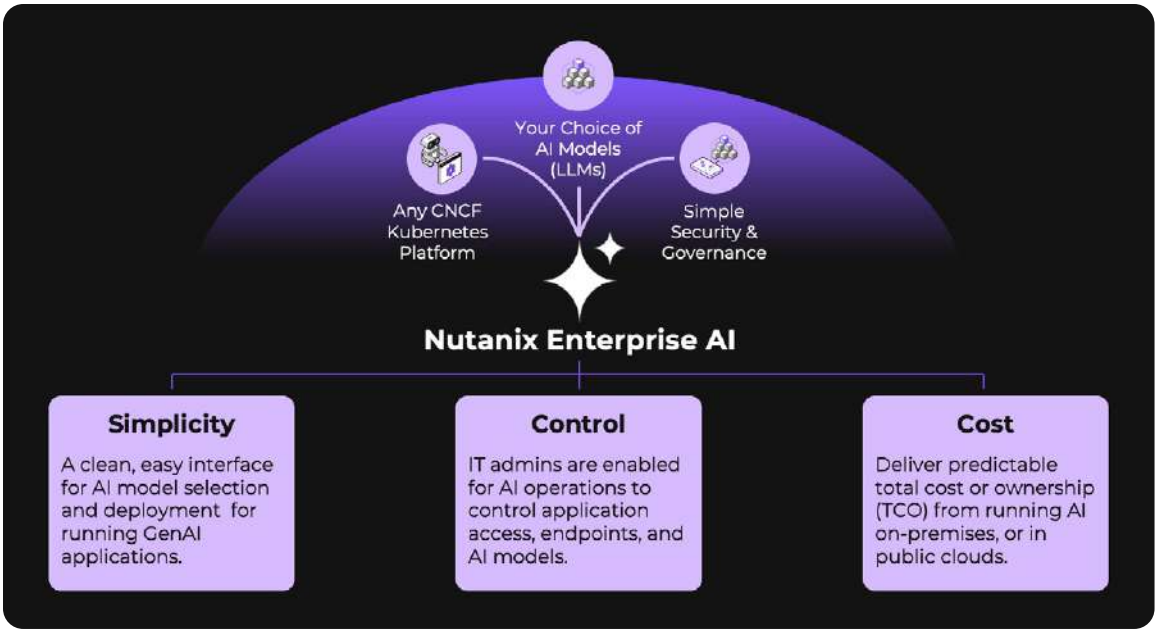
To orchestrate it all, Nutanix's platform facilitates lifecycle management at scale. As AI deployments proliferate across edge, core, and cloud, operational complexity compounds quickly. Firmware updates, Kubernetes upgrades, model version management, and security patches must be coordinated across potentially hundreds of nodes without disrupting production inference.

Nutanix addresses this through its integrated Lifecycle Management (LCM) full-stack update manager that automates updates across the entire footprint from a single control plane. Combined with NKP's container

orchestration and NAI's model registry, institutions can manage AI workloads as a unified fleet rather than a collection of point deployments, supporting Day 2 operations scaling.



This is what a Trusted Stack can look like in production: purpose-built infrastructure that addresses the pilot-to-production gaps by incorporating trust, security, and predictability into every layer of the AI lifecycle.



Conclusion: The Narrowing Window

The institutions that will lead financial services in 2030 are not those with just the most sophisticated models, they are those building the foundational infrastructure that allows AI to scale from pilots to production.

The challenge is architectural: building systems that maintain data sovereignty while leveraging distributed compute, enforcing governance without bureaucratic paralysis to assist in demonstrating regulatory expectations without sacrificing innovation velocity, and delivering predictable economics within performance objectives.

The competitive stakes are substantial. Three U.S. banks already account for 75% of global banking AI patents. Tier 1 institutions report billions in realized value while processing AI workloads for hundreds of thousands of employees daily. Meanwhile, 70% of institutions struggle to realize tangible ROI from AI investments, and 98% face challenges scaling from development to production. The gap between leaders and laggards is widening, and infrastructure is one of the dividing lines.

Nutanix Enterprise AI infrastructure can help customers implement elements of the Trusted Stack with the hybrid flexibility financial services are seeking, the resilience and governance increasingly sought, and economics that institutions can sustain.

As the Chief Data and Analytics offices at a leading global bank advised: "Stop deliberating and move." The question is not whether to act, but whether your infrastructure will be ready when you do.

Where to Go Deeper

This report provides a strategic framework for evaluating AI infrastructure readiness. For institutions ready to move to implementation, Nutanix provides detailed technical resources including reference architectures, sizing guides, and deployment patterns for financial services environments. Contact your Nutanix representative, test drive it for yourself or visit our Enterprise AI page to access more materials.

NUTANIX

info@nutanix.com | www.nutanix.com | @nutanix

PSM-CN-Engineering-the-AI-Trusted-Stack-FinServ-WhitePaper-FY26Q3-V1-040626