cognizant® | NUTANIX

A cloud-native AI fabric for the distributed enterprise

Enable scalable, governed AI across hybrid cloud and edge with Cognizant® Distributed AI Services, powered by Nutanix

Enterprise AI is entering a new phase. While early initiatives were concentrated in centralized data centers and public cloud environments, achieving many of the highest-value AI use cases will depend on moving AI closer to where data is created and consumed. Yet as edge AI applications mature, they introduce a new set of operational challenges. Models, applications and infrastructure must be deployed and managed across dozens, hundreds, or even thousands of locations. IT teams must balance latency, bandwidth, cost, security, compliance and data sovereignty requirements while maintaining consistent governance across highly distributed environments.

In this white paper, we explore how Cognizant and Nutanix help organizations:

- Deploy AI anywhere: Establish a consistent runtime and operating model across edge, datacenter and cloud environments while placing workloads where they perform best.
- Connect AI Services Everywhere: Nutanix AI Gateway provides secure access to private, edge and public AI endpoints.
- Govern at scale: Maintain centralized control over security, compliance, cost and AI lifecycle management across distributed environments.
- Secure the full stack: Apply consistent security policies and controls across infrastructure, data, applications and AI services.
- Accelerate outcomes: Reduce operational complexity while enabling scalable, resilient and production-ready AI deployments.

The AI imperative

In a prototypical centralized AI model, data is assumed to be collected from across the organization, moved to a central location, processed using a data center or public cloud and then delivered back to users and applications. For many emerging AI use cases, that assumption no longer holds. Organizations are increasingly deploying AI in environments where decisions must be made close to the point of data generation. As reported by Forbes, Gartner estimates that up to 75% of all enterprise data is processed outside the traditional data center.

This shift is occurring due to specific operational constraints of AI strategies that are dependent on a traditional, centralized model.

- **Latency and responsiveness.** Many AI use cases derive value from real-time decision-making. Robotics, industrial automation, physical security, autonomous systems and operational monitoring often require responses measured in milliseconds. Transmitting data to a centralized environment for processing can introduce delays that limit effectiveness, while edge computing can reduce latency by up to 90%.
- **Data gravity and sovereignty.** Moving large volumes of sensor, video, operational, or transactional data to centralized environments can introduce significant cost, bandwidth, compliance and data residency challenges. In many cases, processing data locally is both more practical and more economical.
- **Governance across distributed environments.** As organizations deploy AI across hundreds or thousands of locations, maintaining consistent security controls, policies, lifecycle management and compliance standards becomes increasingly difficult when restricted to tools designed for a centralized environment.
- **Operational scale.** As AI scales across locations, IT is faced with managing fleets of applications, models, infrastructure resources and data services across a distributed estate. The complexity of deployment, upgrades, monitoring, security and lifecycle management grows rapidly as AI expands beyond isolated pilots.

Common edge AI environments

In practice, these challenges are particularly visible in industries with highly distributed operations and/or air-gapped security considerations, including:

Manufacturing

Real-time quality inspection, predictive maintenance, worker safety and autonomous operations.

Retail

Loss prevention, inventory optimization, customer analytics and intelligent store operations.

Healthcare

Clinical decision support, medical imaging, patient monitoring and data sovereignty requirements.

Financial services

Fraud detection, risk analysis, branch operations and customer engagement.

Government and defense

Mission systems, remote operations, tactical environments and disconnected or air-gapped deployments.

A unified, cloud native fabric for enterprise AI

Cognizant® Distributed AI Services, powered by Nutanix, provide a consistent foundation for deploying, managing, securing and scaling AI workloads across edge, datacenter and public cloud environments.

Built on Nutanix's complete, open cloud native platform, the fabric brings together Kubernetes orchestration, unified management, enterprise-grade data services and security with Cognizant offerings that help compress delivery time and reduce operational risk by directly addressing the core challenges that come with a centralized AI Stack.

Cost optimization

Local edge processing minimizes expensive data movement, while shared infrastructure maximizes GPU utilization and prevents siloed hardware sprawl.

Speed to innovation

Standardized, self-service environments give developers a well-defined path for moving refined models from central training to global edge deployment in hours, not weeks.

Resilience & business continuity

Edge autonomy keeps critical operations running through network outages, preventing costly operational downtime and limiting the risk of cascading failures.

Trust and security

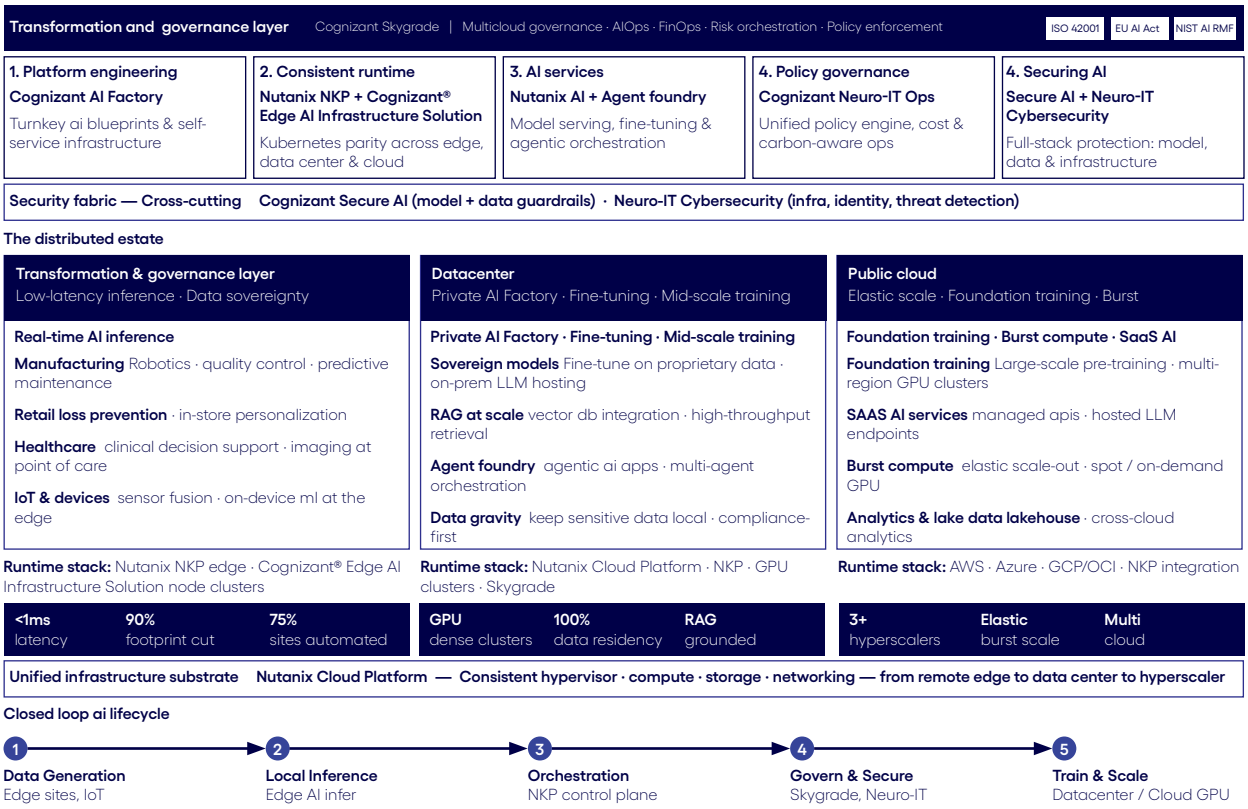
Cognizant Secure AI and Neuro® Cybersecurity deliver continuous, full-stack protection for models, data and infrastructure.

Governance and compliance

Cognizant Skygrade enforces policy and orchestrates risk across the estate, aligning with emerging standards like ISO 42001 and the EU AI Act, making distributed AI auditable by design.

The unified cloud-native AI fabric

A single operating model spanning Edge · Data center · Public cloud — built on five pillars



Five pillars of an effective AI estate

Cognizant® Distributed AI Services, powered by Nutanix, act as a consistent operating layer, unifying build, deployment, security and governance across the complete edge-data center-public cloud ecosystem.

1. Platform engineering for AI

Cognizant AI Factory provides standardized 'Infra for AI' blueprints that form the backbone of AI operations, simplifying deployment, scaling, and management across the lifecycle. This standardization directs a faster path from blueprint to self-service infrastructure with less operational overhead.

2. Consistent runtime across edge–cloud

Workload portability depends on environmental parity across infrastructures. Nutanix Kubernetes Platform (NKP) provides a standardized Kubernetes foundation that accelerates containerization, while Cognizant® Edge AI Infrastructure Solution (a unified edge management control plane) lets enterprises push, version and govern AI models and agents across the entire fleet, eliminating site-by-site manual deployment.

3. API-driven AI services with Nutanix AI

Nutanix Enterprise AI (NAI) hosts and fine-tunes models, exposing them through the AI Gateway as robust, scalable APIs surface. The Gateway's unified endpoint normalizes access across private, edge-hosted and public cloud models, enabling organizations to optimize inference placement without fragmenting governance. Cognizant Agent Foundry adds an enterprise framework to design, build, orchestrate and scale agentic AI systems across business processes, consuming model endpoints and enterprise tools through the Gateway's governed MCP interface

4. Policy-driven governance

At scale, governance has to be enforced from a single interface. Cognizant Neuro® IT Ops unifies AIOps, SecOps and FinOps across cloud, datacenter and edge, delivering proactive reliability engineering, automated incident resolution and unified observability. These advantages translate directly to improved uptime, faster incident response and less need for manual intervention across the distributed estate. The Nutanix AI Gateway feeds token consumption, endpoint utilization and tool-call telemetry directly into this governance plane, completing the operational visibility picture.

5. Securing AI

As AI expands from pilots to distributed execution across edge, datacenter, and public cloud, the attack surface grows across models, data, inference endpoints, agentic workflows, MCP-connected tools, and infrastructure. Securing AI requires a full-stack approach that protects models, data, and infrastructure consistently across every environment.

Cognizant Secure AI protects the model and data layer with controls for prompt injection, model theft, sensitive data exposure, and responsible AI guardrails. Nutanix AI Gateway secures inference endpoints through unified authentication, RBAC, tool-level filtering, MCP governance, and full audit trails. The Gateway enforces zero-trust access at the model and tool call layer, ensuring no agent or application consumes resources beyond its authorized scope. Cognizant Neuro® Cybersecurity secures the runtime and infrastructure with threat detection, zero-trust enforcement, identity governance, and automated response across edge, datacenter, and cloud. Policies are defined once and enforced across model gateways, Kubernetes runtimes, networks, and edge nodes.

A single control pane for AI

Siloed management tools, security policies, deployment processes, and governance frameworks quickly create operational complexity that risks sprawling alongside the AI environment itself. Cognizant® Distributed AI Services address this challenge through a unified control plane that spans the full AI stack. Governance and security operate as cross-cutting layers across infrastructure, cloud-native platforms, and AI services, providing centralized visibility and policy enforcement regardless of where workloads are deployed.

This approach allows organizations to distribute execution while maintaining control. AI services can run where they deliver the greatest business value, while governance, security, operational policies, and lifecycle management remain consistent across the environment.

1. Transformation and governance layer

Cognizant Skygrade

Centralized Multicloud & Edge Governance
Integrated AIOps, SecOps, FinOps Orchestration
Unified policy engine across clouds & edges

Neuro® IT Ops

Proactive incident resolution & failure prevention
AI-driven root cause analysis & remediation

2. Securing AI layer

Cognizant Secure AI

Model and data guardrails - prompt injection, model theft, PII leakage
Responsible AI controls and trust boundaries

Cognizant Neuro® Cybersecurity

Zero-trust enforcement & identity governance
Threat detection, response & runtime security

How AI works from edge to cloud: A closed-loop AI lifecycle

Step 1: Data generation at the edge

Data originates where the physical world meets the digital: IoT sensors on factory floors, real-time interactions in retail stores, and enterprise systems.

Step 2: Local AI inference (Edge)

Models deployed on Edge AI Nodes enable low-latency decisions (stopping a robotic arm in milliseconds), cut bandwidth use, and preserve data sovereignty by processing sensitive information locally.

Step 3: Cloud-native orchestration

A unified Kubernetes-based control plane schedules workloads across the estate. AI services run as portable microservices: the same containerized logic is used on an edge node or a cloud cluster. The AI Gateway routes inference requests dynamically based on availability, data classification policy, and cost rules, ensuring that the right model is always in the right place at the right time.

Step 4: Central governance and security

Cognizant Skygrade enforces security policies, tracks costs, and ensures compliance with regional regulations. The Nutanix AI Gateway surfaces token consumption, endpoint utilization, and MCP tool-call telemetry for FinOps and compliance reporting. In parallel, Cognizant Secure AI and Neuro® Cybersecurity monitor models, data flows, and runtime behavior, detecting adversarial activity, enforcing zero-trust, and containing incidents before they cascade.

Step 5: Cloud for training and scale

The Public Cloud handles what it does best: scale. Aggregated, anonymized data flows back for heavy model training and burst compute; refined models flow back to the edge (Step 2), closing the loop and keeping enterprise intelligence in a cycle of continuous improvement.

Conclusion: A practical framework for the Distributed Intelligence Frontier

Cognizant® Distributed AI Services, powered by Nutanix, provide a defined path forward to powerful, streamlined, and economical edge AI.

- Standardize on cloud-native platforms for workload portability.
- Decentralize execution by bringing AI directly to the data.
- Secure the full stack (model, data, and infrastructure) across every environment.
- Centralize governance to maintain control over cost, risk, and security.

For organizations evaluating their next phase of AI adoption, a phased evaluation and implementation program can help reduce implementation risks while planning for an AI environment tuned to your organization's specific challenges.

1. **Assess & identify value + deployment strategy:** Evaluate the application estate, data landscape, and infrastructure footprint to identify high-value AI use cases and determine where each workload should run based on latency, data sensitivity, security, and cost. Map AI Gateway endpoint topology.
2. **Pilot a private AI factory:** Establish a controlled AI environment with Nutanix GPT-in-a-Box to validate data readiness, model performance, RBAC Configuration governance needs, and operational feasibility before scaling.
3. **Extend to the Edge:** Move AI capabilities closer to the data, systems, and processes they support. Register edge-hosted inference endpoints with the AI Gateway. Enable autonomous agents to operate in distributed environments where real-time context and local execution are critical.
4. **Scale with governance, security & optimization:** Expand from pilot to production with unified management, embedded AI security, policy enforcement, and continuous optimization. Active full MCP governance, tool-level RBAC, and compliance reporting across hybrid multicloud and edge environments.

Ready to scale enterprise intelligence? Get in touch with us at HybridCloudOfferings@cognizant.com

cognizant®

Cognizant (Nasdaq: CTSH) is an AI Builder and technology services provider, bridging the gap between AI investment and enterprise value by building full-stack AI solutions for our clients. Our deep industry, process and engineering expertise enables us to build an organization's unique context into technology systems that amplify human potential, drive tangible outcomes and keep global enterprises ahead in a fast-changing world. See how at cognizant.ai or @cognizant.

World Headquarters
300 Frank W. Burr Blvd.
Suite 36, 6th Floor
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233

European Headquarters
280 Bishopsgate
London
EC2M 4AG
England
Tel: +44 (01) 020 7297 7600

India Corporate Office
Siruseri-Software Technology Park of India (STPI)
SDB Block—Ground Floor North Wing
Plot No H4, SIPCOT IT Park
Chengalpattu District
Chennai 603103, Tamil Nadu
Tel: 1800 208 6999

APAC Headquarters
1 Fusionopolis Link,
Level 5 NEXUS@One-North,
North Tower Singapore 138542
Phone: + 65 6812 4000