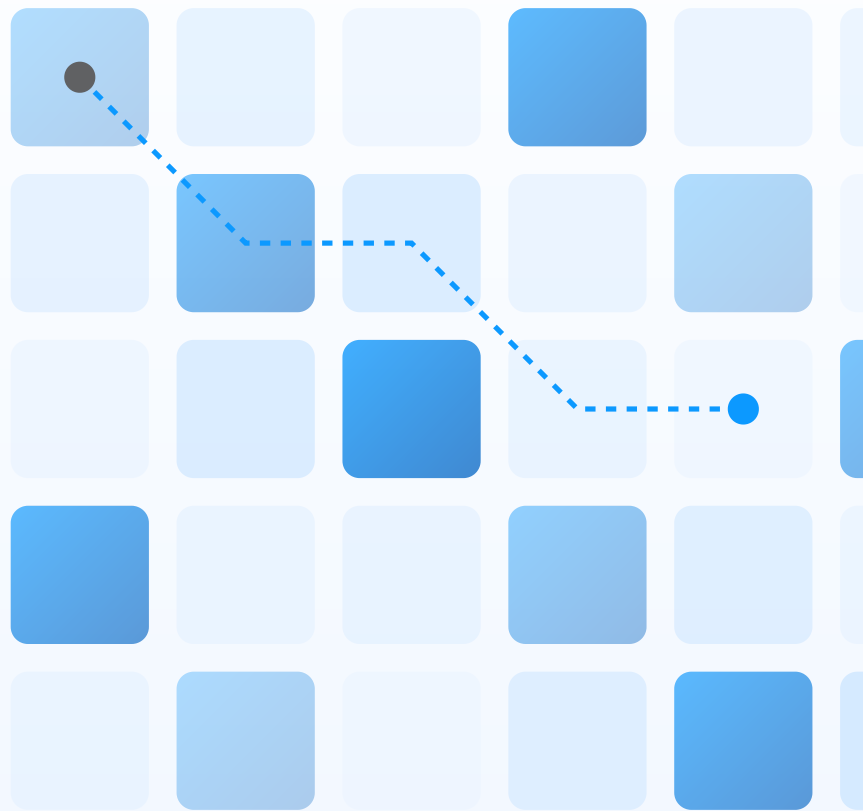




eBOOK • FOR CIOs AND CTOs

From Pilot to Production: The CIO's Guide to Building a Private AI Factory

How enterprises turn proprietary data into production-grade AI on dedicated, GPU-native private cloud infrastructure, without sending sensitive data to a public AI service.

What's inside

01	The AI Mandate Has Outgrown Its Infrastructure	03
02	Why Public AI Services Stall at Enterprise Scale	04
03	What Is an AI Factory?	05
04	Inside UPC AI Factory: The Architecture	06
05	The End-to-End AI Lifecycle, in One Environment	07
06	Five Outcomes That Matter to the Board	08
07	Choosing Your AI Infrastructure Model	09
08	AI Factory in Action: Industry Scenarios	10
09	A 90-Day Path to Your First Production Model	11
10	The CIO's Evaluation Checklist	12

A note to the reader

Most enterprises have already proven that AI works. Proofs of concept have shown value in fraud detection, clinical documentation, customer service and code generation. The harder question now sits with IT leadership: **where will these models run, who controls the data they learn from, and what will they cost at scale?**

This eBook is written for CIOs and CTOs making that decision. It explains the AI Factory model, shows how UPC AI Factory implements it on United Private Cloud®, and closes with a practical roadmap and an evaluation checklist you can use with any provider.

01

The AI mandate has outgrown its infrastructure

AI has moved from the innovation lab to the operating plan. Boards now expect AI to deliver measurable gains in efficiency, speed of research and quality of decisions, and they expect IT to deliver it safely.

Across financial services, healthcare, government and technology, enterprises are racing to operationalize AI and turn their captive global capability centers into innovation engines. Yet the infrastructure required to run AI at scale has not kept pace. The result is a widening gap between what the business wants from AI and what IT can responsibly put into production.



Models are getting bigger

As models grow toward trillions of parameters, general-purpose cloud compute is no longer enough. Training and fine-tuning demand dedicated, AI-optimized infrastructure.



Data is getting more sensitive

The most valuable AI is trained on the data you can least afford to expose: customer records, patient histories, source code and financial transactions.



Costs are getting harder to predict

Metered GPU hours, egress fees and per-token inference charges make AI budgets volatile precisely when programs need to scale.

“Standard” compute will no longer satisfy “AI-optimized” needs. The enterprises that win will be those that own a repeatable, governed way to turn data into models.

The shift CIOs are making

From	To
Isolated AI pilots on rented GPUs	A governed, repeatable AI production capability
Sending data to public AI services	Bringing AI to the data, inside a private environment
Stitching together point tools	One lifecycle toolchain from ingestion to inference
Unpredictable consumption bills	Dedicated capacity with predictable economics

02

Why public AI services stall at enterprise scale

Public clouds offer general-purpose scalability, but they were never truly optimized for the brute-force demands of large-scale AI training, or for the governance demands of regulated enterprises.

Public AI services are an excellent place to experiment. The problems appear when a pilot becomes a production system that touches sensitive data, runs continuously and must answer to auditors and CFOs. Six friction points show up again and again.

01

Data leakage risk

Prompts, fine-tuning datasets and outputs leave your control boundary. For regulated industries, that alone can end a project in legal review.

02

GPU resource contention

Shared multi-tenant environments mean noisy neighbors, queueing for capacity and unpredictable training times.

03

Unpredictable costs

Metered GPU hours, storage, egress and per-token pricing compound as usage grows, making AI hard to budget.

04

Compliance complexity

Proving where data resides, who accessed it and how models were trained is difficult on infrastructure you don't control.

05

Integration sprawl

Data engineering, model development, MLOps and inference often live in separate tools, slowing every release.

06

Vendor lock-in

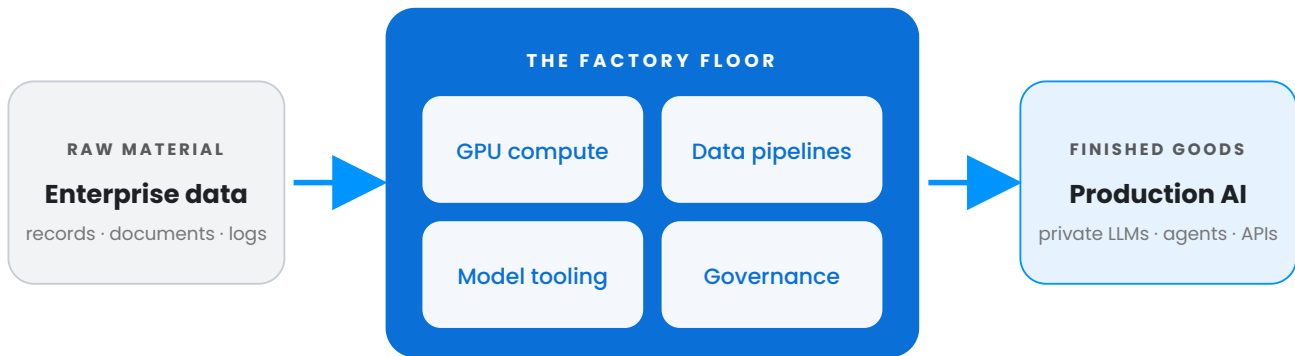
Proprietary AI services tie models, pipelines and skills to one provider's roadmap and pricing.

The CIO's takeaway: the constraint on enterprise AI is rarely the model. It is the infrastructure, data control and operating model around it. That is exactly the problem an AI Factory is designed to solve.

03

What is an AI Factory?

An AI Factory is a dedicated environment that turns raw enterprise data into trained, deployed and continuously improving AI, the way a factory turns raw materials into finished goods: repeatably, at scale and under quality control.



The concept matters because it changes how AI is funded and managed. Instead of every business unit procuring its own GPUs and tools, the enterprise operates one governed capability that every team can use.

Introducing UPC AI Factory

UPC AI FACTORY

UPC AI Factory is part of UnitedLayer's United Private Cloud® suite. It enables enterprises to build and operate private large language models and AI pipelines without sending sensitive data to a public AI service, running on dedicated, single-tenant infrastructure built from the silicon up for high-performance compute, and fully managed by UnitedLayer.

Private

Your data, models and outputs stay in a dedicated environment

Production-grade

Enterprise resilience, observability and compliance built in

Managed

UnitedLayer operates the stack so your teams focus on models

04

Inside UPC AI Factory: the architecture

When you deploy AI workloads on UPC, every component is yours. The platform is built as an integrated stack, so each layer is engineered to feed the one above it without bottlenecks.

AI SERVICES

Private LLM-as-a-Service · custom models trained on your data · AI-driven automation and remediation

AI LIFECYCLE TOOLCHAIN

100+ integrations across data ingestion, data engineering, model development, MLOps and inference

ORCHESTRATION

Native Kubernetes orchestration for scheduling, scaling and containerized AI workloads

DATA & NETWORK FABRIC

RDMA-native networking · NVMe storage for ultra-fast data access · software-defined networking via UnitedConnect®

ACCELERATED COMPUTE

Dedicated NVIDIA GPU clusters (H100, A100, L40S, T4) on single-tenant infrastructure

Cross-cutting

UPC Built-in Cloud Intelligence

Observability, hybrid cloud management and agentic orchestration

Security & governance

Encryption, MFA, intrusion detection, policy-as-code compliance

Resilience

N+M redundancy and automated failover

Three technical differentiators worth asking any provider about

RDMA-native networking

Remote Direct Memory Access lets GPUs exchange data directly across nodes, reducing the communication overhead that slows distributed training.

No resource contention

Dedicated GPU clusters eliminate the noisy-neighbor effect of shared public cloud, delivering predictable performance for every training run.

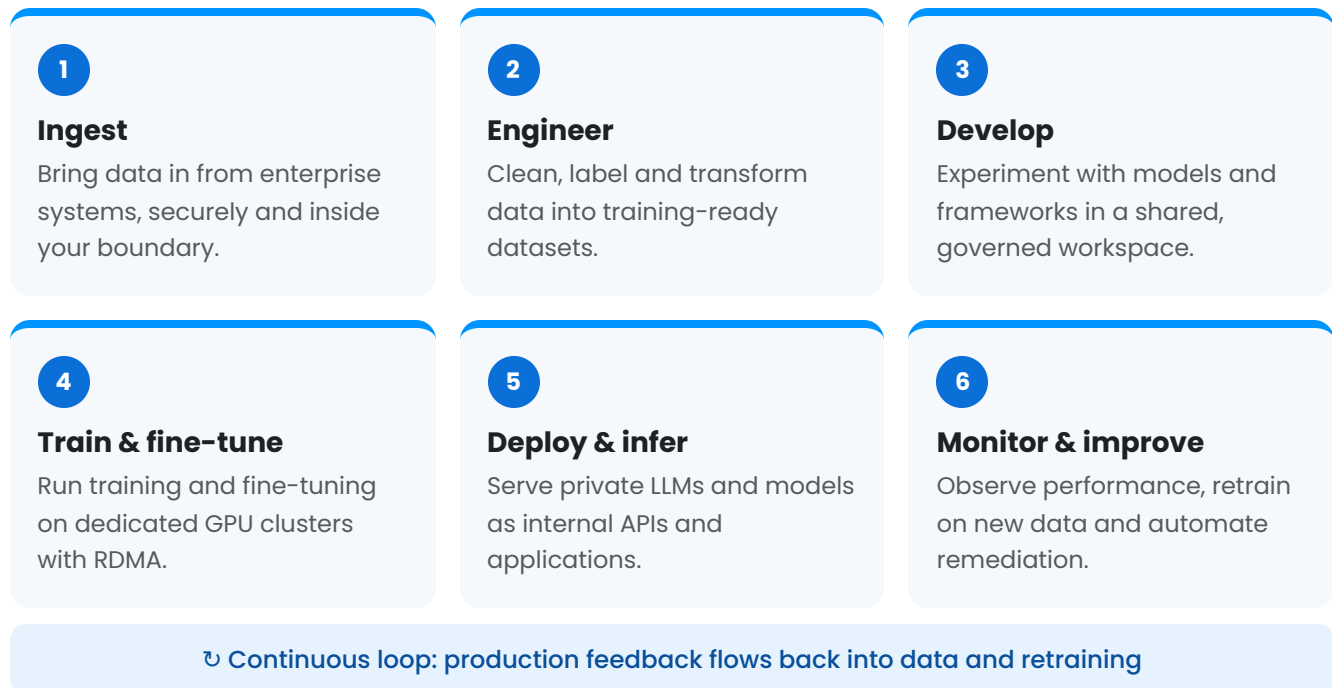
Reserved capacity

Train on readily available NVIDIA H100 and L40S GPUs, or reserve premium GPUs up to a year in advance so capacity is there when programs need it.

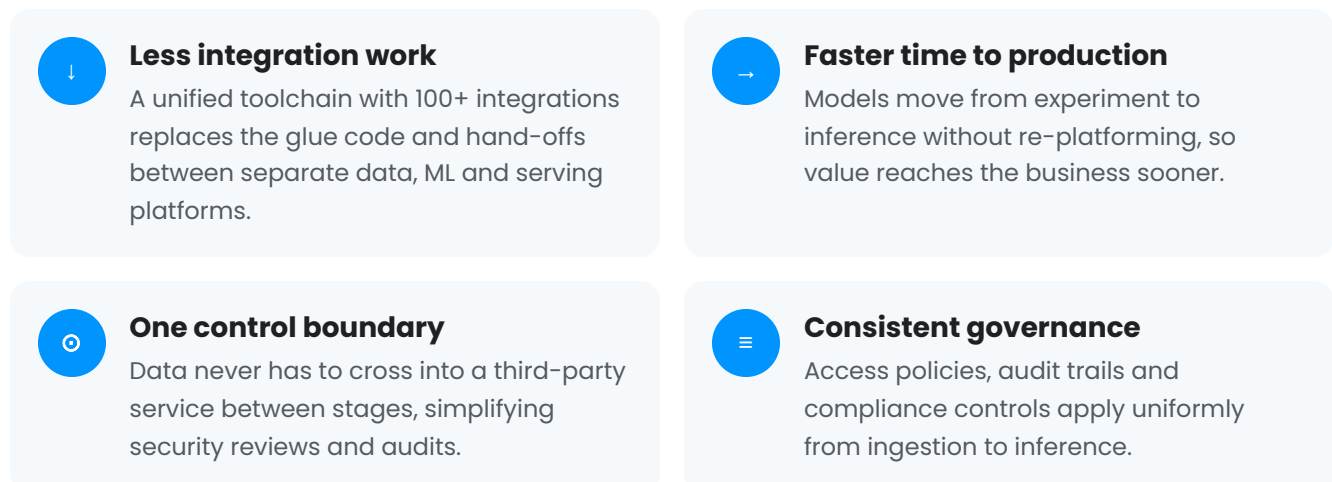
05

The end-to-end AI lifecycle, in one environment

Enterprise AI infrastructure requires more than GPU availability. UPC AI Factory brings every stage of the lifecycle into a single private environment, removing friction between data, models and deployment.



Why a single environment changes the economics



06

Five outcomes that matter to the board

Technology choices earn board support when they map to business outcomes. Here is how UPC AI Factory translates into the language of risk, speed, cost, continuity and compliance.

01 Data sovereignty and control

With a private environment, your data never leaves dedicated infrastructure. Train on proprietary data, host your own LLMs and keep prompts and outputs inside your boundary.

Private

data stays in your environment

02 Performance at AI scale

A GPU-native, single-tenant architecture with RDMA networking and NVMe storage is designed for large-scale training.

40%

faster training vs. hyperscalers*

03 Predictable economics

Dedicated capacity replaces volatile metered GPU hours and per-token charges, making AI programs easier to budget and scale.

Fixed

capacity, not surprise bills

04 Resilience for always-on AI

In AI, downtime is an expensive failure. N+M redundancy, redundant infrastructure and automated failover protect long training runs and live inference.

99.999%

uptime commitment

05 Compliance by design

Built-in policy governance, automated audit trails and support for HIPAA, PCI DSS, GDPR, ISO and SOC frameworks help regulated teams move faster.

Audit-ready

controls built in

*Based on UnitedLayer's published analysis of AI training performance on UPC's dedicated, single-tenant architecture. Actual results vary by model, dataset and configuration.

07

Choosing your AI infrastructure model

Most enterprises weigh three options. Each has a place, but for sensitive, production-scale AI the trade-offs look very different.

	Public AI cloud	DIY on-premises	UPC AI Factory
Data control	Data leaves your boundary to a shared service	Full control	Full control in a dedicated, managed environment
GPU performance	Shared; subject to contention and queueing	Dedicated, but limited by in-house design skills	Dedicated GPU clusters with RDMA and NVMe
Time to capacity	Fast for small jobs; scarce for premium GPUs	Months of procurement and build-out	Available GPUs now, or reserve up to a year ahead
Cost profile	Metered and volatile at scale	High upfront capex and staffing	Predictable, dedicated-capacity economics
Operations burden	Low for infra; high for integration	High: your team runs everything	Managed by UnitedLayer, with AI-driven monitoring
Compliance	Shared-responsibility complexity	You build every control	Built-in governance and audit trails
Lifecycle tooling	Provider-specific, lock-in risk	Assembled from point tools	Unified toolchain, 100+ integrations

Keep public AI for: short-lived experiments on non-sensitive data, and bursty workloads where speed of access outweighs control.

Choose a private AI Factory for: models trained on regulated or proprietary data, continuous inference, and programs where cost predictability and sovereignty matter.

A hybrid approach is common. A private AI Factory becomes the governed core, with public services used selectively at the edge of the program.

AI Factory in action: industry scenarios

Organizations across financial services, healthcare, government and technology choose private cloud because their most valuable data is also their most sensitive. These are representative scenarios for an AI Factory.

Financial services

TYPICAL WORKLOADS

Fraud and risk models trained on transaction data; private LLM assistants for analysts and relationship managers; document intelligence for KYC and lending.

WHY PRIVATE

Keeps customer financial data inside a controlled boundary aligned to PCI DSS and regulatory expectations.

Healthcare & life sciences

TYPICAL WORKLOADS

Clinical documentation assistants, imaging analysis and research models trained on patient and trial data.

WHY PRIVATE

Supports HIPAA-aligned handling of protected health information without exposure to public AI services.

Government & public sector

TYPICAL WORKLOADS

Citizen-service assistants, case-file summarization and secure analytics on sensitive records.

WHY PRIVATE

Dedicated infrastructure supports residency and control requirements for public-sector data.

Technology & GCCs

TYPICAL WORKLOADS

Code assistants trained on internal repositories, IT operations copilots and AI products built on proprietary IP.

WHY PRIVATE

Protects source code and IP while giving engineering teams GPU capacity on demand.

Common thread: in every scenario, the model is only as valuable as the data behind it, and the data is only usable if it stays under enterprise control.

09

A 90-day path to your first production model

The fastest AI programs start narrow, prove value on real data and then scale the factory. This phased plan is a practical starting point for most enterprises.

Days 0–30

Assess & align

- Select one or two high-value use cases with clear owners
- Classify the data involved and its compliance requirements
- Define success metrics the board will recognize
- Size GPU, storage and network needs with UnitedLayer architects

Days 31–60

Build & train

- Provision dedicated GPU capacity on UPC AI Factory
- Stand up data pipelines and the lifecycle toolchain
- Fine-tune or train the first model on private data
- Establish access policies and audit logging

Days 61–90

Deploy & scale

- Serve the model to a pilot user group as an internal API
- Monitor quality, latency and utilization
- Report results against the success metrics
- Build the backlog for the next wave of use cases

Roles that make it work

Role	Responsibility
CIO / CTO	Sponsors the program, owns the infrastructure decision and reports outcomes
Business owner	Defines the use case, success metrics and adoption plan
Data & ML leads	Own data preparation, model development and evaluation
Security & compliance	Approve data handling, access policies and audit requirements
UnitedLayer	Designs, provisions and manages the AI Factory infrastructure

10

The CIO's evaluation checklist

Use these twelve questions to evaluate any AI infrastructure provider, including UnitedLayer. UPC AI Factory is designed to answer “yes” to each.

Data & sovereignty

- ☒ Does sensitive data stay inside a dedicated, single-tenant environment?
- ☒ Can we host our own LLMs rather than calling a public AI service?
- ☒ Are prompts, outputs and training data kept out of third-party systems?

Performance

- ☒ Are GPU clusters dedicated, with no resource contention?
- ☒ Is the network RDMA-capable for distributed training?
- ☒ Can premium GPU capacity be reserved in advance?

Operations

- ☒ Does one toolchain cover ingestion through inference?
- ☒ Is the environment managed, with proactive AI-driven monitoring?
- ☒ Does it run on Kubernetes to avoid lock-in?

Risk & cost

- ☒ Is there N+M redundancy and a clear uptime commitment?
- ☒ Which compliance frameworks are supported, with audit trails?
- ☒ Are costs predictable as usage scales?

Tip: ask each provider to answer in writing and to show you the architecture behind every “yes.” The differences between providers usually appear in the details of data control, contention and cost.

Build your private AI Factory with UnitedLayer

United Private Cloud® is an AI-native, enterprise-grade managed private cloud platform that delivers the agility of the public cloud with the security, compliance and control of dedicated infrastructure. UPC AI Factory extends it with everything enterprises need to build, train and run private AI.

100+

IaaS and PaaS services on UPC

99.999%

uptime commitment

ISG Leader

ISG Provider Lens™ 2026

Talk to a UnitedLayer AI infrastructure architect

Map your first AI Factory use case and size the capacity you need.

unitedlayer.com

UPC AI FACTORY

 UNITEDLAYER®