

CUSTOMER SUCCESS STORY

A global bank put its AI back inside its own walls.

Fraud detection, a copilot, and risk models all ran beautifully in the lab — and nowhere else. Here is how the bank moved them into production on **UPC AI Factory** in twelve weeks, without customer data leaving its boundary.

BUILT ON UPC AI FACTORY

THE SITUATION

Three AI programmes, all stuck at the same gate

The bank's AI team was not short of results: fraud models flagged anomalies in real time, the copilot answered policy questions in seconds, and new risk models beat the legacy scorecards. Production was where each one stopped, and every path forward hit the same four obstacles:

- **Regulatory exposure.** Prompts, embeddings, and fine-tuned models sat on shared infrastructure — a non-starter for customer PII.
- **Budget that would not hold still.** GPU, egress, and API charges made every invoice a surprise, and TCO could not be forecast.
- **Latency the business could feel.** Shuttling data across regions made real-time fraud scoring inconsistent.
- **A fragmented platform.** Kubernetes, GPU operators, vector DBs, and serving stacks ran with no shared view.

THE DECISION

Own the factory, not just the output

Rather than negotiate harder with a hyperscaler, the bank brought the whole estate in-house — operated by UnitedLayer, controlled by the bank.

“If AI is going to be core to our business, it cannot run on borrowed infrastructure. We need our own factory — where we control the raw materials, the production line, and the finished goods.”

— CHIEF INFORMATION OFFICER

CUSTOMER PROFILE

INDUSTRY

Banking & financial services

PROFILE

Global bank, multi-region operations

AI WORKLOADS

Fraud detection · RM copilot · Risk & underwriting

FOOTPRINT

Two primary data centres plus a DR site

TENANCY

Three isolated vDCs — Fraud, Retail Banking, Risk

TIME TO PRODUCTION

12 weeks, end to end

COMMERCIAL MODEL

Single license, fixed monthly charge

SOLUTION

UPC AI Factory, fully managed

WHAT WAS DEPLOYED

One operated stack, inside the bank's boundary

AI services

OpenAI-compatible Router, Model Registry, managed Vector DB, and GPU Workbench — so existing applications pointed at private models without rewrites.

Compute

Dedicated **GPU pools with MIG partitioning**, split between training and inference, on managed Kubernetes with the NVIDIA GPU Operator.

Data

High-performance storage for datasets, model artefacts, and vector indexes, **co-located with GPU** and linked privately to existing sources.

Resilience & operations

Backup, DR, and a **cyber vault** for critical AI assets; full-stack observability and AIOps wired into existing ITSM and security tooling, plus **FinOps dashboards** reporting cost per million tokens.

THE RESULTS · YEAR ONE

↓ **42%**

Lower TCO for AI workloads
Against the prior hyperscaler setup

↑ **40%**

Faster AI training
Single-tenant, RDMA-native GPU fabric (35–40%)

<100ms

p95 inference latency
Real-time fraud scoring and copilot

100%

AI data and models inside the tenant boundary

No external exposure, satisfying internal risk, external auditors, and data-residency obligations in every region.

~60%

Less platform toil for the team

One operated stack and unified dashboards replaced the tooling each group had been maintaining separately.

Fixed

A monthly cost finance can forecast

Cost per million tokens and cost to serve per model, visible by business line, made multi-year AI budgeting possible.

WHY THIS MATTERS IN FINANCIAL SERVICES

Keeping the estate inside the tenant boundary turned AI governance from an open question into evidence an auditor can inspect — data residency, model lineage, and access control all demonstrable in one place.

Customer PII kept in-boundary

Data residency by region

Model lineage & versioning

Auditable access control

Cyber vault for AI assets

SOC 1/2 · PCI · GDPR · FINRA aligned

WHERE THE BANK IS TAKING IT NEXT



AML & transaction monitoring

Extending the fraud pattern to anti-money-laundering, where the same latency and data-residency constraints apply.



KYC & document intelligence

Onboarding and due-diligence documents processed in-boundary, with model lineage available to auditors.



Risk modelling at scale

Fixed GPU capacity makes stress testing and scenario modelling a planning decision rather than a budget question.

BOTTOM LINE

UPC AI Factory turned scattered AI POCs into a scalable, cost-predictable, fully controlled production AI estate.

Ready for enterprise AI without compromise?

Book a walkthrough, or ask us to scope a private AI factory for your workloads.

UPC AI FACTORY



EMAIL

contact@unitedlayer.com



PHONE

+1-888-853-7733

[UNITEDLAYER.COM](https://unitedlayer.com)