

CUSTOMER SUCCESS STORIES

Consumer AI at scale, on infrastructure you own.

Personalization, recommendation, content operations, and network optimization run on the most sensitive asset these businesses hold — first-party consumer data. Three operators moved those workloads onto **UPC AI Factory** and kept the data inside.

53–78%

AI adoption across retail, media, and telecom spans roughly 53–78% — retail toward the lower end, media and telecom higher. These are mid-to-late adopters with production workloads already running, where the question is no longer whether to deploy but what it costs to serve and where the data sits.

BUILT ON **UPC AI FACTORY**

THE COMMON THREAD

High volume, thin margins, sensitive data

The three operators ran different workloads but hit the same four constraints once AI moved from a feature to a dependency:

- **First-party data is the asset.** Purchase history, viewing behaviour, and subscriber records are what differentiate them — and what they least wanted on shared infrastructure.
- **Inference cost scales with traffic.** Recommendation and personalization run on every session, so token and GPU spend tracked revenue rather than headcount.
- **Latency is conversion.** A recommendation that arrives after the page renders does not influence the basket or the watch decision.
- **Privacy commitments got specific.** Consent regimes and regional regulation made "where does this data go" a question with a documented answer.

THE DECISION

Own the serving layer

Each operator moved inference onto dedicated capacity sited close to their customers, operated by UnitedLayer and governed by their own teams.

“Our recommendation engine touches every session we serve. Once that runs on capacity we control, personalization stops being a variable cost and starts being infrastructure.”

— GROUP CTO, MULTI-BRAND RETAIL & MEDIA GROUP

DEPLOYMENT SNAPSHOT

SECTORS

Retail · Media & entertainment · Telecom

AI WORKLOADS

Personalization · Recommendation · Content operations · Network optimization

FOOTPRINT

Two regional data centres plus a DR site, per operator

TENANCY

Isolated vDCs by brand, channel, or network domain

TIME TO PRODUCTION

12 weeks, end to end

GOVERNANCE

Consumer data in-boundary, regional residency

SOLUTION

UPC AI Factory, fully managed

WHAT WAS DEPLOYED

A serving layer built for session-rate traffic

AI services

OpenAI-compatible Router, Model Registry for per-brand and per-region model versions, managed Vector DB for catalogue and content retrieval, and GPU Workbench.

Compute

Dedicated **GPU pools with MIG partitioning**, separating always-on recommendation inference from batch retraining and content generation runs.

Data

High-performance storage for behavioural data, embeddings, and media assets, **co-located with GPU** so session traffic never crosses a region to be scored.

Governance & operations

Regional residency controls, backup, DR, and a **cyber vault**; observability and AIOps into existing tooling, with **FinOps dashboards** showing cost to serve per session.

THE RESULTS · YEAR ONE

↓ **42%**

Lower TCO for AI workloads

Cost to serve that no longer rises with every session

↑ **40%**

Faster model training

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

<100ms

p95 inference latency

Recommendations that land before the page does

100%

Consumer data and models inside the tenant boundary

First-party data stayed first-party, with regional residency documented rather than assumed.

~60%

Less platform toil

One operated stack across brands, channels, and regions replaced separately maintained serving pipelines.

Fixed

Cost to serve, per session and per model

Personalization moved from a variable line that grew with traffic to a planned infrastructure cost.

THREE OPERATORS, THREE STARTING POINTS



Retail · Personalization

Product recommendation and search ranking across brands, scored in-session on shared customer profiles that never leave the group's boundary.



Media · Content operations

Metadata enrichment, localisation, and recommendation over the catalogue, with generation runs scheduled against fixed capacity instead of metered GPU.



Telecom · Network optimization

Traffic prediction and fault detection on subscriber telemetry, processed in-region so operational data stays within the licence footprint.

WHY THIS MATTERS

First-party data kept in-boundary

Regional residency

Consent-aligned processing

Cost to serve per session

Session-rate latency

Multi-brand tenant isolation

BOTTOM LINE

UPC AI Factory turned personalization, content, and network AI from a cost that scaled with traffic into infrastructure these operators control.

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)