CLIENT OVERVIEW

A fast-growing mid-sized online retailer headquartered in the European Union was struggling to compete with larger players whose AI-driven recommendation engines and demand-forecasting systems were delivering measurably better conversion rates and inventory efficiency. The company's data science team had the models — they lacked the compute.

THE CHALLENGE

Procuring and operating a dedicated GPU cluster was out of the question. A single enterprise-grade NVIDIA H100 server costs upward of €300,000, and that figure does not include racking, cooling, power, networking, or the ongoing engineering burden of keeping it running. US-based hyperscalers offered GPU rentals, but European data protection regulations — and the company's own privacy-first brand positioning — made routing customer data outside the EU legally and commercially untenable.

• **No GPU hardware budget.** Capital expenditure for AI compute was not approved for the fiscal year.

• **GDPR & data sovereignty.** Customer purchase history, browsing behaviour, and personal identifiers could not leave EU jurisdiction under any circumstances.

• **Speed to market.** A competitive peak-season window meant the team needed production-grade infrastructure in weeks, not the months a hardware procurement cycle would require.

• **Scalability uncertainty.** Workloads ranged from overnight batch training jobs to real-time inference at checkout — wildly different compute profiles impossible to right-size upfront.

THE SOLUTION

The retailer partnered with **Sapience AI** — Europe's sovereign GPU-as-a-Service platform — to access enterprise NVIDIA GPU capacity hosted inside EU Tier 3 data centres, with zero hardware investment required.

Full AI Infrastructure, Managed

Sapience AI provided bare-metal NVIDIA GPU nodes — including RTX PRO 6000 Blackwell class hardware — pre-configured with a production ML stack (CUDA, cuDNN, PyTorch, Kubernetes via Rafay Systems). The client's team deployed models directly; the infrastructure layer was entirely managed by Sapience.

No GPU Purchase Required

Rather than procuring hardware, the team consumed GPU hours on a pay-as-you-go basis. Training runs were billed by the hour; inference endpoints scaled elastically. Capital expenditure was replaced entirely by predictable operational expenditure — with no minimum commitment.

All GPU instances ran in **VNET Tier 3 data centres within the EU**. Training data, model weights, and inference outputs never left Union borders — guaranteed by infrastructure architecture, not policy statements. The setup passed the company's internal GDPR review in under a week.

RESULTS

0	3 wks	+34%	~60%
GPUs purchased	time to production	recommendation CTR	lower vs. capex path

Within three weeks of onboarding, the retailer's recommendation engine was live in production, serving personalised product suggestions to shoppers at checkout and on category pages. A demand-forecasting model trained on 18 months of historical sales data reduced overstock write-offs by measurably cutting end-of-season surplus.

- **Click-through rate on recommendations rose 34%** within the first 30 days, directly attributable to the collaborative-filtering model deployed on Sapience infrastructure.
- **Inventory overstock decreased by 22%** in the first quarter following deployment of the demand-forecasting model, reducing end-of-season markdowns.
- **Zero compliance incidents.** All customer data remained within EU borders throughout training, fine-tuning, and inference. The legal team signed off before go-live.
- **Infrastructure cost 58% lower** than equivalent AWS or GCP GPU instances, with no capital outlay and no long-term lock-in.

WHY SAPIENCE AI

EU Sovereign Infrastructure — GPU compute hosted exclusively in EU Tier 3 data centres. Data never crosses EU borders — by architecture, not policy.

No Hardware Investment — Pay for GPU hours, not servers. Eliminate CapEx entirely and turn AI compute into a predictable operational line item.

Managed Stack — CUDA, cuDNN, PyTorch, and Kubernetes are pre-configured and kept up to date. Your team ships models — Sapience manages the rest.

Elastic Scale — Scale from a single GPU for experimentation to multi-node clusters for large training runs. Pay only for what you use.

GDPR by Architecture — Single-tenant isolation, hardware-level separation, encryption at rest and in transit. Built for regulated industries from day one.