LARSEN & TOUBRO
vyoma

Implementing a GPU-Backed Private AI Cloud for an Enterprise SaaS Platform

About the Client

The client is a rapidly scaling enterprise SaaS company delivering mission-critical HR and finance solutions to medium and large organizations. Their platform manages highly sensitive employee data, internal workflows, and financial records, operating under strict compliance and data governance expectations.

As the customer base expanded, sensitivity of the data was processed. At the same time, the business needed to embed AI-driven capabilities directly into its core SaaS offerings.

The Challenge

Initial experimentation with third-party AI services helped us validate the cases, but it quickly exposed structural limitations.

1. Firstly, privacy and compliance risks became a primary concern. Customer data was routed through external AI platforms, triggering prolonged reviews from legal and security teams and creating friction in enterprise sales cycles.

2. Cost unpredictability was another issue as consumption-based pricing models made it difficult to forecast AI spend, especially as AI use cases expanded across multiple products and customer segments.

3. Finally, limited control over models and pipelines slowed innovation. Customization was constrained by vendor tooling, and even minor changes required dependency on external providers which prevented product and engineering teams from iterating at the pace the business demanded.

Client Requirements

Client's requirement was primarily AI governance at scale, not only adoption.
And so, they needed solution that would-

- ***Keep sensitive HR and financial data entirely within their own environment***
- ***Support on-demand scalability without sudden or opaque cost escalation***
- ***Enable internal teams to train, optimize, and deploy models without vendor lock-in***

Fundamentally, AI had to become a first-class component of the SaaS architecture, fully governed rather than existing as an external dependency.

The Solution: Private AI Cloud Architecture

To meet these needs, the client implemented a secure private AI cloud purpose-built for large-scale training and inference workloads.

1. The platform was designed around a GPU-optimized compute layer, leveraging high-performance infrastructure capable of parallel processing and memory-intensive AI workloads. A high-throughput, low-latency interconnect, based on InfiniBand networking, enabled efficient GPU-to-GPU communication, significantly reducing training bottlenecks.

2. An AI-optimized storage and network architecture was deployed to prioritize east-west traffic, eliminating I/O constraints during model training and inference, which ensures consistent performance as workloads are scaled.

Crucially, the solution delivered full ownership of data, models, and pipelines. All AI processing remained within the client's private environment, ensuring data sovereignty, and enterprise-grade security and compliance.

Execution Approach

Initial efforts focused on migrating high-value AI workloads into the private environment and validating performance and stability. Once operational baselines were achieved, the focus shifted to cost optimization, and reduced model training cycles.

At the same time, internal engineering teams were trained to run and manage the platform independently. This helped reduce the long-term reliance on external vendors and ensured the AI infrastructure could be sustained and developed internally.

Measurable Results

The transition to a private AI setup delivered clear, quantifiable outcomes-

- ***Accelerated AI training cycles, enabling faster experimentation and improved R&D velocity.***
- ***25–30% lower total cost of ownership compared to hyperscaler-based AI services.***
- ***Predictable operating costs, supporting accurate financial planning and scaling.***
- ***Enhanced security posture, with all customer data remaining within the client's controlled infrastructure.***

Business Impact

From a commercial standpoint, the move strengthened customer trust. The client could clearly articulate how AI capabilities were delivered with privacy, compliance, and control as core principles. AI innovation was no longer constrained by external pricing models, tooling limitations, or vendor timelines, and thus, product and engineering teams gained autonomy

Most importantly, the client now operates on a future-ready AI platform that has helped support long-term product innovation, scalable growth, and strict delivery SLAs without compromising governance.

Conclusion

By transitioning to a private AI cloud, the client addressed immediate challenges around security, cost control, and operational governance. More importantly, they established a robust foundation for embedding AI deeply and sustainably into their SaaS platform.

LARSEN & TOUBRO

vyoma

Start Your Journey With Us

Contact Us