



**From silos to scale:
Operationalizing AI
across the organization**



AI has become a driving force for innovation, but scaling it presents unique challenges. Many organizations begin their AI journey by solving specific use cases within individual teams or departments. While this approach can yield early wins, it often leads to siloed infrastructure, duplicated efforts, and inefficiencies as AI initiatives grow. The real challenge lies in operationalizing AI at scale — transforming isolated successes into a cohesive, organization-wide effort that drives meaningful outcomes.

For service providers and model builders acting as internal service providers for their teams, the stakes are particularly high. Managing the entire AI lifecycle — from model building and training to fine-tuning, inference, and monitoring — requires robust infrastructure, scalability, and reliability. These organizations must support diverse, multi-tenant workloads while ensuring security and governance. Without a strategic, scalable approach, the complexity of scaling AI can quickly undermine its potential to deliver transformative results.

Centralized control for decentralized AI infrastructure

The complexity of AI operations often stems from their decentralized nature. AI workloads span on-premises, edge, and cloud environments, creating silos that hinder collaboration and slow progress. Without a unified approach, organizations risk delays, wasted resources, and missed opportunities to scale AI effectively. The key to overcoming this challenge is centralization of management and control.

A centralized platform provides a unified view of AI environments, enabling organizations to manage resources more effectively and foster collaboration across teams. Cloud-like management capabilities, such as built-in orchestration, observability, and lifecycle automation, simplify operations and ensure visibility into every layer of the AI stack. By breaking down silos and streamlining workflows, businesses can scale their AI operations efficiently, achieving greater innovation and impact. Importantly, centralized control also allows service providers and model builders to deliver secure, multi-tenant AI workloads, supporting diverse teams while maintaining resource isolation and governance.



Optimizing resources for a multi-tenant world

Scaling AI isn't just about building more infrastructure — it's about using what you have intelligently. The demand for AI processing power is growing exponentially, especially for training and inference workloads. But without intelligent workload placement and resource optimization, businesses risk underutilized GPUs, rising costs, and bottlenecks that impede progress. Multi-tenancy is at the heart of addressing this challenge.

Multi-tenancy enables multiple teams or departments to securely share infrastructure while isolating data, models, and security configurations. At a hardware level, GPUs can be allocated dynamically to ensure optimal performance for diverse workloads. Monitoring tools play a critical role, providing insights into the performance and quality of AI models to prevent performance decline and ensure governance. Additionally, resource optimization extends beyond computational power to include considerations for power and cooling, ensuring infrastructure is also efficient. By creating a dynamic, flexible infrastructure that balances performance, reliability, and cost efficiency, organizations can maximize their resources while delivering scalable AI operations tailored to their specific needs.

Simplifying AI lifecycles with automation, observability, and integration

Managing AI pipelines can be complex and time-consuming, with repetitive manual tasks and an overwhelming array of performance metrics slowing down innovation. Automation and observability streamline lifecycle management, allowing organizations to optimize workflows and ensure AI operations remain efficient and reliable. Automation reduces manual overhead by enabling self-service deployment and management of resources, while observability tools provide real-time insights into performance and governance, helping teams detect and resolve issues before they escalate. In multi-tenant environments, metering plays a critical role by tracking resource usage across teams or departments, enabling organizations to allocate costs fairly, monitor consumption, and ensure accountability. Together, these capabilities simplify IT operations across multi-vendor, hybrid, and multi-cloud environments, freeing teams to focus on innovation.

Seamless integration is equally critical to achieving the full potential of AI. By aligning AI operations with existing DevOps and IT tools — such as Kubernetes, MLOps frameworks, and data services — organizations can reduce disruption and accelerate adoption. Integration allows AI services, including training, inference, and monitoring, to function alongside existing infrastructure, cutting time to AI value and lowering operating costs. This synergy ensures that AI initiatives complement broader business operations, empowering organizations to deliver impactful outcomes while maintaining consistency, control, and resource efficiency.

Transforming AI operations at scale

Scaling AI across the organization requires more than just infrastructure. It demands a holistic approach that simplifies complexity, optimizes resources, and drives meaningful outcomes. For organizations building AI factories at large scale or with complex workload needs, HPE delivers customized solutions that operationalize AI across the entire lifecycle, from model building and training to fine-tuning, inference, and monitoring. These solutions are designed to adapt to the unique needs of each organization, providing a flexible, conceptual reference architecture that integrates seamlessly with existing infrastructure and support scalable, multi-tenant AI operations with a GPU-cloud-like experience.

HPE goes beyond technology to deliver unrivaled services and real-world global expertise, ensuring the success of every deployment. HPE Services experts partner with organizations at every step of their AI journey, helping with strategy, design, finance, deployment, education, management, support, and refresh. These services accelerate the implementation of large-scale AI factories, helping ensure that the solution evolves over time, maximizing value, and enabling organizations to achieve AI-driven outcomes faster.

With near-linear scalability, AI factories at scale support workloads ranging from medium to very large clusters while maintaining high performance, cost efficiency, and resource optimization. Built-in orchestration, observability, and lifecycle automation simplify operations and reduce overhead, ensuring reliable and compliant AI model performance. HPE empowers service providers and model builders to scale AI confidently and effectively, paving the way for transformative innovation and long-term success.

Learn more at

[HPE.com/ai/insights](https://hpe.com/ai/insights)

Visit [HPE.com](https://hpe.com)

[Chat now](#)

© Copyright 2025 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

a50013338ENW

HEWLETT PACKARD ENTERPRISE

hpe.com

