# AI Factory Purchasing Guide

## What Is an AI Factory?

An AI factory is a full-stack solution with accelerated computing, networking, and optimized AI software. It's designed to create value from an organization's data by managing the entire AI life cycle—from data ingestion to training and fine-tuning to high-volume inference. It leverages a broad ecosystem of partners with validated designs and reference architectures, enabling the fastest time to market for agentic AI, physical AI, and HPC workloads.

Organizations can deploy AI factories of all sizes to meet their business needs. The modular design of an NVIDIA AI factory allows organizations to easily scale as their AI use cases grow.

## Option 1: Build

**Overview**
Involves investing in and deploying your own hardware, software, and infrastructure within your organization's data centers or private cloud environments.

**Key Features**
Complete control, customizable hardware/software; greater ROI for highly utilized workloads; data stays on-site.

### Why it Makes Sense

- **Security and Compliance:** Ensures maximum protection and regulatory alignment with predictable long-term costs.
- **Full Control:** Enables customization of every layer of your AI infrastructure, from hardware to software.
- **Cost Efficiency at Scale:** Delivers predictable, long-term costs after initial investment for highly utilized, long-term workloads; marginal cost per query is near zero.
- **Data Protection:** Provides enhanced data security and compliance for sensitive or regulated industries.
- **Low Latency:** Supports real-time inference and local workloads.
- **Data Sovereignty:** Preserves full ownership of IP and sensitive data.

### Things to Consider

- **Up-Front Investment:** High initial costs require in-house expertise and maintenance.
- **Infrastructure Commitment:** High initial investment in hardware, software, and facilities, as well as potential additional costs for upgraded power and cooling systems to meet requirements.
- **Scaling Limits:** Adding capacity requires new capital hardware purchases and configuration, which may be confined by data center limitations and take weeks to months to arrive.
- **Ongoing Operations:** Maintenance, upgrades, and monitoring are your responsibility, though managed services are available from many partners.
- **Expertise Required:** Deployment and management require in-house expertise.

### Ideal for

- Large enterprises with stable, sensitive workloads.
- Large enterprises with stable, high-volume AI workloads.
- Industries with strict data privacy, security, or compliance requirements (e.g., healthcare, finance, government).
- Organizations with existing data centers and IT expertise.
- Use cases requiring low latency, high throughput, or data sovereignty

### Use Cases

- Long-term, high-volume AI deployments; regulated industries.
- Enterprise-scale AI model training and inference (e.g., proprietary large language models, sensitive data processing).
- Mission-critical applications (e.g., predictive maintenance, real-time analytics, regulated industries).
- Long-term, high-volume AI deployments where control and cost efficiency are paramount.

## Option 2: Rent

### Overview

Quickly deploy AI workloads across your enterprise by renting an AI factory from Cloud Service Providers (CSPs) or NVIDIA Cloud Partners (NCPs).

### Key Features

Instant GPU access, flexible consumption models, rapid scalability, latest accelerated computing platforms, managed services, cloud-native integrations, multi-cloud portability.

### Why it Makes Sense

- **Speed and Simplicity:** Get started in minutes. No need to worry about power, cooling, or racking hardware. Just bring your data and go. Run proof of concepts (POCs) fast to clear the POC backlog.
- **Scalable and Flexible:** Scale easily from one to thousands of GPUs, making it ideal for spiky or short-term workloads.
- **Try Before You Buy:** Evaluate different NVIDIA Accelerated Computing platforms without long-term commitment.
- **Managed Services:** Offload infrastructure provisioning, management, and maintenance to cloud providers.
- **Ecosystem Integration:** Leverage broad integration with cloud-native services (e.g., storage, orchestration, observability).
- **Resiliency:** Rely on strong uptime service-level agreements and global infrastructure presence.
- **Portability:** Use NVIDIA's unified platform to ensure consistent performance and portability across clouds, enabling hybrid and multi-cloud strategies.

### Things to Consider

- **Cost Accumulation:** Pay-as-you-go pricing can add up quickly for sustained, large-scale usage. Customers may pay for cloud providers' cloud-native platforms, data egress, storage, and additional fees.
- **Limited Control:** You have less say in hardware configuration, infrastructure decisions, and software update schedules.
- **Vendor Lock-in:** Dependency on specific cloud provider's APIs and services
- **Data Sovereignty:** Sensitive data may leave your controlled environment, so there are potential concerns around data residency or compliance.

### Ideal for

- Enterprises, startups, research teams that:
- Need to prototype or scale quickly without infrastructure overhead.
- Have variable or bursty workloads that don't justify fixed infrastructure.
- Are experimenting with new GPUs, frameworks, or services in the cloud.
- Have existing deployments on CSPs or NCPs.

### Use Cases

- Burst workloads, short- to medium-term AI projects, experimentation, and development.
- Burst compute for training, fine-tuning runs.
- Hosted inference with variable demand.
- Hybrid and multi-cloud deployments for portability and resilience.

## Comparison Table

| Feature | Build | Rent |
|---|---|---|
| **Cost Structure** | Significant upfront investment; lower operational cost and greater ROI over time | Pay-as-you-go; flexible cost structure, but costs can increase with heavy continuous use |
| **Speed To Deploy** | Initial AI factory deployment requires procurement, setup, and deployment cycle. For subsequent projects, deployment times are similar to cloud if there is capacity available. | Same-day AI factory access; start in minutes |
| **Scalability** | Fixed capacity, but scalable; adding more requires new investments | Elastic—scale from one to thousands of GPUs |
| **Control and Customization** | Full control over hardware, software, and infrastructure | Limited control over hardware/software stack |
| **Data Security and Sovereignty** | Data remains on-site; ideal for sensitive and regulated workloads | Shared cloud infrastructure; data leaves local environment |
| **Use Case Fit** | Long-term, stable, high-volume production workloads | Short-term, bursty, or experimental workloads |
| **Best For** | Large enterprises with strict compliance or predictable, sensitive workloads | Startups, research teams, or enterprises experimenting or scaling quickly |

## FAQ

**Q: Should I build or rent an AI factory?**
A: Build for cost-predictability, security, compliance, and full control at scale. Rent for speed, experimentation, and agility—great for development, bursts, or variable workloads.

**Q: What are the trade-offs between control and convenience?**
A: Cloud is convenient and elastic, with less control. On-premises provides control and security but needs expertise and planning.

**Q: Which deployment model is best for AI model training?**
A: For computationally intensive AI training that requires massive GPU clusters, cloud deployment is preferred because it provides elastic scaling, managed services, and cost-effectiveness for periodic training cycles.

**Q: Which deployment model is best for AI model inference?**
A: On-premises deployment is often preferred for real-time, low-latency applications, while cloud may be more suitable for variable inference loads. Many enterprises take a hybrid approach: Train in cloud, deploy inference on-premises.

**Q: Can I switch between cloud and on-premises?**
A: Yes, NVIDIA Accelerated Computing and AI software enable on-premises, hybrid, and multi-cloud approaches, supporting transitions as your needs evolve. Hybrid deployment combines the benefits of both models, allowing strategic workload distribution based on specific requirements.

**Q: What is my initial decision framework for deploying an NVIDIA AI factory?**
A: Start by evaluating your data compliance requirements, then consider your usage pattern and total cost of ownership of your planned timeframe. Next, assess your in-house expertise to support. Choose the model (rent/build/hybrid) that provides the best balance of speed, control, security, and cost.

**Q: Where can I explore buying or building options for AI factories?**
A: Explore the NVIDIA Enterprise Marketplace for building on-prem AI Factories with your preferred OEM partner or renting options with your preferred NVIDIA Cloud Partner.

## Choosing the Right Strategy for Your AI Factory

Choosing between renting a cloud-based AI factory or building your own on-premises solution is ultimately a matter of your unique business requirements, existing resources, and future goals. Both approaches offer significant advantages and potential trade-offs in terms of scalability, control, cost, security, and speed to deployment.

The right path depends on how your organization balances these priorities—your workload patterns, compliance needs, in-house expertise, and growth trajectory will all shape the most effective strategy. As AI initiatives evolve in complexity and scale, many enterprises find value in blending both approaches or continually reassessing their deployment models to stay aligned with business objectives and technology advancements.

## Explore AI Factory Purchasing Options

nvidia.com/buy-ai-factory/



NVIDIA