AHEAD

Building a Production-Ready AI Factory with AHEAD

How AHEAD streamlines the end-to-end enterprise AI lifecycle with repeatable, vendor-agnostic architectures

Most organizations have experimented with AI at this point, and many have experienced early wins with individual pilot programs. However, enterprise AI initiatives often get stuck in the initial pilot phase without a practical method to move them to large-scale production. This requires a standardized approach that shifts from fragmented AI projects to a unified model for delivering real AI-driven business solutions.

An AI Factory is a governed, repeatable system for turning data into AI-powered outcomes. It includes the physical infrastructure required to run AI workloads, the software and platform services required to build and deploy them, and the operating model required to keep the whole system aligned with business and security goals.

Like any traditional factory model, this standardizes the way you turn data into decisions, insights, and AI automations. And with those repeatable, governed patterns in place, it becomes easier to improve and integrate AI capabilities into daily operations. The real advantage of the AI Factory model is not just faster deployment, but also the ability to scale AI repeatedly without redesigning the architecture, governance, and operating model for every new use case.

In this eBook, we'll guide you through AHEAD's AI Factory pattern that spans strategy, infrastructure, data, AI platforms, and operations. Our partner-agnostic approach is designed to run on your preferred hardware, cloud, and software stacks, with opinionated frameworks and reference architectures that reduce risk and time-to-value for your own AI Factory roadmap.

An Overview of the AI Factory

Before we get into the individual layers of an AI Factory, it's important to understand this blueprint at a higher level. An effective AI Factory is designed to move workloads or use cases along a standardized production line that includes infrastructure, data, AI platforms, and an operating model. By unifying these layers into a single blueprint, AI initiatives can move more quickly and reliably from exploration to production.

AI Factories address that perpetual, nagging investment question by helping you get more out of your existing resources. That's because this repeatable way to use infrastructure, data, and software for multiple AI use cases minimizes rework to deliver each new initiative. Performance often becomes more predictable even for demanding workloads like large language models, computer vision, digital twins, and high throughput analytics deployed across different on-premise and cloud environments.

An important design principle for the AI Factory is that it's a partner-agnostic blueprint that can be repeated for different use cases with the most appropriate technology stack. That's why AHEAD has worked with leading vendors to architect and integrate AI-optimized solutions across data, cloud, core, and edge for our enterprise clients. Capabilities like platform engineering, a standardized AI operating model, and managed services keep the factory running over the long term.

In the following sections, we'll look at each of the AI Factory layers:

Level 1

Optimized infrastructure to provide the foundation for AI-intensive workloads.

Level 2

Systems and frameworks that ingest, transform, store data for AI use.

Level 3

Centralized platforms to build models and expose AI capabilities to end users.

Level 4

Operations and governance standards that keep AI systems reliable and secure.

LEVEL 1

Physical & Infrastructure Layer

The backbone of the AI Factory is the infrastructure layer because modern AI workloads often put real pressure on compute and storage infrastructure, with very high power and cooling demands. This means designing and optimizing data center layouts, rack integrations, heat dissipation considerations, and capacity planning are critical to the successful deployment of AI systems.

At the infrastructure layer, GPUs, storage, and applications should all be connected by high bandwidth InfiniBand and Ethernet fabrics with the throughput and latency needed for AI training and inferencing. All of this should also be built on top of AI-ready storage platforms to serve as the data foundation.

Training workloads often require dense accelerated compute, high throughput storage, and low latency networking to fully utilize expensive GPUs. This requires careful planning for scalability to avoid bottlenecks that would reduce utilization and drive up costs. In addition, training and deploying computer vision models for image and video workloads often involves additional considerations, especially in the healthcare, manufacturing, and retail industries where these use cases are more common.

Inferencing has different infrastructure requirements from training workloads because the priority is usually responsiveness, efficiency, and support for real-time decision-making. For example, some use cases require low-latency inferencing to make critical decisions immediately, while other situations are better suited to batch or streaming inferencing patterns.

Edge AI — which involves inferencing at the branch, plant floor, or device level — is also a growing deployment pattern that involves specialized hardware, especially in the manufacturing industry (Industrial Internet of Things). Managing large fleets of edge nodes through a centralized hub requires visibility into the full lifecycle of components and consistent configurations to optimize IT operations and simplify hardware refreshes over time.

High-density AI deployments also force practical decisions around rack integrations, power distribution, and cooling design. For example, pre-integrated and validated racks can reduce deployment risk and compress time to go-live by reducing on-site assembly and troubleshooting. These designs may also need to incorporate direct-to-chip liquid cooling or another heat dissipation strategy as AI workloads push rack density beyond the capabilities of traditional data center environments.

LEVEL 2

Data & Integration Layer

Above the infrastructure foundation is the data layer, which is responsible for processing and protecting enterprise data across on-premises, edge, and cloud infrastructure. This keeps your data safe while making sure it's used responsibly and efficiently by AI.

A critical first step is assessing data availability, quality, and access for AI use cases. Then different data domains can be mapped to AI opportunities and risks to create a clear path from enterprise data sources to valid AI use cases. This helps organizations prioritize where to start instead of trying to operationalize every possible use case at once.

When it comes to data platforms, there's no single option that's suitable for every organization or use cases. Some environments require a data lakehouse model, while others may need a combination of a data lake, warehouse, and domain-specific data services. The right architecture depends on existing platforms, compliance requirements, and how data needs to be utilized across on-premises, cloud, and edge environments.

Data pipelines are crucial for efficiently ingesting, transforming, and streaming data for AI workloads. These pipelines may also need to support processes like feature engineering to improve machine learning model performance. An effective data pipeline should also make data available to AI platforms and integrate with existing enterprise systems, such as transactional apps, analytics platforms, SaaS tools, and more.

Finally, data governance and compliance standards should cover data residency, privacy, lineage, retention, and access policies throughout the pipeline. These controls become even more important when AI workloads span on-premises, edge, and cloud environments and data. It's also crucial to align data governance with the compliance requires for your industry — especially in heavily regulated industries like healthcare, financial services, and the public sector.

LEVEL 3

AI Platform Layer

The AI platform layer is the central hub that turns infrastructure and data into reusable AI services. Leading AI platforms can provide broader model and tooling ecosystems for building and deploying AI services as they're created. This AI hub should be a centralized workspace where builders can access compute, data, models, and tools as a self-service capability.

The platform itself should support different model types and deployment patterns. These models may be open, proprietary, or fine-tuned for internal use depending on use case and governance requirements. It's also important not to get locked into a single model development framework because AI technologies are rapidly evolving and different workloads may require unique model types, runtimes, or deployment patterns.

Generative AI and large language models (LLMs) require runtime controls, orchestration, and clear policies for how enterprise data is used. These workloads often depend on centralized platform services for model access, prompt management, and more. It's also ideal if the same platform supports traditional machine learning and deep learning workflows instead of forcing each pattern into a separate silo.

Retrieval augmented generation (RAG) is one of the most common enterprise AI patterns, which involves combining enterprise content with model inferencing to generate more reliable outputs. This requires a pipeline for document ingestion, embedding, retrieval, and LLM orchestration with clear governance over data sources and access.

Another common use case is integrating AI into code assistants and developer productivity platforms. This requires standing up AI Factories that support code generation, refactoring, and review — often in a regulated environment. These environments also need to be integrated with existing IDEs, CI/CD tools, and ticketing systems.

Agentic AI introduces a more dynamic usage model, where systems can reason through tasks, use tools, and take action across workflows. This increases the need for strong controls around orchestration, permissions, validation, and observability to minimize risk as AI systems become more capable of acting autonomously within enterprise environments. It's also still recommended for human-in-the-loop oversight to be integrated into high-impact or higher-risk workflows.

Finally, the AI Factory should support an experience layer by turning AI capabilities into reusable APIs and SDKs for product teams. This makes it easier to integrate AI into existing applications and workflows without rebuilding core capabilities for each use case.

SIDEBAR:

MLOps & Platform Engineering for AI

Surrounding all the other layers is the MLOps and observability capabilities needed to manage the full AI lifecycle. While building out a new AI use case is a great first step, running and optimizing it over time is where you can realize substantial business value.

A MLOps best practice is continuous integration and continuous delivery (CI/CD) for models, data pipelines, and prompts. By standardizing how changes are tested and deployed, organizations can incorporate model registries, experiment tracking, and artifact management to improve consistency. This creates a more repeatable and reliable process for moving AI use cases into production.

Monitoring model performance, drift, data quality, prompt performance, and user behavior can surface issues early. These insights can be fed back into the AI Factory operating model to iteratively improve models and prompts over time. Organizations can also use this data to determine which use cases are ready to scale further.

MLOps ties into platform engineering because the AI Factory benefits from shared workflows, standards, and self-service capabilities help avoid one-off delivery patterns. By treating the AI Factory as a product with clear SLOs and paved roads to success for teams, organizations can reduce friction for builders while still enforcing standards. This could include reusable templates, opinionated pipelines, and shared services that facilitate moving from experimentation to full-scale production.

Reach out to AHEAD if you're interested in learning about our platform engineering solutions for enterprise AI infrastructure.

LEVEL 4

Operating Model & Services Layer

At the very top of the technology stack is the operating model. This is what makes it possible to prioritize use cases, manage risk, and ultimately move from AI experimentation to production in a controlled, repeatable way. It's where governance frameworks live, along with intake processes and day-two operational practices, so that AI initiatives continue to be managed and optimized long term.

An ongoing operating model is just as important as the initial deployment to maximize the business value of AI. As a starting point, defining service tiers for AI workloads, such as experimentation or mission-critical, can help teams know what governance standards to apply to each use case. The operating model can also include workshops, playbooks, and AI literacy programs for builders and business stakeholders to facilitate adoption.

Governance frameworks should define roles, processes, and decision rights for responsible AI across teams and use cases. An effective AI policy framework should include guidance for acceptable usage, model approvals, data classification, and retention requirements. These guardrails make it easier to scale AI adoption safely.

When it comes to security, we strongly recommend incorporating patterns for segmenting AI environments, applying Zero Trust, and hardening data and model pipelines from the start. This secure-by-design architecture approach ensures the AI Factory minimizes risk as new use cases are rolled out at scale.

Many end-users will be unaware of the implications involved with AI, so strong runtime validation for prompts, responses, and tool usage will be critical for reducing unintended risks in production. This becomes even more important as AI capabilities are embedded into workflows and applications across the organization.

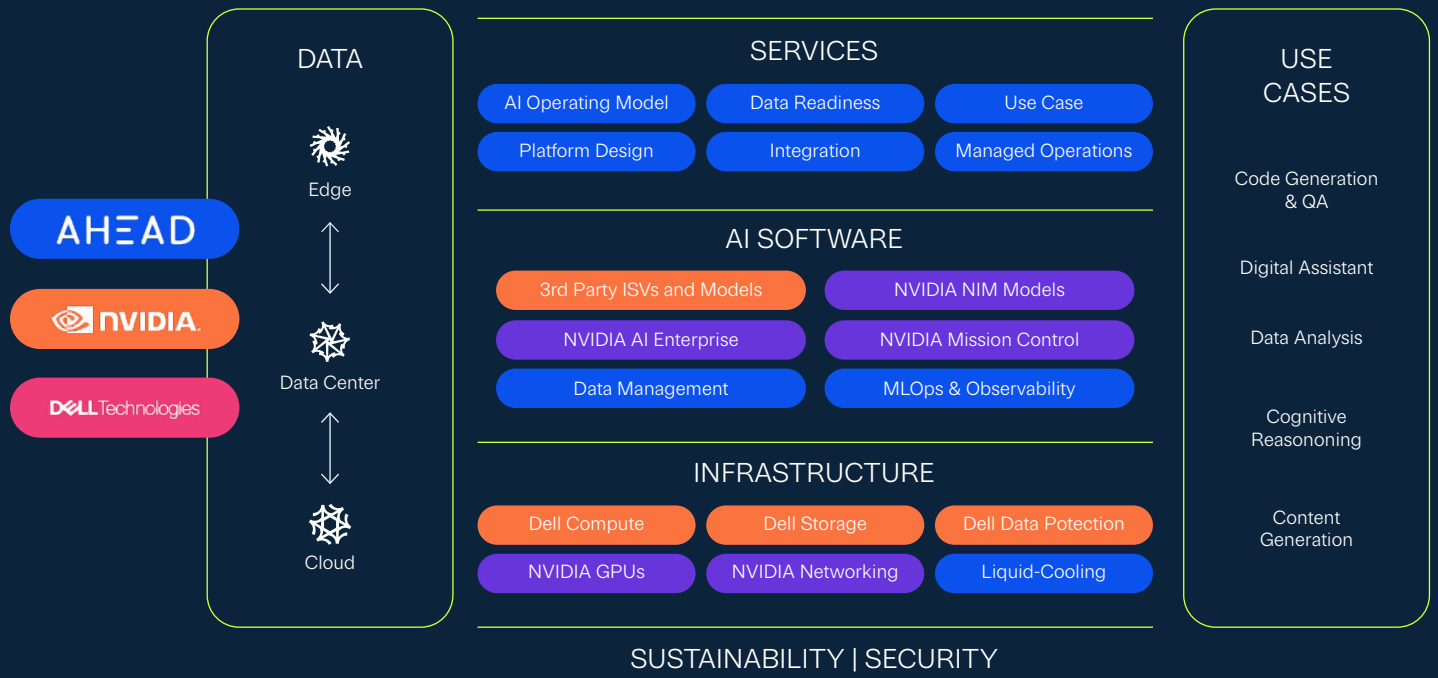
The AI threat, risk, and security management (TRisM) framework can also unify separate frameworks for security, risk management, governance, reliability, and data protection into a unified approach for AI system management. It's often challenging to coordinate responsibilities when they're fragmented across various teams and tools, so the TRisM framework is useful to minimize any gaps. This is especially true for AI Factories, where security and governance need to operate across the full lifecycle.

In short, the operating model and services layer is all about keeping the AI Factory running smoothly. This requires strong governance and security built into the platforms, processes, and operating model from the start.

SIDEBAR:

The AI Factory in Practice with AHEAD, Dell, and NVIDIA

An example of the AI Factory pattern in practice is AHEAD's work with Dell and NVIDIA to design secure and resilient AI architectures. As a reference blueprint with integrated compute, data, software, and services, the Dell AI Factory with NVIDIA helps turn isolated pilots into fully realized business solutions.



The Infrastructure Layer

Integrates Dell PowerEdge servers (including XE platforms) and NVIDIA for accelerated computing. This could then be combined with high-bandwidth networking and AI-ready storage platforms like PowerScale, PowerStore, and the Dell AI Data Platform.

The Data Layer

uses the Dell AI Data Platform to process and protect enterprise data across on-premises, edge, and cloud. This keeps enterprise data safe while making sure it's used responsibly and efficiently by AI.

The AI Platform Layer

leverages NVIDIA's AI Enterprise, NIM, and broader model and tooling ecosystem to provide a standardized software layer for training, fine-tuning, and inferencing workloads across on-premises and cloud environments.

The Operating Model Layer

includes AHEAD managed services to keep the AI Factory secure and compliant while continuing to optimize it over time. This involves stabilizing operating costs, reducing risk across the AI supply chain, and ensuring scalability as the initiative grows.

This specific technology stack is one example of how the broader AI Factory pattern can be applied in practice with innovative AI vendors like Dell and NVIDIA. AHEAD has a strong track record of deploying NVIDIA AI solutions and understands how to translate Dell's capabilities to work efficiently within different enterprise environments. We can adapt this blueprint based on the vendors that are best suited to your specific technology stack.

Reach out to us if you're interested in learning more about building your AI Factory with a Dell and NVIDIA technology stack.

Getting Started – Building Your AI Factory with AHEAD

AI Factories are a powerful way to accelerate your enterprise AI journey. By developing a standardized blueprint for operationalizing AI, your organization can continuously turn data into strategic insights and decisions. This means initial AI pilots will no longer be stuck as proof-of-concepts, and your team can deliver repeatable, production-grade AI workloads with essential governance, security, and operational controls already built in.

However, building an AI Factory takes more than just investing in and assembling the right technology stack. If it were that simple, most organizations wouldn't be stuck in a perpetual AI pilot phase. A standardized AI strategy is essential to avoid wasting time and money on fragmented pilots and one-off infrastructure decisions.

Working with an enterprise solutions provider like AHEAD can help you plan and execute an AI strategy that leverages our partnerships with leading AI innovators — including Dell, NVIDIA, Microsoft, AWS, Snowflake, Vast, and Cisco. We have the expertise to design a comprehensive blueprint model for an enterprise AI Factory that produces measurable value for your organization and programs.

AHEAD can own the AI Factory lifecycle from end to end, beginning with AI strategy, assessment, and use case discovery. Our AI readiness assessments across data, infrastructure, governance, and talent uncovers gaps, dependencies, and constraints that could slow adoption. We also offer collaborative workshops to help you prioritize use cases aligned to your business KPIs and risk posture.

From there, we can help you translate your enterprise AI strategy into a target AI Factory architecture and phased plan. This includes selecting pilot use cases that prove the AI Factory pattern, and designing the technology stack and architecture spanning data, cloud, core, and edge

environments. We ensure your AI Factory isn't just technically sound, but also completely integrated within your existing platforms and operating models.

Next, we'll execute the phased rollout — starting from an initial pilot and moving towards full-scale deployments. AHEAD Hatch, our logistics and lifecycle management platform, provides comprehensive asset and inventory tracking so that you have visibility into all related hardware across sites. From deployment through day-two operations, Hatch gives you the insights needed to keep your AI Factory performing consistently over its lifespan.

Our AHEAD Forge facility also allows us to design and integrate full AI racks — including servers, storage, networking, power, and liquid cooling — tailored to the realities of your edge, colocation, or on-premises data centers. This ensures racks arrive pre-integrated and tested, so you're not delayed by on-site assembly and troubleshooting.

After the AI Factory is live, our managed services and operations teams work to keep it secure and compliant. We will track key metrics related to productivity, risk, and total cost of ownership (TCO) to continue optimizing the architecture and operating model over time as AI capabilities mature.

If you're tired of trying to connect the dots between scattered AI pilots, AHEAD can guide you along a proven path to a trustworthy, scalable enterprise AI Factory. Reach out to us to learn more about our comprehensive enterprise AI services and solutions.

AHEAD

Combining cloud-native capabilities in software and data engineering with an unparalleled track record of modernizing infrastructure, we're uniquely positioned to help accelerate the promise of digital transformation.

Visit us at ahead.com.

National Hubs

CHICAGO
444 W. Lake Street
Suite 3000
Chicago, IL 60606

NEW YORK
500 5th Avenue
Floor 17
New York, NY 10010

ATLANTA
1117 Perimeter Center
W406
Atlanta, GA 30338

SAN FRANCISCO
2000 Crow Canyon Place
Suite 250
San Ramon, CA 94583