EVOLVING SOLUTIONS

IBM Platinum Partner

EBOOK

AI at Enterprise Scale

The Infrastructure Behind the Model

BUILDING THE OPERATIONAL FOUNDATION FOR ENTERPRISE SCALE AI

Table of Contents

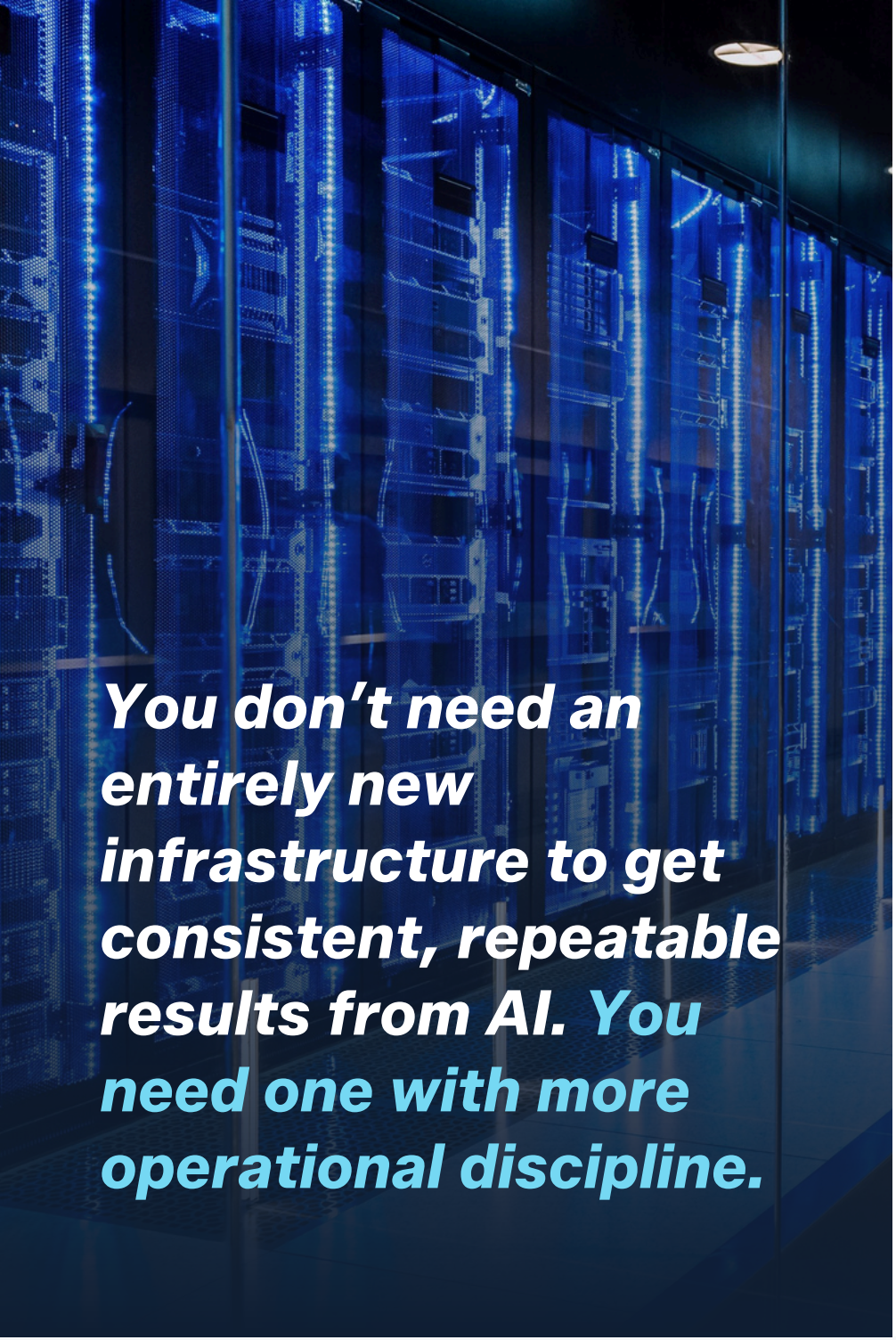
Why AI Pilots Struggle in Production

Artificial intelligence has moved from experimentation to operational expectation.

Leadership expects measurable impact. Business units expect efficiency gains. Competitors are signaling progress. For CIOs and infrastructure leaders, the mandate is clear: enable AI quickly without compromising stability, security, or cost discipline.

Inside most enterprises, adoption is already underway. Teams are running pilots. Developers are connecting to public models. Employees are experimenting with generative tools, often outside formal IT oversight.

The pattern is predictable. Proofs of concept show promise, but production exposes gaps. Infrastructure that works for isolated experimentation struggles at enterprise scale. Governance discussions begin after deployment instead of before. Automation is partial, and security



controls are reactive or simply don't exist. Progress slows, time to value stretches, and AI initiatives struggle to move beyond experimentation.

It's easy to assume the AI model is the constraint. In many cases, the real limitation is the environment where it runs.

You don't need an entirely new infrastructure to get consistent, repeatable results from AI. You need one with more operational discipline. Compute, storage, and networking must be balanced around real use cases. Container-based platforms support portability and scale. Automation and infrastructure as code (IaC) enable repeatable deployment. Governance must be designed into the operating model from day one rather than layered on after risk emerges.

AI often exposes weaknesses in how IT environments are designed and operated. Organizations with strong architectural discipline, automation, and embedded governance will accelerate. Those without it will find AI initiatives stalled not by algorithms, but by operational friction.

AI Ambition Meets Operational Reality

Experimentation Comes First

AI experimentation tends to spread quickly inside large organizations. Public models are widely accessible, development frameworks are mature, and cloud services and advanced AI coding tools make it easy for teams to prototype new applications with relatively little infrastructure investment. A small team can test an idea, build a proof of concept, and demonstrate results in days.

That speed creates momentum. Business leaders see promising early results and expect those capabilities to scale across the organization.

Production Raises the Bar

Moving from experimentation to production is a different challenge. Enterprise environments come with requirements that early pilots don't: integration with existing systems, consistent performance at scale, security controls, data governance, and ongoing operational support.

What worked in an isolated environment now has to function reliably inside a complex IT ecosystem.

The Hidden Constraint

At this stage, organizations often assume the next step is improving the model. In reality, the model is rarely the primary constraint. The real challenge is the infrastructure and operating model that support it.

AI workloads place new demands on infrastructure, automation, and governance. Environments that are loosely managed or dependent on manual processes can support experimentation, but they struggle to support consistent deployment at enterprise scale.

The real challenge is operational readiness. That readiness depends on several interconnected layers of the IT environment, from the underlying infrastructure to the platforms, automation, and governance that allow AI workloads to run reliably in production.

Understanding how those layers work together is the first step toward building infrastructure that supports AI at enterprise scale.

Infrastructure for AI is an Operating Model, not a Hardware Refresh

If the challenge of enterprise AI is operational readiness, what does that mean for the IT environment?

AI infrastructure is often discussed in terms of specialized hardware, accelerated compute, or new software tools. Those elements matter, but focusing on them exclusively misses the bigger picture. For most organizations, the real challenge is ensuring that the environment where AI runs is designed to support reliable, repeatable operation at scale.

Infrastructure for AI is not about a single platform. It's about how several layers work together. At the foundation is the physical infrastructure that provides compute, storage, and networking across on-premises and cloud environments. Above that sits the container platform layer where AI workloads run. Automation and deployment pipelines enable those workloads to move from development into production consistently. Finally, governance and security ensure that AI systems operate within the policies, controls, and risk frameworks required by the enterprise.

Organizations that focus only on hardware performance may still struggle with deployment speed.

Each layer plays a different role, but they are tightly connected. Weakness in any one of them slows the entire system. Organizations that focus only on hardware performance may still struggle with deployment speed. Those that automate infrastructure but neglect governance can create new risks. And environments with strong security but manual processes often find that moving AI workloads into production takes far longer than expected.

Understanding how these layers work together provides a practical way to think about infrastructure for AI. It also clarifies why operational discipline, not just new technology, determines whether AI initiatives scale successfully.

Layered Operating Model for AI Infrastructure

Physical infrastructure
The compute, storage, and networking foundation that provides the performance and capacity AI workloads require across on-premises and cloud environments.

Container platform
The Kubernetes-based runtime environment, such as Red Hat OpenShift, where AI applications and services are deployed, managed, and scaled consistently.

Automation and pipelines
Infrastructure as Code and deployment workflows that allow AI environments to be provisioned, updated, and reproduced reliably at scale.

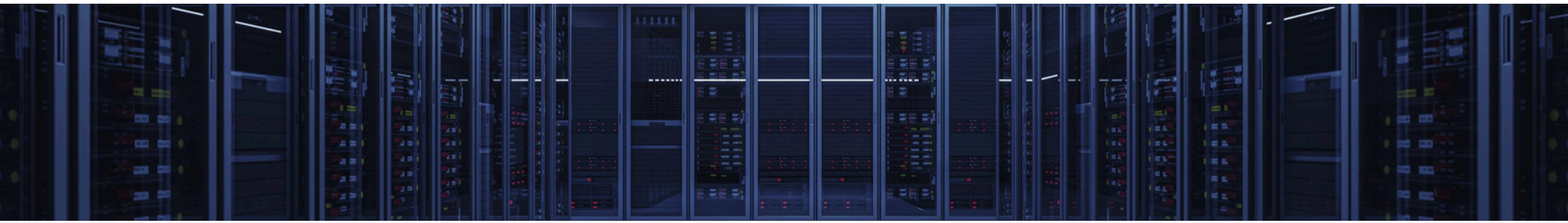
Governance and security
Policies, controls, and monitoring ensure AI systems operate safely, comply with enterprise requirements, and manage risk as they scale.

Designing the Foundation: Balance Over Brute Force

The first layer of the AI operating model is the physical infrastructure that runs AI workloads. For most enterprises, this means environments designed to support production AI applications rather than large-scale model training. That infrastructure relies on balanced compute, storage, and networking resources that enable models to process data, generate responses, and integrate with enterprise systems.

AI infrastructure discussions often start with compute. Graphics processing units and other accelerators get most of the attention, and it's easy to assume that enabling AI simply means adding more hardware. In practice, the challenge is broader. AI workloads place demands on the entire infrastructure stack. Any imbalance across compute, storage, and networking can quickly become a bottleneck.

For example, many AI workloads require large volumes of data to move quickly between storage and compute. If storage performance lags behind compute capacity, processors sit idle waiting for data. Similarly, networking constraints can slow distributed workloads that rely on rapid communication between systems. In these environments, adding more compute rarely solves the problem.



The goal is to design an environment that avoids both overbuilding expensive infrastructure and underbuilding systems that can't support production workloads.

This is why infrastructure planning for AI begins with understanding the use cases the organization intends to support. Some workloads focus on model training and require high-performance compute environments. Others center on inference, where models are deployed to generate predictions or responses for users or applications. In most enterprises, inference workloads dominate. Organizations are far more likely to deploy models to support applications, automation, and decision-making than to train large models from scratch.

Hybrid environments also play a role. Organizations frequently run AI workloads across a combination of on-premises infrastructure, cloud resources, and edge locations. Each environment contributes different strengths, and designing infrastructure that allows workloads to run in the most appropriate environment improves flexibility and cost control.

Once that physical foundation is in place, the next question becomes how those resources are organized and managed as applications move from development into production.

Containers and the Shift to Platform Thinking

With the physical infrastructure in place, the next question is how AI workloads are packaged, deployed, and managed in production. In modern environments, that responsibility falls to the platform layer.

Most modern AI applications are deployed as containerized workloads. Containers package applications and their dependencies, so workloads run consistently across different environments. For teams building and deploying AI-driven services, this consistency reduces the effort required to move applications into production.

The platform layer that manages containerized workloads has become central to enterprise IT environments. It provides the scheduling, scaling, and orchestration needed to run applications that span multiple services and systems. Rather than managing individual servers or virtual machines, teams manage applications at the

platform level, allowing workloads to be deployed and managed consistently across environments.

Most organizations implement this platform layer using an enterprise Kubernetes platform such as Red Hat OpenShift, particularly where workloads must run across multiple clusters, data centers, or cloud providers. Its adoption reflects the broader shift toward platform-based operations in enterprise IT.

This shift changes how infrastructure teams think about operations. Instead of managing infrastructure as a collection of individual systems, they manage a platform that supports applications at scale. AI workloads benefit from this model because they often involve multiple services, data pipelines, and integration points that must work together reliably.

In many ways, AI raises the bar on practices already required for modern cloud operations. Environments that already support container platforms and distributed applications are far better positioned to operationalize AI workloads.

For organizations adopting AI, the container platform becomes the operational layer that connects infrastructure to applications. It's where models are deployed, services scale to meet demand, and updates move safely into production.

Automation Determines Time to Value

A container platform provides the environment where AI workloads run. The speed and reliability with which they move into production, however, depend on automation.

In many enterprise environments, infrastructure is still deployed and managed through manual processes. Systems are configured individually; environments are built through a series of operational steps, and changes require coordination across multiple teams. These approaches can work for traditional applications, but they quickly become a constraint when organizations begin deploying AI workloads that evolve rapidly and require frequent updates.

Automation addresses this challenge by making infrastructure consistent and reproducible. Teams define infrastructure through code that can be versioned, tested, and deployed through automation in a consistent way. Infrastructure as Code (IaC) allows

environments to be created, modified, and reproduced reliably across development, testing, and production systems.

For AI initiatives, this repeatability is critical. Models are updated, data pipelines evolve, and applications integrating AI capabilities often change quickly as teams refine their use cases. Without automation, each update requires manual configuration and increases the risk of errors or configuration drift between environments.

Automation also shortens the path from development to production. Deployment pipelines allow infrastructure and applications to move through standardized workflows, which include automated testing and security checks so changes can be validated, approved, and released in a controlled manner. These processes reduce operational friction and make it easier to scale AI capabilities across the organization.

In this sense, automation becomes a key enabler of AI adoption. Organizations that rely on manual infrastructure management often find that AI initiatives stall as operational complexity grows. Those that invest in automated, code-driven infrastructure can deploy new capabilities more quickly and maintain consistent environments as their AI workloads expand.

Governance as Infrastructure

As infrastructure and deployment processes become more automated, governance becomes even more important. The same practices that allow organizations to deploy AI capabilities quickly also increase the need for clear policies, security controls, and operational oversight.

In many organizations, AI experimentation begins before governance frameworks are fully established. Applications connect to external models, teams test new capabilities, and data begins flowing through systems that were not originally designed for AI workloads. As these experiments move toward production, governance quickly becomes a critical concern.

AI systems often interact directly with enterprise data, customer information, and business processes. They may generate outputs that influence decisions, trigger automated workflows, or provide responses to users. Without appropriate governance, organizations risk exposing sensitive data, violating regulatory requirements, or losing visibility into how AI systems make decisions and access and expose enterprise data.

For this reason, governance must be treated as part of the infrastructure that supports AI. Policies and controls need to be requirements, or losing visibility into how AI systems make decisions and access and expose enterprise data. Access controls,

Well-designed governance allows teams to deploy new AI capabilities with greater confidence.

data management practices, monitoring capabilities, and runtime guardrails should be designed alongside the platforms and automation that support AI applications.

Automation can also strengthen governance when it is implemented thoughtfully. Policy as Code (PaC) allows organizations to embed policy controls directly into deployment processes. Security requirements, configuration standards, and compliance checks can be incorporated into automated workflows so that environments are created consistently and within established guidelines.

Organizations that approach governance this way create a stronger foundation for deploying AI safely. Instead of slowing innovation, well-designed governance allows teams to deploy new AI capabilities with greater confidence. AI initiatives move forward with clearer guardrails, reducing risk while supporting the operational discipline required for enterprise-scale deployment.

Governance also depends on strong security foundations. Identity management, access controls, and data protection practices must already be in place before organizations can confidently deploy AI systems. Without clear visibility into who can access data, how systems authenticate machine and human users, and how sensitive information is protected, AI initiatives can introduce new operational and regulatory risks.

Operational Outcomes of AI-Ready Infrastructure

For business and IT leaders, the impact of mature AI infrastructure shows up in three areas: speed, scale, and control. Organizations can move promising AI initiatives into production faster, scale them as business demand grows, and manage risk without slowing innovation.

Faster Path from Experimentation to Production

AI experiments often succeed because they run in isolated environments with limited operational requirements. Moving those capabilities into production can push infrastructure beyond its limits, which must support higher data volumes, greater scale, and consistent performance.

When infrastructure is automated and managed through standardization, that transition becomes significantly easier. Standardized environments allow applications and models to move from development to testing and production without being rebuilt for each stage. Deployment pipelines allow updates to move through controlled workflows rather than manual operational processes.

For business leaders, the result is shorter time to market for AI-enabled capabilities and faster iteration as use cases evolve.

Infrastructure That Scales with Demand

AI workloads rarely remain static. As organizations integrate AI into applications and workflows, usage can grow quickly. Systems must handle higher transaction volumes, more complex data pipelines, and expanding service dependencies.

Environments designed around balanced infrastructure and container-based platforms are better equipped to handle this growth. Resources can be scaled horizontally, workloads can be distributed across clusters, and applications can run across hybrid environments when additional capacity is required.

This flexibility allows organizations to support growing AI adoption without continually redesigning their infrastructure.

Resilience and Risk Management

AI systems often operate close to core business processes, interacting with enterprise data, automating decisions, or generating outputs used by employees and customers. As these systems scale, reliability and oversight become critical.

Governance, security controls, and monitoring capabilities help ensure that AI systems operate within established policies. Identity management, data protection practices, and runtime guardrails help ensure AI systems operate safely within enterprise environments.

When these capabilities are embedded into the operating environment, organizations can deploy AI capabilities with the resilience, security, and oversight required for enterprise systems.

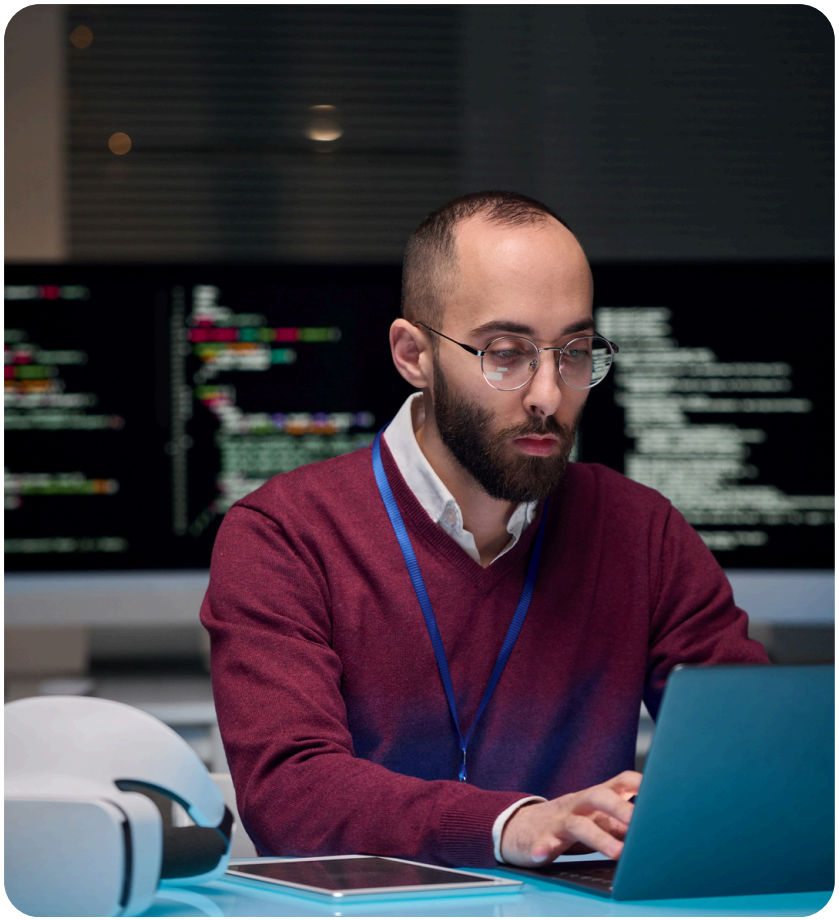
Enabling the Operating Model

The core capabilities of AI-ready infrastructure aren't new to many organizations. Container platforms, automation frameworks, and hybrid infrastructure that support modern applications are already common. What many organizations need is greater consistency in how these capabilities are implemented, so AI workloads can operate reliably across the enterprise.

Implementing these capabilities typically involves a combination of infrastructure automation, service networking, and container platforms.

Terraform for Infrastructure as Code

Infrastructure automation often begins with Infrastructure as Code (IaC). Tools such as Terraform allow teams to define infrastructure configurations in code so environments can be created, updated, and reproduced consistently. This approach reduces manual configuration and helps ensure that development, testing, and production environments remain aligned.



HashiCorp Consul for Service Networking

Service networking becomes increasingly important as AI applications scale. Platforms such as HashiCorp Consul provide service discovery, secure service-to-service communication, and network visibility across complex environments. These capabilities help organizations manage the growing number of services, APIs, and data pipelines that support AI-enabled applications.

IBM Fusion for Standardized Deployment

Organizations also benefit from platforms designed to simplify the deployment and management of containerized workloads across hybrid environments. Solutions such as IBM Fusion provide a unified platform for deploying and managing container-based applications while integrating security, data services, and operational controls. By standardizing how applications are deployed and managed, platforms like these help infrastructure teams maintain consistency as AI workloads move from development into production.

Together, these tools support the broader operating model required for enterprise AI. They help organizations automate infrastructure, manage distributed services, and operate container-based platforms with the operational discipline required for enterprise AI deployments.

Turning AI Ambition into Operational Capability

For many organizations, the challenge of AI adoption isn't hidden in the model itself, but the environment surrounding it. To operate at enterprise scale, infrastructure must support new operational and performance demands. Achieving that level of readiness requires more than fast processors. It requires aligning infrastructure, platforms, automation, and governance into a consistent operating model.

Evolving Solutions helps organizations build that foundation. Through architecture guidance, platform implementation, and operational best practices, Evolving Solutions works with IT teams to modernize infrastructure and prepare environments to support enterprise-scale AI workloads. From container platforms and automation frameworks to governance and security integration, Evolving Solutions helps organizations move AI initiatives from experimentation to reliable production use.

With the right operational foundation in place, AI becomes more than a promising experiment. It becomes a practical capability that supports innovation, efficiency, and better business outcomes.

Contact Evolving Solutions today to learn how your organization can get on the path to AI-ready infrastructure.

LET'S GET TO WORK!

evolvingsol.com