November 2024 | Infographic sponsored by: Red Hat | intel

AI-Ready Infrastructure, The Bedrock of AI-Fueled Business

Building the digital infrastructure that enables the pivot from generative AI (GenAI) experimentation to AI at scale

Lawrence Cheok
Senior Research Analyst, IDC
AI Infrastructure Strategies, AI Healthcare

Critical Focus Areas to Enable Broader AI Adoption in 2025

Platforms: facilitate improved data use, model, and app development and deployment with a unified AI platform.

Infrastructure: build the digital foundation that supports enterprise AI-enabled capabilities.

Data: enable value-driven use of entire information estate comprising structured, unstructured and event streams.

% of Organizations Recognizing the Importance of an AI-Ready Digital Infrastructure

Source: Future Enterprise Resiliency & Spending Survey, Wave 1, IDC, July 2024

What is an AI-Ready Infrastructure

An AI-ready infrastructure enables the use of interoperable, fit-for-purpose infrastructure to support the workload cost and scale while ensuring consistent governance, compliance, and secured management of AI models and data. This is achieved by leveraging open hybrid cloud standards and architectures to ensure maximum interoperability across infrastructure components such as compute, storage, network, and across diverse environments, whether on-premise, in the cloud, or at the edge.

Essential AI-Ready Infrastructure Capabilities

83% Optimize resources by leasing and deploying AI workload in a way that balances cost and performance.

80% Leverage open standards and software-defined innovation.

78% Integrate their company's security posture by ensuring consistent AI management and governance policies for AI models, data, and applications.

68% Provides the scalability their business needs to handle fluctuating computational AI workloads efficiently across on-premise, public cloud, and edge locations.

Source: IDC AI Infrastructure and AI Strategies Survey, 2024, N=1 = 850

Train And Deploy Fit-For-Purpose AI To Balance Cost And Performance

AI ready that are infrastructure offers the compute, storage, and network flexibility that are needed to train and deploy a diverse range of AI models in accordance with specific business needs. The fit-for-purpose approach optimizes resource utilization based on AI use cases and their associated workload types.

AI Workload Types

Building custom models	Fine-tuning open sourced/third-party models	Contextualizing 'off-the-shelf' AI applications	Consuming AI-infused applications 'as-is'
17%	22%	26%	35%
Built from scratch with large data sets, highly tailored.	Adjust pre-trained models using first-party data.	Empowers models with first-party data, prompts, and retrieval-augmented generation (RAG) operations.	Consumed as-is with little reliance on first-party data.

High → Low

Impact on Infrastructure

Demanding high GPU/FPGA power, flash data storage, and high-performance network fabrics to handle vast amounts of structured and unstructured data.

Fine-tuning to adapt to a specific data set that involves moderate computing resources, particularly for data preprocessing and vector database indexing for RAG operations.

Model is consumed as-a-service, with the compute and storage handled by the SaaS provider.

Note: Retrieval augmented generation (RAG) is a technique that combines external data sources (such as databases, documents, or real-time web content) with generative models to provide contextually relevant information, improving the accuracy and quality of the generated output.

Source: IDC AI Infrastructure and AI Strategies Survey, 2024, N=1 = 850

Ensure Consistent AI Management and Governance

AI-ready infrastructure automates infrastructure management to enable scalable, consistent, and secured governance AI deployments across hybrid, multicloud, and edge environments.

Critical Governance Measures For AI-at-Scale

01 Architect for hybrid cloud governance

02 Monitor and optimize model performance

03 Enable secured data sharing practices

04 Create an internal AI CoE to accelerate adoption

05 Establish guidelines for use of open-source GenAI/AI

An AI-ready infrastructure allows my company to

- Establish a secure and scalable foundation for AI across hybrid environments
- Optimize AI cost versus performance
- Provide consistent data protection across distributed systems
- Enforce consistent, responsible AI policies
- Ensure continuous model governance, regardless of deployment choice

Source: IDC AI Infrastructure and AI Strategies Survey, 2024, N=1 = 850

Best-fit Environments for AI Workload

Enterprises that adopt open hybrid cloud standards gain the flexibility to deploy AI data, models, and applications based on a range of differing performance, cost, security, and scalability needs. An AI-ready infrastructure does this by ensuring consistent policies and measures across heterogeneous environments that are integrated and managed via open standards.

Top AI Workload Deployment Criteria (AI Scalers)

01 Infrastructure costs

02 Cybersecurity

03 Access to latest processors/GPUs

04 Network/data transfer costs

05 AIOps and autonomous capabilities

An AI-ready infrastructure allows my company to...

- Optimize cost while scaling AI workloads
- Protect AI models and data from security breaches
- Leverage cutting-edge hardware for faster AI processing
- Utilize AI-at-edge to minimize network and data transfer costs
- Automate AI infrastructure management across hybrid cloud environments

Source: IDC AI Infrastructure and AI Strategies Survey, 2024, N=1 = 850

Bring AI to the Data for Scalability and Performance

AI-ready infrastructure enables location agnostication to deployed AI workloads across different environments based on latency, security, and performance requirements, with edge locations enabling real-time data processing for critical insights.

AI Inference Workload By Location

Edge location	Private cloud	On-premise	Public cloud
17%	27%	29%	30%
Low latency for real-time use cases.	Enforce consistent governance for sensitive data while benefiting from the cloud's scalability.	Ensure high performance infrastructure under tight security and compliance controls.	Provision for large-scale AI tasks requiring significant scalability.

Benefits of Location Choice

Edge Use Cases by Industry

Telecommunication: 5G edge computing
Deploy 5G networks to enable real-time applications (AR/VR, IoT), improving performance by reducing latency at the edge.

Manufacturing: Predictive maintenance
Process sensor data in real time to predict equipment failure, reduce downtime, and schedule maintenance efficiently.

Healthcare: Remote patient monitoring
Process patients using wearable devices that process data locally, triggering real-time alerts without constant connectivity to central servers.

Retail: Personalized CX
Analyze customer behavior in real time to personalize promotions and enhance in-store experiences.

Source: IDC AI Infrastructure and AI Strategies Survey, 2024, N=1 = 850

Pivoting to AI-at-Scale

Organizations that leverage open hybrid cloud architectures can accelerate AI development by taking the following actions:

Use AI-Ready Infrastructure to Train and Deploy Fit-For-Purpose AI Models
Adopt an open hybrid cloud architecture to enable the training and deployment of fit-for-purpose models based on the needs of specific business requirements and use cases.

Ensure Consistent Hybrid Cloud Management and Governance
Automate the deployment and management of AI models and infrastructure to ensure consistent enforcement of AI policies and security measures across edge, on-premises, and cloud environments.

Best-Fit Environments for AI Workloads
Establish hybrid and multicloud environments with a centralized plane across the IT infrastructure to develop, monitor, and manage AI workloads across heterogeneous hybrid cloud environments.

Bring AI to the Data for Optimal Performance
Enhance AI performance by processing data at the edge and its potential benefits lowering latency and enabling real-time decision-making.

Message from the Sponsor

Red Hat | intel

A robust, efficient and flexible AI infrastructure will be built on 5 strategic pillars:

Pillar 1: Bring AI to your Data, not your Data to AI
By deploying AI across a mix of cloud and on-premises environments, we ensure compliance and control, unlocking data's full potential.

Pillar 2: Optimize Your AI, Minimize Complexity
Hybrid AI strategies and the false dichotomy between performance and cost. Deploy AI where it yields the highest ROI.

Pillar 3: Future-Proofing Innovation
Embracing open source frameworks ensures that your AI strategy is as dynamic and adaptable as the technological landscape itself.

Pillar 4: Unify AI and App Development
Unify AI and app development to foster an environment where creativity and technology intersect for intelligent application development.

Pillar 5: Build Trust with Transparent AI Governance
AI solutions will only be trusted by your customers if the AI lifecycle is rooted in transparency, security and accountability at every step.

Click Here to Find Out More on Why Choose Red Hat for AI!

Presented by IDC: Carmen Galasso
IDC Asia/Pacific
Copyright notice: External Use: Minimum 10 point font size required. Contact IDC Asia/Pacific for permission if you plan to use this material in a presentation, press release, or promotional piece.
Exceptions: IDC support Global IDC Corporate Office terms of a Multinational Licence Agreement.
Trademarks: Red Hat is a trademark of Red Hat, Inc. registered in the U.S. and other countries.

IDC Asia
Singapore
T: +65
F: +65