IDC eBook

AI-Ready Infrastructure: Public AI, Private AI... or Both?

Choosing your AI Infrastructure

The disruptive effect of generative AI (GenAI) as a major new workload requiring adoption of AI-ready infrastructure is entering a critical phase: extended buildout.

Encompassing many aspects of deployment including public and private AI, on premises and off, core and edge, hybrid architectures and enterprise size, public AI and private AI provide a suitable framework for considering one's choices for AI-ready infrastructure.

Rapid adoption is moving AI from an emerging software segment to a critical technology at the center of a platform transition and a new era of computing.

Starting in 2024, enterprises will accelerate deployment of new AI-ready hardware and software infrastructure as they invest to drive meaningful gains in business and staff productivity as well as reimagine customer digital experiences.

IDC predicts that by 2025, 70% of enterprises will form strategic ties with cloud providers for GenAI platforms, developer tools, and infrastructure, requiring new corporate controls for data and cost governance.

Source: IDC Perspective: Key Considerations for Making AI-Ready Infrastructure Decisions, IDC #US51907424, March 2024

What is Public AI?

Public AI is the use of cloud provider AI frameworks and relevant cloud PaaS and network services capabilities by enterprise IT practitioners for actioning enterprise-specific AI/ML workflows.

When considering using public AI, tech leaders must determine if public AI-ready infrastructure is right for their organization based on numerous factors:

- **Scale and scope of AI projects.** Public AI infrastructure, offered by leading cloud service providers, is advantageous for projects with fluctuating computational demands or those requiring access to extensive computing resources across geographies.
- **Budget and long-term cost-effectiveness of public AI infrastructure compared with setting up, maintaining, and upgrading private infrastructure.** Public cloud services often offer pay-as-you-go models, enabling cost optimization and flexibility.
- **Level of data sensitivity and regulatory compliance requirements.** While public cloud providers offer various cybertools and security measures, certain industries or organizations with sensitive data may prefer on-premises solutions for greater control and compliance adherence.
- **Expertise and resources available within your organization**. Leveraging public AI infrastructure can reduce the need for in-house IT expertise and infrastructure management, allowing teams to focus more on AI model development and innovation.

Key Next Steps in Selecting Public AI

Upon selecting public AI solutions, several additional considerations must be carefully evaluated to ensure ongoing alignment with organizational goals and objectives. Assessing the provider's AI infrastructure and platform offerings in terms of scalability, interpretability, and integration with existing workloads is crucial. These will determine the required level of interoperability with existing tools. Focus on:

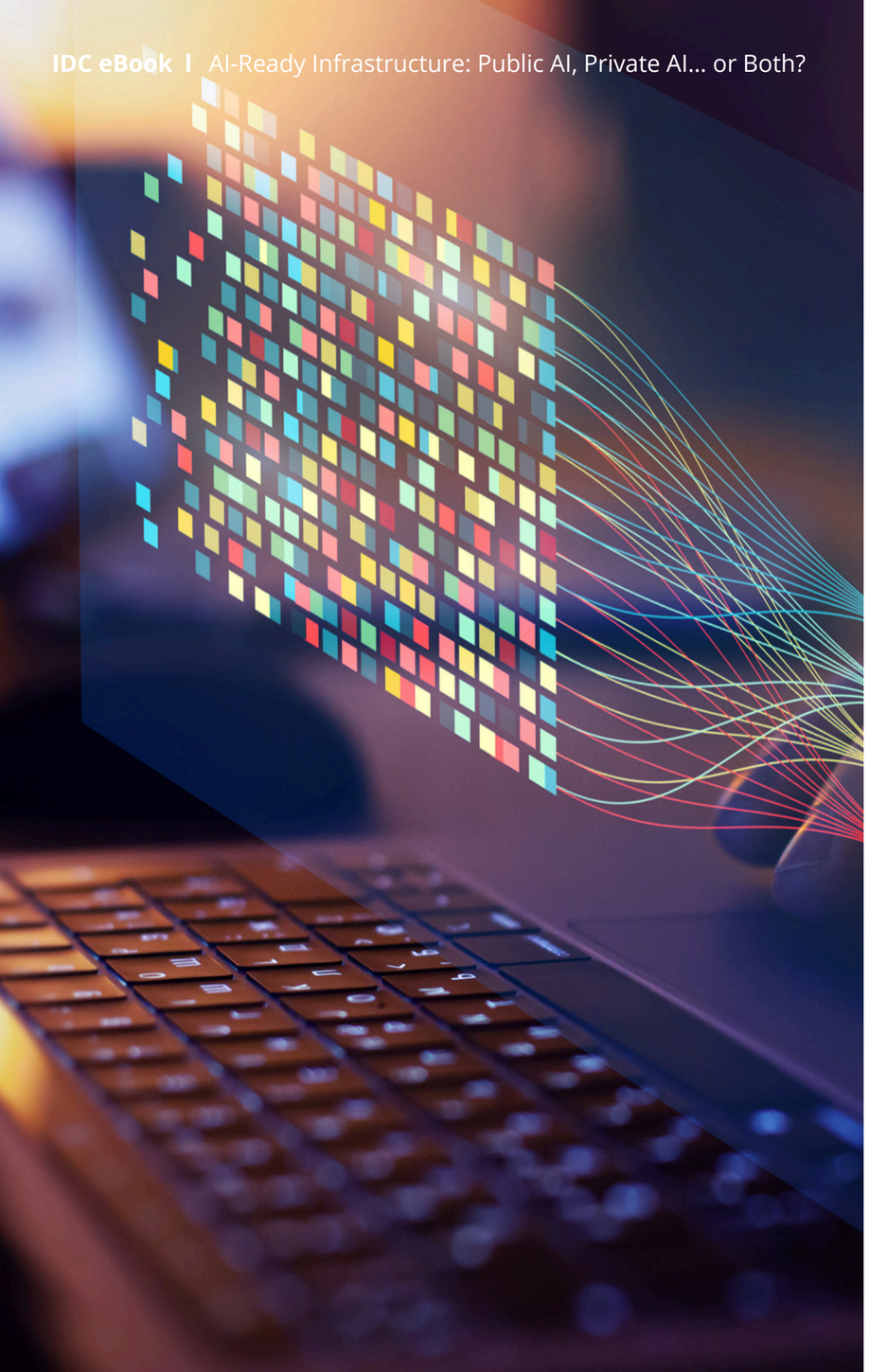
- **Access to expertise:** Choosing an AI platform is not just a matter of deciding which provider has the best foundational model that is currently leapfrogging the others. Most enterprises do not have comprehensive AI tools implemented in production.
- **Community and support:** Speed is essential for enterprises that are in the early stage of choosing an AI strategy or product. Choosing public AI often fosters a larger community of users, documentation and supporting channels where ideas can be exchanged and/or projects can be collaborated.
- **Global reach:** Public AI is usually hosted on cloud platforms running in hyperscale datacenters distributed worldwide across a growing range of geolocations. Enterprises with a presence across many different regions can access AI resources with low latency and high reliability.
- **Complementary infrastructure services:** With leading cloud providers, enterprises can gain access to a wider range of AI services and tools that complement GenAI-specific infrastructure workloads. Focus on the level of connection with data storage, analytics, and developer tools that enable quick deployment and link to existing pretrained models and services.

What is Private AI?

Private AI is the use of enterprise datacenter infrastructure and AI framework capabilities by enterprise IT practitioners for actioning enterprise-specific AI/ML workflows. The datacenter assets could be hosted at interconnect providers or in enterprise-owned facilities. When using private AI, enterprises will also need to look for AI platforms that support hybrid deployment options that allow them to govern model development and use.

When considering private AI, understand the perception that private infrastructure is thought to be safer. But, safety is not the sole reason organizations are considering private AI. Control over data residency, compliance requirements, and the ability to customize security protocols are equally significant. In addition, some organizations prioritize private infrastructure due to specific use cases such as research and development of proprietary algorithms or handling highly classified information that demand absolute control and isolation.

Investing in private AI-ready infrastructure comes with its own set of challenges. These include higher initial costs, the need for highly skilled maintenance and management personnel and, potentially, a much slower ability to scale compared with public AI solutions. Therefore, while sovereignty and perceived safety are key considerations, organizations must weigh these factors against the time/cost benefits and trade-offs of private versus public AI infrastructure based on specific uses cases and long-term priorities.



Key Next Steps in Selecting Private AI

Upon selecting private AI solutions, several considerations should be carefully evaluated to ensure alignment with organizational goals and objectives. A private AI cloud environment will offer greater control over governance, increasing confidence in the ability to be in compliance with regulatory bodies and enabling higher customization of infrastructure configurations to meet specific AI workload requirements. Focus on:

- **Competitive advantage:** Having control over an AI Infrastructure can provide a competitive advantage by enabling faster innovation and better integration with existing systems, allowing the ability to leverage proprietary data more effectively. This can lead to improved decision-making and product development thus enabling better strategic decision-making.
- **Customization and flexibility:** Having a private environment offers the flexibility to customize AI models and algorithms to suit the organization's unique needs. This can be in fine-tuning of hardware specifics, networking configurations, and/or security policies tailored to an organization's needs.
- **Data sovereignty**: The need for greater control and ownership over data and infrastructure is a major driver for private AI infrastructure adoption. While public cloud offerings provide robust security measures, some organizations, particularly those in highly regulated industries or with sensitive data, opt for private AI infrastructure to maintain sovereignty and ensure compliance with data protection regulations.

According to IDC's Future Enterprise Resiliency and Spending (FERS) Survey, Wave 11, only one-third of enterprises are investing significantly in GenAI, and another third are very much in piloting phase. Partnering with a provider of choice can help guide and shape the AI road map for the organization.

Source: IDC Perspective: Key Considerations for Making AI-Ready Infrastructure Decisions, IDC #US51907424, March 2024

Public, Private or Both?

Public and private AI are not exclusive. Each encompasses choices of location and technologies that must be guided by the needs of the end workload. **All considerations for public or private AI must be applied flexibly** and adapted according to workload needs; public and private, on premises and off premises, and core and edge are all choices, and IDC expects most enterprises to adopt a mixture of hybrid solutions.

Public AI infrastructure depends on factors such as scale and score pf AI projects, budget and cost-effectiveness, the level of data sensitivity and regulatory compliance requirements and the expertise and resources available within one's organization.

Private AI infrastructure requires higher initial costs, the needs for skilled personnel for maintenance and management and potentially slower scalability compared with public cloud solutions.

Whether deploying infrastructure for public or private AI, enterprises should be guided by the required balance among AI building blocks: compute (processing) memory, storage, security and networking technologies required by their major workloads.

AI Building Blocks

Compute	Storage	Security	Networking
While both core and edge compute infrastructure play crucial roles in AI, they have distinct characteristics due to their differing purposes and environments.	The type of storage used will depend on performance requirements and the nature of the data. Object storage has become synonymous with cloud-native applications that rely on massive amounts of unstructured data, performance limitations impact the time needed for training large language models.	The exponential growth of AI-ready infrastructure brings with it significant data protection challenges. Unlike traditional IT systems, AI often deals with highly sensitive data. Protecting this data is crucial for ethical and legal reasons.	Networking is an integral part of all cost- and performance-efficient private AI and public AI solutions. While compute technologies receive much attention and investment in the initial phases of AI infrastructure development, networking is critical for data movement that minimizes lag time within and between infrastructure systems.

IDC's Recommendations

Enterprises must pay special attention to major changes in how emerging technologies will be deployed in infrastructure, how technology is designed/bundled and delivered based on infrastructure location, and how the fundamental architectures of the systems deploying are altered based on use of new technologies. The choice between public AI or private AI infrastructure will be based on specific needs and priorities.

Most midmarket enterprises are likely to leverage cloud-based AI platforms offering open source and proprietary foundation models, while large enterprises have a stronger preference for use of a combination of cloud foundation models and pretrained on-premises GenAI models.

Interested in learning more?

Program Subscribers can learn more about how to evaluate the choices, challenges, and trade-offs organizations must consider as they make AI-ready infrastructure decisions, by looking at IDC's new Perspective, "Key Considerations for Making AI-Ready Infrastructure Decisions".

If you would like to find out more about how IDC's data and expertise around AI, contact us today.

Contact Us

May 2024

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets.

IDC is a wholly-owned subsidiary of International Data Group (IDG, Inc.), the world's leading tech media, data and marketing services company, and has been recognized Analyst Firm of the Year by the Institute of Industry Analyst Relations.

Today, our 1,300 global analysts publish thousands of reports annually in over 500+ markets that include global, regional, and local expertise on technology and industry opportunities, helping Technology Leader professionals and business executives make fact-based decisions.

IDC