SPOTLIGHT

Sponsored by: VMware by Broadcom

Some 39% of organizations are investing in infrastructure for generative AI behind their firewall as their data is too sensitive to be used for training or customizing outside the datacenter and because they cannot move proprietary data sets into a cloud for AI training.

For Generative AI, Private Data Is the Differentiator But Poses Security Concerns

January 2024

Written by: Peter Rutten, Research Vice President, Performance-Intensive Computing, Worldwide Infrastructure Research

Introduction

The global business sector has been abuzz with the latest round of AI developments, broadly characterized as generative AI (GenAI). A recent IDC survey shows that this technology is already in use for dozens of use cases, with the top 5 being the following:

- Application development (34%)
- Content writing and marketing communications (32%)
- Chatbots and customer service (31%)
- Code generation and documentation (31%)
- Product development (30%)

Also emerging are multiple use cases in research and development (R&D). In R&D, GenAI is expected to design chips, develop new materials with innovative properties, and discover new classes of drugs. These latter examples are essentially high-performance computing (HPC) use cases in which GenAI is used to enable an alternative, sometimes faster and more promising, discovery strategy, driving a potentially lucrative next innovation round.

AT A GLANCE

KEY STAT

More than a third of organizations are investing in infrastructure for GenAI behind their firewall as the best option because of data sensitivity.

WHAT'S IMPORTANT

Organizations train or customize generative AI models on premises on secure AI infrastructure stacks because they have the following requirements:

- High data privacy requirements
- A consistent AI pipeline of initiatives
- IT Infrastructure skills for AI
- Suitable datacenters for AI
- High AI model accuracy needs
- AI scaling needs that will prove too expensive in a public cloud

The productivity-improving and revenue-generating opportunities promised by this technology stretch across an organization — from marketing to manufacturing and from sales to software development. GenAI is driven by a collection of artificial neural networks called large language models (LLMs). The potential future annual economic value of embedding LLMs into new products and processes is conservatively measured in the hundreds of billions of dollars. In finance, GenAI can unearth investment insights; in manufacturing, it can design products; and in media, it can create characters, scenes, and stories. Across all industries, GenAI can serve as the underlying technology for content and insights-generating virtual assistants to humans, leveraging an organization's proprietary information as well as public domain knowledge bases.

While public cloud is a viable deployment strategy for some GenAI projects, IDC has found that 39% of organizations are investing in infrastructure for GenAI on premises as the best option. Their data may be too sensitive to be used for training outside the datacenter. Financial institutions, healthcare organizations, defense contractors, and governmental departments are all entities that cannot move large proprietary data sets into a cloud to train or customize a GenAI model. In addition, they may not be able to move the applications that have been augmented with GenAI to the cloud for inferencing. National and international regulatory limitations will also play a role here.

Foundation LLMs are trained in performance-intensive computing (PIC) using very large volumes of data. PIC is an increasingly common category of computer, storage, and networking infrastructure that — to varying degrees —is different from the general-purpose infrastructure typically used for workloads such as web serving or enterprise resource management. However, many businesses will not be building LLMs from scratch. They can use standard servers with one or two GPUs for fine-tuning LLMs or customizing them using retrieval-augmented generation (RAG), a technique for improving the accuracy and reliability of GenAI models with data pulled from external sources. IDC research has found that this is the preferred approach for leveraging GenAI for 45% of organizations.

An AI-specific infrastructure software stack is recommended, with orchestration, scheduling, containerization, software libraries, open source frameworks, and software development kits, as well as a robust security and privacy wrapper to isolate and protect proprietary or sensitive data and ensure regulatory compliance.

IDC sees four key factors for the GenAI infrastructure stack that will determine how successful an initiative to train, customize, or fine-tune an LLM model will be:

- **Performance:** The larger the model (in terms of parameters) and the data volumes to train on, the more comprehensive and multifunctional the model will be. However, more infrastructure will be required to train it.
- **Speed:** The faster AI engineers can complete the model training and push the model to production, the sooner ROI can be achieved with the newly developed functionality. Additional infrastructure will be needed to train it.
- **Accuracy:** The higher the model's expected accuracy rate, the better it will serve its purpose and the greater end-user satisfaction rates will be, but again more infrastructure is required to train it.
- **Privacy:** The more private (sensitive or proprietary) the model or the data is, the more differentiating the model will be for the organization. Privacy becomes paramount.

In summary, time to insight, accuracy, model performance, and privacy are conspiring to make developing LLMs difficult and resource intensive.

Public Cloud or On Premises?

Public Cloud

Some might argue that the public cloud can provide relief, and this is true under certain circumstances.

Organizations might opt for training, customizing, or fine-tuning LLMs in the public cloud if they have some combination of the following factors:

- **On and off AI initiatives.** The organization is not managing a consistent pipeline of AI initiatives, so these projects come and go on an as-needed basis.
- **Limited IT skill sets.** The IT staff is not trained to manage PIC infrastructure, especially not the types of clusters that larger GenAI projects require.
- **Limited datacenter floor space, power, and cooling.** PIC can be demanding, requiring extra racks in a differently cooled part of the datacenter and with higher wattage power supplies.
- **No ability to keep utilization rates high.** The amount of GPU capacity that is used is on average less than 50%; investing in expensive GPUs requires enough anticipated load to keep them busy.
- **Preferred vendor that only provides capex models (rare these days).** The preferred vendor only sells or leases its infrastructure solutions and does not provide consumption-based pricing.
- **Fewer model iterations.** The model performance and accuracy expectations are such that fewer training iterations are needed until the model is deemed completed.
- **Less need for scaling.** Once the model is in production, the volume of inferences on it is expected to be low, meaning that there will be no huge end-user demand for the model that would cause an unforeseen cloud scaling need and a subsequent massive cloud bill.

On Premises

Conversely, the on-premises datacenters can be more suitable for training, customizing, or fine-tuning a GenAI model. According to IDC research, 39% of organizations that train AI models from scratch do so on premises, while 30% customize or fine-tune a pretrained model on premises.

These organizations have:

- Ongoing AI initiatives from a team that is sustaining a consistent AI pipeline
- IT infrastructure skills to manage the systems where training/customization takes place
- Suitable datacenter floorspace, power, and cooling capabilities
- The ability to keep utilization rates high (keep expensive coprocessors busy)
- A vendor that can provide consumption-based pricing
- Accuracy needs that require many model training iterations
- Training and inferencing scaling needs that could get expensive in the cloud
- The need to keep their training or customization data private and within known physical boundaries

The Data Privacy Dilemma

An additional factor to consider is the nature of the data used to train the generative AI model, and this may well be the most important consideration. Using proprietary or core enterprise data can improve the differentiating value of a GenAI model once it has been taken into production. Such a model will provide more targeted functionality than a public model that is accessed via an API, or one that has been licensed and used as is, or a model that is customized or fine-tuned using public data.

To offer a simple example: If an aerospace manufacturer wants to create a GenAI model that can design components for a soon-to-be-developed airplane, then that model will need to be trained on the manufacturer's thousands of requirements, specifications, and designs, not on public domain data. However, proprietary or core enterprise data is usually highly sensitive and/or subject to regulatory compliance.

Using such data to train, customize, or fine-tune a GenAI model in the public cloud means that the data could be incorporated into that model with all the associated IP, privacy, and compliance risks of cloud; therefore, using only data that can safely be moved, copied, and mixed with public data would be recommended.

But training a GenAI model on premises is not without its own risk, especially if it is done with data that is highly sensitive or is subject to compliance requirements and/or is proprietary. Just as in the cloud, AI training data sets in the datacenter are being moved, copied, prepared, cleaned, mixed with other data sets, stored in tier 1 storage for training, and then maybe archived in tier 2 or 3. All these actions on the data must be performed with a tight security envelope.

A Comprehensive Private and Secure AI Stack

As stated previously, more than a third of organizations are investing in infrastructure for GenAI behind their firewall as the best option because of data sensitivity. What this means is that they need an AI infrastructure stack with robust features to maintain the privacy of data. Putting all the elements of such a stack together is not trivial, and doing so in an optimized fashion is especially hard. Vendors, therefore, have started offering comprehensive AI stacks usually in partnerships with other vendors in the AI ecosystem (e.g., processor and coprocessor vendors, server vendors, storage vendors, providers of abstraction layers such as virtualization and containerization, and AI model providers). They aim to keep these stacks open source-based as much as possible, but vendors sometimes add their own flavors of solutions where open source leaves gaps, especially with regard to security.

There's no need for an IT organization to reinvent the wheel and build an AI stack from scratch, as this can be time-consuming and cost prohibitive. Especially for GenAI, there are options to take advantage of comprehensively designed stacks that are available either as fully delivered solutions or as reference architectures.

Benefits of a Private and Secure AI Stack

For an organization, a private and secure AI stack offers the following advantages.

Vendors have started offering comprehensive AI stacks usually in partnerships with other vendors in the AI ecosystem.

- It usually will include all required layers (processors, accelerators, operating environments, storage protocols, virtualization layers, development languages, AI frameworks).
- The stack is built as much as possible on open source solutions.
- It can be built around server and storage hardware with consumption-based pricing (opex model).
- The stack can be designed as a private cloud and/or as a multicloud.
- It is optimized for performance by the vendor and other partners in the stack's ecosystem.
- Pretrained models that reduce AI training time can be included.
- It often has a security layer to protect the AI data.

Considering VMware Private AI

In August 2023, VMware introduced VMware Private AI. VMware Private AI is an architectural approach for AI services that enables privacy and control of corporate data, choice of open source and commercial AI solutions, quick time to value, and integrated security and management. VMware Private AI is designed to help organizations leverage GenAI to eliminate redundant tasks and build intelligent process improvement mechanisms.

With VMware Private AI, enterprises get the flexibility to run a range of AI solutions for their environment. VMware Private AI is built on VMware Cloud Foundation and helps organizations achieve the required AI model performance with VMware vSphere and VMware Cloud Foundation GPU integrations. Customers benefit from the fact that VMware has built partnerships with the leading AI providers for its solutions.

VMware Private AI Foundation with NVIDIA

VMware and NVIDIA also announced plans to collaborate to develop a fully integrated GenAI platform called VMware Private AI Foundation with NVIDIA. This platform will enable enterprises to fine-tune LLM models and run inference workloads in their datacenters, addressing privacy, choice, cost, performance, and compliance concerns. The platform will include the NVIDIA NeMo framework, NVIDIA LLMs, and other community models, such as Hugging Face models, running on VMware Cloud Foundation.

VMware's extensive network of partners includes industry leaders such as NVIDIA and Intel and major server OEMs such as Dell, HPE, and Lenovo. VMware also collaborated with various AI vendors and MLOps providers, including Anyscale, Run:ai, and Domino Data Lab, as well as with global systems integrators HCL and Wipro to deliver its solutions to organizations.

VMware Private AI Reference Architecture

VMware has collaborated with NVIDIA, Hugging Face, Ray, PyTorch, and Kubeflow to provide validated reference architectures for building AI models based on open source technology and taking them into production. The key components are:

- **VMware Cloud Foundation** is a turnkey platform for multicloud and modern applications that provides the infrastructure layer on which generative AI models can be deployed.
- **NeMo framework** is one of the key components of this architecture and a part of NVIDIA AI Enterprise, an end-to-end, cloud-native enterprise framework to build, fine-tune, and deploy GenAI models with billions of parameters. NeMo offers a choice of several customization techniques and is optimized for at-scale inference of large-scale models for language and image applications, with multi-GPU and multinode configurations.
- **Major server OEM support** is an architecture supported by major server OEMs such as Dell, Lenovo, and HPE:
  - **Dell:** Collaborative engineering from Dell, VMware, and NVIDIA has resulted in a full stack for GenAI. Dell solutions and professional services help organizations create value with their data faster without losing control of their intellectual property.
  - **HPE:** The HPE AI Inference solution integrates VMware Private AI Foundation and NVIDIA AI Enterprise software suite with the new HPE ProLiant Gen11 systems to streamline AI deployment and accelerate organizations' AI initiatives.
  - **Lenovo:** Leveraging the Lenovo ThinkSystem SR675 V3, VMware Private AI Foundation, and NVIDIA AI Enterprise, Lenovo's newest reference design for GenAI shows businesses how to deploy and commercialize powerful GenAI tools and foundation models using pre-validated, fully integrated, and performance-optimized solutions for enterprise datacenters.

VMware Private AI Ecosystem Expansion

In November 2023, VMware announced further expansion to the VMware Private AI ecosystem with two partners:

- **VMware Private AI with Intel**: VMware and Intel will help organizations build and deploy private and secure AI models running on VMware Cloud Foundation and boost AI performance with Intel's AI software suite, 4th Generation Intel Xeon Scalable Processors with built-in accelerators, and Intel Max Series GPUs.
- **VMware Private AI with IBM**: VMware and IBM are building on VMware Private AI to enable enterprises to access IBM Watsonx.ai in private, on-premises environments, and hybrid cloud for the secure training and fine-tuning of their models with the Watsonx.ai platform. The strategic partnership between IBM and VMware aims to enable mutual clients to easily embrace the hybrid cloud and modernize their mission-critical workloads.

Challenges

For VMware, delivering AI solutions as part of its broader offerings has been a continuous journey. Indeed, VMware was among the first ISVs to build a virtualization layer that integrates AI workloads with other enterprise workloads. This new offering is well targeted toward the data privacy concerns specifically around generative AI, and it brings in big-name partners that organizations are familiar with.

That said, for enterprises that are evaluating the offering and how to integrate it into their overall AI environment, there is the risk of stack overload. Since much of the VMware solution is built around open source software, it can be expected that businesses already have some of that software in their AI environments and possibly they are running different versions of that software. Furthermore, all the participants in a typical AI stack are offering their own AI software stacks, be they processor vendors, server vendors, and storage vendors. It may not be trivial to untangle these existing stacks and then cleanly install VMware Private AI.

For greenfield deployments, this is not a problem, but for brownfield installations, this could introduce some complications, at which point comprehensive support will be required. The question then is, who provides that support? In any scenario where multiple vendors partner to deliver a solution, the support question becomes critical.

Conclusion

The productivity-improving and revenue-generating opportunities with GenAI are extremely promising. However, developing and deploying a GenAI model requires careful consideration. Organizations have a few options:

- They can opt to merely use APIs to get access to foundation models, but this approach will yield very little business differentiation for them.
- Most have decided to customize or fine-tune foundation models with their own data. Such proprietary models provide high value as they will be perfectly tailored to the company's specific needs. They also pose a significant risk as the privacy of the data can be compromised during the many AI life-cycle steps.
- Many organizations prefer to keep generative model development and deployment on premises, behind their firewall. To do so, they need the right infrastructure stacks delivered by reliable vendors and based on as much open source software as possible.

IDC believes that with VMware Private AI, VMware and its various partners are delivering such a full-stack solution for secure on-premises generative AI development and deployment.

About the Analyst

Peter Rutten, *Research Vice President, Performance-Intensive Computing, Worldwide Infrastructure Research*

Peter Rutten is a research vice president within IDC's Worldwide Infrastructure Practice, covering research on computing platforms. Mr. Rutten is IDC's global research lead on performance-intensive computing solutions and use cases.

MESSAGE FROM THE SPONSOR

More About VMware Private AI

VMware Private AI is an architectural approach for AI services which enables privacy and control of corporate data, choice of open source and commercial AI solutions, quick time-to-value, and integrated security and management. Go to this link to learn more about VMware Private AI.

VMware by Broadcom

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com