

LEARNING MADE EASY

Broadcom Special Edition

# Private AI

<sup>for</sup>  
**dummies**<sup>®</sup>  
A Wiley Brand



Enable privacy and  
security of AI models

Simplify infrastructure  
management

Streamline model  
deployment

Compliments  
of

**vmware**<sup>®</sup>  
by Broadcom

**Ulrika Jägaré**  
**Shobhit Bhutani**

## About Broadcom

Broadcom Inc. (NASDAQ: AVGO) is a global technology leader that designs, develops, and supplies a broad range of semiconductor, enterprise software, and security solutions. Broadcom's category-leading product portfolio serves critical markets including cloud, data center, networking, broadband, wireless, storage, industrial, and enterprise software. The company's solutions include service provider and enterprise networking and storage, mobile device and broadband connectivity, mainframe, cybersecurity, and private and hybrid cloud infrastructure. Broadcom is a Delaware corporation headquartered in Palo Alto, California. For more information, go to [www.broadcom.com](http://www.broadcom.com).



# Private AI

Broadcom Special Edition

**by Ulrika Jägare and  
Shobhit Bhutani**

**for  
dummies®**  
A Wiley Brand

# Private AI For Dummies®, Broadcom Special Edition

Published by  
**John Wiley & Sons, Inc.**  
111 River St.  
Hoboken, NJ 07030-5774  
[www.wiley.com](http://www.wiley.com)

Copyright © 2026 by John Wiley & Sons, Inc., Hoboken, New Jersey. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Trademarks:** Wiley, For Dummies, the Dummies Man logo, The Dummies Way, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact [info@dummies.biz](mailto:info@dummies.biz), or visit [www.dummies.com/custom-solutions](http://www.dummies.com/custom-solutions). For information about licensing the *For Dummies* brand for products or services, contact [BrandedRights&Licenses@Wiley.com](mailto:BrandedRights&Licenses@Wiley.com).

ISBN 978-1-394-34789-6 (pbk); ISBN 978-1-394-34790-2 (ebk); ISBN 978-1-394-34791-9 (ebk)

## Publisher's Acknowledgments

**Editor:** Elizabeth Kuball

**Acquisitions Editor:** Traci Martin

**Senior Managing Editor:** Rev Mengle

**Client Account Manager:**  
Cynthia Tweed

**Content Refinement Specialist:**  
Bharaneedharan Murthy

**Special Help:** Justin Murray

# Table of Contents

- INTRODUCTION ..... 1
  - About This Book ..... 1
  - Icons Used in This Book..... 2
  - Beyond the Book..... 2
- CHAPTER 1: **Exploring the Evolution of AI**..... 3
  - Understanding the Next Wave of App Innovation ..... 4
  - Considering AI Usage over the Years..... 5
  - Zeroing In on GenAI and Its Potential..... 6
- CHAPTER 2: **Pinpointing Challenges with AI** ..... 9
  - Safeguarding Privacy and Security Concerns ..... 9
  - Controlling Costs and Obtaining Return on Investment ..... 10
  - Choosing Large Language Models ..... 11
  - Achieving Needed Performance..... 13
  - Ensuring Compliance ..... 14
- CHAPTER 3: **Selecting a Secure and Compliant AI Infrastructure**..... 15
  - Listing Top Risks with AI Model Management ..... 16
  - Comparing Cloud versus On-Premises..... 17
    - Concerns with a public cloud-based approach ..... 18
    - Concerns with an on-premises approach..... 19
  - Optimizing Cost while Staying Compliant ..... 20
  - Selecting the Right AI Infrastructure ..... 21
- CHAPTER 4: **Managing AI Deployments**..... 23
  - Optimizing Resource Sharing ..... 24
  - Monitoring GPUs ..... 25
  - Fine-Tuning Large Language Models and Optimizing Retrieval Augmented Generation Workflows ..... 26
- CHAPTER 5: **Optimizing Infrastructure for AI Workloads**..... 29
  - Virtualizing Your Infrastructure ..... 30
    - Virtualization of graphics processing units..... 30
    - Enabling resource allocation ..... 31
  - Supporting Live Migration..... 31
  - Cloning a Virtual Machine ..... 32

**CHAPTER 6: Unlocking AI with VMware Private AI Foundation with NVIDIA..... 33**

Understand the Architecture of the Platform..... 34

Enable Privacy and Security of AI Models ..... 36

Simplify Infrastructure Management..... 37

Streamline Model Deployment..... 38

Use Full-Stack Software-Defined Architecture through VCF ..... 39

Explore Use Cases ..... 40

    GPU-as-a Service ..... 40

    Agentic AI ..... 40

    Code generation/creation..... 41

    Contact centers resolution experience ..... 41

    Document search and summarization..... 41

    Private/secure content creation..... 42

**CHAPTER 7: Ten Steps to Get Started with VMware Private AI Foundation with NVIDIA ..... 43**

# Introduction

**G**enerative artificial intelligence (GenAI) has transitioned from a research concept for technology companies to a key driver of innovation, creating opportunities for any enterprise that wants to enhance productivity and efficiency. However, many organizations face challenges with implementing GenAI effectively due to a variety of issues. One of the most critical is that artificial intelligence (AI) demands significant infrastructure management capabilities in order to deliver value to enterprises. Although many off-the-shelf hyperscalers' cloud solutions offer comprehensive support, the public cloud approach may cause issues with privacy, security, and compliance, which can be a massive issue for many organizations.

In this book, you find out how to enable your organization to produce and run AI applications efficiently with a fit-for-purpose infrastructure in the area of semiconductors, enterprise software, and privacy constraints, keeping your data safe while meeting regulatory requirements. The right infrastructure not only enhances security by reducing the risk of data breaches but also builds trust with customers and stakeholders. With the private AI approach, data remains secure and is never exposed to third parties. Private AI represents a transformative opportunity for businesses seeking to harness the full potential of AI while addressing both privacy and operational challenges. This book details the methodology for building safe, flexible, scalable, and automated virtualized environments, where organizations can simplify AI operations, optimize resource allocation, and achieve significant real-world outcomes driven by AI.

## About This Book

This short book is packed with useful information about applying a robust AI infrastructure that provides businesses with enhanced control over their data and operational environments. In this book, you not only learn about the values that can be gained in your company, but also what private AI is all about and how to approach infrastructure-related challenges.

This book is an accessible, informative guide on the important aspects to consider when you're choosing an AI infrastructure for your organization's GenAI investment.

## Icons Used in This Book

Throughout the book, we use special icons in the margins to focus your attention. Here's what you find:



REMEMBER

The Remember icon flags information that's worth committing to memory.



TIP

When you see the Tip icon, expect to find suggestions, advice, or observations that will make your life easier (well, at least your work life — we make no promises beyond that).



WARNING

The Warning icon helps you steer clear of potholes, money pits, and other hazards along the way.



TECHNICAL  
STUFF

The Technical Stuff icon highlights technical information you can skip if you're short on time — or savor if you live and breathe this kind of detail.

## Beyond the Book

This book helps you explore strategies for how to select and build the right AI infrastructure in your company. However, it's a relatively short introduction to the subject, so for further reading and deep dives on the topic, check out [www.vmware.com/products/cloud-infrastructure/private-ai-foundation-nvidia](http://www.vmware.com/products/cloud-infrastructure/private-ai-foundation-nvidia).

- » Exploring the next step of application innovation
- » Looking into AI utilization over the years
- » Sorting out details of GenAI and its business potential

# Chapter 1

## Exploring the Evolution of AI

Since ChatGPT-3.5 beta was introduced to the world in November 2022, we've been immersed in a historic feeding frenzy of interest, innovation, and investment in artificial intelligence (AI), specifically generative AI (GenAI).



REMEMBER

GenAI is a broad category of AI models that can create new content, such as text, images, code, and more. GenAI enables computers to learn the underlying pattern related to an input and then use that pattern to generate similar content. This generated content is known as *synthetic data*, because it has been artificially created.

Large language models (LLMs) are a specific type of GenAI that focuses on question answering, summarizing, and translating. With today's LLMs, any language can now be transformed into an actual programming language. The word prompts we give these models are essentially the code they use to compute an answer, and that's as close as we've ever come to a true democratization of programming.

As this next wave of AI innovation accelerates, it will have a deep impact on the global economy. With GenAI, we have an opportunity to reinvent education by addressing variability in learning,

assist doctors in providing clinical diagnoses, help consumers with their investment decisions, and so much more.

In this chapter, you learn about the rise and business potential of AI, from simple algorithmic models to the impressive techniques of GenAI that we have today.

## Understanding the Next Wave of App Innovation

Optimized intelligent apps powered by AI are the future of user-centric technology. With the evolution and use of GenAI app experiences expanding at an aggressive pace, organizations from a wide range of industries are struggling to find the right balance between privacy, security, user experience, innovation, and operational efficiency. To safely take advantage of these rapidly evolving technologies requires careful planning and a clear understanding of the foundational elements that will help you succeed.



TIP

Unlike traditional apps that rely on predefined algorithms, AI-driven apps use machine learning (ML), data analysis, and predictive modeling to create highly responsive, self-learning, and adaptable user experiences.



TIP

AI makes programming easier for all types of users, regardless of previous experience, through *low-code/no-code* technology. This AI trend helps users to create apps, websites, and workflows without prior programming experience — which is very useful for companies that need to develop solutions constantly and quickly.

Another interesting trend in app innovation is the increased usage of *conversational AI*, in which AI techniques can create realistic conversations between humans and machines. Conversational AI-enabled apps allow users to communicate by text or voice as they would with another person. In this way, costs are reduced, productivity is increased, and a better customer experience is offered by providing personalized assistance at any time.



REMEMBER

Another example of app innovation is that GenAI facilitates the creation of new content, such as text, images, or code, by training on existing data. This technique uses neural networks to analyze the patterns and structure of input data to generate new content

with similar characteristics. For example, you could create an app that summarizes complex financial documents, a medical app that generates images showing the future progression of a disease, or an AI assistant that provides information from a company's website to speed up research.



TIP

AI has also become an essential aid in cybersecurity for safeguarding online systems against attacks by cybercriminals. By learning and processing natural language, AI techniques help monitor and analyze behavioral patterns to detect most cyber threats in real time. It also improves the management of user access to apps and websites by analyzing login attempts using biometrics, multifactor authentication (MFA), and behavioral data. These capabilities are key in industries such as finance, insurance, e-commerce, and health care.

Finally, *AI hyperautomation* aims to automate everything possible in an organization, both in terms of data and processes. Hyperautomation collects and analyzes data from different sources, such as sensors, databases, or user input, with greater precision and speed than a human being could. But it doesn't stop there: It also identifies inefficiencies and opportunities for improvement. By automating repetitive and routine tasks and streamlining business processes, hyperautomation allows employees to focus on more strategic and higher-value tasks, thereby improving the organization's efficiency and productivity.

## Considering AI Usage over the Years

AI has come a long way from its conceptual beginnings to becoming a key driver in the technological revolution we're witnessing today. This journey of AI is a fascinating tale of innovation, challenge, and transformation. AI, once a subject of science fiction, has now seeped into various aspects of our lives, reshaping industries and redefining our understanding of what's possible.



REMEMBER

It all started in the 1950s with the development of the first AI programs capable of playing checkers and solving algebra problems. These initial steps were followed by a period in the 1960s and early 1970s, often referred to as the "golden years of AI," characterized by significant funding, enthusiasm, and groundbreaking research. During this time, some of the most fundamental

algorithms and techniques in AI were developed. Innovations in natural language processing (NLP), problem-solving, and early neural networks were some of the highlights.

In the late 1970s and early 1980s, the first AI winter occurred, a period marked by skepticism, reduced funding, and disillusionment in the field of AI. The limitations of the early AI systems had become apparent — they were costly, limited in their capabilities, and unable to scale. This period forced a reevaluation of expectations and approaches in AI research, paving the way for more focused and realistic developments in the future.

The 1990s marked a turning point in the resurgence of AI, thanks largely to advancements in ML. The development of more sophisticated algorithms' increased computational power and availability of large amounts of data due to massive digitalization efforts contributed to this revival. ML, particularly the emergence of deep learning, enabled AI systems to learn from data, improve over time, and make decisions with minimal human intervention. This period set the stage for the modern era of AI.

In the 21st century, AI has become an integral part of our daily lives, from personalized recommendations on streaming services to virtual assistants in our smartphones. Apps enabled by AI are vast and varied. AI apps are facilitating breakthroughs in fields like health care, finance, and autonomous vehicles. Plus, the integration of AI with technologies like the Internet of Things (IoT) has further expanded its capabilities and apps, making it a cornerstone of the current technological landscape and spurring continued app innovation going forward.

## Zeroing In on GenAI and Its Potential

AI has seeped into our lives incrementally, through everything from technology powering our smartphones to autonomous-driving features on cars to the tools retailers use to surprise and delight consumers with potential offers. As a result of this step-by-step integration into our lives, its progress has been almost invisible. Clear milestones — such as when AlphaGo, an AI-based program developed by DeepMind, was created to play the board game Go, defeated a world-champion Go player in 2016 — were celebrated but then quickly faded from the public consciousness.

However, when GenAI apps such as ChatGPT and GitHub Copilot were released, they captured the imagination of people around the world in a way AlphaGo did not, thanks in large part to their broad usefulness and ease of use. Almost anyone can use these GenAI apps to communicate and create content because of their ability to have a conversation with a user. The latest GenAI apps can perform a range of routine tasks, such as reorganizing and classifying data. But it's their ability to write text, compose music, and create digital art that has garnered headlines and persuaded consumers and households to experiment on their own. As a result, a broader set of stakeholders is grappling with GenAI's impact on business and society — without much context to help them make sense of it.

The speed at which GenAI technology is developing isn't making this task any easier. ChatGPT was developed and released by OpenAI in November 2022. Four months later, OpenAI released a new LLM called GPT-4 with markedly improved capabilities.



TECHNICAL  
STUFF

To grasp what lies ahead requires an understanding of the breakthroughs that have enabled the rise of GenAI, which took several decades. From a technical perspective, GenAI apps are typically built using *foundation models*, which contain expansive artificial neural networks inspired by the billions of neurons connected in the human brain. Foundation models are part of the deep learning technique, a term that refers to the many deep layers within neural networks. Deep learning has powered many of the recent advances in AI, but the foundation models powering GenAI apps are a step-change evolution within deep learning. Unlike previous deep learning models, they can process extremely large and varied sets of unstructured data and perform more than one task.

The foundation models have enabled new capabilities and vastly improved existing ones across a broad range of modalities, including images, video, audio, and computer code. AI trained on these models can perform several functions; it can classify, edit, summarize, answer questions, and draft new content, among other tasks.



REMEMBER

Although AI may seem very advanced already, we're still at the beginning of a journey to understand GenAI's power, reach, and capabilities. Recent research suggests that GenAI will transform roles and boost performance across functions such as sales and marketing, customer operations, and software development. In

the process of doing so, it has the potential to unlock trillions of dollars in value across any business sector, not just the tech industry. In “The Economic Potential of Generative AI: The Next Productivity Frontier,” McKinsey & Company predicts that “the total economic potential of generative AI use cases amounts to \$4.4 trillion annually” (see <https://tinyurl.com/39jffn7x> for the full report).

## IN THIS CHAPTER

- » Making privacy and security concerns a priority
- » Taking control of cost aspects and achieving return on investment
- » Picking the right large language models
- » Reaching desired performance levels
- » Ensuring compliance

# Chapter 2

## Pinpointing Challenges with AI

According to McKinsey & Company (see “The Economic Potential of Generative AI: The Next Productivity Frontier”; <https://tinyurl.com/39jffn7x>), there is a potential \$4.4 trillion annual economic value for enterprises with generative artificial intelligence (GenAI). So, it’s no surprise that companies are eager to use this technology to boost productivity across every aspect of their organizations. However, although it’s very likely that GenAI will transform businesses in much the same way that the PC, the internet, and the smartphone did, there’s no sure road to success. Enterprises must tackle several challenges before bringing GenAI into their environment, as I describe in this chapter.

## Safeguarding Privacy and Security Concerns

One basic fact is that, in order to succeed with GenAI, you need to move from risk to trust. Current artificial intelligence (AI) models create significant risks, including privacy issues, legal and

regulatory threats, and intellectual property (IP) leakage. These risks have the potential to damage a company's reputation and brand perception, harm customers and employees, and negatively impact revenue. Many organizations have established policies restricting employee use of GenAI tools in the wake of employees accidentally leaking sensitive internal data into GenAI solutions such as ChatGPT.



WARNING

Today's GenAI systems suffer from a fundamental lack of trust because they can *hallucinate* (create new content that is nonsensical, irrelevant, and/or inaccurate).

To tackle this problem, the AI industry needs to develop a robust set of ethical principles to ensure and reinforce fairness, privacy, accountability, the IP of others, and the transparency of training data. A large and growing ecosystem of organizations is on a quest to address the core issues of AI explainability, data integrity, and data privacy.

## Controlling Costs and Obtaining Return on Investment

You must properly think through your investment in GenAI, especially from an architectural perspective. Without a plan, GenAI can be expensive to architect because it's a rapidly evolving technology.



TIP

Approach your infrastructure choices from the perspective that it should be "built for change" rather than "built to last." With new vendors constantly entering the market and new software continuously launching, deploying, and managing today's GenAI models is both complex and costly.



WARNING

Training GenAI models requires massive amounts of specialized compute power and high-speed networking with lots of memory. There is effectively a one-to-one relationship right now between AI model performance and compute infrastructure, a dynamic that's neither scalable nor sustainable. Andreessen Horowitz (a16z), a venture capital firm in Silicon Valley, describes training a model like ChatGPT as "one of the more computationally intensive tasks humankind has undertaken so far" (<https://news.broadcom.com/leadership/three-challenges-to-generative-ai>). The price

tag for a single training run now ranges anywhere from \$500,000 to \$4.6 million — and training will remain an ongoing expense as models are updated.

Looking at these very high costs, it's easy to jump to the conclusion that our world is going to be limited to a very small number of mega large language models (LLMs) like ChatGPT. But there's another path forward where a typical enterprise is empowered to run its own customized AI models at an affordable price. It comes down to flexibility and choice.



TIP

Many enterprises want to use LLMs for a variety of use cases, but they also want to run numerous, smaller AI models that they can optimize for specific tasks. These models are often based on openly licensed models. Indeed, the sheer volume of innovation in open-source AI models right now is astonishing. Many businesses will likely embrace these models — for example, Llama or Mistral — as their go-to option for many use cases.

These purpose-built models will focus on leveraging an organization's domain-specific data, which is their unique IP. These more compact AI systems could run cost-efficiently on on-premises infrastructure on low-cost central processing units (CPUs) modified to deliver the performance and throughput that AI workloads require.

A GPU is a specialized processor originally designed to handle the complex calculations required for rendering images and videos. GPUs differ from *central processing units* (CPUs), which handle general-purpose computing tasks in a more sequential manner. GPUs are optimized for *parallel processing*, which means they can perform many calculations simultaneously, opening the potential for more efficient work and technology-enabled outcomes.

By driving down costs and building solutions that offer flexibility and choice, it's possible to make AI innovation more accessible to mainstream enterprises.

## Choosing Large Language Models

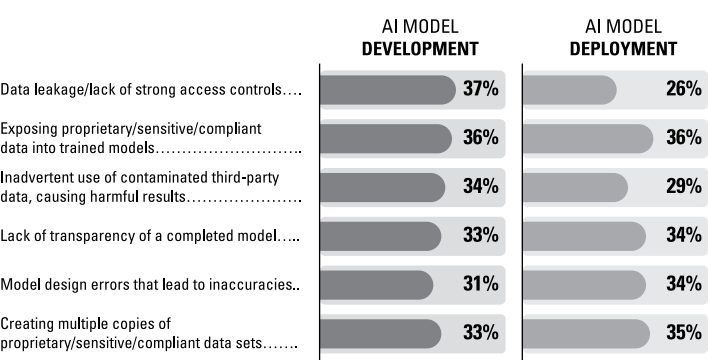


REMEMBER

Not all LLMs are equal. It's important to consider licensing and cost, evaluate inference speed and precision for performance needs, and consider context length and model size. Organizations have varying preferences for sourcing GenAI foundation models.

Public cloud and commercial independent software vendor (ISV) GenAI models can be attractive because of their quick deployment, ease of adoption, scalability, and performance. On the other hand, open-licensed models appeal to organizations that are seeking greater customization, flexibility, and transparency. Open-source models offer the ability to assess, modify, and troubleshoot models more easily, which appeals to teams that require a high degree of control and customization.

However, there’s a universal concern about data privacy and security when using AI models. According to a survey done by IDC in 2024, data leakage and exposing proprietary, sensitive, and compliant data into trained models are emphasized as the top risks associated with developing or deploying AI models (see Figure 2-1).



Source: “On-Premises AI Infrastructure Balances Innovation and Security,” published by IDC ([www.vmware.com/docs/vmw-ids-us51604323](http://www.vmware.com/docs/vmw-ids-us51604323))

**FIGURE 2-1:** Top risks with AI models for development and deployment.



**WARNING**

Enterprises want to choose the LLM that best fits their GenAI use cases, vertical industry requirements, and organizational needs, so access to various LLM options is essential. However, organizations should be wary of exposing proprietary or sensitive data into trained AI models, because this could result in data breaches or inadvertent leakage of confidential information. For example, your private data is going out over the internet to get to the site where the model is running — and you don’t know who may be monitoring that conversation. For many organizations, the concern is not limited to external threats but includes issues such as data leakage and unintentional use of contaminated or biased

third-party data, which could compromise model integrity as well. Consumers of models need a level of assurance, through testing or validation, that the models are safe from this kind of risk.

## Achieving Needed Performance



WARNING

Except for a few niche use cases, AI deployments in the enterprise are still at an early stage. Because of this, enterprise IT functions in organizations will continue to struggle to understand what the ideal AI platform would look like to meet the organization's needs — or if this is even possible. So, what can your enterprise achieve with AI and with open LLMs with a limited collection of servers with limited sets of GPUs that are feasible for you to use in your data center and produce business results early on in adoption? It's important to remember that fine-tuning of models, which refers to the process of adapting a pretrained model for specific tasks or use cases, as well as customizing, deploying, and querying LLMs, can be intensive, and scaling up can be challenging without access to adequate resources.

As these enterprise needs become more apparent, the variety of use cases (and underlying compute characteristics) point to GPUs as the ideal training accelerator. Because of GPU architecture, an enterprise IT organization can generally rely on GPUs to deliver good performance and support for its workloads. This applies even if performance is less than ideal, even if power consumption is a little too high, and even if the system generates a little more heat than hoped for. The net-net is that these generally available GPUs meet the “good enough” criteria to support most enterprise AI deployments and that efficient allocation of GPU resources is critical to ensuring low latency.



REMEMBER

However, the enterprise and hyperscale AI providers that serve it are different. The use cases are well defined, and the infrastructure required to deliver an optimal user experience is well known. Because of these factors and the sheer size of the relevant hyperscale environments, a “good enough” deployment would be costly both financially and in terms of performance. When considering the value of these platforms (defined as performance divided by total cost of ownership), the penalties for an unoptimized AI training platform add up. An organization can

absorb a 10 percent tax on performance and a 12 percent tax on power, though costly, when deploying across, say, 100 servers. These same taxes become extraordinarily burdensome when scaling across 100,000 servers.

## Ensuring Compliance



REMEMBER

For regulated industry segments, maintaining data security and compliance is paramount. One way to approach this is by using a *sovereign cloud* setup. A sovereign cloud is a cloud computing environment that stores each organization's data (including meta-data) on servers located within their local country. This ensures that sensitive data remains within specific geographic boundaries, adhering to local regulations and compliance standards. This aspect is critical for businesses operating in sectors such as health care, finance, and government, where data sovereignty is a legal requirement.



TIP

By utilizing a sovereign cloud, organizations can produce AI apps that keep their data safe while meeting regulatory requirements inherited by the cloud service provider's sovereign status and services. A sovereign cloud not only enhances security by reducing the risk of data breaches but also builds trust with customers and stakeholders. The result is a robust environment where organizations can innovate with confidence, knowing that their data is always secure and compliant.

#### IN THIS CHAPTER

- » Explaining risks with AI model management
- » Understanding the difference between cloud and on-premises
- » Balancing cost while remaining compliant
- » Picking the most suitable AI infrastructure

## Chapter 3

# Selecting a Secure and Compliant AI Infrastructure

**A**s you aim to utilize the transformative potential of artificial intelligence (AI), you need to carefully balance cloud and on-premises resources. Investing in an infrastructure that not only optimizes performance but also meets tough requirements for data security, compliance, and cost control is imperative.

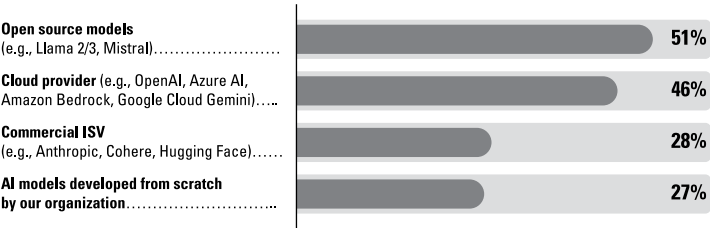
By carefully planning and managing these infrastructure aspects, your organization can build a resilient AI infrastructure that supports your objectives. As AI infrastructure continues to mature, infrastructure choices are becoming more strategically important. It's not only about investing in robust platforms to realize AI's capabilities; it's also about investing in capabilities to be able to proactively manage associated risks.

In this chapter, I explain the fundamental components of a fit-for-purpose, compliant, and secure infrastructure and how to manage associated costs and risks.

# Listing Top Risks with AI Model Management

Organizations have varying preferences for finding generative artificial intelligence (GenAI) foundation models to train and utilize for various use cases. According to a 2024 study from IDC, open-source models are most commonly used (51 percent), closely followed by cloud provider models (46 percent), as shown in Figure 3-1.

**Preferred Sources for AI Models**  
What are your preferred sources for AI models?  
(Percentage of respondents)



Source: “On-Premises AI Infrastructure Balances Innovation and Security,” published by IDC ([www. vmware . com/docs/vmw-idc-us51604323](http://www.vmware.com/docs/vmw-idc-us51604323))

**FIGURE 3-1:** Preferred sources for AI models.

According to the IDC report, public cloud and commercial independent software vendor (ISV) GenAI models are attractive because of speed and ease of deployment, ease of adoption, scalability, and performance. On the other hand, it seems that open-source models appeal to organizations looking for models that are easier to customize, with a high degree of flexibility and transparency. The common perception seems to be that open-source models offer better opportunities to assess, modify, and troubleshoot models more easily. These aspects strongly appeal to teams that require a high degree of control and customization.



**WARNING**

However, in addition to these varied approaches and preferences for sourcing foundation models, there is common concern about data privacy and security when using AI models. Fifty-four percent of the respondents in an IDC survey ([www. vmware . com/docs/vmw-idc-us51604323](http://www.vmware.com/docs/vmw-idc-us51604323)) claim that they’re extremely or very concerned about the possibility that the data used for AI development or deployment could be compromised. Data leakage and exposing

proprietary, sensitive, and compliant data into trained models are mentioned as the top risks associated with developing or deploying AI models.

The IDC report lists the following top risks that an organization associates with developing and/or deploying AI models. Here are the risks, ranked in order based on the respondents' ranking:

- » **Lack of control:** Data leakage or lack of strong access controls
- » **Exposure:** Exposing proprietary, sensitive, and compliant data into trained models
- » **Diluted data:** Inadvertent use of contaminated third-party data, causing harmful results
- » **Lack of transparency:** Lack of transparency of a completed model
- » **Model errors:** Model design errors that lead to inaccuracies
- » **Duplication of datasets:** Creating multiple copies of proprietary, sensitive, and compliant datasets



REMEMBER

Clearly, privacy and security are common concerns for organizations. Companies are wary of exposing proprietary or sensitive data into trained AI models, because this could result in data breaches or inadvertent leakage of confidential information.



WARNING

Plus, for many organizations the security concern isn't limited to external threats. Internally, issues such as data leakage and unintentional use of contaminated or biased third-party data can compromise the integrity of the model(s), and as a result, incorrect output from the model(s) could be generated.

## Comparing Cloud versus On-Premises



REMEMBER

AI is becoming increasingly important to organizational strategy across more and more industries — and not just for traditional technology companies. The landscape for AI infrastructure deployment also continues to expand, bridging on-premises and public cloud environments. *On-premises environments*, in this context, are IT infrastructure hardware and software apps that are hosted on-site — as opposed to IT assets that are hosted by a public cloud platform or remote data center.

Some of the most frequently deployed AI use cases today address private and secure content generation, contact center efficiency, document search and summarization, content creation and marketing, code development, IT operations automation, risk mitigation, and supply chain optimization. These use cases are deployed across both public cloud and on-premises infrastructure, highlighting the dual approach companies often adopt to leverage the flexibility of the cloud, while retaining control over sensitive data on-premises.



REMEMBER

These applications also showcase AI's adaptability across industries and emphasize the increasing demand for optimized AI infrastructure that accommodates both private and public cloud environments.

## Concerns with a public cloud-based approach



WARNING

The choice between public cloud and on-premises environments introduces its own concerns. Top concerns about public cloud environments include data becoming an integral part of the cloud-based AI model and the subsequent loss of control over the AI model, as well as data privacy, data control, and *egress costs* (the network fees most cloud providers charge to move your data out of the cloud).

In that 2024 IDC survey ([www.vmware.com/docs/vmw-idc-us51604323](https://www.vmware.com/docs/vmw-idc-us51604323)), here are the main worries that organizations reported when it comes to developing and deploying AI services in a public cloud (ranked in order of importance):

- » Data becoming an integral part of the cloud-based AI model
- » Data egress costs
- » Data privacy
- » Loss of control over data
- » Loss of control over AI model
- » High cost
- » Connectivity

## Concerns with an on-premises approach

For on-premises AI deployments, companies participating in the IDC survey ([www.vmware.com/docs/vmw-idc-us51604323](http://www.vmware.com/docs/vmw-idc-us51604323)) cited challenges in data accessibility for engineers, limited specialized skill sets, AI/IT personnel shortages, expanding data center maintenance costs, and data silos.



REMEMBER

A *data silo* is a repository of data that's controlled by one department or business unit and isolated from the rest of an organization. Siloed data typically is stored in a stand-alone system and often is incompatible with other datasets. That makes it hard for users in other parts of the organization to access and use company data. And when an organization aspires to be data-driven, siloed data can be a huge obstacle.

Here are the main reservations organizations have regarding developing and deploying AI services on-premises (ranked in order):

- » Lack of personnel with the right specialized skill sets (for example, prompt engineers, data scientists, and data engineers)
- » Ability to make data accessible to engineers and data scientists
- » Lack of employees with skills needed to learn and work with AI
- » Increase in data center maintenance costs
- » Data silos
- » Availability of new hardware
- » Networking



REMEMBER

Making informed decisions about where and how to develop and deploy models effectively, while also addressing security and compliance concerns, is vital to any organization. Although public clouds provide scalability and access to a variety of predefined models, they can raise concerns over data security, control, and unforeseen costs. In contrast, on-premises environments offer a customizable, secure, and cost-efficient framework that prioritizes data privacy and compliance while maximizing performance.

# Optimizing Cost while Staying Compliant

Twenty years ago, most workloads and IT resource organizations resided in their own data centers, with some limited hosting environments for off-premises workloads. During the last 12 years, however, we've seen a rush to public cloud services in the form of three different service types:

- » **Infrastructure as a service (IaaS):** A type of cloud computing service that offers essential compute, storage, and networking resources on demand, on a pay-as-you-go basis
- » **Platform as a service (PaaS):** A complete cloud environment that includes everything developers need to build, run, and manage applications — from servers and operating systems to all the networking, storage, middleware, tools, and more
- » **Software as a service (SaaS):** A cloud-based software delivery model in which individuals or organizations subscribe to software apps available online instead of purchasing and installing them locally

The adoption of public cloud services was already on the rise, but it got an extra boost during the COVID-19 pandemic (starting in 2020) because IT teams had to make split-second decisions and skip critical architecture design and planning steps in an attempt to quickly address the unexpected needs of changing work environments.



REMEMBER

Although public cloud is here to stay, the pendulum is moving toward a need to balance infrastructure investments between on-premises and the public cloud. Today, many enterprises have adopted a private cloud architecture. This natural evolution occurred as companies modernized their legacy apps with containers, microservices and discovered edge apps. Edge app services reduce the volumes of data that must be moved, the consequent traffic, and the distance that data must travel. Edge app services provide lower latency and reduce transmission costs, incentivizing companies to leverage cloud resources for local data centers and in company-operated remote locations.



WARNING

Companies have started to realize that the public cloud isn't an optimal environment for every workload for a variety of reasons, including cost, security, privacy, manageability, and other requirements and expectations. Companies must be agile as the pendulum returns to on-premises while maintaining the flexibility to handle an uncertain future.

According to the previously referenced IDC survey ([www.vmware.com/docs/vmw-idc-us51604323](http://www.vmware.com/docs/vmw-idc-us51604323)), IT skill set and training, platform costs, and data security costs were the top three reasons that public cloud was seen as more expensive for AI model development. When it came to AI model deployment, compute infrastructure costs rose to the top reason, which indicates that organizations are likely concerned about surprise cloud bills due to AI models. In addition to overall costs, organizations also see on-premises AI infrastructure as offering more cost predictability and stability, especially over time.

## Selecting the Right AI Infrastructure

Today, AI workloads, especially GenAI workloads, are viewed as discrete from other workloads by many organizations. However, in the future, organizations want private cloud AI workloads managed similarly to any other enterprise workload. This includes a common platform with a comprehensive feature set.



TIP

Most likely, companies will expect this common platform to include features for resiliency and business continuity, a shared infrastructure, automation, performance management, and security. Furthermore, organizations will also expect virtualization to deliver similar benefits to a private AI infrastructure as it has to traditional servers over the past two decades. Because virtualization of private AI servers offers substantial benefits — such as hardware and maintenance cost savings, improved resource utilization and management, improved disaster recovery and business continuity, and security — this will be a preferred choice.

In the aforementioned IDC survey ([www.vmware.com/docs/vmw-idc-us51604323](http://www.vmware.com/docs/vmw-idc-us51604323)), the respondents were asked about server and storage features that would best help their organization

achieve its goal of optimizing the cost and performance of an on-premises AI infrastructure. The top-rated responses were a hybrid cloud strategy, access to AI development tools, and a shared pool of graphics processing units (GPUs) and central processing units (CPUs).



TECHNICAL  
STUFF

Other important features for optimizing the cost and performance of on-premises AI infrastructure listed by the survey respondents include

- » Best-of-breed accelerators — GPUs/field programmable gate arrays (FPGAs) and more
- » Network capacity
- » Memory
- » Load balancing
- » Alert notifications
- » Autoscaling
- » Infrastructure monitoring tools
- » Heat maps that monitor and identify patterns of usage
- » Virtualization
- » Lifecycle management features

- » Understanding the need for resource sharing
- » Emphasizing the importance of monitoring graphics processing units and performance
- » Exploring how to fine-tune large language models and optimize retrieval augmented generation workflows

# Chapter 4

## Managing AI Deployments

One approach that's getting increased attention is on-premises deployment in a private cloud setup. This approach, known as *private artificial intelligence* (AI), offers a strategic solution, allowing organizations to retain complete control over their critical data and comply with industry regulations. Private AI fosters trust and transparency, ensuring that sensitive information and intellectual property (IP) are securely protected within enterprise environments.

A fundamental decision you make when you adopt Private AI in your data center is to use validated models that are hosted on-premises with business data that is kept within the enterprise. You bring the models to the data instead of potentially exposing your data to the outside world. To achieve this, you first consider how you can use pretrained large language and embedding models in-house.

# Optimizing Resource Sharing



WARNING

However, fine-tuning, customizing, deploying, and using large language models (LLMs) for inference can be resource demanding, and scaling up the capacity can be challenging without access to adequate resources. So, efficient allocation of graphics processing unit (GPU) resources is critical to ensure low latency.



REMEMBER

As you deploy and scale your AI infrastructure, you may encounter several critical infrastructure-specific issues that can hinder adoption of LLMs based on the specific nature of different GenAI use cases. Without addressing these challenges, your system architects will find it difficult to deploy, configure, and reconfigure compute, storage, and networking infrastructure to support the needs of the use cases.



TIP

So, to ensure efficient deployment and operation of your use cases, you'll need fine-grained management of the infrastructure for your workloads. That will accelerate your AI throughput, deliver seamless scaling, and maximize the GPU utilization.



TECHNICAL  
STUFF

Data scientists in the enterprises will want to:

- » **Fine-tune LLM models.** Fine-tuning LLMs is the process of adapting a pretrained language model to a specific task or dataset.
- » **Deploy retrieval-augmented generation (RAG) workflows.** Deploying RAG workflows is the process of optimizing the output of an LLM so it references an authoritative knowledge base outside of its training data sources before generating a response.
- » **Run inference workloads in a secure, controlled, and cost-efficient environment.** Running inference workloads means taking a trained AI model and deploying it into a production environment.

All of these computational needs put demands on the needed support for various AI workloads, so you must be able to simplify AI deployments, and they must be supported by GPU monitoring capabilities so that you have full control over your different AI workloads.



TIP

Be sure to invest in an architectural approach that balances the business gains from AI with your organization's practical privacy and compliance needs. This means investing in a solution for AI services that ensures privacy and control of corporate data, offers a choice of open-source and commercial AI solutions, provides quick time to value, and provides integrated security and management.

When integrating LLMs, expect the right infrastructure setup and resource sharing to assist in simplifying complex tasks. The solution should ensure consistency across various cloud environments and facilitate seamless AI workload deployment. The scalability capabilities of the environment should support the resource demands of the AI tasks at the same time as infrastructure resources are optimally utilized.



REMEMBER

Invest in streamlined lifecycle management capabilities of the platform. Those capabilities will aid in streamlining updates and upgrades, which will minimize disruptions to your AI operations.

## Monitoring GPUs

In addition to training and fine-tuning AI models, GPUs are essential for *inference* (the process by which a trained model is used to make predictions on new data in a live production setting). In AI, training and inference are two distinct phases:

- » **Training:** Training involves teaching the model to recognize patterns and learn from data.
- » **Inference:** After training is complete, inference applies the model to new, unseen data to generate predictions or classifications.

Training can be time-consuming and resource-intensive, but inference needs to happen quickly, especially for real-time applications where delays can have significant consequences (for example, in the case of self-driving vehicles).



REMEMBER

There's an increasing reliance on GPUs for compute-intensive tasks such as machine learning (ML), deep learning, and data processing. This means that both infrastructure administrators and users of GPUs (such as data scientists, ML engineers, and AI app developers) require timely access and insights into performance, efficiency, and the overall health of their GPU resources.



WARNING

GPUs are expensive and energy-intensive resources. Ensuring that GPUs are fully utilized (not underused or overused) is essential for cost-efficiency. Administrators need to monitor GPUs so they can rebalance workloads across multiple GPUs, identify idle GPUs, and right-size resources for different applications, thereby reducing overall infrastructure costs.

GPU metrics provide vital information for diagnosing performance bottlenecks in applications that rely on GPU acceleration. Admins and users should be able to quickly identify performance issues (such as GPU bottlenecks or inefficiencies). Here are some examples of important GPU metrics to monitor:

- » **GPU compute and memory utilization:** Some workloads are compute-dependent, while others are memory-dependent. Monitoring compute and memory utilization helps to allocate the right GPUs for the job.
- » **GPU temperature monitoring:** GPU throttling can degrade performance. Monitoring GPU temperature helps identify when GPUs are throttling.
- » **Monitoring virtualized GPUs (vGPUs):** Similar to GPU metrics, visibility into various aspects of vGPU performance — such as memory usage, compute utilization, and encode and decode utilization — is critical for optimum workload performance.



TIP

In order to manage the GPU situation for your AI workloads, your infrastructure needs a scalable, software-defined architecture. This architectural approach needs applicable performance tools, designed to deliver a self-service unified platform, and should leverage an automated IT environment that simplifies the deployment and management of all AI workloads.

## Fine-Tuning Large Language Models and Optimizing Retrieval Augmented Generation Workflows

LLMs have revolutionized natural language processing (NLP) tasks, enabling machines to understand, generate, and interact with human language in a remarkably human-like manner. These models — such as GPT-5, MPT, Vicuna, and Falcon — have

gained popularity because they can process vast amounts of text data and produce coherent and contextually relevant responses. Behind the scenes, these models rely on complex operational components and processes that work closely together to achieve their impressive capabilities.



WARNING

However, without RAG, the LLM takes user input and creates a response based only on information it was trained on — or what it already knows. RAG introduces an AI framework that combines information retrieval with text generation, allowing a model to fetch relevant external data and use it to generate more accurate and informed responses.

In a nutshell, RAG is the process of triggering a search in a separate data source (for example, a database) to retrieve information relevant to the user's prompt before sending it to an LLM. For example, if a user asks an onboarding chatbot about the company's holiday policy, a search gathers the previously uploaded information on that topic. That data is then packed with the user prompt and supplied to the LLM as context. This is the same effect of copying and pasting the content yourself but in a much faster, integrated, and user-friendly way. Adding this additional context doesn't change the architecture, layers, or parameters of the model.



REMEMBER

Although the basic RAG design preserves the completions and embeddings models with their original parameters, there are situations where you may fine-tune these models for a specific purpose. In the process of fine-tuning a model, the approach is to use a collection of domain-specific data to train a general-purpose LLM for a particular task. Think of medicine or coding, each of which has its own logic and jargon. However, when fine-tuning and RAG techniques are combined, the LLM can be based on both strong domain knowledge and up-to-date information, generating a more correct and trustworthy output.



TIP

From an infrastructure point of view, strive for an architectural approach that enables data scientists to use GPU-enabled self-service workstations with a RAG reference solution capability. Using RAG to retrieve facts from an enterprise knowledge base containing your company's business data will enable data scientists to enhance business value in GenAI solutions by augmenting an existing foundational LLM to fit your business use case. Pay special attention to ways in which you can augment an LLM with your domain-specific business data to create AI apps that are agile and responsive to change.

- » Leveraging a virtualized infrastructure
- » Using live migration capability
- » Duplicating a virtual machine

# Chapter 5

## Optimizing Infrastructure for AI Workloads

To fully harness the power of artificial intelligence (AI) systems, optimizing their performance infrastructure is critical. By leveraging advanced capabilities, you can accelerate AI model training and inference, leading to improved efficiency and speed.

In this chapter, we explain why virtualization, enabling automated resource scheduling, supporting live migration, and cloning virtual machines (VMs) are important infrastructure enablers for accelerating performance.



REMEMBER

*Compute resources* are infrastructure elements, whether hardware or software, that enable problem-solving and solution creation through receiving, analyzing, and storing data. A *workload*, in the most general sense, is the amount of time and computing resources a system or network takes to complete a task or generate a particular output. It refers to the total system demand of all users and processes at a given moment. Compute workloads

are applications or services that require processing power and memory to perform their functions. These can include VMs and containers. Containers can run within VMs for better workload isolation.

## Virtualizing Your Infrastructure

*Virtualization* is a process that allows a computer to share its hardware resources with multiple digitally separated environments. Each virtualized environment runs within its allocated resources, such as memory, processing power, and storage. A VM is a virtual environment that works like a computer within a computer. It runs on an isolated subset of its host computer with its own central processing unit (CPU) power, memory, operating system (OS), and other resources.

### Virtualization of graphics processing units

Graphics processing units (GPUs) are a scarce, expensive, and in-demand resource. Hence, it's important to virtualize GPUs to extract maximum value from them. GPU virtualization lets multiple users or applications share a single powerful GPU. Instead of dedicating one GPU to one workload, virtualization can divide the GPU into isolated “instances,” so different tasks — like AI training, rendering, or analytics — can run on that physical GPU at the same time without interfering with each other. This makes it possible to get more value from the same hardware while giving each user the performance they need. GPU virtualization is done through NVIDIA Virtual GPU (vGPU) technology, which is integrated with VMware Cloud Foundation (VCF).

For organizations, virtualized GPUs mean flexibility and efficiency. Teams can scale up GPU resources on demand, reduce costs by avoiding idle hardware, and enable remote workers to access high-performance computing from anywhere. Whether it's data scientists running AI models, engineers doing simulations, or designers using 3-D applications, GPU virtualization helps deliver the right level of power while keeping infrastructure simpler and more cost-effective.

## Enabling resource allocation

In a computer system, a *cluster* is a group of servers and other resources that act like a single system and enable high availability, load balancing, and parallel processing. For these reasons, live migration is particularly useful for cluster administrators.

With automated resource allocation capability, users can define the rules for the automatic allocation of physical resources among VMs.

VCF has this capability enabled through a utility called the Distributed Resources Scheduler (DRS). DRS can be configured for manual or automatic control. Resource pools can be easily added, removed, or reorganized. *Resource pools* are a native concept to VCF that allow system administrators to set boundaries around a collection of CPU, main memory, networking, and storage resources that allowed VMs will consume those resources from the pool. If the workload on one or more VMs drastically changes, the DRS may decide to redistribute the VMs among the physical servers. DRS increases productivity by allowing one administrator to monitor a large pool of resources. DRS also reduces maintenance downtime. When a physical server is put into maintenance mode, DRS automatically migrates all VMs, including those with vGPU profiles, to other servers without taking those VMs down (that is, live migration).

## Supporting Live Migration

*Live migration* refers to the process of moving a VM running on one physical host to another host without disrupting normal operations or causing any downtime or other adverse effects for the end user. Live migration is considered a major step in virtualization.

By allowing an entire VM to be moved with a running OS, live migration helps accommodate low-disruption fault management and low-level system maintenance.

Live migration is normally performed when a physical host machine (computer or server) needs maintenance or an update, or when a VM must be manually switched to a different host. The process transfers the vCPU contents, VM memory contents,

network connectivity, and in certain circumstances the storage as well. When VMs operate on shared storage such as vSAN, then the storage contents may not move at all.

The live migration process allows a clean separation between hardware and software with a separation of concerns between the users and operator of a data center or cluster.

End application users do not need to be aware that a vMotion is taking place, because their application is not affected by that process. In this sense, there is a separation of concerns between the users and the operations in a data center.

Migrating an entire OS and all its applications as one unit can eliminate many of the difficulties involved in process-level migration approaches. This method eliminates the issue of residual dependencies that require the original host machine to remain available and network-accessible to service memory accesses or system calls on behalf of migrated processes. Migrating at the entire VM level also means that its in-memory state can be transferred consistently and efficiently.

Additionally, live migration supports more efficient load balancing, so systems and CPU resources can be organized and shared for optimal usage. It also allows applications to continue running while the administrator manages maintenance activities, such as security updates, in the background.

## Cloning a Virtual Machine



TIP

Installing a guest OS and applications can be time-consuming. With *clones*, you can conveniently make copies of a virtual machine. In the cloning process, you create an exact copy of your original VM. A VMware clone has the same virtual hardware, software, and other configurations as the original VM. Cloning is useful when deploying several VMs with the same resource allocation and configurations. Cloning is the key to rapid provisioning of environments, whether it is done from templates or from existing VMs.

#### IN THIS CHAPTER

- » Exploring the architectural components of the platform
- » Describing how to handle privacy, security, and compliance of artificial intelligence models
- » Exemplifying important aspects of infrastructure management
- » Looking at organizational use cases

# Chapter 6

## Unlocking AI with VMware Private AI Foundation with NVIDIA

In the rapidly evolving landscape of artificial intelligence (AI), Broadcom and NVIDIA have developed a joint AI platform called VMware Private AI Foundation with NVIDIA. This platform unlocks AI and unleashes productivity with lower total cost of ownership (TCO). VMware Private AI Foundation with NVIDIA enables enterprises to run RAG workflows, fine-tune and customize large language models (LLMs), and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns. In this chapter, you learn more about the specifics of this solution.

# Understand the Architecture of the Platform



WARNING

Organizations often face challenges in efficiently scaling their compute infrastructure, especially when dealing with specialized AI workloads requiring high parallel processing levels. Regular enterprise computing environments may not be optimized for the intense computational demands of AI applications, particularly those requiring graphics processing unit (GPU) acceleration. This can lead to underutilized resources and increased operational costs, especially when considering the cost of deploying GPUs in these complex environments. Without proper operational visibility and the ability to properly allocate these resources across multiple users, valuable GPU resources are wasted.



REMEMBER

These challenges are complex to overcome. Enterprises can get tremendous productivity and business transformation from AI; however, they're also looking for ways to address the concerns outlined earlier. To address this, Broadcom and NVIDIA have collaborated to develop the joint AI platform called VMware Private AI Foundation with NVIDIA. This platform enables enterprises to fine-tune LLM models, deploy RAG workflows, and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns. This joint platform simplifies AI deployments for enterprises by offering Model Store, Air-Gapped Support, Model Runtime, NVIDIA NIM, NVIDIA Blueprints, and more.

As you can see in Figure 6-1, VMware Private AI Foundation with NVIDIA is built and run on the private cloud platform, VMware Cloud Foundation (VCF), and includes VCF Private AI services, NVIDIA AI Enterprise, NVIDIA NIM ([www.nvidia.com/en-us/ai-data-science/products/nim-microservices](http://www.nvidia.com/en-us/ai-data-science/products/nim-microservices)) inference microservices for the latest AI models, and NVIDIA Blueprints (<https://build.nvidia.com/blueprints>). VCF, VMware's full-stack private cloud infrastructure solution, offers a secure, comprehensive, and scalable platform for building and operating AI workloads, providing organizations with the agility, flexibility, and scalability to meet their evolving business needs. Previously sold separately, VMware Private AI services are now included as part of the VCF subscription. NVIDIA AI Enterprise licenses need to be purchased separately from NVIDIA.

## VMware Private AI Foundation with NVIDIA

Unlock AI and Unleash Productivity with Lower TCO

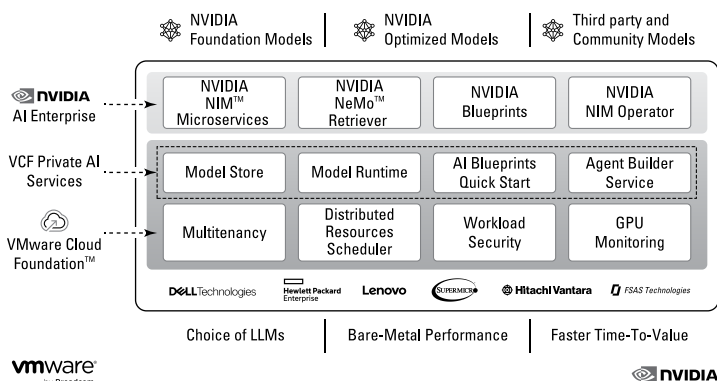


**FIGURE 6-1:** VMware Private AI Foundation with NVIDIA.

### Components of this platform include

- » **VCF:** VCF is a private cloud platform that delivers public cloud scale and agility, private cloud security, resilience and performance, and low overall TCO for AI workloads. The architecture enables cloud admins to utilize different workload domains, which can be customized to support specific workload types, optimizing for workload performance and resource utilization, specifically GPUs.
- » **VCF Private AI services:** Provide capabilities like vector databases, deep learning VMs, Data Indexing and Retrieval service, and AI Agent Builder service to enable privacy and security, to simplify infrastructure management and streamline model deployment.
- » **NVIDIA AI Enterprise:** A secure, end-to-end, cloud-native software platform that accelerates the data science pipeline and streamlines development and deployment of production-grade AI applications, including generative AI, computer vision, and speech AI. NVIDIA NIM allows enterprises to run inference on a range of LLMs from NVIDIA models to community models.
- » **Server original equipment manufacturer (OEM) support:** Major server OEMs, such as Dell, Fsas Technologies, Hitachi Vantara, HPE, Lenovo, and Supermicro, support the platform.

# Enable Privacy and Security of AI Models

A vital concern for most enterprises regards efficient handling of privacy and security aspects of their AI models. In VMware Private AI Foundation with NVIDIA's architectural approach for AI services, this is addressed through integrated capabilities for handling privacy and control of corporate data. Here are the main capabilities in this category:

- » **Model Store:** Because enterprises often lack the governance measures for downloading and deploying LLMs from independent software vendors (ISVs) or other third-party sources, it can pose significant risks to enterprises. LLMs with malware or viruses could be introduced into the environment, knowingly or unknowingly. However, the Model Store capability enables MLOps teams and data scientists to curate and provide secure LLMs with integrated role-based access control (RBAC). This can ensure governance and security for the environment and the privacy of enterprise data and IP.
- » **Air-gap support:** Through VCF Automation, VMware Private AI Foundation with NVIDIA can be deployed in air-gapped environments, supporting the business needs of customers, with data confidentiality and isolation for their critical workloads.
- » **AI software security patching through NVIDIA AI Enterprise:** For NVIDIA AI Enterprise customers, NVIDIA commits to patch critical and high common vulnerabilities and exposures (CVEs) monthly for production branches and quarterly for long-term support branches while maintaining application programming interface (API) compatibility up and down the stack.



REMEMBER

With support of these infrastructural management capabilities, organizations can then develop AI apps that keep the data private and safe while meeting regulatory requirements. These infrastructure capabilities not only enhance security by reducing the risk of potential data breaches, but also generate customer trust in the solutions.

# Simplify Infrastructure Management



REMEMBER

Investing in solid infrastructure management capabilities is a key success factor to fully leverage the business value of AI. However, knowing what to invest in is not an easy task. Many off-the-shelf hyperscale solutions, though comprehensive, cause fragmentation in operations and data privacy, governance, and compliance, which can be a massive issue for many organizations.

Additionally, AI applications are complex and costly to architect because they're rapidly evolving with new vendors, software as a service (SaaS) components, and bleeding-edge AI software continuously launched and deployed. In this complex environment, VMware Private AI Foundation with NVIDIA comes with specifically architected capabilities that help simplify infrastructure management of AI environments and optimize costs:

- » **vGPU Profile Visibility:** This is a capability to view all vGPUs across their GPU footprint in an easy-to-use UI screen in vCenter, eliminating the manual tracking of vGPUs and reducing admin time.
- » **Vector Databases for Enabling RAG Workflows:** Broadcom has enabled Vector databases by leveraging pgvector on PostgreSQL. This capability is managed through Data Services Manager and enables quick deployment of vector databases to support retrieval-augmented generation (RAG) AI applications.
- » **AI Blueprints Quick Start:** With this capability, line of business (LOB) admins can, on Day 0, quickly design, curate, and offer infrastructure catalog objects through VCF's self-service portal (formerly VCF Automation), greatly simplifying the deployment of AI workloads.



REMEMBER

Legacy systems often hinder operational efficiency, leading to increased costs and downtime. This is addressed through VMware Private AI Foundation with NVIDIA, which enables businesses to modernize their infrastructure without the need for a complete overhaul. By gradually updating legacy systems on-premises and implementing microservices architectures managed by the VMware Cloud Foundation platform, organizations can achieve better integration with AI technologies.

There's also a broad set of automation tools available for organizations to scale AI workloads more effectively, reduce operational

complexity, and minimize resource waste. This is particularly important in GPU environments, where the high costs and power consumption associated with these specialized resources demand a careful and strategic approach from an operational standpoint.

## Streamline Model Deployment

To easily deploy and manage AI workloads for data scientists and IT teams, Broadcom and NVIDIA have also enabled capabilities and software to support this. These capabilities include

- » **Model Runtime:** This service enables data scientists to create and manage model endpoints for their applications. This simplifies model usage and the scalability of LLMs.
- » **Agent Builder:** This service allows for GenAI application developers to build AI agents by using resources from the Model Store, Model Runtime, and Data Indexing and Retrieval Service.
- » **NVIDIA NIM:** This is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing across the cloud, data center, and workstations. These prebuilt containers support a broad spectrum of AI models — from open-source community models to NVIDIA AI Foundation models, as well as custom AI models.
- » **Support for NVIDIA Blackwell GPUs:** NVIDIA Blackwell GPUs ([www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/](http://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/)) unlock the potential of generative, agentic, and physical AI by delivering exceptional performance, efficiency, and scale for enterprises. VCF now supports the NVIDIA Blackwell architecture, enabling enterprises to get the industry-leading AI training and inference capabilities at unprecedented scale. The specific supported GPUs are NVIDIA RTX PRO 6000 Blackwell Server Edition ([www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/](http://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/)). For workloads that require the compute density and scale that data centers offer, the NVIDIA RTX PRO 6000 Blackwell Server Edition GPU delivers powerful performance for next-generation AI, scientific, and visual computing applications across industries such as health care, manufacturing, retail, and media and entertainment. Designed for enterprise data center deployments, the RTX PRO 6000 Blackwell Server Edition can be configured with up to eight GPUs per server.

# Use Full-Stack Software-Defined Architecture through VCF

VMware Cloud Foundation (VCF) is the industry's first private cloud platform to deliver public cloud scale and agility with private cloud security, resilience, and performance. VCF offers a full-stack software-defined architecture designed to deliver a self-service unified platform. It leverages an automated IT environment for the deployment and management of AI, non-AI workloads, and containers. This unified approach provides overall lower TCO and improved performance. The recently released benchmark study, compared against bare metal using MLPerf Inference v5.0 standards, shows performance similar to bare metal. So, putting AI workloads on virtualized solutions preserves performance while adding the benefits of virtualization, such as ease of management and enterprise-grade security.

Here are some of the key capabilities in the platform enabled through VCF:

- » **Workload security:** VCF has several built-in capabilities to improve workload security. These include Secure Boot, Virtual TPM, vSphere Trust Authority, and VM Encryption.
- » **Identity and access management:** VCF integrates with various identity and access management solutions, including VMware Identity Manager and third-party identity providers. This ensures that only authorized users and applications can access AI models and datasets.
- » **Network security:** VCF helps protect applications with advanced threat prevention at the network level via software-dedicated firewalls for applications and their associated AI models and datasets.
- » **Distributed Resource Scheduler (DRS):** This industry-leading capability helps achieve excellent cost optimization and workload performance by optimally placing workloads on hosts.
- » **Multi-tenancy:** With VCF 9.0's multi-tenancy capability, cloud service providers and enterprise administrators can enable secure and private environments for tenants on the same infrastructure and achieve high efficiency, scalability, and lower TCO.

» **GPU and vGPU monitoring improvements:** VCF offers strong GPU monitoring capabilities at the host, cluster, and VM levels, giving administrators deep visibility into GPU utilization. These capabilities help identify GPU over-provisioning or under-utilization, optimize TCO, accelerate issue resolution, and enhance overall performance.

## Explore Use Cases

With VMware Private AI Foundation with NVIDIA, enterprises can get excellent productivity to every department while maintaining the privacy and control of corporate data and IP.

The VMware Private AI Foundation with NVIDIA enables many use cases for enterprises. This section provides a number of examples of interesting applications.

### GPU-as-a Service

AI adoption can be accelerated with a scalable, on-demand, GPU infrastructure for enterprises and organizations. GPU as a service enables on-demand provisioning of capabilities like Jupyter Notebooks and models as a service, allowing organizations to accelerate data science and GenAI workflows on high-performance virtual infrastructure. This capability empowers business units to innovate independently while ensuring IT governance and cost control. The GPU as a service capability is enabled through virtualized GPUs, deep learning VM templates, self-service and AI blueprints, and NVIDIA AI Enterprise.

### Agentic AI

Agentic AI is the approach to enable automated real-time actions and accelerated results. Agentic AI is enabled by using the capabilities of the Agent Builder Service. This enables organizations to leverage AI-driven insights for operations like predictive maintenance, automation, and customer personalization while maintaining full control over their data and infrastructure. Agentic AI is enabled by harnessing capabilities such as NIM, NeMo Retriever, NVIDIA Blueprints, Agent Builder service, vector databases, Model Runtime, Data Indexing and Retrieval service, and more.

## Code generation/creation

Assisted code generation accelerates developer speed, but privacy and intellectual property (IP) contamination are of utmost concern when using public AI services. With this Private AI solution, a fine-tuned LLM will allow developer communities to provide confidential code examples and application programming interface (API) specifications while ensuring that the IP stays inside the enterprise. This capability can be enabled by deploying software tools like Tabnine on the VMware Private AI Foundation with NVIDIA platform.

## Contact centers resolution experience

For a contact center, it's vital to provide high-quality content and feedback to customers, which is enabled via summarization of complex documents for customer contact center staff. This can facilitate enhanced relevance and accuracy of responses from the contact centers and dramatically enhance customer experience. Having control of the input prompts to an LLM allows enterprises to protect and have visibility into user interactions (allowing for a better experience), alongside an understanding of the datasets and the foundation model that it was originally trained on. This use case is enabled by chatbots leveraging customized GenAI models on the VMware Private AI Foundation with NVIDIA platform by harnessing capabilities such as NIM, NeMo Retriever, NVIDIA Blueprints, vector databases, and Model Runtime, Data Indexing and Retrieval, and Agent Builder services.

## Document search and summarization

Increased productivity is another critical business value that you want to address infrastructure-wise. VMware Private AI Foundation with NVIDIA supports employee productivity by efficient document search and procedure research. The platform supports RAG chatbots to enable domain-specific trained LLMs to retrieve relevant information from additional authoritative predetermined knowledge sources. This use case is enabled by harnessing capabilities such as NIM, NeMo Retriever, NVIDIA Blueprints, vector databases, model runtime, Data Indexing and Retrieval service, and more.

## Private/secure content creation

VMware Private AI Foundation with NVIDIA enhances content creation in a secure environment by utilizing highly refined LLMs, which facilitates various apps, including creative writing, marketing campaigns, and product specifications. Fine-tuning an enterprise's private domain data helps provide a unique voice in their marketing materials, which is important because enterprises don't want AI to homogenize their communication styles, making them sound identical to their competitors. Like document search and summarization and contact centers resolution, this use case is enabled by harnessing capabilities such as NIM, NeMo Retriever, NVIDIA Blueprints, vector databases, Model Runtime, and Data Indexing and Retrieval.

# Chapter 7

## Ten Steps to Get Started with VMware Private AI Foundation with NVIDIA

Here are ten steps to get started with deployment of VMware Private AI Foundation with NVIDIA:

- » **Analyze the system architecture.** VMware Private AI Foundation with NVIDIA runs on top of VMware Cloud Foundation. However, the system architectural setup contains many different components, so before you jump in, analyze the *technical components* of the system architecture as well as the *management components* of the system architecture.
- » **Adhere to requirements for deployment.** These technical prerequisites include aspects such as VMware software versions, NVIDIA graphics processing unit (GPU) devices, NVIDIA software, and VMware Cloud Foundation setup.
- » **Create a content library.** Deep learning virtual machine (VM) images are distributed in a shared content library published by VMware. A cloud administrator should use a content library to pull specific VM images in the workload domain during VM deployment.

- » **Configure the supervisor on your cluster.** In order to provide DevOps and MLOps engineers and data scientists with the ability to deploy deep learning VMs or VKS clusters with AI container workloads, you must deploy a supervisor on GPU-enabled cluster in a workload domain and create virtual GPU (vGPU)–enabled VM classes.
- » **Set up a private harbor registry.** A *private harbor registry* is a local image registry in a disconnected environment where you host the container images downloaded from the NVIDIA NGC catalog.
- » **Ensure access to NVIDIA AI Enterprise components.** In a disconnected environment, you should upload the components of the NVIDIA AI Enterprise to internal locations.
- » **Initiate VCF Automation.** VCF Automation offers support for self-service catalog items that DevOps and MLOps engineers and data scientists can use to easily provision AI workloads.
- » **Run the Private AI Quickstart process in VCF Automation to generate your Catalog Items.** Configure the Private AI services in your supervisor cluster.
- » **Build your first RAG application.** You can do this by deploying a deep learning VM, provisioning an AI Kubernetes cluster, or using the higher-level services.
- » **Use the Model Runtime, Data Indexing and Retrieval, and Agent Builder services.** This allows you to chunk, index, and vectorize data and make it available through updatable knowledge bases with a configurable refresh policy that ensures that the model output remains current. Furthermore, using the Agent Builder service allows your generative AI (GenAI) application developers to build AI agents by using resources from the Model Store, Model Runtime, and Data Indexing and Retrieval Service.

Visit <https://techdocs.broadcom.com/us/en/vmware-cis/private-ai/foundation-with-nvidia/9-0/private-ai-foundation-9-x.html> for the full implementation guide.

# Private AI that puts you in control of your data?

## Smart.

Learn more at  
[vmware.com/aiml-nvidia](https://vmware.com/aiml-nvidia)



The smarter way  
to cloud.

**vmware**<sup>®</sup>  
by Broadcom

# Unlock AI and unleash productivity with lower TCO

VMware Private AI Foundation with NVIDIA enables enterprises to customize and fine-tune large language models (LLMs) and deploy secure and private models for their internal production workloads. These private models can run inferencing and drive automated decision-making against an enterprise's private data sets. Leveraging these fine-tuned AI models, enterprises across all industries will be able to unleash the productivity in every department of the organization.

## Inside...

- Explore the evolution of AI
- Pinpoint the challenges
- Select a secure and compliant infrastructure
- Manage AI deployments
- Optimize infrastructure
- Unlock AI with VMware Private AI Foundation with NVIDIA
- Get started with AI deployments

**vmware**<sup>®</sup>  
by **Broadcom**

**Ulrika Jägar** is Head of Commercial AI and Data at Scania Connected Vehicles and the author of *Operating AI: Bridging the Gap between Technology and Business* (Wiley). **Shobhit Bhutani**, a marketing leader at Broadcom, has more than 25 years of experience driving innovations in AI and enterprise technology.

Go to **Dummies.com**<sup>™</sup>  
for videos, step-by-step photos,  
how-to articles, or to shop!

ISBN: 978-1-394-34789-6

Not For Resale

**for**  
**dummies**<sup>®</sup>  
A Wiley Brand



# **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.