

# VMware as the *foundation* for AI/ML workloads



**Kobi Shamama**

Sr. Sales Manager, Tanzu Division

VMware Israel



# Set the table



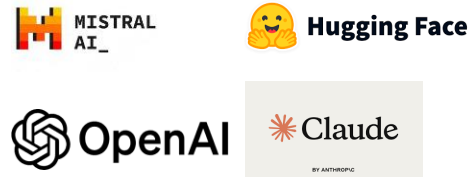
## Generative AI

Sub domain of AI that can **generate new data, like text, images, videos, or even code.**  
It's trained on massive amounts of data. This data could be anything from text articles and books to collections of images or videos. Once trained, the AI can use its knowledge to create new content.



## LLM

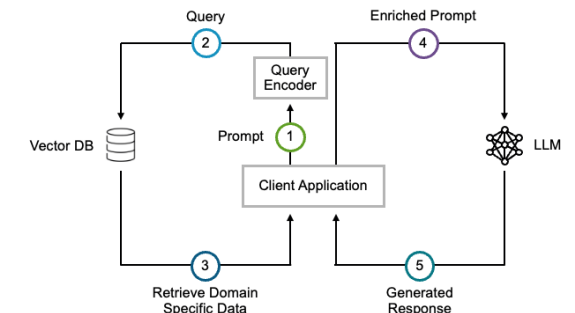
LLMs are a **specific type of generative AI** focused on processing and **generating text.** They're essentially like the language experts of the generative AI world. Trained by massive amounts of text data and consists of a **pre-trained language model.**



## RAG

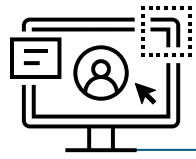
**Business Data + LLM = RAG**  
Retrieval-Augmented Generation. AI technique specifically **designed to improve the accuracy and reliability** of generative models, particularly LLMs.

1. Enrich general LLMs (like open ai) with your domain specific data
2. Leveraging VectorDB as engine for similarity check
3. Disruptive technology that can increase revenue for companies



# Paradigms for inserting knowledge

Retrieval Augmented Generation – feed only relevant information into the prompt



Input Prompt

How the Kneset handle the climate change??

```
[56]: response = query_engine.query("What were the main concerns raised by Michael Eitan regarding the Ministry of Environmental Protection?")
[57]: print(textwrap.fill(str(response)))

The main concerns raised by Michael Eitan regarding the Ministry of Environmental Protection include: 1. The need to reform the current system and not continue with the status quo. He emphasizes, "מוכרחים 2" לעשות בטניין מה סדר ואי-אפשר להמשיך במצב הקיים, אני חושב שאננו 3" Focus on the importance of environmental issues and recognizing the value of the Ministry of Environmental Protection. He highlights, "יש ליפתי הערות. מוכרחים 4" The need for proper oversight and management of environmental concerns. Michael Eitan mentions the lack of staff at the Ministry of Environmental Protection, stating "11 הייתה כוללת 11" All these main concerns were brought up by Michael Eitan during a discussion in the Israeli Kneset plenary session.
```

**HaifaCLGroup KnesetCorpus**

Tasks: Text Classification, Text Generation | Language: Hebrew | Size Categories: 10M+~100M | ArXiv: arXiv:2405.18115

License: cc-by-sa-4.0

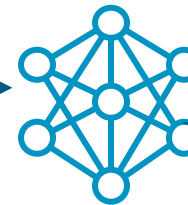
**The Kneset (Israeli Parliament) Proceedings Corpus**

[\[Github Repo\]](#) • [\[Paper\]](#) • [\[ES kibana dashboard\]](#)

**Dataset Description**

An annotated corpus of Hebrew parliamentary proceedings containing over 32 million sentences from all the (plenary and committee) protocols held in the Israeli parliament from 1992 to 2022. Sentences are annotated with various levels of linguistic information, including part-of-speech tags, morphological features, dependency structures, and named entities. They are also associated with detailed meta-information reflecting demographic and political properties of the speakers, based on a large database of parliament members and factions that we compiled.

Curated by: [Gili Goldin \(University of Haifa\)](#), Nick Howell (IAHLT), Noam Ordan (IAHLT), Ella Rabinovich (The Academic College of Tel-Aviv Yaffo), Shuly Wintner (University of Haifa)



LLM

Knowledge-Augmented Prompt

Use this context to answer the question:  
In space flight, "attitude" refers to orientation.

Given the context, answer the following question:  
How the Kneset handle the climate change?

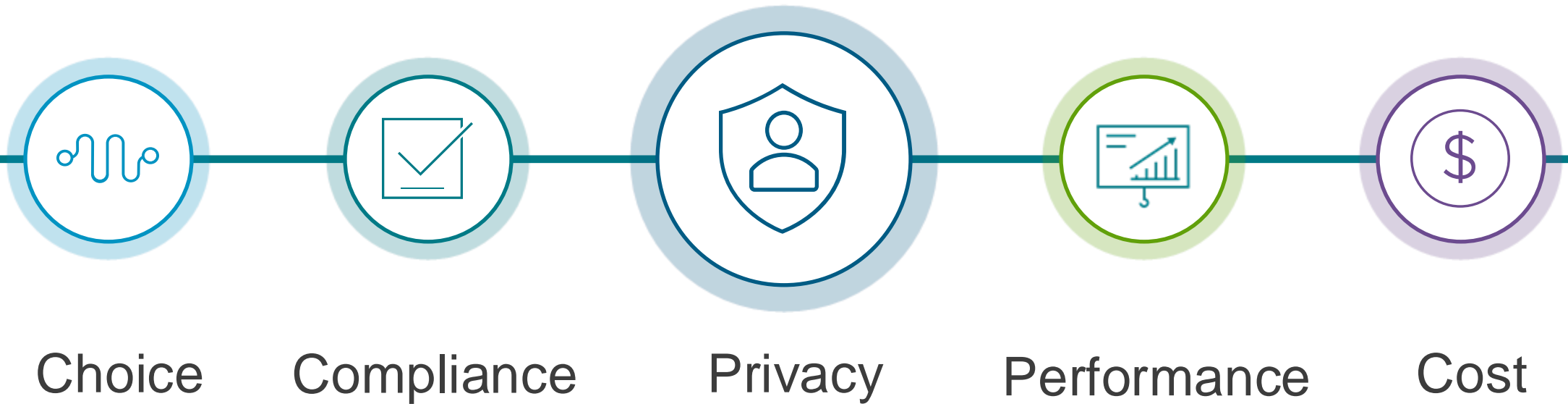


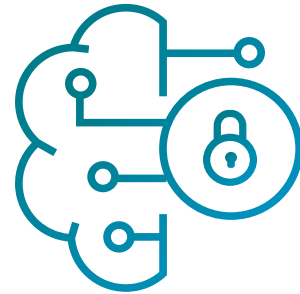
80% of CIOs are investing heavily in Generative AI as their competitive edge for the future

- KPMG Global CEO Outlook 2023

Enterprises that adopt next-generation AI like LLMs and Generative AI are **2.6X more likely to increase revenue by 10%** or more but must invest in their AI infrastructure to fully reap the benefits.

# Key Challenges in Generative AI Today





# VMware Private AI


An architectural approach that balances the business gains from AI with the privacy and compliance needs of the organization.

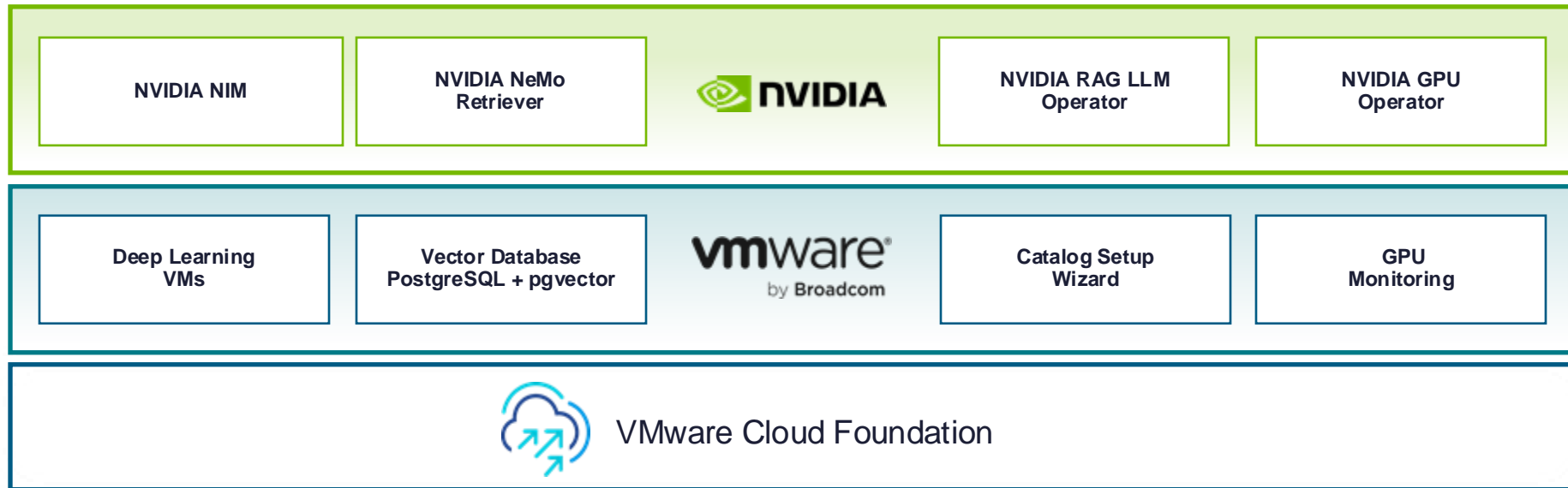
# VMware Private AI Foundation with NVIDIA

General Availability

 NVIDIA Foundation Models

 NVIDIA Fine-Tuned Models

 Third party & Community Models



 Dell Technologies

 Hewlett Packard Enterprise

 Lenovo

 VMware  
by Broadcom

Broadcom Proprietary and Confidential. Copyright © 2024 Broadcom.  
All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.



# Enable Rapid Deployment of Your Deep Learning Projects

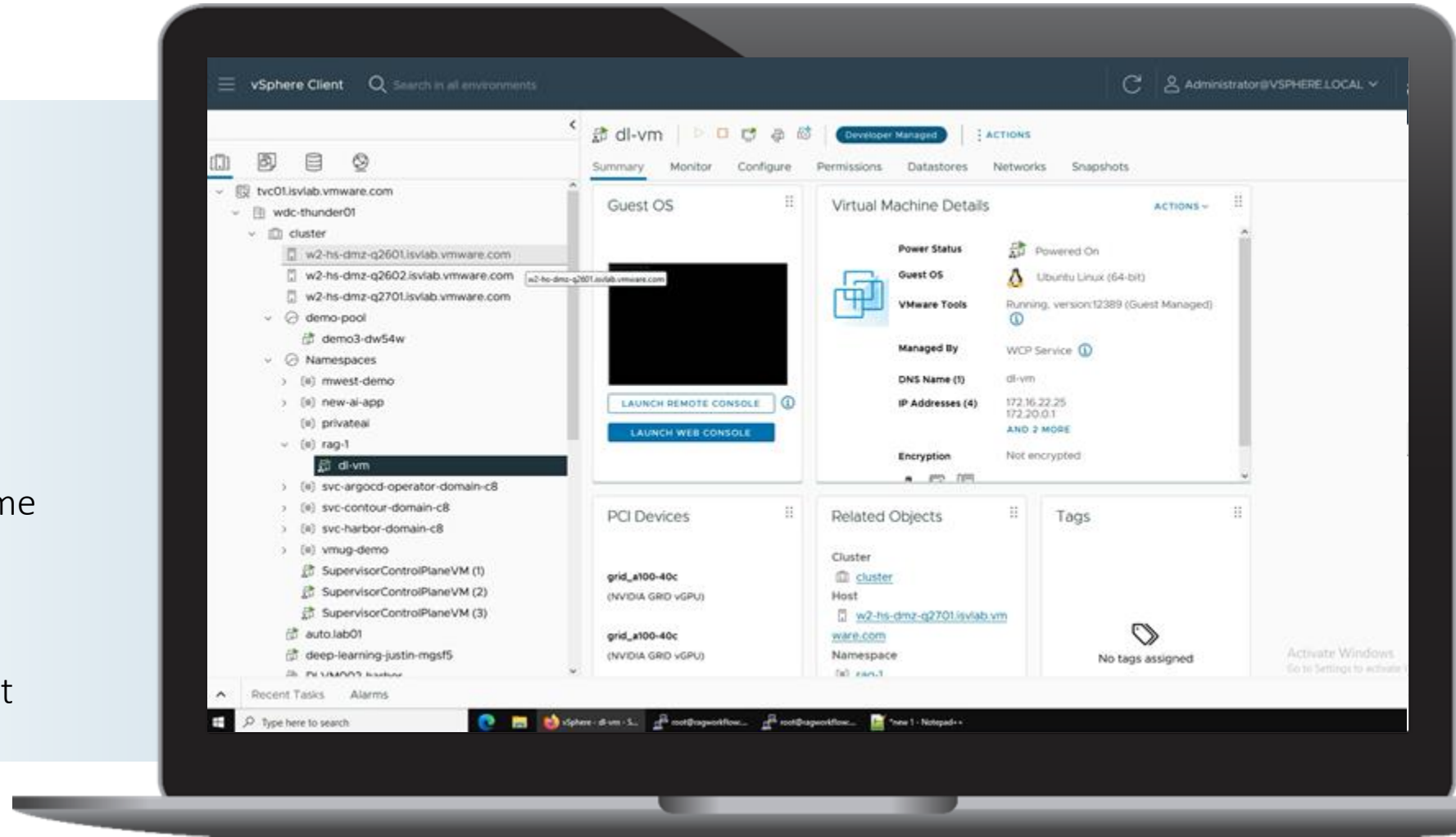
Simple to setup and maintain



Get a scalable and flexible environment for deep learning projects

Use deep learning VMs which come pre-configured for vGPU use.

NVIDIA drivers and specialized containers are downloaded at first boot.



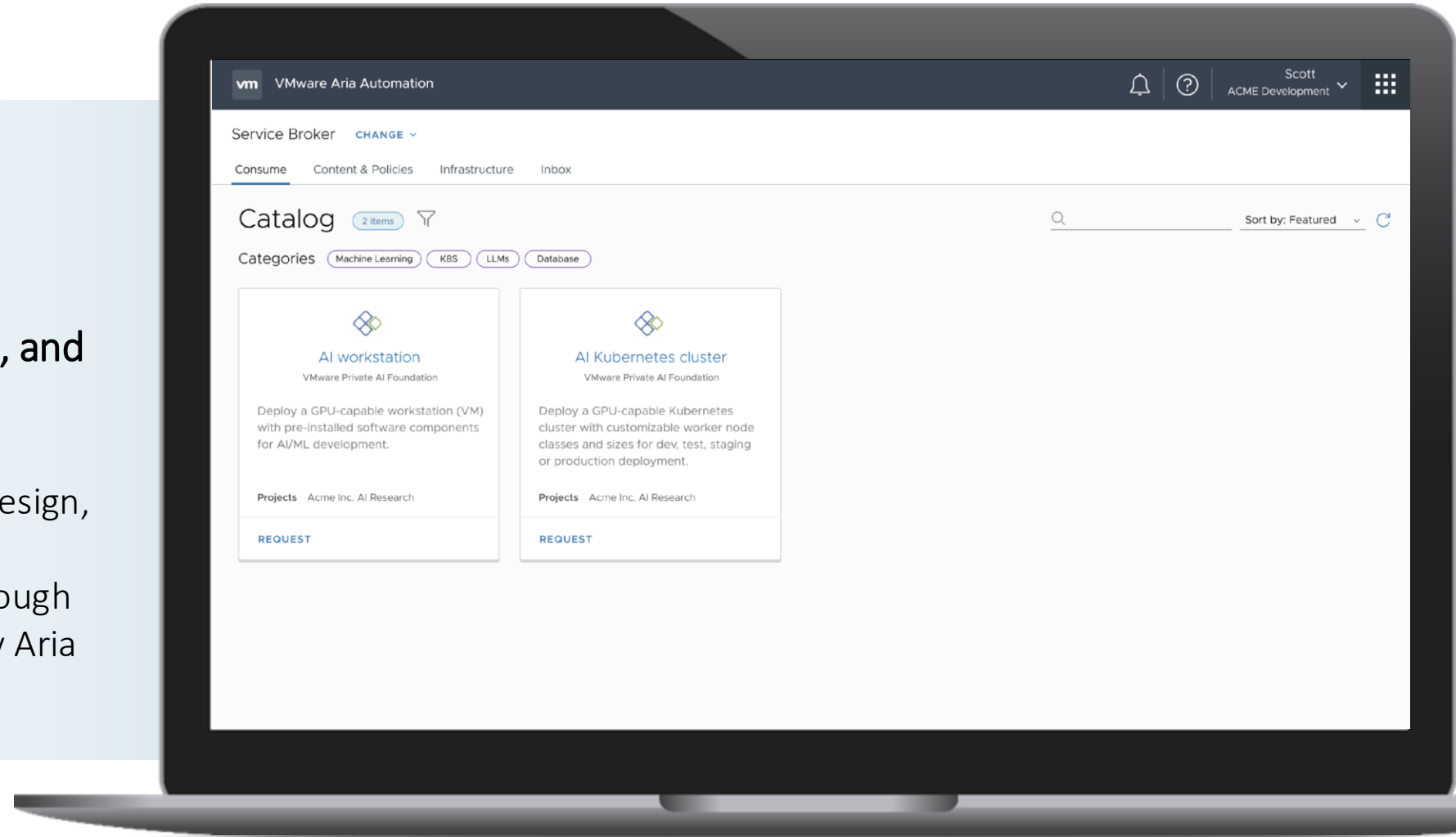
# Quickly Deploy Complex AI Infrastructure Objects

Simple to Consume



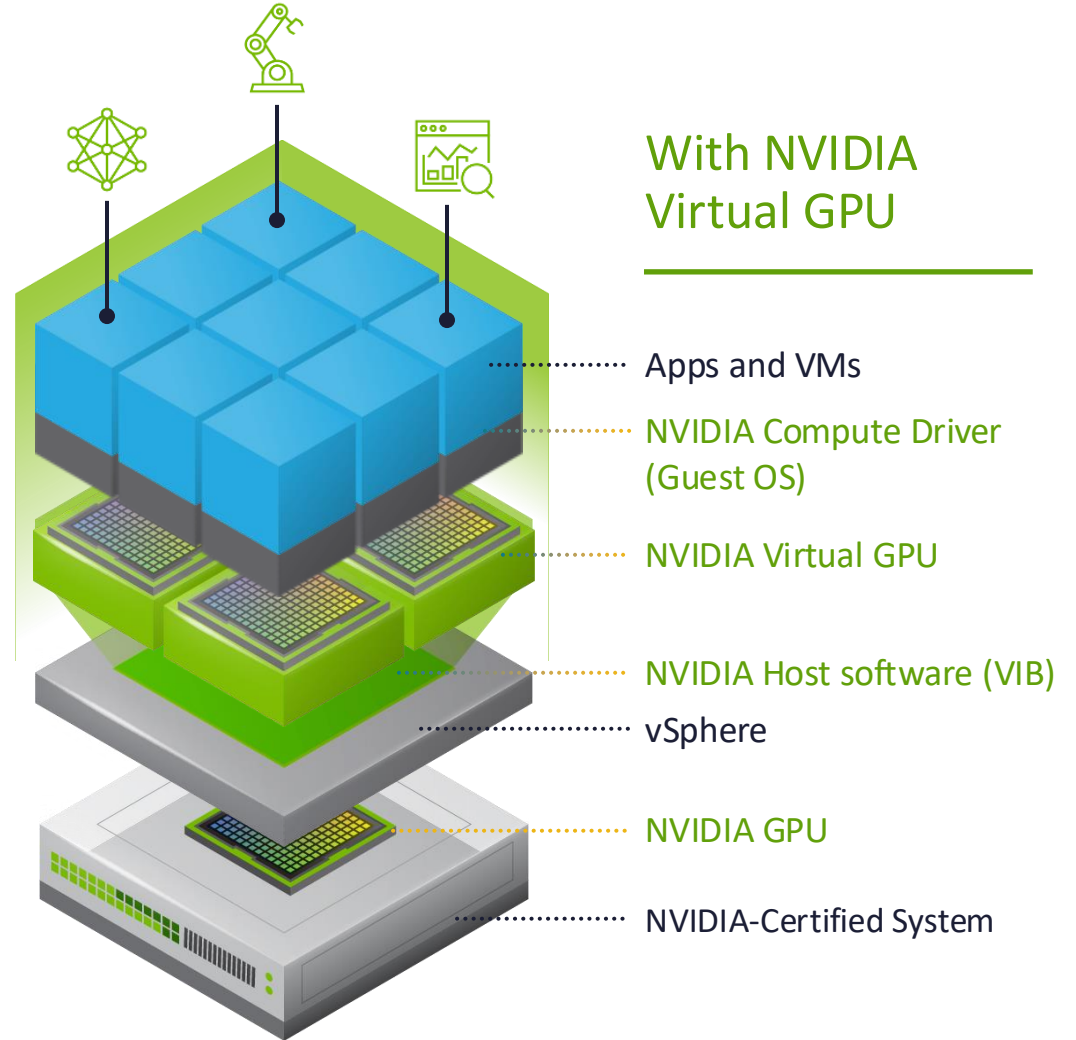
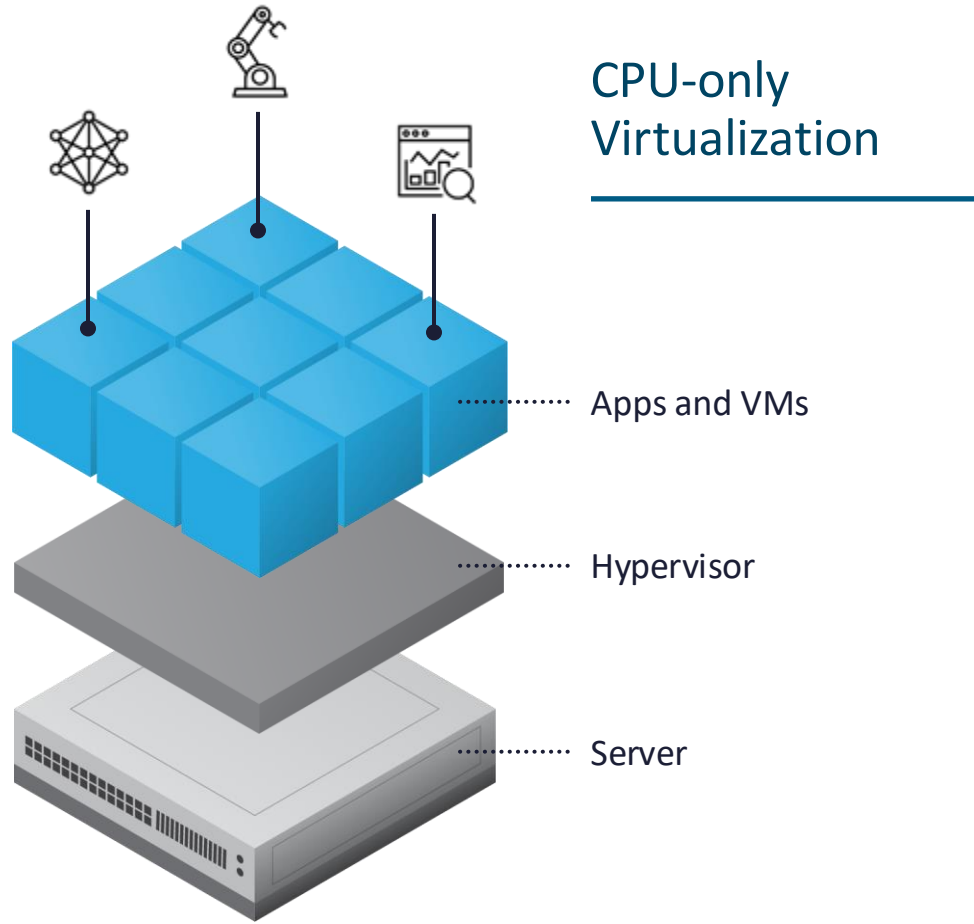
Simplify Day 0 design, curation, and offering of complex objects

Enabling LOB Admins to quickly design, curate, and offer optimized AI infrastructure catalog objects through VCF's self-service portal (formerly Aria Automation).



# GPU Virtualization Accelerates Compute Workloads & Reduce Costs

Reduce cost and increase efficiency



# Vector Database

DB as a Service

PostgreSQL + pgvector Support



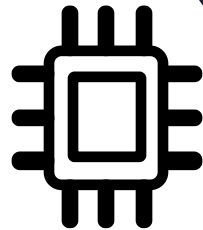
DSM 2.0

Included in VCF



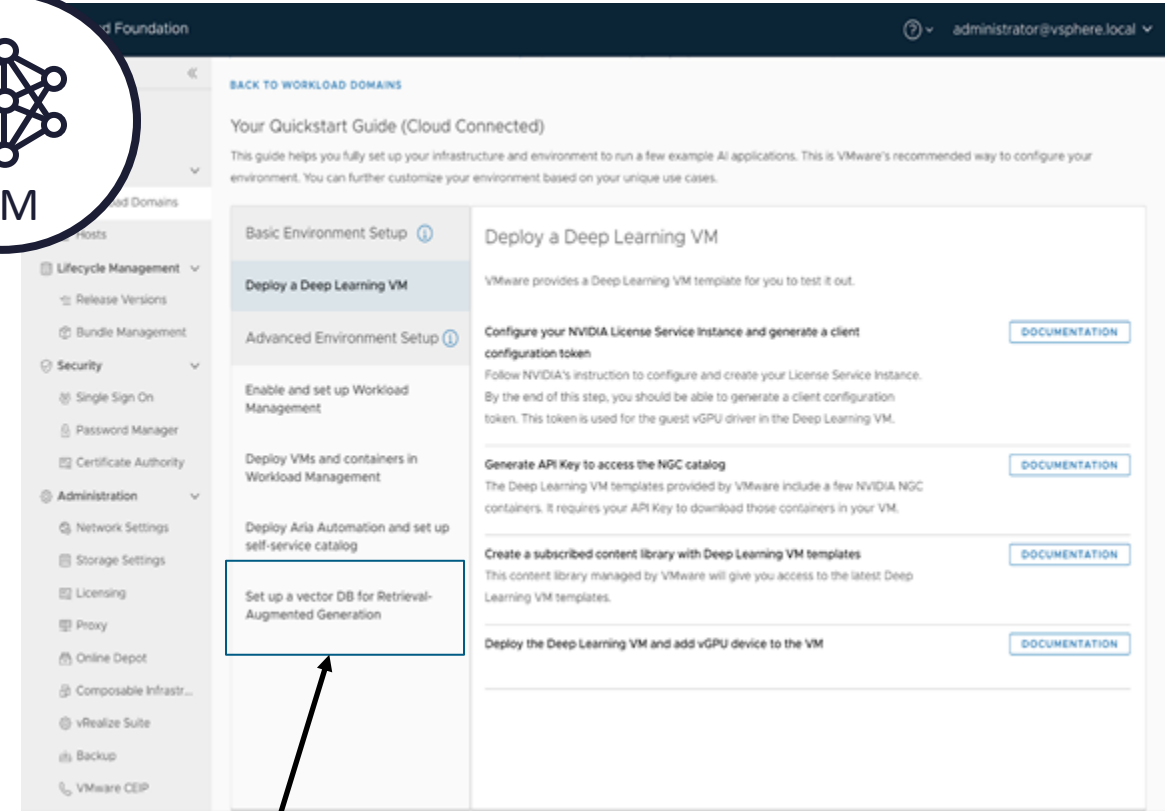
PostgreSQL

Relational Database



pgvector

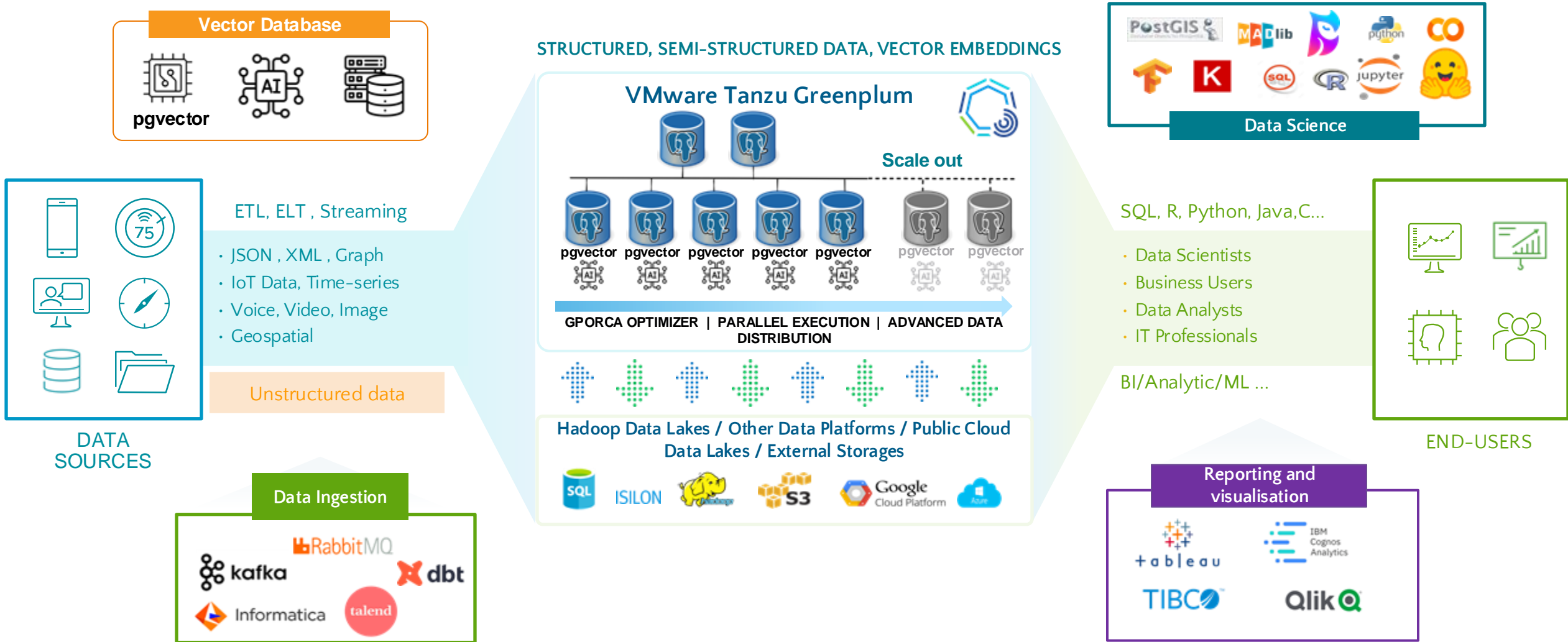
Extension



Automated Setup

# VMware Tanzu Greenplum - Massive Scalability from TB to Multiple PB

Modern Data Platform for Analytics & AI; Scalable VectorDB



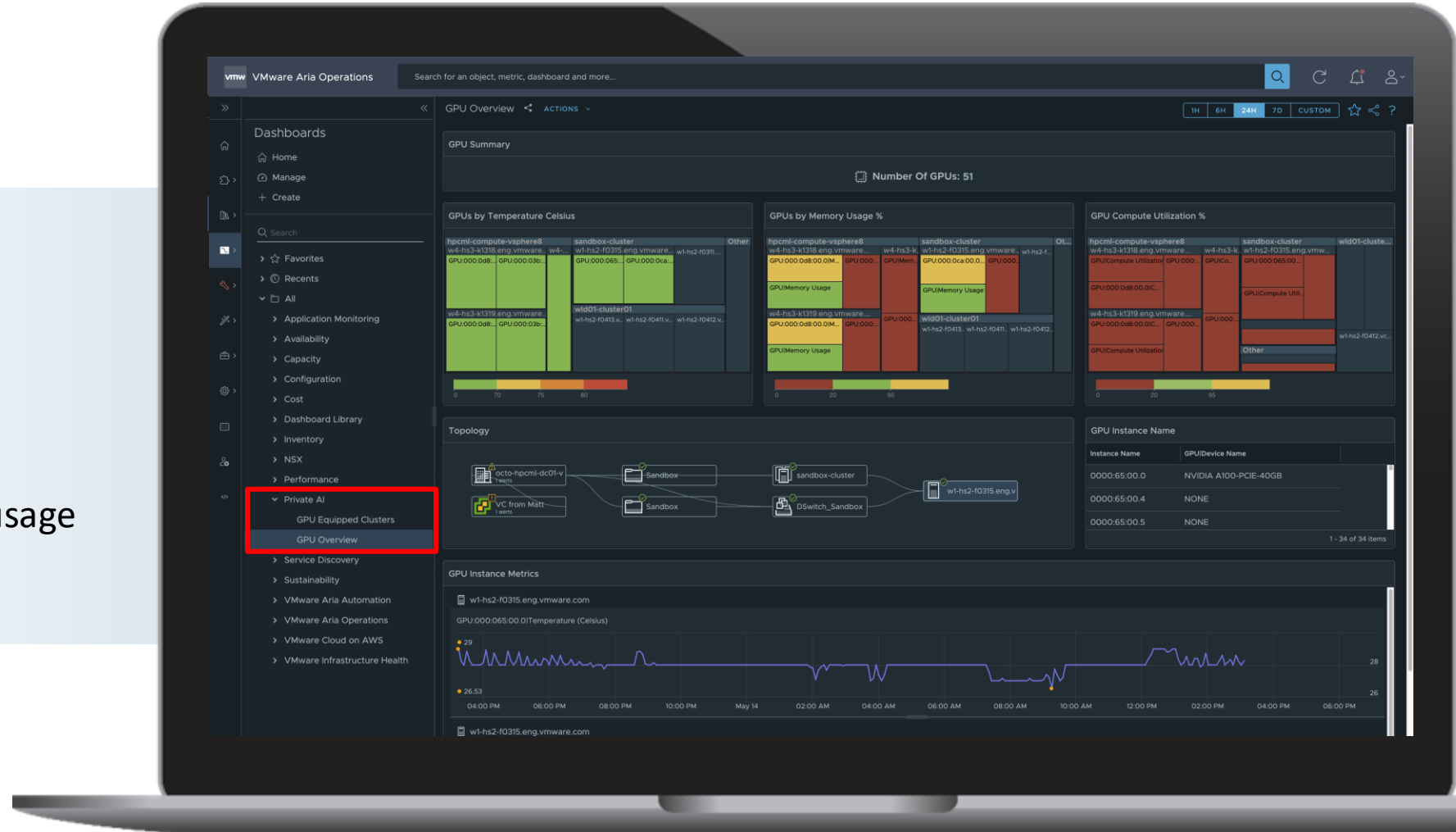
# Improve GPU Usage With Monitoring Capability

## Advanced GPU Monitoring



## Get Enhanced GPU visibility

Get GPU compute and memory usage across hosts and clusters



# Tanzu AI complements Private AI Foundation

Accelerate application delivery for new generative AI-powered applications

Accelerate developer productivity

## VMware Tanzu AI

Empower Java devs to develop powerful AI apps

### Tanzu Spring AI

Developer friendly Java APIs to access LLM models and implement higher level AI app logic patterns

Quickly broker AI models & data stores in Tanzu Platform

### Tanzu Marketplace

Service broker for self-service access to curated software

Securely tune AI model access w/ governance

### Tanzu AI Server (beta)

Enhanced OpenAI compatible API gateway with load balancing, rate limiting, and caching

Choice in enterprise data solutions for your AI apps

### Tanzu Data

Developer-ready, managed database solutions that complement Tanzu Platform

## Nvidia partnership

Nvidia NIM

Nvidia NeMo Retriever

Nvidia RAG LLM Operator

Nvidia GPU Operator

## VMware Private AI Foundation

Deep learning VMs

Vector Database

Catalog Setup Wizard

GPU monitoring



## Key Value Provided by VMware Private AI

- **Reduced Cost**
- **Improved TTM;**
- **Start fast, scale faster**
- **Developer Productivity**
- **Increased Security**



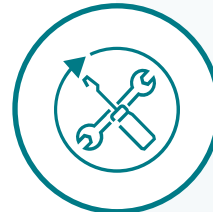
### Agility

Reconfigure or re-allocate hardware in minutes  
Avoid bare-metal silos, share GPU resources



### Scale

Right-size compute to match demand  
Provision new resources quickly  
Rapid configuration of new hardware



### Enhanced Lifecycle Management

Streamline Maintenance, Upgrades, Patches Increase  
Reliability and Uptime  
Self-service and Automation



### Privacy & Security

Secure Boot, Host Lockdown, VM Encryption  
and more...





# Thank You