

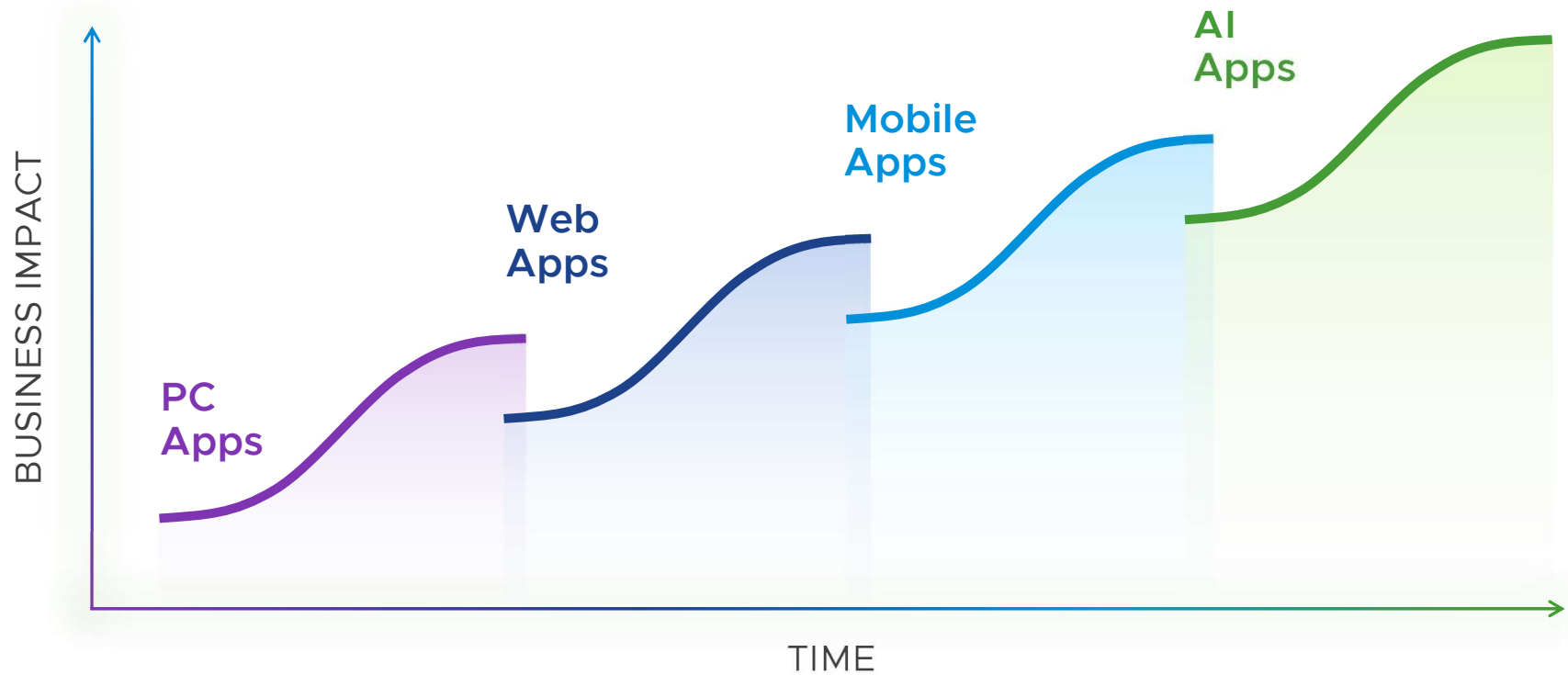


# 기업의 Private AI 구축을 위한 Reference Architecture Model

SEOK KEUN YOO Ph.D.(syoo@vmware.com)  
Solution Engineering

September 13, 2023

# The Next Wave of App Innovation



## PREDICTIVE AI

Data Scientists



Predictive Analytics



# SPECIALIZED AI MODEL

## PREDICTIVE AI

Data Scientists



Predictive Analytics



**SPECIALIZED AI MODEL**

## GENERATIVE AI



Marketing



Supply Chain



Sales



Human Resources



Customer Operations



Research and Development



Legal



Software Development



Manufacturing



Procurement



IT



Finance

NATURAL LANGUAGE INTERACTION

APIs



**LARGE LANGUAGE MODELS**

## GENERATIVE AI



NATURAL LANGUAGE INTERACTION

APIs

∨  
**LARGE LANGUAGE MODELS**

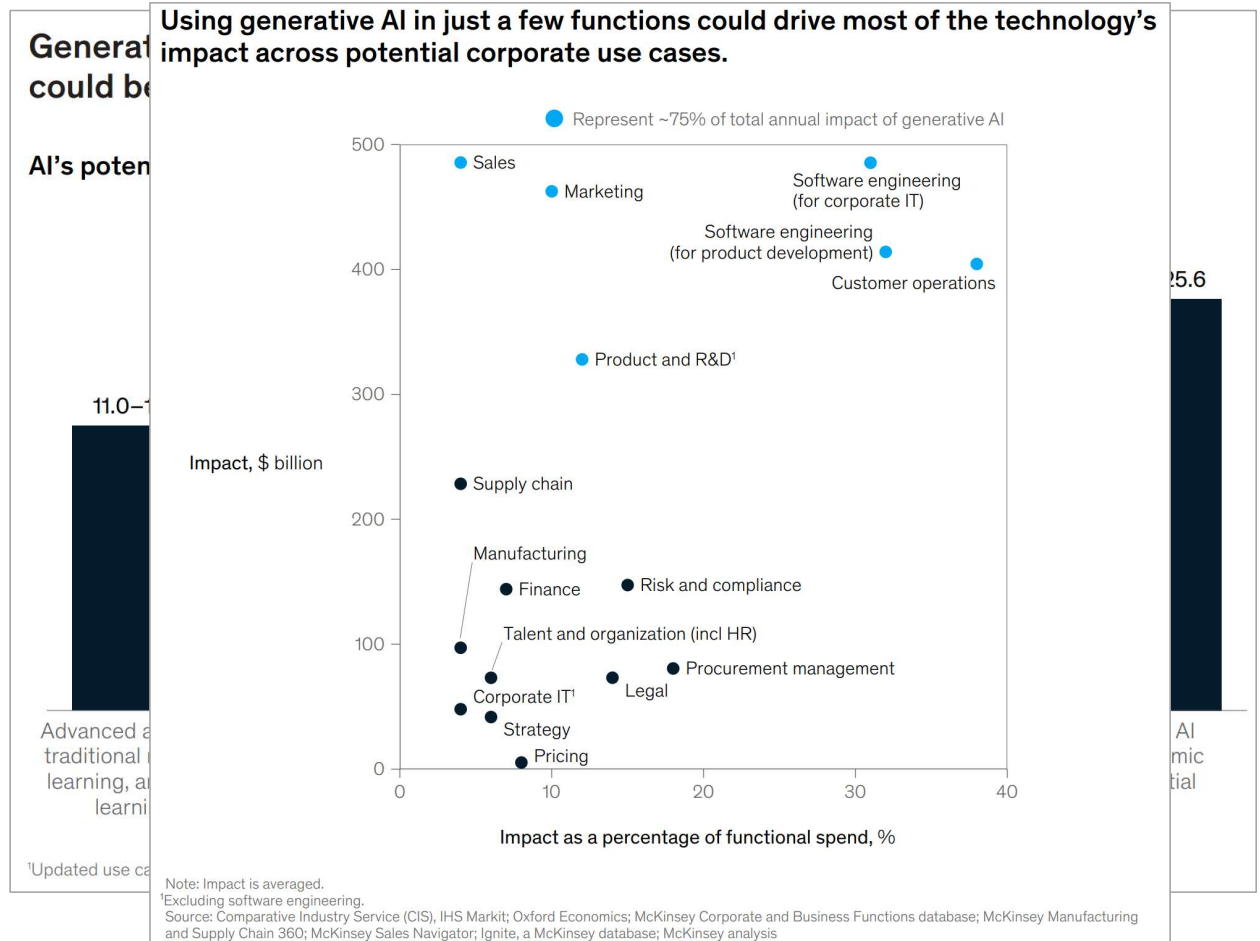
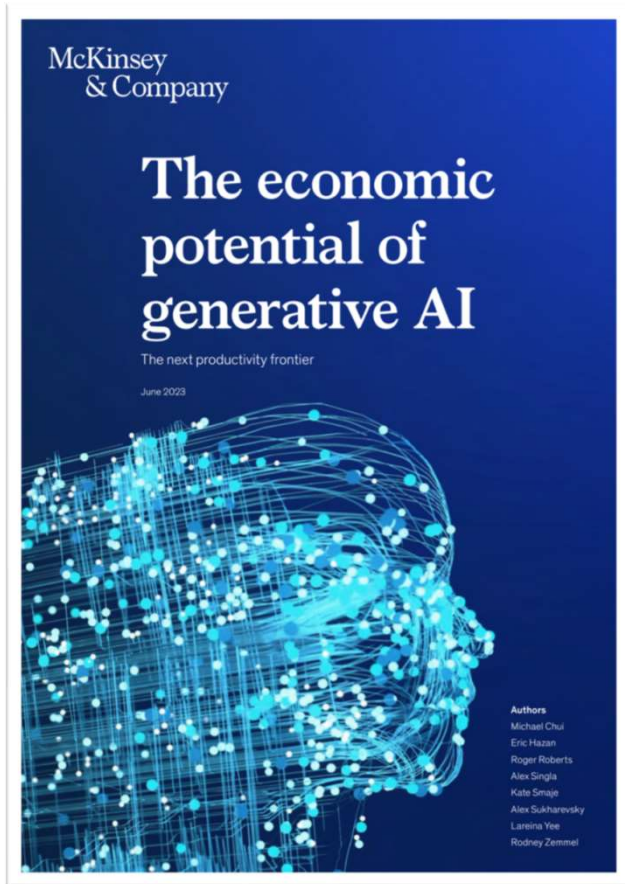
**\$4.4T**  
annual  
economic  
value

# GENERATIVE AI

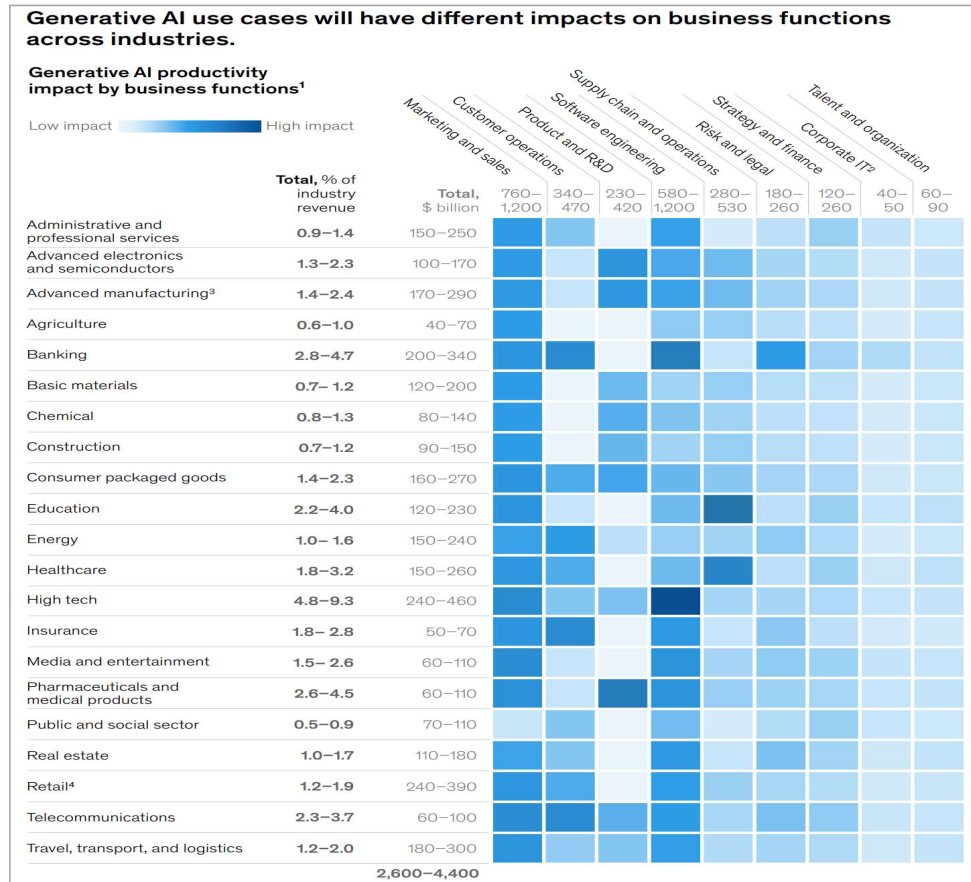
---

## MULTI-CLOUD

# The economic potential of generative AI



# Generative AI use cases

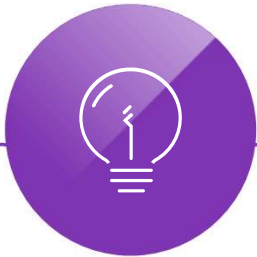


# Generative AI Will Boost Productivity for All Workers



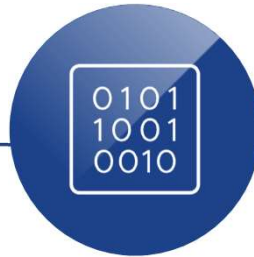
# Core Challenge: Privacy

## PRIVATE INTELLECTUAL PROPERTY



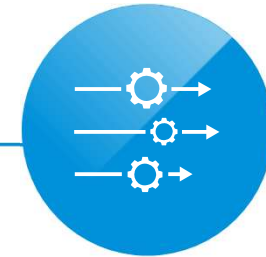
“I need to minimize my intellectual-property risk.”

## PRIVATE DATA



“I need to ensure my private data will never be shared externally.”

## PRIVATE ACCESS



“I need complete control over access to my AI models.”

1

Choose  
a Model

2

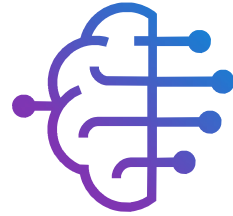
Domain-Specific  
Fine Tuning

3

Inferencing



# VMware AI Labs



# Private AI

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

# Fuel for AI: Data Spread Across Multiple Clouds

1

Choose  
a Model

2

Domain-Specific  
Fine Tuning

3

Inferencing

Data + Computing = AI



Public Cloud



Private Cloud



Edge

# Private AI Addresses Key Enterprise Requirements



PRIVACY



CHOICE



COST



PERFORMANCE



COMPLIANCE

INTRODUCING

# VMware Private AI Foundation WITH NVIDIA

INTRODUCING

# VMware Private AI Foundation

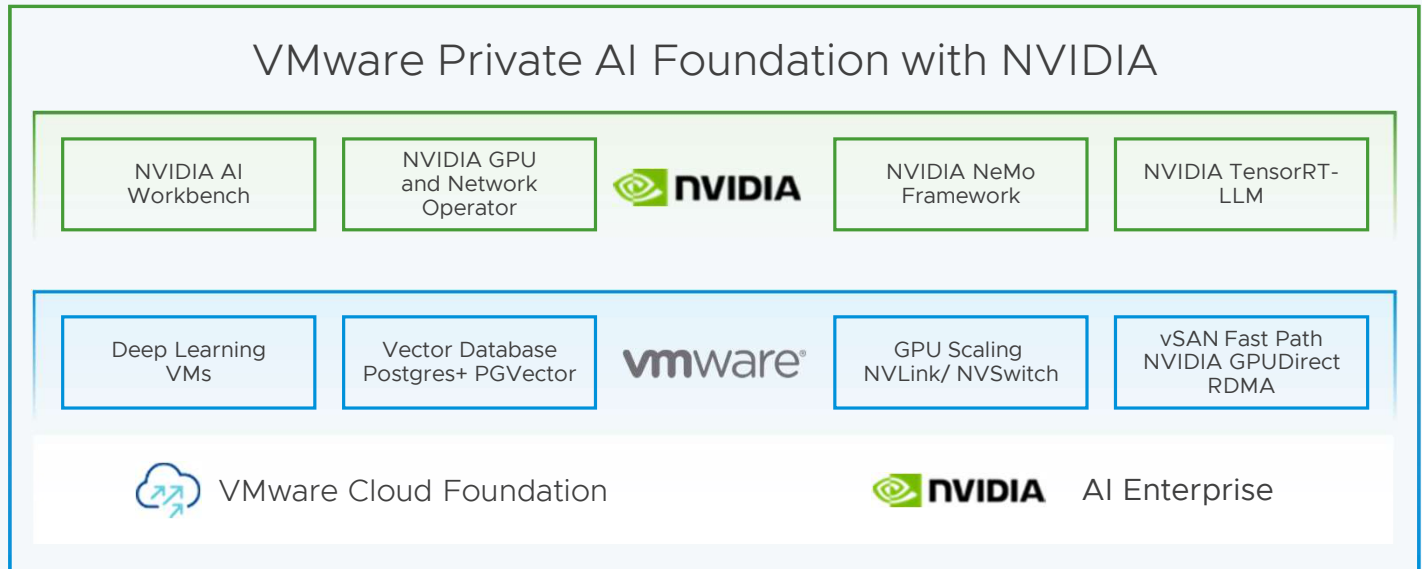
WITH NVIDIA

  
Falcon LLM

  
Llama 2

  
MPT

  
NVIDIA NeMo



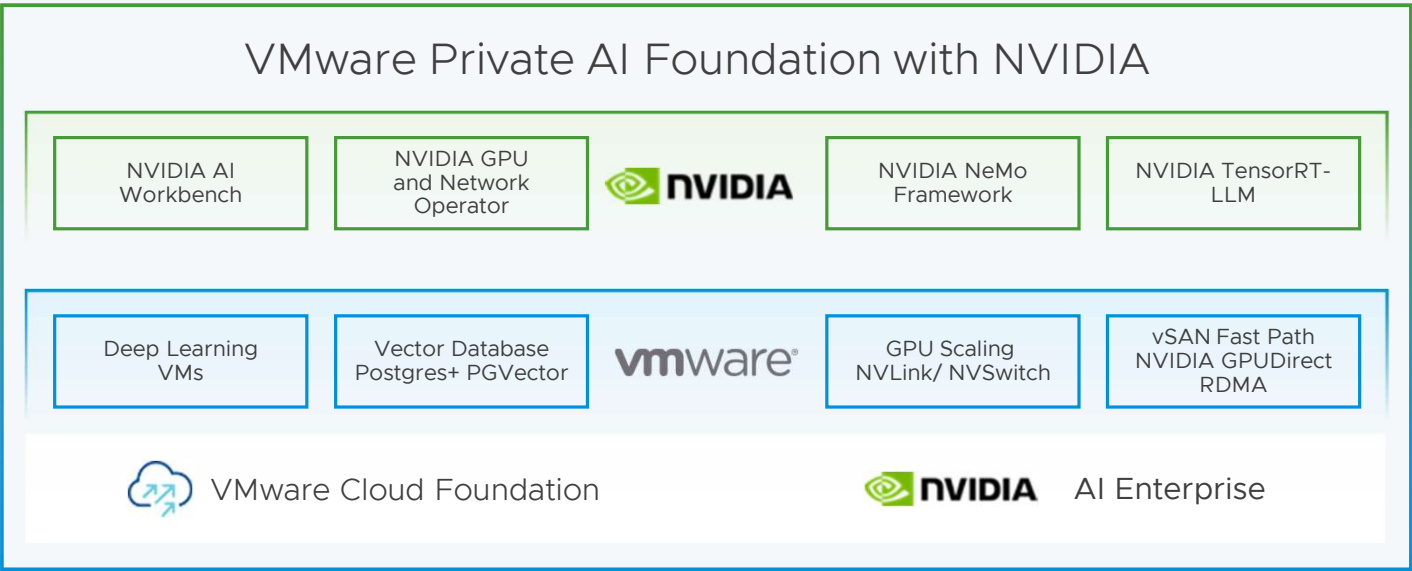



INTRODUCING

# VMware Private AI Foundation WITH NVIDIA

Falcon LLM      Llama 2      MPT      NVIDIA NeMo

## VMware Private AI Foundation with NVIDIA



 VMware Cloud Foundation

 NVIDIA AI Enterprise

 Dell Technologies

  
Hewlett Packard  
Enterprise

 Lenovo

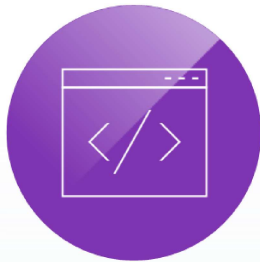
Choice of  
LLMs

Bare-metal  
Performance

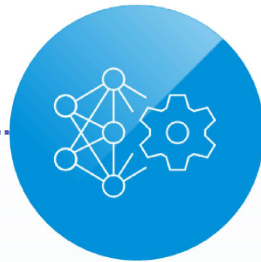
Faster  
time-to-value

# Generative AI from Customization to Deployment

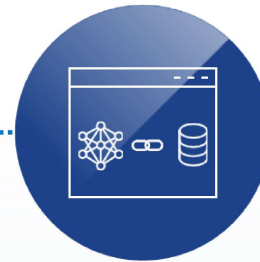
CREATE  
PROJECT



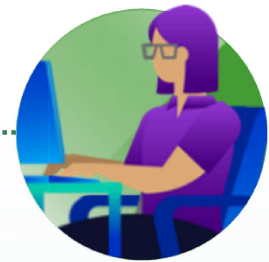
CUSTOMIZE  
MODEL



BUILD AI  
CO-PILOT



DEPLOY AI  
CO-PILOT



VMware Private AI Foundation  
**WITH NVIDIA**

INTRODUCING

# VMware Private AI Foundation

# Reference Architecture for VMware Private AI

 Falcon LLM

 Llama 2

 Hugging Face

 MPT

 NVIDIA NeMo

 RAY

 Kubeflow  
VMware Distribution

 PyTorch

Deep Learning VMs

Vector DB

GPU Scaling

vSAN Fast Path

 VMware Cloud Foundation

  
Compute

  
Storage

  
Network/DPU

  
GPU

# VMware Private AI Open Ecosystem



Dolly



Falcon



Llama 2



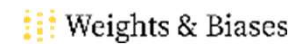
MPT



Platypus 2

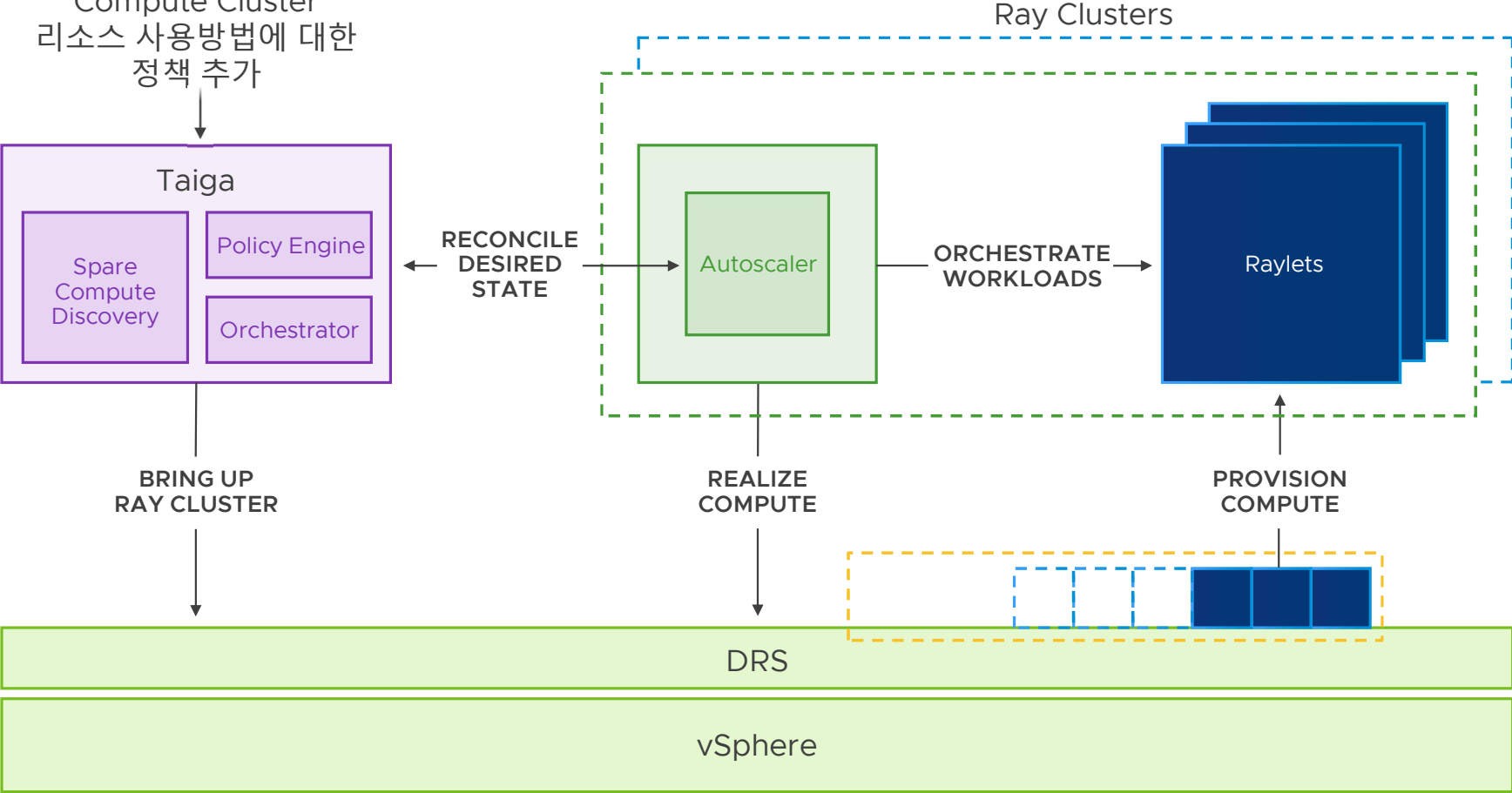


cnvrg.io



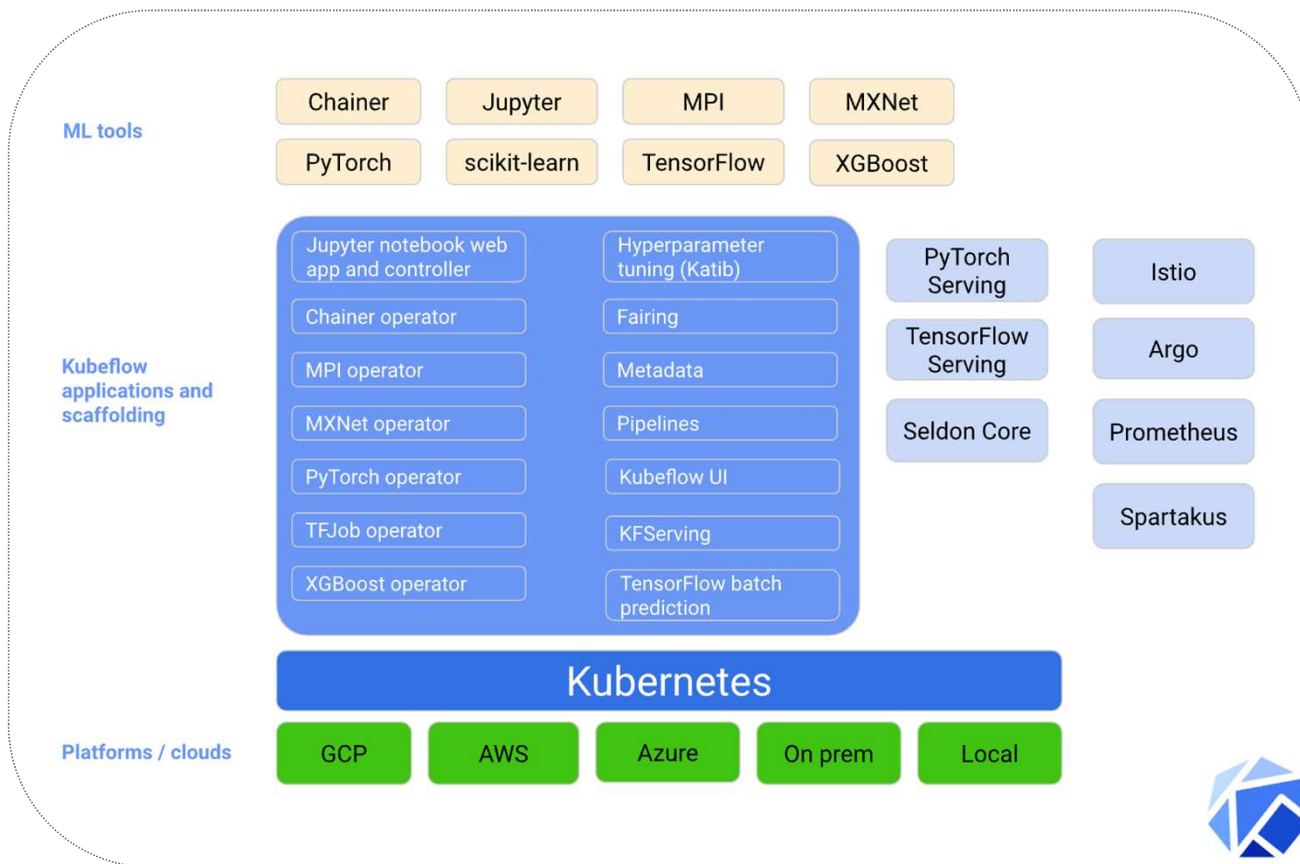
# VMware Project Taiga

VI admin이 Spare Compute Cluster 리소스 사용방법에 대한 정책 추가



# Kubeflow

Kubernetes 기반의 Machine Learning 플랫폼



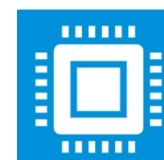
## 누가 Kubeflow를 사용하나요?

- ML 파이프라인을 구축하고 실험하려는 **데이터 과학자**
- 다양한 환경에 ML 시스템을 배포하려는 **ML 엔지니어/운영팀**



## 어떤 툴이 제공되나요?

- 모델 개발/학습/추론을 위한 **Jupyter Notebook**
- 도커 컨테이너 기반 Multi-Step ML 워크플로를 구축/배포/관리하기 위한 **Kubeflow Pipelines**



## 어떻게 Training/Serving 하나요?

- ML 모델의 training을 위한 커스텀 **TensorFlow training job operator** 제공
- 훈련된 TensorFlow 모델을 쿠버네티스로 내보내는 **TensorFlow Serving** 컨테이너 지원

1

Choose  
a Model

2

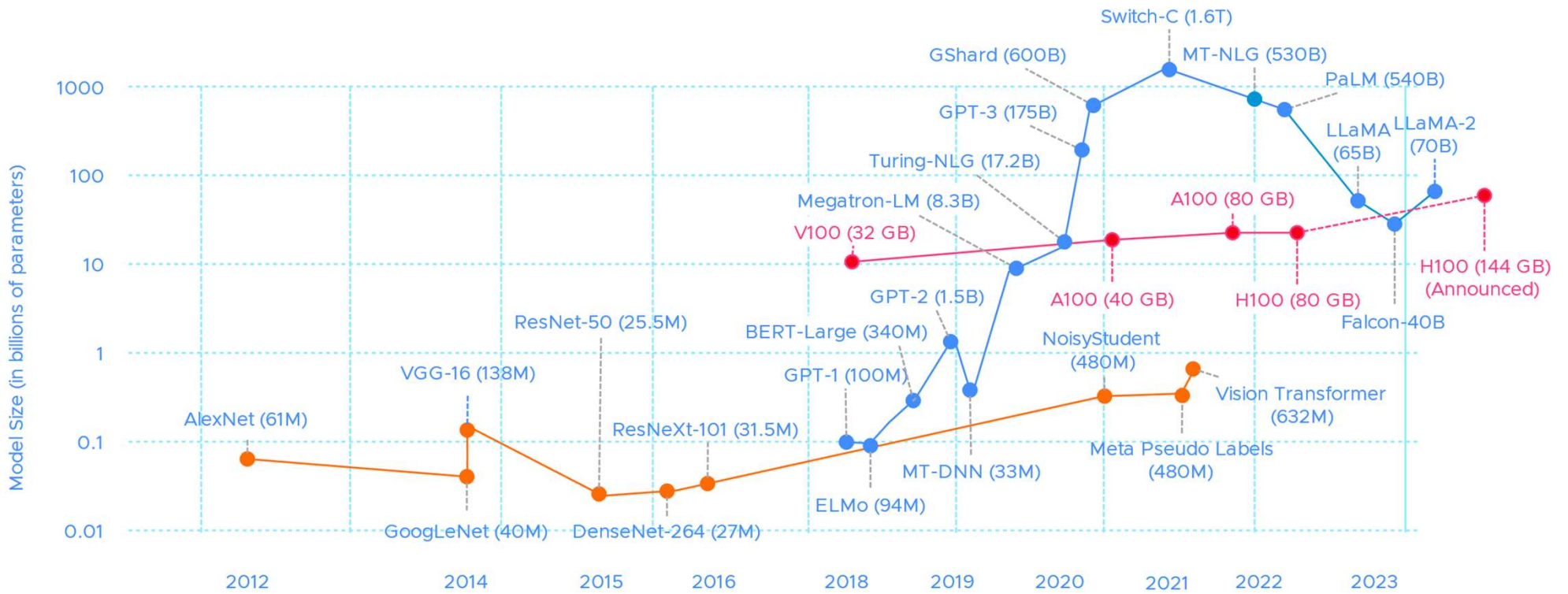
Domain-Specific  
Fine Tuning

3

Inferencing



# Model Size



Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation

Models 302,669 Filter by name

new Full-text search Sort: Trending

- stabilityai/control-lora**  
Text-to-Image • Updated 4 days ago • 304
- meta-llama/Llama-2-7b**  
Text Generation • Updated Jul 19 • 2.01k
- facebook/seamless-m4t-large**  
Updated about 9 hours ago • 133
- Open-Orca/OpenOrca-Platypus2-13B**  
Text Generation • Updated 3 days ago • 15.3k • 143
- diffusers/controlnet-canny-sdxl-1.0**  
Text-to-Image • Updated 11 days ago • 12.1k • 298
- garage-bAInd/Platypus2-70B-instruct**  
Text Generation • Updated 4 days ago • 3.77k • 108
- stabilityai/stablecode-instruct-alpha-3b**  
Text Generation • Updated 15 days ago • 7.02k • 232
- stabilityai/stable-diffusion-xl-refiner-1.0**  
Text-to-Image • Updated 17 days ago • 322k • 723

- stabilityai/stable-diffusion-xl-base-1.0**  
Text-to-Image • Updated 19 days ago • 856k • 2.13k
- Deci/DeciCoder-1b**  
Text Generation • Updated 1 day ago • 2.59k • 146
- meta-llama/Llama-2-7b-chat-hf**  
Text Generation • Updated 14 days ago • 474k • 875
- meta-llama/Llama-2-70b-chat-hf**  
Text Generation • Updated 14 days ago • 174k • 1.1k
- defog/sqlcoder**  
Text Generation • Updated 1 day ago • 430 • 71
- runwayml/stable-diffusion-v1-5**  
Text-to-Image • Updated Jul 4 • 7.82M • 9.02k
- TheBloke/Llama-2-7B-Chat-GGML**  
Text Generation • Updated 29 days ago • 8.42k • 380
- THUDM/chatglm2-6b**  
Updated Jul 19 • 2.46M • 1.55k

1

Choose  
a Model

2

Domain-Specific  
Fine Tuning

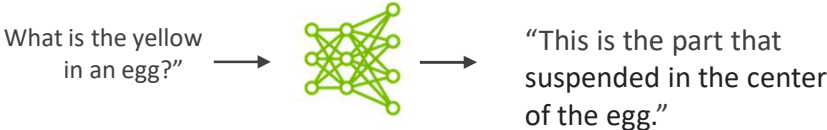
3

Inferencing

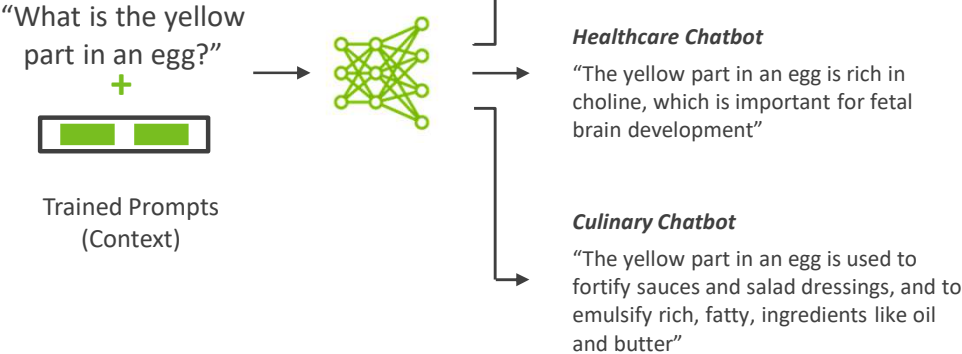
# LLM Customization is Required to Solve Business Problems

## Domain Specific Knowledge - Prompts

No Customization



Customization



1

Choose  
a Model

2

Domain-Specific  
Fine Tuning

3

Inferencing

# Virtualization Is Fast, Even With GPUs

No tradeoff in performance for all the great virtualization features

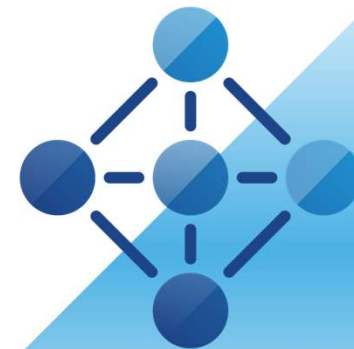
ML inferencing on ESXi  
runs up to **5% faster**  
than bare metal



# Embracing AI/ML

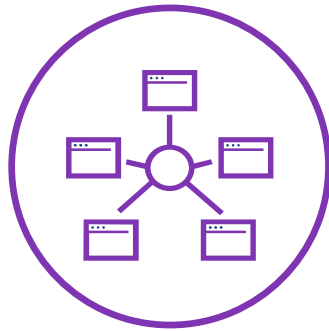
VMware is taking a cloud-smart approach

- Maintain privacy, control, and reduce risks
- Building on top of existing infrastructure for cost effectiveness
- Bring your own LLM to VMware Private AI Foundation
- Deliver accelerated LLM services to transform AI powered apps



# VMware Is Integrating AI Everywhere

## Boosting Engineering Throughput



Improved  
Documentation  
Search

**5.7x** better  
top=5 results



Faster Customer  
Feedback to  
Product

Feedback to  
analysis in **mins**



Increased  
Developer  
Agility

**71%** are coding  
faster



# VMware + Hugging Face Helps Developers Go Faster



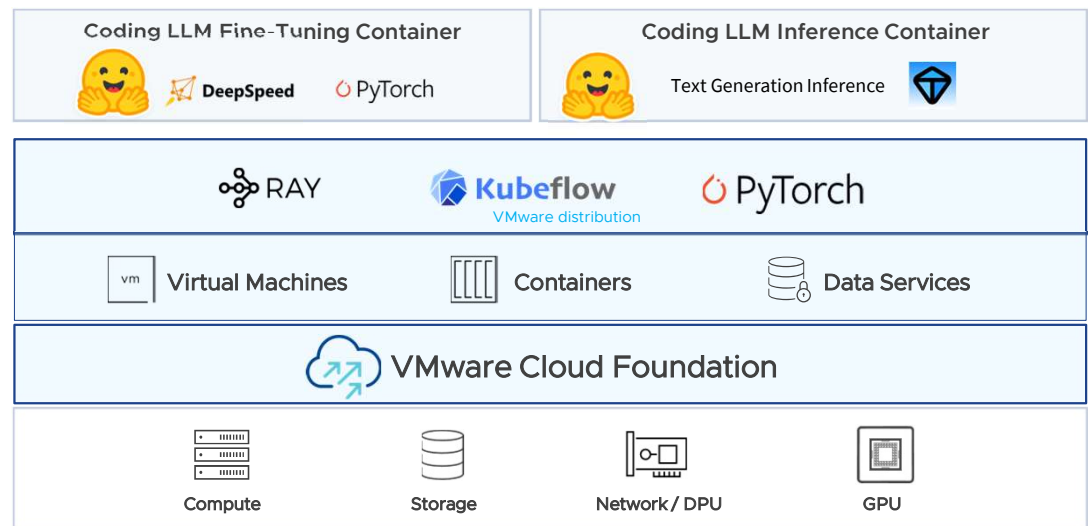
## Developer & DevOps

Better coding experience

More time for system design

Faster app delivery

## VMware Private AI Foundation



# Our Journey with Coding Assistance

VMware + Hugging Face

*“[insert coding assistant] has dramatically accelerated my coding... Still learning to use it but it already writes ~80% of my code, ~80% accuracy. I don't even really code, I prompt. & edit.”*

– Andrej Karpathy



## Use Case: Coding Assistance

Must meet strict data privacy and security requirements without compromising on capability and usability



## Solution: Hugging Face Coding LLM On-prem

Based on the OSS model with support for over 86 programming languages



## Initial Rollout and Lessons Learned

Fine-tuning and inference were run using **just a single A100 40GB** each for our initial pilot



## Results and Next Steps

**85%** of developers found this to be helpful and **93%** would continue using this going forward for all development

NEW

# SafeCoder

on VMware's Private AI Foundation

Fine-tuning and  
inference on VMware  
infrastructure

Coding assistance  
via simple plugin

Code generation  
through natural  
language

Fine-tuned for better  
code completions

Delivered via  
Hugging Face

# EXPLORE YOUR LUCK!

SafeCoder:  
The results are in...

< 2

seconds average  
inference time

71%

are completing  
coding tasks faster

93%

developer  
satisfaction rate



# VMware Tanzu Intelligent Assist

Streamlined on-boarding  
via instant knowledge  
access

Guided deployment via  
conversational insights

Realize cost-saving  
strategies through  
conversational AI

### Intelligent Assist

- Security of our clouds and our data
- Availability and reducing outages and MTTR
- Understanding and optimizing cloud consumption

Security is the top of my mind but so is troubleshooting and preventing outages. Also, our k8s clusters are hard to manage, we just don't have visibility to what is happening with them.

**Intelligent Assist**

OK, so security, availability and Kubernetes support. Anything else?

Not really. Cost is always a concern but that's my boss's problem.

**Intelligent Assist**

I understand. OK thanks for answering my questions. Based on your feedback, we have put together 3 tutorials.

Start tutorial 2 | Start tutorial 3

Start tutorial 1

**Intelligent Assist**

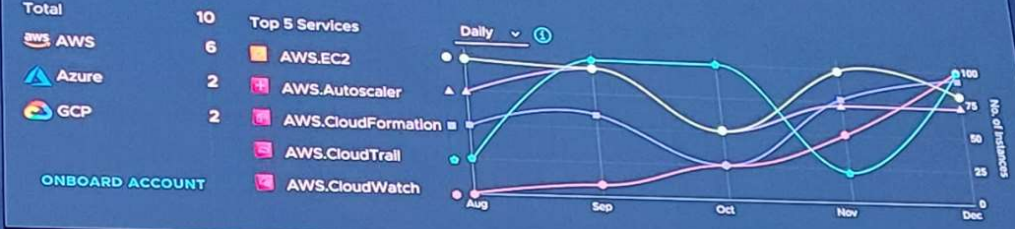
Data collection is completed. Now you can search your cloud inventory

Text to chat.

## Welcome to VMware Tanzu Hub, Kit



### My Accounts



### My Applications

Business 240 | Potential 2015

Recently Modified Applications

- acme-fitness
- lBac338
- lnt-be-infra/helmdall-ingress-api
- aws-mock/aws-mock-prelude-stoyan-dev-setup-service-http
- cas-next-infra/helmdall-ingress-api

MODEL APPLICATIONS

### Security

Findings 128 (▲ 1% last 7 days)

Business Applications 67

Accounts 2

Applications by Risk

12 critical	13 high
8 medium	15 low
	3 no

VIEW APPLICATIONS FINDINGS

Findings based on CIS Benchmarking framework. To view findings based on other frameworks check out VMware Aria Secure Cloud

### Cost

Some data may be missing due to account issues. Fix

Projected cost this month (\$) 3,200 K (▲ 2% over last month)

Potential Savings (\$) 64 K

Cost per month by provider for 2 accounts

Cost (\$): 0, 100K, 200K, 300K

Providers: AWS, Azure, GCP (Coming Soon)

Findings based on CIS Benchmarking framework. To view findings based on other frameworks check out VMware Aria Cost

### Modern Infrastructure

Clusters	146	Nodes	3657
Self Managed	17	Namespaces	502
AWS	121	Services	1728
Azure	6	Pods	8174
GCP	2		

ADD KUBERNETES COLLECTOR

### Cool Things You Can Try

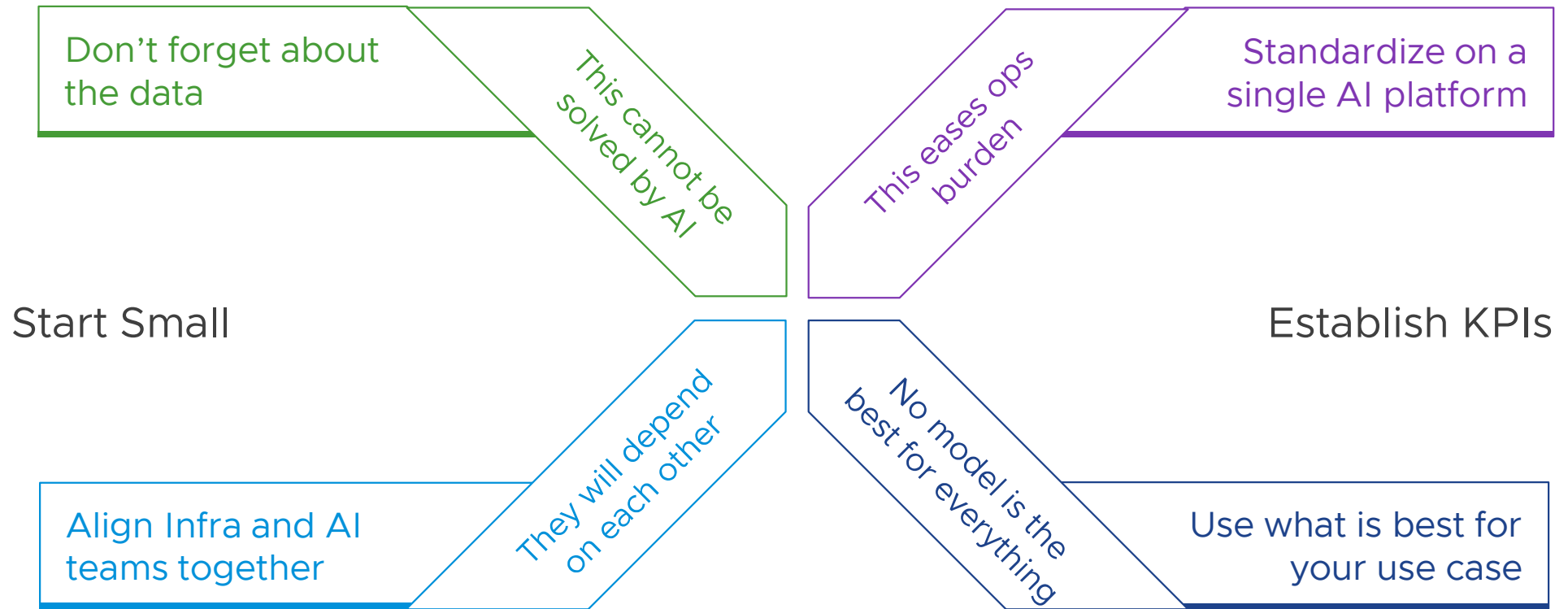
Powerful search capabilities in VMware Tanzu Hub give you answers fast! Try these samples to see how easy it is to find what you need, quickly.

Sample searches

- Running EC2 instances by region
- Running VMs by resource group
- All VMs with a public IP address

# Best Practices Learned at VMware

Converging parts of a whole





Thank You