AI Platform Engineering

Coté – SREDay London, September 19th, 2025

New things platform engineers will likely do with and for AI

- Hosting models.
- Gateways, brokers.
- Application Frameworks
- Center Excellence Bottlenecks.
- Curating models.
- Hosting models with self-service access.
- Cost and performance management for inference.
- Eval, testing, safety.
- Audit and compliance.
- Data access.
- Registries – MCP, prompts, integrations.

A platform treats AI like any other service, adding AI middleware & focusing on new models & frameworks

DEVELOP → OPERATE → OPTIMIZE

App Teams

- Bring their own framework
- Simple command to push to production
- Frictionless data services & content
- Plug into any CI/CD
- AI-Ready dev framework
- AI Self-service model access
- AI Integrate with enterprise data

Platform/AI Teams

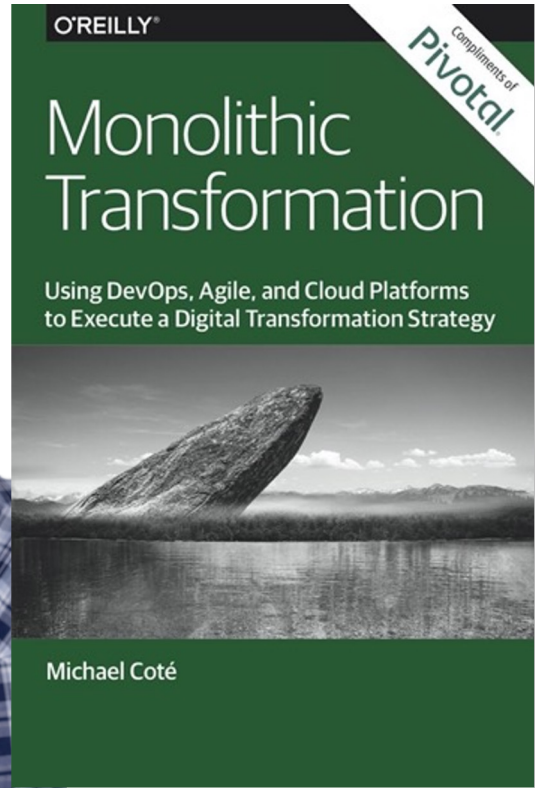
- Autoscaling & load balancing
- Flexibility to run on any Cloud
- Ability to curate services
- AI & platform quota management
- Zero downtime deployments
- AI Model running & updating
- AI Control costs & policy with guardrails

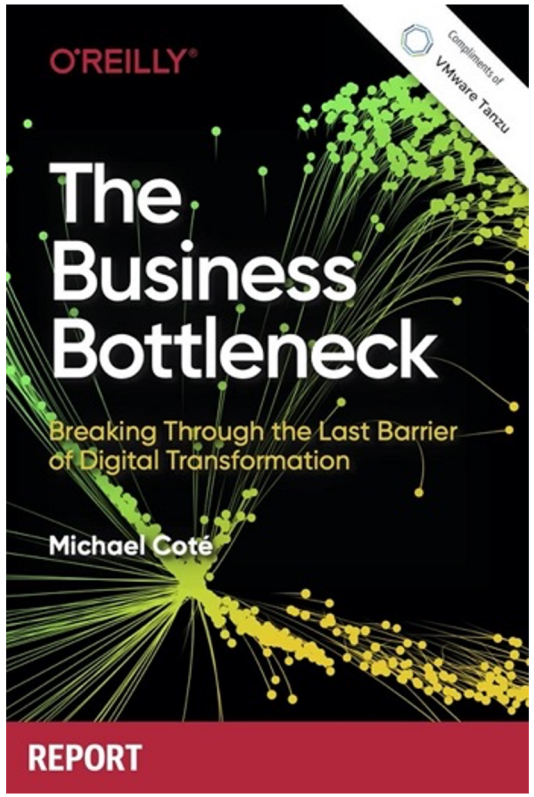
App, Platform/AI Teams

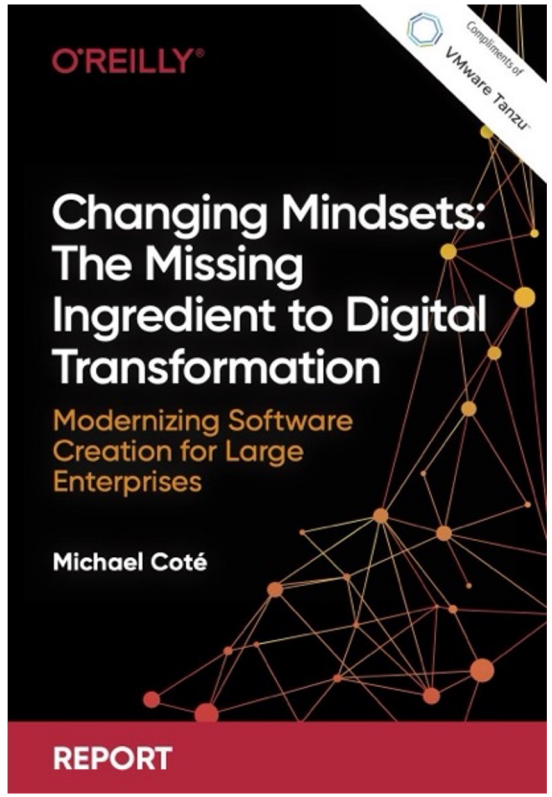
- Rapidly respond to CVEs
- Lifecycle automation & patching
- Credential rotation
- Automated compliance
- Built-in Observability
- AI Integrated model observability
- AI Continuous model curation

Coté

https://newsletter.cote.io/ | cote@broadcom.com







SOFTWARE DEFINED Interviews

SOFTWARE DEFINED TALK

RedMonk

DELL

451 Research S&P Global Market Intelligence

Pivotal

vmware

Tanzu by Broadcom

Home work

Or, acknowledgements and references

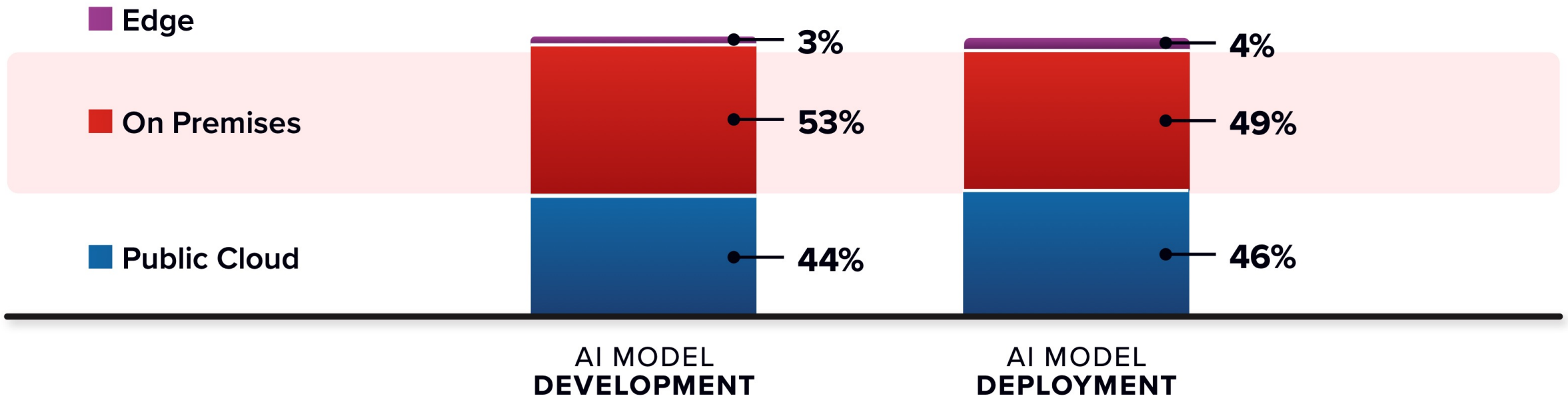
- Tanzu customers and internal use.
- "Tales from Production - Debugging LLMs and GenAI Apps on VMware Tanzu Platform," Nick Kuhn VMware Explore 2025, No. CODEQT1641LV, August, 2025.
- "How platform teams can help scale generative AI application delivery," Manjunath Bhat, Gartner, June 23rd, 2025.
- "Why AI needs a platform team," Patrick Debois, PlatformCon 2025, June 23rd, 2025.
- "AI Platform Engineering," Patrick Debois, Dec 31, 2024.
- "From Experiment to Enterprise: How Block Operationalized MCP at Scale," Angie Jones, Block, MCP Developers Summit, May, 2025.
- "The evolving role of Platform teams in the AI era," Abi Noda and Laura Tacho, DX, August 28th, 2025.
- "The Emperor's New GPT Why Your 'Custom AI' Is a Demo, Not a Product," Keith Townsend, August 18th, 2025.

FIGURE 5

Deployment Location for the Development and Deployment of AI Models

Where does your organization primarily develop and deploy AI models?

(Percentage of respondents)



Note: Public cloud includes public cloud provider, managed services provider, and special-purpose cloud. On premises includes private cloud, traditional infrastructure, and colocation provider. n = 411; Source: IDC's *AI Infrastructure Survey*, July 2024.

For an accessible version of the data in this figure, see Figure 5 Supplemental Data in Appendix 1.

Source: IDC White Paper, sponsored by Broadcom, On-Premises AI Infrastructure Balances Innovation and Security, doc #US52747024 December, 2024. Conducted July, 2024, n=411.

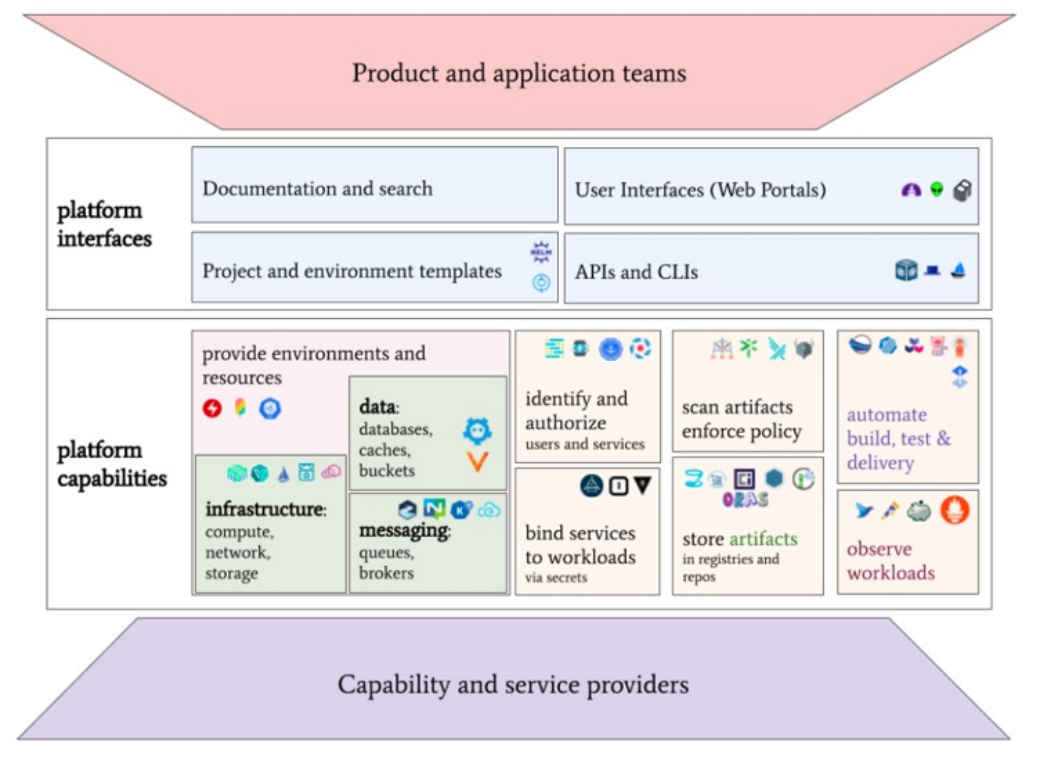
The platform

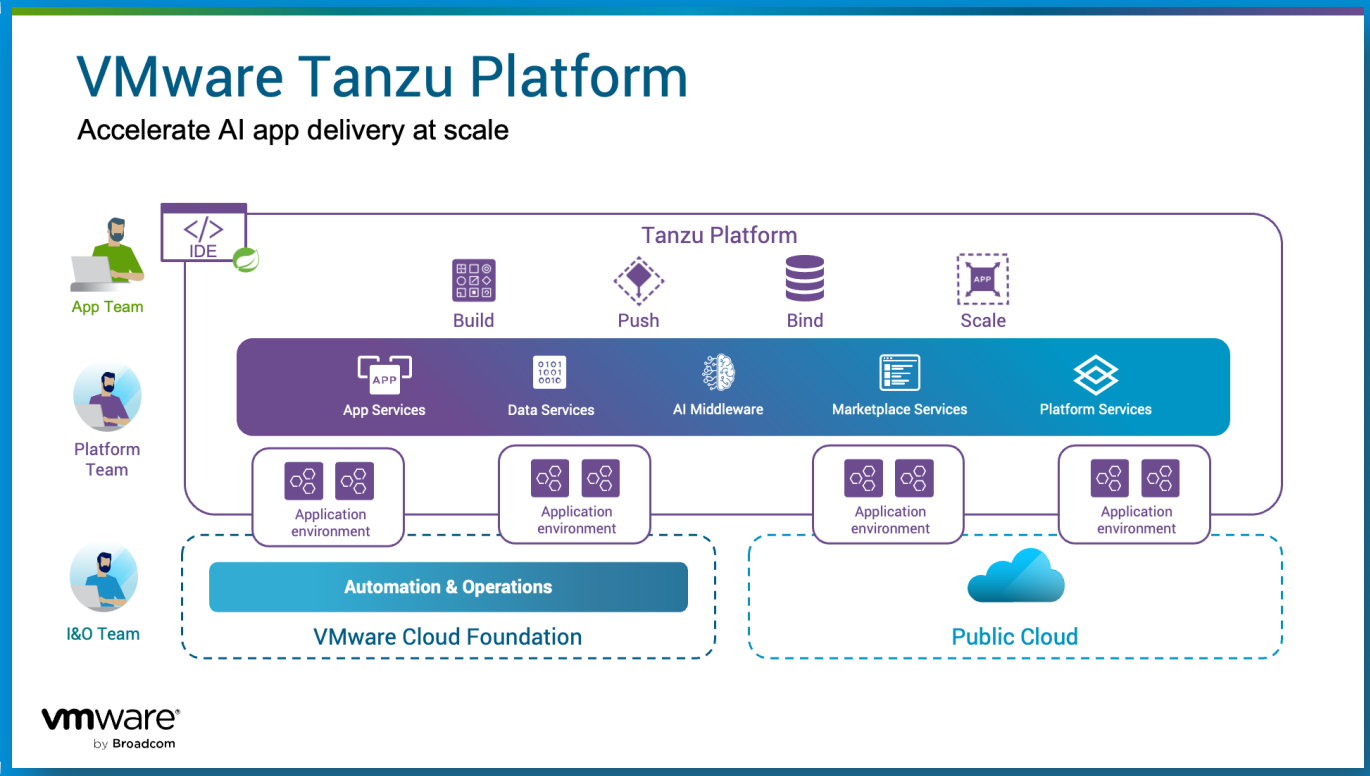
> A *digital platform* is a foundation of self-service APIs, tools, services, knowledge and support which are arranged as a compelling internal product.
>
> [SO THAT] Autonomous delivery teams can make use of the platform to deliver product features at a higher pace, with reduced co-ordination.

Evan Bottcher, March, 2018

What is a platform?

Centralized, standardized stack for building, running, and managing in-house apps.

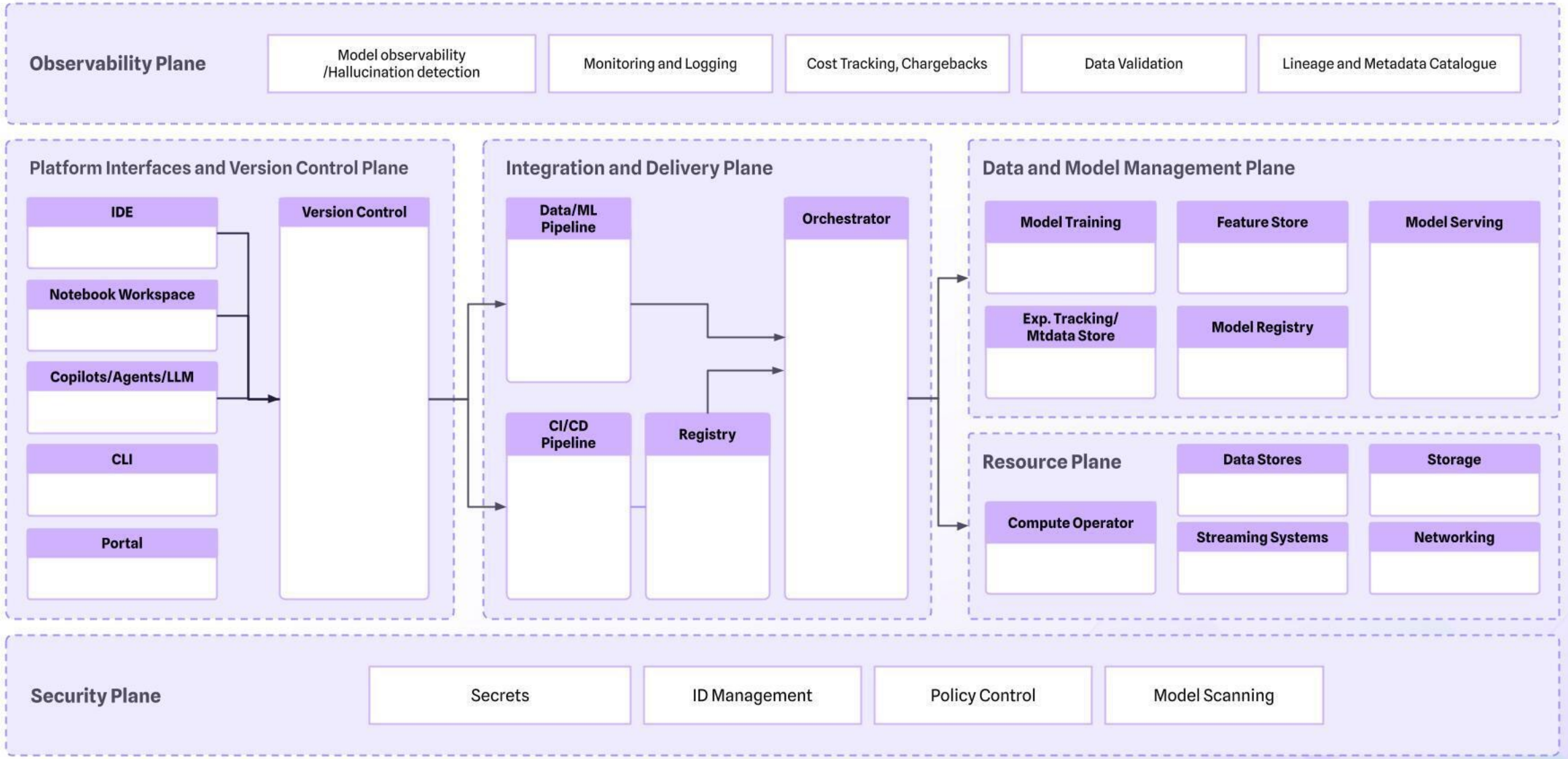




Sources: "CNCF Platforms White Paper," March 2023; VMware Tanzu.

New reference architecture for AI

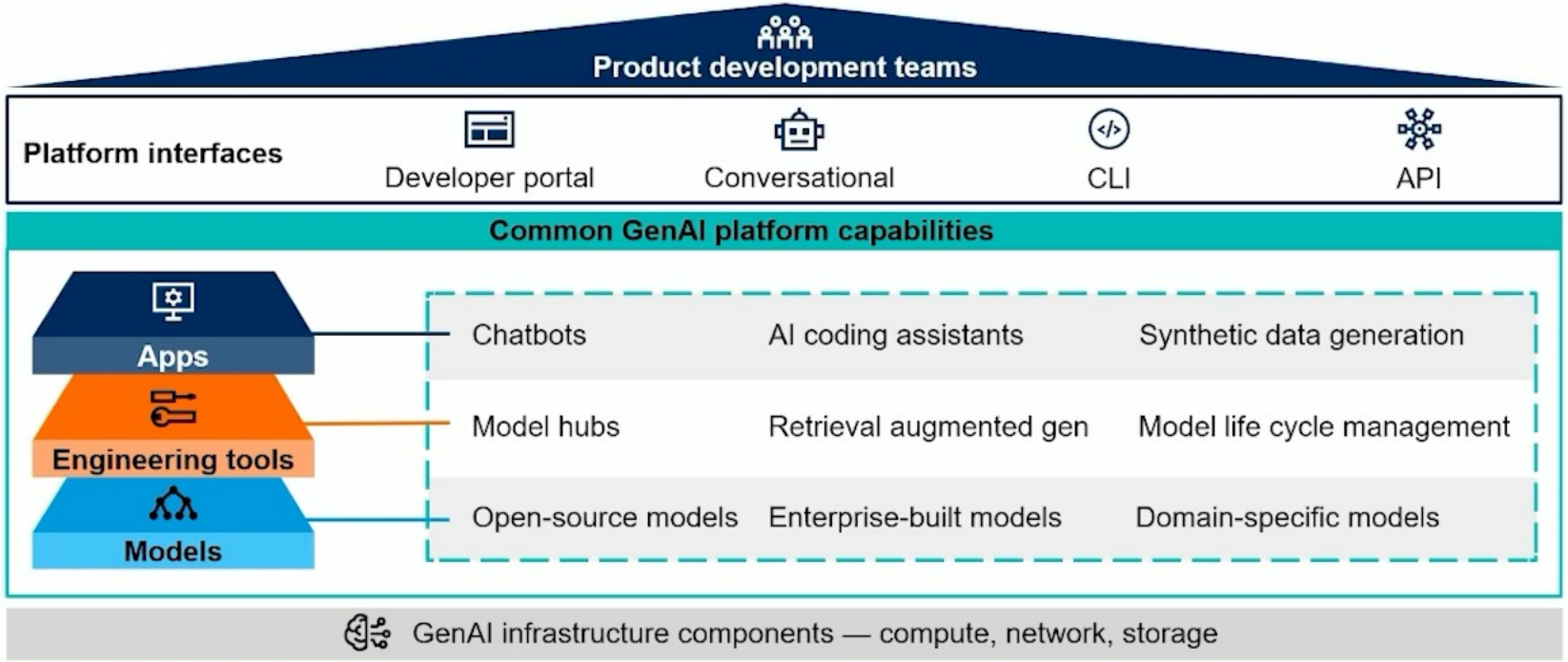
Platform Engineering



Sources: "Platform Engineering is eating the world," Luca Galante, PlatformCon 2025, June 23rd, 2025. "Can Platform Engineering Accelerate AI Adoption?", Jennifer Riggins, The New Stack, August, 2025.

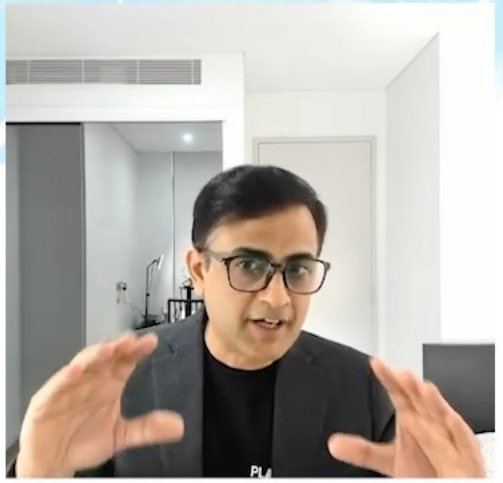
PLATFORM CON 25

Platform Teams Provide Internal Platform Services to Support Common GenAI Needs



[a]TRiSM — AI trust, risk and security management
Source: Gartner

Gartner



Manjunath Bhat

Research VP, Gartner

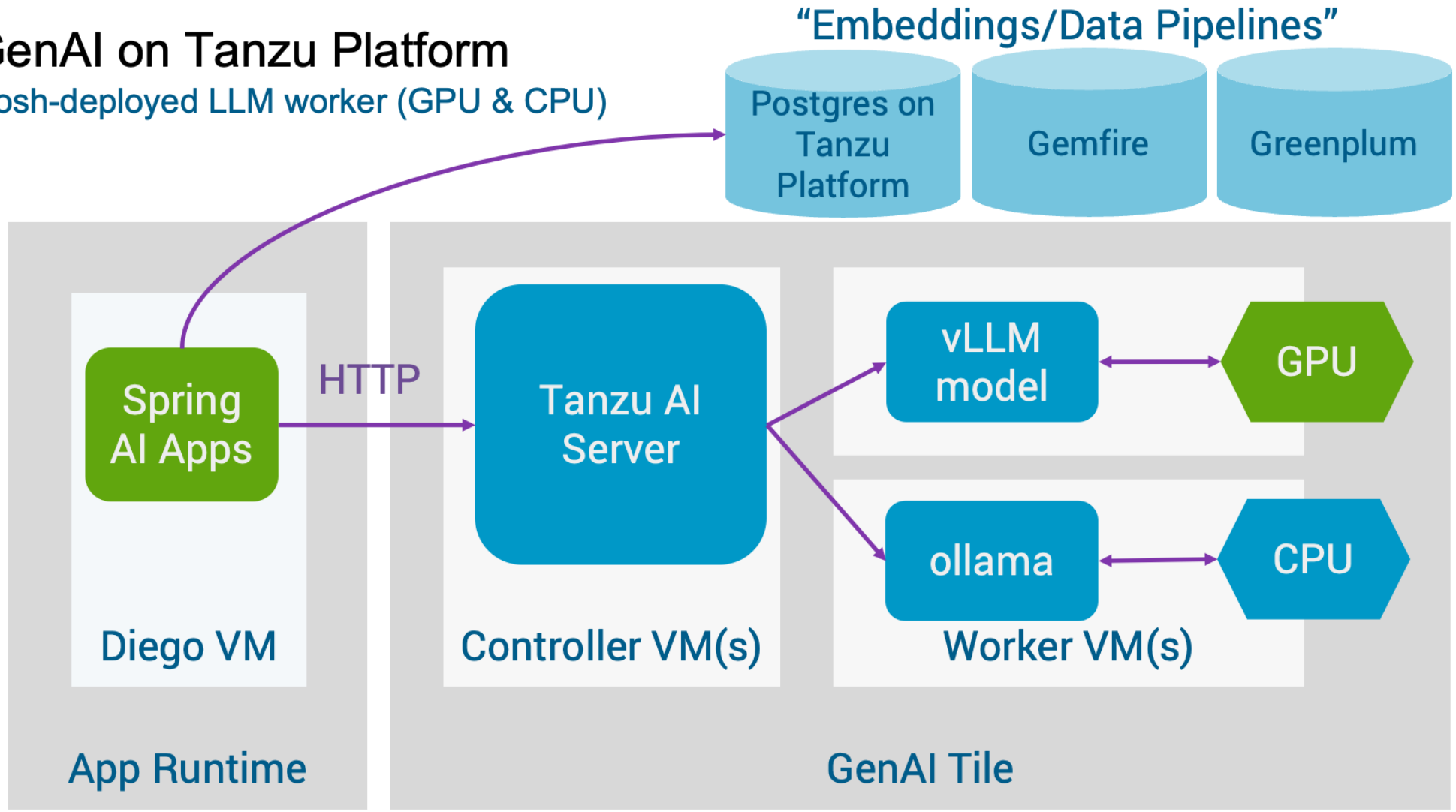
IMPACT

Source: "How platform teams can help scale generative AI application delivery," Manjunath Bhat, Gartner, PlatformCon 2025, June, 2025.

The one thing you must do right now

GenAI on Tanzu Platform

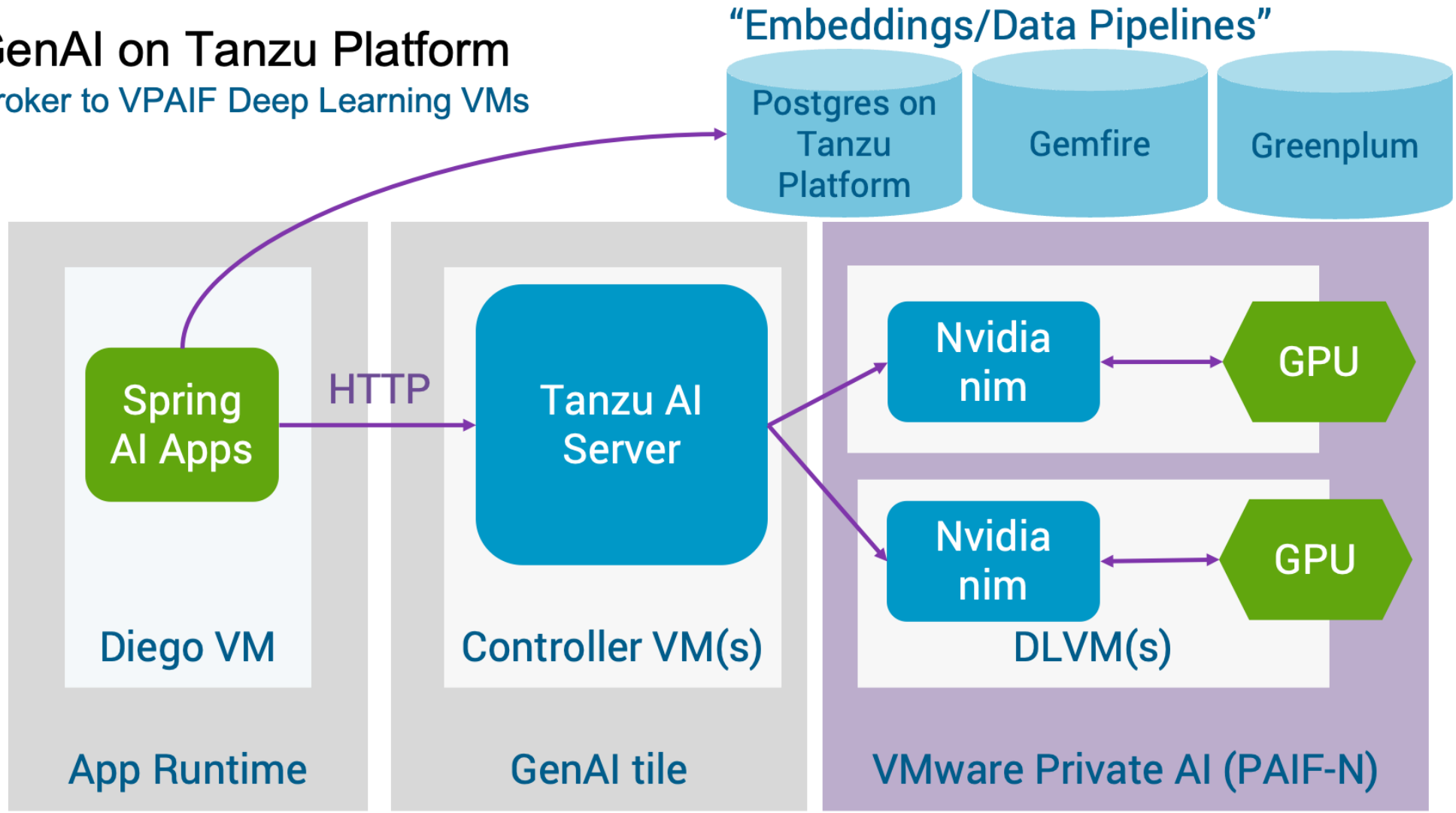
Bosh-deployed LLM worker (GPU & CPU)



Source: "Tales from Production - Debugging LLMs and GenAI Apps on VMware Tanzu Platform," Nick Kuhn VMware Explore 2025, No. CODEQT1641LV, August, 2025.

GenAI on Tanzu Platform

Broker to VPAIF Deep Learning VMs

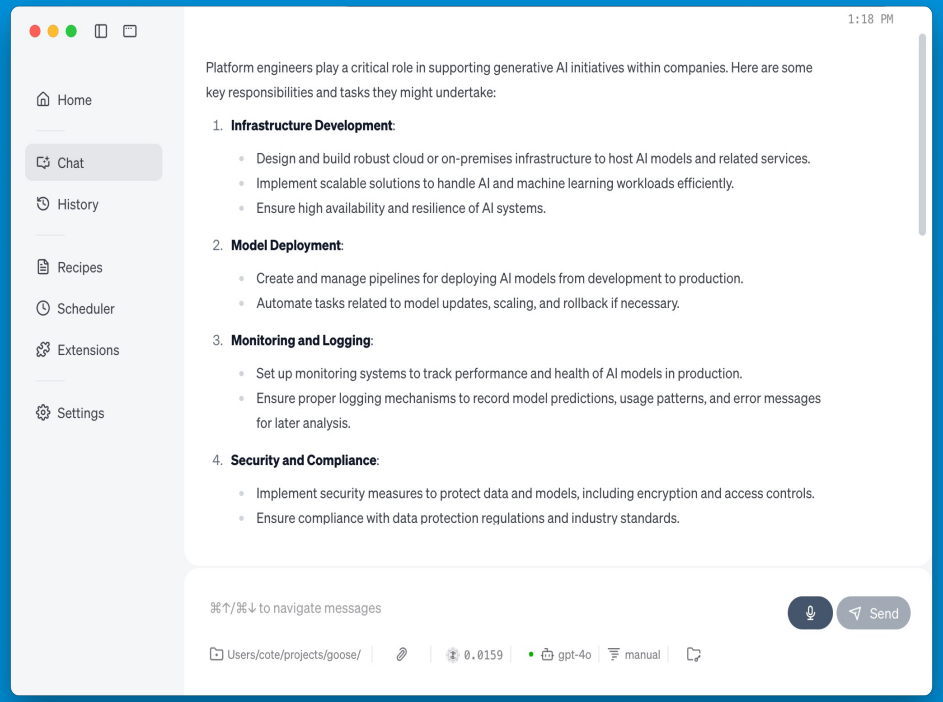


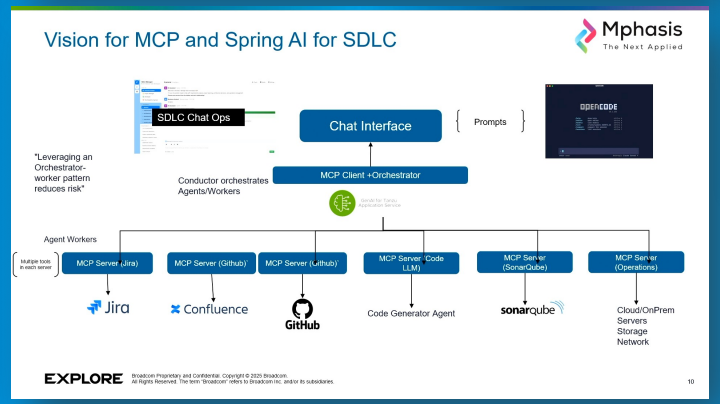
Source: "Tales from Production - Debugging LLMs and GenAI Apps on VMware Tanzu Platform," Nick Kuhn VMware Explore 2025, No. CODEQT1641LV, August, 2025.

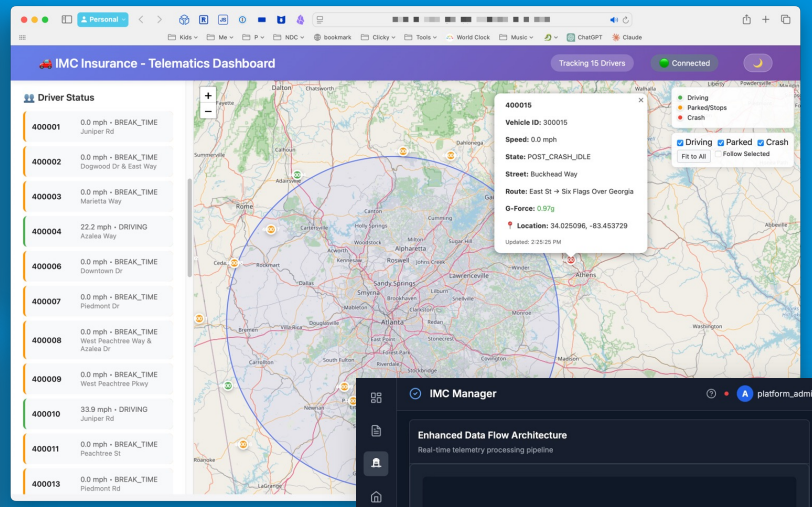
Practices so far

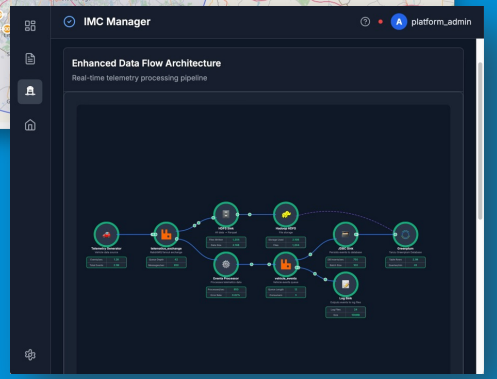
“We are building this platform not for us, we are building it for Mercedes-Benz developers.”

Thomas Müller, Mercedes-Benz









👨‍💼 Chat (for normals)…

- Your own ChatGPT.
- Customer service.
- Better search.
- Chat—as-UI

👩‍💻 Programming…

- New & old code.
- SDLC juicing.
- Trad'l "data science."
- Making pptx?

💾 AI in apps…

- Sales assistants.
- Sloppy integration.
- Science-ing.
- ???

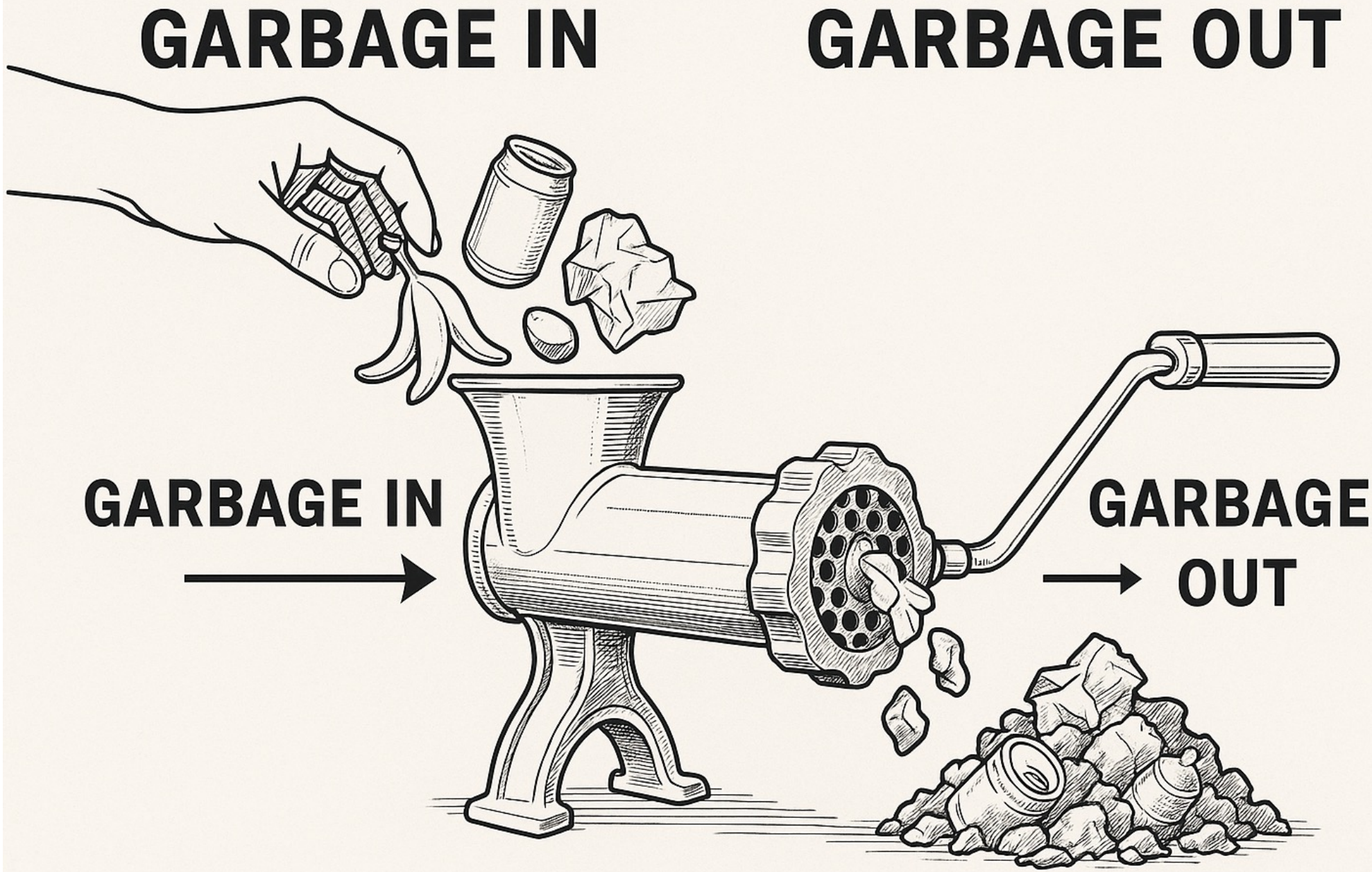
Note: audio, pictures, video are omitted.

Sources: Tanzu customers; AI at Goldman, *FT*, September 14th, 2025; "Leverage Generative AI to Streamline the Software Development Lifecycle," Banu Parasuraman, Andrew Berenato, Explore 2025, August, 2025.

Tales from the Customer...

Meta (llama family)	**Denied**
Mistral Family	**Approved** (*only apache2 licensed models*)
Gemma	**Denied**
Qwen3 + DeepSeek	**Denied**
SaaS (OpenAI,Azure AI)	**Partial (*only public data*)**
OpenAI (gpt-oss)	***Under Review***

Sources: "Tales from Production - Debugging LLMs and GenAI Apps on VMware Tanzu Platform," Nick Kuhn VMware Explore 2025, No. CODEQT1641LV, August, 2025.



Source: ChatGPT, September, 2025.

When you don’t know what you’re doing, do a lot of it quickly (to learn).

Get an AI App Up and Running in Under 2 Weeks

What is it:

Developed by innovators of Cloud Foundry, the AI Starter Kit provides a seamless path to your first AI application on Tanzu Platform.

Who is it for:

Platform engineers, application developers and data teams can utilize AI Starter Kit to develop patterns for scaling AI apps.

Why it's needed:

- Take AI out of the shadows and support internal innovation with minimal friction
- De-risk AI exploration and prepare for scaled AI app rollout
- Ensure your models are ready for primetime with performance, accuracy and ROI

AI Starter Kit for Tanzu Platform

Expected results:

- Frictionless AI Prototyping
- Forecast AI needs more accurately
- Identify optimal delivery strategies
- Optimize production performance
- Reduce GPU dependency

Thanks!

https://newsletter.cote.io/

cote@broadcom.com