

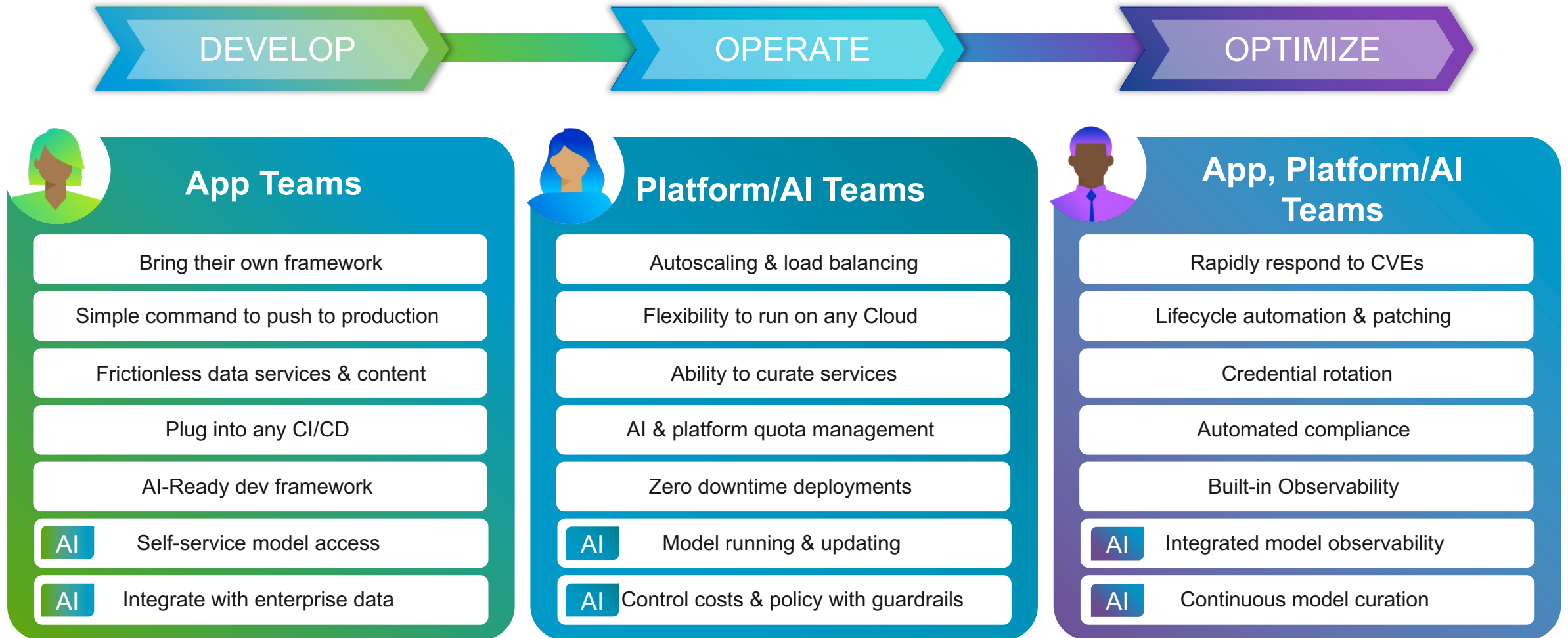
AI Platform Engineering

Coté – VMUG UK, September 18th, 2025

New things platform engineers will likely do with and for AI

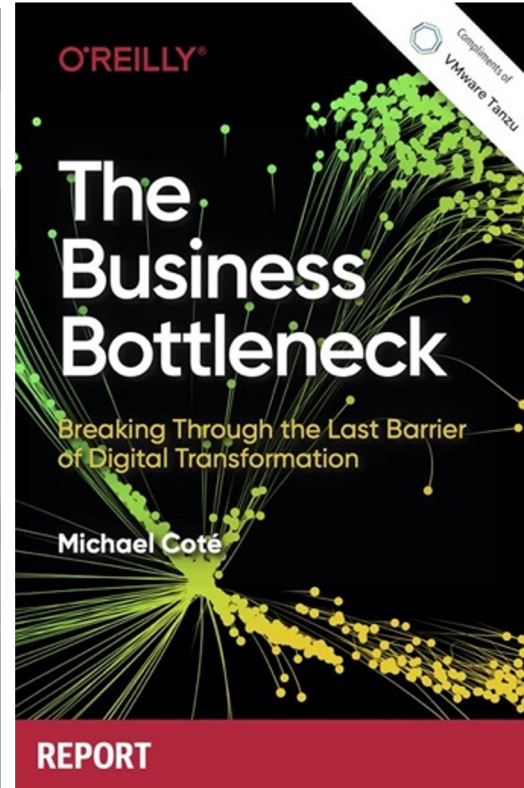
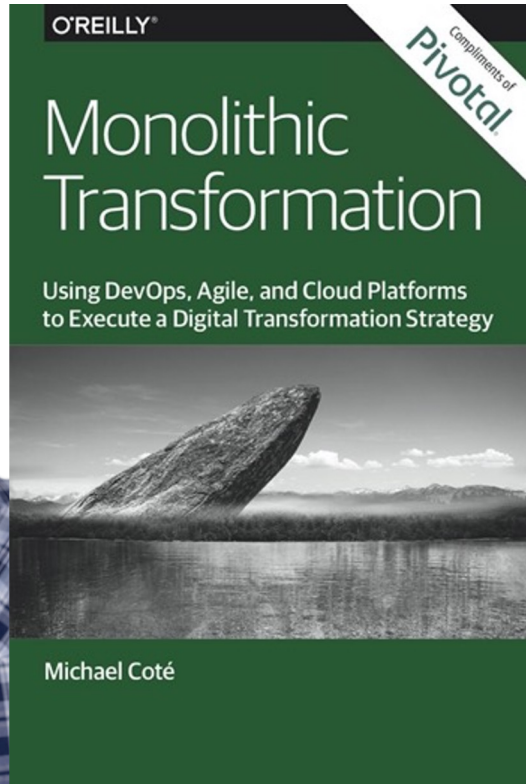
- Hosting models.
- Gateways, brokers.
- Application Frameworks
- Center Excellence
- Bottlenecks.
- Curating models.
- Hosting models with self-service access.
- Cost and performance management for inference.
- Eval, testing, safety.
- Audit and compliance.
- Data access.
- Registries – MCP, prompts,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

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



Coté

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



Home work

Or, acknowledgements and references

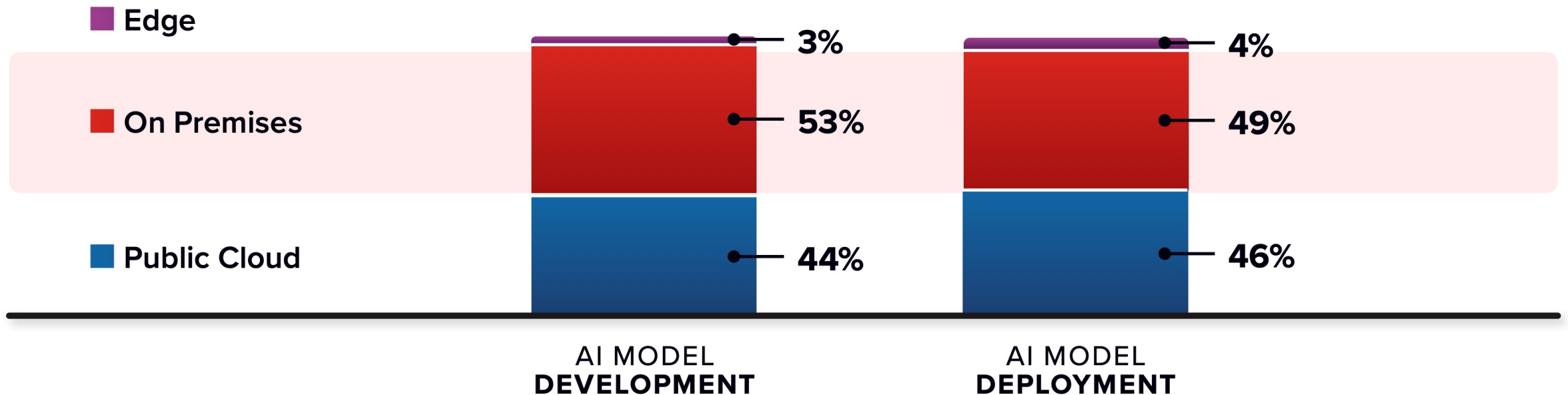
- Tanzu customers and internal use.
- ["Tales from Production - Debugging LLMs and GenAI Apps on VMware Tanzu Platform,"](#) Nick Kuhn VMware Explore 2025, No. CODEQT1641LV, August, 2025.
- ["How platform teams can help scale generative AI application delivery,"](#) Manjunath Bhat, Gartner, June 23rd, 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.

The platform



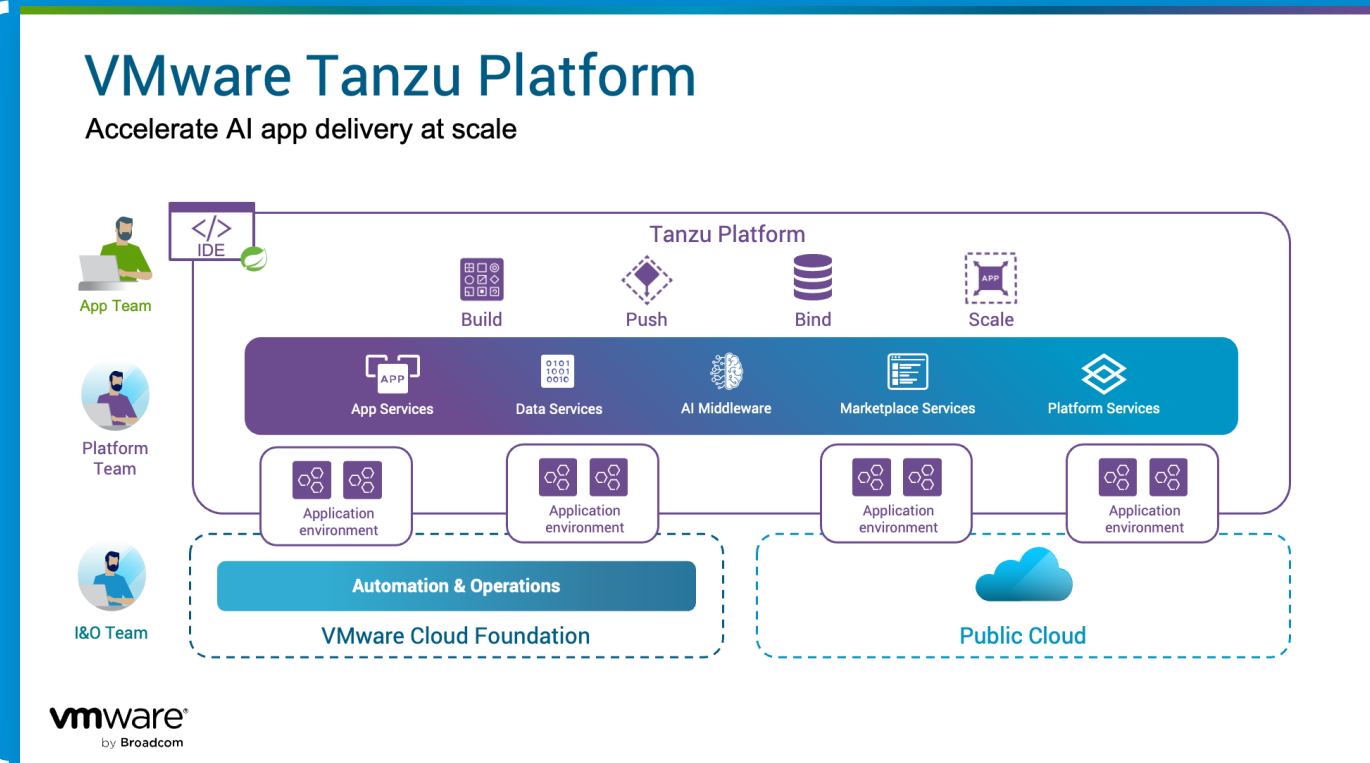
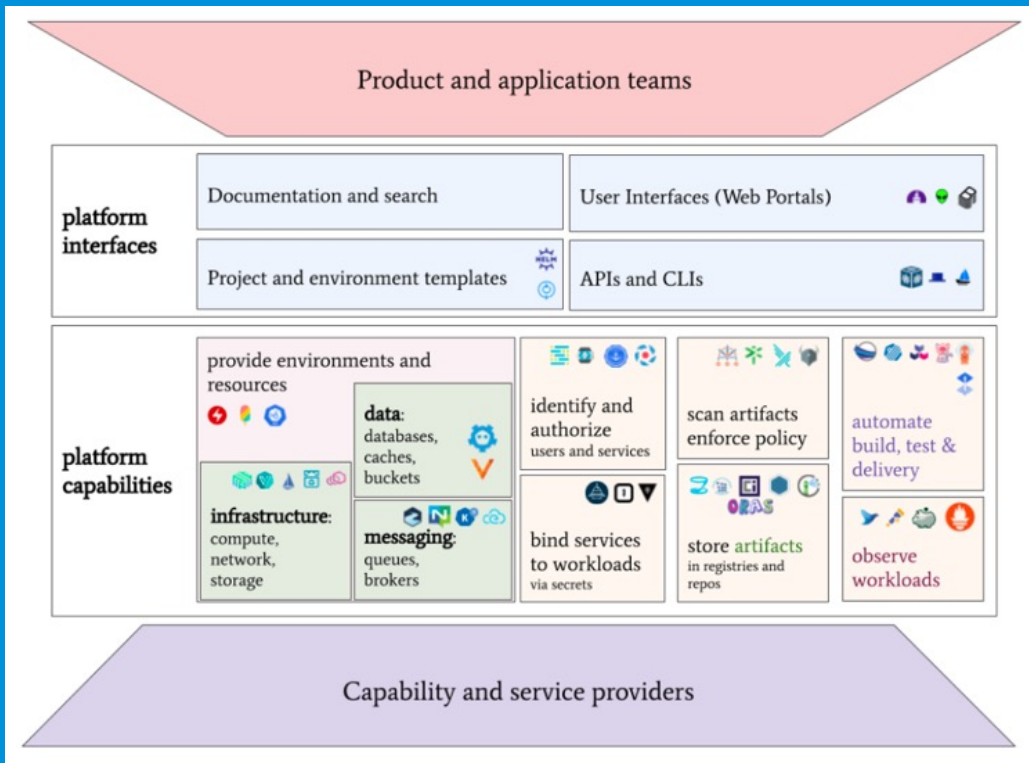
A digital platform is a foundation of self-service APIs, tools, services, knowledge and support which are arranged as a compelling internal product.

[SO THAT] Autonomous delivery teams can make use of the platform to deliver product features at a higher pace, with reduced co-ordination.

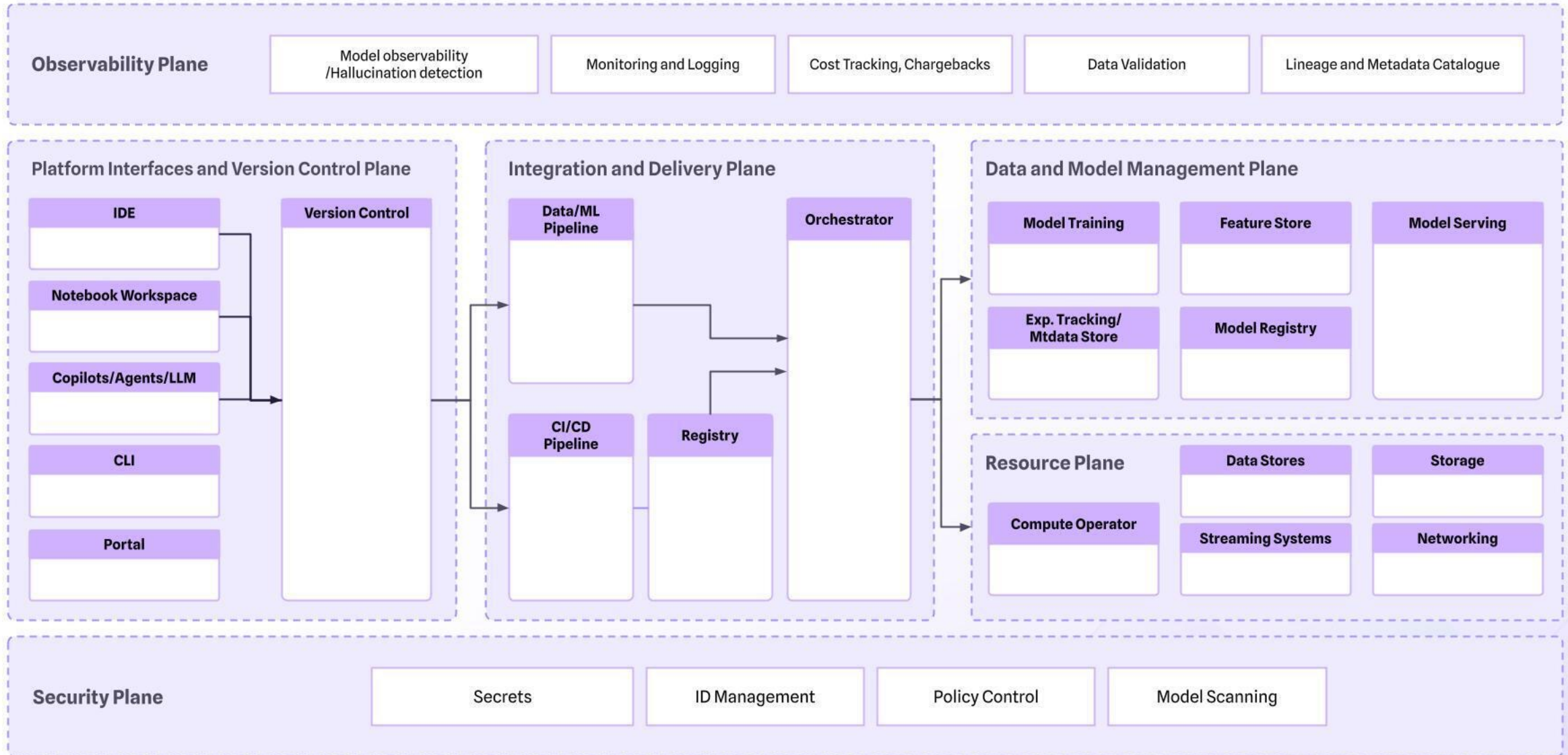
[Evan Bottcher](#), March, 2018

What is a platform?

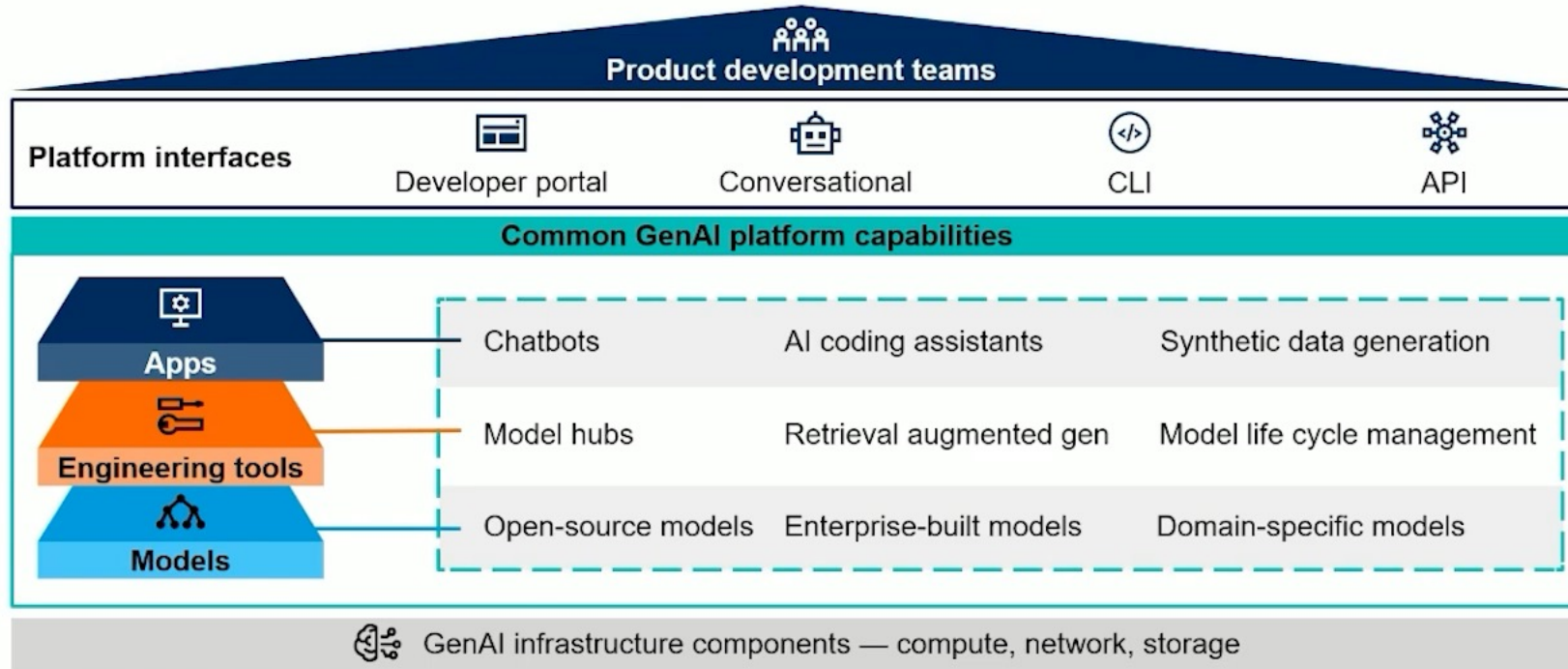
Centralized, standardized stack for building, running, and managing in-house apps.



New reference architecture for AI



Platform Teams Provide Internal Platform Services to Support Common GenAI Needs



*TRISM — AI trust, risk and security management
Source: Gartner

20 © 2025 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates.

Gartner



Manjunath Bhat

Research VP, Gartner

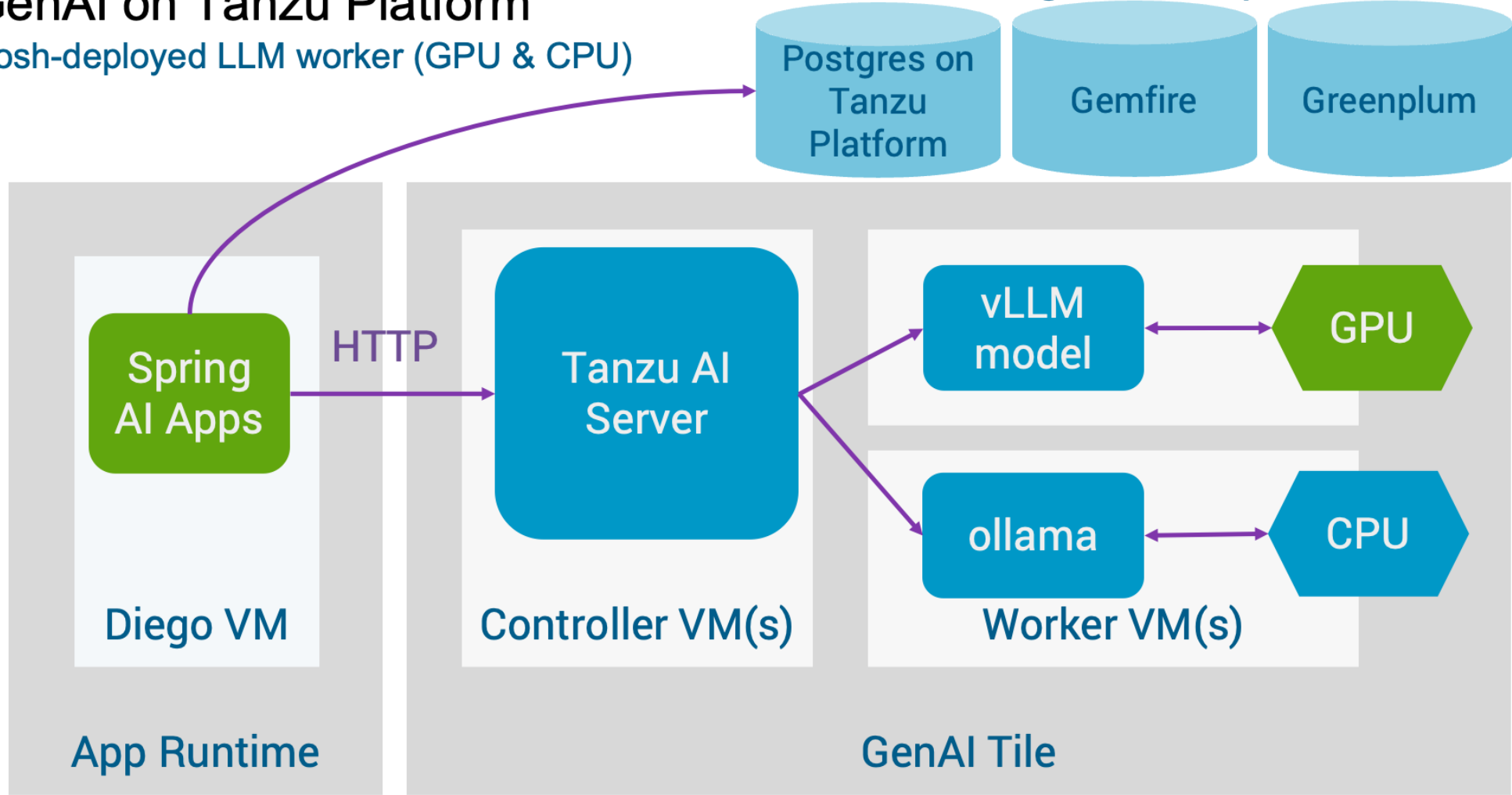
IMPACT

The one thing you must do right now

GenAI on Tanzu Platform

Bosh-deployed LLM worker (GPU & CPU)

“Embeddings/Data Pipelines”



EXPLORE

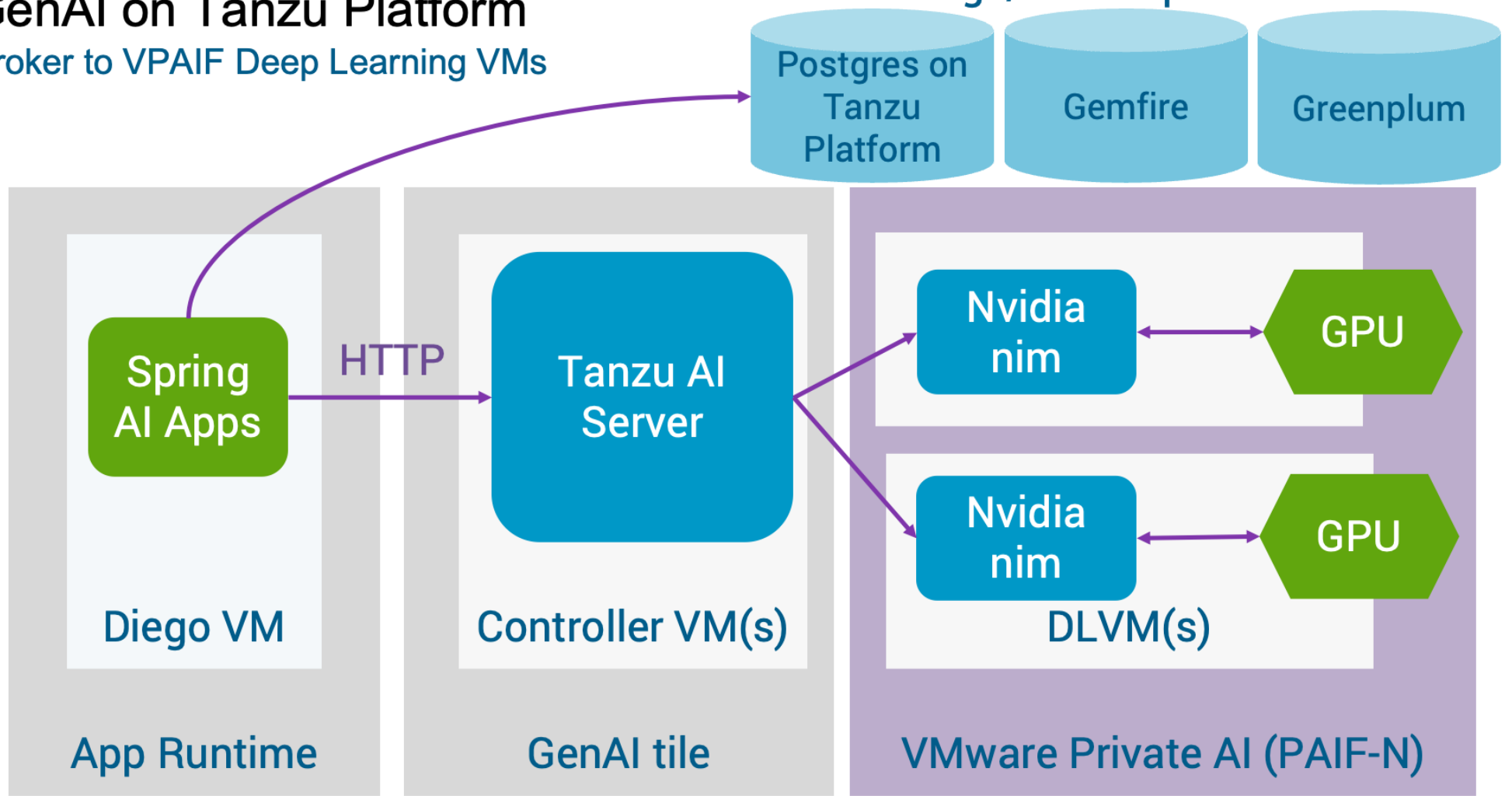
Broadcom Proprietary and Confidential. Copyright © 2025 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc. and/or its subsidiaries.

8

GenAI on Tanzu Platform

Broker to VPAIF Deep Learning VMs

"Embeddings/Data Pipelines"



EXPLORE Broadcom Proprietary and Confidential. Copyright © 2025 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

Practices so far

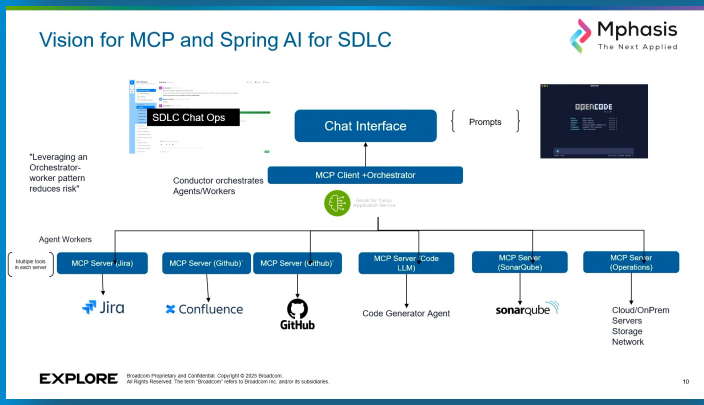
“We are building this platform not for us, we are building it for Mercedes-Benz developers.”

Thomas Müller, Mercedes-Benz



Platform engineers play a critical role in supporting generative AI initiatives within companies. Here are some key responsibilities and tasks they might undertake:

- Infrastructure Development:**
 - Design and build robust cloud or on-premises infrastructure to host AI models and related services.
 - Implement scalable solutions to handle AI and machine learning workloads efficiently.
 - Ensure high availability and resilience of AI systems.
- Model Deployment:**
 - Create and manage pipelines for deploying AI models from development to production.
 - Automate tasks related to model updates, scaling, and rollback if necessary.
- Monitoring and Logging:**
 - Set up monitoring systems to track performance and health of AI models in production.
 - Ensure proper logging mechanisms to record model predictions, usage patterns, and error messages for later analysis.
- Security and Compliance:**
 - Implement security measures to protect data and models, including encryption and access controls.
 - Ensure compliance with data protection regulations and industry standards.



IMC Insurance - Telematics Dashboard

Tracking 15 Drivers Connected

ID	Status	Location
400001	0.0 mph - BREAK_TIME	Juniper Rd
400002	0.0 mph - BREAK_TIME	Dogwood Dr & East Way
400003	0.0 mph - BREAK_TIME	Marietta Way
400004	22.2 mph - DRIVING	Azalea Way
400006	0.0 mph - BREAK_TIME	Downtown Dr
400007	0.0 mph - BREAK_TIME	Piedmont Dr
400008	0.0 mph - BREAK_TIME	West Peachtree Way & Azalea Dr
400009	0.0 mph - BREAK_TIME	West Peachtree Pkwy
400010	33.9 mph - DRIVING	Juniper Rd
400011	0.0 mph - BREAK_TIME	Peachtree St
400013	0.0 mph - BREAK_TIME	Piedmont Rd

Vehicle ID: 300015
 Speed: 0.0 mph
 State: POST_CRASH_IDLE
 Street: Buckhead Way
 Route: East St + Six Flags Over Georgia
 G-Force: 0.97g
 Location: 34.025096, -83.453720
 Updated: 2:25:25 PM

IMC Manager

Enhanced Data Flow Architecture

Real-time telemetry processing pipeline

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.

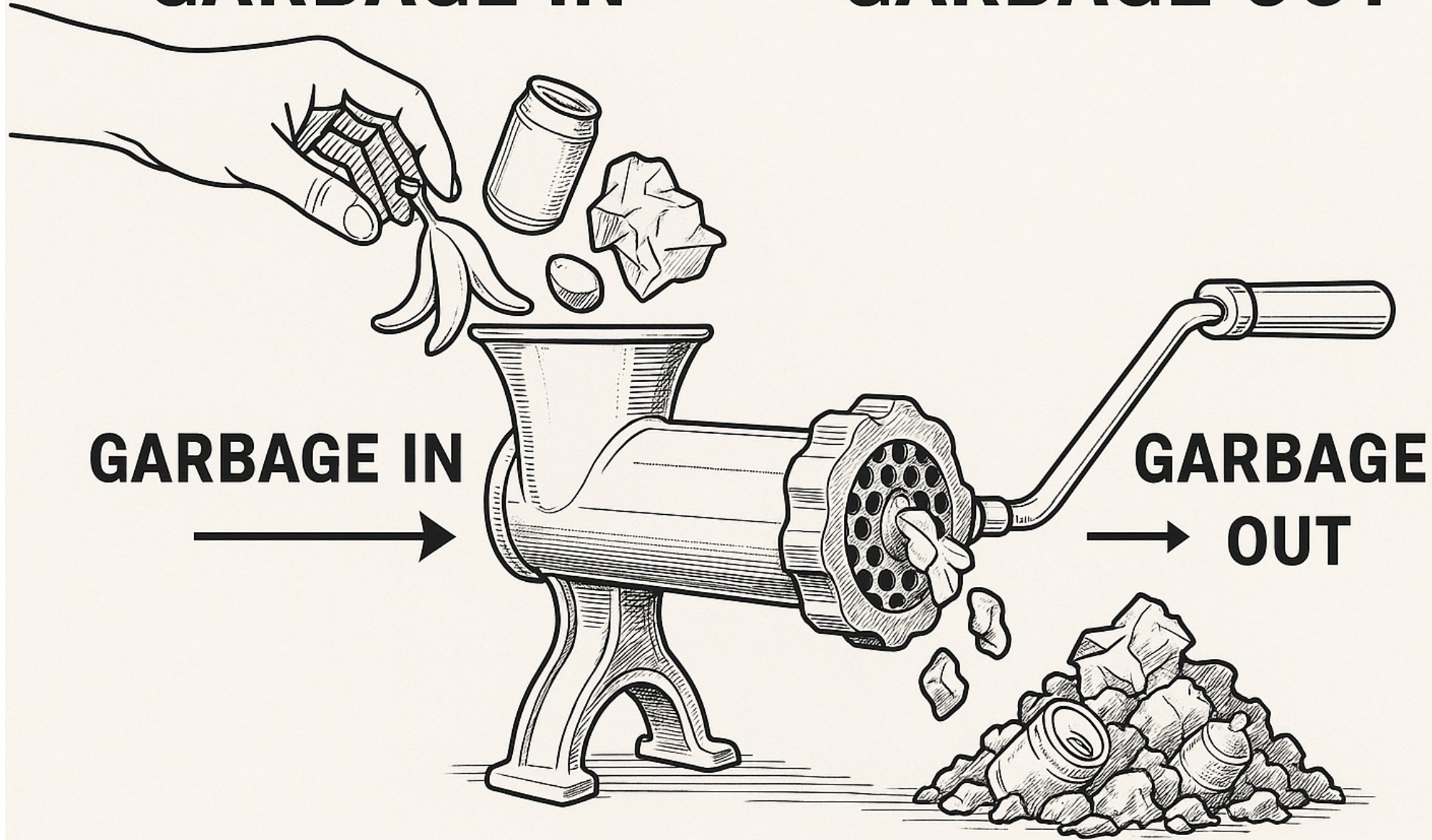
Sources: Taznu customers; [AI at Goldman](#), FT, September 14th, 2025; ["Leverage Generative AI to Streamline the Software Development Lifecycle."](#) Banu Parasuraman, Andrew Berenato, Explore 2025, August, 2025.

Tales from the Customer...

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

GARBAGE IN

GARBAGE OUT



**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.



AI Starter Kit for Tanzu Platform

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

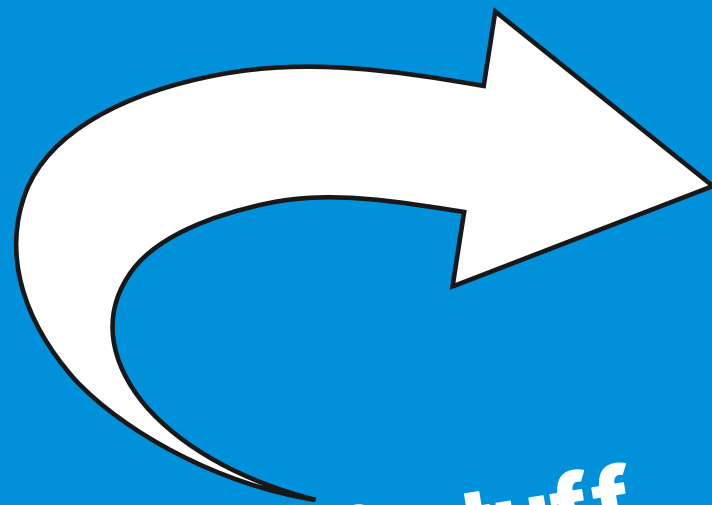
Expected results:

- 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



Slides & stuff

