VMware Tanzu AI Solutions for AI Applications

Safely accelerate agentic and GenAI app delivery with Tanzu Platform

Get started without GPUs today - with Tanzu AI Solutions

Learn more:
vmware.com/products/app-platform/tanzu

"Intelligent apps" incorporate artificial intelligence (AI) to transform customer interactions. By integrating AI into applications, organizations can effectively protect and grow market share, increase customer satisfaction, and enhance brand perception, all while optimizing performance and managing costs.

In the past, incorporating AI into apps required training your own models—a time-consuming and specialized process. However, with the advancement of pre-trained models, the barrier to delivering intelligent apps has largely vanished. Yet, integrating non-deterministic data sources, like models, poses questions: How will this affect your current application delivery patterns? What corporate safety and compliance changes are needed? What if you lack access to GPUs—are they truly essential for deliveringAI applications? Fortunately, an application platform can facilitate adding AI into new and existing apps.

When expanding your platform to incorporate AI applications, it's crucial to consider how you are extending and standardizing protocols for your development team to enable them to accelerate their delivery of these new application types. Application frameworks remain essential to enable effective AI-ready delivery and app platforms help implement policy management at scale and maintain app governance. End-to-end solutions should always be the priority when preparing to deliver intelligent applications.

Adapting to the increased pace of intelligent app experimentation and redeployment requires an app platform. These application types demand continuous auditing and monitoring, allowing for frequent model fine-tuning or model swapping. Additionally, implementing new policy types is vital for enhancing content safety and ensuring response appropriateness. Strict model governance, including role-based access control, rate limiting policies, and safety endpoints will increase your protection from misuse of your apps or attempted attacks. By embedding these requirements and controls into an app platform, you can improve the quality and security of your intelligent applications.

Tanzu Platform offers a scalable and repeatable approach to address AI application concerns, allowing you to efficiently apply enterprise governance to AI applications.

Tanzu AI Solutions offers both an AI dev framework and a platform to accelerate AI app delivery

Faster AI App Development - Developers can use a familiar dev framework for AI app delivery based on Spring. Spring AI provides everything you need to build AI apps including accelerators that provide codified best practices and advisors that provide helpful guidance to deliver AI apps more quickly and without having to reskill.

Safer AI App Delivery - Governance is a key capability needed for delivering AI apps safely. Tanzu AI Solutions ensures that models are safer by providing a curated marketplace of evaluated models directly in the Tanzu Platform and also provides the ability to connect 3rd party tools that are used for policies like automated PII redaction and content safety rules to provide model guardrails.

More Accurate AI Apps - Tanzu AI Solutions includes includes a model and app evaluation and monitoring capability. Observability for AI collects essential metrics to observe agentic and GenAI apps and evaluate model output so you can ensure accuracy and performance while having visibility into token usage so you can optimize for costs.

Quickly Deliver AI Apps with AI-Ready Developer Framework

According to Stack Overflow's 2024 Developer Survey, the majority of enterprise developers favor Java. Java developers and dev team leaders might think that they need to reskill in order to deliver AI in their current or new applications. However, Spring AI can extend your current team to be AI app developers.

Spring AI offers familiar patterns provided by the Spring Framework so that teams can quickly transform and streamline AI app development. In general, AI app delivery can require more rapid iteration and innovation cycles and providing a familiar framework that simplifies the path to production can help organizations meet the need for speed. Developers can also take advantage of the fully integrated and extensive Spring ecosystem, which includes a vast array of tools and libraries that developers can leverage to build applications quickly and effectively. Furthermore, Spring AI offers exceptional portability, allowing developers to easily switch between different AI models as their project requirements evolve.

Spring AI works best with Tanzu Platform which instantiates best practices for AI app delivery and supports developers in creating abstractions so they can focus on innovation. Tanzu Platform simplifies and expedites AI app delivery by creating an API abstraction layer, allowing developers to utilize a consistent API interface regardless of the underlying AI model or the publisher. The AI middleware feature in Tanzu Platform provides developers with a seamless and consistent experience by leveraging abstractions that standardize the behavior of underlying services to resemble OpenAI.

Safer AI Applications with model and application governance and audit

The Tanzu Platform strengthens model safety by managing access and controlling the use of models that power AI applications. It empowers data science teams to curate models tailored to specific application needs, while developers can effortlessly access these approved models through the AI model marketplace. Meanwhile, platform engineering teams maintain oversight and regulate model usage, mitigating risks and preventing overuse by controlling access. This ensures a more secure and efficient intelligent app development process.

These governance capabilities are implemented through AI middleware integrated into the Tanzu Platform. The AI middleware in Tanzu Platform enables role-based-access-control with virtual model keys. The AI middleware also provides real-time monitoring of token usage, delivering clear visibility into spending. This allows teams to optimize resource allocation and adapt to changing needs.

The AI middleware in Tanzu Platform also allows you to seamlessly integrate customizable content safety policies, ensuring that the model adheres to your guidelines. These policies can restrict the model from responding to prompts containing prohibited terms, helping to maintain compliance with legal or

Easier AI app delivery with AI middleware

AI middleware is essential for enterprises integrating AI into their applications, ensuring scalability, governance, and security while improving production efficiency.

Tanzu AI Solutions provides AI middleware that acts as a bridge, streamlining AI application development by offering secure integration, model access controls, and compliance tools. It addresses challenges such as unpredictable AI behaviors, data security risks, and uncontrolled costs through features like private AI hosting, rate limiting, and token monitoring. With AI models evolving rapidly, the AI middleware in Tanzu AI Solutions simplifies model swapping and integration, reducing vendor lock-in by providing a standardized API.

Ultimately, AI middleware is a strategic necessity for organizations looking to adopt AI at scale while maintaining operational control and compliance. Learn how AI middleware in Tanzu Platform simplifies AI app delivery by enhancing security, governance, scalability, and cost control, enabling safer and faster innovation.

Read the blog. /

ethical standards. Additionally, the system can be tailored to permit specific words or phrases based on the unique needs of your business, providing a flexible solution for industries with specialized requirements or sensitive content considerations.

More accurate AI applications with monitoring & evaluation

When AI models are integrated into an application's services, it is essential to implement new monitoring capabilities to understand how to optimize performance and reliability. Managing model drift, which refers to changes in the model's behavior due to evolving data patterns, and maintaining the relevance and accuracy of generated data become continuous monitoring and improvement tasks. Regular updates and adjustments to the models are necessary to align with current organizational data or updates to models. Additionally, model hallucinations, where the AI generates incorrect or nonsensical information, can become problematic if appropriate monitoring and mitigation strategies are not established. Implementing robust safeguards and validation checks are crucial to minimizing these occurrences and enhancing the trustworthiness of AI-generated outputs.

Tanzu AI Solutions features an AI model evaluation utility powered by Spring AI that helps organizations assess the generated content and protect against hallucinated responses. The evaluator enables real-time testing of your models through the AI middleware feature. It also performs model accuracy checks and scans for hallucinations so enterprise organizations can confidently serve AI apps to their customers

More responsive and extensible applications with AI-ready data

VMware Tanzu Greenplum is a powerful data and analytics platform built on MPP Postgres, designed for petabyte-scale data processing. It supports structured, semi-structured, unstructured, vector, geospatial, and graph data. Its MPP architecture ensures efficient handling of large datasets, speeding up AI model training and deployment, including building LLMs. Tanzu Greenplum also offers a robust ecosystem of Postgres-based data science tools for model training, data visualization, ingestion, and embeddings (pgvector).

VMware Tanzu GemFire is a low-latency, high-throughput in-memory data grid that's vector compatible. It efficiently manages and analyzes high-dimensional vector data, making it ideal for applications where traditional databases fall short. Tanzu GemFire is a fast, consistent, Java-native solution designed to improve vector storage and caching for sub-millisecond AI or Retrieval Augmented Generation (RAG) response times. Whether used with proprietary or hosted LLMs, Tanzu GemFire enables caching and real-time analytics essential for AI experimentation.

Summary

AI, whether it is used for predictive, generative, or agentic application patterns, is revolutionizing how enterprises deliver innovative and highly personalized customer experiences with speed and efficiency. Traditionally, adopting AI demanded specialized, custom-trained models, creating significant entry barriers. However, the advent of pre-trained models has simplified the process, making this powerful technology more accessible.

Yet, integrating AI into existing or new applications presents challenges, including concerns around data sources, compliance, and infrastructure (GPUs). This is where Tanzu AI Solutions in Tanzu Platform steps in to eliminate these obstacles. By ensuring AI applications and models are secure, accurate, and optimized for performance, Tanzu AI Solutions empowers teams to build AI applications seamlessly, without the need for specialized skillsets or hardware.

For more information on Tanzu AI Solutions in Tanzu Platform, please visit vmware.com/products/app-platform/tanzu or contact us.

vmware® by Broadcom

Item No: BC_SB_VMware_Tanzu_AI_Solutions_for_GenAI_Applications_converted 4/25