DELL Technologies

Resumen de la solución: Validated Design for Generative AI Inferencing con NVIDIA

Obtenga información de alta calidad más rápido con IA generativa

Implemente rápidamente una solución de pila completa para inferencia de inteligencia artificial generativa (GenAI) con grandes modelos de lenguaje

Mayor productividad y más información

Esta arquitectura conjunta proporciona un diseño modular y flexible que contribuye a multitud de casos de uso y requisitos de computación. Puede elegir diferentes componentes, combinarlos y ampliarlos de forma independiente según las necesidades de su aplicación.

Algunos ejemplos notables de casos de uso de inferencia admitidos son:

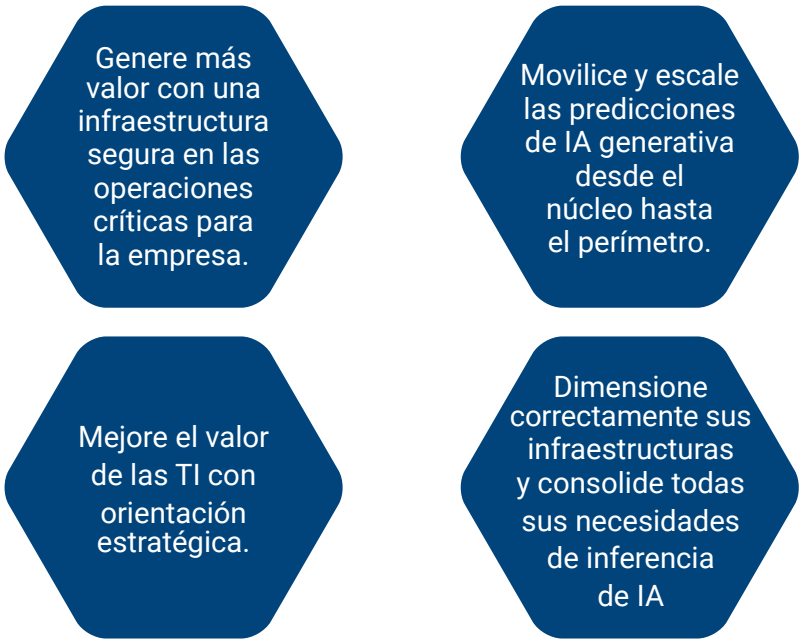
Generación de lenguaje natural: Los modelos generativos se pueden usar para tareas de generación de texto, como la redacción de documentos, la generación de diálogos y tareas de resumen y creación de contenidos.

Chatbots y asistentes virtuales: La IA generativa está detrás de los agentes conversacionales, los chatbots y los asistentes virtuales para generar respuestas en lenguaje natural basadas en las consultas o instrucciones de los usuarios.

Desarrollo de código: Obtenga ayuda en desarrollo de software gracias a funciones como la capacidad de completar código o de generar pruebas unitarias, o bien una función de chat para explicar el código.

Genere predicciones y resultados de mayor calidad y con un tiempo de rentabilización más rápido a la vez que agiliza la toma de decisiones mediante una potente solución de IA generativa de Dell Technologies y NVIDIA. Esta solución, diseñada conjuntamente, aborda los retos de la inferencia como la latencia, la capacidad de respuesta y las exigencias de computación, y ayuda a convertir los datos empresariales en resultados más inteligentes y de alto valor.

Mediante tecnologías innovadoras, servicios profesionales completos y un amplio ecosistema de socios, su organización puede acelerar la IA generativa en toda la empresa. Ahora, organizaciones de TI, científicos de datos y equipos de DevOps de IA pueden proporcionar fácilmente una plataforma modular y escalable para IA generativa e inferencia de LLM.



Reduzca el tiempo hasta la obtención de resultados con una solución probada

Construya infraestructuras rápidamente en sus instalaciones para sus necesidades de aplicaciones con un diseño validado y una arquitectura de referencia diseñados para simplificar su adopción. Reduciendo la complejidad de cada paso del camino, ahora puede obtener más información y las decisiones más rápidas a la vez que potencia la productividad.

NVIDIA

Más información

- Consulte la guía de diseño
- AI InfoHub
- delltechnologies.com/ai
- Dell Technologies y NVIDIA

¿Qué significa “inferencia”?

En IA, “inferencia” se refiere al proceso según el cual se usa un modelo entrenado para crear predicciones, tomar decisiones o generar resultados basados en los datos introducidos. Implica aplicar el conocimiento y los patrones aprendidos durante la etapa de entrenamiento del modelo a datos nuevos, no utilizados previamente.

Durante la inferencia, el modelo entrenado toma los datos de entrada y los procesa mediante algoritmos de computación o una arquitectura de red neuronal para dar lugar a una salida o predicción. El modelo aplica los parámetros, ponderaciones o reglas aprendidos previamente para transformar los datos introducidos en información o acciones relevantes.

La interferencia es una etapa esencial en el ciclo de vida de un sistema de IA. Después de entrenar un modelo con datos etiquetados o no etiquetados para que aprenda patrones y correlaciones, la inferencia le permite al modelo crear generalizaciones basadas en los conocimientos adquiridos y generar predicciones o respuestas sobre datos del mundo real o no vistos.

Genere resultados más rápidamente con nuestra ayuda

Los expertos de Dell Services le ayudan a hacer realidad el valor de la IA generativa más rápidamente con un catálogo de servicios para ayudarle en cada etapa del proceso.

- **Defina estrategias:** elabore su propio roadmap para lograr los objetivos de innovación de las partes interesadas en sus TI y su negocio.
- **Implemente:** establezca su plataforma usando diseños validados por Dell para implementar hardware y software de inferencia con IA generativa.
- **Adopte:** acelere la obtención de valor en sus casos de uso de IA generativa implementando un modelo de inferencia entrenado previamente
- **Escale:** Gestione su cartera de innovaciones con IA generativa con ofertas de expertos técnicos residentes y de formación para desarrollar las habilidades de su equipo.

Especificaciones técnicas

Las configuraciones con Validated Design se basan en los servidores Dell PowerEdge XE y servidoers de montaje en rack más recientes y optimizados para la aceleración de la IA, que aprovechan lo último en las tecnologías de GPU NVIDIA y NVIDIA AI Enterprise, con Triton Inference Server y el entorno de trabajo NeMo. Las cabinas de almacenamiento todo flash o híbridas Dell PowerScale proporciona un almacenamiento en un lago de datos rápido y de grandes dimensiones para la IA generativa.

Computación	Aceleradores	Redes	Software	Almacenamiento
Servidores Dell PowerEdge R760xa	GPU NVIDIA A100 o H100	NVIDIA Networking, Dell PowerSwitch S5232F-ON o S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise con el entorno de trabajo NeMo para LLM y Triton Inference Server; NVIDIA Base Command Manager Essentials	Compatible con Dell PowerScale, ECS y ObjectScale

Dell Technologies y NVIDIA

Dell Technologies y NVIDIA trabajan conjuntamente para impulsar y agilizar las cargas de trabajo de la IA generativa, así como para distribuir hardware y software validado por el equipo de ingenieros que agilicen las cargas de trabajo de la IA, del aprendizaje automático y del aprendizaje profundo para satisfacer las necesidades de los clientes en todos los negocios y los sectores. Con este Validated Design para inferencia de LLM, puede agilizar su transformación digital mediante datos en tiempo real que mejoran la toma de decisiones clave a escala, con soluciones optimizadas para acortar al máximo el tiempo de rentabilización de sus iniciativas de IA.

Más información acerca de las soluciones Dell

Póngase en contacto con un experto de Dell Technologies.

Ver más recursos

Únase a la conversación con #HashTag.

DELL Technologies

NVIDIA