DELL Technologies

Resumo da solução: Validated Design para inferência de IA generativa com a NVIDIA

Obtenha insights valiosos rapidamente usando a IA generativa

Implante rapidamente uma solução de pilha completa para Inteligência Artificial Generativa (GenAI) baseada em grandes modelos de linguagem (LLMs).

Aumente a produtividade e os insights

Essa arquitetura conjunta oferece um design modular e flexível que oferece suporte a uma infinidade de casos de uso e requisitos computacionais. Os componentes podem ser combinados e dimensionados de modo independente conforme as necessidades dos aplicativos.

Alguns exemplos notáveis de casos de uso de inferência compatíveis:

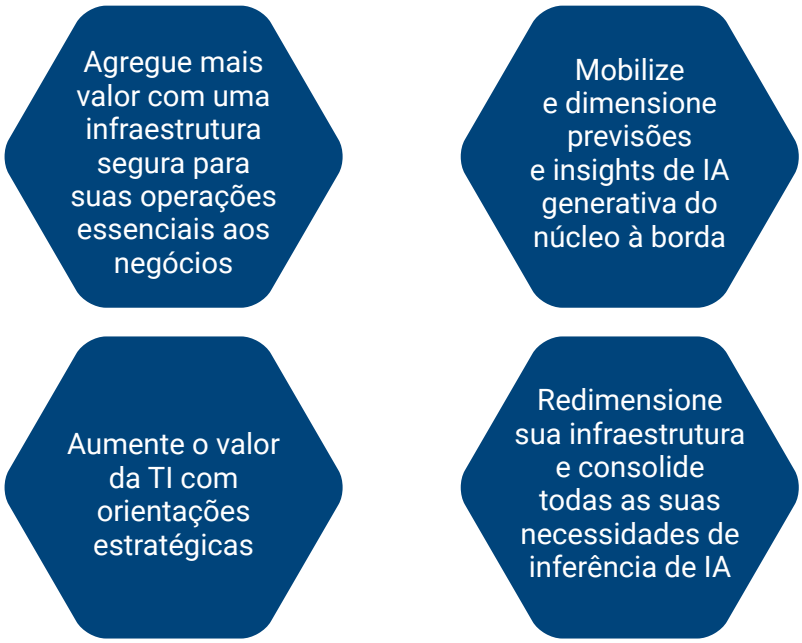
Geração de linguagem natural: os modelos de IA generativa podem ser usados em tarefas de geração de texto, como composição de documentos, geração de diálogos, produção de resumos ou criação de conteúdo

Chatbots e assistentes virtuais: a IA generativa potencializa agentes de conversa, chatbots e assistentes virtuais ao gerar respostas em linguagem natural com base em dúvidas ou instruções de usuários.

Desenvolvimento de código: obtenha assistência no desenvolvimento de software usando recursos como conclusão de código, geração de testes unitários ou uma função de chat para explicar um código.

Gere resultados e previsões de maior qualidade e menor time-to-value, além de acelerar a tomada de decisões com uma avançada solução de IA generativa da Dell Technologies e da NVIDIA. Esta solução projetada em conjunto elimina os desafios da inferência, como latência, capacidade de resposta e demandas computacionais, ajudando a transformar dados empresariais em resultados mais inteligentes e de alto valor.

Com tecnologias inovadoras, serviços profissionais abrangentes e uma ampla rede de parceiros, sua organização pode acelerar a IA generativa em escala empresarial. Agora, organizações de TI, cientistas de dados e AI DevOps podem fornecer facilmente uma plataforma modular e escalável para inferência de IA generativa e de LLM.



Reduza o tempo de obtenção de resultados com uma solução comprovada

Crie rapidamente uma infraestrutura local para os aplicativos de que precisa com um design validado e uma arquitetura de referência projetada para simplificar a adoção. Ao reduzir a complexidade de cada etapa do processo, agora você pode gerar mais insights e tomar decisões com mais rapidez, além de aumentar a produtividade.

© 2023 Dell Inc. ou suas subsidiárias.

NVIDIA

Saiba mais

- Consulte o Guia de design
- InfoHub de IA
- delltechnologies.com/ai
- Dell Technologies e NVIDIA

O que é inferência?

No âmbito da IA, inferência significa o processo de usar um modelo treinado para gerar previsões, tomar decisões ou produzir resultados com base em dados de entrada. Ela envolve a aplicação do conhecimento aprendido e dos padrões adquiridos durante a fase de treinamento do modelo a dados novos e inéditos.

Durante a inferência, o modelo treinado processa os dados de entrada por meio de algoritmos computacionais ou de uma arquitetura de rede neural para gerar um resultado ou uma previsão. O modelo aplica os parâmetros, pesos ou regras aprendidos para transformar os dados de entrada em informações ou ações significativas.

A inferência é uma etapa crucial no ciclo de vida de um sistema de IA. Após treinar um modelo usando dados rotulados ou não rotulados para identificar padrões e correlações, a inferência permite que o modelo gere seu próprio conhecimento e faça previsões ou produza respostas para dados reais ou inéditos.

Conte com nossa ajuda para acelerar a obtenção de resultados

Os especialistas em serviços da Dell ajudam você a perceber o valor da IA generativa para seus dados com mais rapidez por meio de um portfólio de serviços que oferece assistência em todas as etapas da jornada rumo à IA generativa:

- **Criar a estratégia** - crie seu roteiro para cumprir os objetivos de inovação das partes interessadas de TI e de negócios
- **Implementar** - estabeleça sua plataforma utilizando Dell Validated Designs para implementar o hardware e software de inferência de IA generativa
- **Adotar** - amplie o valor de seus casos de uso de IA generativa ao implementar um modelo de inferência pré-treinado
- **Dimensionar** - gerencie seu portfólio de inovação de IA generativa aproveitando especialistas técnicos residentes e ofertas de treinamento para desenvolver as habilidades da sua equipe

Especificações técnicas

As configurações de Validated Design baseiam-se nos mais novos servidores em rack e Dell PowerEdge XE otimizados para aceleração de IA, aproveitando as mais recentes GPUs NVIDIA e o NVIDIA AI Enterprise, com Triton Inference Server e a estrutura NeMo. O armazenamento rápido e amplo de data lake para IA generativa e LLMs é fornecido pelos storage arrays totalmente flash ou híbrido Dell PowerScale.

Computação	Aceleradores	Sistema de rede	Software	Armazenamento
Servidores Dell PowerEdge R760xa	GPUs NVIDIA A100 ou H100	NVIDIA Networking, Dell PowerSwitch S5232F-ON ou S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise com estrutura Nemo Framework para LLMs e Triton Inference Server; NVIDIA Base Command Manager Essentials	Compatível com Dell PowerScale, ECS e ObjectScale

Dell Technologies e NVIDIA

A Dell Technologies e a NVIDIA trabalham em conjunto para habilitar e agilizar as cargas de trabalho de IA generativa, bem como para oferecer hardware e software validados por engenharia para potencializar as cargas de trabalho de IA, ML e DL a fim de atender às necessidades dos clientes em todos os mercados verticais e empresas. Com esse Validated Design para inferência de LLM, você pode acelerar a transformação digital por meio de dados em tempo real que aprimoram a tomada de decisões essenciais em escala, usando soluções otimizadas para reduzir ao máximo o time-to-value das iniciativas de IA.

Saiba mais sobre as soluções Dell

Entre em contato com um especialista da Dell Technologies

Veja mais recursos

Participe da conversa com #HashTag

© 2023 Dell Inc. ou suas subsidiárias. Todos os direitos reservados. Dell e as demais marcas comerciais pertencem à Dell Inc. ou suas subsidiárias. SAP, SAP HANA, SAP S/4HANA e SAP Business One são marcas registradas da SAP SE na Alemanha e em outros países. Outras marcas comerciais podem pertencer aos respectivos proprietários.