DELL Technologies

Solution brief: Validated Design per l'inferenza basata su intelligenza artificiale generativa con NVIDIA

Ottieni informazioni analitiche ad alto valore più velocemente con l'intelligenza artificiale generativa (GenAI)

Implementa rapidamente una soluzione full-stack per l'inferenza LLM (Large Language Model) basata su GenAI

Aumenta la produttività e ottieni informazioni analitiche

Questa architettura congiunta prevede una progettazione modulare e flessibile che supporta molteplici casi d'uso e requisiti computazionali. I componenti sono combinabili e abbinabili tra loro, nonché adattabili in modo indipendente a seconda delle esigenze dell'applicazione.

Di seguito sono riportati alcuni casi d'uso supportati di esempio relativi all'inferenza:

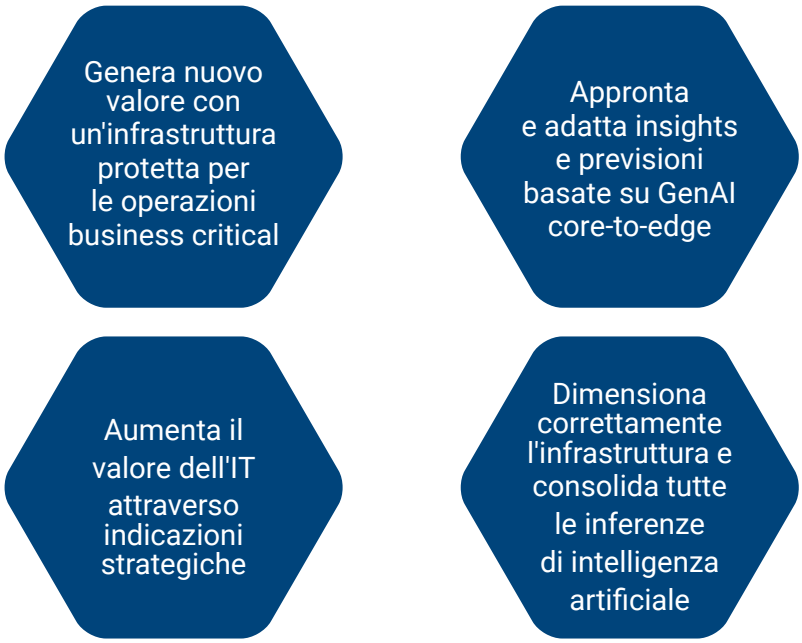
Generazione del linguaggio naturale: utilizza modelli per le attività di generazione del testo quali la scrittura di documenti, la composizione di dialoghi, il riepilogo o la creazione di contenuti.

Chatbot e assistenti virtuali: la GenAI aumenta l'efficienza di agenti conversazionali, chatbox e assistenti virtuali, generando risposte in linguaggio naturale in base alle istruzioni o alle query degli utenti.

Sviluppo del codice: ricevi assistenza per lo sviluppo software con funzionalità quali il completamento del codice, la generazione di test delle unità o chat per la spiegazione del codice.

Genera previsioni e output migliori con time-to-value più rapido, accelerando al contempo il processo decisionale grazie alla potente soluzione di intelligenza artificiale generativa di Dell Technologies e NVIDIA. Questa soluzione progettata congiuntamente risolve le problematiche legate all'inferenza (ad es latenza, reattività e requisiti computazionali) in modo da trasformare i dati aziendali in risultati più intelligenti ad alto valore.

Grazie a tecnologie innovative, servizi professionali a 360° e all'ampio ecosistema dei partner, la tua organizzazione accelera l'intelligenza artificiale generativa a livello aziendale. Organizzazioni IT, data scientist ed esperti in attività di DevOps con AI possono ora distribuire una piattaforma modulare e scalabile per l'inferenza GenAI e LLM.



Ottieni risultati in minor tempo con una soluzione testata

Crea rapidamente un'infrastruttura on-premise per soddisfare le tue esigenze in relazione alle applicazioni con una progettazione convalidata e un'architettura di riferimento pensata per semplificare l'adozione. Riducendo la complessità delle singole fasi del percorso, ora ricavi più informazioni analitiche e prendi decisioni più rapidamente, aumentando al contempo la produttività.

NVIDIA

Ulteriori informazioni

- Leggi la guida alla progettazione
- InfoHub per l'AI
- delltechnologies.com/ai
- Dell Technologies e NVIDIA

Cos'è l'inferenza?

In riferimento all'AI, per inferenza si intende un processo che prevede l'utilizzo di un modello addestrato per effettuare previsioni, prendere decisioni o produrre output in base ai dati di input, nonché la successiva applicazione delle conoscenze e dei pattern acquisiti in fase di addestramento del modello a nuovi dati.

Durante l'inferenza, il modello addestrato acquisisce i dati di input e li elabora tramite i propri algoritmi computazionali o l'architettura di rete neurale per generare un output o una previsione. Il modello applica, quindi, le regole, i pesi o i parametri appresi per trasformare i dati di input in informazioni utili o azioni.

L'inferenza è un passaggio fondamentale nel ciclo di vita di un sistema di intelligenza artificiale. Dopo aver addestrato un modello su dati con o senza etichetta per apprendere i pattern e le correlazioni, l'inferenza permette al modello di generalizzare la propria conoscenza e di fare previsioni o generare risposte su dati reali o sconosciuti.

Fornisci risultati più rapidamente con il nostro aiuto

Affidati agli esperti Dell Services per realizzare più rapidamente il valore della GenAI per i tuoi dati con un portafoglio di servizi pensati per aiutarti in ogni fase del tuo percorso verso l'intelligenza artificiale generativa:

- **Strategia:** crea la roadmap per raggiungere gli obiettivi di innovazione delle entità interessate in ambito IT e aziendale
- **Implementazione:** definisci la piattaforma sfruttando le soluzioni Dell Validated Design per implementare hardware e software di inferenza basata su intelligenza artificiale generativa
- **Adozione:** accelera il valore dei tuoi casi d'uso di GenAI implementando un modello di inferenza precedentemente addestrato
- **Scalabilità:** gestisci il portafoglio di innovazioni GenAI con l'aiuto di tecnici qualificati sul posto e offerte di formazione per sviluppare le competenze del tuo team

Specifiche tecniche

Le configurazioni Validated Design si basano sugli innovativi server rack e server PowerEdge XE ottimizzati per l'accelerazione AI, sfruttando le innovative GPU NVIDIA e la suite NVIDIA AI Enterprise con Triton Inference Server e framework NeMo. Gli array di storage all-flash o ibridi Dell PowerScale forniscono, invece, storage basato su data lake veloce e ad alta capacità per l'intelligenza artificiale generativa e i modelli linguistici di grandi dimensioni.

Elaborazione	Acceleratori	Rete	Software	Storage
Server Dell PowerEdge R760xa	GPU NVIDIA A100 o H100	Rete NVIDIA e Dell PowerSwitch S5232F-ON o S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ, NVIDIA AI Enterprise con framework Nemo per modelli LLM e Triton Inference Server, NVIDIA Base Command Manager Essentials	Supportato da Dell PowerScale, ECS e ObjectScale

Dell Technologies e NVIDIA

Dell Technologies e NVIDIA collaborano per abilitare e accelerare i carichi di lavoro di intelligenza artificiale generativa, oltre a fornire soluzioni hardware e software con progettazione convalidata per accelerare i carichi di lavoro AI, ML e DL, al fine di soddisfare le esigenze del cliente per ogni tipologia di business o settore verticale. Con questa soluzione Validated Design per l'inferenza LLM, acceleri la Digital Transformation attraverso dati in tempo reale che migliorano il processo decisionale su vasta scala e soluzioni ottimizzate per ridurre il time-to-value delle tue iniziative di intelligenza artificiale.

Scopri di più sulle soluzioni Dell

Contatta un esperto Dell Technologies

Visualizza più risorse

Partecipa alla conversazione con #HashTag

DELL Technologies

NVIDIA