DELL Technologies

Lösungsübersicht: Validated Design for Generative AI Inferencing with NVIDIA

Schnellere Gewinnung aussagekräftiger Erkenntnisse mit GenAI

Schnelle Bereitstellung einer Full-Stack-Lösung für das Inferencing großer Sprachmodelle auf Basis von GenAI (Generative Artificial Intelligence)

Produktivität und Erkenntnisse verbessern

Diese gemeinsame Architektur zeichnet sich durch ein modulares und flexibles Design aus, das viele verschiedene Anwendungsfälle und Rechenanforderungen unterstützt. Die Komponenten können kombiniert, angepasst und je nach Anwendungsbedarf unabhängig voneinander skaliert werden.

Einige interessante Beispiele für unterstützte Inferencing-Anwendungsfälle:

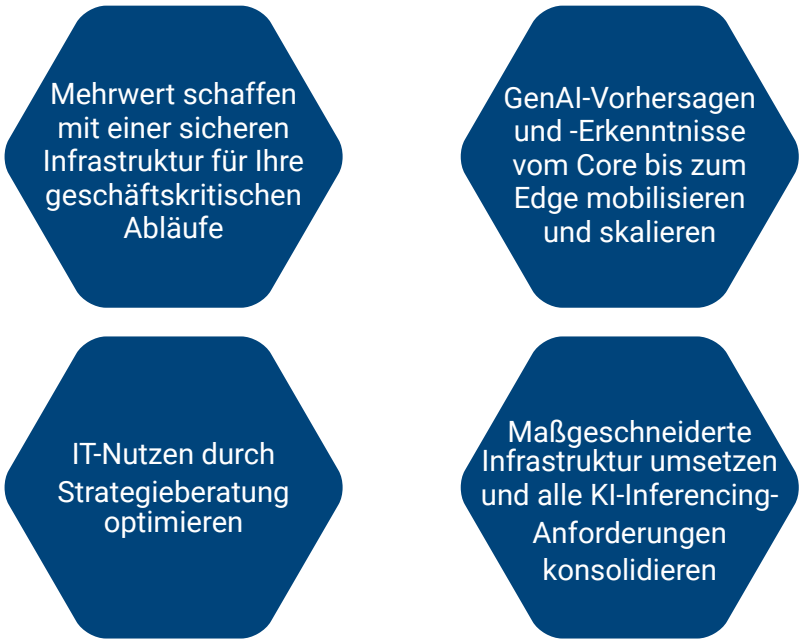
Generierung natürlicher Sprache: Generierungsmodelle können für Textgenerierungsaufgaben wie das Verfassen von Dokumenten und die Erstellung von Dialogen, Zusammenfassungen oder Inhalten genutzt werden.

Chatbots und virtuelle Assistenz: GenAI unterstützt dialogorientierte Agents, Chatbots und virtuelle Assistenz. Dabei werden Antworten auf der Basis von Nutzeranfragen oder -anweisungen in natürlicher Sprache generiert.

Codeentwicklung: Profitieren Sie bei der Softwareentwicklung von Funktionen wie Codevervollständigung, der Möglichkeit, Komponententests zu generieren, oder einer Chatfunktion, um Code zu erläutern.

Generieren Sie qualitativ hochwertigere und schnellere Time-to-Value-Vorhersagen und -Ergebnisse und beschleunigen Sie gleichzeitig die Entscheidungsfindung mit einer leistungsstarken GenAI-Lösung von Dell Technologies und NVIDIA. Diese gemeinsam entwickelte Lösung ist die Antwort auf Inferencing-Herausforderungen wie Latenz, Reaktionsgeschwindigkeit und Rechenanforderungen. Zudem hilft sie Ihnen, Unternehmensdaten in hochqualitative, intelligentere Ergebnisse zu transformieren.

Mit innovativen Technologien, umfassenden Dienstleistungen und einem weit gespannten Partnernetzwerk können Sie GenAI unternehmensweit schneller einführen. IT-Abteilungen, Data Scientists und KI-DevOps sind nun ohne Weiteres in der Lage, eine modulare und skalierbare Plattform für GenAI- und LLM-Inferencing bereitzustellen.



Time-to-Results durch eine bewährte Lösung verkürzen

Realisieren Sie schnell die passende On-Premises-Infrastruktur für Ihre Anwendungsanforderungen, indem Sie auf ein validiertes Design und die einfache Einführung mittels einer Referenzarchitektur setzen. Da die einzelnen Schritte weniger komplex sind, können Sie jetzt mehr Erkenntnisse gewinnen, schnellere Entscheidungen treffen und gleichzeitig die Produktivität steigern.

NVIDIA

Mehr erfahren

- Designleitfaden anzeigen
- KI-InfoHub
- delltechnologies.com/ai
- Dell Technologies und NVIDIA

Was ist Inferencing?

Inferencing bezieht sich in der KI auf die Verwendung eines trainierten Modells, um Vorhersagen und Entscheidungen zu treffen oder Ergebnisse auf der Grundlage von Eingabedaten zu erzielen. Das gelernte Wissen und die Muster, die in der Trainingsphase des Modells ermittelt wurden, werden dabei auf neue, unbekannte Daten angewendet.

Beim Inferencing nutzt das trainierte Modell Eingabedaten und verarbeitet sie durch seine Rechenalgorithmen oder die Architektur des neuronalen Netzes, um ein Ergebnis oder eine Vorhersage zu erhalten. Das Modell wendet seine gelernten Parameter, Gewichtungen oder Regeln an, um die Eingabedaten in nützliche Informationen oder Aktionen zu transformieren.

Inferencing ist eine wichtige Phase im Lebenszyklus eines KI-Systems. Nachdem ein Modell mit gekennzeichneten oder nicht gekennzeichneten Daten trainiert wurde, um Muster und Korrelationen zu erlernen, kann es durch Inferencing sein Wissen verallgemeinern und Vorhersagen treffen bzw. Antworten für reale oder unbekannte Daten generieren.

Mit unserer Unterstützung schneller Ergebnisse erzielen

Dell Services ExpertInnen helfen Ihnen, mit GenAI schneller einen Mehrwert aus Ihren Daten zu generieren. Dafür sorgt ein Serviceportfolio, das Ihnen den Weg zu GenAI in jeder Phase erleichtert:

- **Strategie entwickeln** – Konzipieren Ihrer Roadmap, um die Innovationsziele Ihrer IT- und Business-StakeholderInnen zu erreichen
- **Implementieren** – Einrichten Ihrer Plattform mithilfe von Dell Validated Designs, um die Inferencing-Hardware und -Software für GenAI zu implementieren
- **Einführen** – Schnellere Wertschöpfung bei GenAI-Anwendungsfällen durch die Implementierung eines vorab trainierten Inferencing-Modells
- **Skalieren** – Managen Ihres GenAI-Innovationsportfolios mithilfe von technischen ExpertInnen vor Ort und Schulungsangeboten zur Weiterqualifizierung Ihres Teams

Technische Daten

Die Validated Design-Konfigurationen basieren auf den modernsten, für die KI-Beschleunigung optimierten Dell PowerEdge XE- und Rack-Servern, die die neuesten NVIDIA-GPUs und NVIDIA AI Enterprise mit Triton Inference Server und dem NeMo Framework nutzen. Durch Dell PowerScale-All-Flash- oder -Hybrid-Storage-Arrays werden ein schneller, großer Data-Lake-Storage für generative KI sowie umfangreiche Sprachmodelle bereitgestellt.

Compute	Accelerator	Netzwerke	Software	Storage
Dell PowerEdge R760xa-Server	NVIDIA A100- oder H100-GPUs	NVIDIA Networking, Dell PowerSwitch S5232F-ON oder S5248F-ON	Dell OpenManage Enterprise, Power Manager, CloudIQ. NVIDIA AI Enterprise mit NeMo Framework für LLMs und Triton Inference Server; NVIDIA Base Command Manager Essentials	Unterstützt durch Dell PowerScale, ECS und ObjectScale

Dell Technologies und NVIDIA

Dell Technologies und NVIDIA arbeiten gemeinsam an der Umsetzung und Beschleunigung von Workloads für generative KI. Das Angebot umfasst technisch validierte Hardware und Software für schnellere KI-, ML- und DL-Workloads, um die Kundenanforderungen in allen Unternehmen und Branchen zu erfüllen. Das Validated Design für LLM-Inferencing beschleunigt Ihre digitale Transformation mit Echtzeitdaten, die wichtige Entscheidungen in großem Umfang verbessern. Dies gelingt durch Lösungen, die für die schnellste Time-to-Value Ihrer KI-Initiativen optimiert sind.

Weitere Informationen zu den Lösungen von Dell

Kontakt zu Dell Technologies ExpertInnen

Weitere Ressourcen

Diskutieren Sie mit: #HashTag

DELL Technologies

NVIDIA