Company: KAYTUS Germany GmbH
Company Description: KAYTUS ist ein führender Anbieter von KI-basierten und flüssigkeitsgekühlten End-to-End IT-Infrastrukturlösungen und bietet eine Reihe von innovativen, offenen und umweltfreundlichen Infrastrukturprodukten für Cloud, KI, Edge und andere aufkommende Anwendungsszenarien. Mit einem kundenzentrierten Ansatz reagiert KAYTUS durch sein agiles Geschäftsmodell flexibel auf die Bedürfnisse der Nutzer.
Nomination Category: Kategorie für Neue Produkte und Produktmanagement Awards
Nomination Sub Category: Business Technology Solution - Künstliche Intelligenz (KI)/ /Lösungen für Machine Learning - Andere
Nomination Title: MotusAI AI DevOps Plattform

1. Releasedatum des neuen Produktes oder der neuen Dienstleistung bzw. Releasedatum der neuen Version

   14. Mai 2024

2. Reichen Sie folgende Informationen online ein, um in diesen Kategorien teilzunehmen: Schriftliche Antworten auf die folgenden Fragen ODER ein Video von bis zu fünf (5) Minuten Länge, das alle diese Fragen beantwortet. Bitte wählen Sie eine Möglichkeit aus:

   Schriftliche Beantwortung der Fragen

3. Wenn Sie ein bis zu fünf (5) Minuten langes Video einreichen, geben Sie bitte hier die URL des Videos ein ODER fügen Sie das Video über den obigen Link "Hinzufügen von Anhängen, Videos oder Links zu diesem Beitrag" Ihrer Nominierung bei, und laden Sie so eine Kopie Ihres Videos hoch.

4. Beschreiben Sie die bisherige Marktleistung, kritische Aufnahme und Kundenzufriedenheit mit dem Produkt oder der Dienstleistung. Nennen Sie, wenn möglich, die bisherigen Verkaufszahlen in Geld- oder Stückzahlen und geben Sie an, wie diese im Vergleich zu den Erwartungen oder der bisherigen Leistung ausfallen. Geben Sie Links zu guten Produkt- oder Dienstleistungsbewertungen an. Fügen Sie, falls zutreffend, einige Kundenmeinungen bei (bis zu 350 Wörter).

   Total 301 words used.

   Als aufstrebendes KI-Produkt hat MotusAI dank sejner flexiblen, branchenübergreifenden Lösungen schnell eine globale Verbreitung erreicht und ist in mehreren Marktsegmenten in Thailand, Japan, Südkorea und Europa vertreten. Das Produkt zielt auf hochwertige Branchen wie Bildung, Finanzen, Gesundheitswesen und KI-Startups ab und integriert seine KI-Fähigkeiten erfolgreich in kritische Szenarien wie städtische öffentliche Verwaltung, industrielle Produktionslinien und Bildungsinfrastruktur.

   - Eine wegweisende Forschungsuniversität in Thailand hat MotusAI eingesetzt und durch intelligente Planung und GPU-Sharing eine GPU-Auslastung von über 90 % erreicht. Die Lösung ermöglichte die parallele Verarbeitung von mehr als 30 priorisierten Aufgaben, reduzierte die Vorbereitungszeit für das verteilte Training durch adaptive Parameterabstimmung von Tagen auf Stunden und erreichte durch das Einhalten von Haltepunkten eine Quote von 90 % gültiger Trainingszeit, was wichtige Forschungsvorhaben um das Fünffache beschleunigte.

   -Eine Geschäftsbank in Indonesien setzte MotusAI ein und erreichte eine einheitliche Ressourcenplanung und automatische Fehlerbehebung, wodurch die Trainingszyklen für KI-Modelle von einer Woche auf einen Tag reduziert wurden. Die beschleunigte Bildverteilung und die optimierte Vernetzung steigerten die GPU-Effizienz um das 7-fache und ermöglichten einen schnellen KI-Einsatz in allen Geschäftsszenarien.

   - Ein führendes Unternehmen für autonomes Fahren in Japan setzte MotusAI ein, um eine intelligente Ressourcenplanung sowie GPU-Virtualisierung zu ermöglichen und die GPU-Auslastung auf 90 % zu steigern. Die Trainingseffizienz wurde um 35 % gesteigert, und dank der verteilten Aufgabenbereitstellung im großen Maßstab, erfolgte die Ausführung im Minutentakt. Außerdem wurden 4-5 gleichzeitige Trainingsaufgaben unterstützt und die Entwicklungszyklen um 67 % reduziert, was die Einführung eines ADAS (Advanced Driver Assistance System) für mehrere Fahrzeugtypen beschleunigte.

   - Mit Hilfe von MotusAI konnte ein führender koreanischer 3D-Vision-Technologie-Innovator die Datenübertragungslatenz um 30% reduzieren und gleichzeitig die Effizienz der GPU-Nutzung verbessern. Durch ein einheitliches Plattformmanagement wurde die Einrichtungszeit für verteiltes Training um das Achtfache beschleunigt, die parallele Entwicklung in über 10 Szenarien unterstützt und eine Reduzierung der Modelliterationen um 75 % erreicht.

5. Optional: Verweisen Sie auf alle beigefügten Zusatzmaterialien, die Sie mit Ihrer Nominierung hochladen und erklären Sie, wie diese Ihre Aussagen, die Sie in der Beantwortung der vorangegangenen Fragen getroffen haben, unterstützen (bis zu 250 Wörter).

   Total 69 words used.

   Presse Meldung:

   KAYTUS Released MotusAI, an AI Development Platform for Highly Efficient GPU Resource Scheduling and Task Orchestration

   Media Report:

   - DataCenter Insider, KAYTUS lanciert MotusAI, eine KI-Entwicklungsplattform für hocheffiziente GPU-Ressourcenplanung und Task-Orchestrierung
   - PresseBox – DE, KAYTUS lanciert MotusAI, eine KI-Entwicklungsplattform für hocheffiziente GPU-Ressourcenplanung und Task-Orchestrierung
   - Big Data Insider, Kaytus stellt KI-DevOps-Plattform MotusAI vor
   - DEV Insider, Kaytus stellt KI-DevOps-Plattform MotusAI vor
   - manage it, MotusAI: KI-Entwicklungsplattform für hocheffiziente GPU-Ressourcenplanung und Task-Orchestrierung

6. Beschreiben Sie die Merkmale, Funktionen und Vorteile des nominierten Produkts oder der nominierten Dienstleistung (bis zu 350 Wörter).

   Total 350 words used.

   MotusAI ist eine von KAYTUS entwickelte KI-DevOps-Plattform für das vollumfängliche Lifecycle-Management von KI-Modellen, von der Modellentwicklung bis zur Bereitstellung.

   Mit dem Übergang von GenAI vom Training zur Inferenz steigt die Akzeptanz von KI rapide an. So ermöglicht etwa das Inferenzmodell GPT-4.0 eine einheitliche Interaktion mit Bildern sowie Texten und bietet einen greifbaren Mehrwert durch die Lösung realer Probleme. Nach Angaben von PwC hat fast die Hälfte der Fortune-1000-Unternehmen KI vollständig in ihre Arbeitsabläufe integriert.

   Die Komplexität der KI von der Entwicklung bis zum Einsatz behindert jedoch den Bearbeitungsprozess der Anwender und bringt große Herausforderungen mit sich, darunter geringe Recheneffizienz und Stabilität sowie Komplexität bei Modell-Trainings und Inferenzierungen. Beispielsweise liegt die GPU-Auslastung beim LLM-Training bei 38-43 %, während die Inferenzauslastung bei nur 0,4 % liegt. Während des Trainings von Llama 3 wurde die Hälfte der unerwarteten Ausfälle durch GPU- und Speicherfehler verursacht.

   MotusAI adressiert diese Herausforderungen, indem es den KI-Modell-Workflow über den gesamten Lebenszyklus hinweg zuverlässig und stabil beschleunigt und vereinfacht.

   - MotusAI verbessert die Auslastung der Rechenressourcen durch eine optimierte Planungsstrategie. MotusAI bietet Planungs-Strategien, einschließlich feinkörnigem GPU-Scheduling und GPU-MIG-Instanzen, die eine durchschnittliche Auslastung der Rechenressourcen von über 70 % erreichen . MotusAI verfügt außerdem über mehrere Strategien zur Aufgabenplanung/Task -Scheduling-Strategien, die eine effiziente Task-Orchestrierung ermöglichen und einen 5-mal höheren Datendurchsatz sowie eine 5-mal geringere Latenzzeit bieten.
   - MotusAI gewährleistet einen stabilen Modellbetrieb mit hoher Fehlertoleranz und Verfügbarkeit. Die Plattform nutzt eine hochverfügbare IT-Architektur, und die Microservices werden über Lastausgleichsstrategien aufgerufen, um stabile Anfrageprozesse zu gewährleisten. Sie leitet automatisch Fehlertoleranzprozesse ein, um eine schnellstmögliche Datenwiederherstellung zu gewährleisten, wodurch die Fehlerbearbeitungszeit durchschnittlich um über 90 % reduziert wird.
   - MotusAI optimier KI-Training sowie Inferenzierung und verbessert die Effizienz. Für das Training bietet MotusAI ein Multi-User- und Multi-Organisations-Management. Außerdem werden die Zyklen für das Zwischenspeichern von Daten erheblich reduziert, was zu einer 2-3-fachen Steigerung der Modelleffizienz führt. MotusAI ermöglicht eine schnelle Fehlerbehebung und die automatische Wiederaufnahme von Checkpoints, um die Auswirkungen von Unterbrechungen zu minimieren. Für Inferenzierungen bietet MotusAI Funktionen zur prozessübergreifenden Inferenzbereitstellung mit geringem Codeaufwand und unterstützt verschiedene Anwendungsszenarien für eine einheitliche Bereitstellung, wodurch die Bereitstellungszeit von 2 Tagen auf 5 Minuten reduziert wird.

Attachments/Videos/Links:

MotusAI AI DevOps Plattform

[REDACTED FOR PUBLICATION]