


# RESSOURCENOPTIMIERTES MLOPS MIT MULTI-MODELL-INFERENZ PER CI/CD IN DER PRIVATE CLOUD

Die Innovationsabteilung der BarmeniaGothaer schafft gemeinsam mit SVA skalierbare und zukunftssichere Systeme zur Optimierung der MLOps-Pipelines für künstliche Intelligenz.

## AUF EINEN BLICK

### AUFGABE

Aufbau einer GPU-beschleunigten Private Cloud zur kontinuierlichen Bereitstellung diverser KI-Modelle

### SYSTEME UND SOFTWARE

> Cisco UCS X
> NVIDIA Virtualisierung & Inferenzserver
> Rancher Kubernetes
> GitHub CI/CD

### DIE VORTEILE

> schnelle und einfache Skalierung der Modelle
> optimierte Handhabung für die IT-Administration
> modulare und einfache MLOps-Pipeline zum Ausrollen verschiedener KI-Applikationen

## BARMENIAGOTHAER FINANZHOLDING AG

Die BarmeniaGothaer Versicherungen zählen zu den führenden Versicherungsgesellschaften in Deutschland. Mit innovativen Produkten, nachhaltigem Handeln und einer starken Kundenorientierung bietet die BarmeniaGothaer maßgeschneiderte Lösungen für Gesundheit, Vorsorge und Absicherung.

Die Entwicklung eigener KI-Modelle setzt Maßstäbe für zukunftsorientierte Ansätze. Durch Effizienzsteigerungen in der Sachbearbeitung und innovative Beiträge zur Produktentwicklung werden Individualfokus und Wirtschaftlichkeit der Leistungen gesteigert.

## HERAUSFORDERUNG

KI-Modelle stellen zunehmend steigende und heterogenere Anforderungen an die zugrundeliegende Infrastruktur. Die BarmeniaGothaer setzte sich daher das Ziel, strategische Mehrwerte durch das frühzeitige Aufbrechen von monolithischen Servicearchitekturen zu erzielen. Da jede Applikation dedizierte Ressourcen benötigt, sollte eine nachhaltigere Entwicklungs- und Ressourcennutzung im Produktivbetrieb von KI gewährleistet werden. Hierzu sollten die verschiedenen Technologien mit ihren individuellen Rollen, Skillsets, Wartungen und Strategien standardisiert und so Prozesse vereinfacht werden.

## LÖSUNG

In einer Pilotierungsphase wurden die Anforderungen in enger Zusammenarbeit mit den Experten der SVA erarbeitet. Die auf Basis dieser Anforderungen durchgeführte Marktsichtung berücksichtigte auch etablierte BarmeniaGothaer IT-Strategien, die bisher nicht genutztes Cluster Management mit Rancher Kubernetes und Grafana einschlossen. Da Rechenzentrenbetrieb, Sicherheits- und Storagekonzepte bereits erfolgreich mit Cisco Lösungen liefen, war Cisco UCS der unternehmensweit gesetzte Standard mit einem

Barmenia Gothaer

cisco Partner

NVIDIA Elite Partner

## REIBUNGSLOSE TRANSFORMATION

Featureset, das gut zu den Anforderungen von KI-Applikationen passt. Dies war die Grundlage für die Ausarbeitung einer geeigneten Servicearchitektur und Prozessen zum Betrieb und der Weiterentwicklung mittels CI/CD.

Zunächst standen hierbei die Überführung der Lösungen in Kubernetes-native Microservices und die Sättigung der Infrastrukturanforderungen bei gleichzeitig effizientem Ressourcenmanagement, beschleunigtem MLOps-Zyklus und hoher Servicequalität im Fokus. Die SVA Experten unterstützten bei der Transformation der Infrastruktur durch die Erweiterung der bestehenden Private-Cloud-Umgebung. Zur Bereitstellung von GPU-Ressourcen auf Kubernetes wurde das Cisco UCS X-Datacenter mit GPU-Kapazitäten ausgebaut. Durch die Kombination von identitätslosen UCS X-Systemen mit Virtualisierung von NVIDIA und VMware wird maximale Flexibilität für zukünftige Ressourcenprofile ermöglicht. Die Identitätsprovision zur Laufzeit und die vollständige Verwaltung von GPU-Virtualisierung und -Monitoring mittels Kubernetes Operatoren und CISCO ACI-Komponenten ermöglichen nun den Entwicklern der BarmeniaGothaer die Integration und Steuerung der Ressourcen dort, wo sie benötigt werden: im Applikationslayer.

Eine zentrale Rolle bei der Gestaltung der MLOps-Architektur nehmen die CI/CD-Pipelines ein, welche den Automatisierungsgrad und die Effizienz der Modellbereitstellung verbessern. Die zugehörigen Unternehmensprozesse wurden auf die Rollenverteilung und Strategien der Experten der BarmeniaGothaer Innovationsabteilung abgestimmt. Die Produktivsetzung von Modellen wird daher nun durch ein standardisiertes Konfigurationsmanagement mit GitOps-Pipelines realisiert. Zudem wird durch den Gebrauch von GPU-Frakturierung der Betrieb verschiedener Modelle oder Workloads auf einer GPU möglich, ohne die Performancekriterien zu brechen. Änderungen im System erfolgen kontrolliert und nachvollziehbar über Git-gestützte Workflows, wodurch die Governance verbessert und die Betriebssicherheit gesteigert wird.

Zur transparenten Überwachung und proaktiven Fehlererkennung wurde außerdem das bestehende Monitoringkonzept erweitert. Hierzu zählen die Kubernetes Umgebung, der Inferenzserver und Beschleuniger-spezifische Komponenten wie der Data Center GPU Manager.

## ERFOLGREICHES FAZIT

Durch die enge Zusammenarbeit mit SVA und den Einsatz von Cisco UCS X konnte die BarmeniaGothaer ihre IT-Systeme nachhaltig optimieren. Die neue Infrastruktur ist nicht nur flexibler und skalierbarer, sondern unterstützt auch ressourceneffizientes MLOps, was die Grundlage für zukunftsweisende Anwendungen im Bereich KI bildet. SVA hat mit fachlicher Expertise und der gezielten Auswahl von Technologien dazu beigetragen, dass die BarmeniaGothaer den Schritt in eine moderne IT-Zukunft erfolgreich meistern konnte.

KONTAKT
SVA System Vertrieb Alexander GmbH
Borsigstraße 26
65205 Wiesbaden
Tel. +49 6122 536-0
mail@sva.de
www.sva.de

Barmenia Gothaer

cisco Partner

NVIDIA Elite Partner