Kostentreiber KI

Strategie-Tipps zur Leistungs- und Kostenoptimierung

Unternehmen nutzen vermehrt künstliche Intelligenz zur Optimierung ihrer operativen Effizienz und Produktinnovation. Eine aktuelle Umfrage des Beratungsunternehmens McKinsey [*] zeigt, dass 40 Prozent der befragten Unternehmen aufgrund der rapiden Fortschritte im Bereich der generativen KI ihre Investitionen in KI-Technologien generell erhöhen wollen.

Ein Nachteil des zunehmenden Einsatzes ist jedoch, dass KI – insbesondere generative KI – rechenintensiv ist und die Kosten mit der Menge der Daten steigen, auf denen die KI-Modelle trainiert werden. Es gibt drei Hauptgründe, weshalb KI sich ohne entsprechende Kontrolle rasch zu einem Kostentreiber entwickeln kann:

1. **KI verbraucht zusätzliche Ressourcen:** Die Ausführung von KI-Modellen und die Abfrage von Daten erfordert große Mengen an Rechenressourcen in der Cloud, was zu höheren Cloud-Kosten führt.

2. **KI erfordert mehr Rechenleistung und Speicherplatz:** Das Trainieren von KI-Daten ist ressourcenintensiv und kostspielig aufgrund der erhöhten Anforderungen an Rechenleistung und Speicherplatz.

3. **KI führt häufige Datenübertragungen durch:** Da KI-Anwendungen häufige Datenübertragungen zwischen Edge-Geräten und Cloud-Anbietern erfordern, können zusätzliche Kosten für die Datenübertragung entstehen.

Wenn Unternehmen mit ihrer KI-Einführung erfolgreich sein wollen, müssen diese die Ursachen steigender Kosten verstehen und optimieren. Dies kann durch die Einführung einer soliden FinOps-Strategie geschehen. FinOps ist ein Konzept für die Verwaltung der Public Cloud, das darauf abzielt, die durch die Cloud-Nutzung entstehenden Kosten zu kontrollieren, und bei dem Finanzen und DevOps aufeinander treffen. Darüber hinaus sollten Unternehmen die Observability von KI berücksichtigen.

Grundlagen der KI-Observability

KI-Observability ist der Einsatz künstlicher Intelligenz zur Erfassung von Leistungs- und Kostendaten, die von verschiedenen Systemen in einer IT-Umgebung erzeugt werden. Darüber hinaus liefert KI-Observability IT-Teams auch Empfehlungen, wie sie diese Kosten eindämmen können. So unterstützt die KI-Observability die FinOps-Initiativen in der Cloud, indem sie aufzeigt, wie die Einführung von KI die Kosten aufgrund der erhöhten Nutzung von Speicher- und Rechenressourcen in die Höhe treibt. Da die KI-Observability die Ressourcennutzung in allen Phasen des KI-Betriebs überwacht – vom Modelltraining über die Inferenz bis hin zur Nachverfolgung der Modellleistung – können Unternehmen ein optimales Gleichgewicht zwischen der Genauigkeit ihrer KI-Ergebnisse und der effizienten Nutzung der Ressourcen herstellen und somit die Betriebskosten optimieren.

Best Practices

für die Optimierung der KI-Kosten mit KI-Observability und FinOps

- **Cloud- und Edge-basierter Ansatz für KI:** Cloud-basierte KI ermöglicht es Unternehmen, KI in der Cloud auszuführen, ohne dass diese sich um die Verwaltung, Bereitstellung oder Unterbringung von Servern kümmern müssen. Mit Edge-basierter KI können KI-Funktionen auf Edge-Geräten wie Smartphones, Kameras oder sogar Sensoren ausgeführt werden, ohne dass die Daten in die Cloud übertragen werden müssen. Durch die Einführung eines Cloud- und Edge-basierten KI-Ansatzes können IT-Teams somit von der Flexibilität, Skalierbarkeit und dem Pay-per-Use-Modell der Cloud profitieren und gleichzeitig die Latenz, Bandbreite und Kosten für das Senden von KI-Daten an Cloud-basierte Prozesse reduzieren.

- **Containerisierung:** Die Containerisierung ermöglicht es, KI-Anwendungen und Abhängigkeiten in eine einzige logische

Autor:
Christian Grimm
Director Sales Engineering - EMEA Central
Dynatrace
www.dynatrace.com

Einheit zu verpacken, die auf jedem Server mit den erforderlichen Abhängigkeiten problemlos bereitgestellt werden kann. Anstatt die Infrastruktur statisch auf Spitzenlasten einzustellen, können Unternehmen so eine dynamisch skalierbare Container-Infrastruktur für KI-Anwendungen nutzen und gleichzeitig Kosten optimieren.

- **Kontinuierliche Überwachung der Leistung von KI-Modellen:** Sobald ein Unternehmen KI-Modelle auf Grundlage seiner Daten trainiert, ist es wichtig, die Qualität und Effektivität des Algorithmus kontinuierlich zu überwachen. Die Überwachung von KI-Modellen hilft dabei, Bereiche mit Verbesserungsbedarf und „Drift" zu identifizieren. Im Laufe der Zeit ist oftmals davon auszugehen, dass KI-Modelle von den realen Bedingungen abweichen und dadurch ungenauer werden. IT-Teams müssen die Modelle daher gegebenenfalls anpassen, um neue Datenpunkte zu berücksichtigen. Die Abnahme der Vorhersagekraft als Ergebnis von Veränderungen in realen Umgebungen, die in den Modellen nicht berücksichtigt wurden, muss insofern überwacht werden.

- **Optimierung von KI-Modellen:** Diese Aufgabe geht Hand in Hand mit der kontinuierlichen Überwachung der Modelle. Es geht darum, die Genauigkeit, Effizienz und Zuverlässigkeit der KI eines Unternehmens zu optimieren, indem Techniken wie Datenbereinigung, Modellkomprimierung und Daten-Observability eingesetzt werden, um die Präzision und Aktualität der KI-Ergebnisse zu gewährleisten. Die Optimierung von KI-Modellen kann helfen, Rechenressourcen, Speicherplatz, Bandbreite und Energiekosten zu sparen.

- **Proaktives Management des KI-Lebenszyklus:** Zu den Aufgaben des IT-Teams gehören typischerweise das Erstellen, Bereitstellen, Überwachen und Aktualisieren von KI-Anwendungen. Die Verwaltung des KI-Lebenszyklus gewährleistet, dass KI-Anwendungen stets funktionsfähig, sicher, konform mit Compliance-Vorgaben und relevant sind, indem Tools und Verfahren wie Protokollierung, Auditing, Debugging und Patching eingesetzt werden. Die Verwaltung eines KI-Lebenszyklus hilft, technische Probleme, ethische Dilemmas, rechtliche Probleme und Geschäftsrisiken zu vermeiden.

- **Generative KI in Verbindung mit anderen Technologien:** Generative KI ist ein leistungsstarkes Werkzeug. Ihr volles Potenzial entfaltet sie jedoch erst in der Kombination mit prädiktiver und kausaler KI. Prädiktive KI nutzt maschinelles Lernen, um Muster in vergangenen Ereignissen zu erkennen und Vorhersagen über zukünftige Ereignisse zu treffen. Kausale KI ermöglicht die Ermittlung der genauen Ursachen und Auswirkungen von Ereignissen oder Verhaltensweisen. Kausale KI ist entscheidend, um die Algorithmen, die der generativen KI zugrunde liegen, mit qualitativ hochwertigen Daten zu versorgen. Composite AI bringt kausale, generative und prädiktive KI zusammen, um die kollektiven Erkenntnisse aller drei Verfahren zu verbessern. Bei Composite AI trifft die Präzision der kausalen KI auf die Vorhersagefähigkeiten der prädiktiven KI, um einen wesentlichen Kontext für generative KI-Prompts zu liefern.

Die Einführung von KI ermöglicht Unternehmen mehr Effizienz und Innovation, birgt aber auch die Gefahr ausufernder Kosten. Daher sollten Unternehmen ihre KI-Modelle proaktiv überwachen und verwalten, um sowohl die Datengenauigkeit als auch Kosteneffizienz ihrer KI-Modelle sicherzustellen. Eine Gesamtstrategie, die FinOps und KI-Observability einbezieht, kann Unternehmen dabei unterstützen, die Leistung und Kosten ihrer Systeme stets genau im Blick zu behalten.

[*] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year ◄