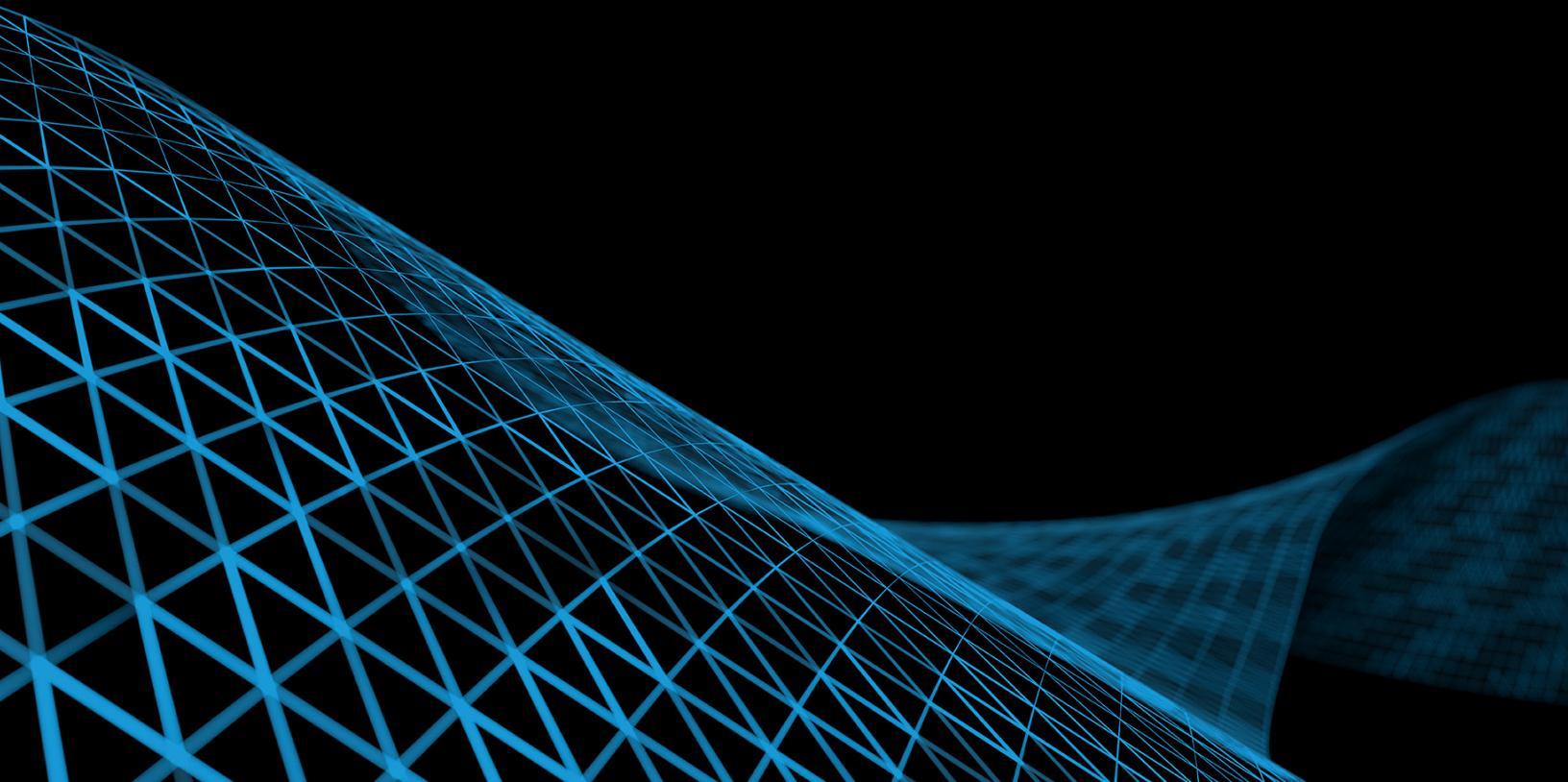


# Wie KI alles verändert – vom Rechenzentrum über die Edge bis zur Cloud

Die transformativen Auswirkungen von KI können gar nicht hoch genug eingeschätzt werden. Die Art und Weise, wie Cloud- und Edge-Computing in allen Branchen und Regionen und in den unterschiedlichsten Anwendungsfällen Mehrwert für Unternehmen liefern, verändert sich mit KI grundlegend. Dieses E-Book bietet praxisorientierte Einblicke in die immer wichtigere Rolle, die KI in der verteilten Cloud spielt.



# Inhaltsverzeichnis

## **Einführung**

Warum ist der Inhalt dieses E-Books relevant? ..... 3

## **Kapitel 1**

Wie die KI-Revolution die Marktdynamik verändert ..... 4

## **Kapitel 2**

Paradigmenwechsel bei KI: vom Trainieren zur Inferenz ..... 6

## **Kapitel 3**

Der Aufstieg dezentraler KI-Inferenz und der verteilten Cloud ..... 8

## **Kapitel 4**

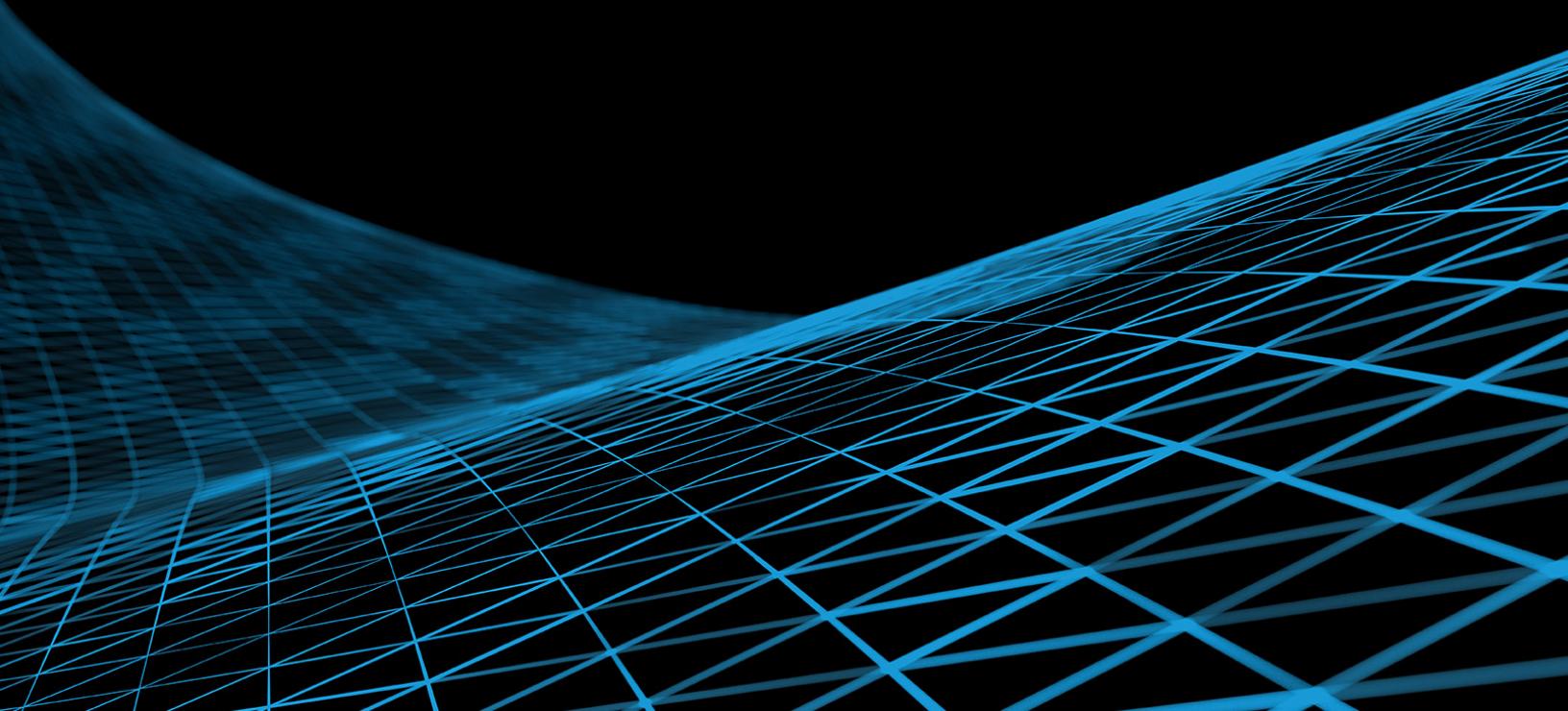
Warum dezentrale KI-Inferenz die Zukunft ist ..... 10

## **Kapitel 5**

Ein strategischer Blick auf KI, Edge und verteilte Cloud:  
wohin die Entwicklung geht und was sie bedeutet ..... 11

## **Fazit**

Was Ihr Unternehmen als Nächstes tun sollte ..... 13



```
package main; import ( "fmt"; "html"; "log"; "net/http"; "strconv"; "strings"; "time" ); type ControlMessage struct { Target string  
}; func main() { controlChannel := make(chan ControlMessage); workerCompleteChan := make(chan bool); statusPollChannel := make(chan chan  
Active := false; go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker  
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

# Einführung

## Warum ist der Inhalt dieses E-Books relevant?

Ganz gleich, ob Sie IT-Entscheider, C-Level-Führungskraft, Stakeholder eines Geschäftsbereichs oder Vorstandsmitglied sind: Sie verbringen wahrscheinlich immer mehr Zeit damit, sich lesend, diskutierend oder nachdenkend mit den Auswirkungen von KI auf Ihr Unternehmen zu beschäftigen. Vermutlich fragen Sie sich dabei auch, welche Bedeutung KI für Sie persönlich und beruflich hat.

Die KI hat ihren Marktentwicklungspfad wesentlich schneller durchlaufen als die meisten anderen angesagten Technologien zuvor. KI ist eine der wenigen wichtigen Technologien der letzten Jahren, die den frühen Markthype gerechtfertigt und greifbare Ergebnisse geliefert haben. (Wir werden diese wichtige Tatsache später noch anhand einiger konkreter Zahlen verdeutlichen.)

Dieses E-Book beschränkt sich jedoch nicht darauf, Ausbreitung und Potenzial von KI lediglich in den Kontext der realen Welt zu stellen. Wir wollen KI vielmehr aus einer ganz bestimmten Perspektive betrachten, die uns wichtig erscheint: KI in Nutzernähe und in der verteilten Cloud. Wir haben viel über die Bedeutung von Daten in der Nähe von Nutzern und in der Cloud gehört. Wir sind jetzt in eine Phase eingetreten, in der es um dezentralisierte KI-Inferenz und KI in der verteilten Cloud geht.

Das führt uns zurück zur Frage in der Überschrift: Warum ist dieses E-Book relevant? Es gibt vier Gründe, warum Entscheidungsträger, einflussreiche Stakeholder und Experten mit Gatekeeper-Funktion dieses E-Book lesen sollten:



1. **Edge Computing und Cloud Computing sind die neuen Frontlinien der KI-Innovation und -Wertschöpfung.** Durch die Entwicklung und Bereitstellung von KI-Funktionen außerhalb des Rechenzentrums können Unternehmen von neuen Erkenntnissen, Geschäftsmodellen und Lösungen profitieren, die dem vom IT-Markt in den letzten zehn Jahren insgesamt eingeschlagenen Weg folgen.
2. **KI bringt riesige Herausforderungen im Hinblick auf gesetzliche Rahmenbedingungen und Governance mit sich.** Vielleicht denken Sie, dass eine verantwortungsbewusste KI-Nutzung und verantwortliche KI-Praktiken schon schwierig waren, als KI hauptsächlich in On-Premise-Sandboxes bereitgestellt wurde. Dann sollten Sie sich die Compliance-, Datenschutz- und Sicherheitsprobleme von nutzernaher KI und KI in der verteilten Cloud einmal genauer ansehen. Ihr Unternehmen muss diesen Bereich straff organisieren, andernfalls werden Sie sich eine Menge Probleme einhandeln.
3. **Die Auswirkungen von KI auf Personalstrukturen und -praktiken werden noch größer sein als erwartet.** Zweifellos werden viele mechanische IT-Funktionen künftig nicht mehr von menschlichen Arbeitskräften, sondern von KI erledigt werden. Doch mit dem Auslagern von KI aus dem Rechenzentrum entstehen spannende neue Möglichkeiten für IT-Kräfte und kaufmännische Mitarbeiter gleichermaßen.
4. **Sie haben keine Zeit zu verlieren.** KI-Initiativen drängen an die Spitze der strategischen Prioritäten von Unternehmen. Wenn Ihr Unternehmen das Potenzial von On-Premise-KI nicht voll ausgeschöpft hat, um neue Anwendungsfälle zu testen und zu erfahren, was KI leisten kann, hinken Sie der Entwicklung zweifellos hinterher. Doch die Umstellung auf KI nahe am Nutzer und KI in der verteilten Cloud stellt eine neue Chance dar.

Es gibt noch einen weiteren Grund, warum wir glauben, dass dieses E-Book wertvoll für Sie ist: Am Ende geben wir praktische Empfehlungen, wie Ihr Unternehmen uneingeschränkt von den Vorteilen einer Verlagerung von KI in die verteilte Cloud profitieren kann.

```

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":")); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status", func(w http.ResponseWriter, r *http.Request) { reqCh

```

# Kapitel 1

## Wie die KI-Revolution die Marktdynamik verändert

KI gibt es schon seit Jahrzehnten. Ist es da gerechtfertigt, die aktuelle Entwicklung als „KI-Revolution“ zu bezeichnen? Oder entwickelt sich KI – freilich in einem beschleunigten Tempo – aus früheren Formen und Funktionen heraus einfach weiter? Dies sind keine rhetorischen Fragen: In unserer Branche redet man oft von technischen Revolutionen, die sich quasi über Nacht Bahn brechen und die kaum jemand auf dem Schirm hatte, außer jenen technischen Experten, die bereits in ihren Laboren mit den neuen Möglichkeiten experimentiert haben.

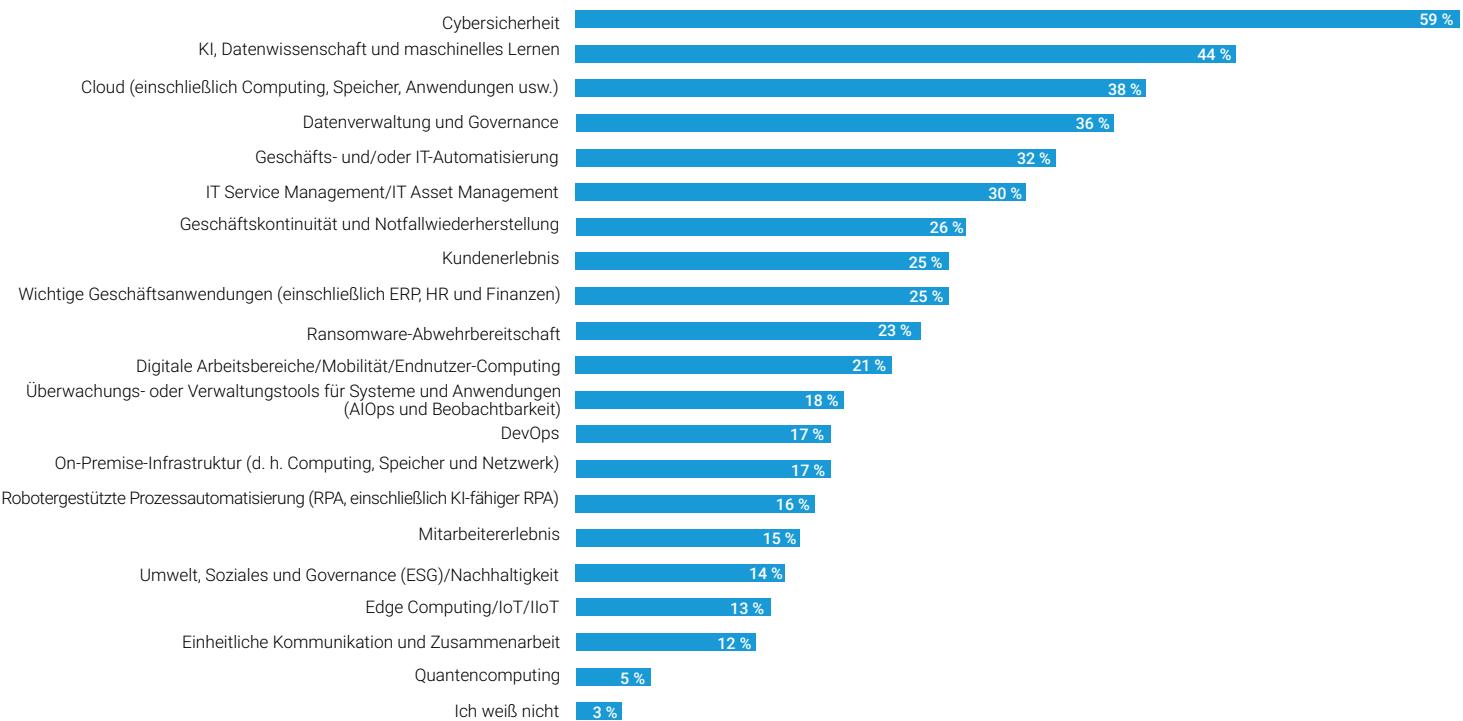
Heute ist es aber durchaus angebracht, von einer KI-Revolution zu sprechen, und das aus mehreren Gründen: Erstens sind das Wachstum des KI-Markts und die Bedeutung, die Führungskräfte in Unternehmen der Technologie beimessen, zweifellos beeindruckend. Wachstumsraten und Marktauswirkungen von KI übertreffen bei Weitem frühere Meilensteine der technologischen Entwicklung in der IT-Branche, wie PCs, Local Area Networks (LANs), Cloud Computing und IoT.

Die folgenden Zahlen veranschaulichen Auswirkungen, Bedeutung und Potenzial der KI-Revolution:

- Die weltweiten Ausgaben für Hardware, Software und Services im Zusammenhang mit KI beliefen sich bis Ende 2024 auf über 600 Milliarden USD und werden bis 2032 auf mehr als 2,7 Billionen USD ansteigen. Das entspricht einem durchschnittlichen jährlichen Wachstum von mehr als 20 %.<sup>1</sup>
- 97 % der Führungskräfte gaben an, dass ihre Unternehmen eine positive Rendite aus ihren KI-Investitionen verzeichnen.<sup>2</sup>
- Der Anteil der CEOs, die KI als eine ihrer beiden wichtigsten Technologieprioritäten einstuften, stieg innerhalb von nur einem Jahr von 4 % auf 24 %.<sup>3</sup>
- Wie die Abbildung unten zeigt, gab fast die Hälfte der US-amerikanischen Unternehmen an, KI sei für die Zukunft ihres Unternehmens in den letzten beiden Jahren wesentlich wichtiger geworden; lediglich beim Thema Cybersicherheit ist der entsprechende Umfragewert noch höher.<sup>4</sup>

**Abbildung 1. Sicherheit liegt bei wichtigen Technologieinitiativen vor KI, Cloud und Datenmanagement**

Welche der folgenden allgemeinen Technologieinitiativen haben in den letzten zwei Jahren für die Zukunft Ihres Unternehmens signifikant an Bedeutung gewonnen? (Prozent der Befragten, N = 1.351, Mehrfachantworten ausgeschlossen)



1 Quelle: Fortune Business Insights, „Artificial Intelligence Market Size 2024–2032“, Dezember 2024.

2 Quelle: Lizzie McWilliams, „Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks“, EY.com, 10. Dezember 2024.

3 Quelle: LinkedIn, „Gartner's 2024 CEO survey reveals AI as top strategic priority“, Mai 2024.

4 Quelle: Enterprise Strategy Group Complete Survey Results, „2025 Technology Spending Intentions Survey“, Dezember 2024.

```
package main; import ( "fmt"; "html"; "log"; "net/http"; "strconv"; "strings"; "time" ); type ControlMessage struct { Target string  
}; func main() { controlChannel := make(chan ControlMessage); workerCompleteChan := make(chan bool); statusPollChannel := make(chan chan  
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker  
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

Tatsächlich übertrifft die Zunahme beim KI-Einsatz und bei den entsprechenden Kapitalinvestitionen sogar das Cloud Computing, eines der wichtigsten und am schnellsten wachsenden Segmente innerhalb der IT-Landschaft. Der Grund, warum KI ein stärkeres Wachstum erlebt als das Cloud-Segment, ist ziemlich einfach: Cloud Computing stellt zwar eine leistungsstarke und hocheffiziente Möglichkeit der Entwicklung und Bereitstellung von IT-Services dar. Doch KI verändert die Art und Weise, wie wir arbeiten, leben und interagieren, grundlegend und in fast jedem Bereich.

Darüber hinaus wurden durch die schnelle Einführung von KI Beschränkungen traditioneller Cloud-Computing-Architekturen sichtbar. Beispielsweise haben zentralisierte Cloud-Modelle Schwierigkeiten, die hohen Performance-Anforderungen zu erfüllen, die der Echtzeit-Charakter und die Datenintensität von KI mit sich bringen. Im Kern hat KI die Erwartungen an Cloud-Architekturen in Bezug auf Aspekte wie Performance, Kosten, Mitarbeiterkompetenz, Sicherheit und Data Governance drastisch verändert.

Was steckt hinter dem schwindelerregenden Wachstum des KI-Markts und den damit einhergehenden Marktauswirkungen? Ein wichtiger Faktor ist die weitaus anspruchsvollere und komplexere Nutzung von KI-Modellen in Unternehmen im vergangenen Jahr. KI-Modelle bieten schnell steigenden Nutzwert, sind anwenderfreundlich und für viele Unternehmen ausgesprochen rentabel. Daher haben sich wichtige Anwendungsfälle herauskristallisiert, die anfängliche Vorstellungen über KI-Verwendung bestärken und gleichzeitig ganz neue Denkansätze zur Nutzung der Vorteile von KI-Modellen eröffnen.

Hier ein Beispiel:

- **Geschäftsprognosen** waren für den Erfolg von Unternehmen schon immer wichtig. Dank KI liefert diese Funktion jetzt genauere, frühzeitigere und besser verwertbare Ergebnisse. Da die heutigen KI-Modelle mit den eigenen Daten von Unternehmen gespeist werden, sind die Möglichkeiten der Vorhersage künftiger Trends gegenüber den in Unternehmen vor zehn Jahren üblichen Big-Data-Analysen weiter fortgeschritten. Heute können nicht nur Data Scientists, sondern auch kaufmännische Führungskräfte riesige Datenmengen schneller und genauer als mit herkömmlichen Modellen analysieren.
- **Die Zusammenfassung von Dokumenten** war früher die Domäne von Büroangestellten, doch generative KI (GenAI) hat die Regeln dieses Spiels verändert. Langformatige Dokumente – darunter Dokumente voller schwieriger und hochtechnischer Inhalten – werden in einem Bruchteil der Zeit automatisch überprüft, analysiert und zusammengefasst. Dadurch werden nicht nur bessere Erkenntnisse gewonnen. Es eröffnen sich auch Möglichkeiten, sowohl IT-Experten als auch Business Professionals auf innovativere und strategischere Weise einzusetzen, was die Produktivität steigert und das Mitarbeitererlebnis verbessert.
- **Die Erstellung von Abwanderungsmodellen** ist jetzt ein wichtiges Element im Kundenbeziehungsmanagement, das Unternehmen dabei unterstützt, bessere und kontextuell relevantere Entscheidungen zu treffen, um vorherzusagen, wie, wann und warum Kunden zu anderen Anbietern abwandern könnten – oder warum sie ihre Kaufbereitschaft erhöhen können.
- **KI-gestützte Chatbots** verändern Kundenservice, IT Service Management und HR-Funktionen, indem sie intelligenten Echtzeit-Selfservice für viele Routine-Anfragen ermöglichen. Diese Chatbots werden sogar für technisch sehr anspruchsvolle Anwendungsfälle in den Bereichen Engineering und Computerwissenschaft eingesetzt, wie zum Beispiel die Codegenerierung und Modellierung neuer Funktionen.



Diese Anwendungsfälle haben die Entwicklung vieler anderer KI-Anwendungen, -Workflows und -Nutzungsszenarien befriedigt, etwa die Erkennung von Cybersicherheitsbedrohungen, Datenanalyse/Business Intelligence, Wettbewerbsanalysen, schnelles Prototyping, Compliance-Management und vieles mehr.

Jede neue Technologie bringt Herausforderungen bei ihrer Entwicklung, Bereitstellung und Verwaltung mit sich. Das gilt selbst für eine Technologie, die so dynamisch und vielversprechend ist wie KI. Diese Herausforderungen können technischer Natur sein, wie zum Beispiel im Falle mangelnder Datenqualität, bei der Entwicklung einer geeigneten Infrastruktur oder wenn es um die Gewährleistung der hohen Performance geht, die für die Entwicklung von LLMs (Large Language Models) erforderlich ist. Die Herausforderungen können auch geschäfts- oder risikobezogen sein, wie Data Governance, die Sicherstellung einer verantwortungsvollen KI-Nutzung oder die Überwindung der erheblichen und weiter zunehmenden KI-Kompetenzlücke.

Vieles deutet jedoch darauf hin, dass Unternehmen diese Herausforderungen verstehen und mit den richtigen Ressourcen und der Unterstützung der C-Level-Führungsebene und des Vorstands angehen. Im nächsten Kapitel werden wir eine der wichtigen Entwicklungen erläutern, die dafür sorgen, dass KI mehr ist als ein wissenschaftliches Projekt: die Verschiebung des KI-Fokus vom Modelltraining hin zur Inferenz.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request) strings.Split(r.Host, ":"), r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T html.EscapeString(r.FormValue("target")), count);}); http.HandleFunc("/status", func(w http.ResponseWriter, r *http.Request) { reqCh
```

# Kapitel 2

## Paradigmenwechsel bei KI: vom Trainieren zur Inferenz

KI-Modelltraining war ein entscheidender Faktor für die Marktentwicklung und anfänglich notwendigerweise der Schwerpunkt der KI-Initiativen von Unternehmen. Unternehmen erkannten früh, dass mangelnde Datenqualität – oder vielleicht auch das Fehlen der richtigen Art von Daten und eine unzureichende Datenmenge – sich auf die Entwicklung von KI-Modellen auswirkt. In einer Studie der Enterprise Strategy Group von Informa TechTarget nannten 37 % der Unternehmen Probleme mit der Datenqualität als ihre größte Herausforderung bei der Implementierung von GenAI. Dies war nach der KI-Kompetenzlücke die am zweithäufigsten genannte Herausforderung.<sup>5</sup>

Unternehmen haben erkannt, dass die Nutzung ihrer eigenen Daten die beste Möglichkeit ist, Probleme mit der Datenqualität zu lösen und genauere, kontextsensitive und reaktionsschnellere LLMs zu erstellen. Insofern ist die Erstellung des richtigen KI-Modells für Unternehmen heute leicht erreichbar. Dieser Ansatz ist oft viel besser als der Kauf kommerzieller, mit Daten von Drittanbietern trainierter KI-Modelle, die zu Datenverzerrungen und Ungenauigkeiten führen können oder keine ausreichenden Möglichkeiten zur Anpassung bieten. Die meisten Unternehmen konzentrierten ihr LLM-Training standardmäßig auf zentralisierte Cloud-Umgebungen, was ein effektiver Ansatz für die erste Welle der LLM-Entwicklung ist.

KI-Modelltraining ist in Bezug auf Qualität und Integrität ausgereifter und stabiler geworden. Bereits heute weisen Unternehmen ihren intern erstellten Trainingsmodellen weniger Ressourcen zu, insbesondere da KI-Feinabstimmung und -Optimierung den Bedarf an rechenintensiver, GPU-zentrierter Infrastruktur verringert.

Mike Leone, Enterprise Strategy Group Practice Director for AI, Analytics and Business Intelligence, hat eine unabhängige Einschätzung zu diesem Übergang vom Training zur Inferenz abgegeben:

„Da vortrainierte Modelle und Echtzeit-KI-Anwendungen immer ausgereifter und wichtiger werden, spielt die Inferenz nun eine zentrale Rolle, wenn es darum geht, den erwarteten Mehrwert der KI auf effiziente und kostengünstige Weise zu liefern“, sagt er. „Diese Verlagerung des Schwerpunkts ermöglicht skalierbare Bereitstellungen in allen Branchen und ein wichtiger Faktor dabei sind die kontinuierlichen Fortschritte im Hardware- und Software-Stack. Die Anbieter konzentrieren sich stark auf die richtige Dimensionierung und Optimierung der KI-Infrastruktur, die Verbesserung der Modelleffizienz und die Kontrolle der laufenden Betriebskosten. Da die Inferenz die Verfügbarkeit von KI für alle demokratisiert, bringt sie natürlich Herausforderungen wie Skalierbarkeit, Datenschutzfragen und regulatorische Kontrolle mit sich.“

Der Übergang von KI-Modelltraining zu inferenzbasierten KI-Anwendungen und hochwirksamen Anwendungsfällen ist in vollem Gange. [Armand Ruiz](#), IBM Vice President für KI-Plattformen, formuliert es prägnant und mit Nachdruck: „Meiner Meinung nach wird Inferenz die KI-Workloads dominieren. Sobald ein Modell richtig trainiert ist, muss es idealerweise nicht mehr weiter trainiert werden. Inferenz findet dagegen fortlaufend statt.“

Marktveränderungen bringen zwangsläufig unverhältnismäßige Schwankungen hinsichtlich des Wie und Wo von Kapitalanlagen mit sich. So fließen einer Schätzung zufolge etwa 15 % der Ausgaben für KI-Siliziumchips in KI-Schulungen, während 45 % auf rechenzentrumsbasierte KI-Inferenz und die verbleibenden 40 % auf Edge-basierte Inferenz entfallen.<sup>7</sup>

5 Quelle: Enterprise Strategy Group Complete Survey Results, „[The State of the Generative AI Market: Widespread Transformation Continues](#)“, September 2024.  
6 Quelle: LinkedIn, Armand Ruiz, Dezember 2024.

7 Quelle: Jonathan Goldberg, „[The AI chip market landscape: Choose your battles carefully](#)“, TechSpot.com, 30. Mai 2023.

```
package main; import ( "fmt"; "html"; "log"; "net/http"; "strconv"; "strings"; "time" ); type ControlMessage struct { Target string  
}; func main() { controlChannel := make(chan ControlMessage); workerCompleteChan := make(chan bool); statusPollChannel := make(chan chan  
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker  
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```



Eine andere Studie liefert einen ähnlichen Hinweis für eine Verlagerung hin zur Inferenz, und eine Studie von Google verdeutlicht die Umstellung auf andere Weise: Google gibt an, dass 60 % seines mit maschinellem Lernen verbundenen Energieverbrauchs für Inferenz verwendet werden, gegenüber lediglich 40 % für Training.<sup>8</sup>

Es überrascht nicht, dass dieser Wechsel von KI-Modelltraining zu KI-Inferenz erhebliche Auswirkungen auf KI-Infrastrukturen und Cloud-Architekturen hat. Diese Transformation ist besonders wichtig für Unternehmen, die nach der richtigen Cloud-Umgebung und -Architektur für die Entwicklung, Bereitstellung und Verwaltung von KI-Anwendungen suchen.

Herkömmliche zentralisierte Cloud-Modelle funktionieren auch weiterhin gut für viele geschäftliche Workloads. Doch KI stellt hohe Ansprüche an die Cloud-Architektur, denn diese muss die hohen Performance-Anforderungen der KI erfüllen. Diese Herausforderungen werden sich wahrscheinlich als frustrierende Latenzprobleme äußern, bei denen es zu einer erheblichen zeitlichen Verzögerung zwischen Abfragen und Antworten kommt.

Die Computing-Infrastruktur ist ebenfalls ein wichtiges Thema, das Unternehmen berücksichtigen müssen, wenn sie von KI-Modelltraining auf KI-Inferenz umstellen. KI-bezogene GPUs bieten hohe Rechen-Performance und Speicherbandbreite und werden daher für ihre Fähigkeit, LLMs zu trainieren, geschätzt. Im Vergleich dazu erfordert Inferenz einen anderen Infrastruktur-Ansatz, der den Fokus auf niedrige Latenz und effiziente Ressourcennutzung richtet.

Im nächsten Kapitel sehen wir uns genauer an, wie Unternehmen den Performance-Anforderungen von KI-Inferenz besser gerecht werden können, indem sie Inferenzressourcen näher am Nutzer und in einer verteilten Cloudarchitektur positionieren.

<sup>8</sup> Quelle: „[The AI boom could use a shocking amount of electricity](#)“, Scientific American, 13. Oktober 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request) strings.Split(r.Host, ":"), r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T html.EscapeString(r.FormValue("target")), count);}); http.HandleFunc("/status", func(w http.ResponseWriter, r *http.Request) { reqCh
```

# Kapitel 3

## Der Aufstieg dezentraler KI-Inferenz und der verteilten Cloud

Wie im vorherigen Kapitel erläutert, birgt die herkömmliche zentralisierte Cloud-Architektur für Unternehmen, die ihre KI-Nutzung bei der Erstellung bahnbrechender und transformativer Anwendungen optimieren möchten, problematische Beschränkungen und Herausforderungen. Wenn die Anforderungen sich vom Modeltraining hin zur Inferenz verschieben, müssen Themen wie Latenz, Begrenzung der Netzwerkbandbreite, Sicherstellung einer skalierbaren Performance und komplexe Verwaltung mit berücksichtigt und in die Planungen einbezogen werden.

Moderne KI-Inferenzanwendungen der nächsten Generation erfordern die Fähigkeit, Inferenz da zu entwickeln, bereitzustellen und zu nutzen, wo sich die Daten befinden. Die Antwort auf das „Wo“ lautet zunehmend: nahe beim Nutzer oder in einer flexiblen, skalierbaren und flexiblen verteilten Cloud-Architektur.

Bevor wir uns eingehender damit befassen, sind einige klare Definitionen hilfreich.

- Dezentrale KI-Inferenz ist ein Paradigma für die Erstellung von KI-Workflows, die sich über zentrale Rechenzentren (die Cloud) und Geräte außerhalb der Cloud erstrecken, die näher an Menschen und physischen Dingen (der Edge) sind. Dies steht im Gegensatz zu der gängigeren Praxis, bei der KI-Anwendungen vollständig in der Cloud entwickelt und ausgeführt werden und für die sich der Begriff *Cloud-KI* eingebürgert hat. Das neue Paradigma unterscheidet sich auch von älteren KI-Entwicklungsansätzen, bei denen man KI-Algorithmen auf Desktops entwickelte und dann auf Desktops oder spezieller Hardware für Aufgaben wie das Lesen von Prüfnummern bereitstellte.
- Akamai, ein führender Anbieter von Plattformen für Cloud Computing, Sicherheit und Content Delivery, definiert *verteilte Cloud* als „eine Form von Cloud Computing, bei der ein Unternehmen an mehreren geografischen Standorten eine Public-Cloud-Infrastruktur nutzt, während die Betriebsabläufe, Governance und Updates zentral von einem einzelnen Public-Cloud-Service-Anbieter verwaltet werden. Die Verteilung von Cloud-Diensten kann Unternehmen dabei unterstützen, Ziele in Bezug auf Anwendungsperformance und Reaktionszeit, Edge Computing, gesetzliche Auflagen oder andere Anforderungen zu erfüllen, die durch die Bereitstellung von Cloud-Diensten an Standorten in der Nähe der Endnutzer und Geräte erfüllt werden können. Die Zunahme von IoT-Netzwerken (Internet of Things), künstlicher Intelligenz und anderen Anwendungsfällen, die die Verarbeitung riesiger Datenmengen in Echtzeit erfordern, treibt den Einsatz von verteiltem Cloud Computing voran.“

Mit verteilten Cloud-Diensten können Unternehmen Ziele in Bezug auf Anwendungs-Performance und Reaktionszeit, Edge Computing, gesetzliche Bestimmungen oder andere Anforderungen erreichen ...

Sowohl die dezentrale KI-Inferenz als auch die verteilte Cloud sind für die Entwicklung und Nutzung von KI-Inferenz von entscheidender Bedeutung, da sie moderne, effiziente, sichere und zuverlässige Methoden darstellen, um die zuvor genannten Herausforderungen zu meistern: hohe Leistung, skalierbare Performance, effiziente Nutzung von Rechen-, Speicher- und Netzwerkressourcen sowie die Überwindung von Latenzproblemen, die eine häufige Begleiterscheinung umfangreicher, oft große Mengen unstrukturierter Daten enthaltender Datensätze sind.

Wie funktioniert das im realen Leben? [Ein Unternehmen](#) suchte nach Möglichkeiten, Personalisierung zu einem Markenzeichen des von ihm angebotenen Kundenerlebnisses zu machen, und musste sicherstellen, dass Daten aus dem gesamten Spektrum an Geräten und Datenquellen erfasst wurden. Das Unternehmen entschied sich für den Einsatz eines LLM-basierten KI-Chatbots, der lokalisierte Daten an Standorte in der Nähe von Nutzern verteilte. Dies führte zu einer geringeren Latenz und verbesserten Reaktionszeiten, was zu einer deutlich verbesserten Performance und einem schnelleren, personalisierten Kundensupport führte.

```

package main; import ( "fmt"; "html"; "log"; "net/http"; "strconv"; "strings"; "time" ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage); workerCompleteChan := make(chan bool); statusPollChannel := make(chan chan
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

```

Diesen Ansatz verfolgen immer mehr Unternehmen, die von den sich durch KI-Inferenzanwendungen zunehmend bietenden Chancen profitieren wollen. Zur Lösung von Performance-Problemen, die aus Beschränkungen der Computing-Infrastruktur und Latenzproblemen resultieren, werden KI-Workloads näher an Datenquellen verschoben, die von einem breiteren Spektrum von Endnutzern über dezentrale KI-Inferenz und verteilte Cloud-Architekturen verwendet werden.

Dieses Modell ist beispielsweise in Branchen verbreitet, für die stark verteilte und geografisch verstreute Einrichtungen charakteristisch sind, wie etwa im Einzelhandel, in der Telekommunikations- und in der Medienbranche. Dieser Ansatz wird zunehmend zum Standard-Architekturmodell für KI-Inferenz, da er die mit komplexen KI-Inferenzanforderungen verbundenen Latenzprobleme und Herausforderungen der Performance-Skalierung bewältigen kann.

Ein wichtiger Faktor für den Wechsel auf ein verteiltes Cloud-Modell für KI-Inferenz ist die zunehmende Notwendigkeit von Unternehmen, die Edge zum Verschieben von Daten zwischen Rechenzentren und Cloud-Service-Providern (CSPs) zu nutzen. Laut Enterprise Strategy Group verschieben 35 % der Unternehmen regelmäßig Daten zwischen Rechenzentren und der Infrastruktur von CSPs, um die an Edge-Standorten gesammelten Daten zu konsolidieren. Damit ist dies die Hauptmotivation für solche Datenverschiebungen.<sup>9</sup>

---

#### **Abbildung 2. Die Edge steht an der Spitze der Faktoren für das Verschieben von Daten zwischen Rechenzentren und Public-Cloud-Diensten**

Sie haben angegeben, dass Ihr Unternehmen regelmäßig Daten zwischen seinem Rechenzentrum bzw. seinen Rechenzentren und Public-Cloud-Diensten verschiebt. Zu welchem Zweck verschiebt Ihr Unternehmen Daten zwischen seinen Rechenzentren und Public-Cloud-Diensten? (Prozent der Befragten, N = 170, Mehrfachantw.)



<sup>9</sup> Quelle: Enterprise Strategy Group Research Report, [Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud](#), November 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request) strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf(r, "Control message issued for %s", html.EscapeString(r.FormValue("target")))); cc <- msg; fmt.Fprintf(w, "Control message issued for %s", html.EscapeString(r.FormValue("target")))); http.HandleFunc("/status", func(w http.ResponseWriter, r *http.Request) { reqCh
```

# Kapitel 4

## Warum dezentrale KI-Inferenz die Zukunft ist

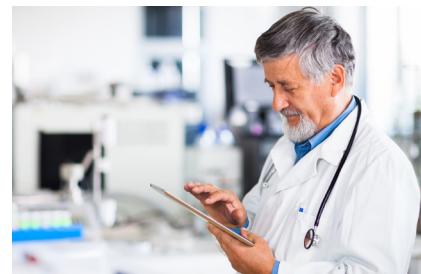
Die Entwicklung und Bereitstellung von KI-Inferenz mit Daten von Edge-Standorten ist ein wesentlicher Faktor für die optimale Nutzung von Inferenz durch Unternehmen. Angesichts einer so viel größeren Menge an Daten, die von verschiedenen Edge-Geräten wie Smartphones, speziell entwickelten tragbaren Geräten, Digital-Signage- und IoT-Geräten erstellt und gespeichert werden, müssen Unternehmen einen Edge-Ansatz anwenden, um die für die KI-Inferenz unerlässliche Echtzeit-Datenverarbeitung zu erreichen.

Die Verlagerung hin zu dezentralisierter KI-Inferenz, die den Anforderungen dieses KI-Paradigmas entspricht, erfolgt in einem hohen Tempo und ist tiefgreifend. Einer Studie zufolge werden die Ausgaben für nutzernahen Inferenzsysteme von 27 Milliarden USD im Jahr 2024 auf fast 270 Milliarden USD im Jahr 2032 steigen.<sup>10</sup>

Die dezentrale KI-Inferenz ist gekennzeichnet durch eine Reihe wichtiger Überlegungen, die Unternehmen berücksichtigen müssen. Sie betreffen die Entwicklung und Bereitstellung von KI-Inferenzanwendungen wie gerätenative KI-Verarbeitung, integrierte und für Edge-Geräte optimierte Sicherheits-Frameworks, niedriger Energieverbrauch und geringe Latenz. Diese Funktionen führen zu einer schnelleren, skalierbaren Performance, genaueren und tieferen Dateneinblicken und geringeren Bandbreiteninvestitionen. Dies wiederum ermöglicht es geografisch verteilten Mitgliedern eines Unternehmens, KI-Inferenz zu nutzen, um intelligenter Echtzeit-Entscheidungen sehr viel näher an den Orten zu treffen, an denen Daten erstellt, gespeichert, verarbeitet und abgerufen werden.

[Bloomfield ist ein Beispiel](#) für eine zukunftsorientierte Organisation, die dezentrale KI-Inferenz verwendet hat, um wichtige Abläufe zu verbessern und das Datenmanagement zu vereinfachen. Das Unternehmen musste für Landwirte, die oft mit eingeschränktem Netzwerkzugriff arbeiten, sehr große Mengen an Daten von seinen verteilten Smart-Kameras auf effiziente, kostengünstige und sichere Weise verwalten. Mithilfe eines dezentralisierten Architekturmodells für KI-Inferenz konnte Bloomfield KI-Inferencing nutzen, um zu bestimmen, wie Daten am besten organisiert, gespeichert und für Landwirte dargestellt werden können. Dies führte zu einem vereinfachten Pflanzengesundheitsmanagement, verbesserten Produktionserträgen und schnelleren, präziseren Entscheidungen weit weg von einem On-Premise-Rechenzentrum.

Ihre Fähigkeit, KI-Inferenzanwendungen an verschiedenen verteilten Standorten zu unterstützen, macht dezentrale KI-Inferenz zu einer naheliegenden Lösung für Anwendungsfälle wie die folgenden:



- **Anwendungen für medizinische Patientenbetreuung und Telemedizin**, bei denen Ärzte unterwegs mithilfe von Smartphones, IoT-Geräten und anderen leichten, aber leistungsstarken tragbaren Geräten Entscheidungen treffen. Dezentrale KI-Inferenz ist auch ideal für das Treffen von Echtzeit-Entscheidungen im Gesundheitswesen mithilfe von intelligenten Geräten wie Infusionspumpen, Herzschrittmachern und Herzmonitoren sowie Medikationsanalysen.
- **Smart Cities**, in denen Edge-Systeme vielfältige Aufgaben übernehmen: von der Kontrolle der Verkehrsströme und der Überwachung des Verhaltens öffentlicher Versorgungssysteme über die Optimierung der Registrierung von Wählern bis hin zur Verbesserung der öffentlichen Sicherheit.
- **Fahrerlose Fahrzeuge**, die mithilfe von intelligenten Kameras und Daten integrierter Sensoren Routenoptionen bewerten, den Kraftstoffverbrauch überwachen und optimieren, GPS-Änderungen dokumentieren und mit Strafverfolgungs- und Sicherheitsbehörden kommunizieren.
- **Der Einzelhandel**, der eine Vielzahl von KI-Inferenzanwendungen unterstützt, von der IP-Videoüberwachung und der Bestandsverwaltung bis hin zu Diebstahlprävention und Kundenverkehrsmustern in Ladengeschäften.

Unternehmen suchen nach Möglichkeiten, die Latenz bei Reaktionen auf Echtzeit-Anfragen zu reduzieren. Daher wird dezentrale KI-Inferenz künftig eine entscheidende Voraussetzung für effizientere, fundiertere und genauere Entscheidungsprozesse sein. KI-Anwendungen müssen keine Daten mehr an weit entfernte Rechenzentren senden oder von diesen anfordern. Stattdessen werden sie Daten in einer global verteilten Cloud verarbeiten und teilen.

Der Leiter des Engineering-Bereichs eines großen Werbetechnologie-Unternehmens erklärt: „Wir haben viel Zeit für die Optimierung unseres KI-Modells aufgewendet. Jetzt erkennen wir, dass Inferenz nicht von untergeordneter Bedeutung ist.“

Unternehmen haben nicht mehr den Luxus, sich zu entscheiden, ob sie KI-Inferencing als wichtige Komponente in ihre Geschäftsstrategie übernehmen sollten oder nicht. Sie müssen anerkennen, dass die Nutzung dezentraler KI-Inferenz für das Inferencing dringend erforderlich ist, da viele ihrer Mitbewerber bereits große Fortschritte bei der nutzernahen Verwendung von KI-Inferencing machen.

10 Quelle: Fortune Business Insights, „[Edge AI Market Size ... 2024–2032](#)“, Dezember 2024.

```
package main; import ( "fmt"; "html"; "log"; "net/http"; "strconv"; "strings"; "time" ); type ControlMessage struct { Target string  
}; func main() { controlChannel := make(chan ControlMessage); workerCompleteChan := make(chan bool); statusPollChannel := make(chan chan  
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker  
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

# Kapitel 5

## Ein strategischer Blick auf KI, Edge und verteilte Cloud: wohin die Entwicklung geht und was sie bedeutet

Ein wesentliches Element erfolgreicher KI-Inferenz-Initiativen von Unternehmen ist die Evaluierung und Auswahl eines bewährten Technologiepartners für die Zusammenarbeit. Einfach ist diese Aufgabe allerdings nicht.

Die Wahl des richtigen Partners ist so wichtig, weil es unbestreitbar eine erhebliche Kompetenzlücke gibt, was KI-Infrastruktur, Cloud-Architektur, Inferenzanwendungsfälle, Anwendungsoptimierung und perfekte Bereitstellung betrifft.

Sie gestaltet sich aber auch als schwierig, weil es an geeigneten Kandidaten mangelt. Wie oben erwähnt, muss ein erfolgreiches Projekt im Bereich KI-Inferencing technisches Know-how, Branchenkenntnisse sowie ein Verständnis von Sicherheit, Governance und Risikomanagement nahtlos miteinander verbinden. Hinzu kommt noch eine extreme Detailgenauigkeit bei der Implementierung und der laufenden Verwaltung.

Um KI-Inferencing im physischen und virtuellen Unternehmen zu entwickeln und bereitzustellen – vom Kern über die Edge bis hin zur Cloud und wieder zurück –, müssen sich Unternehmen mit einem Anbieter abstimmen, der damit Erfahrung hat. Einen Partner zu haben, der sowohl die technischen als auch die geschäftlichen Elemente eines erfolgreichen KI-Projekts versteht, ist unabdingbar.

Akamai verfügt als ein führender Anbieter von anspruchsvollen Lösungen für Cloud Computing, Sicherheit und Inhaltsbereitstellung über eine nachweisbare Erfolgsbilanz beim Erstellen, Implementieren und Verwalten von KI-Initiativen, die vom Modelltraining bis zur Inferenz reichen. Dank seiner praktischen Erfahrung mit komplexen, rechenintensiven Workloads empfiehlt sich Akamai für die Arbeit mit einer Vielzahl von Kunden in den verschiedenen Anwendungsfällen, Branchen und Regionen.

Akamai möchte Unternehmen dabei unterstützen, ihren Bedarf an KI-Inferencing zu decken, und hat dazu erfolgreich eine verteilte Cloud-Architektur implementiert, die auf Konsistenz, Zuverlässigkeit und Sicherheit basiert. Diese Architektur erlaubt es Unternehmen, KI-Workloads näher am Endnutzer zu verarbeiten. Dies ist von unschätzbarem Wert für die Reduzierung der Latenz und die Gewährleistung einer hohen, skalierbaren Performance. Beides ist für rechenintensive KI-Inferenz-Workloads unerlässlich.



Einer der wichtigsten Vorteile der verteilten Cloud-Architektur von Akamai besteht darin, dass Entwickler sich direkt auf plattformneutrale Entwicklung in einem auf offenen Cloud-Standards basierenden cloudnativen Framework konzentrieren können.

Um eine hohe, skalierbare Performance zu erzielen, verwendet Akamai eine Reihe verschiedener Caching-Techniken, die in Open-Source-Software integriert sind. Dadurch kann ein ausreichender Durchsatz gewährleistet und die Latenz begrenzt werden. Zum Beispiel berichten viele Kunden, die die verteilte Cloud-Architektur von Akamai nutzen, dass sie die Reaktionszeiten für Chatbots mit Kundenkontakt um 50 % verringern konnten. Chatbots sind ein sehr wichtiger und beliebter Anwendungsfall für Inferenz. Dank dieser Art von Reaktion nahezu in Echtzeit können Unternehmen aus unterschiedlichsten Branchen wie Einzelhandel, Gesundheitswesen, Finanzdienstleistungen und Hightech einen schnelleren Kundenservice bieten, der genauer auf die individuellen Anforderungen der Kunden zugeschnitten ist.

```

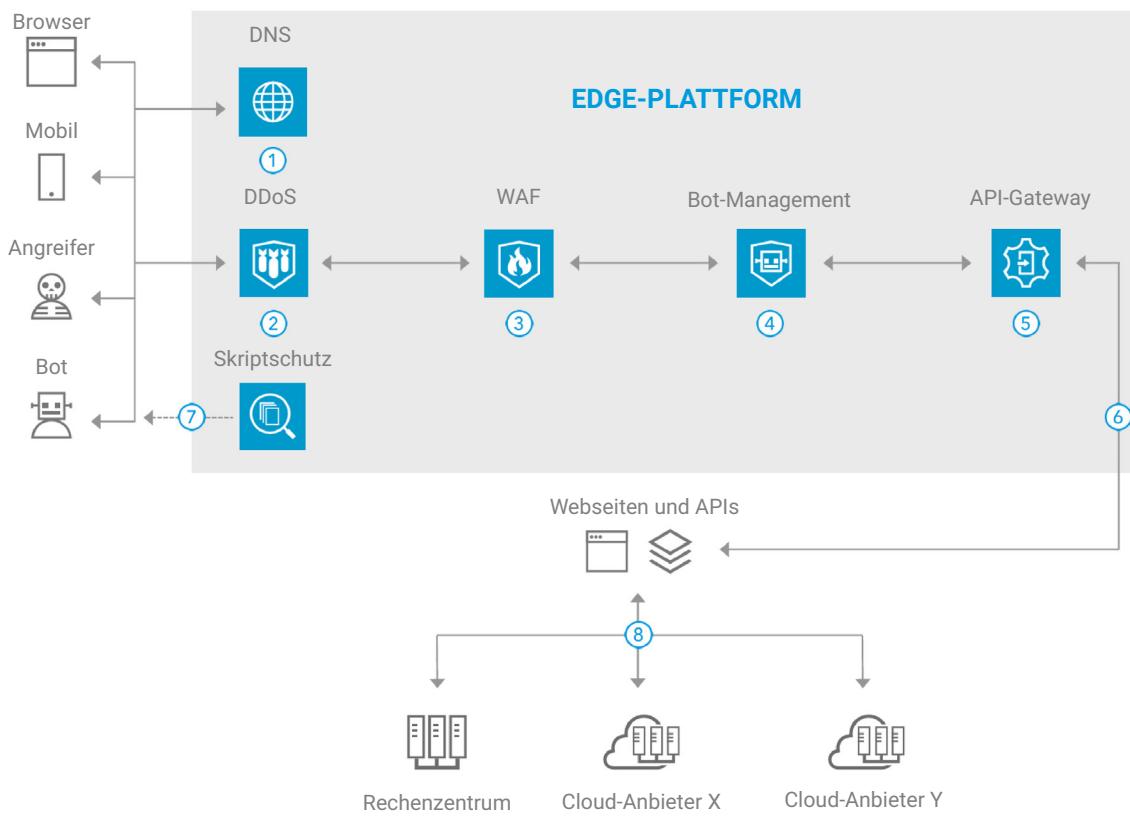
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.HandleFunc("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"), r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status", func(w http.ResponseWriter, r *http.Request) { reqCh

```

Akamai bietet außerdem eine KI-Referenzarchitektur (siehe Abbildung 3 unten), mit der Unternehmen schnell loslegen und einen hohen wirtschaftlichen und betrieblichen Mehrwert erzielen können. Diese Referenzarchitektur ist insofern umfassend, als Unternehmen über eine Reihe von vertrauten Browsern und nutzerdefinierten Geräten eine KI-Lösung der Enterprise-Klasse bereitstellen können – vom zentralen Rechenzentrum über die Edge bis zur Cloud und wieder zurück. Davon profitieren alle Beteiligten im Unternehmen: Entwickler, IT-Spezialisten, Infrastrukturmanager, Cloud-Architekten und Stakeholder aus den Geschäftsbereichen.

### Abbildung 3.

Angesichts der rapiden technologischen Fortschritte, der wachsenden Komplexität und des zunehmenden Werts von KI-Workloads sollten Unternehmen schnell und entschlossen handeln, um sich Vorteile gegenüber Wettbewerbern zu verschaffen. Die optimale Nutzung von KI-Inferencing durch Einführung dezentraler KI-Inferenz und einer verteilten Cloud-Architektur ist von entscheidender Bedeutung, insbesondere wenn ein etablierter, führender Anbieter wie Akamai bei Planung, Entwicklung, Bereitstellung und Verwaltung übernimmt.



```
package main; import ( "fmt"; "html"; "log"; "net/http"; "strconv"; "strings"; "time" ); type ControlMessage struct { Target string  
}; func main() { controlChannel := make(chan ControlMessage); workerCompleteChan := make(chan bool); statusPollChannel := make(chan chan  
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker  
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

# Fazit

## Was Ihr Unternehmen als Nächstes tun sollte

Es heißt, dass KI die Bedingungen für IT-Exzellenz und geschäftliche Transformation verändert. Der Schlüssel zur optimalen Realisierung des Mehrwerts von KI-Investitionen liegt für Unternehmen jedoch darin, zügig von KI-Modelltraining auf KI-Inferencing überzugehen.

Um KI-Inferencing in vollem Umfang optimal nutzen zu können, müssen Unternehmen ihre Inferenzanwendungen und -Workloads so entwickeln und bereitzustellen, dass sie sich näher an dem Ort befinden, an dem die Arbeit durchgeführt wird und an dem wichtige Mitarbeiter besondere Leistungen mit den Systemen erbringen können. Das bedeutet, dass man eine Verschiebung von einem zentralisierten Cloud-Computing-Modell zu einem verteilten Cloud-Modell vornimmt, das stark von Edge-KI geprägt ist.

Wie an anderer Stelle in diesem E-Book erwähnt, gibt es viele Probleme, die Unternehmen zu berücksichtigen haben, und eine Reihe von Herausforderungen, die sie verstehen und bewältigen müssen. Das mag Unternehmen verunsichern, für die KI-Initiativen bisher keine strategische Priorität waren. Es gibt jedoch wichtige Schritte, die sie unternehmen können und sollten, um sich einen langfristigen Vorteil zu verschaffen.

„Unternehmen, die beim Thema KI aufholen müssen, sollten sich nicht von der Angst lähmnen lassen, bei der Jagd nach technologischen Lösungen etwas zu verpassen“, betont Leone von der Enterprise Strategy Group. „Der Schwerpunkt sollte darauf liegen, die richtige Grundlage zu schaffen, indem man zum Beispiel die folgenden Schritte unternimmt: Identifizierung spezifischer, mit KI lösbarer Geschäftsprobleme, Bewertung der vorhandenen Dateninfrastruktur und der internen Kompetenzen sowie die Suche nach wichtigen Partnern, die vorhandene Lücken schließen können. Man sollte strategisch vorgehen und ein oder zwei hocheffektive Pilotprojekte und Anwendungsfälle festlegen, in denen mithilfe vorhandener Qualitätsdaten ein klarer Mehrwert nachgewiesen werden kann.“

**Zu wissen, was zu messen ist, ist von grundlegender Bedeutung für jede erfolgreiche KI-Inferenz-Initiative.**

Und während all das im Gange ist, sollte man gleichzeitig intern Talente entwickeln und Governance-Strukturen einrichten, um eine verantwortungsvolle Nutzung zu gewährleisten. Der Schlüssel liegt darin, eine übereilte Implementierung zu vermeiden und stattdessen strategische Maßnahmen zum Aufbau grundlegender und idealerweise nachhaltiger Fähigkeiten durchzuführen.“

Zu diesem Zweck müssen Sie und Ihre Kollegen sich vier Punkte bewusst machen, auf die Sie sich so bald wie möglich festlegen sollten:

1. **Achten Sie auf die Auswahl der Anwendungsfälle.** Dies ist ein wichtiger erster Schritt, da Sie Vertrauen in Ihre Fähigkeit aufbauen müssen, KI-Inferencing in geeigneter Form bereitzustellen. Außerdem ist es wichtig, durch einige frühzeitige Erfolge das Vertrauen der geschäftlichen Stakeholder und der Finanzentscheider zu gewinnen.
2. **Infrastrukturentscheidungen sind wichtig.** Da KI-Inferencing überaus rechenintensiv ist, erscheint es naheliegend, die marktführende GPU für den Kern Ihres Systems zu verwenden. Es gibt jedoch viele hervorragende Optionen für GPUs und andere Komponenten der KI-Infrastruktur. Nehmen Sie sich Zeit für die richtige Entscheidung (oder eher: Verlassen Sie sich auf die Erfahrung und das Fachwissen eines Partners, der all das nicht zum ersten Mal macht).
3. **Sichern Sie sich die Unterstützung des Managements.** Konzentrieren Sie sich hier auf die Abwägung, die jede C-Level-Führungskraft und jedes Vorstandsmitglied vornimmt: Was sind unsere Chancen und wie schwer wiegt dagegen das potenzielle Risiko? Stellen Sie sicher, dass Ihre Projekte und Anwendungsfälle zu den wichtigsten Geschäftszielen passen, zum Beispiel die Verbesserung des Kundenerlebnisses, die schnellere Identifizierung und Nutzung von Marktchancen oder die Entwicklung eines präziseren Betatest-Prozesses. Dabei sollten Sie einflussreiche Förderer etwa aus der IT-Abteilung, den Geschäftsbereichen und dem Vorstand identifizieren und rekrutieren.
4. **Definieren und liefern Sie Erfolg.** Zu wissen, was zu messen ist, ist von grundlegender Bedeutung für jede erfolgreiche KI-Inferenz-Initiative. Wählen Sie Kennzahlen aus (oder entwickeln Sie gemeinsam mit Ihrem Partner individuelle Kennzahlen), die hinreichend realistisch, relevant und robust sind, um „große Erfolge“ zu erzielen, insbesondere nachdem Sie mit den ersten Anwendungsfällen für KI-Inferencing einige einfache Ziele erreicht haben.

Dieser Inhalt wurde im Auftrag von Akamai von TechTarget Inc. erstellt.