Serverless Use Cases: Inspiration für die Praxis – Teil 3

Pay as you go: Skalieren ohne Grenzen

Florian Lenz

26 März 2025

Artikelserie: Serverless Use Cases: Inspiration für die Praxis

Teil 1: Serverless Edge Computing: Ganz nah am Nutzer

Teil 2: Serverless AI-Chatbot: Hilfe auf Abruf

Teil 3: Pay as you go: Skalieren ohne Grenzen

Plötzlich explodierende Nutzerzahlen oder Lastspitzen durch virale Kampagnen – unvorhersehbarer Traffic stellt viele Unternehmen vor Herausforderungen. Herkömmliche IT-Infrastrukturen stoßen hier schnell an ihre Grenzen: Entweder sind sie überlastet oder es werden teure Serverkapazitäten vorgehalten, die oft ungenutzt bleiben. Doch es gibt eine Lösung: Serverless-Architekturen und Pay-as-you-go-Modelle bieten eine flexible und kosteneffiziente Alternative, um Lastspitzen problemlos abzufangen – ohne sich um das Servermanagement kümmern zu müssen.

Der Black Friday steht vor der Tür. Die Werbekampagne war ein voller Erfolg, die Kunden strömen in Scharen auf die Website – und plötzlich passiert das Unfassbare: Alles bricht zusammen. Die Seiten laden nicht mehr, der Check-out hängt, frustrierte Kunden springen ab. Der Umsatz sollte eigentlich in die Höhe schnellen, stattdessen kommt es zum Desaster.

Dieses Szenario ist der Alptraum vieler Unternehmen – und doch passiert es immer wieder. Herkömmliche Infrastrukturmodelle stoßen an ihre Grenzen. Entweder sind die Kapazitäten zu knapp bemessen und die Server brechen unter der Last zusammen oder die Unternehmen investieren vorab in teure Überprovisionierung, um auf Nummer sicher zu gehen, zahlen aber auch dann, wenn die Server kaum ausgelastet sind. Das Resultat: hohe Kosten, kompliziertes Management und doch keine wirkliche Flexibilität.

Schwankender Traffic und begrenzte Ressourcen

In vielen Branchen, insbesondere E-Commerce, Ticketing, Streaming und Online-Gaming, ist der Traffic nie konstant. Ein Social-Media-Trend, eine erfolgreiche Werbekampagne oder ein saisonales Ereignis wie der Black Friday können dazu führen, dass sich die Besucherzahlen innerhalb weniger Minuten vervielfachen. Unternehmen stehen dann vor einem entscheidenden Problem: Wie kann die IT-Infrastruktur mit diesen Trafficspitzen umgehen, ohne übermäßige Kosten oder Ausfälle zu riskieren?

Viele Unternehmen versuchen, Trafficschwankungen anhand historischer Daten oder geplanter Ereignisse vorherzusagen. Es gibt jedoch Situationen, in denen Lastspitzen kaum vorhersehbar sind. Ein plötzlicher Ansturm auf eine Website oder eine App kann innerhalb von Sekunden auftreten – oft ohne Vorwarnung. Diese Unvorhersehbarkeit macht es nahezu unmöglich, eine starre Serverinfrastruktur effizient zu betreiben.

Im Folgenden gehen wir näher auf die verschiedenen Ursachen für einen unerwarteten Anstieg des Traffics ein und erläutern, warum klassische IT-Architekturen damit oft nicht Schritt halten können.

Durch Social Media und Content-Sharing kann ein Produkt, ein Artikel oder eine Dienstleistung jederzeit viral gehen. Ein einziger Tweet, ein TikTok-Video oder ein YouTube-Clip kann dazu führen, dass Millionen von Menschen gleichzeitig auf eine Website zugreifen.

E-Commerce-Unternehmen veranstalten häufig kurzfristige Rabattaktionen oder Blitzverkäufe (Flash Sales), um die Nachfrage zu steigern. Besonders bekannt ist dieses Phänomen an Tagen wie Black Friday, Cyber Monday oder Singles' Day, an denen Kunden innerhalb weniger Stunden oder sogar Minuten einkaufen wollen. Nicht immer sind die Unternehmen selbst für Lastspitzen verantwortlich. Oft treten sie plötzlich und ohne Vorwarnung auf – ausgelöst durch externe Einflüsse, die nicht vorhersehbar sind.

Sobald die Anzahl der Anfragen das verfügbare Limit überschreitet, beginnen Webseiten langsam zu laden oder brechen im schlimmsten Fall ganz zusammen. Besonders fatal ist dies im E-Commerce, beim Ticketverkauf oder bei Streamingdiensten, wo jede Sekunde Verzögerung direkt Umsatz kostet. Kunden, die mit langen Ladezeiten oder Fehlermeldungen konfrontiert werden, springen ab und suchen sich Alternativen, ein Wettbewerbsnachteil, der das Unternehmen nachhaltig schädigen kann. Im Extremfall können solche Ausfälle zu Verlusten in Millionenhöhe führen, insbesondere wenn sie zu besonders umsatzstarken Zeiten wie dem Black Friday oder einer wichtigen Produkteinführung auftreten [1].

Um dieses Risiko zu vermeiden, wählen viele Unternehmen den gegenteiligen Ansatz: eine großzügige Überprovisionierung der IT-Ressourcen. Dabei werden dauerhaft zusätzliche Serverkapazitäten reserviert, um für Lastspitzen gewappnet zu sein. Diese Strategie ist jedoch teuer und ineffizient. Da ein Großteil der Ressourcen in ruhigeren Zeiten ungenutzt bleibt, entstehen hohe Fixkosten, ohne dass ein echter Mehrwert entsteht. Besonders problematisch ist das für Unternehmen mit stark schwankendem Verkehrsaufkommen, da sie für Kapazitäten bezahlen, die sie nur selten wirklich benötigen.

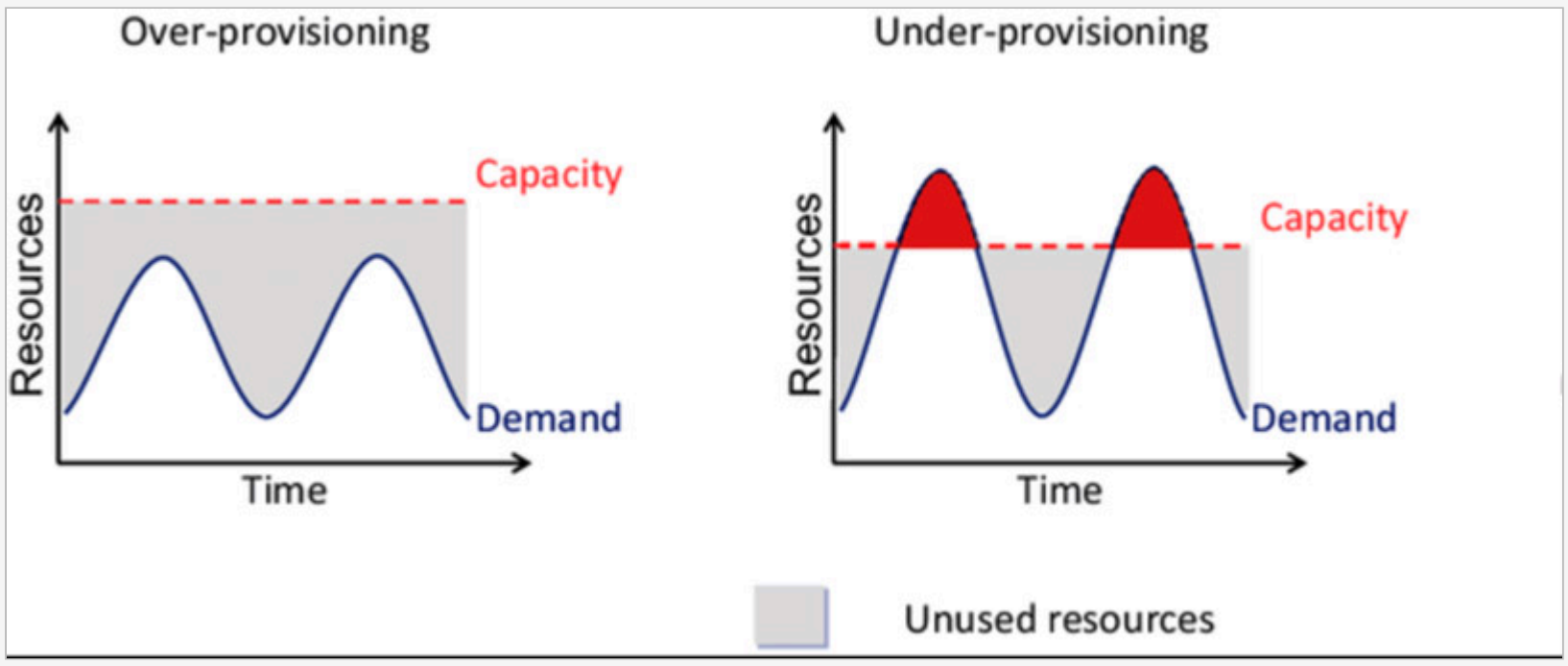

Abb. 1: Over- vs. Under-Provisioning [2]

Auto-Scaling ist ein Fortschritt gegenüber festen Serverkapazitäten, aber nicht die perfekte Lösung. Es reagiert nicht in Echtzeit – neue Instanzen benötigen oft Minuten, um hochzufahren, was bei plötzlichen Trafficspitzen zu Verzögerungen führt. Zudem setzen viele Cloud-Anbieter Obergrenzen für die Skalierung, wodurch extreme Lasten nicht immer abgefangen werden können. Die Einrichtung ist komplex, erfordert genaue Konfigurationen und kann zu unerwartet hohen Kosten führen, wenn sich der Traffic unkontrolliert erhöht.

Dieses Dilemma macht deutlich, dass traditionelle Infrastrukturmodelle den heutigen dynamischen Anforderungen nicht mehr gerecht werden. Unternehmen brauchen eine Lösung, die sich in Echtzeit an die tatsächlichen Nutzerzahlen anpasst – eine Architektur, die bei Bedarf automatisch skaliert und in Zeiten geringer Auslastung keine unnötigen Kosten verursacht. Genau hier setzen Serverless-Architekturen und Pay-as-you-go-Modelle an.

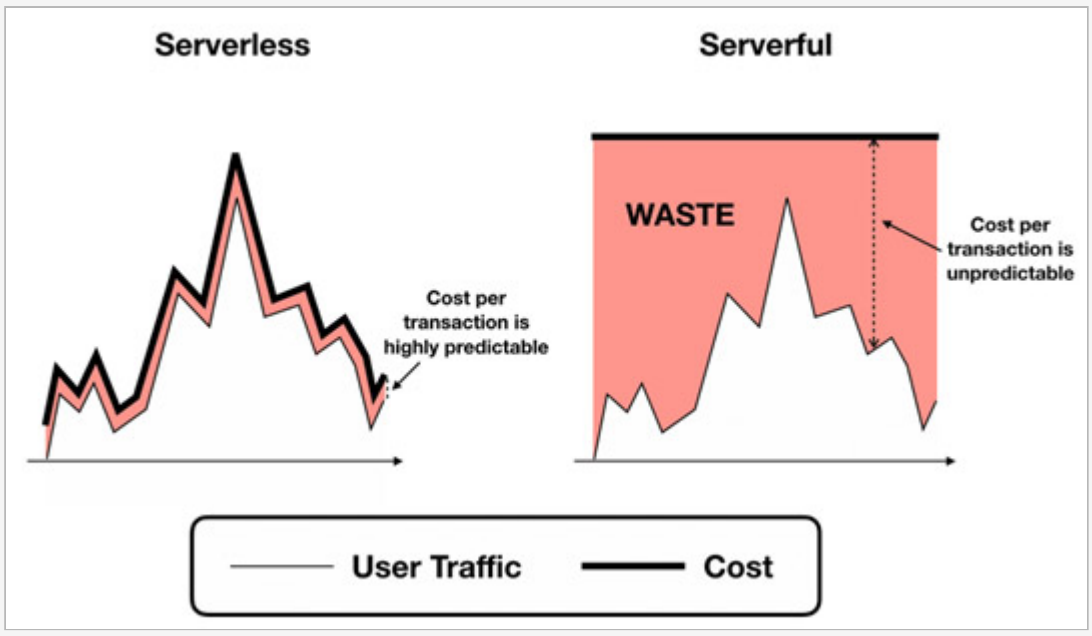

Abb. 2: Serverless (Pay as you go) vs. Serverful [3]

Was ist Serverless Computing?

Serverless Computing ist ein Cloud-Computing-Modell, bei dem die Verwaltung der Serverinfrastruktur vollständig vom Cloud-Anbieter übernommen wird. Im Gegensatz zu traditionellen Hostingmodellen müssen sich Entwickler nicht mehr um Bereitstellung, Wartung und Skalierung von Servern kümmern. Stattdessen können sie sich ausschließlich auf das Schreiben und Deployen von Code konzentrieren. Trotz des Namens sind Server in diesem Modell keineswegs abwesend, vielmehr werden sie dynamisch vom Anbieter bereitgestellt und verwaltet, basierend auf den tatsächlichen Anforderungen der Anwendung.

Ein zentrales Konzept innerhalb des Serverless Computing ist Function as a Service (FaaS). FaaS ermöglicht es Entwicklern, einzelne Funktionen als unabhängige Einheiten zu erstellen und bereitzustellen, die nur dann ausgeführt werden, wenn sie benötigt werden. Diese Funktionen sind in der Regel klein, spezialisiert und erfüllen eine spezifische Aufgabe innerhalb einer Anwendung. Beispielsweise könnte eine Funktion für das Verarbeiten von Zahlungen, eine andere für das Senden von E-Mail-Benachrichtigungen und eine weitere für das Verarbeiten von Bilddateien verantwortlich sein.

Der Hauptvorteil von FaaS liegt in seiner ereignisgesteuerten Natur. Das bedeutet, dass Funktionen nur als Reaktion auf bestimmte Ereignisse ausgelöst werden, wie etwa das Eintreffen einer HTTP-Anfrage, eine Änderung in einer Datenbank oder das Empfangen einer Nachricht in einer Warteschlange. Diese Flexibilität ermöglicht es Anwendungen, effizient und bedarfsorientiert zu arbeiten.

Ein weiteres wesentliches Merkmal von Serverless Computing ist die automatische Skalierung. Bei FaaS werden Funktionen automatisch skaliert, um der aktuellen Nachfrage gerecht zu werden. Wenn die Anzahl der Anfragen zunimmt, stellt der Cloud-Anbieter zusätzliche Instanzen der Funktion bereit, ohne dass eine manuelle Intervention erforderlich ist. So wird gewährleistet, dass die Anwendung stets optimal performt, unabhängig von den aktuellen Anforderungen.

Wie funktioniert Function as a Service?

FaaS basiert auf dem Prinzip, dass jede Funktion eine spezifische Aufgabe innerhalb einer Anwendung erfüllt. Diese Funktionen werden in der Regel in einer serverlosen Umgebung ausgeführt, die von einem Cloud-Anbieter bereitgestellt wird. Der Entwickler schreibt den Code für die Funktion und definiert, welche Ereignisse die Ausführung der Funktion auslösen sollen. Sobald ein definiertes Ereignis eintritt, wird die entsprechende Funktion automatisch gestartet.

Der gesamte Lebenszyklus der Funktion – von der Bereitstellung über die Ausführung bis hin zur Skalierung – wird vom Cloud-Anbieter verwaltet. Das bedeutet, dass Entwickler sich nicht mehr mit der Infrastruktur befassen müssen, sondern sich vollständig auf die Logik und Funktionalität ihrer Anwendungen konzentrieren können. Das führt zu einer erheblichen Reduzierung des operativen Aufwands und ermöglicht eine schnellere Entwicklung und Bereitstellung von Anwendungen.

Ein weiterer Vorteil von FaaS ist die Trennung der Verantwortlichkeiten. Da jede Funktion eine spezifische Aufgabe übernimmt, können Entwickler den Code modular gestalten und einzelne Funktionen unabhängig voneinander entwickeln, testen und deployen. So wird eine hohe Wartbarkeit gefördert und die Fehlerbehebung sowie die kontinuierliche Weiterentwicklung der Anwendung werden erleichtert.

Das Pay-as-you-go-Modell: Flexibilität und Kosteneffizienz vereint

Das Pay-as-you-go-Modell stellt einen Ansatz in der Cloud-Infrastruktur dar, der Unternehmen eine bislang unerreichte Flexibilität und Kosteneffizienz bietet. Im Kern basiert dieses Modell auf der Idee, dass Unternehmen nur für die tatsächlich genutzten Ressourcen bezahlen, anstatt im Voraus für feste Kapazitäten. Das steht im deutlichen Gegensatz zu traditionellen

Modellen, bei denen Unternehmen erhebliche Vorabinvestitionen tätigen müssen, um sicherzustellen, dass genügend Serverkapazitäten vorhanden sind, um auch unerwartete Trafficspitzen abzufangen.

Beim Pay-as-you-go-Modell wird auf Basis der tatsächlichen Nutzung von Rechenleistung, Speicherplatz, Datenübertragungen und anderen Ressourcen abgerechnet. Dadurch müssen Unternehmen keine festen Kosten für ungenutzte Ressourcen tragen. Stattdessen wird jede Ressource, die in Anspruch genommen wird, präzise erfasst und entsprechend berechnet. Diese Transparenz ermöglicht es Unternehmen, ihre Ausgaben genau zu kontrollieren und nur für die Dienste zu zahlen, die sie tatsächlich benötigen und nutzen.

Ein wesentlicher Vorteil des Pay-as-you-go-Modells liegt in seiner Flexibilität. Unternehmen können ihre Infrastruktur dynamisch an den aktuellen Bedarf anpassen, ohne langfristige Verpflichtungen eingehen zu müssen. Wenn der Traffic plötzlich ansteigt, können zusätzliche Ressourcen schnell und unkompliziert aktiviert werden. Sobald die Last abgeklungen ist, werden die nicht mehr benötigten Ressourcen automatisch wieder freigegeben, wodurch keine unnötigen Kosten entstehen.

Darüber hinaus fördert das Pay-as-you-go-Modell Kosteneffizienz, indem es Überprovisionierung vermeidet. In traditionellen Modellen müssen Unternehmen oft große Mengen Serverkapazität vorhalten, um auch selten auftretende Trafficspitzen abdecken zu können. Das führt zu hohen Fixkosten, die selbst bei geringer Auslastung unverändert bleiben. Im Gegensatz dazu ermöglicht das Pay-as-you-go-Modell eine skalierbare Kostenstruktur, bei der die Ausgaben direkt an die tatsächliche Nutzung gekoppelt sind. Unternehmen können ihre IT-Kosten exakt mit ihrem Geschäftsvolumen abstimmen, was insbesondere für Start-ups und wachstumsorientierte Unternehmen von großem Vorteil ist, da sie ihre Ausgaben flexibel anpassen können, ohne finanzielle Engpässe befürchten zu müssen.

Zudem unterstützt das Pay-as-you-go-Modell Innovation und Agilität. Entwickler können neue Features und Anwendungen schnell testen und bereitstellen, ohne sich Sorgen um die zugrunde liegende Infrastruktur machen zu müssen. Das verkürzt die Entwicklungszyklen und ermöglicht es Unternehmen, schneller auf Marktveränderungen und Kundenanforderungen zu reagieren. Die Möglichkeit, Ressourcen nach Bedarf zu nutzen, fördert experimentelle Projekte und Innovationen, da das finanzielle Risiko minimal und die Skalierung unkompliziert ist.

Ein praktisches Beispiel verdeutlicht die Vorteile des Pay-as-you-go-Modells: Ein Onlineshop plant eine Flash-Sale-Aktion. Mit einem traditionellen Modell müsste das Unternehmen im Vorfeld eine große Anzahl von Servern bereitstellen und bezahlen, unabhängig davon, ob diese während der Aktion voll ausgelastet sind oder nicht. Mit dem Pay-as-you-go-Modell hingegen kann der Shop die benötigten Ressourcen sofort skalieren, sobald die Trafficspitze eintritt, und diese wieder reduzieren, sobald die Aktion beendet ist. Dadurch werden nur die tatsächlich genutzten Ressourcen bezahlt, was die Kosten erheblich senkt und gleichzeitig eine reibungslose Kundenerfahrung sicherstellt.

Zusammenfassend lässt sich sagen, dass das Pay-as-you-go-Modell eine ideale Lösung für Unternehmen darstellt, die mit unvorhersehbaren Trafficspitzen konfrontiert sind. Es bietet die notwendige Flexibilität, um sich dynamisch an wechselnde Anforderungen anzupassen und gewährleistet gleichzeitig eine hohe Kosteneffizienz durch die Vermeidung von Überprovisionierung und die präzise Abrechnung nach Nutzung.

Durch die Integration von Serverless-Architekturen mit dem Pay-as-you-go-Modell können Unternehmen nicht nur ihre IT-Infrastruktur optimieren, sondern auch ihre Geschäftsprozesse agiler gestalten und sich Wettbewerbsvorteile sichern.

Beispiel LEGO-Onlineshop – Bausteine für digitalen Erfolg

Die LEGO Group, bekannt für ihre ikonischen Bausteine, stand vor der Herausforderung, ihre Onlinepräsenz zu modernisieren und den steigenden Anforderungen des digitalen Markts gerecht zu werden. Traditionell basierte die E-Commerce-Plattform von LEGO auf einer monolithischen Architektur, die in einem firmeneigenen Rechenzentrum betrieben wurde. Diese Struktur führte während hoher Lastzeiten zu erheblichen Performanceproblemen, einschließlich Serverüberlastungen und Dienstunterbrechungen.

Um diese Herausforderungen zu bewältigen, initiierte LEGO eine umfassende Transformation hin zu einer serverlosen Microservices-Architektur mit AWS Lambda, der Serverless-Lösung von Amazon Web Services (AWS). Der Übergang begann im September 2018 mit der Entkopplung des Backends und der schrittweisen Einführung von AWS-Lambda-Funktionen. Bis Juli 2019 wurde shop.LEGO.com vollständig auf eine serverlose Infrastruktur umgestellt.

In der neuen Architektur werden über 150 Lambda-Funktionen eingesetzt, die in mehr als 25 Microservices organisiert sind. Diese Microservices kommunizieren über 30 API-Gateway-Endpunkte und nutzen verschiedene AWS-Dienste wie Amazon DynamoDB, Amazon S3 und Amazon SQS. Ein typisches Anwendungsbeispiel ist das Hinzufügen eines Artikels zum Warenkorb, das als atomares Anfrage-Antwort-API implementiert ist. Hierbei validiert eine Lambda-Funktion die Anfrage und aktualisiert den Warenkorb in der Datenbank, wodurch eine schnelle und zuverlässige Benutzererfahrung gewährleistet wird.

Ein weiteres Beispiel ist die Verarbeitung von Bestellungen, die als ereignisgesteuerte Pipeline mit Pufferung realisiert wurde. Eingehende Bestellungen werden in eine Warteschlange gestellt und von Lambda-Funktionen asynchron verarbeitet, um eine skalierbare und fehlertolerante Abwicklung sicherzustellen. Zudem implementierte LEGO ein Hub-and-Spoke-Event-Bus-Muster unter Verwendung von Amazon EventBridge, um die Kommunikation zwischen verschiedenen Microservices zu koordinieren und die Systemarchitektur weiter zu entkoppeln.

Durch diese serverlose Transformation konnte LEGO die Skalierbarkeit und Flexibilität seiner E-Commerce-Plattform erheblich verbessern. Die automatische Anpassung der Ressourcen an die aktuelle Nachfrage ermöglicht es, Trafficspitzen effizient zu bewältigen und gleichzeitig die Betriebskosten zu optimieren. Zudem verkürzt sich die Entwicklungszeit für neue Funktionen, da Entwickler sich auf die Geschäftslogik konzentrieren können, während die Infrastruktur von AWS verwaltet wird.

Die serverlose Reise von LEGO zeigt, wie ein traditionelles Unternehmen durch die Einführung moderner Cloudtechnologien seine digitalen Angebote transformieren und den sich wandelnden Anforderungen des Markts gerecht werden kann.

Fazit

Serverless-Architekturen und das Pay-as-you-go-Modell stellen eine Lösung dar, um den Herausforderungen unvorhersehbarer Trafficspitzen und dynamischer Nutzeranforderungen zu begegnen, da sie es Unternehmen ermöglichen, ihre IT-Infrastruktur flexibel und kosteneffizient zu skalieren, ohne dabei in teure Überprovisionierung zu investieren. Durch die automatische Skalierung von Diensten wie Azure Functions wird sichergestellt, dass Ressourcen bedarfsgerecht zur Verfügung stehen, was nicht nur die Leistungsfähigkeit der Anwendungen verbessert, sondern auch die Entwicklungszyklen verkürzt, da sich Entwickler voll und ganz auf die Anwendungslogik konzentrieren können. Dennoch gilt es, mögliche Herausforderungen nicht außer Acht zu lassen: So können sogenannte Cold Starts zu kurzfristigen Verzögerungen führen, und die enge Bindung an einen bestimmten Cloud-Anbieter birgt das Risiko eines Vendor Lock-in, was im Fall eines Plattformwechsels zu Komplikationen führen kann.

Insgesamt bietet der Einsatz von Serverless-Technologien in Kombination mit einem Pay-as-you-go-Modell eine erhebliche Steigerung der Flexibilität und Kosteneffizienz, vorausgesetzt, dass Unternehmen die damit einhergehenden Herausforderungen – wie Latenzprobleme, Sicherheitsaspekte und Abhängigkeiten von einzelnen Anbietern – transparent adressieren und entsprechende Maßnahmen zur Risikominimierung ergreifen.

Links & Literatur

[1] https://www.upguard.com/blog/the-cost-of-downtime-at-the-worlds-biggest-online-retailer

[2] https://www.researchgate.net/profile/Lei-Wei-3/publication/283948945/figure/fig5/AS:668237306003460@1536331597239/The-cases-of-over-provisioning-under-provisioning-and-delay-caused-by-under-provisioning.png

[3] https://lumigo.io/wp-content/uploads/2023/07/1_70O8fdeXplzbok3Fi49g3g.jpg