Skalierung der Amazon EKS-Infrastruktur zur Optimierung von Rechenleistung, Workloads und Netzwerkleistung

AWS Präskriptive Leitlinien

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS Präskriptive Leitlinien: Skalierung der Amazon EKS-Infrastruktur zur Optimierung von Rechenleistung, Workloads und Netzwerkleistung

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Die Handelsmarken und Handelsaufmachung von Amazon dürfen nicht in einer Weise in Verbindung mit nicht von Amazon stammenden Produkten oder Services verwendet werden, durch die Kunden irregeführt werden könnten oder Amazon in schlechtem Licht dargestellt oder diskreditiert werden könnte. Alle anderen Handelsmarken, die nicht Eigentum von Amazon sind, gehören den jeweiligen Besitzern, die möglicherweise zu Amazon gehören oder nicht, mit Amazon verbunden sind oder von Amazon gesponsert werden.

Table of Contents

Einführung 1
Ziele 2
Skalierung berechnen 4
Cluster AutoScaler 4
Cluster-Autoscaler mit Over-Provisioning 5
Karpenter 5
Skalierung der Arbeitslast 7
Horizontal Pod Autoscaler 7
Proportionaler Cluster-Autoscaler 8
Kubernetes-basierter ereignisgesteuerter Autoscaler 9
Netzwerkskalierung 11
Amazon-VPC-CNI-Plug-In für Kubernetes 11
Benutzerdefinierte Netzwerke 12
Präfixdelegierung 13
Amazon VPC Lattice 14
Kostenoptimierung 16
Kubecost 16
Goldlöckchen 17
AWS Fargate 18
Spot Instances 19
Reserved Instances 19
AWS Graviton-Instanzen 20
Nächste Schritte 22
Ressourcen 23
Dokumentverlauf 24
Glossar 25
..... 25
A 26
B 29
C 31
D 35
E 39
F 41
G 43

H 44
I 46
L 49
M 50
O 54
P 57
Q 60
R 61
S 64
T 68
U 70
V 70
W 71
Z 72
..... lxxiii

Skalierung der Amazon EKS-Infrastruktur zur Optimierung von Rechenleistung, Workloads und Netzwerkleistung

Aniket Dekate, Aniket Kurzadkar und Ishwar Chauthaiwale, Amazon Web Services ()AWS

November 2024 (Geschichte der Dokumente)

Amazon Elastic Kubernetes Service (Amazon EKS) ist ein verwalteter Kubernetes-Service. Mit Amazon EKS können Sie Kubernetes-Pods in einer containerisierten Cloud-Umgebung ausführen, ohne Ihre eigene Steuerungsebene installieren und betreiben zu müssen. Durch die AWS Verwaltung der Kontrollebene reduziert Amazon EKS das organisatorische Betriebsmanagement. Zu den weiteren Vorteilen der Verwendung von Amazon EKS gehören Skalierung, Zuverlässigkeit und Sicherheit in der Cloud-Umgebung.

Dieser Leitfaden soll Unternehmen dabei helfen, ihre Amazon EKS-Infrastruktur in den folgenden Bereichen zu optimieren:

- Die Rechenskalierung ist eine wichtige Komponente für die Anwendungsleistung in einer dynamischen Kubernetes-Umgebung:
  - Effiziente Ressourcenzuweisung — Erfahren Sie mehr über Techniken zur dynamischen Zuweisung berechneter Ressourcen, um unterschiedlichen Anforderungen gerecht zu werden.
  - Automatisierungstools — Verschaffen Sie sich einen Überblick über Tools und Services, die die Rechenskalierung automatisieren und so den Bedarf an manuellen Eingriffen reduzieren.
- Durch die Workload-Skalierung wird sichergestellt, dass Anwendungen unterschiedliche Workloads ohne Leistungseinbußen bewältigen können:
  - Horizontaler Pod-Autoscaler — Sehen Sie sich eingehend an, wie ein HPA bei der Skalierung von Workloads auf der Grundlage von Echtzeitmetriken hilft.
  - Cluster Proportional Autoscaler — Erfahren Sie, wie CPA automatisch skaliert und ein proportionales Verhältnis zwischen Knoten und Replikaten beibehält und Workloads bei sich ändernder Clustergröße nach oben oder unten skaliert.
  - Ereignisgesteuerte Skalierung — Informieren Sie sich über Strategien zur Skalierung von Anwendungen als Reaktion auf bestimmte Ereignisse oder Auslöser.
- Die Netzwerkskalierung trägt dazu bei, eine reibungslose Kommunikation zwischen Diensten und einen effizienten Datenfluss in dynamischen Umgebungen aufrechtzuerhalten:

  - Amazon VPC CNI-Plugin — Erfahren Sie, wie das VPC CNI-Plugin skalierbare Netzwerke innerhalb von Amazon EKS-Clustern ermöglicht.
  - Benutzerdefiniertes Netzwerk — Überprüfen Sie die IP-Adressverwaltung und die Trennung des Netzwerkverkehrs auf Amazon EKS-Clustern.
  - Präfix-Delegierung — Verschaffen Sie sich einen Überblick über die Optimierung der IP-Verwaltung in großen und skalierbaren Amazon EKS-Clustern.
  - Amazon VPC Lattice — Verschaffen Sie sich einen Überblick darüber, wie VPC Lattice VPC- und Netzwerkübergreifend verwalten kann, um eine nahtlose Skalierung zu gewährleisten. service-to-service
- Mithilfe der Kostenoptimierung können Unternehmen erkennen, wofür ihre Ressourcen ausgegeben werden, und Ausgaben entsprechend Abteilungen oder Projekten zuordnen:
  - Richtige Dimensionierung von Ressourcen — Überlegen Sie, wie Sie Cloud-Ressourcen entsprechend der Arbeitslast dimensionieren können.
  - Kostenüberwachung und -kontrolle — Informieren Sie sich über Tools und bewährte Methoden zur Nachverfolgung und Optimierung von Cloud-Ausgaben.

Jeder Abschnitt konzentriert sich auf bestimmte Ziele, die für die Schaffung einer zuverlässigen, effektiven und erschwinglichen Cloud-Umgebung erforderlich sind.

Ziele

Dieser Leitfaden kann Ihnen und Ihrem Unternehmen dabei helfen, die folgenden Geschäftsziele zu erreichen:

- Verbesserte Ressourceneffizienz — Erzielen Sie eine optimale Ressourcennutzung, indem Sie Rechen-, Workloads und Netzwerkressourcen dynamisch auf der Grundlage von Echtzeitanforderungen skalieren.

  Dieses Ziel unterstreicht, wie wichtig es ist, Ressourcen entsprechend den tatsächlichen Nutzungsmustern nach oben und unten zu skalieren. Tools wie horizontale Pod-Autoscaler und das Amazon VPC CNI-Plugin helfen Unternehmen dabei, nur die Ressourcen zu nutzen, die sie benötigen, wodurch Verschwendung minimiert und die Leistung maximiert wird.

- Verbesserte Anwendungsleistung — Sorgen Sie für eine hohe Leistung und Reaktionsfähigkeit der Anwendungen, auch bei schwankenden Workloads und Datenverkehrsmustern.

Dieses Ziel konzentriert sich auf Strategien, mit denen sichergestellt werden soll, dass Anwendungen Spitzenverkehr und hohe Arbeitslasten bewältigen können, ohne die Leistung zu beeinträchtigen. Techniken wie ereignisgesteuerte Workload-Skalierung, effiziente Rechenzuweisung und skalierbare Netzwerkarchitekturen sind entscheidend, um dieses Ziel zu erreichen.

- Nahtlose Skalierbarkeit — Ermöglichen Sie eine reibungslose Skalierung von Infrastrukturkomponenten und ermöglichen so ein müheloses Wachstum und eine Anpassung an sich ändernde Geschäftsanforderungen.

  Eine nahtlose Skalierbarkeit ist für Unternehmen, die Wachstum erwarten oder unterschiedliche Datenverkehrszahlen verzeichnen, von entscheidender Bedeutung. Dieses Ziel trägt der Bedeutung der Implementierung skalierbarer Lösungen für Rechen-, Workload- und Netzwerkressourcen Rechnung, sodass die Skalierung automatisch, effizient und transparent erfolgen kann.

- Kostenoptimierung — Minimierung der Cloud-Kosten bei gleichbleibender oder verbesserter Leistung und Skalierbarkeit.

  Die Kostenoptimierung kann die Senkung der Ausgaben umfassen, z. B. durch die richtige Dimensionierung von Ressourcen, den Einsatz kostengünstiger Skalierungslösungen und die Überwachung der Ausgaben. Ziel ist es, Kosteneinsparungen mit dem Bedarf an hoher Leistung und Skalierbarkeit in Einklang zu bringen.

Skalierung berechnen

Die Skalierung von Rechenleistung ist eine wichtige Komponente für die Anwendungsleistung in einer dynamischen Kubernetes-Umgebung. Kubernetes reduziert Verschwendung durch die dynamische Anpassung der Rechenressourcen (wie CPU und Speicher) an die Nachfrage in Echtzeit. Diese Funktion trägt dazu bei, eine Über- oder Unterbereitstellung zu vermeiden, wodurch auch Betriebskosten eingespart werden können. Kubernetes macht manuelle Eingriffe effektiv überflüssig, da die Infrastruktur zu Spitzenzeiten automatisch hochskaliert und zu Nebenzeiten herunterskaliert werden kann.

Die allgemeine Rechenskalierung von Kubernetes automatisiert den Skalierungsprozess, wodurch die Flexibilität und Skalierbarkeit der Anwendung erhöht und ihr fehlertolerantes Verhalten verbessert wird. Letztlich verbessern die Funktionen von Kubernetes die betriebliche Exzellenz und Produktivität.

In diesem Abschnitt werden die folgenden Arten der Rechenskalierung beschrieben:

- Cluster-Autoscaler
- Cluster-Autoscaler mit Over-Provisioning
- Zimmerer

Cluster AutoScaler

Je nach den Anforderungen der Pods passt das Cluster Autoscaler-Tool die Größe automatisch an, indem es bei Bedarf Knoten hinzufügt oder Knoten entfernt, wenn sie nicht benötigt werden und nicht ausgelastet sind.

Betrachten Sie das Cluster Autoscaler-Tool als Skalierungslösung für Workloads, bei denen die Nachfrage schrittweise steigt und die Latenz bei der Skalierung kein großes Problem darstellt.

Das Cluster Autoscaler-Tool bietet die folgenden Hauptfunktionen:

- Skalierung — Skaliert Knoten dynamisch nach oben und unten als Reaktion auf den tatsächlichen Ressourcenbedarf.
- Pod-Planung — Hilft sicherzustellen, dass jeder Pod betriebsbereit ist und über die Ressourcen verfügt, die er benötigt, um zu funktionieren, und verhindert so Ressourcenknappheit.
- Kosteneffizienz — Eliminiert die unnötigen Kosten für den Betrieb nicht ausgelasteter Knoten, indem diese eliminiert werden.

Cluster-Autoscaler mit Over-Provisioning

Cluster-Autoscaler mit Over-Provisioning funktioniert ähnlich wie Cluster-Autoscaler, da er Knoten effizient bereitstellt und Zeit spart, indem Pods mit niedriger Priorität auf den Knoten ausgeführt werden. Bei dieser Technik wird der Datenverkehr als Reaktion auf plötzliche Bedarfsspitzen in diese Pods umgeleitet, sodass die Anwendung ohne Unterbrechung weiterarbeiten kann.

Cluster Autoscaler mit Over-Provisioning bietet die Funktionen von Dummy-Pods, mit denen Knoten einfach bereitgestellt und ausgeführt werden können, wenn die Arbeitslast sehr groß ist, keine Latenz erforderlich ist und die Skalierung schnell erfolgen muss.

Cluster Autoscaler mit Over-Provisioning bietet die folgenden Hauptfunktionen:

- Bessere Reaktionsfähigkeit — Da überschüssige Kapazität ständig verfügbar ist, nimmt die Skalierung des Clusters als Reaktion auf Nachfragespitzen weniger Zeit in Anspruch.
- Ressourcenreservierung — Der effektive Umgang mit unerwarteten Verkehrsspitzen unterstützt die korrekte Verwaltung mit nur geringen Ausfallzeiten.
- Reibungslose Skalierung — Die Minimierung von Verzögerungen bei der Ressourcenzuweisung ermöglicht einen reibungsloseren Skalierungsprozess.

Karpenter

Karpenter for Kubernetes übertrifft das herkömmliche Cluster Autoscaler-Tool in Bezug auf Open Source, Leistung und Anpassbarkeit. Mit Karpenter können Sie automatisch nur die benötigten Rechenressourcen starten, um die Anforderungen Ihres Clusters in Echtzeit zu erfüllen. Karpenter wurde entwickelt, um eine effizientere und reaktionsschnellere Skalierung zu ermöglichen.

Anwendungen mit extrem variablen oder komplexen Workloads, bei denen schnelle Skalierungsentscheidungen unerlässlich sind, profitieren stark von der Verwendung von Karpenter. Es lässt sich integrieren und bietet eine verbesserte Bereitstellung und Optimierung der Knotenauswahl. AWS

Karpenter umfasst die folgenden Hauptfunktionen:

- Dynamische Bereitstellung — Karpenter stellt die richtigen Instanzen und Größen für diesen Zweck bereit und stellt neue Knoten dynamisch auf der Grundlage der speziellen Anforderungen von Pods bereit.

- Erweitertes Scheduling — Mithilfe einer cleveren Pod-Platzierung ordnet Karpenter die Knoten so an, dass Ressourcen wie GPU, CPU, Arbeitsspeicher und Speicher so effektiv wie möglich genutzt werden.
- Schnelle Skalierung — Karpenter kann schnell skalieren und reagiert häufig innerhalb von Sekunden. Diese Reaktionsfähigkeit ist hilfreich bei plötzlichem Datenverkehr oder wenn die Arbeitslast eine sofortige Skalierung erfordert
- Kosteneffizienz — Durch die sorgfältige Auswahl der effektivsten Instance können Sie die Betriebskosten senken und zusätzliche kostensparende Alternativen wie On-Demand-Instances AWS, Spot-Instances und Reserved Instances nutzen.

Skalierung der Arbeitslast

Die Skalierung der Arbeitslast in Kubernetes ist für die Aufrechterhaltung der Anwendungsleistung und Ressourceneffizienz in dynamischen Umgebungen unerlässlich. Durch Skalierung wird sichergestellt, dass Anwendungen unterschiedliche Workloads ohne Leistungseinbußen bewältigen können. Kubernetes bietet die Möglichkeit, Ressourcen auf der Grundlage von Echtzeitmetriken automatisch nach oben oder unten zu skalieren, sodass Unternehmen schnell auf Änderungen im Datenverkehr reagieren können. Diese Elastizität verbessert nicht nur die Benutzererfahrung, sondern optimiert auch die Ressourcennutzung und trägt so dazu bei, die Kosten zu minimieren, die mit zu wenig genutzten oder übermäßig bereitgestellten Ressourcen verbunden sind.

Darüber hinaus unterstützt eine effektive Skalierung der Arbeitslast eine hohe Verfügbarkeit und stellt sicher, dass Anwendungen auch in Zeiten mit hoher Nachfrage reaktionsschnell bleiben. Die Workload-Skalierung in Kubernetes ermöglicht es Unternehmen, Cloud-Ressourcen besser zu nutzen, indem sie die Kapazität dynamisch an aktuelle Bedürfnisse anpassen.

In diesem Abschnitt werden die folgenden Arten der Workload-Skalierung beschrieben:

- Horizontaler Pod-Autoscaler
- Proportionaler Cluster-Autoskalierer
- Kubernetes-basierter ereignisgesteuerter Autoscaler

Horizontal Pod Autoscaler

Der Horizontal Pod Autoscaler (HPA) ist eine Kubernetes-Funktion, die die Anzahl der Pod-Replikate in einer Bereitstellung, einem Replikationscontroller oder einem Stateful-Set automatisch auf der Grundlage der beobachteten CPU-Auslastung oder anderer ausgewählter Metriken anpasst. Der HPA stellt sicher, dass Anwendungen schwankende Datenverkehrs- und Arbeitslasten bewältigen können, ohne dass manuelles Eingreifen erforderlich ist. Die HPA bietet die Möglichkeit, eine optimale Leistung aufrechtzuerhalten und gleichzeitig die verfügbaren Ressourcen effektiv zu nutzen.

In Kontexten, in denen die Nutzernachfrage im Laufe der Zeit stark schwanken kann, wie z. B. bei Web-Apps, Microservices usw. APIs, ist HPA besonders hilfreich.

Der Horizontal Pod Autoscaler bietet die folgenden Hauptfunktionen:

- Automatische Skalierung — HPA erhöht oder verringert automatisch die Anzahl der Pod-Replikate als Reaktion auf Echtzeitmetriken und stellt so sicher, dass Anwendungen skaliert werden können, um den Benutzeranforderungen gerecht zu werden.
- Metrikbasierte Entscheidungen — Standardmäßig skaliert HPA auf der Grundlage der CPU-Auslastung. Es können jedoch auch benutzerdefinierte Messwerte wie die Speichernutzung oder anwendungsspezifische Metriken verwendet werden, was maßgeschneiderte Skalierungsstrategien ermöglicht.
- Konfigurierbare Parameter — Sie können die minimale und maximale Anzahl von Replikaten sowie die gewünschten Nutzungsprozentsätze wählen, sodass Sie selbst bestimmen können, wie stark die Skalierung sein soll.
- Integration mit Kubernetes — Um Ressourcen zu überwachen und zu ändern, arbeitet HPA mit anderen Elementen des Kubernetes-Ökosystems zusammen, darunter dem Metrics Server, der Kubernetes-API und benutzerdefinierten Metrikadaptern.
- Bessere Ressourcennutzung — HPA hilft dabei, sicherzustellen, dass Ressourcen effektiv genutzt werden, wodurch Kosten gesenkt und die Leistung verbessert werden, indem die Anzahl der Pods dynamisch geändert wird.

Proportionaler Cluster-Autoscaler

Der Cluster Proportional Autoscaler (CPA) ist eine Kubernetes-Komponente, mit der die Anzahl der Pod-Replikate in einem Cluster automatisch an die Anzahl der verfügbaren Knoten angepasst wird. Im Gegensatz zu herkömmlichen Autoscalern, die auf der Grundlage von Kennzahlen zur Ressourcennutzung (wie CPU und Arbeitsspeicher) skalieren, skaliert CPA Workloads proportional zur Größe des Clusters selbst.

Dieser Ansatz ist besonders nützlich für Anwendungen, die ein gewisses Maß an Redundanz oder Verfügbarkeit im Verhältnis zur Clustergröße aufrechterhalten müssen, wie CoreDNS und andere Infrastrukturdienste. Zu den wichtigsten Anwendungsfällen für CPA gehören die folgenden:

- Übermäßige Bereitstellung
- Skalieren Sie die wichtigsten Plattformdienste
- Skalieren Sie Workloads, da CPA keinen Metrikserver oder Prometheus-Adapter benötigt

Durch die Automatisierung des Skalierungsprozesses unterstützt CPA Unternehmen dabei, eine ausgewogene Verteilung der Arbeitslast aufrechtzuerhalten, die Ressourceneffizienz zu

steigern und sicherzustellen, dass Anwendungen entsprechend bereitgestellt werden, um die Benutzeranforderungen zu erfüllen.

Der Cluster Proportional Autoscaler bietet die folgenden Hauptfunktionen:

- Knotenbasierte Skalierung — CPA skaliert Replikate entsprechend der Anzahl der Clusterknoten, die geplant werden können, sodass Anwendungen proportional zur Größe des Clusters erweitert oder verkleinert werden können.
- Proportionale Anpassung — Um sicherzustellen, dass die Anwendung entsprechend den Änderungen der Clustergröße skaliert werden kann, stellt der Autoscaler ein proportionales Verhältnis zwischen der Anzahl der Knoten und der Anzahl der Replikate her. Diese Beziehung wird verwendet, um die gewünschte Anzahl von Replikaten für eine Arbeitslast zu berechnen.
- Integration mit Kubernetes-Komponenten — CPA funktioniert mit Standard-Kubernetes-Komponenten wie dem Horizontal Pod Autoscaler (HPA), konzentriert sich jedoch speziell auf die Anzahl der Knoten und nicht auf die Kennzahlen zur Ressourcennutzung. Diese Integration ermöglicht eine umfassendere Skalierungsstrategie.
- Golang-API-Clients — Um die Anzahl der Knoten und ihre verfügbaren Kerne zu überwachen, verwendet CPA Golang-API-Clients, die innerhalb von Pods ausgeführt werden und mit dem Kubernetes-API-Server kommunizieren.
- Konfigurierbare Parameter — Mithilfe von können Benutzer Schwellenwerte und Skalierungsparameter festlegen`ConfigMap`, anhand derer CPA sein Verhalten ändert, und sicherstellen, dass der beabsichtigte Skalierungsplan eingehalten wird.

Kubernetes-basierter ereignisgesteuerter Autoscaler

Der auf Kubernetes basierende Event Driven Autoscaler (KEDA) ist ein Open-Source-Projekt, mit dem Kubernetes-Workloads auf der Grundlage der Anzahl der zu verarbeitenden Ereignisse skaliert werden können. KEDA verbessert die Skalierbarkeit von Anwendungen, indem es ihnen ermöglicht, dynamisch auf unterschiedliche Workloads zu reagieren, insbesondere auf solche, die ereignisgesteuert sind.

Durch die Automatisierung des Skalierungsprozesses auf der Grundlage von Ereignissen unterstützt KEDA Unternehmen dabei, die Ressourcennutzung zu optimieren, die Anwendungsleistung zu verbessern und die Kosten im Zusammenhang mit übermäßiger Bereitstellung zu senken. Dieser Ansatz ist besonders nützlich für Anwendungen mit unterschiedlichen Datenverkehrsmustern, wie z. B. Microservices, serverlose Funktionen und Echtzeit-Datenverarbeitungssysteme.

KEDA bietet die folgenden Hauptfunktionen:

- Ereignisgesteuerte Skalierung — Mit KEDA können Sie Skalierungsregeln auf der Grundlage externer Ereignisquellen wie Nachrichtenwarteschlangen, HTTP-Anfragen oder benutzerdefinierten Metriken definieren. Diese Funktion trägt dazu bei, dass Anwendungen entsprechend der Nachfrage in Echtzeit skaliert werden.
- Leichte Komponente — KEDA ist eine einfache Komponente mit einem einzigen Zweck, für deren einfache Integration in bestehende Kubernetes-Cluster weder viel Einrichtung noch Mehraufwand erforderlich ist.
- Integration mit Kubernetes — KEDA erweitert die Funktionen von Kubernetes-nativen Komponenten wie dem Horizontal Pod Autoscaler (HPA). KEDA erweitert diese Komponenten um ereignisgesteuerte Skalierungsfunktionen und verbessert sie, anstatt sie zu ersetzen.
- Support für mehrere Ereignisquellen — KEDA ist mit einer Vielzahl von Ereignisquellen kompatibel, darunter beliebte Messaging-Plattformen wie RabbitMQ, Apache Kafka und andere. Aufgrund dieser Anpassungsfähigkeit können Sie die Skalierung an Ihre einzigartige ereignisgesteuerte Architektur anpassen.
- Benutzerdefinierte Skalierer — Mithilfe von benutzerdefinierten Skalierern können Sie spezifische Metriken festlegen, anhand derer KEDA Skalierungsaktionen als Reaktion auf bestimmte Geschäftslogik oder Anforderungen einleiten kann.
- Deklarative Konfiguration — Gemäß den Kubernetes-Prinzipien können Sie KEDA verwenden, um das Skalierungsverhalten deklarativ zu beschreiben, indem Sie benutzerdefinierte Kubernetes-Ressourcen verwenden, um zu definieren, wie die Skalierung erfolgen soll.

Netzwerkskalierung

Die Netzwerkskalierung in Kubernetes ist entscheidend für die Aufrechterhaltung einer reibungslosen Kommunikation zwischen Diensten und die Unterstützung eines effizienten Datenflusses in dynamischen Umgebungen. Durch die Skalierung der Netzwerkinfrastruktur wird sichergestellt, dass der Cluster unterschiedliche Datenverkehrsmengen bewältigen kann, ohne dass Engpässe oder Latenzprobleme auftreten. Kubernetes bietet Tools und Mechanismen zur Skalierung von Netzwerkressourcen, sodass Unternehmen bei sich ändernden Verkehrsmustern eine optimale Leistung aufrechterhalten können.

Diese Flexibilität bei der Netzwerkskalierung verbessert das allgemeine Benutzererlebnis, indem schnelle und zuverlässige Verbindungen gewährleistet werden. Die Netzwerkskalierung optimiert auch die Nutzung von Netzwerkressourcen und trägt so dazu bei, die Kosten zu senken, die mit nicht ausgelasteten oder überlasteten Netzwerkkomponenten verbunden sind.

Darüber hinaus ist eine effektive Netzwerkskalierung für die Unterstützung von Hochverfügbarkeit und Ausfallsicherheit von entscheidender Bedeutung. Durch die dynamische Anpassung der Netzwerkkapazität und des Routing können Unternehmen sicherstellen, dass Dienste auch in Zeiten hoher Nachfrage oder unerwarteter Verkehrsspitzen zugänglich und reaktionsschnell bleiben. Dieser Ansatz ermöglicht eine bessere Nutzung der Cloud-Netzwerkressourcen und stellt sicher, dass die Infrastruktur stets den aktuellen Anforderungen entspricht.

In diesem Abschnitt werden die folgenden Arten der Netzwerkskalierung beschrieben:

- Amazon VPC CNI-Plugin für Kubernetes
- Benutzerdefiniertes Netzwerk
- Präfix-Delegierung
- Amazon VPC Lattice

Amazon-VPC-CNI-Plug-In für Kubernetes

Das Amazon VPC Container Network Interface (CNI) -Plugin für Kubernetes ist eine wichtige Komponente in Amazon EKS. Das VPC CNI-Plugin bietet erweiterte Netzwerkfunktionen durch die Integration von Kubernetes-Pods in Amazon VPC. Mit diesem Plugin wird jedem Pod eine eindeutige IP-Adresse aus der Virtual Private Cloud (VPC) zugewiesen, wodurch die Netzwerkisolierung und Leistung verbessert wird. Da Cluster wachsen und die Netzwerkanforderungen schwanken, spielt das

Amazon VPC CNI-Plugin eine wichtige Rolle bei der Sicherstellung eines effizienten und skalierbaren Netzwerkbetriebs.

Das Plugin verwaltet automatisch die Zuweisung und das Routing von IP-Adressen innerhalb der VPC, vereinfacht so die Netzwerkverwaltung und reduziert das Risiko von IP-Konflikten. Es unterstützt Funktionen wie die Präfix-Delegierung, was ein flexibleres IP-Management ermöglicht.

Das VPC CNI Plugin hilft Unternehmen dabei, die Netzwerkleistung zu optimieren, die Sicherheit zu erhöhen und das Risiko einer IP-Erschöpfung zu verringern. Diese Funktionen sind besonders wertvoll für große, dynamische Umgebungen, in denen die Netzwerkanforderungen schwanken, wie z. B. Microservices-Architekturen, Workloads mit hoher Dichte und Mehrmandantenanwendungen.

Das Amazon VPC CNI-Plugin bietet die folgenden Hauptfunktionen:

- Verbessertes Netzwerk — Das VPC-CNI-Plugin ermöglicht es jedem Pod, seine eigene IP-Adresse direkt von der VPC zu erhalten, was für eine starke Isolierung und Netzwerkleistung sorgt. Dieser Ansatz ist entscheidend für Workloads, die einen hohen Netzwerkdurchsatz und eine geringe Latenz erfordern.
- Präfix-Delegierung — Um Probleme mit der Erschöpfung von IP-Adressen in großen Clustern zu vermeiden, weist die Präfix-Delegierung dynamisch größere Blöcke von Knoten IPs zu, die dann für die Pod-Nutzung unterteilt werden. Dieser Ansatz gewährleistet eine effiziente IP-Nutzung und vereinfacht die Netzwerkskalierung.
- Benutzerdefiniertes Netzwerk — Benutzer können benutzerdefinierte Netzwerkschnittstellen (ENIs) für Pods konfigurieren, wodurch der Pod-Verkehr auf mehrere Schnittstellen verteilt wird, wodurch Netzwerküberlastungen reduziert und die Skalierbarkeit verbessert werden.
- Support für IPv6 — Durch die Aktivierung IPv6 in Amazon EKS-Clustern können Benutzer den verfügbaren IP-Adressraum erheblich erweitern und so die Skalierung großer, verteilter Anwendungen ohne IPv4 Einschränkungen erleichtern.
- Integration mit Kubernetes — Das VPC CNI-Plugin arbeitet nahtlos mit Kubernetes-Netzwerkkomponenten zusammen und stellt sicher, dass diese effizient über Pods, Dienste und externe Endpunkte hinweg verwaltet IPs werden, und es unterstützt erweiterte Funktionen wie Sicherheitsgruppen für Pods.

Benutzerdefinierte Netzwerke

Benutzerdefinierte Netzwerke in Amazon EKS ermöglichen die Zuweisung bestimmter Netzwerkschnittstellen zu Pods und bieten so eine verbesserte Kontrolle über die IP-

Adressverwaltung und den Netzwerkverkehr. Dieser Ansatz ist besonders nützlich in Szenarien, in denen die Erschöpfung von IP-Adressen ein Problem darstellt oder wenn der Netzwerkverkehr aus Sicherheits-, Compliance- oder Leistungsgründen getrennt werden muss. Benutzerdefinierte Netzwerke helfen Unternehmen dabei, den IP-Adressraum effizient zu verwalten, den Datenverkehr zu trennen und eine skalierbare Netzwerkleistung sicherzustellen.

Mit benutzerdefinierten Netzwerken können Administratoren Netzwerkressourcen effizienter verwalten. Administratoren können benutzerdefinierte Netzwerke verwenden, um sicherzustellen, dass Pods über die erforderliche Netzwerkisolierung verfügen und dass der Cluster skaliert werden kann, ohne auf IP-Adressbeschränkungen zu stoßen.

Benutzerdefiniertes Netzwerk bietet die folgenden Hauptfunktionen:

- Verbessertes IP-Management — Benutzerdefiniertes Netzwerk ermöglicht die Zuweisung bestimmter Netzwerkschnittstellen (ENIs) zu Pods und trägt so dazu bei, die Erschöpfung der IP-Adressen zu bewältigen, indem der Pod-Verkehr auf mehrere ENIs Pods verteilt wird. Diese Funktion ist besonders wichtig in Clustern mit Workloads mit hoher Dichte.
- Trennung des Datenverkehrs — Mit benutzerdefinierten Netzwerkschnittstellen können Sie den Pod-Verkehr nach bestimmten Kriterien wie Anwendungstyp oder Sicherheitsanforderungen trennen. Dieser Ansatz bietet eine bessere Kontrolle darüber, wie der Datenverkehr innerhalb und außerhalb des Clusters fließt.
- Support für IPv6 — Benutzerdefinierte Netzwerke in Amazon EKS werden ebenfalls unterstützt IPv6 und bieten eine Lösung für die Einschränkungen von IPv4 Adressen. Das Netzwerk kann effizient und ohne IP-Adresskonflikte skaliert werden, selbst bei großen Bereitstellungen.
- Skalierbarkeit und Flexibilität — Mit der Skalierung des Clusters ermöglichen benutzerdefinierte Netzwerke die dynamische Verwaltung von Netzwerkschnittstellen. Neuen Pods werden ohne manuelles Eingreifen die entsprechenden Netzwerkressourcen zugewiesen. Dieser Ansatz trägt zur Aufrechterhaltung einer flexiblen und skalierbaren Netzwerkumgebung bei, die sich an wechselnde Workloads anpassen kann.

Präfixdelegierung

Die Präfix-Delegierung in Kubernetes, insbesondere innerhalb von Amazon EKS, wurde entwickelt, um die IP-Adressverwaltung zu rationalisieren und zu optimieren, wenn Cluster skalieren. Durch die dynamische Zuweisung größerer Blöcke von IP-Adressen (Präfixe) zu Knoten reduziert die Präfix-Delegierung das Risiko einer IP-Erschöpfung und vereinfacht die Verwaltung des IP-Bereichs.

Dieser Ansatz verbessert die Netzwerkeffizienz, minimiert die Fragmentierung und hilft Clustern, reibungslos zu skalieren, ohne dass der IP-Bereich manuell angepasst werden muss. Die Präfix-Delegierung ist besonders nützlich für umfangreiche Bereitstellungen, Workloads mit hoher Dichte und Umgebungen, in denen flexibles, dynamisches IP-Management für die Aufrechterhaltung der Netzwerkleistung und Skalierbarkeit entscheidend ist.

Die Präfix-Delegierung bietet die folgenden Hauptfunktionen:

- Effiziente IP-Adressverwaltung — Die Präfix-Delegierung ermöglicht die dynamische Zuweisung von IP-Bereichen, wodurch das Risiko einer IP-Erschöpfung verringert und eine effiziente Nutzung des verfügbaren IP-Speicherplatzes gewährleistet wird.
- Vereinfachtes Netzwerkmanagement — Da die Knoten ihre eigenen IP-Zuweisungen verwalten können, minimiert die Präfix-Delegierung die Netzwerkfragmentierung und vereinfacht den Routing-Prozess, sodass Cluster leichter nach Bedarf skaliert werden können.
- Support für umfangreiche Bereitstellungen — In großen Clustern mit Workloads mit hoher Dichte ermöglicht die Präfix-Delegierung eine nahtlose Skalierung, da neue Knoten dem Cluster beitreten können, ohne dass der IP-Bereich manuell angepasst werden muss.

Amazon VPC Lattice

Amazon VPC Lattice ermöglicht eine effiziente und sichere service-to-service Kommunikation innerhalb und zwischen ihnen VPCs, insbesondere in Microservices-Architekturen. VPC Lattice verwendet zusätzlich zur (IAM) -Integration Sicherheitsmaßnahmen wie Sicherheitsgruppen und Netzwerkzugriffskontrolllisten AWS Identity and Access Management (Netzwerk ACLs) für eine differenzierte Anwendungsauthentifizierung. Ein Layer-7-Proxy-Service als Herzstück von VPC Lattice bietet Verbindung, Lastenausgleich, Authentifizierung, Autorisierung, Beobachtbarkeit, Verkehrsmanagement und Serviceerkennung.

Durch die Vereinfachung von Netzwerk- und Sicherheitskonfigurationen hilft VPC Lattice Unternehmen dabei, das Datenverkehrsmanagement zu optimieren, die Anwendungsleistung zu verbessern und nahtlos über mehrere Netzwerke hinweg zu skalieren. VPCs AWS-Regionen Dies ist besonders nützlich für verteilte Anwendungen, die konsistente und zuverlässige Netzwerke erfordern, wie z. B. Mikroservices, regionsübergreifende Bereitstellungen und komplexe Cloud-native Umgebungen.

Amazon VPC Lattice bietet die folgenden Hauptfunktionen:

- Service-to-service Netzwerk — VPC Lattice vereinfacht die Netzwerk- und Sicherheitskonfiguration zwischen Diensten innerhalb einer Microservices-Architektur. Es bietet eine einheitliche Plattform für das Kommunikationsmanagement, sodass Dienste unabhängig voneinander skaliert werden können und gleichzeitig hohe Leistung und Sicherheit gewährleistet werden.
- VPC-übergreifende Netzwerke — VPC Lattice ist entscheidend für die Verwaltung des Datenverkehrs über mehrere oder Regionen hinweg. VPCs Es bietet ein konsistentes Netzwerkframework, das es Diensten ermöglicht, unabhängig von ihrem physischen Standort nahtlos zu kommunizieren. Diese Funktion ist besonders wichtig für umfangreiche Anwendungen, die sich über mehrere VPCs oder geografische Regionen erstrecken.
- Verbessertes Sicherheitsmanagement — Durch die direkte Integration von Sicherheitsrichtlinien in die Netzwerkebene unterstützt VPC Lattice eine sichere und effiziente service-to-service Kommunikation. Diese Funktion reduziert die Komplexität der Sicherheitsverwaltung in einer verteilten Umgebung und ermöglicht so eine einfachere Skalierung und einen geringeren Betriebsaufwand.
- Vereinfachtes Verkehrsmanagement — VPC Lattice bietet erweiterte Funktionen für das Verkehrsmanagement, darunter Routing, Lastenausgleich und Failover-Mechanismen. Mit diesen Funktionen wird der Datenverkehr effizient auf die Dienste verteilt, wodurch die Netzwerkleistung optimiert und die Skalierbarkeit der Anwendung verbessert wird.

Kostenoptimierung

Um eine effektive Ressourcenkontrolle zu unterstützen, ist die Kostenminimierung von Kubernetes für Unternehmen, die diese Container-Orchestrierungstechnologie verwenden, von entscheidender Bedeutung. Aufgrund ihrer Komplexität, die mehrere Komponenten wie Pods und Nodes umfasst, ist es schwierig, die Ausgaben in Kubernetes-Umgebungen richtig nachzuverfolgen. Durch die Anwendung von Techniken zur Kostenoptimierung können Unternehmen erkennen, wofür ihre Ressourcen ausgegeben werden, und die Ausgaben entsprechend Abteilungen oder Projekten zuordnen.

Dynamische Skalierung hat zwar Vorteile, kann aber zu unvorhergesehenen Ausgaben führen, wenn sie nicht richtig verwaltet wird. Ein effizientes Kostenmanagement hilft dabei, Ressourcen nur dann zuzuweisen, wenn sie wirklich benötigt werden, wodurch unerwartete Ausgabensteigerungen vermieden werden.

In diesem Abschnitt werden die folgenden Ansätze zur Kostenoptimierung erörtert:

- Kubecost
- Goldlöckchen
- AWS Fargate
- Spot Instances
- Reserved Instances
- AWS Graviton-Instanzen

Kubecost

Kubecost ist eine Kostenmanagementlösung, mit der Unternehmen ihre Ausgaben für die Cloud-Infrastruktur verfolgen, kontrollieren und maximieren können. Sie wurde speziell für Kubernetes-Cluster entwickelt. Kubecost bietet Ihnen Einblicke in die Ressourcennutzung und das Kostenbewusstsein in Echtzeit, sodass Sie besser verstehen können, wo und wie viel Ihrer Cloud-Ressourcen genutzt werden. Mit diesen Erkenntnissen können Sie Ihre Infrastrukturausgaben optimieren, die Ressourceneffizienz verbessern und fundiertere Entscheidungen über Ihre Cloud-Investitionen treffen.

Kubecost bietet die folgenden Hauptfunktionen:

- Kostenzuweisung — Kubecost bietet eine gründliche Kostenzuweisung für Kubernetes-Ressourcen, einschließlich Workloads, Services, Namespaces und Labels. Diese Funktion hilft Teams dabei, die Kosten nach Umgebung, Projekt oder Team zu überwachen.
- Kostenüberwachung in Echtzeit — Sie bietet eine Echtzeitüberwachung der Cloud-Kosten, gibt Unternehmen sofortige Einblicke in die Ausgabenmuster und hilft, unerwartete Kostenüberschreitungen zu vermeiden.
- Optimierungsempfehlungen — Kubecost bietet praktische Vorschläge zur Minimierung der Ressourcennutzung, einschließlich der Reduzierung ungenutzter Ressourcen, der richtigen Dimensionierung von Workloads und der Maximierung der Speicherkosten.
- Budgetierung und Benachrichtigungen — Kubecost-Benutzer können Budgets erstellen und Erinnerungen erhalten, wenn sich eine Ausgabe vordefinierten Kriterien nähert oder diese übertrifft. Diese Funktion hilft Teams dabei, finanzielle Einschränkungen einzuhalten.

Goldlöckchen

Goldilocks ist ein Kubernetes-Hilfsprogramm, das Benutzern helfen soll, ihre Ressourcenanfragen und Grenzwerte für Kubernetes-Workloads zu optimieren. Es enthält Empfehlungen zur Konfiguration der CPU- und Speicherressourcen für Container, die in einem Kubernetes-Cluster ausgeführt werden. Mithilfe dieser Empfehlungen können Sie sicherstellen, dass Anwendungen über die richtige Anzahl an Ressourcen verfügen, um effizient und ohne Verschwendung zu arbeiten. Diese Optimierung kann zu Kosteneinsparungen, verbesserter Leistung und einer effizienteren Nutzung von Kubernetes-Clustern führen.

Goldilocks bietet die folgenden Hauptfunktionen:

- Ressourcenempfehlungen — Goldilocks ermittelt die idealen Einstellungen für Ressourcenanfragen und Einschränkungen, indem es frühere CPU- und Speicherverbrauchsstatistiken für Kubernetes-Workloads analysiert. Auf diese Weise wird es einfacher, Unter- oder Überprovisionierung zu vermeiden, was zu Leistungsproblemen und Ressourcenverschwendung führen kann.
- VPA-Integration — Goldilocks nutzt den Kubernetes Vertical Pod Autoscaler (VPA), um Daten zu sammeln und Empfehlungen abzugeben. Es läuft in einem „Empfehlungsmodus“, was bedeutet, dass es die Ressourceneinstellungen nicht wirklich ändert, sondern eine Anleitung dazu bietet, wie diese Einstellungen aussehen sollten.

- Namespace-basierte Analyse — Goldlöckchen gibt Ihnen die Möglichkeit, genau zu regulieren, welche Workloads optimiert und überwacht werden, indem Sie bestimmte Namespaces für die Analyse gezielt auswählen können.
- Visuelles Dashboard — Das webbasierte Dashboard zeigt vorgeschlagene Ressourcenanfragen und Einschränkungen visuell an, sodass Sie die Daten leicht verstehen und entsprechend handeln können.
- Störungsfreier Betrieb — Goldlöckchen verändert das Setup des Clusters nicht, da es im Empfehlungsmodus arbeitet. Wenn Sie möchten, können Sie die empfohlenen Ressourceneinstellungen manuell anwenden, nachdem Sie die Empfehlungen gelesen haben.

AWS Fargate

Im Kontext von Amazon EKS https://docs.aws.amazon.com/eks/latest/userguide/fargate.htmlAWS Fargate können Sie Kubernetes-Pods ausführen, ohne die zugrunde liegenden EC2 Amazon-Instances verwalten zu müssen. Es handelt sich um eine serverlose Compute-Engine, mit der Sie sich auf die Bereitstellung und Skalierung von containerisierten Anwendungen konzentrieren können, ohne sich Gedanken über die Infrastruktur machen zu müssen.

AWS Fargate bietet die folgenden Hauptfunktionen:

- Kein Infrastrukturmanagement — Fargate macht die Bereitstellung, Verwaltung oder Skalierung von EC2 Amazon-Instances oder Kubernetes-Knoten überflüssig. AWS kümmert sich um das gesamte Infrastrukturmanagement, einschließlich Patching und Skalierung.
- Isolierung auf Pod-Ebene — Im Gegensatz zu Worker-Knoten, die auf Amazon basieren EC2, bietet Fargate eine Isolierung auf Aufgaben- oder Pod-Ebene. Jeder Pod wird in seiner eigenen isolierten Computerumgebung ausgeführt, was die Sicherheit und Leistung verbessert.
- Automatische Skalierung — Fargate skaliert Kubernetes-Pods automatisch je nach Bedarf. Sie müssen keine Skalierungsrichtlinien oder Knotenpools verwalten.
- Abrechnung pro Sekunde — Sie zahlen nur für die vCPU- und Speicherressourcen, die von jedem Pod genau für die Dauer seiner Ausführung verbraucht werden. Dies ist eine kostengünstige Option für bestimmte Workloads.
- Geringerer Overhead — Durch den Wegfall der Verwaltung von EC2 Instanzen können Sie sich mit Fargate auf die Erstellung und Verwaltung Ihrer Anwendungen konzentrieren, anstatt sich auf den Betrieb der Infrastruktur zu konzentrieren.

Spot Instances

Spot-Instances bieten erhebliche Einsparungen gegenüber den Preisen für On-Demand-Instances und sind eine erschwingliche Option für den Betrieb von EC2 Amazon-Worker-Knoten in einem Amazon EKS-Cluster. Spot-Instances AWS können jedoch unterbrochen werden, falls On-Demand-Instance-Kapazität benötigt wird. AWS kann Spot-Instances innerhalb von 2 Minuten zurückfordern, wenn die Kapazität benötigt wird, wodurch sie für kritische, statusbehaftete Workloads weniger zuverlässig sind.

Für kostenempfindliche Workloads, die Störungen aushalten können, sind Spot-Instances in Amazon EKS eine gute Option. Durch die Verwendung einer Kombination aus Spot-Instances und On-Demand-Instances in einem Kubernetes-Cluster können Sie Geld sparen, ohne die Verfügbarkeit wichtiger Workloads zu beeinträchtigen.

Spot Instances bietet die folgenden Hauptfunktionen:

- Kosteneinsparungen — Spot-Instances können günstiger sein als die Preise für On-Demand-Instances und eignen sich daher ideal für kostensensible Workloads.
- Ideal für fehlertolerante Workloads — Gut geeignet für statuslose, fehlertolerante Workloads wie Batch-Verarbeitung, CI/CD-Jobs, maschinelles Lernen oder umfangreiche Datenverarbeitung, bei der Instanzen ohne größere Unterbrechungen ersetzt werden können.
- Auto-Scaling-Gruppenintegration — Amazon EKS integriert Spot-Instances mit Kubernetes Cluster Autoscaler, der unterbrochene Spot-Instance-Knoten automatisch durch andere verfügbare Spot-Instances oder On-Demand-Instances ersetzen kann.

Reserved Instances

In Amazon EKS sind Reserved Instances ein Preismodell für die EC2 Amazon-Worker-Knoten, auf denen Ihre Kubernetes-Workloads ausgeführt werden. Durch die Nutzung von Reserved Instances verpflichten Sie sich, bestimmte Instance-Typen für eine Laufzeit von 1 oder 3 Jahren zu verwenden. Im Gegenzug erhalten Sie Kosteneinsparungen im Vergleich zu den Preisen für On-Demand-Instances. Das Reservieren von Instances in Amazon EKS ist eine kostengünstige Möglichkeit, konsistente, langfristige Workloads auf EC2 Amazon-Worker-Knoten auszuführen.

Reserved Instances werden häufig für Amazon verwendet EC2. Allerdings können auch die Worker-Knoten in Ihrem Amazon EKS-Cluster (bei denen es sich um EC2 Instances handelt) von diesem

kostensparenden Modell profitieren, vorausgesetzt, die Arbeitslast erfordert eine langfristige, vorhersehbare Nutzung.

Produktionsservices, Datenbanken und andere statusbehaftete Anwendungen, die hohe Verfügbarkeit und konsistente Leistung benötigen, sind Beispiele für stabile Workloads, die sich gut für Reserved Instances eignen.

Reserved Instances bietet die folgenden Hauptfunktionen:

- Kosteneinsparungen — Reserved Instances bieten Einsparungen im Vergleich zu On-Demand-Instances, abhängig von der Laufzeit (1 oder 3 Jahre) und dem Zahlungsplan (All Upfront, Partial Upfront oder No Prefront).
- Langfristiges Engagement — Sie verpflichten sich zu einer Laufzeit von 1 oder 3 Jahren für einen bestimmten Instance-Typ, eine bestimmte Größe und. AWS-Region Dies ist ideal für Workloads, die stabil sind und im Laufe der Zeit kontinuierlich ausgeführt werden.
- Vorhersehbare Preise — Da Sie an eine bestimmte Laufzeit gebunden sind, bieten Reserved Instances vorhersehbare monatliche oder Vorabkosten, sodass Sie leichter für langfristige Workloads budgetieren können.
- Instance-Flexibilität — Mit Convertible Reserved Instances können Sie den Instance-Typ, die Instance-Familie oder die Größe während des Reservierungszeitraums ändern. Convertible Reserved Instances bieten mehr Flexibilität als Standard Reserved Instances, die keine Änderungen zulassen.
- Garantierte Kapazität — Reserved Instances stellen sicher, dass Kapazität in der Availability Zone verfügbar ist, in der die Reservierung vorgenommen wurde. Dies ist entscheidend für kritische Workloads, die eine konstante Rechenleistung benötigen.
- Kein Unterbrechungsrisiko — Im Gegensatz zu Spot-Instances unterliegen Reserved Instances keiner Unterbrechung durch AWS. Dadurch eignen sie sich ideal für die Ausführung geschäftskritischer Workloads, für die eine garantierte Verfügbarkeit erforderlich ist.

AWS Graviton-Instanzen

AWS Graviton ist eine Familie von ARM-basierten Prozessoren, die entwickelt wurden, AWS um eine verbesserte Leistung und Kosteneffizienz für Cloud-Workloads zu bieten. Im Kontext von Amazon EKS können Sie Graviton-Instances als Worker-Knoten für die Ausführung Ihrer Kubernetes-Workloads verwenden, was zu erheblichen Leistungssteigerungen und Kosteneinsparungen führt.

Graviton-Instances sind eine hervorragende Option für cloudnative und rechenintensive Anwendungen, da sie ein besseres Preis-Leistungs-Verhältnis als x86-Instances bieten. Wenn Sie jedoch die Einführung von Graviton-Instances in Betracht ziehen, sollten Sie die ARM-Kompatibilität berücksichtigen.

AWS Graviton-Instanzen bieten die folgenden Hauptfunktionen:

- ARM-basierte Architektur — AWS Graviton-Prozessoren basieren auf der ARM-Architektur, die sich von herkömmlichen x86-Architekturen unterscheidet, aber für viele Workloads hocheffizient ist.
- Kosteneffizient — EC2 Amazon-Instances, die auf Graviton basieren, bieten im Vergleich zu x86-basierten Instances in der Regel ein besseres Preis-Leistungs-Verhältnis. EC2 Dies macht sie zu einer attraktiven Option für Kubernetes-Cluster, auf denen Amazon EKS ausgeführt wird.
- Leistung — Graviton2-Prozessoren, die zweite Generation von AWS Graviton, bieten erhebliche Verbesserungen in Bezug auf Rechenleistung, Speicherdurchsatz und Energieeffizienz. Sie eignen sich ideal für rechenintensive und speicherintensive Workloads.
- Verschiedene Instance-Typen — Graviton-Instances gibt es in verschiedenen Familien, z. B. t4g, m7g, c7g und r7g, die eine Reihe von Anwendungsfällen abdecken, von allgemeinen bis hin zu rechenoptimierten, speicheroptimierten und burstfähigen Workloads.
- Amazon EKS-Knotengruppen — Sie können Knotengruppen, die von Amazon EKS verwaltet werden, oder selbstverwaltete Knotengruppen so konfigurieren, dass sie Graviton-basierte Instances enthalten. Mit diesem Ansatz können Sie Workloads, die für die ARM-Architektur optimiert sind, auf demselben Kubernetes-Cluster zusammen mit x86-basierten Instances ausführen.

Nächste Schritte

Dieses Handbuch enthält Informationen, die Sie bei der Optimierung von Amazon EKS in Bezug auf Rechenskalierung, Workload-Skalierung, Netzwerkskalierung und Kostenoptimierung unterstützen. Durch das Verständnis und die Anwendung dieser Konzepte können Unternehmen eine hocheffiziente, skalierbare und kostengünstige Cloud-Umgebung einrichten, die ihren dynamischen Anforderungen entspricht.

Eine effektive Implementierung von Rechen- und Workload-Skalierung trägt dazu bei, dass Ressourcen effizient genutzt werden und die Anwendungen auch in Spitzenzeiten eine hohe Leistung beibehalten. Der Einsatz von Techniken zur Netzwerkskalierung, wie z. B. benutzerdefinierte Netzwerke und Präfix-Delegierung, unterstützt die Verwaltung von Netzwerkressourcen und ermöglicht eine nahtlose Skalierbarkeit. Die Betonung der Kostenoptimierung hilft Unternehmen dabei, Leistung und finanzielle Effizienz in Einklang zu bringen.

Die Integration dieser Leitlinien in Ihre Cloud-Strategie kann Ihnen helfen, die Leistung und Skalierbarkeit Ihrer Infrastruktur zu verbessern und Kosteneinsparungen zu erzielen. Mit diesem umfassenden Ansatz können Sie eine robuste Cloud-Umgebung aufbauen, die das Wachstum Ihres Unternehmens unterstützt und sich an die sich ständig ändernden Geschäftsanforderungen anpasst.

Ressourcen

AWS Blogs

- Wir setzen auf Kostenoptimierung und Ausfallsicherheit für EKS mit Spot-Instances
- Kombination von AWS Graviton mit x86 CPUs zur Optimierung von Kosten und Stabilität mithilfe von Amazon EKS

AWS Dokumentation

- Amazon VPC CNI
- Amazon Elastic Kubernetes Service (AWS Whitepaper: Überblick über die Bereitstellungsoptionen auf) AWS
- Leitfaden zu bewährten Methoden für Amazon EKS
- Zimmerer
- Erfahre mehr über Kubecost
- Vereinfachen Sie die Rechenverwaltung mit AWS Fargate

Sonstige Ressourcen

- Cluster-Autoscaling (Kubernetes-Dokumentation)
- Goldilocks: Ein Open-Source-Tool zur Empfehlung von Ressourcenanfragen (Fairwinds Blog)
- Horizontale automatische Pod-Skalierung (Kubernetes-Dokumentation)
- Kubecost (Kubecost-Dokumentation)
- Kubernetes Ereignisgesteuertes Autoscaling (KEDA-Dokumentation)

Dokumentverlauf

In der folgenden Tabelle werden wichtige Änderungen an diesem Handbuch, Skalierung der Amazon EKS-Infrastruktur zur Optimierung von Rechenleistung, Workloads und Netzwerkleistung, beschrieben. Um Benachrichtigungen über zukünftige Aktualisierungen zu erhalten, können Sie einen RSS-Feed abonnieren.

Änderung	Beschreibung	Datum
Erste Veröffentlichung	—	11. November 2024

AWS Glossar zu präskriptiven Leitlinien

Die folgenden Begriffe werden häufig in Strategien, Leitfäden und Mustern von AWS Prescriptive Guidance verwendet. Um Einträge vorzuschlagen, verwenden Sie bitte den Link Feedback geben am Ende des Glossars.

Zahlen

7 Rs

Sieben gängige Migrationsstrategien für die Verlagerung von Anwendungen in die Cloud. Diese Strategien bauen auf den 5 Rs auf, die Gartner 2011 identifiziert hat, und bestehen aus folgenden Elementen:

- Faktorwechsel/Architekturwechsel – Verschieben Sie eine Anwendung und ändern Sie ihre Architektur, indem Sie alle Vorteile cloudnativer Feature nutzen, um Agilität, Leistung und Skalierbarkeit zu verbessern. Dies beinhaltet in der Regel die Portierung des Betriebssystems und der Datenbank. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank auf die Amazon Aurora PostgreSQL-kompatible Edition.
- Plattformwechsel (Lift and Reshape) – Verschieben Sie eine Anwendung in die Cloud und führen Sie ein gewisses Maß an Optimierung ein, um die Cloud-Funktionen zu nutzen. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank zu Amazon Relational Database Service (Amazon RDS) für Oracle in der. AWS Cloud
- Neukauf (Drop and Shop) – Wechseln Sie zu einem anderen Produkt, indem Sie typischerweise von einer herkömmlichen Lizenz zu einem SaaS-Modell wechseln. Beispiel: Migrieren Sie Ihr CRM-System (Customer Relationship Management) zu Salesforce.com.
- Hostwechsel (Lift and Shift) – Verschieben Sie eine Anwendung in die Cloud, ohne Änderungen vorzunehmen, um die Cloud-Funktionen zu nutzen. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank zu Oracle auf einer EC2 Instanz in der. AWS Cloud
- Verschieben (Lift and Shift auf Hypervisor-Ebene) – Verlagern Sie die Infrastruktur in die Cloud, ohne neue Hardware kaufen, Anwendungen umschreiben oder Ihre bestehenden Abläufe ändern zu müssen. Sie migrieren Server von einer lokalen Plattform zu einem Cloud-Dienst für dieselbe Plattform. Beispiel: Migrieren Sie eine Microsoft Hyper-V Anwendung zu AWS.
- Beibehaltung (Wiederaufgreifen) – Bewahren Sie Anwendungen in Ihrer Quellumgebung auf. Dazu können Anwendungen gehören, die einen umfangreichen Faktorwechsel erfordern und

die Sie auf einen späteren Zeitpunkt verschieben möchten, sowie ältere Anwendungen, die Sie beibehalten möchten, da es keine geschäftliche Rechtfertigung für ihre Migration gibt.

- Außerbetriebnahme – Dekommissionierung oder Entfernung von Anwendungen, die in Ihrer Quellumgebung nicht mehr benötigt werden.

A

ABAC

Siehe attributbasierte Zugriffskontrolle.

abstrahierte Dienste

Weitere Informationen finden Sie unter Managed Services.

ACID

Siehe Atomarität, Konsistenz, Isolierung und Haltbarkeit.

Aktiv-Aktiv-Migration

Eine Datenbankmigrationsmethode, bei der die Quell- und Zieldatenbanken synchron gehalten werden (mithilfe eines bidirektionalen Replikationstools oder dualer Schreibvorgänge) und beide Datenbanken Transaktionen von miteinander verbundenen Anwendungen während der Migration verarbeiten. Diese Methode unterstützt die Migration in kleinen, kontrollierten Batches, anstatt einen einmaligen Cutover zu erfordern. Es ist flexibler, erfordert aber mehr Arbeit als eine aktiv-passive Migration.

Aktiv-Passiv-Migration

Eine Datenbankmigrationsmethode, bei der die Quell- und Zieldatenbanken synchron gehalten werden, aber nur die Quelldatenbank Transaktionen von verbindenden Anwendungen verarbeitet, während Daten in die Zieldatenbank repliziert werden. Die Zieldatenbank akzeptiert während der Migration keine Transaktionen.

Aggregatfunktion

Eine SQL-Funktion, die mit einer Gruppe von Zeilen arbeitet und einen einzelnen Rückgabewert für die Gruppe berechnet. Beispiele für Aggregatfunktionen sind `SUM` und`MAX`.

AI

Siehe künstliche Intelligenz.

AIOps

Siehe Operationen im Bereich künstliche Intelligenz.

Anonymisierung

Der Prozess des dauerhaften Löschens personenbezogener Daten in einem Datensatz. Anonymisierung kann zum Schutz der Privatsphäre beitragen. Anonymisierte Daten gelten nicht mehr als personenbezogene Daten.

Anti-Muster

Eine häufig verwendete Lösung für ein wiederkehrendes Problem, bei dem die Lösung kontraproduktiv, ineffektiv oder weniger wirksam als eine Alternative ist.

Anwendungssteuerung

Ein Sicherheitsansatz, bei dem nur zugelassene Anwendungen verwendet werden können, um ein System vor Schadsoftware zu schützen.

Anwendungsportfolio

Eine Sammlung detaillierter Informationen zu jeder Anwendung, die von einer Organisation verwendet wird, einschließlich der Kosten für die Erstellung und Wartung der Anwendung und ihres Geschäftswerts. Diese Informationen sind entscheidend für den Prozess der Portfoliofindung und -analyse und hilft bei der Identifizierung und Priorisierung der Anwendungen, die migriert, modernisiert und optimiert werden sollen.

künstliche Intelligenz (KI)

Das Gebiet der Datenverarbeitungswissenschaft, das sich der Nutzung von Computertechnologien zur Ausführung kognitiver Funktionen widmet, die typischerweise mit Menschen in Verbindung gebracht werden, wie Lernen, Problemlösen und Erkennen von Mustern. Weitere Informationen finden Sie unter Was ist künstliche Intelligenz?

Operationen mit künstlicher Intelligenz (AIOps)

Der Prozess des Einsatzes von Techniken des Machine Learning zur Lösung betrieblicher Probleme, zur Reduzierung betrieblicher Zwischenfälle und menschlicher Eingriffe sowie zur Steigerung der Servicequalität. Weitere Informationen zur Verwendung in der AWS Migrationsstrategie finden Sie im Operations Integration Guide. AIOps

Asymmetrische Verschlüsselung

Ein Verschlüsselungsalgorithmus, der ein Schlüsselpaar, einen öffentlichen Schlüssel für die Verschlüsselung und einen privaten Schlüssel für die Entschlüsselung verwendet. Sie können den öffentlichen Schlüssel teilen, da er nicht für die Entschlüsselung verwendet wird. Der Zugriff auf den privaten Schlüssel sollte jedoch stark eingeschränkt sein.

Atomizität, Konsistenz, Isolierung, Haltbarkeit (ACID)

Eine Reihe von Softwareeigenschaften, die die Datenvalidität und betriebliche Zuverlässigkeit einer Datenbank auch bei Fehlern, Stromausfällen oder anderen Problemen gewährleisten.

Attributbasierte Zugriffskontrolle (ABAC)

Die Praxis, detaillierte Berechtigungen auf der Grundlage von Benutzerattributen wie Abteilung, Aufgabenrolle und Teamname zu erstellen. Weitere Informationen finden Sie unter ABAC AWS in der AWS Identity and Access Management (IAM-) Dokumentation.

autoritative Datenquelle

Ein Ort, an dem Sie die primäre Version der Daten speichern, die als die zuverlässigste Informationsquelle angesehen wird. Sie können Daten aus der maßgeblichen Datenquelle an andere Speicherorte kopieren, um die Daten zu verarbeiten oder zu ändern, z. B. zu anonymisieren, zu redigieren oder zu pseudonymisieren.

Availability Zone

Ein bestimmter Standort innerhalb einer AWS-Region , der vor Ausfällen in anderen Availability Zones geschützt ist und kostengünstige Netzwerkkonnektivität mit niedriger Latenz zu anderen Availability Zones in derselben Region bietet.

AWS Framework für die Cloud-Einführung (AWS CAF)

Ein Framework mit Richtlinien und bewährten Verfahren, das Unternehmen bei der Entwicklung eines effizienten und effektiven Plans für den erfolgreichen Umstieg auf die Cloud unterstützt. AWS AWS CAF unterteilt die Leitlinien in sechs Schwerpunktbereiche, die als Perspektiven bezeichnet werden: Unternehmen, Mitarbeiter, Unternehmensführung, Plattform, Sicherheit und Betrieb. Die Perspektiven Geschäft, Mitarbeiter und Unternehmensführung konzentrieren sich auf Geschäftskompetenzen und -prozesse, während sich die Perspektiven Plattform, Sicherheit und Betriebsabläufe auf technische Fähigkeiten und Prozesse konzentrieren. Die Personalperspektive zielt beispielsweise auf Stakeholder ab, die sich mit Personalwesen (HR), Personalfunktionen und Personalmanagement befassen. Aus dieser Perspektive bietet AWS CAF Leitlinien für Personalentwicklung, Schulung und Kommunikation, um das Unternehmen auf eine erfolgreiche

Cloud-Einführung vorzubereiten. Weitere Informationen finden Sie auf der AWS -CAF-Webseite und dem AWS -CAF-Whitepaper.

AWS Workload-Qualifizierungsrahmen (AWS WQF)

Ein Tool, das Workloads bei der Datenbankmigration bewertet, Migrationsstrategien empfiehlt und Arbeitsschätzungen bereitstellt. AWS WQF ist in () enthalten. AWS Schema Conversion Tool AWS SCT Es analysiert Datenbankschemas und Codeobjekte, Anwendungscode, Abhängigkeiten und Leistungsmerkmale und stellt Bewertungsberichte bereit.

B

schlechter Bot

Ein Bot, der Einzelpersonen oder Organisationen stören oder ihnen Schaden zufügen soll.

BCP

Siehe Planung der Geschäftskontinuität.

Verhaltensdiagramm

Eine einheitliche, interaktive Ansicht des Ressourcenverhaltens und der Interaktionen im Laufe der Zeit. Sie können ein Verhaltensdiagramm mit Amazon Detective verwenden, um fehlgeschlagene Anmeldeversuche, verdächtige API-Aufrufe und ähnliche Vorgänge zu untersuchen. Weitere Informationen finden Sie unter Daten in einem Verhaltensdiagramm in der Detective-Dokumentation.

Big-Endian-System

Ein System, welches das höchstwertige Byte zuerst speichert. Siehe auch Endianness.

Binäre Klassifikation

Ein Prozess, der ein binäres Ergebnis vorhersagt (eine von zwei möglichen Klassen). Beispielsweise könnte Ihr ML-Modell möglicherweise Probleme wie „Handelt es sich bei dieser E-Mail um Spam oder nicht?“ vorhersagen müssen oder „Ist dieses Produkt ein Buch oder ein Auto?“

Bloom-Filter

Eine probabilistische, speichereffiziente Datenstruktur, mit der getestet wird, ob ein Element Teil einer Menge ist.

Blau/Grün-Bereitstellung

Eine Bereitstellungsstrategie, bei der Sie zwei separate, aber identische Umgebungen erstellen. Sie führen die aktuelle Anwendungsversion in einer Umgebung (blau) und die neue Anwendungsversion in der anderen Umgebung (grün) aus. Mit dieser Strategie können Sie schnell und mit minimalen Auswirkungen ein Rollback durchführen.

Bot

Eine Softwareanwendung, die automatisierte Aufgaben über das Internet ausführt und menschliche Aktivitäten oder Interaktionen simuliert. Manche Bots sind nützlich oder nützlich, wie z. B. Webcrawler, die Informationen im Internet indexieren. Einige andere Bots, sogenannte bösartige Bots, sollen Einzelpersonen oder Organisationen stören oder ihnen Schaden zufügen.

Botnetz

Netzwerke von Bots, die mit Malware infiziert sind und unter der Kontrolle einer einzigen Partei stehen, die als Bot-Herder oder Bot-Operator bezeichnet wird. Botnetze sind der bekannteste Mechanismus zur Skalierung von Bots und ihrer Wirkung.

branch

Ein containerisierter Bereich eines Code-Repositorys. Der erste Zweig, der in einem Repository erstellt wurde, ist der Hauptzweig. Sie können einen neuen Zweig aus einem vorhandenen Zweig erstellen und dann Feature entwickeln oder Fehler in dem neuen Zweig beheben. Ein Zweig, den Sie erstellen, um ein Feature zu erstellen, wird allgemein als Feature-Zweig bezeichnet. Wenn das Feature zur Veröffentlichung bereit ist, führen Sie den Feature-Zweig wieder mit dem Hauptzweig zusammen. Weitere Informationen finden Sie unter Über Branches (GitHub Dokumentation).

Zugang durch Glasbruch

Unter außergewöhnlichen Umständen und im Rahmen eines genehmigten Verfahrens ist dies eine schnelle Methode für einen Benutzer, auf einen Bereich zuzugreifen AWS-Konto , für den er in der Regel keine Zugriffsrechte besitzt. Weitere Informationen finden Sie unter dem Indikator Implementation break-glass procedures in den AWS Well-Architected-Leitlinien.

Brownfield-Strategie

Die bestehende Infrastruktur in Ihrer Umgebung. Wenn Sie eine Brownfield-Strategie für eine Systemarchitektur anwenden, richten Sie sich bei der Gestaltung der Architektur nach den Einschränkungen der aktuellen Systeme und Infrastruktur. Wenn Sie die bestehende Infrastruktur erweitern, könnten Sie Brownfield- und Greenfield-Strategien mischen.

Puffer-Cache

Der Speicherbereich, in dem die am häufigsten abgerufenen Daten gespeichert werden.

Geschäftsfähigkeit

Was ein Unternehmen tut, um Wert zu generieren (z. B. Vertrieb, Kundenservice oder Marketing). Microservices-Architekturen und Entwicklungsentscheidungen können von den Geschäftskapazitäten beeinflusst werden. Weitere Informationen finden Sie im Abschnitt Organisiert nach Geschäftskapazitäten des Whitepapers Ausführen von containerisierten Microservices in AWS.

Planung der Geschäftskontinuität (BCP)

Ein Plan, der die potenziellen Auswirkungen eines störenden Ereignisses, wie z. B. einer groß angelegten Migration, auf den Betrieb berücksichtigt und es einem Unternehmen ermöglicht, den Betrieb schnell wieder aufzunehmen.

C

CAF

Weitere Informationen finden Sie unter Framework für die AWS Cloud-Einführung.

Bereitstellung auf Kanaren

Die langsame und schrittweise Veröffentlichung einer Version für Endbenutzer. Wenn Sie sich sicher sind, stellen Sie die neue Version bereit und ersetzen die aktuelle Version vollständig.

CCoE

Weitere Informationen finden Sie im Cloud Center of Excellence.

CDC

Siehe Erfassung von Änderungsdaten.

Erfassung von Datenänderungen (CDC)

Der Prozess der Nachverfolgung von Änderungen an einer Datenquelle, z. B. einer Datenbanktabelle, und der Aufzeichnung von Metadaten zu der Änderung. Sie können CDC für verschiedene Zwecke verwenden, z. B. für die Prüfung oder Replikation von Änderungen in einem Zielsystem, um die Synchronisation aufrechtzuerhalten.

Chaos-Technik

Absichtliches Einführen von Ausfällen oder Störungsereignissen, um die Widerstandsfähigkeit eines Systems zu testen. Sie können AWS Fault Injection Service (AWS FIS) verwenden, um Experimente durchzuführen, die Ihre AWS Workloads stress, und deren Reaktion zu bewerten.

CI/CD

Siehe Continuous Integration und Continuous Delivery.

Klassifizierung

Ein Kategorisierungsprozess, der bei der Erstellung von Vorhersagen hilft. ML-Modelle für Klassifikationsprobleme sagen einen diskreten Wert voraus. Diskrete Werte unterscheiden sich immer voneinander. Beispielsweise muss ein Modell möglicherweise auswerten, ob auf einem Bild ein Auto zu sehen ist oder nicht.

clientseitige Verschlüsselung

Lokale Verschlüsselung von Daten, bevor das Ziel sie AWS-Service empfängt.

Cloud-Exzellenzzentrum (CCoE)

Ein multidisziplinäres Team, das die Cloud-Einführung in der gesamten Organisation vorantreibt, einschließlich der Entwicklung bewährter Cloud-Methoden, der Mobilisierung von Ressourcen, der Festlegung von Migrationszeitplänen und der Begleitung der Organisation durch groß angelegte Transformationen. Weitere Informationen finden Sie in den CCoE-Beiträgen im AWS Cloud Enterprise Strategy Blog.

Cloud Computing

Die Cloud-Technologie, die typischerweise für die Ferndatenspeicherung und das IoT-Gerätemanagement verwendet wird. Cloud Computing ist häufig mit Edge-Computing-Technologie verbunden.

Cloud-Betriebsmodell

In einer IT-Organisation das Betriebsmodell, das zum Aufbau, zur Weiterentwicklung und Optimierung einer oder mehrerer Cloud-Umgebungen verwendet wird. Weitere Informationen finden Sie unter Aufbau Ihres Cloud-Betriebsmodells.

Phasen der Einführung der Cloud

Die vier Phasen, die Unternehmen bei der Migration in der Regel durchlaufen AWS Cloud:

- Projekt – Durchführung einiger Cloud-bezogener Projekte zu Machbarkeitsnachweisen und zu Lernzwecken
- Fundament — Tätigen Sie grundlegende Investitionen, um Ihre Cloud-Einführung zu skalieren (z. B. Einrichtung einer landing zone, Definition eines CCo E, Einrichtung eines Betriebsmodells)
- Migration – Migrieren einzelner Anwendungen
- Neuentwicklung – Optimierung von Produkten und Services und Innovation in der Cloud

Diese Phasen wurden von Stephen Orban im Blogbeitrag The Journey Toward Cloud-First & the Stages of Adoption im AWS Cloud Enterprise Strategy-Blog definiert. Informationen darüber, wie sie mit der AWS Migrationsstrategie zusammenhängen, finden Sie im Leitfaden zur Vorbereitung der Migration.

CMDB

Siehe Datenbank für das Konfigurationsmanagement.

Code-Repository

Ein Ort, an dem Quellcode und andere Komponenten wie Dokumentation, Beispiele und Skripts gespeichert und im Rahmen von Versionskontrollprozessen aktualisiert werden. Zu den gängigen Cloud-Repositorys gehören GitHub oderBitbucket Cloud. Jede Version des Codes wird Zweig genannt. In einer Microservice-Struktur ist jedes Repository einer einzelnen Funktionalität gewidmet. Eine einzelne CI/CD-Pipeline kann mehrere Repositorien verwenden.

Kalter Cache

Ein Puffer-Cache, der leer oder nicht gut gefüllt ist oder veraltete oder irrelevante Daten enthält. Dies beeinträchtigt die Leistung, da die Datenbank-Instance aus dem Hauptspeicher oder der Festplatte lesen muss, was langsamer ist als das Lesen aus dem Puffercache.

Kalte Daten

Daten, auf die selten zugegriffen wird und die in der Regel historisch sind. Bei der Abfrage dieser Art von Daten sind langsame Abfragen in der Regel akzeptabel. Durch die Verlagerung dieser Daten auf leistungsschwächere und kostengünstigere Speicherstufen oder -klassen können Kosten gesenkt werden.

Computer Vision (CV)

Ein Bereich der KI, der maschinelles Lernen nutzt, um Informationen aus visuellen Formaten wie digitalen Bildern und Videos zu analysieren und zu extrahieren. Amazon SageMaker AI bietet beispielsweise Bildverarbeitungsalgorithmen für CV.

Drift in der Konfiguration

Bei einer Arbeitslast eine Änderung der Konfiguration gegenüber dem erwarteten Zustand. Dies kann dazu führen, dass der Workload nicht mehr richtlinienkonform wird, und zwar in der Regel schrittweise und unbeabsichtigt.

Verwaltung der Datenbankkonfiguration (CMDB)

Ein Repository, das Informationen über eine Datenbank und ihre IT-Umgebung speichert und verwaltet, inklusive Hardware- und Softwarekomponenten und deren Konfigurationen. In der Regel verwenden Sie Daten aus einer CMDB in der Phase der Portfolioerkennung und -analyse der Migration.

Konformitätspaket

Eine Sammlung von AWS Config Regeln und Abhilfemaßnahmen, die Sie zusammenstellen können, um Ihre Konformitäts- und Sicherheitsprüfungen individuell anzupassen. Mithilfe einer YAML-Vorlage können Sie ein Conformance Pack als einzelne Entität in einer AWS-Konto AND-Region oder unternehmensweit bereitstellen. Weitere Informationen finden Sie in der Dokumentation unter Conformance Packs. AWS Config

Kontinuierliche Bereitstellung und kontinuierliche Integration (CI/CD)

Der Prozess der Automatisierung der Quell-, Build-, Test-, Staging- und Produktionsphasen des Softwareveröffentlichungsprozesses. CI/CD is commonly described as a pipeline. CI/CDkann Ihnen helfen, Prozesse zu automatisieren, die Produktivität zu steigern, die Codequalität zu verbessern und schneller zu liefern. Weitere Informationen finden Sie unter Vorteile der kontinuierlichen Auslieferung. CD kann auch für kontinuierliche Bereitstellung stehen. Weitere Informationen finden Sie unterKontinuierliche Auslieferung im Vergleich zu kontinuierlicher Bereitstellung.

CV

Siehe Computer Vision.

D

Daten im Ruhezustand

Daten, die in Ihrem Netzwerk stationär sind, z. B. Daten, die sich im Speicher befinden.

Datenklassifizierung

Ein Prozess zur Identifizierung und Kategorisierung der Daten in Ihrem Netzwerk auf der Grundlage ihrer Kritikalität und Sensitivität. Sie ist eine wichtige Komponente jeder Strategie für das Management von Cybersecurity-Risiken, da sie Ihnen hilft, die geeigneten Schutz- und Aufbewahrungskontrollen für die Daten zu bestimmen. Die Datenklassifizierung ist ein Bestandteil der Sicherheitssäule im AWS Well-Architected Framework. Weitere Informationen finden Sie unter Datenklassifizierung.

Datendrift

Eine signifikante Variation zwischen den Produktionsdaten und den Daten, die zum Trainieren eines ML-Modells verwendet wurden, oder eine signifikante Änderung der Eingabedaten im Laufe der Zeit. Datendrift kann die Gesamtqualität, Genauigkeit und Fairness von ML-Modellvorhersagen beeinträchtigen.

Daten während der Übertragung

Daten, die sich aktiv durch Ihr Netzwerk bewegen, z. B. zwischen Netzwerkressourcen.

Datennetz

Ein architektonisches Framework, das verteilte, dezentrale Dateneigentum mit zentraler Verwaltung und Steuerung ermöglicht.

Datenminimierung

Das Prinzip, nur die Daten zu sammeln und zu verarbeiten, die unbedingt erforderlich sind. Durch Datenminimierung im AWS Cloud können Datenschutzrisiken, Kosten und der CO2-Fußabdruck Ihrer Analysen reduziert werden.

Datenperimeter

Eine Reihe präventiver Schutzmaßnahmen in Ihrer AWS Umgebung, die sicherstellen, dass nur vertrauenswürdige Identitäten auf vertrauenswürdige Ressourcen von erwarteten Netzwerken zugreifen. Weitere Informationen finden Sie unter Aufbau eines Datenperimeters auf. AWS

Vorverarbeitung der Daten

Rohdaten in ein Format umzuwandeln, das von Ihrem ML-Modell problemlos verarbeitet werden kann. Die Vorverarbeitung von Daten kann bedeuten, dass bestimmte Spalten oder Zeilen entfernt und fehlende, inkonsistente oder doppelte Werte behoben werden.

Herkunft der Daten

Der Prozess der Nachverfolgung des Ursprungs und der Geschichte von Daten während ihres gesamten Lebenszyklus, z. B. wie die Daten generiert, übertragen und gespeichert wurden.

betroffene Person

Eine Person, deren Daten gesammelt und verarbeitet werden.

Data Warehouse

Ein Datenverwaltungssystem, das Business Intelligence wie Analysen unterstützt. Data Warehouses enthalten in der Regel große Mengen historischer Daten und werden in der Regel für Abfragen und Analysen verwendet.

Datenbankdefinitionssprache (DDL)

Anweisungen oder Befehle zum Erstellen oder Ändern der Struktur von Tabellen und Objekten in einer Datenbank.

Datenbankmanipulationssprache (DML)

Anweisungen oder Befehle zum Ändern (Einfügen, Aktualisieren und Löschen) von Informationen in einer Datenbank.

DDL

Siehe Datenbankdefinitionssprache.

Deep-Ensemble

Mehrere Deep-Learning-Modelle zur Vorhersage kombinieren. Sie können Deep-Ensembles verwenden, um eine genauere Vorhersage zu erhalten oder um die Unsicherheit von Vorhersagen abzuschätzen.

Deep Learning

Ein ML-Teilbereich, der mehrere Schichten künstlicher neuronaler Netzwerke verwendet, um die Zuordnung zwischen Eingabedaten und Zielvariablen von Interesse zu ermitteln.

defense-in-depth

Ein Ansatz zur Informationssicherheit, bei dem eine Reihe von Sicherheitsmechanismen und -kontrollen sorgfältig in einem Computernetzwerk verteilt werden, um die Vertraulichkeit, Integrität und Verfügbarkeit des Netzwerks und der darin enthaltenen Daten zu schützen. Wenn Sie diese Strategie anwenden AWS, fügen Sie mehrere Steuerelemente auf verschiedenen Ebenen der AWS Organizations Struktur hinzu, um die Ressourcen zu schützen. Ein defense-in-depth Ansatz könnte beispielsweise Multi-Faktor-Authentifizierung, Netzwerksegmentierung und Verschlüsselung kombinieren.

delegierter Administrator

In AWS Organizations kann ein kompatibler Dienst ein AWS Mitgliedskonto registrieren, um die Konten der Organisation und die Berechtigungen für diesen Dienst zu verwalten. Dieses Konto wird als delegierter Administrator für diesen Service bezeichnet. Weitere Informationen und eine Liste kompatibler Services finden Sie unter Services, die mit AWS Organizations funktionieren in der AWS Organizations -Dokumentation.

Bereitstellung

Der Prozess, bei dem eine Anwendung, neue Feature oder Codekorrekturen in der Zielumgebung verfügbar gemacht werden. Die Bereitstellung umfasst das Implementieren von Änderungen an einer Codebasis und das anschließende Erstellen und Ausführen dieser Codebasis in den Anwendungsumgebungen.

Entwicklungsumgebung

Siehe Umgebung.

Detektivische Kontrolle

Eine Sicherheitskontrolle, die darauf ausgelegt ist, ein Ereignis zu erkennen, zu protokollieren und zu warnen, nachdem ein Ereignis eingetreten ist. Diese Kontrollen stellen eine zweite Verteidigungslinie dar und warnen Sie vor Sicherheitsereignissen, bei denen die vorhandenen präventiven Kontrollen umgangen wurden. Weitere Informationen finden Sie unter Detektivische Kontrolle in Implementierung von Sicherheitskontrollen in AWS.

Abbildung des Wertstroms in der Entwicklung (DVSM)

Ein Prozess zur Identifizierung und Priorisierung von Einschränkungen, die sich negativ auf Geschwindigkeit und Qualität im Lebenszyklus der Softwareentwicklung auswirken. DVSM erweitert den Prozess der Wertstromanalyse, der ursprünglich für Lean-Manufacturing-Praktiken

konzipiert wurde. Es konzentriert sich auf die Schritte und Teams, die erforderlich sind, um durch den Softwareentwicklungsprozess Mehrwert zu schaffen und zu steigern.

digitaler Zwilling

Eine virtuelle Darstellung eines realen Systems, z. B. eines Gebäudes, einer Fabrik, einer Industrieanlage oder einer Produktionslinie. Digitale Zwillinge unterstützen vorausschauende Wartung, Fernüberwachung und Produktionsoptimierung.

Maßtabelle

In einem Sternschema eine kleinere Tabelle, die Datenattribute zu quantitativen Daten in einer Faktentabelle enthält. Bei Attributen von Dimensionstabellen handelt es sich in der Regel um Textfelder oder diskrete Zahlen, die sich wie Text verhalten. Diese Attribute werden häufig zum Einschränken von Abfragen, zum Filtern und zur Kennzeichnung von Ergebnismengen verwendet.

Katastrophe

Ein Ereignis, das verhindert, dass ein Workload oder ein System seine Geschäftsziele an seinem primären Einsatzort erfüllt. Diese Ereignisse können Naturkatastrophen, technische Ausfälle oder das Ergebnis menschlichen Handelns sein, z. B. unbeabsichtigte Fehlkonfigurationen oder ein Malware-Angriff.

Disaster Recovery (DR)

Die Strategie und der Prozess, die Sie verwenden, um Ausfallzeiten und Datenverluste aufgrund einer Katastrophe zu minimieren. Weitere Informationen finden Sie unter Disaster Recovery von Workloads unter AWS: Wiederherstellung in der Cloud im AWS Well-Architected Framework.

DML

Siehe Sprache zur Datenbankmanipulation.

Domainorientiertes Design

Ein Ansatz zur Entwicklung eines komplexen Softwaresystems, bei dem seine Komponenten mit sich entwickelnden Domains oder Kerngeschäftszielen verknüpft werden, denen jede Komponente dient. Dieses Konzept wurde von Eric Evans in seinem Buch Domaingesteuertes Design: Bewältigen der Komplexität im Herzen der Software (Boston: Addison-Wesley Professional, 2003) vorgestellt. Informationen darüber, wie Sie domaingesteuertes Design mit dem Strangler-Fig-Muster verwenden können, finden Sie unter Schrittweises Modernisieren älterer Microsoft ASP.NET (ASMX)-Webservices mithilfe von Containern und Amazon API Gateway.

DR

Siehe Disaster Recovery.

Erkennung von Driften

Verfolgung von Abweichungen von einer Basiskonfiguration. Sie können es beispielsweise verwenden, AWS CloudFormation um Abweichungen bei den Systemressourcen zu erkennen, oder Sie können AWS Control Tower damit Änderungen in Ihrer landing zone erkennen, die sich auf die Einhaltung von Governance-Anforderungen auswirken könnten.

DVSM

Siehe Abbildung des Wertstroms in der Entwicklung.

E

EDA

Siehe explorative Datenanalyse.

EDI

Siehe elektronischer Datenaustausch.

Edge-Computing

Die Technologie, die die Rechenleistung für intelligente Geräte an den Rändern eines IoT-Netzwerks erhöht. Im Vergleich zu Cloud Computing kann Edge Computing die Kommunikationslatenz reduzieren und die Reaktionszeit verbessern.

elektronischer Datenaustausch (EDI)

Der automatisierte Austausch von Geschäftsdokumenten zwischen Organisationen. Weitere Informationen finden Sie unter Was ist elektronischer Datenaustausch.

Verschlüsselung

Ein Rechenprozess, der Klartextdaten, die für Menschen lesbar sind, in Chiffretext umwandelt.

Verschlüsselungsschlüssel

Eine kryptografische Zeichenfolge aus zufälligen Bits, die von einem Verschlüsselungsalgorithmus generiert wird. Schlüssel können unterschiedlich lang sein, und jeder Schlüssel ist so konzipiert, dass er unvorhersehbar und einzigartig ist.

Endianismus

Die Reihenfolge, in der Bytes im Computerspeicher gespeichert werden. Big-Endian-Systeme speichern das höchstwertige Byte zuerst. Little-Endian-Systeme speichern das niedrigwertigste Byte zuerst.

Endpunkt

Siehe Service-Endpunkt.

Endpunkt-Services

Ein Service, den Sie in einer Virtual Private Cloud (VPC) hosten können, um ihn mit anderen Benutzern zu teilen. Sie können einen Endpunktdienst mit anderen AWS-Konten oder AWS Identity and Access Management (IAM AWS PrivateLink -) Prinzipalen erstellen und diesen Berechtigungen gewähren. Diese Konten oder Prinzipale können sich privat mit Ihrem Endpunktservice verbinden, indem sie Schnittstellen-VPC-Endpunkte erstellen. Weitere Informationen finden Sie unter Einen Endpunkt-Service erstellen in der Amazon Virtual Private Cloud (Amazon VPC)-Dokumentation.

Unternehmensressourcenplanung (ERP)

Ein System, das wichtige Geschäftsprozesse (wie Buchhaltung, MES und Projektmanagement) für ein Unternehmen automatisiert und verwaltet.

Envelope-Verschlüsselung

Der Prozess der Verschlüsselung eines Verschlüsselungsschlüssels mit einem anderen Verschlüsselungsschlüssel. Weitere Informationen finden Sie unter Envelope-Verschlüsselung in der AWS Key Management Service (AWS KMS) -Dokumentation.

Umgebung

Eine Instance einer laufenden Anwendung. Die folgenden Arten von Umgebungen sind beim Cloud-Computing üblich:

- Entwicklungsumgebung – Eine Instance einer laufenden Anwendung, die nur dem Kernteam zur Verfügung steht, das für die Wartung der Anwendung verantwortlich ist. Entwicklungsumgebungen werden verwendet, um Änderungen zu testen, bevor sie in höhere Umgebungen übertragen werden. Diese Art von Umgebung wird manchmal als Testumgebung bezeichnet.
- Niedrigere Umgebungen – Alle Entwicklungsumgebungen für eine Anwendung, z. B. solche, die für erste Builds und Tests verwendet wurden.

- Produktionsumgebung – Eine Instance einer laufenden Anwendung, auf die Endbenutzer zugreifen können. In einer CI/CD-Pipeline ist die Produktionsumgebung die letzte Bereitstellungsumgebung.
- Höhere Umgebungen – Alle Umgebungen, auf die auch andere Benutzer als das Kernentwicklungsteam zugreifen können. Dies kann eine Produktionsumgebung, Vorproduktionsumgebungen und Umgebungen für Benutzerakzeptanztests umfassen.

Epics

In der agilen Methodik sind dies funktionale Kategorien, die Ihnen helfen, Ihre Arbeit zu organisieren und zu priorisieren. Epics bieten eine allgemeine Beschreibung der Anforderungen und Implementierungsaufgaben. Zu den Sicherheitsthemen AWS von CAF gehören beispielsweise Identitäts- und Zugriffsmanagement, Detektivkontrollen, Infrastruktursicherheit, Datenschutz und Reaktion auf Vorfälle. Weitere Informationen zu Epics in der AWS - Migrationsstrategie finden Sie im Leitfaden zur Programm-Implementierung.

ERP

Siehe Enterprise Resource Planning.

Explorative Datenanalyse (EDA)

Der Prozess der Analyse eines Datensatzes, um seine Hauptmerkmale zu verstehen. Sie sammeln oder aggregieren Daten und führen dann erste Untersuchungen durch, um Muster zu finden, Anomalien zu erkennen und Annahmen zu überprüfen. EDA wird durchgeführt, indem zusammenfassende Statistiken berechnet und Datenvisualisierungen erstellt werden.

F

Faktentabelle

Die zentrale Tabelle in einem Sternschema. Sie speichert quantitative Daten über den Geschäftsbetrieb. In der Regel enthält eine Faktentabelle zwei Arten von Spalten: Spalten, die Kennzahlen enthalten, und Spalten, die einen Fremdschlüssel für eine Dimensionstabelle enthalten.

schnell scheitern

Eine Philosophie, die häufige und inkrementelle Tests verwendet, um den Entwicklungslebenszyklus zu verkürzen. Dies ist ein wichtiger Bestandteil eines agilen Ansatzes.

Grenze zur Fehlerisolierung

Dabei handelt es sich um eine Grenze AWS Cloud, z. B. eine Availability Zone AWS-Region, eine Steuerungsebene oder eine Datenebene, die die Auswirkungen eines Fehlers begrenzt und die Widerstandsfähigkeit von Workloads verbessert. Weitere Informationen finden Sie unter Grenzen zur AWS Fehlerisolierung.

Feature-Zweig

Siehe Zweig.

Features

Die Eingabedaten, die Sie verwenden, um eine Vorhersage zu treffen. In einem Fertigungskontext könnten Feature beispielsweise Bilder sein, die regelmäßig von der Fertigungslinie aus aufgenommen werden.

Bedeutung der Feature

Wie wichtig ein Feature für die Vorhersagen eines Modells ist. Dies wird in der Regel als numerischer Wert ausgedrückt, der mit verschiedenen Techniken wie Shapley Additive Explanations (SHAP) und integrierten Gradienten berechnet werden kann. Weitere Informationen finden Sie unter Interpretierbarkeit von Modellen für maschinelles Lernen mit AWS.

Featuretransformation

Daten für den ML-Prozess optimieren, einschließlich der Anreicherung von Daten mit zusätzlichen Quellen, der Skalierung von Werten oder der Extraktion mehrerer Informationssätze aus einem einzigen Datenfeld. Das ermöglicht dem ML-Modell, von den Daten profitieren. Wenn Sie beispielsweise das Datum „27.05.2021 00:15:37“ in „2021“, „Mai“, „Donnerstag“ und „15“ aufschlüsseln, können Sie dem Lernalgorithmus helfen, nuancierte Muster zu erlernen, die mit verschiedenen Datenkomponenten verknüpft sind.

Eingabeaufforderung mit wenigen Klicks

Bereitstellung einer kleinen Anzahl von Beispielen, die die Aufgabe und das gewünschte Ergebnis veranschaulichen, bevor das LLM aufgefordert wird, eine ähnliche Aufgabe auszuführen. Bei dieser Technik handelt es sich um eine Anwendung des kontextbezogenen Lernens, bei der Modelle anhand von Beispielen (Aufnahmen) lernen, die in Eingabeaufforderungen eingebettet sind. Bei Aufgaben, die spezifische Formatierungs-, Argumentations- oder Fachkenntnisse erfordern, kann die Eingabeaufforderung mit wenigen Handgriffen effektiv sein. Siehe auch Zero-Shot Prompting.

FGAC

Siehe detaillierte Zugriffskontrolle.

Feinkörnige Zugriffskontrolle (FGAC)

Die Verwendung mehrerer Bedingungen, um eine Zugriffsanfrage zuzulassen oder abzulehnen.

Flash-Cut-Migration

Eine Datenbankmigrationsmethode, bei der eine kontinuierliche Datenreplikation durch Erfassung von Änderungsdaten verwendet wird, um Daten in kürzester Zeit zu migrieren, anstatt einen schrittweisen Ansatz zu verwenden. Ziel ist es, Ausfallzeiten auf ein Minimum zu beschränken.

FM

Siehe Fundamentmodell.

Fundamentmodell (FM)

Ein großes neuronales Deep-Learning-Netzwerk, das mit riesigen Datensätzen generalisierter und unbeschrifteter Daten trainiert wurde. FMs sind in der Lage, eine Vielzahl allgemeiner Aufgaben zu erfüllen, z. B. Sprache zu verstehen, Text und Bilder zu generieren und Konversationen in natürlicher Sprache zu führen. Weitere Informationen finden Sie unter Was sind Foundation-Modelle.

G

generative KI

Eine Untergruppe von KI-Modellen, die mit großen Datenmengen trainiert wurden und mit einer einfachen Textaufforderung neue Inhalte und Artefakte wie Bilder, Videos, Text und Audio erstellen können. Weitere Informationen finden Sie unter Was ist Generative KI.

Geoblocking

Siehe geografische Einschränkungen.

Geografische Einschränkungen (Geoblocking)

Bei Amazon eine Option CloudFront, um zu verhindern, dass Benutzer in bestimmten Ländern auf Inhaltsverteilungen zugreifen. Sie können eine Zulassungsliste oder eine Sperrliste verwenden,

um zugelassene und gesperrte Länder anzugeben. Weitere Informationen finden Sie in der Dokumentation unter Beschränkung der geografischen Verteilung Ihrer Inhalte. CloudFront

Gitflow-Workflow

Ein Ansatz, bei dem niedrigere und höhere Umgebungen unterschiedliche Zweige in einem Quellcode-Repository verwenden. Der Gitflow-Workflow gilt als veraltet, und der Trunk-basierte Workflow ist der moderne, bevorzugte Ansatz.

goldenes Bild

Ein Snapshot eines Systems oder einer Software, der als Vorlage für die Bereitstellung neuer Instanzen dieses Systems oder dieser Software verwendet wird. In der Fertigung kann ein Golden Image beispielsweise zur Bereitstellung von Software auf mehreren Geräten verwendet werden und trägt zur Verbesserung der Geschwindigkeit, Skalierbarkeit und Produktivität bei der Geräteherstellung bei.

Greenfield-Strategie

Das Fehlen vorhandener Infrastruktur in einer neuen Umgebung. Bei der Einführung einer Neuausrichtung einer Systemarchitektur können Sie alle neuen Technologien ohne Einschränkung der Kompatibilität mit der vorhandenen Infrastruktur auswählen, auch bekannt als Brownfield. Wenn Sie die bestehende Infrastruktur erweitern, könnten Sie Brownfield- und Greenfield-Strategien mischen.

Integritätsschutz

Eine allgemeine Regel, die dazu beiträgt, Ressourcen, Richtlinien und die Einhaltung von Vorschriften in allen Unternehmenseinheiten zu regeln (OUs). Präventiver Integritätsschutz setzt Richtlinien durch, um die Einhaltung von Standards zu gewährleisten. Sie werden mithilfe von Service-Kontrollrichtlinien und IAM-Berechtigungsgrenzen implementiert. Detektivischer Integritätsschutz erkennt Richtlinienverstöße und Compliance-Probleme und generiert Warnmeldungen zur Abhilfe. Sie werden mithilfe von AWS Config, AWS Security Hub, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector und benutzerdefinierten AWS Lambda Prüfungen implementiert.

H

HEKTAR

Siehe Hochverfügbarkeit.

Heterogene Datenbankmigration

Migrieren Sie Ihre Quelldatenbank in eine Zieldatenbank, die eine andere Datenbank-Engine verwendet (z. B. Oracle zu Amazon Aurora). Eine heterogene Migration ist in der Regel Teil einer Neuarchitektur, und die Konvertierung des Schemas kann eine komplexe Aufgabe sein. AWS bietet AWS SCT, welches bei Schemakonvertierungen hilft.

hohe Verfügbarkeit (HA)

Die Fähigkeit eines Workloads, im Falle von Herausforderungen oder Katastrophen kontinuierlich und ohne Eingreifen zu arbeiten. HA-Systeme sind so konzipiert, dass sie automatisch ein Failover durchführen, gleichbleibend hohe Leistung bieten und unterschiedliche Lasten und Ausfälle mit minimalen Leistungseinbußen bewältigen.

historische Modernisierung

Ein Ansatz zur Modernisierung und Aufrüstung von Betriebstechnologiesystemen (OT), um den Bedürfnissen der Fertigungsindustrie besser gerecht zu werden. Ein Historian ist eine Art von Datenbank, die verwendet wird, um Daten aus verschiedenen Quellen in einer Fabrik zu sammeln und zu speichern.

Holdout-Daten

Ein Teil historischer, beschrifteter Daten, der aus einem Datensatz zurückgehalten wird, der zum Trainieren eines Modells für maschinelles Lernen verwendet wird. Sie können Holdout-Daten verwenden, um die Modellleistung zu bewerten, indem Sie die Modellvorhersagen mit den Holdout-Daten vergleichen.

Homogene Datenbankmigration

Migrieren Sie Ihre Quelldatenbank zu einer Zieldatenbank, die dieselbe Datenbank-Engine verwendet (z. B. Microsoft SQL Server zu Amazon RDS für SQL Server). Eine homogene Migration ist in der Regel Teil eines Hostwechsels oder eines Plattformwechsels. Sie können native Datenbankserviceprogramme verwenden, um das Schema zu migrieren.

heiße Daten

Daten, auf die häufig zugegriffen wird, z. B. Echtzeitdaten oder aktuelle Translationsdaten. Für diese Daten ist in der Regel eine leistungsstarke Speicherebene oder -klasse erforderlich, um schnelle Abfrageantworten zu ermöglichen.

Hotfix

Eine dringende Lösung für ein kritisches Problem in einer Produktionsumgebung. Aufgrund seiner Dringlichkeit wird ein Hotfix normalerweise außerhalb des typischen DevOps Release-Workflows erstellt.

Hypercare-Phase

Unmittelbar nach dem Cutover, der Zeitraum, in dem ein Migrationsteam die migrierten Anwendungen in der Cloud verwaltet und überwacht, um etwaige Probleme zu beheben. In der Regel dauert dieser Zeitraum 1–4 Tage. Am Ende der Hypercare-Phase überträgt das Migrationsteam in der Regel die Verantwortung für die Anwendungen an das Cloud-Betriebsteam.

I

IaC

Sehen Sie sich Infrastruktur als Code an.

Identitätsbasierte Richtlinie

Eine Richtlinie, die einem oder mehreren IAM-Prinzipalen zugeordnet ist und deren Berechtigungen innerhalb der AWS Cloud Umgebung definiert.

Leerlaufanwendung

Eine Anwendung mit einer durchschnittlichen CPU- und Arbeitsspeicherauslastung zwischen 5 und 20 Prozent über einen Zeitraum von 90 Tagen. In einem Migrationsprojekt ist es üblich, diese Anwendungen außer Betrieb zu nehmen oder sie On-Premises beizubehalten.

IIoT

Siehe Industrielles Internet der Dinge.

unveränderliche Infrastruktur

Ein Modell, das eine neue Infrastruktur für Produktionsworkloads bereitstellt, anstatt die bestehende Infrastruktur zu aktualisieren, zu patchen oder zu modifizieren. Unveränderliche Infrastrukturen sind von Natur aus konsistenter, zuverlässiger und vorhersehbarer als veränderliche Infrastrukturen. Weitere Informationen finden Sie in der Best Practice Deploy using immutable infrastructure im AWS Well-Architected Framework.

Eingehende (ingress) VPC

In einer Architektur AWS mit mehreren Konten ist dies eine VPC, die Netzwerkverbindungen von außerhalb einer Anwendung akzeptiert, überprüft und weiterleitet. Die AWS Security Reference Architecture empfiehlt, Ihr Netzwerkkonto mit eingehendem und ausgehendem Datenverkehr und Inspektion einzurichten, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

Inkrementelle Migration

Eine Cutover-Strategie, bei der Sie Ihre Anwendung in kleinen Teilen migrieren, anstatt eine einziges vollständiges Cutover durchzuführen. Beispielsweise könnten Sie zunächst nur einige Microservices oder Benutzer auf das neue System umstellen. Nachdem Sie sich vergewissert haben, dass alles ordnungsgemäß funktioniert, können Sie weitere Microservices oder Benutzer schrittweise verschieben, bis Sie Ihr Legacy-System außer Betrieb nehmen können. Diese Strategie reduziert die mit großen Migrationen verbundenen Risiken.

Industrie 4.0

Ein Begriff, der 2016 von Klaus Schwab eingeführt wurde und sich auf die Modernisierung von Fertigungsprozessen durch Fortschritte in den Bereichen Konnektivität, Echtzeitdaten, Automatisierung, Analytik und KI/ML bezieht.

Infrastruktur

Alle Ressourcen und Komponenten, die in der Umgebung einer Anwendung enthalten sind.

Infrastructure as Code (IaC)

Der Prozess der Bereitstellung und Verwaltung der Infrastruktur einer Anwendung mithilfe einer Reihe von Konfigurationsdateien. IaC soll Ihnen helfen, das Infrastrukturmanagement zu zentralisieren, Ressourcen zu standardisieren und schnell zu skalieren, sodass neue Umgebungen wiederholbar, zuverlässig und konsistent sind.

industrielles Internet der Dinge (T) IIo

Einsatz von mit dem Internet verbundenen Sensoren und Geräten in Industriesektoren wie Fertigung, Energie, Automobilindustrie, Gesundheitswesen, Biowissenschaften und Landwirtschaft. Weitere Informationen finden Sie unter Aufbau einer digitalen Transformationsstrategie für das industrielle Internet der Dinge (IIoT).

Inspektions-VPC

In einer Architektur AWS mit mehreren Konten eine zentralisierte VPC, die Inspektionen des Netzwerkverkehrs zwischen VPCs (in demselben oder unterschiedlichen AWS-Regionen), dem Internet und lokalen Netzwerken verwaltet. In der AWS Security Reference Architecture wird empfohlen, Ihr Netzwerkkonto mit eingehendem und ausgehendem Datenverkehr sowie Inspektionen einzurichten, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

Internet of Things (IoT)

Das Netzwerk verbundener physischer Objekte mit eingebetteten Sensoren oder Prozessoren, das über das Internet oder über ein lokales Kommunikationsnetzwerk mit anderen Geräten und Systemen kommuniziert. Weitere Informationen finden Sie unter Was ist IoT?

Interpretierbarkeit

Ein Merkmal eines Modells für Machine Learning, das beschreibt, inwieweit ein Mensch verstehen kann, wie die Vorhersagen des Modells von seinen Eingaben abhängen. Weitere Informationen finden Sie unter Interpretierbarkeit von Modellen für maschinelles Lernen mit. AWS

IoT

Siehe Internet der Dinge.

IT information library (ITIL, IT-Informationsbibliothek)

Eine Reihe von bewährten Methoden für die Bereitstellung von IT-Services und die Abstimmung dieser Services auf die Geschäftsanforderungen. ITIL bietet die Grundlage für ITSM.

T service management (ITSM, IT-Servicemanagement)

Aktivitäten im Zusammenhang mit der Gestaltung, Implementierung, Verwaltung und Unterstützung von IT-Services für eine Organisation. Informationen zur Integration von Cloud-Vorgängen mit ITSM-Tools finden Sie im Leitfaden zur Betriebsintegration.

BIS

Weitere Informationen finden Sie in der IT-Informationsbibliothek.

ITSM

Siehe IT-Servicemanagement.

L

Labelbasierte Zugangskontrolle (LBAC)

: Eine Implementierung der Mandatory Access Control (MAC), bei der den Benutzern und den Daten selbst jeweils explizit ein Sicherheitslabelwert zugewiesen wird. Die Schnittmenge zwischen der Benutzersicherheitsbeschriftung und der Datensicherheitsbeschriftung bestimmt, welche Zeilen und Spalten für den Benutzer sichtbar sind.

Landing Zone

: Eine landing zone ist eine gut strukturierte AWS Umgebung mit mehreren Konten, die skalierbar und sicher ist. Dies ist ein Ausgangspunkt, von dem aus Ihre Organisationen Workloads und Anwendungen schnell und mit Vertrauen in ihre Sicherheits- und Infrastrukturumgebung starten und bereitstellen können. Weitere Informationen zu Landing Zones finden Sie unter Einrichtung einer sicheren und skalierbaren AWS -Umgebung mit mehreren Konten..

großes Sprachmodell (LLM)

: Ein Deep-Learning-KI-Modell, das anhand einer riesigen Datenmenge vorab trainiert wurde. Ein LLM kann mehrere Aufgaben ausführen, z. B. Fragen beantworten, Dokumente zusammenfassen, Text in andere Sprachen übersetzen und Sätze vervollständigen. Weitere Informationen finden Sie unter Was sind. LLMs

Große Migration

: Eine Migration von 300 oder mehr Servern.

SCHWARZ

: Weitere Informationen finden Sie unter Label-basierte Zugriffskontrolle.

Geringste Berechtigung

: Die bewährte Sicherheitsmethode, bei der nur die für die Durchführung einer Aufgabe erforderlichen Mindestberechtigungen erteilt werden. Weitere Informationen finden Sie unter Geringste Berechtigungen anwenden in der IAM-Dokumentation.

Lift and Shift

: Siehe 7 Rs.

Little-Endian-System

: Ein System, welches das niedrigwertigste Byte zuerst speichert. Siehe auch Endianness.

LLM

Siehe großes Sprachmodell.

Niedrigere Umgebungen

Siehe Umgebung.

M

Machine Learning (ML)

Eine Art künstlicher Intelligenz, die Algorithmen und Techniken zur Mustererkennung und zum Lernen verwendet. ML analysiert aufgezeichnete Daten, wie z. B. Daten aus dem Internet der Dinge (IoT), und lernt daraus, um ein statistisches Modell auf der Grundlage von Mustern zu erstellen. Weitere Informationen finden Sie unter Machine Learning.

Hauptzweig

Siehe Filiale.

Malware

Software, die entwickelt wurde, um die Computersicherheit oder den Datenschutz zu gefährden. Malware kann Computersysteme stören, vertrauliche Informationen durchsickern lassen oder sich unbefugten Zugriff verschaffen. Beispiele für Malware sind Viren, Würmer, Ransomware, Trojaner, Spyware und Keylogger.

verwaltete Dienste

AWS-Services für die die Infrastrukturebene, das Betriebssystem und die Plattformen AWS betrieben werden, und Sie greifen auf die Endgeräte zu, um Daten zu speichern und abzurufen. Amazon Simple Storage Service (Amazon S3) und Amazon DynamoDB sind Beispiele für Managed Services. Diese werden auch als abstrakte Dienste bezeichnet.

Manufacturing Execution System (MES)

Ein Softwaresystem zur Nachverfolgung, Überwachung, Dokumentation und Steuerung von Produktionsprozessen, bei denen Rohstoffe in der Fertigung zu fertigen Produkten umgewandelt werden.

MAP

Siehe Migration Acceleration Program.

Mechanismus

Ein vollständiger Prozess, bei dem Sie ein Tool erstellen, die Akzeptanz des Tools vorantreiben und anschließend die Ergebnisse überprüfen, um Anpassungen vorzunehmen. Ein Mechanismus ist ein Zyklus, der sich im Laufe seiner Tätigkeit selbst verstärkt und verbessert. Weitere Informationen finden Sie unter Aufbau von Mechanismen im AWS Well-Architected Framework.

Mitgliedskonto

Alle AWS-Konten außer dem Verwaltungskonto, die Teil einer Organisation in sind. AWS Organizations Ein Konto kann jeweils nur einer Organisation angehören.

DURCHEINANDER

Siehe Manufacturing Execution System.

Message Queuing-Telemetrietransport (MQTT)

Ein leichtes machine-to-machine (M2M) -Kommunikationsprotokoll, das auf dem Publish/Subscribe-Muster für IoT-Geräte mit beschränkten Ressourcen basiert.

Microservice

Ein kleiner, unabhängiger Dienst, der über genau definierte Kanäle kommuniziert APIs und in der Regel kleinen, eigenständigen Teams gehört. Ein Versicherungssystem kann beispielsweise Microservices beinhalten, die Geschäftsfunktionen wie Vertrieb oder Marketing oder Subdomains wie Einkauf, Schadenersatz oder Analytik zugeordnet sind. Zu den Vorteilen von Microservices gehören Agilität, flexible Skalierung, einfache Bereitstellung, wiederverwendbarer Code und Ausfallsicherheit. Weitere Informationen finden Sie unter Integration von Microservices mithilfe serverloser Dienste. AWS

Microservices-Architekturen

Ein Ansatz zur Erstellung einer Anwendung mit unabhängigen Komponenten, die jeden Anwendungsprozess als Microservice ausführen. Diese Microservices kommunizieren mithilfe von Lightweight über eine klar definierte Schnittstelle. APIs Jeder Microservice in dieser Architektur kann aktualisiert, bereitgestellt und skaliert werden, um den Bedarf an bestimmten Funktionen einer Anwendung zu decken. Weitere Informationen finden Sie unter Implementierung von Microservices auf. AWS

Migration Acceleration Program (MAP)

Ein AWS Programm, das Beratung, Unterstützung, Schulungen und Services bietet, um Unternehmen dabei zu unterstützen, eine solide betriebliche Grundlage für die Umstellung auf

die Cloud zu schaffen und die anfänglichen Kosten von Migrationen auszugleichen. MAP umfasst eine Migrationsmethode für die methodische Durchführung von Legacy-Migrationen sowie eine Reihe von Tools zur Automatisierung und Beschleunigung gängiger Migrationsszenarien.

Migration in großem Maßstab

Der Prozess, bei dem der Großteil des Anwendungsportfolios in Wellen in die Cloud verlagert wird, wobei in jeder Welle mehr Anwendungen schneller migriert werden. In dieser Phase werden die bewährten Verfahren und Erkenntnisse aus den früheren Phasen zur Implementierung einer Migrationsfabrik von Teams, Tools und Prozessen zur Optimierung der Migration von Workloads durch Automatisierung und agile Bereitstellung verwendet. Dies ist die dritte Phase der AWS - Migrationsstrategie.

Migrationsfabrik

Funktionsübergreifende Teams, die die Migration von Workloads durch automatisierte, agile Ansätze optimieren. Zu den Teams in der Migrationsabteilung gehören in der Regel Betriebsabläufe, Geschäftsanalysten und Eigentümer, Migrationsingenieure, Entwickler und DevOps Experten, die in Sprints arbeiten. Zwischen 20 und 50 Prozent eines Unternehmensanwendungsportfolios bestehen aus sich wiederholenden Mustern, die durch einen Fabrik-Ansatz optimiert werden können. Weitere Informationen finden Sie in Diskussion über Migrationsfabriken und den Leitfaden zur Cloud-Migration-Fabrik in diesem Inhaltssatz.

Migrationsmetadaten

Die Informationen über die Anwendung und den Server, die für den Abschluss der Migration benötigt werden. Für jedes Migrationsmuster ist ein anderer Satz von Migrationsmetadaten erforderlich. Beispiele für Migrationsmetadaten sind das Zielsubnetz, die Sicherheitsgruppe und AWS das Konto.

Migrationsmuster

Eine wiederholbare Migrationsaufgabe, in der die Migrationsstrategie, das Migrationsziel und die verwendete Migrationsanwendung oder der verwendete Migrationsservice detailliert beschrieben werden. Beispiel: Rehost-Migration zu Amazon EC2 mit AWS Application Migration Service.

Migration Portfolio Assessment (MPA)

Ein Online-Tool, das Informationen zur Validierung des Geschäftsszenarios für die Migration auf das bereitstellt. AWS Cloud MPA bietet eine detaillierte Portfoliobewertung (richtige Servergröße, Preisgestaltung, Gesamtbetriebskostenanalyse, Migrationskostenanalyse) sowie Migrationsplanung (Anwendungsdatenanalyse und Datenerfassung, Anwendungsgruppierung,

Migrationspriorisierung und Wellenplanung). Das MPA-Tool (Anmeldung erforderlich) steht allen AWS Beratern und APN-Partnerberatern kostenlos zur Verfügung.

Migration Readiness Assessment (MRA)

Der Prozess, bei dem mithilfe des AWS CAF Erkenntnisse über den Cloud-Bereitschaftsstatus eines Unternehmens gewonnen, Stärken und Schwächen identifiziert und ein Aktionsplan zur Schließung festgestellter Lücken erstellt wird. Weitere Informationen finden Sie im Benutzerhandbuch für Migration Readiness. MRA ist die erste Phase der AWS - Migrationsstrategie.

Migrationsstrategie

Der Ansatz, der verwendet wurde, um einen Workload auf den AWS Cloud zu migrieren. Weitere Informationen finden Sie im Eintrag 7 Rs in diesem Glossar und unter Mobilisieren Sie Ihr Unternehmen, um umfangreiche Migrationen zu beschleunigen.

ML

Siehe maschinelles Lernen.

Modernisierung

Umwandlung einer veralteten (veralteten oder monolithischen) Anwendung und ihrer Infrastruktur in ein agiles, elastisches und hochverfügbares System in der Cloud, um Kosten zu senken, die Effizienz zu steigern und Innovationen zu nutzen. Weitere Informationen finden Sie unter Strategie zur Modernisierung von Anwendungen in der AWS Cloud.

Bewertung der Modernisierungsfähigkeit

Eine Bewertung, anhand derer festgestellt werden kann, ob die Anwendungen einer Organisation für die Modernisierung bereit sind, Vorteile, Risiken und Abhängigkeiten identifiziert und ermittelt wird, wie gut die Organisation den zukünftigen Status dieser Anwendungen unterstützen kann. Das Ergebnis der Bewertung ist eine Vorlage der Zielarchitektur, eine Roadmap, in der die Entwicklungsphasen und Meilensteine des Modernisierungsprozesses detailliert beschrieben werden, sowie ein Aktionsplan zur Behebung festgestellter Lücken. Weitere Informationen finden Sie unter Evaluierung der Modernisierungsbereitschaft von Anwendungen in der AWS Cloud.

Monolithische Anwendungen (Monolithen)

Anwendungen, die als ein einziger Service mit eng gekoppelten Prozessen ausgeführt werden. Monolithische Anwendungen haben verschiedene Nachteile. Wenn ein Anwendungs-Feature stark nachgefragt wird, muss die gesamte Architektur skaliert werden. Das Hinzufügen oder

Verbessern der Feature einer monolithischen Anwendung wird ebenfalls komplexer, wenn die Codebasis wächst. Um diese Probleme zu beheben, können Sie eine Microservices-Architektur verwenden. Weitere Informationen finden Sie unter Zerlegen von Monolithen in Microservices.

MPA

Siehe Bewertung des Migrationsportfolios.

MQTT

Siehe Message Queuing-Telemetrietransport.

Mehrklassen-Klassifizierung

Ein Prozess, der dabei hilft, Vorhersagen für mehrere Klassen zu generieren (wobei eines von mehr als zwei Ergebnissen vorhergesagt wird). Ein ML-Modell könnte beispielsweise fragen: „Ist dieses Produkt ein Buch, ein Auto oder ein Telefon?“ oder „Welche Kategorie von Produkten ist für diesen Kunden am interessantesten?“

veränderbare Infrastruktur

Ein Modell, das die bestehende Infrastruktur für Produktionsworkloads aktualisiert und modifiziert. Für eine verbesserte Konsistenz, Zuverlässigkeit und Vorhersagbarkeit empfiehlt das AWS Well-Architected Framework die Verwendung einer unveränderlichen Infrastruktur als bewährte Methode.

O

OAC

Siehe Origin Access Control.

OAI

Siehe Zugriffsidentität von Origin.

COM

Siehe organisatorisches Change-Management.

Offline-Migration

Eine Migrationsmethode, bei der der Quell-Workload während des Migrationsprozesses heruntergefahren wird. Diese Methode ist mit längeren Ausfallzeiten verbunden und wird in der Regel für kleine, unkritische Workloads verwendet.

OI

Siehe Betriebsintegration.

OLA

Siehe Vereinbarung auf operativer Ebene.

Online-Migration

Eine Migrationsmethode, bei der der Quell-Workload auf das Zielsystem kopiert wird, ohne offline genommen zu werden. Anwendungen, die mit dem Workload verbunden sind, können während der Migration weiterhin funktionieren. Diese Methode beinhaltet keine bis minimale Ausfallzeit und wird in der Regel für kritische Produktionsworkloads verwendet.

OPC-UA

Siehe Open Process Communications — Unified Architecture.

Offene Prozesskommunikation — Einheitliche Architektur (OPC-UA)

Ein machine-to-machine (M2M) -Kommunikationsprotokoll für die industrielle Automatisierung. OPC-UA bietet einen Interoperabilitätsstandard mit Datenverschlüsselungs-, Authentifizierungs- und Autorisierungsschemata.

Vereinbarung auf Betriebsebene (OLA)

Eine Vereinbarung, in der klargestellt wird, welche funktionalen IT-Gruppen sich gegenseitig versprechen zu liefern, um ein Service Level Agreement (SLA) zu unterstützen.

Überprüfung der Betriebsbereitschaft (ORR)

Eine Checkliste mit Fragen und zugehörigen bewährten Methoden, die Ihnen helfen, Vorfälle und mögliche Ausfälle zu verstehen, zu bewerten, zu verhindern oder deren Umfang zu reduzieren. Weitere Informationen finden Sie unter Operational Readiness Reviews (ORR) im AWS Well-Architected Framework.

Betriebstechnologie (OT)

Hardware- und Softwaresysteme, die mit der physischen Umgebung zusammenarbeiten, um industrielle Abläufe, Ausrüstung und Infrastruktur zu steuern. In der Fertigung ist die Integration von OT- und Informationstechnologie (IT) -Systemen ein zentraler Schwerpunkt der Industrie 4.0-Transformationen.

Betriebsintegration (OI)

Der Prozess der Modernisierung von Abläufen in der Cloud, der Bereitschaftsplanung, Automatisierung und Integration umfasst. Weitere Informationen finden Sie im Leitfaden zur Betriebsintegration.

Organisationspfad

Ein Pfad, der von erstellt wird und in AWS CloudTrail dem alle Ereignisse für alle AWS-Konten in einer Organisation protokolliert werden. AWS Organizations Diese Spur wird in jedem AWS-Konto , der Teil der Organisation ist, erstellt und verfolgt die Aktivität in jedem Konto. Weitere Informationen finden Sie in der CloudTrail Dokumentation unter Einen Trail für eine Organisation erstellen.

Organisatorisches Veränderungsmanagement (OCM)

Ein Framework für das Management wichtiger, disruptiver Geschäftstransformationen aus Sicht der Mitarbeiter, der Kultur und der Führung. OCM hilft Organisationen dabei, sich auf neue Systeme und Strategien vorzubereiten und auf diese umzustellen, indem es die Akzeptanz von Veränderungen beschleunigt, Übergangsprobleme angeht und kulturelle und organisatorische Veränderungen vorantreibt. In der AWS Migrationsstrategie wird dieses Framework aufgrund der Geschwindigkeit des Wandels, der bei Projekten zur Cloud-Einführung erforderlich ist, als Mitarbeiterbeschleunigung bezeichnet. Weitere Informationen finden Sie im OCM-Handbuch.

Ursprungszugriffskontrolle (OAC)

In CloudFront, eine erweiterte Option zur Zugriffsbeschränkung, um Ihre Amazon Simple Storage Service (Amazon S3) -Inhalte zu sichern. OAC unterstützt alle S3-Buckets insgesamt AWS-Regionen, serverseitige Verschlüsselung mit AWS KMS (SSE-KMS) sowie dynamische `PUT` und `DELETE` Anfragen an den S3-Bucket.

Ursprungszugriffsidentität (OAI)

In CloudFront, eine Option zur Zugriffsbeschränkung, um Ihre Amazon S3 S3-Inhalte zu sichern. Wenn Sie OAI verwenden, CloudFront erstellt es einen Principal, mit dem sich Amazon S3 authentifizieren kann. Authentifizierte Principals können nur über eine bestimmte Distribution auf Inhalte in einem S3-Bucket zugreifen. CloudFront Siehe auch OAC, das eine detailliertere und verbesserte Zugriffskontrolle bietet.

ORR

Weitere Informationen finden Sie unter Überprüfung der Betriebsbereitschaft.

NICHT

Siehe Betriebstechnologie.

Ausgehende (egress) VPC

In einer Architektur AWS mit mehreren Konten eine VPC, die Netzwerkverbindungen verarbeitet, die von einer Anwendung aus initiiert werden. Die AWS Security Reference Architecture empfiehlt die Einrichtung Ihres Netzwerkkontos mit eingehendem und ausgehendem Datenverkehr sowie Inspektion, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

P

Berechtigungsgrenze

Eine IAM-Verwaltungsrichtlinie, die den IAM-Prinzipalen zugeordnet ist, um die maximalen Berechtigungen festzulegen, die der Benutzer oder die Rolle haben kann. Weitere Informationen finden Sie unter Berechtigungsgrenzen für IAM-Entitys in der IAM-Dokumentation.

persönlich identifizierbare Informationen (PII)

Informationen, die, wenn sie direkt betrachtet oder mit anderen verwandten Daten kombiniert werden, verwendet werden können, um vernünftige Rückschlüsse auf die Identität einer Person zu ziehen. Beispiele für personenbezogene Daten sind Namen, Adressen und Kontaktinformationen.

Personenbezogene Daten

Siehe persönlich identifizierbare Informationen.

Playbook

Eine Reihe vordefinierter Schritte, die die mit Migrationen verbundenen Aufgaben erfassen, z. B. die Bereitstellung zentraler Betriebsfunktionen in der Cloud. Ein Playbook kann die Form von Skripten, automatisierten Runbooks oder einer Zusammenfassung der Prozesse oder Schritte annehmen, die für den Betrieb Ihrer modernisierten Umgebung erforderlich sind.

PLC

Siehe programmierbare Logiksteuerung.

PLM

Siehe Produktlebenszyklusmanagement.

policy

Ein Objekt, das Berechtigungen definieren (siehe identitätsbasierte Richtlinie), Zugriffsbedingungen spezifizieren (siehe ressourcenbasierte Richtlinie) oder die maximalen Berechtigungen für alle Konten in einer Organisation definieren kann AWS Organizations (siehe Dienststeuerungsrichtlinie).

Polyglotte Beharrlichkeit

Unabhängige Auswahl der Datenspeichertechnologie eines Microservices auf der Grundlage von Datenzugriffsmustern und anderen Anforderungen. Wenn Ihre Microservices über dieselbe Datenspeichertechnologie verfügen, kann dies zu Implementierungsproblemen oder zu Leistungseinbußen führen. Microservices lassen sich leichter implementieren und erzielen eine bessere Leistung und Skalierbarkeit, wenn sie den Datenspeicher verwenden, der ihren Anforderungen am besten entspricht. Weitere Informationen finden Sie unter Datenpersistenz in Microservices aktivieren.

Portfoliobewertung

Ein Prozess, bei dem das Anwendungsportfolio ermittelt, analysiert und priorisiert wird, um die Migration zu planen. Weitere Informationen finden Sie in Bewerten der Migrationsbereitschaft.

predicate

Eine Abfragebedingung, die `true` oder zurückgibt`false`, was üblicherweise in einer Klausel vorkommt. `WHERE`

Prädikat Pushdown

Eine Technik zur Optimierung von Datenbankabfragen, bei der die Daten in der Abfrage vor der Übertragung gefiltert werden. Dadurch wird die Datenmenge reduziert, die aus der relationalen Datenbank abgerufen und verarbeitet werden muss, und die Abfrageleistung wird verbessert.

Präventive Kontrolle

Eine Sicherheitskontrolle, die verhindern soll, dass ein Ereignis eintritt. Diese Kontrollen stellen eine erste Verteidigungslinie dar, um unbefugten Zugriff oder unerwünschte Änderungen an Ihrem Netzwerk zu verhindern. Weitere Informationen finden Sie unter Präventive Kontrolle in Implementierung von Sicherheitskontrollen in AWS.

Prinzipal

Eine Entität AWS , die Aktionen ausführen und auf Ressourcen zugreifen kann. Bei dieser Entität handelt es sich in der Regel um einen Root-Benutzer für eine AWS-Konto, eine IAM-Rolle oder einen Benutzer. Weitere Informationen finden Sie unter Prinzipal in Rollenbegriffe und -konzepte in der IAM-Dokumentation.

Datenschutz von Natur aus

Ein systemtechnischer Ansatz, der den Datenschutz während des gesamten Entwicklungsprozesses berücksichtigt.

Privat gehostete Zonen

Ein Container, der Informationen darüber enthält, wie Amazon Route 53 auf DNS-Abfragen für eine Domain und deren Subdomains innerhalb einer oder mehrerer VPCs Domains antworten soll. Weitere Informationen finden Sie unter Arbeiten mit privat gehosteten Zonen in der Route-53-Dokumentation.

proaktive Steuerung

Eine Sicherheitskontrolle, die den Einsatz nicht richtlinienkonformer Ressourcen verhindern soll. Diese Steuerelemente scannen Ressourcen, bevor sie bereitgestellt werden. Wenn die Ressource nicht mit der Steuerung konform ist, wird sie nicht bereitgestellt. Weitere Informationen finden Sie im Referenzhandbuch zu Kontrollen in der AWS Control Tower Dokumentation und unter Proaktive Kontrollen unter Implementierung von Sicherheitskontrollen am AWS.

Produktlebenszyklusmanagement (PLM)

Das Management von Daten und Prozessen für ein Produkt während seines gesamten Lebenszyklus, vom Design, der Entwicklung und Markteinführung über Wachstum und Reife bis hin zur Markteinführung und Markteinführung.

Produktionsumgebung

Siehe Umgebung.

Speicherprogrammierbare Steuerung (SPS)

In der Fertigung ein äußerst zuverlässiger, anpassungsfähiger Computer, der Maschinen überwacht und Fertigungsprozesse automatisiert.

schnelle Verkettung

Verwendung der Ausgabe einer LLM-Eingabeaufforderung als Eingabe für die nächste Aufforderung, um bessere Antworten zu generieren. Diese Technik wird verwendet, um eine komplexe Aufgabe in Unteraufgaben zu unterteilen oder um eine vorläufige Antwort iterativ zu verfeinern oder zu erweitern. Sie trägt dazu bei, die Genauigkeit und Relevanz der Antworten eines Modells zu verbessern und ermöglicht detailliertere, personalisierte Ergebnisse.

Pseudonymisierung

Der Prozess, bei dem persönliche Identifikatoren in einem Datensatz durch Platzhalterwerte ersetzt werden. Pseudonymisierung kann zum Schutz der Privatsphäre beitragen. Pseudonymisierte Daten gelten weiterhin als personenbezogene Daten.

publish/subscribe (pub/sub)

Ein Muster, das asynchrone Kommunikation zwischen Microservices ermöglicht, um die Skalierbarkeit und Reaktionsfähigkeit zu verbessern. In einem auf Microservices basierenden MES kann ein Microservice beispielsweise Ereignismeldungen in einem Kanal veröffentlichen, den andere Microservices abonnieren können. Das System kann neue Microservices hinzufügen, ohne den Veröffentlichungsservice zu ändern.

Q

Abfrageplan

Eine Reihe von Schritten, wie Anweisungen, die für den Zugriff auf die Daten in einem relationalen SQL-Datenbanksystem verwendet werden.

Abfrageplanregression

Wenn ein Datenbankserviceoptimierer einen weniger optimalen Plan wählt als vor einer bestimmten Änderung der Datenbankumgebung. Dies kann durch Änderungen an Statistiken, Beschränkungen, Umgebungseinstellungen, Abfrageparameter-Bindungen und Aktualisierungen der Datenbank-Engine verursacht werden.

R

RACI-Matrix

Siehe verantwortlich, rechenschaftspflichtig, konsultiert, informiert (RACI).

LAPPEN

Siehe Erweiterte Generierung beim Abrufen.

Ransomware

Eine bösartige Software, die entwickelt wurde, um den Zugriff auf ein Computersystem oder Daten zu blockieren, bis eine Zahlung erfolgt ist.

RASCI-Matrix

Siehe verantwortlich, rechenschaftspflichtig, konsultiert, informiert (RACI).

RCAC

Siehe Zugriffskontrolle für Zeilen und Spalten.

Read Replica

Eine Kopie einer Datenbank, die nur für Lesezwecke verwendet wird. Sie können Abfragen an das Lesereplikat weiterleiten, um die Belastung auf Ihrer Primärdatenbank zu reduzieren.

neu strukturieren

Siehe 7 Rs.

Recovery Point Objective (RPO)

Die maximal zulässige Zeitspanne seit dem letzten Datenwiederherstellungspunkt. Damit wird festgelegt, was als akzeptabler Datenverlust zwischen dem letzten Wiederherstellungspunkt und der Serviceunterbrechung gilt.

Wiederherstellungszeitziel (RTO)

Die maximal zulässige Verzögerung zwischen der Betriebsunterbrechung und der Wiederherstellung des Dienstes.

Refaktorierung

Siehe 7 Rs.

Region

Eine Sammlung von AWS Ressourcen in einem geografischen Gebiet. Jeder AWS-Region ist isoliert und unabhängig von den anderen, um Fehlertoleranz, Stabilität und Belastbarkeit zu gewährleisten. Weitere Informationen finden Sie unter Geben Sie an, was AWS-Regionen Ihr Konto verwenden kann.

Regression

Eine ML-Technik, die einen numerischen Wert vorhersagt. Zum Beispiel, um das Problem „Zu welchem Preis wird dieses Haus verkauft werden?“ zu lösen Ein ML-Modell könnte ein lineares Regressionsmodell verwenden, um den Verkaufspreis eines Hauses auf der Grundlage bekannter Fakten über das Haus (z. B. die Quadratmeterzahl) vorherzusagen.

rehosten

Siehe 7 Rs.

Veröffentlichung

In einem Bereitstellungsprozess der Akt der Förderung von Änderungen an einer Produktionsumgebung.

umziehen

Siehe 7 Rs.

neue Plattform

Siehe 7 Rs.

Rückkauf

Siehe 7 Rs.

Ausfallsicherheit

Die Fähigkeit einer Anwendung, Störungen zu widerstehen oder sich von ihnen zu erholen. Hochverfügbarkeit und Notfallwiederherstellung sind häufig Überlegungen bei der Planung der Ausfallsicherheit in der. AWS Cloud Weitere Informationen finden Sie unter AWS Cloud Resilienz.

Ressourcenbasierte Richtlinie

Eine mit einer Ressource verknüpfte Richtlinie, z. B. ein Amazon-S3-Bucket, ein Endpunkt oder ein Verschlüsselungsschlüssel. Diese Art von Richtlinie legt fest, welchen Prinzipalen der Zugriff gewährt wird, welche Aktionen unterstützt werden und welche anderen Bedingungen erfüllt sein müssen.

RACI-Matrix (verantwortlich, rechenschaftspflichtig, konsultiert, informiert)

Eine Matrix, die die Rollen und Verantwortlichkeiten aller an Migrationsaktivitäten und Cloud-Operationen beteiligten Parteien definiert. Der Matrixname leitet sich von den in der Matrix definierten Zuständigkeitstypen ab: verantwortlich (R), rechenschaftspflichtig (A), konsultiert (C) und informiert (I). Der Unterstützungstyp (S) ist optional. Wenn Sie Unterstützung einbeziehen, wird die Matrix als RASCI-Matrix bezeichnet, und wenn Sie sie ausschließen, wird sie als RACI-Matrix bezeichnet.

Reaktive Kontrolle

Eine Sicherheitskontrolle, die darauf ausgelegt ist, die Behebung unerwünschter Ereignisse oder Abweichungen von Ihren Sicherheitsstandards voranzutreiben. Weitere Informationen finden Sie unter Reaktive Kontrolle in Implementieren von Sicherheitskontrollen in AWS.

Beibehaltung

Siehe 7 Rs.

zurückziehen

Siehe 7 Rs.

Retrieval Augmented Generation (RAG)

Eine generative KI-Technologie, bei der ein LLM auf eine maßgebliche Datenquelle verweist, die sich außerhalb seiner Trainingsdatenquellen befindet, bevor eine Antwort generiert wird. Ein RAG-Modell könnte beispielsweise eine semantische Suche in der Wissensdatenbank oder in benutzerdefinierten Daten einer Organisation durchführen. Weitere Informationen finden Sie unter Was ist RAG.

Drehung

Der Vorgang, bei dem ein Geheimnis regelmäßig aktualisiert wird, um es einem Angreifer zu erschweren, auf die Anmeldeinformationen zuzugreifen.

Zugriffskontrolle für Zeilen und Spalten (RCAC)

Die Verwendung einfacher, flexibler SQL-Ausdrücke mit definierten Zugriffsregeln. RCAC besteht aus Zeilenberechtigungen und Spaltenmasken.

RPO

Siehe Recovery Point Objective.

RTO

Siehe Ziel der Wiederherstellungszeit.

Runbook

Eine Reihe manueller oder automatisierter Verfahren, die zur Ausführung einer bestimmten Aufgabe erforderlich sind. Diese sind in der Regel darauf ausgelegt, sich wiederholende Operationen oder Verfahren mit hohen Fehlerquoten zu rationalisieren.

S

SAML 2.0

Ein offener Standard, den viele Identitätsanbieter (IdPs) verwenden. Diese Funktion ermöglicht föderiertes Single Sign-On (SSO), sodass sich Benutzer bei den API-Vorgängen anmelden AWS Management Console oder die AWS API-Operationen aufrufen können, ohne dass Sie einen Benutzer in IAM für alle in Ihrer Organisation erstellen müssen. Weitere Informationen zum SAML-2.0.-basierten Verbund finden Sie unter Über den SAML-2.0-basierten Verbund in der IAM-Dokumentation.

SCADA

Siehe Aufsichtskontrolle und Datenerfassung.

SCP

Siehe Richtlinie zur Dienstkontrolle.

Secret

Interne AWS Secrets Manager, vertrauliche oder eingeschränkte Informationen, wie z. B. ein Passwort oder Benutzeranmeldedaten, die Sie in verschlüsselter Form speichern. Es besteht aus dem geheimen Wert und seinen Metadaten. Der geheime Wert kann binär, eine einzelne Zeichenfolge oder mehrere Zeichenketten sein. Weitere Informationen finden Sie unter Was ist in einem Secrets Manager Manager-Geheimnis? in der Secrets Manager Manager-Dokumentation.

Sicherheit durch Design

Ein systemtechnischer Ansatz, der die Sicherheit während des gesamten Entwicklungsprozesses berücksichtigt.

Sicherheitskontrolle

Ein technischer oder administrativer Integritätsschutz, der die Fähigkeit eines Bedrohungsakteurs, eine Schwachstelle auszunutzen, verhindert, erkennt oder einschränkt. Es gibt vier Haupttypen von Sicherheitskontrollen: präventiv, detektiv, reaktionsschnell und proaktiv.

Härtung der Sicherheit

Der Prozess, bei dem die Angriffsfläche reduziert wird, um sie widerstandsfähiger gegen Angriffe zu machen. Dies kann Aktionen wie das Entfernen von Ressourcen, die nicht mehr benötigt werden, die Implementierung der bewährten Sicherheitsmethode der Gewährung geringster Berechtigungen oder die Deaktivierung unnötiger Feature in Konfigurationsdateien umfassen.

System zur Verwaltung von Sicherheitsinformationen und Ereignissen (security information and event management – SIEM)

Tools und Services, die Systeme für das Sicherheitsinformationsmanagement (SIM) und das Management von Sicherheitsereignissen (SEM) kombinieren. Ein SIEM-System sammelt, überwacht und analysiert Daten von Servern, Netzwerken, Geräten und anderen Quellen, um Bedrohungen und Sicherheitsverletzungen zu erkennen und Warnmeldungen zu generieren.

Automatisierung von Sicherheitsreaktionen

Eine vordefinierte und programmierte Aktion, die darauf ausgelegt ist, automatisch auf ein Sicherheitsereignis zu reagieren oder es zu beheben. Diese Automatisierungen dienen als detektive oder reaktionsschnelle Sicherheitskontrollen, die Sie bei der Implementierung bewährter AWS Sicherheitsmethoden unterstützen. Beispiele für automatisierte Antwortaktionen sind das Ändern einer VPC-Sicherheitsgruppe, das Patchen einer EC2 Amazon-Instance oder das Rotieren von Anmeldeinformationen.

Serverseitige Verschlüsselung

Verschlüsselung von Daten am Zielort durch denjenigen AWS-Service , der sie empfängt.

Service-Kontrollrichtlinie (SCP)

Eine Richtlinie, die eine zentrale Steuerung der Berechtigungen für alle Konten in einer Organisation in ermöglicht AWS Organizations. SCPs Definieren Sie Leitplanken oder legen Sie Grenzwerte für Aktionen fest, die ein Administrator an Benutzer oder Rollen delegieren kann. Sie können sie SCPs als Zulassungs- oder Ablehnungslisten verwenden, um festzulegen, welche Dienste oder Aktionen zulässig oder verboten sind. Weitere Informationen finden Sie in der AWS Organizations Dokumentation unter Richtlinien zur Dienststeuerung.

Service-Endpunkt

Die URL des Einstiegspunkts für einen AWS-Service. Sie können den Endpunkt verwenden, um programmgesteuert eine Verbindung zum Zielservice herzustellen. Weitere Informationen finden Sie unter AWS-Service -Endpunkte in der Allgemeine AWS-Referenz.

Service Level Agreement (SLA)

Eine Vereinbarung, in der klargestellt wird, was ein IT-Team seinen Kunden zu bieten verspricht, z. B. in Bezug auf Verfügbarkeit und Leistung der Services.

Service-Level-Indikator (SLI)

Eine Messung eines Leistungsaspekts eines Dienstes, z. B. seiner Fehlerrate, Verfügbarkeit oder Durchsatz.

Service-Level-Ziel (SLO)

Eine Zielkennzahl, die den Zustand eines Dienstes darstellt, gemessen anhand eines Service-Level-Indikators.

Modell der geteilten Verantwortung

Ein Modell, das die Verantwortung beschreibt, mit der Sie gemeinsam AWS für Cloud-Sicherheit und Compliance verantwortlich sind. AWS ist für die Sicherheit der Cloud verantwortlich, während Sie für die Sicherheit in der Cloud verantwortlich sind. Weitere Informationen finden Sie unter Modell der geteilten Verantwortung.

SIEM

Siehe Sicherheitsinformations- und Event-Management-System.

Single Point of Failure (SPOF)

Ein Fehler in einer einzelnen, kritischen Komponente einer Anwendung, der das System stören kann.

SLA

Siehe Service Level Agreement.

SLI

Siehe Service-Level-Indikator.

ALSO

Siehe Service-Level-Ziel.

split-and-seed Modell

Ein Muster für die Skalierung und Beschleunigung von Modernisierungsprojekten. Sobald neue Features und Produktversionen definiert werden, teilt sich das Kernteam auf, um neue Produktteams zu bilden. Dies trägt zur Skalierung der Fähigkeiten und Services Ihrer Organisation bei, verbessert die Produktivität der Entwickler und unterstützt schnelle Innovationen. Weitere Informationen finden Sie unter Schrittweiser Ansatz zur Modernisierung von Anwendungen in der AWS Cloud.

SPOTTEN

Siehe Single Point of Failure.

Sternschema

Eine Datenbank-Organisationsstruktur, die eine große Faktentabelle zum Speichern von Transaktions- oder Messdaten und eine oder mehrere kleinere dimensionale Tabellen zum Speichern von Datenattributen verwendet. Diese Struktur ist für die Verwendung in einem Data Warehouse oder für Business Intelligence-Zwecke konzipiert.

Strangler-Fig-Muster

Ein Ansatz zur Modernisierung monolithischer Systeme, bei dem die Systemfunktionen schrittweise umgeschrieben und ersetzt werden, bis das Legacy-System außer Betrieb genommen werden kann. Dieses Muster verwendet die Analogie einer Feigenrebe, die zu einem etablierten Baum heranwächst und schließlich ihren Wirt überwindet und ersetzt. Das Muster wurde eingeführt von Martin Fowler als Möglichkeit, Risiken beim Umschreiben monolithischer Systeme zu managen. Ein Beispiel für die Anwendung dieses Musters finden Sie unter Schrittweises Modernisieren älterer Microsoft ASP.NET (ASMX)-Webservices mithilfe von Containern und Amazon API Gateway.

Subnetz

Ein Bereich von IP-Adressen in Ihrer VPC. Ein Subnetz muss sich in einer einzigen Availability Zone befinden.

Aufsichtskontrolle und Datenerfassung (SCADA)

In der Fertigung ein System, das Hardware und Software zur Überwachung von Sachanlagen und Produktionsabläufen verwendet.

Symmetrische Verschlüsselung

Ein Verschlüsselungsalgorithmus, der denselben Schlüssel zum Verschlüsseln und Entschlüsseln der Daten verwendet.

synthetisches Testen

Testen eines Systems auf eine Weise, die Benutzerinteraktionen simuliert, um potenzielle Probleme zu erkennen oder die Leistung zu überwachen. Sie können Amazon CloudWatch Synthetics verwenden, um diese Tests zu erstellen.

Systemaufforderung

Eine Technik, mit der einem LLM Kontext, Anweisungen oder Richtlinien zur Verfügung gestellt werden, um sein Verhalten zu steuern. Systemaufforderungen helfen dabei, den Kontext festzulegen und Regeln für Interaktionen mit Benutzern festzulegen.

T

tags

Schlüssel-Wert-Paare, die als Metadaten für die Organisation Ihrer Ressourcen dienen. AWS Mit Tags können Sie Ressourcen verwalten, identifizieren, organisieren, suchen und filtern. Weitere Informationen finden Sie unter Markieren Ihrer AWS -Ressourcen.

Zielvariable

Der Wert, den Sie in überwachtem ML vorhersagen möchten. Dies wird auch als Ergebnisvariable bezeichnet. In einer Fertigungsumgebung könnte die Zielvariable beispielsweise ein Produktfehler sein.

Aufgabenliste

Ein Tool, das verwendet wird, um den Fortschritt anhand eines Runbooks zu verfolgen. Eine Aufgabenliste enthält eine Übersicht über das Runbook und eine Liste mit allgemeinen Aufgaben, die erledigt werden müssen. Für jede allgemeine Aufgabe werden der geschätzte Zeitaufwand, der Eigentümer und der Fortschritt angegeben.

Testumgebungen

Siehe Umgebung.

Training

Daten für Ihr ML-Modell bereitstellen, aus denen es lernen kann. Die Trainingsdaten müssen die richtige Antwort enthalten. Der Lernalgorithmus findet Muster in den Trainingsdaten, die die Attribute der Input-Daten dem Ziel (die Antwort, die Sie voraussagen möchten) zuordnen. Es gibt ein ML-Modell aus, das diese Muster erfasst. Sie können dann das ML-Modell verwenden, um Voraussagen für neue Daten zu erhalten, bei denen Sie das Ziel nicht kennen.

Transit-Gateway

Ein Netzwerk-Transit-Hub, über den Sie Ihre Netzwerke VPCs und Ihre lokalen Netzwerke miteinander verbinden können. Weitere Informationen finden Sie in der Dokumentation unter Was ist ein Transit-Gateway. AWS Transit Gateway

Stammbasierter Workflow

Ein Ansatz, bei dem Entwickler Feature lokal in einem Feature-Zweig erstellen und testen und diese Änderungen dann im Hauptzweig zusammenführen. Der Hauptzweig wird dann sequentiell für die Entwicklungs-, Vorproduktions- und Produktionsumgebungen erstellt.

Vertrauenswürdiger Zugriff

Gewährung von Berechtigungen für einen Dienst, den Sie angeben, um Aufgaben in Ihrer Organisation AWS Organizations und in deren Konten in Ihrem Namen auszuführen. Der vertrauenswürdige Service erstellt in jedem Konto eine mit dem Service verknüpfte Rolle, wenn diese Rolle benötigt wird, um Verwaltungsaufgaben für Sie auszuführen. Weitere Informationen finden Sie in der AWS Organizations Dokumentation unter Verwendung AWS Organizations mit anderen AWS Diensten.

Optimieren

Aspekte Ihres Trainingsprozesses ändern, um die Genauigkeit des ML-Modells zu verbessern. Sie können das ML-Modell z. B. trainieren, indem Sie einen Beschriftungssatz generieren, Beschriftungen hinzufügen und diese Schritte dann mehrmals unter verschiedenen Einstellungen wiederholen, um das Modell zu optimieren.

Zwei-Pizzen-Team

Ein kleines DevOps Team, das Sie mit zwei Pizzen ernähren können. Eine Teamgröße von zwei Pizzen gewährleistet die bestmögliche Gelegenheit zur Zusammenarbeit bei der Softwareentwicklung.

U

Unsicherheit

Ein Konzept, das sich auf ungenaue, unvollständige oder unbekannte Informationen bezieht, die die Zuverlässigkeit von prädiktiven ML-Modellen untergraben können. Es gibt zwei Arten von Unsicherheit: Epistemische Unsicherheit wird durch begrenzte, unvollständige Daten verursacht, wohingegen aleatorische Unsicherheit durch Rauschen und Randomisierung verursacht wird, die in den Daten liegt. Weitere Informationen finden Sie im Leitfaden Quantifizieren der Unsicherheit in Deep-Learning-Systemen.

undifferenzierte Aufgaben

Diese Arbeit wird auch als Schwerstarbeit bezeichnet. Dabei handelt es sich um Arbeiten, die zwar für die Erstellung und den Betrieb einer Anwendung erforderlich sind, aber dem Endbenutzer keinen direkten Mehrwert bieten oder keinen Wettbewerbsvorteil bieten. Beispiele für undifferenzierte Aufgaben sind Beschaffung, Wartung und Kapazitätsplanung.

höhere Umgebungen

Siehe Umgebung.

V

Vacuuming

Ein Vorgang zur Datenbankwartung, bei dem die Datenbank nach inkrementellen Aktualisierungen bereinigt wird, um Speicherplatz zurückzugewinnen und die Leistung zu verbessern.

Versionskontrolle

Prozesse und Tools zur Nachverfolgung von Änderungen, z. B. Änderungen am Quellcode in einem Repository.

VPC-Peering

Eine Verbindung zwischen zwei VPCs , die es Ihnen ermöglicht, den Verkehr mithilfe privater IP-Adressen weiterzuleiten. Weitere Informationen finden Sie unter Was ist VPC-Peering? in der Amazon-VPC-Dokumentation.

Schwachstelle

Ein Software- oder Hardwarefehler, der die Sicherheit des Systems beeinträchtigt.

W

Warmer Cache

Ein Puffer-Cache, der aktuelle, relevante Daten enthält, auf die häufig zugegriffen wird. Die Datenbank-Instance kann aus dem Puffer-Cache lesen, was schneller ist als das Lesen aus dem Hauptspeicher oder von der Festplatte.

warme Daten

Daten, auf die selten zugegriffen wird. Bei der Abfrage dieser Art von Daten sind mäßig langsame Abfragen in der Regel akzeptabel.

Fensterfunktion

Eine SQL-Funktion, die eine Berechnung für eine Gruppe von Zeilen durchführt, die sich in irgendeiner Weise auf den aktuellen Datensatz beziehen. Fensterfunktionen sind nützlich für die Verarbeitung von Aufgaben wie die Berechnung eines gleitenden Durchschnitts oder für den Zugriff auf den Wert von Zeilen auf der Grundlage der relativen Position der aktuellen Zeile.

Workload

Ein Workload ist eine Sammlung von Ressourcen und Code, die einen Unternehmenswert bietet, wie z. B. eine kundenorientierte Anwendung oder ein Backend-Prozess.

Workstream

Funktionsgruppen in einem Migrationsprojekt, die für eine bestimmte Reihe von Aufgaben verantwortlich sind. Jeder Workstream ist unabhängig, unterstützt aber die anderen Workstreams im Projekt. Der Portfolio-Workstream ist beispielsweise für die Priorisierung von Anwendungen, die Wellenplanung und die Erfassung von Migrationsmetadaten verantwortlich. Der Portfolio-Workstream liefert diese Komponenten an den Migrations-Workstream, der dann die Server und Anwendungen migriert.

WURM

Sehen Sie einmal schreiben, viele lesen.

WQF

Siehe AWS Workload-Qualifizierungsrahmen.

einmal schreiben, viele lesen (WORM)

Ein Speichermodell, das Daten ein einziges Mal schreibt und verhindert, dass die Daten gelöscht oder geändert werden. Autorisierte Benutzer können die Daten so oft wie nötig lesen, aber sie können sie nicht ändern. Diese Datenspeicherinfrastruktur gilt als unveränderlich.

Z

Zero-Day-Exploit

Ein Angriff, in der Regel Malware, der eine Zero-Day-Sicherheitslücke ausnutzt.

Zero-Day-Sicherheitslücke

Ein unfehlbarer Fehler oder eine Sicherheitslücke in einem Produktionssystem. Bedrohungsakteure können diese Art von Sicherheitslücke nutzen, um das System anzugreifen. Entwickler werden aufgrund des Angriffs häufig auf die Sicherheitsanfälligkeit aufmerksam.

Zero-Shot-Aufforderung

Bereitstellung von Anweisungen für die Ausführung einer Aufgabe an einen LLM, jedoch ohne Beispiele (Schnappschüsse), die ihm als Orientierungshilfe dienen könnten. Der LLM muss sein vortrainiertes Wissen einsetzen, um die Aufgabe zu bewältigen. Die Effektivität von Zero-Shot Prompting hängt von der Komplexität der Aufgabe und der Qualität der Aufforderung ab. Siehe auch Few-Shot-Prompting.

Zombie-Anwendung

Eine Anwendung, deren durchschnittliche CPU- und Arbeitsspeichernutzung unter 5 Prozent liegt. In einem Migrationsprojekt ist es üblich, diese Anwendungen außer Betrieb zu nehmen.

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.